



MANARAT INTERNATIONAL UNIVERSITY

A Centre of Academic and Moral Excellence

Department of Computer Science and Engineering
Artificial Intelligence (CSE – 411)

Report: House Price Prediction Competitions

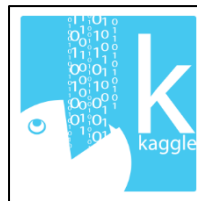
TEAM : ALPHA

CONTESTATS NAME :

01. Asifur Rahaman ID: 1640CSE00473

02. Minhajul Abedin ID: 1539CSE00462

03. Mir Adnan Bin Hira ID: 1640CSE00526



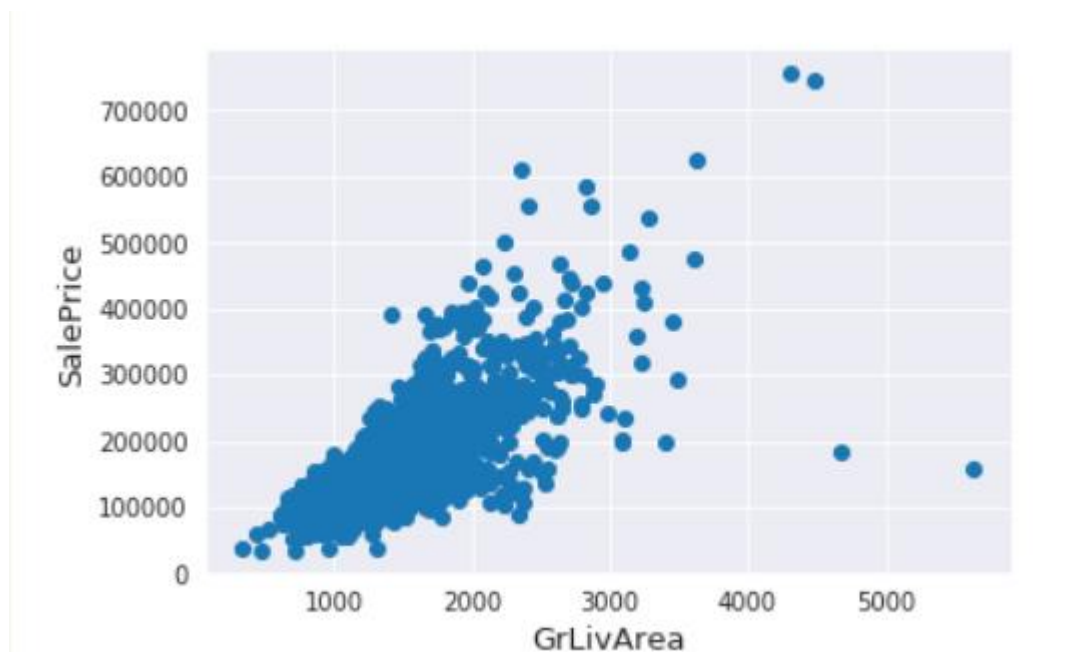
Project Goal:

The goal of this competition is to apply what we have learned in the "Artificial Intelligence" Course , to build a regression model in the competition "House Prices: Advanced Regression Techniques" in Kaggle. 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 :

Handling outliers:

In this graph we see there are outliers with low SalePrice and high GrLivArea . We need to remove it.



Handle Missing Data:

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

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

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

For imputation We 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.

Separate Dataframes:

Might be useful when we consider features of different data types

Features Engineering:

- Note that some of the features have quite high correlation with the target. These features are really significant.
- Of these the features with correlation value >0.5 are really important. Some features like GrLivArea etc.. are even more important.
- We will consider these features (i.e. GrLivArea) etc


Modeling Methods :

- Regression Models:
We used various regression models from the scikit.
- Gradient Boosting Machine (GBM)

Results & Discussion:

All	Successful	Selected	
Submission and Description		Public Score	Use for Final Score
submission.csv 16 days ago by ASIFUR RAHAMAN 4 th		0.11548	<input type="checkbox"/>
sample_submission.csv 16 days ago by ASIFUR RAHAMAN 3 rd subbmision		0.40890	<input type="checkbox"/>
submission.csv 22 days ago by ASIFUR RAHAMAN 2 nd submission		0.12269	<input type="checkbox"/>
submission.csv a month ago by ASIFUR RAHAMAN 1st submission		0.12269	<input type="checkbox"/>

Position : 459

459	ASIFUR RAHAMAN		0.11548	4	16d
-----	----------------	--	---------	---	-----