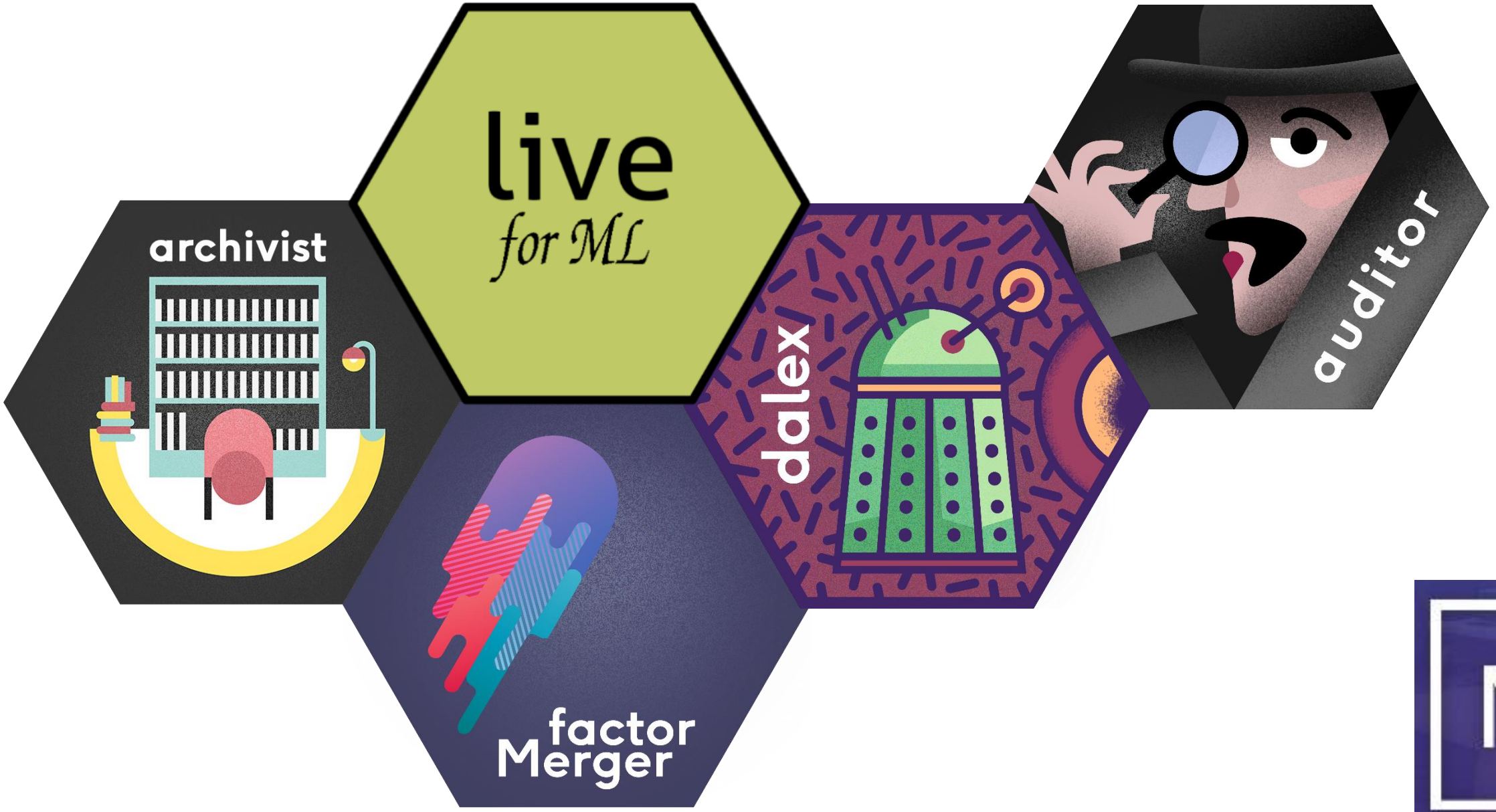
auditor: an R package and methodology for validation of any statistical model

Alicja Gosiewska

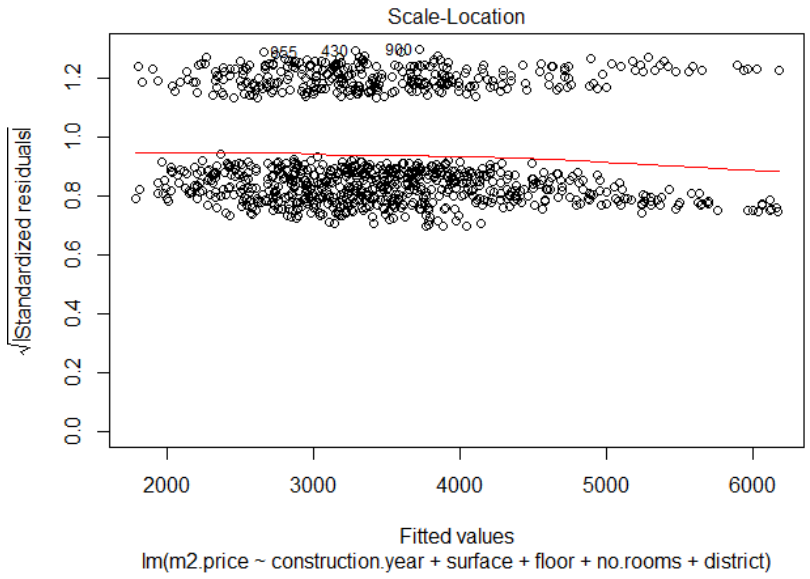
MI2 DataLab

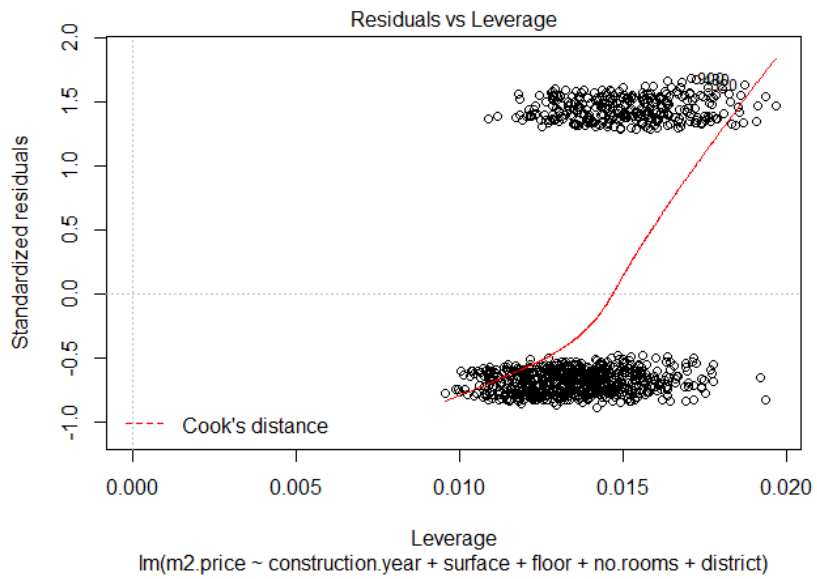
Faculty of Mathematics and Information Science,
Warsaw University of Technology

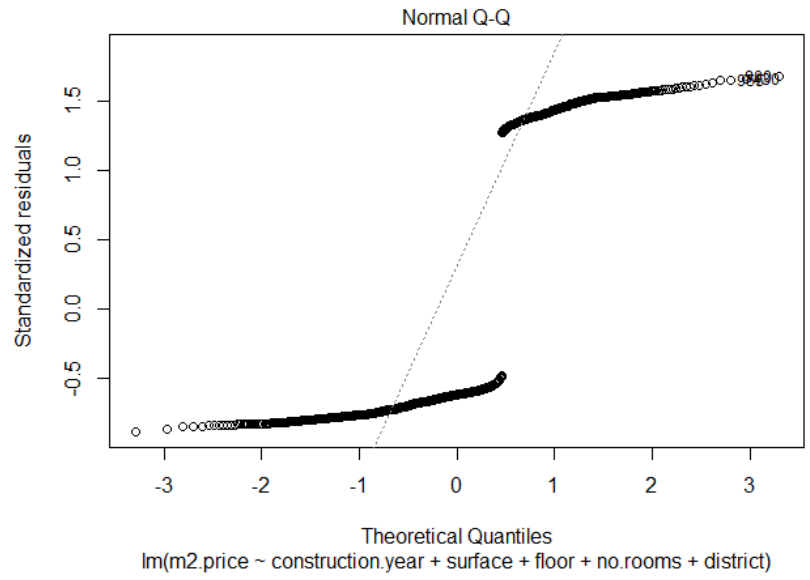
WhyR? 2018
Wrocław, 04.07.2018

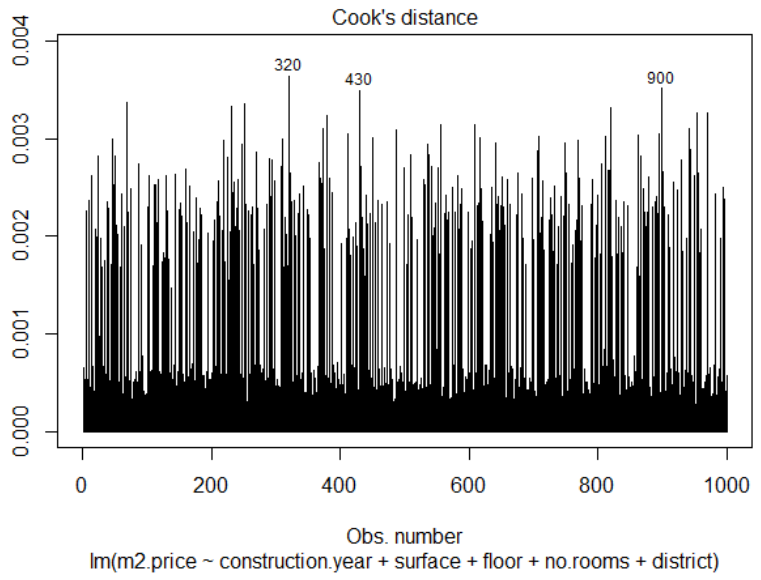

archivist
live
for ML
dalex
auditor
factor
Merger

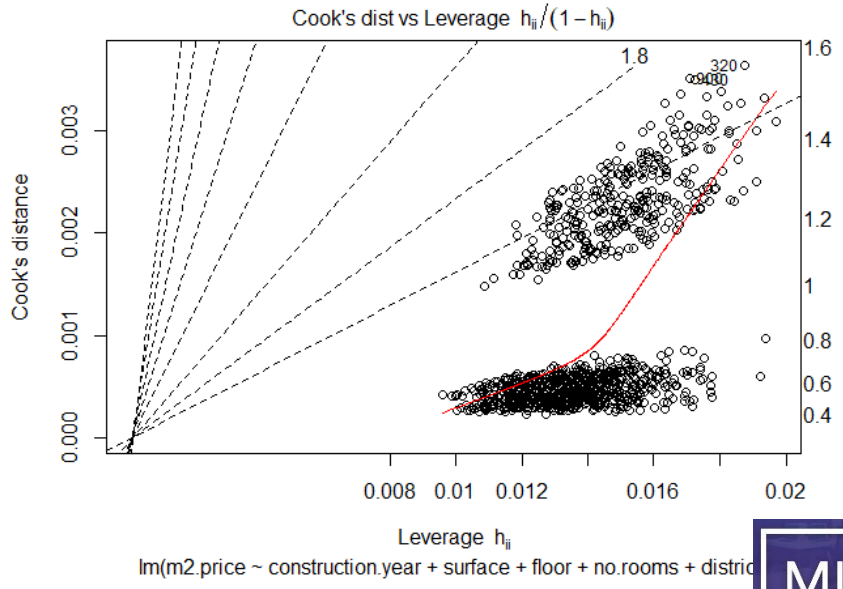
Motivation – linear models











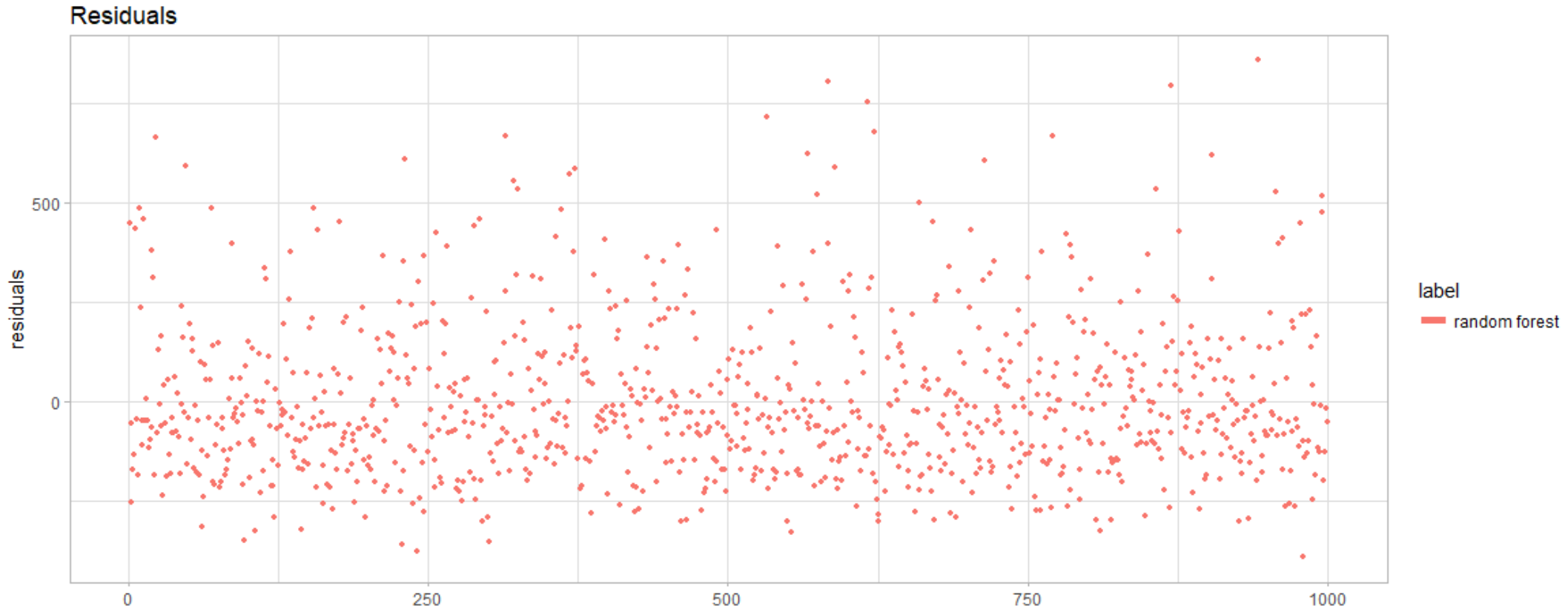
auditor

- Model-agnostic
- Model comparisons
- Consistency

```
model %>% audit() %>% plot(type = )
```

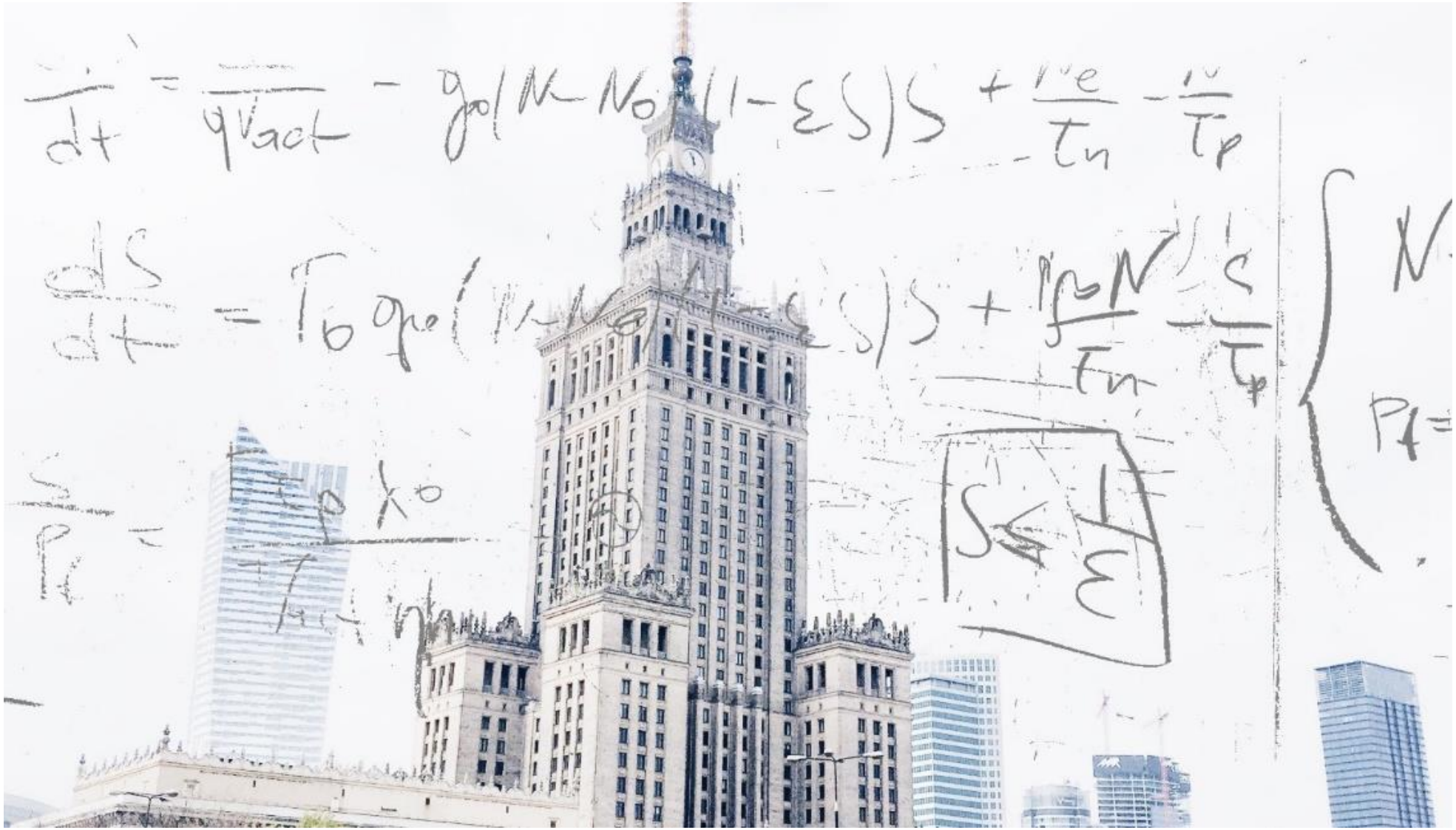
model %>% audit() %>% plot(type =)


```
library(auditor)

apartments_rf_model <- randomForest(m2.price ~ ., data = apartments)

au_rf <- audit(apartments_rf_model, data = apartmentsTets, y = apartmentsTest$m2.price)

plot(au_rf, type = "Residual")
```



Do we really need model audit?

A story about two models!



Do we really need model audit?

A story about two models!

m2.price	construction.year	surface	floor	no.rooms	district
5897	1953	25	3	1	Srodmiescie
1818	1992	143	9	5	Bielany
3643	1937	56	1	2	Praga
3517	1995	93	7	3	Ochota
3013	1992	144	6	5	Mokotow

MI

Do we really need model audit?

A story about two models!

m2.price	construction.year	surface	floor	no.rooms	district
5897	1953	25	3	1	Srodmiescie
1818	1992	143	9	5	Bielany
3643	1937	56	1	2	Praga
3517	1995	93	7	3	Ochota
3013	1992	144	6	5	Mokotow

```
apartments_lm_model <- lm(m2.price ~ construction.year + surface + floor +
                          no.rooms + district, data = apartments)


apartments_rf_model <- randomForest(m2.price ~ construction.year + surface + floor +
                        no.rooms + district, data = apartments)
```

Accuracy is not enough!

Root Mean Square error

```
predicted_mi2_lm <- predict(apartments_lm_model, apartmentsTest)
sqrt(mean((predicted_mi2_lm - apartmentsTest$m2.price)^2))
```

```
[1] 283.0865
```

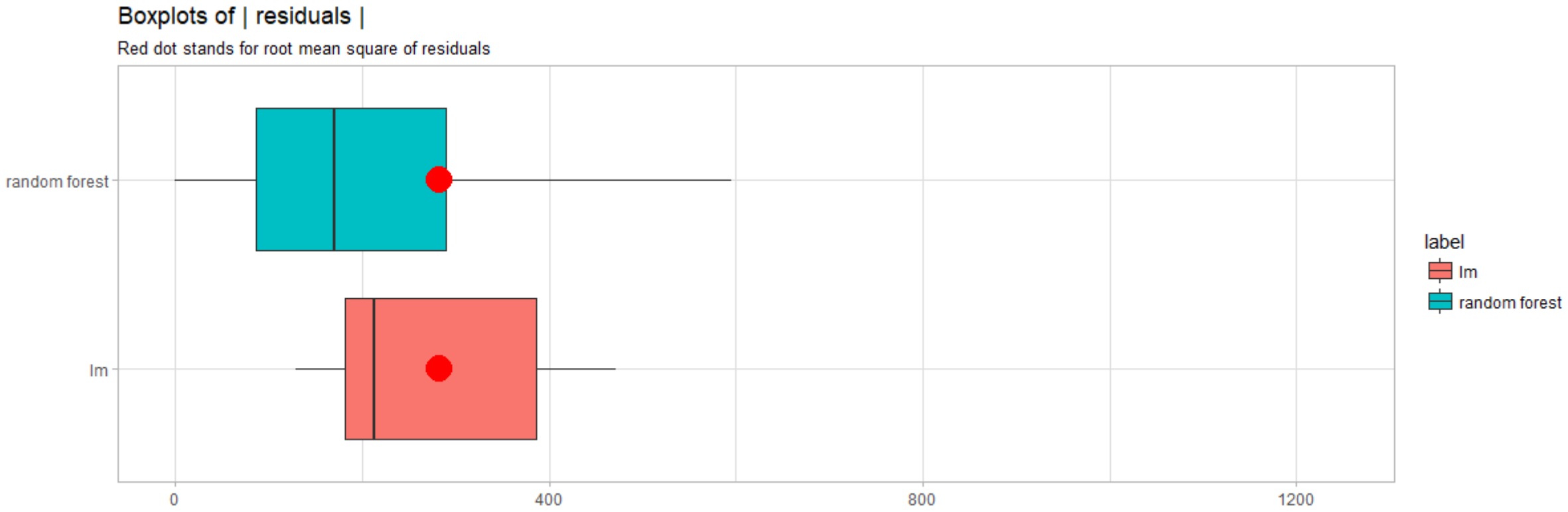
```
predicted_mi2_rf <- predict(apartments_rf_model, apartmentsTest)
sqrt(mean((predicted_mi2_rf - apartmentsTest$m2.price)^2))
```

```
[1] 283.1138
```

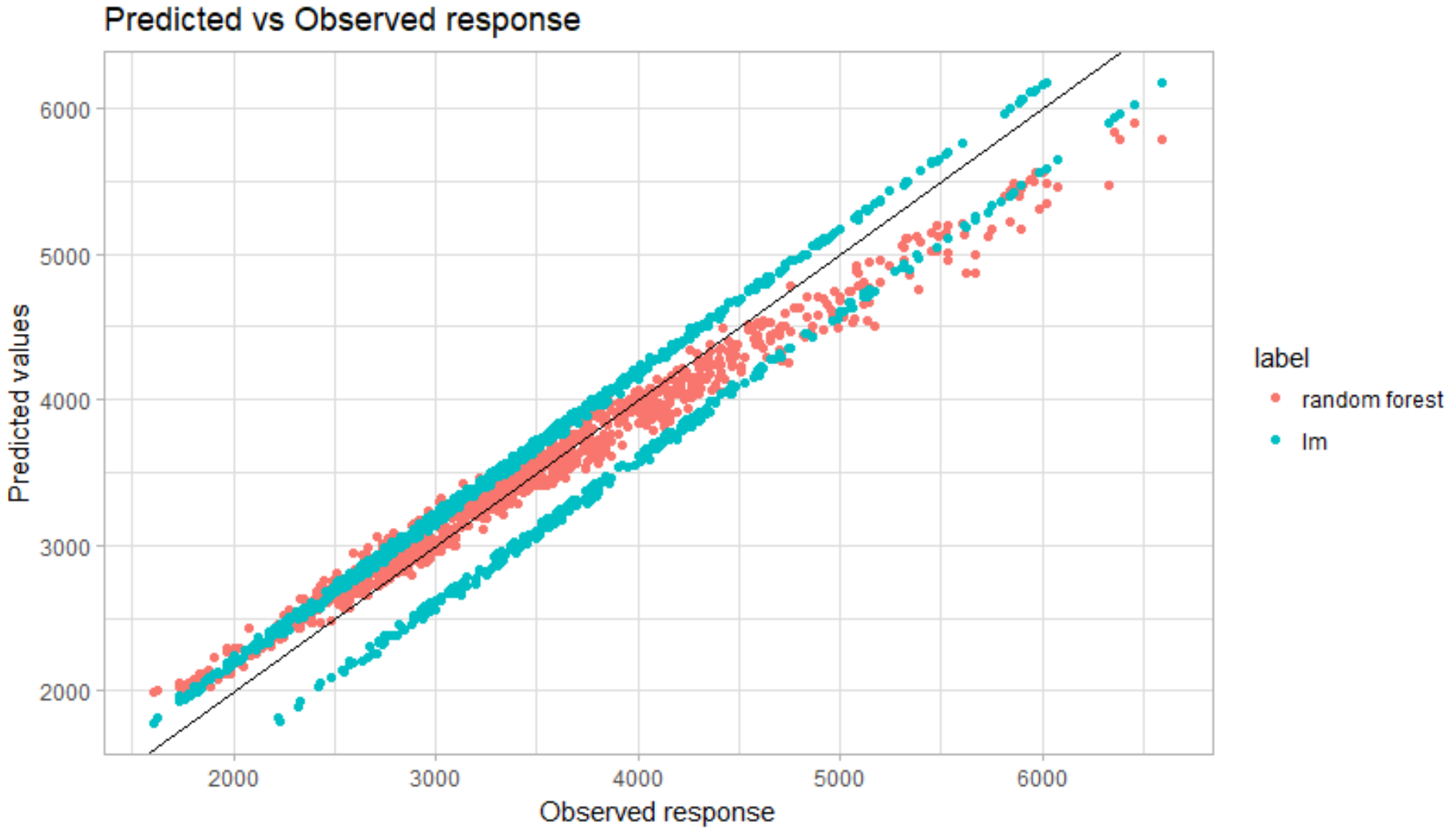
MI

Model Performance

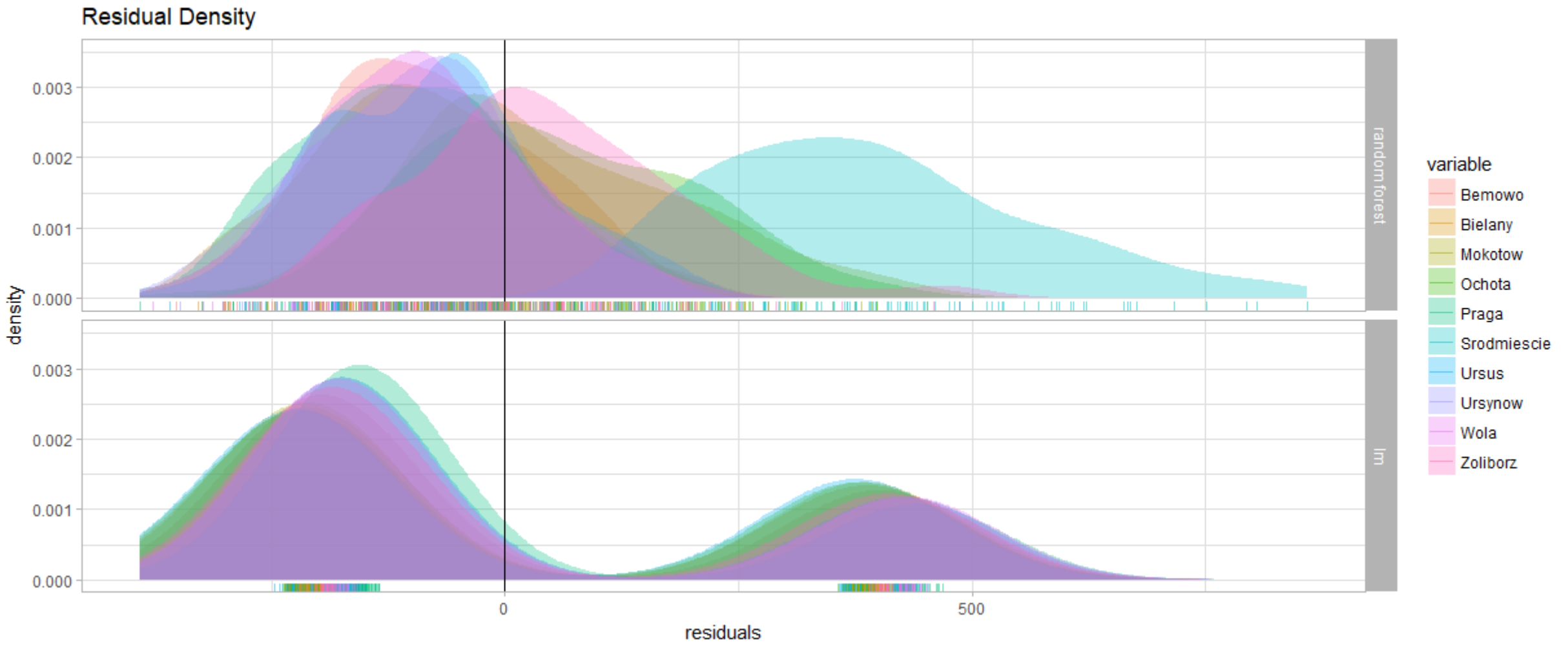

```
au_rf <- audit(apartments_rf_model, data = apartments, y = apartments$m2.price)
au_lm <- audit(apartments_lm_model, data = apartments, y = apartments$m2.price)

plot(au_rf, au_lm, type = "ResidualBoxplot")
```



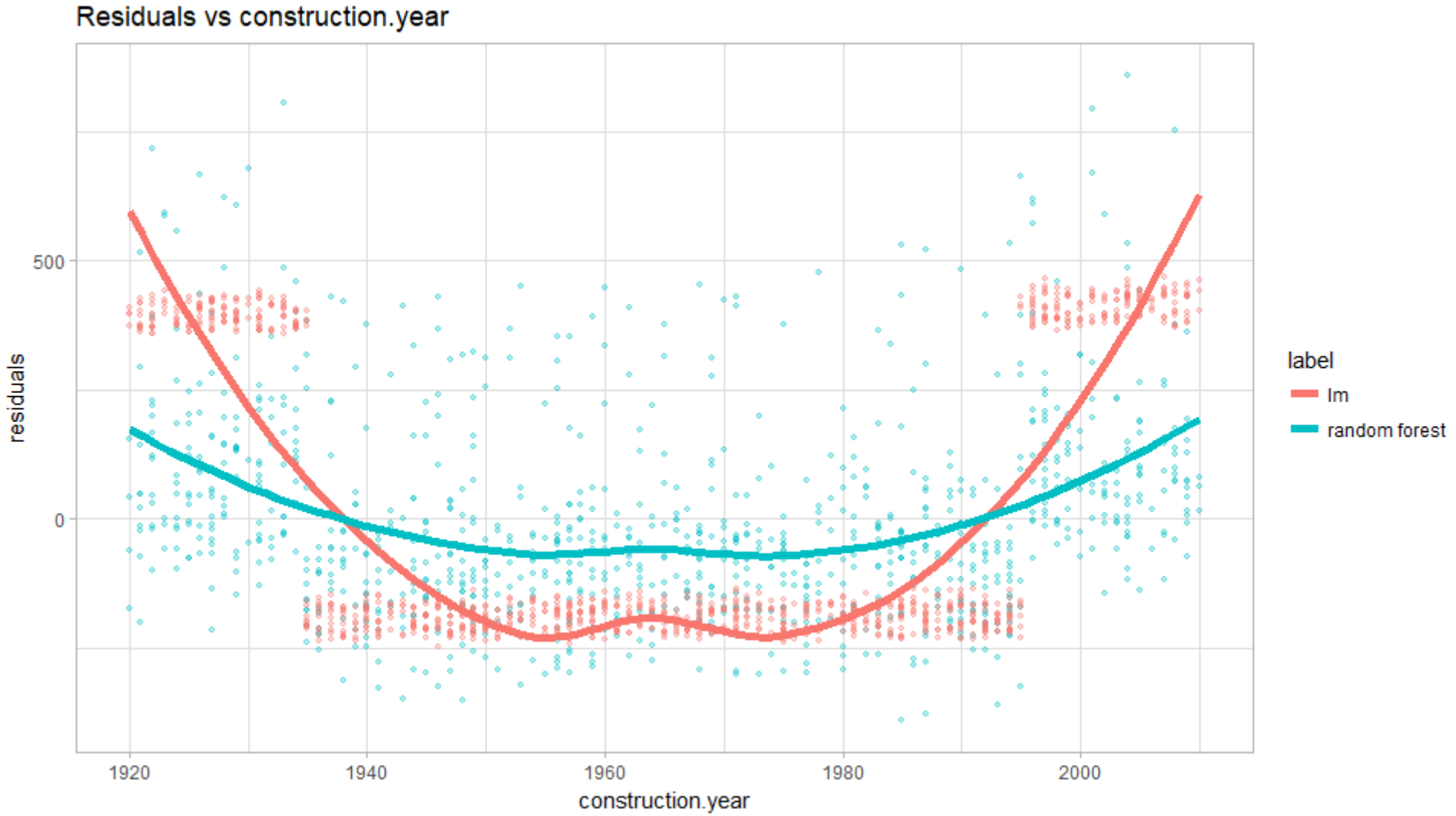
```
plot(au_rf, au_lm, type = "Performance")
```



```
plot(au_rf, au_lm, type = "ResidualDensity", variable = "district")
```

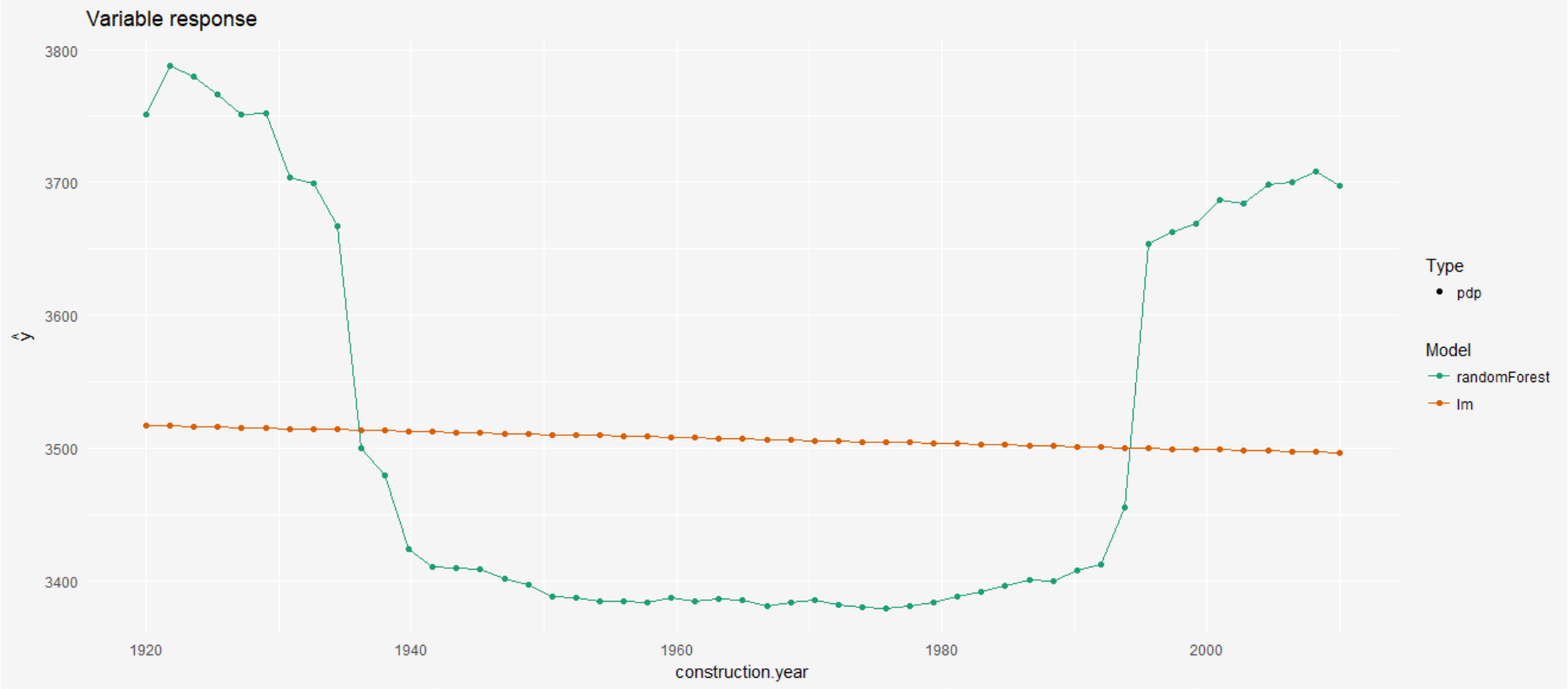


```
plot(au_rf, au_lm, type = "Residual", variable = "construction.year")
```



Partial Dependence Plot

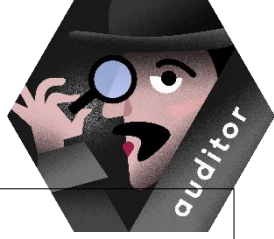
```
library(DALEX)

exp_rf <- explain(audit(apartments_rf_model, data = apartments, y = apartments$m2.price)
exp_lm <- explain(audit(apartments_lm_model, data = apartments, y = apartments$m2.price)


plot_model_performance(exp_rf, exp_lm, type = "pdp", variable = "construction.year")
```



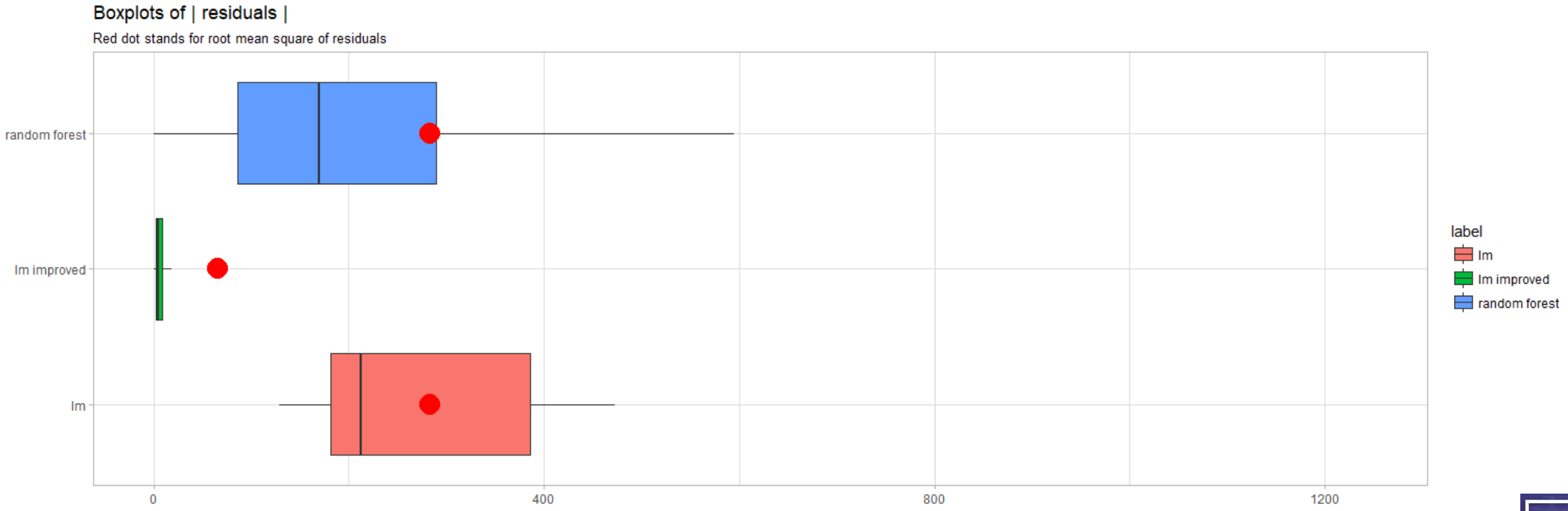
Let's sum up!

- Both models have similiar performance.
- Random forest model has smaller residuals than the linear model but there is a small fraction of very large residuals.
- Random forest model under-predicts expensive apartments. It is not a model that we would like to employ.
- The relation between construction year and the price of square meter is non linear.

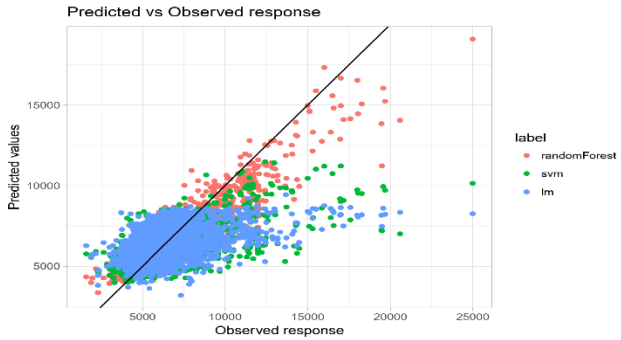
Let's improve our model!

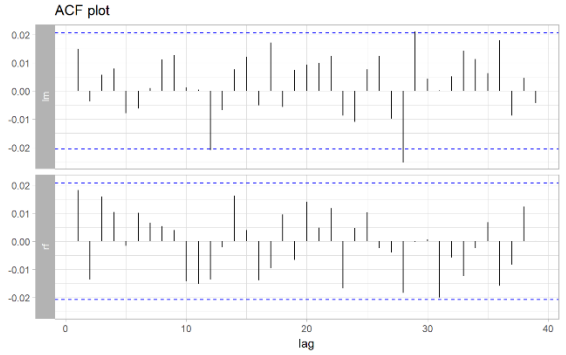


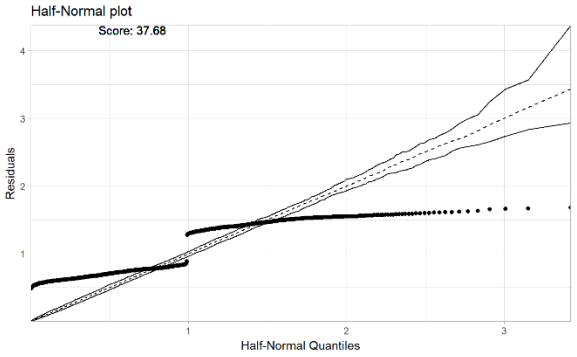

```
apartments_imp_model <- lm(m2.price ~ I(construction.year < 1935 | construction.year > 1995) +
                                      surface + floor + no.rooms + district, data = apartments)

au_imp <- audit(apartments_imp_model, label = "lm improved", data = apartments, y = apartments$m2.price)

plot(au_lm, au_rf, au_imp, type = "ResidualBoxplot")
```

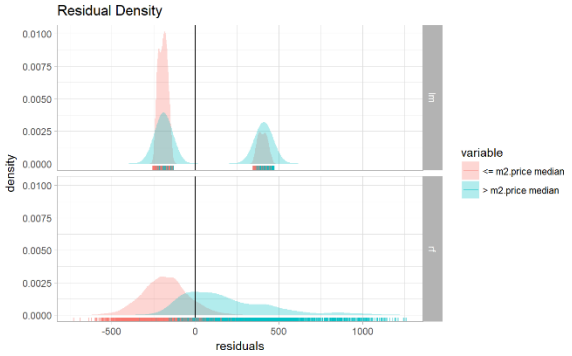


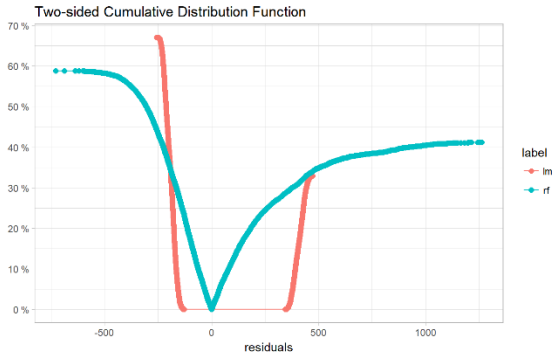
What's more?

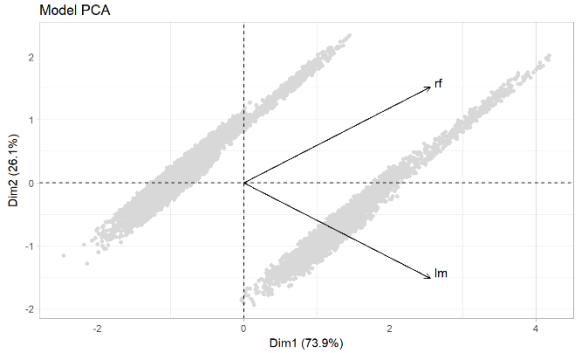


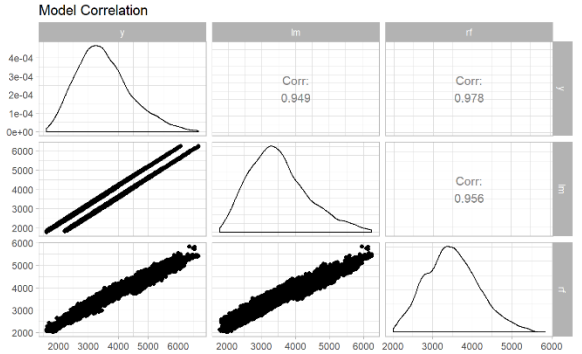


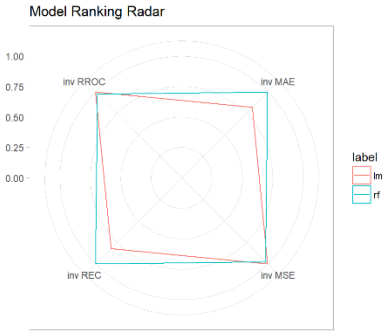












GitHub https://github.com/mi2-warsaw/auditor

Thank you!

auditor

GitHub https://github.com/mi2-warsaw/auditor

Alicja Gosiewska

http://gosiewska.com

alicjagosiewska@gmail.com

MI2 DataLab

http://mi2.mini.pw.edu.pl

auditor package was financially supported by the 'NCN Opus grant 2016/21/B/ST6/02176'.