Explanatory Model Analysis

Explore, Explain and Examine Predictive Models

Przemyslaw Biecek and Tomasz Burzykowski

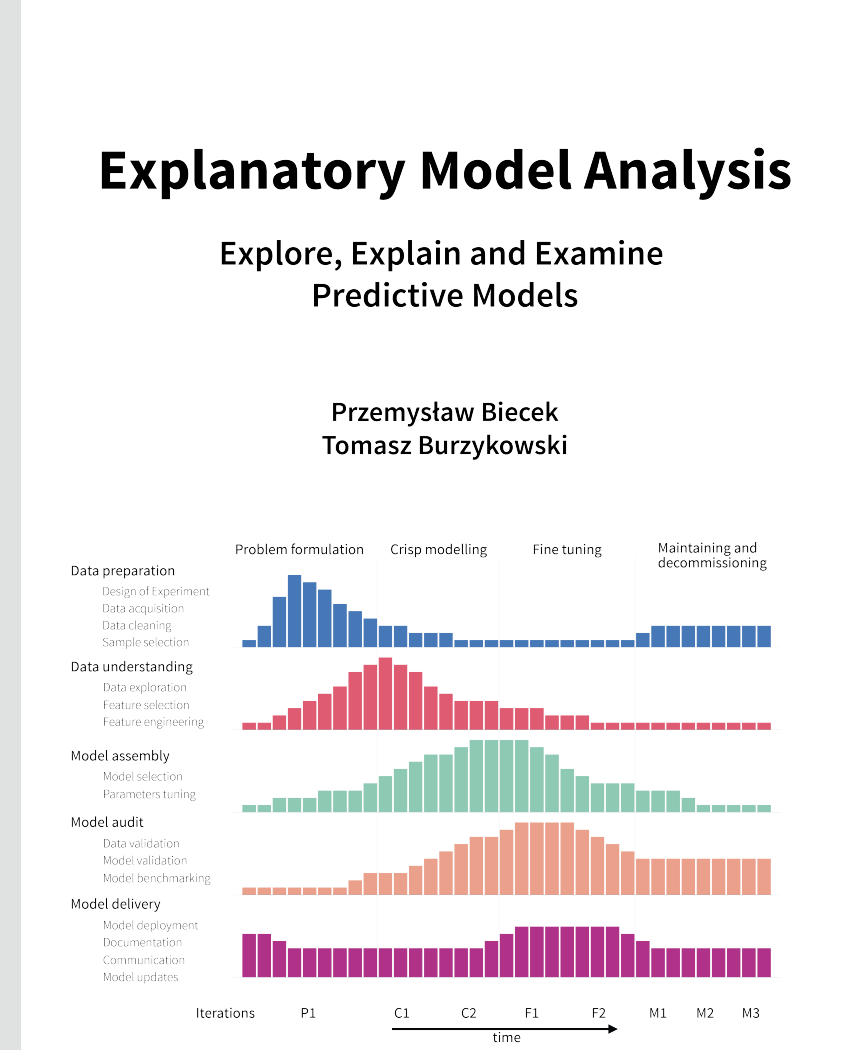
2020-01-27

Contents

List of Tables

List of Figures

Preface


Explanatory Model Analysis
Explore, Explain and Examine
Predictive Models
Przemysław Biecek
Tomasz Burzykowski
Problem formulation
Crisp modelling
Fine tuning
Maintaining and decommissioning
Data preparation
Design of Experiment
Data acquisition
Data cleaning
Sample selection
Data understanding
Data exploration
Feature selection
Feature engineering
Model assembly
Model selection
Parameters tuning
Model audit
Data validation
Model validation
Model benchmarking
Model delivery
Model deployment
Documentation
Communication
Model updates
Iterations
P1
C1
C2
F1
F2
M1
M2
M3
time

0.1 Introduction

0.1.1 Notes to readers

A note to readers: this text is a work in progress.

We've released this initial version to get more feedback. Feedback can be given at the GitHub repo https://github.com/pbiecek/ema/issues. We are primarily interested in the organization and consistency of the content, but any comments will be welcommed.

We'd like to thank everyone that contributed feedback, found typos, or ignited discussions while the book was being written, including GitHub contributors: agosiewska, Rees Morrison, kasiapekala, hbaniecki, AsiaHenzel, kozaka93.

0.1.2 The aim of the book

Predictive models are used to guess (statisticians would say: predict) values of a variable of interest based on other variables. As an example, consider prediction of sales based on historical data, prediction of risk of heart disease based on patient characteristics, or prediction of political attitudes based on Facebook comments.

Predictive models have been constructed through the enitre human history. Ancient Egyptians, for instance, used observations of the rising of Sirius to predict flooding of the Nile. A more rigorous approach to model construction may be attributed to the method of least squares, published more than two centuries ago by Legendre in 1805 and by Gauss in 1809. With time, the number of applications in economy, medicine, biology, and agriculture has grown. The term *regression* was coined by Francis Galton in 1886. Initially, it was referring to biological applications, while today it is used for various models that allow prediction of continuous variables. Prediction of nominal variables is called *classification*, and its beginning may be attributed to works of Ronald Fisher in 1936.

During the last century, many statistical models that can be used for predictive purposes have been developed. These include linear models, generalized linear models, regression and classification trees, rule-based models, and many others. Developments in mathematical foundations of predictive models were boosted by increasing computational power of personal computers and availability of large datasets in the era of „big data" that we have entered.

With the increasing demand for predictive models, model features such as flexibility, ability to perform internally variable selection (feature engineering), and high precision of predictions are of interest. To obtain robust models, ensembles of models are used. Techniques like bagging, boosting, or model stacking combine hundreds or thousands of small models into a one super-model. Large deep neural models have over a billion parameters.

There is a cost of this progress. Complex models may seem to operate like „black boxes'". It may be difficult, or even impossible, to understand how thousands of coefficients affect the model prediction. At the same time, complex models may not work as well as we would like them to. An overview of real problems with massive-scale black-box models may be found in an excellent book of Cathy O'Neil (O'Neil, 2016) or in her TED Talk „*The era of blind faith in big data must end*". There is a growing number of examples of predictive models with performance that deteriorated over time or became biased in some sense. For instance, IBM's Watson for Oncology was criticized by oncologists for delivering unsafe and inaccurate recommendations (Ross and Swetliz, 2018). Amazon's system for CV screening was found to be biased against women (Dastin, 2018). The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm for predicting recidivism, developed by Northpointe (now Equivant), is accused to be biased against blacks (Larson et al., 2016). Algorithms beyond Apple Credit Card are accused to be gender-biased (Duffy, 2019). Some tools for sentiment analysis are susspected to be age-biased (Diaz et al., 2018). These are examples of models and algorithms that led to serious violations of fairness and ethical principles. An example of situation when data drift led to deterioration in model

performance is the Google Flu model, which gave worse predictions after two years than at baseline (Salzberg, 2014), (Lazer et al., 2014).

A reaction to some of these examples and problems are new regulations, like the General Data Protection Regulation (GDPR, 2018). Also, new civic rights are being formulated (Goodman and Flaxman, 2016), (Casey et al., 2018), (Ruiz, 2018). A noteworthy example is the *„Right to Explanation"*, i.e., the right to be provided an explanation for an output of an automated algorithm (Goodman and Flaxman, 2016). To exercise the right, we need new methods for verification, exploration, and explanation of predictive models.

Figure 1 shows how the increase in the model complexity affects the relative importance of domain understanding vs. modeling vs. validation. Simplest models are usually built on top of a good understanding of the domain. Domain knowledge helps to create and select most important variables that can be transformed into predictive scores. Machine learning exploits the tradeoff between availability of data and domain knowledge. Flexible models can use massive data to learn good features and filter out bad ones. The effort is shifted from a deep understanding of the domain towards computationally heavy training of models. The validation part is of an increased importance because it creates a feedback loop with the modeling. Results from model validation lead to next decisions related to model training. This is diffenrent than in case of statistical hypothesis testing. Statistical hypotheses shall be stated in advance of data analysis and obtained p-values shall not interfere in the way how data or models were prepared.

What will be next? The increasing automation in the EDA (Exploratory Data Analysis) and modeling part of the process shift the focus towards the validation of models. The purpose of validation is not only to measure how good is the model but also what other risks are associated with models. Risks like concept drift, gender, age or race bias. This book is about new methods that can be used for validation and justification.

Out of this we can conclude that, today, the true bottleneck in

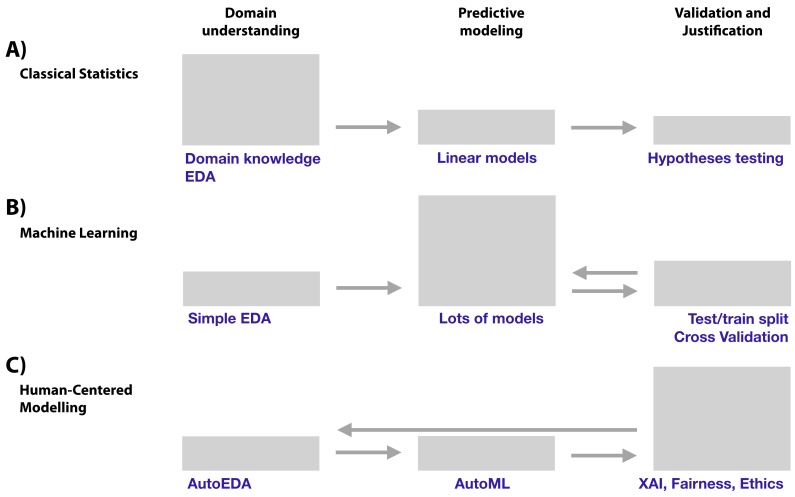


FIGURE 1 Shift in the relative importance and effort put in different phases of the data-driven modeling. (A) Statistical modeling is often based on deep understanding of the domain. Manual data exploration, consultations with domain experts, variable transformations lead to good models. Structures of models are often based on (generalized) linear models. Model verification is done through hypothesis testing. (B) Machine learning modeling is often based on elastic models fitted to large volumes of data. Domain exploration is often shallow while the focus is based on predictive performance. Lots of attention is put in cross validation and other strategies that deal with overfitting. (C) What will be next? Human-centered modeling? Better tools for auto EDA and auto ML will shift focus into the part related with validation against the domain knowledge like fairness, bias or new techniques for data exploration. Arrows show feedback loops in the modeling process. The feedback loop is even larger now, as the results from model validation are helping also in the domain understanding.

predictive modelling is not the lack of data, nor the lack of computational power, nor inadequate algorithms, nor the lack of flexible models. It is the lack of tools for model validation, model exploration, and explanation of model decisions. Thus, in this book, we present a collection of methods that may be used for this purpose. As development of such methods is a very active area of research and new methods become available almost on a continuous basis, we do not aim at being exhaustive. Rather, we present the mindset, key problems, and several examples of methods that can be used in model exploration.

0.1.3 A bit of philosophy: three laws of model explanation

Seventy-six years ago, Isaac Asimov forumlated Three Laws of Robotics:

1) a robot may not injure a human being,
2) a robot must obey the orders given it by human beings, and
3) a robot must protect its own existence.

Today's robots, like cleaning robots, robotic pets, or autonomous cars are far from being conscious enough to fall under Asimov's ethics. However, we are more and more surrounded by complex predictive models and algorithms used for decision making. Artificial Intelligence models are used in health care, politics, education, justice, and many other areas. The models and algorithms have a far larger influence on our lives than physical robots. Yet, applications of such models are left unregulated despite examples of their potential harmfulness. See *Weapons of Math Destruction* by Cathy O'Neil (O'Neil, 2016) for an excellent overview of selected problems.

It's clear that we need to control the models and algorithms that may affect us. Thus, Asimov's laws are referred to in the context of the discussion around Ethics of Artifical Intelligence. Initiatives to formulate principles for AI development have been undertaken,

for instance, in the UK [Olhede & Wolfe, Significance 2018, 15: 6-7]. Following Asimov's approach, we propose three requirements that any predictive model should fulfill:

- **Prediction's validation**. For every prediction of a model, one should be able to verify how strong is the evidence that confirms the prediction.
- **Prediction's justification**. For every prediction of a model, one should be able to understand which variables affect the prediction and to what extent.
- **Prediction's speculation**. For every prediction of a model, one should be able to understand how the model prediction would change if input variables changed.

We see two ways to comply with these requirements. One is to use only models that fulfill these conditions by design. There are so called interpretable by design models like linear models, rule based models or classification trees with small number of parameters (Molnar, 2019). However, the price for transparency may be a reduction in performance. Another way is to use tools that allow, perhaps by using approximations or simplifications, to „explain" predictions for any model. In our book, we will focus on the latter approach.

0.1.4 The structure of this book

This book is split in two major parts. In the part *Instance-level explainers*, we present techniques for exploration and explanation of model predictions for a single observation. On the other hand, in the part *Dataset-level explainers*, we present techniques for exploration and explanation of a model for an entire dataset.

Before embarking on the description of the methods, in Chapter 0.2, we provide a short introduction to the process of data exploration and model assembly along with notation and definition of key concepts that areused in consecutive chapters. In chapters 0.3 and 0.4, we provide a short description of R and python tools and packages that are necessary to replicate the results presented in

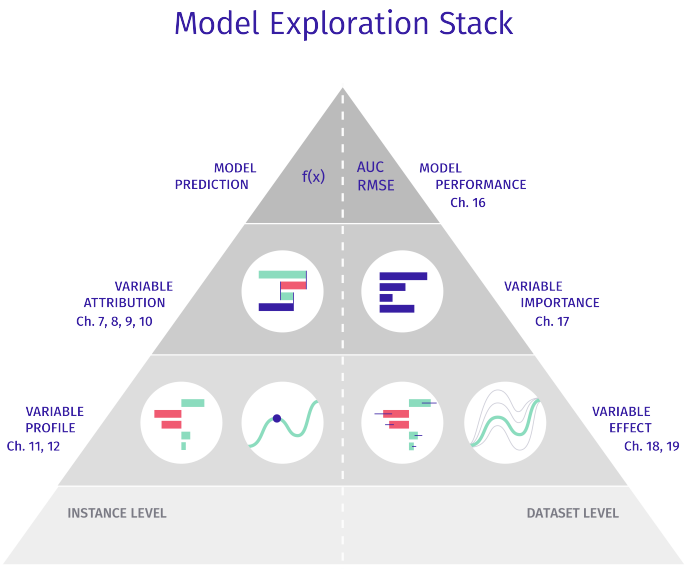


FIGURE 2 Stack with model exploration methods presented in this book. Left side is focused on instance-level explanation while the right side is focused on dataset-level explanation. Consecutive layers of the stack are linked with a deeper level of model exploration. These layers are linked with law's of model exploration introduced in Section efthree-single-laws

this book. In Chapter 0.5, we describe two datasets that are used throughout the book to illustrate the presented methods and tools.

Rest of the book is structured in Figure 2.

The **Instance-level** part of the book consists of Chapters 0.7-0.14. Chapters 0.7-0.9 present methods to decompose model predictions into variable contributions. In particular, Chapter 0.7 introduces Break-down (BD) plots for models with additive effects. On the other hand, Chapter 0.8 presents a method that allows for interactions. Finally, Chapter 0.9 describes SHAP (Lundberg and Lee, 2017) an alternative method for decomposing model predictions that is closely linked with Shapley values (Shapley, 1953) developed originally for cooperative games. Chapter 0.10 presents a

different approach to explanation of single-instance predictions. It is based on a local approximation of a black-box model by a simpler, glass-box one. In this chapter, we discuss the Local Interpretable Model-Agnostic Explanations (LIME) method (Ribeiro et al., 2016). These chapters corresponds to the second layer of the stack in Figure 2.

In Chapters 0.11-0.13 we present methods based on Ceteris-paribus (CP) profiles. The profiles show the change of model-based predictions induced by a change of a single variable. These profiles are introduced in Chapter 0.11 while Chapter 0.12 presents a CP-profile-based measure that summarizes the impact of a selected variable on model's predictions. This measure can be used to determine the order of variables in model exploration. It is particulary important for models with large numbers of explanatory variables. Chapter 0.13 is focused on model diagnostic. It describes local-fidelity plots that are useful to investigate the sources of a poor prediction for a particular single observation. The final chapter of the first part, Chapter 0.14 compares various instance-level explainers.

The **Dataset-level explainers** part of the book consists of Chapters 0.15-0.20. These chapters present methods in the same order as appered in the Model Exploration Stack in Figure 2. Chapter 0.16 shows selected measures for model benchmarking along with performance measures for classification and regression models. On top of these measures, the Chapter 0.17 presented an algorithm for assessment of importance of variables based on selected performance measure. This method is model agnostic and can be used for cross models comparisons. Next layer of the Model Exploration Stack is presented in Chapters 0.18 and 0.19. Here we introduce Partial Dependency and Accumulated Dependency methods for univariate exploration of variable effects. This part of the book is closed with the Chapter 0.20 that summarises diagnostic techniques for model residuals.

To make the exploration of the book easier, in each Chapter we introduce a single method and each chapter has the same structure:

- Section *Introduction* explains the goal of and the general idea behind the method.
- Section *Method* shows mathematical or computational details related to the method. This subsection can be skipped if you are not interested in the details.
- Section *Example* shows an exemplary application of the method with discussion of results.
- Section *Pros and cons* summarizes the advantages and disadvantages of the method. It also provides some guideance regarding when to use the method.
- Section *Code snippets* shows the implementation of the method in R and Python. This subsection can be skipped if you are not interested in the implementation.

0.1.5 Terminology

It is worth noting that, when it comes to predictive models, the same concepts have often been given different names in statistics and in machine learning. For instance, in the statistical-modelling literature, one refers to „explanatory variables," with „independent variables," „predictors," or „covariates" as often-used equivalents. Explanatory variables are used in the model as means to explain (predict) the „dependent variable," also called „predicted" variable or „response." In machine-learning terminology, „input variables" or „features" are used to predict the „output" or „target" variable. In statistical modelling, models are fit to the data that contain „observations'', whereas in the machine-learning world a dataset may contain „instances" or „cases''. When we talk about values that define a single instance of a model in statistical modelling we refer to model „coefficients" while in machine-learning it is more common to use phrase model „parameters''. In statistics it is common to say that model coefficients are „estimated" while in machine learning it is more common to say that parameters are „trained" or are obtained in the process of „model training".

To the extent possible, in our book we try to consistently use the statistical-modelling terminology. However, the reader may

find references to a „feature" here and there. Somewhat inconsistently, we also introduce the term „instance-level" explanation. Instance-level explanation methods are designed to extract information about the behavior of the model related to a specific observation (or instance). On the other hand, „dataset-level" explanation techniques allow obtaining information about the behavior of the model for an entire dataset.

We consider models for dependent variables that can be continuous or nominal/categorical. The values of a continuous variable can be represented by numbers with an ordering that makes some sense (zip codes or phone numbers are not considered as continuous variables while age, number of children are). A continuous variable does not have to be continuous in the mathematical sense; counts (number of floors, steps, etc.) will be treated as continuous variables as well. A nominal/categorical variable can assume only a finite set of values that are not numbers in the mathematical sense, i.e. it makes no sense to substract or divide these values.

In this book we focus on „black-box" approach. We discuss this approach in a bit more detail in the next section.

0.1.6 Glass-box models vs. black-box models

Black-box models are models with a complex structure that is hard to understand by humans. Usually this refers to a large number of model coefficients or complex mathematical transformations. As people vary in their capacity to understand complex models, there is no strict threshold for the number of coefficients that makes a model a black-box. In practice, for most people this threshold is probably closer to 10 than to 100.

A „glass-box" (sometimes called white-box or transparent-box) model, which is opposite to a „black-box" one, is a model that is easy to understand (though maybe not by every person). It has a simple structure and a limited number of coefficients.

The most common classess of glass-box models are decision or regression trees, as an example in Figure 3, rules, or models with

an explicit compact structure, like the following model for obesity based on the BMI index.

$$BMI = \frac{mass_{kg}}{height_{m^2}}.$$

In the model, two explanatory variables are used, mass in kilograms and height in meters. Based on them a BMI index is derived that commonly used for classification into *Underweight* (BMI $<$ 18), *Normal* (18 $<$ BMI $<$ 25) or *Overweight* (BMI $>$ 25) categories. Having the model in a compact form it is easy to understand how changes in one variable affect the model output.

The structure of a glass-box model is, in general, easy to understand. It may be difficult to collect the necessary data, build the model, fit it to the data, or perform model validation, but once the model has been developed its interpretation and mode of working is straightforward.

Why is it important to understand the model structure? There are several important advantages. If the model structure is clear, we can easily see which variables are included in the model and which are not. Hence, for instance, we may be able to, question the model when a particular explanatory variable was excluded from it. Also, in the case of a model with a clear structure and a limited number of coefficients, we can easily link changes in model predictions with changes in particular explanatory variables. This, in turn, may allow us to challenge the model against domain knowledge if, for instance, the effect of a particular variable on predictions is inconsistent with previously established results. Note that linking changes in model predictions with changes in particular explanatory variables may be difficult when there are many variables and/or coefficients in the model. For instance, a classification tree with hundreds of nodes is difficult to understand, as is a linear regression model with hundreds of cofficients.

Note that some glass-box models, like the decision tree model presented in Figure 3 by design satisfies explainability laws introduced in Section 0.1.3. For *Prediction's validation* we see in each node

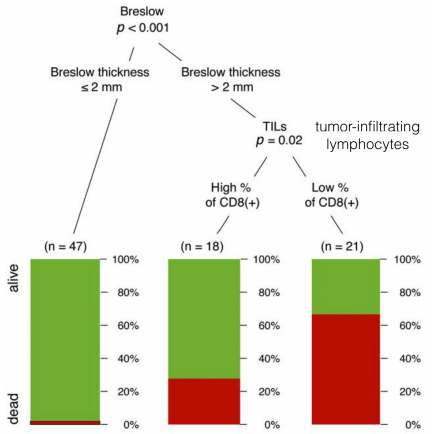


FIGURE 3 Example classification tree model for melanoma risk patients based on [@BILLCD8]. The model is based on two explanatory variables, Breslow thickness and Tumor infiltration lymphocytes. These two variables lead to three groups of paritents with different odds of survival.

how many patients fall in a given category. For *Prediction's justification* we see which variables are used in every decision path. For *Prediction's speculation* we can trace how changes in particular variables will affect the model prediction. We can, of course, argue if the model is good or not, but obviously the model structure is transparent.

Comprehending the performance of a black-box models presents more challenges. The structure of a complex model, such as a neural-network model, may be far from transparent. Consequently, we may not understand which features influence the model decisions and by how much. Consequently, it may be difficult to decide whether the model is consistent with our domain knowledge. In our book we present tools that can help in extracting the information necessary for the evaluation of complex models.

0.1.7 Model-agnostic vs. model-specific approach

Interest in model interpretability is as old as the statistical modeling itself. Some classes of models have been developed for a long period of time or have attracted intensive research. Consequently, those classes of models are equipped with excellent tools for model exploration or visualisation. For example:

- There are many tools for diagnostics and evaluation of linear models, see for example (Galecki and Burzykowski, 2013) or (Faraway, 2002). Model assumptions are formally defined (normality, linear structure, homogenous variance) and can be checked by using normality tests or plots (normal qq-plot), diagnostic plots, tests for model structure, tools for identification of outliers, etc.
- For many more advanced models with an additive structure, like the proportional hazards model, many tools can be used for checking model assumptions, see for example (Harrell Jr, 2018) or (Sheather, 2009).
- Random-forest models are equipped with the out-of-bag method of evaluating performance and several tools for measuring variable importance (Breiman et al., 2018). Methods have been developed to extract information from the model structure about possible interactions (Paluszynska and Biecek, 2017). Similar tools have been developed for other ensembles of trees, like boosting models (xgboost, gbm). See (Foster, 2017) or (Karbowiak and Biecek, 2019).
- Neural networks enjoy a large collection of dedicated model-explanation tools that use, for instance, the layer-wise relevance propagation technique (Bach et al., 2015), or saliency maps technique (Simonyan et al., 2013), or a mixed approach. Broader summary is presented in (Samek et al., 2017) and (Alber et al., 2018).
- BERT family of models leads to high-performance models in Natural Language Processing. The exBERT method (Hoover et al., 2019) is designed to visualize the activation of attention heads in this model.

Of course, the list of model classes with dedicated collections of

model-explanation and/or diagnostics methods is much longer. This variety of model-specific approaches does lead to issues, though. For instance, one cannot easily compare explanations for two models with different structures. Also, every time a new architecture or a new ensemble of models is proposed, one needs to look for new methods of model exploration. Finally, for brand-new models no tools for model explanation or diagnostics may be immediately available.

For these reasons, in our book we focus on model-agnostic techniques. In particular, we prefer not to assume anything about the model structure, as we may be dealing with a black-box model with an unspecified structure. Often we do not have access to model parameters just to a specified Application Programming Interface (API) that allows for quering remote models (for exmaple in Microsoft Cognitive Services (Azure, 2019)). In that case, the only operation that we may be able to perform is the evaluation of a model for a specified data.

However, while we do not assume anything about the structure of the model, we will assume that the model operates on p-dimensional vector of variables/features and, for a single observation, it returns a single value (score/probability) which is a real number. This assumption holds for a broad range of models for data such as tabular data, images, text data, videos, etc. It may not be suitable for, e.g., models with memory like sequence-to-sequence models (Sutskever et al., 2014) or Long Short Term Memory models (Hochreiter and Schmidhuber, 1997) in which the model output depends also on sequence of previous inputs or generative models that output text of images.

0.1.8 What is in this book and what is not

The area of model exploration and explainability is quickly growing and is present in many different flavors. Instead of showing every existing method (is it really possible?) we rather selected a subset of consistent tools that are a good starting set for model exploration. Our focus was on the impact of the model exploration

and explanation tools rather than on selected methods. We believe that once we become aware of potential beyond visual model exploration, once we will learn a language of model explanation, we will improve our process of data modeling.

Taking this goal into account **in this book, we do show**

- how to determine features that affect model prediction for a single observation. In particular, we present the theory and examples of methods that can be used to explain prediction like Break Down plots, Ceteris Paribus profiles, local-model approximations, or Shapley values;
- techniques to examine fully-trained machine-learning models as a whole. In particular, we review the theory and examples of methods that can be used to explain model performance globally, like partial-dependency plots, variable-importance plots, and others;
- charts that can be used to present key information in a quick way;
- tools and methods for model comparison;
- code snippets for R and Python that explain how to use the described methods.

On the other hand, **in this book, we do not focus on**

- any specific model. The techniques presented are model agnostic and do not make any assumptions related to the model structure;
- data exploration. There are very good books on this topic, like *R for Data Science* by Garrett Grolemund and Hadley Wickham (Grolemund and Wickham, 2019) or *Python for Data Analysis* (Wes, 2012) by Wes McKinney or an excellent *Exploratory Data Analysis* by John Tukey (Tukey, 1977);
- the process of model building. There are also very good books on this topic, see *Modern Applied Statistics with S* by W. Venables and B. Ripley (Venables and Ripley, 2002), *An Introduction to Statistical Learning* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (James et al., 2014) or *Computer Age Statistical Inference* by Bradley Efron and Trevor Hastie (Efron and Hastie, 2016);
- any particular tools for model building. These are discussed, for

instance, in *Applied Predictive Modeling* by Max Kuhn and Kjell Johnson (Kuhn and Johnson, 2013a).

0.1.9 Acknowledgements

This book has been prepared using the `bookdown` package (Xie, 2018), created thanks to the amazing work of Yihui Xie. Figures and tables are created in R language for statistical computing (R Core Team, 2018) with numerous libraries that support predictive modeling. Just to name few frequently used in this book `randomForest` (Liaw and Wiener, 2002a), `ranger` (Wright and Ziegler, 2017), `rms` (Harrell Jr, 2018), `gbm` (Ridgeway, 2017) or `caret` (from Jed Wing et al., 2016). For statistical graphics we used the `ggplot2` library (Wickham, 2009) and for model governance we used `archivist` (Biecek and Kosinski, 2017).

Przemek's work on interpretability started during research trips within the RENOIR (H2020 grant no. 691152) secondments to Nanyang Technological University (Singapour) and Davis University of California (USA). So he would like to thank Prof. Janusz Holyst for the chance to take part in this project. Przemek would also like to thank Prof. Chris Drake for her hospitality. This book would have never been created without perfect conditions that Przemek found at Chris's house in Woodland.

0.2 Model Development

0.2.1 Introduction

In this book we present methods that can be used for exploration and explanation of predictive models. But before we can explore a model, first we need to train one.

In this part of the book we overview the process of model development and introduce steps that lead to a model creation. It is not

a comprehensive manual „how to train a model in 5 steps". The goal of this chapter is to show what needs to be performed before we can do any diagnostic or exploration of a trained model.

Predictive models are created for different purposes. Sometimes it is a team of data scientists that spend months on a single model that will be used for model scoring in a big financial company. Every detail is important for models that operate on large scale and have long-term consequences. Another time it is an in-house model trained for prediction of a demand for pizza. The model is developed by a single person in few hours. If model will not perform well it will be updated, replaced or removed.

Whatever it is a large model or small one, similar steps are to be taken during model development.

0.2.2 The Process

Several approaches are proposed in order to describe the process of model development. Their main goal is to standardize the process. And the standardisation is important because it helps to plan resources needed to develop and maintain the model and also to not miss any important step.

The most known methodology for data science projects is CRISP-DM (Chapman et al., 1999), (Wikipedia, 2019) which is a tool agnostic procedure. The key component of CRISP-DM is the break down of the whole process into six phases, that are iterated: business understanding, data understanding, data preparation, modeling, evaluation and deployment. CRISP-DM is general, it was designed for any data science project. For predictive models some methodologies are introduced in (Grolemund and Wickham, 2019) and (Hall, 2019).Both are focused on iterative repetitions of some phases. Figure 4 presents a variant of iterative process divided into five steps. Data preparation is needed prior to any modeling. Better data is needed for better models. On the other hand, garbage-in garbage-out. Once the data is gathered, steps that are usually highlighted are Data understanding, Model assembly and Model audit.

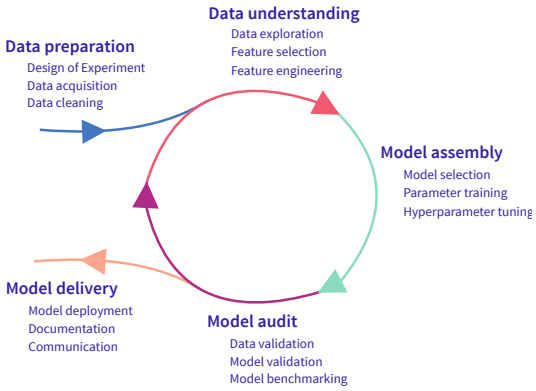


FIGURE 4 Lifecycle of predictive model can be decomposed into five tasks. First we need data that is poured into the model development cycle. The model development is highly iterative, learn something new about the data, assemble a new model based on current understanding, and validate the new model. Repeat these steps as long as needed to be satisfied with model performance. Once the model is created we can deliver the model to the production along with required tests and documentation.

This is the common thinking about model development. Repeat these steps until some convergence, e.g. repeat until best model is identified.

In this book we use *Model Development Process* introduced in (Biecek, 2019). It is motivated by Rational Unified Process for Software Development (Kruchten, 1998), (Jacobson et al., 1999), (Boehm, 1988). One can think about MDP as an extension of process introduced in Figure 4. What is important is to notice that consecutive iterations are not identical. Our knowledge increases during the process and consecutive iterations are performed with different goals in mind.

This is why MDP is build as an untangled version of Figure 4. The MDP process is shown in Figure 5. Each vertical stripe is a single run of the cycle. First iterations are usually focused on *formulation of the problem*. Sometimes the problem is well stated, however it's a rare situation valid maybe only for kaggle competitions. In

FIGURE 5 Overview of the Model Development Process. Horizontal axis show how time passes from the problem formulation to the model decommissioning. Vertical axis shows tasks are performed in a given phase. Each veritical strip is a next iteration of cycle presented in Figure ef(fig:MDPwashmachine)

most real-life problems the problem formulation requires lots of discussions and experiments. Once the problem is defined we can start building first prototypes, first *crisp versions of models*. These initial versions of models are needed to verify if the problem can be solved and how far we are form the solution. Usually we gather more information and go for the next phase, the *fine tuning*. We repeat these iterations until a final version of a model is developed. Then we move to the last phase *maintenance and* (one day) *decommissioning*.

Having in mind the map of model development we can point places where one can use methods presented in this book.

As suggested in the title of this book, three primary applications are: exploration, explanation and debugging. *Exploration* refers to situations in which we better understand the data and the domain.

Presented techniques can be used to speed up the variable engineering or variable selection. *Explanation* refers to situations in which we are interested in decision paths beyond particular predictions. *Debugging* refers to situations in which we want to understand week points of a model and correct them. These applications target phases Data understanding, Model assembly and Model audit.

In this book we present various examples based on three use cases. Two introduced in Chapter 0.5 (binary classification in surviving Titanic sinking and regression in apartments pricing) and one in Chapter 0.21 (estimation of soccer player value based on its skills). Due to space limitation we do not show the full life cycle of these problems, but we are focused on phases Crisp modeling and Fine tuning.

Rest of this chapter is focused on a brief overview of the notation and commonly used methods for data exploration, model training and model validation.

0.2.3 Notation

Methods described in this book were developed by different authors, who used different mathematical notations. We try to keep the mathematical notation consistent throughout the entire book. In some cases this may result in formulas with a fairly complex system of indices.

In this section, we provide a general overview of the notation we use. Whenever necessary, parts of the notation will be explained again in subsequent chapters.

We assume that the data consist n observations/instances. Each observation is described by p explanatory variables. Thus data is described as a set of points on a **p-dimensional input space** $\mathcal{X} \equiv \mathcal{R}^p$. By $x \in \mathcal{X}$ we will refer to a single point in this input space. By x_i we refer to the i-th observation in this dataset. Of course, $x_i \in \mathcal{X}$. By X we denote a matrix $n \times p$ with rows corresponding to consecutive observations.

Some methods of model exploration are constructed around an observation of interest which will be denoted by x_*. The observation may not necessarily belong to the analyzed dataset; hence, the use of the asterisk in the index. Of course, $x_* \in \mathcal{X}$.

Points in $\mathcal{X}$ are p dimensional vectors. We refer to the j-th coordinate by using j in superscript. Thus, x_i^j deontes the j-th coordinate of the i-th observation from the analyzed dataset. If $\mathcal{J}$ denotes a subset of indices, then $x^{\mathcal{J}}$ denotes the elements of vector x corresponding to the indices included in $\mathcal{J}$.

We will use the notation x^{-j} to refer to a vector that results from removing the j-th coordinate from vector x. By $x^{j|=z}$, we denote a vector with the values at all coordinates equal to the values in x, except of the j-th coordinate, which is set equal to z. So, if $w = x^{j|=z}$, then $w^j = z$ and $\forall_{k \neq j} w^k = x^k$. In other words $x^{j|=z} = (x^1, ..., x^{j-1}, z, x^{j+1}, ..., x^p)$.

In this book, a model is a function $f : \mathcal{X} \to \mathcal{R}$ that transforms a point from $\mathcal{X}$ into a real number. In most cases, the presented methods can be used directly for multi-variate dependent variables; however, we use examples with uni-variate responses to simplify the notation. Typically, during the model development, we create many competing models. Formally we shall index models to refer to a specific version of a trained model. But for the sake of simplicity we omit these indexes where they are not important.

Later in this book we will use the term **model residual** as the the difference between the observed value of the dependent variable Y for the i-th observation from a particular dataset and the model prediction for the observaton

$$r_i = y_i - f(x_i) = y_i - \hat{y}_i. \tag{0.1}$$

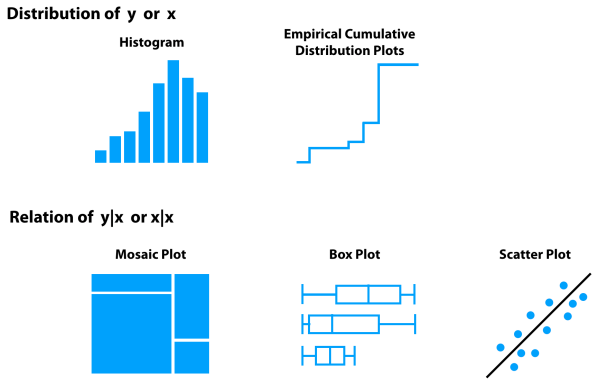


FIGURE 6 Basic methods for visual exploration. Histogram for distribution of continuous or categorical variables, empirical cumulative distribution for continuous variables. Mosaic plot for relation between two categorical variables, boxplots for relation between continuous and categorical variables or scatterplot for relation between two continuous variables.

0.2.4 Data exploration

Before we start the modeling we need to understand the data. Visual, tabular and statistical tools for data exploration are used depending on the character of variables.

The most know introduction to data exploration is the famous book by John Tukey (Tukey, 1977). It introduces new tools for data exploration, like for example boxplots or stem-and-leaf plots. Availability of computational tools makes the process of data exploration easier and more interactive. Find a good overview of techniques for data exploration in (Nolan and Lang, 2015) or (Wickham and Grolemund, 2017).

In this book we will relay on five visual methods for data exploration presented in Figure 6. Two of them are used to present distribution of explanatory or target variables; three others are used to explore pairwise relations between variables.

Distribution of categorical variable is summarized with a barplot,

distribution of numerical variable is summarized with a histogram or empirical cumulative distribution function.

Primary goal for exploration of target variable is to decide if some variable transformation is needed (e.g. if the variable is skewed or with fat tails) or to verify if target variable is balanced (because some methods are not working well with unbalanced data). Exploration of dependent variables is performed mainly to decide if any variable transformation is needed.

Relations between two variables, mostly between a single dependent variable and target variable, are visualized with mosaic plots (for two categorical variables), boxplots (for numerical and categorical variable) and scatter plots (for two numerical variables). Such exploration may provide some insights for variable selection/filtering (if the variable is not related with the target then variable may be removed from the model) or variable engineering (if from the exploration we gain information how a variable may be transformed).

0.2.5 Model training

In predictive modeling, we are interested in a distribution of a dependent variable Y given vector x_*. The latter contains values of explanatory variables. In the ideal world, we would like to know the conditional distribution of Y given x_*. In practical applications, however, we usually do not predict the entire distribution, but just some of its characteristics like the expected (mean) value, a quantile, or variance. Without loss of generality we will assume that we model the conditional expected value $E_Y(Y|x_*)$.

Assume that we have got model $f()$, for which $f(x_*)$ is an approximation of $E_Y(Y|x_*)$, i.e., $E_Y(Y|x_*) \approx f(x_*)$. Note that we do not assume that it is a "good" model, nor that the approximation is precise. We simply assume that we have a model that is used to estimate the conditional expected value and to form predictions of the values of the dependent variable. Our interest lies in the evaluation of the quality of the predictions. If the model offers a

"good" approximation of the conditional expected value, it should be reflected in its satisfactory predictive performance.

Usually the available data is split into two parts. One will be used for model training (estimation of model parameters), second will be used for model validation. The splitting may be repeated as in k-fold cross validation or repeated k-fold cross validation (see for example (Kuhn and Johnson, 2013b)). We leave the topic of model validation for Chapter 0.16.

Training procedures are different for different models, but most of them can be written as an optimization problem. Let Θ be a space for possible model parameters. Model training is a procedure of selection a $\theta \in \Theta$ that maximize some loss function $L(y, f_\theta(X))$. For models with large parameter spaces it is common to add additional term $\lambda(\theta)$ that control the model complexity.

$$\hat{\theta} = \arg\min_{\theta \in \Theta} L(y, f_\theta(X)) + \lambda(\theta). \tag{0.2}$$

For statistical models it is common to assume some family of probability distributions for $y|x$. In such case the loss function L may be defined as a minus log-likelihood function for θ. Likelihood is probability of observing $y|x$ as a function of parameter θ.

For example, in linear regression we assume that that observed vector of values y follow a multidimensional Gaussian distribution

$$y \sim \mathcal{N}(X\beta, I\sigma^2),$$

where $\theta = (\beta, \sigma^2)$. In this case equation (0.2) become

$$\hat{\theta} = \arg\min_{\theta \in \Theta} ||y - X\beta||_2 + \lambda(\beta). \tag{0.3}$$

For linear regression, the penalty term $\lambda(\beta)$ is equal to 0, and optimal parameters β in equation (0.3) have close analytical solution $\hat{\beta} = (X^T X)^{-1} X^T y$. In ridge regression the penalty $\lambda(\beta) = \lambda||\beta||_2$ and also (0.3) have analytical solution $\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y$. For

LASSO regression the penalty $\lambda(\beta) = \lambda||\beta||_1$ and β are estimated through a numerical optimization.

For classification, the natural choice for distribution of y is a Binomial distribution. This leads to logistic regression and logistic loss function. For multi label classification frequent choice is the cross-entropy loss function.

Apart from linear models for y there is a large variety of predictive models. Find a good overview of different techniques for model development in (Venables and Ripley, 2010) or (Kuhn and Johnson, 2013b).

0.2.6 Model understanding

Usually the model development starts with some crisp early versions that are refined in consecutive iterations. In order to train a final model we need to try numerous candidate models that will be explored, examined and diagnosed. In this book we will introduce techniques that:

- summarise how good is the current version of a model. Section 0.16 overviews measures for model performance. These measures are usually used to trace the progress in model development.
- assess the feature importance. Section 0.17 shows how to assess influence of a single variable on model performance. Features that are not important are usually removed from a model during the model refinement.
- shows how a single feature affects the model response. Sections 0.18 – 0.19 present Partial Dependency Profiles, Accumulated Local Effects and Marginal Profiles. All these techniques help to understand how model consumes particular features.
- identifies potential problems with a model. Section 0.20 shows techniques for exploration of model residuals. Looking closer on residuals often help to improve the model. This is possible with tools for local model exploration which are presented in the fist part of the book.
- performs sensitivity analysis for a model. Section 0.11 introduces

Ceteris Paribus profiles that helps in a what-if analysis for a model.
- validated local fit for a model. Section 0.13 introduces techniques for assessment if for a single observation the model support its prediction.
- decompose model predictions into pieces that can be attributed to particular variables. Sections 0.7 – 0.10 show different techniques like SHAP, LIME or Break Down for local exploration of a model.

0.3 Do-it-yourself With R

In this book we introduce various methods for instance-level and dataset-level explanation and exploration of predictive models. In each chapter, there is a section with code snippets for R and python that shows how to use a particular method. In this chapter we provide a short description of steps that are needed to set-up the environment with required libraries.

0.3.1 What to install?

Obviously, the R software (R Core Team, 2018) is needed. It is always a good idea to use the newest version. At least R in version 3.6 is recommended. R can be downloaded from the CRAN website https://cran.r-project.org/.

A good editor makes working with R much easier. There is a plenty of choices, but, especially for beginners, it is worth considering the RStudio editor, an open-source and enterprise-ready tool for R. It can be downloaded from https://www.rstudio.com/.

Once R and the editor are available, the required packages should be installed.

The most important one is the DALEX package in version 1.0 or

newer. It is the entry point to solutions introduced in this book. The package can be installed by executing the following command from the R command line:

```
install.packages("DALEX")
```

Installation of DALEX will automatically take care about installation of other hard requirements (packages required by it), like the ggplot2 package for data visualization.

To repeat all examples in this book, two additional packages are needed: ingredients and iBreakDown. The easiest way to get them, including other useful weak dependencies, is to execute the following command:

```
DALEX::install_dependencies()
```

0.3.2 How to work with DALEX?

To conduct model exploration with DALEX, first, a model has to be created. Then the model has got to be prepared for exploration.

There are many packages in R that can be used to construct a model. Some packages are structure-specific, like randomForest for Random-Forest Classification and Regression models (Liaw and Wiener, 2002b), gbm for Generalized Boosted Regression Models (Ridgeway, 2017), extensions for Generalized Linear Models (Harrell Jr, 2018), or many others. There is also a number of packages that can be used for constucting models with different structures. These include the h2o package (LeDell et al., 2019), caret (from Jed Wing et al., 2016) and its successor parsnip (Kuhn and Vaughan, 2019), a very powerful and extensible framework mlr (Bischl et al., 2016), or keras that is a wrapper to Python library with the same name (Allaire and Chollet, 2019).

While it is great to have such a large choice of tools for constructing models, the downside is that different packages have different interfaces and different arguments. Moreover, model-objects created with different packages may have different internal structures. The main goal of the DALEX package is to create a level of abstraction

around a model that makes it easier to explore and explain the model.

Function `DALEX::explain` is THE function for model wrapping. The function requires five arguments:

- `model`, a model-object;
- `data`, a data frame with validation data;
- `y`, observed values of the dependent variable for the validation data; it is an optional argument, required for explainers focused on model validation and benchmarking.
- `predict_function`, a function that returns prediction scores; if not specified, then a default `predict()` function is used. Note that, for some models, the default `predict()` function returns classes; in such cases you should provide a function that will return numerical scores.
- `label`, a name of a model; if not specified, then it is extracted from the `class(model)`. This name will be presented in figures, so it is recommended to make the name informative.

For an example, see Section 0.5.1.6.

0.3.3 How to work with `archivist`?

As we will focus on exploration of predictive models, we prefer not to waste space nor time on replication of the code necessary for model development. This is where the `archivist` packages helps.

The `archivist` package (Biecek and Kosinski, 2017) is designed to store, share, and manage R objects. We will use it to easily access pretrained R models and precalculated explainers. To install the package, the following command should be executed in the R command line:

```
install.packages("archivist")
```

Once the package has been installed, function `aread()` can be used to retrieve R objects from any remote repository. For this book, we use a GitHub repository `models` hosted at https://github.com/

pbiecek/models. For instance, to download a model with the md5 hash `ceb40`, the following command has to be executed:

```
archivist::aread("pbiecek/models/ceb40")
```

Since the md5 hash `ceb40` uniquely defines the model, referring to the repository object results in using exactly the same model and the same explanations. Thus, in the subsequent chapters, pre-constructed model explainers will be accessed with `archivist` hooks. In following sections we will also use `archivist` hooks in references to datasets.

0.4 Do-it-yourself With Python

0.5 Data sets and models

We illustrate the methods presented in this book by using two datasets:

- Predicting odds of survival out of *Sinking of the RMS Titanic*
- Predicting prices for *Apartments in Warsaw*

The first dataset will be used to illustrate the application of the techniques in the case of a predictive model for a binary dependent variable. The second one will provide an example for models for a continuous variable.

In this chapter, we provide a short description of each of the datasets, together with results of exploratory analyses. We also introduce models that will be used for illustration purposes in subsequent chapters.

FIGURE 7 Titanic sinking by Willy Stöwer

0.5.1 Sinking of the RMS Titanic

Sinking of the RMS Titanic is one of the deadliest maritime disasters in history (during peacetime). Over 1500 people died as a consequence of collision with an iceberg. Projects like *Encyclopedia titanica* `https://www.encyclopedia-titanica.org/` are a source of rich and precise data about Titanic's passengers. The `stablelearner` package includes a data frame with some passenger characteristics. The dataset, after some data cleaning and variable transformations, is also avaliable in the `DALEX` package. In particular, the 'titanic' data frame contains 2207 observations (for 1317 passengers and 890 crew members) and nine variables:

- *gender*, person's (passenger's or crew member's) gender, a factor (categorical variable) with two levels (categories) `male` (78%) and `female` (22%);
- *age*, person's age in years, a numerical variable; the age is given in (integer) years, range 0 – 74 years;
- *class*, the class in which the passenger travelled, or the duty class

of a crew member; a factor with seven levels: `1st` (14.7%), `2nd` (12.9%), `3rd` (32.1%), `deck crew` (3%), `engineering crew` (14.7%), `restaurant staff` (3.1%), `victualling crew` (19.5%);
- *embarked*, the harbor in which the person embarked on the ship, a factor with four levels, `Belfast` (8.9%), `Cherbourg` (12.3%), `Queenstown` (5.6%), `Southampton` (73.2%);
- *country*, person's home country, a factor with 48 levels, the most common are `England` (51%), `United States` (12%), `Ireland` (6.2%) and `Sweden` (4.8%);
- *fare*, the price of the ticket (only available for passengers; 0 for crew members), a numerical variable range 0 – 512;
- *sibsp*, the number of siblings/spouses aboard the ship, a numerical variable range 0 – 8;
- *parch*, the number of parents/children aboard the ship, a numerical variable range 0 – 9;
- *survived*, a factor with two levels `yes` (67.8%), `no` (32.2%), indicating whether the person survived or not.

The R code below provides more info about the contents of the dataset, values of the variables, etc.

```
library("DALEX")
head(titanic, 2)
```

```
##   gender age class    embarked       country  fare sibsp parch survived
## 1   male  42   3rd Southampton United States  7.11     0     0       no
## 2   male  13   3rd Southampton United States 20.05     0     2       no
```

Models considered for this dataset will use *survived* as the (binary) dependent variable.

0.5.1.1 Data exploration

It is always advisable to explore data before modelling. However, as this book is focused on model exploration, we will limit the data exploration part.

Before exploring the data, we first do some pre-processing. In particular, the value of variables *age*, *country*, *sibsp*, *parch*, and *fare*

is missing for a limited number of observations (2, 81, 10, 10, and 26, respectively). Analyzing data with missing values is a topic on its own (Little and Rubin 1987; Schafer 1997; Molenberghs and Kenward 2007). An often-used approach is to impute the missing values. Toward this end, multiple imputation should be considered (Schafer 1997; Molenberghs and Kenward 2007; van Buuren 2012). However, given the limited number of missing values and the intended illustrative use of the dataset, we will limit ourselves to, admittedly inferior, single imputation. In particular, we replace the missing *age* values by the mean of the observed ones, i.e., 30. Missing *country* will be coded by "X". For *sibsp* and *parch*, we replace the missing values by the most frequently observed value, i.e., 0. Finally, for *fare*, we use the mean fare for a given *class*, i.e., 0 pounds for crew, 89 pounds for the 1st, 22 pounds for the 2nd, and 13 pounds for the 3rd class. The R code presented below implements the imputation steps.

- missing `age` is replaced by its average, that is 30

```
titanic$age[is.na(titanic$age)] = 30
```

- missing `country` is replaced by `"X"`

```
titanic$country <- as.character(titanic$country)
titanic$country[is.na(titanic$country)] = "X"
titanic$country <- factor(titanic$country)
```

- missing `fare` is replaced by within `class` average, that is 89, 22 and 13 correspondingly

```
titanic$fare[is.na(titanic$fare) & titanic$class == "1st"] = 89
titanic$fare[is.na(titanic$fare) & titanic$class == "2nd"] = 22
titanic$fare[is.na(titanic$fare) & titanic$class == "3rd"] = 13
```

- missing `sibsp` and `parch` are replaced by 0

```
titanic$sibsp[is.na(titanic$sibsp)] = 0
titanic$parch[is.na(titanic$parch)] = 0
```

After imputing the missing values, we investigate the association between survival status and other variables. Figures 9-12 present

graphically the proportion non- and survivors for different levels of the other variables with the use of mosaic plots. The height of the bars (on the y-axis) reflects the marginal distribution (proportions) of the observed levels of the variable. On the other hand, the width of the bars (on the x-axis) provides the information about the proportion of non- and survivors. Note that, to construct the graphs for *age* and *fare*, we categorized the range of the observed values.

Figure 9 indicates that the proportion of survivors was larger for females and children below 5 years of age. This is most likely the result of the "women and children first" principle that is often evoked in situations that require evacuation of persons whose life is in danger. The principle can, perhaps, partially explain the trend seen in Figure 10, i.e., a higher proportion of survivors among those with 1-3 parents/children and 1-2 siblings/spouses aboard. Figure 11 indicates that passengers travelling in the first and second class had a higher chance of survival, perhaps due to the proximity of the location of their cabins to the deck. Interestingly, the proportion of survivors among crew deck was similar to the proportion of the first-class passengers. It also shows that the proportion of survivors increased with the fare, which is consistent with the fact that the proportion was higher for passengers travelling in the first and second class. Finally, Figure 12 does not suggest any noteworthy trends.

0.5.1.2 Logistic regression model

The dependent variable of interest, *survival*, is binary. Thus, a natural choice is to start the predictive modelling with logistic regression model. As there is no reason to expect a linear relationship between age and odds of survival, we use linear tail-restricted cubic splines, available in the `rcs()` function of the `rms` package (Harrell Jr, 2018), to model the effect of age. We also do not expect linear relation for the `fare` variable, but because of it's skewness, we do not use splines for this variable. The results of the model are stored in model-object `titanic_lmr_v6`, which will be used in subsequent chapters.

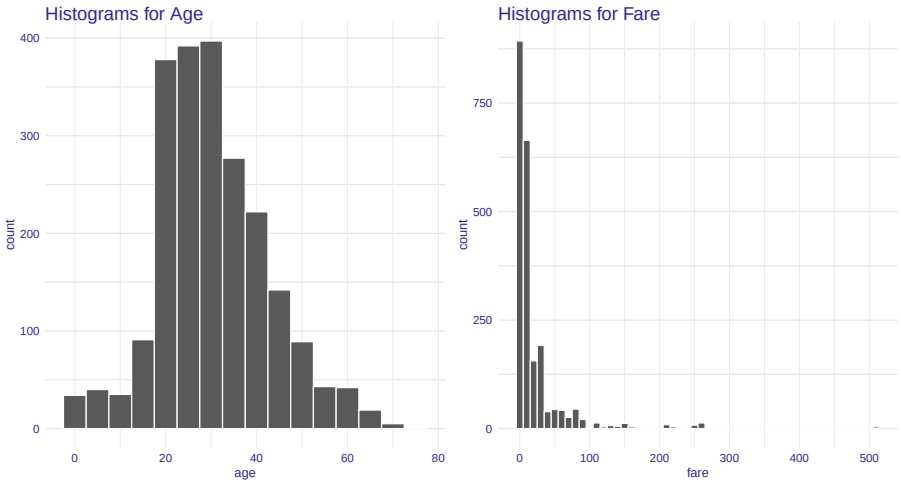


FIGURE 8 Histogram of Age and Fare for the Titanic data.

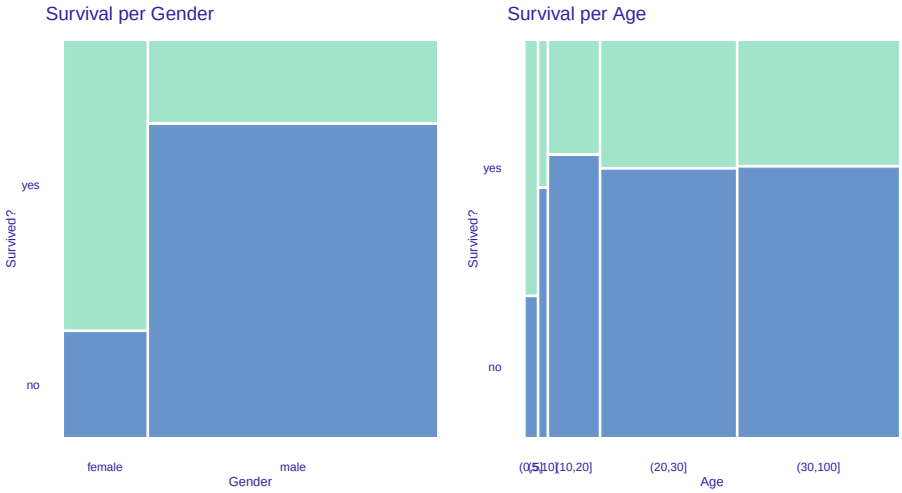


FIGURE 9 Survival status in groups defined be Gender and Age for the Titanic data.

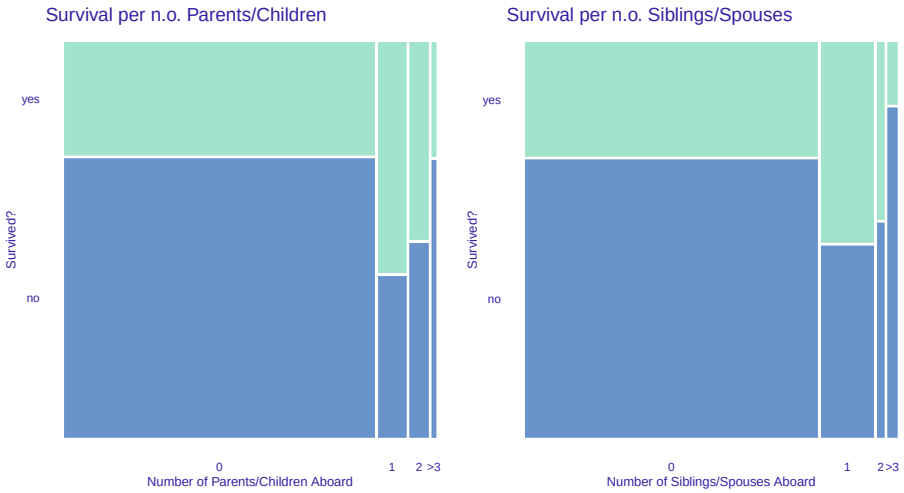


FIGURE 10 Survival according to the number of parents/children and siblings/spouses in the Titanic data.

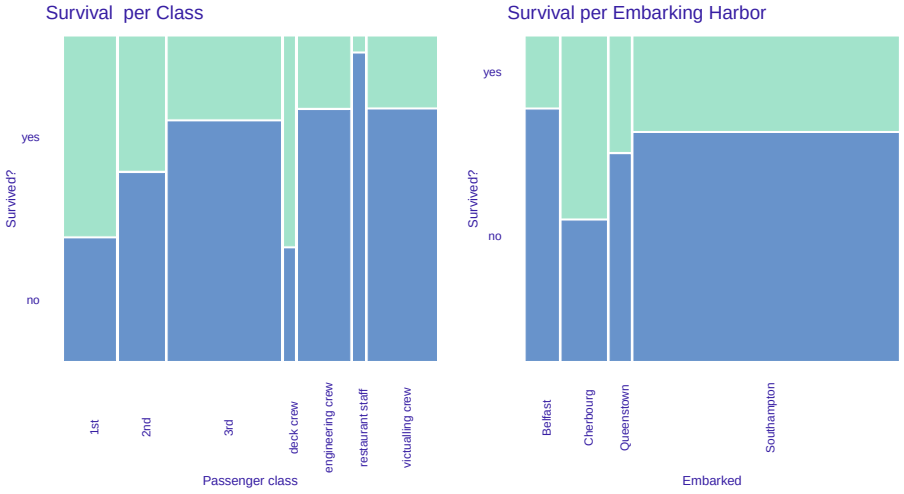


FIGURE 11 Survival according to the class and port of embarking in the Titanic data.

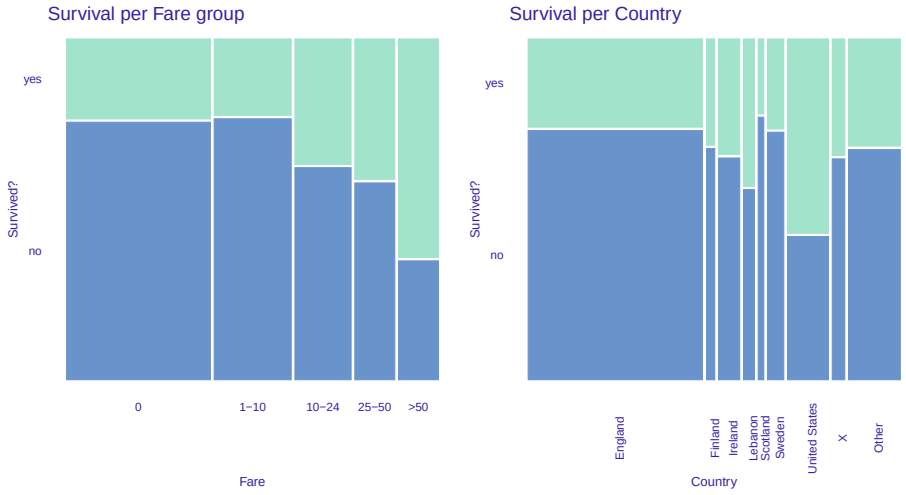


FIGURE 12 Survival according to fare and country in the Titanic data.

```
library("rms")
set.seed(1313)
titanic_lmr_v6 <- lrm(survived == "yes" ~ gender + rcs(age) + class + sibsp +
                    parch + fare + embarked, titanic)
titanic_lmr_v6
```

```
## Logistic Regression Model
##
##  lrm(formula = survived == "yes" ~ gender + rcs(age) + class +
##      sibsp + parch + fare + embarked, data = titanic)
##
##                         Model Likelihood     Discrimination    Rank Discrim.
##                            Ratio Test           Indexes           Indexes
##  Obs           2207    LR chi2     752.06    R2       0.404    C       0.817
##   FALSE        1496    d.f.            17    g        1.647    Dxy     0.635
##   TRUE          711    Pr(> chi2) <0.0001    gr       5.191    gamma   0.636
##  max |deriv| 0.0001                          gp       0.282    tau-a   0.277
##                                              Brier    0.146
##
##                       Coef    S.E.   Wald Z Pr(>|Z|)
```

```
##  Intercept               4.5746 0.5480   8.35 <0.0001
##  gender=male            -2.7687 0.1586 -17.45 <0.0001
##  age                    -0.1180 0.0221  -5.35 <0.0001
##  age'                    0.6313 0.1628   3.88 0.0001
##  age''                  -2.6583 0.7840  -3.39 0.0007
##  age'''                  2.8977 1.0130   2.86 0.0042
##  class=2nd              -1.1390 0.2501  -4.56 <0.0001
##  class=3rd              -2.0627 0.2490  -8.28 <0.0001
##  class=deck crew         1.0672 0.3498   3.05 0.0023
##  class=engineering crew -0.9702 0.2648  -3.66 0.0002
##  class=restaurant staff -3.1712 0.6583  -4.82 <0.0001
##  class=victualling crew -1.0877 0.2596  -4.19 <0.0001
##  sibsp                  -0.4504 0.1006  -4.48 <0.0001
##  parch                  -0.0871 0.0987  -0.88 0.3776
##  fare                    0.0014 0.0020   0.70 0.4842
##  embarked=Cherbourg      0.7881 0.2836   2.78 0.0055
##  embarked=Queenstown     0.2745 0.3409   0.80 0.4208
##  embarked=Southampton    0.2343 0.2119   1.11 0.2689
##
```

Note that our prime interest is not in the assessment of model performance, but rather in the understanding of model behavior. This is why we do not split the data int train/test subsets. The model is trained and will be explained on the whole dataset.

0.5.1.3 Random forest model

As a challanger to the logistic regression model, we consider a random forest model. Random forest is known for good predictive performance, is able to grasp low-order variable interactions, and is quite stable (Breiman, 2001). To fit the model, we apply the `randomForest()` function, with default settings, from the package with the same name (Liaw and Wiener, 2002a).

In the first instance, we fit a model with the same set of explanatory variables as the logistic regression model. The results of the model are stored in model-object `titanic_rf_v6`.

```
library("randomForest")
set.seed(1313)
titanic_rf_v6 <- randomForest(survived ~ class + gender + age + sibsp + parch + fare + embarke
                              data = titanic)
titanic_rf_v6
```

```
##
## Call:
##  randomForest(formula = survived ~ class + gender + age + sibsp +      parch + fare + embar
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 18.62%
## Confusion matrix:
##       no yes class.error
## no  1393 103  0.06885027
## yes  308 403  0.43319269
```

For comparison purposes, we also consider a model with only three explanatory variables: *class*, *gender*, and *age*. The results of the model are stored in model-object `titanic_rf_v3`.

```
titanic_rf_v3 <- randomForest(survived ~ class + gender + age, data = titanic)
titanic_rf_v3
```

```
##
## Call:
##  randomForest(formula = survived ~ class + gender + age, data = titanic)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 1
##
##         OOB estimate of  error rate: 21.02%
## Confusion matrix:
##       no yes class.error
## no  1367 129  0.08622995
## yes  335 376  0.47116737
```

0.5.1.4 Gradient boosting model

Finally, we consider yet another challanger – the gradient-boosting model (Friedman, 2000). The tree based boosting models are known for being able to accomodate higher-order interactions between variables. We use the same set of six explanatory variables as for the logistic regression model. To fit the gradient-boosting model, we use function `gbm()` from the `gbm` package (Ridgeway, 2017). The results of the model are stored in model-object `titanic_gbm_v6`.

```
library("gbm")
set.seed(1313)
titanic_gbm_v6 <- gbm(survived == "yes" ~ class + gender + age + sibsp + parch + fare + embark
                      data = titanic, n.trees = 15000, distribution = "bernoulli")
titanic_gbm_v6
```

```
## gbm(formula = survived == "yes" ~ class + gender + age + sibsp +
##     parch + fare + embarked, distribution = "bernoulli", data = titanic,
##     n.trees = 15000)
## A gradient boosted model with bernoulli loss function.
## 15000 iterations were performed.
## There were 7 predictors of which 7 had non-zero influence.
```

0.5.1.5 Model predictions

Let us now compare predictions that are obtained from the three different models. In particular, we will compute the predicted probability of survival for an 8-year-old boy who embarked in Belfast and travelled in the 1-st class with no parents nor siblings and with a ticket costing 72 pounds.

First, we create a data frame `johny_d` that contains the data describing the passenger.

```
johny_d <- data.frame(
            class = factor("1st", levels = c("1st", "2nd", "3rd", "deck crew", "engineering cr
            gender = factor("male", levels = c("female", "male")),
            age = 8,
```

```
            sibsp = 0,
            parch = 0,
            fare = 72,
            embarked = factor("Southampton", levels = c("Belfast","Cherbourg","Queenstown","So
)
```

Subsequently, we use the generic function `predict()` to get the predicted probability of survival for the logistic regression model.

```
(pred_lmr <- predict(titanic_lmr_v6, johny_d, type = "fitted"))
```

```
##         1
## 0.7677036
```

The predicted probability is equal to 0.77.

We do the same for the random forest and gradient boosting models.

```
(pred_rf <- predict(titanic_rf_v6, johny_d, type = "prob"))
```

```
##      no   yes
## 1 0.578 0.422
## attr(,"class")
## [1] "matrix" "votes"
```

```
(pred_gbm <- predict(titanic_gbm_v6, johny_d, type = "response", n.trees = 15000))
```

```
## [1] 0.6632574
```

As a result, we obtain the predicted probabilities of 0.42 and 0.66, respectively.

The models lead to different probabilities. Thus, it might be of interest to understand the reason for the differences, as it could help us to decide which of the predictions we might want to trust.

Note that for some examples we will use another observation (instance) with lower chances of survival. Let's call this passenger Henry.

```
henry <- data.frame(
            class = factor("1st", levels = c("1st", "2nd", "3rd", "deck crew", "engineering cr
```

```
            gender = factor("male", levels = c("female", "male")),
            age = 47,
            sibsp = 0,
            parch = 0,
            fare = 25,
            embarked = factor("Cherbourg", levels = c("Belfast","Cherbourg","Queenstown","Sout
)
predict(titanic_lmr_v6, henry, type = "fitted")
```

```
##         1
## 0.4318245
```

```
predict(titanic_rf_v6, henry, type = "prob")
```

```
##      no   yes
## 1 0.754 0.246
## attr(,"class")
## [1] "matrix" "votes"
```

```
predict(titanic_gbm_v6, henry, type = "response", n.trees = 15000)
```

```
## [1] 0.3073358
```

0.5.1.6 Model adapters

Model-objects created with different machine learning libraries may have different internal structures. Thus, first, we have got to create an adapter for the model that provides an uniform interface. Toward this end, we use the `explain()` function from the `DALEX` package (Biecek, 2018). The function requires five arguments:

- `model`, a model-object;
- `data`, a validation data frame;
- `y`, observed values of the dependent variable for the validation data;
- `predict_function`, a function that returns prediction scores; if not specified, then a default `predict()` function is used;
- `label`, an unique name of the model; if not specified, then it is extracted from the `class(model)`.

Each adapter containes all elements needed to create a model explanation, i.e., a suitable `predict()` function, validation data set, and the model object. Thus, in subsequent chapters we will use the explainers instead of the model objects to keep code snippets more concise.

```
explain_titanic_lmr_v6 <- explain(model = titanic_lmr_v6,
                                  data = titanic[, -9],
                                  y = titanic$survived == "yes",
                                  label = "Logistic Regression v6")
explain_titanic_rf_v6 <- explain(model = titanic_rf_v6,
                                  data = titanic[, -9],
                                  y = titanic$survived == "yes",
                                  label = "Random Forest v6")
explain_titanic_rf_v3 <- explain(model = titanic_rf_v3,
                                  data = titanic[, -9],
                                  y = titanic$survived == "yes",
                                  label = "Random Forest v3")
explain_titanic_gbm_v6 <- explain(model = titanic_gbm_v6,
                                  data = titanic[, -9],
                                  y = titanic$survived == "yes",
                                  label = "Generalized Boosted Regression v6")
```

0.5.1.7 List of objects for the `titanic` example

In the previous sections we have built four predictive models for the `titanic` data set. The models will be used in the rest of the book to illustrate the model explanation methods and tools.

For the ease of reference, we summarize the models in Table 0.1. The binary model-objects can be downloaded by using the indicated `archivist` hooks (Biecek and Kosinski, 2017). By calling a function specified in the last column of the table, one can restore a selected model in its local R environment.

TABLE 0.1: Predictive models created for the `titanic` dataset.

Model name	Model generator	Variables	Archivist hooks
`titanic_lmr_v6`	`rms:: lmr` v.5.1.3	gender, age, class, sibsp, parch, fare, embarked	Get the model: `archivist:: aread("pbiecek/models/ceb40")`. Get the explainer: `archivist:: aread("pbiecek/models/51c50")`
`titanic_rf_v6`	`randomForest:: randomForest` v.4.6.14	gender, age, class, sibsp, parch, fare, embarked	Get the model: `archivist:: aread("pbiecek/models/1f938")`. Get the explainer: `archivist:: aread("pbiecek/models/42d51")`
`titanic_rf_v3`	`randomForest:: randomForest` v.4.6.14	gender, age, class	Get the model: `archivist:: aread("pbiecek/models/855c1")`. Get the explainer: `archivist:: aread("pbiecek/models/0e5d2")`
`titanic_gbm_v6`	`gbm:: gbm` v.2.1.5	gender, age, class, sibsp, parch, fare, embarked	Get the model: `archivist:: aread("pbiecek/models/24e72")`. Get the explainer: `archivist:: aread("pbiecek/models/3d514")`

Table 0.2 summarizes the data frames that will be used in examples in the subsequent chapters.

TABLE 0.2: Data frames created for the `titanic` example.

Description	No. rows	Variables	Link to this object
`titanic` dataset with imputed missing values	2207	gender, age, class, embarked, country, fare, sibsp, parch, survived	`archivist::aread("pbiecek/models/27e5c")`
`johny_d` 8-year-old boy that travelled in the 1st class without parents	1	class, gender, age, sibsp, parch, fare, embarked	`archivist::aread("pbiecek/models/e3596")`
`henry` 47-year-old male passenger from the 1st class, paid 25 pounds and embarked at Cherbourg	1	class, gender, age, sibsp, parch, fare, embarked	`archivist::aread("pbiecek/models/a6538")`

0.5.2 Apartment prices

Predicting house prices is a common exercise used in machine-learning courses. Various datasets for house prices are available at websites like Kaggle (https://www.kaggle.com) or UCI Machine Learning Repository (https://archive.ics.uci.edu).

In this book, we will work with an interesting variant of this problem. The `apartments` dataset is an artificial dataset created to match key characteristics of real apartments in Warsaw, the capital of

FIGURE 13 Warsaw skyscrapers by Artur Malinowski Flicker

Poland. However, the dataset is created in a way that two very different models, namely linear regression and random forest, have almost exactly the same accuracy. The natural question is then: which model should we choose? We will show that the model-explanation tools provide important insight into the key model characteristics and are helpful in model selection.

The dataset is available in the DALEX package (Biecek, 2018). It contains 1000 observations (apartments) and six variables:

- *m2.price*, apatments price per meter-squared (in EUR), a numerical variable range 1607 – 6595;
- *construction.year*, the year of construction of the block of flats in which the apartment is located, a numerical variable range 1920 – 2010;
- *surface*, apartment's total surface in squared meters, a numerical variable range 20 – 150;
- *floor*, the floor at which the apartment is located (ground floor

taken to be the first floor), a numerical integer variable with values from 1 to 10;
- *no.rooms*, the total number of rooms, a numerical variable with values from 1 to 6;
- *distric*, a factor with 10 levels indicating tha distric of Warszawa where the apartment is located.

The R code below provides more info about the contents of the dataset, values of the variables, etc.

```
library("DALEX")
head(apartments, 2)
```

```
##   m2.price construction.year surface floor no.rooms    district
## 1     5897              1953      25     3        1 Srodmiescie
## 2     1818              1992     143     9        5     Bielany
```

Models considered for this dataset will use *m2.price* as the (continuous) dependent variable.

Model predictions will be obtained for a set of six apartments included in data frame `apartments_test`.

```
head(apartments_test)
```

```
##      m2.price construction.year surface floor no.rooms    district
## 1001     4644              1976     131     3        5 Srodmiescie
## 1002     3082              1978     112     9        4     Mokotow
## 1003     2498              1958     100     7        4     Bielany
## 1004     2735              1951     112     3        5        Wola
## 1005     2781              1978     102     4        4      Bemowo
## 1006     2936              2001     116     7        4      Bemowo
```

0.5.2.1 Data exploration

Note that `apartments` is an artificial dataset created to illustrate and explain differences between random forest and linear regression. Hence, the structure of the data, the form and strength of association between variables, plausibility of distributional assumptions, etc., is better than in a real-life dataset. In fact, all these

characteristics of the data are known. Nevertheless, we conduct some data exploration to illustrate the important aspects of the data.

The variable of interest is *m2.price*, the price per meter-squared. The histogram presented in Figure **??** indicates that the distribution of the variable is slightly skewed to the right.

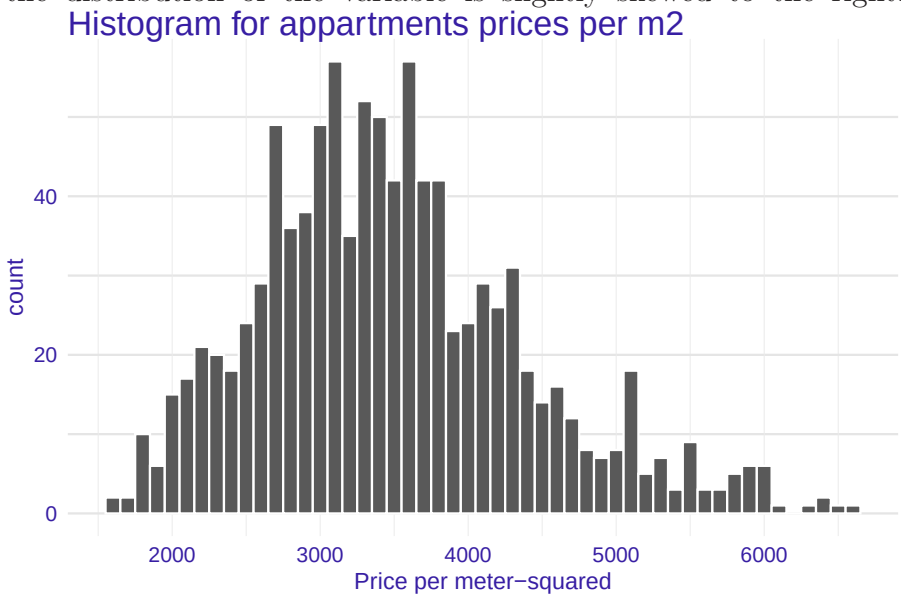


Figure 14 suggests (possibly) a nonlinear relation between *construction.year* and *m2.price* and a linear relation between *surface* and *m2.price*.

Relation between *floor* and *m2.price* is also close to linear, as well as relation between *no.rooms* and *m2.price* as seen in Figure 15.

Surface and number of rooms are correlated and prices depend on district. Boxplots plots in Figure 16 indicate that the highest prices per meter-squared are observed in Srodmiescie (Downtown).

0.5.2.2 Linear regression model

The dependent variable of interest, *m2.price*, is continuous. Thus, a natural choice to build a predictive model is the linear regression. We treat all the other variables in the `apartments` data frame as

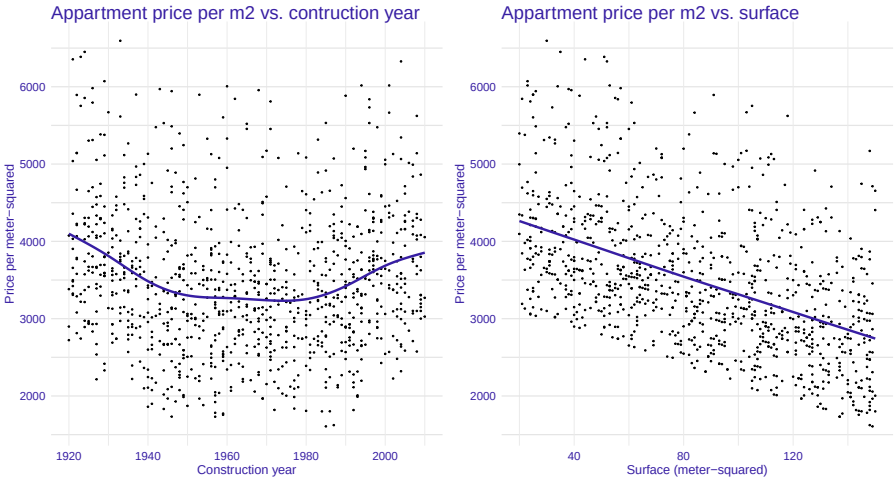


FIGURE 14 Left panel shows appartment price per m2 vs. year of construction, right panel shows price vs. square footage

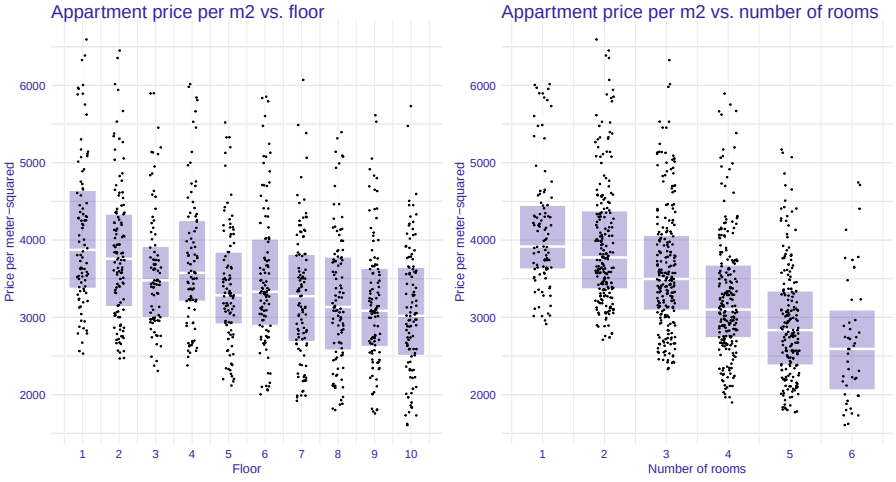


FIGURE 15 Price per meter-squared vs. floor and vs. number of rooms.

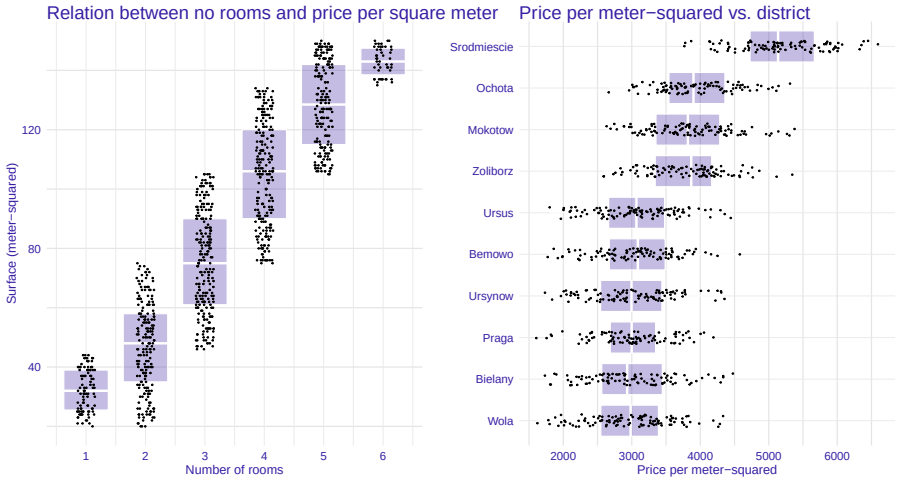


FIGURE 16 Left panel: surface vs. number of rooms. Right panel: price per meter-squared for different districts

explanatory and include them in the model. The results of the model are stored in model-object `apartments_lm_v5`.

```
apartments_lm_v5 <- lm(m2.price ~ ., data = apartments)
anova(apartments_lm_v5)
```

```
## Analysis of Variance Table
##
## Response: m2.price
##                    Df    Sum Sq   Mean Sq  F value    Pr(>F)
## construction.year   1   2629802   2629802   33.233 1.093e-08 ***
## surface             1 207840733 207840733 2626.541 < 2.2e-16 ***
## floor               1  79823027  79823027 1008.746 < 2.2e-16 ***
## no.rooms            1    956996    956996   12.094  0.000528 ***
## district            9 451993980  50221553  634.664 < 2.2e-16 ***
## Residuals         986  78023123     79131
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

0.5.2.3 Random forest model

As a challanger to linear regression, we consider a random forest model. To fit the model, we apply the `randomForest()` function, with default settings, from the package with the same name (Liaw and Wiener, 2002a).
The results of the model are stored in model-object `apartments_rf_v5`.

```
library("randomForest")
set.seed(72)
apartments_rf_v5 <- randomForest(m2.price ~ ., data = apartments)
apartments_rf_v5
```

```
##
## Call:
##  randomForest(formula = m2.price ~ ., data = apartments)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 1
##
##           Mean of squared residuals: 79789.39
##                     % Var explained: 90.28
```

0.5.2.4 Model predictions

The `predict()` function calculates predictions for a specific model. In the example below we use model-object `apartments_lm_v5` to calculate predictions for prices for first six rows.

```
predict(apartments_lm_v5, apartments_test[1:6, ])
```

```
##     1001     1002     1003     1004     1005     1006
## 4820.009 3292.678 2717.910 2922.751 2974.086 2527.043
```

In the example below we calculate predictive performance for `apartments_lm_v5` and `apartments_rf_v5` as the square root of the average of squared errors (RMSE).

```
predicted_apartments_lm <- predict(apartments_lm_v5, apartments_test)
(rmsd_lm <- sqrt(mean((predicted_apartments_lm - apartments_test$m2.price)^2)))
```

```
## [1] 283.0865
```

```
predicted_apartments_rf <- predict(apartments_rf_v5, apartments_test)
(rmsd_rf <- sqrt(mean((predicted_apartments_rf - apartments_test$m2.price)^2)))
```

```
## [1] 282.9519
```

For the random forest model, the root-mean-square of the mean squared difference is equal to 283. It is almost identical as root-mean-square for the linear regression model 283.1. Thus, the question we may face is: should we choose the more complex, but flexible random-forest model, or the simpler and easier to interpret linear model? In the subsequent chapters we will try to provide an answer to this question.

As we will show in following chapters, a proper model exploration helps to understand which model we should choose. And even more, it helps to understand weak and strong sides of both models and in consequence we can create a new model better than these two.

0.5.2.5 Model adapters

In similar spirit to the Section 0.5.1.6 we will use explainers also for predictive models created for the `apartments` dataset.

```
explain_apartments_lm_v5 <- explain(model = apartments_lm_v5,
                                    data = apartments_test,
                                    y = apartments_test$m2.price,
                                    label = "Linear Regression v5")
explain_apartments_rf_v5 <- explain(model = apartments_rf_v5,
                                    data = apartments_test,
                                    y = apartments_test$m2.price,
                                    label = "Random Forest v5")
```

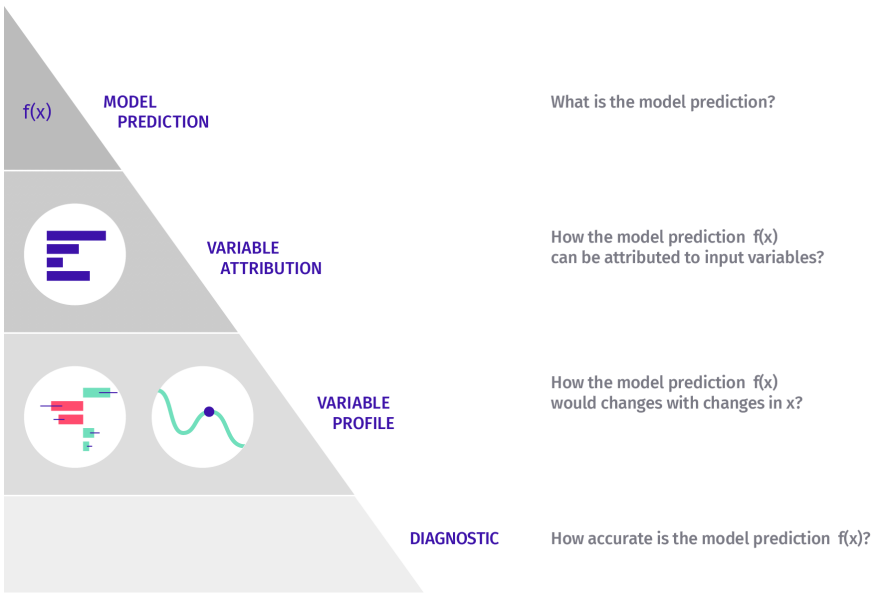
0.5.2.6 List of objects for the `apartments` example

In Sections 0.5.2.2 and 0.5.2.3 we have built two predictive models for the `apartments` data set. The models will be used in the rest of the book to illustrate the model explanation methods and tools.

For the ease of reference, we summarize the models in Table 0.3. The binary model-objects can be downloaded by using the indicated `archivist` hooks (Biecek and Kosinski, 2017). By calling a function specified in the last column of the table, one can restore a selected model in a local R environment.

TABLE 0.3: Predictive models created for the `apartments` dataset.

Model name	Model generator	Variables	Archivist hooks
`apartments_lm_v5`	`stats:: lm` v.3.5.3	construction .year, surface, floor, no.rooms, district	Get the model: `archivist:: aread("pbiecek/models/55f19")`. Get the explainer: `archivist:: aread("pbiecek/models/f49ea")`
`apartments_rf_v5`	`randomForest:: randomForest` v.4.6.14	construction .year, surface, floor, no.rooms, district	Get the model: `archivist:: aread("pbiecek/models/fe7a5")`. Get the explainer: `archivist:: aread("pbiecek/models/569b0")`



0.6 Introduction to Instance Level Exploration

Instance-level methods help to understand how a model yields a prediction for a single observation. We can think about the following situations as examples:

- We may want to evaluate the effects of explanatory variables on model predictions. For instance, we may be interested in predicting the risk of heart attack based on person's age, sex, and smoking habits. A model may be used to construct a score (for instance, a linear combination of the explanatory variables representing age, sex, and smoking habits) that could be used for the purposes of prediction. For a particular patient we may want

to learn how much the different variables contribute to the score of an individual patient?

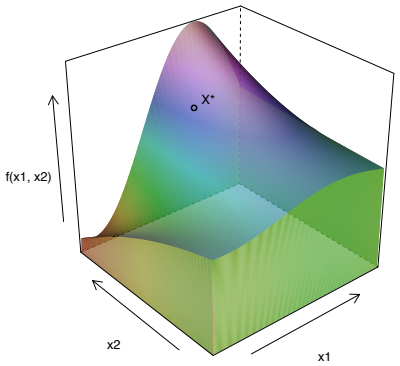
- We may want to understand how models predictions would change if values of some of the explanatory variables changed. For instance, what would be the predicted risk of heart attack if the patient cut the number of cigarettes smoked per day by half?
- We may discover that the model is providing incorrect predictions and we may want to find the reason. For instance, a patient with a very low risk-score experienced a heart attack. What has driven the wrong prediction?

A model is a function with a p-dimensional vector x as an argument. Instance level methods are designed to explore the model around a single point of interest x^*. In the following sections we will describe the most popular approaches to such exploration. They can be divided into three classes.

- One approach is to analyze how the model prediction for point x^* is different from the average model prediction and how the difference can be distributed among explanatory variables. It is often called the „variable attributions" approach. An example is provided in panel A of Figure 17. Chapters 0.7-0.9 present various methods implementing this approach.
- Another approach is to analyze the curvature of the response surface around the point of interest x^*. Treating the model as a function, we are interested in the local behavior of this function around x^*. In case of a black-box model, we may approximate it with a simpler glass-box model around x^*. An example is provided in panel B of Figure 17. Chapter 0.10 presents the Local Interpretable Model-agnostic Explanations (LIME) method that exploits the concept of a „local model".
- Yet another approach is to investigate how the model prediction changes if the value of a single explanatory variable changes. The approach is useful in the so-called „What-If" analyses. In particular, we can construct plots presenting the change in model-based predictions induced by a change of a single variable. Such plots are usually called Ceteris-paribus (CP) profiles. An example is

provided in panel A of Figure 17. Chapters 0.11-0.13 introduce the CP profiles and methods based on them.

Each method has its own merits and limitations. They are briefly discussed in the corresponding chapters. Chapter 0.14 offers a comparison of the methods.



0.7 Break-down Plots for Additive Variable Attributions

Probably the most common question related to the explanation of model prediction for a single instance is: *which variables contributed to this result the most?*

Unfortunately, there is no silver bullet. Fortunately, there are some bullets. In this chapter we introduce Break-down (BD) plots, which offer a solution to this problem. Next two chapters are related to extensions of BD plot. Finally, Chapter 0.10 offer a different approach to this problem. The goal for BD plots is to show "variables attributions" i.e., the decomposition of the model prediction among explanatory variables.

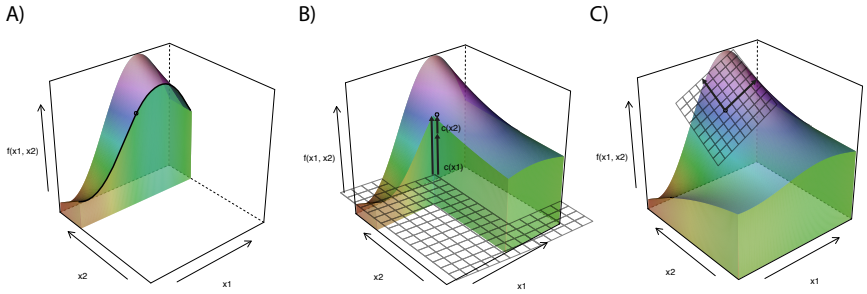


FIGURE 17 Response surface for a model that is a function of two variables. We are interested in understanding the response of a model in a single point x*. Illustration of different approaches to instance-level explanation. Panel A illustrates the concept of variable attributions like Break Down or SHAP. Additive effects of each variable show how the model response differs from the average. Panel B illustrates the concept of explanations through local models e.g. LIME. A simpler glass-box model is fitted around the point of interest. It describes the local behaviour of the black-box model. Panel C presents a What-If analysis with Ceteris-paribus profiles. The profiles show the model response as a function of a value of a single variable, while keeping the values of all other explanatory variables fixed.

0.7.1 Intuition

The underlying idea is to calculate contribution of an explanatory variable x^i to model's prediction $f(x)$ as a shift in the expected model response after conditioning on other variables.

This idea is illustrated in Figure 18. Consider an example related to the prediction for the random-forest model `model_rf_v6` for Titanic data (see Section 0.5.1.3). We are interested in chances of survival for `johny_d` - an 8-years old pasanger from first class. Panel A shows distribution of model predictions for all 2207 instances from dataset X. The row `all data` shows the vioplot of the predictions for the entire dataset. The red dot indicates the average and it is an estimate of the expected model prediction $E_X[f(X)]$ over the distribution of all explanatory variables. In this example the average model response is 23.5%.

To evaluate the contribution of the explanatory variables to the particular instance prediction, we trace changes in model predictions when fixing the values of consecutive variables. For instance, the row `class=1st` in Panel A of Figure 18 presents the distribution of the predictions obtained when the value of the `class` variable has been fixed to the `1st` class. Again, the red dot indicates the average of the predictions. The next row `age=8` shows the distribution and the average predictions with the value of variable `class` set to `1st` and `age` set to `8`, and so on. With this procedure after p steps every row in X will be filled up with variable values of `johny_d`. All predictions for these rows will be equal, so the last row in the Figure corresponds to the prediction for `model response for`johny_d‘.

The thin black lines in Panel A show how the inidividual prediction for a single person changes after the value of the j-th variable has been replaced by the value indicated in the name of the row.

As we see from lines between first and the second row, the conditioning over `class=1st` has different effect on different instances. For some the model prediction has not changes (probably these passengers were already in the 1st class). For some the model pre-

diction increase (probably they were in 2nd or 3rd class) while for other passenger the model prediction decreases (probably these were desk crew members).

Eventually, however, we may be interested in the average predictions, as indicated in Panel B of Figure 18, or even only in the changes of the averages, as shown in Panel C. In Panel C, positive changes are presented with green bars, while negative differences are marked with red bar. The changes sum up to the final prediction, which is illustrated by the violet bar at the bottom of Panel C.

What can be learned from Break-down plots? In this case we have concise summary of effects of particular variables on expected model response. First, we see that average model response is 23.5 percent. These are odds of survival averaged over all people on Titanic. Note that it is not the fraction of people that survived, but the average model response, so for different models one can get different averages. The model prediction for Johny D is 42.2 percent. It is much higher than an average prediction. Two variables that influence this prediction the most are class (=1st) and age (=8). Setting these two variables increase average model prediction by 33.5 percent points. Values in all other variables have rather negative effect. Low fare and being a male diminish odds of survival predicted by the model. Other variables do not change model predictions that much. Note that value of variable attribution depends on the value not only a variable itself. In this example the `embarked = Southampton` has small effect on average model prediction. It may be because the variable `embarked` is not important or it is possible that variable `embarked` is important but `Southampton` has an average effect out of all other possible values of the `embarked` variable.

0.7.2 Method

First, let's see how variable attribution works for linear models. Because of the simple and additive structure of linear models it will be easier to build some intuitions.

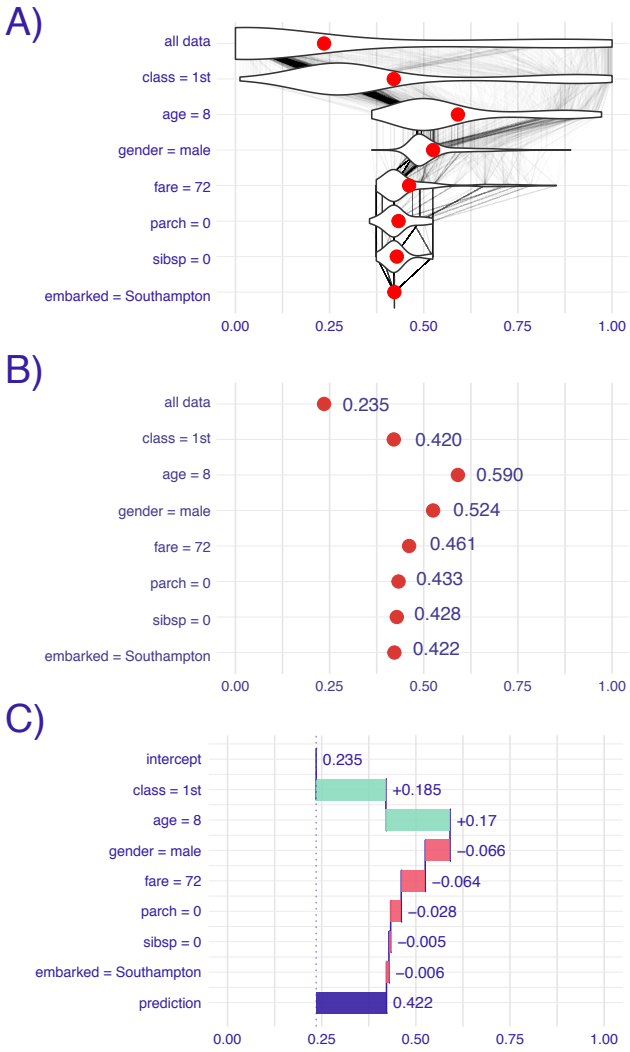


FIGURE 18 Break-down plots show how the contribution of individual explanatory variables change the average model prediction to the prediction for a single instance (observation). Panel A) The first row shows the distribution and the average (red dot) of model predictions for all data. The next rows show the dirstribution and the average of the predictions when fixing values of subseqeunt explanatory variables. The last row shows the prediction for a particular instance of interest. B) Red dots indicate the average predictions from Panel B. C) The green and red bars indicate, resspectively, positive and negative changes in the average predictions (variable contributions).

0.7.2.1 Break-down for linear models

Assume a classical linear model for response y with p explanatory variables collected in the vector $X = (X^1, X^2, \ldots, X^p)$ and coefficients $\beta = (\beta^0, \beta^1, .., \beta^p)$, where β^0 is the intercept. The prediction for y at point $x = (x^1, x^2, \ldots, x^p)$ is given by the expected value of Y conditional on $X = x$. For a linear model, the expected value is given by the following linear combination:

$$E_Y(y|x) = f(x) = \beta^0 + x^1\beta^1 + \ldots + x^p\beta^p.$$

Now assume that we selected a single point from the input space $x_* \in \mathcal{R}^p$. We are interested in the contribution of the i-th explanatory variable to model prediction $f(x_*)$ for a single observation described by x_*. Because of additive structure of the linear model we expect that this contribution will be somehow linked to $x_*^i\beta^i$, because the i-th variable occurs only in this term. As it will become clear in the sequel, it is easier to interpret the variable's contribution if x^i is centered by subtracting a constant $\hat{x}^i$ (usually, the mean of x^i). This leads the following, proposition for the variable attribution:

$$v(i, x_*) = \beta^i(x_*^i - \bar{x}^i). \tag{0.4}$$

Here $v(x_*, i)$ is the contribution of the i-th explanatory variable to the prediction of model $f()$ at point x_*. Assume that $E_Y(y|x_*) \approx f(x_*)$, where $f(x_*)$ is the value of the model at x_*. A possible approach to define $v(x_*, i)$ is to measure how much the expected model response changes after conditioning on x_*^i:

$$v(i, x_*) = E_Y(y|x_*) - E_{X^i}\{E_Y[y|(x_*^1, \ldots, x_*^{i-1}, X^i, x_*^{i+1}, x_*^p)]\} \approx f(x_*) - E_{X^i}[f(x_*^{-i})], \tag{0.5}$$

where x_*^{-i} indicates that variable X^i in vector x_* is treated as random. For the classical linear model, if the explanatory variables are independent, $v(x_*, i)$ can be expressed as follows:

$$v(i, x_*) = f(x_*) - E_{X^i}[f(x_*^{-i})] = \beta^0 + x_*^1\beta^1 + \ldots + x_*^p\beta^p - E_{X^i}[\beta^0 + x_*^1\beta^1 + \ldots + \beta^i X^i \ldots + x_*^p\beta^p] = \beta \tag{0.6}$$

In practice, given a dataset, the expected value of X_i can be estimated by the sample mean $\bar{x}_i$. This leads to

$$v(i, x_*) = \beta_i(x_*^i - \bar{x}^i). \tag{0.7}$$

Note that the linear-model-based prediction may be re-expressed in the following way:

$$f(x_*) = [\beta^0 + \bar{x}^1\beta^1 + ... + \bar{x}^p\beta^p] + [(x_*^1 - \bar{x}^1)\beta^1 + ... + (x_*^p - \bar{x}^p)\beta^p]$$

$$\equiv [average\ prediction] + \sum_{j=1}^{p} v(i, x_*). \tag{0.8}$$

Thus, the contributions of the explanatory variables $b(i, x_*)$ summ up to the difference between the model prediction for x_* and the average model prediction.

NOTE for careful readers

Obviously, sample mean $\bar{x}^i$ is an estimator of the expected value $E_{X^i}(X^i)$, calculated using a training data. For the sake of simplicity we do not emphasize these differences in the notation. Also, we ignore the fact that, in practice, we never know the true model coefficients and we work with an estimated coefficients.

0.7.2.2 Break-down for a general case

Note that the method is similar to the `EXPLAIN` algorithm introduced in (Robnik-Šikonja and Kononenko, 2008) and implemented in the `ExplainPrediction` package (Robnik-Šikonja, 2018).

Again, let $v(j, x_*)$ denote the variable-importance measure of the j-th variable and instance x_*, i.e., the contribution of the j-th variable to prediction at x_*.

We would like the sum of the $v(j, x_*)$ for all explanatory variables to be equal to the instance prediction (property called *local accuracy*), so that

$$f(x_*) = v_0 + \sum_{j=1}^{p} v(j, x_*), \tag{0.9}$$

where v_0 denotes the average model response. If we re-write the equation above as follows:

$$E_X[f(X)|X^1 = x_*^1, \ldots, X^p = x_*^p] = E_X[f(X)] + \sum_{j=1}^{p} v(j, x_*), \tag{0.10}$$

then a natural proposal for $v(j, x_*)$ is

$$v(j, x_*) = E_X[f(X)|X^1 = x_*^1, \ldots, X^j = x_*^j] - E_X[f(X)|X^1 = x_*^1, \ldots, X^{j-1} = x_*^{j-1}]. \tag{0.11}$$

In other words, the contribution of the j-th variable is the difference between the expected value of the prediction conditional on setting the values of the first j variables equal to their values in x_* and the expected value conditional on setting the values of the first $j-1$ variables equal to their values in x_*.

Note that the definition does imply the dependence of $v(j, x_*)$ on the order of the explanatory variables that is reflected in their indices.

To consider more general cases, let J denote a subset of K ($K \leq p$) indices from $\{1, 2, \ldots, p\}$, i.e., $J = \{j_1, j_2, \ldots, j_K\}$ where each $j_k \in \{1, 2, \ldots, p\}$. Furthermore, let L denote another subset of M ($M \leq p - K$) indices from $1, 2, \ldots, p$ distinct from J. That is, $L = \{l_1, l_2, \ldots, l_M\}$ where each $l_m \in \{1, 2, \ldots, p\}$ and $J \cap L = \emptyset$. Let us define now

$$\begin{aligned}\Delta^{L|J}(x_*) &\equiv E_X[f(X)|X^{l_1} = x_*^{l_1}, \ldots, X^{l_M} = x_*^{l_M}, X^{j_1} = x_*^{j_1}, \ldots, X^{j_K} = x_*^{j_K}] \\ &- E_X[f(X)|X^{j_1} = x_*^{j_1}, \ldots, X^{j_K} = x_*^{j_K}]. \end{aligned} \tag{0.12}$$

In other words, $\Delta^{L|J}(x_*)$ is the change between the expected model prediction when setting the values of the explanatory variables with indices from the set $J \cup L$ equal to their values in x_* and the expected prediction conditional on setting the values of the explanatory variables with indices from the set J equal to their values in x_*.

In particular, for the l-th explanatory variable, let

$$\begin{aligned}\Delta^{l|J}(x_*) \equiv \Delta^{\{l\}|J}(x_*) &= E_X[f(X)|X^{j_1} = x_*^{j_1}, \ldots, X^{j_K} = x_*^{j_K}, X^l = x_*^l] \\ &- E_X[f(X)|X^{j_1} = x_*^{j_1}, \ldots, X^{j_K} = x_*^{j_K}]. \end{aligned} \tag{0.14}$$

Thus, $\Delta^{l|J}$ is the change between the expected prediction when setting the values of the explanatory variables with indices from the set $J \cup \{l\}$ equal to their values in x_* and the expected prediction conditional on setting the values of the explanatory variables with indices from the set J equal to their values in x_*. Note that, if $J = \emptyset$, then

$$\Delta^{l|\emptyset}(x_*) = E_X[f(X)|X^l = x_*^l] - E_X[f(X)]. \tag{0.16}$$

It follows that

$$v(j, x_*) = \Delta^{j|\{1,\ldots,j-1\}}(x_*). \tag{0.17}$$

Unfortunately, for non-additive models (that include interactions), the value of so-defined variable-importance measure depends on the order, in which one sets the values of the explanatory variables. Figure 19 presents an example. We fit the random forest model to predict whether a passenger survived or not, then, we explain the model's prediction for a 2-year old boy that travels in the second class. The model predicts survival with a probability of 0.964. We

would like to explain this probability and understand which factors drive this prediction. Consider two explanations.

Explanation 1: The passenger is a boy, and this feature alone decreases the chances of survival. He traveled in the second class which also lower survival probability. Yet, he is very young, which makes odds higher. The reasoning behind such an explanation on this level is that most passengers in the second class are adults, therefore a kid from the second class has high chances of survival.

Explanation 2: The passenger is a boy, and this feature alone decreases survival probability. However, he is very young, therefore odds are higher than adult men. Explanation in the last step says that he traveled in the second class, which make odds of survival even more higher. The interpretation of this explanation is that most kids are from the third class and being a child in the second class should increase chances of survival.

Note that the effect of *the second class* is negative in explanations for scenario 1 but positive in explanations for scenario 2.

There are three approaches that can be used to address the issue of the dependence of $v(j, x_*)$ on the order, in which one sets the values of the explanatory variables.

In the first approach, one chooses an ordering according to which the variables with the largest contributions are selected first. In this chapter, we describe a heuristic behind this approach.

In the second approach, one identifies the interactions that cause a difference in variable-importance measure for different orderings and focuses on those interactions. This approach is discussed in Chapter 0.8.

Finally, one can calculate an average value of the variance-importance measure across all possible orderings. This approach is presented in Chapter 0.9.

To choose an ordering according to which the variables with the largest contributions are selected first, one can apply a two-step procedure. In the first step, the explanatory variables are ordered.

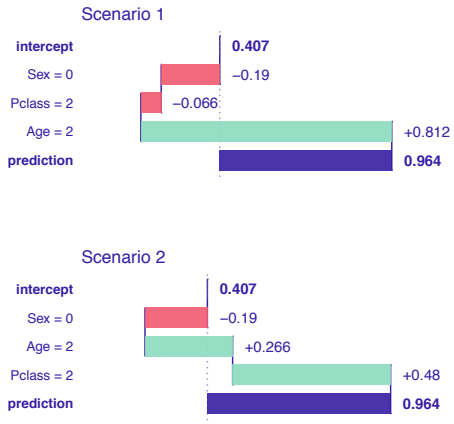


FIGURE 19 An illustration of the order-dependence of the variable-contribution values. Two *Break-down* explanations for the same observation from Titanic data set. The underlying model is a random forest. Scenarios differ due to the order of variables in *Break-down* algorithm. Blue bar indicates the difference between the model's prediction for a particular observation and an average model prediction. Other bars show contributions of variables. Red color means a negative effect on the survival probability, while green color means a positive effect. Order of variables on the y-axis corresponds to their sequence used in *Break-down* algorithm.

In the second step, the conditioning is applied according to the chosen order of variables.

In the first step, the ordering is chosen based on the decreasing value of the scores equal to $|\Delta^{k|\emptyset}|$. Note that the absolute value is needed, because the variable contributions can be positive or negative. In the second step, the variable-importance measure for the j-th variable is calculated as

$$v(j, x_*) = \Delta^{j|J},$$

where

$$J = \{k : |\Delta^{k|\emptyset}| < |\Delta^{j|\emptyset}|\},$$

that is, J is the set of indices of explanatory variables that have scores $|\Delta^{k|\emptyset}|$ smaller than the corresponding score for variable j.

The time complexity of each of the two steps of the procedure is $O(p)$, where p is the number of explanatory variables.

0.7.3 Example: Titanic data

Let us consider the random-forest model `titanic_rf_v6` (see Section 0.5.1.3 and passenger `johny_d` (see Section 0.5.1.5) as the instance of interest in the Titanic data.

The average of model predictions for all passengers is equal to $v_0 = 0.2353095$. Table 0.4 presents the scores $|\Delta^{j|\emptyset}|$ and the expected values $E[f(X|X^j = x^j_*)]$. Note that $\Delta^{j|\emptyset} = E[f(X)|X^j = x^j_*] - v_0$ and, since for all variables $E[f(X)|X^j = x^j_*] > v_0$, we have got $E[f(X|X^j = x^j_*)] = |\Delta^{j|\emptyset}| + v_0$.

TABLE 0.4: Expected values $E[f(X)|X^j = x^j_*]$ and scores $|\Delta^{j|\emptyset}|$ for the random-forest model `titanic_rf_v6` for the Titanic data and `johny_d`. The scores are sorted in the decreasing order.

variable j	$E[f(X)\|X^j = x^j_*]$	$\|\Delta^{j\|\emptyset}\|$
age	0.7407795	0.5051210

variable j	$E[f(X)\|X^j = x^j_*]$	$\|\Delta^{j\|\emptyset}\|$
class	0.6561034	0.4204449
fare	0.6141968	0.3785383
sibsp	0.4786182	0.2429597
parch	0.4679240	0.2322655
embarked	0.4602620	0.2246035
gender	0.3459458	0.1102873

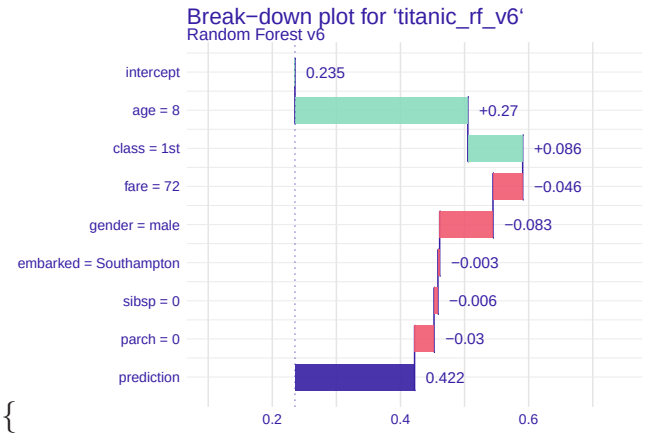
Based on the ordering defined by the scores $|\Delta^{j|\emptyset}|$ from Table 0.4, we can compute the variable-importance measures based on the sequential contributions $\Delta^{j|J}$. The computed values are presented in Table 0.5.

TABLE 0.5: Variable-importance measures $\Delta^{j|\{1,...,j\}}$ for the random-forest model `titanic_rf_v6` for the Titanic data and `johny_d` computed by using the ordering of variables defined in Table 0.4.

variable j	$E[f(X)\|X^{\{1,...,j\}} = x^{\{1,...,j\}}_*)]$	$\Delta^{j\|\{1,...,j\}}$
intercept	0.2353095	0.2353095
age = 8	0.5051210	0.2698115
class = 1st	0.5906969	0.0855759
fare = 72	0.5443561	-0.0463407
gender = male	0.4611518	-0.0832043
embarked = Southampton	0.4584422	-0.0027096
sibsp = 0	0.4523398	-0.0061024
parch = 0	0.4220000	-0.0303398
prediction	0.4220000	0.4220000

Results from Table 0.5 are presented as a waterfall plot in Figure 0.7.3.

\begin{figure}



{

}

\caption{Break-down plot for the `titanic_rf_v6` model and `johny_d` for the Titanic data.} \end{figure}

0.7.4 Pros and cons

Break-down plots offer a model-agnostic approach that can be applied to any predictive model that returns a single number for a single instance. The approach offers several advantages. The plots are easy to understand. They are compact; results for many variables may be presented in a small space. The approach reduces to an intuitive interpretation for the generalized-linear models. Numerical complexity of the Break-down algorithm is linear in the number of explanatory variables.

Break-down plots for non-additive models may be misleading, as they show only the additive contributions. An important issue is the choice of the ordering of the explanatory variables that is used in the calculation of the variable-importance measures. Also, for models with a large number of variables, the Break-down plot may be complex and include many variables with small contributions to the instance prediction.

0.7.5 Code snippets for R

In this section, we present key features of the `iBreakDown` R package (Gosiewska and Biecek, 2019a) which is a part of the `DrWhy.AI` universe. The package covers all methods presented in this chapter. It is available on CRAN and GitHub.

For illustration purposes, we use the `titanic_rf_v6` random-forest model for the Titanic data developed in Section 0.5.1.3. Recall that it is developed to predict the probability of survival from sinking of Titanic. Instance-level explanations are calculated for a single observation: `henry` - a 47-year-old passenger that travelled in the 1st class.

`DALEX` explainers for the model and the `henry` data are retrieved via `archivist` hooks as listed in Section 0.5.1.7.

```
library("randomForest")
explain_rf_v6 <- archivist::aread("pbiecek/models/9b971")

library("DALEX")
henry <- archivist::aread("pbiecek/models/a6538")
henry
```

```
##   class gender age sibsp parch fare  embarked
## 1   1st   male  47     0     0   25 Cherbourg
```

0.7.5.1 Basic use of the `break_down()` function

The `iBreakDown::break_down()` function calculates the variable-importance measures for a selected model and the instance of interest. The result of applying the `break_down()` function is a data frame containg the calculated measures. In the simplest call, the function requires only two arguments: the model explainers and the data frame for the instance of interest. The call below essentailly re-creates the variable-importance values ($\Delta^{j|\{1,...,j\}}$) presented in Table 0.5.

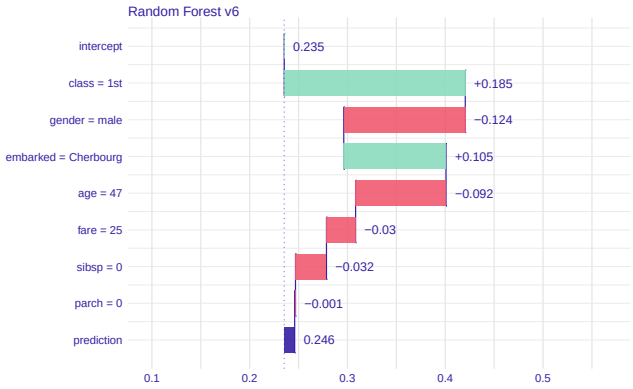


FIGURE 20 Generic plot() function for the BreakDown method calculated for 'henry'.

```
library("iBreakDown")
bd_rf <- break_down(explain_rf_v6, henry)
bd_rf
```

```
##                                          contribution
## Random Forest v6: intercept                     0.235
## Random Forest v6: class = 1st                   0.185
## Random Forest v6: gender = male                -0.124
## Random Forest v6: embarked = Cherbourg          0.105
## Random Forest v6: age = 47                     -0.092
## Random Forest v6: fare = 25                    -0.030
## Random Forest v6: sibsp = 0                    -0.032
## Random Forest v6: parch = 0                    -0.001
## Random Forest v6: prediction                    0.246
```

Applying the generic `plot()` function to the object resulting from the application of the `break_down()` function creates a BD plot. In this case, it is the plot from Figure 20.

```
plot(bd_rf)
```

Now we can compare contributions calculated for `johny_d` presented in Figure 0.7.3 with contributions calculated for `henry` presented in 20. Both explanations refer to the same model `model_rf_v6`. In both cases the `class=1st` increases chances of

survival. For `johny_d` young age increases chances of survival while for `henry` the `age=47` decreases chances of survival.

0.7.5.2 Advanced use of the `break_down()` function

The function `break_down()` allows more arguments. The most commonly used are:

- `x` - a wrapper over a model created with function `DALEX::explain()`,
- `new_observation` - an observation to be explained is should be a data frame with structure that matches the training data,
- `order` - a vector of characters (column names) or integers (column indexes) that specify order of explanatory variables that is used for computing the variable-importance measures. If not specified (default), then a one-step heuristic is used to determine the order,
- `keep_distributions` - a logical value; if `TRUE`, then additional diagnostic information about conditional distributions is stored in the resulting object and can be plotted with the generic `plot()` function.

In what follows we illustrate the use of the arguments.

First, we will specify the ordering of the explanatory variables. Toward this end we can use integer indexes or variable names. The latter option is prerferable in most cases because of transparency. Additionally, to reduce clutter in the plot, we set `max_features = 3` argument in the `plot()` function.

```
library("iBreakDown")
bd_rf_order <- break_down(explain_rf_v6,
                  henry,
                  order = c("class", "age", "gender", "fare", "parch", "sibsp", "embarked"))
plot(bd_rf_order, max_features = 3)
```

We can use the`keep_distributions = TRUE` argument to enrich the resulting object with additional information about conditional distributions. Subsequently, we can apply the `plot_distributions = TRUE` argument in the `plot()` function to present the distributions as violin plots. Red dots in the plots indicate the average model

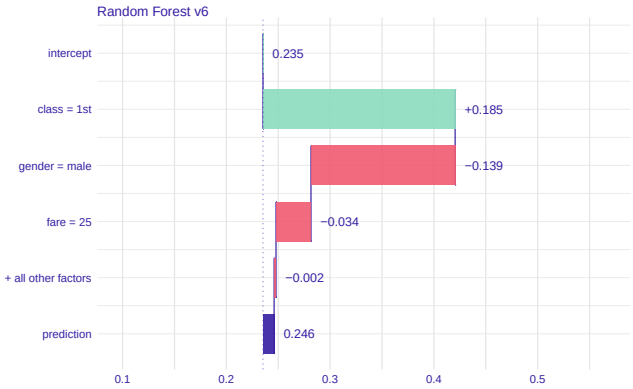


FIGURE 21 Break Down plot for top three variables.

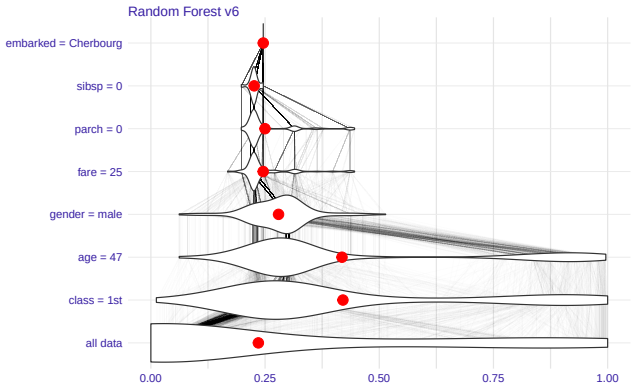


FIGURE 22 Break Down plot with distributions for a defined order of variables.

predictions. Thin black lines between violin plots correspond to predictions for individual observations. They can be used to trace how model predictions change after consecutive conditionings.

```
bd_rf_distr <- break_down(explain_rf_v6,
                          henry,
                          order = c("class", "age", "gender", "fare", "parch", "sibsp", "embarked"),
                          keep_distributions = TRUE)
plot(bd_rf_distr, plot_distributions = TRUE)
```

0.8 Break-down Plots for Models with Interactions (iBreak-down Plots)

In Chapter 0.7, we presented a model-agnostic approach to evaluation of the importance of an explanatory variable for model predictions. An important issue is that, for some models, e.g. models with interactions, the estimated value of the variable-importance measure depends on the ordering of the explanatory variables that is used when computing the measure.

In this chapter, we present an algorithm that addresses the issue. In particular, the algorithm identifies interactions between pairs of variables and takes them into account when constructing Break-down (BD) plots. In our presentation we focus on interactions that involve pairs of explanatory variables, but the algorithm can be easily extended to interactions involving a larger number of variables.

0.8.1 Intuition

Lack of additivness means that the effect of an explanatory variable depends on the value(s) of other variable(s). To illustrate such a situation, we will consider the Titanic dataset (see Section 0.5.1. For the sake of simplicity, we consider only two variables, `age` and `class`. In the data `age` is a continuous variable, but we will use a dichotomized version of it, with two levels: boys (0-16 years old) and adults (17+ years old). Also, we will consider just two classes: the 2nd and "other".

Table 0.6 shows percentages of survivors for boys and adult men in the 2nd class and other classes on Titanic. Overall, the proportion of survivors among males is 20.5%. However, among boys in the 2nd class, the proportion is 91.7%. How do age and class contribute to this higher survival probability? Let us consider the following two decompositions.

- Decomposition 1: The overall probability of survival for males is 20.5%, but for the male passengers from the 2nd class the probability is even lower, i.e. 13.5%. Thus, the effect of the 2nd class is negative, as it decreases the probability of survival by 7%. Now, if, for male passengers of the 2nd class, we consider age, we see that the survival probability for boys increases by 78.2%, from 13.5% (for a male in the 2nd class) to 91.7%. Thus, by considering first the effect of the class, and then the effect of age, we can conclude the effect of -7% for the 2nd class and +78.2% for age (being a boy).
- Decomposition 2: The overall probability of survival for males is 20.5%, but for boys the probability is higher, i.e., 40.7%. Thus, the effect of age (being a boy) is positive, as it increases the survival probability by 20.2%. On the other hand, for boys, travelling in the 2nd class increases the probability further, from 40.7% overall to 91.7%. Thus, by considering first the effect of age, and then the effect of class, we can conclude the effect of +20.2% for age (being a boy) and +51% for the 2nd class.

TABLE 0.6: Proportions of survivors for men on Titanic.

Class	Boys (0-16)	Adults (>16)	Total
2nd	11/12 = 91.7%	13/166 = 7.8%	24/178 = 13.5%
other	22/69 = 31.9%	306/1469 = 20.8%	328/1538 = 21.3%
Total	33/81 = 40.7%	319/1635 = 19.5%	352/1716 = 20.5%

By considering effects of class and age in different order, we get very different contributions. This is because there is an interaction: the effect of class depends on the age and `vice versa`. In particular, from Table 0.6 we could conclude that the overall effect of 2nd class is negative (-7%), as it decreases the probability of survival from 20.5% to 13.5%. On the other hand, the overall effect of age (being a boy) is positive (+20.2%), as it increases the probability of survival from 20.5% to 40.7%. Based on those effects, we would expect a probability of 20.5%-7%+20.2%=33.7% for a boy in the 2nd class. However, the

actually observed proportion is much higher, 90.7%. The difference of 90.7%-33.7%=57% is the interaction effect. We can interpret it as an additional effect of the 2nd class specific for boys, or as an additional effect of age (being a boy) for the 2nd class male passengers.

The example illustrates that interactions complicate the evaluation of the importance of explanatory variables to model predictions. In what follows we present an algorithm to include interactions in the BD plots.

0.8.2 Method

Identification of interactions in the model is performed in three steps (Gosiewska and Biecek, 2019a):

1. Calculate the variable-importance measure separately for each explanatory variable. In particular, for each variable, compute $\Delta^{j|\emptyset}(x_*)$ (see Section 0.7.2).
2. Calculate the measure for each pair of variables. Subtract the obtained value from the sum of the measures for the particular variables to obtain a contribution attributable to an interaction. In particular, for each pair of variables, compute $\Delta^{\{i,j\}|\emptyset}$ (see Section 0.7.2) and then

$$\Delta_I^{\{i,j\}}(x_*) \equiv \Delta^{\{i,j\}|\emptyset}(x_*) - \Delta^{i|\emptyset}(x_*) - \Delta^{j|\emptyset}(x_*). \tag{0.18}$$

3. Rank the so-obtained importance measures for the "main" and interaction effects to determine the final ordering for computing the variable-importance measures. Using the ordering, compute variable-importance measures $v(j, x_*) = \Delta^{j|\{1,\ldots,j-1\}}(x_*)$ (see Section 0.7.2).

The time complexity of the first step is $O(p)$, where p is the number of explanatory variables. For the second step, the

complexity is $O(p^2)$, while for the third step it is $O(p)$. Thus, the time complexity of the entire procedure is $O(p^2)$.

0.8.3 Example: Titanic data

Let us consider the random-forest model `titanic_rf_v6` (see Section 0.5.1.3) and passenger `johny_d` (see Section 0.5.1.5) as the instance of interest in the Titanic data.

Table 0.7 presents the expected model predictions $E_X[f(X)|X^i = x^i_*, X^j = x^j_*]$, single-variable effects $\Delta^{\{i,j\}|\emptyset}(x_*)$ (see Equation (0.16)), and interaction effects $\Delta_I^{\{i,j\}}(x_*)$ (see Equation (0.18)) for each explanatory variable and each pair of variables. All the measures are calculated for `johny_d`, the instance of interest. The rows in the table are sorted according to the absolute value of the net impact of the variable or net impact of the interaction between two variables. For a single variable the net impact is simply measured by $\Delta^{\{i,j\}}(x_*)$ while for the pairs of variables the net impact is measured by $\Delta_I^{\{i,j\}}(x_*)$. This way if two variables are important but there is no interaction, then the net effect of interaction $\Delta_I^{\{i,j\}}(x_*)$ is smaller than additive effect of each variable and the interaction will be lower in the table, see `age` and `gender`. Contrary, is the interaction is important then its net effect will be higher than each variable separately, see `fare` and `class`.

Based on the ordering of the rows, the following sequence of variables is identified as informative:

- `age` because it has largest net effect 0.270,
- then `fare:class` because the net effect of the interaction is −0.231,
- then `gender` because its net effect if 0.125 and single variables like `class` or `fare` are already used in the interaction,
- then `embarked` because of its net effect −0.011,
- then `sibsp`, and `parch` as variables with lowest net effects but still larger than effect of their interaction.

TABLE 0.7: Expected model predictions $E_X[f(X)|X^i = x^i_*, X^j = x^j_*]$, single-variable effects $\Delta^{\{i,j\}|\emptyset}(x_*)$ (see Equation (0.16)), and interaction effects $\Delta_I^{\{i,j\}}(x_*)$ (see Equation (0.18)) for the random-forest model `titanic_rf_v6` and passenger `johny_d` in the Titanic data. The rows are sorted according to the absolute value of the net impact of the variable or net impact of the interaction between two variables. For a single variable the net impact is defined as $\Delta^{\{i,j\}}(x_*)$ while for the pairs of variables the net impact is equal to $\Delta_I^{\{i,j\}}(x_*)$.

Variable	$E_X[f(X)\|X^i = x^i_*, X^j = x^j_*]$	$\Delta^{\{i,j\}\|\emptyset}(x_*)$	$\Delta_I^{\{i,j\}}(x_*)$
age	0.505	0.270	
fare:class	0.333	0.098	-0.231
class	0.420	0.185	
fare:age	0.484	0.249	-0.164
fare	0.379	0.143	
gender	0.110	-0.125	
age:class	0.591	0.355	-0.100
age:gender	0.451	0.215	0.070
fare:gender	0.280	0.045	0.027
embarked	0.225	-0.011	
embarked:age	0.504	0.269	0.010
parch:gender	0.100	-0.136	-0.008
sibsp	0.243	0.008	
sibsp:age	0.520	0.284	0.007
sibsp:class	0.422	0.187	-0.006
embarked:fare	0.374	0.138	0.006
sibsp:gender	0.113	-0.123	-0.005
fare:parch	0.380	0.145	0.005
parch:sibsp	0.236	0.001	-0.004
parch	0.232	-0.003	
parch:age	0.500	0.264	-0.002
embarked:gender	0.101	-0.134	0.002

Variable	$E_X[f(X)\|X^i = x^i_*, X^j = x^j_*]$	$\Delta^{\{i,j\}\|\emptyset}(x_*)$	$\Delta_I^{\{i,j\}}(x_*)$
embarked:parch	0.223	-0.012	0.001
fare:sibsp	0.387	0.152	0.001
embarked:class	0.409	0.173	-0.001
gender:class	0.296	0.061	0.001
embarked:sibsp	0.233	-0.002	0.001
parch:class	0.418	0.183	0.000

Table 0.8 presents the variable-importance measures computed by using the sequence of variables `age`, `fare:class`, `gender`, `embarked`, `sibsp`, and `parch`.

TABLE 0.8: Variable-importance measures $\Delta^{j|\{1,...,j\}}(x_*)$ computed by using the sequence of variables `age`, `fare:class`, `gender`, `embarked`, `sibsp`, and `parch` for the random-forest model `titanic_rf_v6` for the Titanic data and `johny_d`.

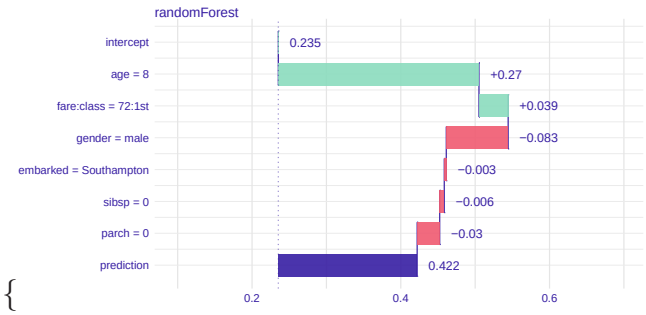
Variable	$\Delta^{j\|\{1,...,j\}}(x_*)$	$E_X[f(X)\|X^{\{1,...,j\}} = x_*^{\{1,...,j\}})]$
intercept		0.235
age = 8	0.269	0.505
fare:class = 72:1st	0.039	0.544
gender = male	-0.083	0.461
embarked = Southampton	-0.002	0.458
sibsp = 0	-0.006	0.452
parch = 0	-0.030	0.422

Figure 0.8.3 presents the BD plot corresponding to the results from Table 0.8.

As we see the interaction between `fare` and `class` is included in the plot as a single bar. As effects of these two variables cannot

be disentangled, the plot shows combination of both variables as a single contribution. From Table 0.7 we can read that `class` alone would increase average prediction by 0.185, `fare` would increase average prediction by 0.143, but together they increase the average prediction only by 0.098. It's because the `fare=72` is a high value on average, but is below median when it comes for the 1st class passengers. So these two values combined `fare:class=72:1st` signal a cheaper version of the fist class, this is why its contribution to model prediction is smaller than contribution of `class` and `fare` separately.

\begin{figure}



{

}

\caption{Break-down plot with interactions for the `titanic_rf_v6` model and `johny_d` for the Titanic data.} \end{figure}

0.8.4 Pros and cons

iBD plots share many pros and cons of BD plots for models without interactions (see section 0.7.4). However, in case of interactions, the iBD plots provide more correct explanations.

Though the numerical complexity of the iBD procedure is quadratic, it may be time-consuming in case of models with a large number of explanatory variables. If p stands for the number of variables, then we need to estimate $p * (p + 1)/2$ net effects for single variables and pair of variables. For datasets with small number of observations calculations of net effects will suffer from

larger variance and therefore larger randomness in the ranking of effects. The identification of interactions in the presented procedure is not based on a formal statistical significance test. Thus, for small sample sizes, the procedure may be prone to errors.

0.8.5 Code snippets for R

For illustration purposes, we use the `titanic_rf_v6` random-forest model for the Titanic data developed in Section 0.5.1.3. Recall that it is developed to predict the probability of survival from sinking of Titanic. Instance-level explanations are calculated for a single observation: `henry` - an 47-years old passenger that travelled in the 1st class.

`DALEX` explainers for the model and the `henry` data are retrieved via `archivist` hooks as listed in Section 0.5.1.7.

```
library("randomForest")
explain_rf_v6 <- archivist::aread("pbiecek/models/9b971")

johny_d <- archivist::aread("pbiecek/models/e3596")
henry
```

```
##   class gender age sibsp parch fare  embarked
## 1   1st   male  47     0     0   25 Cherbourg
```

The key function to construct iBD plots is the `iBreakDown::break_down()` function from the `iBreakDown` R package (Gosiewska and Biecek, 2019a). The use of the function has already been explained in Section 0.7.5. The additional necessary argument is `interactions = TRUE`.

```
library("DALEX")
library("iBreakDown")
bd_rf <- break_down(explain_rf_v6, henry, interactions = TRUE)
bd_rf
```

```
##                                          contribution
```

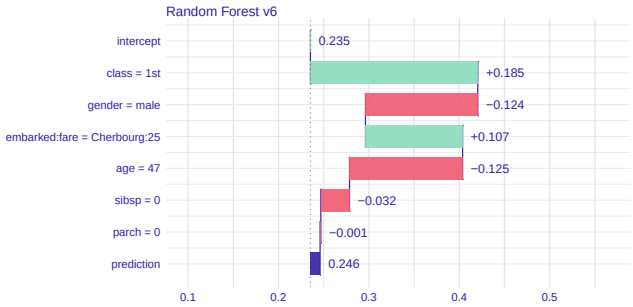


FIGURE 23 Generic plot() function for the iBreakDown method calculated for 'henry'

```
## Random Forest v6: intercept                          0.235
## Random Forest v6: class = 1st                        0.185
## Random Forest v6: gender = male                     -0.124
## Random Forest v6: embarked:fare = Cherbourg:25       0.107
## Random Forest v6: age = 47                          -0.125
## Random Forest v6: sibsp = 0                         -0.032
## Random Forest v6: parch = 0                         -0.001
## Random Forest v6: prediction                         0.246
```

Now we can plot this object with the generic `plot()` function.

```
plot(bd_rf)
```

The Figure 23 shows iBD plot for `henry` while Figure 0.8.3 shows iBD plot for `johny_d`. In this case different variables were identified as an interaction. As `fare=25` for `henry` is much lower than `fare=72` for `johny_d` effect of `class` was not modified by `fare`.

0.9 Shapley Additive Explanations (SHAP) and Average Variable Attributions

In Chapter 0.7, we introduced Break-down (BD) plots, a method of assessment of local variable-importance based on the contribution of an explanatory variable to model's prediction. We

also indicated that, in the presence of interactions, the computed value of the contribution depends on the order of explanatory covariates that is used in calculations. One solution to the problem is to find an ordering in which the most important variables are placed at the beginning. Another solution, described in Chapter 0.8, is to identify interactions and explicitly present their contributions to the predictions.

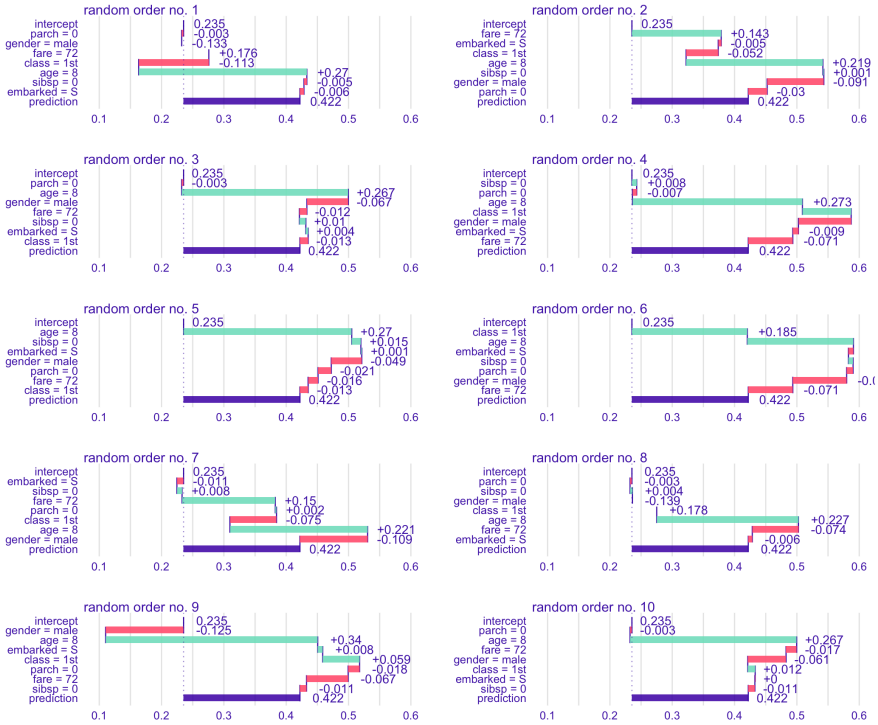
In this chapter, we introduce yet another approach to address the ordering issue. It is based on the idea of averaging the value of a variable's contribution over all, or a large number of, possible orderings. The idea is closely linked to „Shapley values" (Shapley, 1953), developed originally for cooperative games.

The approach was first introduced in (Štrumbelj and Kononenko, 2010) and (Štrumbelj and Kononenko, 2014). It was widely adopted after the publication of the 2017 paper (Lundberg and Lee, 2017) and Python's library SHAP (Lundberg, 2019). The authors of SHAP (SHapley Additive exPlanations) introduced an efficient algorithm for tree-based models (Lundberg et al., 2018). They also showed that SHAP values can be presented an unification of a collection of different commonly used techniques for model explanations (Lundberg and Lee, 2017).

0.9.1 Intuition

Figure 0.9.1 presents BD plots for ten random orderings (indicated by the order of the rows in each plot) of explanatory variables for the prediction for `johny_d` (see Section 0.5.1.5) for the random-forest model (see Section 0.5.1.3) for the Titanic dataset. The plots show clear differences in the contributions of various variables for different orderings. The most remarkable differences can be observed for variables `fare` and `class`, with contributions changing the sign depending on the ordering.

\begin{figure}



{

}

\caption{(fig:shap10orderings) Break-down plots for ten random orderings of explanatory variables for the prediction for `johny_d` for the random-forest model for the Titanic dataset. Each panel presents a single ordering, indicated by the order of the rows in the plot} \end{figure}

To remove the influence of the ordering of the variables, we can compute an average value of the contributions. Figure 24 presents the average contributions, calculated over the ten orderings presented in Figure 0.9.1. Red and green bars present, respectively, the negative and positive averages. Violet box-plots summarize the distribution of the contributions for each explanatory variable across different orderings. The plot indicates that the most important variables, from the point of view of the prediction for `johny_d` are `age` and `gender`.

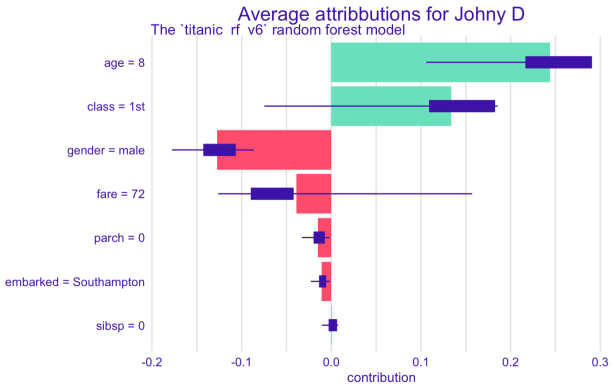


FIGURE 24 Average contributions for ten random orderings. Red and green bars present the averages. Box-plots summarize the distribution of contributions for each explanatory variable across the orderings.

0.9.2 Method

SHapley Additive exPlanations (SHAP) are based on „Shapley values," a concept in cooperative game theory developed by Lloyd Shapley (Shapley, 1953). Note that the notation may be confusing at the first glance. Shapley values are introduced for cooperative games. SHAP is an acronym for a method designed for ML models. We will use the name Shapley values.

Consider the following problem. A coalition of players cooperates, and obtains a certain overall gain from the cooperation. Players are not identical, and different players may have different importance. Cooperation is beneficial, because it may bring more benefit than individual actions. The problem to solve is how to distribute the generated surplus among the players? The Shapley value provides one possible fair answer to this question (Shapley, 1953).

Now let's translate this problem to the context of model predictions. Explanatory variables are the players, while model $f()$ plays the role of the coalition. The payoff from the coalition

is the model prediction. The problem to solve is how to distribute the model prediction across particular variables?

The idea of using Shapley values for evaluation of local variable-importance was introduced in (Štrumbelj and Kononenko, 2010). We define them here in the notation introduced in Section 0.7.2.

Let us consider a permutation J of the set of indices $\{1, 2, \ldots, p\}$ corresponding to an ordering of p explanatory variables included in model $f()$. Denote by $\pi(J, j)$ the set of the indices of the variables that are positioned in J before the j-th variable. Note that, if the j-th variable is placed as the first, then $\pi(J, j) = \emptyset$. Consider the model prediction $f(x_*)$ for a particular instance of interest x_*. The Shapley value is defined as follows:

$$\varphi(x_*, j) = \frac{1}{p!} \sum_{J} \Delta^{j|\pi(J,j)}(x_*), \tag{0.19}$$

where the sum is taken over all $p!$ possible permutations (orderings of explanatory variables) and the variable-importance measure $\Delta^{j|J}(x_*)$ was defined in Section 0.7.2. Essentially, $\varphi(x_*, j)$ is the average of the variable-importance measures across all possible orderings of explanatory variables.

It is worth noting that the value of $\Delta^{j|\pi(J,j)}(x_*)$ is constant for all ordering J that share with the same subset $\pi(J, j)$. It follows that equation (0.19) can be expressed in an alternative form:

$$\begin{aligned} \varphi(x_*, j) &= \frac{1}{p!} \sum_{s=0}^{p-1} \sum_{\substack{S \subseteq \{1,\ldots,p\} \setminus \{j\} \\ |S|=s}} \left[s!(p-1-s)! \Delta^{j|S}(x_*) \right] \\ &= \frac{1}{p} \sum_{s=0}^{p-1} \sum_{\substack{S \subseteq \{1,\ldots,p\} \setminus \{j\} \\ |S|=s}} \left[\binom{p-1}{s}^{-1} \Delta^{j|S}(x_*) \right], \end{aligned} \tag{0.20}$$

where $|S|$ denotes the cardinal number (size) of set S and the second sum is taken over all subsets S of explanatory variables, excluding the j-th one, of size s.

Note that the number of all subsets of sizes from 0 to $p-1$ is $2^p - 1$, i.e., it is much smaller than number of all permutations $p!$. Nevertheless, for a large p, it may not be feasible to compute the Shapley values from (0.19) nor (0.20). In that case, an estimate based on a sample of permutations may be considered. A Monte Carlo estimator was introduced in (Štrumbelj and Kononenko, 2014). An efficient implementation of computations of Shapley values was introduced in (Lundberg and Lee, 2017).

From the properties of Shapley values for cooperative games it follows that, in the context of predictive models, they enjoy the following properties:

- Symmetry: if two explanatory variables j and k are interchangeable, i.e., for any set of explanatory variables $S \subseteq \{1, \ldots, p\} \setminus \{j, k\}$ we have got

$$\Delta^{j|S}(x_*) = \Delta^{k|S}(x_*),$$

then their Shapley values are equal:

$$\varphi(x_*, j) = \varphi(x_*, k).$$

- Dummy feature: if an explanatory variable j does not contribute to any prediction for any set of explanatory variables $S \subseteq \{1, \ldots, p\} \setminus \{j\}$, that is,

$$\Delta^{j|S}(x_*) = 0,$$

then its Shapley value is equal to 0:

$$\varphi(x_*, j) = 0.$$

- Additivity: if model $f()$ is a sum of two other models $g()$ and $h()$, then the Shapley value calculated for model $f()$ is a sum of Shapley values for models $g()$ and $h()$.
- Local accuracy: the sum of Shapley values is equal to the model predicition, that is,

$$f(x_*) - E_X[f(X)] = \sum_{j=1}^{p} \varphi(x_*, j).$$

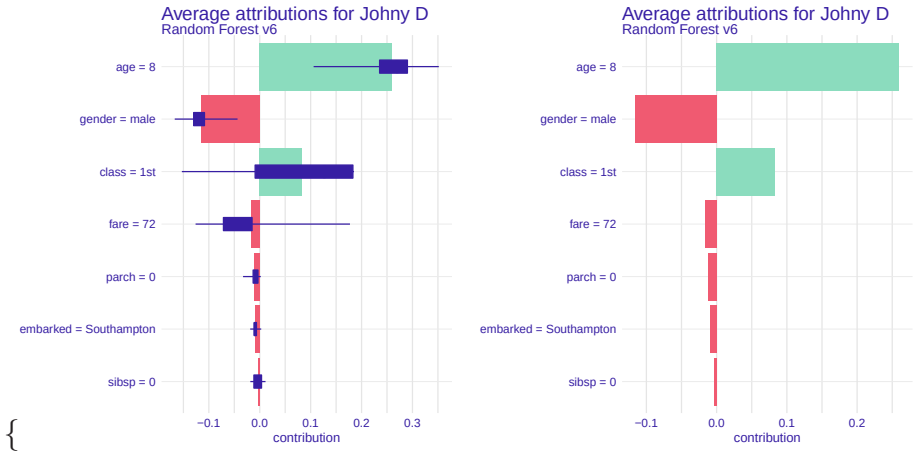
0.9.3 Example: Titanic data

Let us consider the random-forest model `titanic_rf_v6` (see Section 0.5.1.3 and passenger `johny_d` (see Section 0.5.1.5) as the instance of interest in the Titanic data.

Box-plots in Figure 0.9.3 present the distribution of the contributions $\Delta^{j|\pi(J,j)}(x_*)$ for each explanatory variable of the model for 25 random orderings of the explanatory variables. Red and green bars represent, respectively, the negative and positive Shapley values across the orderings. It is clear that the young age of Johny D results in a positive contribution for all orderings. The Shapley value is equal to 0.2525. On the other hand, the effect of gender is in all cases negative, with the Shapley value equal to -0.0908.

The picture for `fare` and `class` is more complex, as their contributions can even change the sign, depending on the ordering. While Figure 0.9.3 presents the Shapley values separately for each of the variables, it is worth noting that, by using the iBD plot in Section 0.8.3 the pair was identified as one for each an interaction effect was present. Hence, the effect of the variables should not be separated.

\begin{figure}



{

}

\caption{Variable contributions for the prediction for `johny_d` for the random-forest model `titanic_rf_v6` and the Titanic data for 25 random orderings. Left plot: Box-plots summarize the distribution of the contributions for each explanatory variable across the orderings. Red and green bars present the Shapley values. Right plot: Average attributions without boxplots.}
\end{figure}

In most applications the detailed information about the distribution of variable contributions across the considered orderings of explanatory variables will not be necessary. Thus, one could simplify the plot by presenting only the Shapley values, as in right panel in Figure 0.9.3. Table 0.9 presents the Shapley values underlying this plot.

TABLE 0.9: Shapley values for the prediction for `johny_d` for the random-forest model `titanic_rf_v6` and the Titanic data for 25 random orderings.

Variable	Shapley value
age = 8	0.2525
class = 1st	0.0246
embarked = Southampton	-0.0032
fare = 72	0.0140

Variable	Shapley value
gender = male	-0.0943
parch = 0	-0.0097
sibsp = 0	0.0027

0.9.4 Pros and cons

Shapley values provide a uniform approach to decompose model predictions into parts that can be attributed additively to different explanatory variables. In (Lundberg and Lee, 2017) it is shown that the method unifies different approaches to additive features attribution, like DeepLIFT (Shrikumar et al., 2017), Layer-Wise Relevance Propagation (Binder et al., 2016), or LIME (Ribeiro et al., 2016). The method has got a strong formal foundation derived from the cooperative games theory. It also enjoys an efficient implementation in Python, with ports or re-implementations in R.

An important drawback of the Shapley values is that they are additive attributions of variable effects. If the model is not additive, then the Shapley values may be misleading. This issue can be seen as arising from the fact that, in the cooperative games, the goal is to distribute the payoff among payers. However, in the predictive modeling context, we want to understand how do the players affect the payoff? Thus, we are not limited to independent payoff-splits for players.

It is worth noting that, for an additive model, the approaches presented in Chapters 0.7, 0.8, and in the current one lead to same variable contributions. It is because for additive models different orderings lead to same attributions. And since Shapley values can bee seen as an average across all ordering it's an average from identical values.

An important practical limitation of the method is that, for large models, the calculation of the Shapley values is time consuming. However, sub-sampling can be used to address the issue.

0.9.5 Code snippets for R

In this section, we present the key features of the `iBreakDown` R package (Gosiewska and Biecek, 2019a) which is a part of the `DrWhy.AI` universe. The package covers all methods presented in this chapter. Note that there are also other R packages that offer similar functionality, like `shapper` (Gosiewska and Biecek, 2019b), which is a wrapper for the Python library `SHAP` (Lundberg, 2019), and `iml` (Molnar et al., 2018).

For illustration purposes, we use the `titanic_rf_v6` random-forest model for the Titanic data developed in Section 0.5.1.3. Recall that it is developed to predict the probability of survival from sinking of Titanic. Instance-level explanations are calculated for a single observation: `henry` - an 42-year-old passenger that travelled in the 1st class.

`DALEX` explainers for the model and the `jonhy_d` data are retrieved via `archivist` hooks as listed in Section 0.5.1.7.

```
library("randomForest")
explain_rf_v6 <- archivist::aread("pbiecek/models/9b971")

library("DALEX")
henry <- archivist::aread("pbiecek/models/e3596")
henry
```

```
##   class gender age sibsp parch fare  embarked
## 1   1st   male  47     0     0   25 Cherbourg
```

We obtain the model prediction for this instance with the help of the 'predict()' function.

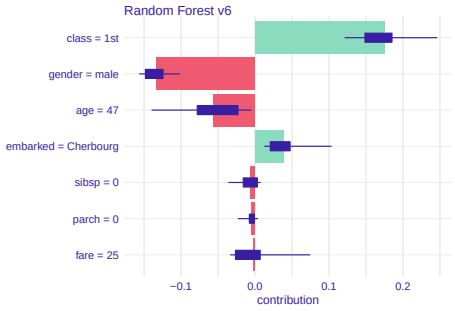
```
predict(explain_rf_v6, henry)
```

```
## [1] 0.246
```

With the help of function `shap()` from the `iBreakDown` we can re-create Figure **??**. The function is applied to the explainer, created with the `explain()` function from the `DALEX` package, and a data frame for the instance of interest. Additionally, in the `B=25`

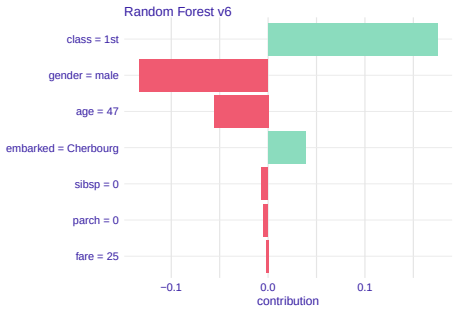
argument we indicate that we want to select 25 random orderings of explanatory variables for which the Shapley values are to be computed. The resulting object is a data frame with variable contributions computed for every ordering. Applying the generic function `plot()` to the object constructs the plot that includes the Shapley values and the corresponding box-plots.

```
library("iBreakDown")

shap_henry <- shap(explain_rf_v6, henry, B = 25)
plot(shap_henry)
```



To obtain a plot with only Shapley values, we can use the `show_boxplots=FALSE` argument in the `plot()` function call.

```
plot(shap_henry, show_boxplots = FALSE)
```



When we compare this plot with the `johny_d` in Figure 0.9.3 the largest difference is related to effect of `age`. Young `johny_d` has larger than average chances of survival, much larger than 47 years old `henry`.

The object obtained as a result of the application of function

`shap()` allows to compute other summary statistics beyond the average.

```
shap_henry
```

```
##                                                 min           q1       median
## Random Forest v6: age = 47             -0.13987947 -0.078103308 -0.030241504
## Random Forest v6: class = 1st           0.12112732  0.148666516  0.180635705
## Random Forest v6: embarked = Cherbourg  0.01245129  0.020617580  0.038341078
## Random Forest v6: fare = 25            -0.03361214 -0.026465791 -0.009642048
## Random Forest v6: gender = male        -0.15670412 -0.148171500 -0.132698686
## Random Forest v6: parch = 0            -0.02331491 -0.007689624 -0.004721142
## Random Forest v6: sibsp = 0            -0.03593203 -0.015895786 -0.002776167
##                                                mean            q3           max
## Random Forest v6: age = 47             -0.055820752 -0.0230765745 -0.004967830
## Random Forest v6: class = 1st           0.174623543  0.1851354780  0.246304486
## Random Forest v6: embarked = Cherbourg  0.038645908  0.0473429995  0.103480743
## Random Forest v6: fare = 25            -0.002075034  0.0069737200  0.074689624
## Random Forest v6: gender = male        -0.133158423 -0.1244390575 -0.101295877
## Random Forest v6: parch = 0            -0.004866842 -0.0008337109  0.004019030
## Random Forest v6: sibsp = 0            -0.006657870  0.0031898505  0.007650204
```

0.10 Local Interpretable Model-agnostic Explanations (LIME)

0.10.1 Introduction

Break-down (BD) and Shapley plots, introduced in Chapters 0.7 and 0.9, respectively, are most suitable for models with a small or moderate number of explanatory variables.

None of those approaches is well-suited for models with a very large number of explanatory variables. In genomics or image recognition, models with hundreds of thousands or milions of input variables are not uncommon. In such cases, sparse explainers with small number of non zero effects offer a useful

alternative. The most popular example of such sparse explainers are Local Interpretable Model-agnostic Explanations (LIME) and their modifications.

The LIME method was originally proposed in (Ribeiro et al., 2016). The key idea behind this method is to locally approximate a black-box model by a simpler glass-box model, which is easier to interpret. In this chapter, we describe this approach.

0.10.2 Intuition

The intuition behind the LIME method is explained in Figure 25. We want to understand factors that influence a complex black-box model around a single instance of interest. Areas presented in Figure 25 correspond to decision regions for a binary classifier, i.e., it pertains to a binary dependent variable. The axes represent the values of two continuous explanatory variables. The colored areas correspond to the decision regions, i.e., they indicate for which combinations of the variables the model classifies the observation to one of the two classes. The instance of interest is marked with the large black dot. By using an artificial dataset around the instance of interest, we can use a simpler glass-box model that will locally approximate the predictions of the black-box model. The glass-box model may then serve as a ''local explainer" for the more complex model.

We may select different classes of glass-box models. The most typical choices are regularized linear models like LASSO regression (Tibshirani, 1994) or decision trees (Hothorn et al., 2006). The important point is to limit the complexity of the models, so that they are easier to explain.

0.10.3 Method

As an explanation, we want to find a model that locally approximates a black-box model $f()$ around the instance of interest x_*. Consider class G of interpretable models (linear

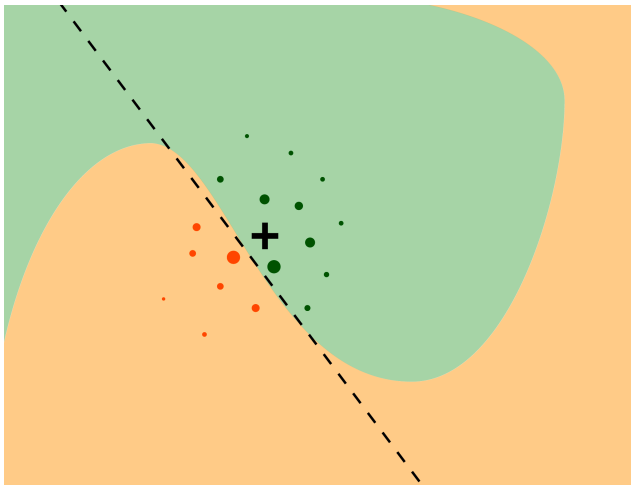


FIGURE 25 The idea behind LIME approximation with local glass-box model. The colored areas correspond to decision regions for a complex binary classification model. The black cross corresponds to the instance of interest x^*. Small dots correspond to the generated new data. Size of dots corresponds to proximity Π to the instance of interest, i.e. to weights w'. Dashed line correspond to a simple linear model fitted for the artificial data. It approximates the black box model around the instance of interest. The simple linear model „explains" local behaviour of the black box model.

moedls or decision trees). To find the required approximation, we consider the following „loss function"

$$\hat{g} = \arg\min_{g \in G} L(f, g, \Pi_{x_*}) + \Omega(g),$$

where model $g()$ belongs to class G, Π_{x_*} defines a neighborhood of x_* in which approximation is sought, $L()$ is a fidelity measure between models $f()$ and $g()$, and $\Omega(g)$ is a penalty for the complexity of model $g()$. The penalty is used to select simple models from class G.

Note that the models $f()$ and $g()$ may operate on different variable spaces. The black-box model (function) $f(x) : \mathcal{X} \to \mathcal{R}$ is defined on the original, large, p-dimensional space $\mathcal{X}$. The glass-box model (function) $g : \mathcal{X}' \to \mathcal{R}$ applies to a lower q-dmensional, interpretable space $\mathcal{X}'$, and usually $q << p$. We will present some examples of $\mathcal{X}'$ in the next section. For now we will just assume that some function $h()$ transforms $\mathcal{X}$ into $\mathcal{X}'$.

If we limit class G to sparse linear models with K non zero coefficients, the following algorithm may be used to find an interpretable glass-box model $g()$ that includes K most important, interpretable explanatory variables:

```
Input: x* - observation to be explained
Input: N  - sample size for the glass-box model
Input: K  - complexity, number of variables for the glass-box model
Input: similarity - distance function in the original input space
1. Let x' = h(x*) be a version of x* in the interpretable space
2. for i in 1...N {
3.   z'[i] <- sample_around(x')
     # prediction for a new observation z'[i]
4.   y'[i] <- f(z[i])
5.   w'[i] <- similarity(x', z'[i])
6. }
7. return K-LASSO(y', x', w')
```

In Step 7, $K - LASSO(y', x', w')$ stands for a weighted LASSO

linear-regression that selects K variables based on new dataset (y', x') with weights w'.

The practical implementation of this idea involves three important steps, which are discussed in the subsequent sub-sections.

0.10.3.1 Interpretable data representation

As it has been mentioned, the black-box model $f()$ and the glass-box model $g()$ operates on different data spaces. For example, let's consider a VGG16 neural network (Simonyan and Zisserman, 2015) trained for ImageNet data (Deng et al., 2009). The model uses an image of the size of 244×244 pixels as input and predicts to which of 1000 potential categories does the image belong to. The original data space is of dimension $3 \times 244 \times 244$ (three single-color channels *red, green, blue* for a single pixel $\times 244 \times 244$ pixels), i.e., the input space is 178,608-dimensional. Explaining predictions in such a high-dimensional space is difficult. Instead, the space can be transformed into superpixels, which are treated as binary features that can be turned on or off. Figure 26 presents an example of 100 superpixels created for an ambiguous picture. Thus, in this case the black-box model $f()$ operates in principle on data space $\mathcal{X} = R^{178,608}$, while the glass-box model $g()$ works on space $\mathcal{X}' = \{0, 1\}^{100}$.

It is worth noting that superpixels are frequent choices for image data. For text data, words are frequently used as interpretable variables. To reduce to complexity of the data space, continuous variables are often discretized to obtain interpretable tabular data. In case of categorical variables, combination of categories is often used. We will present examples in the next section.

0.10.3.2 Sampling around the instance of interest

To develop the locally-approximation glass-box model, we need new data points in the interpretable space around the instance of interest. It may not be enough to sample points from the original

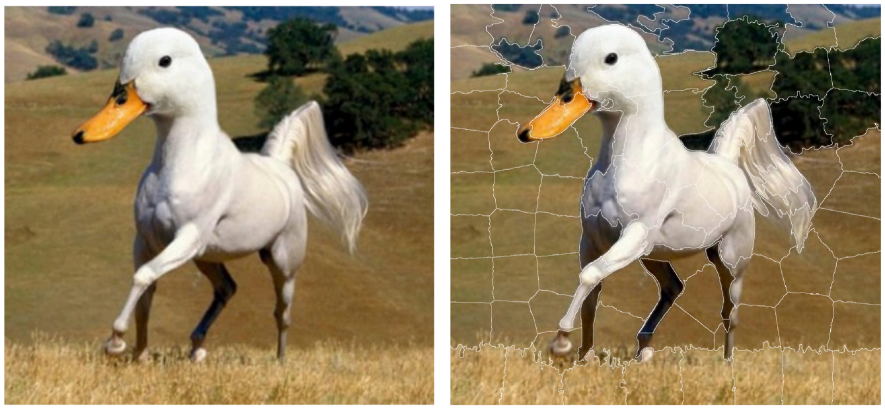


FIGURE 26 The left panel shows an ambiguous picture, half-horse and half-duck. The right panel shows 100 superpixels identified for this figure. Source: https://twitter.com/finmaddison/status/352128550704398338

dataset, because in a high-dimensional data space the data are usually very sparse and data points are „far" from each other. We need new artificial datapoints in the interpretable space. For this reason, the data for the development of the glass-box model are often created by using perturbations of the instance of interest.

For a set of binary variables in the interpretable space, the common choice is to flip (from 0 to 1 or from 1 to 0) the value of a randomly-selected number of variables describing the instance of interest.

For continuous variables, various proposals are introduced in different papers. For example (Molnar et al., 2018) and (Molnar, 2019) adds some Gaussian noise to continuous variables. In (Pedersen and Benesty, 2019) continuous variables are discretized with the use of quintiles and the perturbations are don on discretized variables. In (Staniak et al., 2019) continuous variables are discretized based on segmentation of local Ceteris Paribus profiles.

In the example of the duck-horse in Figure 26, the perturbations

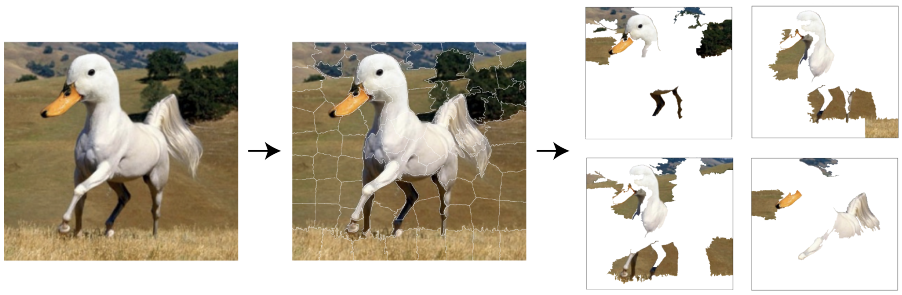


FIGURE 27 In the original input space image is described by RGB colors for each pixel (left panel). The image is transformed into the interpretable input space with 100 super pixels (central panel). The artificial data is a subset of superpixels (right panel).

of the image would be created by randomly including or excluding some of the superpixels. See an example in Figure 27.

0.10.3.3 Developing the glass-box model

Once the new data were sampled around the instance of interest, we may attempt to develop an interpretable glass-box model $g()$ from class G.

The most common choices for G are generalized linear models. To get sparse models, i.e., models with a limited number of variables, LASSO (Tibshirani, 1994) or similar regularization-modelling techniques are used. For instance, in the algorithm presented in Section 0.10.3, the K-LASSO method has been mentioned. An alternative choice are classification-and-regression trees (Breiman et al., 1984).

The VGG16 network for each picture predicts 1000 probabilities that corresponds to the 1000 classes used for training. For the duck-horse picture the two most likely classes are *'standard poodle'* and *'goose'*. Figure 28 presents LIME explanations for these top two classes. The explanations were obtained with the K-LASSO method which selected K superpixels that were the most influential from the model-prediction point of view. Here we

FIGURE 28 LIME for two predictions ('standard poodle' and 'goose') obtained by the VGG16 network with ImageNet weights for the half-duck, half-horse image.

show results for $K = 15$. For each of the selected two classes, the K superpixels with non-zero coefficients are highlighted. It is interesting to observe that the superpixel which contains the beak is influential for the prediction *'goose'*, while the superpixels linked with the white colour are influential for the prediction *'standard poodle'*. This is aligned with the intiotion thus such additional validation increses trust in model prediction.

0.10.4 Example: Titanic data

Most examples of LIME method are related to the text or image data. Here we present examples for tabular data to facilitate comparisons between methods introduced in different chapters. Let us consider the random-forest model `titanic_rf_v6` (see Section 0.5.1.3) and passenger `johny_d` (see Section 0.5.1.5) as the instance of interest in the Titanic data.

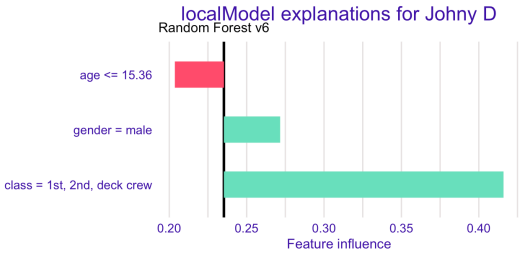
First, we need to define an interpertable input space. One option would be to gather similar variables into larger constructs

corresponding to concepts. For example `class` and `fare` variables can be combines into a concept `wealth`, `age` and `gender` into a concept `demography` and so on. In this example we have relatively small number of variables so we will use a simpler interpretable data representation in the form of a binary vector. Each variable is dychotomized into two levels. For example `age` is transformed into a binary variable `<=/>` than 15, `class` is transformed into a binary variable `1st/2nd/deck crew` and so on. The LIME algorithm is applied to this interpretable feature space and the K-LASSO method with K=3 is used to identify 3 most important variables that will be transformed into an explanation.

Once the interpretable variable space is defined, we need to transform `johny_d` to this space and generate a new dataset that will be used for K-LASSO approximations of random forest model. Figure 0.10.4 shows coefficients estimated in this K-LASSO model.

The three variables that are identified as the most influential are: `age`, `gender`, and `class`. Note that, for age, a dichotomized version of the orignially conitinuous variable is used. On the other hand, for class, a dichotomized version based on the combination of several original categories is used.

\begin{figure}



{

}

\caption{LIME method for the prediction for `johny_d` for the random-forest model `titanic_rf_v6` and the Titanic data. Presented values are beta coefficients in the K-LASSO model

fitted locally to the response from the original model.}
\end{figure}

The interpretable features can be defined in a many different ways. One idea would to be use quartiles for the feature of interest. Another idea is to use Ceteris Paribus profiles (see Chapter 0.11 and change-point method (Picard, 1985) to find a instance specific discretization. Different implementations of LIME differ in the way how the interpretable feature space is created.

0.10.5 Pros and cons

As mentioned by (Ribeiro et al., 2016), the LIME method - is *model-agnostic*, as it does not imply any assumptions on the black-box model structure, - offers an *interpretable representation*, because the original data space is transformed into a more interpretable lower-dimension space (like tranformation from individual pixels to super pixels for image data), - provides *local fidelity*, i.e., the explanations are locally well-fitted to the black-box model.

The method has been widely adopted in text and image analysis, in part due to the interpretable data representation. Also, explanations are delivered as a subset of an image/text and our brain is good in the justification of such explanations. The underlying intuition for the method is easy to understand: a simpler model is used to approximate a more complex one. By using a simpler model, with a smaller number of interpretable explanatory variabes, predictions are easier to explain. The LIME method can be applied to complex, high-dimensional models.

But there are several important limitations. For instance, despite several proposals, the issue of finding interpretable representations for continuous and categorical variables is not solved yet. Also, because the glass-box model is selected to approximate the black-box model, and the data themselves, the method does not control the quality of the local fit of the

glass-box model to the data. Thus, the latter model may be misleading.

Finally, in high-dimensional data, data points are sparse. Defining a ''local neighborhood" of the instance of interest may not be straightforward. Importance of the local neighbourhood is presented for example in the article (Alvarez-Melis and Jaakkola, 2018). Sometimes even slight changes in the neighbourhood affects strongly obtained explanations.

To summarise, the most useful applications of LIME are limited to high dimensional data for which one can defined a low-dimensional interpretable data representation, as in image analysis, text analysis or genomics.

0.10.6 Code snippets for R

LIME and similar methods are implemented in various R and Python packages. For example, `lime` (Pedersen and Benesty, 2018) is a port of the LIME Python library (Lundberg, 2019), while `live` (Staniak and Biecek, 2018), `localModel` (Staniak et al., 2019), and `iml` (Molnar et al., 2018) are separate R packages that implements this method from scratch.

Different implementations of LIME offer different algorithms for extraction of interpretable features, different methods for sampling, and different methods of weighting. For instance, regarding transformation of continuous variables into interpretable features, `lime` performs global discretization using quartiles, `localModel` performs local discretization using CP profiles, while `live` and `iml` work directly on continuous variables. Due to these differences, the packages yield different results (explanations).

In what follows, for illustration purposes, we use the `titanic_rf_v6` random-forest model for the Titanic data developed in Section 0.5.1.3. Recall that it is developed to predict the probability of survival from sinking of Titanic. Instance-level explanations are calculated for a single observation: `johny_d` - an

8-year-old passenger that travelled in the 1st class. DALEX explainers for the model and the johny_d data are retrieved via archivist hooks as listed in Section 0.5.1.7.

```
library("DALEX")
library("randomForest")

titanic <- archivist::aread("pbiecek/models/27e5c")
titanic_rf_v6 <- archivist::aread("pbiecek/models/31570")
johny_d <- archivist::aread("pbiecek/models/e3596")
```

0.10.6.1 The lime package

The key elements of the lime package are functions lime(), which creates an explainer, and explain(), which evaluates explanations.

The detailed results for the titanic_rf_v6 random-forest model and johny_d are presented below. First we need to specifiy that we will work with a model for classification.

```
library("lime")
model_type.randomForest <- function(x, ...) "classification"
```

Second we need to create an explainer - an object with all elements needed for calculation of explanations. This can be done with the lime function, the dataset and the model.

```
lime_rf <- lime(titanic[,colnames(johny_d)], titanic_rf_v6)
```

In the last step we generate an explanation. The n_features set the K for K-LASSO method. Here we ask for explanations not larger than 4 variables. The n_permutations argument defines how many points are to be sampled for a local model approximation. Here we use a set of 1000 artificial points for this.

```
lime_expl <- lime::explain(johny_d, lime_rf, labels = "yes",
                           n_features = 4, n_permutations = 1000)
lime_expl

#      model_type case label label_prob  model_r2 model_intercept model_prediction
```

```
#1 classification    1    no      0.602 0.5806297       0.5365448       0.5805939
#2 classification    1    no      0.602 0.5806297       0.5365448       0.5805939
#3 classification    1    no      0.602 0.5806297       0.5365448       0.5805939
#4 classification    1    no      0.602 0.5806297       0.5365448       0.5805939
#  feature feature_value feature_weight  feature_desc                 data   prediction
#1    fare            72     0.00640936  21.00 < fare 1, 2, 8, 0, 0, 72, 4 0.602, 0.398
#2  gender             2     0.30481181 gender = male 1, 2, 8, 0, 0, 72, 4 0.602, 0.398
#3   class             1    -0.16690730   class = 1st 1, 2, 8, 0, 0, 72, 4 0.602, 0.398
#4     age             8    -0.10026475     age <= 22 1, 2, 8, 0, 0, 72, 4 0.602, 0.398
```

In this table the `feature_weight` column has coefficients for the K-LASSO method in the explanation. In the column `case` one will find an index of observation for which the explanation is calculated. Here it's 1 since we asked for explanation for only one observation. The `feature_weight` columns shows the β coefficients in the K-LASSO model, `feature` column points out which variables have non zero coefficients in the K-LASSO method. The `feature_value` column denotes values for the selected features for the observation of interest. The `feature_description` column shows how the original feature was transformed into a interpretable feature.

This implementation of the LIME method dichotomizes continuous variables by using quartiles. Hence, in the output we get a binary variable `age < 22`.

The corresponding local white box model is

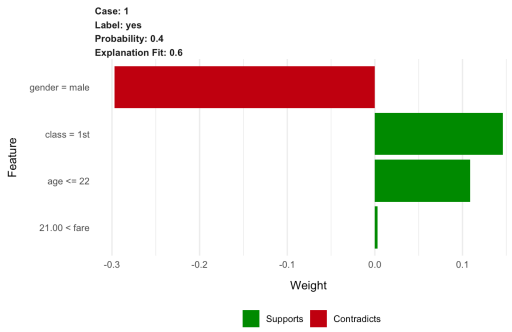
$$\hat{y} = 0.00640936 * 1_{fare>21} + 0.30481181 * 1_{gender=male} - 0.16690730 * 1_{class=1st} - 0.10026475 * 1_{age<22}$$

Figure 0.10.6.1 shows the graphical presentation of the results, obtained by applying the generic `plot()` function.

Color corresponds to the sign of the β coefficient while length of the bar corresponds to the absolute value of β coefficient in the K-LASSO method.

```
plot_features(lime_expl)
```

\begin{figure}



{

}

\caption{(fig:limeExplLIMETitanic) LIME-method results for the prediction for `johny_d` for the random-forest model `titanic_rf_v6` and the Titanic data, generated by the `lime` package. } \end{figure}

0.10.6.2 The localModel package

The `localModel` package operates on `DALEX::explain()` object. The main function in this package is `individual_surrogate_model()` which trains the local glass-box model.

The detailed results for the `titanic_rf_v6` random-forest model and `johny_d` are presented below.

```
library("localModel")

explainer_titanic_rf <- DALEX::explain(model = titanic_rf_v6,
                      data = titanic[,colnames(johny_d)])
local_model_rf <- individual_surrogate_model(explainer_titanic_rf, johny_d,
                                          size = 1000, seed = 1313)
local_model_rf
#   estimated                    variable dev_ratio response
#1 0.23479837                (Model mean) 0.6521442
```

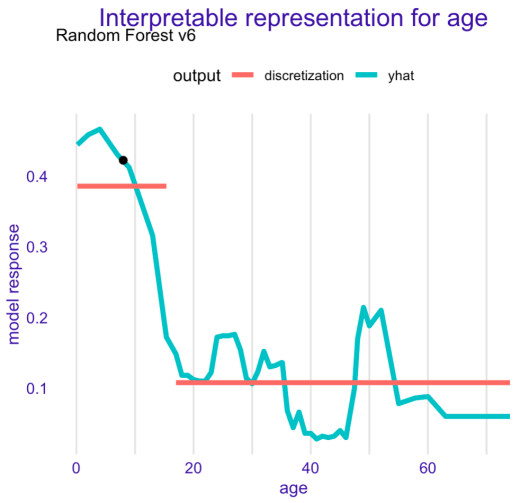


FIGURE 29 Interpretable instance-level discretisation of age variable. Based on the Ceteris Paribus profiles we may estimate an optimal change-point as 15 years.

```
#2 0.14483341                (Intercept) 0.6521442
#3 0.08081853 class = 1st, 2nd, deck crew 0.6521442
#4 0.00000000     gender = female, NA, NA 0.6521442
#5 0.23282293               age <= 15.36 0.6521442
#6 0.02338929               fare > 31.05 0.6521442
```

In the column `estimated` one will find β coefficients for LASSO logistic regression while in the `variable` column one will find corresponding values.

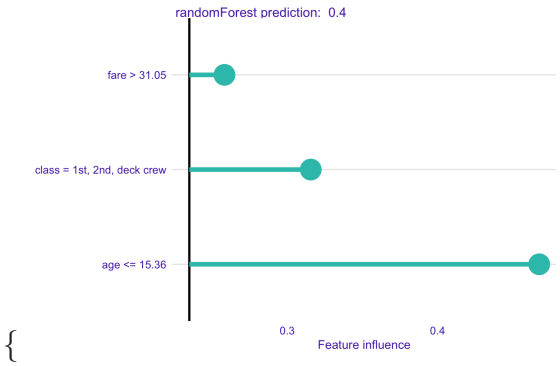
The implemented version of LIME dichotomizes continuous variables by using CP profiles. The CP profile for `johny_d`, presented in Figure 0.11.6.2 in Chapter 0.11, indicated that, for age, the largest drop in the predicted probability of survival was observed for the age increasing beyond 15 years. Hence, in the output of the `individual_surrogate_model()`, we see a binary variable `age < 15.36`.

Figure 29 illustrates how the two levels for age can be extracted from the Ceteris Paribus profile.

The graphical presentation of the results, obtained by applying the generic `plot()` function is provided in Figure 0.10.6.2. Bars correspond to β coefficients in the LASSO model.

```
plot(local_model_rf)
```

\begin{figure}



{

}

\caption{LIME-method results for the prediction for `johny_d` for the random-forest model `titanic_rf_v6` and the Titanic data, generated by the `localModel` package. } \end{figure}

0.10.6.3 The iml package

The key elements of the `iml` package are functions `Predictor$new()`, which creates an explainer, and `LocalModel$new()`, which develops the local glass-box model.

The detailed results for the `titanic_rf_v6` random-forest model and `johny_d` are presented below.

```
library("iml")
iml_rf = Predictor$new(titanic_rf_v6, data = titanic[,colnames(johny_d)])
iml_glass_box = LocalModel$new(iml_rf, x.interest = johny_d, k = 6)
iml_glass_box
#Interpretation method:  LocalModel
#
#Analysed predictor:
```

```
#Prediction task: unknown
#
#Analysed data:
#Sampling from data.frame with 2207 rows and 7 columns.
#
#Head of results:
#           beta x.recoded     effect  x.original              feature
#1 -0.158368701         1 -0.1583687         1st            class=1st
#2  1.739826204         1  1.7398262        male          gender=male
#3  0.018515945         0  0.0000000           0                sibsp
#4 -0.001484918        72 -0.1069141          72                 fare
#5  0.131819869         1  0.1318199 Southampton embarked=Southampton
#6  0.158368701         1  0.1583687         1st            class=1st
```

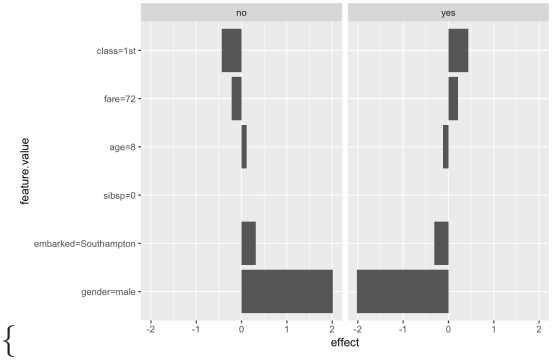
In the `effect` column on can read β coefficients for the LASSO method.

The implemented version of LIME does not transform continuous variables. The CP profile for `johny_d`, presented in Figure 0.11.6.2 in Chapter 0.11, indicated that, for boys younger than 15-year-old, the predicted probability of survival did not change very much. Hence, in the printed output, age does not appear as an important variable.

The graphical presentation of the results, obtained by applying the generic `plot()` function to the object resulting from the `application of the`explain()‘ function, is provided in Figure 0.10.6.3. Note that only first 6 rows are listed in the table above. The whole table has 12 coefficients that corresponds to bars in the plot.

```
plot(iml_glass_box)
```

\begin{figure}



{

}

\caption{(fig:limeExplIMLTitanic) LIME-method results for the prediction for `johny_d` for the random-forest model `titanic_rf_v6` and the Titanic data, generated by the `iml` package. }
\end{figure}

0.11 Ceteris-paribus Profiles and What-If Analysis

0.11.1 Introduction

Chapters 0.7 – 0.10 are related to the decomposition of a prediction $f(x)$ into parts linked with particular variables. In this chapter we focus on methods that analyse an effect of selected variables on model response. These techniques may be used for sensitivity analysis, how stable is the model response, or to what-if analysis, how model response would change if input is changes. It is important to remember that the what-if analysis is performed not in the sense of causal modeling, but in the sense of model exploration. We need causal model to do causal inference for the real-world phenomena. Here we focus on explanatory analysis of the model behaviour. To show the difference between these two things, think about a model for survival for lung-cancer patients based on some treatment parameters. We need causal model to say how the survival would change if the treatment is

changed. Techniques presented in this chapter will explore how the model result will change if the treatment is changed.

Ceteris paribus is a Latin phrase meaning "other things held constant" or "all else unchanged." In this chapter, we introduce a technique for model exploration based on the *Ceteris paribus* principle. In particular, we examine the influence of each explanatory variable, assuming that effects of all other variables are unchanged. The main goal is to understand how changes in a single explanatory variable affects model predictions.

Explanation tools (explainers) presented in this chapter are linked to the second law introduced in Section 0.1.3, i.e. the law of "Prediction's speculation." This is why the tools are also known as *What-If model analysis* or *Individual Conditional Expectations* (Goldstein et al., 2015). It appears that it is easier to understand how a black-box model is working if we can explore the model by investigating the influence of explanatory variables separately, changing one at a time.

0.11.2 Intuition

Ceteris-paribus profiles show how the model response would change if a single variable is changed. For example, panel A of Figure 30 presents response (prediction) surface for the `titanic_lmr_v6` model for two explanatory variables, *age* and *class*, from the *titanic* dataset (see Section 0.5.1). We are interested in the change of the model prediction induced by each of the variables. Toward this end, we may want to explore the curvature of the response surface around a single point with *age* equal to 47 and *class* equal to "1st," indicated in the plot. Ceteris-paribus (CP) profiles are one-dimensional profiles that examine the curvature across each dimension, i.e., for each variable. Panel B of Figure 30 presents the CP profiles corresponding to *age* and *class*. Note that, in the CP profile for *age*, the point of interest is indicated by the black dot. In essence, a CP profile shows a conditional expectation of the dependent variable (response) for the particular explanatory variable.

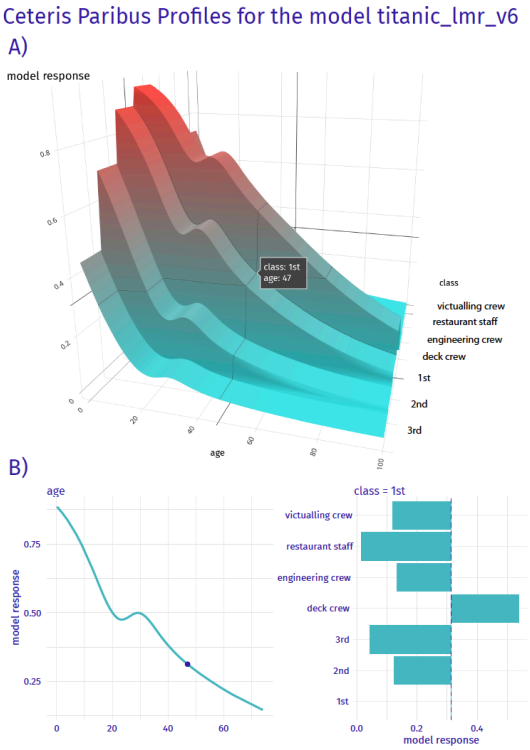


FIGURE 30 Panel A) Model response (prediction) surface. Ceteris-paribus (CP) profiles marked with black curves help to understand the curvature of the surface while changing only a single explanatory variable. Panel B) CP profiles for individual variables, age (continuous) and class (categorical).

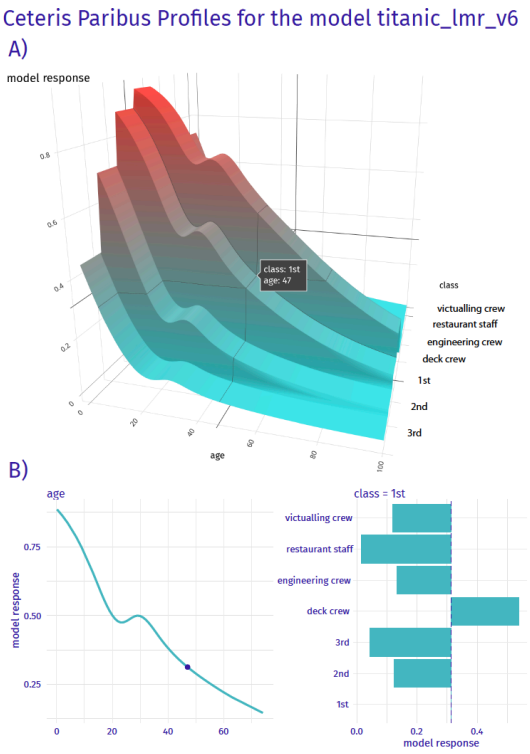


FIGURE 31 Animated model response for 2D surface as in ef(fig:modelResponseCurveLine).

CP belongs to the class of techniques that examine local curvature of the model response surface. Other very popular technique from this class called LIME is presented in Chapter 0.10. The difference between these two methods lies in the fact that LIME approximates the model of interest locally with a simpler glass-box model. Usually, the LIME model is sparse, i.e., contains fewer explanatory variables. Thus, one needs to investigate a plot across a smaller number of dimensions. On the other hand, the CP profiles present conditional predictions for a single variable and, in most cases, are easier to interpret. More detailed comparison of these techniques is presented in the Chapter 0.14.

0.11.3 Method

In this section, we introduce more formally one-dimensional CP profiles.

Recall (see Section 0.2.3) that we use x_i to refer to the vector corresponding to the i-th observation in a dataset. Let x_*^j denote the j-th element of x_*, i.e., the j-th explanatory variable. We use x_*^{-j} to refer to a vector resulting from removing the j-th element from x_*. By $x_*^{j|=z}$, we denote a vector resulting from changing the value of the j-th element of x_* to (a scalar) z.

We define a one-dimensional CP profile $h()$ for model $f()$, the j-th explanatory variable, and point x_* as follows:

$$h_{x_*}^{f,j}(z) := f(x_*|^j = z).$$

CP profile is a function that provides the dependence of the approximated expected value (prediction) of Y on the value z of the j-th explanatory variable. Note that, in practice, z is taken to go through the entire range of values typical for the variable, while values of all other explanatory variables are kept fixed at the values specified by x_*.

Note that in the situation when only a single model is considered,

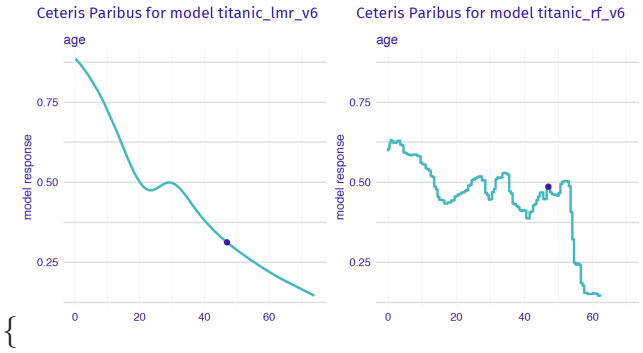
we will skip the model index and we will denote the CP profile for the j-th explanatory variable and the point of interest x_* by $h^j_{x_*}(z)$.

0.11.4 Example: Titanic

For continuous explanatory variables, a natural way to represent the CP function is to use a profile plot similar to the ones presented in Figure 0.11.4. In the figure, the dot on the curves marks an instance prediction, i.e., prediction $f(x_*)$ for a single observation x_*. The curve itself shows how the prediction would change if the value of a particular explanatory variable changed.

Figure 0.11.4 presents CP profiles for the *age* variable in the logistic regression `titanic_lmr_v6` and random forest model `titanic_rf_v6` for the Titanic dataset (see Sections 0.5.1.2 and 0.5.1.3, respectively). It is worth observing that the profile for the logistic regression model is smooth, while the one for the random forest model shows more variability. For this instance (observation), the prediction for the logistic regression model would increase substantially if the value of *age* became lower than 20. For the random forrest model, a substantial increase would be obtained if *age* became lower than 13 or so.

\begin{figure}



{

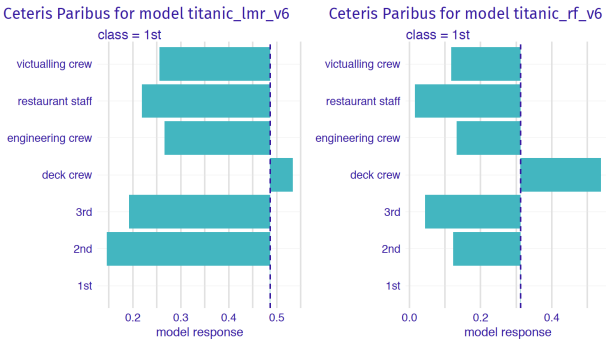
}

\caption{Ceteris-paribus profiles for variable `age` for the logistic

regression (`titanic_lmr_v6`) and random forest (`titanic_rf_v6`) models that predict the probability of surviving based on the Titanic data. Black dotcorresponds to the passanger `johny_d`.} \end{figure}

For a categorical explanatory variable, a natural way to represent the CP function is to use a barplot similar to the ones presented in Figure 0.11.4. The barplots in Figure 0.11.4 present CP profiles for the *class* variable in the logistic regression and random forest models for the Titanic dataset (see Sections 0.5.1.2 and 0.5.1.3, respectively). For this instance (observation), the predicted probability for the logistic regression model would decrease substantially if the value of *class* changed to "2nd". On the other hand, for the random forest model, the largest change would be marked if *class* changed to "restaurant staff".

\begin{figure}



{

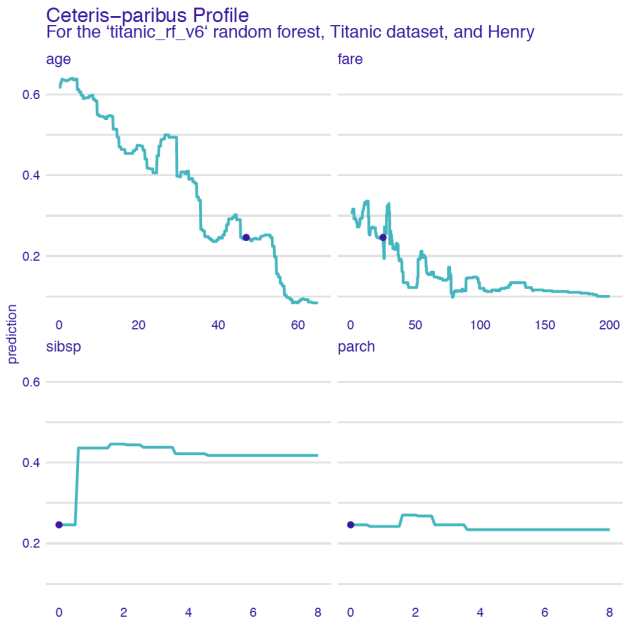
}

\caption{Ceteris-paribus profiles for variable `class` for the logistic regression (`titanic_lmr_v6`) and random forest (`titanic_rf_v6`) models that predict the probability of surviving based on the Titanic data} \end{figure}

Usually, black-box models contain a large number of explanatory variables. However, CP profiles are legible even for tiny subplots, created with techniques like sparklines or small multiples (Tufte, 1986). In this way we can display a large number of profiles at the same time keeping profiles for consecutive variables in separate

panels, as shown in Figure 0.11.4 for the random forest model for the Titanic dataset. It helps if these panels are ordered so that the most important profiles are listed first. We discuss a method to assess the importance of CP profiles in the next chapter.

\begin{figure}



{

}

\caption{Ceteris-paribus profiles for all continuous explanatory variables for the random forest (`titanic_rf_v6`) model for the `titanic` dataset} \end{figure}

0.11.5 Pros and cons

One-dimensional CP profiles, as presented in this chapter, offer a uniform, easy to communicate and extendable approach to model exploration. Their graphical representation is easy to understand and explain. It is possible to show profiles for many variables or models in a single plot. CP profiles are easy to compare, thus we can juxtapose two or more models to better understand

differences between models. We can also compare two or more instances to better understand model stability. CP profiles are also a useful tool for sensitivity analysis.

But. There are several issues related to the use of the CP profiles. If explanatory variables are correlated, then changing one variable implies a change in the other. In such case, the application of the *Ceteris paribus* principle may lead to unrealistic settings, as it is not possible to keep one variable fixed while varying the other one. For example, apartment's price prediction features like surface and number of rooms are correlated thus it is unrealistic to consider very small apartments with extremly large number of rooms. Special cases are interactions, which require the use of two-dimensional CP profiles that are more complex than one-dimensional ones. Also, in case of a model with hundreds or thousands of variables, the number of plots to inspect may be daunting. Finally, while barplots allow visualization of CP profiles for factors (categorical explanatory variables), their use becomes less trivial in case of factors with many nominal (unordered) categories (like, for example, a ZIP-code).

0.11.6 Code snippets for R

In this section, we present key features of the R package `ingredients` (Biecek et al., 2019) which is a part of `DrWhy.AI` universe and covers all methods presented in this chapter.

Note that there are also other R packages that offer similar functionality, like `condvis` (O'Connell et al., 2017), `pdp` (Greenwell, 2017), `ICEbox` (Goldstein et al., 2015), `ALEPlot` (Apley, 2018), `iml` (Molnar et al., 2018).

For illustration, we use two classification models developed in Chapter 0.5.1, namely the logistic regression model `titanic_lmr_v6` (Section 0.5.1.2) and the random forest model `titanic_rf_v6` (Section 0.5.1.3). They are developed to predict the probability of survival after sinking of Titanic. Instance-level explanations are

calculated for a single observation `henry` - a 47 years old male passenger that travelled in the 1st class.

`DALEX` explainers for both models and the `henry` data frame are retrieved via the `archivist` hooks as listed in Section 0.5.1.7.

```
library("rms")
explain_lmr_v6 <- archivist::aread("pbiecek/models/2b9b6")

library("randomForest")
explain_rf_v6 <- archivist::aread("pbiecek/models/9b971")

library("DALEX")
henry <- archivist::aread("pbiecek/models/a6538")
henry
```

```
##   class gender age sibsp parch fare  embarked
## 1   1st   male  47     0     0   25 Cherbourg
```

0.11.6.1 Basic use of the `ceteris_paribus` function

The easiest way to create and plot CP profiles is to call `ceteris_paribus()` function and then the generic `plot()` function. By default, profiles for all variables are being calculated and all numeric features are being plotted. One can limit the number of variables that should be considered with the `variables` argument.

To obtain CP profiles, the `ceteris_paribus()` function requires the explainer-object and the instance data frame as arguments. As a result, the function yields an object od the class `ceteris_paribus_explainer`. It is a data frame with model predictions.

```
library("ingredients")
cp_titanic_rf <- ceteris_paribus(explain_rf_v6, henry)
cp_titanic_rf
```

```
## Top profiles    :
##               class gender age sibsp parch fare  embarked _yhat_ _vname_ _ids_
## 1               1st   male  47     0     0   25 Cherbourg  0.246   class     1
```

```
## 1.1              2nd   male  47     0     0   25 Cherbourg  0.054   class     1
## 1.2              3rd   male  47     0     0   25 Cherbourg  0.100   class     1
## 1.3        deck crew   male  47     0     0   25 Cherbourg  0.454   class     1
## 1.4 engineering crew   male  47     0     0   25 Cherbourg  0.096   class     1
## 1.5 restaurant staff   male  47     0     0   25 Cherbourg  0.092   class     1
##              _label_
## 1   Random Forest v6
## 1.1 Random Forest v6
## 1.2 Random Forest v6
## 1.3 Random Forest v6
## 1.4 Random Forest v6
## 1.5 Random Forest v6
##
##
## Top observations:
##   class gender age sibsp parch fare  embarked _yhat_          _label_ _ids_
## 1   1st   male  47     0     0   25 Cherbourg  0.246 Random Forest v6     1
```

To obtain a graphical representation of CP profiles, the generic `plot()` function can be applied to the data frame returned by the `ceteris_paribus()` function. It returns a `ggplot2` object that can be processed further if needed. In the examples below, we use the `ggplot2` functions, like `ggtitle()` or `ylim()`, to modify plot's title or the range of the Y-axis.

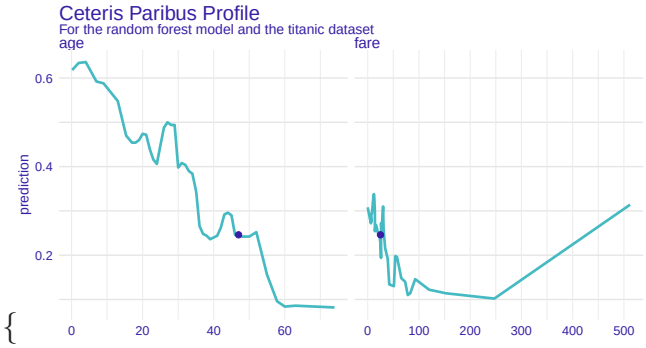
The resulting plot can be enriched with additional data by applying functions `ingredients::show_rugs` (adds rugs for the selected points), `ingredients::show_observations` (adds dots that shows observations), or `ingredients::show_aggreagated_profiles`. All these functions can take additional arguments to modify size, color, or linetype.

Below we show an R snippet that can be used to replicate plots presented in the upper part of Figure 0.11.4.

```
library("ggplot2")
plot(cp_titanic_rf, variables = c("age", "fare")) +
  show_observations(cp_titanic_rf, variables = c("age", "fare")) +
```

```
  ggtitle("Ceteris Paribus Profile",
          "For the random forest model and the titanic dataset")
```

\begin{figure}



{

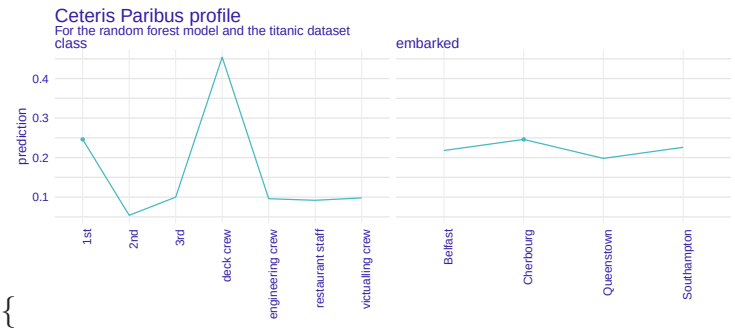
}

\caption{Ceteris-paribus profiles for `age` and `fare` variables and the `titanic_rf_v6` model.} \end{figure}

By default, all numerical variables are plotted. To plot CP profiles for categorical variables, we have got to add the `variable_type = "categorical"` argument to the `plot()` function. The code below an be used to recreate the right-hand-side plot from Figure 0.11.4.

```
plot(cp_titanic_rf, variables = c("class", "embarked"), variable_type = "categorical") +
  ggtitle("Ceteris Paribus profile",
          "For the random forest model and the titanic dataset")
```

\begin{figure}



{

}

\caption{Ceteris-paribus profiles for `class` and `embarked` variables and the `titanic_rf_v6` model.} \end{figure}

0.11.6.2 Advanced use of the `ceteris_paribus` function

The `ceteris_paribus()` is a very flexible function. To better understand how it can be used, we briefly review its arguments.

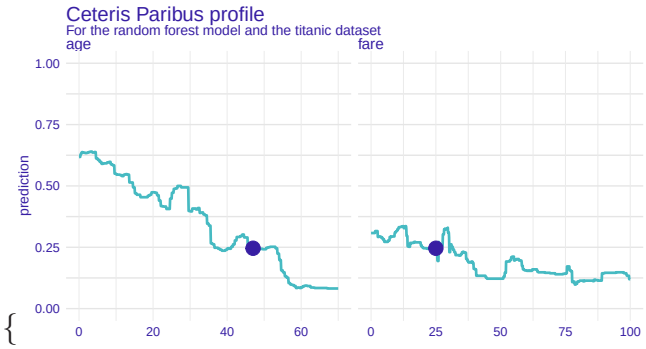
- `x`, `data`, `predict_function`, `label` - information about a model. If `x` is created with the `DALEX::explain` function, then other arguments are extracted from `x`; this is how we use the function in this chapter. Otherwise, we have got to specify directly the model, the validation data, the predict function, and the model label.
- `new_observation` - instance (one or more), for which we want to calculate CP profiles. It should be a data frame with same variables as in the validation data.
- `y` - observed value of the dependent variable for `new_observation`. The use of this argument is illustrated in Section 0.13.1.
- `variables` - names of explanatory variables, for which CP profiles are to be calculated. By default, the profiles will be constructed for all variables, which may be time consuming.
- `variable_splits` - a list of values for which CP profiles are to be calculated. By default, these are all values for categorical variables. For continuous variables, uniformly-placed values are selected; one can specify the number of the values with the `grid_points` argument (the default is 101).

The code below allows to obtain the plots in the upper part of Figure 0.11.4. The argument `variable_splits` specifies the variables (`age` and `fare`) for which CP profiles are to be calculated, together with the list of values at which the profiles are to be evaluated.

```
cp_titanic_rf <- ceteris_paribus(explain_rf_v6, henry,
              variable_splits = list(age = seq(0, 70, 0.1),
                                     fare = seq(0, 100, 0.1)))
```

```
plot(cp_titanic_rf) +
  show_observations(cp_titanic_rf, variables = c("age", "fare"), size = 5) +
  ylim(0, 1) +
  ggtitle("Ceteris Paribus profile",
          "For the random forest model and the titanic dataset")
```

\begin{figure}



}

\caption{Ceteris-paribus profiles for `class` and `embarked` variables and the `titanic_rf_v6` model. Blue dot stands for `henry`.}
\end{figure}

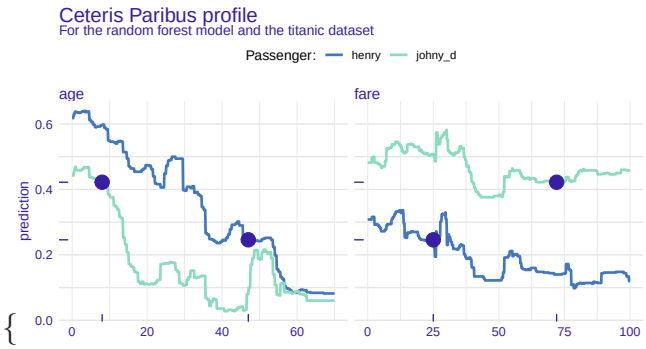
To enhance the plot, additional functions can be used. The generic `plot()` function creates a `ggplot2` object with a single `geom_line` layer. Function `show_observations` adds `geom_point` layer, `show_rugs` adds `geom_rugs`, while `show_profiles` adds another `geom_line`. All these functions take, as the first argument, an object created with the `ceteris_paribus` function. They can be combined freely to superimpose profiles for different models or observations.

In the example below, we present the code to create CP profiles for two passengers, `henry` and `johny_d`. Their profiles are included in a plot presented in Figure 0.11.6.2. We use the `scale_color_manual` function to add names of passengers to the plot, and to control colors and positions.

```
johny_d <- archivist::aread("pbiecek/models/e3596")
cp_titanic_rf2 <- ceteris_paribus(explain_rf_v6, rbind(henry, johny_d))
```

```
plot(cp_titanic_rf2, color = "_ids_") +
  show_observations(cp_titanic_rf2, size = 5, variables = c("age", "fare")) +
  show_rugs(cp_titanic_rf2, sides = "bl", variables = c("age", "fare")) +
  scale_color_manual(name = "Passenger:", breaks = 1:2,
            values = c("#4378bf", "#8bdcbe"), labels = c("henry" , "johny_d")) +
  ggtitle("Ceteris Paribus profile",
            "For the random forest model and the titanic dataset")
```

\begin{figure}



}

\caption{Ceteris-paribus profiles for the `titanic_rf_v6` model. Profiles for different passangers are color-coded.} \end{figure}

0.11.6.3 Champion-challenger analysis

One of the most interesting uses of the explainers is comparison of CP profiles for two or more of models.

To illustrate this possibility, first, we have go to construct profiles for the models. In our illustration, for the sake of clarity, we limit ourselves just to two models: the logistic regression and random forest models for the Titanic data. Moreover, we only consider the `age` and `fare` variables. We use `henry` as the instance, for which predictions are of interest.

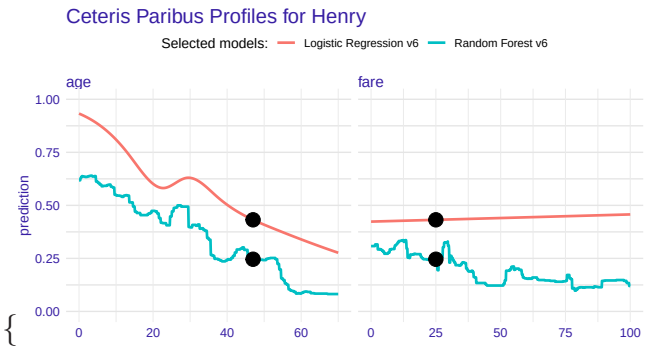
```
cp_titanic_rf <- ceteris_paribus(explain_rf_v6, henry)
cp_titanic_lmr <- ceteris_paribus(explain_lmr_v6, henry)
```

Subsequently, we construct the plot. The result is shown in Figure 0.11.6.3. Predictions for `henry` are slightly different, logistic regression returns in this case higher predictions then random forest. For `age` variable profiles of both models are similar, in both models we see decreasing dependency. While for `fare` the logistic regression model is slightly positive while random forest is negative. The larger the `fare` the larger is difference between these models. Such analysis helps us to which degree different models agree on what if scenarios.

Note that every `plot` and `show_*` function can take a collection of explainers as arguments. Profiles for different models are included in a single plot. In the presented R snippet, models are color-coded with the help of the argument `color = "_label_"`, where `_label_` refers to the name of the column in the CP explainer that contains the model label.

```
plot(cp_titanic_rf, cp_titanic_lmr, color = "_label_") +
  show_observations(cp_titanic_rf, cp_titanic_lmr, color = "black", variables = c("age", "fare
  scale_color_discrete(name = "Selected models:") + ylim(0,1) +
  ggtitle("Ceteris Paribus Profiles for Henry")
```

\begin{figure}



{

}

\caption{Champion-challenger comparison of the `titanic_lmr_v6`

and titanic_rf_v6 models. Profiles for different models are color-coded.} \end{figure}

0.12 Ceteris-paribus Oscillations and Local Variable-importance

0.12.1 Introduction

Visual examination of Ceteris-paribus (CP) profiles is insightful, but for a model with a large number of explanatory variables we may end up with a large number of plots which may be overwhelming. To prioritize between the profiles we need a measure that would summarize the impact of a selected variable on model's predictions. In this chapter we describe a solution closely linked with CP profiles. An alternative instance-level variable importnace is discussed in the Chapters 0.7 (Break Down), 0.9 (SHAP) and 0.10 (LIME).

0.12.2 Intuition

To assign importance to CP profiles, we can use the concept of profile oscillations. In particular, the larger influence of an explanatory variable on prediction at a particular instance, the larger the fluctuations along the corresponding CP profile. For a variable that exercises little or no influence on model prediction, the profile will be flat or will barely change. In other words, the values of the CP profile should be close to the value of the model prediction for the particular instance. Consequently, the sum of differences between the profile and the value of the prediction, take across all possible values of the explanatory variable, should be close to zero. The sum can be graphically depicted by the area between the profile and the horizontal line representing the instance prediction. On the other hand, for an explanatory variable with a large influence on the prediction, the area should

be large. Figure 0.12.2 illustrates the concept. Panel A of the Figure corresponds to the CP profiles presented in Figure 0.11.4. The larger the highlighted area in Figure 0.12.2, the more important is the variable for the particular prediction.

\begin{figure}

{

}

\caption{The value of the colored area summarizes the Ceteris-paribus-profile oscillations and provides the mean of the absolute deviations between the CP profile and the instance prediction. Panel A shows plots for continuous explanatory variables, while panel B shows plots for categorical variables in the `titanic_rf_v6` model.} \end{figure}

0.12.3 Method

Let us formalize this concept now. Denote by $g^j(z)$ the probability density function of the distribution of the j-th explanatory variable. The summary measure of the variable's importance for model prediction at point x_*, $vip^j_{CP}(x_*)$, computed based on the variable's CP profile, is defined as follows:

$$vip^j_{CP}(x_*) = \int_{\mathcal{R}} |h^j_{x_*}(z) - f(x_*)| g^j(z) dz = E_{X^j}\left[|h^j_{x_*}(X^j) - f(x_*)|\right]. \tag{0.21}$$

Thus, $vip^j_{CP}(x_*)$ is the expected absolute deviation of the CP profile from the model prediction for x_* over the distribution $g^j(z)$ for the j-th explanatory variable.

The true distribution of j-th explanatory variable is, in most cases, unknown. Thus, there are several options how to calculate (0.21).

One is to calculate just the area under the CP curve, i.e., to assume that $g^j(z)$ is a uniform distribution for the range of variable x^j. It folows then that a straightforward estimator of $vip^{j,uni}_{CP}(x_*)$ is

$$\widehat{vip}^{j,uni}_{CP}(x_*) = \frac{1}{k}\sum_{l=1}^{k} |h^j_{x_*}(z_l) - f(x_*)|, \tag{0.22}$$

where z_l $(l = 1, \ldots, k)$ are the selected values of the j-th explanatory variable. For instance, one can select use all unique values of x^j in the considered dataset. Alternatively, for a continuous variable, one can use an equi-distant grid of values.

Another approach is to use the empirical distribution for x^j. This leads to the estimator of $vip^{j,emp}_{CP}(x_*)$ defined as

$$\widehat{vip}^{j,emp}_{CP}(x_*) = \frac{1}{n}\sum_{i=1}^{n} |h^j_{x_*}(x^j_i) - f(x_*)|, \tag{0.23}$$

where index i goes through all observations in a dataset.

The use of of $\widehat{vip}^{j,emp}_{CP}(x_*)$ is preferred when there are enough data to accurately estimate the empirical distribution and when the distribution is not uniform. On the other hand, $\widehat{vip}^{j,uni}_{CP}(x_*)$ is in

most cases quicker to compute and, therefore, it is preferred if we look for fast approximations.

It is worth noting that the importance of an explanatory variable for instance prediction may be very different for different points x_*. For example, consider model

$$f(x_1, x_2) = x_1 * x_2,$$

where x_1 and x_2 take values in $[0, 1]$. Consider prediction for an observation described by vector $x_* = (0, 1)$. In that case, the importance of X_1 is larger than X_2. This is because the CP profile for the first variable, given by the values of function $f(z, 1) = z$, will have oscillations. On the other hand, the profile for the second variable will show no oscillations, because the profile is given by function $f(0, z) = 0$. Obviously, the situation is reversed for $x_* = (1, 0)$.

0.12.4 Example: Titanic

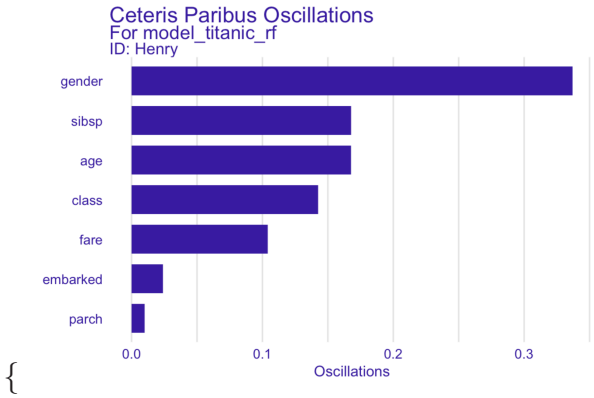
Figure 0.12.4 provides a barplot of variable importance measures for different continuous explanatory variables for the random forest model `titanic_rf_v6` for `johny_d`.

The longer the bar, the larger the CP-profile oscillations for a particular explanatory variable. Thus, Figure 0.12.4 indicates that the most important variable for prediction for the selected observation are `gender` and `sibsp`, followed by `age`.

From the Ceteris Paribus one can read that if Henry were older, this would significantly lower the chance of survival. One the other hand, were Henry not travelling alone, this would increase the chance.

From the oscillation's plot one can only read which features are important but one cannot read how they influence the prediction. This is why profile oscillations shall be accompanied by Ceteris Paribus profiles.

\begin{figure}



{

}

\caption{Variable-importance measures calculated for Ceteris-paribus oscillations for `johny_d` based on the `titanic_rf_v6` model} \end{figure}

0.12.5 Pros and cons

Oscillations of CP profiles are easy to interpret and understand. By using the average of oscillations, it is possible to select the most important variables for an instance prediction. This method can easily be extended to two or more variables. In such cases one needs to integrate the equation (0.22) over larger number of variables.

There are several issues related to the use of the CP oscillations. For example, the oscillations may not be of help in situations when the use of CP profiles may itself be problematic (e.g., in the case of correlated explanatory variables or interactions - see Section 0.11.5). An important issue is that the CP based local variable importance do not sum up to the instance prediction for which they are calculated, oposite to Break Down (Chapter 0.7) and Shapley values (Chapter 0.9).

0.12.6 Code snippets for R

In this section, we present key features of R package `ingredients` which is a part of the `DrWhy.AI` universe and covers all methods presented in this chapter.

For illustration purposes we use the random forest model `titanic_rf_v6` (see Section **??**). Recall that it is developed to predict the probability of survival from sinking of Titanic. Instance-level explanations are calculated for a single observation: `henry` - a 47-year-old passenger that travelled in the 1st class.

`DALEX` explainers for both models and the Henry data are retrieved via `archivist` hooks as listed in Section 0.5.1.7.

```
library("randomForest")
explain_rf_v6 <- archivist::aread("pbiecek/models/9b971")

library("DALEX")
henry <- archivist::aread("pbiecek/models/a6538")
henry
```

0.12.6.1 Basic use of the `calculate_oscillations` function

To calculate CP oscillations, we have got to calculate CP profiles for the selected observation. We use `henry` as the instance prediction of interest.

CP profiles are calculated by applying the `ceteris_paribus()` function to the wrapper object.

```
library("ingredients")
library("ggplot2")

cp_titanic_rf <- ceteris_paribus(explain_rf_v6, henry)
```

The resulting object can subsequently be processed with the `calculate_oscillations()` function to calculate the oscillations and the estimated value of the variable-importance measure (0.21).

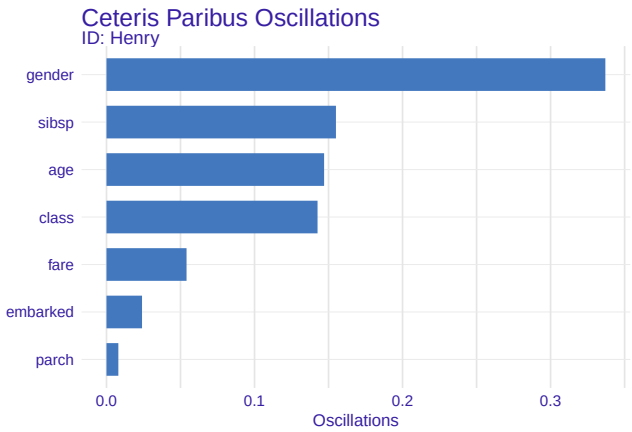
```
oscillations_titanic_rf <- calculate_oscillations(cp_titanic_rf)
oscillations_titanic_rf
```

```
##    _vname_ _ids_ oscillations
## 2   gender     1   0.33700000
## 4    sibsp     1   0.15500000
## 3      age     1   0.14700000
## 1    class     1   0.14257143
## 6     fare     1   0.05407273
## 7 embarked     1   0.02400000
## 5    parch     1   0.00800000
```

Note that, by default, `calculate_oscillations()` estimates $vip^{j}_{CP}(x_*)$ by $\widehat{vip}^{j,uni}_{CP}(x_*)$, given in (0.22), using all unique values of the explanatory variable as the grid points.

The `calculate_oscillations()` function returns an object of class `ceteris_paribus_oscillations`, which has a form of a data frame, but has also an overloaded `plot()` function. We can use the latter function to plot the local variable-importance measures for the instance of interest.

```
oscillations_titanic_rf$`_ids_` <- "Henry"
plot(oscillations_titanic_rf) + ggtitle("Ceteris Paribus Oscillations")
```



0.12.6.2 Advanced use of the `calculate_oscillations` function

As mentioned in the previous section, `calculate_oscillations()` estimates $vip^{j}_{CP}(x_*)$ by $\widehat{vip}^{j,uni}_{CP}(x_*)$ using all unique values of the explanatory variable as the grid points. However, other approaches are also possible.

One is to use $\widehat{vip}^{j,uni}_{CP}(x_*)$, but assuming an equi-distant grid of values for a continuous explanatory variable. Toward this aim, we have got to explicitly specify a dense uniform grid of values for such a variable. The `variable_splits` argument can be used for this purpose.

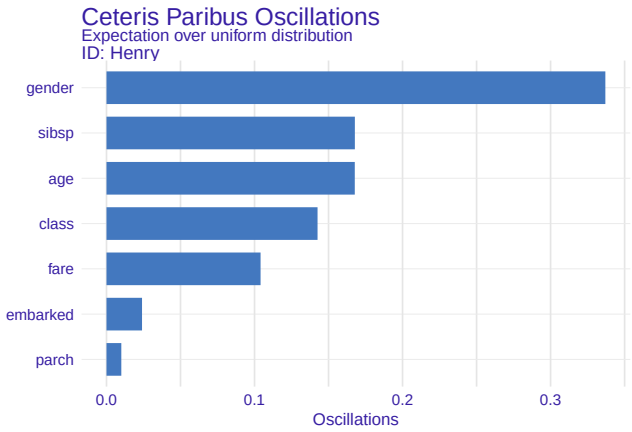
```
cp_titanic_rf_uniform <- ceteris_paribus(explain_rf_v6, henry,
              variable_splits = list(age = seq(0, 65, 0.1),
                                     fare = seq(0, 200, 0.1),
                                     sibsp = seq(0, 8, 0.1),
                                     parch = seq(0, 8, 0.1),
                                     gender = unique(titanic$gender),
                                     embarked = unique(titanic$embarked),
                                     class = unique(titanic$class)))
```

Subsequently, we apply the `calculate_oscillations()` function to compute the oscillations and the variable-importance measures.

```
oscillations_uniform <- calculate_oscillations(cp_titanic_rf_uniform)
oscillations_uniform$`_ids_` <- "Henry"
oscillations_uniform
```

```
##    _vname_ _ids_ oscillations
## 5   gender Henry    0.3370000
## 3    sibsp Henry    0.1677778
## 1      age Henry    0.1677235
## 7    class Henry    0.1425714
## 2     fare Henry    0.1040790
## 6 embarked Henry    0.0240000
## 4    parch Henry    0.0100000
```

```
plot(oscillations_uniform) + ggtitle("Ceteris Paribus Oscillations", "Expectation over uniform
```



Another approach is to calculate the expectation (0.21) over the empirical distribution of a variable, i..e, to use $\widehat{vip}^{j,emp}_{CP}(x_*)$, given in (0.23). Toward this aim, we use the `variable_splits` argument to explicitly specify the validation-data sample to define the grid of values.

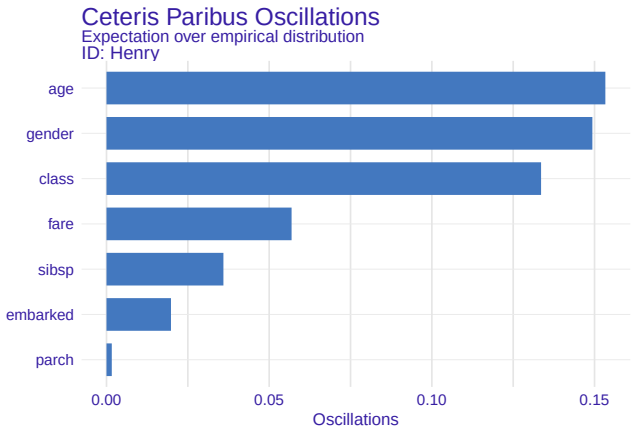
```
titanic <- na.omit(titanic)

cp_titanic_rf_empirical <- ceteris_paribus(explain_rf_v6, henry,
              variable_splits = list(age = titanic$age,
                                     fare = titanic$fare,
                                     sibsp = titanic$sibsp,
                                     parch = titanic$parch,
                                     gender = titanic$gender,
                                     embarked = titanic$embarked,
                                     class = titanic$class))
```

```
oscillations_empirical <- calculate_oscillations(cp_titanic_rf_empirical)
oscillations_empirical$`_ids_` <- "Henry"
oscillations_empirical
```

```
##    _vname_ _ids_ oscillations
## 1      age Henry  0.153323969
## 5   gender Henry  0.149336656
## 7    class Henry  0.133567739
## 2     fare Henry  0.056883552
## 3    sibsp Henry  0.035932034
```

```
## 6 embarked Henry  0.019818758
## 4    parch Henry  0.001623924
```

```
plot(oscillations_empirical) + ggtitle("Ceteris Paribus Oscillations", "Expectation over empir
```



0.13 Local Diagnostics Plots

0.13.1 Introduction

It may happen that, while the global predictive performance of a model is good, the model predictions for some observations are very misfitted. We often say that the model does not cover well some areas of the input space.

For example, a model calibrated for typical patients in a certain hospital may not do well with exceptionally young patients. Or a model calibrated for the credit risk of spring holiday consumer loans may not work well on a group of autumn loans for Christmas holiday gifts. For this reason, we should not be satisfied with global measures of model performance. For important decisions, it is good to check how the model behaves for observations similar to the instance of interest.

In this chapter, we present two local-diagnostics techniques that

address this issue, namely, *local fidelity plots* that show local performance around observation of interest and *local stability plots* that show the local stability around observation of interest.

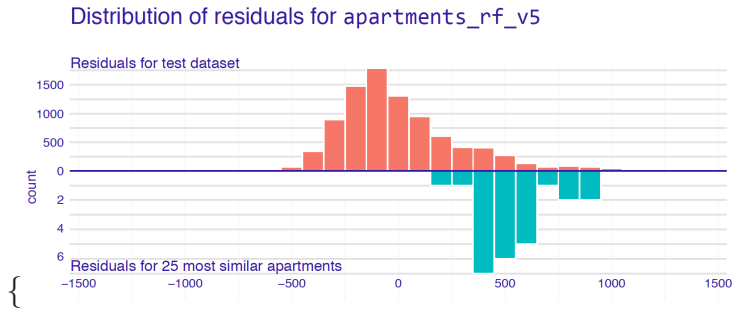
The general idea behind fidelity plots is to select a number of observations ("neighbors") from the validation dataset that are closest to the instance (observation) of interest. Then, for the selected observations, we plot CP profiles and check how stable they are. Additionally, if we know true values of the dependent variable for the selected neighbors, we may add residuals to the plot to evaluate the local fit of the model.

0.13.2 Intuition

Assume that we have identified a set of observations from the training data similar in terms of dependent variables to the observation of interest. The basic idea behind local fidelity plots is to compare distribution of residuals for these similar cases against distribution of all residuals.

Figure 0.13.2 presents histograms of residuals for the entire dataset and the selected neighbors for the random forest model for the Apartments dataset (Section 0.5.2.3). The distribution of residuals for the entire dataset is rather symmetric and centered around 0, suggesting a reasonable average performance of the model. On the other hand, the residuals for the selected neighbors are centered around the value of 500. This suggests that for the apartment of interest around this apartment the model is biased towards values smaller than observed (residuals are positive, so on average y is higher than $\hat{y}$, see (0.1)).

\begin{figure}



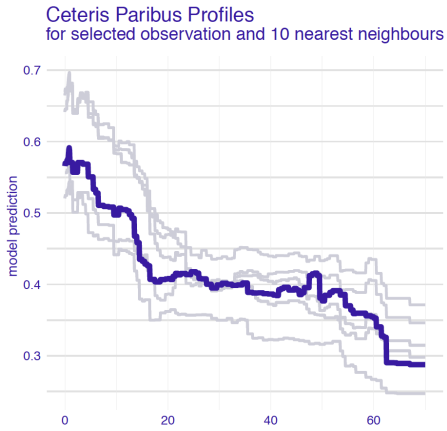
{

}

\caption{Histograms of residuals for the `apartments_rf_v5` model for the Apartments dataset. Upper panel: residuals calculated for all observations from the dataset. Bottom panel: residuals calculated for 25 nearest neighbors of the instance of interest.}
\end{figure}

Another approach to local model diagnostics is to examine how stable is model behaviour around the observation of interest. Figure 0.13.2 presents CP profiles for variable `age` for the instance of interest and its 10 nearest neighbors for the random forest model for the Titanic dataset (Section 0.5.1.3). The profiles are almost parallel and very close to each other. This suggests that model predictions are stable around the instance of interest, because small changes in the explanatory variables (represented by the nearest neighbors) have not got much influence on the predictions.

\begin{figure}



{

}

\caption{Ceteris-paribus profiles for a selected instance (dark violet line) and 10 nearest neighbors (light grey lines) for the `titanic_rf_b6` model. The profiles are almost parallel and close to each other what suggests the stability of the model.} \end{figure}

Of course CP profiles for different variables may be very different so a natural question arises which variables shall we examine. The most natural choice is to explore the most important variables according to results from the Break Down, SHAP, LIME od CP Oscillations methods.

0.13.3 Method

The proposed method is based on three steps:

- first, we need to select observations nearest to the observation of interest,
- for fidelity analysis we need to calculate and compare residuals for the neighbors.
- for stability analysis we need to calculate and visualize CP profiles for the selected neighbors.

In what follows we discuss each of the elements in more detail.

0.13.3.1 Nearest neighbors

There are two important questions related to the selection of the neighbors "nearest" to the instance (observation) of interest:

- How many neighbors should we choose?
- What metric should be used to measure the "proximity" of observations?

The answer to both questions is *it depends*.

- The smaller the number of neighbors, the more local is the analysis. However, a very small number will lead to a larger variability of the results. In many cases we found that 20 neighbors works fine. However, one should always take into account computational time (smaller number of neighbors results in quicker calculations) and the size of the dataset (for a small dataset, smaller sets of neighbors may be preferred).
- The metric is very important. The more explanatory variables, the more important is the choice. In particular, the metric should be capable of accommodating variables of different nature (categorical, continuous). Our default choice is the Gower similarity measure:

$$d_{gower}(x_i, x_j) = \frac{1}{p}\sum_{k=1}^{p} d^k(x_i^k, x_j^k),$$

where x_i is a p-dimensional vector of explanatory covariates for the i-th observation and $d^k(x_i^k, x_j^k)$ is the distance between values of the k-th variable for the i-th and j-th observations. Note that $d^k()$ depends on the nature of the variable. For instance, for a continuous variable it is equal to $|x_i^k - x_j^k|/\{max(x_1^k, \ldots, x_n^k) - min(x_1^k, \ldots, x_n^k)\}$, i.e., the absolute difference scaled by the observed range of the variable. On the other hand, for a categorical variable, it is simply $I(x_i^k = x_j^k)$, where $I()$ is the indicator function. Note that p may be equal to the number of all explanatory variables included in the model, or only a subset of them. An advantage of Gower similarity measure is that it "deals" with heterogeneous vectors with both categorical and continuous variables. The disadvantage of Gower similarity measure is that it does not take into account neither variable correlation nor

variable importance. For high dimensional setting an interesting alternative would be the proximity measure in Random Forest (Breiman, 2001). It takes into account variable importance but requires a fitted Random Rorest model.

Once we have decided on the number of neighbors, we can use the chosen metric to select the required number observations "closest" to the one of interest.

0.13.3.2 Local-fidelity plot

Figure 0.13.2 illustrates two distribution of residuals, for the whole dataset and for neighbours of the observation of interest.

For a typical observation these two distributions shall be similar. An alarming situation would be if the residuals for neighbours will be shifted towards the extremely positive of negative values.

Apart from visual examination we may also use some statistical tests that compares these two distributions. Since we cannot assume any distribution for residuals we can use a nonparametric test like Wilcoxon test or Kolmogorov-Smirnov test.

[TODO: maybe we need a better test for the stochastic dominance, or it is enough to have a test for location parameter?]

0.13.3.3 Local-stability plot for neighbors

Once nearest neighbors have been identified, we can graphically compare CP profiles for selected (or all) variables.

For a model with a large number of variables, we may end up with a large number of plots. In such a case a better strategy is to focus only on K most important variables, selected by using the variable-importance measure (see for example Chapter 0.12).

CP profiles are helpful to assess the model stability. In addition, we can enhance the plot by adding residuals to it to allow evaluation of the local model fit. For model $f()$ and observation i described by the vector of explanatory variables x_i, the residual

is the difference between the observed and predicted value of the dependent variable Y_i. Let us recall the definition (0.1):

$$r_i = y_i - f(x_i).$$

Note that, for a binary variable, the residual is the difference between the value of 0 or 1, depending on how we code "success," and the value of the predicted probability of "success." This defintion also applies to categorical responses, as it is common to define, in such case, a binary "success" indicator and compute the predicted probaiblity of "success" for each category separately.

The plot that includes CP profiles for the nearest neighbors and the corresponding residuals is called a local-fidelity plot. See an example in Figure 0.13.2.

0.13.4 Example: Titanic

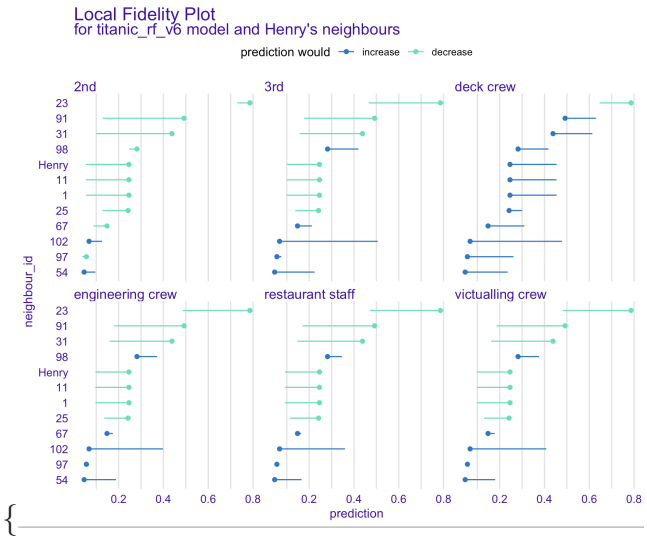
As an example, we will use the predictions for the random forest model for the Titanic data (see Section 0.5.1.3).

Figure 32 presents a detailed explanation of the elements of a local-fidelity plot for *age*, a continuous explanatory variable. The plot includes eight nearest neighbors of Henry (see Section 0.5.1.5). Profiles are quite apart from each other, which indicates potential instability of model predictions. However, the residuals included in the plots are positive and negative, indicating that, on average, the instance prediction should not be biased.

Figure 0.13.4 presents a local-fidelity plot for the categorical explanatory variable `class`. Henry and his neighbors traveled in the `1st` class. In different panels we see how the predicted probabilityof survival changes if the `1st` class is replaced, for instance, by the `2nd` (in most cases, they probability will be reduced) or the `deck crew` (in most cases, the probability will increase). Such plots can help to detect interactions, as we see that the same change (let's say, from the `1st` to the `3rd` class) results in a different change of the model prediction.

FIGURE 32 Elements of a local-stability plot for a continuous explanatory variable. The green line shows the Ceteris-paribus profile for the instance of interest. Profiles of the nearest neighbors are marked with grey lines. The vertical intervals correspond to residuals; the shorter the interval, the smaller the residual and the more accurate prediction of the model. Blue intervals correspond to positive residuals, red intervals to negative intervals. Stable model will have profiles close to each other; additive model will have parallel lines.

\begin{figure}



{

}

\caption{The local-stability plot for the categorical explanatory variable `class` in the random effects model for the Titanic data, `johny_d`, and his 10 neighbors. Each panel indicates how the model prediction would change if the class changed from `1st` to another one. Dots indicate original model predictions for the neighbors; the end of the interval corresponds to model prediction after changing the class. The top-lef panel indicates that, for the majority of the neighbors, the change from the `1st` to the `2nd` class reduces the predicted value of the probability of survival. On the other hand, the top-right panel indicates that changing the lass to `deck crew` members increases the predicted probability.} \end{figure}

0.13.5 Pros and cons

Local fidelity and stability plots may be very helpful to check if

- the model is locally additive, as for such models the CP profiles should be parallel;

- the model is locally stable, as in that case the CP profiles should be close to each other;
- the model fit for the instance of interest is good, as in that case the residuals should be small and their distribution should be balanced around 0.

The drawback is that such plots are quite complex and lack objective measures of the quality of the model fit. Thus, they are mainly suitable for an exploratory analysis.

0.13.6 Code snippets for R

In this section, we show how to use the R package `DALEX` (Biecek, 2018) to construct local-fidelity plots.

We use the random forest model `titanic_rf_v6` developed for the Titanic dataset (see Section 0.5.1.3) as the example. Recall that we try to address a classification problem for a binary dependent variable - we want to predict the probability of survival for a selected passenger.

`DALEX` explainers for the model and the `henry` data frame are retrieved via `archivist` hooks, as listed in Section 0.5.1.7.

```
library("randomForest")
library("DALEX")
explain_rf_v6 <- archivist::aread("pbiecek/models/9b971")
# recreate object to have predict function
explain_rf_v6 <- explain(explain_rf_v6$model,
                         data = explain_rf_v6$data,
                         y = explain_rf_v6$y,
                         verbose = FALSE)
henry <- archivist::aread("pbiecek/models/a6538")
henry
```

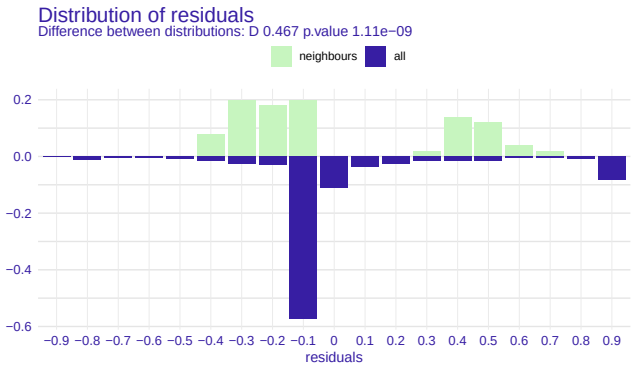
```
##   class gender age sibsp parch fare  embarked
## 1   1st   male  47     0     0   25 Cherbourg
```

We will show how to construct fidelity plot as in Figure 0.13.2. Toward this aim we need some number of passengers most similar

to `henry`. Here we are using `residuals_distribution` function from the `DALEX` package. First argument is an explainer, second the instance of interest, optional arguments are `neighbours` (number of neighbours) and `distance` (by default, the Gower distance is used).

This function needs to calculate residuals, so explainer shall be created with the `y` argument and also the `residual_function` argument.

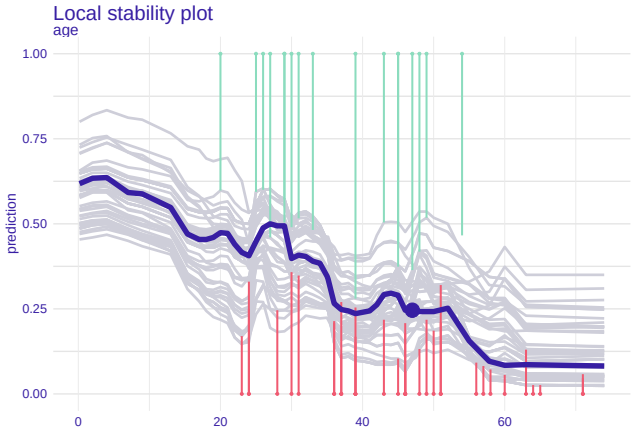
```
residuals_distribution(explain_rf_v6,
                       henry,
                       neighbours = 100)
```



The function `residuals_distribution()` can be also used for a local stability plot as in Figure 32 and **??**. To do this we need to also specify an `variables` argument.

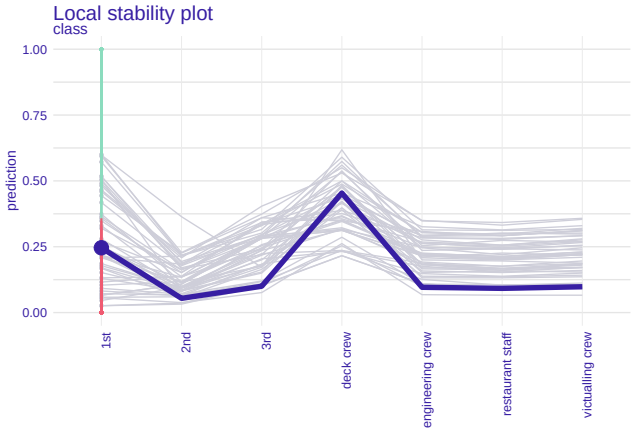
Toward this aim, we use the `y` argument in the `ceteris_paribus()` function. The argument takes numerical values. Our binary dependent variable `survived` assumes values `yes/no`; to convert them to numerical values, we use the `survived == "yes"` expression.

```
residuals_distribution(explain_rf_v6,
                       henry,
                       neighbours = 10,
                       variables = "age")
```



As we see the 10 passangers closest to `henry` are all from the `1st` class with age span between 20 and 60. Profiles for both `age` and `class` looks stable.

```
residuals_distribution(explain_rf_v6,
                       henry,
                       neighbours = 10,
                       variables = "class")
```

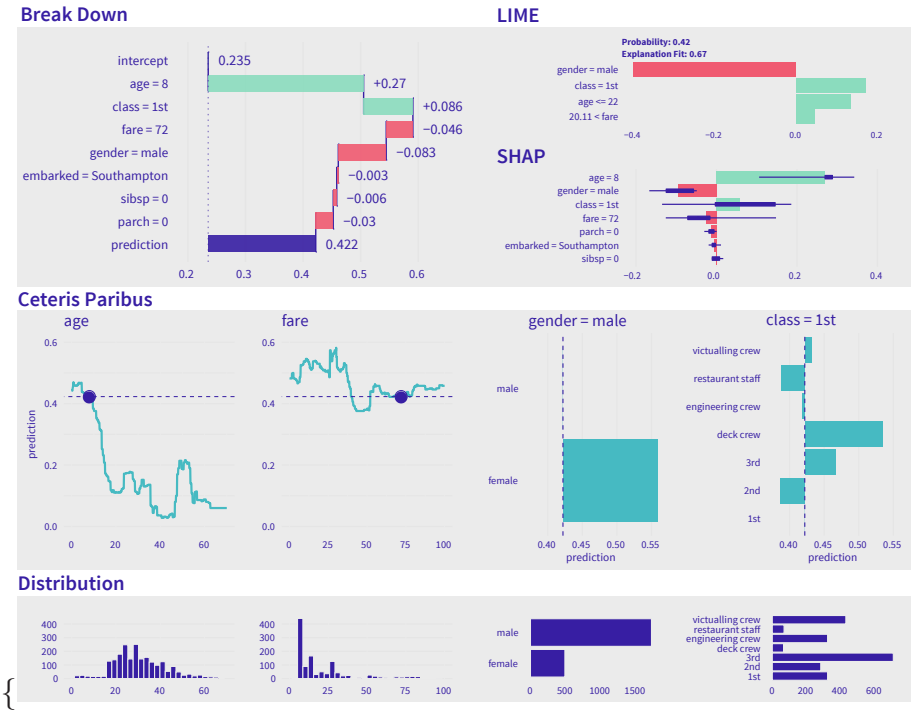


0.14 Summary of Instance-level Explainers

In the first part of the book, we introduced a number of techniques for exploration and explanation of model predictions for individual instances. In each chapter we introduced and presented a single technique. But in practice these techniques rarely shall be used separately. It's more informative to combine different views offered by each technique into a more holistic overview.

See an example in Figure 0.14. Four different approaches to the explanation of the random forest model are used. First row shows results from variable attribution methods like LIME, SHAP and Break Down. All these method agree that the most important variables for `johny_d` are his `age`, `gender`, `class` and `fare`. Since `fare` and `class` are correlated and possibly `age` is in the interaction with `gender` then the additive decomposition is ambiguous. Second row shows Ceteris Paribus profiles for these four most important variables. We see that higher age or being in the `2nd` class in he `restaurant staff` would decrease the model response while lower fare (which is counter intuitive), being a female or in the `deck crew` would increase the model response. Third row show univariate distributions of particular variables. We see that `fare=72` is very high as for a ticket and that only small fraction of people on the titanic were children (no kids in the crew). Combination of different perspectives supplement each other.

\begin{figure}



{

}

\caption{Instance-level explanations of the fandom forest model for the titanic data and `johny_d` an 8-years old boy that travels in the 1ts class. } \end{figure}

In the Chapter 0.21 we show an example how instance level explanations may be combined with dataset level explanations on a new use-case related to FIFA 19 data.

On one hand it is good to supplement different techniques for explanation with each other, but on another hand these techniques are different and may be more or less suitable for some selected problems. Below we discuss some differences.

0.14.1 Number of explanatory variables in the model

One of the most important criteria for selection of model exploration and explaination methods is the number of explanatory variables in the model.

0.14.1.1 Low to medium number of explanatory variables

A low number of variables usually implies that the particular variables have a very concrete meaning and interpretation. An example are models for the Titanic data presented in Sections 0.5.1.2-0.5.1.4.

In such a situation, the most detailed information about the influence of the variables on the model predictions is provided by the CP profiles. In particular, the variables that are most influential for model predictions are selected by considering CP-profile oscillations (see Chapter 0.12) and then illustrated graphically with the help of individual-variable CP profiles (see Chapter 0.11).

0.14.1.2 Medium to large number of explanatory variables

In models with a medium or large number of variables, it is still possible that most (or all) of them are interpretable. An example of such a model is a car-insurance pricing model in which we need to estimate the value of an insurance based on behavioral data that includes 100+ variables about characteristics of the driver and characteristics of the car.

When the number of explanatory variables increases, it becomes harder to show CP profile for each individual variable. In such situation, the most common approach is to use BD plots, presented in Chapter 0.7, or plots of Shapley values, discussed in Cahpter 0.9). They allow a quick evaluation whether a particular variable has got a positive or negative effect on model's prediction; we can also judge the size of the effect. If necessary, it

is possible to limit the plots only to the variables with the largest effects.

0.14.1.3 Very large number of explanatory variables

When the number of explanatory variables is very large, it may be difficult to interpret the role of each single variable. An example of such situation are models for processing of images or texts. In that case, explanatory variables may be individual pixels in image processing or individual characters in text analysis. As such, their individual interpretation is limited. Due to additional issues with computational complexity, it is not feasible to use CP profiles, BD plots, nor Shapley values to evaluate influence of individual values on model's predictions. Instead, the most common approach is to use LIME, presented in Chapter 0.10, which works on context-relevant groups of variables.

0.14.2 Correlated explanatory variables

When we derived some properties for presented methods we assumed that explanatory variables are independent. Obviously, this is not always the case. For instance, in the case of the data on aparement prices (see Chapter 0.5.2), the number of rooms and surface of an apartment will most likely be positively associated same is true for the class variable and fare for titanic data.

Of course all presented methods can be applied for correlated features, however sometimes it may be harder to analyze these features independently from each other.

To address the issue, the two most common approaches are: * to create new features that are independent (sometimes it is possible due to domain knowledge; sometimes it can e achieve by using principal components analysis or a similar technique), * construct two-dimensional extensions for CP plots (model response is plotted as a 2d surface) or permute variables in blocks to preserve the correlation structure of variables.

0.14.3 Models with interactions

In models with interactions, the effect of one explanatory variable may depend on values of other variables. For example, the probability of survival on Titanic may decrease with age, but the effect may be different for different classes of passengers. In such a case, to explore and explain model's predictions, we have got to consider not individual variables, but sets of variables included in interactions. To identify interactions, we can use BD plots as described in Chapter 0.8. To show effects of an interaction we may use a set of CP profiles. For the Titanic example we may use CP profiles for age with to instances that differ only in gender. The less parallel are such profiles the higher the effect of an interaction.

0.14.4 Sparse explanations

Predictive models may use hundreds of explanatory variables to yield a prediction for a particular instance. However, for a meaningful interpretation and illustration, most of human beings can handle only a very limited (say, less than 10) number of variables. Thus, sparse explanations are of interest. The most common method that is used to construct such explanations is LIME (Chapter 0.10). However, constructing a sparse explanation for a complex model is not trivial and may be misleading. Hence, care is needed when applying LIME to very complex models.

0.14.5 Additional uses of model exploration and explanation

In the previous chapters we focused on the application of the presented methods to exploration and explanation of predictive models. However, the methods can also be used to other aims:

- Model improvement. If a model prediction is particularly bad for a selected observation, then the investigation of the reasons for such a bad performance may provide some hints about how to

improve the model. In case of instance predcitions it is easier to note that a selected explanatory variable should have a different effect than the observed one.

- Additional domain-specific validation. Understanding which factors are important for model predictions helps in evalaution of the plausibility of the model. If the effects of some variables on the predictions are inconsistent with the domain knowledge, then this may provide a ground for criticising the model and, eventually, replacing it by another one. On the other hand, if the influence of the variables on model predictions is consistent with prior expectations, the user may become more confident with the model. Such a confidence is fundamental when the model predictions are used as a support for taking decisions that may lead to serious consequences, like in the case of, for example, predictive models in medicine.
- Model selection. In case of multiple candidate models, one may use results of the model explanation techniques to select one of the candidates. It is possible that, even if two models are similar in terms of a global model fit, the fit of one of them is locally much better. Consider the following, highly hypothethical example. Assume that a model is sought to predict whether it will rain on a particular day in a region where it rains on a half of the days. Two models are considered: one which simply predicts that it will rain every other day, and another that predicts that it will rain every day since October till March. Arguably, both models are rather unsophisticated (to say the least), but they both predict that, on average, half of the days wll be rainy. However, investigation of the instance predictions (for individual days) may lead to a preference for one of them.

0.14.6 Champion Challenger analysis

The techniques for explaining and exploring models have many applications. One of them is the opportunity to compare models.

Why compare models? One scenario is the Champion-Challenger

analysis. Let's assume that some institution uses a predictive model but wants to know if they could get a better model using other modeling techniques. For example, the risk department in a bank uses logistical regression to assess credit risk. The model has some efficiency and is the so-called champion - the best model considered in the class of logistic regression models. However, it is worth checking whether using more complex models, so called challengers, e.g. boosting or random trees, will not be more effective. And if they are more effective, the question will arise as to how these challengers differ from the champion.

Another reason why we want to compare models is because of the iterativness of the modeling process itself (see 0.2.2). During the modeling process many versions of the models are created, often with different structures, sometimes with very similar efficiency. Comparative analysis allows for better understanding how these models differ from each other.

Below is an example of champion-challenger analysis for Random Forest model `model_titanic_rf`, logistic regression model `model_titanic_lmr`, boosting model of `model_titanic_gbm` and support-vector machines (SVM) model of `model_titanic_svm`.

Each of these models has a different way of functioning. Random forest and boosting models are on trees, so the response curves will be stepped one. The logistic regression and booster models have continuous and smooth response curves.

Figure 33 shows the Shapley values for the four models built in chapter 0.5.1 using the example of `john_d`. For three models, namely random forest, boosting and logistic regression, similar variables are indicated as important: `class`, `age` and `gender`. For the SVM model the most important variable is `gender`, followed by `age` and `parch`.

Shapley values show an additive distribution for model predictions. In the chapter 0.8 we discussed what to do if the add-on attribute may not reflect the exact behaviour of the model. Figure 34 compares Break Down plots with interactions for the four models under consideration.

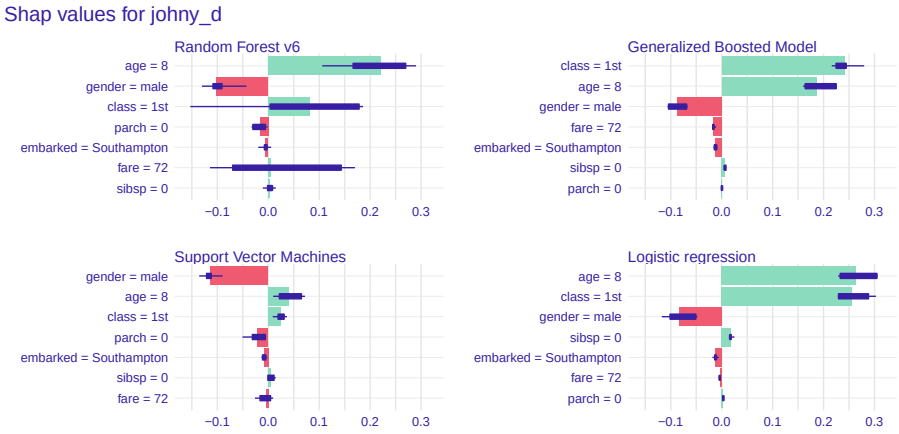


FIGURE 33 SHAP plots for four different models for the Titanic data.

Each of these models obviously has a different estimate for the chances of survival for `johny_d`. The highest estimate has the logistic regression model 0.764 while the lowest estimate has the random forest model 0.441. For the SVM model, the most important variable is `gender` and for the other models, `age` and `class`. The Random Forest model included interactions of `fare:class` and the SVM model included interactions of `fare:age`.

Figure 35 shows Ceteris Paribus profiles for the four models considered for the `age` and `fare` variables. The logistic regression and GBM models behave in a similar way. Random forest and SVM models are much less sensitive to the `age` variable.

Each of the four models under consideration has a different structure and each of them is for some reason a complex model. Random forest and boosting models are complex due to the large number of trees used for prediction. The SVM model is complex due to the non-linear function of the kernel and the logistic regression model due to spline transformations.

The compilation of the operating profile of the models side-by-side allows for a better understanding of the similarities and differences in the signals that these models have learned.

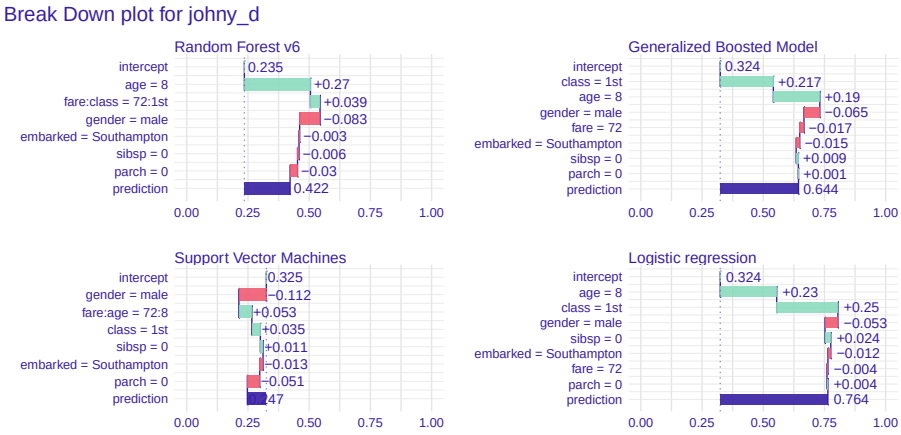


FIGURE 34 Break Down plots for four different models for the Titanic data.

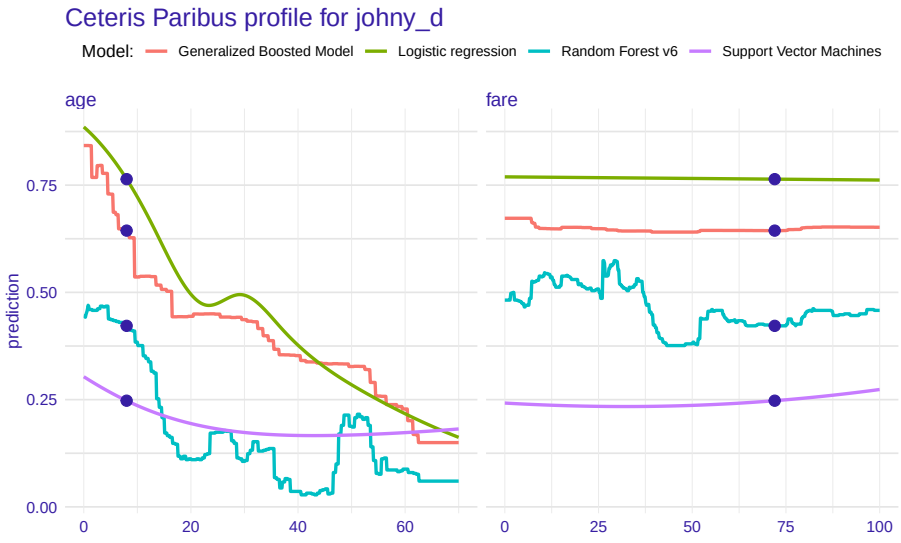


FIGURE 35 Ceteris Paribus profiles for four different models for the Titanic data.

Dataset Level

0.15 Model-level exploration

In Part I, we focused on instance-level explainers, which help to understand how a model yields a prediction for a single observation (instance).

In Part II, we concentrate on model-level explainers, which help to understand how model's predictions perform overall, for a set of observations. Assuming that the observations form a representative sample from a general population, model-level explainers can provide an information about the quality of predictions for the population.

The following examples illustrate situations in which model-level explainers may be useful:

- We may want to learn which variables are ''important" in the model. For instance, we may be interested in predicting the risk of heart attack by using explanatory variables that are obtained based on results of some medical examinations. If some of the variables do not influence model's predictions, we could simplify the model by removing the variables.
- We may want to understand how a selected variable influences model's predictions. For instance, we may be interested in predicting prices of apratments. Apartment's location is an important factor, but we may want to know which locations lead to higher prices?
- We may want to discover whether there are any observations, for which the model yields wrong predictions. For instance, for a model predicting the probability of survival after a risky treatment, we might know whether there are patients for whom the model predictions are extremely wrong. Identifying such a group

of patients might point to, for instance, an incorrect form of a explanatory variable or even a missed variable.

Model-level explainers focus on four main aspects of a model:

- Variable's importance: which explantory variables are ''important", and which are not?
- Variable's effect: how does a variable influence average model's predictions?
- Model's performance: how ''good" is the model? Is one model ''better" than another?
- Model's fit: which observations are misfitted by the model, where residual are the largest?

In all cases, measures capturing a particular aspect of the model have to be defined. We will discuss them in subsequent chapters. In particular, in Chapter 0.16, we discuss measures that are useful for the evaluation of the overall predictive model performance. In Chapter 0.17, we focus on methods that allow evaluation of a variable's effect on model's predictions. Chapter 0.18 XXXX Chapter 0.19 XXX. Chapter 0.20 presents an overview of the classical residual-diagnostics tools. Finally, in Chapter **??**, we present an example of an analysis that illustrates the use of the model-level explainers introduced in the previous chapters.

[PBI: IDEAS TO BE EXPANDED]

- Comparison of models with different sets of explanatory variables
- Drift in a model performance in time

0.16 Model Performance Measures

0.16.1 Introduction

In this chapter, we present measures that are useful for the evaluation of the overall performance of a predictive model. They may be applied for several purposes:

- model evaluation: we may want to know how good is the model, i.e., how reliable are the model predictions (how frequent and how large errors we may expect);
- model comparison: we may want to compare two or more models in order to choose between them;
- out-of-sample and out-of-time comparisons: we may want to check model's performance when applied to new data to evaluate if the performance has not worsened.

Depending of the nature of the dependent variable (continuous, binary, categorical, count, etc.), different model performance measures may be used. Moreover, the list of useful measures is growing as new applications emerge. In this chapter, we focus on a selected set of measures that are used in model-level exploration techniques that are introduced in subsequent chapters.

0.16.2 Intuition

Most model performance measures are based on comparison of the model predictions with the (known) values of the dependent variable in a dataset. For an ideal model, the predictions and the dependent-variable values should be equal. In practice, it is never the case, and we want to quantify the disagreement.

In applications, we can weigh differently the situation when the prediction is, for instance, larger than the true value, as compared to the case when it is smaller. Depending on the

decision how to weigh different types of disagreement, we may need different performance measures.

When assessing model's overall performance, it is important to take into account the risk of overestimation of the quality of the performance when considering the data that were used for developing of the model. To mitigate the risk, various assessment stategies, such as cross-validation, have been proposed (see (Kuhn and Johnson, 2013b)). In what follows, we consider the simple train-test-split strategy, i.e., we assume that the available data are split into a training set and a testing set. Model is created on the training set, and the testing set is used to assess the model's performance.

In the best possible scenario we can specify a single model performance measure before the model is created and then we optimize model for this measure. But in practice the more common scenario is to have few performance measures that are often selected after the model is created.

0.16.3 Method

Assume that we have got a testing dataset with n observations on p explanatory variables and on a dependent variable Y. Let x_i denote the (column) vector of values of the explanatory variables for the i-th observation, and y_i the corresponding value of the dependent variable. Denote by $\widehat{y}_i = f(x_i)$ model's $f()$ prediction corresponding to y_i. Let $X = (x_1', \ldots, x_n')$ denote the matrix of explanatory variables for all n observations, and $y = (y_1, \ldots, y_n)'$ denote the (column) vector of the values of the dependent variable.

0.16.3.1 Continuous dependent variable

The most popular model performance measure for models for a continuous dependent variable is the mean squared-error, defined as

$$MSE(f, X, y) = \frac{1}{n}\sum_{i}^{n}\{f(x_i) - y_i\}^2 = r_i^2, \tag{0.24}$$

where $r_i = f(x_i) - y_i$ is the residual for the i-th observation. Thus, MSE can be seen as a sum of squared residuals. MSE is a convex differentiable function, which is important from an optimization point of view. As the measure weighs all differences equally, large residuals have got a high impact on MSE. Thus, the measure is sensitive to outliers. For a ''perfect" predictive model, which predicts all y_i exactly, $MSE = 0$.

Note that MSE is constructed on a different scale than the dependent variable. Thus, a more interpretable variant of this measure is the root-mean-squared-error (RMSE), defined as

$$RMSE(f, X, y) = \sqrt{MSE(f, X, y)}. \tag{0.25}$$

A popular variant of RMSE is its normalized version, R^2, defined as

$$R^2(f, X, y) = 1 - \frac{MSE(f, X, y)}{MSE(f_0, X, y)}. \tag{0.26}$$

In (0.26), $f_0()$ denotes a ''baseline" model. For instance, in the case of the classical linear regression, $f_0()$ is the model that includes only the intercept, which implies the use of the average value of Y as a prediction for all observations. R^2 is normalized in the sense that the ''perfect" predictive model leads to $R^2 = 1$, while $R^2 = 0$ means that we are not doing better than the baseline model. In the context of the classical linear regresssion, R^2 is the familiar coefficient of determination and can be interpreted as the fraction of the total variance of Y explained by model $f()$.

Given sensitivity of MSE to outliers, sometimes the mean absolute-error (MAE) is considered as a model performance measure:

$$MAE(f, X, y) = \frac{1}{n}\sum_{i}^{n}|f(x_i) - y_i| = \frac{1}{n}\sum_{i}^{n}|r_i|. \tag{0.27}$$

MAE is more robust to outliers than MSE. A disadvantage of MAE are its less favorable mathematical properties.

0.16.3.2 Binary dependent variable

To introduce model performance measures, we, somewhat arbitrarily, label the two possible values of the dependent variable as ''success" and ''failure''. (Of course, in a particular application, the meaning of the''success" outcome does not have to be positive nor optimistic; in diagnostic tests ''success" often means detection of a diseases.) We also assume that model prediction $f(x_i)$ takes the form of the predicted probability of success.

If, additionally, we assign the value of 1 to success and 0 to failure, it is possible to use MSE, RMSE, and MAE, as defined in (0.24), (0.25), (0.27), respectively, as a model performance measure. In practice, however, those summary measures are not often used. One of the main reasons is that they penalize too mildly for wrong predictions. In fact, the maximum penalty for an individual prediction is equal to 1 (if, for instance, the model yields zero probability for an actual success).

To address this issue, the log-likelihood function based on the Bernoulli distibution can be used:

$$l(f, X, y) = -\sum_{i=1}^{n}[y_i \ln\{f(x_i)\} + (1 - y_i)\ln\{1 - f(x_i)\}]. \tag{0.28}$$

Note that, in the machine-learning world, often $l(f, X, y)/n$ is

considered (sometimes also with ln replaced by $\log_2$) and termed ''logloss" or ''cross-entropy''. The log-likelihood heavily''penalizes" the cases when the model-predicted probability of success $f(x_i)$ is high for an actual failure ($y_i = 0$) and low for an actual success ($y_i = 1$).

In many situations, however, a consequence of a prediction error depends on the form of the error. For this reason, performance measures based on the (estimated values of) probability of correct/wrong prediction are more often used. To introduce some of those measures, we assume that, for each observation from the testing dataset, the predicted probability of success $f(x_i)$ is compared to a fixed cut-off threshold, C say. If the probability is larger than C, then we assume that the model predicts success; otherwise, we assume that it predicts failure. As a result of such a procedure, the comparison of the observed and predicted values of the dependent variable for the n observations in the testing dataset can be summarized in the following table:

	True value: `success`	True value: `failure`	T
Predicted: `success`	True Positive: TP	False Positive (type I error): FP	P
Predicted: `failure`	False Negative (type II error): FN	True Negative: TN	N
Total	S	F	n

In machine-learning world, the table is often referred to as the ''confusion table" or ''confusion matrix''. In statistics, it is often called the''decision table". The counts TP and TN on the diagonal of the table correspond to the cases when the predicted and observed value of the dependent variable Y coincide. FP is the number of cases in which failure is predicted as success. These are false-positive, or type I error, cases. On the other hand, FN is the count of false-negative, or type II error, cases, in which success is predicted as failure. Marginally, there are P predicted successes and N predicted failures, with $P + N = n$. In the testing dataset, there are S observed sucesses and F observed failures, with $S + N = n$.

The simplest measure of model performance is **accuracy**, defined as

$$ACC = \frac{TP + TN}{n}.$$

It is the fraction of correct predictions in the entire testing dataset. Accuracy is of interest if true positives and true negatives are more important than their false counterparts. However, accuracy may not be very informative when one of the binary categories is much more prevalent. For example, if the testing data contain 90% of successes, a model that would always predict a success would reach accuracy of 0.9, although one could argue that this is not a very useful model.

There may be situations when false positives and/or false negatives may be of more concern. In that case, one might want to keep their number low. Hence, other measures, focused on the false results, might be of interest.

In the machine-learning world, two other measures are often considered: **precision** and **recall**. Precision is defined as

$$Precision = \frac{TP}{TP + FP} = \frac{TP}{P}.$$

Precision is also referred to as the positive predictive value. It is the fraction of correct predictions among the predicted successes. Precision is high if the number of false positives is low. Thus, it is a useful measure when the penalty for committing the type I error (false positive) is high. For instance, consider the use of a genetic test in cancer diagnostics, with a positive result of the test taken as an indication of an increased risk of developing a cancer. A false positive result of a genetic test might mean that a person would have to unnecessarily cope with emotions and, possibly, medical procedures, related to the fact of being evaluated as having a high risk of developing a cancer. We might want to avoid this situation more than the false negative case.

The latter would mean that the genetic test gives a negative result for a person that, actually, might be at an increased risk of developing a cancer. However, an increased risk does not mean that the person will develop cancer. And even so, we could hope that we could detect it in due time.

Recall is defined as

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{S}.$$

Recall is also referred to as sensitivity or true positive rate. It is the fraction of correct predictions among the true successes. Recall is high if the number of false negatives is low. Thus, it is a useful measure when the penalty for committing the type II error (false negative) is high. For instance, consider the use of an algorithm that predicts whether a bank transaction is fraudulent. A false negative result means that the algorithm accepts a fraudulent transaction as a legitimate one. Such a decision may have immediate and unpleasant consequences for the bank, because it may imply a non-recoverable loss of money. On the other hand, a false positive result means that a legitimate transaction is considered as fraudulent one and is blocked. However, upon further checking, the legitimate nature of the transaction can be confirmed with, perhaps, annoyed client as the only consequence for the bank.

The harmonic mean of these two measures defines the **F1 score**:

$$F1\ score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}.$$

F1 score tends to give a low value if either precision or recall is low, and a high value if both precision and recall are high. For instance, if precision is 0, F1 score will also be 0 irrespectively of the value of recall. Thus, it is a useful measure if we have got to seek a balance between precision and recall.

In statistics, and especially in applications in medicine, the popular measures are **sensitivity** and **specificity**. Sensitivity is simply another name for recall. Specificity is defined as

$$Specificity = \frac{TN}{TN + FP} = \frac{TN}{F}.$$

Specificity is also referred to as true negative rate. It is the fraction of correct predictions among the true failures. Specificity is high if the number of false positives is low. Thus, as precision, it is a useful measure when the penalty for committing the type I error (false positive) is high.

The reason why sensitivity and specificity may be more often used outside the machine-learning world is related to the fact that their values do not depend on the proportion S/n (sometimes termed ''prevalence") of true successess. This means that, once estimated in a sample obtained from a population, they may be applied to other populations, in which the prevalence may be different. This is not true for precision, because one can write

$$Precision = \frac{Sensitivity \cdot \frac{S}{n}}{Sensitivity \cdot \frac{S}{n} + Specificity \cdot \left(1 - \frac{S}{n}\right)}.$$

All the measures depend on the choice of cut-off C. To assess the form and the strength of dependence, a common approach is to construct the Receiver Operating Characteristic (ROC) curve. The curve plots the $sensitivity$ in function of $1 - specificity$ for all possible, ordered values of C. Figure 36 presents the ROC curve for the random-forest model for the Titanic dataset (see Section 0.5.1.3). Note that the curve indicates an inverse relationship between sensitivity and specifcity: by increasing one measure, the other is decreased.

The ROC curve is very informative. For a model that predicts successes and failures at random, the corresponding ROC curve

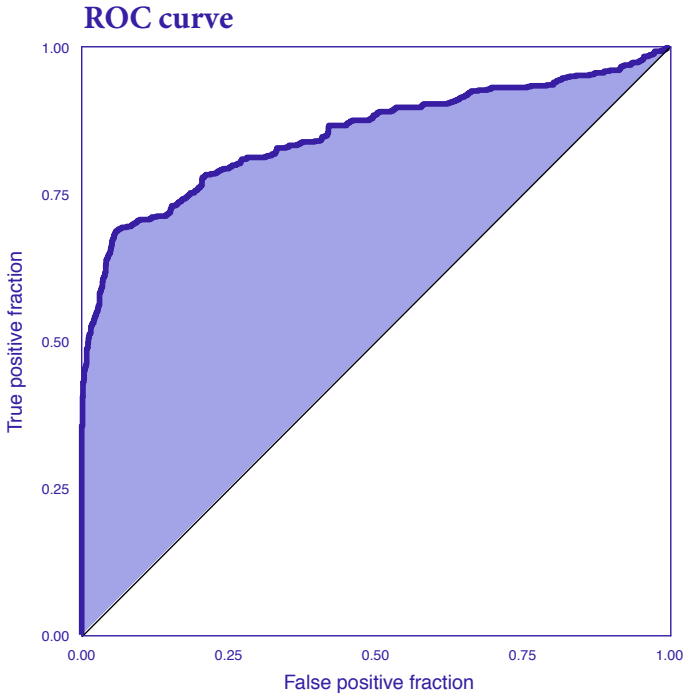


FIGURE 36 (fig:exampleROC) ROC curve for the random-forest model for the Titanic dataset. The Gini ceofficient can be calculated as 2 x area between the ROC curve and the diagonal (this area is highlighted).

will be equal to the diagonal line. On the other hand, for a model that yields perfect predictions, the ROC curve reduces to a two intervals that connect points (0,0), (0,1), and (1,1).

Often, there is a need to summarize the ROC curve and, hence, model's performance. A popular measure that is used toward this aim is the area under the curve (AUC). For a model that predicts successes and failures at random, AUC is the area under the diagonal line, i.e., it is equal to 0.5. For a model that yields perfect predictions, AUC is equal to 1.

Another ROC-curve-based measure that is often used is the Gini coefficient G. It is closely related to AUC; in fact, it can be calculated a $G = 2 \times AUC - 1$. For a model that predicts successes and failures at random, $G = 0$; for a perfect-prediction model, $G = 1$.

The value of Gini's coefficient or, equivalently, of $AUC - 0.5$ allow a comparison of the model-based predictions with random guessing. A measure that explicitly compares a prediction model with a baseline (or null) model is the **lift**. Commonly, random guessing is considered as the baseline model. In that case,

$$Lift = \frac{\frac{TP}{P}}{\frac{S}{n}} = \frac{Precision}{\frac{S}{n}}.$$

Note that S/n can be seen as the estimated probability of a correct prediction of a success for random guessing. On the other hand, TP/P is the estimated probability of a correct prediction a succcess given that the model predicts a success. Hence, informally speaking, the lift indicates how many more (or less) times the model does better in predicting success than random quessing. As other measures, the lift depends on the choice of cut-off C. The plot of the lift as a function of C is called the lift chart.

There are many more measures aimed at measuring performance of a predictive model for a inary dependent variable. An overview can be found in, e.g., (Berrar D. Performance Measures for

Binary Classification. Encyclopedia of Bioinformatics and Computational Biology Volume 1, 2019, Pages 546-560). [TOMASZ: INCLUDE IN THE REFERENCE LIST.]

0.16.3.3 Categorical dependent variable

To introduce model performance measures for a categorical dependent variable, we assume that y_i is now a vector of K elements. Each element y_{ik} $(k = 1, \ldots, K)$ is a binary variable indicating whether the k-th category was observed for the i-th observation. We assume that for each observation only one category can be observed. Thus, all elements of y_i are equal to 0 except of one that is equal to 1. Furthermore, We assume that model prediction $f(x_i)$ takes the form of a vector of the predicted probabilities for each of the K categories. The predicted category is the one with the highest predicted probability.

The log-likelihood function (0.28) can be adapted to the categorical dependent variable case as follows:

$$l(f, X, y) = -\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik} \ln\{f(x_i)_k\}. \tag{0.29}$$

It is essentially the log-likelihood function based on a multinomial distibution.

It is also possible to extend the performance measures like accuracy, precision, etc., introduced in Section 0.16.3.2. Toward this end, first, a confusion table is created for each category k, treating the category as ''success" and all other categories as ''failure". Let us denote the counts in the table by TP_k, FP_k, TN_k, and FN_k. Based on the counts, we can compute the average accuracy across all classes as follows:

$$\overline{ACC} = \frac{1}{K}\sum_{k=1}^{K} \frac{TP_k + TN_k}{n}. \tag{0.30}$$

Similarly, one could compute the average precision, average sensitivity, etc. In machine-learning world, this approach is often termed ''macro-averaging". The averages computed in that way treat all classes equally.

An alternative approach is to sum the appropriate counts from the confusion tables for all classes, and then form a measure based on the so-computed cumulative counts. For instance, for precision, this would lead to

$$\overline{Precision}_{\mu} = \frac{\sum_{k=1}^{K} TP_k}{\sum_{k=1}^{K} (TP_k + FP_k)}. \tag{0.31}$$

In machine-learning world, this approach is often termed ''micro-averaging" (hence subscript μ for ''micro" in $Precision_{\mu}$ in (0.31)). Note that, for accuracy, this computation still leads to (0.30). The measures computed in that way favor classes with larger numbers of observations.

0.16.3.4 Count dependent variable

In case of counts, one could consider using any of the measures for a continuous dependent variable mentioned in Section 0.16.3.1. However, a particular feature of a count dependent variable is that, often, its variance depends on the mean value. Consequently, weighing all contributions to MSE equally, as in (0.24), is not appropriate, because the same residual value r_i indicates a larger discrepancy for a smaller count y_i than for a larger one. Therefore, a popular measure is of performance of a predictive model for counts is Pearson's statistic:

$$\chi^2(f, X, y) = \sum_{i}^{n} \left\{ \frac{f(x_i) - y_i}{\sqrt{f(x_i)}} \right\}^2 = \sum_{i}^{n} \left\{ \frac{r_i}{\sqrt{f(x_i)}} \right\}^2. \tag{0.32}$$

From (0.32) it is clear that, if the same residual value is obtained

for two different observed counts, it is assigned a larger weight for the count for which the predicted value is smaller.

0.16.4 Example

0.16.4.1 Apartment prices

Let us consider the linear regression model `apartments_lm_v5` (see Section 0.5.2.2) and the random-forest model `apartments_rf_v5` (see Section 0.5.2.3) for the data on the apartment prices (see Section 0.5.2). Recall that, for these data, the dependent variable, the price, is continuous. Hence, we can use the performance measures presented in Section 0.16.3.1. In particular, we consider MSE and MAE. The values of the two measures for the two models are presented below.

```
## Model label:  Linear Regression v5
##          score name
## mse 78023.1235  mse
## mae   260.0254  mae

## Model label:  Random Forest v5
##          score name
## mse 36669.1954  mse
## mae   144.0888  mae
```

Both MSE and MAE indicate that, overall, the random-forest model performs better than the linear regression model.

0.16.4.2 Titanic data

Let us consider the random-forest model `titanic_rf_v6` (see Section 0.5.1.3 and the logistic regression model `titanic_lmr_v6` (see Section 0.5.1.2) for the Titanic data (see Section 0.5.1). Recall that, for these data, the dependent variable is binary, with success defined as survival of the passenger.

First, we will take a look at the accuracy, F1 score, and AUC for the models.

```
## Model label:  Logistic Regression v6
##         score name
## auc 0.8196991  auc
## f1  0.6589018   f1
## acc 0.8046689  acc

## Model label:  Random Forest v6
##         score name
## auc 0.8566304  auc
## f1  0.7289880   f1
## acc 0.8494521  acc
```

Overall, the random-forest model is performing better, as indicated by the larger values of all the measures.

Figure 37 presents ROC curves for both models. The curve for the random-forest model lies above the one for the logistic regression model for the majority of the cut-offs C, except for the very high values.

Figure 38 presents lift curves for both models. Also in this case the curve for the random-forest suggests a better performance than for the logistic regression model, except for the very high values of cut-ff C. [TOMASZ: THIS CURVE IS NOT CONSISTENT WITH THE DEFINITION OF THE LIFT. EXPLAIN/CHANGE?]

0.16.5 Pros and cons

All model performance measures presented in this chapter face some limitations. For that reason, many measures are available, as the limitations of a particular measure were addressed by developing an alternative. For instance, RMSE is frequently used and reported for linear regression models. However, as it is sensitive to outliers, MAE was proposed. In case of predictive models for a binary dependent variable, the measures like

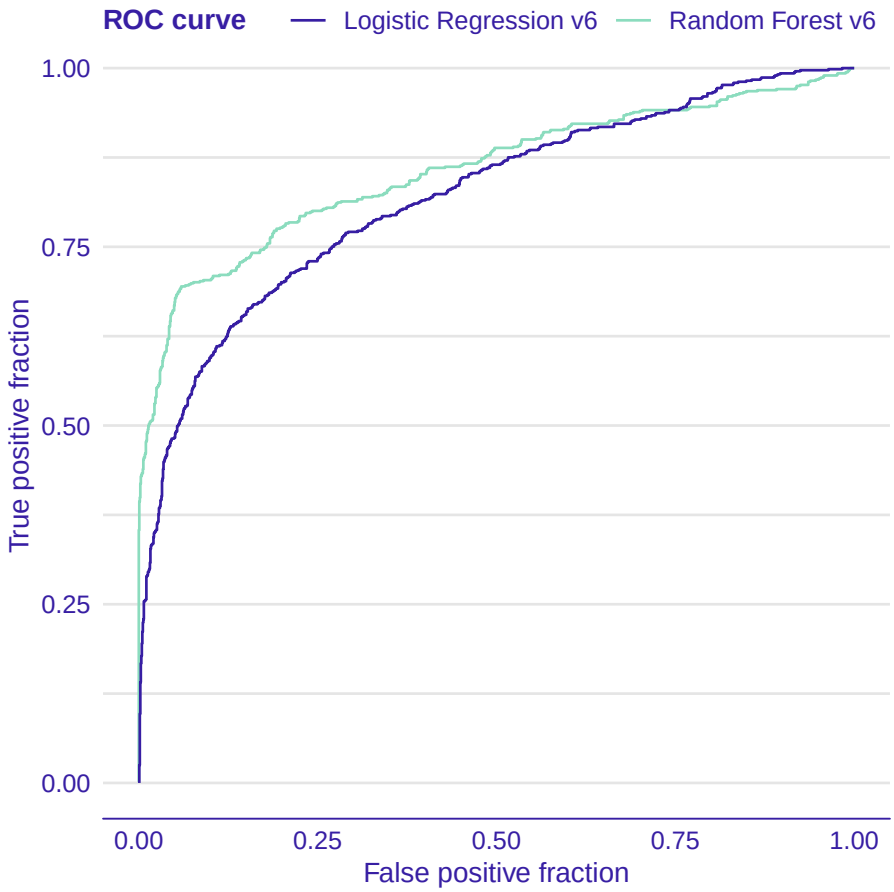


FIGURE 37 (fig:titanicROC) ROC curves for the random-forest model and the logistic regression model for the Titanic dataset.

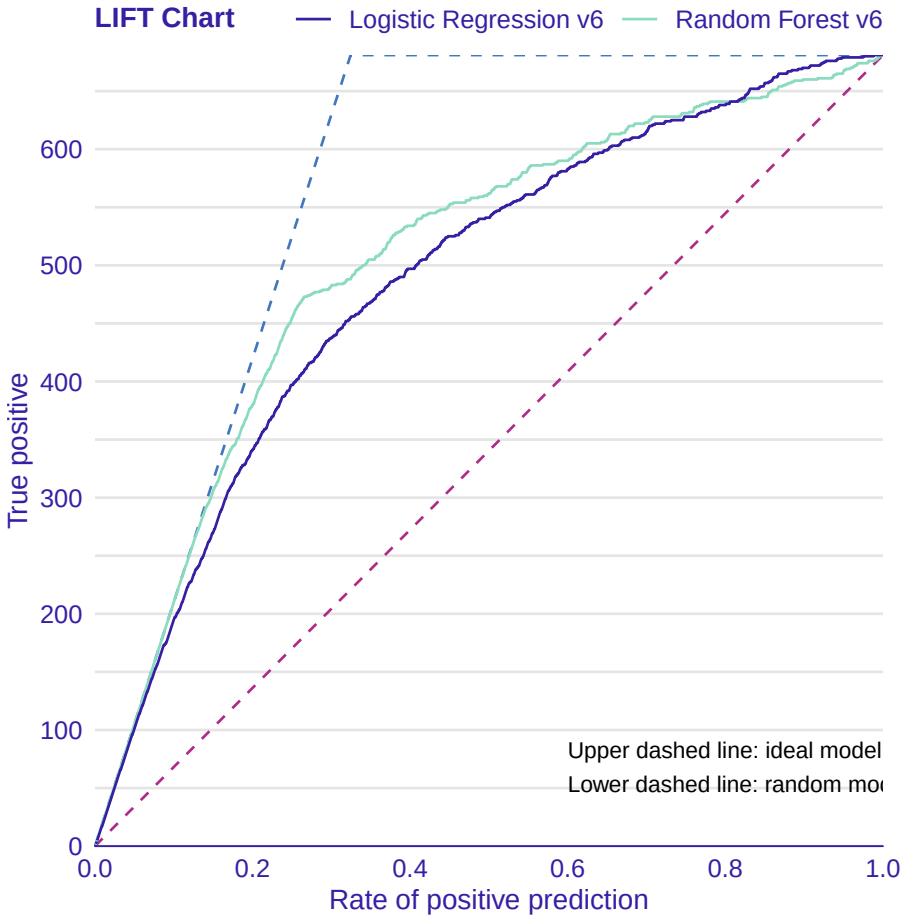


FIGURE 38 (fig:titanicLift) Lift curves for the random-forest model and the logistic regression model for the Titanic dataset.

accuracy, F1 score, sensitivity, and specificity, are often considered depending on the consequences of corrrect/incorrect predictions in a particular application. However, the value of those measures depends on the cut-off value used for creating the predictions. For this reason, ROC curve and AUC have been developed and have become very popular. They are not easily extanded to the case of a categircal dependent variable, though.

Given the advantages and disadvantages of various measures, and the fact that each may reflect a different aspect of the predicitve performance of a model, it is customary to report and compare several of them when evaluating a model's performance.

0.16.6 Code snippets for R

In this section, we present the key features of the `auditor` R package (**?**) which is a part of the DrWhy.AI universe. The package covers all methods presented in this chapter. It is available on CRAN and GitHub. More details and examples can be found at https://modeloriented.github.io/auditor/.

Note that there are also other R packages that offer similar functionality. These include, for instance, packages `mlr` (Bischl et al., 2016), `caret` (from Jed Wing et al., 2016), `tidymodels` (Max and Wickham, 2018), and `ROCR` (Sing et al., 2005).

For illustration purposes, we use the random-forest model `titanic_rf_v6` (see Section 0.5.1.3 and the logistic regression model `titanic_lmr_v6` (see Section 0.5.1.2) and the random-forest model `titanic_rf_v6` (see Section 0.5.1.3) for the Titanic data (see Section 0.5.1). Consequently, the functions from the `auditor` package are applied in the context of a binary classification problem. However, the same functions can be used for, e.g., linear regression problems.

To illustrate the use of the functions, we first load explainers for both models.

```
library("auditor")
library("randomForest")

explainer_titanic_rf <- archivist:: aread("pbiecek/models/51c50")
explainer_titanic_lr <- archivist:: aread("pbiecek/models/42d51")
```

Function `auditor::model_performance()` calculates selected model performance measures. The `score` argument is used to select the desired measures. The `data` argument serves for specification of the test dataset, for which the selected measures are to be computed. Note that, by default, the data are extracted from the explainer object. Finally, it is possible to use the `cutoff` argument to specify the cut-off value to obtained cut-off-dependent measures like F1 score or accuracy.

```
model_performance(explainer_titanic_rf, score = c("auc", "f1", "acc"))
```

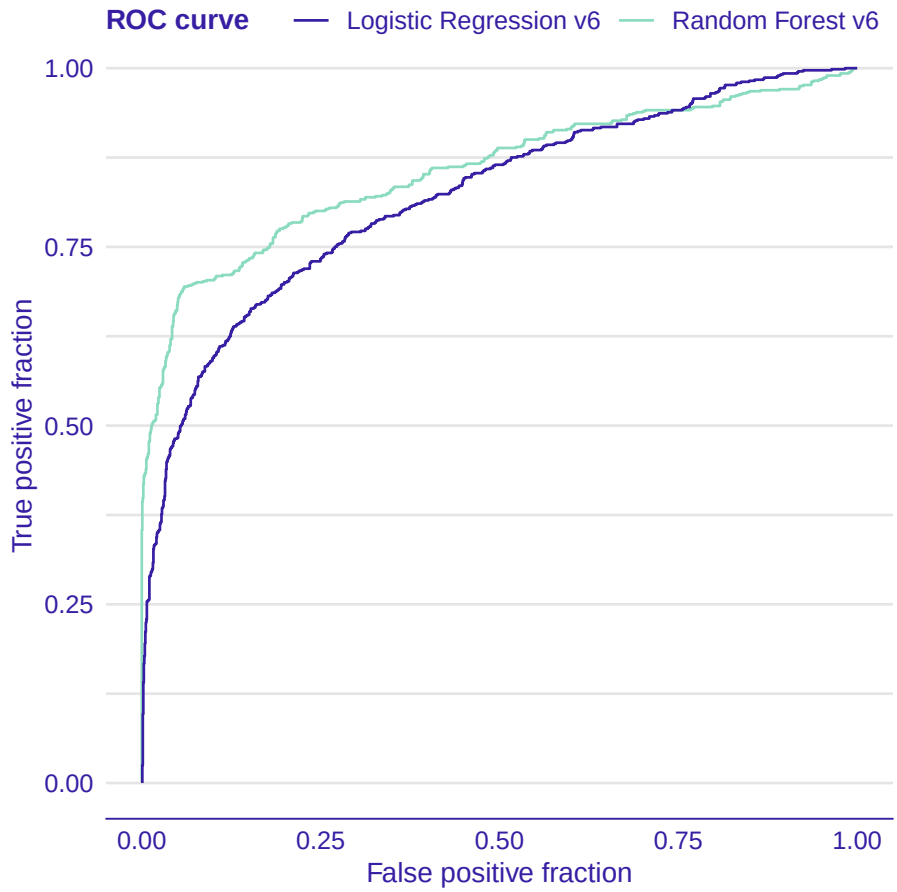
```
## Model label:  Logistic Regression v6
##         score name
## auc 0.8196991  auc
## f1  0.6589018   f1
## acc 0.8046689  acc
```

```
model_performance(explainer_titanic_lr, score = c("auc", "f1", "acc"))
```

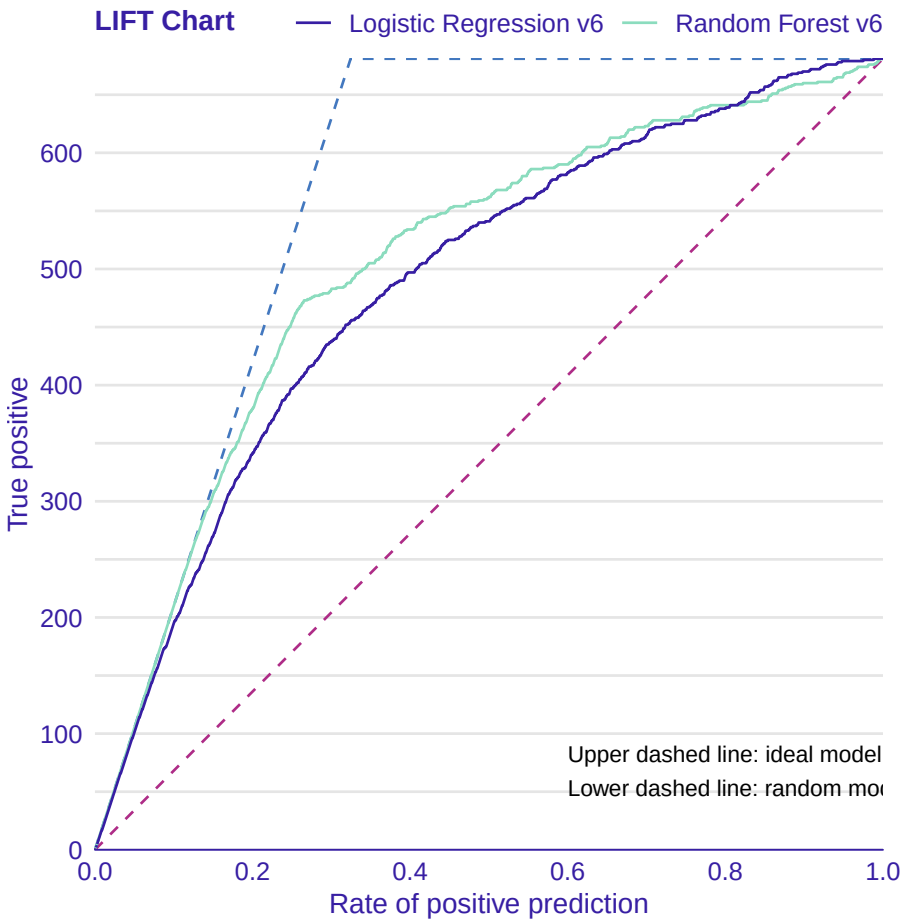
```
## Model label:  Random Forest v6
##         score name
## auc 0.8566304  auc
## f1  0.7289880   f1
## acc 0.8494521  acc
```

ROC or lift curves can be constructed by, first, using the `model_evaluation()` function. [TOMASZ: WHAT DOES IT DO?] Subsequently, the resulting object is used in the `plot_roc()` or `plot_lift()` function calls. Both plot functions return `ggplot2` objects and can take one or more explainer objects as arguments. In the latter case, the profiles for each explainer are superimposed on one plot.

```
eva_rf <- model_evaluation(explainer_titanic_rf)
eva_lr <- model_evaluation(explainer_titanic_lr)
plot_roc(eva_rf, eva_lr)
```




```
plot_lift(eva_rf, eva_lr)
```



The resulting plots are shown in Figures 37 and 38. Both plots can be supplemented with boxplots for residuals. Toward this end, the residuals have got to be computed and added to the explainer object with the help of the `model_residual()` function. Subsequently, the `plot_residual_boxplot()` can be applied to the resulting object.

```
mr_rf <- model_residual(explainer_titanic_rf)
mr_lm <- model_residual(explainer_titanic_lr)

plot_residual_boxplot(mr_rf, mr_lm)
```

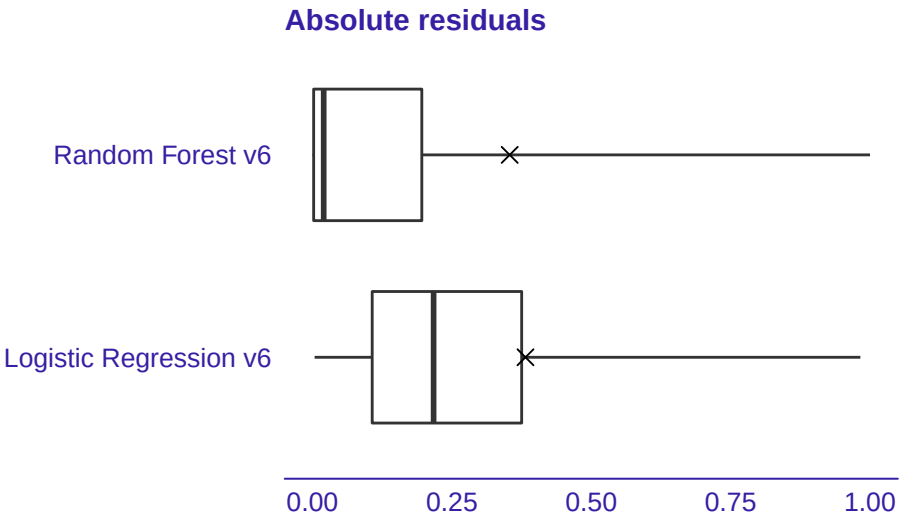


FIGURE 39 (fig:titanicBoxplots) Boxplots for residuals for two models on Titanic dataset.

0.17 Variable's Importance

0.17.1 Introduction

In this chapter, we present methods that are useful for the evaluation of an explanatory variable's importance. The methods may be applied for several purposes.

- Model simplification: variables that do not influence model's predictions may be excluded from the model.
- Model exploration: comparison of a variable's importance in different models may help in discovering interrelations between the variables.Also, ordering of variables in function of their importance is helpful in deciding in what order should we perform further model exploration.
- Domain-knowledge-based model validation: identification of the most important variables may be helpful in assessing the validity of the model based on the domain knowledge.
- Knowledge generation: identification of the most important vari-

ables may lead to discovery of new factors involved in a particular mechanism.

The methods for assessment of feature importance can be divided, in general, into two groups: model-specific and model-agnostic.

For models like linear models, random forest, and many others, there are methods of asssing of variable's importance that exploit particular elements of the structure of the model. These are model-specific methods. For instance, for linear models, one can use the value of the normalized regression coefficient or its corresponding p-value as the variable-importance measure. For tree-based ensembles, such a measure may be based on the use of a particular variable in particular trees (see, e.g., (Foster, 2017) for gradient boosting and (Paluszynska and Biecek, 2017) for random forest).

In this book we focus on model-agnostic methods. These methods do not assume anything about the model structure. Therefore, they can be applied to any predictive model or ensemble of models. Moreover, and perhaps even more importantly, they allow comparing variable's importance between models with different structures.

0.17.2 Intuition

We focus on the method described in more detail in (Fisher et al., 2018). The main idea is to measure how much the model fit decreases if the effect of a selected explanatory variable or of a group of variables is removed. The effect is removed by means of perturbations like resampling from an empirical distribution of just permutation of the values of the variable.

The idea is in some sense borrowed from variable important measure proposed by **??**randomForestBreiman for random forest. If a variable is important, then, after permutation, model's performance should become worse. The larger drop in the performance, the more important is the variable.

The method can be used to measure importance of a single explanatory variable or of a group of variables. The latter is useful for aspects - groups of variables that are complementary. Consider for example a behavioral model in credit scoring in which some aggregate, let say number of loans, is calculated for different intervals. In one column there is an aggregate for 1 month, in the next one is for 6 months and in another in one year. If we are interested in a question, how important is the aspect 'number of loans' then we can measure the drop in performance if all variables in a given aspect are perturbated.

Despite the simplicity of definition, the model agnostic feature importance is a very powerful tool for model exploration. Values of feature importance may be compared between different models. So one can compare how different models use correlated variables. Models like random forest are expected to spread importance across every variable while in regression models coefficients for one correlated feature may dominate over coefficients for other variables.

0.17.3 Method

Consider a set of n observations for a set of p explanatory variables. Denote by $\widetilde{y} = (f(x_1), \ldots, f(x_n))$ the vector of predictions for model $f()$ for all the observations. Let y denote the vector of observed values of the dependent variable Y.

Let $\mathcal{L}(\widetilde{y}, y)$ be a loss function that quantifies goodness of fit of model $f()$ based on $\widetilde{y}$ and y. For instance, $\mathcal{L}$ may be the value of likelihood. Consider the following algorithm:

1. For each explanatory variable X^j included in the model, do steps 2-5
2. Replace vector x^j of observed values of X^j by vector $x^{*,-j}$ of resampled or permuted values.
3. Calculate model predictions $\widetilde{y}^{*,-j}$ for the modified data.
4. Calculate the value of the loss function for the modified

data:

$$L^{*,-i} = \mathcal{L}(\tilde{y}^{*,-j}, y)$$

5. Variable's importance is calculated as $vip_A(x^j) = L^{*,-j} - L$ or $vip_R(x^j) = L^{*,-j}/L$, where L is the value of the loss function for the original data.

Note that resampling or permuting data, used in Step 2, involves randomness. Thus, the results of the procedure may depend on the actual configuration of resampled/permuted values. Hence, it is advisable to repeat the procedure several times. In this way, the uncertainty related to the calculated variable-importance values can be assessed.

The calculations in Step 5 "normalize" the value of the variable's importance measure with respect to L. However, given that l is a constant, the normalization has no effect on the ranking of variables according to $vip_A(x^j)$ or $vip_R(x^j)$. Thus, in practice, often the values of $L^{*,-i}$ are simly used to quantify variable's importance.

0.17.4 Example: Titanic data

In this section, we illustrate the use of the permutation-based variable-importance method by applying it to the random forest model for the Titanic data (see Section 0.5.1.3).

Consider the random forest model for the Titanic data (see Section 0.5.1.3). Recall that the goal is to predict survival probability of passengers based on their sex, age, cityplace of embarkment, class in which they travelled, fare, and the number of persons they travelled with.

Figure **??** shows the values of $L^{*,-j}$ after permuting, in turn, each of the variables included in the model. [TOMASZ: WHICH LOSS FUNCTION? WHY COUNTRY IS INCLUDED IN THE PLOT?] Additionally, the plot indicates the value of L by the vertical dashed line at the left-hand-side of the plot.

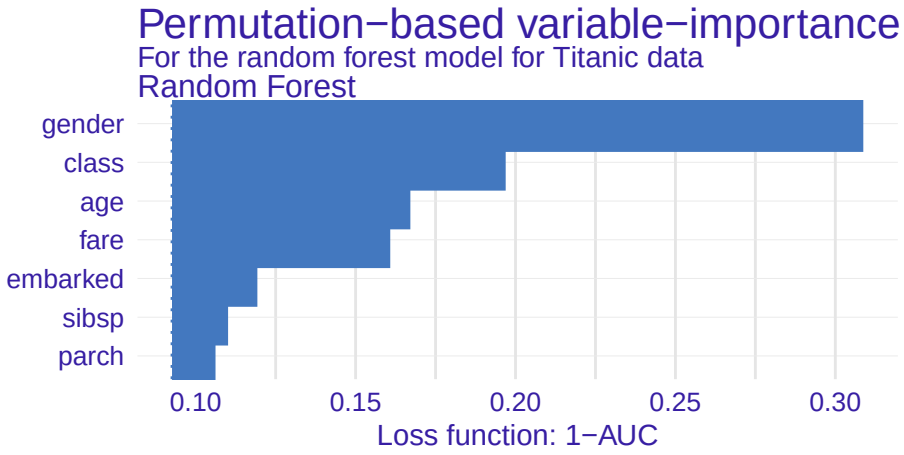


FIGURE 40 (fig:TitanicRFFeatImp) Variable importance. Each interval presents the difference between the loss function for the original data (vertical dashed line at the left) and for the data with permuted observation for a particular variable.

The plot in Figure **??** suggests that the most important variable in the model is gender. This agrees with the conclusions drawn in the exploratory analysis presented in Section 0.5.1.1. The next three important variables are class of the travel (first-class patients had a higher chance of survival), age (children had a higher chance of survival), and fare (owners of more expensive tickets had a higher chance of survival).

To take into account the uncertainty related to the use of permutations, we can consider computing the average values of $L^{*,-j}$ over a set of, say, 10 permutations. The plot in Figure **??** presents the average values. [TOMASZ: IT WOULD BE GOOD TO GET THE SAME X-AXIS IN BOTH PLOTS.] The only remarkable difference, as compared to Figure **??**, is the change in the ordering of the `sibsp` and `parch` variables.

The plots similar to those presented in Figures Figure **??** and Figure **??** are useful for comparisons of variable importance for different models. Figure **??** presents the single-permutation [TOMASZ: CORRECT?] results for the random forest, gradient

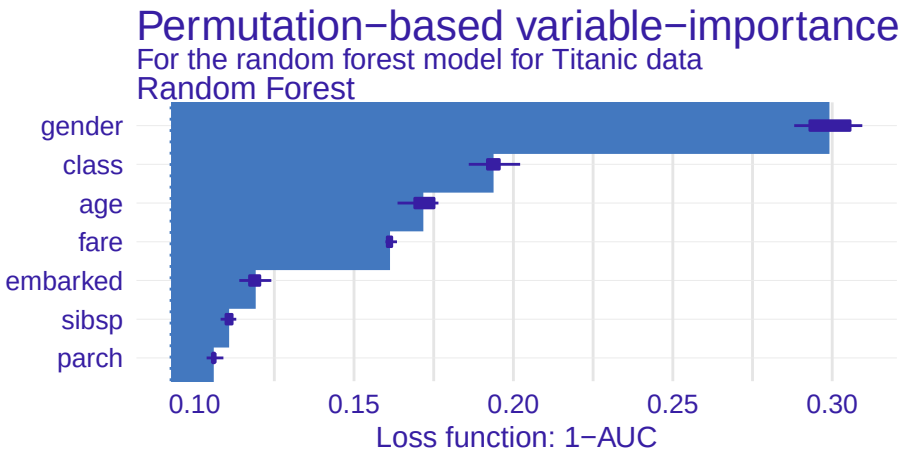


FIGURE 41 (fig:TitanicRFFeatImp10) Average variable importance based on 10 permutations.

boosting (see Section 0.5.1.4), and logistic regression (see Section 0.5.1.2) models. [TOMASZ: WHAT LOSS FUNCTION?] The best result, in terms of the smalles value of the goodness-of-fit function L, are obtained for the random forest model. Note, however, that this model includes more variables than the other two. For instance, variable `fare` variable, which is highly correlated with the travel class, is used neither in the gradient boosting nor in the logistic regression model, but is present in the random forest model. [TOMASZ: BUT, IN CHAPTER 4, ALL MODELS WERE BUILT USING THE SAME SET OF VARIABLES. ARE WE USING DIFFERENT MODELS HERE? THIS IS CONFUSING.]

The plots in Figure **??** indicate that `gender` is the most important variable in all three models, followed by `class`.

[TOMASZ: STOPPED HERE WITH RE-WRITING]

0.17.5 Pros and cons

[TOMASZ: TO POPULATE]

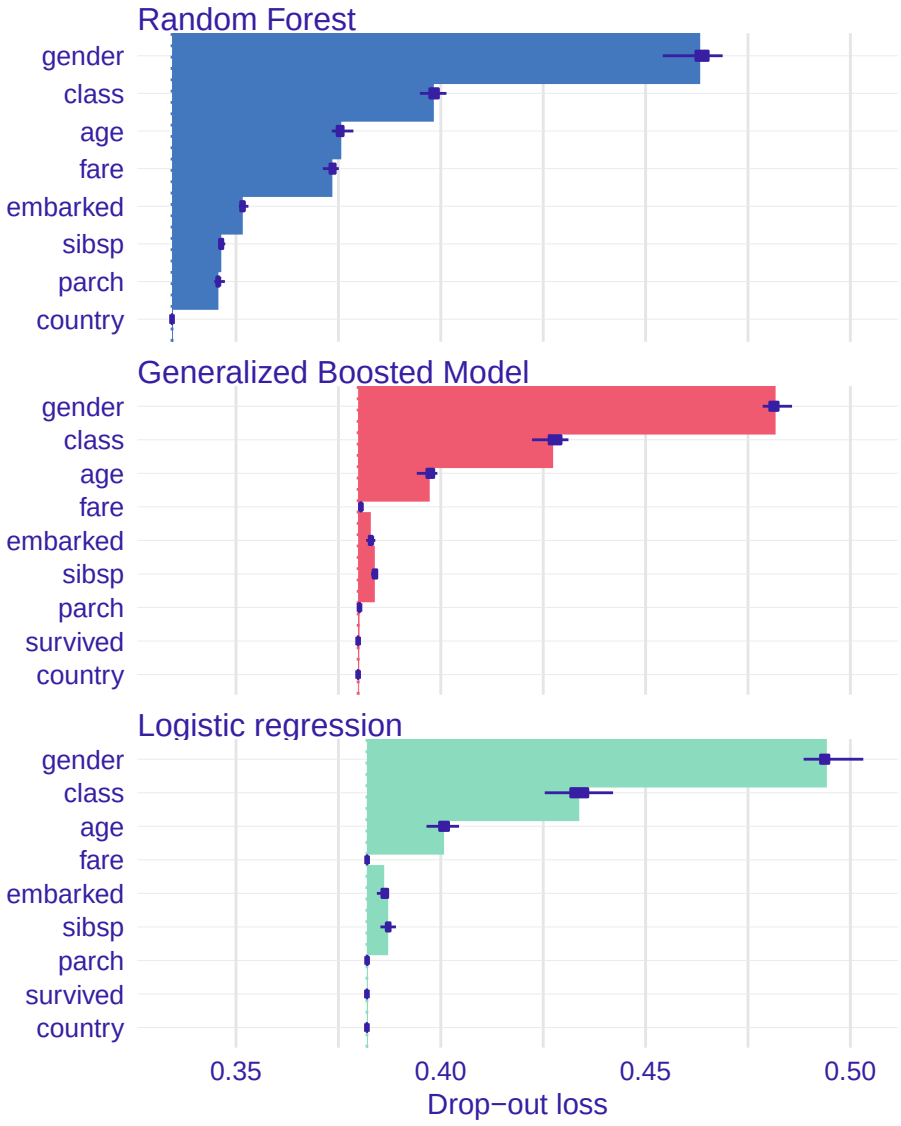


FIGURE 42 (fig:TitanicFeatImp) Variable importance for the random forest, gradient boosting, and logistic regression models for the Titanic data.

0.17.6 Code snippets for R

For illustration, We will use the random forest model for the apartment prices data (see Section 0.5.2.3.

Let's create a regression model for prediction of apartment prices.

```
library("DALEX")
library("randomForest")
set.seed(59)
model_rf <- randomForest(m2.price ~ construction.year + surface + floor +
                           no.rooms + district, data = apartments)
```

A popular loss function for regression model is the root mean square loss

$$L(x, y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}$$

```
loss_root_mean_square(
  predict(model_rf, apartments),
  apartments$m2.price
)
```

```
## [1] 193.8477
```

Let's calculate feature importance

```
explainer_rf <- explain(model_rf,
            data = apartmentsTest[,2:6], y = apartmentsTest$m2.price,
                                colorize = FALSE)
```

```
## Preparation of a new explainer is initiated
##   -> model label       :  randomForest  (  default  )
##   -> data              :  9000  rows  5  cols
##   -> target variable   :  9000  values
##   -> model_info        :  package randomForest , ver. 4.6.14 , task regression (  default
##   -> predict function  :  yhat.randomForest  will be used (  default  )
##   -> predicted values  :  numerical, min =  1977.609 , mean =  3511.789 , max =  5839.917
##   -> residual function :  difference between y and yhat (  default  )
##   -> residuals         :  numerical, min =  -1970.395 , mean =  -0.2658962 , max =  1944.71
```

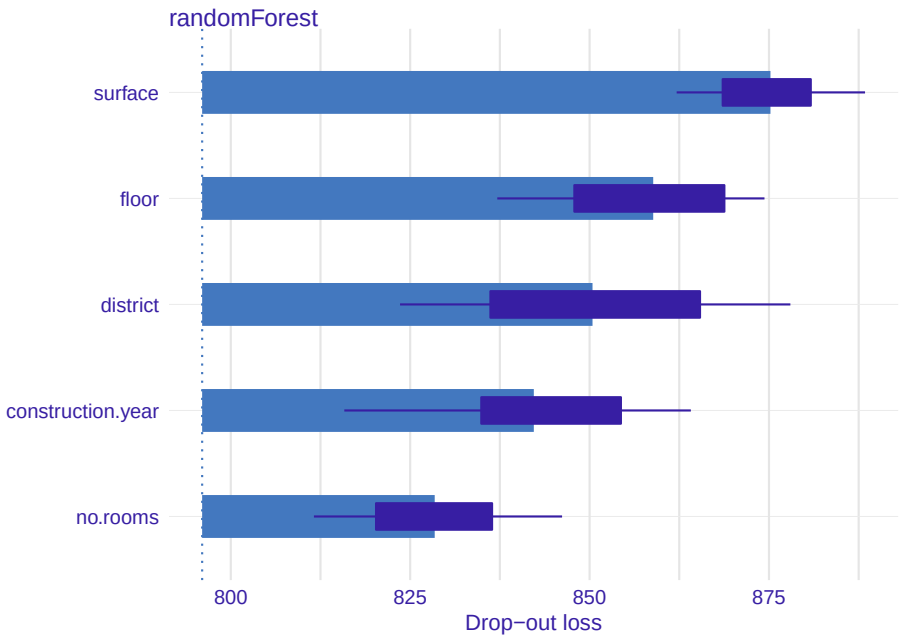
```
##   A new explainer has been created!
```

```
vip <- variable_importance(explainer_rf,
            loss_function = loss_root_mean_square)
vip
```

```
##             variable mean_dropout_loss        label
## 1      _full_model_          796.0100 randomForest
## 2          no.rooms          828.4179 randomForest
## 3 construction.year          842.2287 randomForest
## 4          district          850.4096 randomForest
## 5             floor          858.8663 randomForest
## 6           surface          875.2063 randomForest
## 7        _baseline_         1118.4724 randomForest
```

On a diagnostic plot is useful to present feature importance as an interval that start in a loss and ends in a loss of perturbed data.

```
plot(vip)
```



0.17.7 More models

[TOMASZ: WE SHOULD ONLY USE MODELS THAT WERE INTRODUCED EARLIER.]

Much more can be read from feature importance plots if we compare models of a different structure. Let's train three predictive models trained on `apartments` dataset from the `DALEX` package. Random Forest model (Breiman et al., 2018) (elastic but biased), Support Vector Machines model (Meyer et al., 2017) (large variance on boundaries) and Linear Model (stable but not very elastic). Presented examples are for regression (prediction of square meter price), but the CP profiles may be used in the same way for classification.

Let's fit these three models.

```
library("DALEX")
model_lm <- lm(m2.price ~ construction.year + surface + floor +
                      no.rooms + district, data = apartments)

library("randomForest")
set.seed(59)
model_rf <- randomForest(m2.price ~ construction.year + surface + floor +
                      no.rooms + district, data = apartments)

library("e1071")
model_svm <- svm(m2.price ~ construction.year + surface + floor +
                        no.rooms + district, data = apartments)
```

For these models we use `DALEX` explainers created with `explain()` function. These explainers wrap models, predict functions and validation data.

```
explainer_lm <- explain(model_lm,
                      data = apartmentsTest[,2:6], y = apartmentsTest$m2.price,
                             colorize = FALSE)
```

```
## Preparation of a new explainer is initiated
##   -> model label       :  lm  (  default  )
```

```
##   -> data              :  9000  rows  5  cols
##   -> target variable   :  9000  values
##   -> model_info        :  package stats , ver. 3.6.1 , task regression (  default  )
##   -> predict function  :  yhat.lm  will be used (  default  )
##   -> predicted values  :  numerical, min =  1792.597 , mean =  3506.836 , max =  6241.447
##   -> residual function :  difference between y and yhat (  default  )
##   -> residuals         :  numerical, min =  -257.2555 , mean =  4.687686 , max =  472.356
##   A new explainer has been created!
```

```
vip_lm <- variable_importance(explainer_lm,
            loss_function = loss_root_mean_square)
vip_lm
```

```
##            variable mean_dropout_loss label
## 1      _full_model_          281.8345    lm
## 2 construction.year          281.7864    lm
## 3          no.rooms          293.7945    lm
## 4             floor          486.0535    lm
## 5           surface          614.4047    lm
## 6          district         1018.8827    lm
## 7        _baseline_         1262.6592    lm
```

```
explainer_rf <- explain(model_rf,
                        data = apartmentsTest[,2:6], y = apartmentsTest$m2.price,
                               colorize = FALSE)
```

```
## Preparation of a new explainer is initiated
##   -> model label       :  randomForest  (  default  )
##   -> data              :  9000  rows  5  cols
##   -> target variable   :  9000  values
##   -> model_info        :  package randomForest , ver. 4.6.14 , task regression (  default
##   -> predict function  :  yhat.randomForest  will be used (  default  )
##   -> predicted values  :  numerical, min =  1977.609 , mean =  3511.789 , max =  5839.917
##   -> residual function :  difference between y and yhat (  default  )
##   -> residuals         :  numerical, min =  -1970.395 , mean =  -0.2658962 , max =  1944.71
##   A new explainer has been created!
```

```
vip_rf <- variable_importance(explainer_rf,
            loss_function = loss_root_mean_square)
vip_rf
```

```
##            variable mean_dropout_loss        label
## 1      _full_model_          802.9422 randomForest
## 2          no.rooms          834.9660 randomForest
## 3 construction.year          851.9975 randomForest
## 4          district          852.5380 randomForest
## 5             floor          874.3987 randomForest
## 6           surface          880.9620 randomForest
## 7        _baseline_         1110.6190 randomForest
```

```
explainer_svm <- explain(model_svm,
                         data = apartmentsTest[,2:6], y = apartmentsTest$m2.price,
                                colorize = FALSE)
```

```
## Preparation of a new explainer is initiated
##   -> model label       :  svm  (  default  )
##   -> data              :  9000  rows  5  cols
##   -> target variable   :  9000  values
##   -> model_info        :  package e1071 , ver. 1.7.2 , task regression (  default  )
##   -> predict function  :  yhat.svm  will be used (  default  )
##   -> predicted values  :  numerical, min =  1692.954 , mean =  3493.954 , max =  6256.247
##   -> residual function :  difference between y and yhat (  default  )
##   -> residuals         :  numerical, min =  -1553.981 , mean =  17.56927 , max =  2452.467
##   A new explainer has been created!
```

```
vip_svm <- variable_importance(explainer_svm,
            loss_function = loss_root_mean_square)
vip_svm
```

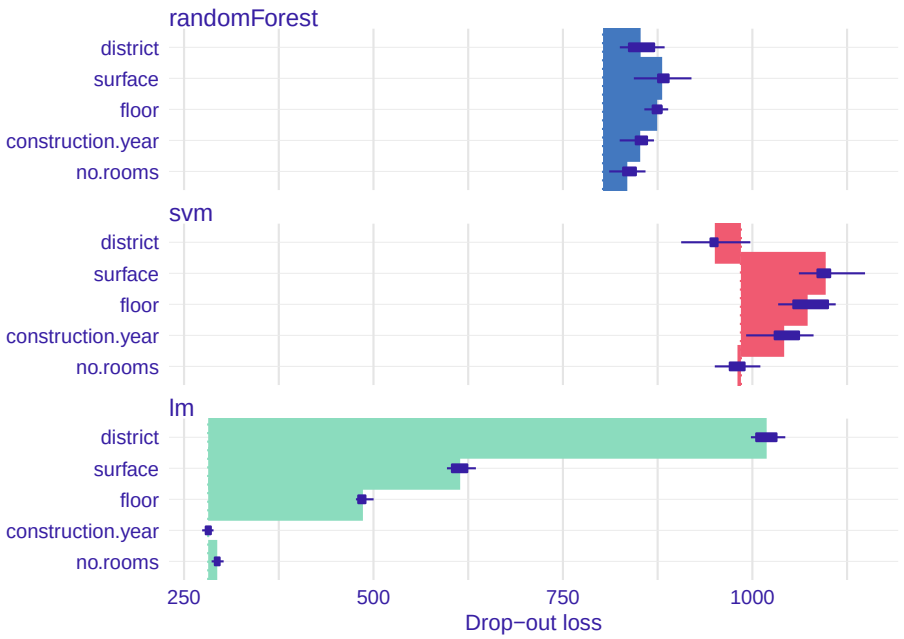
```
##            variable mean_dropout_loss label
## 1      _full_model_          984.9034   svm
## 2          district          950.4622   svm
## 3          no.rooms          980.3698   svm
## 4 construction.year         1041.9925   svm
## 5             floor         1072.9481   svm
## 6           surface         1096.7851   svm
## 7        _baseline_         1237.6861   svm
```

Let's plot feature importance for all three models on a single plot.

Intervals start in a different values, thus we can read that loss for SVM model is the lowest.

When we compare other features it looks like in all models the `district` is the most important feature followed by `surface` and `floor`.

```
plot(vip_rf, vip_svm, vip_lm)
```



There is interesting difference between linear model and others in the way how important is the `construction.year`. For linear model this variable is not importance, while for remaining two models there is some importance.

In the next chapter we will see how this is possible.

0.17.8 Level frequency

What does the feature importance mean? How it is linked with a data distribution.

0.18 Partial Dependency Profiles

0.18.1 Introduction

One of the first and the most popular tools for inspection of black-box models on the global level are Partial Dependence Plots (sometimes called Partial Dependence Profiles).

PDP were introduced by Friedman in 2000 in his paper devoted to Gradient Boosting Machines (GBM) - new type of complex yet effective models (Friedman, 2000). For many years PDP as sleeping beauties stay in the shadow of the boosting method. But this has changed in recent years. PDP are very popular and available in most of data science languages. In this chapter we will introduce key intuitions, explain the math beyond PDP and discuss strengths and weaknesses.

General idea is to show how the expected model response behaves as a function of a selected feature. Here the term „expected" will be estimated simply as the average over the population of individual Ceteris Paribus Profiles introduced in Chapter 0.11.

0.18.2 Intuition

Ceteris paribus profiles intoriduces in the Section 0.11 show profile of model response for a single observation. Partial dependency profile is an average profile for all observations.

For additive models all ceteris paribus profiles are parallel. Same shape, just shifted up or down. But for complex models these profiles may be different. Still, the average will be some crude summary how (in general) the model respond for changes in a given variable.

Left panel in the figure 43 shows ceteris paribus profiles for 25 sample observations for Titanic data for random forest model

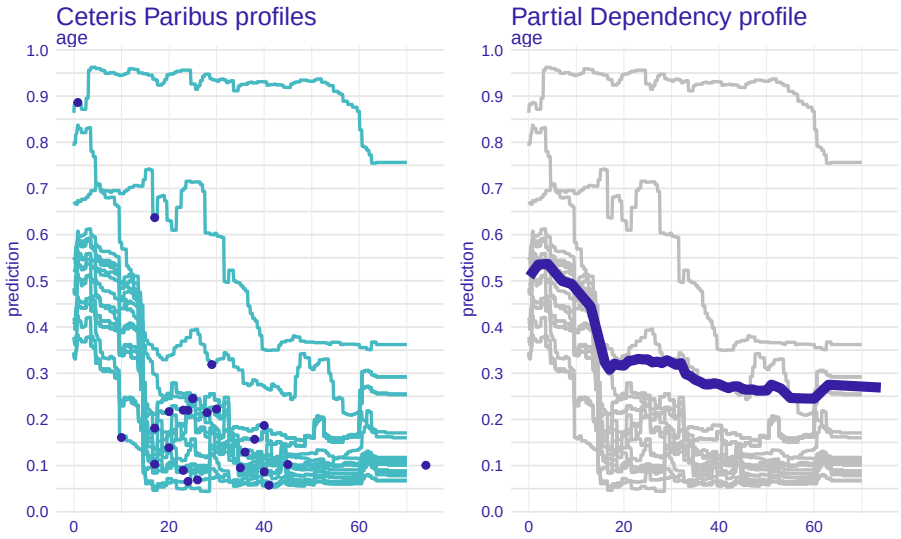


FIGURE 43 (fig:pdpIntuition) Left panel: Ceteris Paribus profiles for selected 25 observations. Blue points stand for selected observations while cyan lines stand for ceteris paribus profiles. Right panel: Grey lines stand for Ceteris paribus profiles as presented in left panel, blue line stands for its average - Partial dependency profile

`titanic_rf_v6`. The right panels show the average over CP profiles - it's an estimate of partial dependency profile

0.18.3 Method

0.18.3.1 Partial Dependency Profiles

Partial Dependency Profile for for a model f and a variable x^j is defined as

$$g_{PD}^{f,j}(z) = E[f(x^j = z, X^{-j})] = E[f(x|^j = z)]. \tag{0.33}$$

So it's an expected value for $x^j = z$ over **marginal** distribution

X^{-j} or equivalently expected value of f after variable x^j is set to z.

The expectation cannot be calculated directly as we do not know fully neither the distribution of X^{-j} nor the analytical formula of f. Yet this value may be estimated by as average from CP profiles.

$$\hat{g}^{f,j}_{PD}(z) = \frac{1}{n}\sum_{i=1}^{N} f(x_i^j = z, x_i^{-j})] = \frac{1}{n}\sum_{i=1}^{N} f(x_i|^j = z). \tag{0.34}$$

This formula comes from two steps.

1. Calculate ceteris paribus profiles for observations from the dataset.

As it was introduced in 0.11 ceteris paribus profiles show how model response change is a selected variable in this observation is modified.

$$h^{f,j}_x(z) := f(x|^j = z).$$

So for a single model and a single variable we get a bunch of *what-if* profiles. In the figure **??** we show an example for 100 observations. Despite some variation (random forest are not as stable as we would hope) we see that most profiles are decreasing. So the older the passengers is the lower is the survival probability.

2. Aggregate Ceteris Paribus into a single Partial Dependency Profile

Simple pointwise average across CP profiles. If number of CP profiles is large, it is enough to sample some number of them to get resonably accurate PD profiles. This way we get the formula (0.34).

0.18.3.2 Clustered Partial Dependency Profiles

Partial pependency profile is a good summary if ceteris paribus profiles have similar shape, i.e. are parallel. But it may happen that the variable of interest is in interaction with some other variable. Not all profiles are parallel because the effect of variable of interest depends on some other variables.

If individual profiles have different shapes then simple average may be misleading. To deal with this problem we propose to cluster Ceteris Paribus profiles and calculate average aggregate separately for each cluster.

The most straightforward approach would be to use a method for clustering and see how these cluster of profiles behave. For clustering one may use standard algorithm like k-means or hierarchical clustering. Once clusters are established we can aggregate within clusters in the same way as in case of Ppartial dependency plots.

So for a single model and a single variable we get k profiles average withing clusters. See an example in Figure 44 created for random forest model. It is easier to notice that ceteris paribus profiles can be grouped in three clusters. Group of passengers with a very large drop in the survival (cluster 1), moderate drop (cluster 2) and almost no drop in survival (cluster 3). Here we do not know what other factors are linked with these clusters, but some additional exploratory analysis can be done to identify these factors.

0.18.3.3 Grouped Partial Dependency Profiles

Once we see that variable of interest may be in interaction with some other variable, it is tempting to look for the factor that distinguish clusters.

The most straightforward approach is to use some other variable as a grouping variable. Instead of clustering we may aggregate groups of CP profiles defined a a selected variable of itnerest.

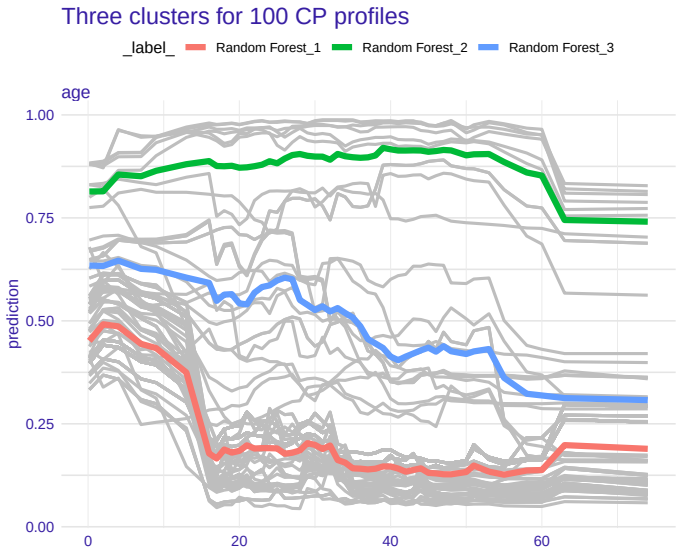


FIGURE 44 (fig:pdpPart4) Grey lines stand for ceteris paribus profiles for 100 sample observations. These profiles were clusterd into 3 groups and blue, green and red lines show corresponding averages

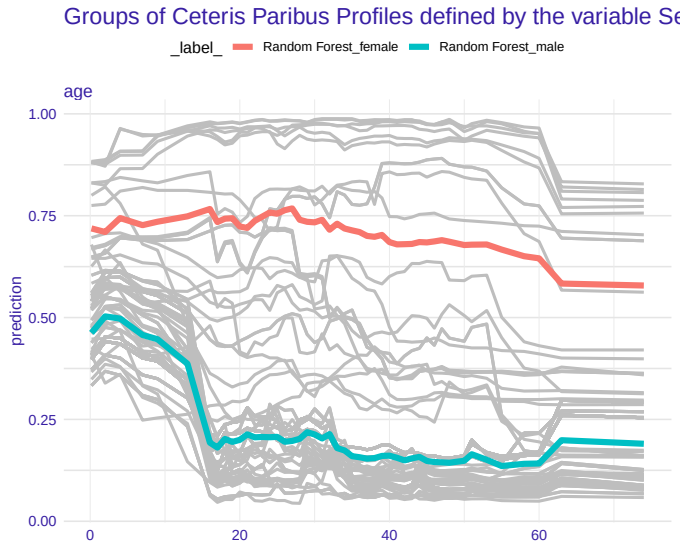


FIGURE 45 Grouped profiles with respect to the gender variable

See an example in Figure 45. PD profiles are calculated separately for each gender. Clearly there is an interaction between Age and Sex. The survival for woman is more stable, while for man there is more sudden drop in Survival for older passengers.

0.18.3.4 Contrastive Partial Dependency profiles

In previous sections we compared PD profiles calculated for a single models but in groups either defined via clustering or via some dependent variable. Comparison of such aggregates overlayed in a single plot may be very insightful. Contrastive comparisons of Partial Dependency Plots are useful not only for subgroups of observations but also for comparisons of different models.

Why one would like to compare models? There are at least three reasons for it.

- *Agreement of models will be reassuring.* Some models are known to be more stable other to be more elastic. If profiles for models

from these two classes are not far from each other we can be more convinced that elastic model is not over-fitted.

- *Disagreement of models suggest how to improve one of them.* If simpler interpretable model disagree with an elastic model, this may suggest a feature transformation that can be used to improve the interpretable model. For example if random forest learned non linear relation then it can be captures by a linear model after suitable transformation.
- *Validation of boundary conditions.* Some models are know to have different behavior on the boundary, for largest or lowest values. Random forest is known to shrink predictions towards the average, while support vector machines are known to have larger variance at edges. Contrastive comparisons may help to understand differences in boundary behavior.

See an example in Figure 46. Random forest model is compared with generalized linear model (logistic regression) with splines. Both models agree when it comes to a general relation between Age and chances of survival (the younger the better) but the curve for random forest is more flat. Difference between both models is largest for lowest values of the variable age. This observation is along out expectations that random forest model in general shrink towards an average and is not so good for interpolation outside the training domain.

0.18.4 Example: Apartments data

In this section we will use random forest model `apartments_rf_v5` trained on `apartments` data in order to predict the price per square meter of an apartment. See section 0.5.2.3 for more details. This example is focused on two dependent variables `surface` and `construction.year`.

0.18.4.1 Partial Dependency Profiles

Figure 47 presents CP profiles for 25 sample apartments along with the average PD profile. It is interesting to see that relation

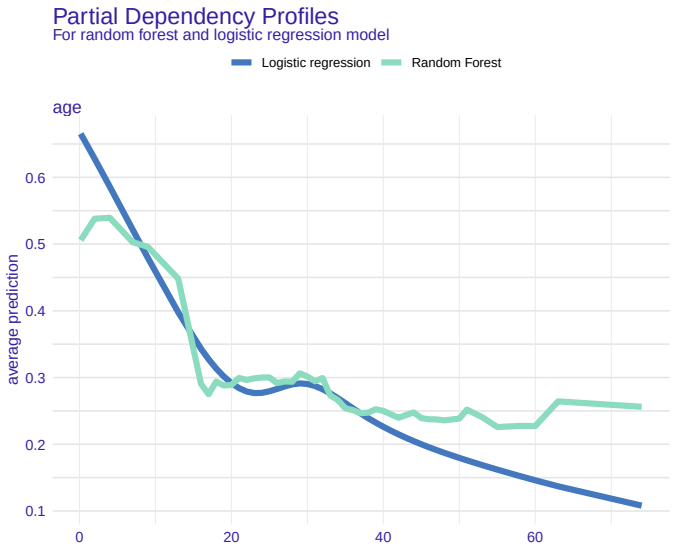


FIGURE 46 Comparison of two predictive models with different structures traind on the same dataset 'titanic'.

between `surface` and the target variable is almost linear while relation between `construction.year` and the target variable is U-shaped. The most expensive are apartments very new or very old. The data is artificial but it was constructed in a way to reassemble effect of lower quality of building materials used in housing construction after II world war.

0.18.4.2 Clustered Partial Dependency Profiles

A natural question would be to ask if the U-shape response profile for construction year is typical for all observations. Figure 48 shows average profiles in three clusters derived from the CP profiles.

Averages in clusters differ slightly in the size of oscillations, but all three shapes are similar. So far we do not have reasons to expect strong interactions in the model.

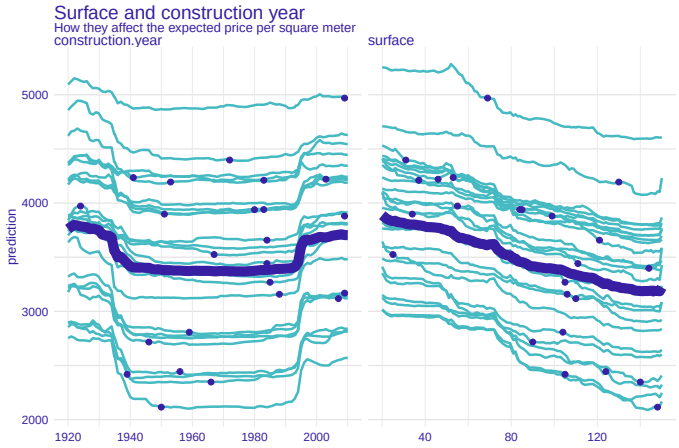


FIGURE 47 Ceteris Paribus profiles for 25 sample apartments and the partial dependency profile for the random forest model

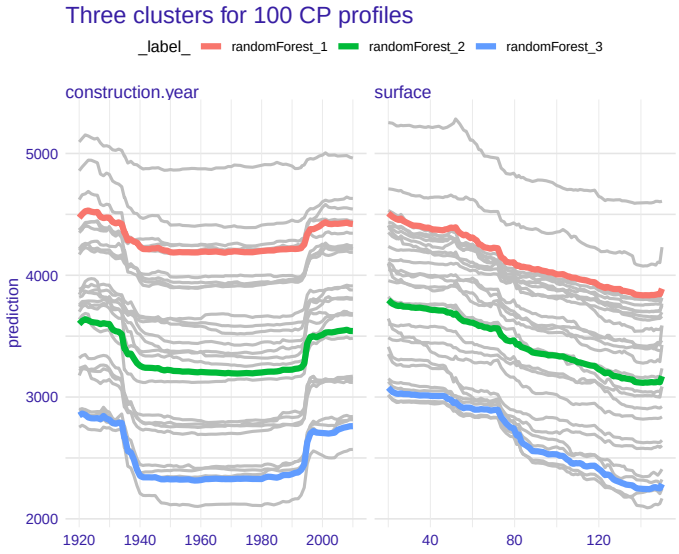


FIGURE 48 (fig:pdpApartment1clustered) Grey lines stand for ceteris paribus profiles for 25 sample observations. These profiles were clusterd into 3 groups and blue, green and red lines show corresponding averages

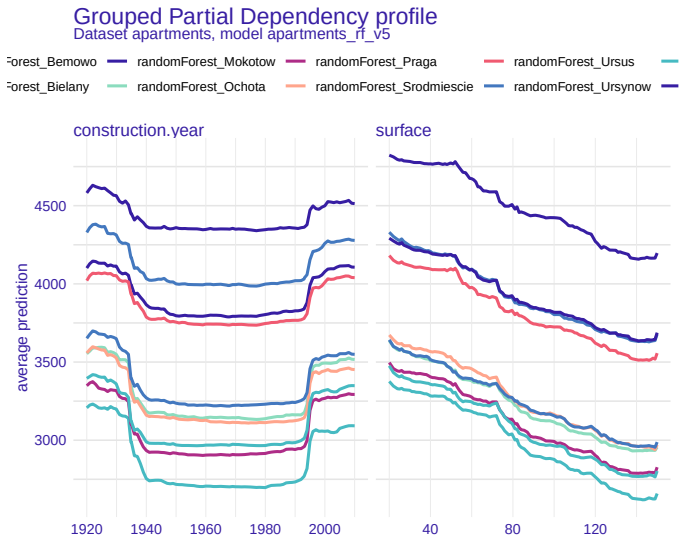


FIGURE 49 Partial dependency profiles calculated for separate districts.

0.18.4.3 Grouped Partial Dependency Profiles

One of categorical variables in the `apartments` dataset is the `district`. In this subsection we will check if the model behavior is similar for all districts. To to this we will calculate average ceteris paribus profiles for each district separately.

Figure 49 shows PD profiles calculated independently for each district. There are some interesting things to see. First is that some profiles are higher than others, so for example apartments in `Srodmiescie` (downtown) are more expensive than in other districts. Second observation is that profiles are parallel, thus the effect of surface and construction year are similar in each district. Third is that these profiles constitute three groups of districts, the `Srodmiescie` (downtown), followed by three districts close to `Srodmiescie` (namely `Mokotow`, `Ochota` and `Ursynow`) followed by all other districts.

0.18.4.4 Contrastive Partial Dependency profiles

One of the biggest challenges in modeling for complex model is if the model structure is flexible enough to capture relations present in the data, but not too flexible to avoid over fitting.

One approach to investigate this direction is to compare what has been learned by models with different structures. For example, figure 50 shows PD profiles calculated for linear model and random forest model.

Here the story is very interesting. The linear model cannot of course capture the non monotonic relation between `construction.year` and the price per square meter. In case of the `surface` variable both models captured linear relation, but the one derived by `lm` model is steeper. It is expected for the random forest model to be biased towards the mean.

So one may say that both models missed something because of their structure. Linear model missed the U-shaped relation between construction year and apartment price while the random forest model shrink too much the effect of the surface over the apartment price. Both these observations lead to the conclusion that we could build a better model that will capture both these relations.

0.18.5 Pros and cons

Partial Dependency profiles, as presented in this chapter, offer a simple way to summaries an effect of a particular variable on the model response.

This method has numerous advantages. Just to name a few

- Partial Dependency profiles are quite popular and are implemented in variety of packages for R, python or other languages
- Partial Dependency profiles are easy to explain and intuitive,
- It is easy to extend PD profiles for different models or groups of observations.

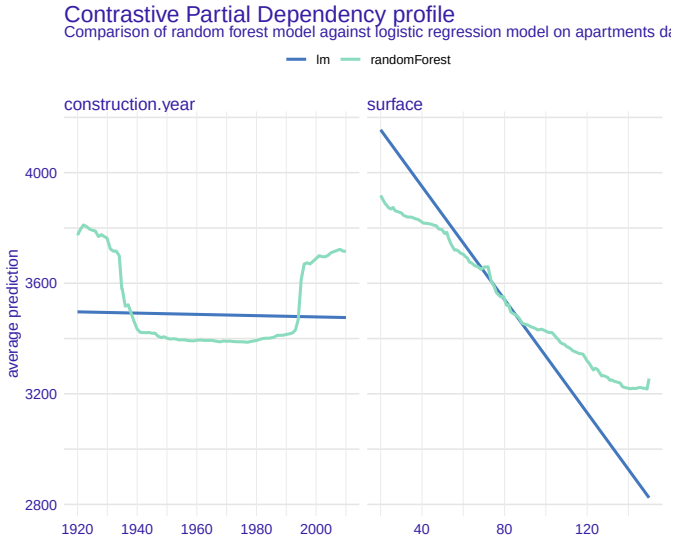


FIGURE 50 (fig:pdpApartment3) Comparison of PD profiles for linear model and random forest model.

Yet there are also some disadvantages. They are mostly inherited from ceteris paribus profiles that are being aggregated.

- For correlated features the rule „all other things being constant" makes no sense. For the dataset `apartments` changes in `surface` should go along with changes in `number of rooms`. This issue will be discussed in the next chapter.
- For non additive models the average across ceteris paribus profiles may be a crude and misleading simplification.

0.18.6 Code snippets for R

Here we show partial dependency profiles calculated with `ingredients` package (Biecek et al., 2019). You will also find similar functions in the `pdp` package (Greenwell, 2017), `ALEPlots` package (Apley, 2018) or `iml` (Molnar et al., 2018) package.

The easiest way to calculate PD profiles is to use the function `ingredients::partial_dependency`. The only required argument is

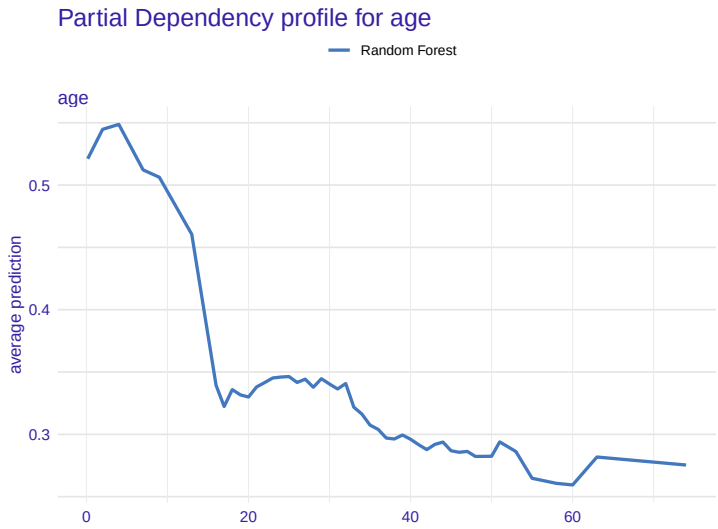


FIGURE 51 Partial Dependency profile for age.

the explainer and by default PD profiles are calculated for all variables.

Below we use `variables` argument to limit list of variables for which PD profiles are calculated. Here we need profiles only for the `age` variable.

```
pdp_rf <- partial_dependency(explain_titanic_rf, variables = "age")
plot(pdp_rf) +
  ggtitle("Partial Dependency profile for age")
```

PD profiles can be plotted on top of standard CP profiles. This is a very useful feature if we want to know how crude is the averaging and how similar are individual profiles to the average.

```
selected_passangers <- select_sample(titanic, n = 25)
cp_rf <- ceteris_paribus(explain_titanic_rf, selected_passangers, variables = "age")

plot(cp_rf, variables = "age") +
  show_aggregated_profiles(pdp_rf, variables = "age", size = 3) +
  ggtitle("Ceteris Paribus and Partial Dependency profiles for age")
```

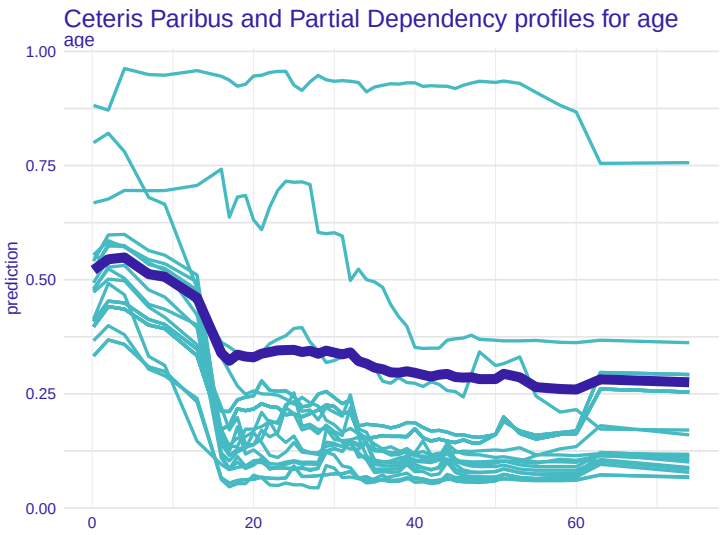


FIGURE 52 Ceteris Paribus and Partial Dependency profiles for age.

0.18.6.1 Clustered Partial Dependency profiles

In order to calculate clustered profiles we need to first calculate CP profiles with the `ceteris_paribus` function.

Then we can use the `cluster_profiles` function, which performs k-means clustering in ceteris paribus profiles.

The clustered profiles can be the plotted with the `plot` function.

```
selected_passangers <- select_sample(titanic, n = 100)
cp_rf <- ceteris_paribus(explain_titanic_rf, selected_passangers, variables = "age")

clust_rf <- cluster_profiles(cp_rf, k = 3, center = TRUE)

plot(cp_rf, color = "grey") +
  show_aggregated_profiles(clust_rf, size = 2, color = "_label_") +
  ggtitle("Clustered Partial Dependency profiles.")
```

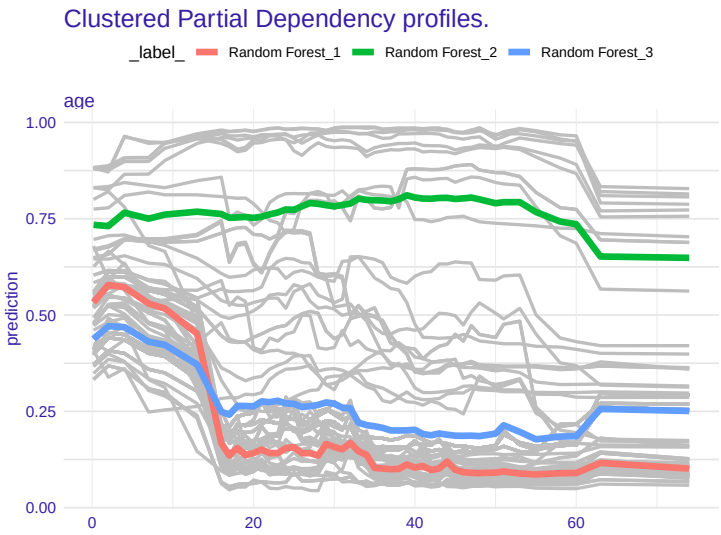


FIGURE 53 Clustered Partial Dependency profiles.

0.18.6.2 Grouped Partial Dependency profiles

The `partial_dependency` has argument `groups`. It is enough to set this argument to some categorical variable to calculate and plot Grouped Partial Dependency profiles.

In the example below we plot groups separately for each gender.

```
pdp_sex_rf <- partial_dependency(cp_rf, variables = "age",
                groups = "gender")

plot(cp_rf, variables = "age") +
  show_aggregated_profiles(pdp_sex_rf, variables = "age", size = 3) +
  ggtitle("Grouped Partial Dependency profiles")
```

0.18.6.3 Contrastive Partial Dependency profiles

As in previous functions, in order to overlay explanations for two or model models in a single plot one can use the generic `plot()` function.

In the example below we create PD profiles for `explain_titanic_rf`

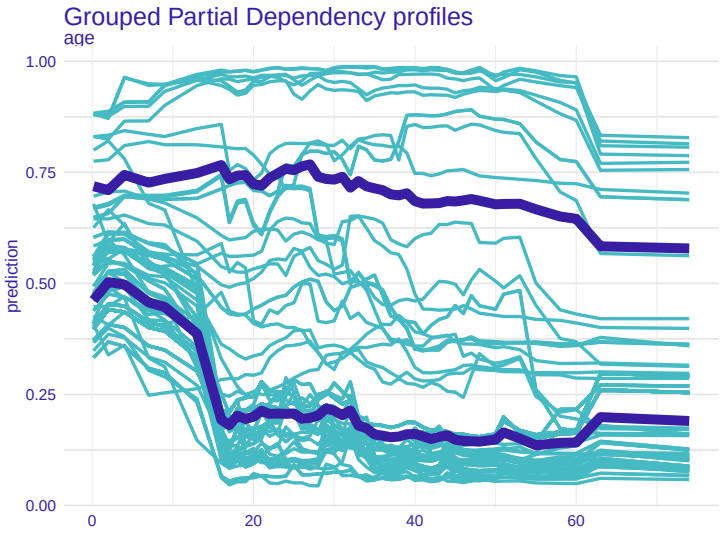


FIGURE 54 Grouped Partial Dependency profiles.

and `explain_titanic_lmr` models and then they are plotted together in a single plot.

```
pdp_rf <- partial_dependency(explain_titanic_rf, variables = "age")
pdp_lmr <- partial_dependency(explain_titanic_lmr, variables = "age")

plot(pdp_rf, pdp_lmr) +
  ggtitle("Contrastive Partial Dependency profiles")
```

0.19 Accumulated Local Profiles

0.19.1 Introduction

One of the largest advantages of the Partial Dependency Profiles is that they are easy to explain, as they are just an average across Ceteris Paribus profiles. But one of the largest disadvantages lies in expectation over marginal distribution which implies that x^j is independent from x^{-j}. In many

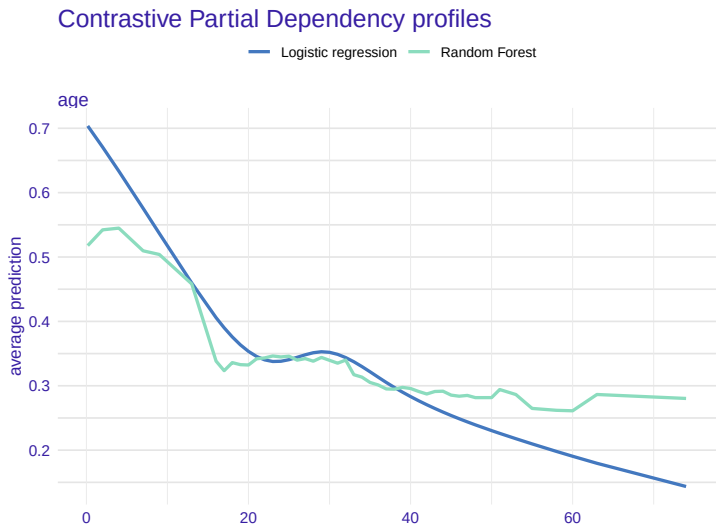


FIGURE 55 Contrastive Partial Dependency profiles.

applications this assumption is violated. For example, for the `apartments` dataset one can expect that features like `surface` and `number.of.rooms` are strongly correlated as apartments with larger number of rooms usually have larger surface. It may makes no sense to consider an apartment with 10 rooms and 20 square meters, so it may be misleading to change $x^{surface}$ independently from $x^{number.of.rooms}$. In the `titanic` dataset we shall expect correlation between `fare` and `passenger class` as tickets in the 1st class are the most expensive.

There are several attempts to fix this problem. In this chapter we present two of them. Local Dependency Profiles and Accumulated Local Profiles, both introduces in the (Apley, 2018). The general idea behind Local Dependency Profiles is to use conditional distribution instead of marginal distribution to accommodate for the dependency between x^j and x^{-j}. The general idea behind Accumulated Local Profiles is to accumulate local changes in model response affected by single feature x^j.

0.19.2 Intuition

Intuition behind Partial Dependency profiles and their extensions is presented in Figure 56.

First, let's consider a simple model

$$f(x_1, x_2) = x_1 * x_2 + x_2 \tag{0.35}$$

Moreover, let's assume that variables x_1 and x_2 have uniform distribution $x_1, x_2 \sim U[-1, 1]$ and are perfectly correlated, i.e. $x_2 = x_1$.

For this example, suppose that we have dataset with 8 points.

i	1	2	3	4	5	6	7	8
x_1	-1	-0.71	-0.43	-0.14	0.14	0.43	0.71	1
x_2	-1	-0.71	-0.43	-0.14	0.14	0.43	0.71	1

Panel A in Figure 56 shows ceteris paribus profiles calculated for selected 8 points.

Bottom part of the panel B shows Partial Dependency profile. It's an average from all ceteris paribus profiles (as shown in the top panel).

The idea behind extensions of partial dependency profiles is to use not all profiles, but only parts that are relevant (as shown in the top panels). Local Dependency Profiles (panel C) are calculated as averages from these selected relevant parts of ceteris paribus profiles. Accumulated Local Profiles (panel D) are calculated as accumulated changes from these selected relevant parts of ceteris paribus profiles.

For example, for the `apartments` dataset one can expect that features like `surface` and `number.of.rooms` are correlated but we can also imagine that each of these variables affect the apartment

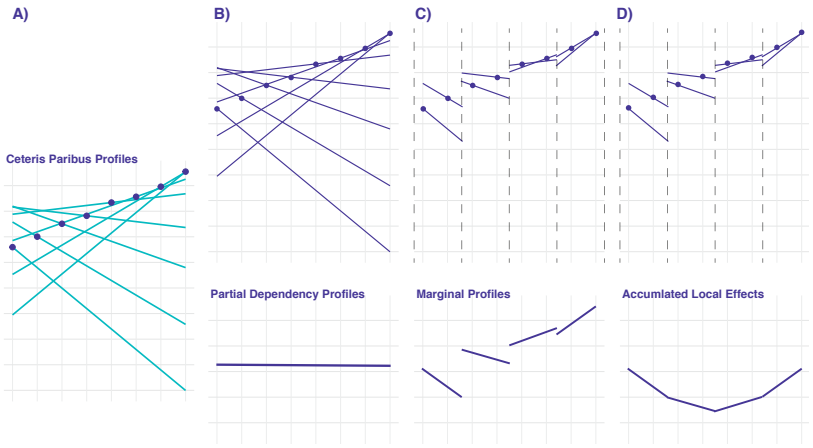


FIGURE 56 (fig:accumulatedLocalEffects) Differences between Partial Dependency, Marginal and Accumulated Local Effects profiles. Panel A) shows Ceteris Paribus Profiles for 8 points. Panel B) shows Partial Dependency profiles, i.e. an average out of these profiles. Panel C shows Marginal profiles, i.e. an average from profiles similar to the point that is being explained. Panel D shows Accumulated Local Effects, i.e. effect curve that takes into account only changes in the Ceteris Paribus Profiles.

price somehow. Partial Dependency Profiles show how the average price changes as a function of surface, keeping all other variables unchanged. Conditional Dependency Profiles show how the average price changes as a function of surface adjusting all other variables to the current value of the surface. Accumulated Local Profiles show how the average price changes as a function of surface adjusting all other variables to the current value of the surface but extracting changes caused by these other features.

0.19.3 Method

0.19.3.1 Partial Dependency Profile

Partial Dependency Profile is defined as an expected value from Ceteris Paribus Profiles.

$$g_i^{PD}(z) = E_{X_{-i}}[f(x|^i = z, x^{-i})]. \tag{0.36}$$

And can be estimated as average from CP profiles.

$$\hat{g}_i^{PD}(z) = \frac{1}{n}\sum_{j=1}^{n} f(x|^i = z, x_j^{-i}). \tag{0.37}$$

As it is shown in Figure 56 panel B, PD profiles are averages from all CP profiles.

0.19.3.2 Conditional Dependency Profile

As it was said, if there is some dependency between X_i and X_{-i} it makes no sense to average CP profiles over marginal X_{-i} bacuse „all other things kep unchanges" is not a reliable approach. Instead, an intuitive approach would to use a conditional distribution $X_{-i}|X_i = x_i$ (which is of course unknown).

Conditional Dependency Profile for a model f and a variable x^j is defined as

$$g_{f,i}^{CD}(z) = E_{X_{-i}|X_i=x_i}[f(x|^i = z, x^{-i})]. \tag{0.38}$$

So it's an expected value over **conditional** distribution $(X^j, X^{-j})|X^j = z$.

For example, let $f(x_1, x_2) = x_1 + x_2$ and distribution of (x_1, x_2) is given by $x_1 \sim U[0, 1]$ and $x_2 = x_1$. In this case $g_{f,1}^{CD}(z) = 2 * z$.

The natural estimator for Conditional Dependency Profiles introduced in (Apley, 2018) is

$$\hat{g}_i^{CD}(z) = \frac{1}{|N_i|} \sum_{j \in N_i} f(x|^i = z, x_j^{-i}). \tag{0.39}$$

where N_i is the set of observations with x_i close to z. This set will be used to estimate distribution $X_{-i}|X_i = x_i$.

In Figure 56 panel C the range of variable x_i is divided into 4 separable intervals. The set N_i contains all observations that fall into the same interval as observation x_i. The final CD profile is an average from closest pieces of CP profiles.

Note that in general the $\hat{g}_i^{CD}(z)$ is neither smooth, nor continuous in boundaries between N_i subsets. Thus here we propose another smooth estimator for g_i^{CD}

$$\tilde{g}_i^{CD}(z) = \frac{1}{\sum_k w_k(z)} \sum_{j=1}^{n} w_j(z) f(x|^i = z, x_j^{-i}). \tag{0.40}$$

Weights $w_j(z)$ correspond to the distance between z and observation x_j. For categorical variables we may use simple

indicator function $w_j(z) := 1_{z==x_j^i}$ while for continuous variables we may use Gaussian kernel

$$w_j(z) = \phi(z - x_j^i; 0; s),$$

where s is a smoothing factor.

0.19.3.3 Accumulated Local Profile

Accumulated Local Profile for a model f and a variable x^j is defined as

$$g_{f,j}^{ALE}(z) = \int_{z_0}^{z} E\left[\frac{\partial f(X^j, X^{-j})}{\partial x_j} | X^j = v\right] dv + c, \tag{0.41}$$

where z_0 if the lower boundary of x^j. The profile $g_{f,j}^{ALE}(z)$ is calculated up to some constant c. Usually the constant c is selected to set average $g_{f,j}^{ALE}$ equal to 0 or an average of $f(x)$.

The equation may be a bit complex, but the intuition is not that complicated. Instead of averaging Ceteris Paribus profiles we just look locally how quickly local CP profiles are changing. And ALE profile is reconstructed from such local partial changes as cumulative derivative over changes.

So it's a cumulated expected change of the model response along where the expected values are calculated over **conditional** distribution $(X^j, X^{-j})|X^j = z$.

For example, let $f(x_1, x_2) = x_1 + x_2$ and distribution of (x_1, x_2) is given by $x_1 \sim U[0, 1]$ and $x_2 = x_1$. In this case $g_{f,1}^{ALE}(z) = z$.

The natural estimator for Accumulated Local Profiles introduced in (Apley, 2018) is

$$\hat{g}_i^{ALE}(z) = \sum_{k=1}^{k_i(z)} \frac{1}{|N_i(k)|} \sum_{j \in N_i} \left[f(x|^i = z_k) - f(x|^i = z_{k-1})\right] + c \tag{0.42}$$

where $k_i(z)$ is the index of interval with point z, N_i is the set of observations with x_i in this interval. The difference $f(x|^i = z_k) - f(x|^i = z_{k-1})$ correspond to the difference of CP profiles in interval k, and this difference is averaged and accumulated.

In Figure 56 panel D the range of variable x_i is divided into 4 separable intervals. The set N_i contains all observations that fall into the same interval as observation x_i. The final ALE profile is constructed from accumulated differences of local CP profiles.

Note that in general the $\hat{g}_i^{CD}(z)$ is not smooth in boundaries between N_i subsets. Thus here we propose another smooth estimator for g_i^{ALE}

$$\tilde{g}_i^{ALE}(z) = \sum_{k \in z_0,...,z} \frac{1}{\sum_j w_j(k)} \sum_{j=1}^{n} w_j(k) \left[f(x|^i = z_k) - f(x|^i = z_k - \Delta) \right] + c \tag{0.43}$$

The set $z_0, ..., z$ is a uniform grid of points between z_0 and z with the step Δ. Weights $w_i(k)$ correspond to the distance between point k and observation x_i. For categorical variables we may use simple indicator function $w_j(k) := 1_{k==x_j^i}$ while for continuous variables we may use Gaussian kernel

$$w_j(k) = \phi(k - x_j^i; 0; s),$$

where s is a smoothing factor.

0.19.3.4 Comparison of Explainers for Feature Effects

In previous sections we introduced different was to calculate model level explainers for feature effects. A natural question is how these approaches are different and which one should we choose.

An example that illustrate differences between these approaches is presented in Figure 56. Here we have a model

$f(x_1, x_2) = x_1 * x_2 + x_2$ and what is important features are correlated $x_1 \sim U[-1, 1]$ and $x_2 = x_1$.

Panel A) shows Ceteris Paribus for 8 data points, the feature x_1 is on the OX axis while f is on the OY. Panel B) shows Partial Dependency Profiles calculated as an average from CP profiles.

$$g_{f,1}^{PD}(z) = E[z * x^2 + x^2] = 0$$

Panel C) shows Conditional Dependency Profiles calculated as an average from conditional CP profiles. In the figure the conditioning is calculated in four bins, but knowing the formula for f we can calculated it directly as.

$$g_{f,1}^{CD}(z) = E[X^1 * X^2 + X^2 | X^1 = z] = z^2 + z$$

Panel D) shows Accumulated Local Effects calculated as accumulated changes in conditional CP profiles. In the figure the conditioning is calculated in four bins, but knowing the formula for f we can calculated it directly as.

$$g_{f,1}^{AL}(z) = \int_{z_0}^{z} E\left[\frac{\partial(X^1 * X^2 + X^2)}{\partial x_1} | X^1 = v\right] dv = \int_{z_0}^{z} E\left[X^2 | X^1 = v\right] dv = \frac{z^2 - 1}{2},$$

0.19.4 Example: Apartments data

In this section we will use random forest model `apartments_rf_v5` trained on `apartments` data in order to predict the price per square meter of an apartment. See section 0.5.2.3 for more details. This example is focused on two dependent variables `surface` and `no.rooms`. What is more important is that these two variables are correlated.

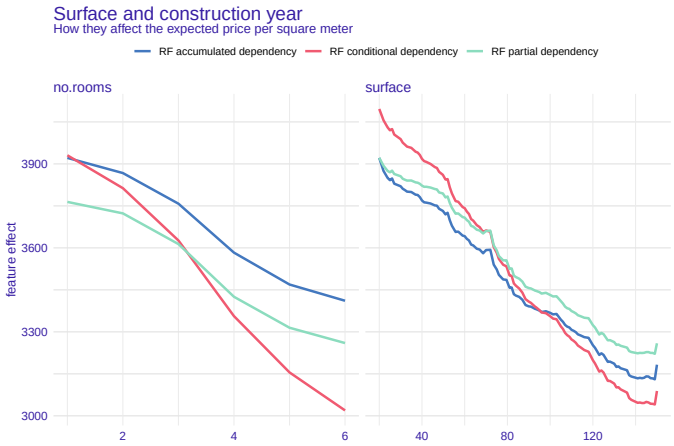


FIGURE 57 Partial Dependency, Conditional Dependency and Accumulated Local profiles for the random forest model and apartments data.

Figure 57 shows Partial Dependency, Conditional Dependency and Accumulated Local profiles for the random forest model.

Number of rooms and surface are two correlated variables, moreover both have some effect on the price per square meter. As we see profiles calculated with different methods are different. One we take into account the correlation between variables, the feature effects are less steep.

0.19.5 Pros and cons

In this chapter we introduced tools for extraction of the information between model response and individual model inputs. These tools are useful to summarize how „in general" model responds to the input of interest. All presented approaches are based on Ceteris Paribus Profiles introduced in Chapter 0.11 but they differ in a way how individual profiles are merged into a global model response.

We use the term „feature effect" to refer to global model response as a function of single or small number of model features.

Methods presented in this chapter are useful for extraction information of feature effect, i.e. how a feature is linked with model response. There are many possible applications of such methods, for example:

- Feature effect may be used for feature engineering. The crude approach to modeling is to fit some elastic model on raw data and then use feature effects to understand the relation between a raw feature and model output and then to transform model input to better fit the model output. Such procedure is called surrogate training. In this procedure an elastic model is trained to learn about link between a feature and the target. Then a new feature is created in a way to better utilized the feature in a simpler model (Gosiewska et al., 2019). In the next chapters we will show how feature effects can be used to transform a continuous variable in to a categorical one in order to improve the model behavior.
- Feature effect may be used for model validation. Understanding how a model utilizes a feature may be used as a validation of a model against domain knowledge. For example if we expect monotonic relation or linear relation then such expectations can be verified. Also if we expect smooth relation between model and its inputs then the smoothness can be visually examined. In the next chapters we will show how feature effects can be used to warn a model developer that model is unstable and should be regularized.
- In new domains an understanding of a link between model output and the feature of interest may increase our domain knowledge. It may give quick insights related to the strength or character of the relation between a feature of interest and the model output.
- The comparison of feature effects between different models may help to understand how different models handle particular features. In the next chapters we will show how feature effects can be used learn limitations of particular classes of models.

0.19.6 Code snippets for R

Here we show partial dependency profiles calculated with `ingredients` package (Biecek et al., 2019). You will also find similar functions in the `ALEPlots` package (Apley, 2018).

Partial dependency profiles can be calculated with the function `ingredients::partial_dependency`. Conditional dependency profiles can be calculated with the function `ingredients::conditional_dependency`. Accumulated local profiles can be calculated with the function `ingredients::accumulated_dependency`.

In all these cases the only required argument is the explainer and by default profiles are calculated for all variables.

Below we use `variables` argument to limit list of variables for which profiles are calculated. Here we need profiles only for the `no.rooms` and `surface` variables.

```
explain_apartments_rf <- explain(model_apartments_rf,
                                 data = apartments,
                                 verbose = FALSE)

pd_rf <- partial_dependency(explain_apartments_rf, variables = c("no.rooms", "surface"))

plot(pd_rf) + ylab("Partial dependency") +
  ggtitle("Surface and number of rooms", "Partial dependency for random forest model")
```

```
ac_rf <- accumulated_dependency(explain_apartments_rf, variables = c("no.rooms", "surface"))

plot(ac_rf) + ylab("Accumulated dependency") +
  ggtitle("Surface and number of rooms", "Accumulated dependency for random forest model")
```

```
pd_rf <- conditional_dependency(explain_apartments_rf, variables = c("no.rooms", "surface"))

plot(pd_rf) + ylab("Conditional dependency") +
  ggtitle("Surface and number of rooms", "Conditional dependency for random forest model")
```

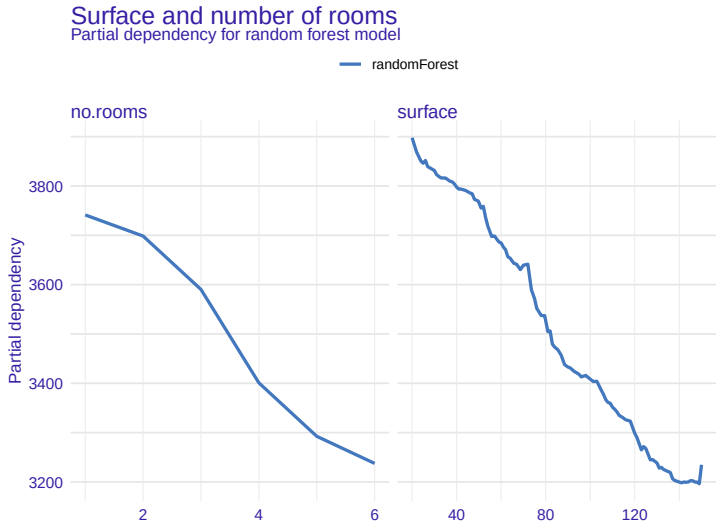


FIGURE 58 Partial Dependency profile for surface and number of rooms

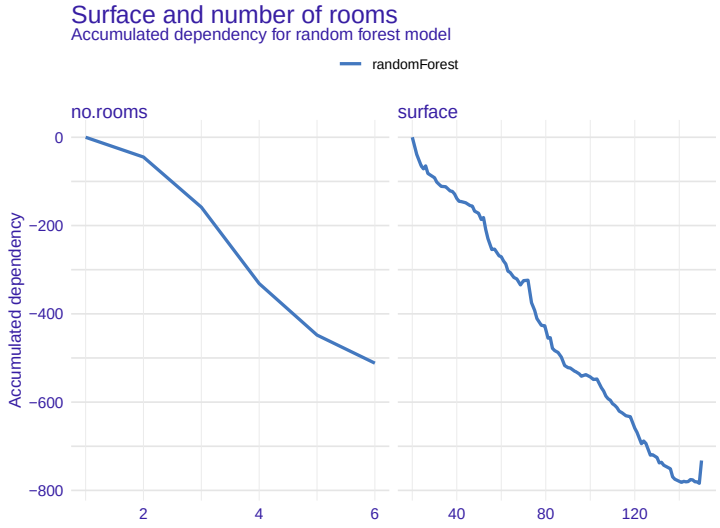


FIGURE 59 Accumulated dependency profile for surface and number of rooms

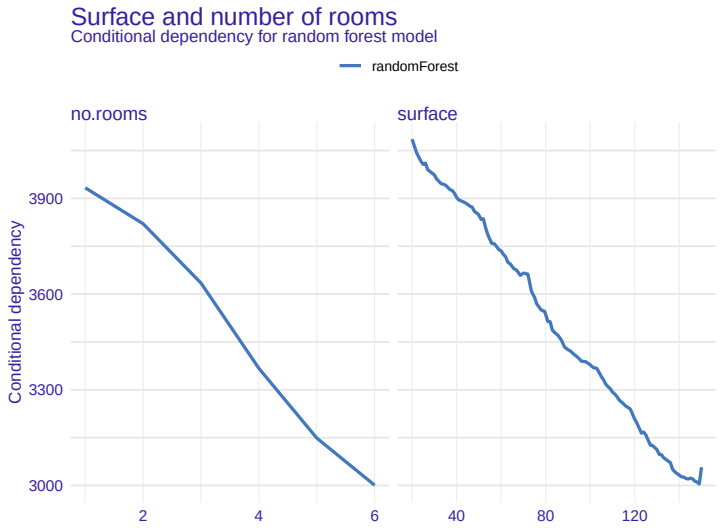


FIGURE 60 Conditional dependency profile for surface and number of rooms

0.20 Residual Diagnostic

0.20.1 Introduction

In this chapter, we present methods that are useful for the detailed examination of model residuals. These methods may be used for several purposes.

- Identification of hard cases; in the Section 0.16 we discussed measures to globally summarize the model performance, but sometimes we are more interested in cases with the largest mispredictions;
- Identification of structural problems in a model; for most models we assume that residuals are random. If we find any structure then maybe there is some problem with a model.
- Identification of cases for details local-level examination. In the first part of this book we discussed tools for examination of single predictions. For debugging purposes it make sense to first identify

largest errors and then use local-methods to understand which factors contributes most to these errors.

0.20.2 Intuition

As it was defined in Section 0.16, the residual r_i is the difference between model prediction and the true value of target variable

$$r_i = y_i - f(x_i).$$

For the perfect model we will expect that all residuals are equal to zero, but perfect models do not exists. For good models we assume that residuals are small, random and symmetric. In fact residuals may violate these assumptions in many different ways. So we need tools for exploration of residuals.

0.20.3 Code snippets for R

In this section, we present the key features of the `auditor` R package (**?**) which is a part of the DrWhy.AI universe. The package covers all methods presented in this chapter. It is available on CRAN and GitHub. More details and examples can be found at https://modeloriented.github.io/auditor/ or in (Gosiewska and Biecek, 2018).

First we load explainers for two models created in Section 0.5.2.5 for the `apartments` data.

```
library("DALEX")
library("auditor")
library("randomForest")

explainer_apartments_lr <- archivist:: aread("pbiecek/models/f49ea")
explainer_apartments_rf <- archivist:: aread("pbiecek/models/569b0")
```

For residual diagnostic we need to calculate residuals for both

explainers. This can be done with the `model_residual()` function. Now we are ready to explore residuals for both models for apartments dataset.

```
mr_lr <- model_residual(explainer_apartments_lr)
mr_rf <- model_residual(explainer_apartments_rf)
```

Figures 61 and 62 shows distribution of residuals for both models. As we know from the Section 0.16.4.1 the RMSE for both model is very similar. But when we compare distributions of residuals we see that these models are very different. The linear regression model tends to have residuals around +- 400 while for random forest model the residuals are on average equal to 0 but have large variation.

We know from previous chapters that the reason for the behavior of the linear model is that it does not capture the nonlinear relation between the price of apartment and the year of construction.

From these plots alone we see that random forest model has more frequently smaller residuals than the linear regression model. But for small fraction of observations residuals for random forest are very large and these extremes balance the RMSE.

```
plot_residual_density(mr_rf, mr_lr)
```

```
plot_residual_boxplot(mr_rf, mr_lr)
```

Figures 63 and 64 show diagnostic plots that link model predictions with other variables. In the first case it's a relation between the true (X axis) and predicted (Y axis) values. For perfect model we would expect a strait line. Here the model is biased towards the average, so we see that for large values of target variables the predictions are shifted towards the average. Same for very low values of target variable.

The second plot shows predictions as a function of the ordering of observations. If observations are randomly collected then we shall not see any relation.

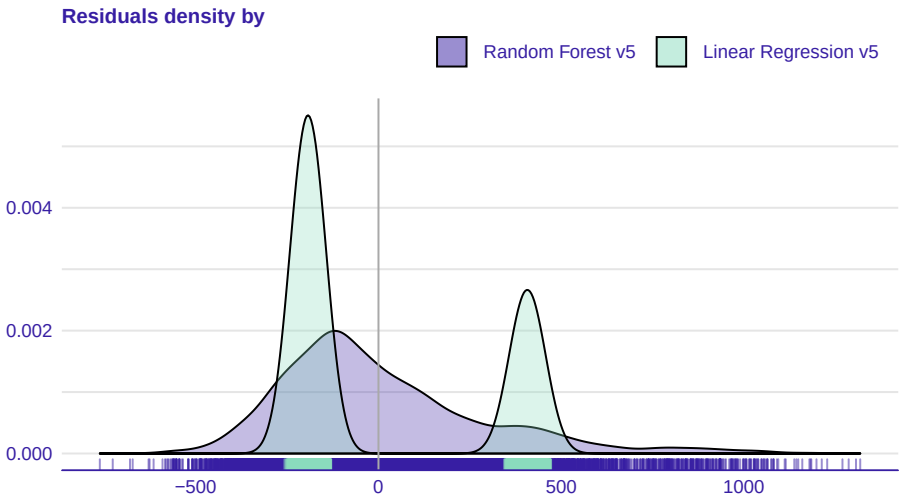


FIGURE 61 (fig:plotResidualDensity1) Density plot for residuals for two models created for apartments dataset. RMSE for both models is very similar, but we see that residuals for linear regression are concentrated around +- 400. For the random forest model residuals are concentrated at 0 but have large variance.

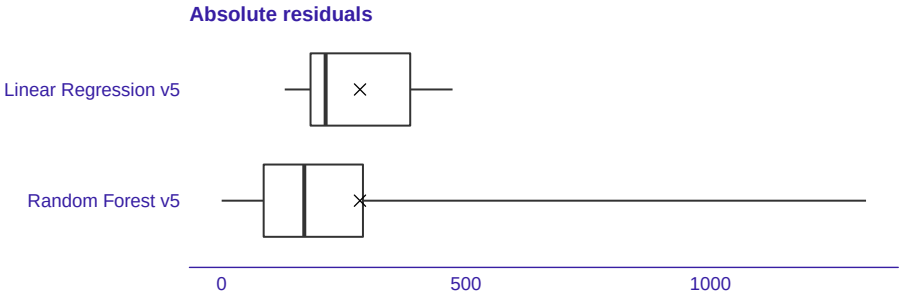


FIGURE 62 (fig:plotResidualBoxplot1) Boxplot for absolute values of residuals for two models created for apartments dataset. The cross shows the average value which corresponds to RMSE (similar for both models).

```
plot_prediction(mr_rf, abline = TRUE)
```

```
plot_prediction(mr_rf, variable = NULL, abline = TRUE)
```

Figures 65 and 66 and 67 are devoted to diagnostics that link residuals with other variables.

Figure 65 shows that for the random forest model residuals are linked with true values of the target variable. We already know that the model is biased it just confirms that predictions are shifted towards the average. Same can be read from the Figure 67.

Figure 66 investigates the relation between residuals and ordering of observations. In this case there is no relation as shall be expected.

```
plot_residual(mr_rf)
```

```
plot_residual(mr_rf, variable = NULL)
```

```
plot_residual(mr_rf, variable = "_y_hat_")
```

Figures presented so far were focused on shifts or biases in model predictions. Figure 68 helps to find problems in the variance of residuals. In many cases we expect that residuals will have constant variance. This can be verified on the scale-location plot. On the X axis there are model predictions while on the Y axis there are square roots from absolute values of residuals.

Smoothed average correspond to the standard deviation of residuals. Flat constant trend confirms homogeneity of the variance. In this example we see that variance of residuals is larger for extreme model predictions.

```
plot_scalelocation(mr_rf, variable = "_y_hat_", smooth = TRUE)
```

Another way of checking if there are problems in model structure related to the ordering of observation is the autocorrelation plot. Example of the autocorrelation is presented in Figure 69. Here we do not see a strong autocorrelation.

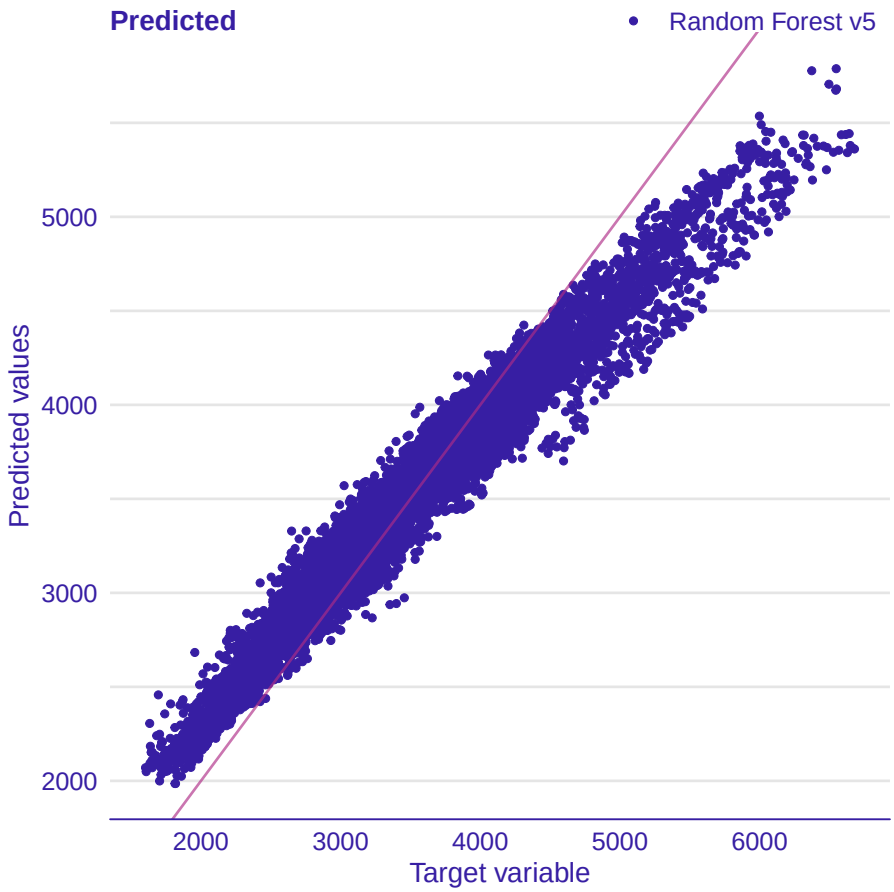


FIGURE 63 (fig:plotPrediction1) Predicted versus true values for the random forest model for apartments data. Red line stands for the baseline. One can read that model predictions are biased towards the mean.

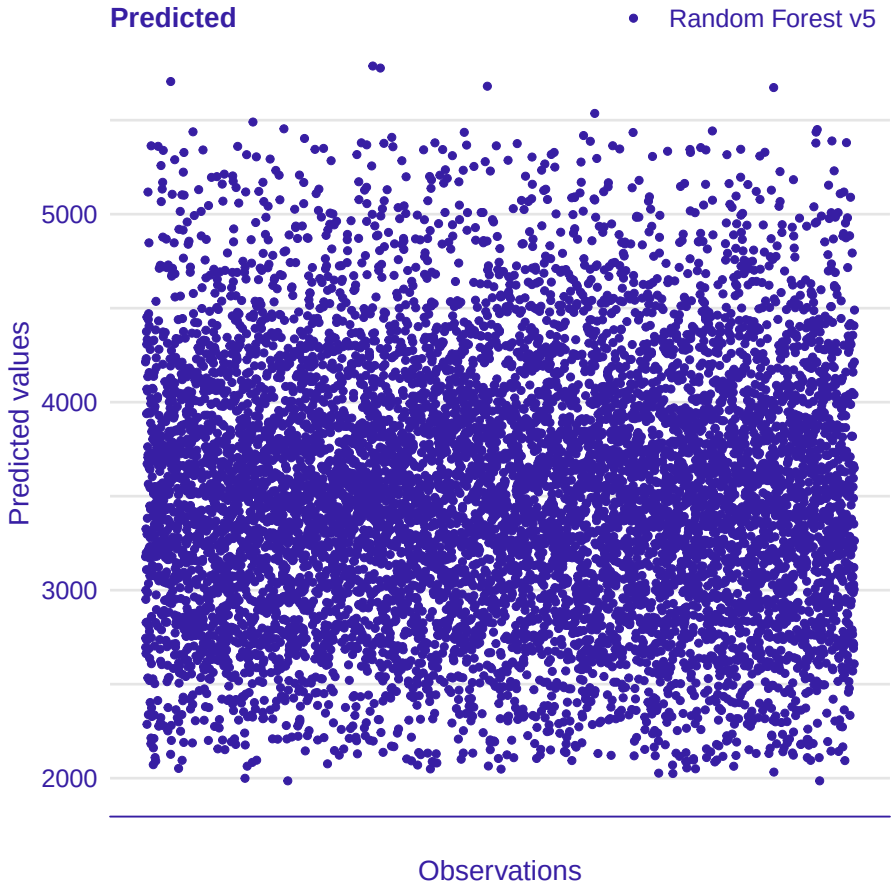


FIGURE 64 (fig:plotPrediction2) Predicted values versus ordering of observations.

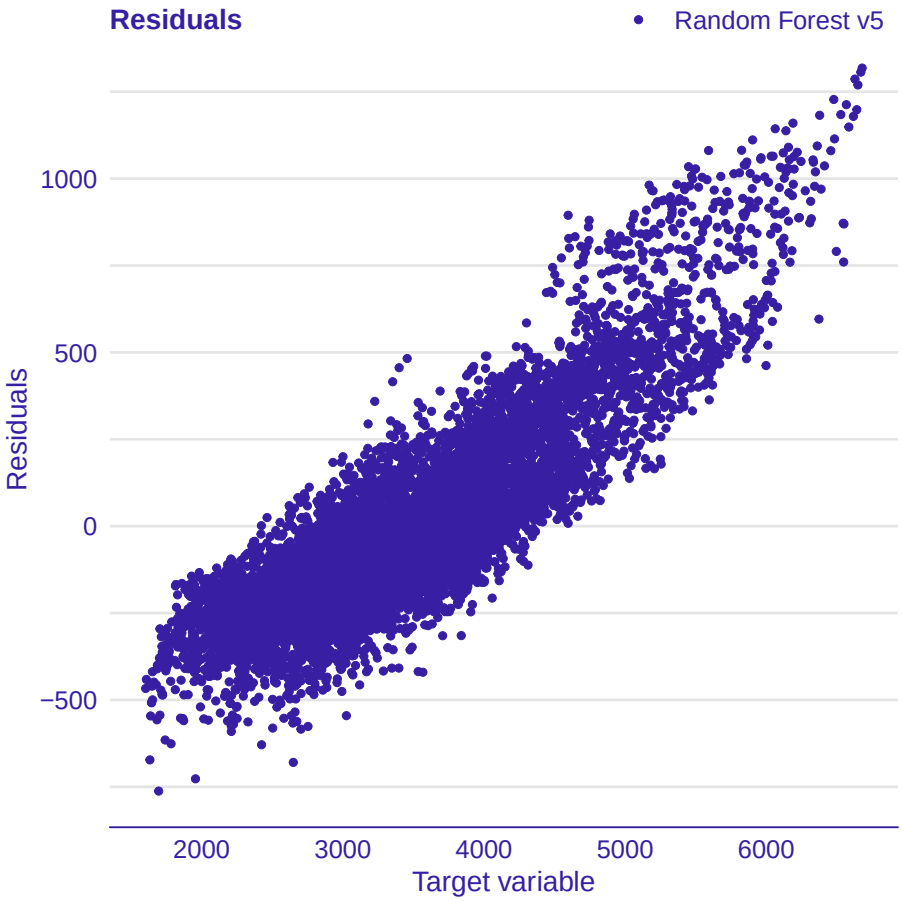


FIGURE 65 (fig:plotResidual1) Residuals versus true values for the random forest model for apartments data. Random forest model is biased towards the mean so for low values of the target variable we see negative residuals while for large values we see large positive residuals.

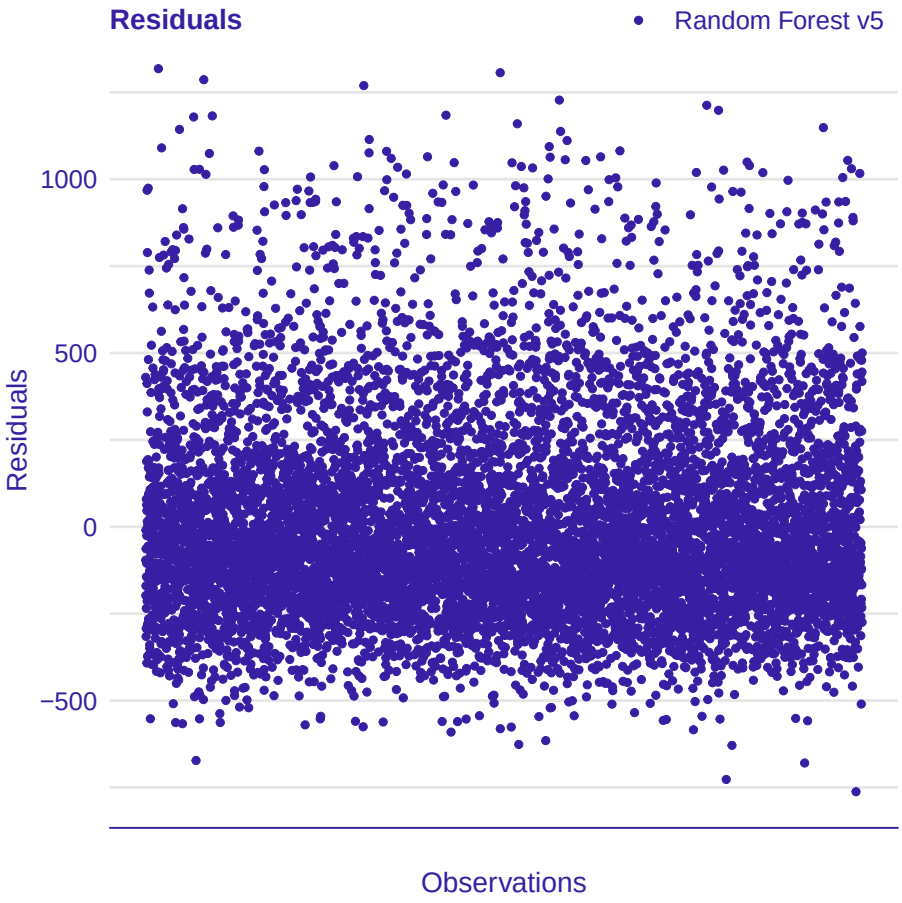


FIGURE 66 (fig:plotResidual2) Residuals versus order of observations.

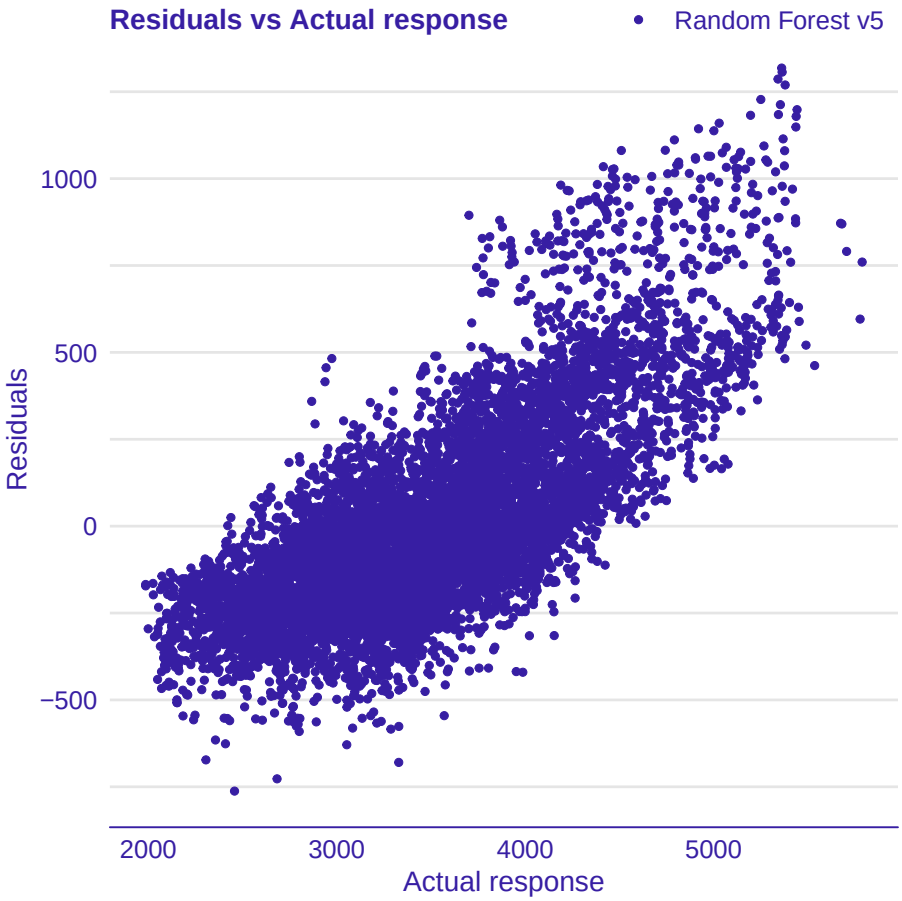


FIGURE 67 (fig:plotResidual3) Residuals versus predicted values for the random forest model for apartments data. Random forest model is biased towards the mean so for low predictions we see negative residuals while for large predictions we see large positive residuals.

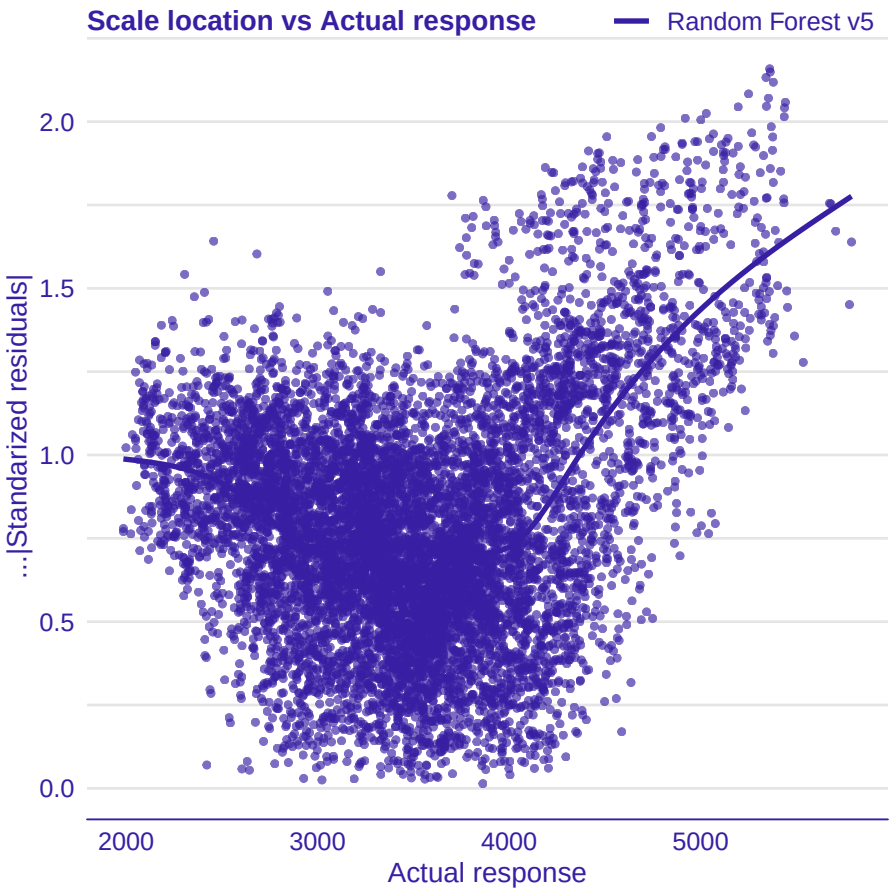


FIGURE 68 (fig:plotScaleLocation1) The scale-location plot for the random forest model for apartments data. On the X axis there are predicted values while on the Y axis there are square roots from absolute values of residuals. Any pattern in the data suggests that variance of residuals is related with predicted variables. It's the case here, since model is biased towards the average and variance of residuals is larger at extremes of the target variable.

```
plot_autocorrelation(mr_rf)
```

Use Cases

0.21 FIFA 19

In previous chapters we introduced a number of methods for instance level exploration of predictive models. In consecutive chapter we showed how to use Ceteris Poribus profiles, SHAP values, LIME or Break Down plots for models created on the dataset `titanic`. These examples we introduced and discussed separately as each of them was focused on a single method described in a given chapter.

In this chapter we present an example of full process for model development along the process introduces in chapter 0.2. We will use a new dataset for FIFA 19 soccer game. Based on it we will tour through the process of data preparation, model assembly and model understanding. In each phase we show how to combine results from different methods of exploration.

The main goal of this chapter is to show how different techniques complement each other. Some phases, like data preparation, are simplified in order to leave space for the method for visual exploration and explanation of predictive models.

0.21.1 Introduction

The story is following. The `https://sofifa.com/` portal is a reliable website for FIFA ratings of football players. Data from this website was scrapped and make available at the Kaggle webpage `https://www.kaggle.com/karangadiya/fifa19`.

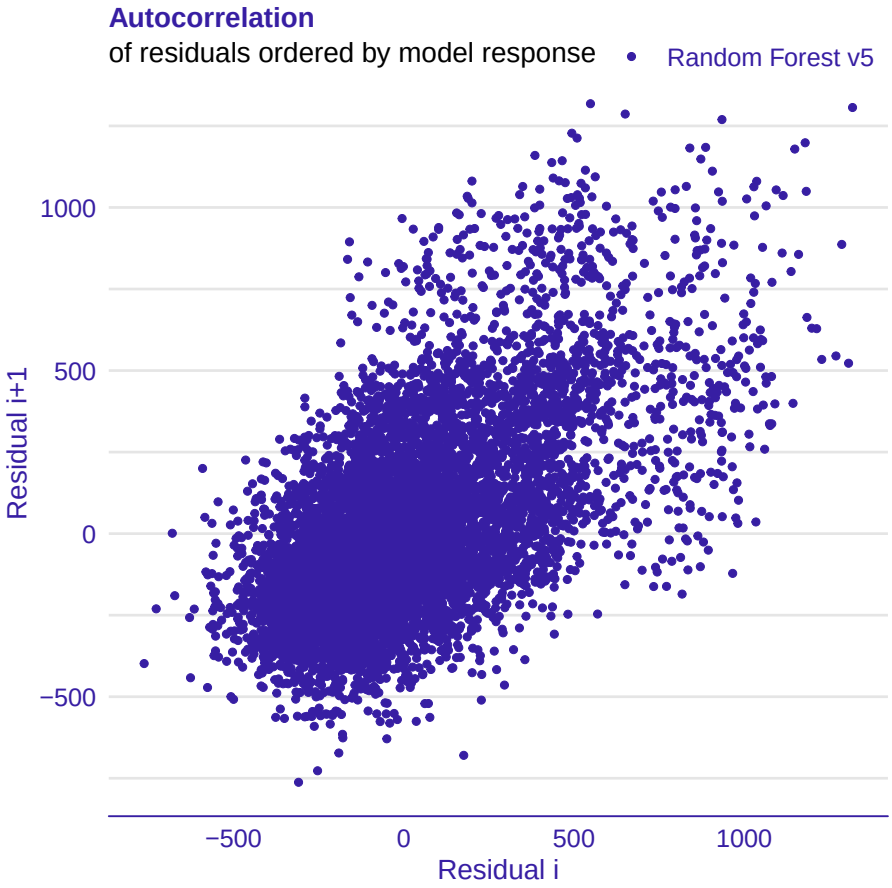


FIGURE 69 (fig:plotAutocorrelation1) The autocorrelation plot for the random forest model for apartments data. On the X axis there are residuals for observation i, while on the Y axis there are residuals for observation i+1.

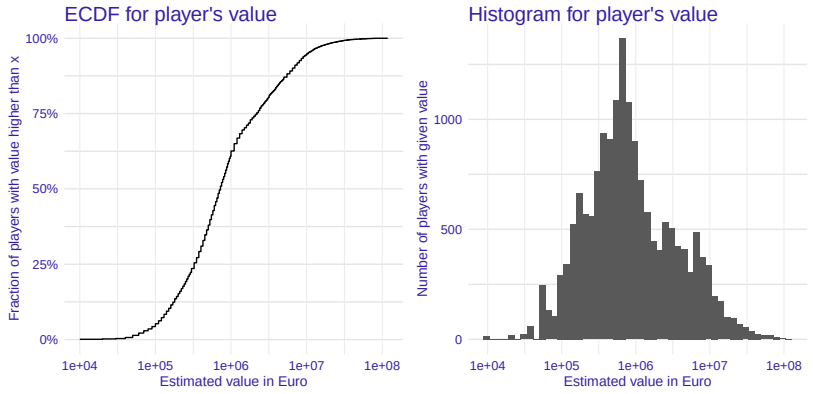


FIGURE 70 Empirical cummulative distribution function and histogram for values of players. The OX axis is in the log10 transformation.

We will use this data to build a predictive model for assessment of player value. Once the model will be created we will use methods for exploration and explanation to better understand how it is working and also to better understand which factors and how influence the player value.

0.21.2 Data preparation

The scrapped data contains 89 columns, and variuos information about players along with photo, club, nationality and others. Here we will focus on 40 players statistics and the way how they influence model predictions.

The data set contains statistics for 16924 players. First, let's see distribution of selected variables from this dataset.

Players values are heavily skewed. Half of players have estimated value between 0.3 and 2.2 millions of Euro. But few players have estimated values higher than 100 millions of Euro. Figure 70 presents empirical cumulative distribution function and histogram with log transformation of the OX axis.

Due to a large number of player characteristics we are not going

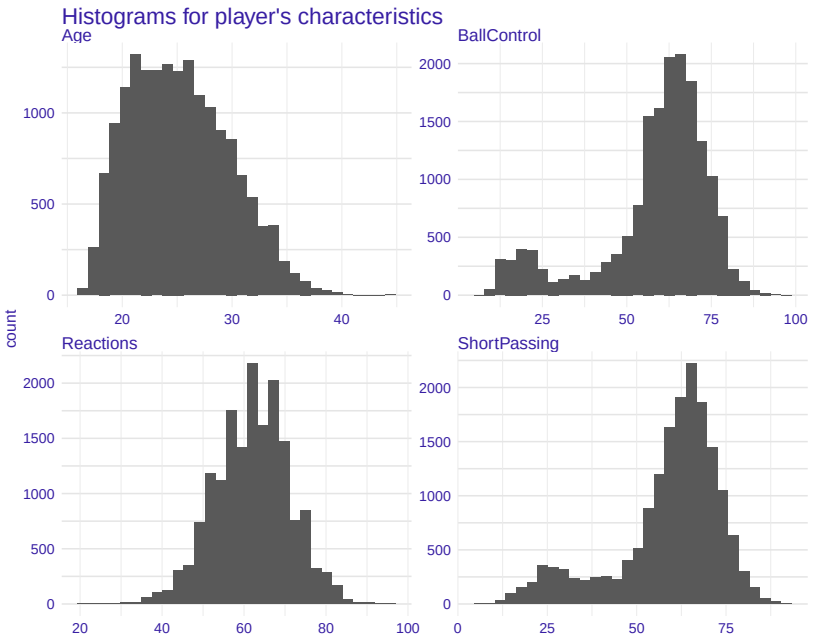


FIGURE 71 Histograms for selected characteristics of players. Note that BallControl and ShortPassing have bimodal distributions

to explore all of them but rather we will focus on four that will be discussed later in this chapter, namely: Age, Reactions, BallControl and ShortPassing.

Figure 71 presents distributions for these variables. For `Age` we see that most players are between 20 and 30 years old. What is interesting in `BallControl` and `ShortPassing` is that they have bimodal distribution. The reason for that is that these characteristics are very low for goalkeepers but higher for other players. The variable `Reactions` has Gaussian shaped distributtion with average 62 and standard deviation 9.

0.21.3 Data understanding

Time to see how these variables are linked with player's value. Figure 72 shows scatterplots for selected four characteristics.

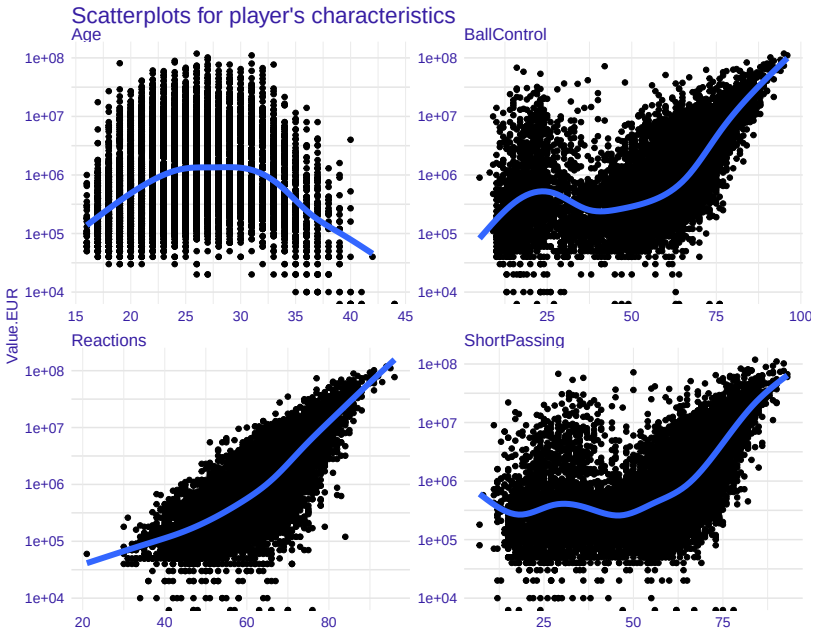


FIGURE 72 (fig:distFIFA19scatter) Scatterplot for realtion between selected four players characteristics and values of players.

Because of the skewness of player's value the OY value is presented after log transformation.

For `Age` it looks like the relation is not monotonic, there is some optimal age in which players value is the highest, between 24 and 28 years. Value of youngest players are on average 10x lower. Same with olders players.

For variables `BallControl` and `ShortPassing` the relation is not monotonic. In general the larger value of these coefficients the higher value of a player and most expensive are players with top characteristics. But among players with very low scores in `BallControl` and `ShortPassing` some are very expensive too. As it was suggested earlier, these players are probably goalkeepers.

For variable `Reactions` the link with player's value is monotonic. As expected, the higher `Reactions` the higher player's value.

Figure 73 shows pairwise scatterplots for dependent variables.

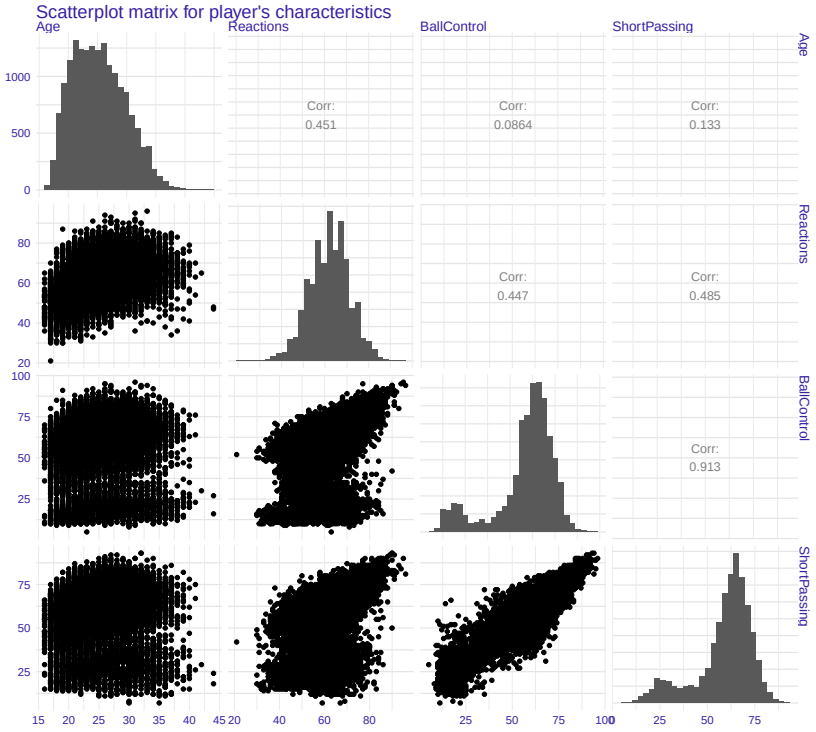


FIGURE 73 (fig:distFIFA19scatter2) Scatterplot for realtion between selected four players characteristics and values of players.

Three observations are clear from these scatterplots. One is that `Age` has positive correlation with other variables. On average older players have higher skills. Second is that skills are positively correlated, the correlation between `BallControl` and `ShortPassing` is higher than 0.9. Third is that goalkeepers' characteristics are different than rest of players.

Let's compare results from this data exploration with exploration of predictive models that will be fitted on this data.

0.21.4 Model assembly

Time to build a predictive model for players' value based on selected characteristics. We will use a trained elastic model to

explore the relation between players' characteristics and players' values.

Having clean data then model assembly is easy. For FIFA 19 data we will try four models with different structures that are able to catch different types of relations. One model would be enough, but we will try four different models to see if they catch similar relations.

Considered models are:

- boosting model with 250 trees 1 level depth as implemented in package `gbm` (Ridgeway, 2017),
- boosting model with 250 trees 4 levels depth, this model shall be able to catch interactions between features,
- linear model with spline transformation of dependent variables implemented in package `rms` (Harrell Jr, 2018),
- random forest model with 250 trees as implemented in package `ranger` (Wright and Ziegler, 2017).

```
library("gbm")
fifa_gbm_shallow <- gbm(LogValue~., data = fifa19small, n.trees = 250, interaction.depth = 1,

fifa_gbm_deep <- gbm(LogValue~., data = fifa19small, n.trees = 250, interaction.depth = 4, dis

library("ranger")
fifa_rf <- ranger(LogValue~., data = fifa19small, num.trees = 250)

library("rms")
fifa_ols <- ols(LogValue ~ rcs(Age) + rcs(International.Reputation) + rcs(Skill.Moves) + rcs(C
```

Before we can explore model behavior we need to create explainers with the `DALEX::explain` function. These explainers will be later used to asses model performance.

Note that models are trained on logarithm of the value, but it will be much more natural to operate on values in Euro. This is why in explainers we specified a user defined predict function that transforms log value to the value in Euro.

Each explainer got also a unique `label` and corresponding `data` and `y` arguments.

```
library("DALEX")
fifa_gbm_exp_deep <- explain(fifa_gbm_deep,
                         data = fifa19small, y = 10^fifa19small$LogValue,
                         predict_function = function(m,x) 10^predict(m, x, n.trees = 250),
                         label = "GBM deep")

fifa_gbm_exp_shallow <- explain(fifa_gbm_shallow,
                         data = fifa19small, y = 10^fifa19small$LogValue,
                         predict_function = function(m,x) 10^predict(m, x, n.trees = 250),
                         label = "GBM shallow")

fifa_rf_exp <- explain(fifa_rf,
                         data = fifa19small, y = 10^fifa19small$LogValue,
                         predict_function = function(m,x) 10^predict(m, x)$predictions,
                         label = "RF")

fifa_rms_exp <- explain(fifa_ols,
                         data = fifa19small, y = 10^fifa19small$LogValue,
                         predict_function = function(m,x) 10^predict(m, x),
                         label = "RMS")
```

0.21.5 Model audit

We have created four models. Let's see which model is better. Figure 74 compares distributtions of absolute model residuals. Crosses corresponds to average, which correspond to RMSE. On average, smallest residuals are for the Random Forest model.

```
library("auditor")

fifa_mr_gbm_shallow <- model_residual(fifa_gbm_exp_shallow)
fifa_mr_gbm_deep <- model_residual(fifa_gbm_exp_deep)
fifa_mr_gbm_rf <- model_residual(fifa_rf_exp)
fifa_mr_gbm_rms <- model_residual(fifa_rms_exp)
```

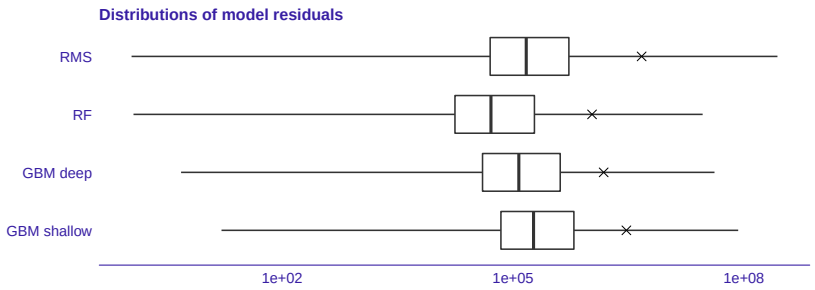


FIGURE 74 Distribution of absolute values of residuals. The smaller are values the better is the model. Crosses stand for averages.

```
plot_residual_boxplot(fifa_mr_gbm_shallow, fifa_mr_gbm_deep, fifa_mr_gbm_rf, fifa_mr_gbm_rms)
  scale_y_log10() +
  ggtitle("Distributions of model residuals")
```

But performance is not everything. Figure 75 show diagnostic plots for every model. Each scatterplot shows true target variable against model predictions. The random forest model has predictions closest to the true target values.

Extreme predictions (lowest and highest) are biased towards the mean, what is typical for such type of models. This means that Random Forest models learned factors that influence players' values, but for the most expensive players these values will be underestimated.

```
plot_prediction(fifa_mr_gbm_shallow, fifa_mr_gbm_deep,
                fifa_mr_gbm_rf, fifa_mr_gbm_rms, abline = TRUE) +
  scale_y_log10() +  scale_x_log10() +
  facet_wrap(~`_label_`) + theme(legend.position = "none") +
  ggtitle("Diagnostic plot Predicted vs True target values")
```

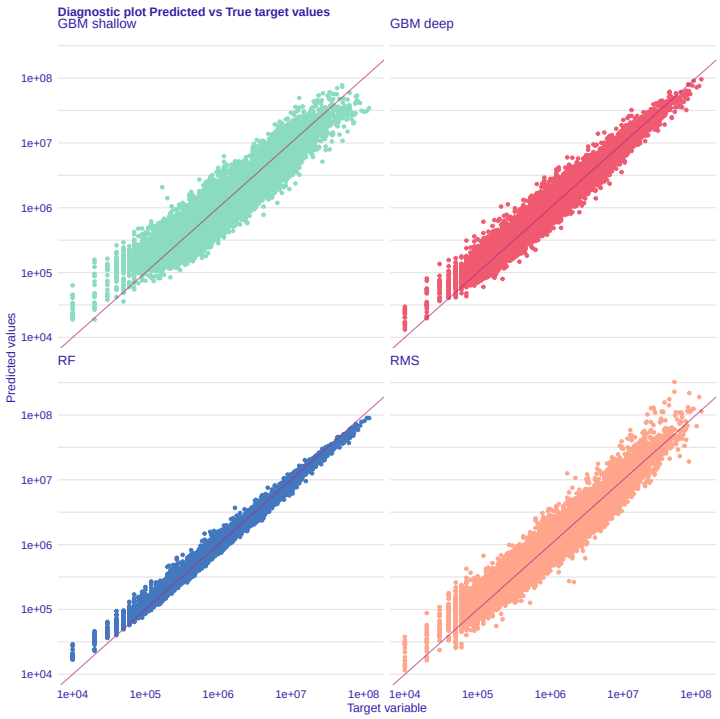


FIGURE 75 Diagnostic plots Predicted vs. True target values. Points correspond to particular players. The closer to the diagonal the better is the model.

0.21.6 Model understanding

Figure 76 shows variable importance plots for four selected models. Only 12 most important variables in each model are presented.

Some variables are important for all models, like `Reactions` or `BallControl`. Importance of other variables may be very different. All models except random forest are using some characteristics of goalkeepers.

```
library("ingredients")

fifa_feat_gbm_shallow <- ingredients::feature_importance(fifa_gbm_exp_shallow)
fifa_feat_gbm_deep <- ingredients::feature_importance(fifa_gbm_exp_deep)
fifa_feat_rf <- ingredients::feature_importance(fifa_rf_exp)
fifa_feat_rms <- ingredients::feature_importance(fifa_rms_exp)

plot(fifa_feat_rf, fifa_feat_rms,
     fifa_feat_gbm_deep, fifa_feat_gbm_shallow,
     max_vars = 12, bar_width = 4) + ylab("RMSE") + facet_wrap(~label, ncol = 1, scales = "fre
```

Figure 77 shows Partial Dependency profiles for the most important variables. They show average relation between particular variable and players value.

The general direction of relation in all models is the same. The larger the player characteristic the higher is the price. With a single exception – variable Age.

Random forest model has smallest range of average model responses. All tree-based models stabilize average predictions at the ends of variables ranges.

The most interesting difference between Exploratory Data Analysis presented in Figure 72 and Exploratory Model Analysis presented in Figure 77 is related with variable `Age`. In Figure 72 the relation was non-monotonic while in Figure 77 its monotonically decreasing. How we can explain this difference? One explanation is following: Youngest players have lower values

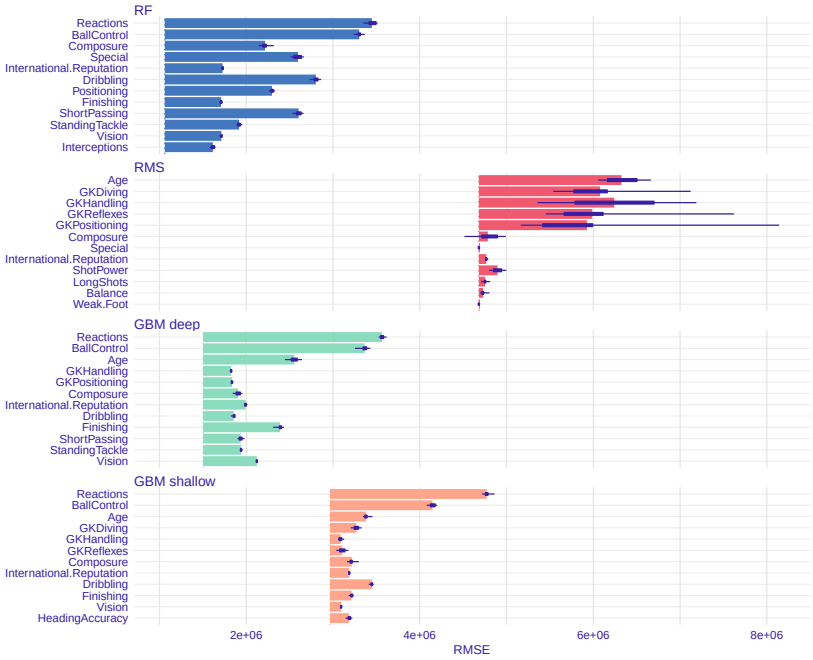


FIGURE 76 Variable importance plots for four considered models. Each bar starts in a RMSE for the model and ends in a RMSE calculated for data with permutated single variable.

not because of their age but because of lower skills that are correlated with Age. The EDA analysis cannot entangle these effects, thus for youngest players we see lover values also because their lower skills. But models learned that once we take skills into account, the effect of age is only decreasing.

This example also shows, that proper *exploration of models may be more insightful than exploration of raw data*. Variable `Age` is correlated with other confounding variables. This entangle was visible in the EDA analysis. But models learned to disentangle these effects.

```
selected_variables <- c("Age", "Reactions","BallControl", "ShortPassing")

fifa19_pd_shallow <- ingredients::partial_dependency(fifa_gbm_exp_shallow, variables = selecte
fifa19_pd_deep <- ingredients::partial_dependency(fifa_gbm_exp_deep, variables = selected_vari
fifa19_pd_rf <- ingredients::partial_dependency(fifa_rf_exp, variables = selected_variables)
fifa19_pd_rms <- ingredients::partial_dependency(fifa_rms_exp, variables = selected_variables)

plot(fifa19_pd_shallow, fifa19_pd_deep, fifa19_pd_rf, fifa19_pd_rms) +
  scale_y_log10() +
  ggtitle("Partial Dependency profiles for selected variables")
```

0.21.7 Instance understanding

Time to see how the model behaves for a single observation / player This can be done for any player, but for this example we will use *Robert Lewandowski*, the most valuable polish football player.

Here are his characteristics in the FIFA 19 database.

```
fifa19small["R. Lewandowski",]
```

```
##                Age Special Preferred.Foot International.Reputation Weak.Foot
## R. Lewandowski  29    2152          Right                         4         4
##                Skill.Moves Crossing Finishing HeadingAccuracy ShortPassing
## R. Lewandowski           4       62        91              85           83
##                Volleys Dribbling Curve FKAccuracy LongPassing BallControl
```

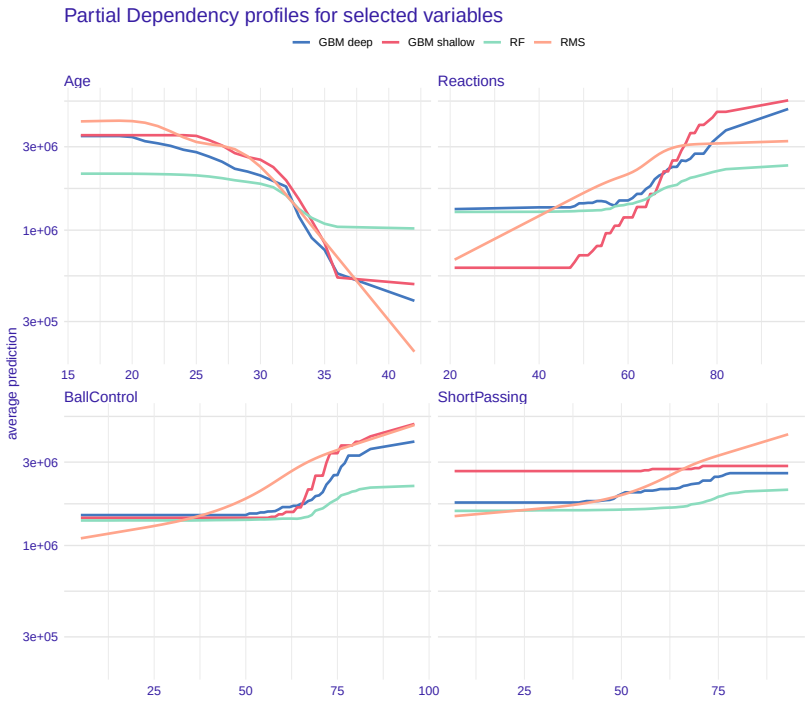


FIGURE 77 Partial dependency profiles for four selected variables and four considered models.

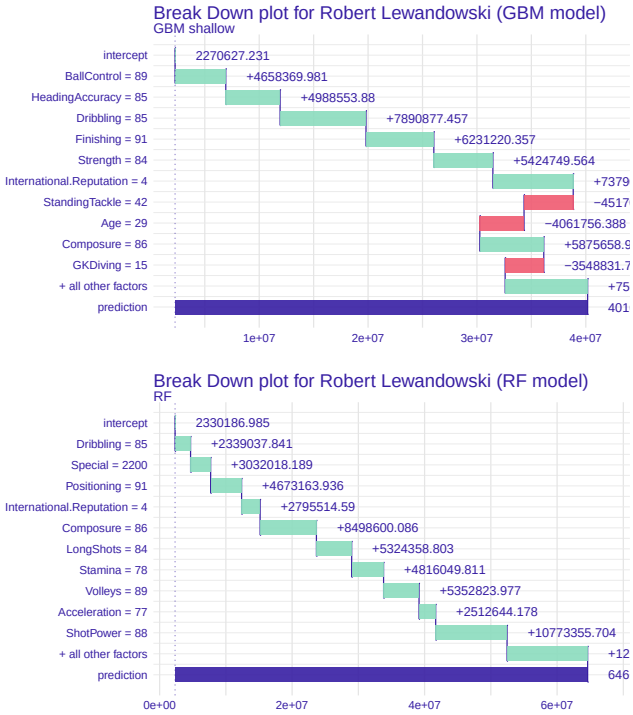


FIGURE 78 Break down plot for Robert Lewandowski. Results for GBM and RF model.

```
## R. Lewandowski      89        85    77         86          65          89
##                Acceleration SprintSpeed Agility Reactions Balance ShotPower
## R. Lewandowski           77          78      78        90      78        88
##                Jumping Stamina Strength LongShots Aggression Interceptions
## R. Lewandowski      84      78       84        84         80            39
##                Positioning Vision Penalties Composure Marking StandingTackle
## R. Lewandowski          91     77        88        86      34              42
##                SlidingTackle GKDiving GKHandling GKKicking GKPositioning
## R. Lewandowski            19       15          6        12              8
##                GKReflexes LogValue
## R. Lewandowski         10 7.886491
```

In the chapter 0.7 we showed a Break Down plots for presentation of variable attributions. In the Figure 78 we show Break Down plots for Robert Lewandowski predictions.

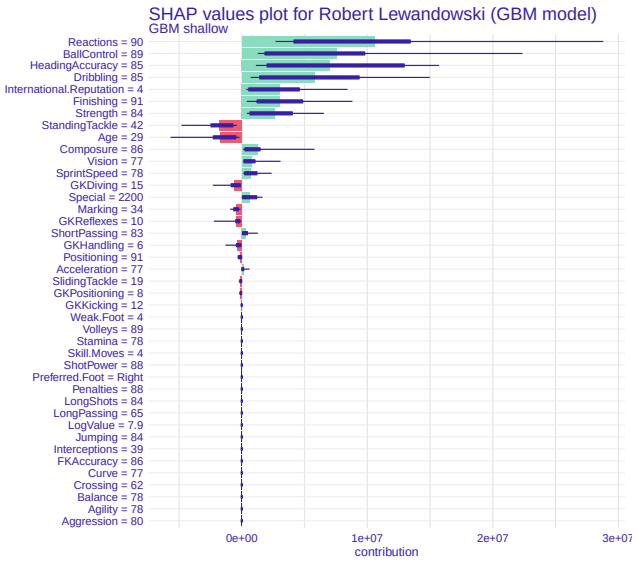


FIGURE 79 (fig:usecaseFIFASHAP) SHAP values for GBM model.

In the chapter 0.9 we showed a SHAP values for presentation of variable attributions. In the Figure 79 we show SHAP plots for Robert Lewandowski predictions. As it was expected, these explanations are consistent.

Robert Lewandowski is a striker. It makes sense that his most valuable characteristics are Reactions and BallControl.

How these plots will look like for goalkeepers? Figure 80 show Break Down plots for Wojciech Szczęsny - most valuable polish goal keeper. As we see the most important coefficients make sense, most of them are liked with properties of goalkeepers.

In chapter 0.11 we introduced Ceteris Paribus profiles. These are more details steps of the model exploration. Based on an example of Robert Lewandowski, let's see how change in one characteristic affects model value.

All models give Robert best scores when it comes to `Reactions`, `BallControl` or `Dribling`. When it comes to Age we see that the predicted value is just before a larger drop in value prediction.

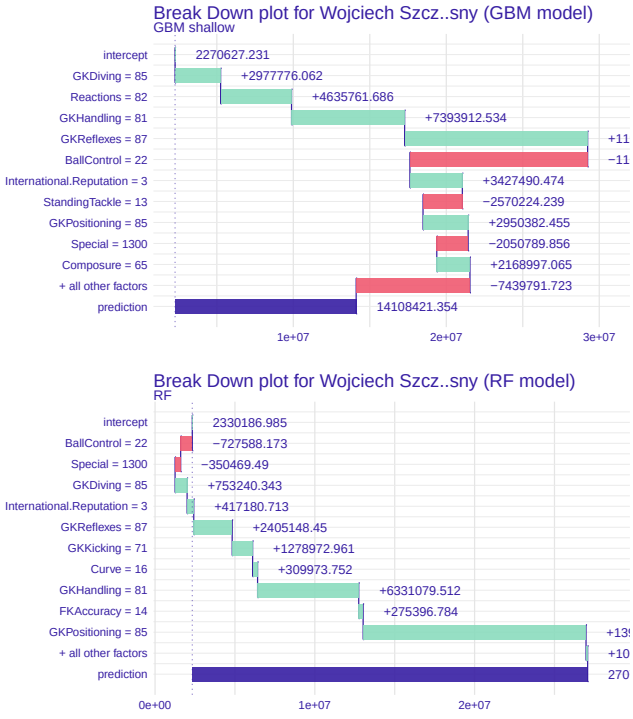


FIGURE 80 Break down plot for Wojciech Szczęsny. Results for GBM and RF model.

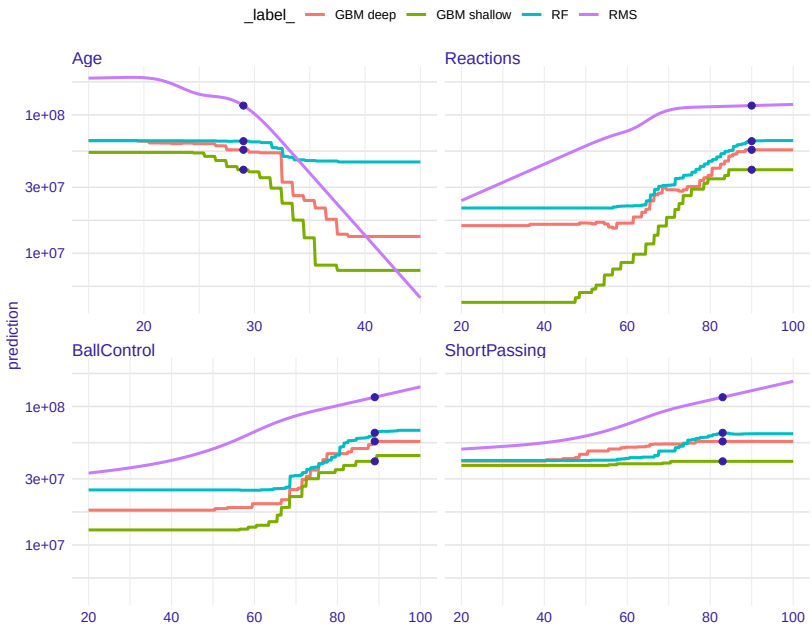


FIGURE 81 Ceteris Paribus profiles for Robert Lewandowski for four selected observations.

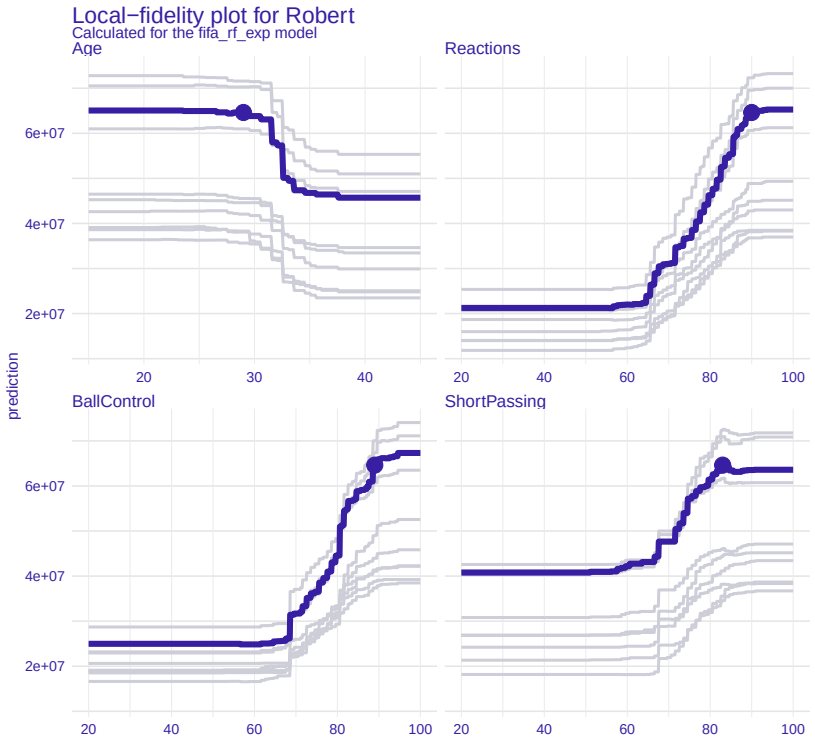


FIGURE 82 Ceteris Paribus profiles for neighbours of Robert Lewandowski.

Bibliography

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., Samek, W., Müller, K.-R., Dähne, S., and Kindermans, P.-J. (2018). innvestigate neural networks!

Allaire, J. and Chollet, F. (2019). *keras: R Interface to 'Keras'*. R package version 2.2.4.1.

Alvarez-Melis, D. and Jaakkola, T. S. (2018). On the Robustness of Interpretability Methods. *arXiv e-prints*, page arXiv:1806.08049.

Apley, D. (2018). *ALEPlot: Accumulated Local Effects (ALE) Plots and Partial Dependence (PD) Plots*. R package version 1.1.

Azure (2019). Microsoft Cognitive Services.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140.

Biecek, P. (2018). *DALEX: Explainers for Complex Predictive Models in R*.

Biecek, P. (2019). Model development process. *CoRR*, abs/1907.04461.

Biecek, P., Baniecki, H., Izdebski, A., and Pekala, K. (2019). *ingredients: Effects and Importances of Model Ingredients*. https://ModelOriented.github.io/ingredients/, https://github.com/ModelOriented/ingredients.

Biecek, P. and Kosinski, M. (2017). archivist: An R package for managing, recording and restoring data analysis results. *Journal of Statistical Software*, 82(11):1–28.

Binder, A., Montavon, G., Bach, S., Müller, K., and Samek, W. (2016). Layer-wise relevance propagation for neural networks with local renormalization layers. *CoRR*, abs/1604.00825.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., Casalicchio, G., and Jones, Z. M. (2016). mlr: Machine learning in r. *Journal of Machine Learning Research*, 17(170):1–5.

Boehm, B. (1988). *A Spiral Model of Software Development and Enhancement*.

Breiman, L. (2001). Random forests. In *Machine Learning*, volume 45, pages 5–32.

Breiman, L., Cutler, A., Liaw, A., and Wiener, M. (2018). *randomForest: Breiman and Cutler's Random Forests for Classification and Regression*. R package version 4.6-14.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA.

Casey, B., Farhangi, A., and Vogl, R. (2018). Rethinking explainable machines: The gdpr's 'right to explanation' debate and the rise of algorithmic audits in enterprise. *Berkeley Technology Law Journal*.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (1999). *The CRISP-DM 1.0 Step-by-step data mining guide*.

Dastin, J. (2018). Amazon scraps secret ai recruiting tool that showed bias against women.

Deng, J., Dong, W., Socher, R., Li, L., Kai Li, and Li Fei-Fei (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Diaz, M., Johnson, I., Lazar, A., Piper, A. M., and Gergle, D. (2018). Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 412:1–412:14, New York, NY, USA. ACM.

Duffy, C. (2019). Apple co-founder steve wozniak says apple card discriminated against his wife.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press, New York, NY, USA, 1st edition.

Faraway, J. (2002). *Practical Regression and ANOVA using R*.

Fisher, A., Rudin, C., and Dominici, F. (2018). Model class reliance: Variable importance measures for any machine learning model class, from the 'rashomon' perspective. *Journal of Computational and Graphical Statistics*.

Foster, D. (2017). *xgboostExplainer: An R package that makes xgboost models fully interpretable*. R package version 0.1.

Friedman, J. H. (2000). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232.

from Jed Wing, M. K. C., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., and Candan., C. (2016). *caret: Classification and Regression Training*. R package version 6.0-64.

Galecki, A. and Burzykowski, T. (2013). *Linear Mixed-Effects Models Using R: A Step-by-Step Approach*. Springer Publishing Company, Incorporated.

GDPR (2018). The eu general data protection regulation (gdpr) is the most important change in data privacy regulation in 20 years.

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65.

Goodman, B. and Flaxman, S. (2016). European union regulations on algorithmic decision-making and a "right to explanation". *Arxiv*.

Gosiewska, A. and Biecek, P. (2018). auditor: an R package for model-agnostic visual validation and diagnostic. *ArXiv e-prints*.

Gosiewska, A. and Biecek, P. (2019a). iBreakDown: Uncertainty of Model Explanations for Non-additive Predictive Models.

Gosiewska, A. and Biecek, P. (2019b). *shapper: Wrapper of Python Library 'shap'*. R package version 0.1.0.

Gosiewska, A., Gacek, A., Lubon, P., and Biecek, P. (2019). Safe ml: Surrogate assisted feature extraction for model learning.

Greenwell, B. M. (2017). pdp: An R Package for Constructing Partial Dependence Plots. *The R Journal*, 9(1):421–436.

Grolemund, G. and Wickham, H. (2019). *R for Data Science*.

Hall, P. (2019). *On Explainable Machine Learning Misconceptions and A More Human-Centered Machine Learning*.

Harrell Jr, F. E. (2018). *rms: Regression Modeling Strategies*. R package version 5.1-2.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Hoover, B., Strobelt, H., and Gehrmann, S. (2019). exbert: A visual analysis tool to explore learned representations in transformers models.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674.

Jacobson, I., Booch, G., and Rumbaugh, J. (1999). *The Unified Software Development Process*.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

Karbowiak, E. and Biecek, P. (2019). *EIX: Explain Interactions in Gradient Boosting Models*. R package version 1.0.

Kruchten, P. (1998). *The Rational Unified Process: An Introduction*.

Kuhn, M. and Johnson, K. (2013a). Applied predictive modeling.

Kuhn, M. and Johnson, K. (2013b). *Applied Predictive Modeling*. Springer. ISBN 978-1461468486.

Kuhn, M. and Vaughan, D. (2019). *parsnip: A Common API to Modeling and Analysis Functions*. R package version 0.0.2.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How we analyzed the compas recidivism algorithm.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of google flu: Traps in big data analysis. *Science*, 343(6176):1203–1205.

LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., Kraljevic, T., Nykodym, T., Aboyoun, P., Kurka, M., and Malohlava, M. (2019). *h2o: R Interface for 'H2O'*. R package version 3.22.1.1.

Liaw, A. and Wiener, M. (2002a). Classification and regression by randomforest. *R News*, 2(3):18–22.

Liaw, A. and Wiener, M. (2002b). Classification and regression by randomforest. *R News*, 2(3):18–22.

Lundberg, S. (2019). *SHAP (SHapley Additive exPlanations)*. Python package.

Lundberg, S. M., Erion, G. G., and Lee, S. (2018). Consistent individualized feature attribution for tree ensembles. *CoRR*, abs/1802.03888.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Max, K. and Wickham, H. (2018). *tidymodels: Easily Install and Load the 'Tidymodels' Packages*. R package version 0.0.2.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2017). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.6-8.

Molnar, C. (2019). *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/.

Molnar, C., Bischl, B., and Casalicchio, G. (2018). iml: An R package for Interpretable Machine Learning. *JOSS*, 3(26):786.

Nolan, D. and Lang, D. T. (2015). *Data Science in R: A Case Studies Approach to Computational Reasoning and Problem Solving*. Chapman & Hall/CRC.

O'Connell, M., Hurley, C., and Domijan, K. (2017). Conditional visualization for statistical models: An introduction to the condvis package in r. *Journal of Statistical Software, Articles*, 81(5):1–20.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, New York, NY, USA.

Paluszynska, A. and Biecek, P. (2017). *randomForestExplainer: A set of tools to understand what is happening inside a Random Forest*. R package version 0.9.

Pedersen, T. L. and Benesty, M. (2018). *lime: Local Interpretable Model-Agnostic Explanations*. R package version 0.4.0.

Pedersen, T. L. and Benesty, M. (2019). *lime: Local Interpretable Model-Agnostic Explanations*. R package version 0.5.0.

Picard, D. (1985). Testing and estimating change-points in time series. *Advances in Applied Probability*, 17(4):841–867.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). *Why Should I Trust You?: Explaining the Predictions of Any Classifier*, page 1135–1144. ACM Press.

Ridgeway, G. (2017). *gbm: Generalized Boosted Regression Models*. R package version 2.1.3.

Robnik-Šikonja, M. (2018). *ExplainPrediction: Explanation of Predictions for Classification and Regression Models*. R package version 1.3.0.

Robnik-Šikonja, M. and Kononenko, I. (2008). Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):589–600.

Ross, C. and Swetliz, I. (2018). Ibm's watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show.

Ruiz, J. (2018). Machine learning and the right to explanation in gdpr.

Salzberg, S. (2014). Why google flu is a failure.

Samek, W., Wiegand, T., and Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models.

Shapley, L. S. (1953). A value for n-person games. In Kuhn, H. W. and Tucker, A. W., editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton.

Sheather, S. (2009). *A Modern Approach to Regression with R*. Springer Texts in Statistics. Springer New York.

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. *CoRR*, abs/1704.02685.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):7881.

Staniak, M. and Biecek, P. (2018). *live: Local Interpretable (Model-Agnostic) Visual Explanations*. R package version 1.5.7.

Staniak, M., Biecek, P., Igras, K., and Gosiewska, A. (2019). *localModel: LIME-Based Explanations with Interpretable Inputs Based on Ceteris Paribus Profiles*. R package version 0.3.11.

Štrumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 58:267–288.

Tufte, E. R. (1986). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Venables, W. N. and Ripley, B. D. (2010). *Modern Applied Statistics with S*. Springer Publishing Company, Incorporated.

Štrumbelj, E. and Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11:1–18.

Wes, M. (2012). *Python for Data Analysis*. O'Reilly Media, Inc., 1 edition.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H. and Grolemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, Inc., 1st edition.

Wikipedia (2019). *CRISP DM: Cross-industry standard process for data mining*.

Wright, M. N. and Ziegler, A. (2017). *ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R*.

Xie, Y. (2018). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.7.