Warsztaty Wyjaśnialne Uczenie Maszynowe (XAI)

Anna Kozak
Przemysław Biecek

Agenda

- dlaczego XAI?
- use case - dane finansowe FICO
- use case - wycena nieruchomości
- analiza XAI w R

Apple Card and Goldman Sachs accused of gender discrimination in credit card algorithm

Posted earlier today at 6:38am

PHOTO: Apple Card has caused a Twitter storm and even prompted a regulator inquiry. (Reuters: Stephen Lam)

Apple launched its own credit card in the United States a few months ago, selling it on the ability to help people keep track of their spending while protecting their privacy.

RELATED STORY: Apple wants you to read, watch and play its new services … using its new credit card

REUTERS World Business Markets Politics TV

Midterm Elections | Imprisoned In Myanmar | Sectors Up Close | Breakingviews | Investing | Future of Money | Charged: The Future of Aut

BUSINESS NEWS OCTOBER 10, 2018 / 5:12 AM / A MONTH AGO

Amazon scraps secret AI showed bias against wom

Jeffrey Dastin

SAN FRANCISCO (Reuters) - Amazon.com Inc's specialists uncovered a big problem: their

The group created 500 computer models focused on specific job functions and locations. They taught each to recognize some 50,000 terms that showed up on past candidates' resumes. The algorithms learned to assign little significance to skills that were common across IT applicants, such as the ability to write various computer codes, the people said.

Instead, the technology favored candidates who described themselves using verbs more commonly found on male engineers' resumes, such as "executed" and "captured," one person said.

Amazon trained a sexism-fighting, resume-screening AI with sexist hiring data, so the bot became sexist

THE VERGE TECH SCIENCE C

TECH AMAZON ARTIFICIAL INTELLIGENCE

21

Amazon reportedly scraps internal AI recruiting tool that was biased against women

The secret program penalized applications that contained the word "women's"

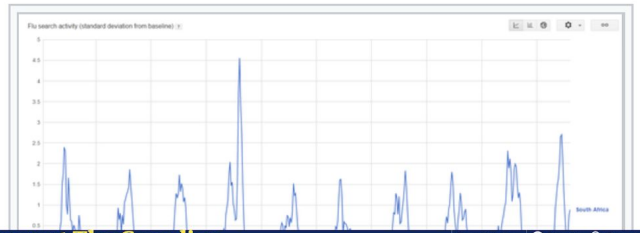
By James Vincent | @jjvincent | Oct 10, 2018, 7:09am EDT

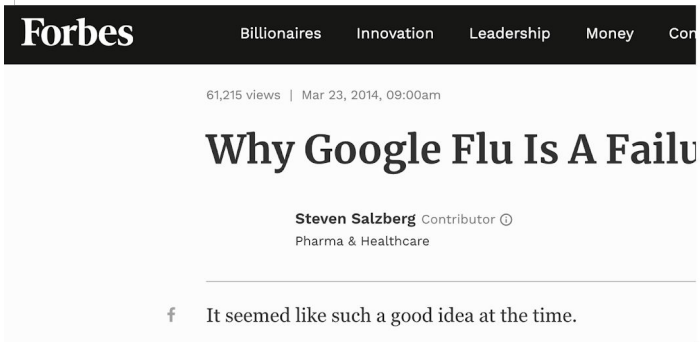
Google Flu Trends

From Wikipedia, the free encyclopedia

Google Flu Trends was a web service operated by Google. It provided estimates of influenza activity for more than 25 countries. By aggregating Google Search queries, it attempted to make accurate predictions about flu activity. This project was first launched in 2008 by Google.org to help predict outbreaks of flu.[1]

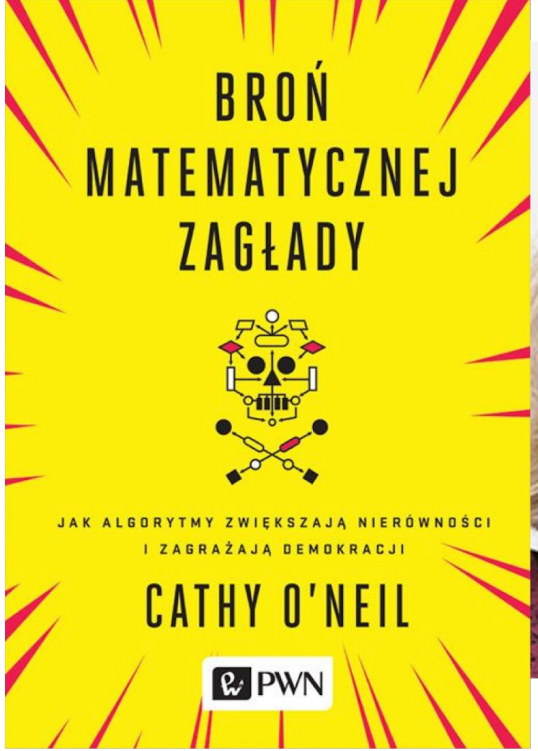
Google Flu Trends is now no longer publishing current estimates. Historical estimates are still available for download, and current data are offerec





https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

Cathy O'Neil:
The era of blind faith in ~~big data~~ black boxes must end



- “You don’t see a lot of skepticism,” she says. “The algorithms are like shiny new toys that we can’t resist using. We trust them so much that we project meaning on to them.”
- Ultimately algorithms, according to O’Neil, reinforce discrimination and widen inequality, “using people’s fear and trust of mathematics to prevent them from asking questions”.

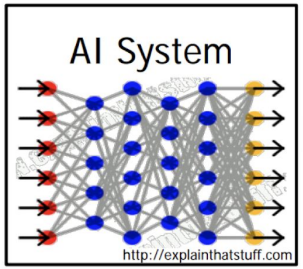
Machine Learning is Creating a Crisis in Science

The adoption of machine-learning techniques is contributing to a worrying number of research findings that cannot be repeated by other researchers.

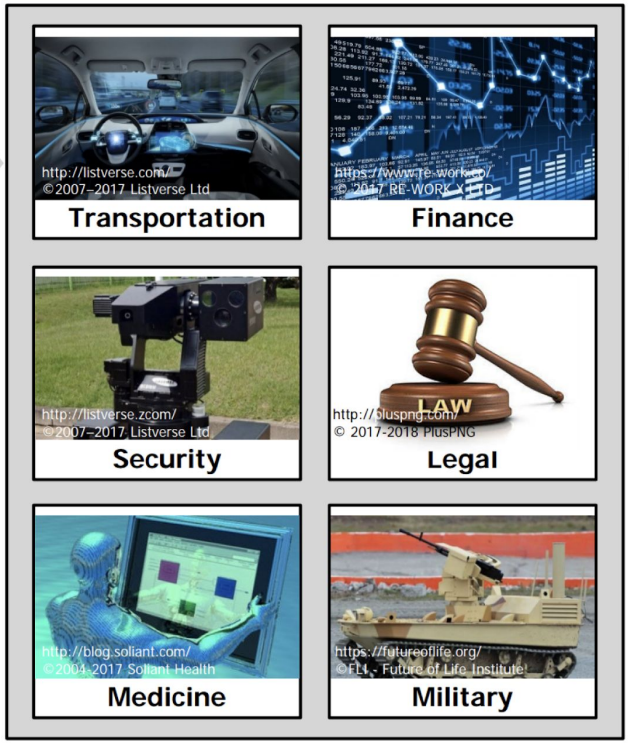
Kevin McCaney
Wed, 02/27/2019 - 11:28

Photo credit: metamorworks/iStock

The Need for Explainable AI



- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand



- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

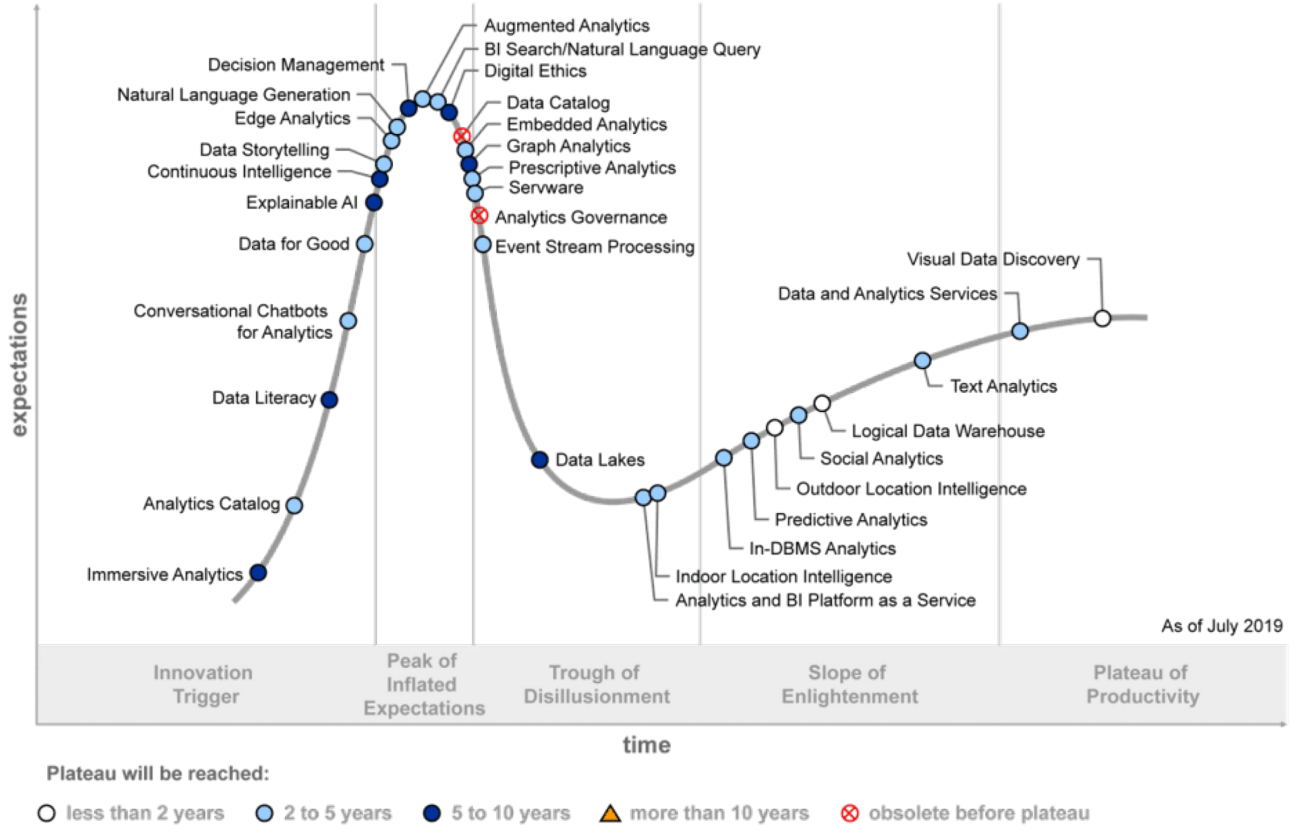
- The current generation of AI systems offer tremendous benefits, but their effectiveness will be limited by the machine's inability to explain its decisions and actions to users
- Explainable AI will be essential if users are to understand, appropriately trust, and effectively manage this incoming generation of artificially intelligent partners

Nowe regulacje?

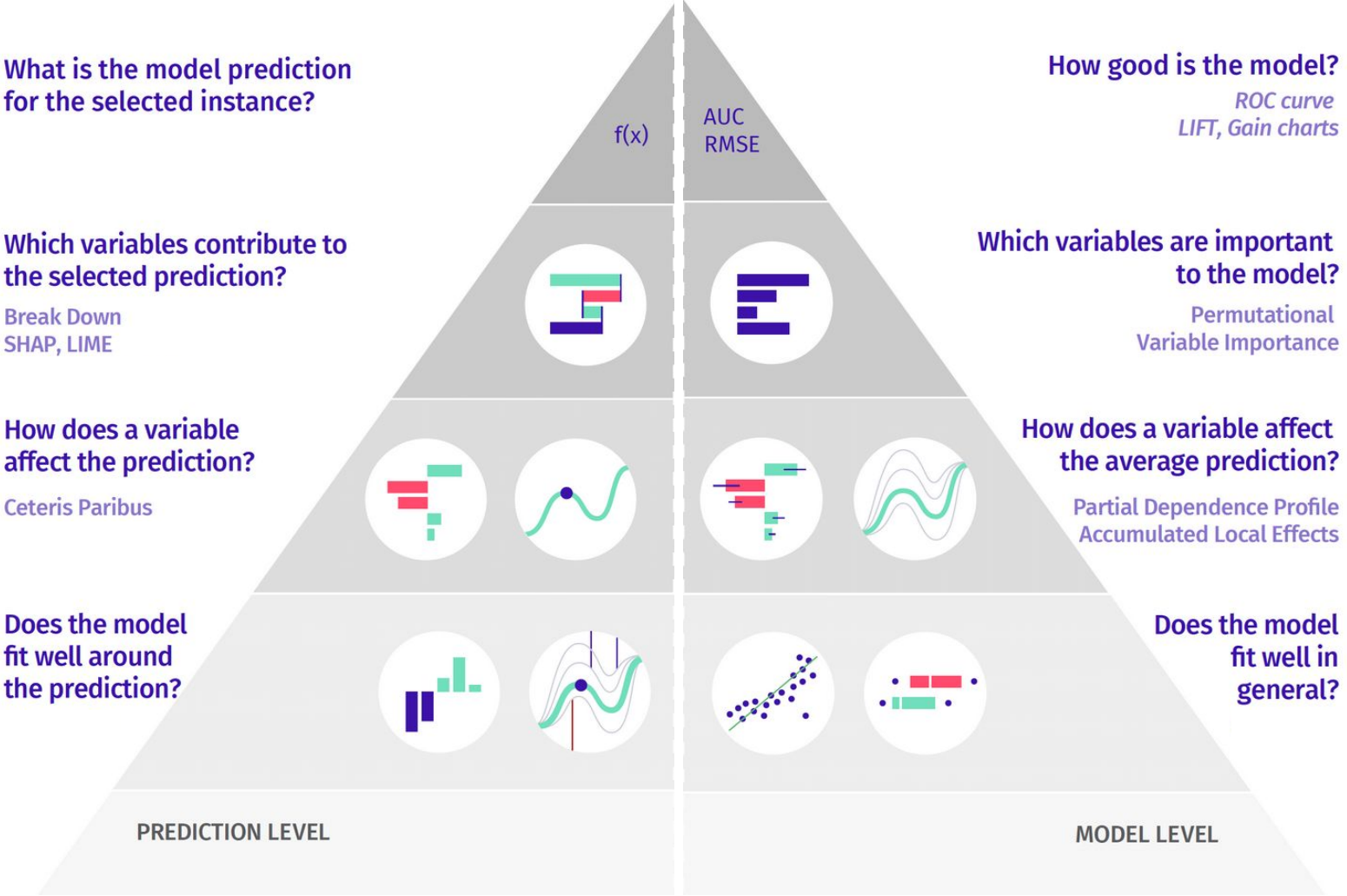
czy

Nowe możliwości?

Hype Cycle for Analytics and Business Intelligence, 2019



Source: Gartner
ID: 369713


What is the model prediction for the selected instance?
f(x)
AUC
RMSE
How good is the model?
ROC curve
LIFT, Gain charts
Which variables contribute to the selected prediction?
Break Down
SHAP, LIME
Which variables are important to the model?
Permutational
Variable Importance
How does a variable affect the prediction?
Ceteris Paribus
How does a variable affect the average prediction?
Partial Dependence Profile
Accumulated Local Effects
Does the model fit well around the prediction?
Does the model fit well in general?
PREDICTION LEVEL
MODEL LEVEL

FICO

FICO

From Wikipedia, the free encyclopedia

For other uses, see FICO (disambiguation).

FICO (legal name: **Fair Isaac Corporation**), originally **F**air, **I**saac and **Co**mpany, is a data analytics company based in San Jose, California focused on credit scoring services. It was founded by Bill Fair and Earl Isaac in 1956.[2] Its FICO score, a measure of consumer credit risk,[3] has become a fixture of consumer lending in the United States.

Predictor variables

The predictor variables are all quantitative or categorical, and come from anonymized credit bureau data.

Please refer to the data dictionary for full descriptions of the variables. Note that there are various special values in the dataset, which require careful handling. The descriptions of the special values are provided in a separate tab in the data dictionary.

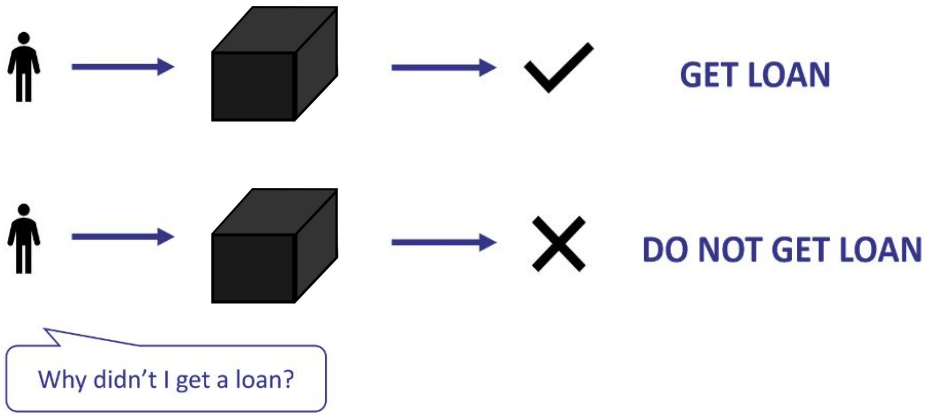
Predictions

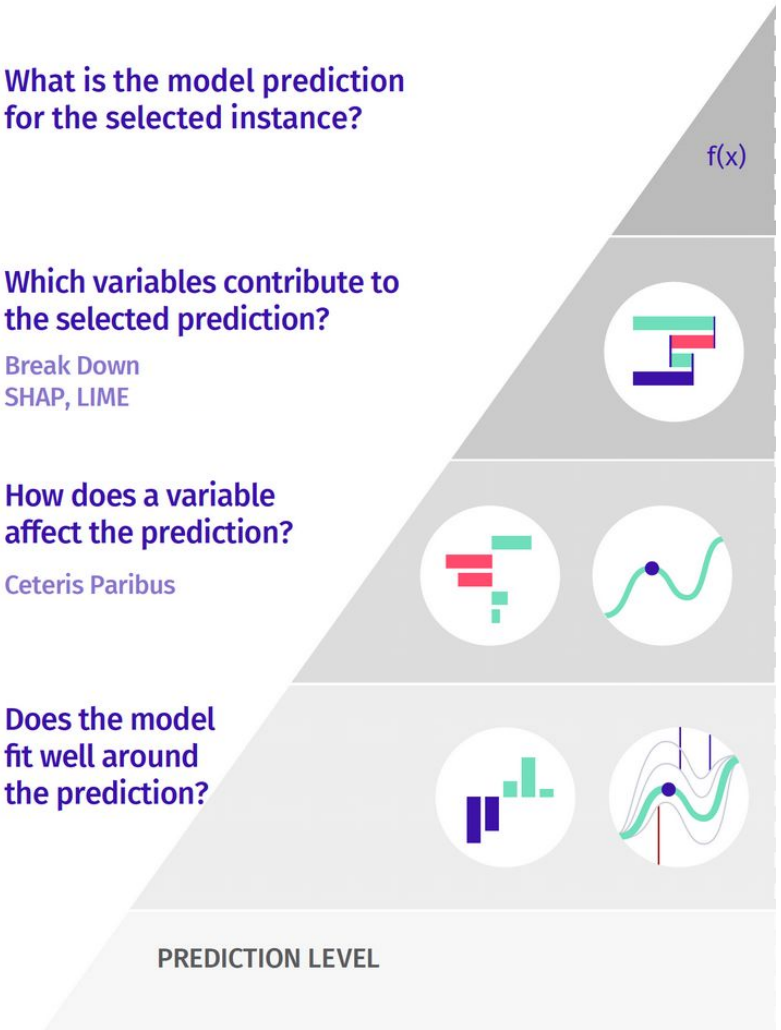
The target variable to predict is a binary variable called RiskPerformance. The value “Bad” indicates that a consumer was 90 days past due or worse at least once over a period of 24 months from when the credit account was opened. The value “Good” indicates that they have made their payments without ever being more than 90 days overdue.

- `ExternalRiskEstimate` - consolidated indicator of risk markers (equivalent of polish BIK's rate)
- `MSinceOldestTradeOpen` - number of months that have elapsed since first trade
- `MSinceMostRecentTradeOpen` - number of months that have elapsed since last opened trade
- `AverageMInFile` - average months in file
- `NumSatisfactoryTrades` - number of satisfactory trades
- `NumTrades60Ever2DerogPubRec` - number of trades which are more than 60 past due
- `NumTrades90Ever2DerogPubRec` - number of trades which are more than 90 past due
- `PercentTradesNeverDelq` - percent of trades, that were not delinquent
- `MSinceMostRecentDelq` - number of months that have elapsed since last delinquent trade
- `MaxDelq2PublicRecLast12M` - the longest delinquency period in last 12 months
- `MaxDelqEver` - the longest delinquency period
- `NumTotalTrades` - total number of trades
- `NumTradesOpeninLast12M` - number of trades opened in last 12 months
- `PercentInstallTrades` - percent of installments trades
- `MSinceMostRecentInqexcl7days` - months since last inquiry (excluding last 7 days)
- `NumInqLast6M` - number of inquiries in last 6 months
- `NumInqLast6Mexcl7days` - number of inquiries in last 6 months (excluding last 7 days)
- `NetFractionRevolvingBurden` - revolving balance divided by credit limit
- `NetFractionInstallBurden` - installment balance divided by original loan amount
- `NumRevolvingTradesWBalance` - number of revolving trades with balance
- `NumInstallTradesWBalance` - number of installment trades with balance
- `NumBank2NatlTradesWHighUtilization` - number of trades with high utilization ratio (credit utilization ratio - the amount of a credit card balance compared to the credit limit)
- `PercentTradesWBalance` - percent of trades with balance

Wyjaśnienia lokalne:

- dla pojedynczej obserwacji
- możemy ocenić co wpływa na predykcję modelu (jaki wpływ mają zmienne)
- możemy ocenić czy dla wybranej obserwacji zmiana wartości jednej ze zmiennej będzie wpływać na wartość predykcji





Klient

=

RiskPerformance	Bad
ExternalRiskEstimate	62
MSinceOldestTradeOpen	131
MSinceMostRecentTradeOpen	4
AverageMInFile	59
NumSatisfactoryTrades	16
NumTrades60Ever2DerogPubRec	4
NumTrades90Ever2DerogPubRec	2
PercentTradesNeverDelq	79
MSinceMostRecentDelq	23
MaxDelq2PublicRecLast12M	6
MaxDelqEver	5
NumTotalTrades	20
NumTradesOpeninLast12M	4
PercentInstallTrades	58
MSinceMostRecentInqexcl7days	0
NumInqLast6M	4
NumInqLast6Mexcl7days	4
NetFractionRevolvingBurden	25
NetFractionInstallBurden	91
NumRevolvingTradesWBalance	3
NumInstallTradesWBalance	4
NumBank2NatlTradesWHighUtilization	1
PercentTradesWBalance	70
NoBureau	0
NoValid_MSinceOldestTradeOpen	0
NoValid_MSinceMostRecentDelq	0
No_MSinceMostRecentDelq	0
NoValid_MSinceMostRecentInqexcl7days	0
No_MSinceMostRecentInqexcl7days	0
NoValid_NetFractionRevolvingBurden	0
NoValid_NetFractionInstallBurden	0
NoValid_NumRevolvingTradesWBalance	0
NoValid_NumInstallTradesWBalance	0
NoValid_NumBank2NatlTradesWHighUtilization	0

What is the model prediction for the selected instance?

f(x)

Which variables contribute to the selected prediction?

Break Down
SHAP, LIME

How does a variable affect the prediction?

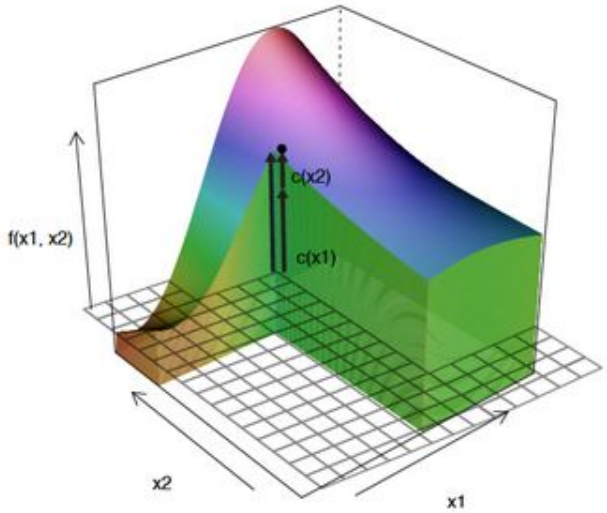
Ceteris Paribus

Does the model fit well around the prediction?

PREDICTION LEVEL

Klient

Dlaczego?



What is the model prediction for the selected instance?

f(x)

Which variables contribute to the selected prediction?

Break Down
SHAP, LIME

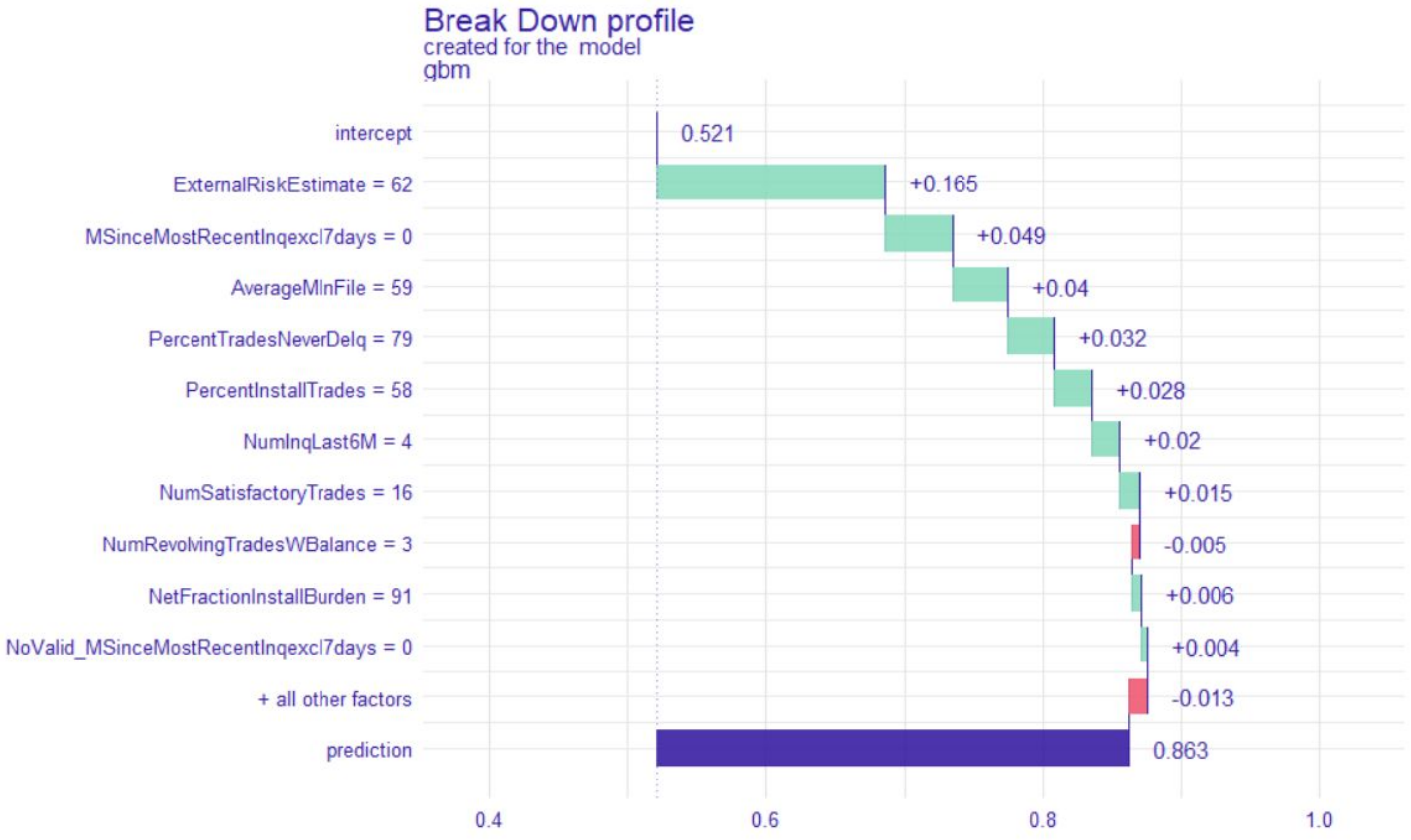
How does a variable affect the prediction?

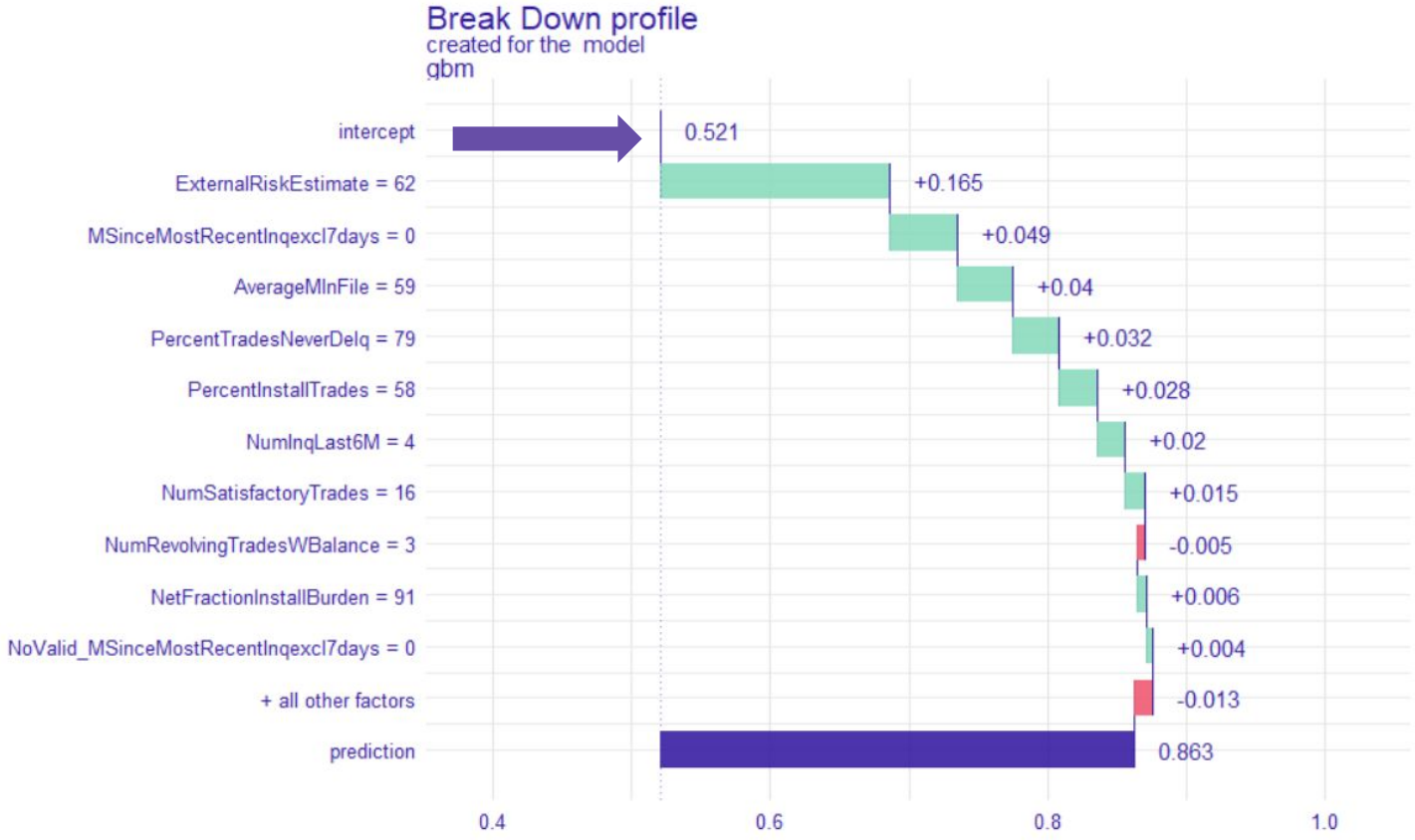
Ceteris Paribus

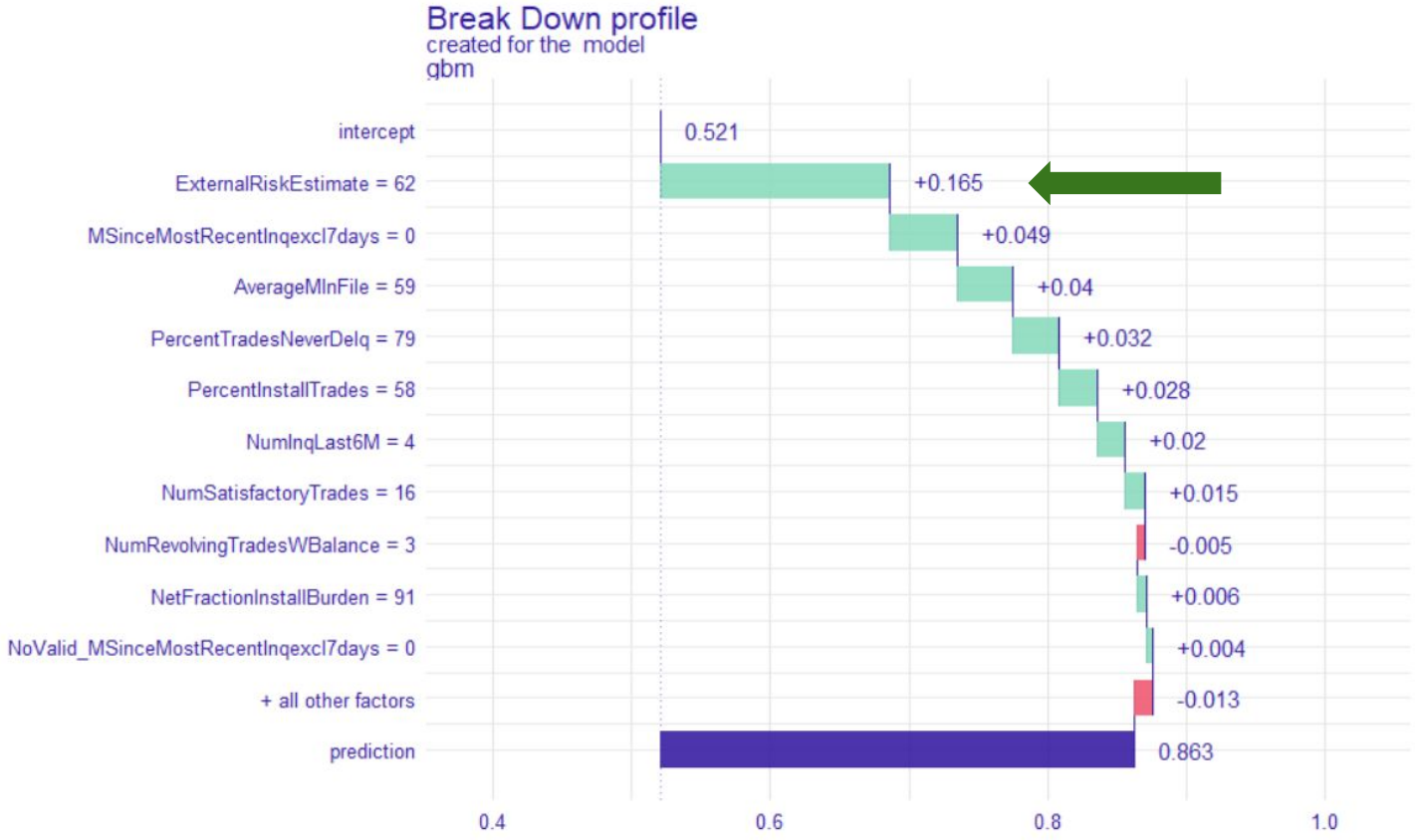
Does the model fit well around the prediction?

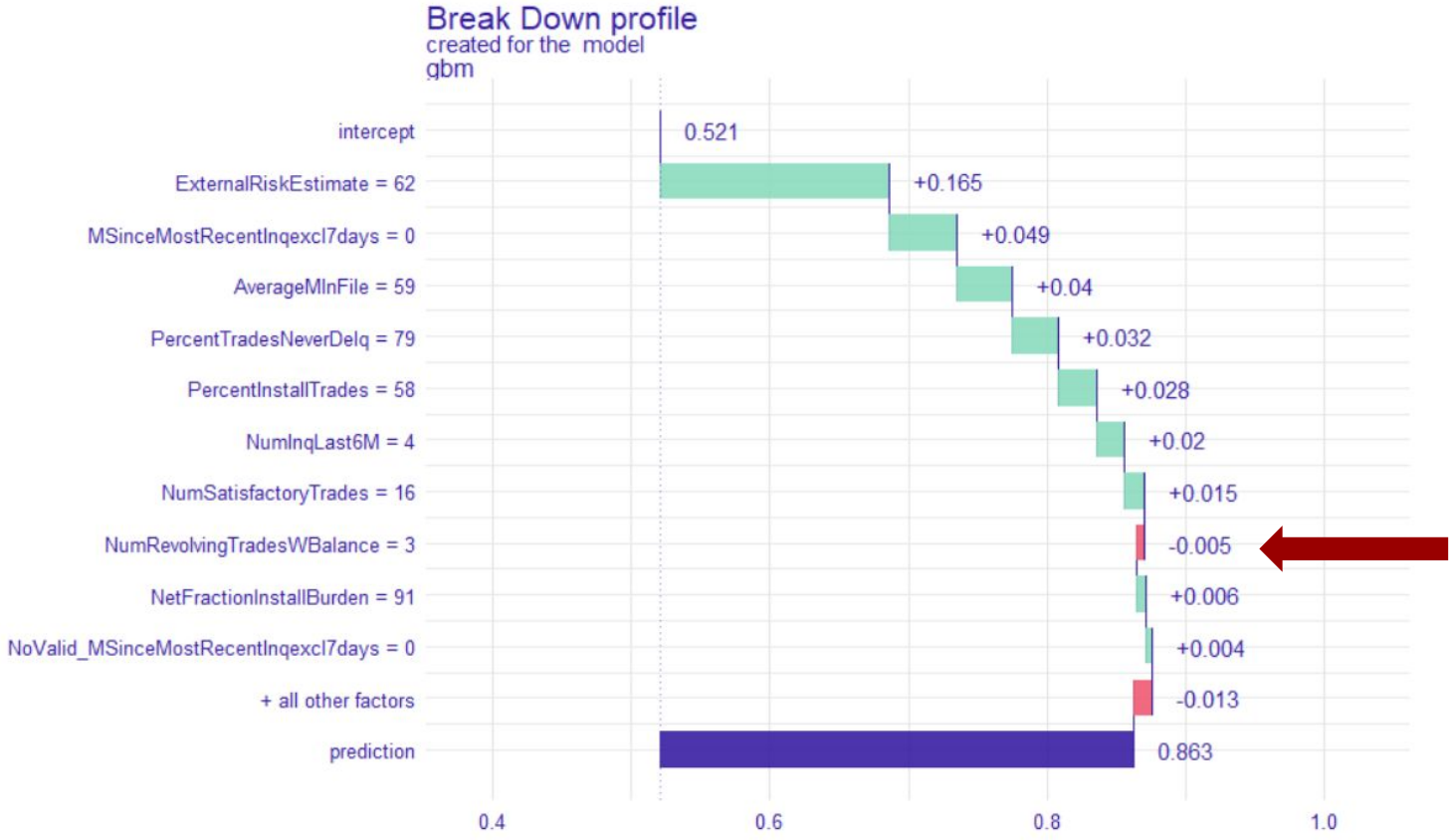
PREDICTION LEVEL

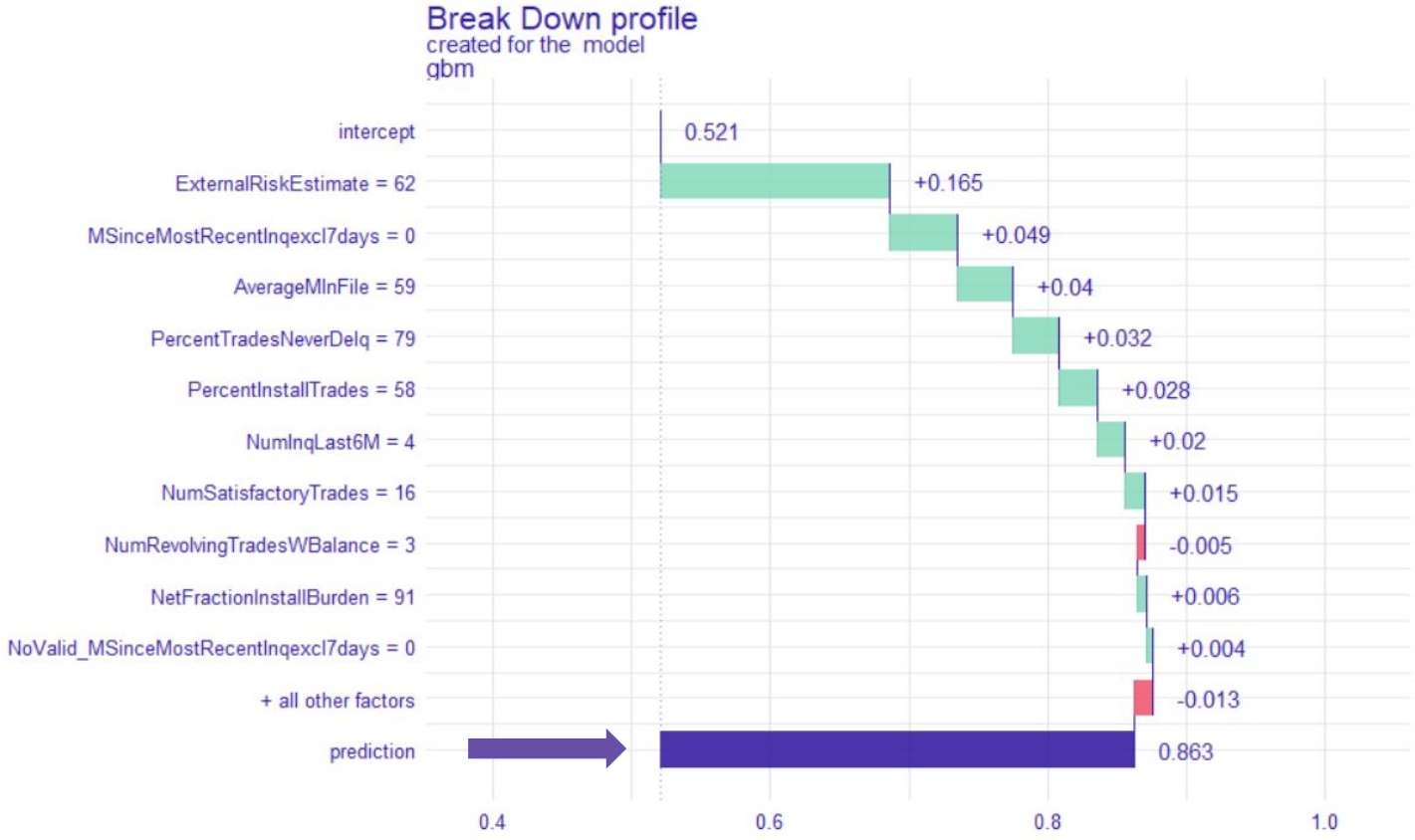
Metoda Break Down




Break Down profile
created for the model
gbm
intercept
ExternalRiskEstimate = 62
MSinceMostRecentInqexcl7days = 0
AverageMInFile = 59
PercentTradesNeverDelq = 79
PercentInstallTrades = 58
NumInqLast6M = 4
NumSatisfactoryTrades = 16
NumRevolvingTradesWBalance = 3
NetFractionInstallBurden = 91
NoValid_MSinceMostRecentInqexcl7days = 0
+ all other factors
prediction
0.521
+0.165
+0.049
+0.04
+0.032
+0.028
+0.02
+0.015
-0.005
+0.006
+0.004
-0.013
0.863
0.4
0.6
0.8
1.0

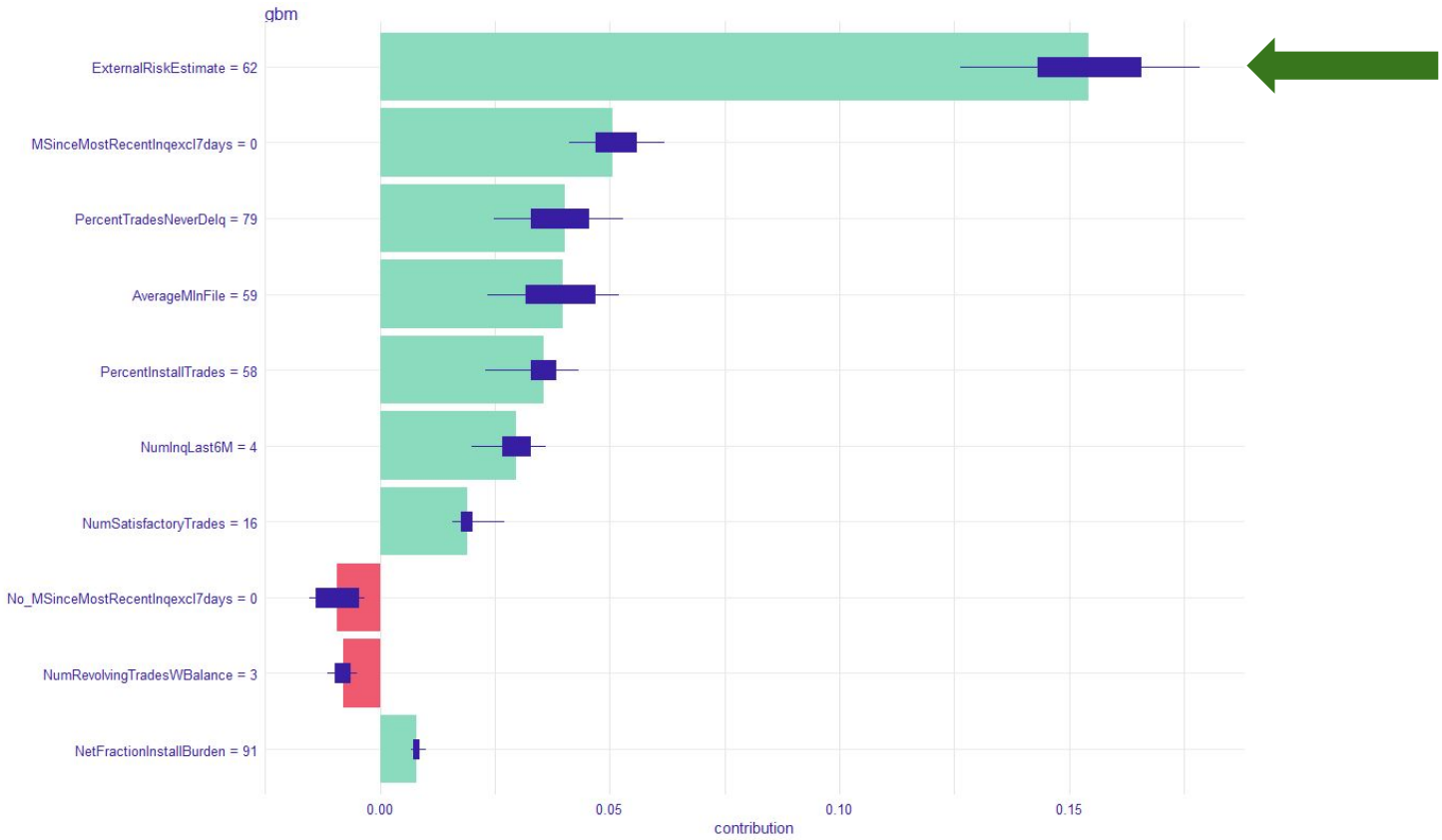

Break Down profile
created for the model
gbm
intercept
ExternalRiskEstimate = 62
MSinceMostRecentInqexcl7days = 0
AverageMInFile = 59
PercentTradesNeverDelq = 79
PercentInstallTrades = 58
NumInqLast6M = 4
NumSatisfactoryTrades = 16
NumRevolvingTradesWBalance = 3
NetFractionInstallBurden = 91
NoValid_MSinceMostRecentInqexcl7days = 0
+ all other factors
prediction
0.521
+0.165
+0.049
+0.04
+0.032
+0.028
+0.02
+0.015
-0.005
+0.006
+0.004
-0.013
0.863
0.4
0.6
0.8
1.0

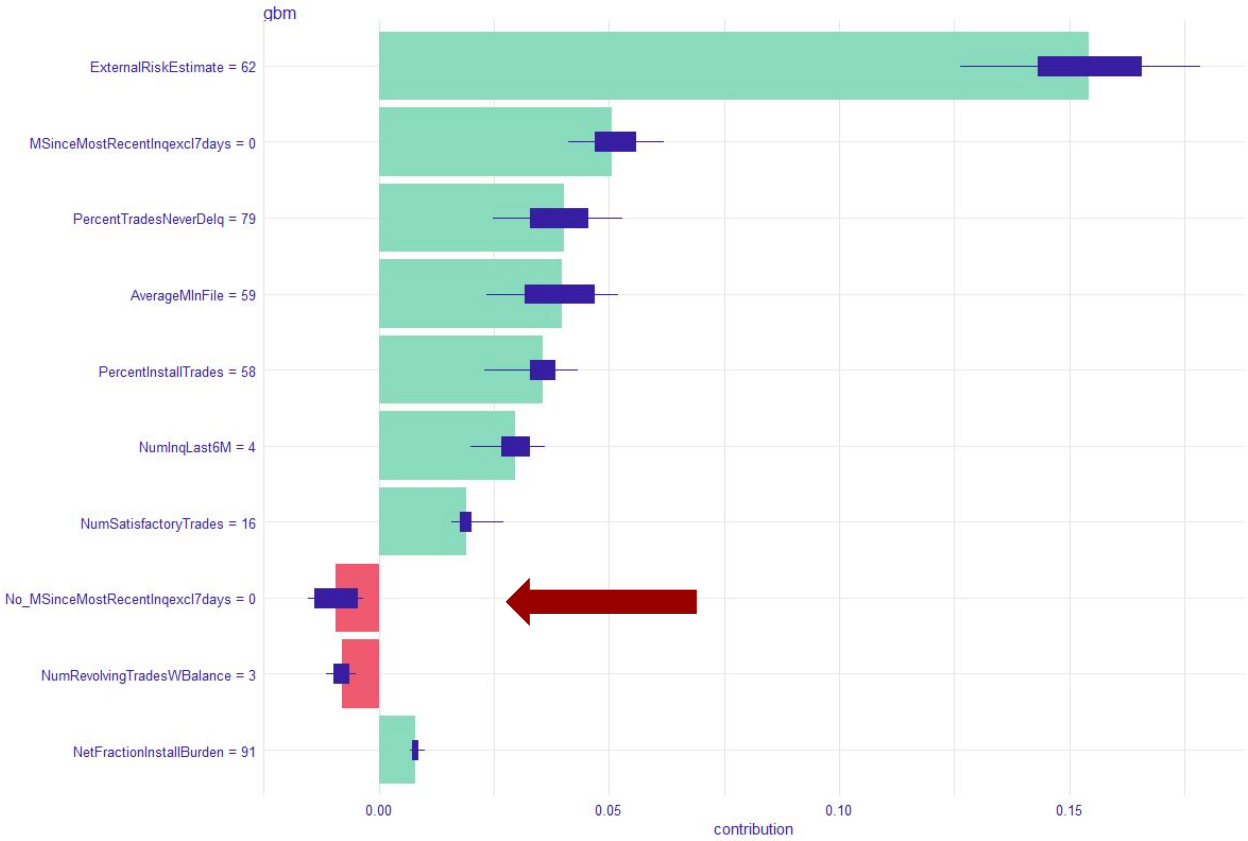

Break Down profile
created for the model
gbm
intercept
ExternalRiskEstimate = 62
MSinceMostRecentInqexcl7days = 0
AverageMInFile = 59
PercentTradesNeverDelq = 79
PercentInstallTrades = 58
NumInqLast6M = 4
NumSatisfactoryTrades = 16
NumRevolvingTradesWBalance = 3
NetFractionInstallBurden = 91
NoValid_MSinceMostRecentInqexcl7days = 0
+ all other factors
prediction
0.521
+0.165
+0.049
+0.04
+0.032
+0.028
+0.02
+0.015
-0.005
+0.006
+0.004
-0.013
0.863
0.4
0.6
0.8
1.0

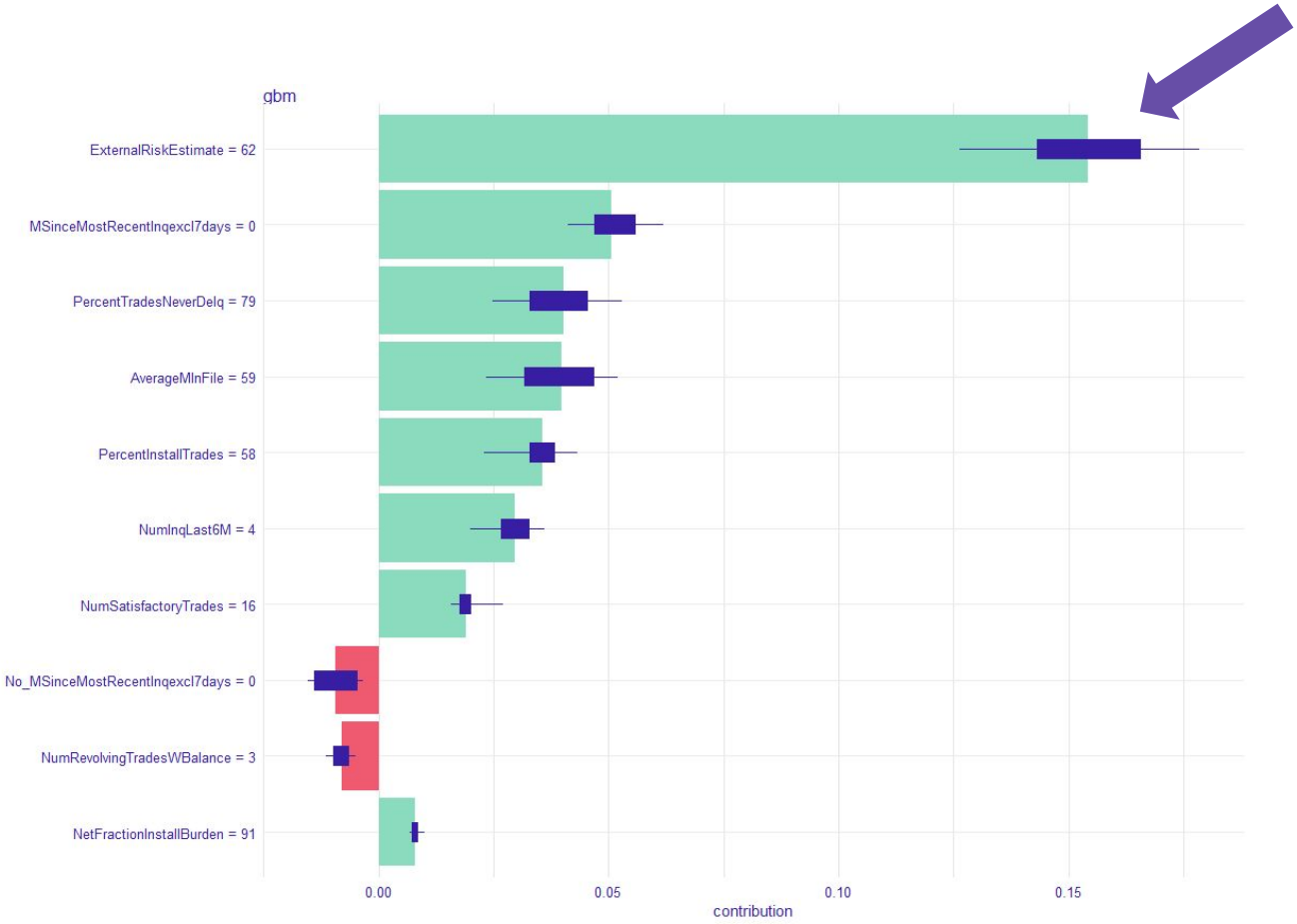

Break Down profile
created for the model
gbm
intercept
ExternalRiskEstimate = 62
MSinceMostRecentInqexcl7days = 0
AverageMInFile = 59
PercentTradesNeverDelq = 79
PercentInstallTrades = 58
NumInqLast6M = 4
NumSatisfactoryTrades = 16
NumRevolvingTradesWBalance = 3
NetFractionInstallBurden = 91
NoValid_MSinceMostRecentInqexcl7days = 0
+ all other factors
prediction
0.521
+0.165
+0.049
+0.04
+0.032
+0.028
+0.02
+0.015
-0.005
+0.006
+0.004
-0.013
0.863
0.4
0.6
0.8
1.0

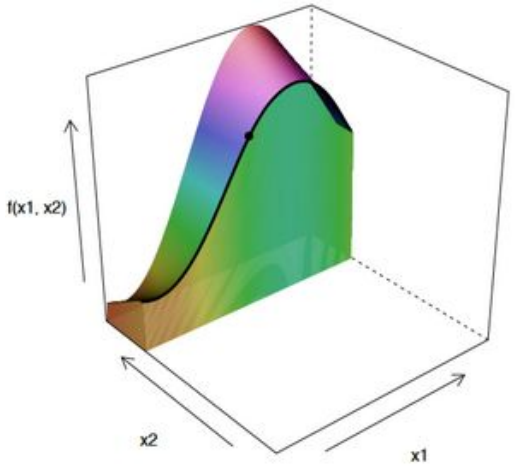
Metoda SHAP


gbm
ExternalRiskEstimate = 62
MSinceMostRecentInqexcl7days = 0
PercentTradesNeverDelq = 79
AverageMInFile = 59
PercentInstallTrades = 58
NumInqLast6M = 4
NumSatisfactoryTrades = 16
No_MSinceMostRecentInqexcl7days = 0
NumRevolvingTradesWBalance = 3
NetFractionInstallBurden = 91
0.00
0.05
0.10
0.15
contribution


gbm
ExternalRiskEstimate = 62
MSinceMostRecentInqexcl7days = 0
PercentTradesNeverDelq = 79
AverageMInFile = 59
PercentInstallTrades = 58
NumInqLast6M = 4
NumSatisfactoryTrades = 16
No_MSinceMostRecentInqexcl7days = 0
NumRevolvingTradesWBalance = 3
NetFractionInstallBurden = 91
0.00
0.05
0.10
0.15
contribution


gbm
ExternalRiskEstimate = 62
MSinceMostRecentInqexcl7days = 0
PercentTradesNeverDelq = 79
AverageMInFile = 59
PercentInstallTrades = 58
NumInqLast6M = 4
NumSatisfactoryTrades = 16
No_MSinceMostRecentInqexcl7days = 0
NumRevolvingTradesWBalance = 3
NetFractionInstallBurden = 91
0.00
0.05
0.10
0.15
contribution


gbm
ExternalRiskEstimate = 62
MSinceMostRecentInqexcl7days = 0
PercentTradesNeverDelq = 79
AverageMInFile = 59
PercentInstallTrades = 58
NumInqLast6M = 4
NumSatisfactoryTrades = 16
No_MSinceMostRecentInqexcl7days = 0
NumRevolvingTradesWBalance = 3
NetFractionInstallBurden = 91
0.00
0.05
0.10
0.15
contribution



What is the model prediction for the selected instance?

f(x)

Which variables contribute to the selected prediction?

Break Down
SHAP, LIME

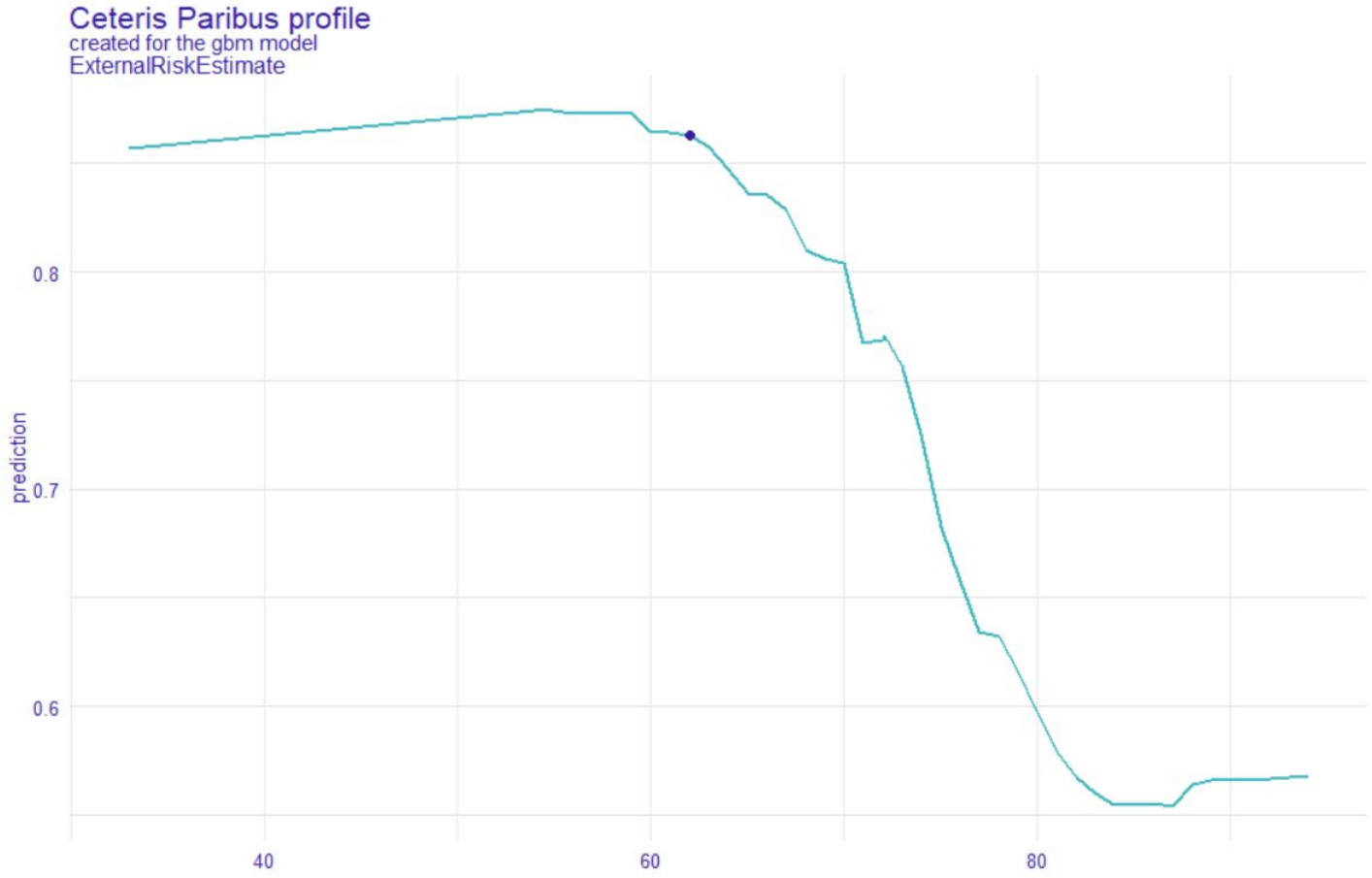
How does a variable affect the prediction?

Ceteris Paribus

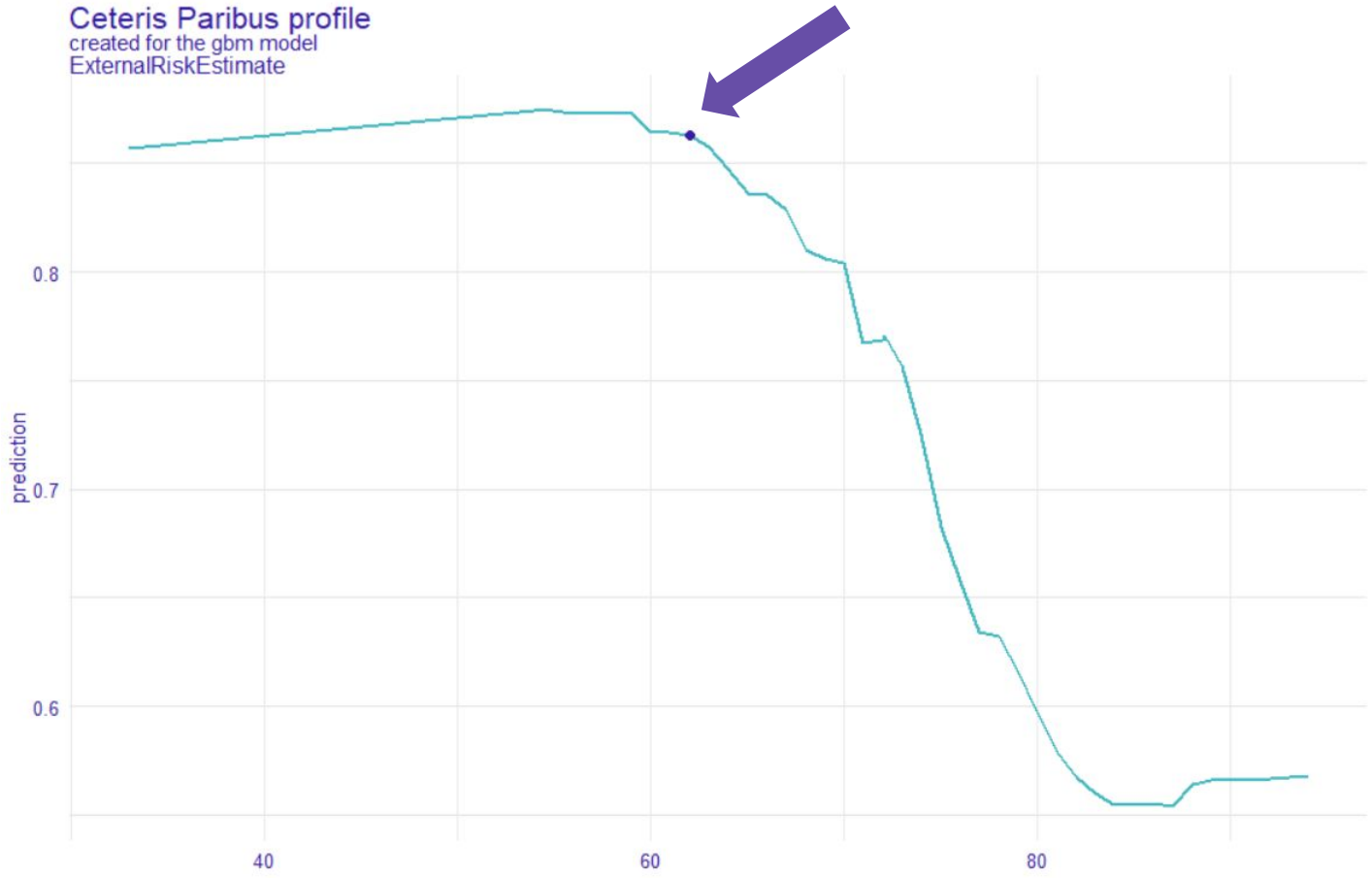
Does the model fit well around the prediction?

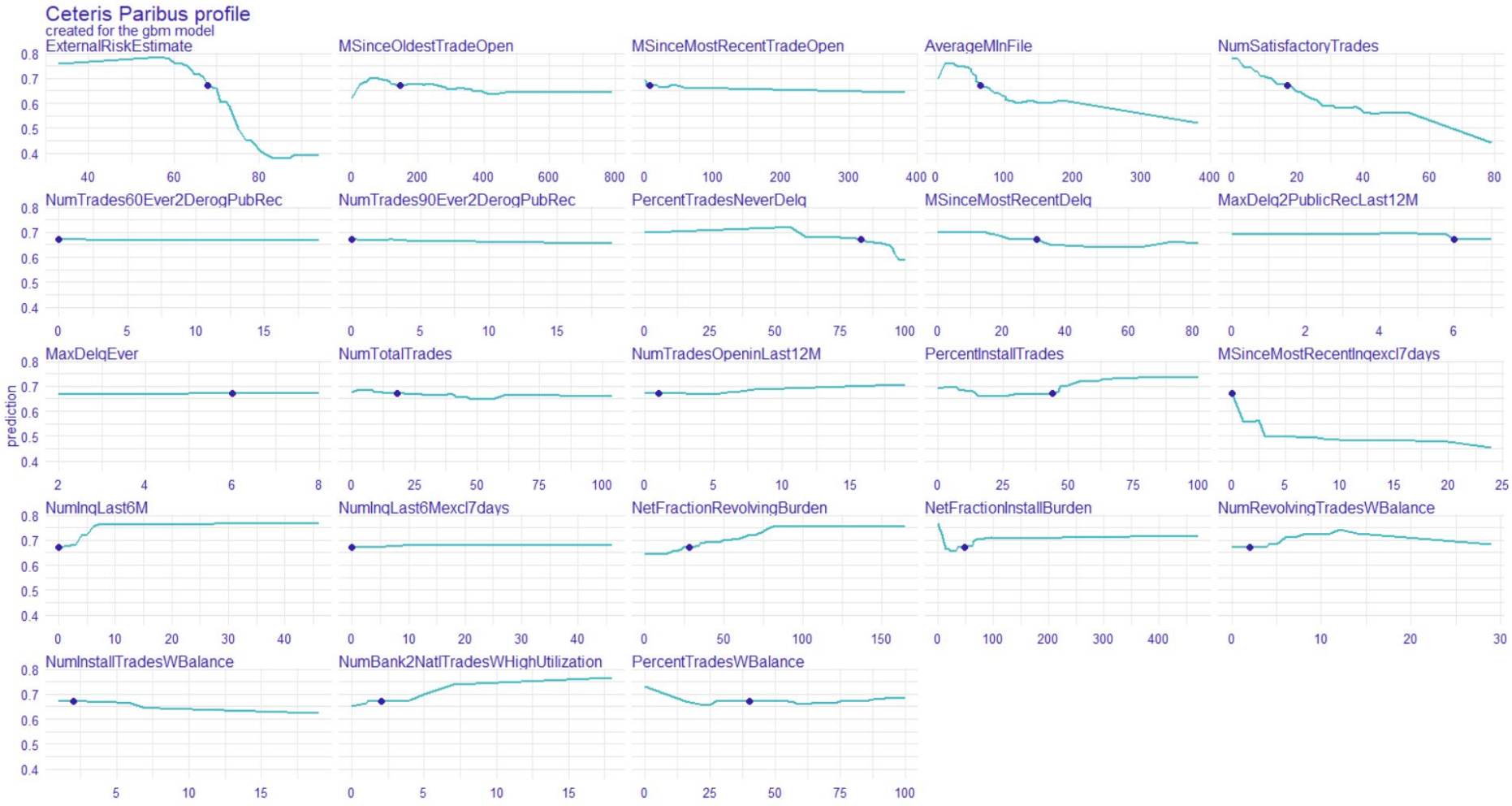
PREDICTION LEVEL

Metoda Ceteris Paribus



Metoda Ceteris Paribus




Ceteris Paribus profile
created for the gbm model
ExternalRiskEstimate
MSinceOldestTradeOpen
MSinceMostRecentTradeOpen
AverageMInFile
NumSatisfactoryTrades
NumTrades60Ever2DerogPubRec
NumTrades90Ever2DerogPubRec
PercentTradesNeverDelq
MSinceMostRecentDelq
MaxDelq2PublicRecLast12M
MaxDelqEver
NumTotalTrades
NumTradesOpeninLast12M
PercentInstallTrades
MSinceMostRecentInqexcl7days
NumInqLast6M
NumInqLast6Mexcl7days
NetFractionRevolvingBurden
NetFractionInstallBurden
NumRevolvingTradesWBalance
NumInstallTradesWBalance
NumBank2NatlTradesWHighUtilization
PercentTradesWBalance
prediction

What is the model prediction for the selected instance?

f(x)

Which variables contribute to the selected prediction?

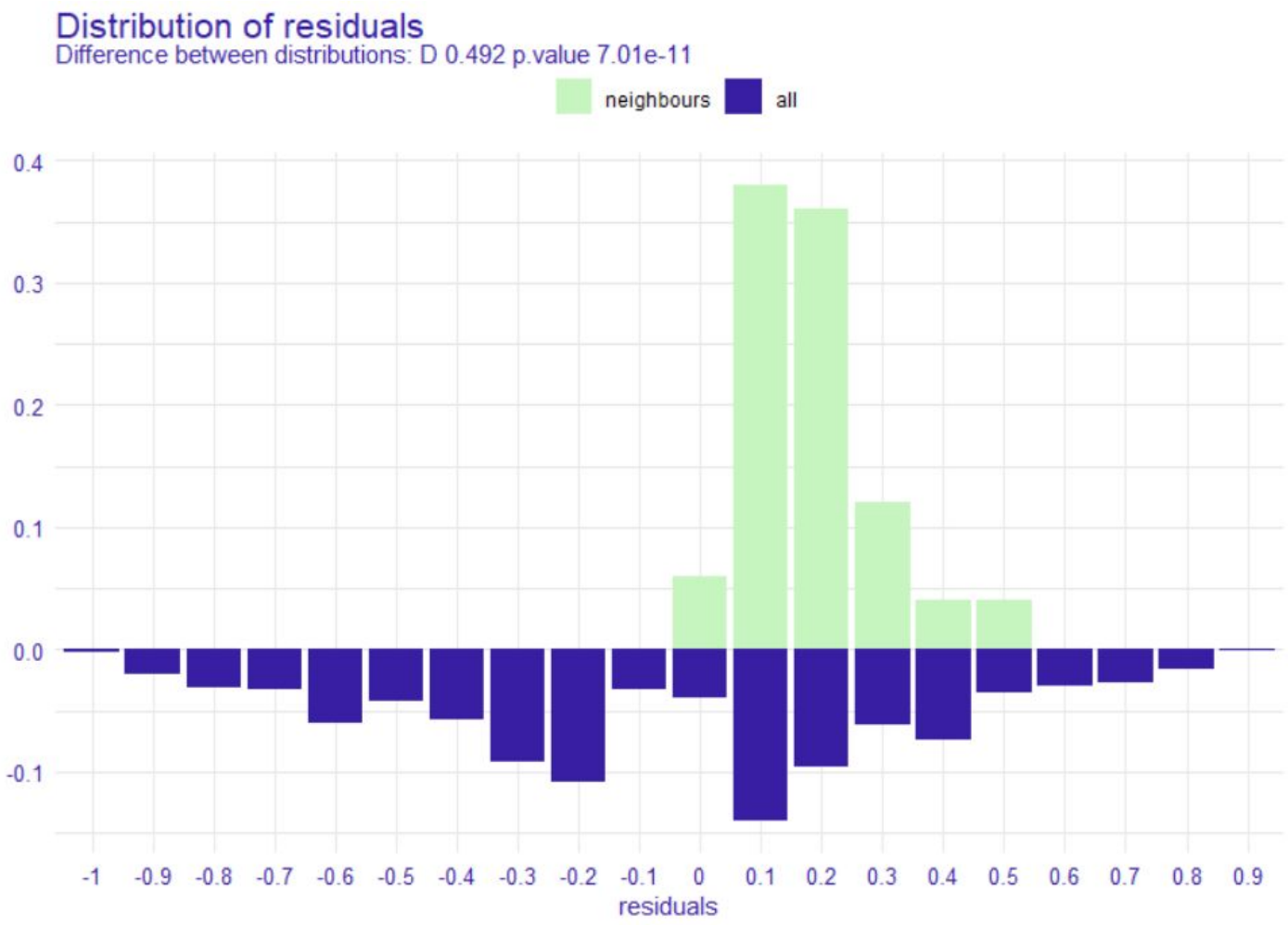
Break Down
SHAP, LIME

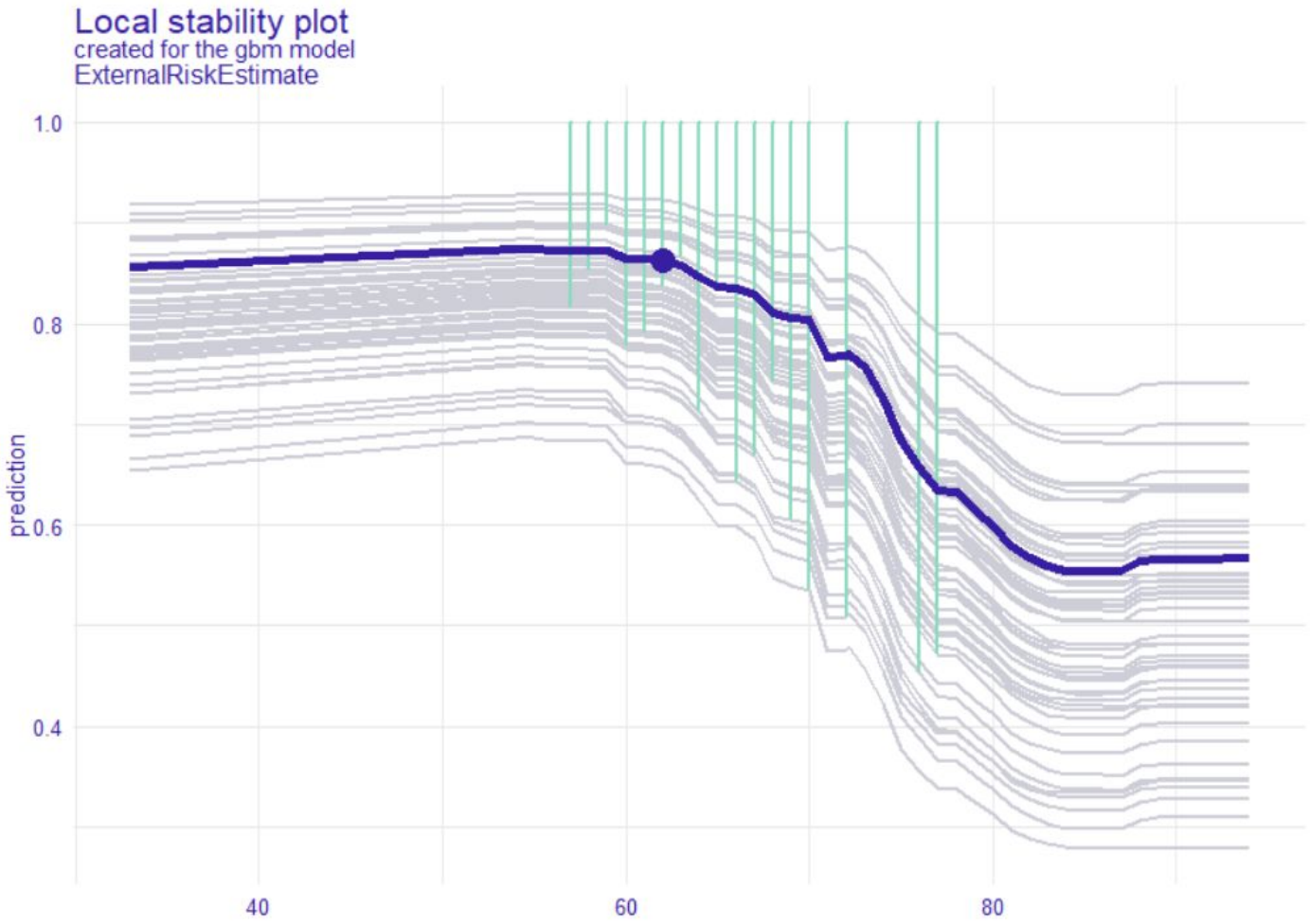
How does a variable affect the prediction?

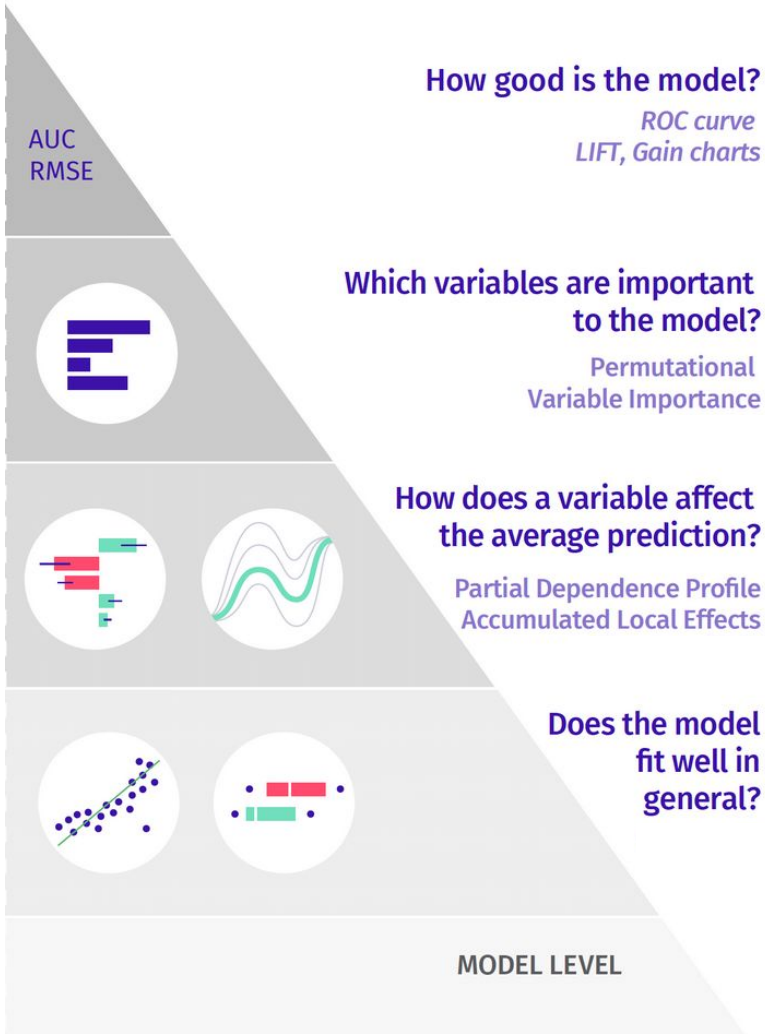
Ceteris Paribus

Does the model fit well around the prediction?

PREDICTION LEVEL


Distribution of residuals
Difference between distributions: D 0.492 p.value 7.01e-11
neighbours
all
0.4
0.3
0.2
0.1
0.0
-0.1
-1
-0.9
-0.8
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
residuals

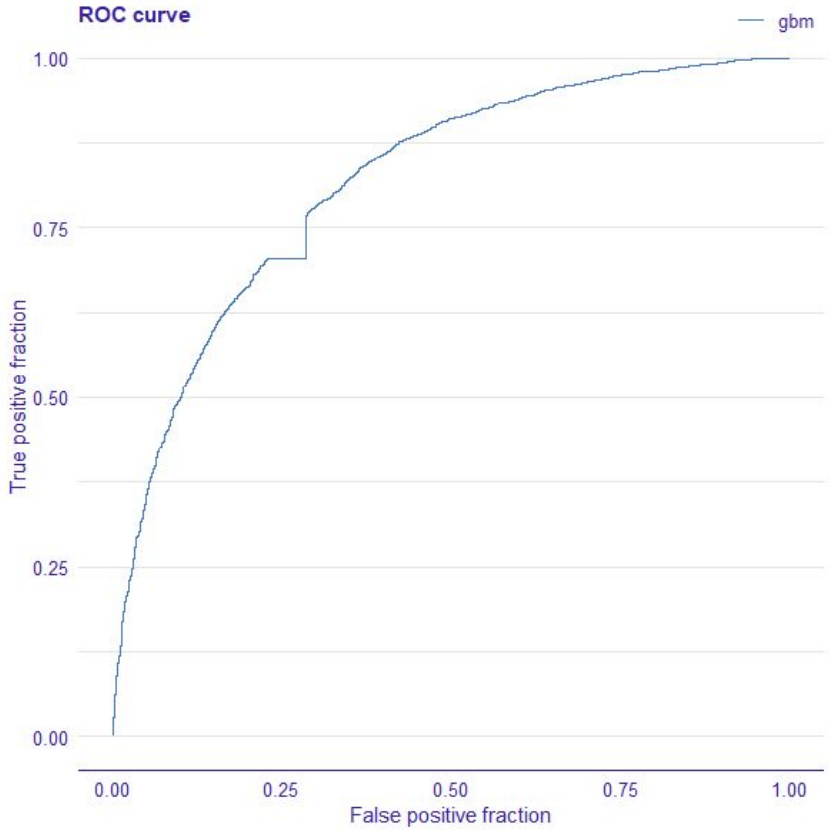

Local stability plot
created for the gbm model
ExternalRiskEstimate
1.0
0.8
0.6
0.4
prediction
40
60
80



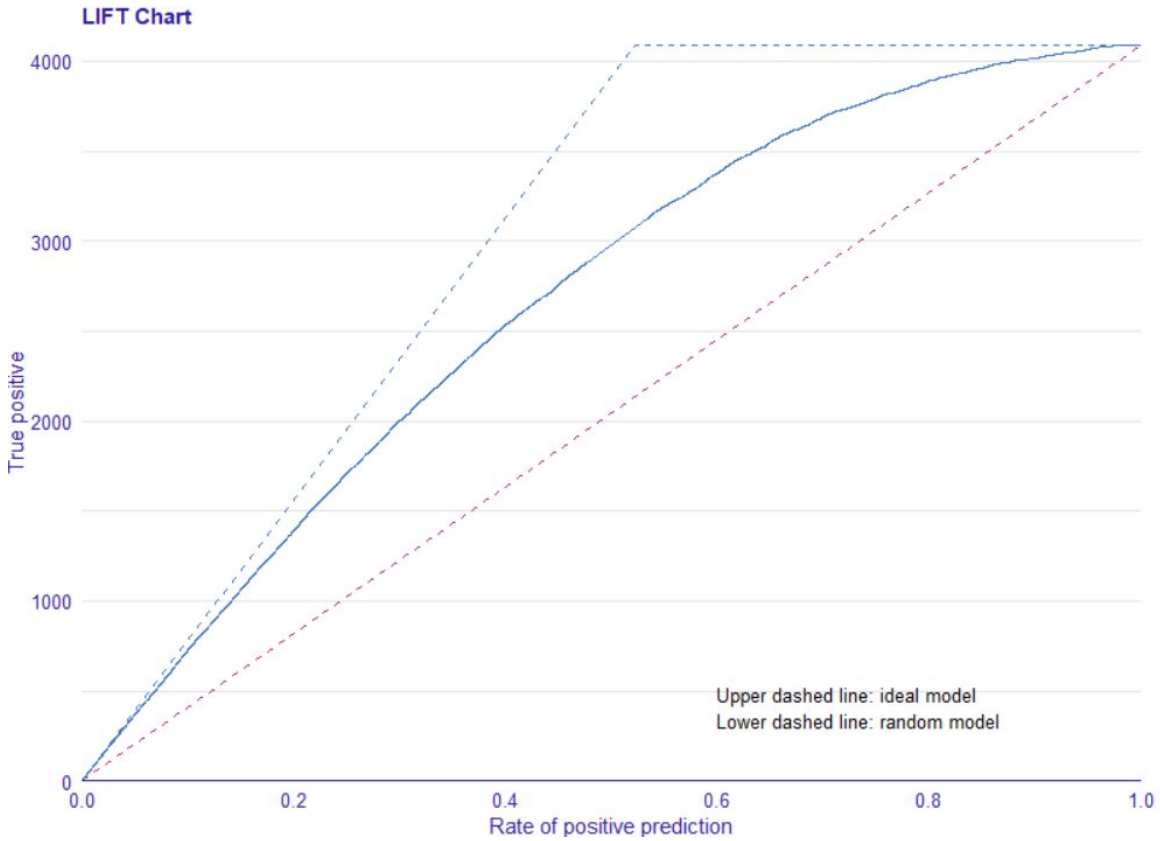
Wyjaśnienia globalne:

- dla zbioru danych
- możemy ocenić jakie zmienne mają największy wpływ na predykcję modelu

Krzywa ROC


ROC curve
gbm
1.00
0.75
0.50
0.25
0.00
True positive fraction
0.00
0.25
0.50
0.75
1.00
False positive fraction

Krzywa LIFT


LIFT Chart
4000
3000
2000
1000
0
True positive
0.0
0.2
0.4
0.6
0.8
1.0
Rate of positive prediction
Upper dashed line: ideal model
Lower dashed line: random model

AUC
RMSE

How good is the model?

ROC curve
LIFT, Gain charts

Which variables are important to the model?

Permutational
Variable Importance

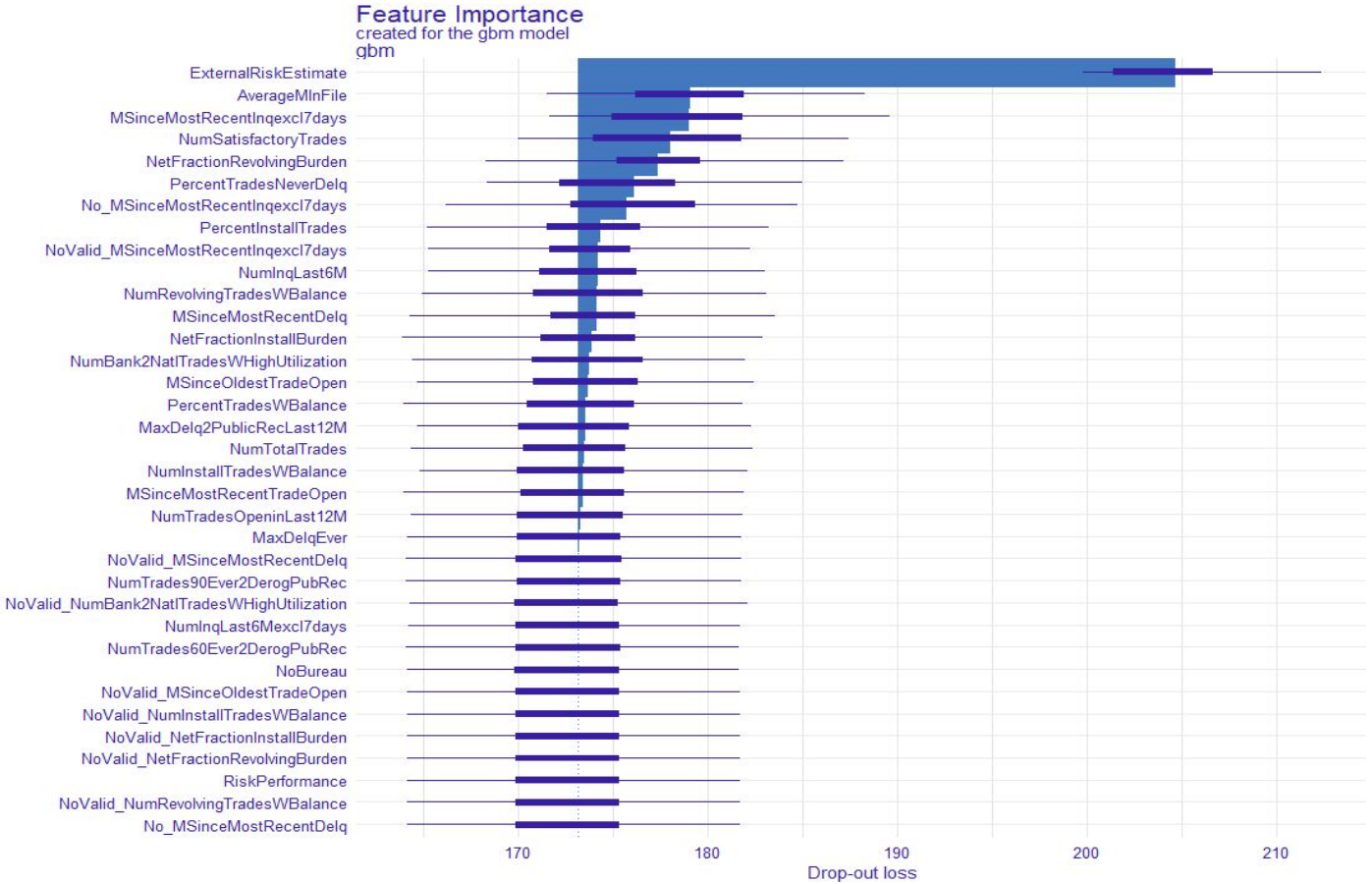
How does a variable affect the average prediction?

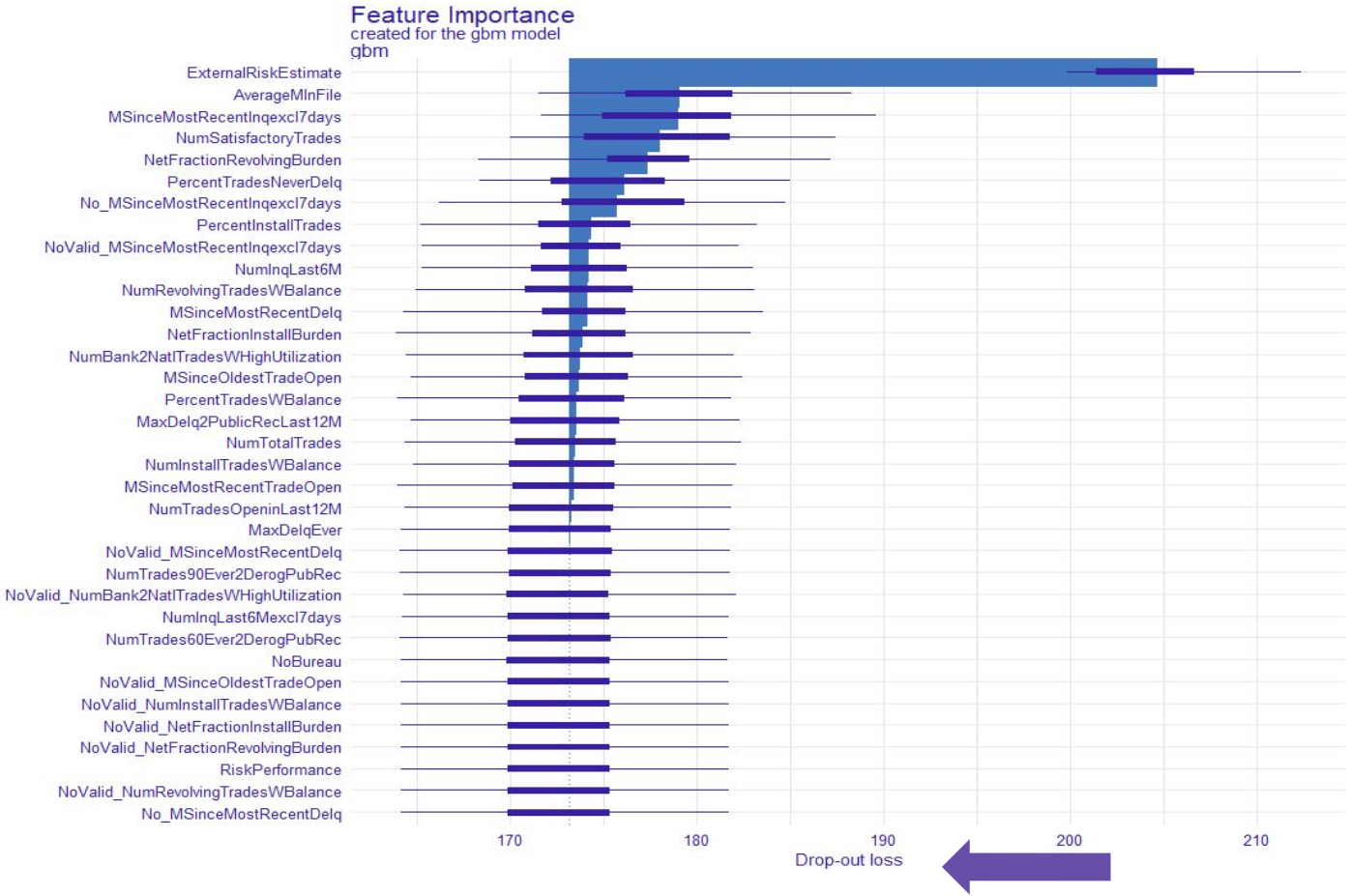
Partial Dependence Profile
Accumulated Local Effects

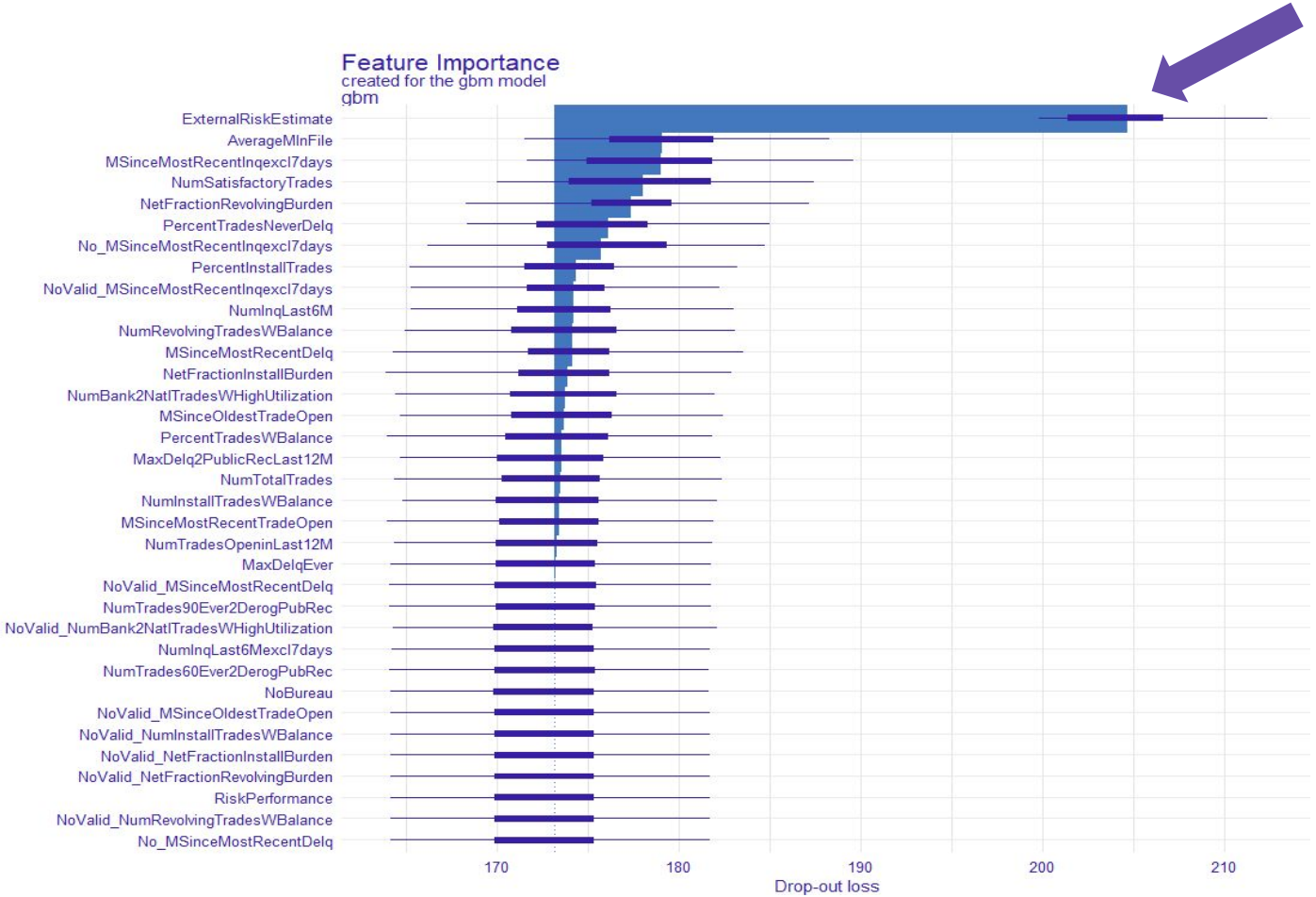
Does the model fit well in general?

MODEL LEVEL

Permutacyjna ważność zmiennych




Feature Importance
created for the gbm model
gbm
ExternalRiskEstimate
AverageMInFile
MSinceMostRecentInqexcl7days
NumSatisfactoryTrades
NetFractionRevolvingBurden
PercentTradesNeverDelq
No_MSinceMostRecentInqexcl7days
PercentInstallTrades
NoValid_MSinceMostRecentInqexcl7days
NumInqLast6M
NumRevolvingTradesWBalance
MSinceMostRecentDelq
NetFractionInstallBurden
NumBank2NatlTradesWHighUtilization
MSinceOldestTradeOpen
PercentTradesWBalance
MaxDelq2PublicRecLast12M
NumTotalTrades
NumInstallTradesWBalance
MSinceMostRecentTradeOpen
NumTradesOpeninLast12M
MaxDelqEver
NoValid_MSinceMostRecentDelq
NumTrades90Ever2DerogPubRec
NoValid_NumBank2NatlTradesWHighUtilization
NumInqLast6Mexcl7days
NumTrades60Ever2DerogPubRec
NoBureau
NoValid_MSinceOldestTradeOpen
NoValid_NumInstallTradesWBalance
NoValid_NetFractionInstallBurden
NoValid_NetFractionRevolvingBurden
RiskPerformance
NoValid_NumRevolvingTradesWBalance
No_MSinceMostRecentDelq
170
180
190
200
210
Drop-out loss


Feature Importance
created for the gbm model
gbm
ExternalRiskEstimate
AverageMInFile
MSinceMostRecentInqexcl7days
NumSatisfactoryTrades
NetFractionRevolvingBurden
PercentTradesNeverDelq
No_MSinceMostRecentInqexcl7days
PercentInstallTrades
NoValid_MSinceMostRecentInqexcl7days
NumInqLast6M
NumRevolvingTradesWBalance
MSinceMostRecentDelq
NetFractionInstallBurden
NumBank2NatlTradesWHighUtilization
MSinceOldestTradeOpen
PercentTradesWBalance
MaxDelq2PublicRecLast12M
NumTotalTrades
NumInstallTradesWBalance
MSinceMostRecentTradeOpen
NumTradesOpeninLast12M
MaxDelqEver
NoValid_MSinceMostRecentDelq
NumTrades90Ever2DerogPubRec
NoValid_NumBank2NatlTradesWHighUtilization
NumInqLast6Mexcl7days
NumTrades60Ever2DerogPubRec
NoBureau
NoValid_MSinceOldestTradeOpen
NoValid_NumInstallTradesWBalance
NoValid_NetFractionInstallBurden
NoValid_NetFractionRevolvingBurden
RiskPerformance
NoValid_NumRevolvingTradesWBalance
No_MSinceMostRecentDelq
170
180
190
200
210
Drop-out loss

AUC
RMSE

How good is the model?

ROC curve
LIFT, Gain charts

Which variables are important to the model?

Permutational
Variable Importance

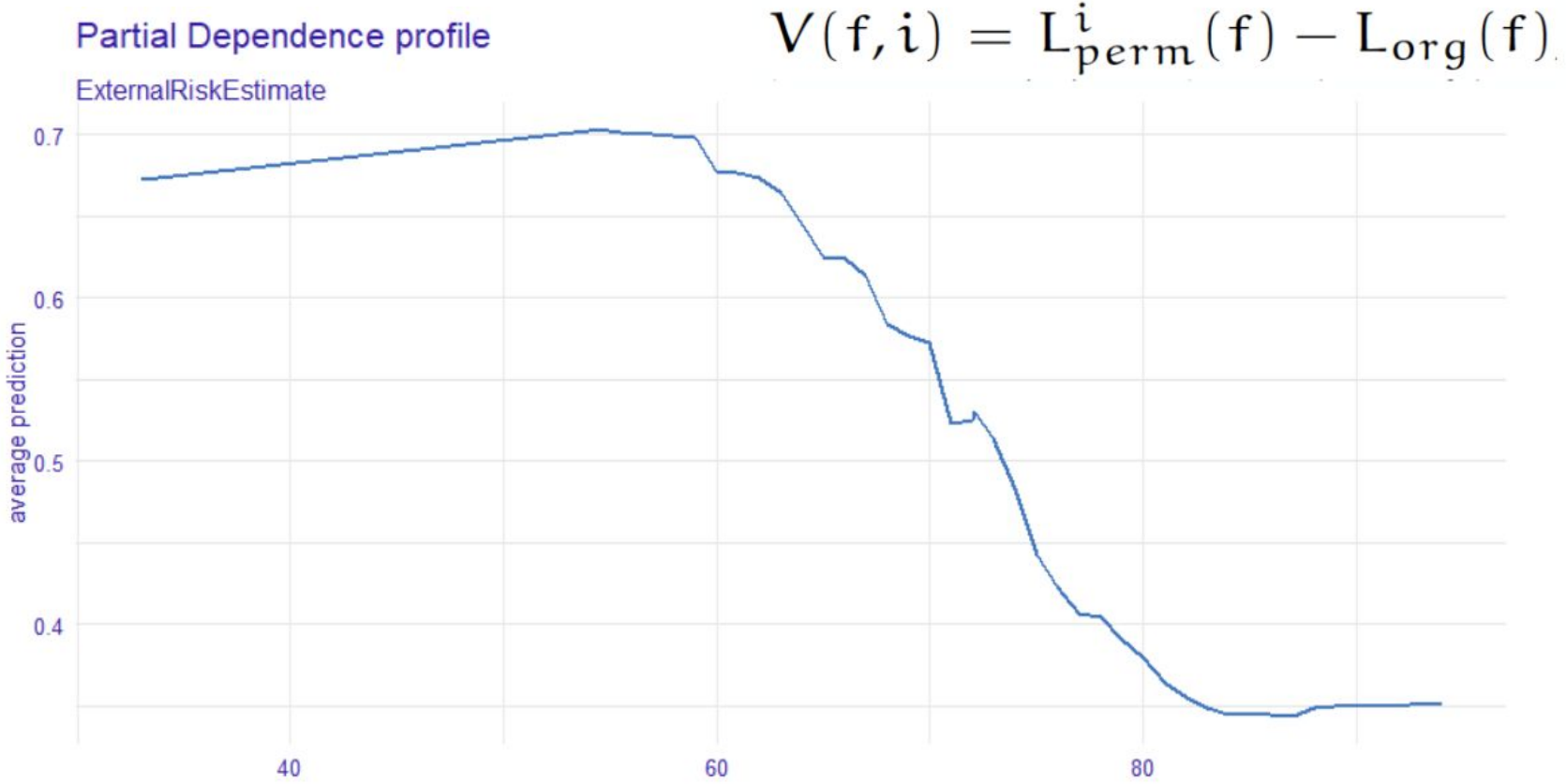
How does a variable affect the average prediction?

Partial Dependence Profile
Accumulated Local Effects

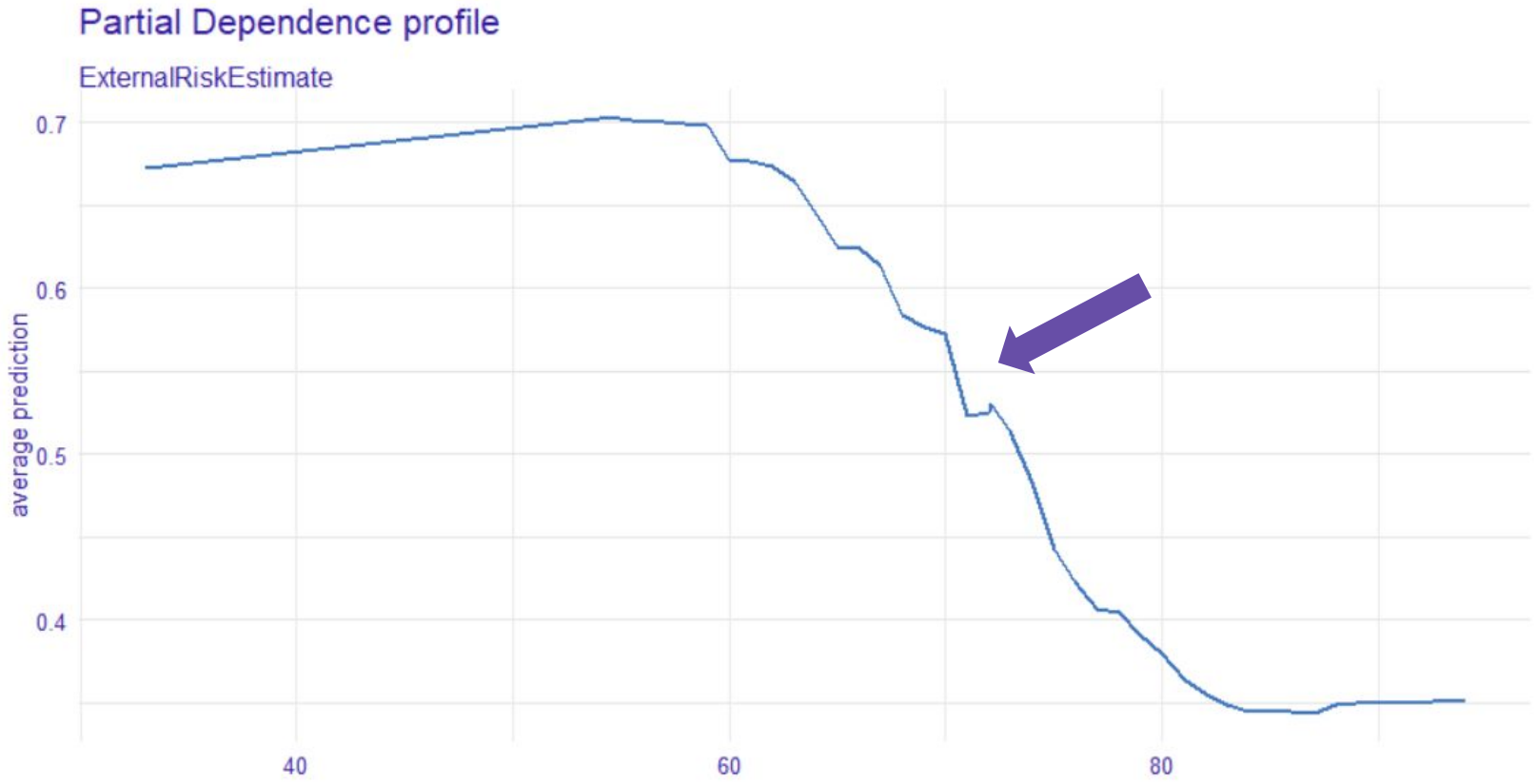
Does the model fit well in general?

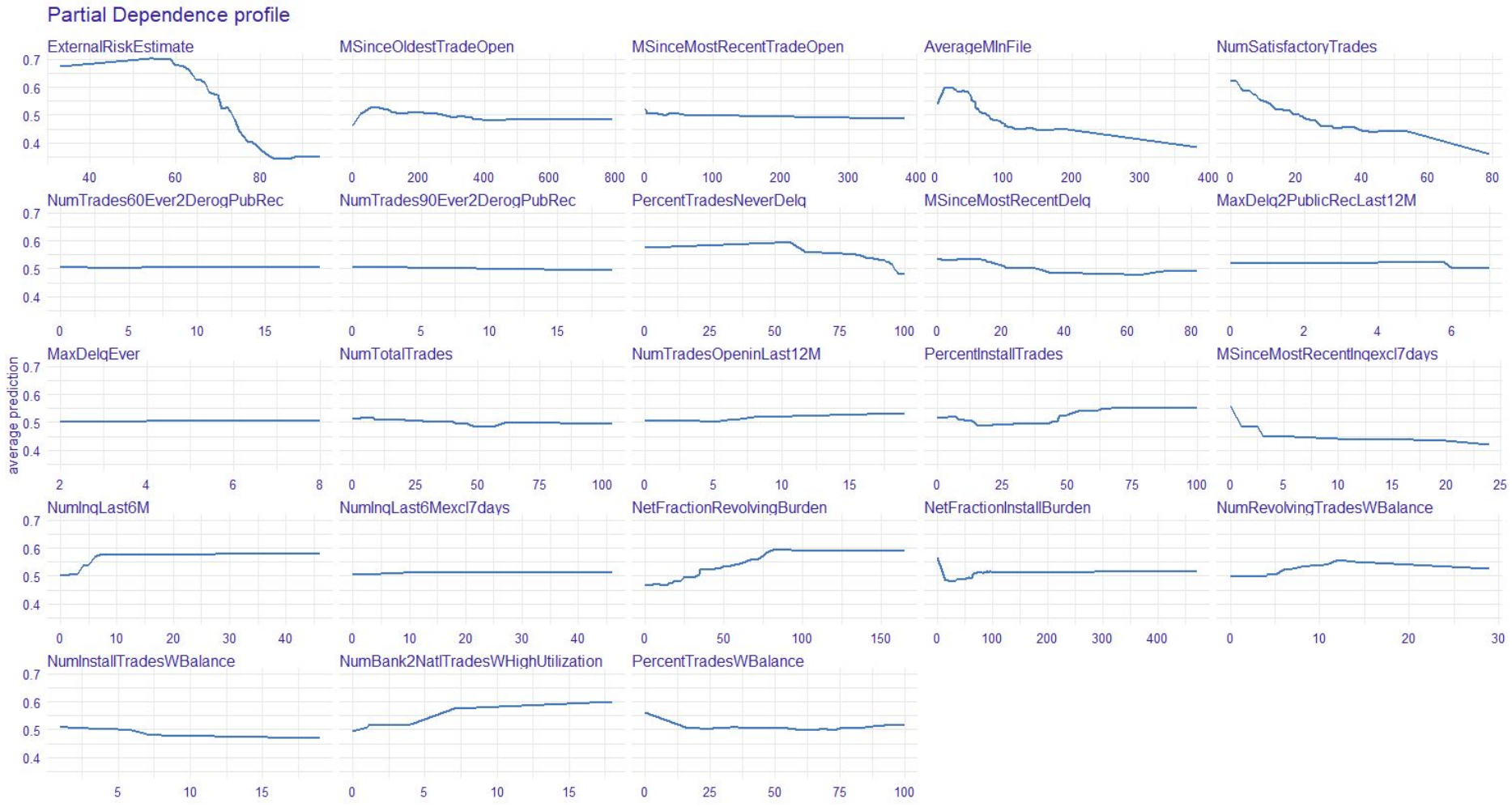
MODEL LEVEL

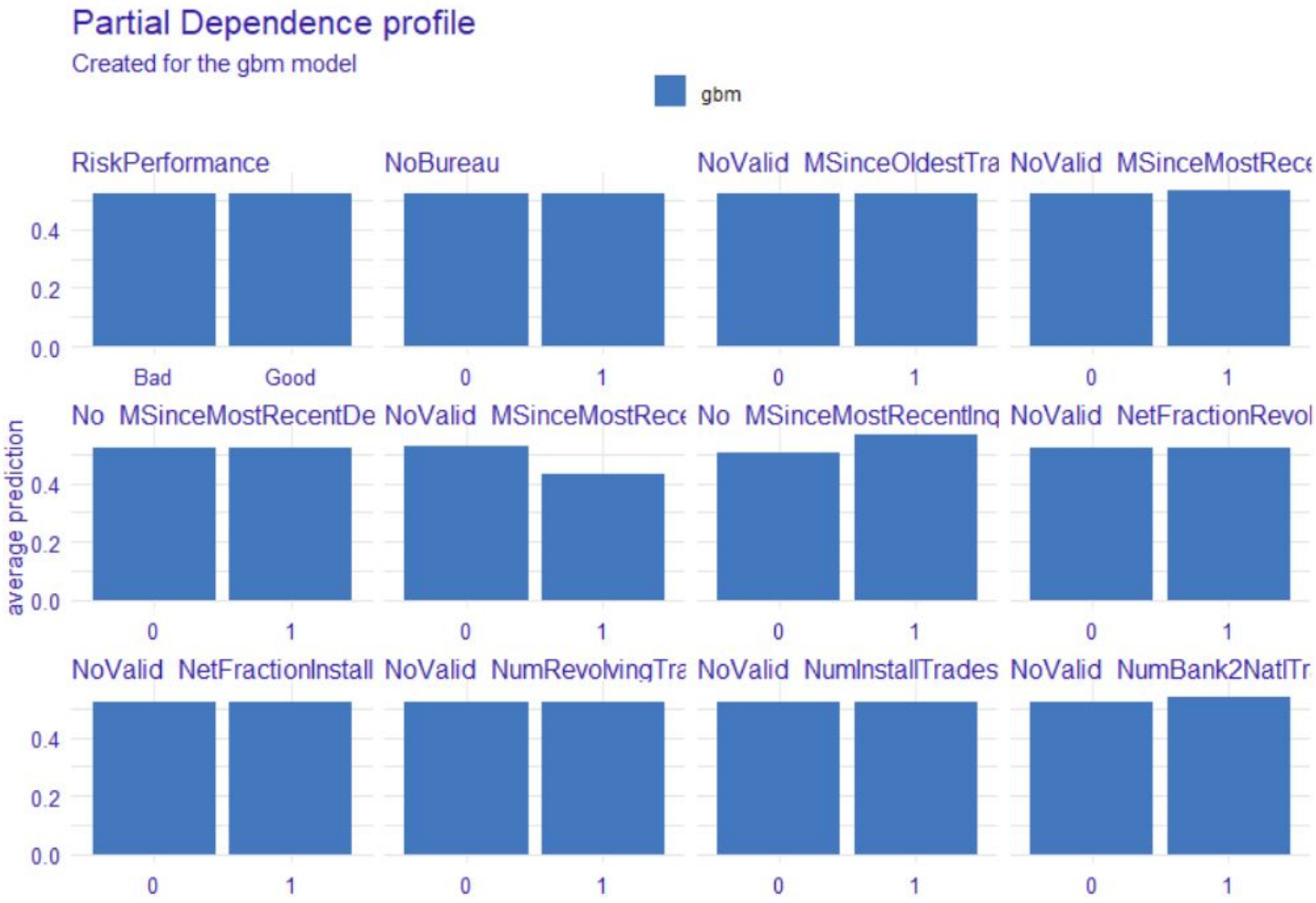
Profil Partial Dependence



$$V(f,i) = L^{i}_{perm}(f) - L_{org}(f)$$


Partial Dependence profile
ExternalRiskEstimate
average prediction
0.7
0.6
0.5
0.4
40
60
80


Partial Dependence profile
ExternalRiskEstimate
MSinceOldestTradeOpen
MSinceMostRecentTradeOpen
AverageMInFile
NumSatisfactoryTrades
NumTrades60Ever2DerogPubRec
NumTrades90Ever2DerogPubRec
PercentTradesNeverDelq
MSinceMostRecentDelq
MaxDelq2PublicRecLast12M
MaxDelqEver
NumTotalTrades
NumTradesOpeninLast12M
PercentInstallTrades
MSinceMostRecentInqexcl7days
NumInqLast6M
NumInqLast6Mexcl7days
NetFractionRevolvingBurden
NetFractionInstallBurden
NumRevolvingTradesWBalance
NumInstallTradesWBalance
NumBank2NatlTradesWHighUtilization
PercentTradesWBalance
average prediction


Partial Dependence profile
Created for the gbm model
gbm
RiskPerformance
NoBureau
NoValid
MSinceOldestTra
NoValid
MSinceMostRece
0.4
0.2
0.0
Bad
Good
0
1
0
1
0
1
No
MSinceMostRecentDe
NoValid
MSinceMostRece
No
MSinceMostRecentInq
NoValid
NetFractionRevol
average prediction
0.4
0.2
0.0
0
1
0
1
0
1
0
1
NoValid
NetFractionInstall
NoValid
NumRevolvingTra
NoValid
NumInstallTrades
NoValid
NumBank2NatlTr
0.4
0.2
0.0
0
1
0
1
0
1
0
1

AUC
RMSE

How good is the model?

ROC curve
LIFT, Gain charts

Which variables are important to the model?

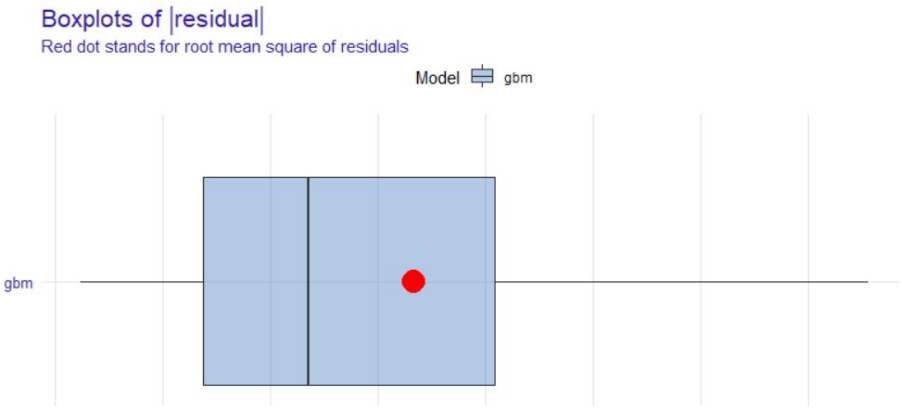
Permutational
Variable Importance

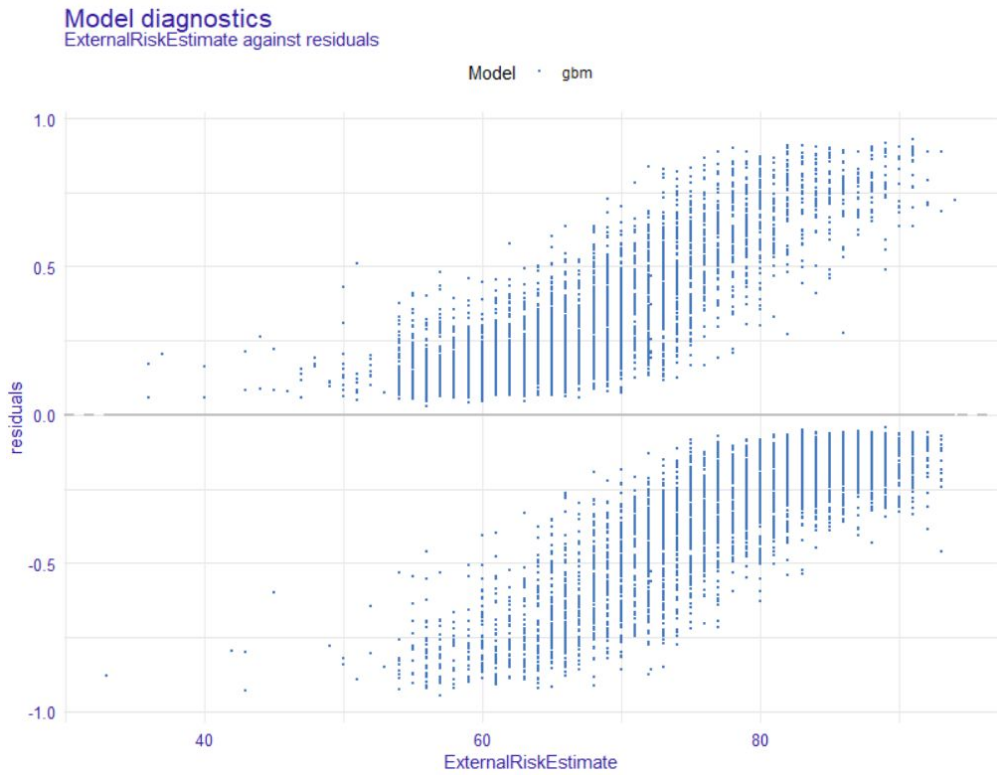
How does a variable affect the average prediction?

Partial Dependence Profile
Accumulated Local Effects

Does the model fit well in general?

MODEL LEVEL

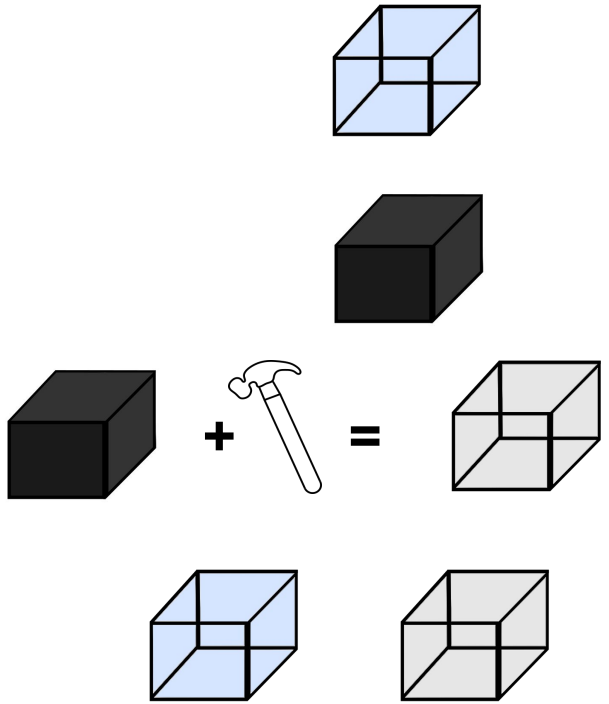

Boxplots of |residual|
Red dot stands for root mean square of residuals
Model
gbm
gbm

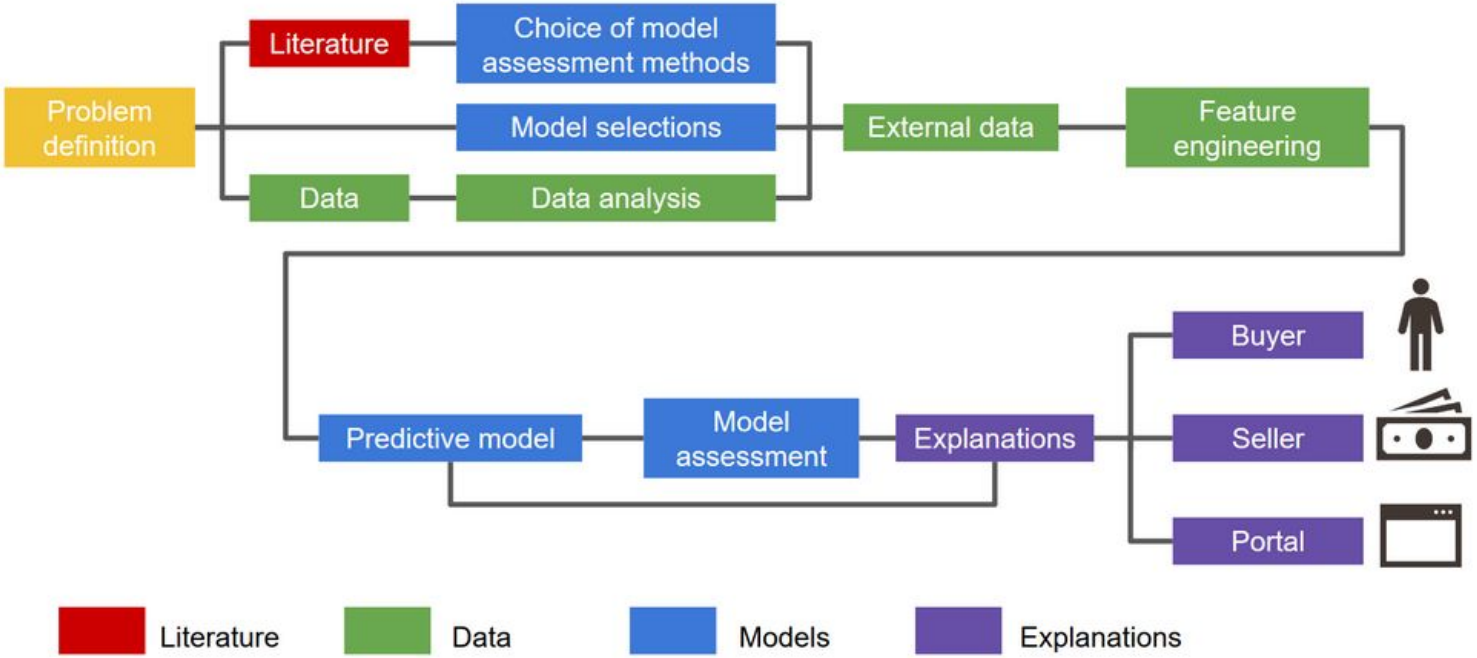

Model diagnostics
ExternalRiskEstimate against residuals
Model
gbm
1.0
0.5
0.0
-0.5
-1.0
residuals
40
60
80
ExternalRiskEstimate

House Sale Prices

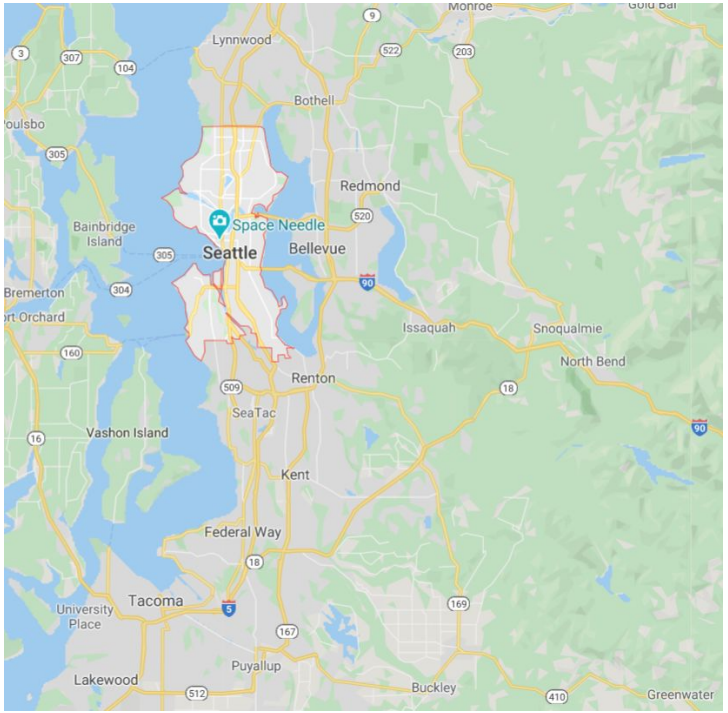
Problem solving strategy

- Build econometric model and understand it.
- Build machine learning models…
- ...and explain them.
- Compare the conclusions.

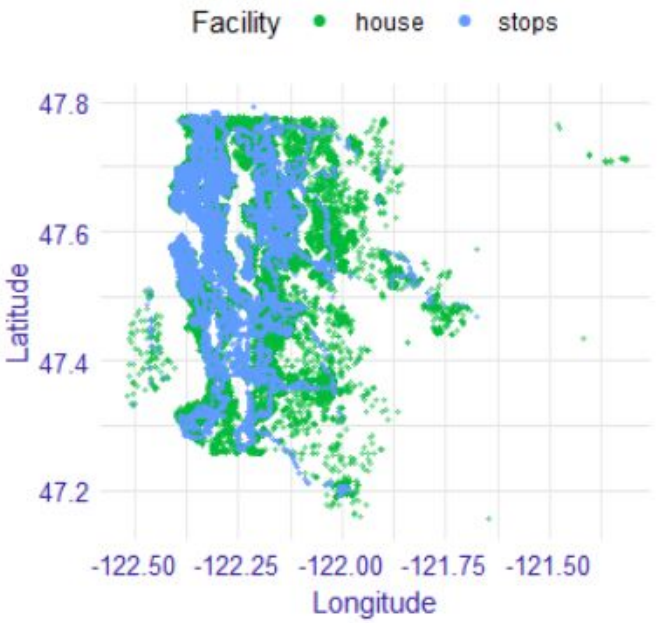


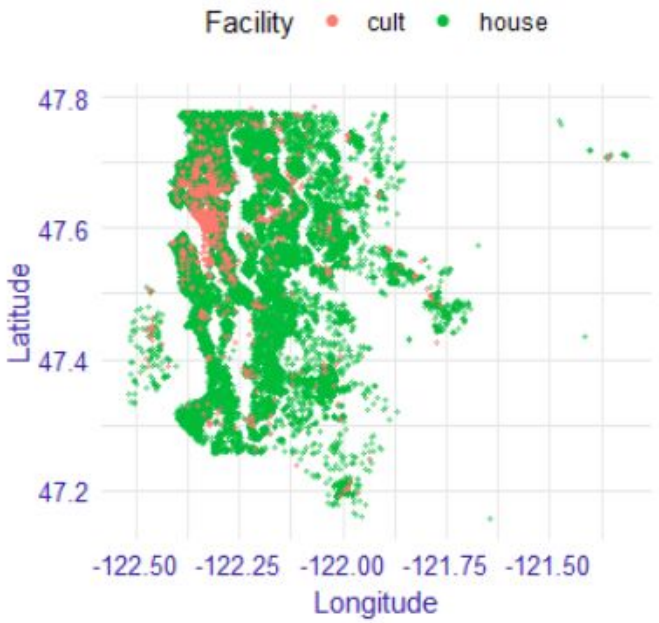

Problem definition
Literature
Choice of model assessment methods
Model selections
Data
Data analysis
External data
Feature engineering
Predictive model
Model assessment
Explanations
Buyer
Seller
Portal
Literature
Data
Models
Explanations

Variable	Description
id	unique ID for each house sold
date	date of the house sale
price	price of each house sold
bedrooms	number of bedrooms
bathrooms	number of bathrooms, where .5 accounts for a room with a toilet but no shower
sqft_living	square footage of the apartments interior living space
sqft_lot	square footage of the land space
floors	number of floors
waterfront	apartment was overlooking the waterfront or not
view	how good the view of the property was
condition	condition of the apartment
grade	level of construction and design
yr_built	the year the house was initially built
yr_renovated	the year of the house's last renovation
zipcode	zipcode area
lat	lattitude
long	longitude



External data



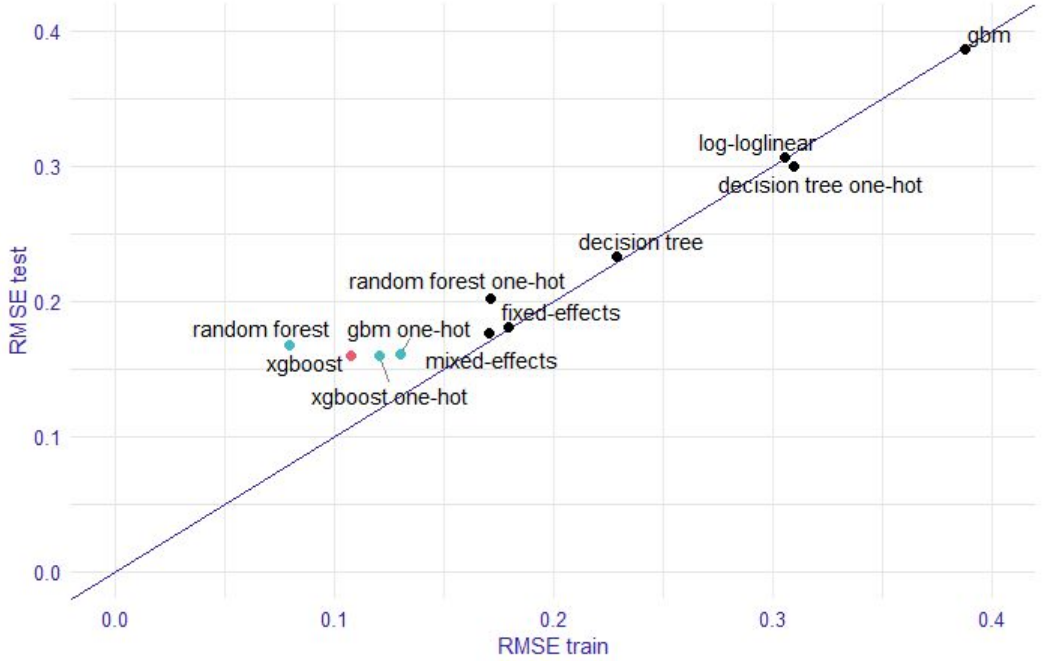


Model selection

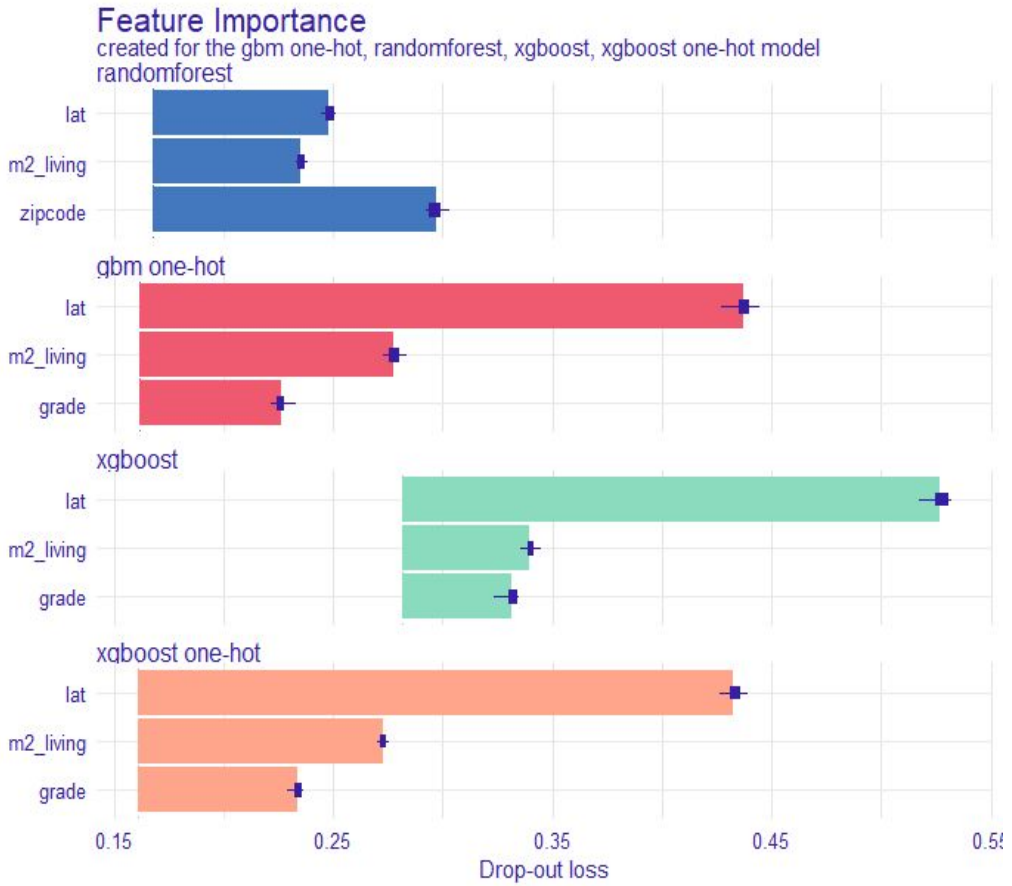
- Random Forest
- XGBoost
- XGBoost one-hot
- GBM one-hot

XAI

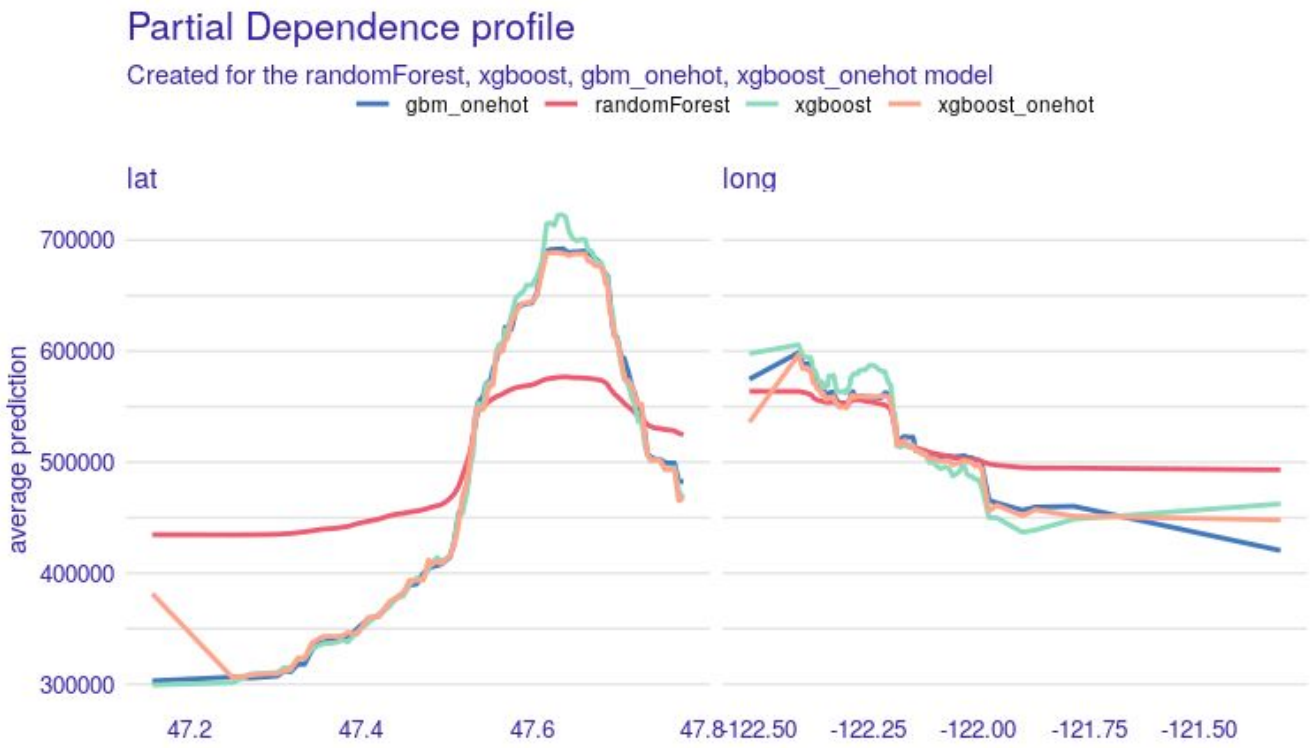
Model of choice



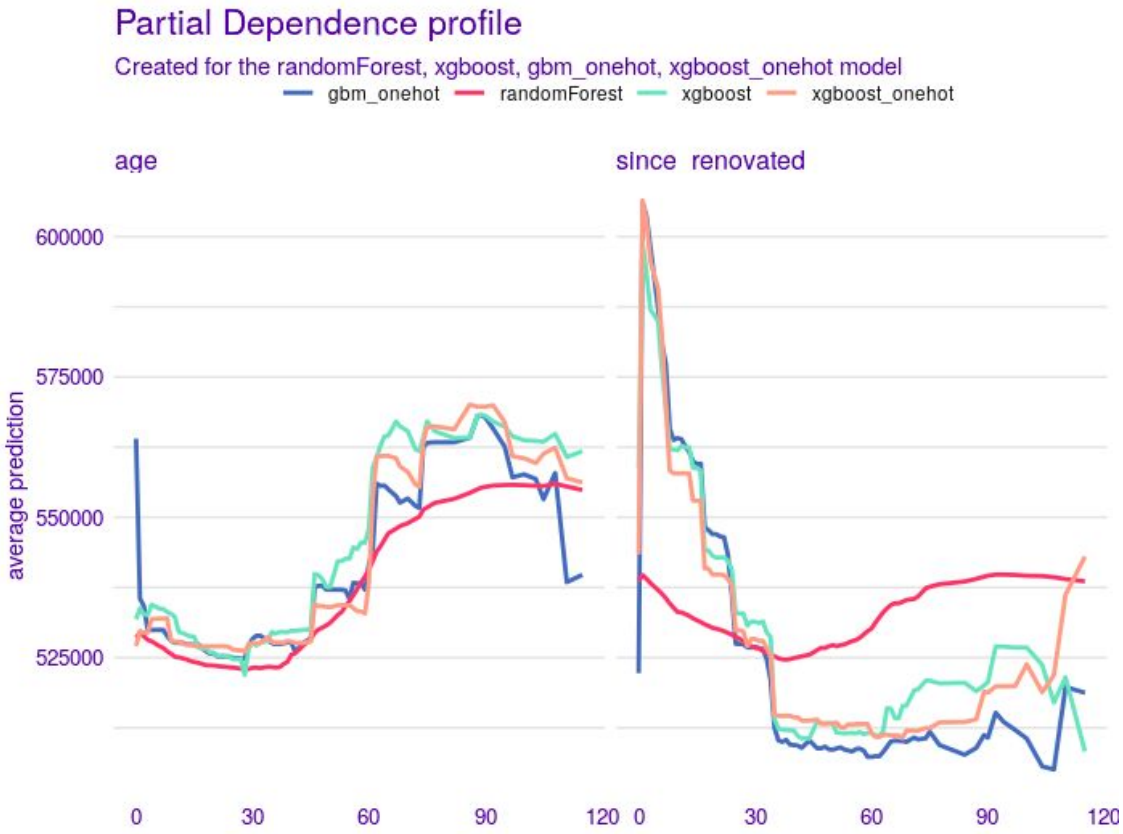
Model selection

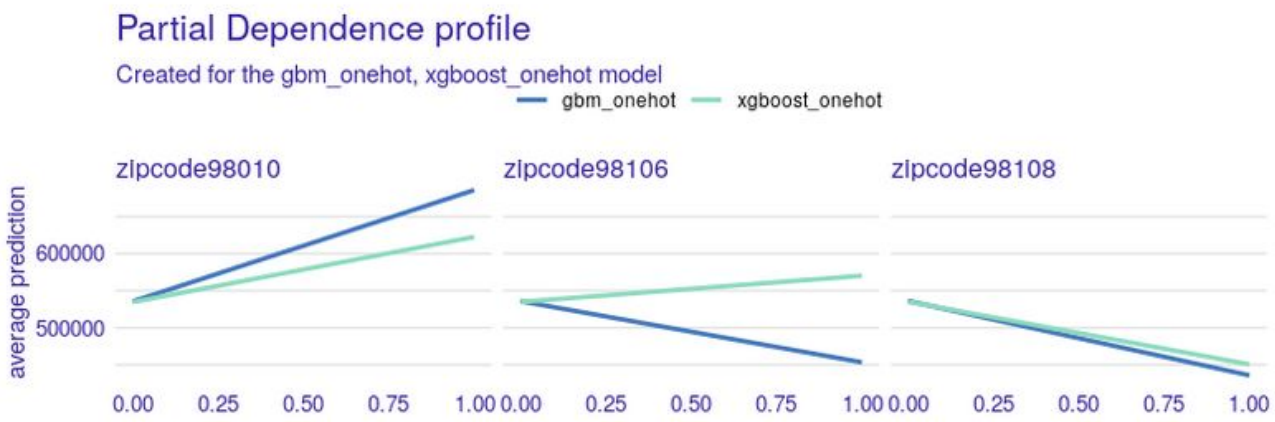


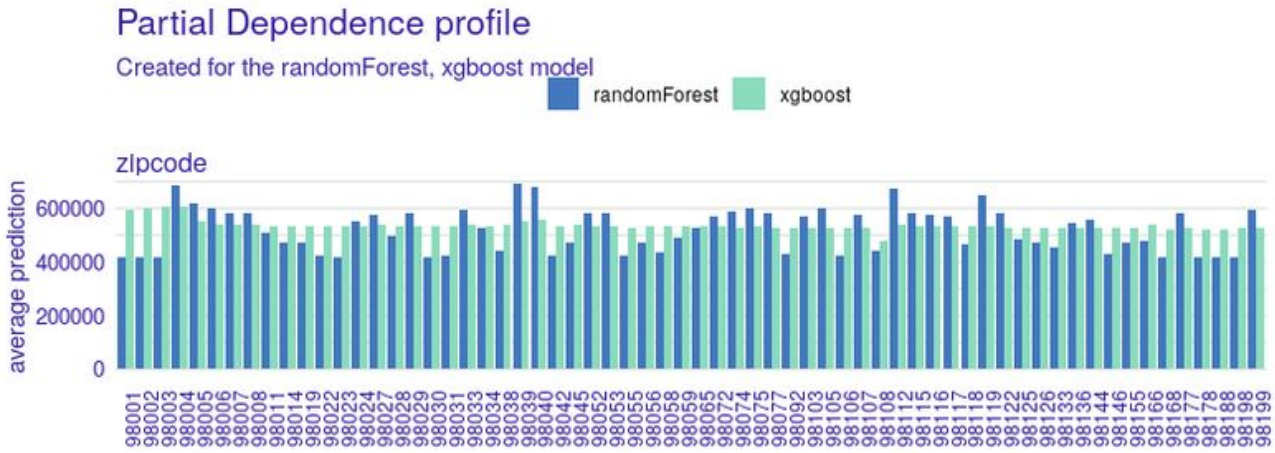
Model selection with PDP



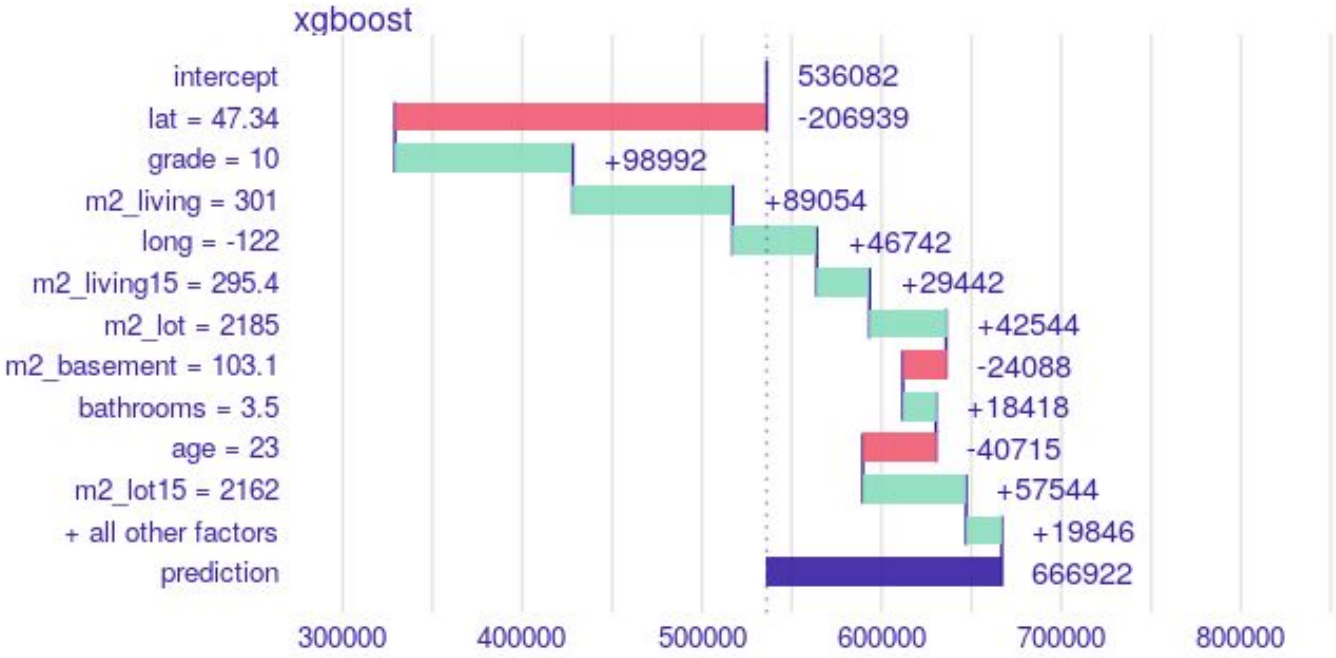
Model selection with PDP



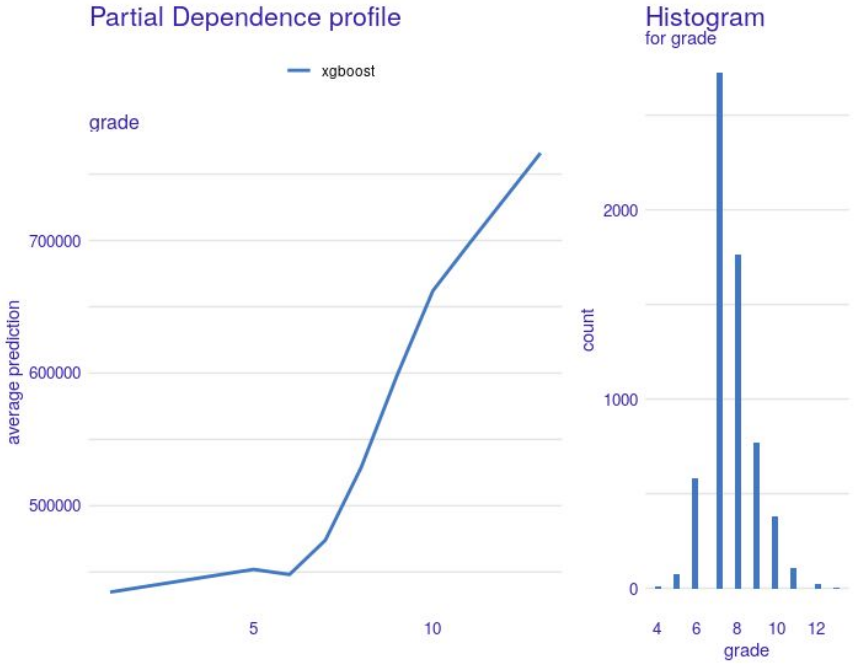

Partial Dependence profile
Created for the gbm_onehot, xgboost_onehot model
gbm_onehot
xgboost_onehot
zipcode98010
zipcode98106
zipcode98108
average prediction
600000
500000
0.00
0.25
0.50
0.75
1.00

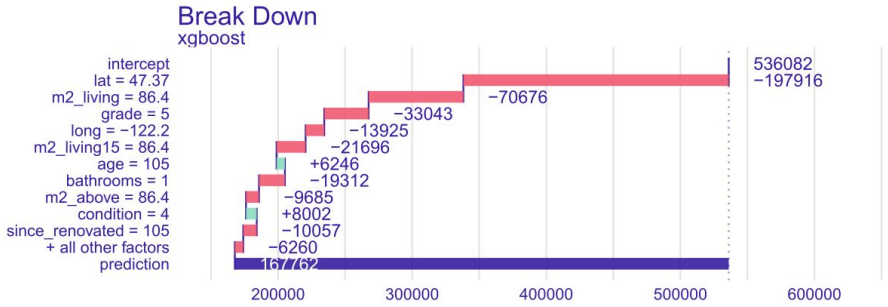

Partial Dependence profile
Created for the randomForest, xgboost model
randomForest
xgboost
zipcode
average prediction
600000
400000
200000
0
98001
98002
98003
98004
98005
98006
98007
98008
98011
98014
98019
98022
98023
98024
98027
98028
98029
98030
98031
98033
98034
98038
98039
98040
98042
98045
98052
98053
98055
98056
98058
98059
98065
98072
98074
98075
98077
98092
98103
98105
98106
98107
98108
98112
98115
98116
98117
98118
98119
98122
98125
98126
98133
98136
98144
98146
98155
98166
98168
98177
98178
98188
98198
98199

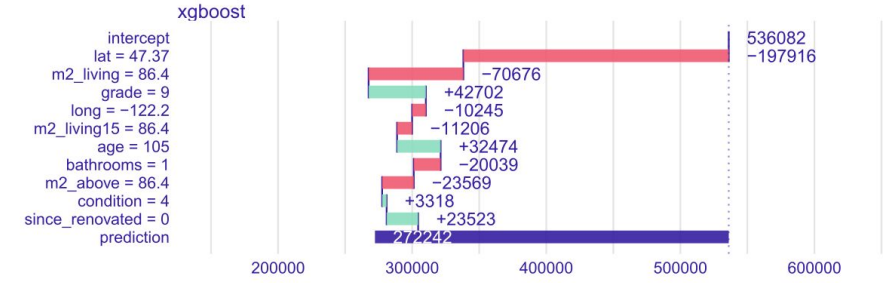
Data Scientist perspective



Financial use case



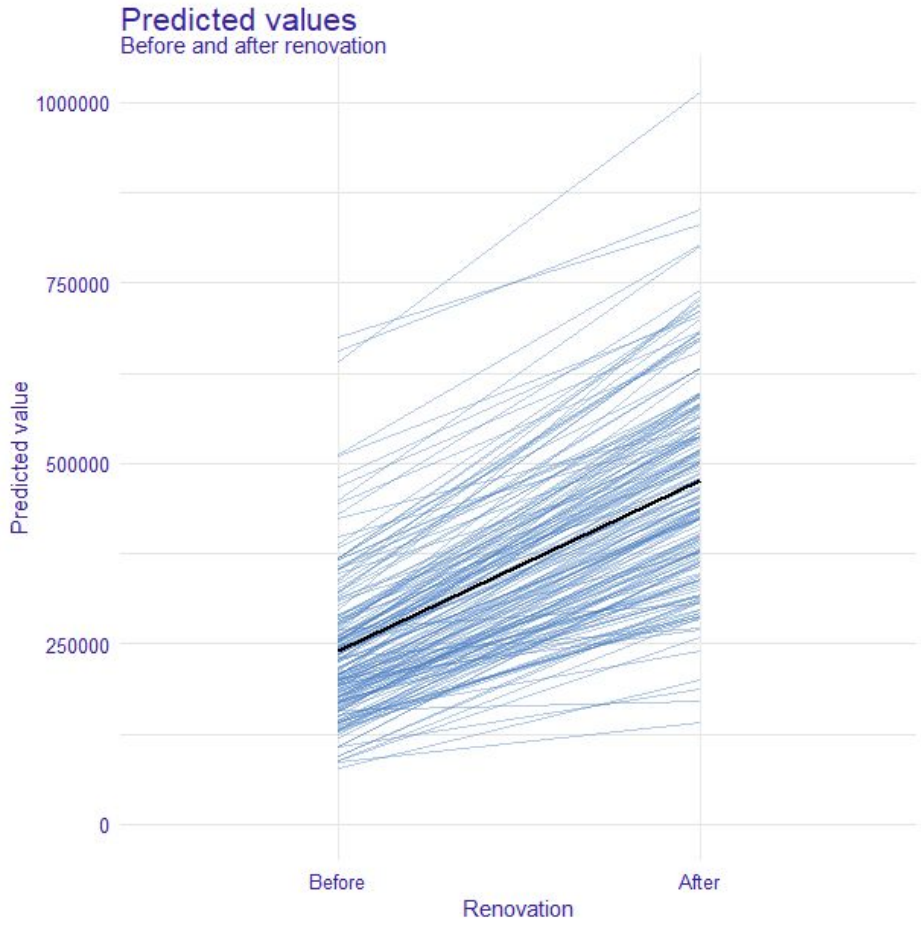




Financial use case

Average uplift:

- +235k$
- +109%



Analiza XAI w R

Model Oriented

MI2DataLab @ Warsaw University of Technology https://mi2-warsaw.github.io/

Repositories 42 Packages People 21 Teams 2 Projects

Pinned repositories

DALEX

moDel Agnostic Language for Exploration and eXplanation

Python ☆ 596 87

DrWhy

DrWhy is the collection of tools for eXplainable AI (XAI). It's based on shared principles and simple grammar for exploration, explanation and visualisation of predictive models.

R ☆ 364 49

randomForestExplainer

A set of tools to understand what is happening inside a Random Forest

R ☆ 161 25

modelStudio

Interactive Studio for Explanatory Model Analysis

R ☆ 126 16

modelDown

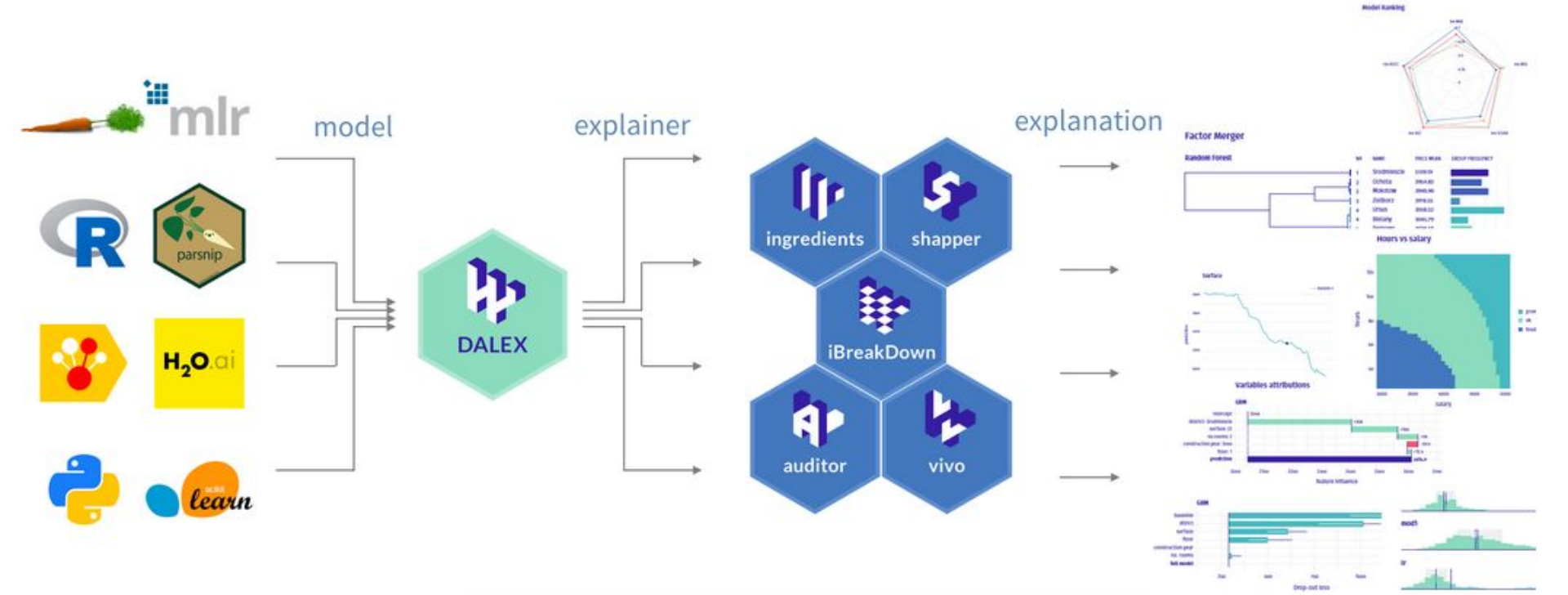
modelDown generates a website with HTML summaries for predictive models

R ☆ 99 11

iBreakDown

Break Down with interactions for local explanations (SHAP, BreakDown, iBreakDown)

R ☆ 50 8


mlr
model
explainer
explanation
parsnip
H2O.ai
scikit learn
DALEX
ingredients
shapper
iBreakDown
auditor
vivo
Model Ranking
Factor Merger
Random Forest
Hours vs salary
Variables attributions

```r
### model gradient boosting - mlr [R]

mod_gbm

### explainer
explain_gbm <- DALEXtra::explain_mlr(mod_gbm,
                                     train,
                                     y = train$RiskPerformance == "Bad",
                                     label = "gbm")
```

DALEXtra

```
>   ### explainer
>   explain_gbm <- DALEXtra::explain_mlr(mod_gbm,
+                                        train,
+                                        y = train$RiskPerformance == "Bad",
+                                        label = "gbm")
Preparation of a new explainer is initiated
  -> model label       :  gbm
  -> data              :  7844  rows  35  cols
  -> target variable   :  7844  values
  -> predict function  :  yhat.WrappedModel  will be used (  default  )
  -> predicted values  :  numerical, min =  0.04138851 , mean =  0.5212069 , max =  0.9708614
  -> model_info        :  package mlr , ver. 2.17.1 , task classification (  default  )
  -> residual function :  difference between y and yhat (  default  )
  -> residuals         :  numerical, min =  -0.9454267 , mean =  0.000210794 , max =  0.9274428
  A new explainer has been created!
>
```

```
>   ### explainer
>   explain_gbm <- DALEXtra::explain_mlr(mod_gbm,
+                                        train,
+                                        y = train$RiskPerformance == "Bad",
+                                        label = "gbm")
Preparation of a new explainer is initiated
  -> model label       :  gbm
  -> data              :  7844  rows  35  cols
  -> target variable   :  7844  values
  -> predict function  :  yhat.WrappedModel  will be used
  -> predicted values  :  numerical, min =  0.04138851 , m                    614
  -> model_info        :  package mlr , ver. 2.17.1 , task
  -> residual function :  difference between y and yhat (
  -> residuals         :  numerical, min =  -0.9454267 , m                    74428
  A new explainer has been created!
>
```

explainer

model

data: data.frame

y: numeric

y_hat: numeric

predict_function: function (model, data)

residuals: numeric

residual_function: function(model, data, y)

weights: numeric

model_info: list(package, ver, type)

class: character

label: character

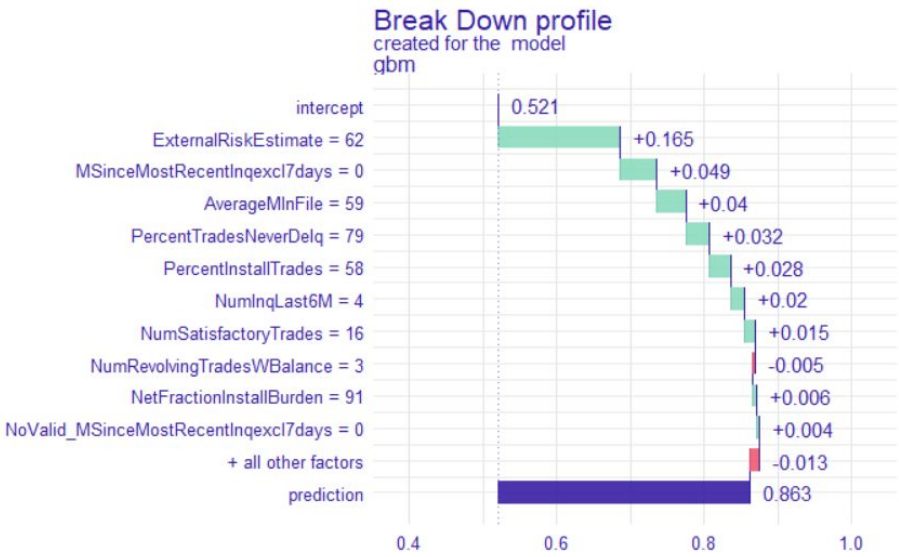
```
> mp_gbm <- model_performance(explain_gbm)
> mp_gbm
Measures for:  classification
recall     : 0.7867971
precision  : 0.7360476
f1         : 0.7605767
accuracy   : 0.7417134
auc        : 0.8196526

Residuals:
         0%         10%         20%         30%         40%         50%         60%         70%
-0.94542666 -0.56170984 -0.37598467 -0.23643189 -0.14372034  0.08493434  0.15816946  0.23813542
        80%         90%        100%
 0.36884241  0.53009286  0.92744281
>
```

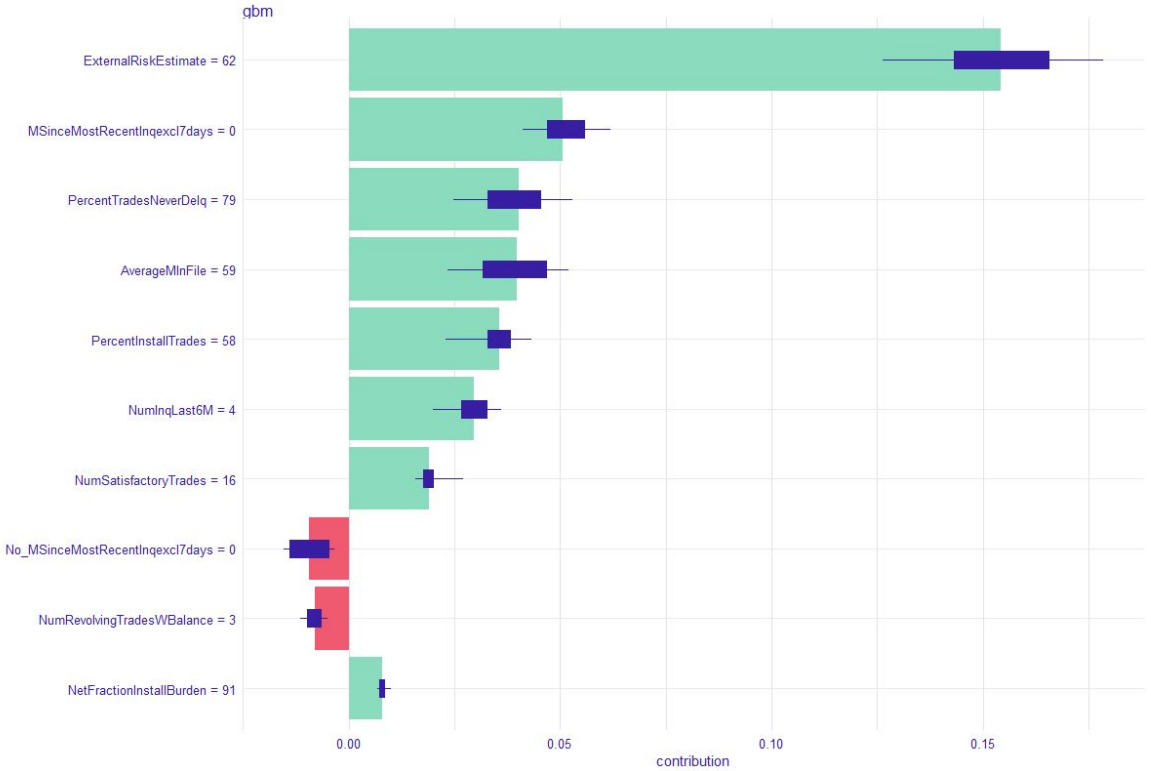
```
> selected_obs <- test[2,]
> predict(explain_gbm, selected_obs)
[1] 0.862629
>
```

```
>
> bd_gbm <- predict_parts(explain_gbm, selected_obs, type = "break_down")
> bd_gbm
                                                    contribution
gbm: intercept                                             0.521
gbm: ExternalRiskEstimate = 62                             0.165
gbm: MSinceMostRecentInqexcl7days = 0                      0.049
gbm: AverageMInFile = 59                                   0.040
gbm: PercentTradesNeverDelq = 79                           0.032
gbm: PercentInstallTrades = 58                             0.028
gbm: NumInqLast6M = 4                                      0.020
gbm: NumSatisfactoryTrades = 16                            0.015
gbm: NetFractionRevolvingBurden = 25                      -0.001
gbm: No_MSinceMostRecentInqexcl7days = 0                  -0.003
gbm: NumRevolvingTradesWBalance = 3                       -0.005
gbm: NetFractionInstallBurden = 91                         0.006
gbm: PercentTradesWBalance = 70                           -0.002
gbm: MSinceMostRecentDelq = 23                            -0.002
gbm: MaxDelq2PublicRecLast12M = 6                         -0.002
gbm: NumTrades60Ever2DerogPubRec = 4                      -0.003
gbm: NoValid_MSinceMostRecentInqexcl7days = 0              0.004
gbm: NumInstallTradesWBalance = 4                         -0.002
gbm: MSinceOldestTradeOpen = 131                           0.002
gbm: NumTotalTrades = 20                                   0.001
gbm: NumTradesOpeninLast12M = 4                           -0.001
gbm: NumBank2NatlTradesWHighUtilization = 1                0.002
gbm: NoValid_NumBank2NatlTradesWHighUtilization = 0       -0.001
gbm: NumInqLast6Mexcl7days = 4                             0.000
gbm: NumTrades90Ever2DerogPubRec = 2                      -0.001
gbm: MSinceMostRecentTradeOpen = 4                         0.000
gbm: MaxDelqEver = 5                                      -0.001
gbm: NoValid_MSinceMostRecentDelq = 0                      0.000
gbm: NoBureau = 0                                          0.000
gbm: NoValid_NumInstallTradesWBalance = 0                  0.000
gbm: NoValid_MSinceOldestTradeOpen = 0                     0.000
gbm: NoValid_NetFractionInstallBurden = 0                  0.000
gbm: NoValid_NetFractionRevolvingBurden = 0                0.000
gbm: RiskPerformance = Bad                                 0.000
gbm: No_MSinceMostRecentDelq = 0                           0.000
gbm: NoValid_NumRevolvingTradesWBalance = 0                0.000
gbm: prediction                                            0.863
>
```

```
>
> bd_gbm <- predict_parts(explain_gbm, selected_obs, type = "break_down")
> bd_gbm
                                                    contribution
gbm: intercept                                             0.521
gbm: ExternalRiskEstimate = 62                             0.165
gbm: MSinceMostRecentInqexcl7days = 0                      0.049
gbm: AverageMInFile = 59                                   0.040
gbm: PercentTradesNeverDelq = 79                           0.032
gbm: PercentInstallTrades = 58                             0.028
gbm: NumInqLast6M = 4                                      0.020
gbm: NumSatisfactoryTrades = 16                            0.015
gbm: NetFractionRevolvingBurden = 25                      -0.001
gbm: No_MSinceMostRecentInqexcl7days = 0                  -0.003
gbm: NumRevolvingTradesWBalance = 3                       -0.005
gbm: NetFractionInstallBurden = 91                         0.006
gbm: PercentTradesWBalance = 70                           -0.002
gbm: MSinceMostRecentDelq = 23                            -0.002
gbm: MaxDelq2PublicRecLast12M = 6                         -0.002
gbm: NumTrades60Ever2DerogPubRec = 4                      -0.003
gbm: NoValid_MSinceMostRecentInqexcl7days = 0              0.004
gbm: NumInstallTradesWBalance = 4                         -0.002
gbm: MSinceOldestTradeOpen = 131                           0.002
gbm: NumTotalTrades = 20                                   0.001
gbm: NumTradesOpeninLast12M = 4                           -0.001
gbm: NumBank2NatlTradesWHighUtilization = 1                0.002
gbm: NoValid_NumBank2NatlTradesWHighUtilization = 0       -0.001
gbm: NumInqLast6Mexcl7days = 4                             0.000
gbm: NumTrades90Ever2DerogPubRec = 2                      -0.001
gbm: MSinceMostRecentTradeOpen = 4                         0.000
gbm: MaxDelqEver = 5                                      -0.001
gbm: NoValid_MSinceMostRecentDelq = 0                      0.000
gbm: NoBureau = 0                                          0.000
gbm: NoValid_NumInstallTradesWBalance = 0                  0.000
gbm: NoValid_MSinceOldestTradeOpen = 0                     0.000
gbm: NoValid_NetFractionInstallBurden = 0                  0.000
gbm: NoValid_NetFractionRevolvingBurden = 0                0.000
gbm: RiskPerformance = Bad                                 0.000
gbm: No_MSinceMostRecentDelq = 0                           0.000
gbm: NoValid_NumRevolvingTradesWBalance = 0                0.000
gbm: prediction                                            0.863
> plot(bd_gbm)
```



```
> shap_gbm <- predict_parts(explain_gbm, selected_obs, type = "shap")
```

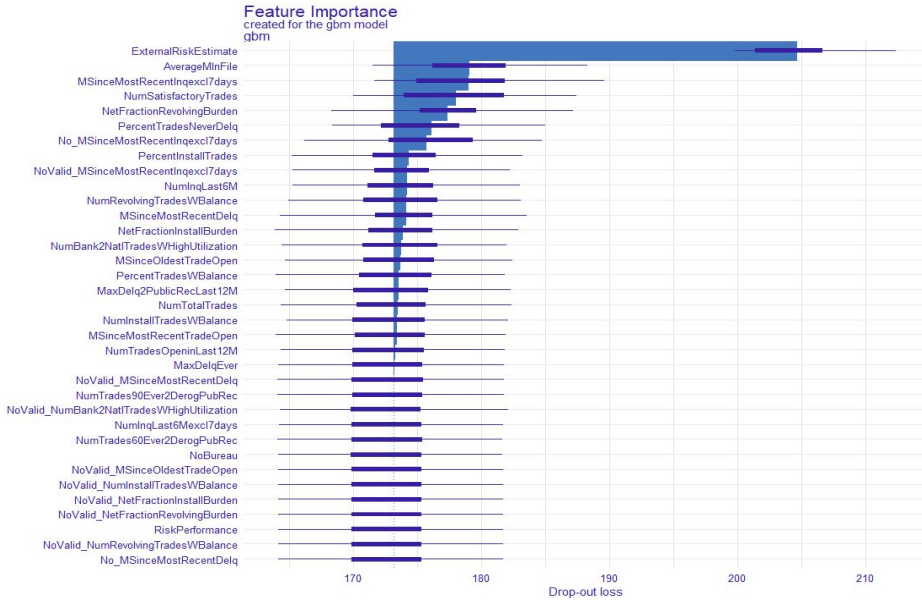


```
> mp_gbm <- model_parts(explain_gbm)
> mp_gbm
                                  variable mean_dropout_loss label
1                                _full_model_         0.1815725   gbm
2                                   NoBureau          0.1814739   gbm
3           NoValid_NetFractionRevolvingBurden         0.1815493   gbm
4             NoValid_NetFractionInstallBurden         0.1815602   gbm
5                NoValid_MSinceOldestTradeOpen         0.1815721   gbm
6                             RiskPerformance          0.1815725   gbm
7                     No_MSinceMostRecentDelq          0.1815725   gbm
8           NoValid_NumRevolvingTradesWBalance         0.1815725   gbm
9             NoValid_NumInstallTradesWBalance         0.1815812   gbm
10                    NumInqLast6Mexcl7days          0.1815878   gbm
11               NoValid_MSinceMostRecentDelq          0.1816706   gbm
12                NumTrades60Ever2DerogPubRec         0.1816710   gbm
13                     NumTradesOpeninLast12M         0.1817206   gbm
14                               MaxDelqEver          0.1817529   gbm
15                  NumInstallTradesWBalance         0.1817647   gbm
16                NumTrades90Ever2DerogPubRec         0.1818872   gbm
17 NoValid_NumBank2NatlTradesWHighUtilization         0.1819068   gbm
18                             NumTotalTrades         0.1823150   gbm
19                   MaxDelq2PublicRecLast12M         0.1823267   gbm
20                 MSinceMostRecentTradeOpen          0.1823741   gbm
21                    PercentTradesWBalance         0.1826900   gbm
22          NumBank2NatlTradesWHighUtilization         0.1829283   gbm
23                    MSinceOldestTradeOpen         0.1831786   gbm
24                     MSinceMostRecentDelq         0.1832626   gbm
25                  NetFractionInstallBurden         0.1834114   gbm
26       NoValid_MSinceMostRecentInqexcl7days         0.1835627   gbm
27                             NumInqLast6M          0.1837569   gbm
28                 NumRevolvingTradesWBalance         0.1846691   gbm
29                    PercentInstallTrades          0.1855981   gbm
30            No_MSinceMostRecentInqexcl7days         0.1861655   gbm
31                   PercentTradesNeverDelq         0.1866492   gbm
32                 NetFractionRevolvingBurden         0.1896944   gbm
33                    NumSatisfactoryTrades         0.1917023   gbm
34                           AverageMInFile          0.1935911   gbm
35               MSinceMostRecentInqexcl7days         0.1953012   gbm
36                    ExternalRiskEstimate          0.2644948   gbm
37                               _baseline_         0.4862608   gbm
>
```

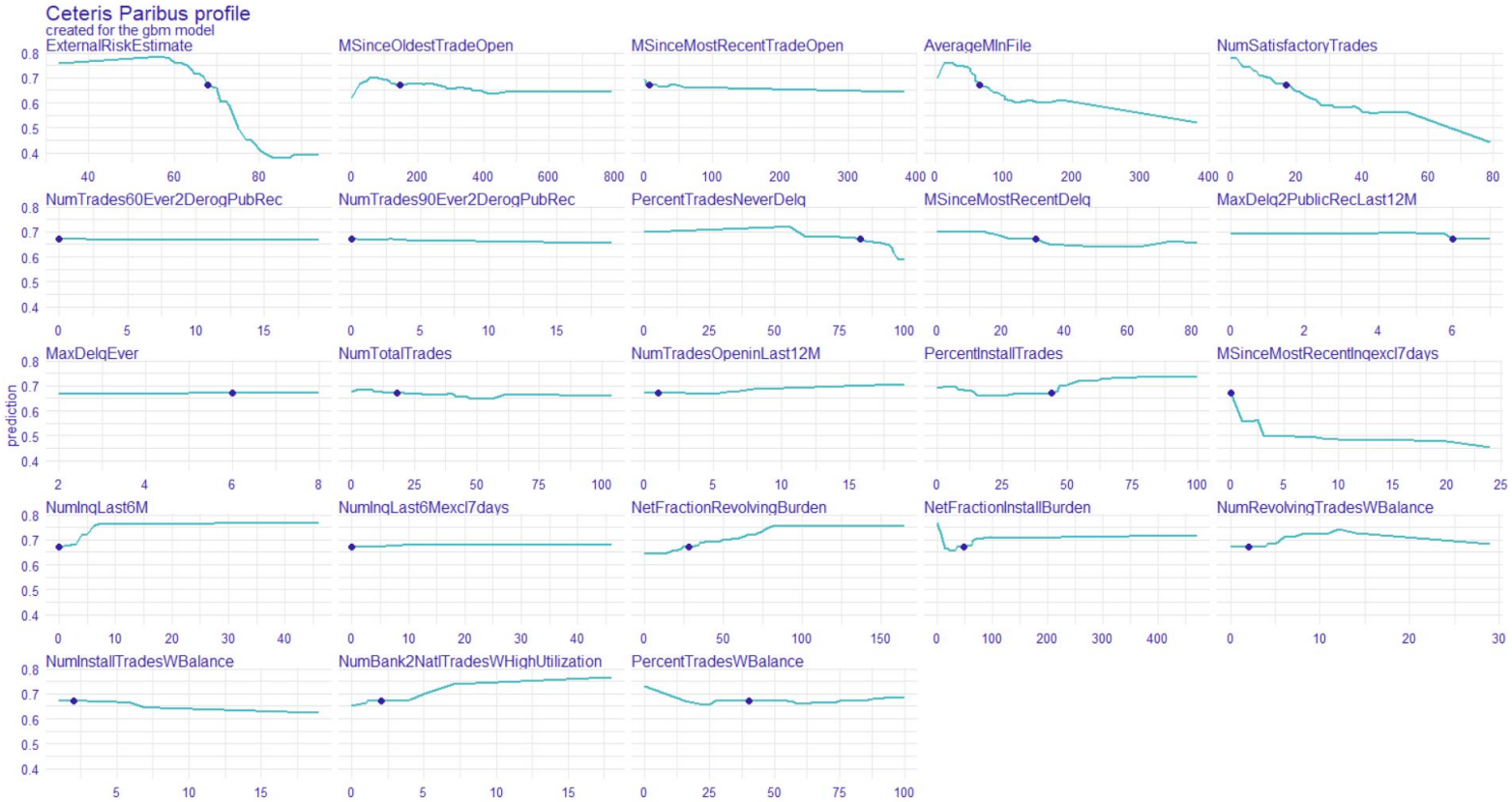
```
> mp_gbm <- model_parts(explain_gbm)
> mp_gbm
```

	variable	mean_dropout_loss	label
1	_full_model_	0.1815725	gbm
2	NoBureau	0.1814739	gbm
3	NoValid_NetFractionRevolvingBurden	0.1815493	gbm
4	NoValid_NetFractionInstallBurden	0.1815602	gbm
5	NoValid_MSinceOldestTradeOpen	0.1815721	gbm
6	RiskPerformance	0.1815725	gbm
7	No_MSinceMostRecentDelq	0.1815725	gbm
8	NoValid_NumRevolvingTradesWBalance	0.1815725	gbm
9	NoValid_NumInstallTradesWBalance	0.1815812	gbm
10	NumInqLast6Mexcl7days	0.1815878	gbm
11	NoValid_MSinceMostRecentDelq	0.1816706	gbm
12	NumTrades60Ever2DerogPubRec	0.1816710	gbm
13	NumTradesOpeninLast12M	0.1817206	gbm
14	MaxDelqEver	0.1817529	gbm
15	NumInstallTradesWBalance	0.1817647	gbm
16	NumTrades90Ever2DerogPubRec	0.1818872	gbm
17	NoValid_NumBank2NatlTradesWHighUtilization	0.1819068	gbm
18	NumTotalTrades	0.1823150	gbm
19	MaxDelq2PublicRecLast12M	0.1823267	gbm
20	MSinceMostRecentTradeOpen	0.1823741	gbm
21	PercentTradesWBalance	0.1826900	gbm
22	NumBank2NatlTradesWHighUtilization	0.1829283	gbm
23	MSinceOldestTradeOpen	0.1831786	gbm
24	MSinceMostRecentDelq	0.1832626	gbm
25	NetFractionInstallBurden	0.1834114	gbm
26	NoValid_MSinceMostRecentInqexcl7days	0.1835627	gbm
27	NumInqLast6M	0.1837569	gbm
28	NumRevolvingTradesWBalance	0.1846691	gbm
29	PercentInstallTrades	0.1855981	gbm
30	No_MSinceMostRecentInqexcl7days	0.1861655	gbm
31	PercentTradesNeverDelq	0.1866492	gbm
32	NetFractionRevolvingBurden	0.1896944	gbm
33	NumSatisfactoryTrades	0.1917023	gbm
34	AverageMInFile	0.1935911	gbm
35	MSinceMostRecentInqexcl7days	0.1953012	gbm
36	ExternalRiskEstimate	0.2644948	gbm
37	_baseline_	0.4862608	gbm

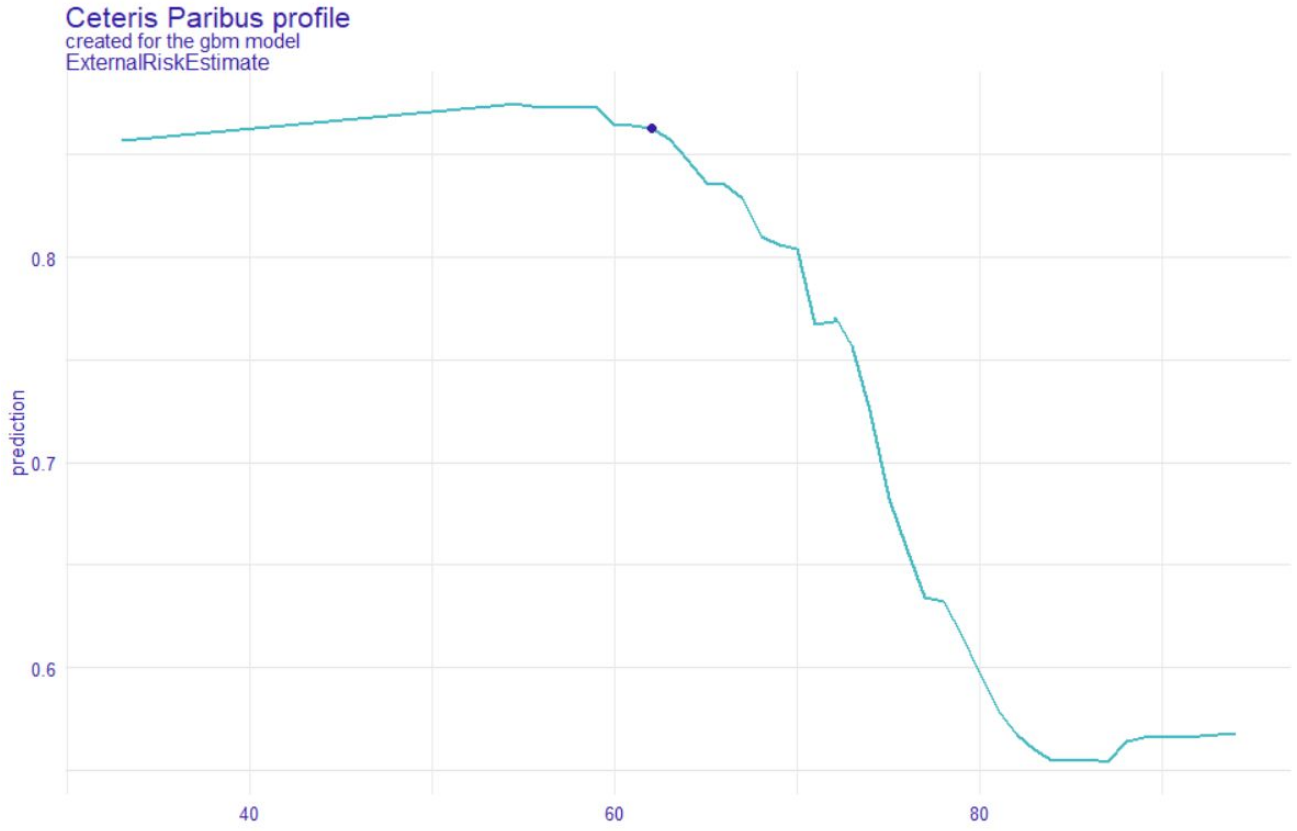
```
> plot(mp_gbm)
```



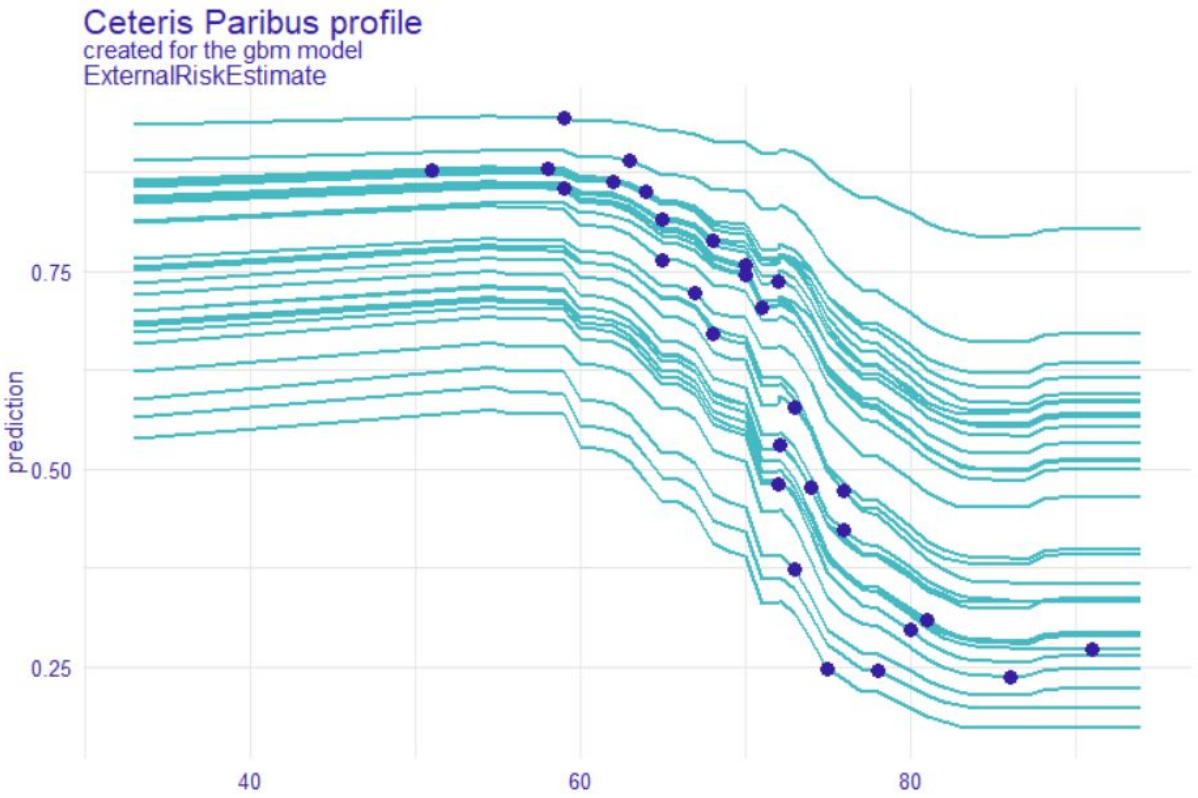
```
> cp_gbm <- predict_profile(explain_gbm, selected_obs)
> cp_gbm
```



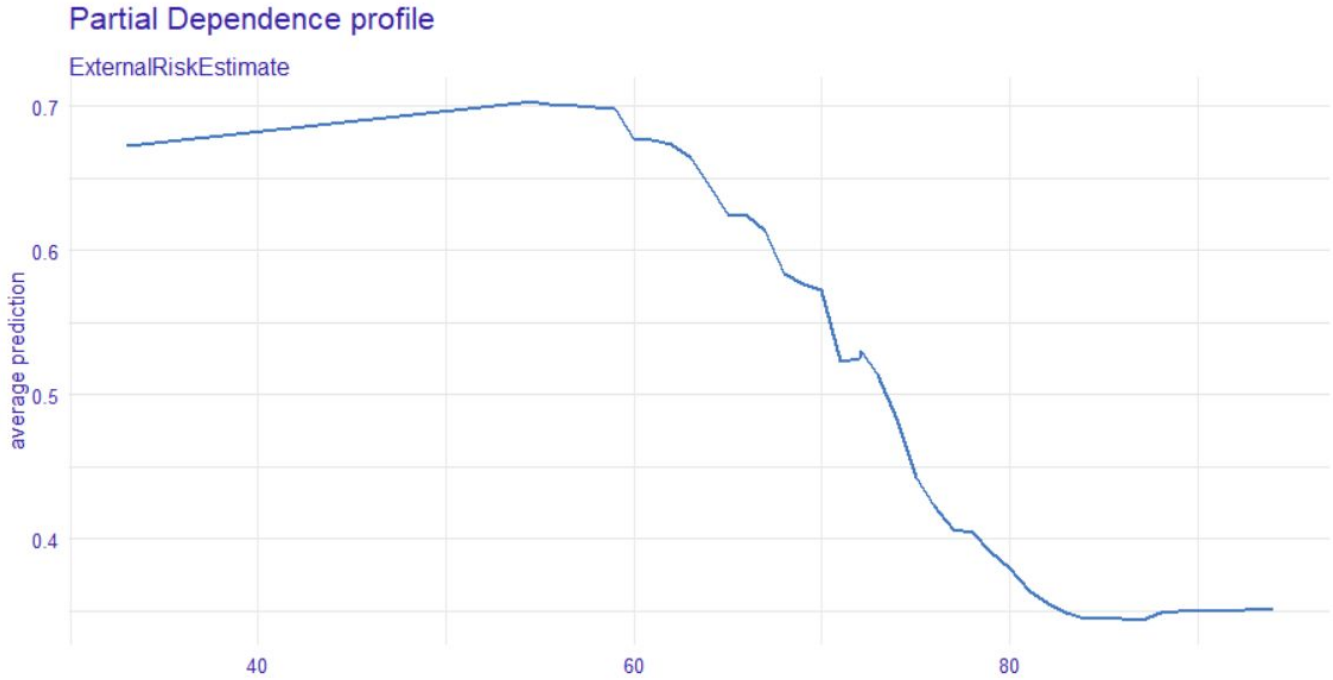
```
> cp_gbm_ere <- predict_profile(explain_gbm, selected_obs, variables = c("ExternalRiskEstimate"))
> plot(cp_gbm_ere)
```



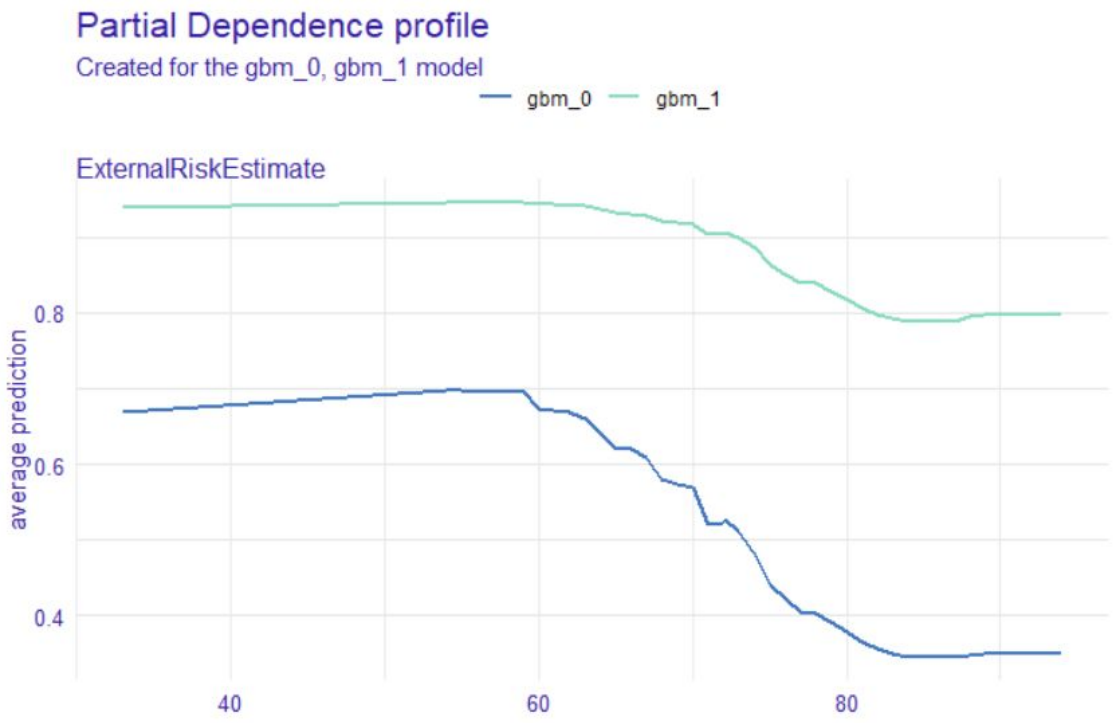
```
> pp <- predict_profile(explain_gbm, test[1:30,])
> plot(pp, variables = c("ExternalRiskEstimate"))
```



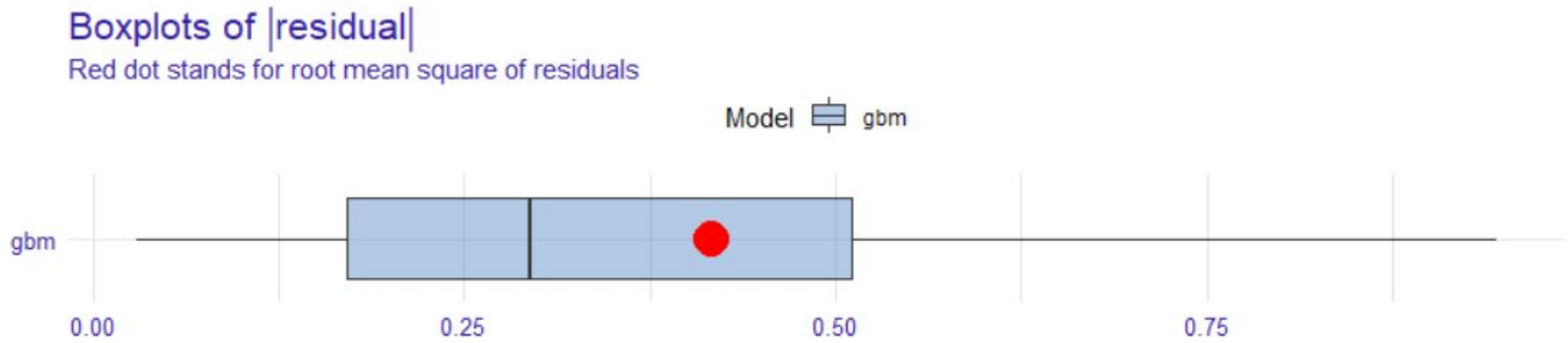
```
> pdp_gbm <- model_profile(explain_gbm)
> plot(pdp_gbm, variables = c("ExternalRiskEstimate"))
```



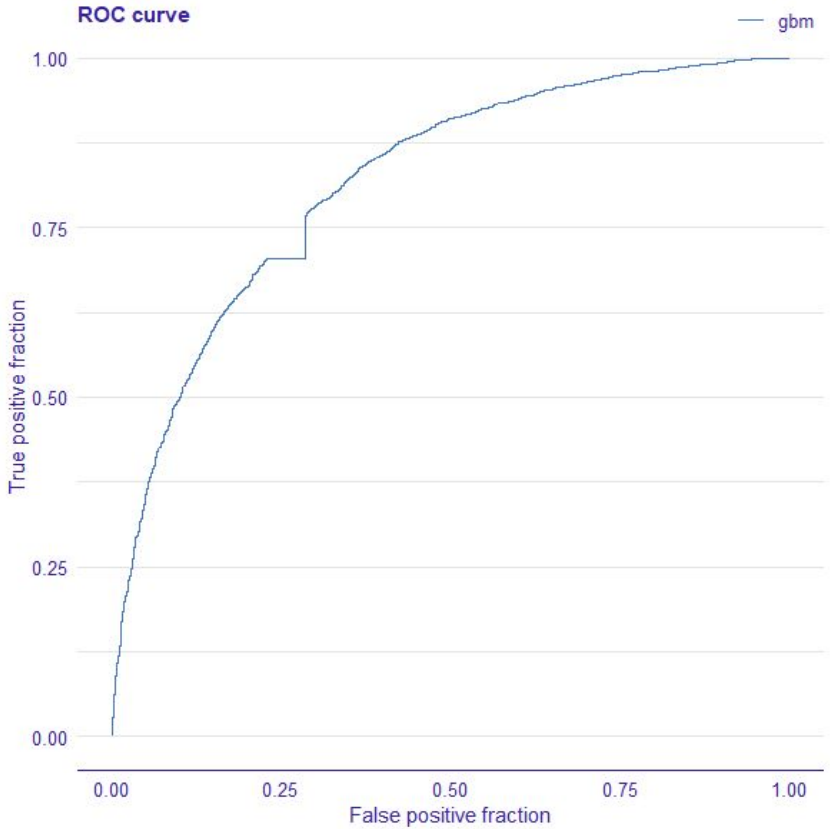
```
> pdp_gbm <- model_profile(explain_gbm, groups = "NoValid_NumRevolvingTradesWBalance")
> plot(pdp_gbm, variables = c("ExternalRiskEstimate"))
```

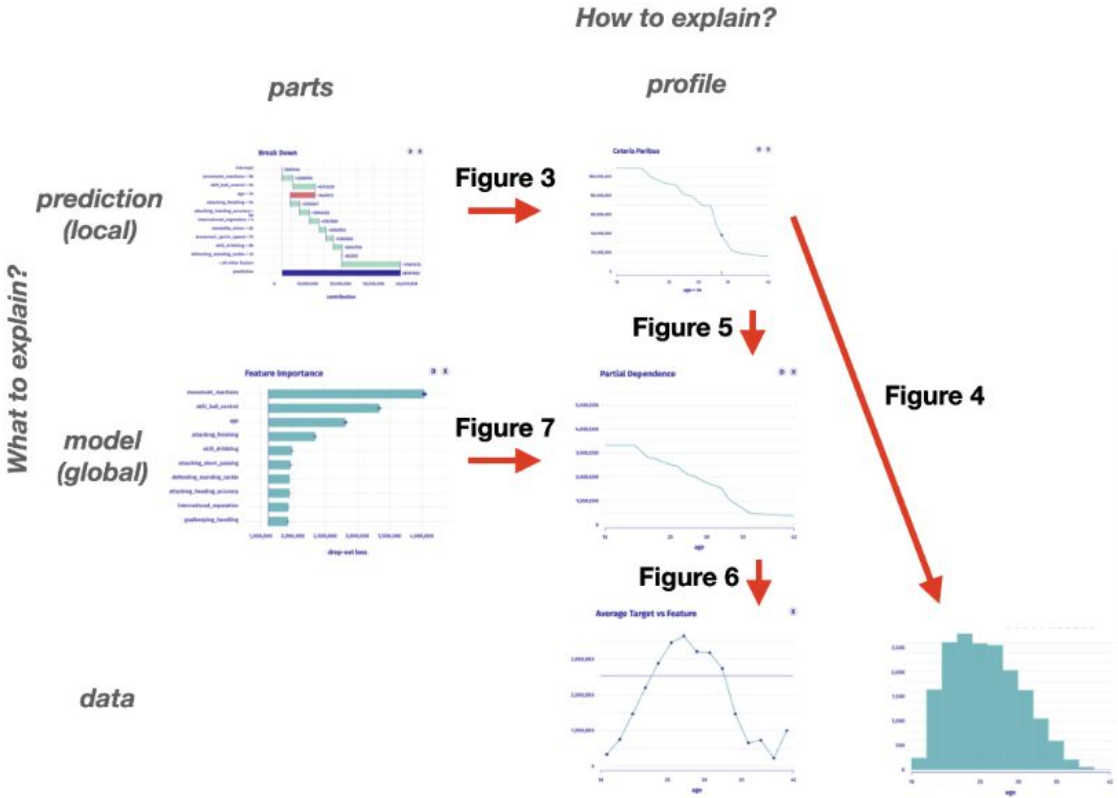


```
plot(mp_gbm, geom = "boxplot")
```



```
plot(mp_gbm, geom = "roc")
```

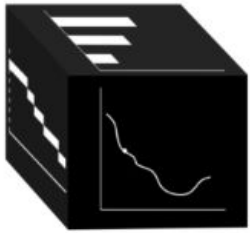



How to explain?
parts
profile
What to explain?
prediction (local)
model (global)
data
Figure 3
Figure 4
Figure 5
Figure 6
Figure 7
Break Down
Ceteris Paribus
Feature Importance
Partial Dependence
Average Target vs Feature

Black Box Model

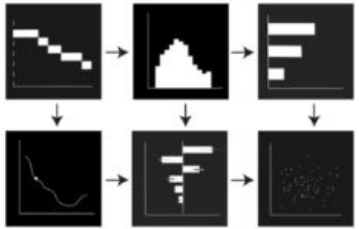
I generation explanations

(single aspect
model explanation)



II generation explanations

(interactive explanatory
model analysis)



modelStudio (https://pbiecek.github.io/explainFIFA20/)