Manarat International University

Department of Computer Science and Engineering

Artificial Intelligence (CSE – 411)

Contest On

House Price Prediction Problem

Team: GirlsPower

- Kazi Tanjee Tamanna (ID 1640CSE00498)
- Raisa Akter (ID 1640CSE00515)
- Sumaiya Afrin (ID 1640CSE00527)

Kaggle Account: https://www.kaggle.com/girlspower

Git Repository Link:
https://github.com/RaisaAkter/House_price_prediction_problem

Project Report: House Price Prediction Problem

Introduction

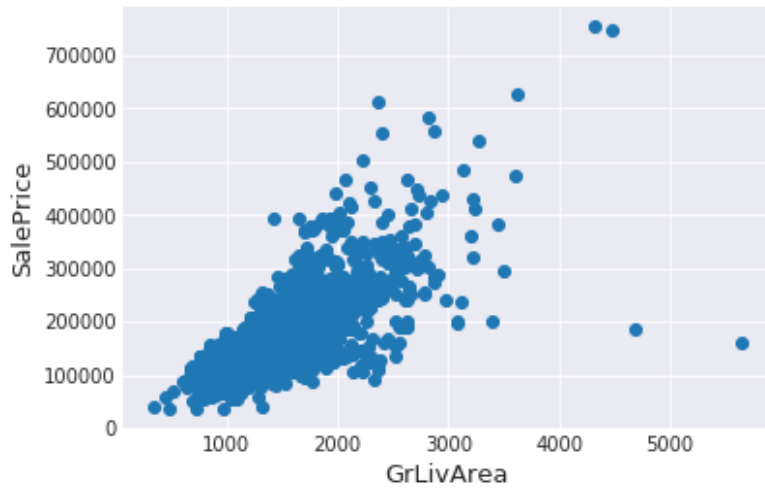
- *Project Goal:* The goal of this project is to develop a model which predict the price of house based on some given features. So in this project, we developed and evaluate the performance and the predictive power of the model. By this project we also learned how to participate a contest and also how to analysis a real problem.

- *Problem Statement*: The house price prediction problem is a problem about predicting price of a house based on some given features. The features are about the house and we have to predict the price as much accurately as we can. In this problem almost 80 unique feature about house is given which we will use to solve the problem.

Data Preprocessing

To develop a model, first we have process our data so that they will be workable in developing the model. The steps of process data are given below.

- Outlier Handling: Outliers are extreme values that fall a long way outside of the other observations. Outliers decrease the performance of the model so we need to remove the outliers for better performance. For example, we took the feature "GrLivArea" and saw the outliers. 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.



Missing & Redundant Value Handling: To increase the predictive power of the model, we searched the missing value in the features.

```
list of total missing data (in percentage)
               Missing Ratio
PoolQC               99.691
MiscFeature          96.400
Alley                93.212
Fence                80.425
FireplaceQu          48.680
LotFrontage          16.661
GarageFinish          5.451
GarageYrBlt           5.451
GarageQual            5.451
GarageCond            5.451
GarageType            5.382
BsmtExposure          2.811
BsmtCond              2.811
BsmtQual              2.777
BsmtFinType2          2.743
BsmtFinType1          2.708
MasVnrType            0.823
MasVnrArea            0.788
MSZoning              0.137
BsmtFullBath          0.069
BsmtHalfBath          0.069
Utilities             0.069
Functional            0.069
Exterior2nd           0.034
Exterior1st           0.034
SaleType              0.034
BsmtFinSF1            0.034
BsmtFinSF2            0.034
BsmtUnfSF             0.034
Electrical            0.034
```

So we fill up the missing value by coding.

```python
### imputing missing values
# PoolQC --> NA means missing houses have no Pool in general so "None"
all_data['PoolQC'] = all_data['PoolQC'].fillna("None")

# MiscFeature --> NA means no misc. features so "No"
all_data['MiscFeature'] = all_data['MiscFeature'].fillna("None")

# Alley :  NA means "no alley access"
all_data['Alley'] = all_data['Alley'].fillna("None")

# Fence: NA means "no fence"
all_data['Fence'] = all_data['Fence'].fillna("None")

# FireplaceQu: NA means "no fireplace"
all_data['FireplaceQu'] = all_data['FireplaceQu'].fillna("None")


# GarageType, GarageFinish, GarageQual and GarageCond: NA means "None"
for col in ('GarageType', 'GarageFinish', 'GarageQual', 'GarageCond'):
    all_data[col] = all_data[col].fillna("None")

# GarageYrBlt, GarageArea and GarageCars : NA means o
for col in ('GarageYrBlt', 'GarageArea', 'GarageCars'):
    all_data[col] = all_data[col].fillna(0)

```

Feature Engineering:

In feature engineering, we have done some steps like

- Transforming some numerical variables that are really categorical.
- Label encoding to categorical feature.
- Adding extra feature to increase predictive power. For example we created "TotalSF", "Total_porch_sf" e.t.c new features.
- Delete the features that are really unimportant to reduction dimensionality.

Modeling Method:

As the contest is about regression problem, we used models related to regression problem. We used linear regression model (lasso,ridge), SVR model, gradient boosting algorithm (LightGBM), ElasticNet regression.

- Cross-validation: We have done k-fold cross validation to develop our model. We took the value of k=10 as a result the data is divided into 10 subsets. Now the holdout method is repeated 10 times, such that each time, one of the k subsets is used as the test set/ validation set and the other k-1 subsets are put together to form a training set. The error estimation is averaged over all k trials to get total effectiveness of our model.

- Ensemble: Ensemble is the technique that combines several base models in order to produce one optimal predictive model. We used various model in our project and ensemble to increase the predictive power. We have seen that after using ensemble method there was a great improvement in RMSE score (decreased which we wanted).

Result & Discussion

We submitted our solution 15 times and by each submission the rmse score decreased gradually. We did a lot of experiment in our solution for this reason the number of submission increased. All submission and their RMSE score are given below

Submission and Description	Public Score	Use for Final Score
submission.csv 15 days ago by GirlsPower add submission details	0.11420	☐
submission.csv 15 days ago by GirlsPower add submission details	1.83201	☐
submission.csv 15 days ago by GirlsPower add submission details	40.03412	☐
submission.csv 15 days ago by GirlsPower add submission details	Error	☐
submission.csv 15 days ago by GirlsPower add submission details	0.11382	☑
submission.csv 15 days ago by GirlsPower add submission details	0.11426	☐
submission.csv 23 days ago by GirlsPower add submission details	0.11440	☐
submission.csv 23 days ago by GirlsPower add submission details	0.11413	☐
submission.csv 23 days ago by GirlsPower add submission details	0.11475	☐
submission.csv 23 days ago by GirlsPower add submission details	0.11470	☐
submission.csv 23 days ago by GirlsPower add submission details	0.11435	☐
submission.csv a month ago by GirlsPower add submission details	0.11393	☐
submission.csv a month ago by GirlsPower add submission details	0.12590	☐
submission.csv a month ago by GirlsPower add submission details	0.11585	☐
submission.csv a month ago by GirlsPower add submission details	0.11445	☐
submission.csv a month ago by GirlsPower First submission	0.12260	☑