Explainable Machine Learning

Kasia Kulma, Senior Data Scientist
Hannah Frick, Senior Data Scientist

Agenda

- Introductions
- Motivation
- DALEX

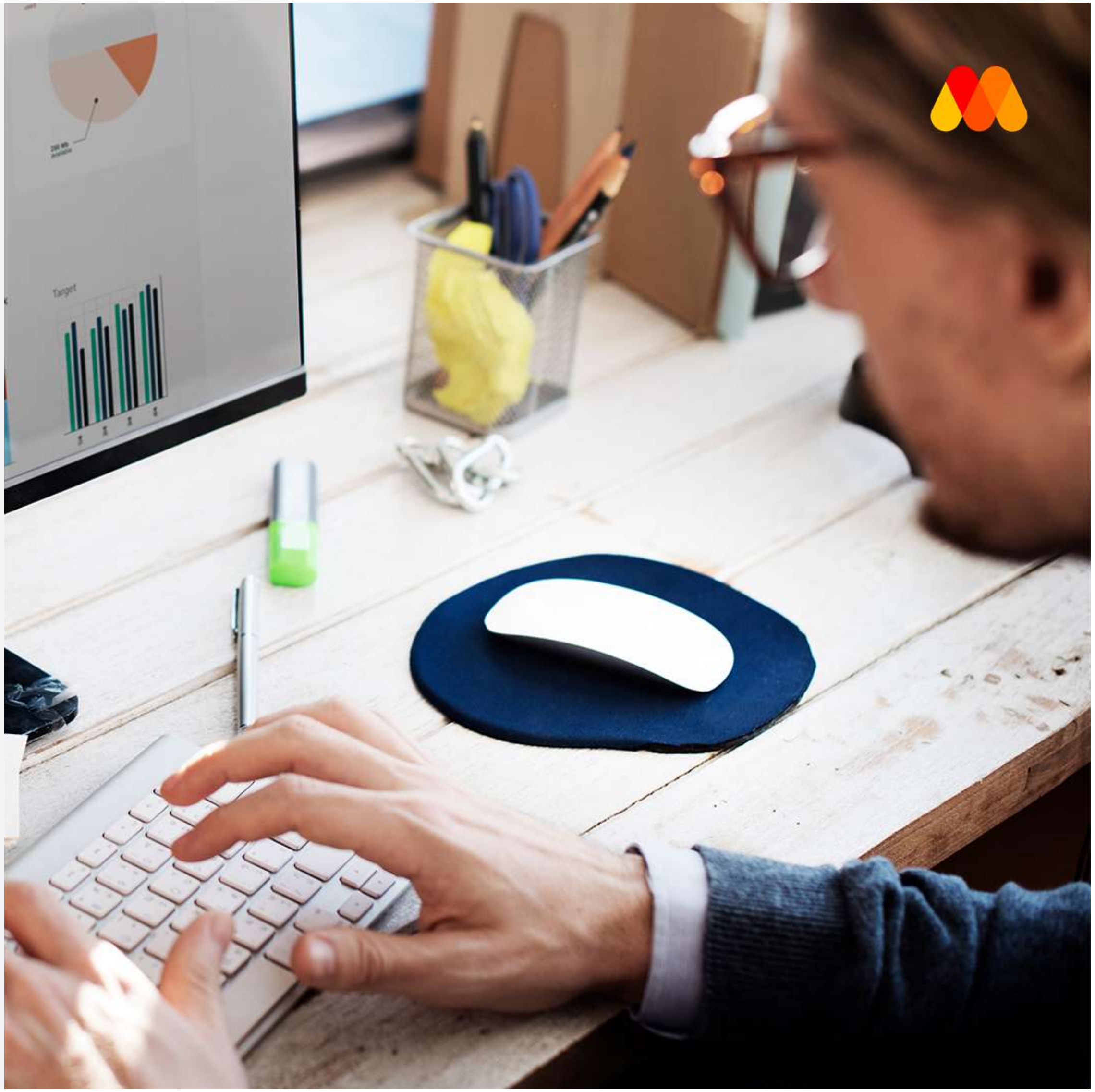
Mango Solutions

- Data science services, advice, training
- Cross-sector
- ~ 80 people
- London and Chippenham
- We ❤ R, Python and Spark

Motivation

Business and Government Decisions Increasingly Rely on Machine Learning and AI

- Healthcare
- Banking
- Crime
- Education
- Recruitment
- …



We Choose to Trust ML Algorithms based on their Accuracy

=

ACCURACY

TRUST

More Accurate ML Algorithms are also **Less** Interpretable

INTERPRETIBILITY

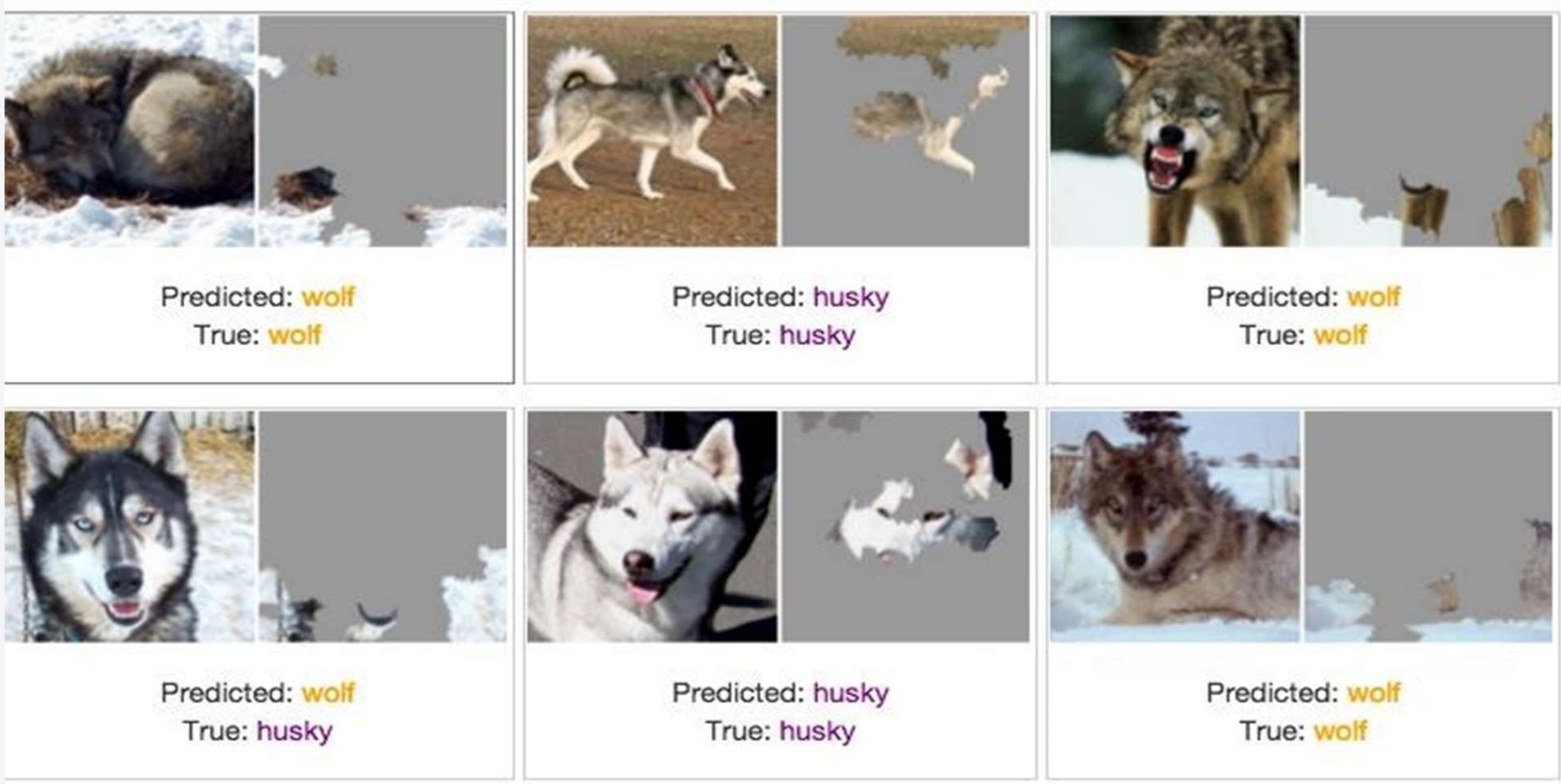
ACCURACY

- Overfitting & Noise
- Correlation
- Data Leakage
- Truth

Data Leakage in Action

Data Leakage in Action

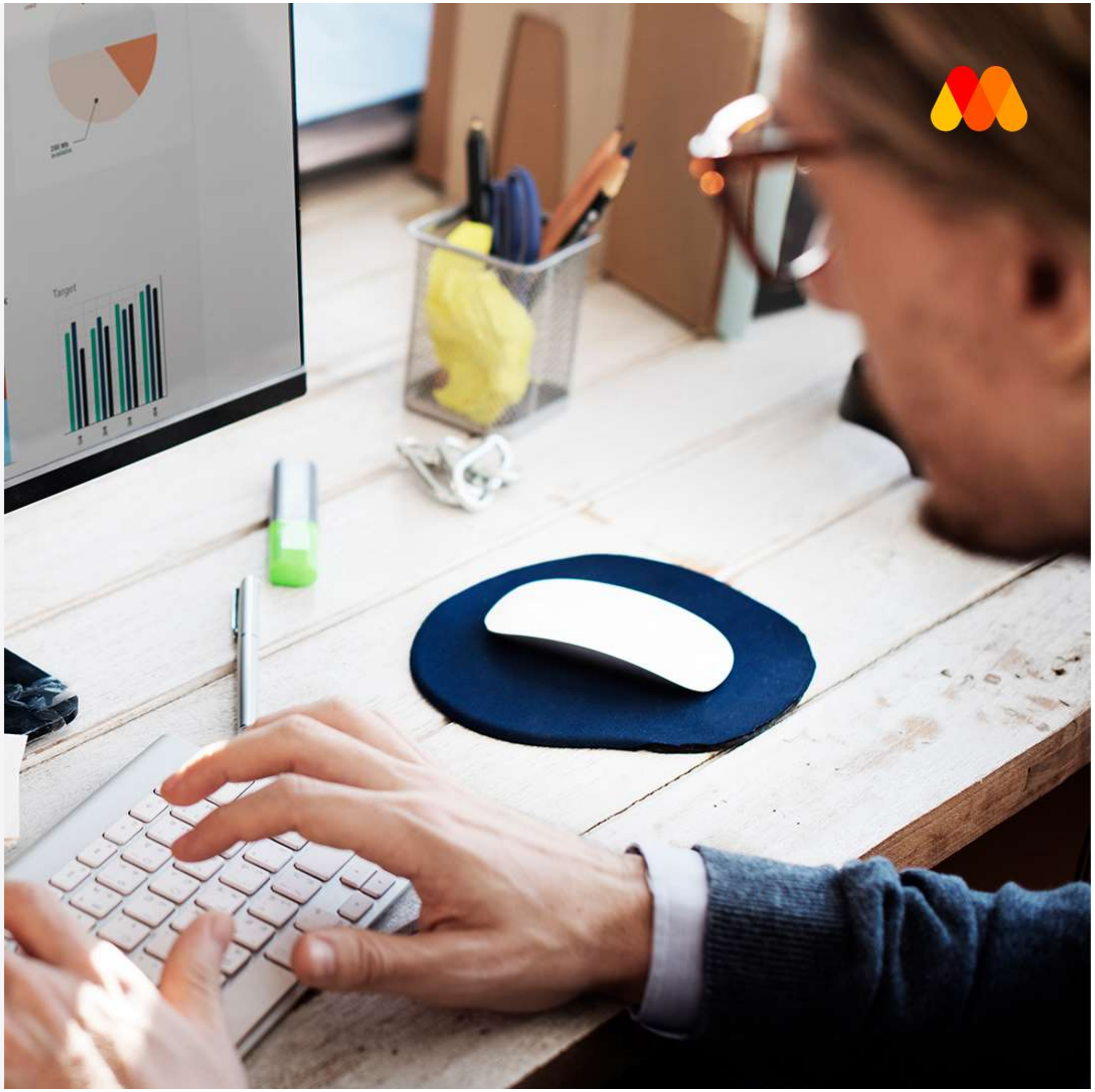
Interpretable Explanations



Interpretable Models are Key in High-Stake Decisions…

- **Healthcare:** cancer detection
- **Banking**: loan lending
- **Crime:** detention and bail
- **Education:** teachers’ promotion / redundancy
- **Recruitment:** interviews

Check out Weapons of Math Destruction by Cathy O’Neil!



.. And in Low-Risk Decisions, too!

TRUST

- Sanity check
- Generalizability
- Fairness

PREDICT

- Foresight of model behaviour

IMPROVE

- Feature and model improvement

Explainable Machine Learning and AI (XML/XAI)

Techniques in Artificial Intelligence [and Machine Learning] that (...) make model predictions easily understood by humans. It contrasts with the concept of the black box in machine learning where even their designers cannot explain why the AI arrived at a specific decision. [1]

[1] https://en.m.wikipedia.org/wiki/Explainable_artificial_intelligence

DALEX

DALEX:
Descriptive mAchine Learning EXplanations

DALEX is a set of tools that help understand how complex models are working

Developed by Przemyslaw Biecek

Github: https://github.com/pbiecek/DALEX

WHAT DO YOU WANT TO UNDERSTAND?

ENTIRE MODEL

How good is the model?

What **metrics** are you interested in?
- model performance audit
- goodness of fit

How does it work?

What **variables** are you interested in?
- All
- A categorical
- A continuous

SINGLE PREDICTION

Variable **attribution**?
(decompose prediction)

Variable **importance**?
(identify key features)

Based on https://raw.githubusercontent.com/pbiecek/DALEX_docs/master/images/Explain.png

WHAT DO YOU WANT TO UNDERSTAND?

ENTIRE MODEL

How good is the model?

What **metrics** are you interested in?
- model performance audit
- goodness of fit

How does it work?

What **variables** are you interested in?
- All
- A categorical
- A continuous

SINGLE PREDICTION

Variable **attribution**?
(decompose prediction)

Variable **importance**?
(identify key features)

DALEX

LIME

Based on https://raw.githubusercontent.com/pbiecek/DALEX_docs/master/images/Explain.png

Before You Start

Regression problem – predict apartment prices in Warsaw, Poland

```
library(DALEX)
library(randomForest)

# train random forest and linear model
str(apartments)
set.seed(519)
apartments_rf_model <- randomForest::randomForest(m2.price ~ ., data = apartments)
predicted_rf <- predict(apartments_rf_model, apartmentsTest)

apartments_lm_model <-  lm(m2.price ~ ., data = apartments)
predicted_lm <- predict(apartments_lm_model, apartmentsTest)
```

Start with the Explainer

Compare model performance

```
# root mean square

sqrt(mean((predicted_rf -apartmentsTest$m2.price)^2))

sqrt(mean((predicted_lm - apartmentsTest$m2.price)^2))
```

Run DALEX explainer

```
explainer_lm <- DALEX::explain(model = apartments_lm_model,

                               data = apartmentsTest[,2:6], y = apartmentsTest$m2.price)

explainer_rf <- DALEX::explain(model = apartments_rf_model,

                               data = apartmentsTest[,2:6], y = apartmentsTest$m2.price)
```

Start with the Explainer

DALEX explainer attaches relevant meta data to the algorithms and unifies model interfacing

```
> explainer_lm
Model label:  train
Model class:  train
Data head  :
     construction.year surface floor no.rooms    district
1001              1976     131     3        5 Srodmiescie
1002              1978     112     9        4     Mokotow
```

WHAT DO YOU WANT TO UNDERSTAND?

ENTIRE MODEL

How good is the model?

What **metrics** are you interested in?
- model performance audit
- goodness of fit

How does it work?

What **variables** are you interested in?
- All
- A categorical
- A continuous

SINGLE PREDICTION

Variable **attribution**?
(decompose prediction)

Variable **importance**?
(identify key features)

Based on https://raw.githubusercontent.com/pbiecek/DALEX_docs/master/images/Explain.png

WHAT DO YOU WANT TO UNDERSTAND?

ENTIRE MODEL

How good is the model?

What **metrics** are you interested in?

- **model performance audit**
- goodness of fit

How does it work?

What **variables** are you interested in?

- All
- A categorical
- A continuous

SINGLE PREDICTION

Variable **attribution**?
(decompose prediction)

Variable **importance**?
(identify key features)

Based on https://raw.githubusercontent.com/pbiecek/DALEX_docs/master/images/Explain.png

How Good is the Model?

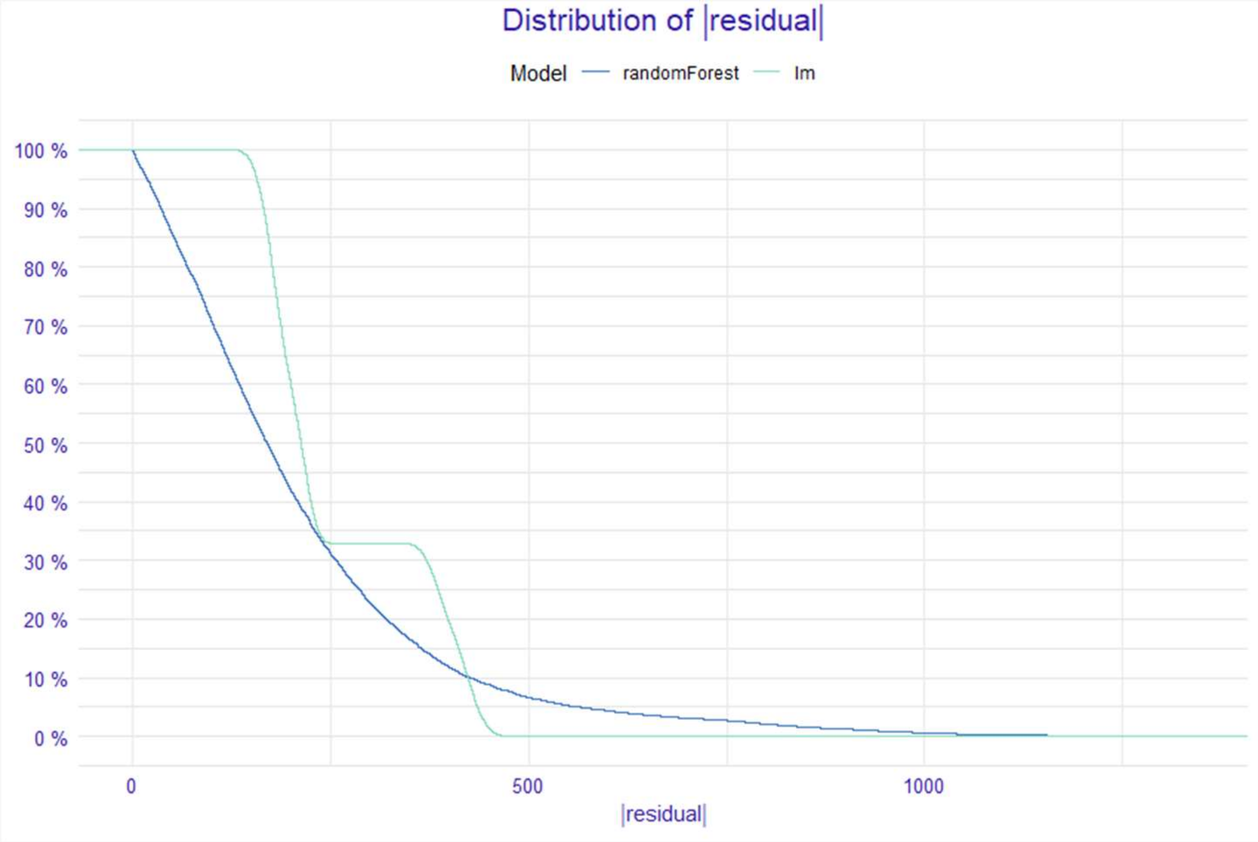
Explainer for model performance gives more information in a consistent form

The function `model_performance()` calculates predictions and residuals for validation data `apartments_test`

```
mp_lm <- model_performance(explainer_lm)

mp_rf <- model_performance(explainer_rf)
```

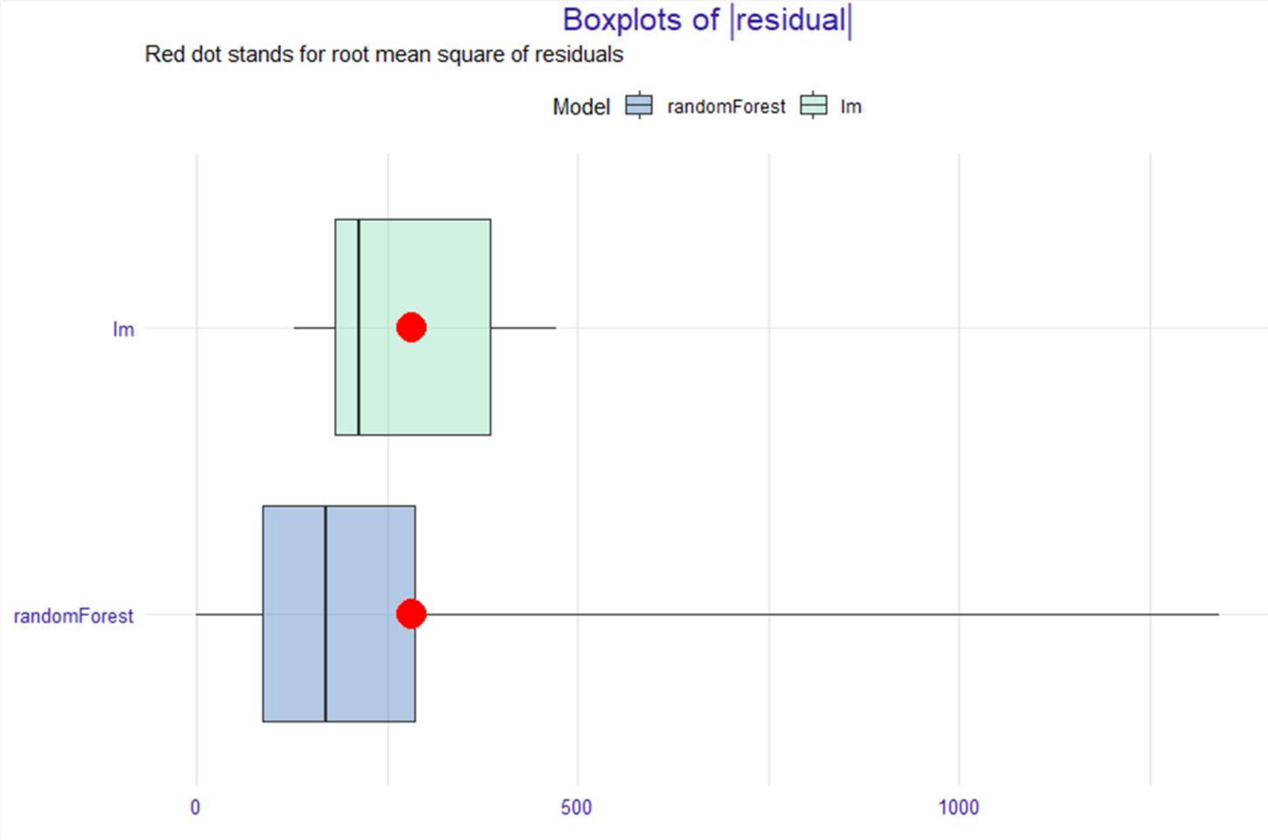
How Good is the Model?

```
# plot model performance

plot(mp_lm, mp_rf)
```



How Good is the Model?

```
# plot model performance

plot(mp_lm, mp_rf,

     geom = "boxplot")
```



WHAT DO YOU WANT TO UNDERSTAND?

ENTIRE MODEL

How good is the model?

What **metrics** are you interested in?
- model performance audit
- **goodness of fit**

How does it work?

What **variables** are you interested in?
- All
- A categorical
- A continuous

SINGLE PREDICTION

Variable **attribution**?
(decompose prediction)

Variable **importance**?
(identify key features)

Based on https://raw.githubusercontent.com/pbiecek/DALEX_docs/master/images/Explain.png

How Good is the Model?

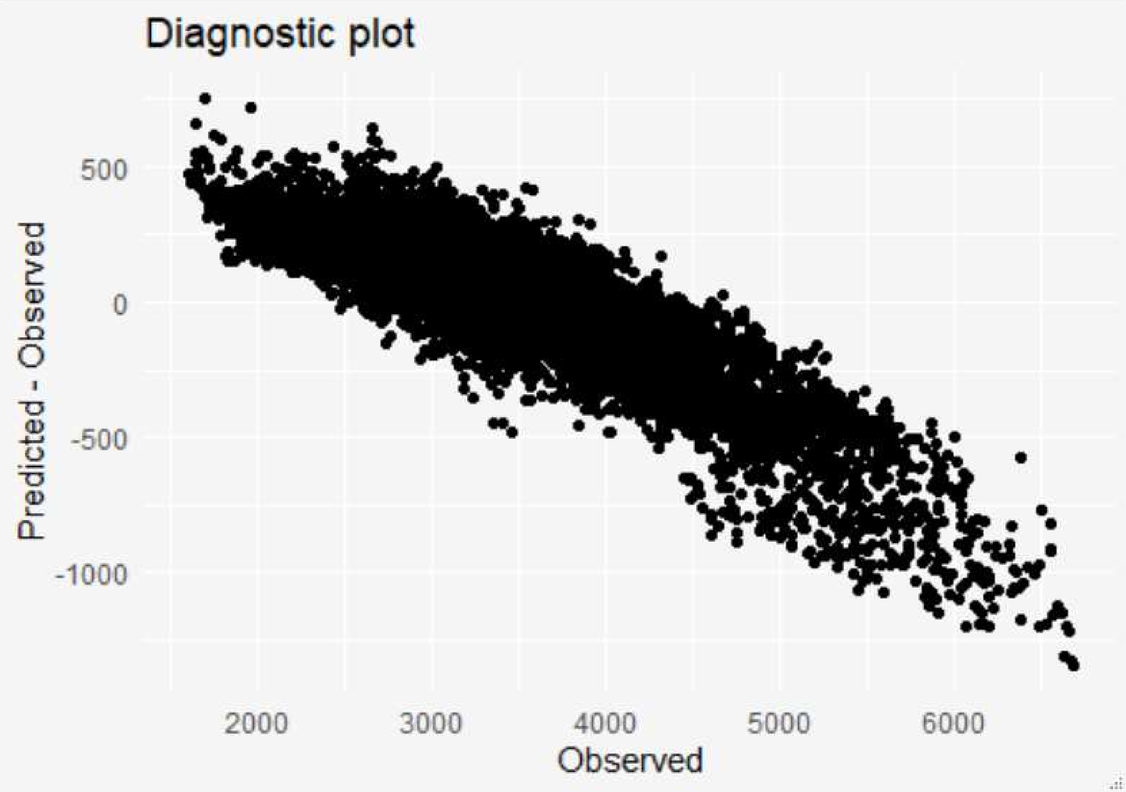
Explainer for model performance gives more information in a consistent form

The function `model_performance()` calculates predictions and residuals for validation data `apartments_test`

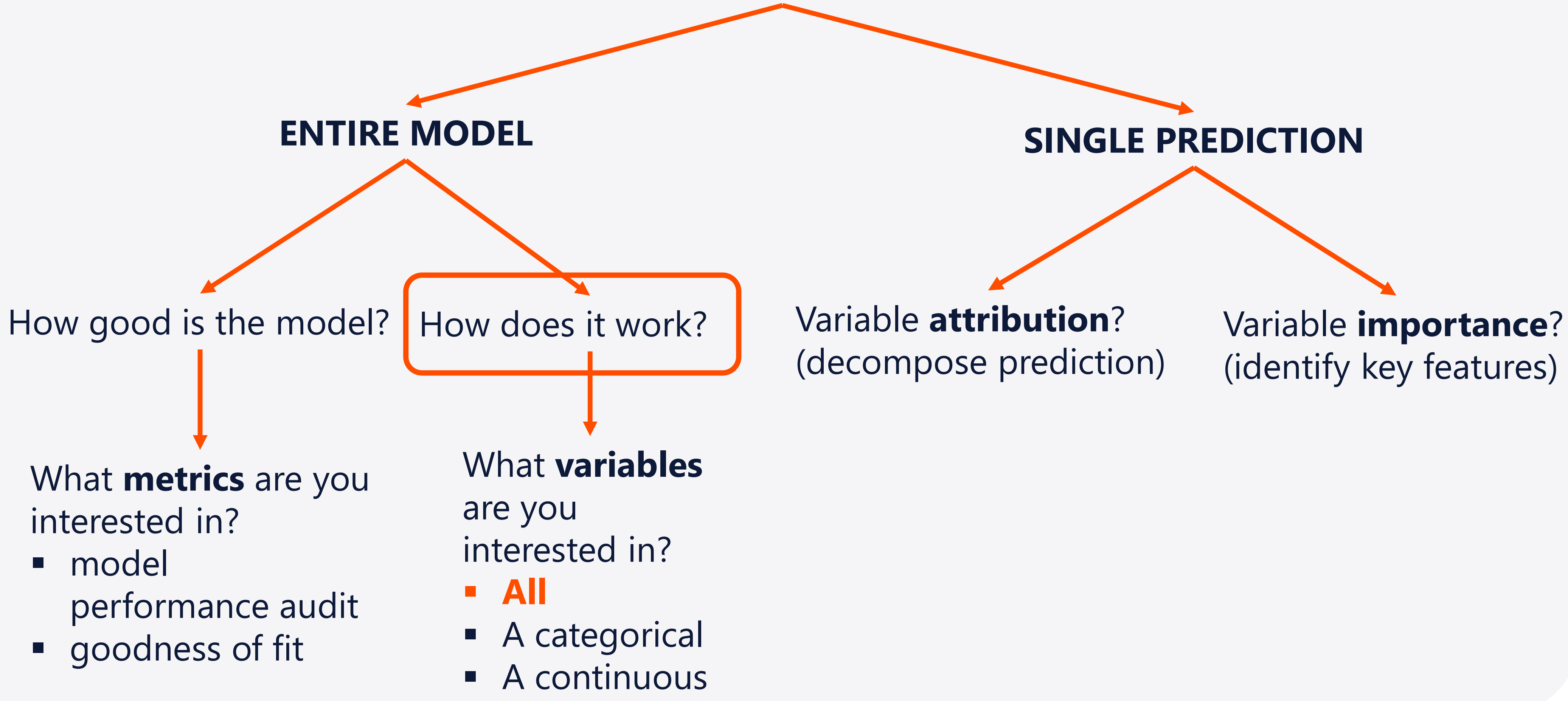
```
mp_rf <- model_performance(explainer_rf)

mp_rf$observed

mp_rf$predicted

mp_rf$diff # predicted - observed
```

How Good is the Model?

```
ggplot(mp_rf, aes(observed, diff) ) +
  stat_density_2d(
    aes(fill = ..level..),
    geom = "polygon",
    colour = "white") +
  scale_fill_gradient(name = "density") +
  xlab("Observed") +
  ylab("Predicted - Observed") +
  ggtitle("Diagnostic plot") +
  theme_mi2()
```



WHAT DO YOU WANT TO UNDERSTAND?



Based on https://raw.githubusercontent.com/pbiecek/DALEX_docs/master/images/Explain.png

How Does the Model Work?

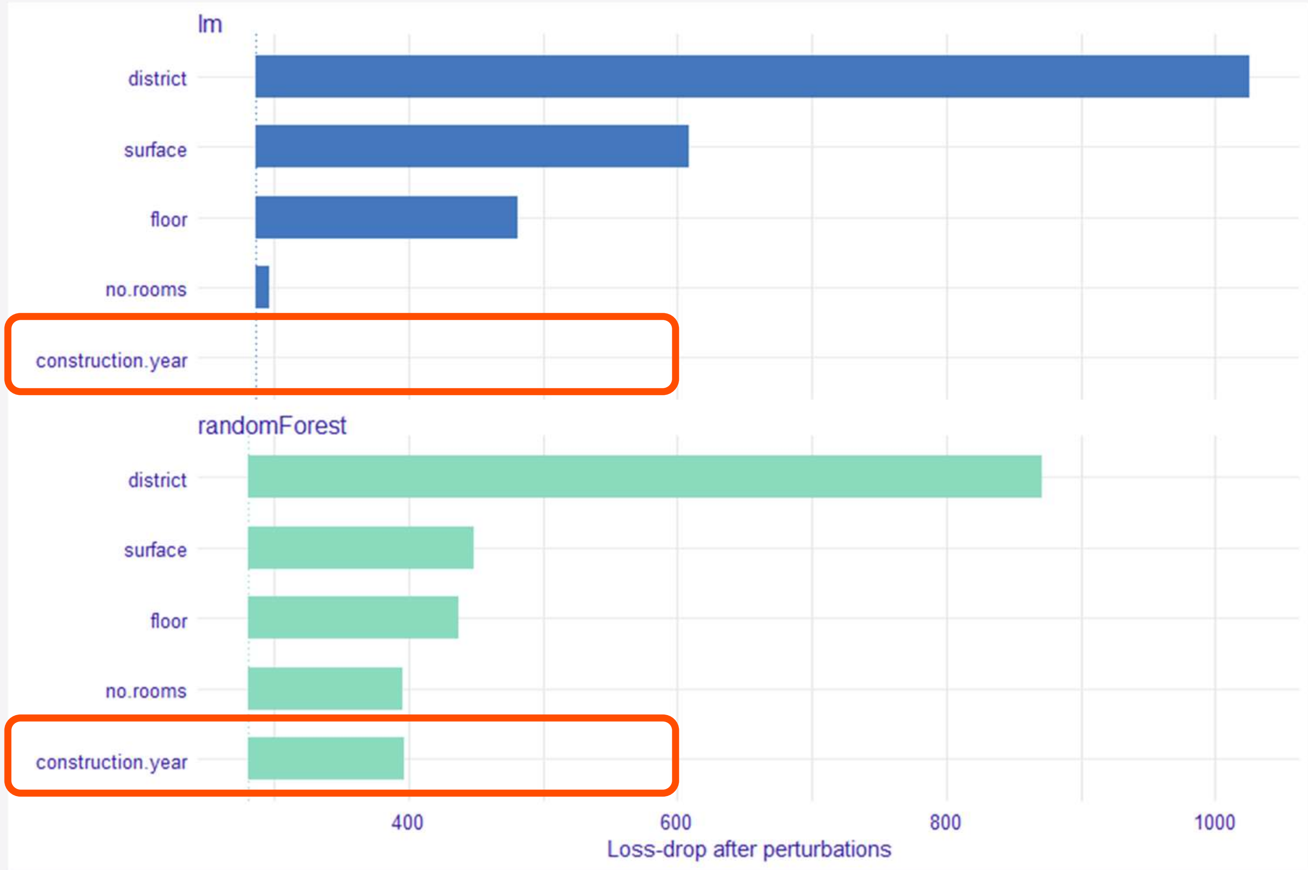
Variable Importance

- Variable importance helps us validate the model and increase our understanding of the domain.
- The function `variable_importance()` provides model agnostic variable importance (as opposed to model-specific).

```
vi_rf <- variable_importance(explainer_rf, loss_function = loss_root_mean_square)

vi_lm <- variable_importance(explainer_lm, loss_function = loss_root_mean_square)
```

How Does the Model Work?

```
plot(vi_rf, vi_lm)
```



WHAT DO YOU WANT TO UNDERSTAND?

ENTIRE MODEL

How good is the model?

What **metrics** are you interested in?
- model performance audit
- goodness of fit

How does it work?

What **variables** are you interested in?
- All
- A categorical
- **A continuous**

SINGLE PREDICTION

Variable **attribution**?
(decompose prediction)

Variable **importance**?
(identify key features)

Based on https://raw.githubusercontent.com/pbiecek/DALEX_docs/master/images/Explain.png

How Does the Model Work?

A single continuous variable

Partial Dependence Plots (PDP) show the expected output conditional on the selected variable.

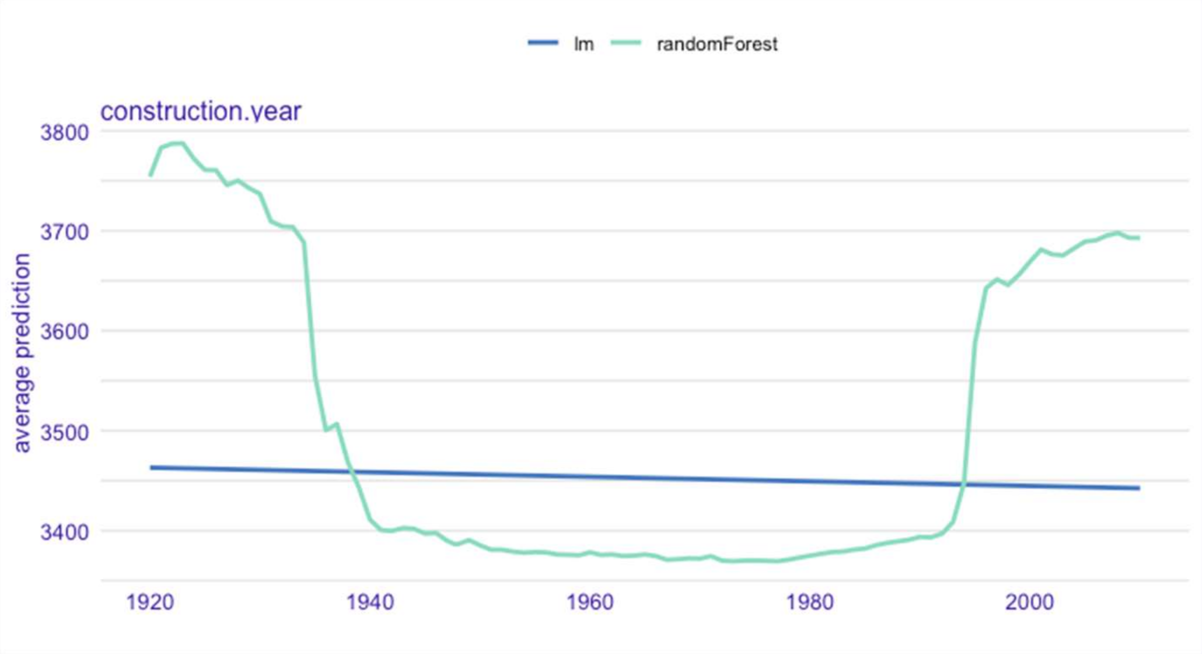
```
pdp_rf <- ingredients::partial_dependency(explainer_rf,

                 variables = "construction.year", variable_type = "numerical")

pdp_lm <- ingredients::partial_dependency(explainer_lm,

                 variables = "construction.year", variable_type = "numerical")
```

How Does the Model Work?

Partial Dependence Plots

```
plot(pdp_rf, pdp_lm)
```

The linear model is unable to capture a non-linear relationship between construction year and apartment price.



WHAT DO YOU WANT TO UNDERSTAND?

ENTIRE MODEL

How good is the model?

What **metrics** are you interested in?

- model performance audit
- goodness of fit

How does it work?

What **variables** are you interested in?

- All
- **A categorical**
- A continuous

SINGLE PREDICTION

Variable **attribution**?
(decompose prediction)

Variable **importance**?
(identify key features)

Based on https://raw.githubusercontent.com/pbiecek/DALEX_docs/master/images/Explain.png

How Does the Model Work?

A single categorical variable

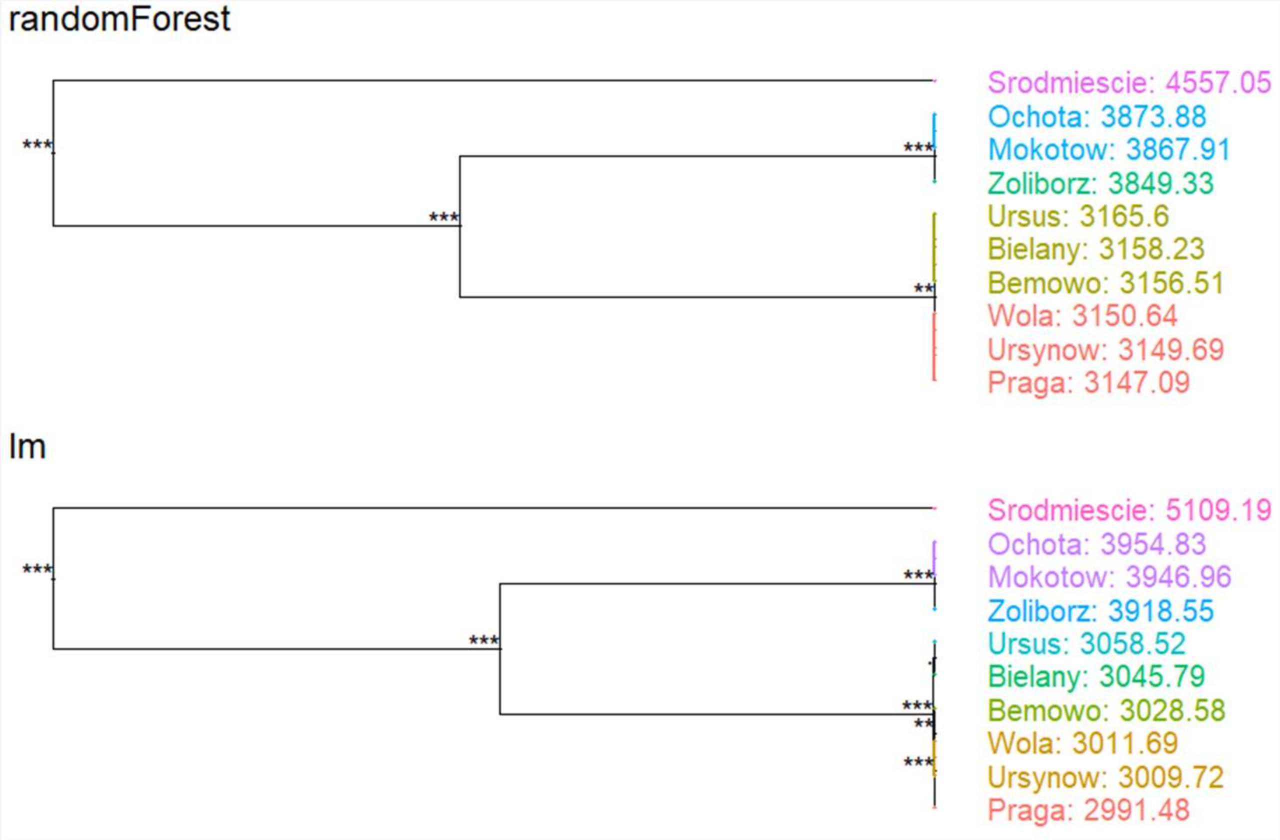
```
svd_rf  <- single_variable(explainer_rf, variable = "district", type = "factor")

svd_lm  <- single_variable(explainer_lm, variable =  "district ", type = "factor")
```

How Does the Model Work?

Merging Path Plots

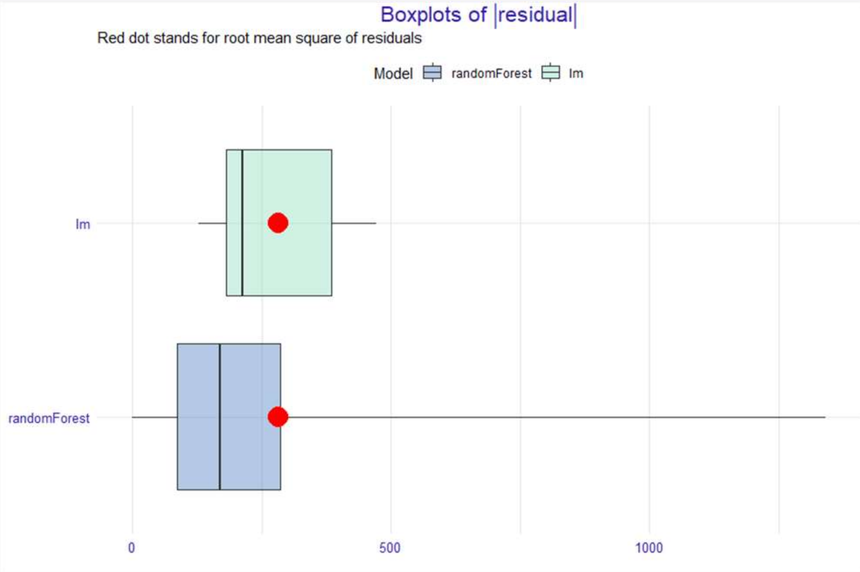
```
plot(svd_rf, svd_lm)
```

Can you identify the three distinct clusters?



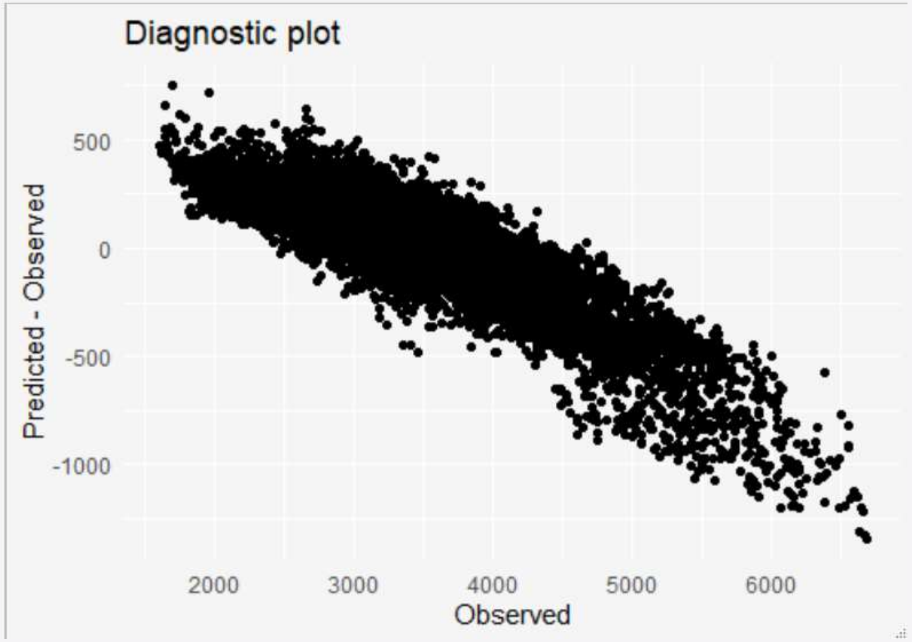
In summary

- Both models have a very similar accuracy



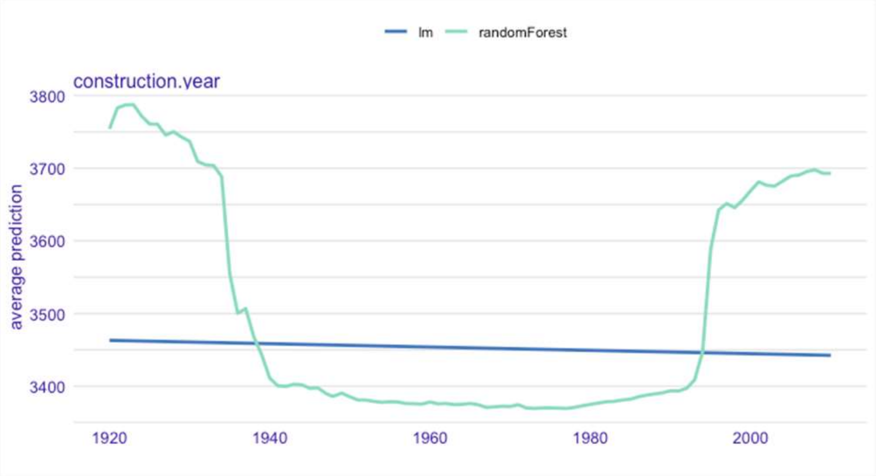
In summary

- Both models have a very similar accuracy
- Random forest is more accurate for lower- and mid-value flats, but not for the high-value ones



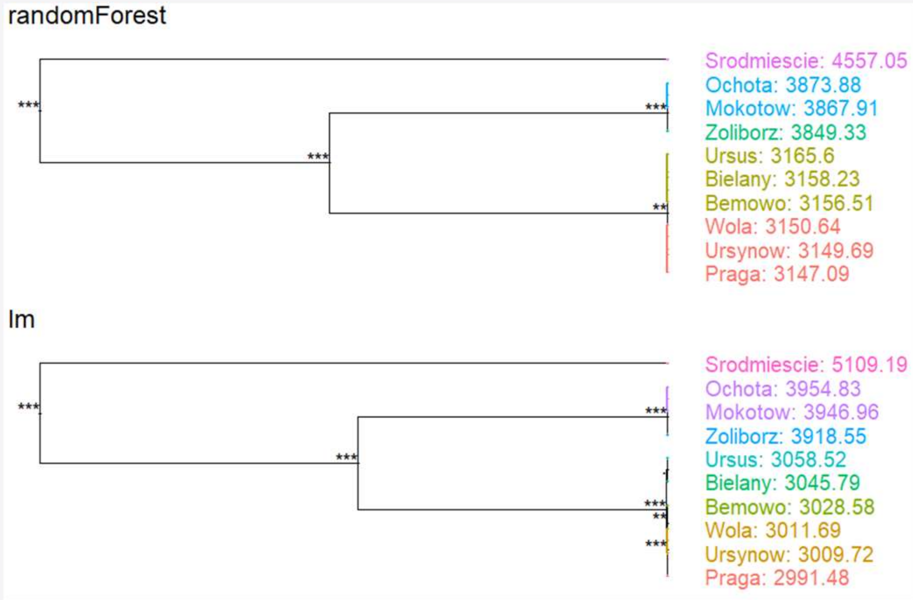
In summary

- Both models have a very similar accuracy
- RF is more accurate for lower- and mid-value flats, but not for the high-value ones
- RF correctly captured a non-linear relationships
- RF tends to under-value most expensive flats by not attributing location enough



In summary

- Both models have a very similar accuracy
- Random forest is more accurate for lower- and mid-value flats, but not for the high-value ones
- Random forest correctly captured a non-linear relationship between the construction year and the flat price
- Still, random forest tends to under-value most expensive flats by not attributing location enough



Material / References

- Materials at

https://github.com/MangoTheCat/explainable-machine-learning-workshop

- DALEX: https://pbiecek.github.io/DALEX/
- LIME: Ribeiro et al. “Why Should I Trust You? Explaining the Predictions of Any Classifier” (ACM SIGKDD, 2016)
  - Python: https://github.com/marcotcr/lime
  - R: https://github.com/thomasp85/lime
- SHAP: Lundberg, Lee (2017). “A Unified Approach to Interpreting Model Predictions.” (NeurIPS, 2017)
  - ShapleyR: https://github.com/redichh/ShapleyR
  - iml: https://github.com/christophM/iml
  - shapper: https://github.com/ModelOriented/shapper