**House Price Predictions**

*Pavit Chawla, Gabriela Cortes-Mejia, Beatriz Ruiz, Abubakir Siedahmed, Sana Shaikh*

**Table of Contents**

[**Project Description, Details, and Goals:**](#_hmbliepag57v) **1**

[**Details about the Data:**](#_a42ip1ruw5iz) **1**

[**The developed methods, algorithms, and tools to address the projects requirements:**](#_menk4x9rj6ao) **2**

[Figure 1: 1stFlrSF vs SalePrice](#_c7g6lfdtemg) 4

[Figure 2: YrSold vs SalePrice](#_17c3rkd7od9b) 4

[Figure 3: Prediction value (Linear Regression) vs Label(SalePrice)](#_ouhek1c8nlb) 4

[Figure 4: Prediction value(KNN) vs Label(SalePrice)](#_y45whohasxei) 5

[Figure 5: Prediction value(Random Forest) vs Label(SalePrice)](#_kpgi0ssd3xhs) 5

[Figure 6: Prediction value(Decision Tree) vs Label(SalePrice)](#_o0n0yf9vs4ew) 5

[**The developed codes and final results:**](#_7xwfkv8mtc3l) **6**

[**Errors:**](#_ipq7w299m9z) **6**

[**Team Responsibilities:**](#_ygpazphdqbqg) **7**

# **Project Description, Details, and Goals:**

The goal of our project, House Price Predictions, was to allow us to apply the tools and techniques we learned in our Data Science class. This involved implementing the machine learning algorithms we learned in class in order to extract knowledge from large-scale data. We used the house price dataset that involved many different numerical and categorical features. We used several regression machine learning algorithms, namely Linear Regression, K-Nearest Neighbor, Decision Tree Regressor and Random Forest. We also used Cross Validation to test our model’s ability to predict new data that wasn’t used in estimating it. Finally, our end goal is to predict the final sale price for all house prices listed in our dataset.

# **Details about the Data:**

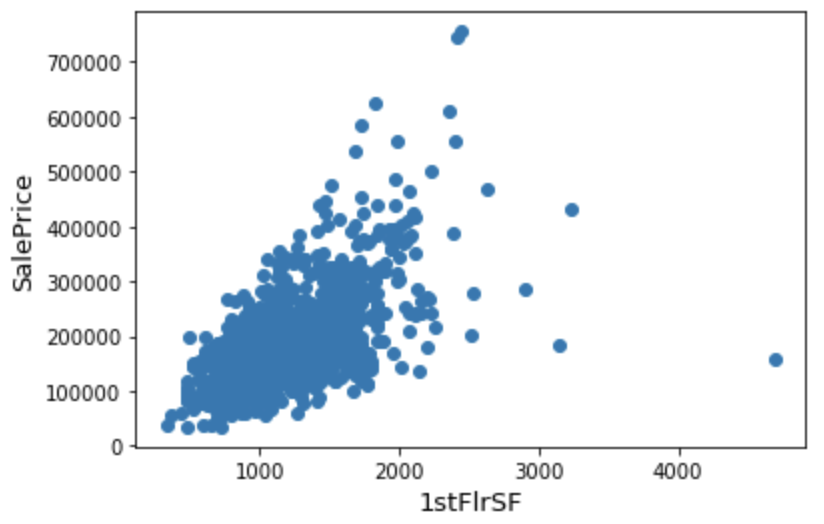
The dataset we chose for this project consisted of multiple numerical and categorical features that consisted of providing details on the specifications of the houses, the type of neighborhood, and sale conditions. Our data set originally had 1460 rows and 81 columns, but once we selected our features our data set shrunk to 1460 rows and 8 columns. For our features we chose six numerical features and two categorical features which had been preprocessed to not include any null values. The six numerical features included MoSold, YrSold, BsmtFinSF1, LotArea, TotalBsmtSF, and 1stFlrSF. We chose two date features, Mosold and YrSold which describe the month and year that the houses were sold. We chose two basement measurement features, BsmtFinSF1 and TotalBsmtSF which measure the final and total square feet of the basement area, respectively. We also chose LotArea which is the lot size in square feet. For our categorical features, we also chose Street and CentralAir. Street is categorized between pave and gravel which we then performed hot-encoding to convert them to numerical values. CentralAir is categorized between yes and no and we also converted them to numerical values by hot encoding them. All of our features were finally scaled and we selected our label which is SalePrice that describes the property’s sale price in US dollars which we used in our prediction. After setting up these features and labels we applied multiple regression algorithms that will help us predict the sale price for the houses.

# **The developed methods, algorithms, and tools to address the projects requirements:**

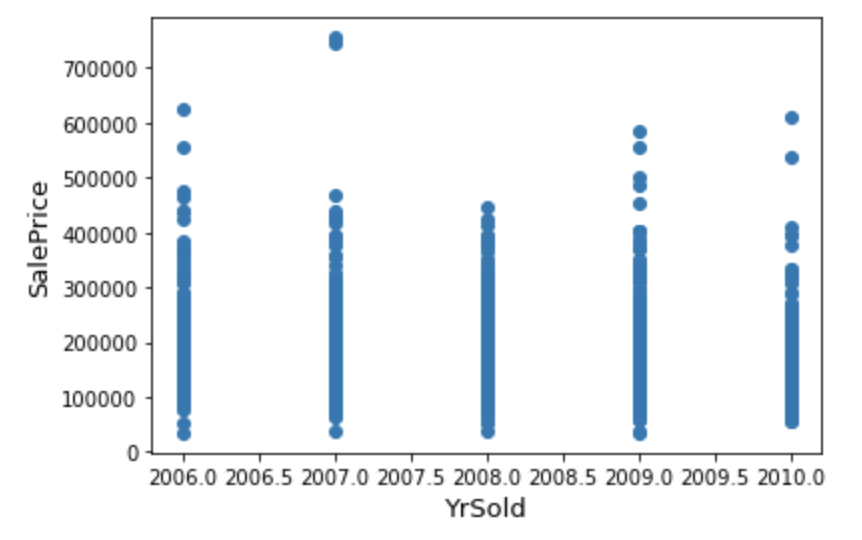
To reach our goal of predicting house sale prices, we selected our features, preprocessed our features, and implemented four different Machine Learning algorithms. When we preprocessed our data set we dropped any null values by using the pandas library function. Furthermore, we selected our features based on how the data was distributed against the label. After selecting our features we hot-encoded our two selected categorical features, and finally we scaled all the features.

From our four machine learning algorithms, we first used the linear regression algorithm and we were able to determine the best and worst features by calculating the coefficients and intercepts. We discovered that our best feature was 1stFlrSF since the coefficient was 23,068.73 and there was a high correlation between the SalePrice and 1stFlrSF that is shown in Figure 1. Our worst feature was YrSold since the coefficient was 443.7 and there was since there is some missing data because there weren’t any houses sold in the second half of the years 2006, 2007, 2008, and 2009 as shown in Figure 2.

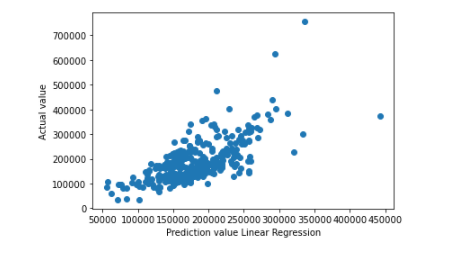
Next, we calculated the root mean square error for linear regression and got 59,098.06. We also used the KNN algorithm with n\_neighbor = 5 and calculated the root mean square error and got 62,466.21. The last two algorithms we implemented were Random Forest with n\_estimators of 3 and random state of 9 and a Decision Tree algorithm with a random state of 5. We calculated RMSE value for Random Forest regressor and the result was 65,574.55 and the RMSE for Decision Tree was 73,217.31.



## Figure 1: 1stFlrSF vs SalePrice



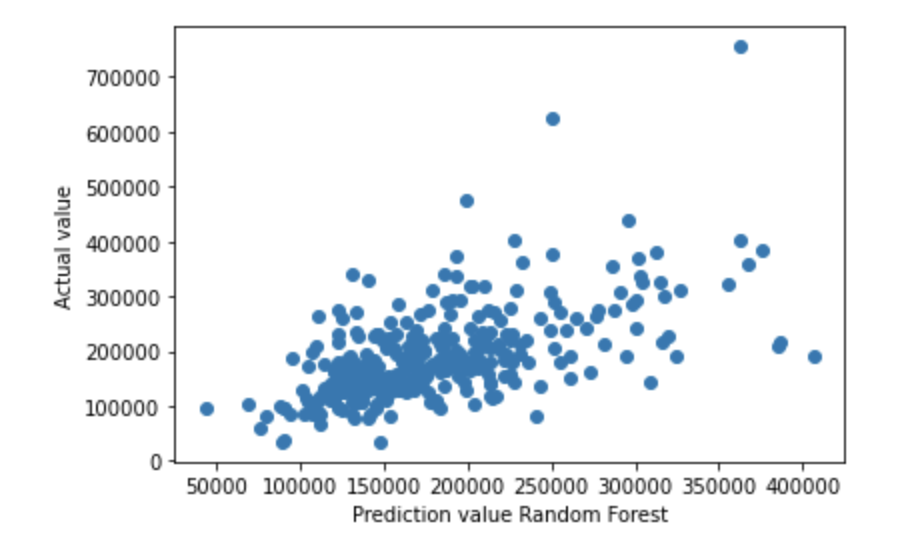
## Figure 2: YrSold vs SalePrice



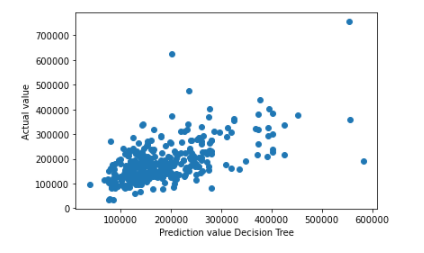
## *Figure 3: Prediction value (Linear Regression) vs Label(SalePrice)*

## 

## Figure 4: Prediction value(KNN) vs Label(SalePrice)



## Figure 5: Prediction value(Random Forest) vs Label(SalePrice)



## Figure 6: Prediction value(Decision Tree) vs Label(SalePrice)

# **The developed codes and final results:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Machine Learning Algorithms** | **RMSE** | **Cross Validation RMSE** | **Error** |
| Linear Regression | 59,098.06 | 59,617.63 | 32% |
| K-Nearest Neighbor | 62,466.21 | 61,025.45 | 33% |
| Random Forest | 65,574.55 | 63,716.70 | 35% |
| Decision Tree | 73,217.31 | 79,676.10 | 35% |

*Figure 7: Data Comparisons Table*

The cross validation RMSE score for the **Linear Regression** algorithm is the lowest at 59,098.06 with the lowest error at 32% making Linear Regression the best algorithm for our model. On the other hand, the worst algorithm for our model is **Decision Tree** with a cross validation RMSE score of 79.676.10 which comes out to an error of 35%.

From our final results seen in *Figure 3* that compares our *Linear Regression* *Prediction Value* against the *Actual Value*, we can conclude that the results for our *Prediction Values* cluster around the best fit line. This means that we were successful in our house sale predictions because they were close to the actual price.

# 

# **Errors:**

Some of the errors that we encountered in our project was the amount of null values in the data set that we had to exclude in order to derive a close prediction. Our data also included many features that contained outliers which if not removed could cause a bias in our model estimates since regression models tend to be sensitive to outliers.

# **Team Responsibilities:**

Among the members of our team, Beatriz was responsible for extracting the features we chose from the large-scale dataset, excluding the null values, splitting the dataset into training and testing, as well as applying the linear regression model. Sana was responsible for scaling the dataset and using the KNN algorithm for the regressor model. Gabriela was responsible for researching and applying the Random Forest algorithm to our dataset. Abubakir was responsible for researching and applying the DecisionTreeRegressor model. Pavit was responsible for applying the cross validation as well as calculating the MSE and RMSE for all the algorithms used in our code.