**Model Traning and Evaluation**

* **Model 1 – Random Forest Regression**

The exploratory data analysis suggests that the data will be best fit for a tree based model instead of linear regression. Now we choose Random Forest over Decision Tress as there are higher chances of overfitting on Decision Tree whereas Random Forest being part of ensemble learning process the overtiffing is low.Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

A Random Forest Model is fit on the data and to improve the accuracy and decrease the error following hyperparameters are tuned

1. n\_estimators = 100
2. min\_samples\_split = 100

**Evaluation of the model 1**

* Mean Sqaured Error for Train Data : 7.31
* Mean Sqaured Error for Test Data : 7.37
* Model 1 Score = 92.08 %
* **Model 2 – Random Forest Regression with new Features**

Now, three more features are added into the model which will help to improve the performance of the model. Feature engineering refers to the process of using domain knowledge to select and transform the most relevant variables from raw data when creating a predictive model using machine learning or statistical modeling. the three new features are -

1. CTR = Clicks / Impressions
2. CPC = Cost / Clicks
3. CPA = Cost/ Conversions

These were the features that were highly correlated and will help to improve the model. Now the total number of features have incresed from 10 to 13.

Training Random Forest Regression model with new features and hyper parameters.

1. n\_estimators = 100
2. min\_samples\_split = 50

**Evaluation of the model 2**

* Mean Sqaured Error for Train Data : 6.06
* Mean Sqaured Error for Test Data : 7.16
* Model 2 Score = 93.42 %

Model 2 performed better than model 1 despite they are same model as we have added new features which lead to decrease in the error and increase in the accuracy.

* **Model 3 – Linear Regression**

Linear regression is used for finding linear relationship between target and one or more predictors. After the addition of the new features the data has become more suitable for linear regression. Therefore applying linear regression on the new dataset with 13 features.

**Evaluation of the model 3**

* Mean Sqaured Error for Train Data : 7.68
* Mean Sqaured Error for Test Data : 7.02
* Model 2 Score = 91.67 %

As evident from the exploratory data analysis this model did not perform better than the random forest regression despite having new features.

* **Model 4 – Support Vector Regression**

Support Vector Regression (SVR) uses the same principle as SVM, but for regression problems. The problem of regression is to find a function that approximates mapping from an input domain to real numbers on the basis of a training sample.

**Evaluation of the model 4**

* Mean Sqaured Error for Train Data : 9.32
* Mean Sqaured Error for Test Data : 9.72
* Model 2 Score = 89.89 %

Support Vector Regression performs most poorly with the datset. SVR gave least score and highest Mean Sqaured Error.

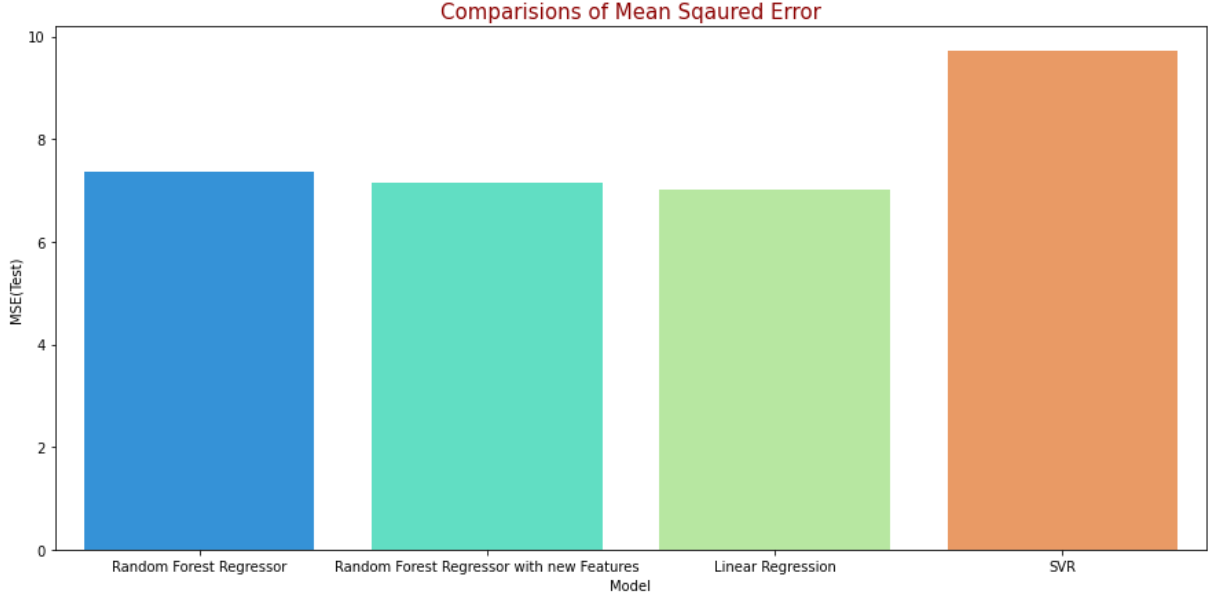
**Comparison of the Models**

The models can be compared with help of a table as follows

|  |  |  |  |
| --- | --- | --- | --- |
| Model Name | MSE(Train) | MSE(Test) | Model Score(%) |
| Random Forest Regression | 7.31 | 7.37 | 92.80 |
| Random Forest Regressions with new Features | 6.06 | 7.12 | 93.42 |
| Linear Regression | 7.68 | 7.02 | 91.67 |
| Support Vector Regression | 9.32 | 9.72 | 89.89 |

**Comparing the models visually.**

**Comparison Based on Mean Sqaured Error :-**



**Comparison Based on Model Score :-**



**Conculsions**

* The attributes were not much linearly related therefore Random Forest Regression performed better than other models.
* Support Vector Machine Performed worst among all models with accuracy of 89.89% and mean squared error of 9.71
* Adding three new features usind feature engineering improved the score of the Random Forest Regression Model.
* Highest Model Accuracy is given by Random Forest Regression with added features which is 93.42 % and lowest mean squared error of 7.12.
* All the models were hypertuned well to overcome overfitting and the error for the train and test data are nearly equal for all models.