Name Osinowo Michael
Reg No: 170153326

MACHINE LEARNING REPORT

The goal of this project is to create a model that predicts the price of a house which is a continuous value. I will be employing the framework from the lecture on building an effective machine learning pipeline to analyse and solve the problem. The following sections contain a breakdown of that process.

Step 1(Visualisation and Cleaning):

Visualisation for a dataset of this nature will be tedious because of the mix of different types of features (Ordinal, Categorical and Numerical). A simple way to approach this problem without creating a plot for every feature(figure 1) is to use the correlation function to find numeric variables that have a strong relationship with the sales price.

Looking through the correlation plot for the dataset as shown in figure 2. A lot of feature/columns in the dataset are not useful in the prediction of the sales price and are clearly derivatives of other features in the dataset. This is easy to spot by comparing the correlation of every other feature with the sales price as shown in Figure 2.

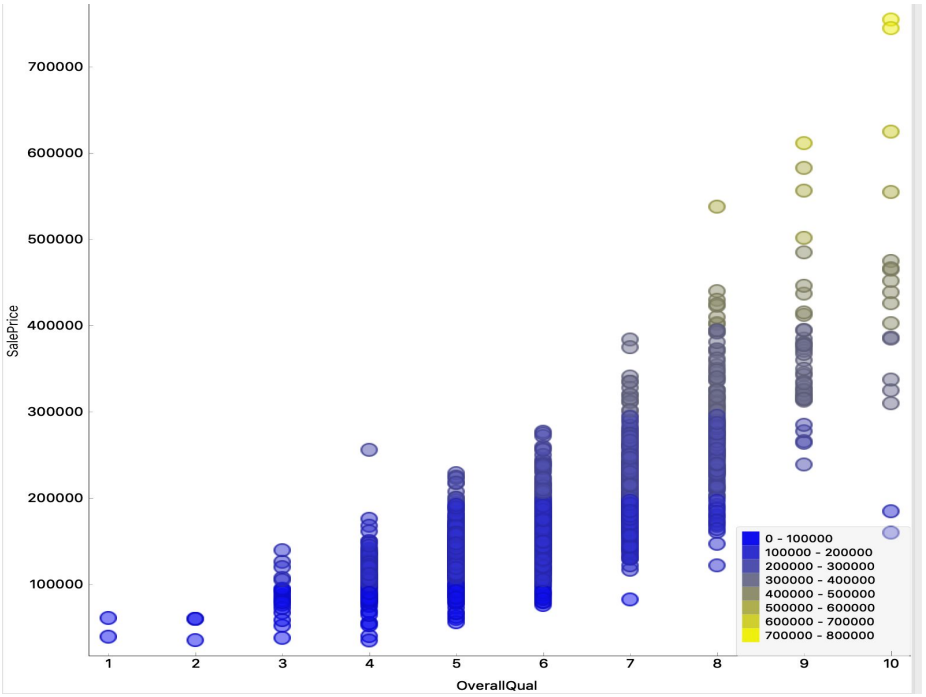

Figure 1. A scatter plot of the sales price against the overall quality which is the most correlated numerical feature in the dataset

1	+0.810	OverallQual	SalePrice
2	+0.731	GrLivArea	SalePrice
3	+0.691	GarageCars	SalePrice
4	+0.653	SalePrice	YearBuilt
5	+0.649	GarageArea	SalePrice
6	+0.636	FullBath	SalePrice
7	+0.603	SalePrice	TotalBsmtSF
8	+0.575	1stFlrSF	SalePrice
9	+0.571	SalePrice	YearRemodAdd
10	+0.565	GarageYrBlt	SalePrice
11	+0.533	SalePrice	TotRmsAbvGrd
12	+0.519	Fireplaces	SalePrice
13	+0.478	OpenPorchSF	SalePrice

Figure 2. Feature ranking showing the strength of the relationship between the numerical variables and the target variable (SalesPrice).

Features with correlation values below 0.45 will be removed to have a more lean dataset to work with(13 features).

Categorical features: Since correlation accounts for the relationship between the target and numerical variables, The ordinal and categorical variables have to be encoded. For instance columns like kitchen quality, garage quality, central air and heating quality would make sense to be factors that determine the price of a house as shown in figure 3 and 4 below. By including these features we make the model decide if they are really useful.

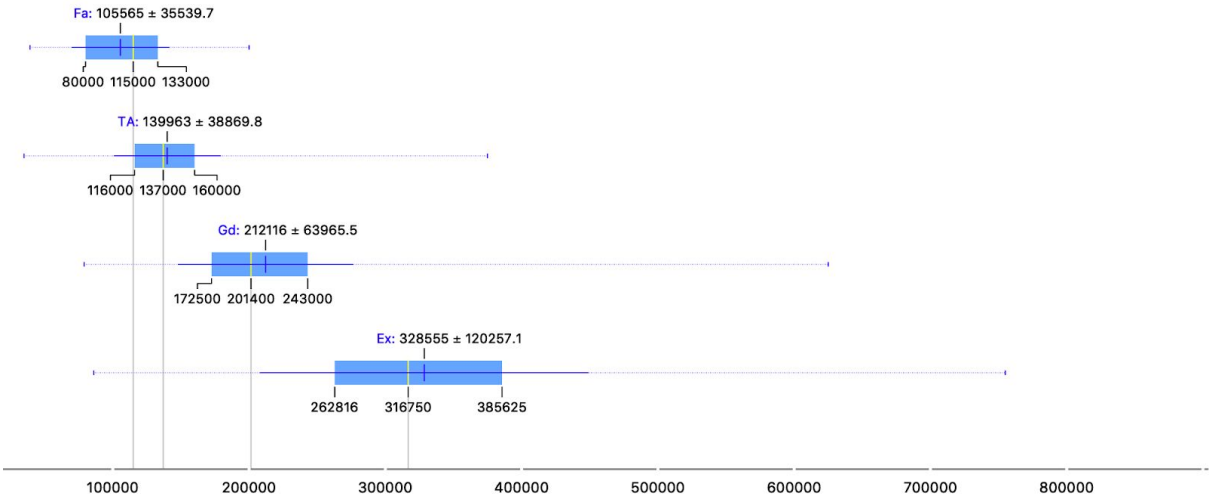

Figure 3: Box plot of the relationship between the categorical variable(Kitchen Quality) and the Sales Price.

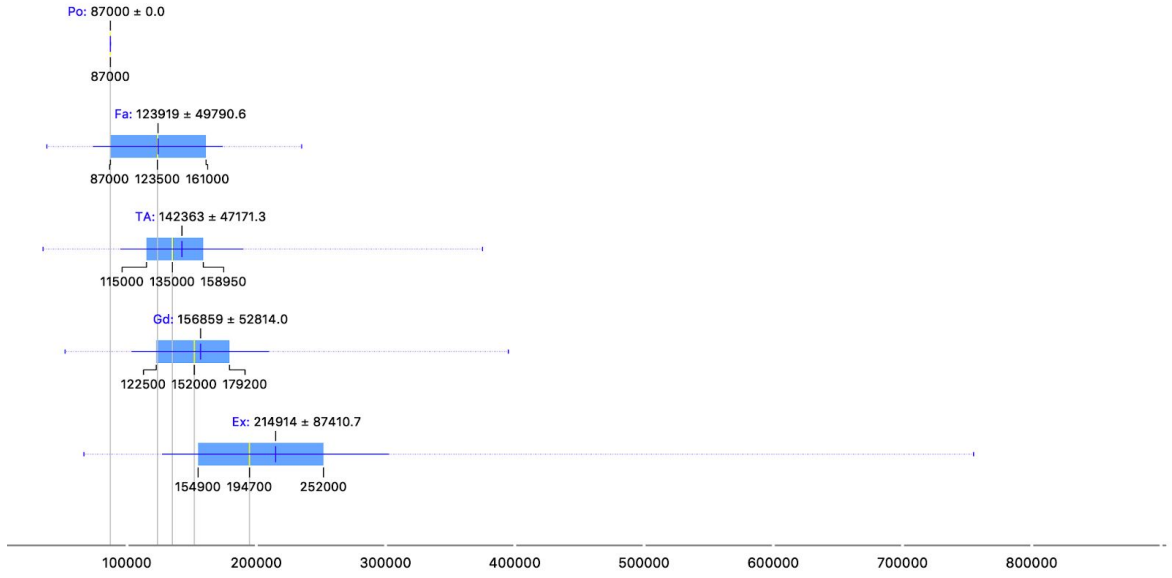

Figure 4: Box plot of the relationship between the categorical variable(Heating Quality) and the Sales Price.

After going through the box plots, the categorical features that seem to have a significant relationship with the sales price are :

- Kitchen Quality
- Heating Quality
- Basement Height
- Neighbourhood

To make the categorical variables useful to the model, they have to be numerically encoded.
Features with information on the garage's size, area and car capacity, seems to have a significant relationship with the sales price.

Step 2 (Preprocessing and Feature Engineering):

The following preprocessing procedures will be used for this task:

Missing values: Columns with too many missing values will be removed e.g(Alley) because they are of no use for prediction.

Outliers: After cleaning the data and selecting the useful columns there were no outliers in the numerical columns.

Standardization: The features used for regression are standardised i.e. scaled between [0,1], to avoid features with relatively larger values from rendering other features with relatively smaller values useless in the model.

Dimensionality Reduction: After cleaning the dataset, there were still 18 features and to select features that have a more significant impact in the predicted sales price, principal component analysis was applied to select features with larger variance, meaning they carry more information than others. The PCA method did not improve the performance so the normalized clean 18 features were used.

Train/test split: the dataset is split 80/20 for train/test sets and the rows are selected at random. The two train test split ratio has no significant effect on performance.

Step 3 (Model Implementation and evaluation):

Linear Regression Model:

The linear regression model included the ridge regularization and a range 30 of values between 0.0001 and 1 for the regularization term was used to experiment for the best model as shown in Figure 5. The pipeline for this model is shown in figure 6

Hyperparameter tuning:

The regularization term clearly has no significant effect on the performance on the linear regression model as long as lambda is below 1.

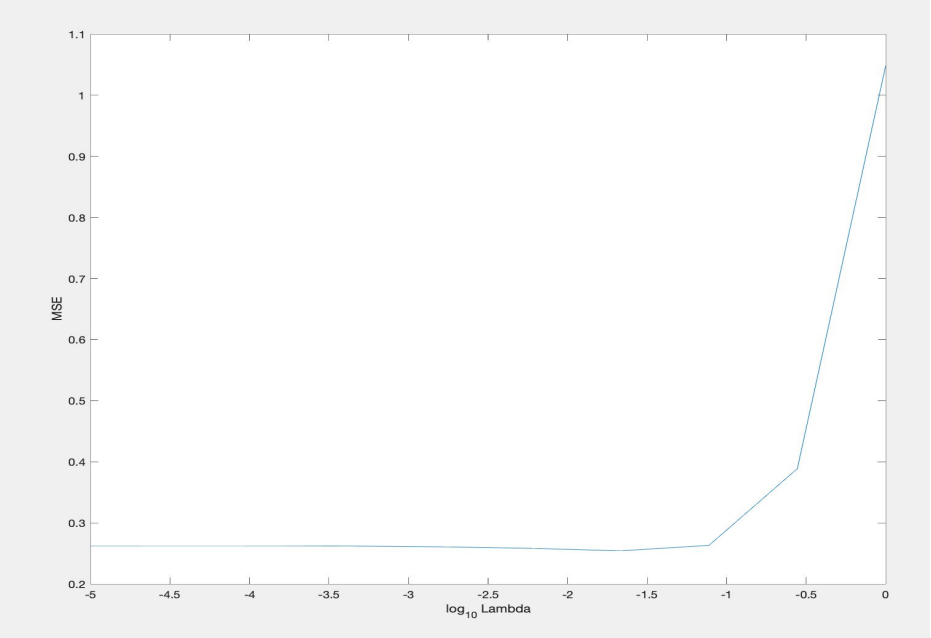

Figure 5. A plot of mean squared error against the range of regularization terms selected

Validation curve:

- Relatively high bias
- Low variance

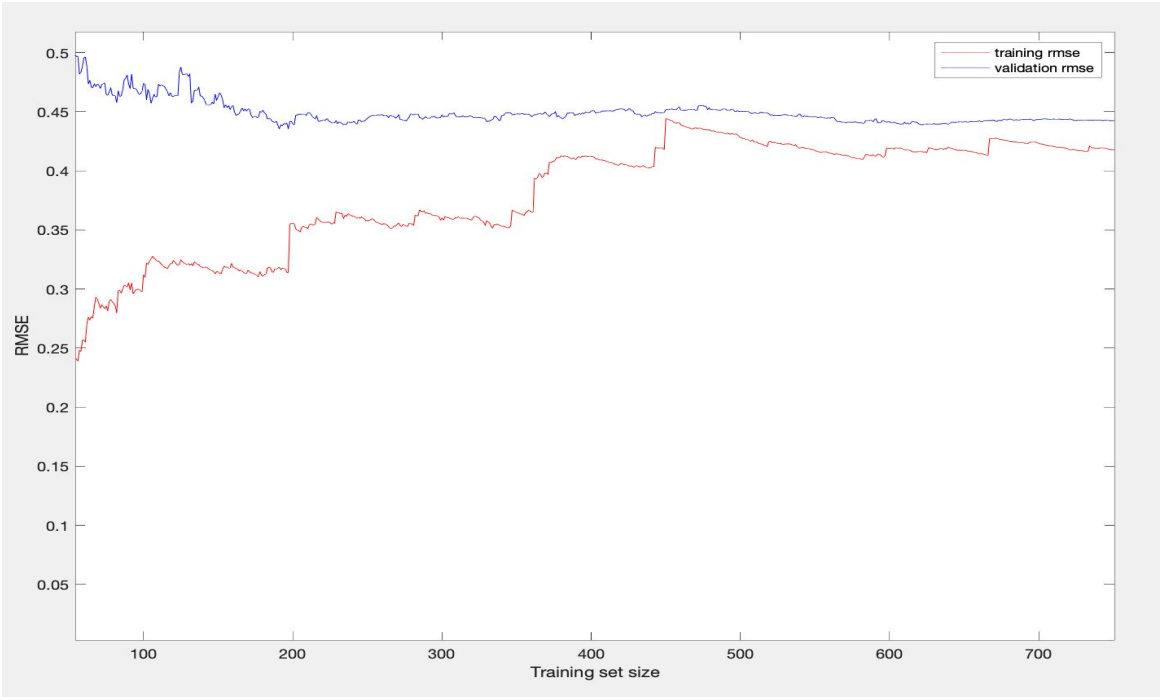

Figure 6.1. Validation curve for Linear Regression model.

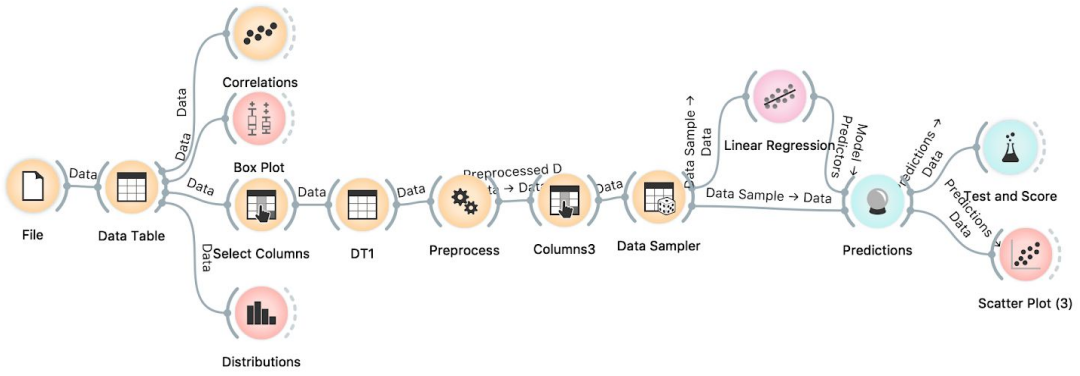

Figure 6.2. The pipeline of the linear regression model.

Model	MSE	RMSE	MAE	R2
Linear Regression	19797.728	0.836	17.788	

Figure 6.3. Performance metric for Linear Regression model

Random Forest Model:

The random forest model has a better performance using the orange pipeline as shown in figure 7.1. 13 trees were used for the random forest model with a maximum depth of 5 as shown in figure 7. Similarly, with the linear regression model, the performance of the random forest model performs poorly without

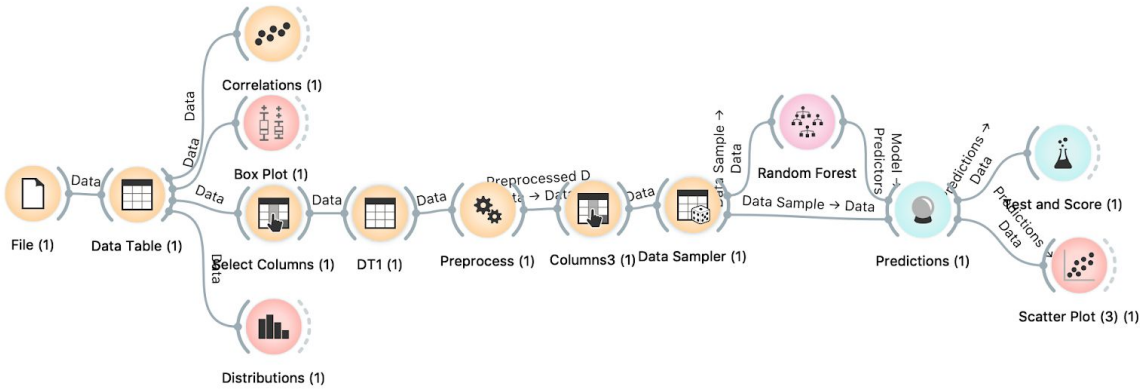

Figure 7.1. The pipeline of the random forest model.

Model	MSE	RMSE	MAE	R2
Random Forest	9561.222	0.965	8.444	

Figure 7.2. Performance metric for Random Forest model

Validation curve Random forest:

- Low bias
- medium variance

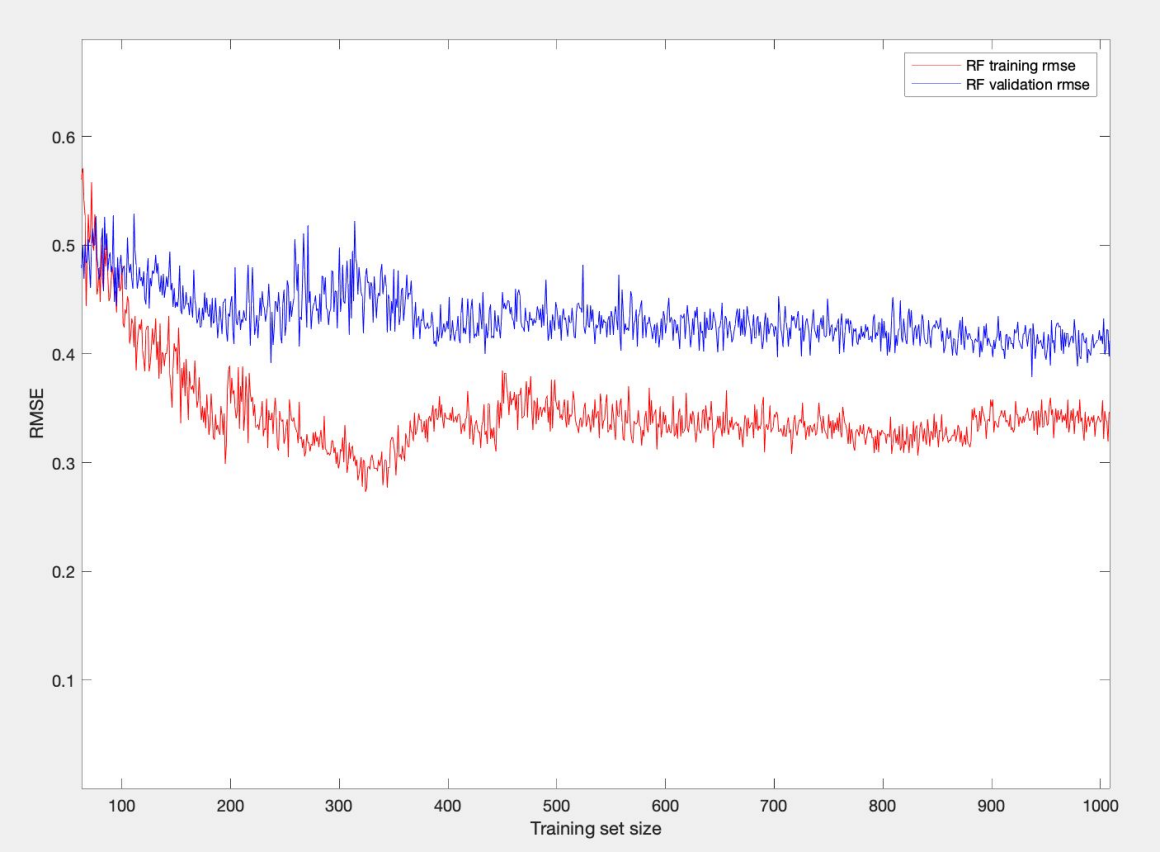

Figure 7.3. Validation curve for Random Forest model.

Conclusion :

The random forest is a better choice of model to predict the sales price more accurately. This can be seen by comparing the metrics in Figure 6.3 and 7.2 and the scatter plot in Figure 8.1 and 8.2. The random forest model is expected to perform better because of multiple decision tree models are run in parallel and then selects the best model for regression. It is more robust to non-linear data and outliers compared to the linear regression model. The two models had similar variance but the Linear Regression model had more bias which is also shown in the correlation plot in figure 8.2. Plotted against the actual values, the linear model had a regression coefficient of 0.83 compared to 0.91 for the random forest, which further proves the conclusion.

To avoid the problem of over-fitting for the models chosen, I used cross-validation with 20 folds and shuffled data randomly before the train/test split. 80/20 split ratio was used.

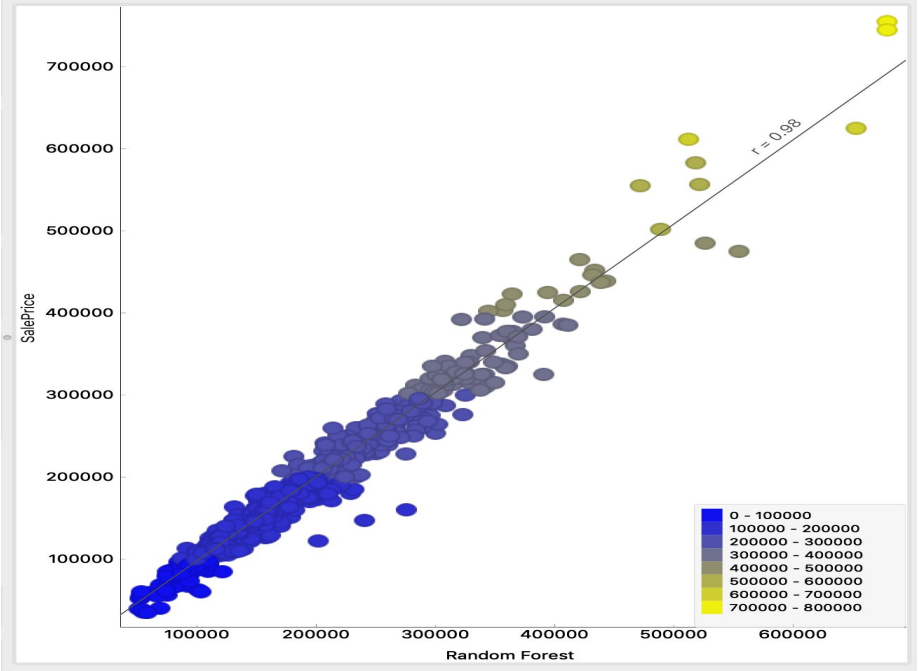

Figure 8.1. Sales price prediction against the random forest prediction.

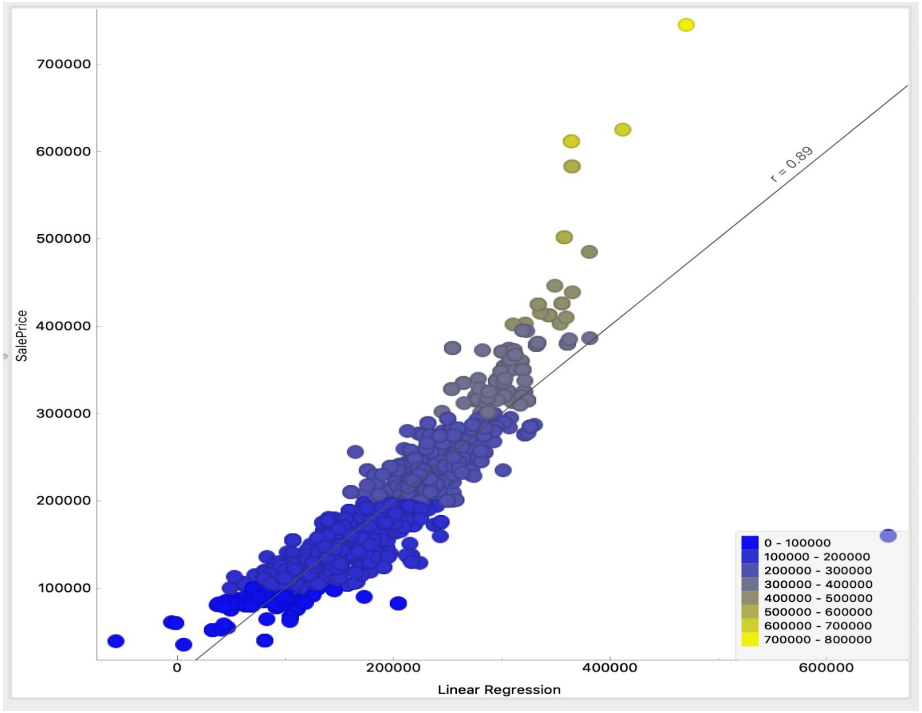

Figure 8.2. Sales price against predicted Linear regression values

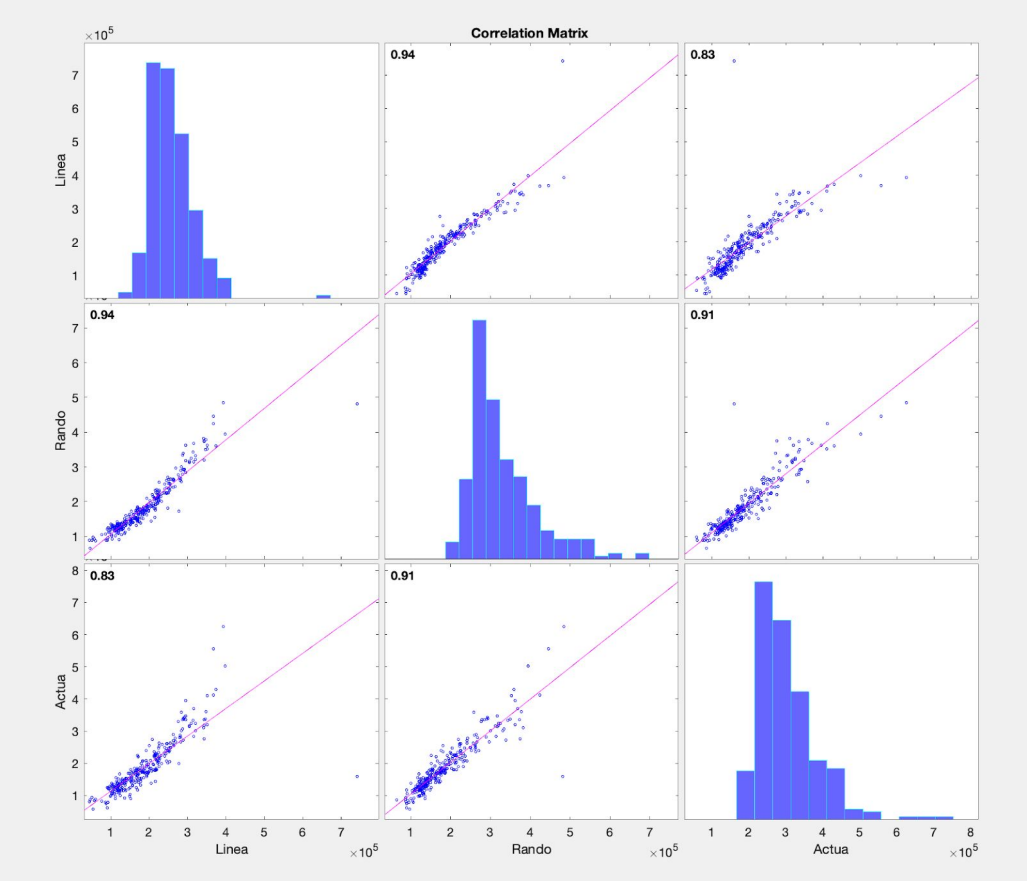

Figure 8.3. A cross-correlation plot of the predicted values from the linear regression model, the random forest model and the actual values from the test set.