Predicting Real-Estate Pricing

Data Analysis for Economic Policy 4 | CENTRAL EUROPEAN UNIVERSITY

Luca Keresztesi

How should we price flats in new projects?

- **Advertised price of flats available in the 13th district of Budapest, from August 2016.**
- **We predict future prices based on past data (6,747 observations) by using regression and RandomForest models.**
- **Target variable is price per square meter (*psqm*). We originally had 47 predictors such as number of floors, air conditioner, parking options, heating type, condition, etc.**
- **Probably due to repeated advertisment, we had to filter out duplicates by only keeping observations with no or little duplication. After this step 6,005 observations were kept.**
  - Flagged as maybe duplicate if the price, sqm, floor and heating are the same: 742
  - Flagged as duplicate if all the above match and also price, lift, airconditioner dummy, balcony dummy, balcony size, concrete blockflat dummy, orientation, view and parking are the same: 1,459
  - No duplicates: 4,546

Objective

Exploratory Data Analysis

Feature Engineering

Modeling Summary

RandomForest

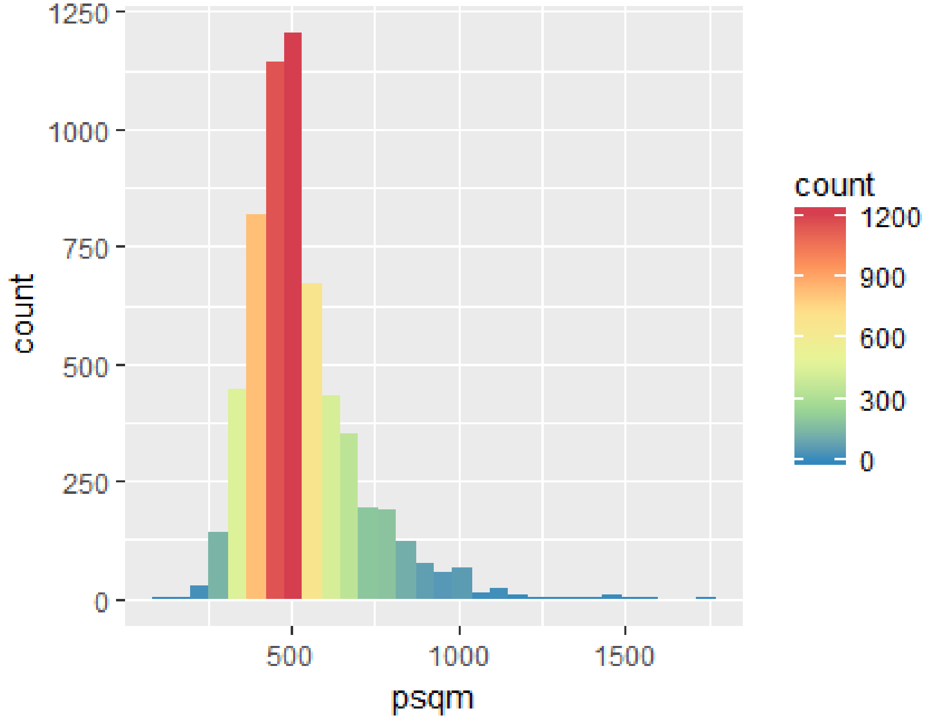
Technical Appendix

Exploratory data analysis

Descriptive Statistics for the numeric variables in the prediction

Statistic	Type	N	Mean	St. Dev.	Min	Max
floor	Numeric	6,005	2.800	2.0002	0	11
balcony	Numeric	6,005	5.280	11.455	0	99
P (million HUF)	Numeric	6,005	34.28	22.19	5.88	419.00
nrooms	Numeric	6,002	1.899	0.930	1	10
nhalfrooms	Numeric	6,005	0.583	0.778	0	14
Psqm (thousand HUF)	Numeric	6,005	529.76	165.00	120.00	1,750.00
hasbalcony	Binary	6,005	0.493	0.500	0	1
concrete_blockflat_d	Binary	6,005	0.105	0.306	0	1
floor2	Numeric	6,005	2.766	1.932	-1	8
aircond_d	Binary	6,005	0.140	0.347	0	1
lift_d	Binary	6,005	0.675	0.468	0	1
ln_psqm	Numeric	6,005	6.230	0.284	4.787	7.467
new_flat	Binary	6,005	0.421	0.494	0	1

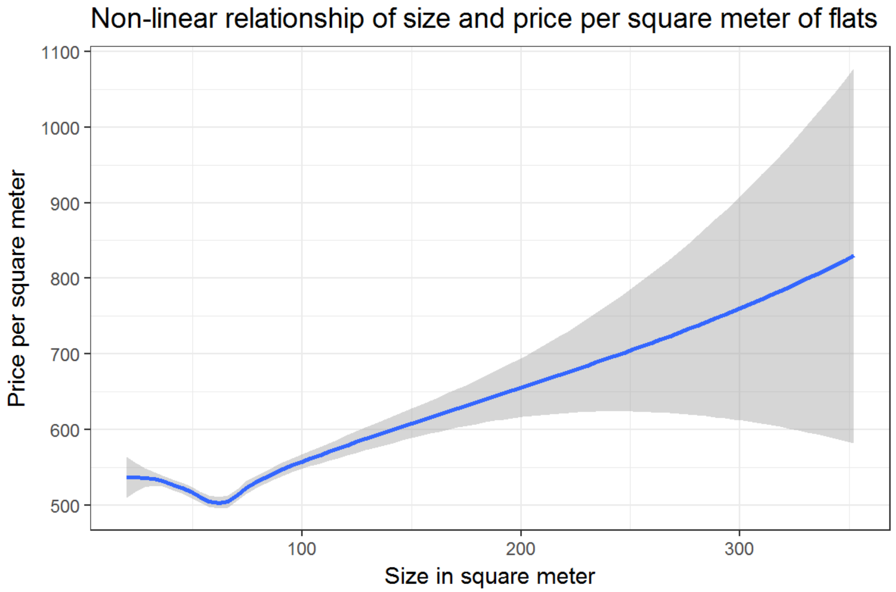
Distribution of Price per square meter

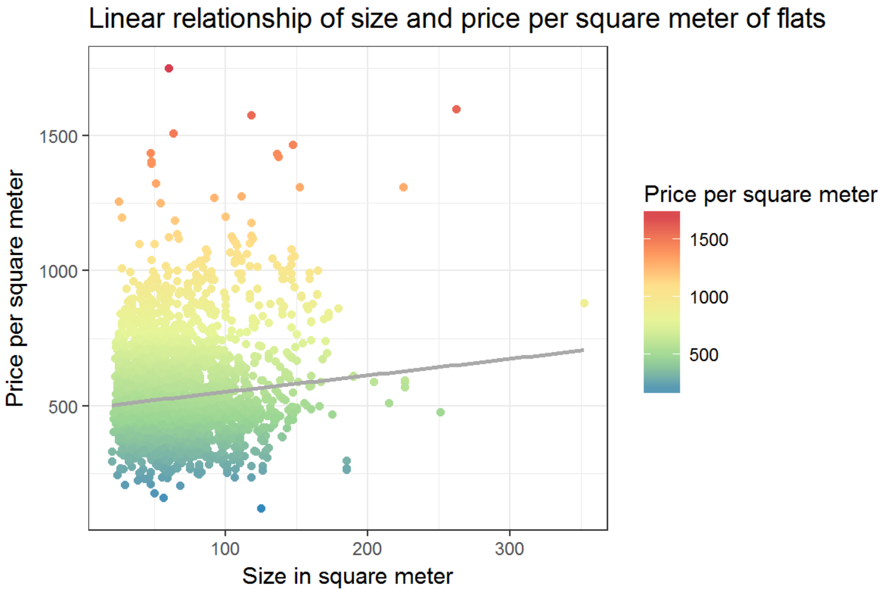


Feature engineering – Size in square meter

Based on the loess regression of Price per square meter on Size in square meter, until approx 60 m² the average relationship is negative, while above the knot the average relationship is positive and has a large confidence interval. Extreme high prices per square meter can be found for basically any size.

- The 32 observations above 180 m² were dropped.
- sqm_sp2060, a new variable was created to signal observations with sqm larger than 60 m². Also log form lnsqm_sp2060 was created.
- Log form of sqm variable was created with further variations.





Feature engineering – Further steps

Floor variable

- Due to non-integer values, floor variable was transformed and decoded into floor2 variable.
- Flag variable was created for values larger than 5 and at floor 0.

Suter variable

- Only observations with suter = 0 and no missing value were kept.

Condition and Heating variables

- Both variables are factors with 7 and 13 categories respectively and a high proportion of missing values (7.5% and 10.9%).
- The factor levels were decoded for both in two alternative ways to create cleaner categories: condition with 8 levels, condition_broad with 6 levels, heating with 14 levels and heating_broad with 6 levels were created.
- NA values were decoded into others for heating to avoid dropping observations automatically.

Feature engineering – Newly added variables

View, Orientation

- Missing values were decoded with NA as a category itself, stored into orient_new variable.

Parking

- 3078 missing values, which is 50% of observations., therefore missing values were decoded with NA as a category itself:
  - Parking_1 variable was created with categories: garage for sale, garage included, street parking, outdoor parking.
  - Parking_2 variable was created with categories : garage, outdoor.

Balcony

- No missing values, originally a string variable with numbers, decoded into categories in two ways:
  - Balcony_1 variable was created with categories: no balcony, french (0-1 m^2), small (1-5 m^2), medium (5-15 m^2), large (>15 m^2).
  - Balcony_2 variable was created with categories : no balcony, normal (0-20 m^2), large (20-60 m^2), extra large (>60 m^2).

Model summary I.

	Description	Interactions	5-fold CV	Best RMSE	Best BIC (Levels)
Model 1 REG	• 6 regressions: reg_1 – reg_6 • Trained on the entire dataset, predictions based on the test set • Trying RHS and LHS variables both in levels and logs • Trying scaling of parking and balcony both in complex and simple form	NO	NO	-	-
Model 2 REG	• 2 regressions: reg_7 – reg_8 • From this point on, trained on the training set, predictions based on the test set • Trying RHS and LHS variables both in levels and logs • Parking and balcony only added in complex form from now on.	NO	NO	128.105 REG 7	59,997.215 REG 7
Model 3 REG	• 2 regressions: reg_9 – reg_10 • Trying RHS and LHS variables both in levels and logs • reg_9: Levels, interactions with 1) nrooms, 2) new_flat, 3) concrete_blockflat_d • reg_10: Logs, interactions with condition_broad for the rest of the variables	YES	NO	125.386 REG 9	60,148.704 REG 9
Model 4 REG	• 3 regressions: reg_7, reg_9 and reg_10 (best performers previously, all in Levels) • 5-fold cross-validation used for all models	YES	YES	125.386 REG 9	60,321.898 REG 10 CV
Model 5 RF	• 100 trees • No. of predictors: tried models with 4, 8 and default = sqrt(p) predictors per split • All potentially useful variables were used from the dataset.	DEFAULT	NO	111.50 RF (8)	-

Training and test set were created randomly in an 80% - 20% ratio for regressions, and 60% - 40% for RF

Model summary II.

For new variables added (view, orientation, bacony, parking) complex scaling performed slightly better than simple scales, therefore it is advised to use more complex scaling for these factor variables.

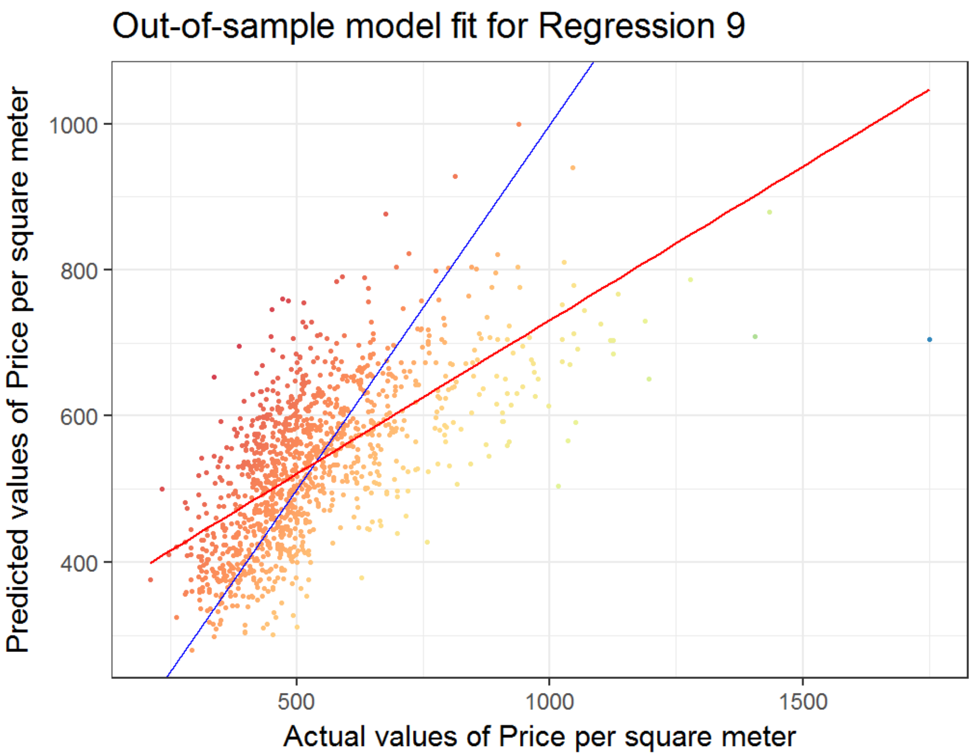
In regressions with no interactions Log models performed slightly better than levels, however, they brought unnecessary model complexity and exposure to massive overfitting. In regressions with interactions Log models did not really improve our prediction.

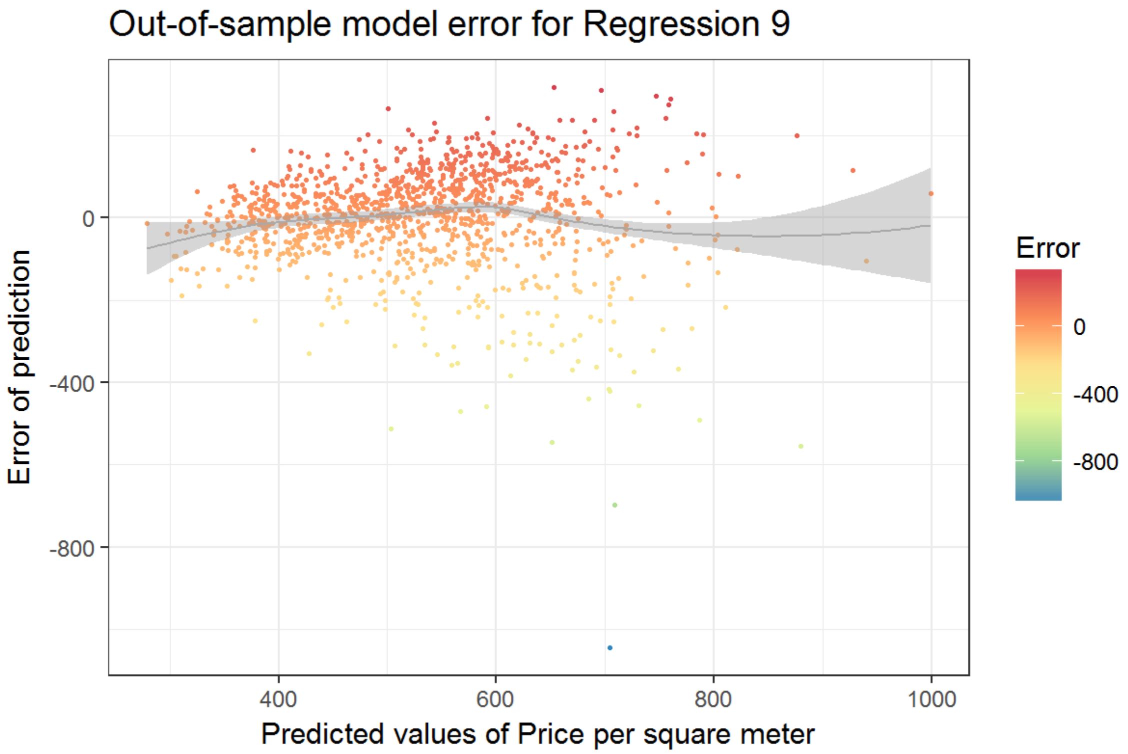
From BIC perspective, Reg 10 with cross-validation is the best because of its predictive power. From the RMSE perspective, Reg 9 without cross-validation provided the lowest error rate, however, both models are rather complex. If our aim is to provide the best prediction possible, we will have to use one of the more complex models.

Out of the models listed, the preferred one is Reg 9, because it performs the best without cross-validation as well and has the lowest error (125.39) and a high-enough BIC (60,148).

A Random Forest model was also estimated with an error rate of 111.50, which is significantly lower than all of the regression models above and also provided meaningful insight into variable importance. Therefore, the preferred method of choice for the prediction would be Random Forest. The most important variables based on error reduction are the concrete_blockflat dummy, the size in square meter spline and heating - the best predictors of price per square meter of flats in the dataset.

Error and Fit – Reg 9

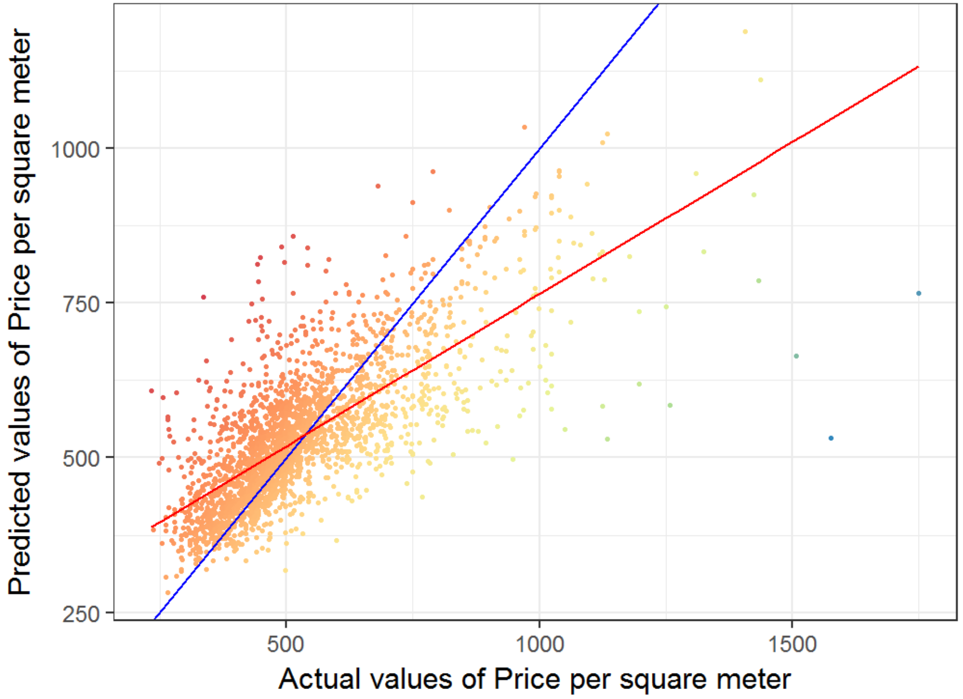




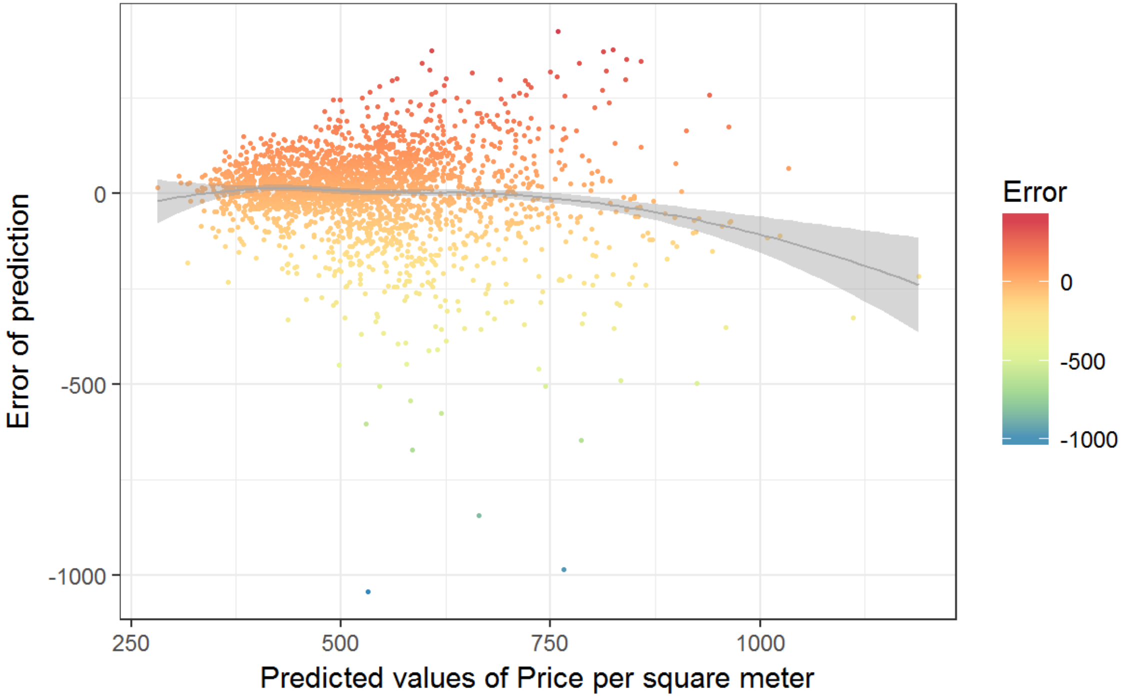
In the first graph the actual and the predicted values are plotted and colour marks the size of the error. The blue line is a 45 degree line (slope equals to R²), the red line is a fitted linear regression. The closer the blue line is to the red one, the smaller is the error rate, the more of the effect is explained by the linear regression and the better the prediction. R² of this model is 42%, which is meaningless, but useful to compare with the rest.

Error and Fit – Random Forest

Random Forest model fit for 8 predictors



Random Forest model error for 8 predictors



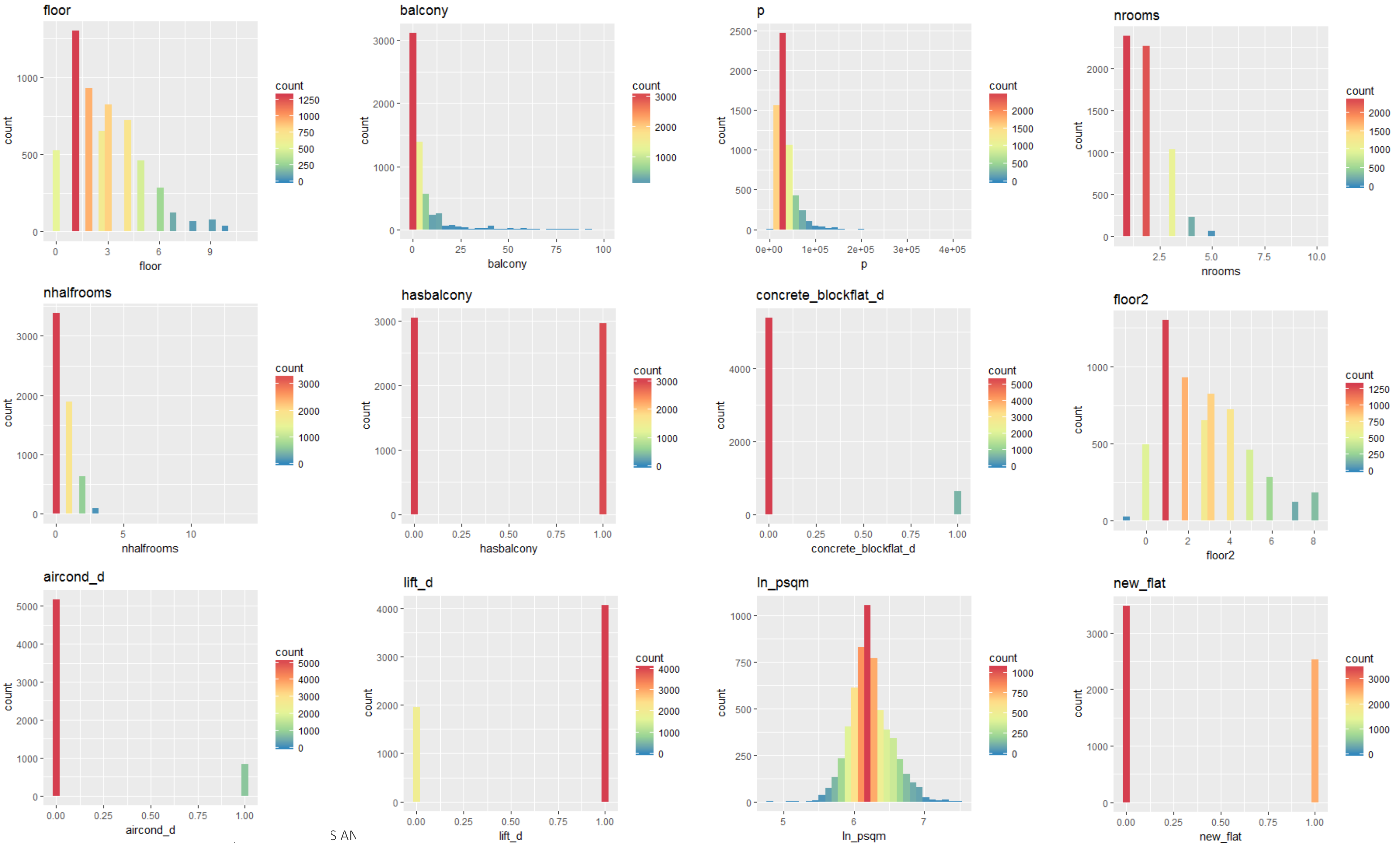
The blue line is closer to 45 degrees, because the slope of the fitted linear regression is approximately 0.5, which is larger than 0.42 of Reg 9. This indicates that the error rate is smaller for the Random Forest than for the best regression. There is still significant unobserved heterogeneity in our model, with relatively large errors above 500.

Technical Appendix

Data Analysis for Economic Policy 4 | CENTRAL EUROPEAN UNIVERSITY

Luca Keresztesi

Distribution of key predictors



Model 1

Key conclusion: Reg_3 and Reg_4 provides best model performance (smallest RMSE), however log brings unnecessary model complexity in reg_4. For new variables complex scaling performs slightly better than simple.

Multiple regression models trained on the entire dataset

Dependent variable: Price per square meter

	Reg 1	Reg 2	Reg 3	Reg 4	Reg 5	Reg 6
New controls added	No	No	Yes	Yes	Yes	Yes
Scaling of new controls			Complex	Complex	Simple	Simple
Dep. and expl. variables	Levels	Logs	Levels	Logs	Levels	Logs
BIC	75,136.894	-949.451	74,888.214	-1,160.274	74,968.228	-1,092.788
RMSE test	133.917	133.956	127.348	127.54	128.7	128.808
Observations	5,940	5,940	5,940	5,940	5,940	5,940
R^2	0.34	0.38	0.39	0.42	0.37	0.41

New controls orientation, parking, balcony and view

Model 2

Key conclusion: Reg_7 provides better model performance (smallest RMSE), while log brings unnecessary model complexity in Reg_8. BIC values are not comparable across Level and Log.

Levels and Logs Out-of-sample models

Dependent variable: Price per square meter

	Reg 3	**Reg 7**	**Reg 8**
Dataset	IS	OS	OS
Scaling of new controls	Complex	Complex	Complex
Dep. and expl. variables	Levels	Levels	Logs
BIC	74,888.214	59,997.215	-818.483
RMSE train	128.37	128.287	128.6
RMSE test	127.348	128.105	128.233
Observations	5,940	4,753	4,753

New controls orientation, parking, balcony and view

Model 3

Key conclusion: Reg_9 provides better model performance (smaller RMSE) but has a larger BIC. This means that neither Logs do not improve our prediction a lot, nor the additional interactions with conditions_broad make our prediction better.

Levels and Logs Out-of-sample models with interactions

Dependent variable: Price per square meter

	Reg 3	Reg 7	Reg 8	Reg 9	Reg 10
Dataset	IS	OS	OS	OS	OS
Scaling of new controls	Complex	Complex	Complex	Complex	Complex
Dep. and expl. variables	Levels	Levels	Logs	Levels	Logs
BIC	74,888.214	59,997.215	-818.483	60,148.704	-457.556
RMSE train	128.37	128.287	128.6	124.117	126.635
RMSE test	127.348	128.105	128.233	125.386	134.682
Number of controls	41	41	41	96	132
Observations	5,940	4,753	4,753	4,753	4,753

Model 4

Key conclusion: Cross-validated datasets provided lower RMSE and larger BIC in case of Reg 10, while for Reg 7 and Reg 9 RMSE from the model run on the original train and test set performed better. Overall, Reg 9 and Reg 10 CV provided nearly the same results: RMSE is lower for Reg 9 while BIC is larger for Reg 10.

Out-of-sample models with interactions and cross-validation

Dependent variable: Price per square meter

	Reg 7	Reg 7 CV	Reg 9	Reg 9 CV	Reg 10	Reg 10 CV
Dataset	OS	CV	OS	CV	OS	CV
BIC	59,997.215	59,985.734	60,148.704	60,124.881	-457.556	**60,321.898**
RMSE train	128.287	128.059	124.117	123.920	126.635	122.360
RMSE test	128.105	129.058	**125.386**	126.207	134.682	125.726
Observations	4,753	4,752	4,753	4,752	4,753	4,752

Model 5

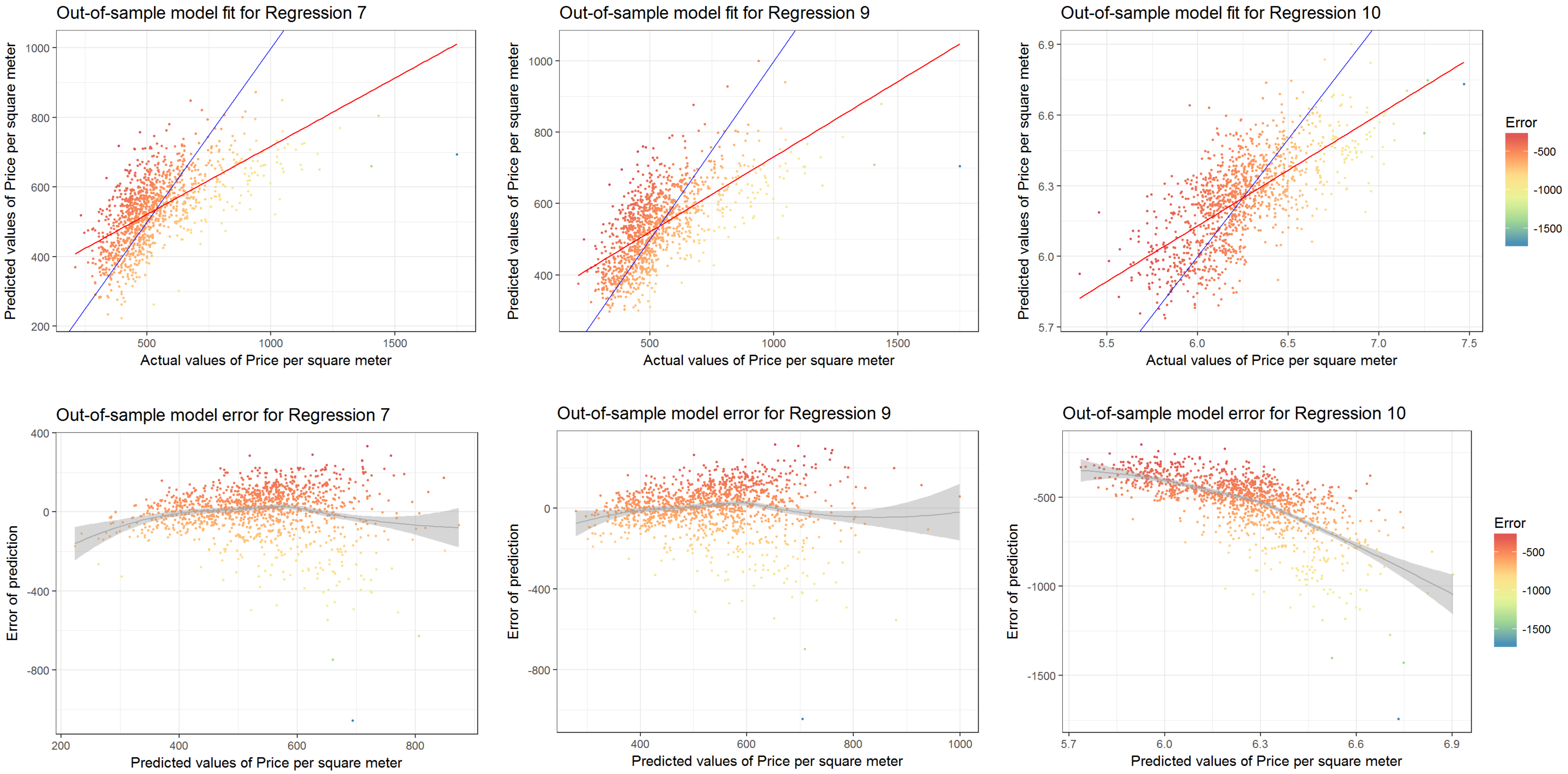
Key conclusion: The RF model with 8 predictors at each split provided the lowest error rate and could explain most of the variance. The most important variables based on the % of MSE reduction are the concrete_blockflat dummy, the size in square meter spline and heating - the best predictors of price per square meter of flats in the dataset.

Random Forest models

Dependent variable: Price per square meter

	RF 1	RF 2	RF 3
No. of trees	100	100	100
No. of predictors	4	8	Sqrt(p) = 6
RMSE	112.35	111.50	112.14
Variance explained	52.65%	53.36%	52.83%
TOP 5 most important variables	• concrete_blockflat_d: 24.16% • heating_broad: 20.37% • sqm_sp2060: 20.18% • sqm_sp60p: 14.32% • elevator: 14.03%	• concrete_blockflat_d: 34.46% • sqm_sp2060: 25.89% • heating_broad: 22.52% • balcony_1: 18.85% • sqm_sp60p: 18.01%	• concrete_blockflat_d: 27.91% • heating_broad: 23.93% • sqm_sp2060: 22.48% • view: 18.36% • balcony_1: 18.10%

Error and Fit – most important models



Out-of-sample model fit for Regression 7

Out-of-sample model fit for Regression 9

Out-of-sample model fit for Regression 10

Out-of-sample model error for Regression 7

Out-of-sample model error for Regression 9

Out-of-sample model error for Regression 10

Code and more detailed Technical Appendix

Github: https://github.com/lucakeresztesi/CEU_Data_Analysis/tree/master/DA4