The Machine Warehouse

Tomasz Siudalski

Mikołaj Gałkowski

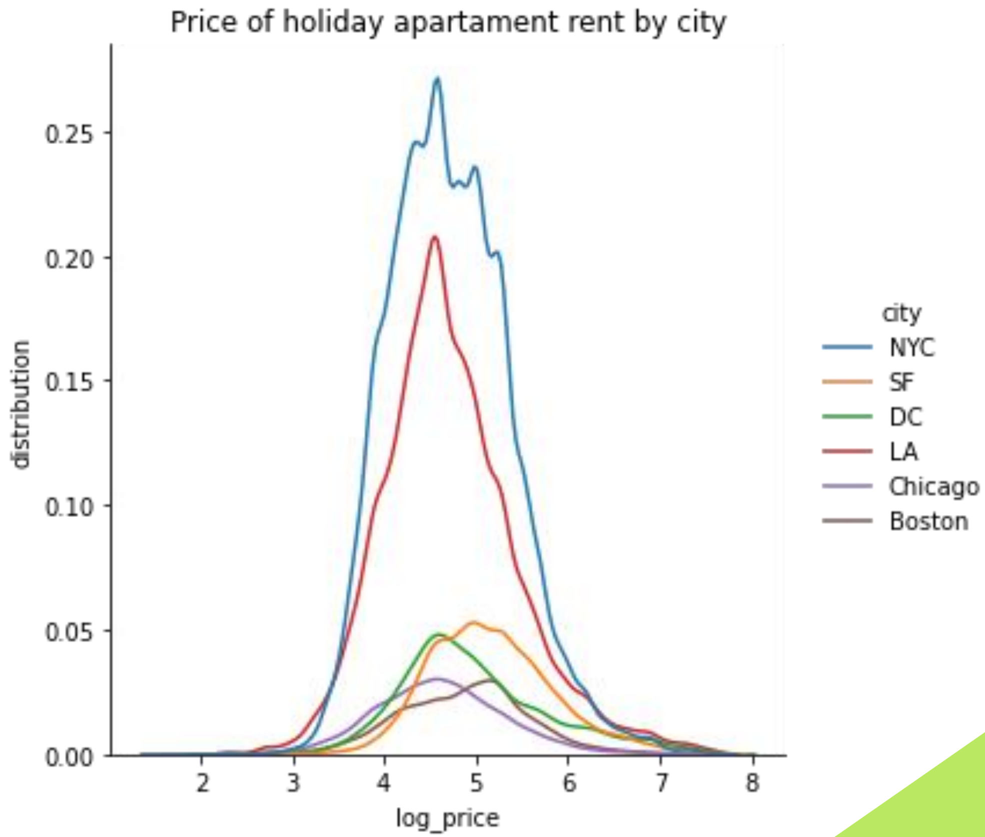
Wiktor Jakubowski

Agenda of presentation

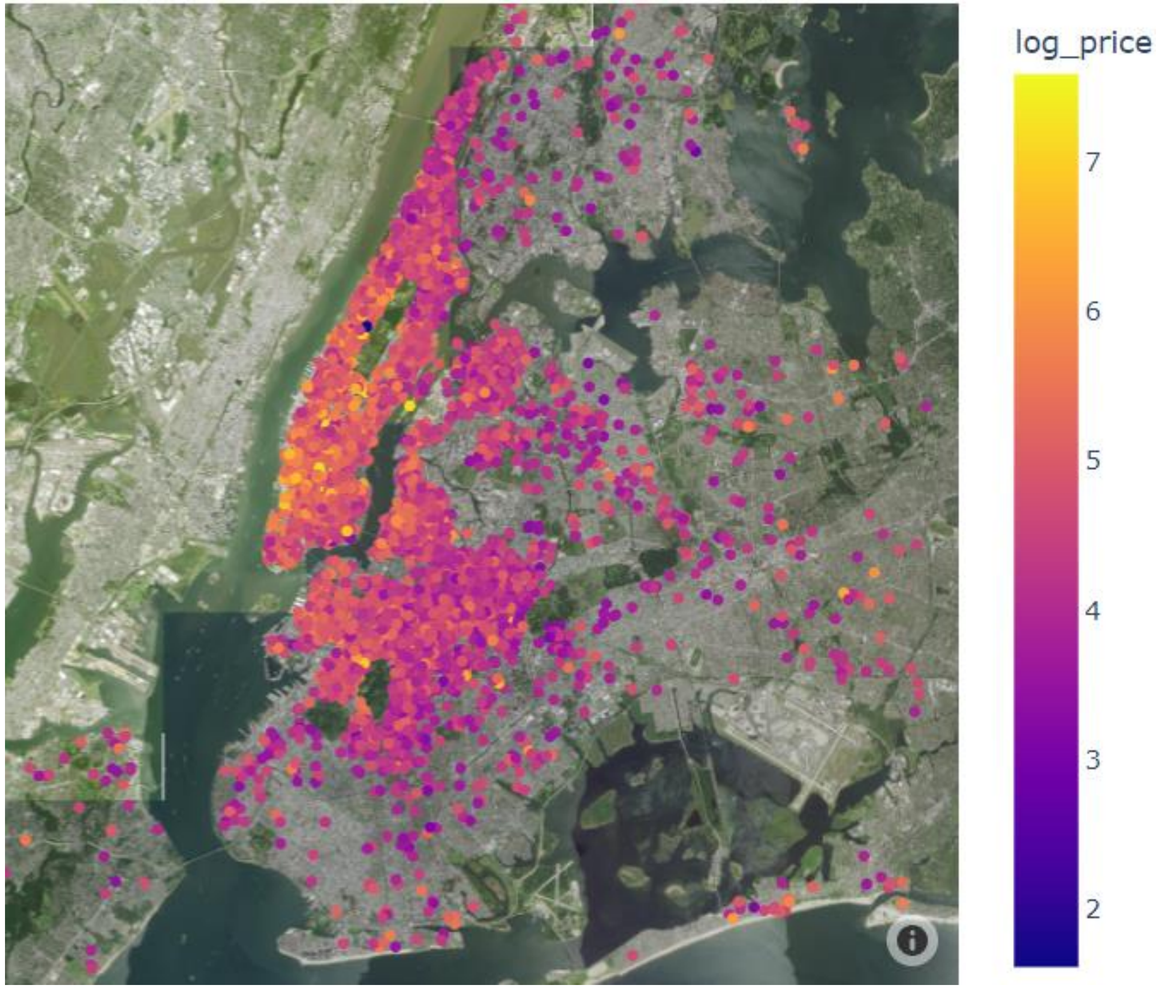
- problem description
- preprocessing
- model selection
- hyperparameters optimisation
- model evaluation
- intepretability
- business approach

Problem description

- Predicting rent prices for holiday stay in properties located in six cities in USA
- Dataset:
  - about 74 000 observations
  - nearly 30 features
  - categorical, numerical and datetime variables



Visualisation



Map of property holiday rent in New York City

- Data visualization based on coordinates
- Six separate areas (New York, Los Angeles, Chicago, Boston, San Francisco, Washington D.C.)
- Strong difference between prices in city centre and suburbs

Creating datetime columns and filling NaN’s

Amenities transformation

Imputing NaN’s (KNN, Iterative)

One-hot and ordinal encoding categorical variables

Metrics used for evaluating models performance

- MAE : mean absolute error
- RMSE : root-mean-square error
- R^2

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

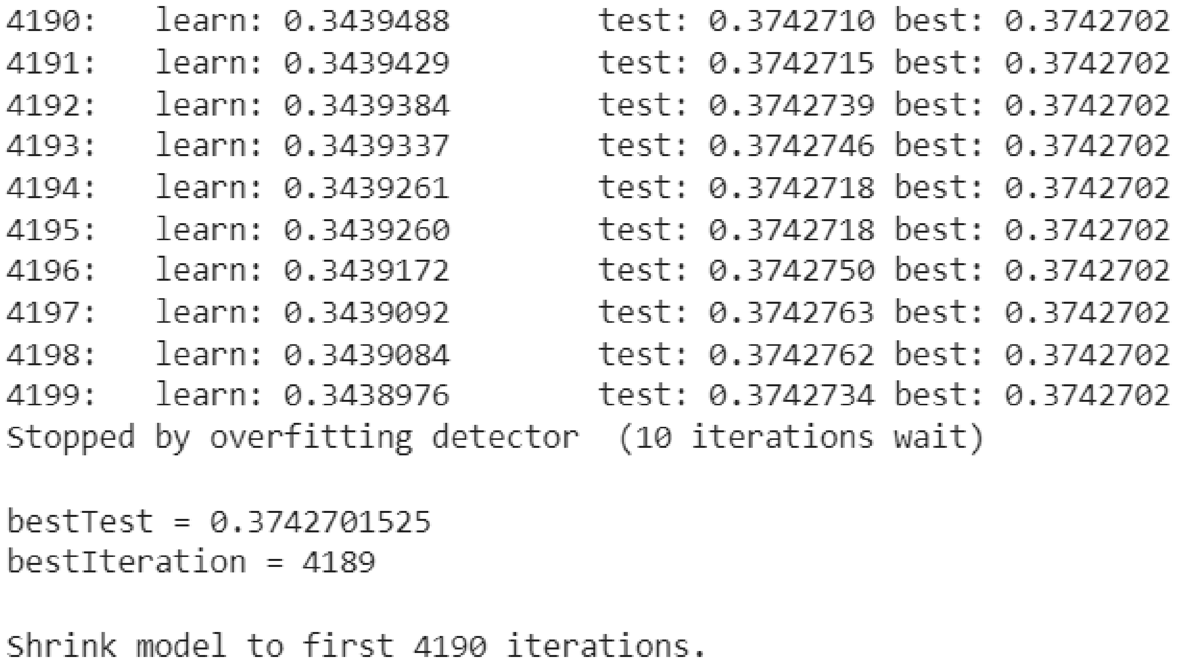
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

Model selection

Default model	*RMSE train*	*RMSE valid*	*MAE train*	*MAE valid*	*R^2 train*	*R^2 valid*
XGBoost	0.314	0.375	0.231	0.273	0.807	0.729
Random Forest	**0.145**	0.382	**0.104**	0.276	**0.959**	0.719
Linear Regression	0.457	0.454	0.343	0.343	0.592	0.602
CatBoost	0.351	**0.369**	0.254	**0.268**	0.76	**0.737**

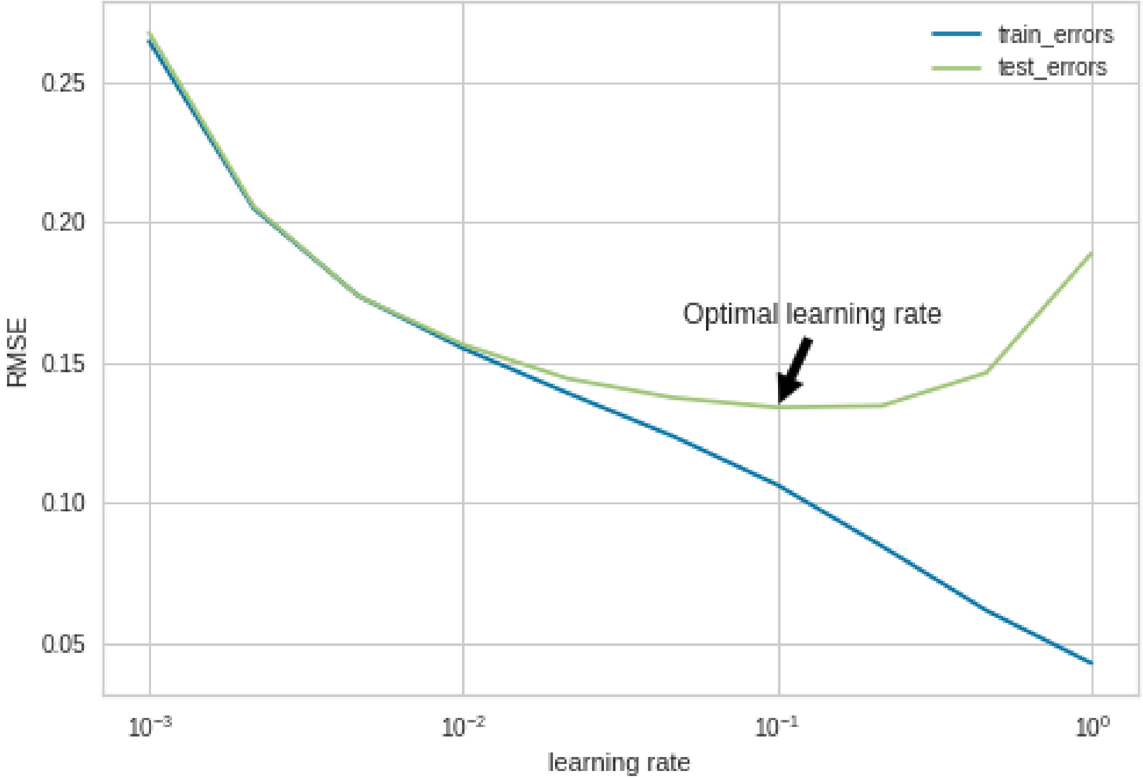
Hyperparameters optimisation

Optimising number of iterations



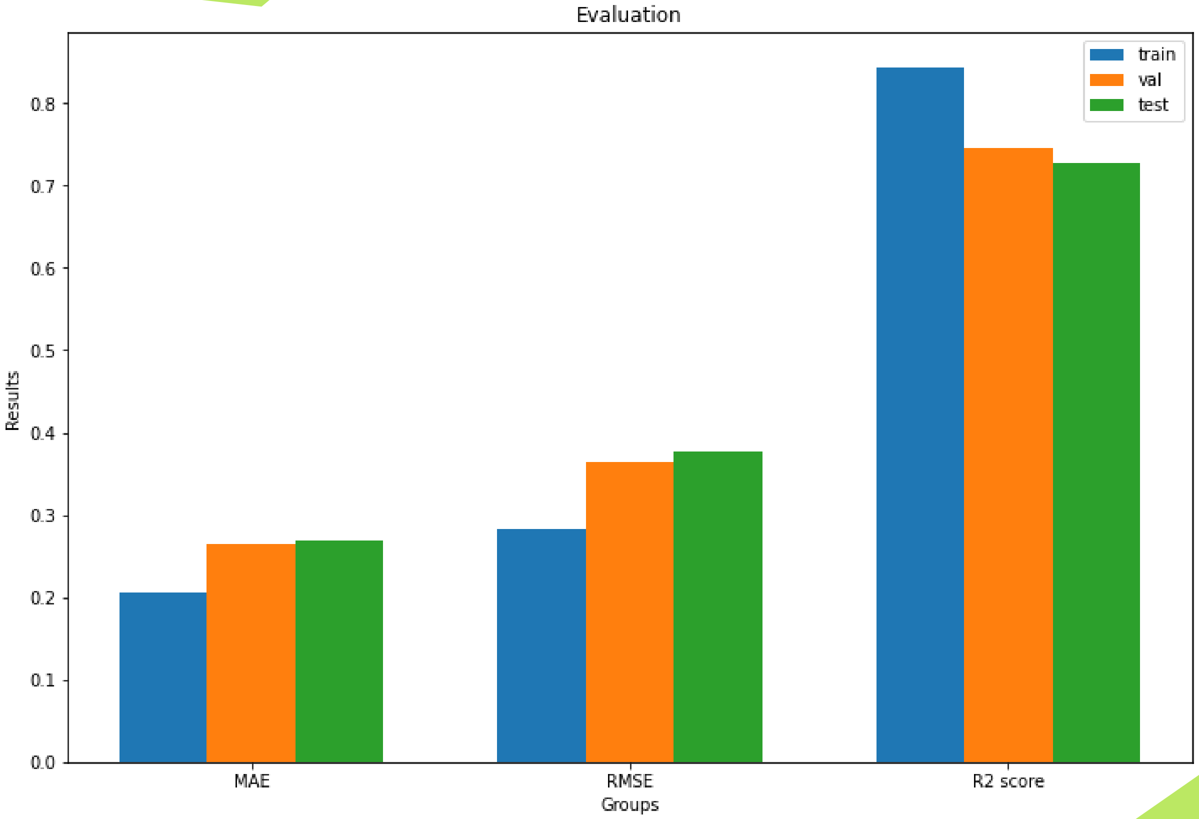
```
4190:   learn: 0.3439488        test: 0.3742710 best: 0.3742702
4191:   learn: 0.3439429        test: 0.3742715 best: 0.3742702
4192:   learn: 0.3439384        test: 0.3742739 best: 0.3742702
4193:   learn: 0.3439337        test: 0.3742746 best: 0.3742702
4194:   learn: 0.3439261        test: 0.3742718 best: 0.3742702
4195:   learn: 0.3439260        test: 0.3742718 best: 0.3742702
4196:   learn: 0.3439172        test: 0.3742750 best: 0.3742702
4197:   learn: 0.3439092        test: 0.3742763 best: 0.3742702
4198:   learn: 0.3439084        test: 0.3742762 best: 0.3742702
4199:   learn: 0.3438976        test: 0.3742734 best: 0.3742702
Stopped by overfitting detector  (10 iterations wait)

bestTest = 0.3742701525
bestIteration = 4189

Shrink model to first 4190 iterations.
```

Optimising learning rate



Model evaluation

- Values on independent test dataset:
  - MAE: 0.268
  - RMSE: 0.376
  - R^2 score: 0.727
- Metrics outcome slightly vary over different datasets



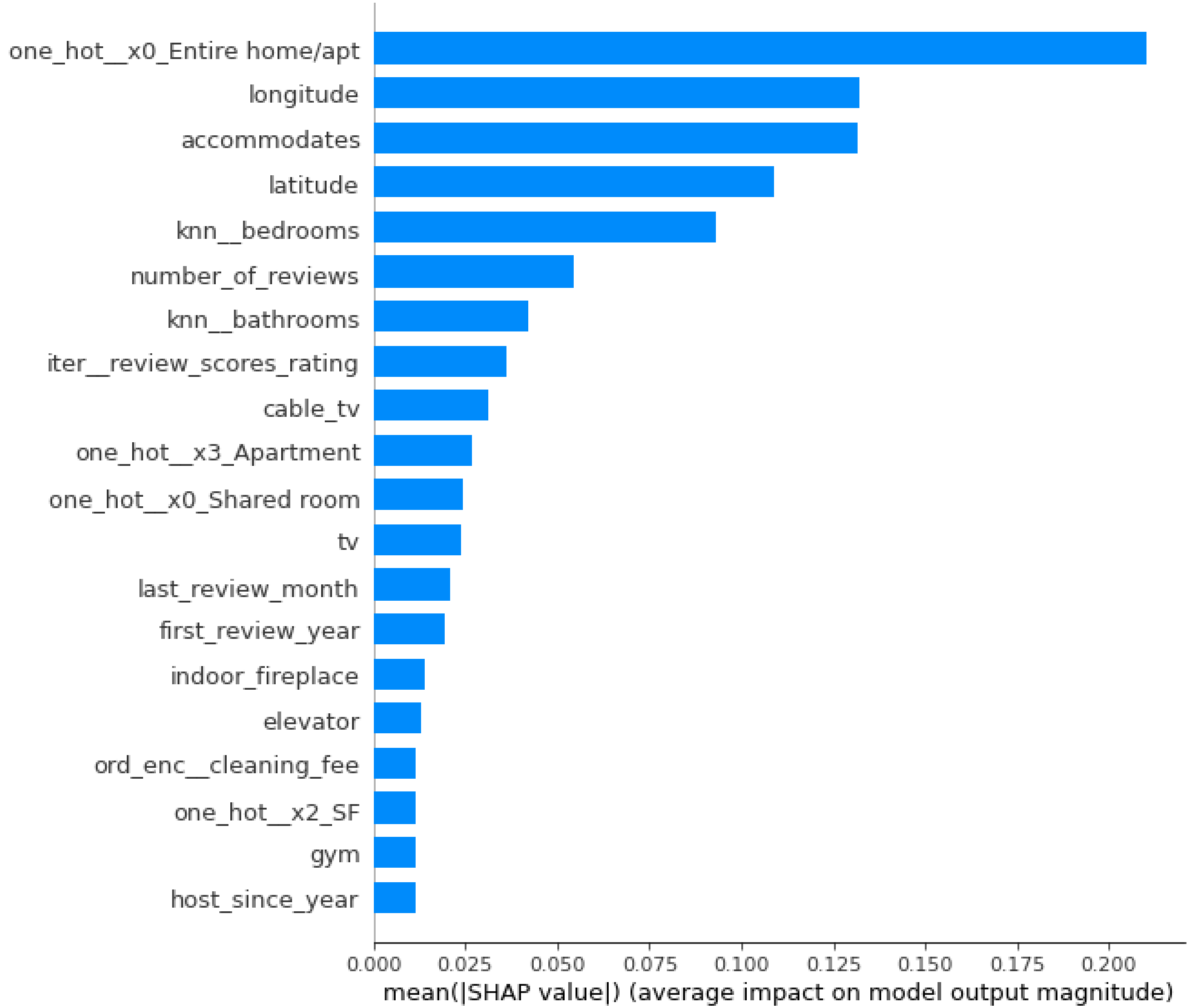
Feature importances

- Most important features:
  - Is the entire property for rent
  - Longitude and latitude
  - Number of accommodates
  - Number of bedrooms

- Least important features:
  - Ground floor access
  - Roll in shower with chair

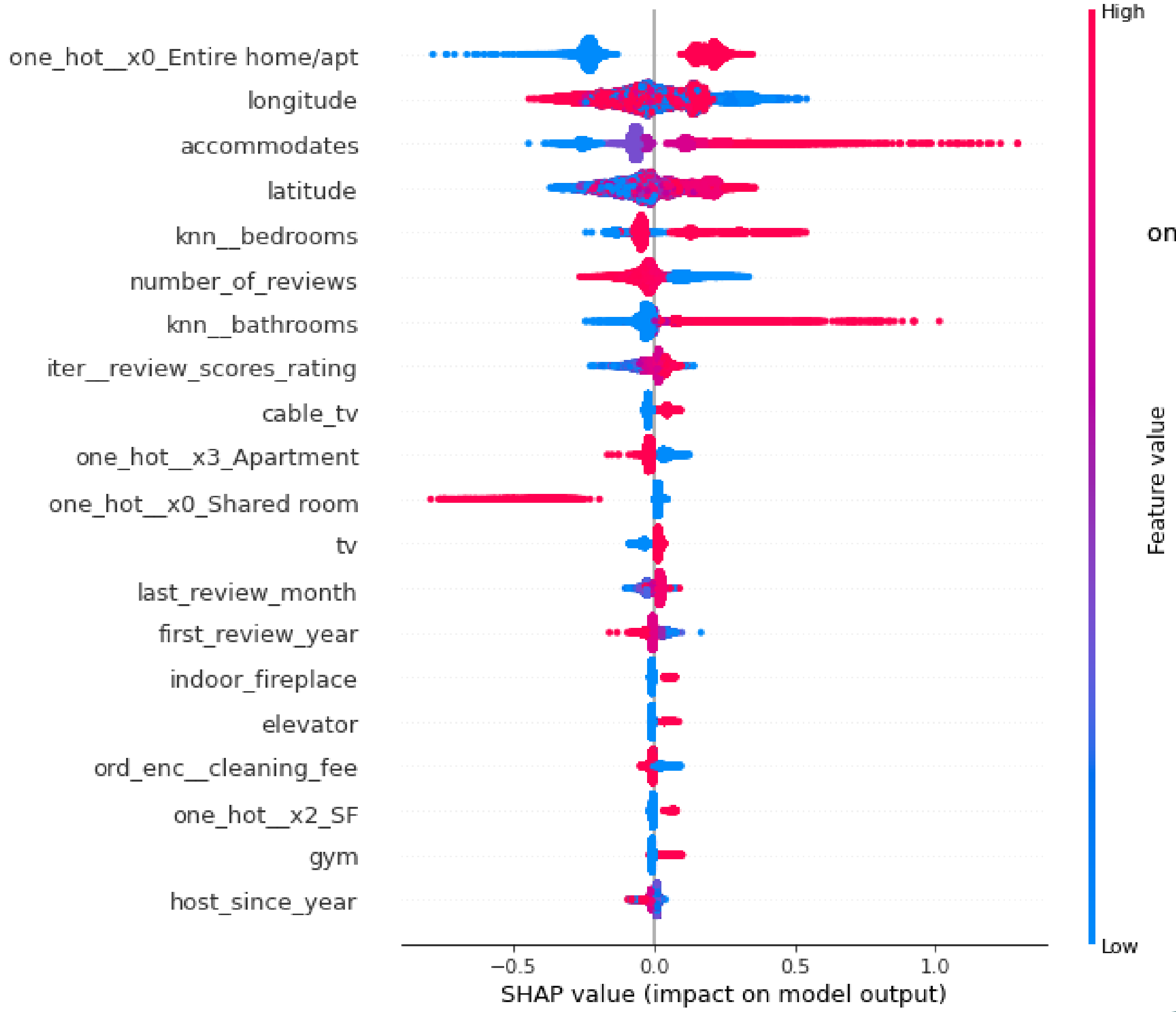
  They were removed during feature engineering

Top 20 most important features by their importance in model

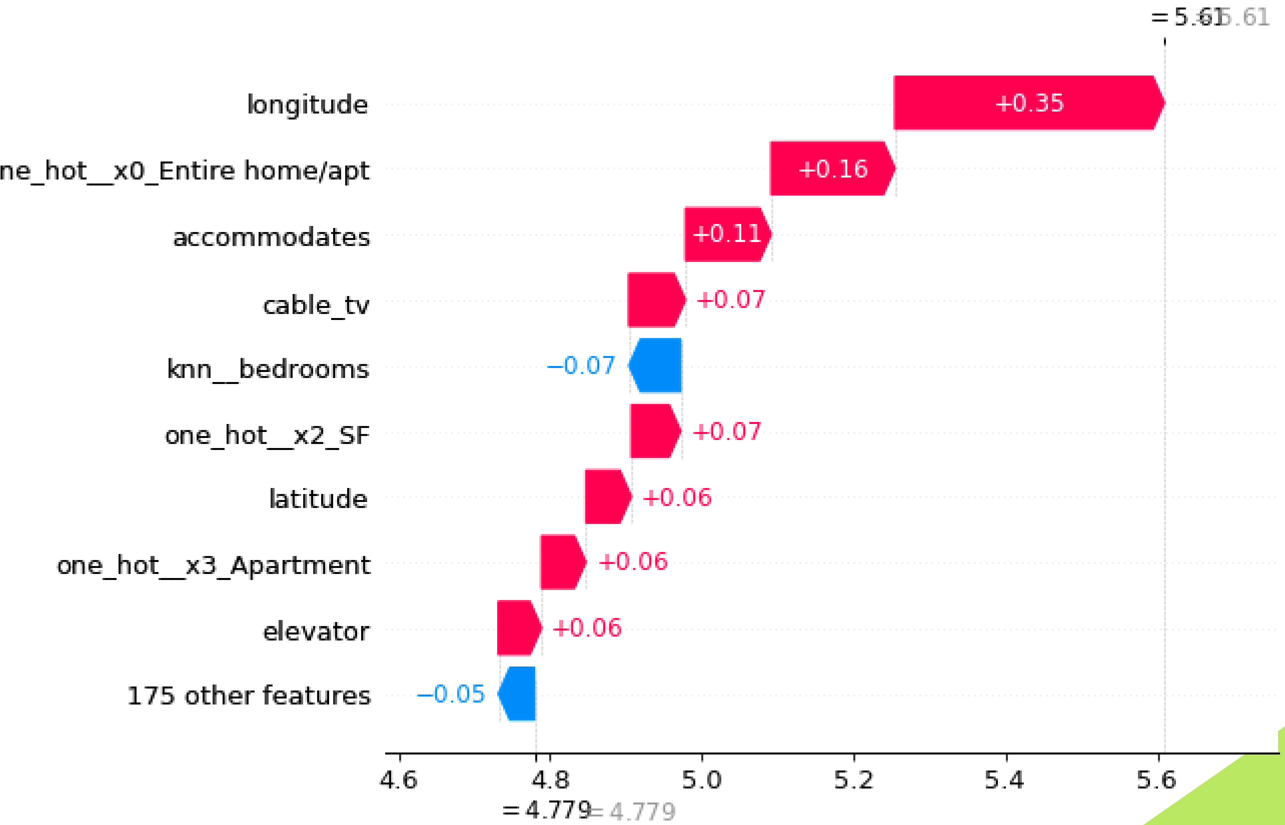


Interpretability

Features impact on model output



Features values impact on shifting of model's expected value



Thank you for your attention