What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Mechanics** section will look under the hood of machine learning to give you an understanding of what happens to the data inputted to an algorithm and how it becomes an insightful output.

Data Science Workflow walks through a day in the life of a data scientist, to show the complete process of how machine learning is used to solve business problems.

Data Science Workflow

Problem Formulation

Problem Formulation is a vital part of good data science. This is the time to get with business partners and subject matter experts to discuss the problem at hand and some ways you might go around solving it with the resources currently at your disposal or that you are able to require.

At this stage, it is not necessary to get into the nitty-gritty of how machine learning algorithms might work, or which tools and data processes sound best suited for the project. Rather, the data scientist should focus on deeply understanding what the business partners value, what kind of timeline they are on, any previous work that has been completed on this subject, and more. Then, looping in a subject matter expert will help give an understanding to the complexity of the problem, and what a supposed solution might look like.

The goal of problem formulation is to walk away with a clear goal as to an end product, a first thought as to what some relevant features might be, and an idea of what kind of data will be needed. Additionally, the data scientist should have a good feel to the management parts of the project - how long will this take, how many people will be involved, and what kind of resources and tools might be needed.

Problem forumulation requires deep and critical conversations with clients. The ability to understand and communicate a buisness partner's needs is a vital skillset for a data scientist to have. Click here to learn more about problem formulation.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Mechanics** section will look under the hood of machine learning to give you an understanding of what happens to the data inputted to an algorithm and how it becomes an insightful output.

Data Science Workflow walks through a day in the life of a data scientist, to show the complete process of how machine learning is used to solve business problems.

Data Science Workflow

Finding and Collecting Data

After formulating a well-defined problem, the next step is to acquire data relevant to solving that problem. This is a great step to involve a subject matter expert as they will be the most knowledgable as to what is most important to collect and know about.

Often, there are three ways for a corporation to get data. First, they might pull from existing resources within their company. This could be data from a previous project or data generated from some process or product. This is a popular way to get data because most often than not it is free. Next, a corporation might try to generate or collect the data on its own. For example, this might mean installing sensors where necessary or creating the correct data pipelines to scrape data. Finally, corporations may look externally for data. They may find some through open-source parties, contracted partners, or paid third-party data services. This is typically a more expensive option, but can be beneficial because the data may already be clean and well formatted, and is quick to obtain.

Once the right sources have been found, the data must all be pulled and stored in the same location, and then joined together to create one cohesive dataset.

Data can be collected from many different sources in many different ways. Click here to learn more about innovative ways that your company can collect data.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Mechanics** section will look under the hood of machine learning to give you an understanding of what happens to the data inputted to an algorithm and how it becomes an insightful output.

Data Science Workflow walks through a day in the life of a data scientist, to show the complete process of how machine learning is used to solve business problems.

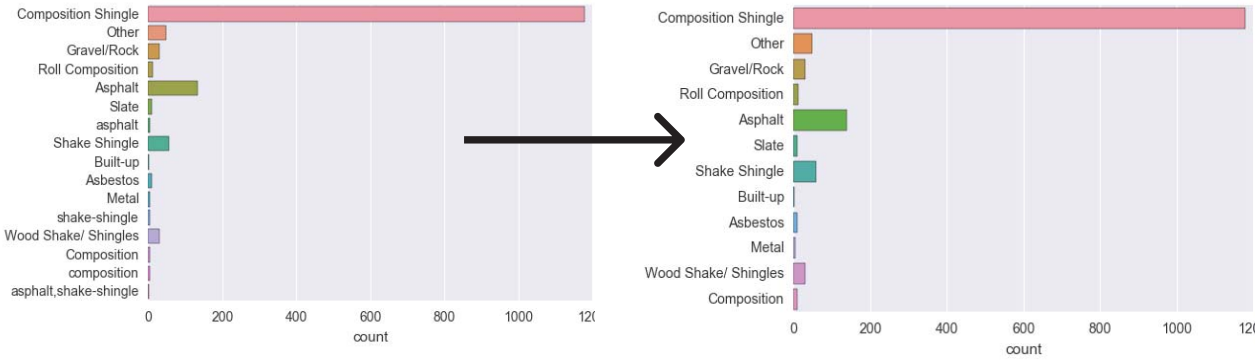
Data Science Workflow

Data Cleaning

Data cleaning is an important but often tedious and monotonous part of data science. Data cleaning refers to modifying, reformating, or deleting certain parts of a dataset to get the most unified and complete dataset possible.

For example, a data scientist might want to remove unwanted observations such as duplicates or data irrelevant to the problem. Alternatively, there might be formatting erroes such as the wrong data type, typos, or inconsistent captialization or punctuation among text or numerical values. Missing data will also need to be handled in a consisten way. Often values or given a null value to show that they are missing.

Sometimes, it may be relevant to filter out unwanted outliers. There must always be a good reason to remove the outlier that provides a factual basis for why you think it is falsely representative. Otherwise, you may be construing the model to fit your previous assumptions. Removing outliers is also often done in the next step, data exploration, when the data scientist has a better feel for the data and can make better decisions about the impact of a given outlier.



In this example, different spelling of the word "Composition" are combined into a single category. For more information about data cleaning, click here.

What am I looking at?

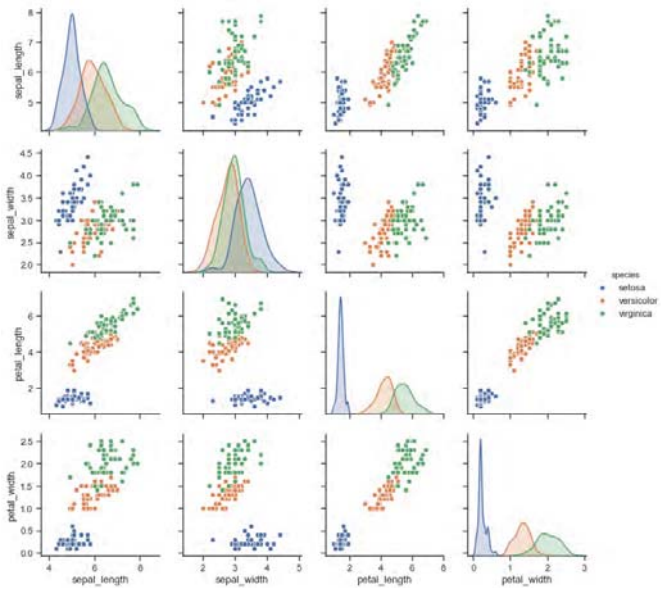
The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Mechanics** section will look under the hood of machine learning to give you an understanding of what happens to the data inputted to an algorithm and how it becomes an insightful output.

Data Science Workflow walks through a day in the life of a data scientist, to show the complete process of how machine learning is used to solve business problems.

Data Science Workflow



Scatterplot Matrices are a good way to visualize pairwise correlation.

Data Exploration

Data exploration is a very important part of machine learning. Often, data scietists are working with very large amounts of relatively unstructured data. Hence, it can be very hard to get an understanding what what that data means, what kinds of trends it shows, and which features appear to be the most important.

For some data scientists, data exploration is quite freeform based on intuition about the problem. For others, it might follow a more structured approach. For example, a data scientist might start by making histograms of each individual numerical value to get a feel for its distribution.

Next, a data scientist might look for correlation between any two of the features. If a feature is highly correlated to the target variable, this might indicate that it is a good candidate for feature selection. If a feature is highly correlated with another feature, this might indicate that they are redundant in some way and it might be best to remove one.

Data exploration is also the place to look for areas where transformations might make sense. The most common transformations are logarithm or exponential in nature. transformations help linearize or normalized data that might otherwise be hard for a machine learning model to handle, without losing any of the model's integrity.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Mechanics** section will look under the hood of machine learning to give you an understanding of what happens to the data inputted to an algorithm and how it becomes an insightful output.

Data Science Workflow walks through a day in the life of a data scientist, to show the complete process of how machine learning is used to solve business problems.

Data Science Workflow

Modelling & Validation

This is the step where machine learning does the heavy lifting.

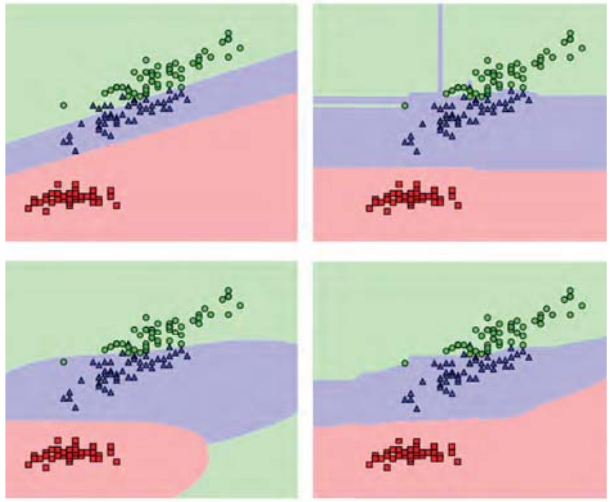
After formulating the problem, collecting and cleaning data, and doing some preliminary data exploration to get a feel for the most important features of the data, it is time to test out some machine learning algorithms.

Typically, data scientists try multiple models and compare their results in order to select the model best suited for the task at hand. One way to do this is to select completely different models. For example, if you were faced with a classification problem you could try both K-Nearest Neighbors and Naive Bayes.

Alternatively, most models have one or more tuning parameters. In K-Nearest Neighbors, this would be K, the number of neighbors the algorithm is looking for. Trying different values for these parameters is another way to test different models.

Once multiple models have been trained, it is time to compare the models against eachother. There are many, many metrics in which to compare, for example, accuracy, false positives, false negatives, f1-score, etc. Which one a data scientist selects depends on the task at hand.

These metrics are typically evaluated by running the algorithm against a 'test' or 'development' dataset, which is a portion of the input data set aside specifically for this purpose. The test data typically has an expected result or label already attached to it, making it easy to directly assess how well each model performed.



Each of these models presents both advantages and disadvantages. For more about model evaluation and selection, click here.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Mechanics** section will look under the hood of machine learning to give you an understanding of what happens to the data inputted to an algorithm and how it becomes an insightful output.

Data Science Workflow walks through a day in the life of a data scientist, to show the complete process of how machine learning is used to solve business problems.

Data Science Workflow

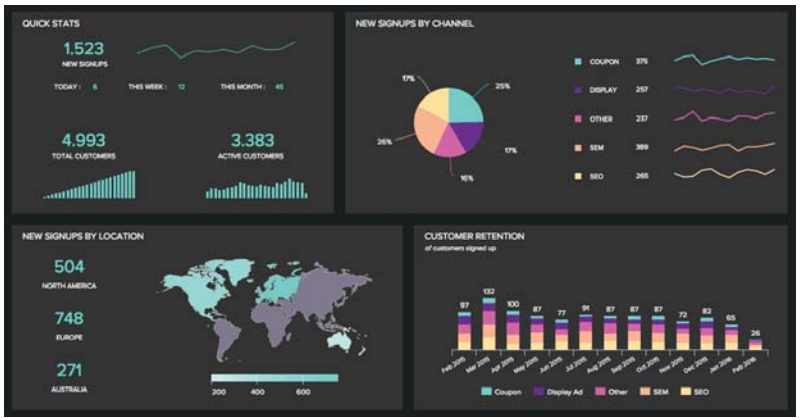
Prediction & Communicating Results

Once a model has been created and validated, it is time to put it to work. Typicallly this means creating some kind of predictions and working with a business partner to make business decisions based off those predictions.

In some cases, predictions may only be needed once. More often than not, however, a data scientist's model will be put into some kind of production environment where it can be called on a routine basis to batch new predictions. This is typically where a data engineer would get involved to create a reliable and robust computation platform.

Data scientists might also be involved in constructing visualizations or dashboard platforms for customers to view there information in an interactive and intuitive manner.

While data science and machine learning are incredibly helpful tools for informing business decisions, there still remains an art to the process, involving both an ability to communicate and collaborate with business partners, as well as a deep insight into the domain from which the data reigns.



For a great blog post about dashboarding best practices, click here.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Mechanics** section will look under the hood of machine learning to give you an understanding of what happens to the data inputted to an algorithm and how it becomes an insightful output.

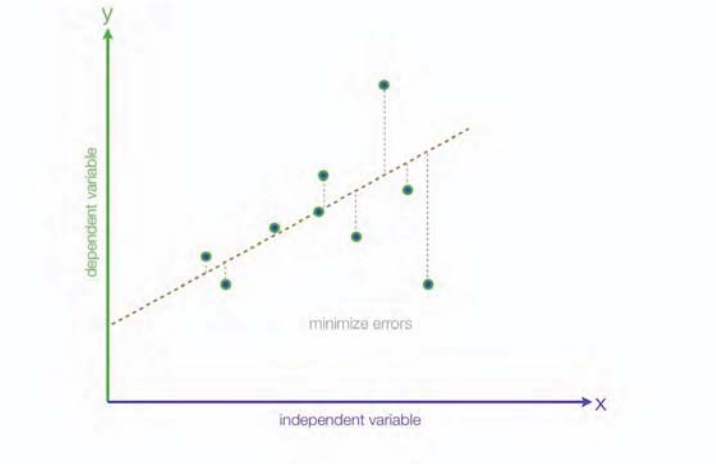
Common ML Algorithms walks through the intricacies and specific applications of several common machine learning algorithms**.**

Common Machine Learning Algorithms

Linear Regression

Linear regression is a very simple but powerful algorithm. It involves creating a linear model that defines the realtionship between two or more variables. If you have done any secondary or college-level mathematics classes, odds are you have done some kind of linear regression.

The goal of linear regression is to minimize the distance between all observations and the straight line that will be used to represent the data in the model. Linear regression is great because it is so simple and easy to understand, however it is limited in that it can only represent a linear relationship and is quite difficult to use with categorical or other non-numeric variables.



Click here for a great YouTube video about Linear Regression.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Mechanics** section will look under the hood of machine learning to give you an understanding of what happens to the data inputted to an algorithm and how it becomes an insightful output.

Common ML Algorithms walks through the intricacies and specific applications of several common machine learning algorithms.

Common Machine Learning Algorithms

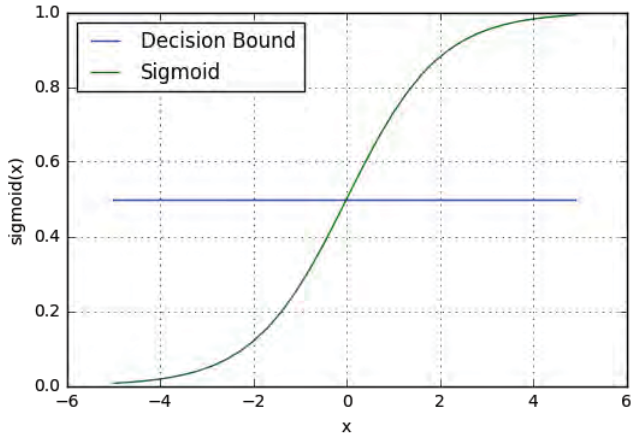
Logistic Regression

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Unlike linear regression which fits a line of best fit in the predictor variables and outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes.

This makes logistic regression is particularly useful for binary classification problems. Linear regression suffers from bias when predicting binary classification problems because the predictions are not limiting to just a 0 or 1 bound, so nonsensical predictions that are outside the bound can occur. Logistic regression is also useful in more complex classification problems - multinomial logistic regression is useful for multiple classification problems, and ordinal logistic regression is useful for classification problems with classification of an ordered group.

The main advantage of logistic regression when compared to more complex models such as ensemble methods or SVMs (Support Vector Machines) is that it is easy to interpret the results, and there is much less of a black box for how the classification probabilities are calculated.

One weakness of logistic regression is that it deals very poorly with categorical variables. A logistic regression model cannot accept a categorical variable as an input, so any categorical variables must be manually converted to binary variable(s) before it can be included in the model.



The image to the left shows a sigmoid function and the decision boundary used for most algorithms, at 0.5. Click here to learn more about Logistic Regression.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Mechanics** section will look under the hood of machine learning to give you an understanding of what happens to the data inputted to an algorithm and how it becomes an insightful output.

Common ML Algorithms walks through the intricacies and specific applications of several common machine learning algorithms.

Common Machine Learning Algorithms

Naive Bayes

Naive Bayes works on probabilities: the probability of each class in general, and a conditional probability of each class given some other information. These probability are calculated during training. When new data is used to create a prediction, the probabilities can be combined in a way to determine the most likely class, which will be selected as the final predicted value.

Naive Bayes is great because it is very simple yet surprisingly effective. However, it does operate under the assumption that all input variables are independent of eachother, which more often than not is untrue.

$$P(\text{class} \mid \text{data}) = \frac{P(\text{data} \mid \text{class}) \times P(\text{class})}{P(\text{data})}$$

The above shows Bayes Theorem, as seen in the context of Machine Learning. Click here for a great tutorial on Naive Bayes in Python.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Mechanics** section will look under the hood of machine learning to give you an understanding of what happens to the data inputted to an algorithm and how it becomes an insightful output.

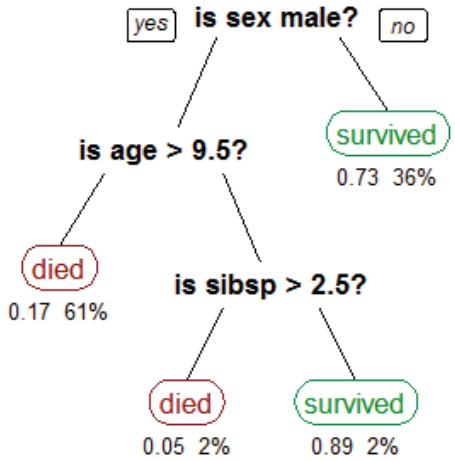
Common ML Algorithms walks through the intricacies and specific applications of several common machine learning algorithms.

Common Machine Learning Algorithms

Decision Trees

Decision tree learning uses a decision tree to make predicitons. Tree models are like dichotemies - they have a series of nodes and decision that are traversed from the root (starting points) to the leaves (prediction).

Consider that we own a bike shop and are trying to predict how much revenue a new bicycle model will give us. We could create a tree evaluating the following sequence of questions: Is the bicycle from the current model year, does the bicycle have more than 5 gears, is the bicycle painted black. Each combination of yes and no for these three questions would lead to a different previced reveue.



Above is an example decision tree for the Titanic survival problem, which we will explore fully in the Example tab.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Mechanics** section will look under the hood of machine learning to give you an understanding of what happens to the data inputted to an algorithm and how it becomes an insightful output.

Common ML Algorithms walks through the intricacies and specific applications of several common machine learning algorithms**.**

Common Machine Learning Algorithms

Ensemble Learning

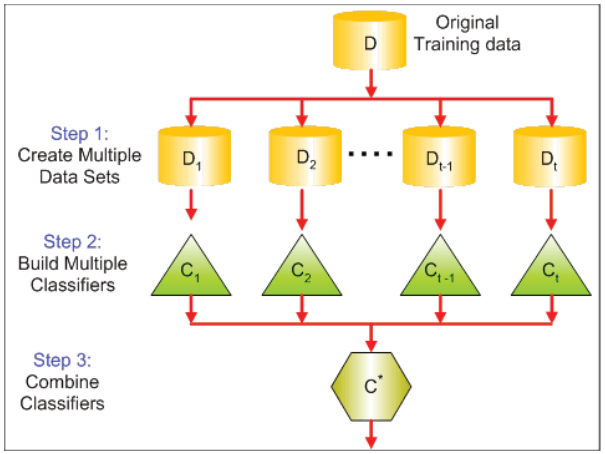
Ensemble learning is using a collection of decision tree classifiers run together, as opposed to a single predictor such as one decision tree or regression model. Within ensemble learning, there are 2 main subtypes: bagging and boosting.

Bagging involves selecting a sample of observations and running a model which contains a subset of the features on a different sample, repeated many times. The final model then involves taking an average of all the sub-models. A popular form of bagging is random forest.

Boosting involves running models sequentially, such that the results of each model is used to determine the features that the next model is run on. The way it does this is to give more weight the observations that are incorrect in a model for the next model, and use the observation with the adjusted weights to train the next model. In this way, the later models fix the errors found by the earlier models.

The functions from each model can be combined to give a model that fits the data better than any single classification model could have done.

Bagging and boosting are both helpful in reducing the variance of the model, which results in a model with better stability. Bagging is better for overfitting, since it takes an average of different trees on different subfeatures and samples so it will not overfit to the shape of the data. Boosting could still have a problem with overfitting as it builds on prior trees (which could already be overfitted) but it is better for generating a model with lower errors, as each model builds on reducing the error from the previous model.



Click here to learn more about ensemble learning.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Mechanics** section will look under the hood of machine learning to give you an understanding of what happens to the data inputted to an algorithm and how it becomes an insightful output.

Common ML Algorithms walks through the intricacies and specific applications of several common machine learning algorithms**.**

Common Machine Learning Algorithms

Catboost

Catboost is a boosting ensemble model developed by Yandex that is at the frontier of gradient boosting - it has extremely strong performance in classification problems even when compared to other ensemble methods, and category variables can be used directly in the model without needing them to be converted into numeric variables. Catboost does this by conducting one-hot encoding to map the categorical variable values to a hash table - this uses the fewest number of dummies. This solves the issue of the user needing to create dummy variables themselves, since trees generally work on thresholds for numeric variables.

Catboost also is better preventing overfitting than some gradient boosting methods. Catboost uses oblivious decision trees, where the same splitting criterion is used for an entire level of the tree. Such trees are more balanced, less prone to overfitting, and gives a faster prediction during the testing.

Catboost also suffers from some disadvantages. Firstly, Catboost is much more difficult to grasp conceptually than simpler methods such as simple decision trees or logistic regression. Understanding Catboost involves understanding the basics of ensemble learning, then the differences that Catboost has with other ensemble learning models. Secondly, Catboost's settings are also not necessarily intuitive, and time is needed to go through what each setting means and which settings are optimal. Here we give an informative overview of these settings and the defaults to choose.

Yandex CatBoost

You can check out CatBoost directly on their GitHub page by clicking here.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

Machine Learning Interpretability

In this chapter, we will learn about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

Specifically, we will cover the following topics:

What is Interpretability?
This section discusses the concept and origins of interpretability, as well as why it is a valueable tool to use in tandem with machine learning.

Local Interpretability
This section discusses the methods behind interpreteting results for a single observation.

Global Interpretability
This section discusses the methods behind interpreting results for the entire model as a whole.

Interpretability Tools
This section introduces some of the latest and greatest interpretabilities tools.

Select a topic from the left to begin.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

What is Interpretability? discusses the concept and origins of interpretability, as well as why it is a valueable tool to use in tandem with machine learning.

What is Interpretability?

Interpretability means different things for different people.

For the different roles involved in a data science or machine learning workflow, interpretability is a tool to help to solve the following:

Data Scientist To understand the model better, to see cases where the model does well or badly and why. This understanding helps the data scientist to build more robust models.	**Business Stakeholder** To gain a deeper understanding of why a system made a particular decision to ensure fairness and to protect its users and brand.
User To understand why a model made a decision and to allow for meaningful challenge if the model made a mistake.	**Expert or Regulator** To audit the AI system and follow the decision trail especially when things go wrong.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

What is Interpretability? discusses the concept and origins of interpretability, as well as why it is a valueable tool to use in tandem with machine learning.

What is Interpretability?

The Drivers Behind Interpretability

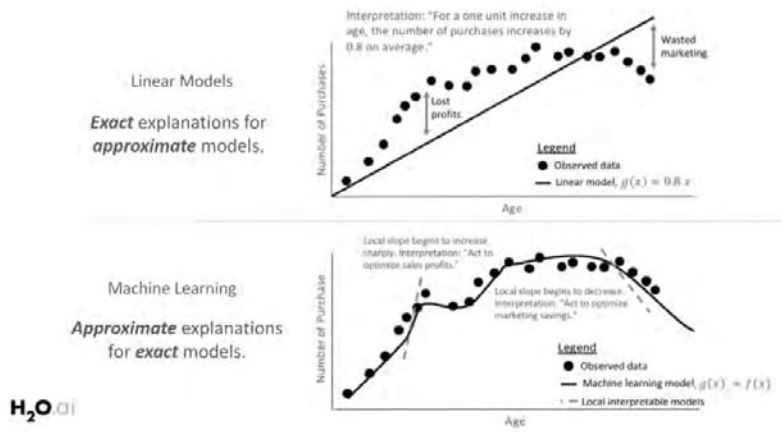
Trust – Interpretability is a prerequisite for trust – understanding how the model classifies in certain situations allows us to trust the model.

Causality – By interpreting certain models, we could generate hypotheses that scientists could then test.

Transferability – Humans exhibit a large capacity to generalize, so it is important that models are interpretable so it is clear what can be transferred and what cannot.

Informativeness – Whether the interpretation could be helpful for the user and useful for some decision.

Fair and Ethical Decision Making – Interpretations are needed to assess whether algorithms confirm to ethical standards. This is particularly relevant in light of the European Union's GDPR guidelines.



This picture is a great visualization for why machine learning interpretability is so difficult. Click here to read more.

What am I looking at?

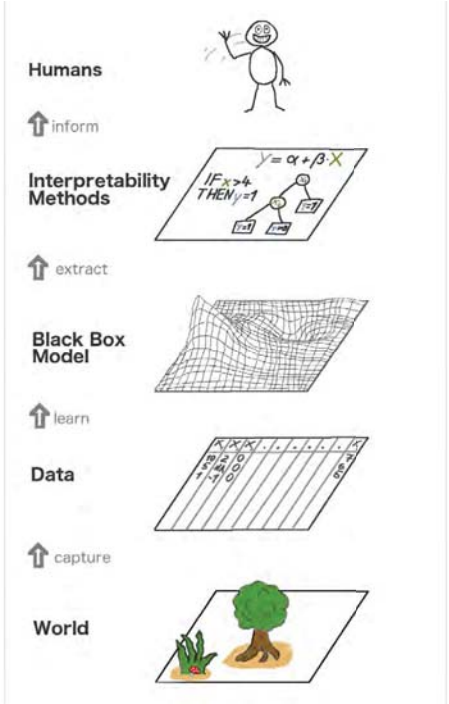
The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

What is Interpretability? discusses the concept and origins of interpretability, as well as why it is a valueable tool to use in tandem with machine learning.

What is Interpretability?



Click here for more on interpretability.

Interpretability Methods

"The Mythos of Model Interpretability" by Zachary Lipton considers two main types of transparency:

Transparency in the original model:
a. Simulatability
b. Decomposability
c. Algorithmic transparency

Post-hoc Interpretability - the application of interpretation methods after model training:
a. Text explanations
b. Visualization
c. Local explanations
d. Explanations by example

Our Primer focuses more on post-hoc interpretability – providing local explanations, examples, and explaining how to interpret the results. In contrast, decomposability or algorithmic transparency focuses more on the design of the algorithm, which is also a very important goal but is not covered here.

Click here to view Lipton's paper.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

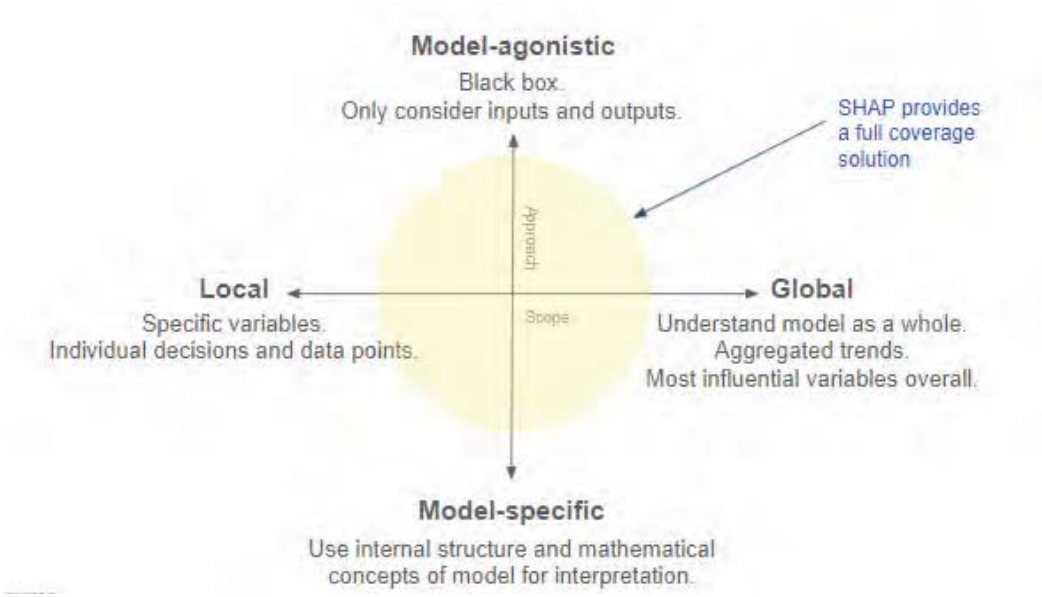
We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

What is Interpretability? discusses the concept and origins of interpretability, as well as why it is a valueable tool to use in tandem with machine learning.

What is Interpretability?

Taxonomy of Interpretability



What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

What is Interpretability? discusses the concept and origins of interpretability, as well as why it is a valueable tool to use in tandem with machine learning.

What is Interpretability?

General Notes on Interpretability

Certain machine models are quite simple to interpret. For example, linear and logistic regression are fairly easy to interpret. The model already provides the coefficient for each variable. The magnitude of every coefficient can be used to determine the impact of the variables, and each coefficient is translatable to an effect on the data. For example, a coefficient of 1 on a male binary variable indicates that being a male leads to a 1 unit log odds increase in the probability compared to being a female.

For certain more advanced models, interpretability is not a straightforward problem. While these models may give very accurate predictions, explaining the relationship between each feature and predicted value is not easy even from a global perspective. It is even more difficult to do so for each individual prediction at the local level. Furthermore, explaining whether a model helps identify which features are actually important, and whether or not it is affected by features that may just be noise, are very important for any user to understanding the model is behaving as intended.

Before diving into the interpretability of individual observations or features, it is important to understand the overall effectiveness of the model. A simple method is just to determine the overall percentage of correct classifications. However, for incorrect observations we often also wish to determine the difference between an incorrect prediction of true, called a False Positive, or an incorrect prediction of false, called a False Negative. Often these will have drastically different consequences - for example in the diagnosis of disease or detection of Fraud.

The false positive, false negative, true positive and true negative can be shown in a 2x2 table called a confusion matrix. This will give the full picture of how the incorrect predictions are allocated. The number in each square can represent either the count or the percentage of observations that make up that category, and a higher number means a greater percentage of the total is in that category.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

What is Interpretability? discusses the concept and origins of interpretability, as well as why it is a valueable tool to use in tandem with machine learning.

What is Interpretability?

General Notes on Interpretability

In addition to the Confusion Matrix, there are additional useful metrics that can be created from the False Positive and False Negative Numbers. The main metrics and their interpretations are below:

Accuracy: Percentage of true positives and negatives over all observations (ie the percentage of correct observations). This is useful as a general indicator of how the model is performing. However, in case of class imbalance (such as fraud) - it does not give enough info on the class of interest, and the model could be performing much worse than it appears.

Precision: Percentage of true positives of all positives. This gives the percentage of positives which are accurately predicted, and consequently, what the false positive rate is. This is useful for applications where false positives can be a big issue (such as cancer detection).

Recall: Percentage of true positives of observations that should be true (true positives and false negatives). This gives the percentage of true observations that should be captured, and is useful for applications where false negatives can be a big issue (such as incoming missile detection).

F1 Score: Calculated as 2 times (Precision multiplied by Recall) divided by (Precision plus Recall). This metric combines both precision and recall and gives an indicator of how the model is performing taking into account potential class imbalance. The F1 score scale goes from 0 to 1, with 0 being very poor and 1 being perfect.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive} \qquad \text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative} \qquad F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Click here for more about these metrics.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

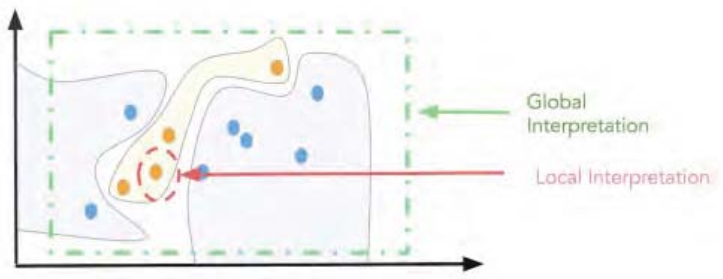
Local Interpretability discusses the methods behind interpreteting results for a single observation.

Local Interpretability

Introduction to Local Interpretability

One of the key things we are interested in for interpretability is what drives a specific observation. In particular, not only is it important to know about the prediction for each observation, it is also important to know what are the features that are driving that prediction, and how much they influence the prediction.

This is particularly important for black box predictions, and/or for situations in which someone who does not have access to the model may be interested in the prediction. As an example, credit scores are something that is calculated using an algorithm that has far reaching implications for any individual. It is thus important to know for a specific individual, what positively impacts their credit score, what negatively impacts it, and which factors have the most impact. For many black box models, it is impossible to directly determine the top features that drive the prediction for any one observation, so we use local interpretability tools to determine proxies for what features have the most impact, and in doing so can determine for a person why their credit score is a certain number.



Click here for a great article about the types of Interpretability.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

Local Interpretability discusses the methods behind interpreteting results for a single observation.

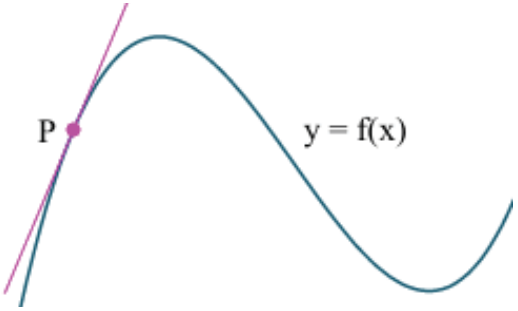
Local Interpretability

Local Interpretability Methods

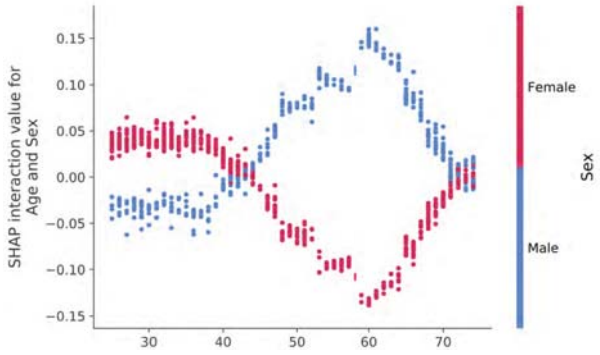
Major ways in which local interpretability can be calculated include:

Perturbation methods (for example LIME)
ICE plots
SHAP

We cover the LIME and ICE plots in the Local Interpretability Section, and SHAP in both the Interpretability Tools section as well as the Tutorial.



Local Interpretability is like fitting a line to a curved prediction - the line gives the local approximation and is much easier to understand.



SHAP Joint Feature Plot

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

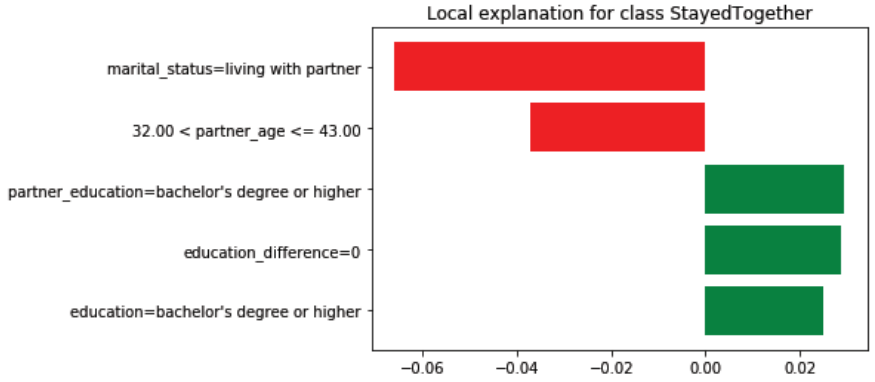
The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

Local Interpretability discusses the methods behind interpreteting results for a single observation.

Local Interpretability

LIME Methodology

LIME is one way of calculating local interpretability, developed by Rubeiro, Singh and Guestrin in 2016. This is the main method by which LIME creates its local prediction: Choose the instance of interest for which we want a local prediction. Perturb the dataset to generate new points Get the local predictions for these new points. Draw samples from the new perturbed predictions, with the samples weighted by their proximity to the instance of interest. Fit a weighted, interpretable model on the dataset with the variations. LIME attempts to find the local model which has the closest predictions to the original model that is interpretable by humans. Explain prediction by interpreting the local model.



The above image shows a sample LIME output

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

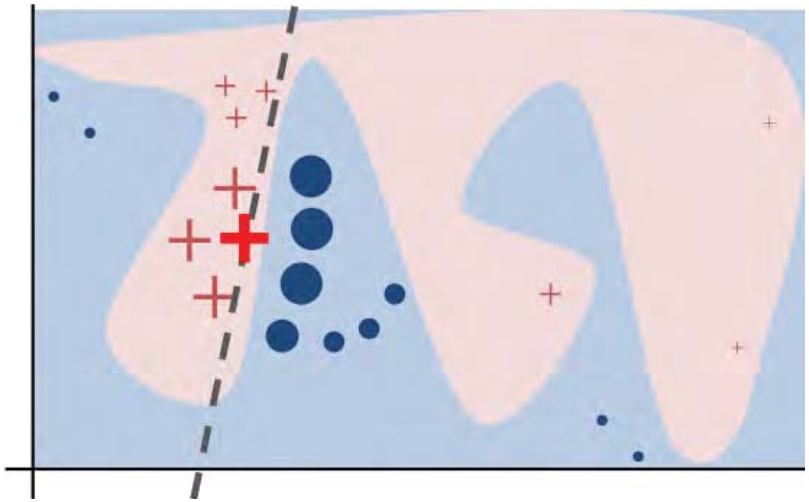
Local Interpretability discusses the methods behind interpreteting results for a single observation.

Local Interpretability

LIME Visualization

Here is a demonstration of how the LIME methodology works:

The overall model (unknown) is represented by the blue/pink background.
The bold red cross is the instance being explained.
The proximity to the instance is represented here by size.
The dashed line is the learned explanation that is locally (but not globally) faithful.



Click here to learn more about LIME.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

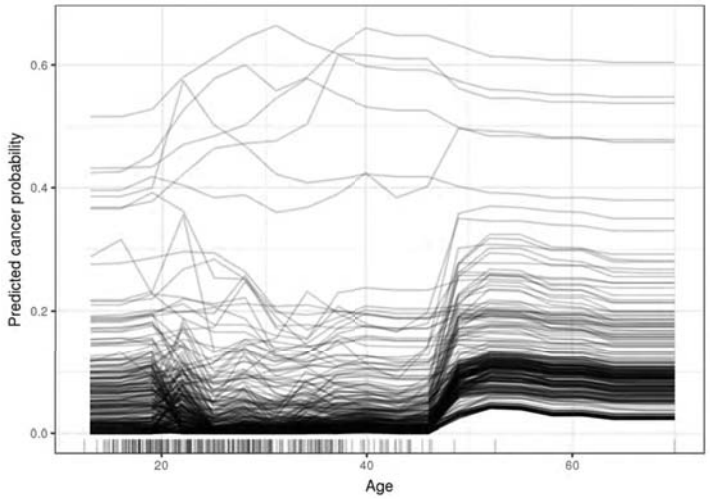
The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

Local Interpretability discusses the methods behind interpreteting results for a single observation.

Local Interpretability

Individual Conditional Expectation

Another method for providing local interpretability is Individual Conditional Expectation (ICE). An ICE plot visualizes the dependence of the prediction on a feature for each instance separately, resulting in one line per instance. The values for a line (and one instance) can be computed by keeping all other features the same, while making predictions with model for other values of this one feature. An individual conditional expectation (ICE) plot is the equivalent to a PDP for individual data instances. This gives a localized interpretation for each observation.



Click here for more about ICE plots.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

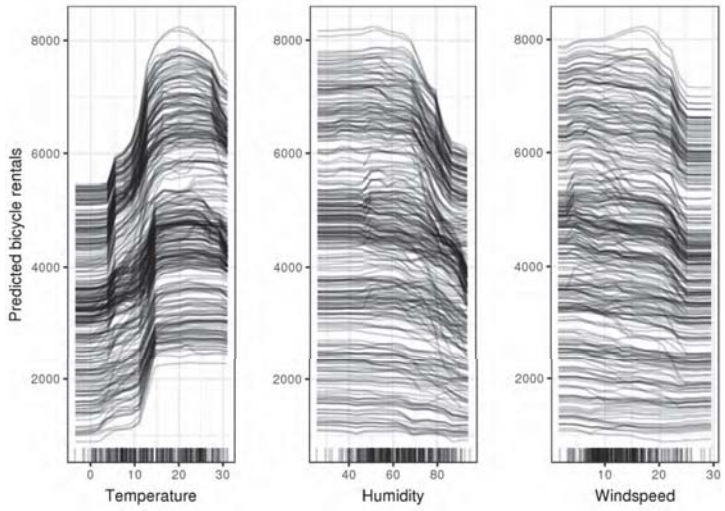
Local Interpretability discusses the methods behind interpreteting results for a single observation.

Local Interpretability

Advantages and Disadvanteges of ICE

ICE plots are very intuitive to understand - one line represents the predictions for one instance if we vary the feature of interest. Also, they can uncover heterogeneous relationships - different feature impacts for different observations can be revealed by the ICE plot.

On the flip side, ICE plots can only display one feature meaningfully and even for one feature, the plot can become overcrowded. It is also difficult to isolate a specific observation that you are interested in from the plot.



Click here for more about ICE plots.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

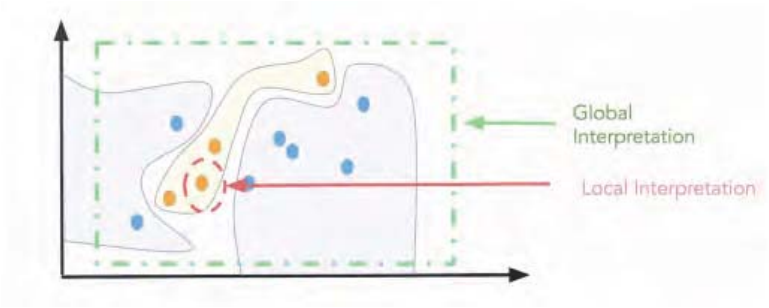
Global Interpretability discusses the methods behind interpreting results for the entire model as a whole.

Global Interpretability

Introduction to Global Interpretability

There are also times when we are interested in what factors are predicted as being most important overall, and how much impact does each factor have. Sometimes we are concerned with the whole population - for example, what factors best drive test score performance? Such questions can allow us to focus on the important factors for improving education, as well as explaining potential issues with how the test is designed.

To do this, global interpretability models can be used. These allow us to determine the magnitude of features impact on the average prediction for the whole population, which the most impactful features. In addition, global interpretability is also very useful for feature selection. For most models, there are a lot of features to choose from, and even more features can be created via feature engineering. In order to improve the accuracy and run time of the model, it is important to identify the key fields in the data so that the model can give a prediction with less noise. Global Interpretability tools allow us to find the most impactful fields which allow us to conduct the necessary feature selection.



Click here for a great article about the types of Interpretability.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

Global Interpretability discusses the methods behind interpreting results for the entire model as a whole.

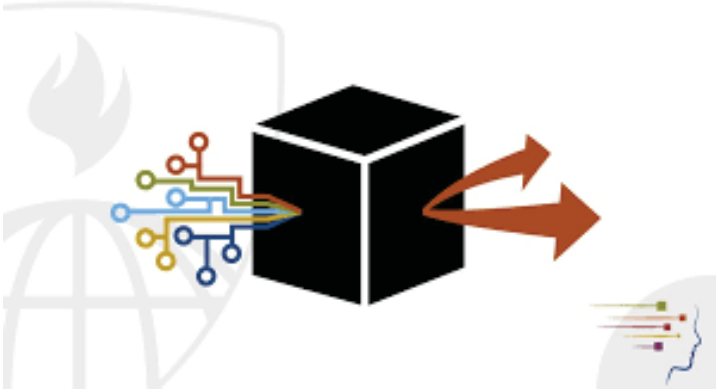
Global Interpretability

Global Interpretability Methods

There are a few key Global Interpretability methods to help explore the relationship of features and their impact on the overall model.

Major methods that we explore are Partial Dependence Plots (PDP) and Feature Importance.

In the next section, we will talk about some tools to help with both Local and Global Interpretability, such as SHapely Additive exPlanations (SHAP).



Global interpretability is finding the overall impact of a feature for a black box model

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

Global Interpretability discusses the methods behind interpreting results for the entire model as a whole.

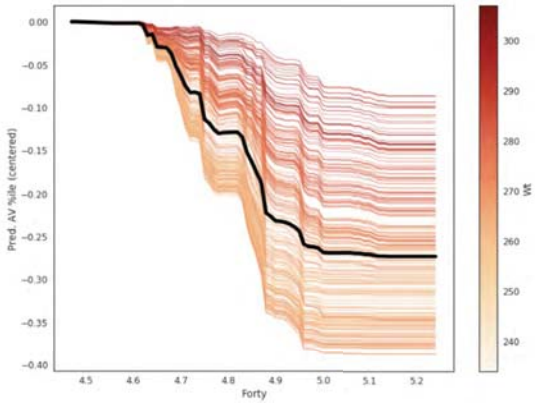
Global Interpretability

Patial Dependency Plots

Partial Dependence Plots are a global interpretability method, as they give insights for all observations in a dataset.

PDPs show the marginal effect one or two features have on the predicted outcome of a machine learning model. For classification problems, the partial dependence plot displays the probability for a certain class given different values for a feature or features.

This is useful because it helps isolate the influence of one feature on the model prediction, so you can better understand what role each feature plays.



Partial Depedence Plot combined with a centered ICE plot

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

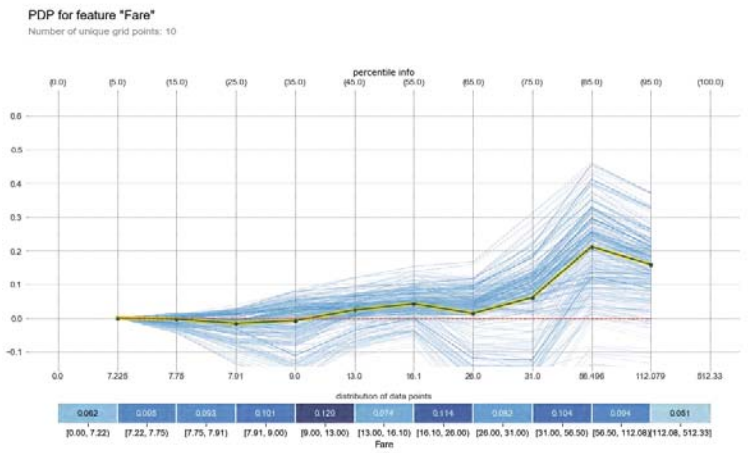
Global Interpretability discusses the methods behind interpreting results for the entire model as a whole.

Global Interpretability

Advantages and Disadvantages of PDP

PDPs are an easy way to understand the effect of one feature on the model. They are also intuitive to understand - it is easy to track the impact of one feature based on how values of that feature change. PDPs can also be combined with ICE plots to give a global and local picture.

As for negatives, PDPs operate under the assumption of independence between features, which in reality is hardly ever true. PDPs are also not able to show heterogenous (differentiated) effects.



Another Partial Dependence Plot and ICE plot example.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

Global Interpretability discusses the methods behind interpreting results for the entire model as a whole.

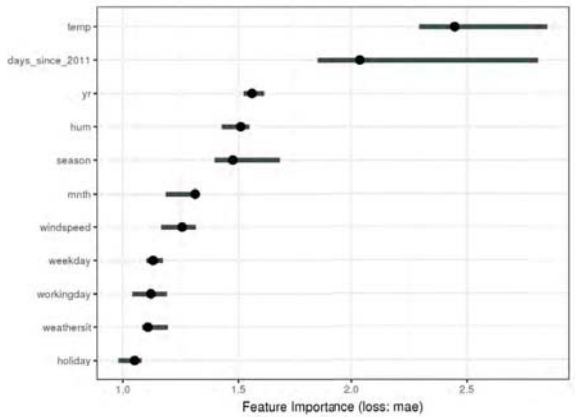
Global Interpretability

Feature Importance

Another global interpretability method is Feature Importance. Feature importance is defined as follows: the increase in the prediction error of the model after we permuted the feature's values. A feature is "important" if shuffling its values increases the model error, because in this case the model relied on the feature for the prediction. A feature is "unimportant" if shuffling its values leaves the model error unchanged, because in this case the model ignored the feature for the prediction.

There are two main ways to measure feature importance:
Mean Decrease Impurity (MDI)
Mean Decrease Accuracy (MDA) or Permutation Importance

From Feature Importance, we can gain an understanding of the top features by importance for any dataset.



For this model about bike rentals, the temperature field has the most importance and the holiday field has the least. Click here to learn more.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

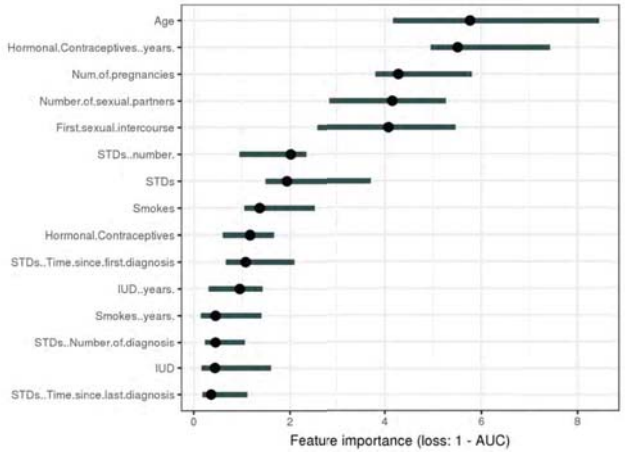
Global Interpretability discusses the methods behind interpreting results for the entire model as a whole.

Global Interpretability

Advantages and Disadvantages of Feature Importance

Feature importance is a strong method because it provides a highly compressed, global insight into the model's behavior. Feature importance measurements are comparable across different problems. when using permutation, feature importance takes into account both the main feature effect and the interaction effects on model performance. Permutation feature importance does not require retraining the model.

Feature importance does have some draw backs. For example, it is very unclear whether you should use training or test data to compute the feature importance as you need access to the true outcome. Permutation feature importance depends on shuffling the feature, which adds randomness to the measurement. Adding a correlated feature can decrease the importance of the associated feature by splitting the importance between both features. This reduces the accuracy of the ordering of feature importance.



For this model about cervical cancer, the age field has the most importance. Click here to learn more.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

Interpretability Tools introduces some of the latest and greatest interpretabilities tools.

Interpretability Tools

Interpretability Tool Examples

Interpretability has only recently become a topic of interest – initial methods were first pioneered in 2014 and new methods have been introduced each year. In our primer, we cover the main methodology of these different interpretability tools, as well as explore then state of the art in interpretability. We then do a deep dive in particular of SHAP, which is not only a state of the art method but also relatively simple to explain in contrast to other methods

There are many different tools which help with both local and Global Interpretability. Some of the main ones are as follows:

LIME – LIME interprets individual model predictions based on locally approximating the model around a given prediction. LIME minimizes an objective function to find its weights.

DeepLIFT – attributes to each input x a value that represents the effect of that input being set to a reference value as opposed to its original value.

Layer-Wise Relevance Propagation – Interprets the predictions of deep networks. This method is equivalent to DeepLIFT with the reference activations of all neurons fixed to zero.

Classic Shapley Value Estimation – Three methods use game theory to compute explanations of model predictions

For our Primer we focus mainly on **SHAP**, which is a state of the art method introduced by Scott Lundberg which combines the tools mentioned above..

What am I looking at?

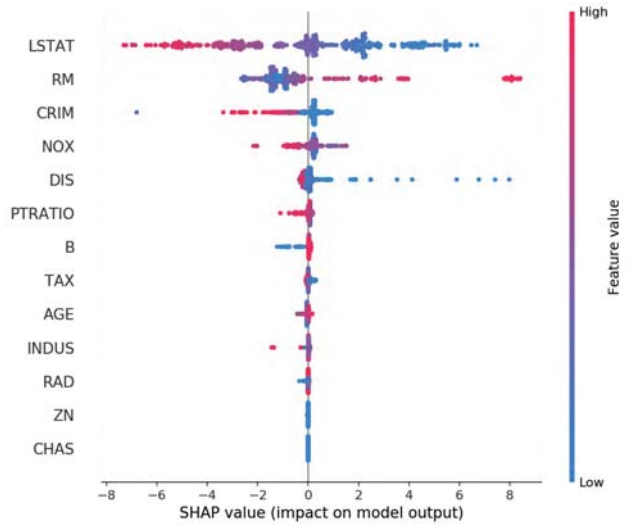
The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

Interpretability Tools introduces some of the latest and greatest interpretabilities tools.

Interpretability Tools



SHAP feature importance plot

What is SHAP?

SHAP (short for SHapley exPlanatory values) is a method of explaining feature importance developed by Scott Lundberg in 2017. SHAP uses Shapley values to determine the marginal contribution of a feature across all feature combinations.

Some "SHAP value" assignment axioms:
Local Accuracy: when approximating the original model for a specific input, the explanation model matches the original model
Missingness: Features that do not contribute receive a SHAP value of 0
Consistency: If a simplified input's contribution increases or stays the same in the model, that input's attribution should not decrease.

One advantage of SHAP is that the difference between the prediction and the average prediction is fairly distributed among the feature values of the instance – the Efficiency property of Shapley values. This is in contrast to LIME, which does not guarantee that the prediction is fairly distributed among the features. Also, the Shapley value allows contrastive explanations. Instead of comparing a prediction to the average prediction of the entire dataset, you could compare it to a subset or even to a single data point.

As for disadvantages, shapley values require a lot of computing time. All permutations need to be calculated which can take a significant amount of time, particularly if there are a large number of features. The number of permutations increase exponentially with the number of features.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

Interpretability Tools introduces some of the latest and greatest interpretabilities tools.

Interpretability Tools

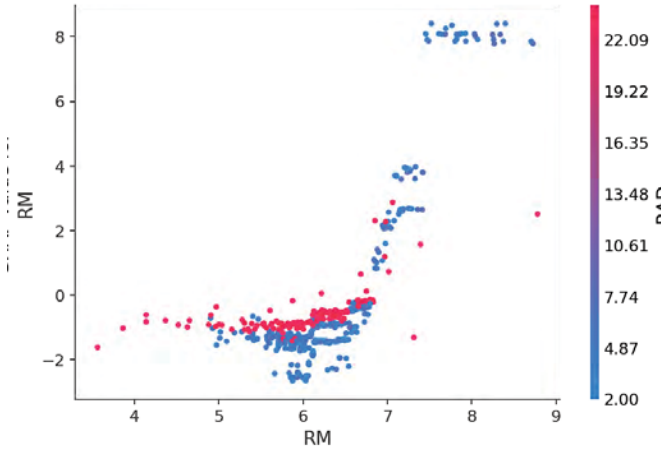
The calculation of SHAP can be split into two components - the calculation of the contribution, and the calculation across all feature combinations. The calculation of the contribution for each feature can be done by sampling feature values from the data and determining the contribution of the feature for each sample. The samples can be repeated and averaged to get an average contribution for that permutations of feature combinations.

As an example calculation for a given sample, if including just age of 30 results in a 10% survival rate prediction and when gender of female is added there is a 30% survival rate, then female is assigned 20%. From the total list of features, there is a set of permutations for the feature list. As a example, assume there are 3 features, gender, age and class. We wish to determine the SHAP values for gender.

There are then the following permutations when gender is not included, for which Shapley values must be calculated for: no features, age, class, and age & class.

For each of these four features the predicted values are computed for males and females, to get the marginal contribution of being male. The Shapley value used by SHAP is calculated as the weighted average of these marginal contributions.

We've now explained how the basics of how SHAP works. SHAP has a variety of tools which can be used to assist with both global and local interpretability. To really see how useful SHAP can be, we have prepared a tutorial on the different interpretability tools that can be provided in SHAP. Click on the Tutorial tab to navigate to the tutorial.



SHAP Joint Feature Importance Graph

SHAP Force Plot

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

Interpretability Tools introduces some of the latest and greatest interpretabilities tools.

Interpretability Tools

What is DeepLIFT?

DeepLIFT (Deep Learning Important FeaTures) is an algorithm to assign importance score to the inputs for a given output.

DeepLIFT frames the question of importance in terms of differences from a 'reference' state, where the 'reference' is chosen by the user according to what is appropriate for the problem at hand. In contrast to most gradient-based methods, using a difference-from-reference allows DeepLIFT to propagate an importance signal even in situations where the gradient is zero and avoids artifacts caused by discontinuities in the gradient.

The choice of a reference input is critical for obtaining insightful results from DeepLIFT. In practice, choosing a good reference would rely on domain-specific knowledge, and in some cases it may be best to compute DeepLIFT scores against multiple different references.

Click here to view DeepLIFT on GitHub.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

Interpretability Tools introduces some of the latest and greatest interpretabilities tools.

Interpretability Tools

Interpretability for Tree-Based Models

Interpretability for tree-based models is introduced in the paper "Interpretable Predictions of Tree Based Ensemble Models via Actionable Feature Tweaking", by Tolomei, Silvestri, Haines, Lalmas.

The paper explores how a feature vector can be changed to modify the prediction, and it presents a technique that exploits the internals of a tree-based ensemble classifier to offer recommendations for transforming true negative instances into positively predicted ones.

The methodology used is first a tree based algorithm is built for binary classification – and then a feature tweaking algorithm is built to determine the cost that is needed to convert the negative instance into a positive one, based on a set of input features that indicate a true negative instance. This cost is then used to evaluate various recommendations.

Click here to view the paper.

What am I looking at?

The Machine Learning Primer is a flipbook of concepts and information to introduce you to machine learning.

We want you to walk away from this primer with increased trust and understanding in machine learning.

The **Machine Learning Interpretability** section will teach you about the two types of interpretability and how we can use tools to aid us in understanding the results of our machine learning models.

Interpretability Tools introduces some of the latest and greatest interpretabilities tools.

Interpretability Tools

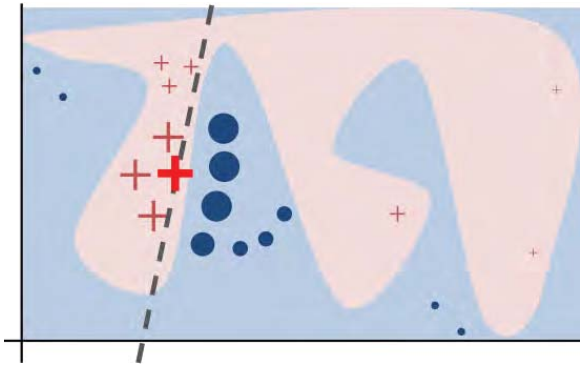
What is LIME?

LIME was a method that was introduced in a paper published in 2016 by Ribeiro, Singh and Guestrin, titled "Why Should I Trust You?" Explaining the Predictions of Any Classifier. As detailed in their paper, the main purpose of LIME is to explain the predictions of any classifier or regression in a faithful way, by approximately it locally with an interpretable model.

LIME then tries to find the model which has the closest predictions for that locality compared to the overall algorithm, while making sure that the model is still interpretable by humans.

LIME was a method that was introduced in a paper published in 2016 by Ribeiro, Singh and Guestrin, titled "Why Should I Trust You?" Explaining the Predictions of Any Classifier. As detailed in their paper, the main purpose of LIME is to explain the predictions of any classifier or regression in a faithful way, by approximately it locally with an interpretable model. The methodology involves finding a model which can approximate the overall algorithm, around a specified locality. LIME then tries to find the model which has the closest predictions for that locality compared to the overall algorithm, while making sure that the model is still interpretable by humans.

They also introduce SP-Lime, which selects a set of representative instances with explanations to address the global problem, via submodular optimization.



Recall the LIME example we discussed in the Local Interpretability section. Click here to learn more about LIME.

Welcome to the Machine Learning Example Walkthrough!

We have covered many of the techniques that can be done to help understand a machine learning model.
Now we wish to show exactly how these techniques can be applied.

To demonstrate these interpretability tools, here is a walkthrough of how these interpretability tools can be used to provide both local interpretability and global interpretability.

In this walkthrough we will cover the entire process of fitting a model to a classification problem, then showcasing both the global and local interpretation of our model results.

Dataset

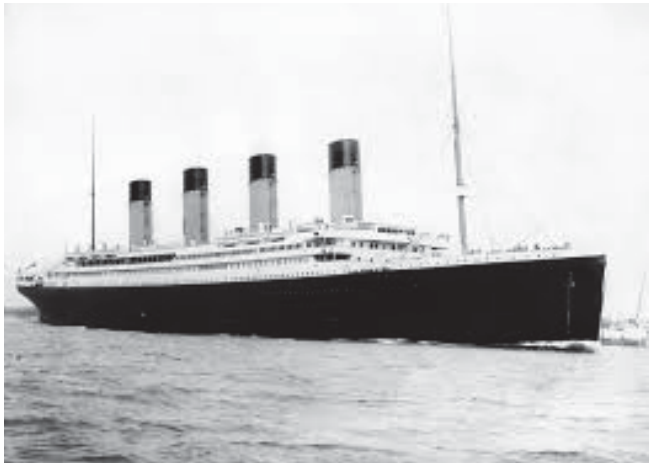
We use the Titantic Dataset as our example dataset. This is one of the introductory machine learning datasets, and contains fairly straightforward features. The data contains a field on whether a person survived, as well as two groups of fields, on person characteristics and trip characteristics.

The fields on the person characteristics include gender, age, siblings and name.
The fields on the trip characteristics include class, cabin, fare, cabin, embarked and ticket.

A snapshot of the dataset is given here:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Goal & Methodology



Our goal is to showcase a classification problem, where we aim to classify whether a passenger survived based on the person characteristics and trip characteristics are.

We use the Catboost model to make the prediction, since it is one of the leading classification algorithms and it also contains the option to automatically process the categorical variables so they can be included in the model without additional feature engineering. (More information on how Catboost works and treats categorical variables can be found on the Machine Learning Algorithm section on Catboost).

We use a few main methods to provide interpretability for our model. Firstly, we look at the F1-score of our model to determine the general accuracy in its predictions. We then use the Permutation Importance plot in the Eli5 ("Explain Like I'm 5") package to determine the features with the greatest impact in our model to help with global interpretability. Finally, we use SHAP to provide a variety of plots that help give both local and global interpretability for our model.

Cleaning & Setup

While the Titanic dataset is not particularly complex, we do conduct some simple dataset cleaning and feature selection.

Missing observations for Cabin are filled in with “Undefined”, and missing observations for Embarked are filled in with “S” for Southhampton
An observation with missing fare has the fare imputed with an average value.

Certain fields in the data are unique or almost by observation, and thus will not be helpful for our analysis. As a result, these fields are removed before the model stage. The list of fields include: Passenger ID, Name, Ticket.

Testing & Optimization

Next, we wish to optimize our model. Catboost has a large list of parameters, and it is important to make sure that we put in the optimal values for these parameters so that our model has the highest accuracy possible. To do this, we conduct a hyperopt optimization, where we are able to try a range of parameter inputs for the Catboost function, and the hyperopt optimization will choose the parameters with the best loss values. The hyperopt optimization will conduct its own cross validation in order to determine the value of the loss function.

Once we have determined the top model, we can fit it to a test dataset to get the accuracy of the model. We can also extract the sample parameters, in our case the parameters are as follows:

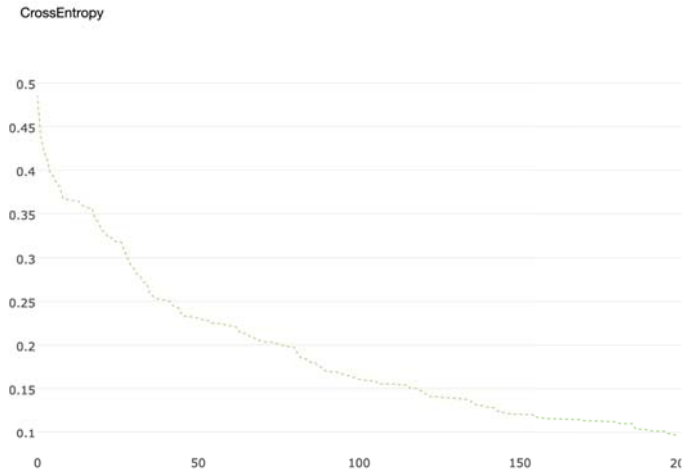
```
print(best_params)
```

```
{'l2_leaf_reg': 1.0, 'learning_rate': 0.4164833841251649, 'max_depth': 8}
```

Overall Model Accuracy

We can then test the overall model accuracy using the F1-score. The F1-score is covered in more detail in our interpretability primer, and we use it because it gives a better accuracy when the class sizes might be imbalanced. We also output the cross entropy loss, which shows how the loss function changes as the predicted probability diverges from the actual label. The F1-Score and Cross Entropy for our model is:

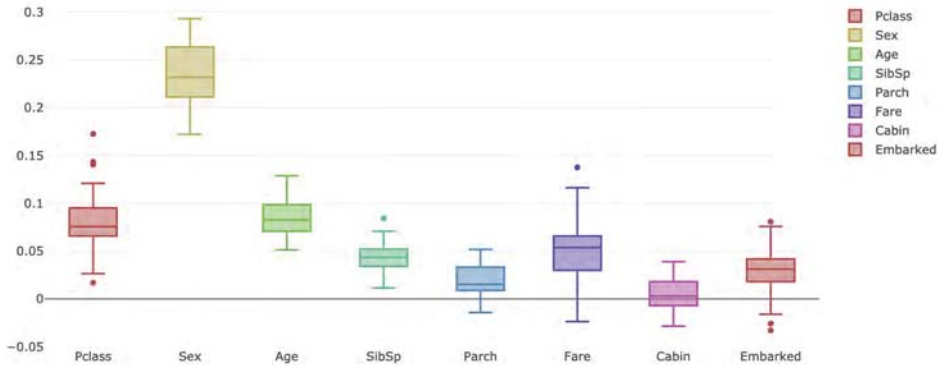
```
Test Dataset
-------------
F1-Score: 0.738
Accuracy: 0.793
ROC-AUC: 0.786
-------------
True Positives: 52
False Positive: 20
True Negative: 90
False Negative: 17
Precision: 0.72
Recall: 0.75
```



Permutation Importance

Now that we have a overall sense of how the model does, we would like to use the permutation importance as a measure of the global impact for each feature. The permutation importance assigns a impact for each feature based on the decrease in the score (which in our case is defined as the F1 score) after the feature is removed.

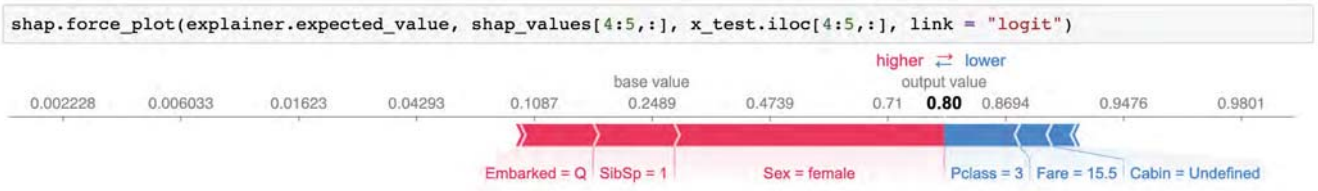
For the Titianic dataset, sex has significantly more impact than the other fields.
Some fields such as fare and Passenger class have much larger range of impact than other fields such as Sibsp
Permutation importance is much more useful for global insights than local insights on what features impact one observation.



SHAP - Force Plot

Now we will transition from the global insights provided by permutation importance into more localized insights. The SHAP force plot is an ideal graph for showing the main factors that impact the prediction for a single observation. Below is the default format provided by the output of the SHAP Force Plot - it provides the prediction probability for an individual observation as well as the magnitude of each feature impact and how they affected the prediction.

For this sample observation, the predicted probability of survival was 80%, and this was largely driven by gender, number of siblings and embark location. The passenger class and fare were negative factors but have a smaller impact in this prediction.

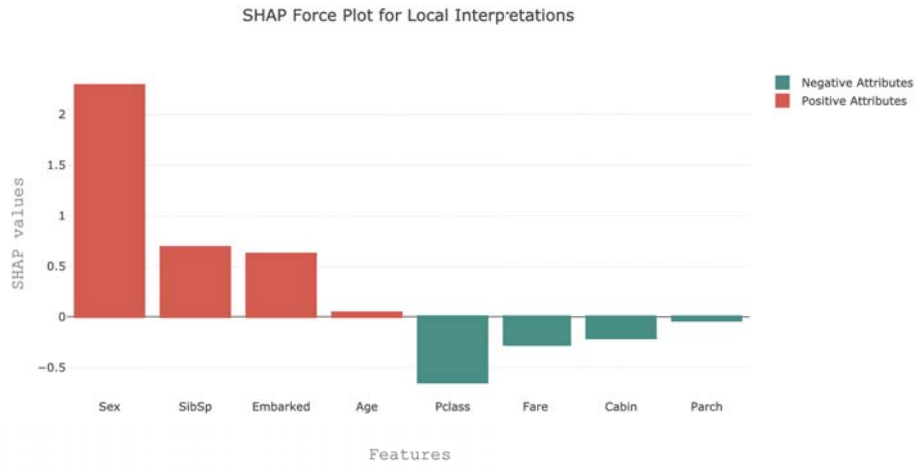


Legend

RED = features that *increase* the probability
BLUE = features that *decrease* the probability
LENGTH = magnitude of the influence (SHAP value)
VALUE = specific value features impacting this prediction

SHAP - Force Plot

One drawback we noticed with the default SHAP graph is that it is difficult to compare the magnitude of different features using the scale that was provided. As a result, we have reformatted the output - this allows significantly better comparison of the positive and negative factors:

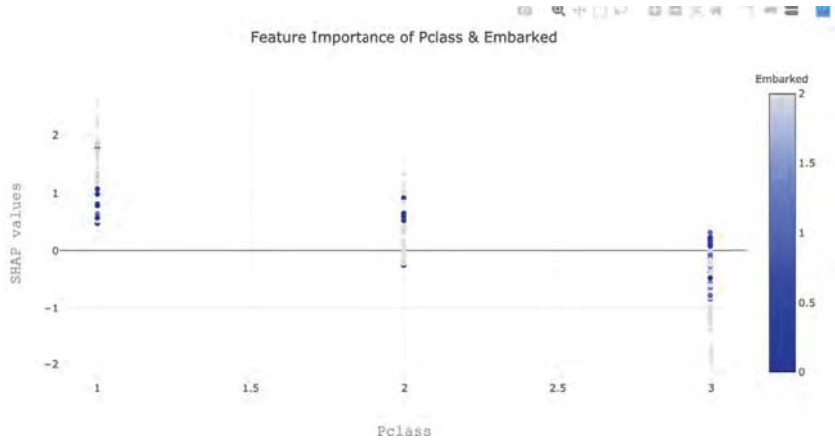


SHAP - Pairwise Joint Feature Importance

We have covered both global interpretability and local interpretability for features individually - that is how would a feature in isolation impact the overall prediction or prediction for one observation. However, in some situations there is significant feature interaction, which means that some features may have a strong or weak impact dependent on another feature. An example of this could be the interaction of work experience and age when predicting income - for people who are in their 20s, work experience may be very important, but for people in their 40s and 50s, work experience likely has a much smaller impact.

To capture this effect, we output joint feature importance graphs for the top 3 feature interactions. The joint feature importance graphs display any impact that one feature has on the impact of another feature. Here is an example of one of the joint feature importance graphs for the Titanic Dataset.

In this graph, we see that while 1st class generally leads to positive SHAP values and 3rd class generally leads to negative SHAP values as expected, there is a difference in impact depending on where a passenger embarked. Passengers who embarked at location 0 (which corresponds to Cherboug) tend to have SHAP values that have significantly less magnitude that those who embarked at location 2 (which corresponds to Southampton). This suggests that for some reason, class has a large impact on the survival of passengers from Cherboug, but not those from Southampton.



UNRAVEL
Machine Learning, Explained

Welcome to the Machine Learning Interpretability Tool!

We have covered many of the techniques that can be done to help understand a machine learning model.
We have shown exactly how these techniques can be applied.
Now we want to put the power of machine learning into your hands.

The Machine Learning Interpretability Tool will ingest your own data and create a machine learning model that fits for your problem. Then, the tool will walk through some of the interpretability methods that we have learned in order to give you an indepth understanding of your data.

Currently, the tool is a minimum viable product that supports supervised machine learning problems only.

Click the button below to launch the tool.

Begin

Nice to meet you!

We are the team behind UnravelML.

Unravel was created as a capstone project for the Masters of Information and Data Science (MIDS) program at UC Berkeley.

Please click on one of us to learn more and to reach out with any questions and comments.

	Kevin Pang	Kathryn Hamilton	Tianhao Xu
Location	New York	Michigan	Bay Area
Industry	Insurance	Automotive	Consulting
Background	Engineering, Operations Research	Engineering, Mathematics	Economics, Mathematics