IBM Big Data and Machine Learning Prodegree

Capstone Project: Project Report

PROJECT REPORT

Predicting Purchase Behaviour on E-Commerce Data

Submitted towards the partial fulfillment of the criteria for award of IBM Big Data and Machine Learning Prodegree by Imarticus

Submitted By:

Shifali Suvarna

Manju P

Shawaf Mohammed

Course and Batch: ML-MAR-02

IBM Big Data and Machine Learning Prodegree

Capstone Project: Project Report

Abstract

Market Basket Analysis is a technique used by retailers to understand customer behavior while purchasing from their stores. In the process of online shopping, you have probably seen a section called "suggestions for you" or "customers who bought this item also bought" in which Market Basket Analysis plays an important role. The implementation of this analysis was aided by the initiation of electronic point of sales systems. Store owners used handwritten records and digital records of the customer transactions which were generated by point of sales system. This was effectively used to analyze ample amount of data to know about customer purchasing behavior and pattern.

In this report we are going to understand and help GroceryKart to make use of their customer transaction data and focus on descriptive analysis on the customer purchase patterns, items which are bought together and units that are highly purchased from the store to facilitate reordering and maintaining adequate product stock.

We have created features related to user_id's and the products purchased by them, to predict whether a product will be reordered or not. This will be used in the modeling and validation stages.

IBM Big Data and Machine Learning Prodegree

Capstone Project: Project Report

Acknowledgements

We are using this opportunity to express our gratitude to everyone who supported us throughout the course of this group project. We are thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

Further, we were fortunate to have Gagan and Priyanshu Panda as our tutors. He/She has readily shared his immense knowledge in data analytics and guide us in a manner that the outcome resulted in enhancing our data skills.

We wish to thank, all the faculties, as this project utilized knowledge gained from every course that formed the Big data and Machine Learning program.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Date: May 30,2020 Shifali Suvarna

Place: Bangalore Manju P

Shawaf Mohammed

IBM Big Data and Machine Learning Prodegree

Capstone Project: Project Report

Certificate of Completion

I hereby certify that the project titled “Predicting Purchase Behaviour on E-Commerce Data” was undertaken and completed under my supervision by Shifali Suvarna, Manju P and Shawaf Mohammed from the batch of ML-MAR-02.

Date: May 30, 2020

Place – Bangalore

IBM Big Data and Machine Learning Prodegree

Capstone Project: Project Report

Table of Contents

Tables

IBM Big Data and Machine Learning Prodegree

Capstone Project: Project Report

CHAPTER 1: INTRODUCTION

1.1 Objective of the project

In the modern world, with the advancements coming in the Internet of Things, the necessity for innovation is in high demand. According to Stastica, the Statistic portal, 1.8 billion people worldwide purchased goods online in 2018, which will continue to increase in the following years. More research works and innovations are in progress in order to meet people’s demand and their satisfaction. One of the key techniques used to understand the customer purchasing behavior is the analysis on their transaction details. Understanding the transaction is a must to any form of business and its effect will lead to increase in sales. Especially in a retail store, it can be achieved by understanding the purchasing pattern of customers and related products which were sold together. This enables impulse buying from customers and also to understand their usual purchasing pattern and their effects towards the retail market.

GroceryKart is an e-commerce website that allows users to shop for groceries from a local grocery store online, and then sends a GroceryKart personal shopper to pick up and deliver the orders made by users the same day. These processes allow retailers to conduct analysis on purchase iterations by users but understanding the customer purchasing patterns and behaviors can become tedious and challenging.

1.2 Need of the project

If the predictions are made correctly like which products the customers is going to buy before the customer orders them, it can give companies a huge advantage in terms of warehouse stocking, delivery times, marketing strategy, improving customer experience on their app etc.

1.3 Data Description

Six datasets have been provided. Those are

a) aisles.csv

This dataset provides information on the aisles such as aisle ID and aisle names, through which the products were organized .

Variables	Description
Aisle ID	Labels the ID of the aisles
Aisle name	Mentions the aisle name in the retail stores

Table 1.1 : Aisle

b) departments.csv

This dataset provides information on the departments such as department names and department Id.

Variables	Description
Department ID	Labels the ID of the departments
Department name	Mentions the department name in the retail stores

Table 1.2 : Department

d) order_products_train.csv

This dataset is the same as order_products_prior and it is a trained dataset.

Variables	Description
Order ID	Labels the ID of the order made by customer
Product ID	Labels the ID of the products purchased by customers

Add to cart order	Sequence of the order placed in the cart
Reordered	Denotes whether the products are reordered or not

Table 1.3 : order_products_train

c) order_products_prior.csv

This dataset gives information on the orders, products, and reordered products

e) orders.csv

This dataset has information about the customer orders like order ID, order number, week day of the order, hour of the order, user ID and days since prior order.

Variables	Description
Order ID	Labels the ID of the order made by customers
User ID	Labels the ID of the users who made the purchase
Order number	Denotes the order number made by the customer
Order_dow	Denotes the day of the week, the order made by the customer
Order hour of day	Denotes the hour of the day, the order made by the customer
Days since prior order	Denotes the number of days since last order

Table 1.4 : orders

g) products.csv

This dataset gives information on the products such as product name, product ID, aisle and departments, which were sold to the customer .

Variables	Description
Product ID	Labels the ID of the products purchased by customers
Product Name	Denotes the product name purchased by the customer
Aisle ID	Labels the ID of the aisles
Departments ID	Labels the ID of the departments

Table 1.5 : products

1.4 Techniques used

Tools: Google Colab

Programming Language: Python

Libraries: pandas, numpy, matplotlib, seaborn, scikit learn

CHAPTER 2: DATA PREPARATION AND UNDERSTANDING

One of the first steps we engaged in was to outline the sequence of steps that we will be following for our project. Each of these steps are elaborated below

2.1 Data Extraction and Cleaning:

- As we have been provided with six datasets, we have to combine the required datasets to get our final dataframe for training.
- We have combined the prior, train, orders and product datasets.
- After combining we found that the merged dataset had missing values only in the "*days_since_prior_order*" column. As the null values are only 6% of the dataset, we can remove those values.
- While checking for outliers the "*order number*" column had negative values, which is not possible, therefore we converted the negative values to positive values.

2.2 Exploratory Data Analysis:

- Customers mostly order on the 0th day of the week.
- Customers normally order between the 10th and 15th hour of the day.
- Most reordered products are bananas, organic bananas, organic strawberries, organic baby spinach and organic hass avocado.
- Most of the ordered products are from the produce department.

2.3 Feature Engineering:

- To improve the performance of the machine learning models we created new dataframes that can be later be added as columns earlier merged dataset, thus forming our final dataset
- "*Total_no_of_orders* " dataframe shows the total number of orders for each user as per the products purchased by them.
- "*average_days"* dataframe implies the average days taken by the user to order their products.
- "*average_reorder" dataframe* shows the average of the reorder for each product bought by that user
- "*product_num_everytime"* dataframe tells the position of the product in cart, as preferred by the user
- "*product_time"* dataframe tells the hour of the day the user normally orders that product.

- “*product_dow'* represents the day of the week the user normally orders that product.
- *“product_order_number*” dataframe tells the maximum orders placed by the user for that product.

2.4 Splitting and Training:

- The “*reordered”* column is taken as the target.
- As the dataset size is very large, we took a sample of 30% from the data for modeling and validation.
- The sample dataset is then split into training and testing with 70% of the data being used for training.

CHAPTER 3: FITTING MODELS TO DATA AND VALIDATION

We have applied machine learning algorithms like logistic regression, decision tree, random forest, AdaBoost Classifier and XGB Classifier. The best model is accessed and validated with the help of a model comparison table comparing all the results of the different models used.

MODEL	ACCURACY SCORE	ROC SCORE
Logistic Regression	0.883552	0.865316
Decision tree with entropy	0.933571	0.929250
Decision tree with gini	0.932288	0.928187
Random Forest	0.948789	0.937771
AdaBoost Classifier	0.928748	0.911965
XGB Classifier	0.929561	0.909759

Table 3.1 Comparing all the models

As can be seen from the table above, accuracy and ROC score is the highest for Random Forest.

K-fold cross-validation is performed on the Random Forest model, where the number of splits taken is 15. Here the validation score is 0.9488.

CHAPTER 4: CONCLUSION

Based on the models that predict the reorder of products, some of the recommendations have been made:

- Based on the prediction of the next product, customers can be given additional offers by bundling the products together for a lesser price and customize those products .
- Based on the reordering model, personalized communications can be made by reminding the customers to reorder the products or can be added to the cart automatically based on the customer preferences.
- A suggestion list can be provided when they make their purchase in order to enhance the customer experience.

CHAPTER 5: REFERENCES

- Journal article - A.A. Raorane, R.V. Kulkarni, B.D. Jitkar, Association Rule – Extracting Knowledge Using Market Basket Analysis, Research Journal of Recent Sciences, 1 (2) (2012), pp. 19-27

- Journal article - A. Herman, L.E. Forcum, Joo Harry. Using Market Basket Analysis in Management Research, Journal of Management, 39 (7) (2013), pp. 1799-1824

- Website - Megaputer blog, An introduction to market basket analysis. Retrieved from https://www.megaputer.com/introduction-to-market-basket-analysis/

- Website - Margaret Rouse, Basic understanding of Market basket analysis. Retrieved from https://searchcustomerexperience.techtarget.com/definition/market-basket-analysis