Final Report: Kaggle House Price Prediction

Team Name: Techduo

Name: Lameya Islam Isha

Sahira Salam

ID: 1537CSE00368

1640CSE00488

Kaggle account: https://www.kaggle.com/techduo

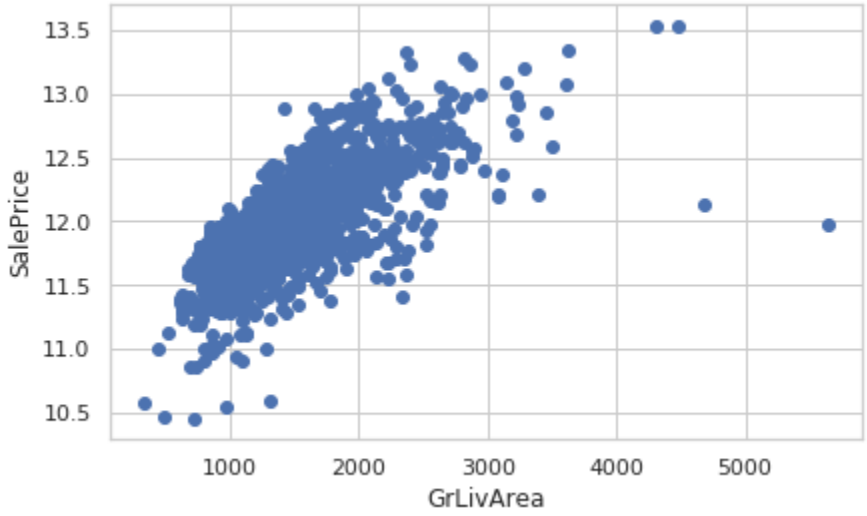
Git repository: https://github.com/SahiraSalam/Techduo

Project Goal:

The goal of this competition is to predict the sales prices of houses based on these variables. We collaborated on certain parts of the project and completed other parts individually as if it were a research project. The goal of the project, as aspiring data scientists, was to utilize our arsenal of machine learning knowledge to predict housing prices.

Data Description and Preprocessing

House price prediction dataset contains 1461 train data and 1459 test data. Each sample is represented by 2919 features for finding required house.



Separate Dataframes

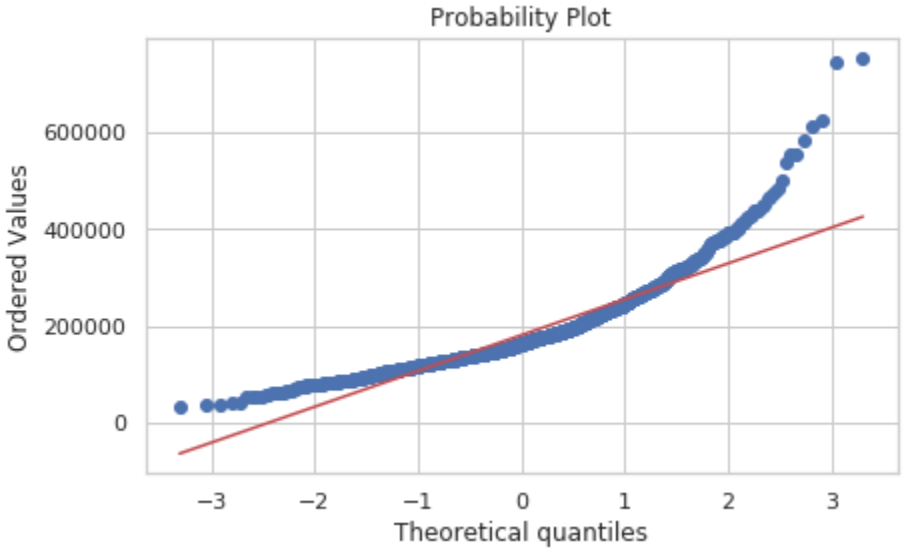
Might be useful when we consider features of different data types.

Analyzing the Target i.e. 'SalePrice'

We are analyzing the target variable is 'Saleprice'

The distribution of target is a bit right skewed. Hence taking the 'log transform' is a reasonable option.

Get also the quantile-quantile plot.



Taking 'log transform' of the target

Most Related Features to the Target

1. Note that some of the features have quite high correlation with the target. These features are really significant.

2. Of these the features with correlation value >0.5 are really important. Some features like GrLivArea etc.. are even more important.
3. We will consider these features (i.e. GrLivArea,OverallQual) etc. In more detail in subsequent sections during univariate and bivariate analysis.

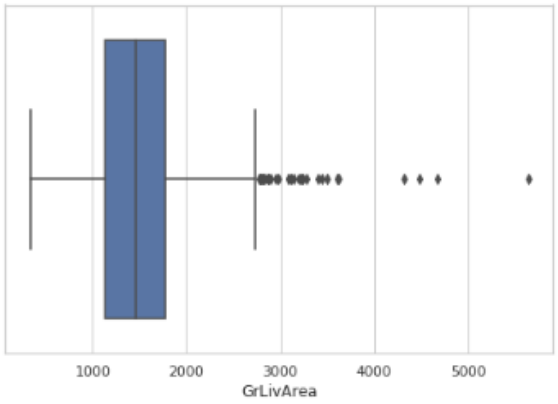
Univariate Analysis

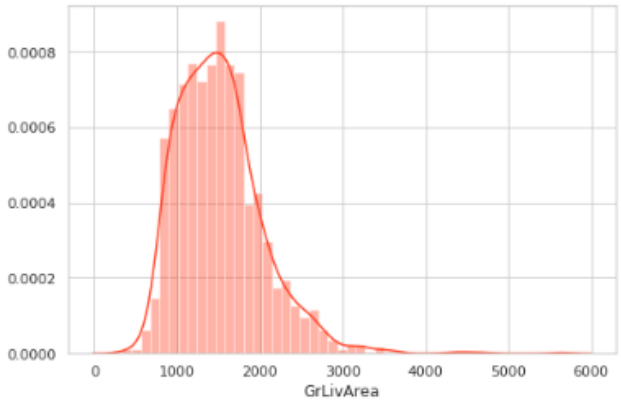
In this section the univariate analysis is performed; More importantly I have considered the features that are more important with the 'Target' that have high correlation with the Target.

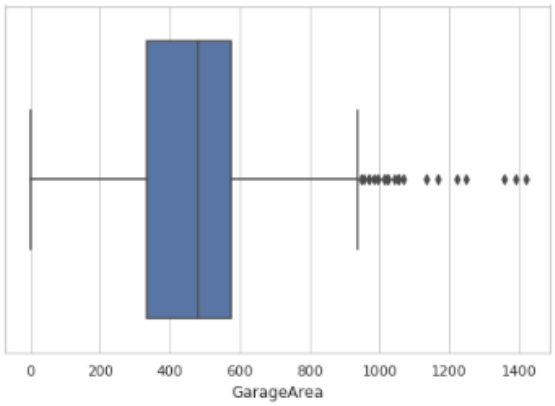
For the numeric features I have used a 'distplot' and 'boxplot' to analyze their distribution.

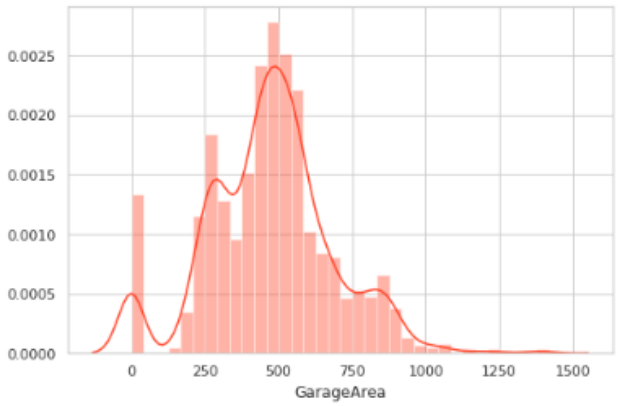
Similarly for categorical features the most reasonable way to visualize the distribution is to use a 'countplot' which shows the relative counts for each category or class. Can use a pie-plot also to be a bit more fancy.

NUMERIC FEATURES









Bivariate Analysis

In this section the Bivariate Analysis have been done. I have plotted various numeric as well as categorical features against the target is 'SalePrice'.

Missing Values Treatment¶

In this section I have handled the missing values in the columns.

Firstly I have dropped a couple of columns that have a really high % of missing values.

For other features I have analyzed if it that feature is important or not and accordingly either have drooped it or imputed the values in it.

For imputation I have considered the meaning of the corresponding feature from the description. Like for a categorical feature if values are missing I have imputed "None" just to mark a separate category meaning absence of that thing. Similarly for a numeric feature I have imputed with 0 in case the missing value implies the 'absence' of that feature.

In all other cases I have imputed the categorical features with 'mode' i.e the most frequent class and with 'mean' for the numeric features.

Handling Skewness¶

For handling skewnesss I will take the log transform of the features with skewness > 0.5.

You can also try the BoxCox transformation as mentioned before.

Regression Models

Lastly it is the time to apply various regression models and check how are we doing. I have used various regression models from the scikit.

Parameter tuning using GridSearchCV is also done to improve performance of some algos.

The evaluation metric that I have used is the Root Mean Squared Error between the 'Log of the actual price' and 'Log of the predicted value' which is also the evaluation metric used by the kaggle.

To get a better idea one may also use the K-fold cross validation instead of the normal holdout set approach to cross validation.