**Technical Report**

From the provided instructions and the dataset, I have understood to solve two different problems out of the provided dataset. The two problems are as follows:

1.To predict the Sales individually for each of all 3 Strategies for the past 2 years i.e., for 2013 and 2014.

2.To calculate the extent of loss in sales for the year 2015 i.e., compare the sales from the start of 2015 till the mid of 2015 and from the mid of 2015 to end of 2015.

So I have solved the above two problem statement in two separate Jupyter notebooks one for each of the above problems.

For Problem Statement1:

I have followed the life cycle of the Data Science project starting with the Exploratory Data Analysis (EDA) part of the provided data set so I can have a better understanding to figure out the hidden insights present in the dataset. As part of my EDA, I have further divided the month column in order to extract the year and day from it so I can able to have a better understanding of the data with respect to sales feature. I have plotted various other plots such as Histograms, Q-Q and Box plots for all continuous features for better visual understanding.

I have observed there was a sudden rise in Sales from the start of year 2014 and from months from 3 to 10 have been increase in Sales than the start and end months of all the 3 years.

I have observed as the amount of account size, account targets increase the sales also increases.

The feature quantity had the best correlation with the target feature Sales. When I plotted the grap between the quantity and sales I have observed a straight line which mean they both are highly correlated with each other.

The Sales have been on the decreasing part at the end as per the Strategy1 and Strategy3 but on the increasing part with respect to Strategy 2.

The Sales have been on the increasing part at the end as per the salesVisit1, salesVisit2, salesVisit5 but on the decreasing end as per the salesVisit3 and salesVisit4.

I have observed out of all account types the Hospital has more amount of data related to sales later followed by the Polyclinic, Pharmacy and Private Clinic.

I have observed the presence of Outliers for all the account types and other features that are present in the dataset when I plotted a box plot, but I chose not to remove them as when tried the final data set had very less rows when I tried to eliminate the outliers by using inter quantile range proximity rule and the machine learning model accuracy is not up to the mark so I haven’t removed the outliers for this purpose as the dataset has few rows and columns.

Later, I divided the dataset into 3 separate datasets each containing one strategy so I can predict the sales individually for each strategy so I can better have a conclusion on which strategy performed the best with respect to dependent Sales feature.

In order to choose the features that should exist in those 3 datasets I have visualized the entire dataset using Heat map so I can better have a visual understanding of the data so from that heat map table I figured out the features that are not having high correlations. So by this approach I can eliminate the multi-collinearity problem and so as the dimensionality reduction happens as well.

As mentioned to eliminate the multi-collinearity for example in the given data set the month and year columns have same correlation values with each other but not with the dependent feature Sales. So I choose to eliminate month features and selected year feature to be a part of next process.

For the purpose of applying the machine learning algorithms I have divided each data set into independent and dependent features namely X and y and performed the Train-Test-Split approach. As the problem statement is about the prediction of Sales feature so its considered as a Regression problem statement. So I have applied various Machine Learning algorithms such as Linear Regression , Lasso Regressor, Support Vector Regressor, Decision Tree Regressor, Random Forest Regressor and XGBoost machine learning models for all the 3 datasets.

During the process of applying the machine learning algorithms on the dataset, I have built a Pipeline which includes that respective machine learning model and a Standard Scaler inside the pipeline. The Standard Scaler was used in order to scale all the values present in the dataset in the range -3 to +3 so for better achieving the model accuracy.

I have calculated the R2 Score and Root Mean Square Error(RMSE) values for each of the above algorithms and plotted the learning curves for each of the machine learning model along with the training and test values. I have plotted all the predicted results with respect to original results of Sales for each of the above mentioned machine learning algorithms.

I have also performed the Hyper parameter tuning using Grid Search CV technique in order to optimize the machine learning algorithms performance by passing various values as parameters inside a parameter grid and applied the Grid Search CV along with that machine learning algorithm.

After continuing the above steps for all 3 datasets which are consisting of all 3 strategies individually, I have observed which machine learning model had the best accuracy for each of all datasets. For the model which had the best accuracy, I calculated the mean of the predicted Sales feature in order to calculate the percentage for the machine learning algorithm had the best accuracy and stored in a variable for all 3 strategies.

I have observed out of all machine learning algorithms I have used the Lasso Regressor and Random Forest have achieved the best accuracy for all 3 strategies. For all

I have observed that Strategy3 had slightly better impact on the prediction of Sales than Strategy1,Strategy2.

The order of all 3 strategies are in the order of Strategy1, Strategy2, Strategy3 where each had the following mean values of their final best prediction.

Percentage\_of\_sales\_strategy1= 0.26881720430107536

Percentage\_of\_sales\_strategy2= 0.26881720430107525

Percentage\_of\_sales\_strategy3= 0.26881720430107525

For Problem Statement2:

In order to predict the extent of loss in potential sale in the year 2015 from mid of 2015 I have calculated the Prediction of Sales from start of 2015 to mid of 2015 and from mid of 2015 to end of 2015 separately.

Firstly I have sorted the first dataset consisting of before mid of 2015 data and performed the Exploratory Data Analysis part for the dataset, I have observed that Strategy3 had no impact the Sales prediction whereas the Strategy1 and Strategy2 had been both ups and downs in their pattern for Sales feature. Later I have repeated the same steps for the second dataset which consists of all the records from after mid of 2015 and found all 3 strategies are having ups and downs but tending towards decreasing order of Sales.

I have applied all the previously mentioned machine learning algorithms to both the datasets and found out Lasso Regressor had performed the well with very good accuracy than rest of all models.

I have calculated the mean value for the final prediction done by the best model for each dataset and stored them in Percentage\_of\_sales\_before\_mid\_2015 and Percentage\_of\_sales\_after\_mid\_2015 variables in order to store the final percentage value of the predicted Sales features.

Finally in order to calculate the extent of loss of potential sales due to a new competitor drug entrance into the market since mid-2015, I have calculated the mean of the percentage of sales before mid of 2015 with percentage of sales after mid of 2015 and observed “52% “ difference in of loss in potential sales due to a new competitor drug entrance.

Observations of Features in the provided dataset.

1.The dataset consists of int, float and object type of data in various features.

2.There are no NULL values that are existed in the dataset.

3.The feature “accID” is having very high cardinality among rest of all features. So I did not include them during the machine learning model training.

4. As there was only one district in the dataset there is nothing much we can use it to predict the Sales.

5.When I plotted all 3 Strategies for 1st problem statement for past two years, I have observed only Strategy1 has been on the increasing order of Sales whereas the Strategy2 and Strategy3 are on the decreasing side of Sales feature when plotted.