RealLeads Machine Learning Process

RealLeads Machine Learning Process

Overview

RealLeads is a data analytics group that is working for determine trends and solve problems within the real estate market in New Castle County, Delaware. There are three questions that we explored using data analytics: 1. Can we predict the listing price of a house? 2. Can we predict the price the house would be sold for? 3. Can we determine the date range the house will be sold on? Using machine learning, we can predict the answer to these questions, which will provide insight for home owners looking to sell their home.

For the machine learning model, we created a structure that has an order to how we answer these questions. We first predict the listing price, then the sold price, followed by the date range. This is because each machine learning builds off of the previous in that there is required information from previous models used in the later models. We first train and predict the listing price. Next, we then use the original list price as well as the same columns as the listed price to determine the sold price. Lastly, we take both the original and sold prices to determine if a house will sell in less than or more than two months using the date range model.

Preprocessing

The first step to creating machine learning models is to provide clean and accurate data to the model. If the data is not cleaned or is inacurate, then the machine learning model will perform poorly and provide incorrect information. We begin preprocessing by importing tables from a local database which contains a large quantity of information related to the New Castle County housing market such as listing prices, sold prices, days on market, number of bedrooms, and much more. Once the tables are imported, we begin to merge relevant tables together and drop any unnecessary columns that will not provide value to the machine learning model. Including invaluable columns in the

machine learning model can break the model and/or provide very poor results. For example, if we include a column named *MLSNumber* in the machine learning model, it will not provide any insight to help the machine learning model predict prices or day on market. This is because each row contains a different MLSNumber as they are the unique identifiers that similar to an ID. After dropping and cleaning the data, we removed over 60 columns to only include relevant columns in the machine learning models.

Testing Different Machine Learning Models Stage 1 (List Price)

Before applying our data to a machine learning model, we first have to split the data into training data and testing data. Eighty percent of the data is being used to train the model, while the remaining twenty percent is used to test the model. Next, we begin testing on multiple models. The below models are aiming to predict the suggested listing price of a house. For the listing price model, we used regression models.

RandomForestRegressor

RandomForestRegressor is the first model we used. This model provided the best score compared to other models.

```
{'Training MAE': 13918.640703125378,
 'Valid MAE': 36000.482241047226,
 'Training RMSLE': 0.05892504314025553,
 'Valid RMSLE': 0.14449068100075413,
 'Training R^2': 0.9319242507194563,
 'Valid R^2': 0.7662662073353244}
```

This model scored a 76% accuracy with a mean absolute error of 36000. Our goal was to attain a model that was in the range of 75-85% accuracy because this shows that the model is not overfitted to the training data and can handle any new data added into the dataset.

RidgeRegression

The next model we tested was RidgeRegression. This was done to determine if we could get a better score than the 76% given by the RandomForestRegressor.

```
{'Training MAE': 54234.54448027509,
 'Valid MAE': 54776.6712138502,
 'Training RMSLE': 0.22701393167799477,
 'Valid RMSLE': 0.2233302885756457,
 'Training R^2': 0.3778312783019443,
 'Valid R^2': 0.5470330419945886}
```

Unlike the RandomForestRegressor, the Ridge model performed poorly with an accuracy of 54% and a mean absolute error of 54000.

Linear Regression

We then tried Linear Regression, which is a standard model that is commonly used to perform a basic prediction but after training it on the model, it performed better than the Ridge model but not as well as the RandomForest model.

```
{'Training MAE': 50107.73711437222,
 'Valid MAE': 50247.51694407945,
 'Training RMSLE': 0.20385248922177895,
 'Valid RMSLE': 0.20542860874220684,
 'Training R^2': 0.4312283506583967,
 'Valid R^2': 0.5977780391427383}
```

The Linear Regression model performed with 59% accuracy and a mean absolute error of 50000.

Lasso

Lasso is another type of machine learning model that we used to train the model, but it performed similarly to the Linear Regression Model.

```
{'Training MAE': 50107.653529046314,
 'Valid MAE': 50247.47905151865,
 'Training RMSLE': 0.203851626725064,
 'Valid RMSLE': 0.2054272249748278,
 'Training R^2': 0.4312283478546358,
 'Valid R^2': 0.5977798663258096}
```

The Lasso model performed with 59% accuracy and a mean absolute error of 50000.

Elastic Net

```
{'Training MAE': 51343.51878851374,
 'Valid MAE': 51866.431747701004,
 'Training RMSLE': 0.21146741406449818,
 'Valid RMSLE': 0.21104379302564175,
 'Training R^2': 0.41375687834513364,
 'Valid R^2': 0.5818660027234267}
```

The Elastic Net model performed with 58% accuracy and a mean absolute error of 51866.

Bayesian Ridge

```
{'Training MAE': 50101.30937751924,
 'Valid MAE': 50263.66617660255,
 'Training RMSLE': 0.2038489648815703,
 'Valid RMSLE': 0.20534557332029352,
 'Training R^2': 0.4310472902536484,
 'Valid R^2': 0.5975992390240552}
```

The Bayesian Ridge model performed with 58% accuracy and a mean absolute error of 50000.

Machine Learning Stage 2 (List Price)

With RandomForestRegressor model being the best option, we decided to attempt one more method which is scaling our data using the StandardScaler method. StandardScaler is used to scale data to the point where the all numbers are in close range with each other. For example, house prices can be listed as 300,000 dollars, but the number of bedrooms are 3, which indicates a wide range of numerical values, although the data columns are not the same. Scaling tackles this issue by changing all the numbers to very small numbers to keep the data normalized, which improves machine learning model capabilities.

We applied the RandomForestRegressor to the scaled data, and resulted with similar scores as the unscaled model.

```
{'Training MAE': 13980.207375384318,
 'Valid MAE': 36191.55754065041,
 'Training RMSLE': 0.059310705683357653,
 'Valid RMSLE': 0.1479598434134952,
 'Training R^2': 0.9317341276848287,
 'Valid R^2': 0.7502608760686673}
```

Though the scaled model performed marginally close to unscaled model, we decided to use this scaled model for the rest of the predictions to maintain consistency.

Machine Learning Stage 3 (Sold Price)

The next step in the analysis was to predict the sold price of the house. We added the original list price to our existing data table to help determine the prediction for what price the house would be sold for. The reason we did not use the predicted list price, for this case, is because we wanted the model to use the existing data given to provide the best learning experience. Once the model learned from the exising data, in production, a home owner can then predict their listed price. Based on the list price prediction, we will then be able to predict the expected sold price of the house.

RandomForestRegressor

We used the RandomForestRegressor again to predict the sold price because of how well it performed on the previous prediction. In addition, the RandomForestRegressor has some useful built in functions that help visualize important correlations between the data. After training the model with our scaled data which included the original list price, it performed remarkably well.

```
{'Training MAE': 9536.110439029704,
 'Valid MAE': 25638.233789198606,
 'Training RMSLE': 0.04081425100179144,
 'Valid RMSLE': 0.10493354561205767,
 'Training R^2': 0.9815139753111272,
 'Valid R^2': 0.8620385960793743}
```

The model had an accuracy of 86% and a mean absolute error of 25000.

Machine Learning Stage 4 (Days On Market)

The next step in the analysis was to predict the days on market for the house sale. After attempting to find the days on market as a value, the accuracy was too low with the given data, so we decided to bucket days on market data. With trial and error, we found that bucketing into two groups, less than/more than 2 months provided an accurate prediction. We split the data into training data and testing data with an 80/20 split, and began testing on multiple classification models. Balanced Random Forest Classifier after dropping columns was the model that provided the most accuracy.

Balanced Random Forest Classifier

The Balanced Random Forest Classifier model performed with 70% accuracy after dropping unncessary columns. Prior to dropping these columns, the accuracy was 63%

Easy Ensemble Classifier

The Easy Ensemble Classifier model performed with 44% accuracy, and only 37% accuracy after dropping columns.

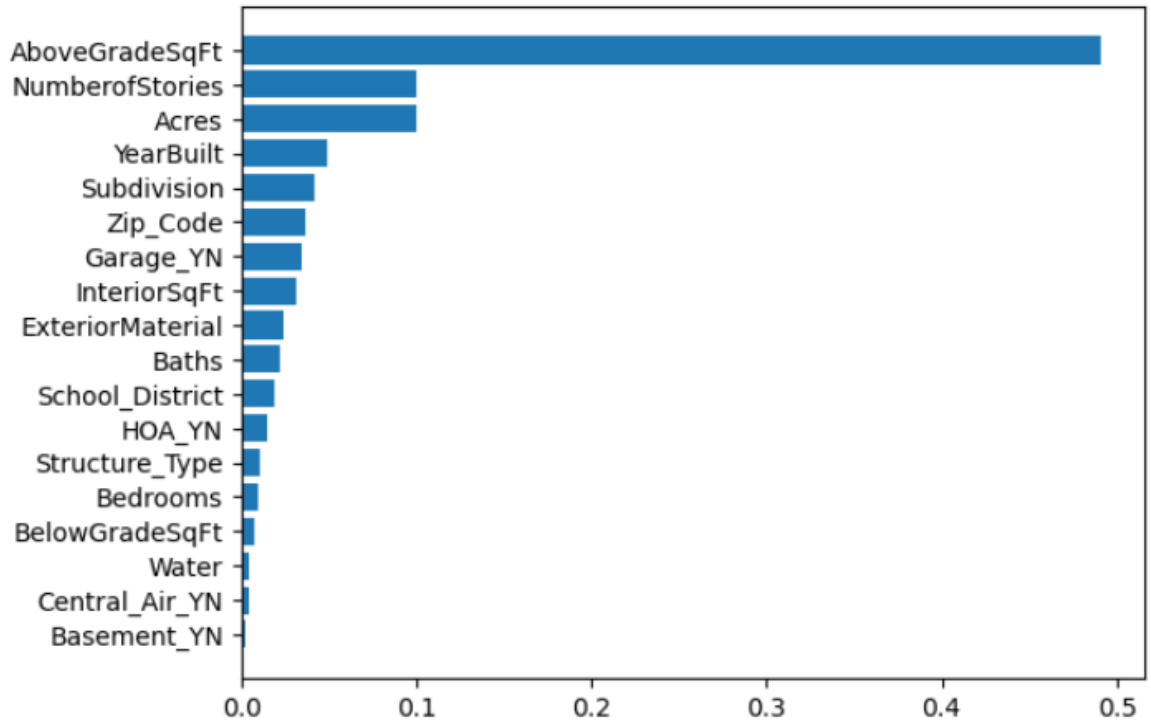
Machine Learning Stage 5 (Unsupervised Learning K-Means)

This step was optional for the RealLeads team, but we were determined to find any correlation with all data points by clustering using unsupervised learning. We used a different data table than the supervised macine learning models in that we included majority of columns for the unsupervised model. Once the data was cleaned, we used a method called the Principal Component Analysis (PCA), which performs a dimensionality reduction on our data table and normalizes the data points. The PCA had 3 components that it reduced the data to which then was used to determine the amount of clusters that would be optimal using the elbow curve method. The elbow curved method uses the K-Means model and goes through a range of clusters to find which number cluster is the best for the data. In this case, four clusters was the most optimal.

After analyzing the clusters, it was difficult to indicate if there was any correlation with the data. The clusters were close together, but were separated appropiately. After looking at some visualizations to explain the data, we were unable to determine any strong correlations between the data.

Results

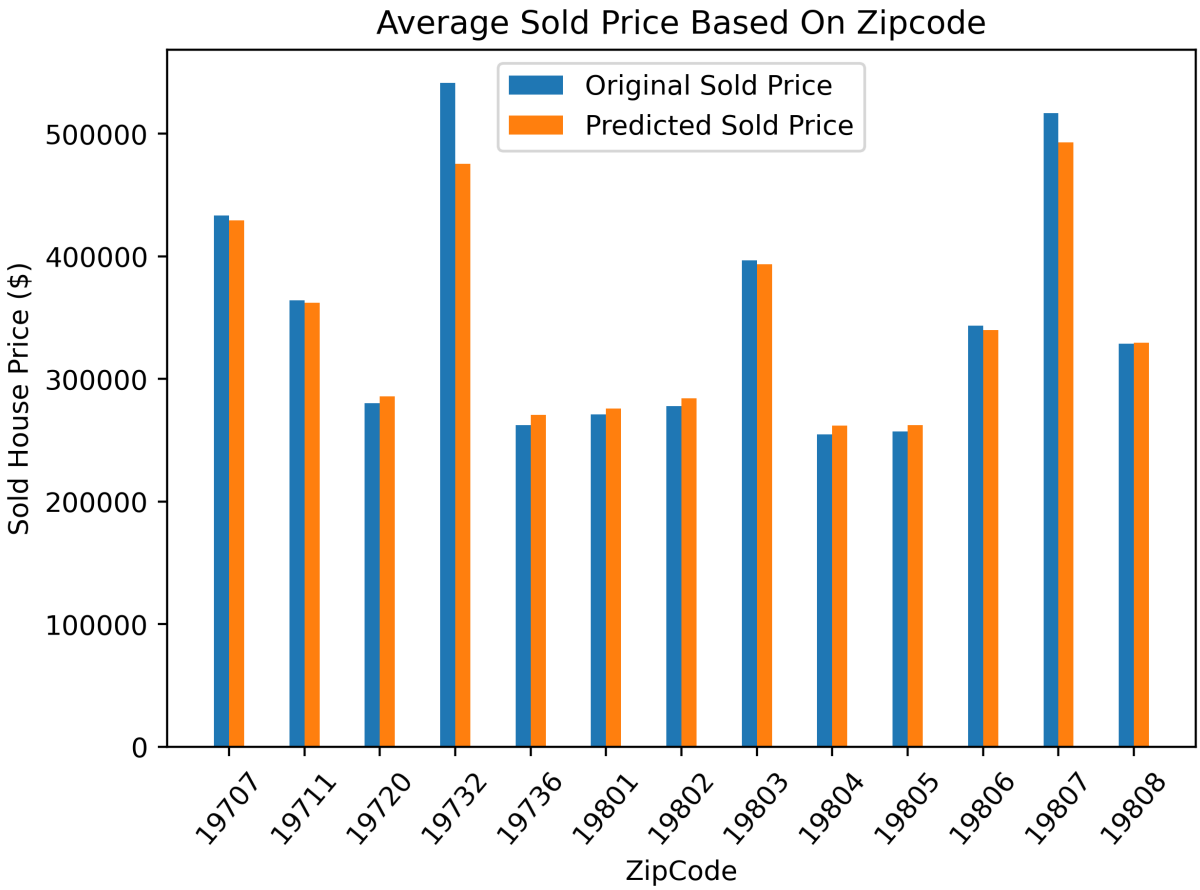
Here are some snippets of the results of all the machine learning model stages:

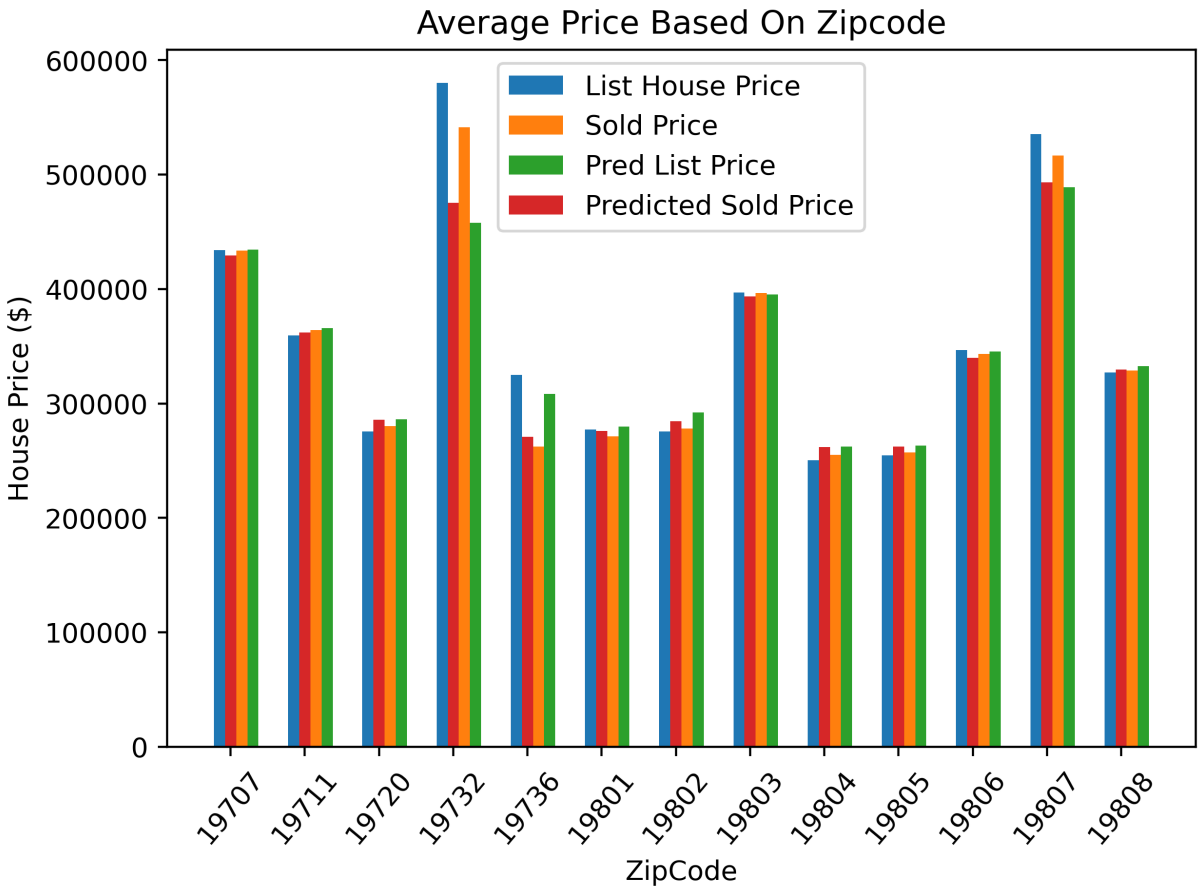


This shows the what features that had the most significant impact on predicting the list price.

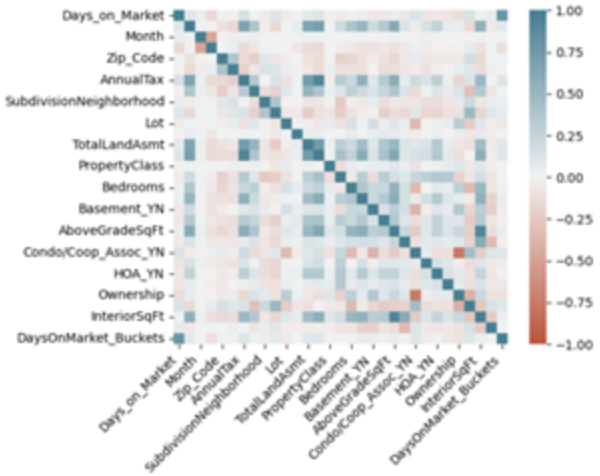
Zip_Code	Orig_List_Price	Pred_Orig_List_Price	Sold_Price	Pred_Sold_Price	Average_Diff_Orig_Price	Average_Diff_Sold_Price
19707	434042.989091	434512.721479	433419.174545	429190.638255	-469.732388	4228.536291
19711	359394.520101	365734.074506	364365.370603	362193.238575	-6339.554405	2172.132028
19720	275412.582503	286382.239500	280407.118494	285786.571963	-10969.656997	-5379.453469
19732	580000.000000	457743.610000	541333.333333	475516.333333	122256.390000	65817.000000
19736	325000.000000	308262.200000	262500.000000	270759.000000	16737.800000	-8259.000000
19801	277116.125000	279961.370278	271316.947917	275937.475417	-2845.245278	-4620.527500
19802	275578.648536	291996.678940	278018.184100	284432.641450	-16418.030404	-6414.457350
19803	396896.794286	395115.382825	396682.449524	393590.894743	1781.411460	3091.554781
19804	250353.475783	262391.690503	254941.646724	261974.714716	-12038.214720	-7033.067992
19805	254758.148699	263105.338451	257122.949814	262531.638483	-8347.189752	-5408.688669
19806	346723.528169	345367.519073	343392.080986	339920.776231	1356.009096	3471.304755
19807	535367.924528	489039.441132	516774.528302	493167.010943	46328.483396	23607.517358
19808	327078.900468	332683.859578	328655.306792	329551.717477	-5604.959110	-896.410685

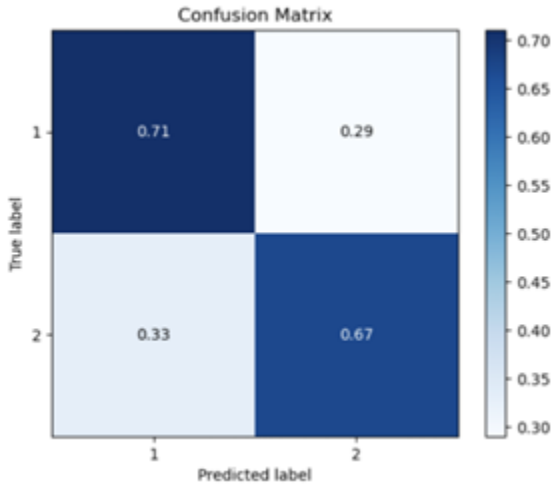
This is an overview of the average prices of houses based on zipcode. The table also shows the difference between the original prices of the houses and the predicted prices from the model to show the difference between them.

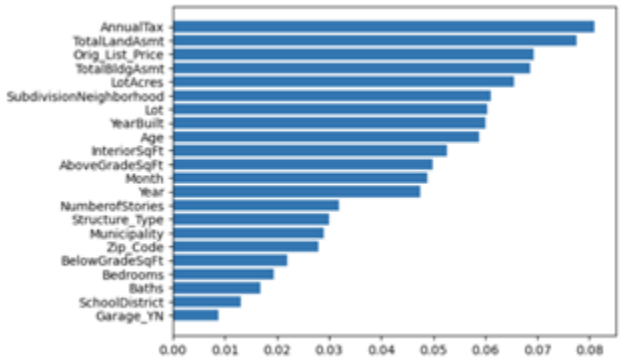

Average Sold Price Based On Zipcode
Original Sold Price
Predicted Sold Price
Sold House Price ($)
0
100000
200000
300000
400000
500000
19707
19711
19720
19732
19736
19801
19802
19803
19804
19805
19806
19807
19808
ZipCode



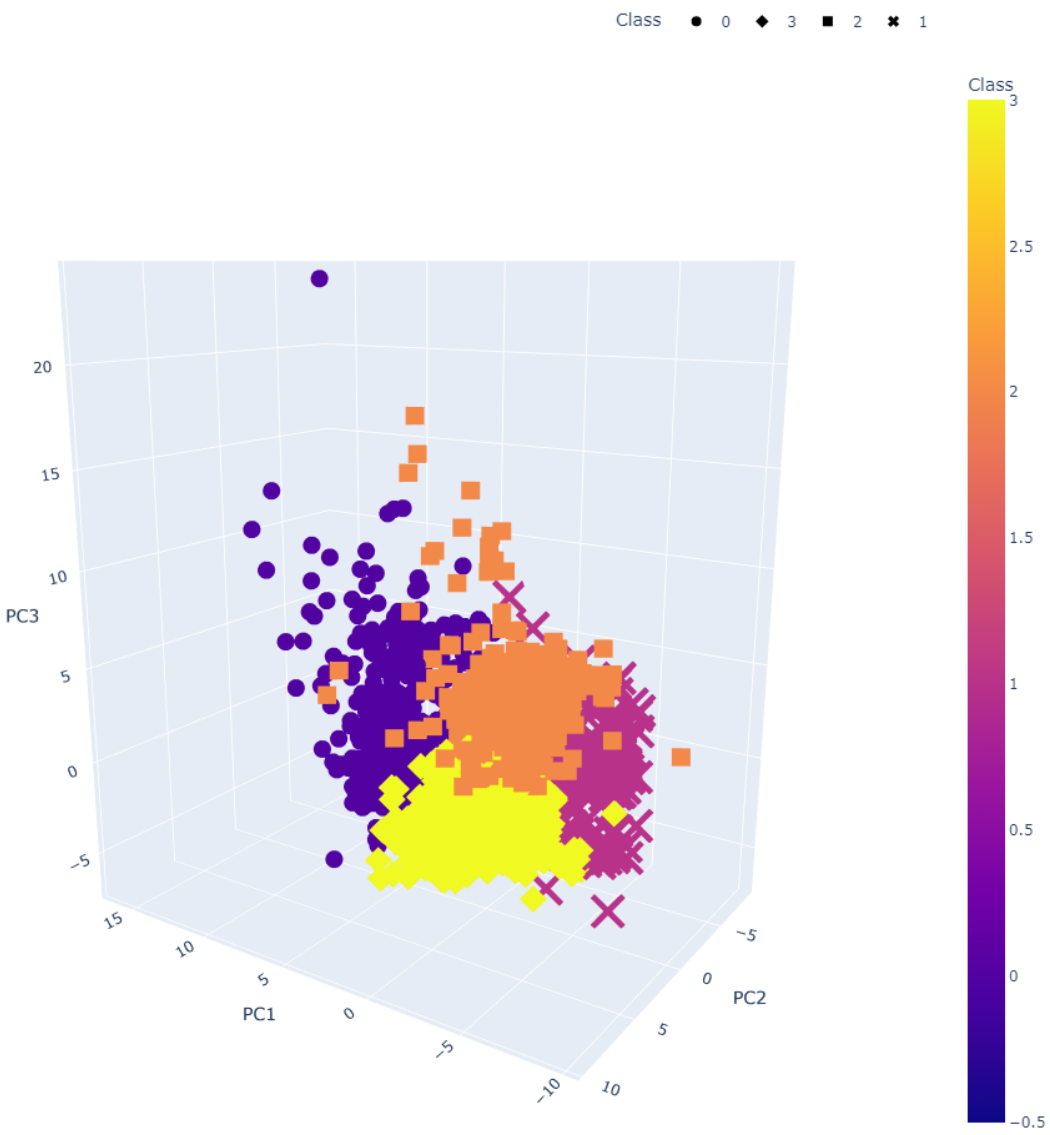
The above three graphs show the difference between the original prices and the predicited prices.



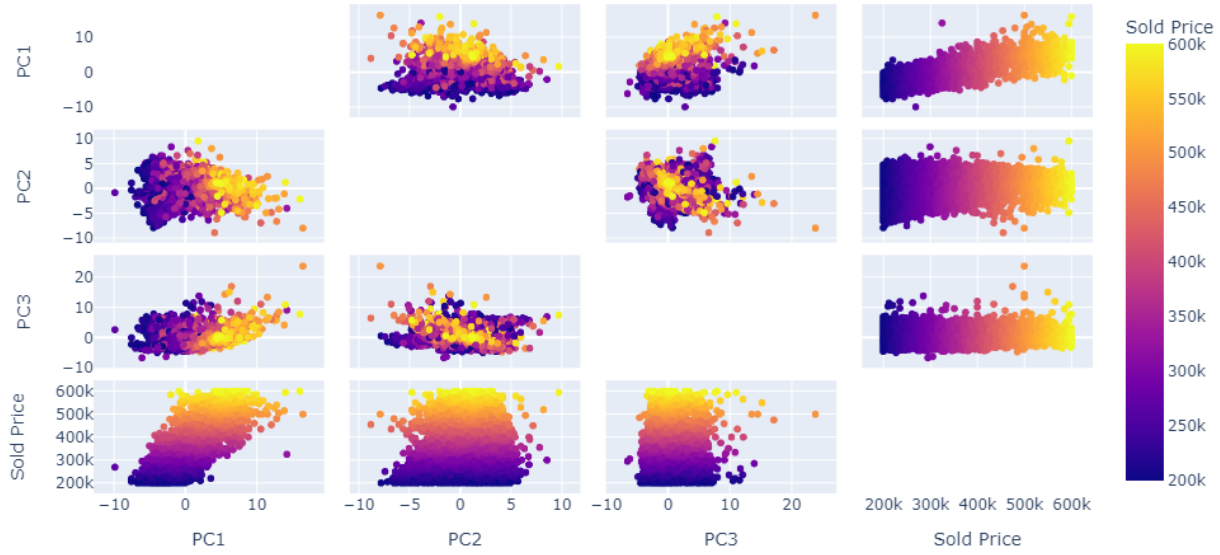




The above three graphs shows the results of the Days On Market Machine learning model.



This cluster graph shows all the datapoints in the data table and the appropiate classes they belong to.



This shows the sold price of houses, each PCA component and correlation.