Notes to Feature Selection

Contents

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. As an example the original data contained a column `bathrooms_text` containing phrases like *'Three and a half shared bathrooms'*. We split this variable into two new features: One of numeric type containing the number of bathrooms (here 3.5) and one boolean variable if the bathroom(s) are shared. Conditioned on these two new columns the original variable `bathrooms_text` did not contribute any new information and was 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 visual-

ization, 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. 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 will go in much more detail about those automatic feature selectors in the next section.

For the auxiliary regression we decided to keep 30 variables. Out of the predictors with largest absolute coefficients most were already chosen by the first step described above. Nonetheless, we could detect some additional strong relationships to the apartment price and, based on these findings, expanded the resulting feature set of step 1.

As a final remark, 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. Consequently, if this observation is used during model fitting, the predicted price during deployment will be extraordinarily high for any other observation of this property type leading to potentially large bias if the original Houseboat was particularly expensive within this property class by chance.

In such cases we combined all rare categories together to a larger `Other` category and used the mean price of all observations contained in this category to avoid strong influences of outliers.

- One-Hot Encoding of Categorical Variables, Standardization of Numeric Variables
- Experimenting with different ways of algorithm-based feature selection / dimensionality reduction
- Focus on `PCA` as most theoretically supported procedure and `RFE` as procedure we chose
- PCA has advantage of reducing dimensionality and simultaneously producing *uncorrelated* features, disadvantage of producing linear combinations of original features which are harder to interpret
- `RFE` showed best performance and selects subset of original features, can be immediately interpreted as potentially most important features
- Briefly explain how `RFE` works