**JOB-A-THON**

Approach By: **Vishwanath Kulkarni**



**Approach:**

**EDA:**

Exploratory Data Analysis is the first step of any machine learning lifecycle. It involves steps like Univariate Analysis, Bi-Variate analysis, checking for missing values, outliers, missing value imputation, etc. Steps followed and some observations were as follows,

1. **Univariate Analysis:**
2. There are no missing values in the dataset.
3. The training dataset is for 1.5 years, and the testing dataset is for two months. There were no overlaps between training and testing datasets.
4. The order column is present in the training data but not in the testing data because we cannot exactly predict the number of orders received in the future.
5. There are 365 unique store id values, and each store id has data for 516 days.
6. There are 4 Store types of which S1 is the most popular and S3 is the least.
7. There are 5 location types of which L1 is the most popular and L4 is the least.
8. The training dataset had 68 holidays in total whereas the test dataset had 2 holidays.
9. There were 501 days in the training dataset on which there was a discount and the testing dataset had 61 such days.
10. There are 19 records where number of orders received, and the sales is 0.
11. Dependent variable is slightly positively skewed with skewness value of 1.24.
12. There are not many outliers in the dependent variable.
13. **Bi-Variate Analysis:**
14. Orders and Sales variables are highly correlated, as the number of orders increase so does the sales. The scatterplot below describes the relationship between two variables,

Chart, scatter chart

Description automatically generated

1. The sales variable is does not follow a particular overall trend, there is no seasonality displayed either. The below line plot describes the sales as the time progresses,

Chart, line chart

Description automatically generated

1. The orders variable does not display any kind of seasonality or an overall trend, the line plot below makes it clear,

Chart, line chart

Description automatically generated

1. Average monthly sales do not follow a pattern, average sales are highest from May to July, followed by January and December. The bar plot below describes the average monthly sales.
2. Saturday and Sunday have the highest average sales in the weekdays. Friday has the least average sales.
3. Average sales are higher on weekends and lower on weekdays. Sales are also lower when there is a holiday. Sales are significantly higher when there is a discount as compared to the no discount days.

Icon

Description automatically generated Icon

Description automatically generated

1. **Correlation:**

There is just one numerical variable in the dataset apart from the dependent variable which is orders. From the correlation plot it can be observed that it is highly correlated to the dependent variable. No other variable has such high dependency to the dependent variable.

A screenshot of a computer

Description automatically generated with medium confidence

1. **Missing Value Treatment & Feature Generation:**
2. Although there were no missing values in the dataset, there were observations where Sales and Orders were 0, these rows were deleted from the datasets.
3. Features were generated by concatenating the categorical variables such as Store Type, Location Type, Discount etc.
4. Some group by features were also generated by using the categorical variables.
5. Label Encoding was used to encode the categorical variables. One hot encoding resulted in a lower leaderboard score.
6. Orders variable was not used in the modelling as the exact number of orders in the future is not known. Although I tried to predict the number of orders first and then predict the sales, this approach failed as the leaderboard score was low.

**Modelling:**

Modelling always comes second to feature engineering and exploratory data analysis, having said that, it plays an important role in the entire predictive modelling purpose.

1. LightGBM and XGBoost were tried individually butXGBoost performed better than LightGBM.
2. Since KFold cross validation or StratifiedKFold cross validation techniques shuffle the data for validation purposes, it wouldn’t be of use in the current problem as it is time dependent. The dataset was sorted based on the date variable and the last two months (April and May) were used as validation datasets whereas the rest of the training dataset was used for training.
3. Hyperparameter tuning was extensively performed tuning the values of learning rate, max depth, subsample, column samples, etc.
4. Regularization was also performed so that the model does not overfit on the training data.

**Room for Improvement:**

There is always room for improvement in the field of machine learning.

1. Generating more features based on domain knowledge always helps the cause.
2. Time series models such as AR, ARMA, ARIMA should also be tried to see the performance.