Ames Housing dataset

Following Kaggle's Getting Started Prediction Competition, the goal for this discover project is to predict the sales price for each house. For each Id in the test set, you must predict the value of the **SalePrice** variable.

The performance metric for your prediction model is the *Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price*. (Taking logs means that errors in predicting expensive houses and cheap houses will affect the result equally.)

Looking at the public leaderboard, the top 2% have RMSLE of 0.00044, whilst 25th-percentile and and median performance is at 0.125 and 0.14, respectively.

As an extra challenge, you can try to trade-off the number of predictors (less is better) vs. performance. Can you make the top 10% (RMSLE 0.123) with the least number of predictors?

Exercises

Exercise 1

Provide a table with descriptive statistics for all included variables and check:

- Classes of each of the variables (e.g. factors or continuous variables).
- Descriptive/summary statistics for all continuous variables (e.g. mean, SD, range) and factor variables (e.g. frequencies).
- Explore missing values

Exercise 2

There a several missing values in the dataset, which need to be tackled before we can proceed with the rest of the analysis. There are many ways to impute missing values, but for now, impute missing values as follows:

- Use the median for numeric variables
- Use the label "100" in all factor variables

Exercise 3

The variable "SalePrice" refers to the price at which a property was sold and hence is the variable of interest for our prediction model ("Y" or dependent variable). Explore Y in terms of:

- Descriptive/summary statistics (e.g. mean, SDs, range)
- Visualize the distribution of Y (e.g. matplotlib or seaborn)
- Visualize the distribution of Y by looking at various subgroups (e.g. create boxplot or scatterplot using matplotlib or seaborn)
- Look at differences between neigbourhoods
- Look at differences between housing style
- Draw a correlation plot to see all correlations between Y and the independent (numeric) variables (Hint: use `df.plotting.scatter_matrix` or `seaborn.pairplot`)

Exercise 4

- Estimate a LASSO model and a kNN model
- Assess which model performs best