School of CMS

Cover Sheet for Coursework

Module Code: CO3093

Assignment: CW1 Regression and clustering

Surnames (in CAPITALS): MONUDDIN, OSIBEMEKUN, MARTINOVA

First names (in CAPITALS): TANIA, FEMI, MARTINA

ID Numbers: 209017703, 209029399, 219045339

I understand that this is a piece of coursework. I confirm that this work is mine and that I am fully aware of the Academic Integrity declaration within the school of Informatics accessible here.

Date: 23/03/2023

Signatures: TANIA, FEMI, MARTINA

REPORT

Introduction

In this report we will be analysing and exploring the Housing dataset that records the sale prices of the properties in Manhattan between the years of August 2012 and August 2013. Our aim is to predict the sale prices of the houses showing the relevant features of the dataset and the k-means approach. This will improve the performance of the regression model by creating cluster-based regression models. We will develop a model that will predict the sale prices of the houses based on the data we have.

Data Cleaning and Transformation

Data cleaning is the first and one of the most crucial steps to create an accurate predictive model since it ensures that the data is accurate and reliable. The quality of the result is heavily dependent on the quality of the data that is being analysed. Data cleaning typically involves treating missing values, dropping duplicates, removing outliers, etc.

After inspecting the data, we noticed that the first three rows were filled with irrelevant text, as shown in Figure 1. As a result, we started the data cleaning procedure by eliminating those three rows and using the fourth row as the data frame's index. This allowed us to give an appropriate header to each column, making the data frame more readable and usable (Figure 2).

Moreover, some of the available columns were misspelt or were not named appropriately. Those columns have been renamed. An example of such a case is “SALE/nPRICE”, which has been renamed to “SALES PRICE”.

Figure 1 - Data Frame before the cleaning of the data

	Manhattan Rolling Sales File. All Sales From August 2012 - August 2013.	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8
0	Sales File as of 08/30/2013 Coop Sales Files ...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	Neighborhood Name 09/06/13, Descriptive Data i...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	Building Class Category is based on Building C...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	BOROUGH	NEIGHBORHOOD	BUILDING CLASS CATEGORY	TAX CLASS AT PRESENT	BLOCK	LOT	EASE-MENT	BUILDING CLASS AT PRESENT	ADDRESS
4	1		13 CONDOS - ELEVATOR APARTMENTS		738	1306			345 WEST 14TH STREET

Figure 2 - DataFrame after cleaning of the data

3	BOROUGH	NEIGHBORHOOD	BUILDING CLASS CATEGORY	TAX CLASS AT PRESENT	BLOCK	LOT	EASE-MENT	BUILDING CLASS AT PRESENT	ADDRESS
4	1		13 CONDOS - ELEVATOR APARTMENTS		738	1306			345 WEST 14TH STREET
5	1		13 CONDOS - ELEVATOR APARTMENTS		738	1307			345 WEST 14TH STREET
6	1		13 CONDOS - ELEVATOR APARTMENTS		738	1308			345 WEST 14TH STREET
7	1		13 CONDOS - ELEVATOR APARTMENTS		738	1309			345 WEST 14TH STREET
8	1		13 CONDOS - ELEVATOR APARTMENTS		738	1310			345 WEST 14TH STREET

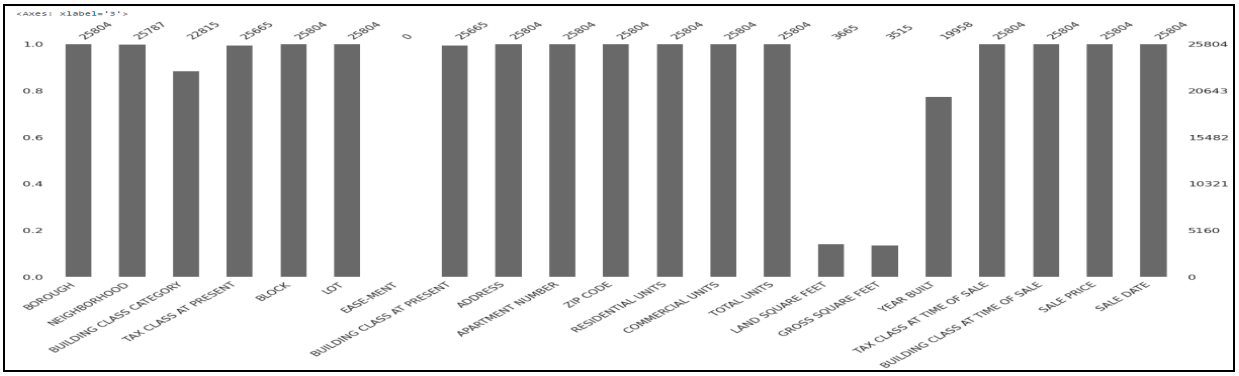
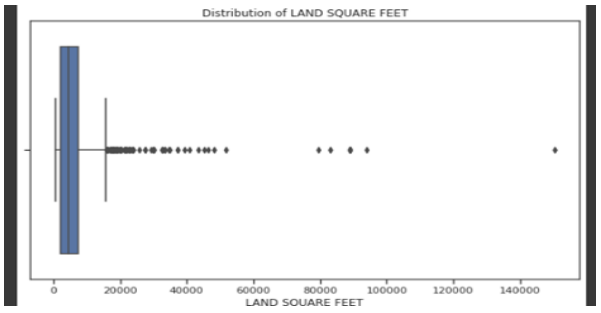


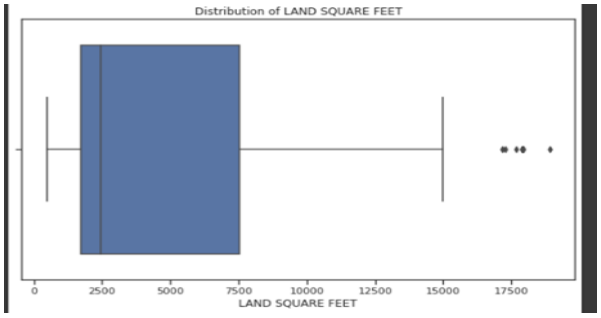
Figure 3 - Graph showing the missing values prior to cleaning

We then continued our cleaning process by dropping columns that would not help the model. The data frame can then be made better suited for analysis as a result. Then, each column's data types are changed from "object" data types to numerical or categorical data types. Examples of such data types include integers, floats and categories as shown in (Figure 6).

In some cases, empty values have been replaced in categorical columns and numerical columns with “np.nan”, which allows us to identify them as missing values and therefore treat them; in other cases we have replaced them with the median, for example in “YEAR BUILT”. This was done to reduce the impact of the outliers. It is important to treat the missing values since the predictive model will not be able to handle them and the model would therefore not function as intended.

Figures 4 ,5 - Boxplot of Distribution of ‘LAND SQUARE FEET’ before and after, respectively, removal of outliers





Outliers were also removed from the data frame in order to increase the accuracy of the model and remove the extreme values in the dataset, these mostly occur due to errors in data collection or are extremely rare values and these can cause the estimates of the model to be biassed or inaccurate. Removal of the outliers increases the visualisation of the data and makes It easier for us to identify the patterns and trends in it. To show the effects of the removal of outliers, two box plots have been shown below in Figures 4 and 5, to illustrate the before and after the removal of outliers. All the extreme data have been removed which makes the box plot of the distribution easier to read and understand. The minimum, medium and medium values are easily identifiable. If the data set has outliers

the box chart may not show the minimum or maximum value. Instead, the ends of the whiskers represent one and a half times the interquartile.

Figure 6 - The Data Frame after dropping columns and defining data types

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 17893 entries, 4 to 27395
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   RESIDENTIAL UNITS  17893 non-null  int64
 1   COMMERCIAL UNITS   17893 non-null  int64
 2   TOTAL UNITS        17893 non-null  int64
 3   LAND SQUARE FEET   2246 non-null   float64
 4   GROSS SQUARE FEET  2251 non-null   float64
 5   YEAR BUILT         17893 non-null  float64
 6   log of SALE PRICE  17893 non-null  float64
dtypes: float64(4), int64(3)
memory usage: 1.1 MB
None
```

We removed columns that would not benefit the model in order to make the data frame more suited for analysis, as shown in Figure 6, since our goal was for the dataset to determine how the square foot of the land and other categories impact the sale price. This was then followed by the change of data types (as explained in the first section of the report) in order to minimise errors while cross-validating the data frame's columns for the model's accuracy, and creating graphs for visualisation purposes.

'SALE PRICE' column was removed and replaced by a new column: 'log of SALE PRICE'. The purpose of this transformation was to make the data more relevant for analysis and modelling.

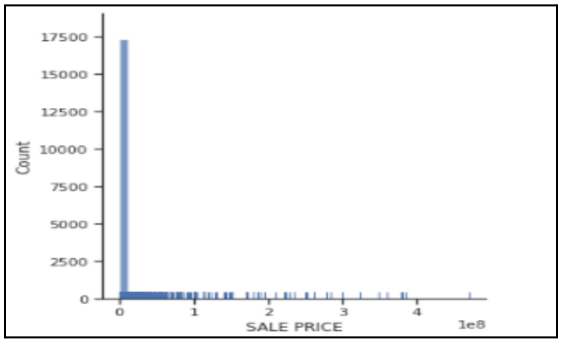


Figure 7 - Graph Sale Price vs Count

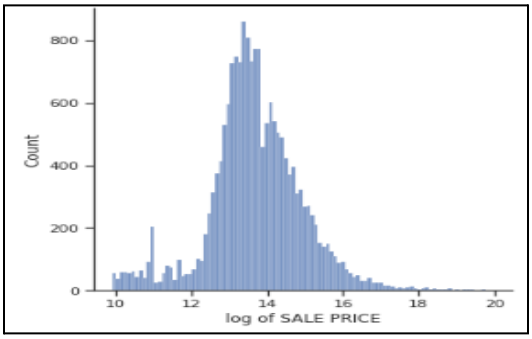


Figure 8 - Graph log of Sale Price vs Count

Data Exploration

Data exploration is the summarising of data based on its graphical displays, bar charts and scatter plots. It is used to show patterns, trends and characteristics in the data which can be helpful for creating an accurate model by spotting potential outliers and ensuring the model is based on relevant information. The graphs below will show some of the patterns found by visualising the data.

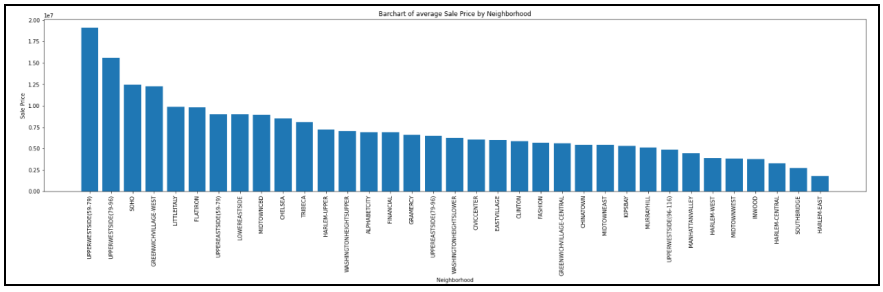


Figure 9 - Prices across neighbourhood

Figure 10 - Prices over time

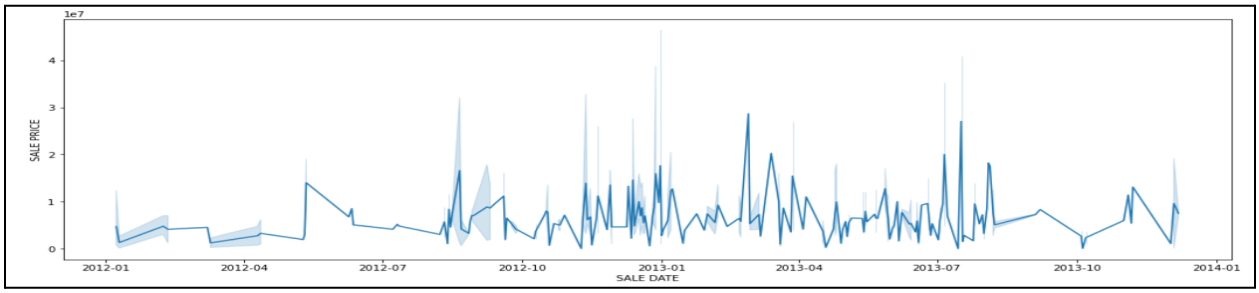


Figure 11 - Scatter Plot of log of Sale Price in Neighborhood

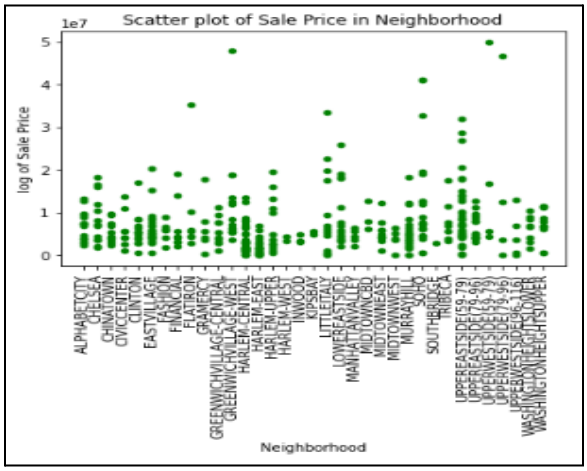
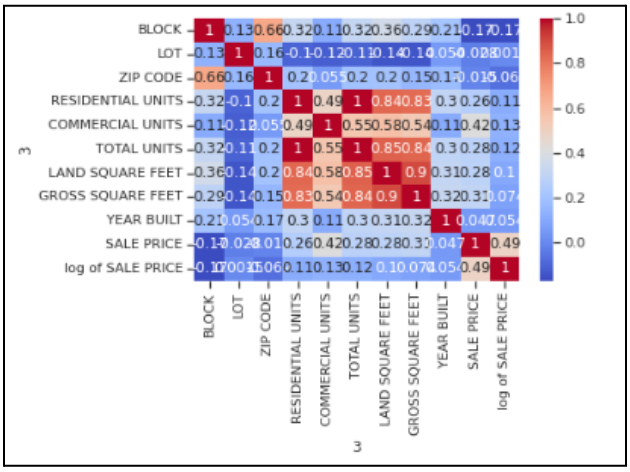


Figure 12 - Correlation Matrix



Some graphs we used were scatter plots (Figure 11) and correlation matrices (Figure 12). They are useful tools for visualising relationships between variables. The main difference being that scatter plot provides visual correlation between two variables while correlation matrices give numerical measure between two or more variables.
Overall, data exploration is a crucial step in the data analysis process, as it helps to identify and understand the characteristics and limitations of the data in a data frame, and to make informed decisions about how to preprocess, analyse, and model the data.

Trends

From Figure 9 the bar chart shows the average sales price in each neighbourhood in Manhattan, this shows the Upper East Side (59-79) has the highest sale price making it the richest neighbourhood in the city while Harlem-East which has the lowest sale price is the poorest neighbourhood due to its cheaper property prices.

Normalisation

For the data processing portion of our project, we employ a normalisation technique called the standard scaler. It makes use of the data's mean and standard deviation, which are less susceptible to the existence of outliers than other scaling techniques like the Min-Max scaler. It

is an important part for data processing as it reduces redundancy, improves the consistency and enhances its integrity. The process organises the data into multiple tables with unique purposes which ensures the information is stored in one place without any risk of errors.

The Model

When building a model for a data frame, it is important to take into account the model's primary goals as well as the specifics of the prediction that will be made, the data's quality, the appropriate features to use, how the data will be cleaned, the model's evaluation, and most importantly, how the model will be used.

The **main objective** of this project is to develop predictive models for houses in Manhattan to predict their sale prices and evaluate their performance. We identified **useful features** with missing values, such as "GROSS SQUARE FEET" and treated them using a **random forest model**. Specifically, we trained the random forest model on the dataset that has no missing values for the features we want to impute and then used it to predict the missing values. By doing so, we were able to ensure that all features had no missing values, resulting in more accurate results. We used Scikit-learn's feature_selection function to identify the most important features for predicting the log of sale price.

Prediction of Sale Prices using Linear Regression Model

Results from Figure 13 show that the linear model did not produce accurate results.

```
Cross-validation scores: [0.1651606  0.13565222 0.12948502 0.14287406 0.15710712]
Average cross-validation score: 0.14605580149113
Mean squared error: 0.8814289931956645
R squared for the training data: 0.14782239292032817
R-squared for test data: 0.13723091726073344
```

Figure 13 - Linear Regression Model Results

Prediction of Sale Prices using Random Forest

Results from Figure 14 show that the random forest model produced much more accurate results.

```
Cross-validation scores: [0.8708541  0.87894054 0.87834223 0.87014485 0.8589123 ]
Average cross-validation score: 0.8714388057306257
Mean squared error: 0.12553943160992737
R squared for the training data: 0.9819804779837583
R-squared for test data: 0.87711824651352
Out-of-bag R-2 score estimate: 0.881
```

Figure 14 - Random Forest Model Results

We will create a cluster-based model and contrast the outcomes to evaluate whether our model can be further improved.

Cluster-Based Model

Choosing our clusters

In order to choose an optimal number of clusters for our dataset when using the KMeans clustering algorithm, we have plotted an elbow method (shown in Figure 15), using which we identified the elbow point and determined that the optimal number of clusters for our dataset is 5 since, from that point onwards, the WCSS (Within-Cluster Sum of Square) starts to decrease significantly, meaning that adding additional clusters will not make a significant improvement.

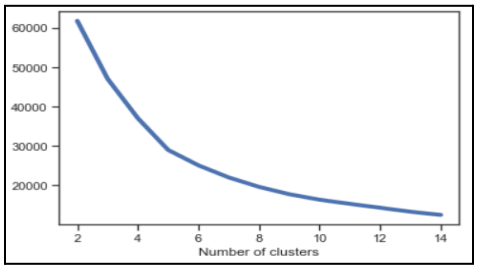


Figure 15 - Elbow Method

Figure 16 - Histogram of Clusters

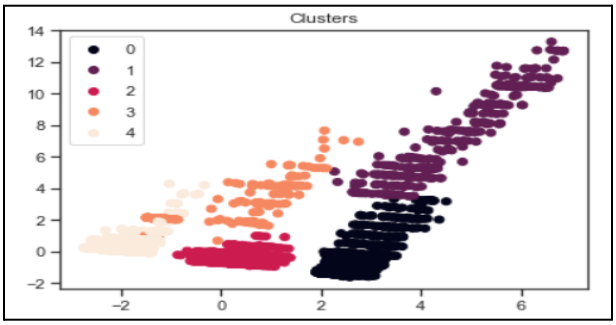


Figure 17 - Distribution of Clusters

As compared to the overall distribution in the second plot, the clusters in the first plot appear to have different forms and sizes of distribution. This shows that using certain underlying patterns or features in the data, the clustering algorithm was successful in grouping comparable sale prices together.

As we can see, some of the clusters in the plot shown in Figure 19, have distributions that are comparable to the overall distribution in the plot in Figure 18, while other clusters have entirely different distributions. This implies that while some clusters might reflect subgroups or population segments with comparable sale prices, others might represent outliers or different groupings with unique characteristics.

Figure 18 - Overall Distribution of Log of Sale Price

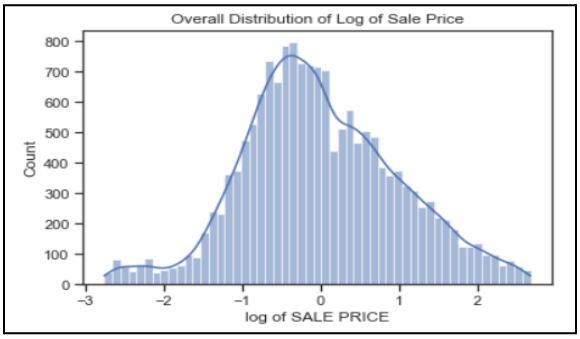
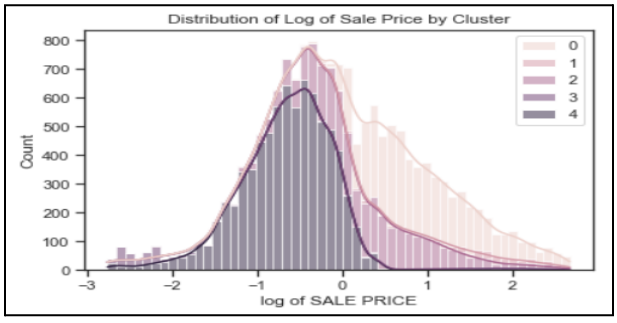


Figure 19 - Distribution of Sale Price by Clusters



Evaluation of the model

We used Random Forest Regressor to evaluate the model instead of linear regression because random forest can handle linear relationships between the features and the target variables and could capture the interactions between features. Linear regression assumes that there is a linear relationship between the features and the targets which is not ideal in our data set. Random forest was the better option because it is more robust to outliers and missing data compared to linear regression.

We created a regression model to evaluate the dataset using R-Squared and the mean squared error (MSE) metrics. Then we created three for loops. The first loop creates a dictionary of the of the regression models that predicts the log of the sales prices based on the other features of the dataset, the second loop prints out the number of clusters we used and an their data points while the third loop evaluates the performance of each model by splitting each of the clusters into the test and training sets. It fits the regression model on the training set, predicts the log of the sale price with the testing set and computes the R Squared and the MSE for the predictions. The results are shown in Figure 20.

```
Cluster 0 has 4546 data points.
Cluster 1 has 7702 data points.
Cluster 2 has 3338 data points.
Cluster 3 has 423 data points.
Cluster 4 has 340 data points.


Cluster 0:
 Training R-squared: 0.9644808700473634
 Testing R-squared: 0.959414618227719
 MSE: 0.01727542734472862

Cluster 1:
 Training R-squared: 0.9869236336549969
 Testing R-squared: 0.9866908382828129
 MSE: 0.0036265137742700583

Cluster 2:
 Training R-squared: 0.9880885089299426
 Testing R-squared: 0.9868238090832647
 MSE: 0.00662862474974737

Cluster 3:
 Training R-squared: 0.972610212165724
 Testing R-squared: 0.9816544210379259
 MSE: 0.07080492890665047

Cluster 4:
 Training R-squared: 0.8900027870311845
 Testing R-squared: 0.8863738040444679
 MSE: 0.07375092505855295


Average R-squared: 0.9601914981352382
Average MSE: 0.03441728396678989
```

Figure 20 - Cluster-Based Regression Model Results

How to further improve the model

There are a number of techniques that may be used to enhance the performance of our model. Increasing the quantity of data the model has access to is one approach. This can be done by including categorical variables, such as ‘NEIGHBOURHOOD’ and ‘TAX CLASS AT TIME OF SALE’, which might benefit the model by enabling it to understand how certain sale prices may change based on the area.

By including this type of data, the model could be better able to capture the subtleties and complexity of the issue being handled, producing predictions that are more accurate. In addition to that, trying different methods to fill in missing values may improve the model even further.

Predictions

To determine how accurate your predictive model will be, it is crucial to contrast the real data with the predicted data when developing one. When trying to get accurate predictions our initial idea was to use a linear regression model to calculate the predicted data. However, upon getting the results we realised that we needed to use another model as the prediction accuracy was below 20%, making it not suitable due to its high inaccuracies. A visual graph of how the predicted data correlates to the actual data when using the linear regression model is shown to the right, in Figure 21 .

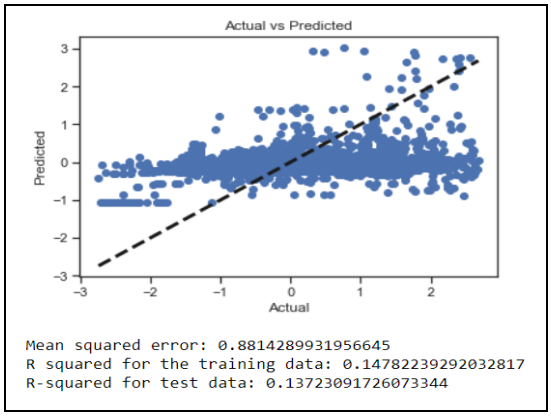


Figure 21 - Linear Regression Actual vs Predicted

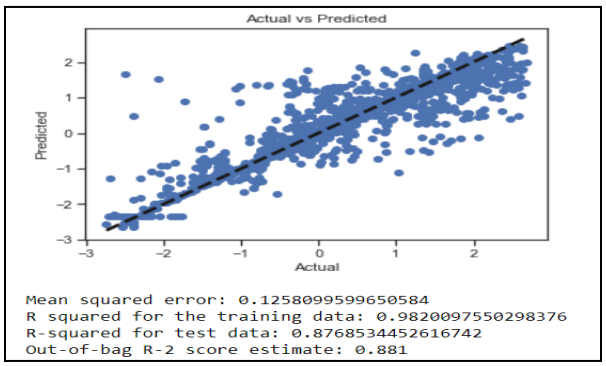


Figure 22 - Random Forest Actual vs Predicted

From our results from the linear regression model, we can conclude that the relationship is more complex than a linear one, hence we decided to use random forest to see if it increases the accuracy. Figure 22 of our data, with an accuracy of over 80%, clearly illustrates the improvements in the relationship between the actual and predicted values.

We use a cluster-based random forest regression model to further enhance our results and obtain more accurate results. Figure 23 to the right illustrates that the predictive model's accuracy against the training data is over 90%.

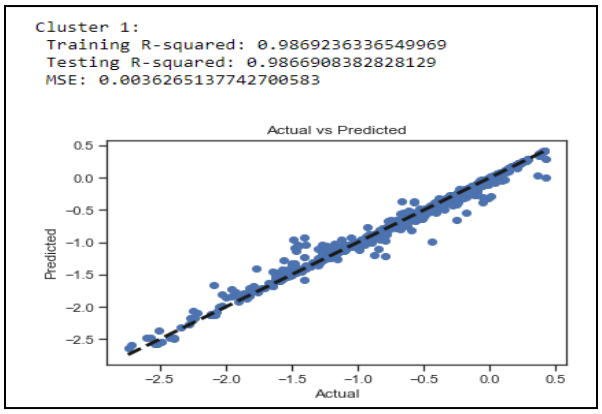


Figure 23 - Cluster-Based Random Forest Regression Actual vs Predicted

Conclusion

In summary, the clusters-based regression model performs better overall than the regression model obtained in Part 2.1, has a lower mean squared error, and has R-squared values for the testing and training data that are more closely matched.