Machine Learning Engineer Nanodegree

Capstone Proposal

Name of Project: Black Friday Sales Prediction

Sreyashi Dutta

March 18th, 2019

Reference: https://datahack.analyticsvidhya.com/contest/black-friday/

Domain Background

Domain: Retail

Retail industry is one of the blooming industries where Machine Learning is heavily used these days. It is also a highly revenue generating industry and hence companies are investing heavily on machine learning to increase the sales of their products.

Black Friday is celebrated a day after the Thanksgiving in USA. Typically, retail companies and small shops offer black Friday sales to their customers and it has been observed as highly popular amongst the customers which in turn brings in a lot of revenue for the companies. As such, companies are now using machine learning algorithms to come up with recommendations of products for the customer in order to increase the sales

Problem Statement

A retail company "ABC Private Limited" wants to understand the customer purchase behavior (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 (age, gender, marital status, city_type, stay_in_current_city), product details (product_id and product category) and Total purchase_amount from last month.

Now, they want to build a model to predict the purchase amount of customer against various products which will help them to create personalized offer for customers against different products.

Datasets and Inputs

The dataset consists of train and test sets. The model performance will be evaluated on the basis of prediction of the purchase amount for the test data (test.csv), which contains similar data-points as train except for their purchase amount.

The variables and their definition in the dataset are mentioned below:

Variable	Definition
User_ID	User ID
Product_ID	Product ID
Gender	Sex of User
Age	Age in bins

Occupation	Occupation (Masked)
City_Category	Category of the City (A,B,C)
Stay_In_Current_City_Years	Number of years stay in current city
Marital_Status	Marital Status
Product_Category_1	Product Category (Masked)
Product_Category_2	Product may belongs to other category also (Masked)
Product_Category_3	Product may belongs to other category also (Masked)
Purchase	Purchase Amount (Target Variable)

Solution Statement

Firstly, as a part of data exploration I will observe NA values as well as unique values. Next we will use LabelEncoder to convert all text or categorical data to numerical for our model to interpret it better.

Ultimately I will compare different model output such as linear regression , XGBoost to see which is giving more accurate result

Benchmark Model

XGBoost .

A good documentation I came across about XGBoost is this :

https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

Evaluation Metrics

This is a problem defined in AnalyticsVidhya website. So I will take into consideration my leaderboard score as my evaluation.

AnalyticsVidhya will evaluate my submission based on the root mean squared error(RMSE). Compared to Mean Absolute Error, RMSE punishes larger errors.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Where y hat is the predicted value and y is the original value.

Project Design

I choose this project as it is very similar to a use case that I have been given to work upon in my current organization.

First I will read the train and test data and find out the NAs and unique values. I will also use LabelEncoder to convert all text or categorical data to numerical for our model to interpret it better. I will perform some visualizations to see my target variable behavior in train set and remove outliers if required. Finally I will use some models and fit them on the train data. Ultimately we will use the model to predict on our test data. Using that, we will write the data on CSV file with attributes – User ID, Product ID and Purchase amount as submission file. I will submit it on AnalyticsVidhya evaluation platform and see how accurate the prediction has been.