House Prices Prediction

Group 8
Helen Zheng, Naman Agrawal, Sneha Thomas, Daniel Sampreeth Eadara.
Department of Software Engineering, San José State University

San José, CA

helen.zheng@sjsu.edu, naman.agrawal@sjsu.edu , sneha.thomas@sjsu.edu ,
danielsampreethreddy.eadara@sjsu.edu

Abstract

Everyone looks to buy themselves a dream house. But there are a lot of factors involved which affect the decision of buying a house. There is a need for obtaining the sale prices of houses based on factors such as location, facilities etc.This is an open problem to predict housing prices that can be found on Kaggle. This is a classic problem to predict data based on a given dataset and can be extended to other problems. We hope to study how various methods such as linear regression, multivariate linear regression, random forest regressor, etc. can aid in the prediction using the code. In addition, we will attempt to incorporate additional information such as crime data, school data etc., and use the models on different housing datasets.

Introduction

In this problem, which can be found on Kaggle as an open problem, the goal is to predict housing prices based on an existing dataset with attributes that describe the houses, such as garage size, neighborhood, condition, etc. Sales price is given as part of the training dataset, and the goal is to predict the sales price of the test dataset.

This is a classic problem to predict data based on a given dataset and features data to train a model as well as data to test on. We are planning to apply various methods such as linear regression, multivariate linear regression, random forest regressor etc. to see which of the methods gives a better result.

Currently, there are many notebooks in Kaggle revolving around various preprocessing methods, feature engineering and prediction algorithms. Many algorithms such as Ridge regression, Lasso regression, ElasticNet regression, Kernel Ridge regression and ensemble methods such as gradient boosting, XGBoost, lightGBM etc were used.

In addition to few of the above mentioned models, we will attempt to incorporate additional information, such as crime rate, distance to schools, and income level. We will also attempt to use the models on different housing datasets, such as in Boston and Manhattan.

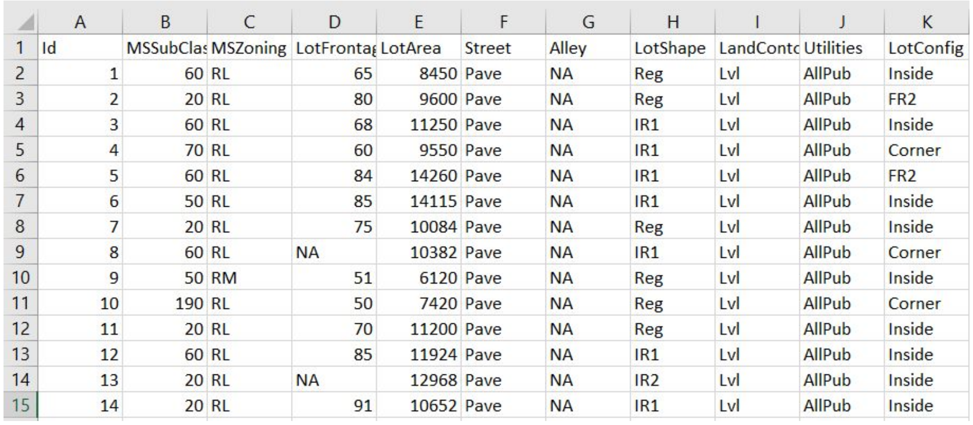
Methods

Data

We found the coordinate data of schools and the houses and used a Google API to find the distance between each data. We incorporated this distance data to the Ames data set. We also plotted the sales price by month and tried to compare it to the average income in Iowa and the U.S. We also used yearly crime rate data of Ames.

The primary data set consists of a training data set and a testing dataset. Both the training and testing datasets have about 80 features like quality, neighbourhood, area etc. The training dataset additionally contains the sale price of each house.

Dataset:

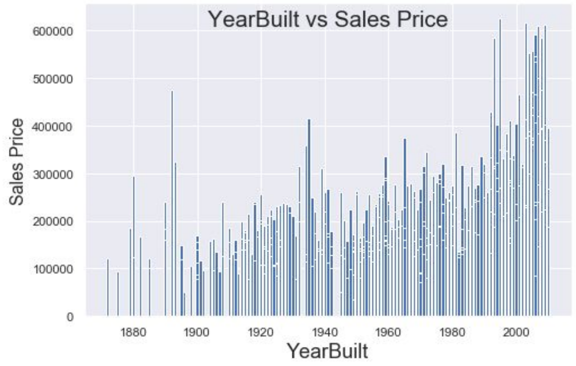


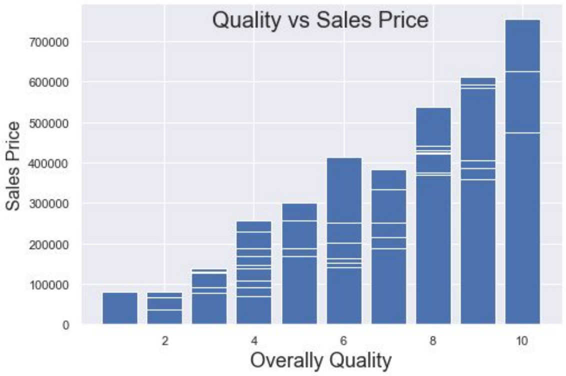
	A	B	C	D	E	F	G	H	I	J	K
1	Id	MSSubClas	MSZoning	LotFrontag	LotArea	Street	Alley	LotShape	LandContc	Utilities	LotConfig
2	1	60	RL	65	8450	Pave	NA	Reg	Lvl	AllPub	Inside
3	2	20	RL	80	9600	Pave	NA	Reg	Lvl	AllPub	FR2
4	3	60	RL	68	11250	Pave	NA	IR1	Lvl	AllPub	Inside
5	4	70	RL	60	9550	Pave	NA	IR1	Lvl	AllPub	Corner
6	5	60	RL	84	14260	Pave	NA	IR1	Lvl	AllPub	FR2
7	6	50	RL	85	14115	Pave	NA	IR1	Lvl	AllPub	Inside
8	7	20	RL	75	10084	Pave	NA	Reg	Lvl	AllPub	Inside
9	8	60	RL	NA	10382	Pave	NA	IR1	Lvl	AllPub	Corner
10	9	50	RM	51	6120	Pave	NA	Reg	Lvl	AllPub	Inside
11	10	190	RL	50	7420	Pave	NA	Reg	Lvl	AllPub	Corner
12	11	20	RL	70	11200	Pave	NA	Reg	Lvl	AllPub	Inside
13	12	60	RL	85	11924	Pave	NA	IR1	Lvl	AllPub	Inside
14	13	20	RL	NA	12968	Pave	NA	IR2	Lvl	AllPub	Inside
15	14	20	RL	91	10652	Pave	NA	IR1	Lvl	AllPub	Inside

Data Visualization

Python libraries such as seaborn and matplotlib were used for data visualization using bar graphs and histograms, distplot, scatter and pairplot.

They were used for comparison of sales price with features such as year built, number of rooms, area, quality etc. The skewness of the data was visualized using distplot and pairplot was used to visualize the most correlated features to the sales price.

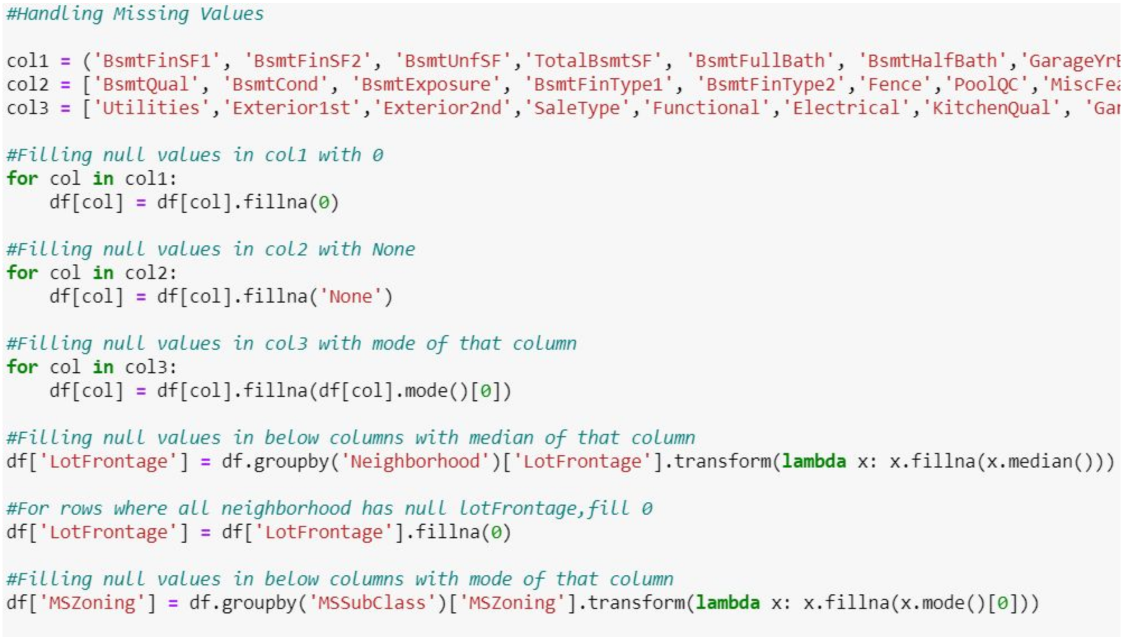




Preprocessing

The dataset we used as input had a lot of null values,outliers and was showing some skewness.Since we want data that is as normally distributed as possible,we tried to reduce the skewness of those columns. Also all the null values, missing values and outliers were handled accordingly.

Code snippet of preprocessing:



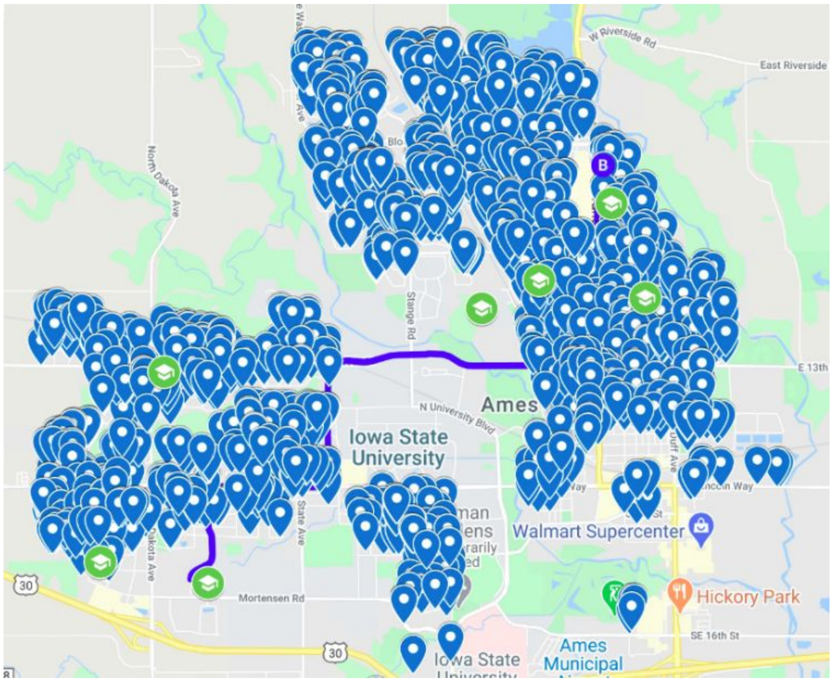
```
#Handling Missing Values

col1 = ('BsmtFinSF1', 'BsmtFinSF2', 'BsmtUnfSF','TotalBsmtSF', 'BsmtFullBath', 'BsmtHalfBath','GarageYr
col2 = ['BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtFinType2','Fence','PoolQC','MiscFe
col3 = ['Utilities','Exterior1st','Exterior2nd','SaleType','Functional','Electrical','KitchenQual', 'Ga

#Filling null values in col1 with 0
for col in col1:
    df[col] = df[col].fillna(0)

#Filling null values in col2 with None
for col in col2:
    df[col] = df[col].fillna('None')

#Filling null values in col3 with mode of that column
for col in col3:
    df[col] = df[col].fillna(df[col].mode()[0])

#Filling null values in below columns with median of that column
df['LotFrontage'] = df.groupby('Neighborhood')['LotFrontage'].transform(lambda x: x.fillna(x.median()))

#For rows where all neighborhood has null lotFrontage,fill 0
df['LotFrontage'] = df['LotFrontage'].fillna(0)

#Filling null values in below columns with mode of that column
df['MSZoning'] = df.groupby('MSSubClass')['MSZoning'].transform(lambda x: x.fillna(x.mode()[0]))
```

Feature Hacking/Selection

Since we have almost 80 features to base our house price prediction,we are using PCA(Principal Component Analysis) and LDA(Linear Discriminant Analysis) to reduce the curse of dimensionality and improve our model accuracy.We are also using correlation matrix to identify the relationship between different features and to identify the top features that contribute to sale price of the house.These three will help us in significantly improve our model accuracy by focusing more on the best features.

School Data using Google API

We found out the geo locations of the houses from the PID (parcel ID) given in Property search forms of the city of Ames and hence we were able to get geocoded data of the houses. We found out the district schools in the city of Ames and then plotted the houses and schools on google map. We calculated the driving distance of each house from all the schools using the distance matrix API of Google Maps.



Prediction Model

Predicting house prices is a regression problem wherein we will have to predict the house prices,given a set of attributes like quality,year built,square feet etc.So we will be using the below regression algorithms to solve this problem:

1. **Multivariable Linear Regression**:This algorithm will help us in utilizing the linear relationship between the independent variables and dependent variable ‘SalePrice’.Even though it's a simple algorithm it can provide good

accuracy rates.In addition,we can use regularization methods with this algorithm to prevent overfitting of data.

2. **Gradient boosting regression:**This algorithm will help in boosting weaker models by optimizing the loss functions.
3. **Random Forest Regressor :**RandomForestRegressor uses multiple decision trees and averages the value of these trees to generate a final output.Since this algorithm uses random sampling of data and random subfeatures to generate trees and their nodes,it is less prone to overfitting.
4. **Support vector regressor :** Support Vector Regression has a high capability of generalization and can result in high accuracy rates on unseen data.
5. **XGBoost**: It is an implementation of gradient boosted decision trees. It involves a repeating cycle of calculating errors, building models, predicting errors and adding the last model to the ensemble.
6. **Ridge, Lasso:**These algorithms are used to regularize the linear regression model and to improve the prediction accuracy of the model.

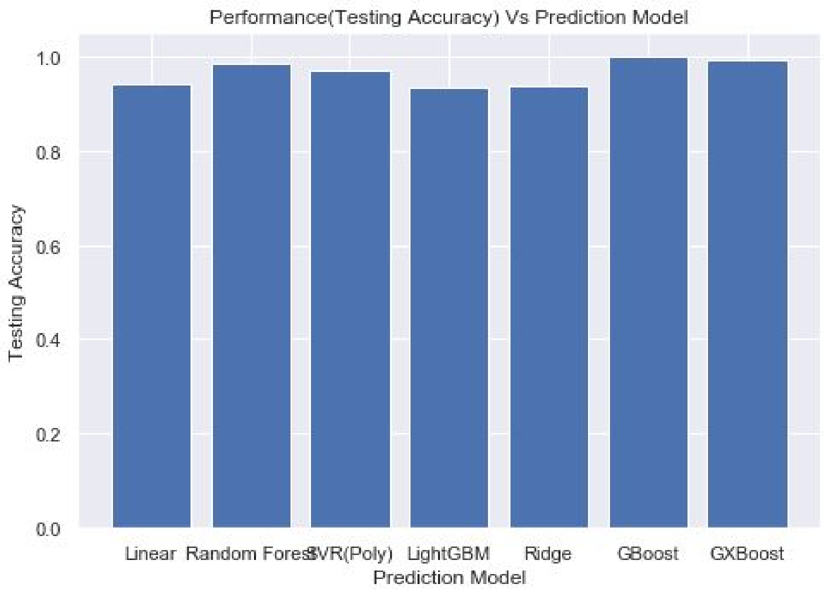
Model Training and Evaluation

We will be splitting the data in such a way that 80% data will be used for training and remaining 20% data will be used for testing.We will also be doing cross validation with 10 folds to make sure that the model will not show overfitting.

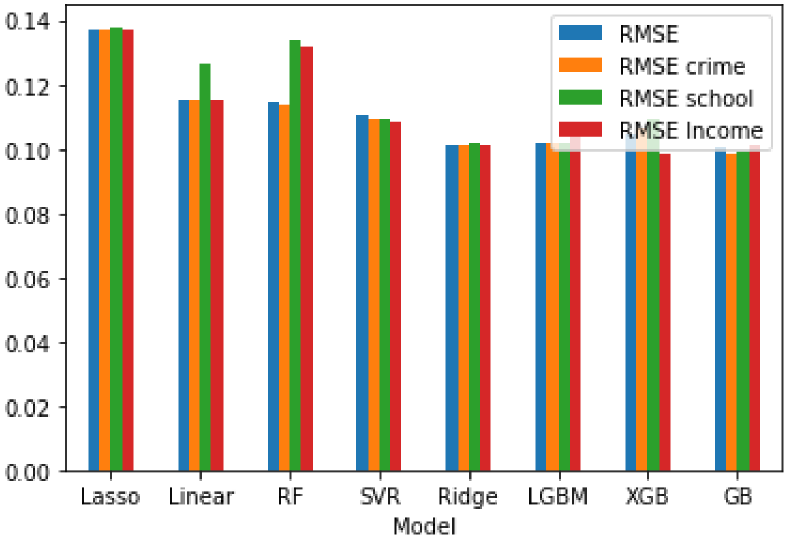
Accuracy scores like R2,RMSE will be used to evaluate the performance of the model.

Results

The results obtained for the testing accuracy of the models are as follows:



The following are the basic RMSE scores, RMSE score with crime data included and the RMSE score with both crime and school data included :



Comparisons

The table portraying the comparison between results obtained by our models and the results obtained by various users on kaggle using similar algorithms.

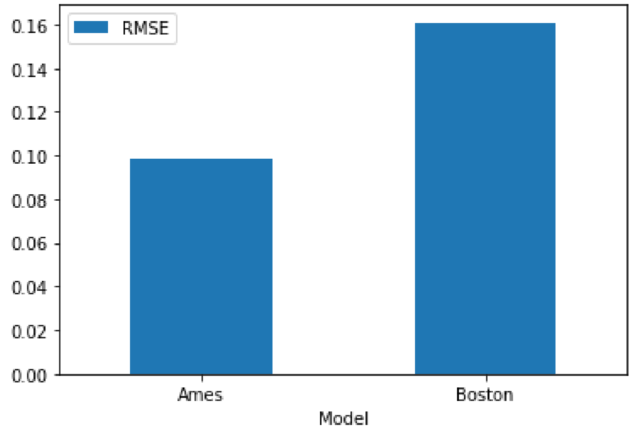
Models/Sl.No	1	2	3	4	Our Results
Multivariable Linear Regression	-	-	-	-	0.1153071381
Gradient Boosting Regression	0.112148	-	0.1177	0.113159	0.0985546494
Random Forest Regression	0.136618	-	-	--	0.1139642362
Support Vector Regression	0.109356	0.1015	-		0.1092381899
XGBoost	0.136361	-	-	0.112392	0.1064203367
Ridge	0.110064	0.1017	0.1153	0.113991	0.101373739934
Lasso	-	0.1016	0.1115	0.111625	

Our comparison of the results of the different algorithms using the Ames Housing data from Kaggle, using the school data, using crime data, and using income data.

	Just Kaggle data	W/ crime data	W/ crime and school data	W/ crime and income data
LightGBM	0.1016237847	0.1018637193	0.1016640585	0.1039711204
Multivariable Linear Regression	0.1153762967	0.11530713811	0.1267046808	0.1153237292
Gradient Boosting Regression	0.1002697702	0.0985546494	0.0989899802	0.1009715625
Random Forest	0.1141464594	0.1139642362	0.1340453253	0.1317121959

Regression				
Support Vector Regression	0.1101738564	0.1092381899	0.1091355600	0.1084691013
XGBoost	0.1047074694	0.1064203367	0.10902111802	0.0985106949
Ridge	0.1013292411	0.1013636882	0.1015310877	0.1013739205
Lasso	0.1372425052	0.1372425052	0.1379613792	0.1370774764

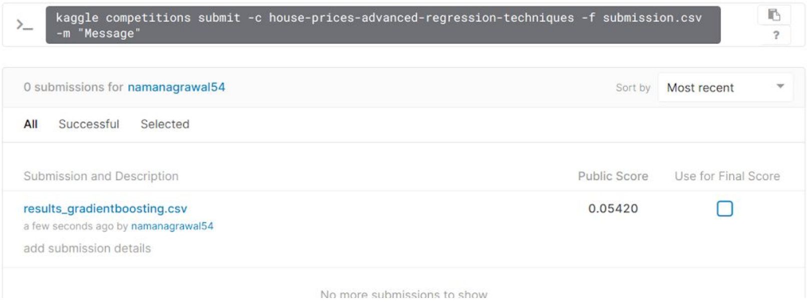
Our implementation of the gradient boosting algorithm on Ames dataset and boston dataset:



Kaggle Competition Result

Our best model submission to kaggle competition landed us on 54th rank out of 4966 entries.

https://www.kaggle.com/c/house-prices-advanced-regression-techniques/leaderboard

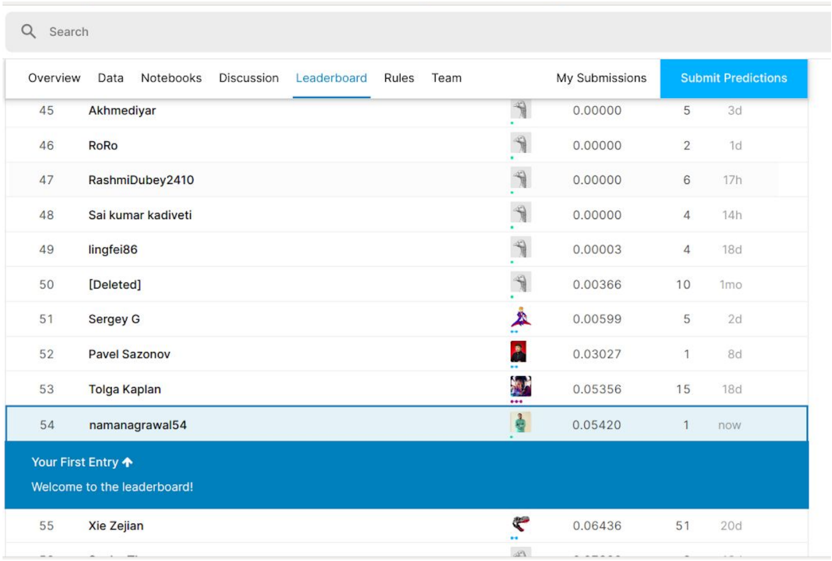


```
kaggle competitions submit -c house-prices-advanced-regression-techniques -f submission.csv
-m "Message"
```

0 submissions for namanagrawal54 Sort by Most recent

Submission and Description	Public Score	Use for Final Score
results_gradientboosting.csv a few seconds ago by namanagrawal54 add submission details	0.05420	□

No more submissions to show



45	Akhmediyar	0.00000	5	3d
46	RoRo	0.00000	2	1d
47	RashmiDubey2410	0.00000	6	17h
48	Sai kumar kadiveti	0.00000	4	14h
49	lingfei86	0.00003	4	18d
50	[Deleted]	0.00366	10	1mo
51	Sergey G	0.00599	5	2d
52	Pavel Sazonov	0.03027	1	8d
53	Tolga Kaplan	0.05356	15	18d
54	namanagrawal54	0.05420	1	now
55	Xie Zejian	0.06436	51	20d

Your First Entry ↑
Welcome to the leaderboard!

Conclusions

We implemented and compared the solutions for the Ames housing data using various algorithms and various additional data. From the results we obtained, we see that the Gradient boosting regression consistently performs better than the other algorithms. This is mostly true despite the different additional data that we added. In fact, adding school data or income data resulted in a less accurate score for the random forest regression algorithm. The best results for Ames were calculated with crime data.

As such, we conclude that for the Ames data set, it is best to use the gradient boosting regression algorithm. As such, we decided to use the gradient boosting regression algorithm for the Boston housing data and compare the results to the Ames housing data. In the comparison between Ames housing data and Boston housing data, we noticed that Ames performed better. This could be due to the fact that we picked an algorithm that was deemed fit for the Ames data. It is possible that for the Boston data, a different algorithm would give the best results for Boston data. It could also be that it is more difficult to predict the sales price for the Boston housing.