WOMart Supplement Sales Prediction

Problem Statement

Supplement Sales Prediction

WOMart follows a multi-channel distribution strategy with 350+ retail stores spread across 100+ cities. Effective forecasting for store sales gives essential insight into upcoming cash flow, meaning WOMart can more accurately plan the cashflow at the store level. Sales data for 18 months from 365 stores of WOMart is available along with information on Store Type, Location Type for each store, Region Code for every store, Discount provided by the store on every day, Number of Orders everyday etc. Your task is to predict the store sales for each store in the test set for the next two months.

Problem solving Methodology

Step 1: Importing the Relevant Libraries

Step 2: Data Inspection

This step involves looking at data anomalies, outlier detection, normalization and data consistency.

a)Looking for null values

b)looking for numeric and categorical variables

Step 3: Data Cleaning

Data cleansing involves removing or correcting inaccurate records.

a)Dropping unused variables or columns in dataframe

b)Combining str methods with numpy to clean coloums

Step 4: Exploratory Data Analysis

EDA involves analyzing the dataset to summarize main characteristics to discover patterns, spot anomalies etc. Here in this dataset we used several different exploratory analyses to identify key variables for regression equation such as correlation plots, heat maps histograms etc.

a)There were no missing values in any columns

b)looking at store type distribution and location type distribution

c)Creating dummy variables for store type, location type, Region code and discount in training as well as test dataset

Step 5: Scaling of Variables

Min-Max Scaler is applied to all the columns except yes-no and dummy variables and is used to convert different scales to a standard scale it make it easier for algorithms.

Step 6: Building Model

The model is built by learning and generalizing from training data, then applying that knowledge to new data that it has never seen before to make predictions.

Linear regression, random forest and decision trees are the models used in this case

Step 7: Prediction of Sales in Test Data Set

Predicting sales across multiple stores for the future weeks is covered here. The performance of various models are explored like linear regression, decision tree, random forest and the one with higher accuracy is least RMSE is selected. Decision tree Model is selected for future sales prediction for test dataset

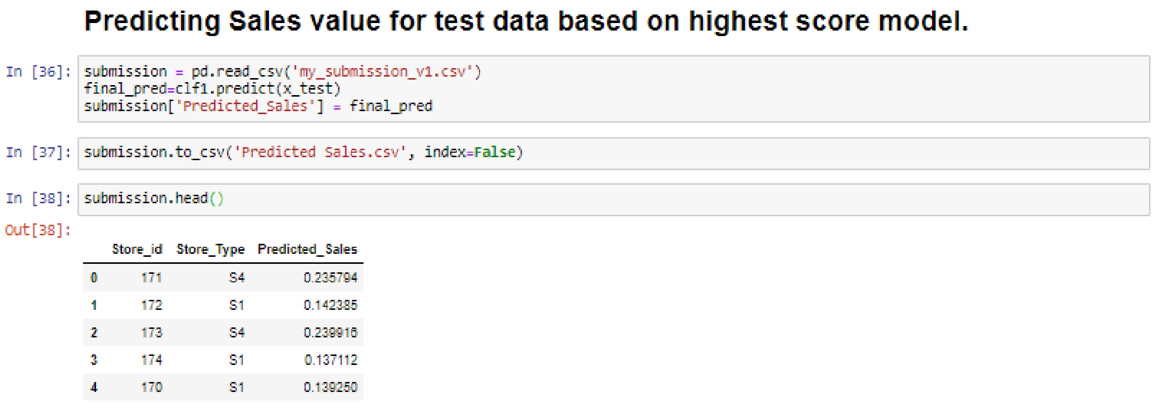
What data-pre-processing / feature engineering ideas really worked? How did you discover them?

- Initial model which was built had Month and year was extracted from Date to add as variables for the prediction. But finally it was avoided.
- Dummy variables was created for Store_Type, Location_Type_test, Region_Code_test, Discount_test in both train and test dataset.
- Scaling of the numerical value was conducted, but its not a mandatory step since only sales data is numeric.

What does your final model look like? How did you reach it?

3 models were built using Linear Regression, Decision Trees and Random forests.

One with least RMSE and best accuracy was used to predict test dataset.



Predicting Sales value for test data based on highest score model.

```
In [36]: submission = pd.read_csv('my_submission_v1.csv')
         final_pred=clf1.predict(x_test)
         submission['Predicted_Sales'] = final_pred

In [37]: submission.to_csv('Predicted Sales.csv', index=False)

In [38]: submission.head()

Out[38]:
```

	Store_id	Store_Type	Predicted_Sales
0	171	S4	0.235794
1	172	S1	0.142385
2	173	S4	0.239916
3	174	S1	0.137112
4	170	S1	0.139250