Notes to Feature Selection

Not all features out of the new combined and extended feature set are equally valuable to the model in terms of predictive power. In fact, some of the features could not even be transformed to meaningful predictors such as multiple columns containing some kind of `id` information. Others were completely redundant in presence of a combination of original and/or manually constructed features and were thus dropped from the feature set.

Before starting the modeling we intended to reduce the feature set even further to avoid strong correlations among the input predictors and possibly improve generalization performance to out-of-sample data. Thus, we agreed on a two-step strategy.

First, we manually selected features based on three criteria:

1. The variable in consideration and the apartment price must have some connection based on human intuition and background knowledge. For instance, we included variables containing information about the apartment's *size* such as the number of `accomodates`, the number of `bathrooms` and the number of `bedrooms`.

2. There has to exist some correlation with the price in a bivariate visualization, e.g. a *barplot* in case of categorical predictors or a *scatterplot* in case of numeric features.

3. Since our dataset was comparably small with roughly 3000 observations, we had to take care about missing values. Therefore, each variable whose missing values could not be imputed in a meaningful and uncontroversial manner was either selected or dropped based on the trade-off of the number of missing values reducing the size of the data set and its additional predictive value.

Up to this point the feature selection process was solely based on bivariate relationships and self-chosen (arguably arbitrary) selection criteria. There was a high chance of exluding important predictors that shine in combination with other variables rather than on their own.

Hence, in a second step we fit an auxiliary Linear Regression Model including *all* of the available features (except for the trivial ones such as the `picture_url`) and analyzed the absolute magnitude of the coefficients to

get the relevant coefficient here. Since all variables are standardized these magnitudes are within the same range and unit-independent and can thus be compared.

This approach, of course, is not perfect as well since multiple highly correlated features having a strong *combined* impact on price could end up with rather small individual estimated coefficients due to *sharing* their predictive / explanatory power. To circumvent this potential issue, we relied on *algorithmic* feature selectors provided by the `scikit-learn` library that are (ideally) capable of separating the effects of strongly correlated features and select only a small subset of them.

We decided to apply the following algorithms: *RFE* (Recursive Feature Selection), *SelectFromModel*, *SelectKBest*, *SequentialFeatureSelector* and *VarianceThreshold*. The detailed results are shown in the appendix. Because the *RFE* showed the best performance and the selected subset of original features can be immediately interpreted as potentially most important features, so we performed an even more detailed analysis with this feature selector afterwards.

The overall results of these algorithms lead to additionally adding the variable `property_type`. But thereby some of the observed connections had to be taken with care. For example, by far the largest coefficient was attributed to the category *Houseboat* of the `property_type` variable. However, this observation turned out to be the only Houseboat contained in the dataset. To avoid strong influences of these outliers, we combined all rare categories together to a larger `Other` category and used the mean price of all observations contained in this category.

The last step consists of data preprocessing. Therefore we implemented a function column_transformer, which included the One-Hot Encoding of the categorical variables as well as the standardization of the numeric variables.