MANARAT INTERNATIONAL UNIVERSITY

TOPIC

House Prices: Advanced Regression Techniques

TEAM

Ibrahim Kardi

CONTESTANTS NAME

01. Tariqul Islam Toriq
    ID: 1640CSE00534
02. Muhammad Ibrahim
    ID: 1640CSE00514
03. Rayhan Nasir
    ID: 1640CSE00510
04. Tamjid Arafin

www.github.com/ikardi420/house_price_prediction

kaggle www.kaggle.com/ikardi

1 Project Goal

The goal of this project is to apply what we have learned in the Class "artificial intelligence" to build a regression model in the competition "House Prices: Advanced Regression Techniques" in Kaggle

The main objective of the competition is to predict sales prices and practice feature engineering, RFs, and gradient boosting.

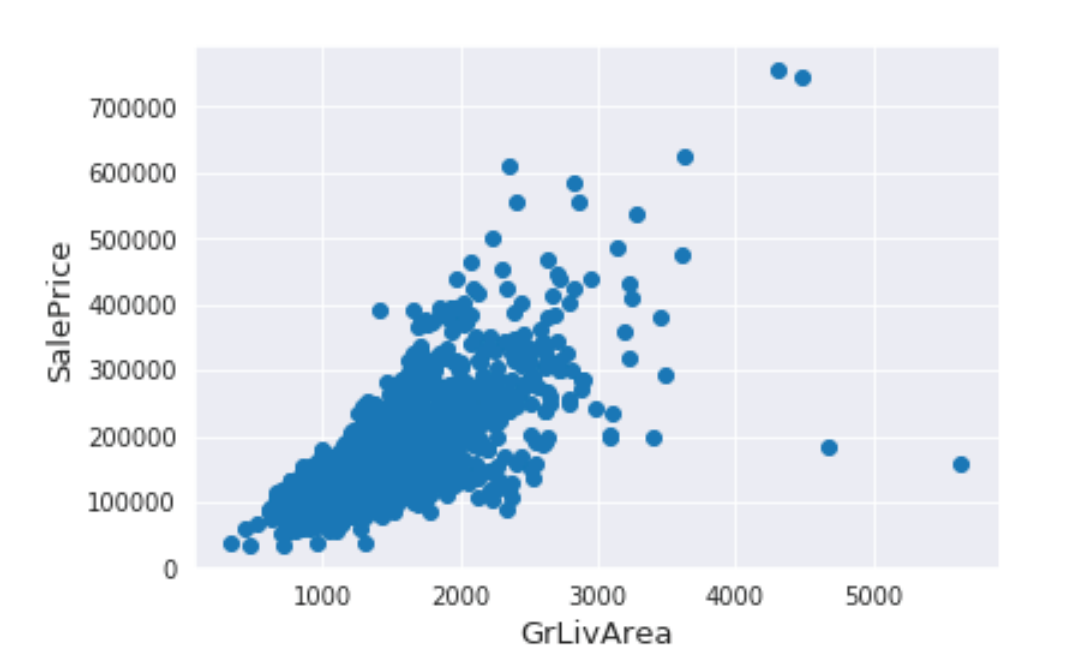
1.1 Problem Statement:

Submissions are evaluated on 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.)

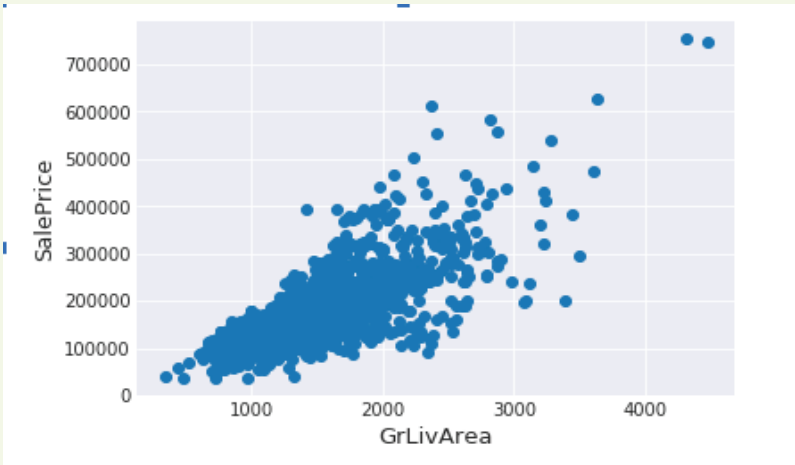
2 Data Description and Preprocessing:

Handling outliers:

We can see that there are outlinear with low SalePrice and high GrLivArea. This looks odd. We need to remove it.

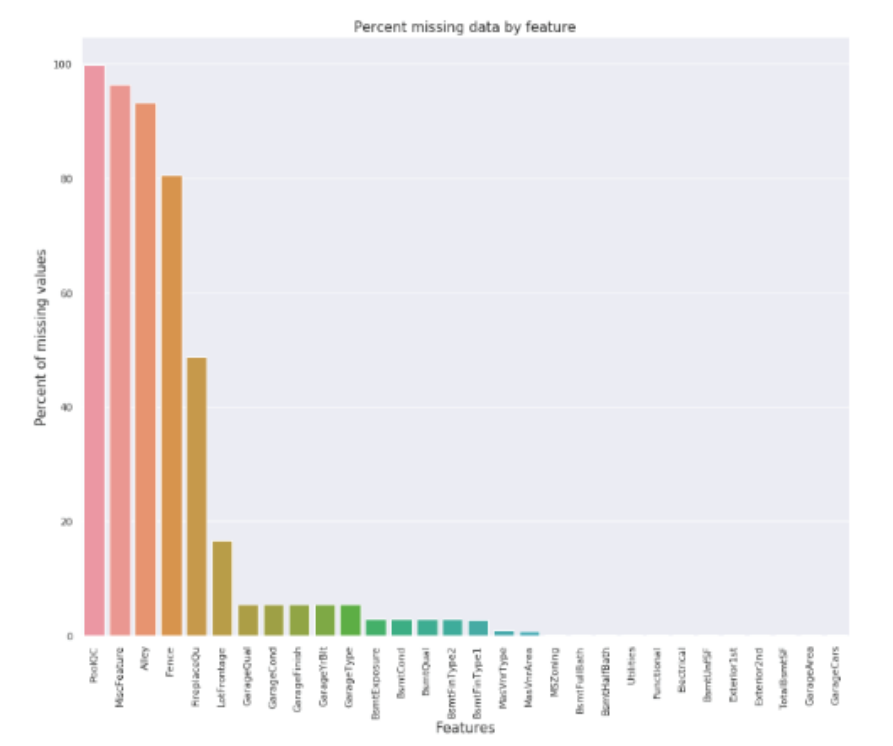


We can see at the bottom right two with extremely large GrLivArea that are of a low price. These values are huge outliers. Therefore, we can safely delete them.



Handle Missing Data:

Missing Data Percentage:



We have removed all missing values. For example:
Since PoolQC has the highest null values according to the data documentation says null values means 'No Pool.Since majority of houses has no pool. So we will replace those null values with 'None'.

Features Engineering:

- Concatenation the train and test data in the same dataframe
- Create features
- Check how the features work with the model.
- Log transformation

Modeling Methods :

- Ridge Regression
- Lasso Regression
- Gradient Boosting Machine (GBM)
- Linear Model with Forward Stepwise
- Models Ensembling

Results & Discussion:

First Submission:

Date: 10-7-19

Score : 0.12225

Position: 1321

Second Submission:

Date : 16-7-19

Score : 0.11548

Position: 495

Third Submission:

Date : 16-7-19

Score : 0.10766

Position: 159

Final submission:

Score :0.10649

Position: 103