Black Friday

A study of Sales Through Consumer behaviors

Sumit Mishra

1. Problem Statement

A retail company wants to understand the customer purchase behaviour (specifically, purchase amount) against various products of different categories. They have shared purchase summary of various customers for selected high-volume products from last month. The data set also contains customer demographics like age, Gender, Marital Status etc. and Total purchase amount from last month.

2. Dataset

The dataset comes from a competition hosted by Analytics Vidhya. In this Data set we have been given with the purchase summary of the customers from the last month. The Demographics of the customers are also given along with the purchase summary. The Dataset has about 538,000 rows and 12 columns along with the Customer ID. The Dataset is also available on the Kaggle.

3. Features and Preprocessing

Age and Genders are very important features because the purchase amount highly depends on the Age and Gender of the customer. They have direct Correlation with the purchase amount. I have used Product ID, Gender, Age, Occupation, Current City Staying, City Category Marital Status and Product Category 1,2,3 as Feature for the models.

Since the Product Category 1 and 2 have Null Values so I have filled the them with the mean of their respective category. Age, Gender, City Category and Stay in Current city are given as a string i.e. Categorical value so I have converted them to Numerical Values for the Model. The Product ID is also categorical hence I have used Label Encoder to change it to the Numerical Values. Purchase column is the target variable. I have used train test split to split up the Data Set into train and test set.

4. Models

The model performance is judged on the basis of its RMSE Score.

- **RIDGE REGRESSION** - I have used Ridge Regression to train a model but it gave a big RMSE score.

- **POLYNOMIAL REGRESSION –** The polynomial regression model of degree 2 and 3 gave a decent RMSE score but still not in the domain I wanted.

- **LASSO REGRESSION –** The lasso Regression also gave a decent RMSE score but no improvement is seen in the RMSE score even after changing the max_iter and alpha parameters.

- **BOOSTED GRADIENT DESCENT –** Boosted Gradient decent regressor gave a good RMSE score of 25XX. I varied the parameters of the XGB regressor to get a good RMSE score and settled on the parameters (objective ='reg: linear', colsample_bytree = 0.3, learning_rate = 0.05, max_depth = 10, alpha = 10, n_estimators = 1000)

5. RESULT

After all the preprocessing and modelling, the Boosted Gradient Descent is the model I have used for the final predictions.