Mekktronix Sales Forecasting Challenge

Prakhar Partha

Approach

- Data Preparation-

  Main challenge of the problem was to create a training set on which model would be fitted. As the data was not present at a place. It was mainly divided into 3 parts the promotional expense data, the holidays data and the previous sales data at weekly level. So the main challenge was to prepare these 3 separately and then merge them.

- # Data Preprocessing-

Few notable **Data Preprocessing** steps used were:

1. Sales data and expense price data had a large no. of outliers which were removed by applying log transformation. The resultant values also resembled the normal distribution curve.
2. To consider the demand of product and economic status of people extra feature of GDP per capita income for each country was added at yearly level.

Trends Observed in the data:

1. Sales increases with the expense cost (Pearson Cor= 0.98).
2. Sales distribution varies atCountry and ProductID level. (Supported by results of ANOVA test with Pvalue of $1.62x10^{-237}$and$1.14x10^{-9}$ resp.)

- # Modelling-

Model Selection:

I used 7 features for fitting the model, out of which 4 were categorical and 3 were continuous. So a tree based model was expected to perform better. After comparing the results of various algorithms like linear regression, random forest, single tree regressor, SVM regressor, as expected random forest was giving the best performance on the test set.

Further I applied hyperparameter tuning to enhance the result.

I finally tried applying boosting using Xtreme Gradient boosting with tree learners as:

1. No. of features<< No. of training examples
2. We have a mix of categorical and numeric features

XGBoost gave the best results with a very low test MAPE error. So I decided to proceed with XGBoost.

Tuning the Model:

1. Hyperparameter tuning- I further tuned the various parameters like colsample_bytree, learning_rate, max_depth, n_estimators, subsample.
2. L1 and L2 regularization- Used proper L1 and L2 regularization coeffiecients (gamma, alpha and lambda) to avoid overfitting.
3. Feature Selection – I explicitly removed least significant features from my model and fitted models with different no of features. I observed that the model with 6 features gave the best accuracy and the least test MAPE error.

- Model Interpretation-

Top 5 most Significant Variables in the dataset I have used for creating the Model:

1. Expense Price
2. Month
3. Year
4. Product_ID
5. GDP per capita

Expected MAPE

35

Tools: Python

IDE: Spyder