IDS 561 Analytics for Big Data - Project Report

Black Friday Sales Prediction using PySpark

Gnana Teja Peddi (UIN: 674047493)

Shubham Sharma (UIN: 675565126)

Problem Setting

The objective of this project is to analyze the data and build a model to predict the purchase amount of customer against various products which will help the company to create a personalized offer for customers against different products in the upcoming Black Friday sale.

Data Description

The dataset is extracted from the Kaggle website and belongs to a retail company "ABC Private Limited". This is an ongoing Kaggle competition. We decided to take this challenge and tried to find the solution using PySpark. We built models to understand the customer purchase behavior (specifically, purchase amount) against various products of different categories. The purchase summary of various customers for selected high-volume products from last month is available.

The data set contains 550069 rows and 12 columns. The dataset 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.

- The dataset here is a sample of the transactions made in a retail store.
- The store wants to know better the customer purchase behavior against different products.
- Specifically, here the problem is a regression problem where we are trying to predict the dependent variable (the amount of purchase) with the help of the information contained in the other variables.
- There are seven categorical variables to analyze
- the dataset contains two short type variables: Product_Category_2 and Product_Category_3

Variable	Description
User_ID	Customer unique identifier
Product_ID	Product unique identifier
Gender	Sex of user
Age	Age in bins
Occupation	Occupation of customer (Masked)
City_Category	Category of city (A, B, C)
Stay_in_current_city_years	Number of years stay in current city
Marital_Status	Marital Status of customer
Product_Category_1	Product Category 1- Clothes
Product_Category_2	Product Category 2 - Electronics
Product_Category_3	Product Category 3 – Home goods
Purchase	Purchase amount (Target Variable)

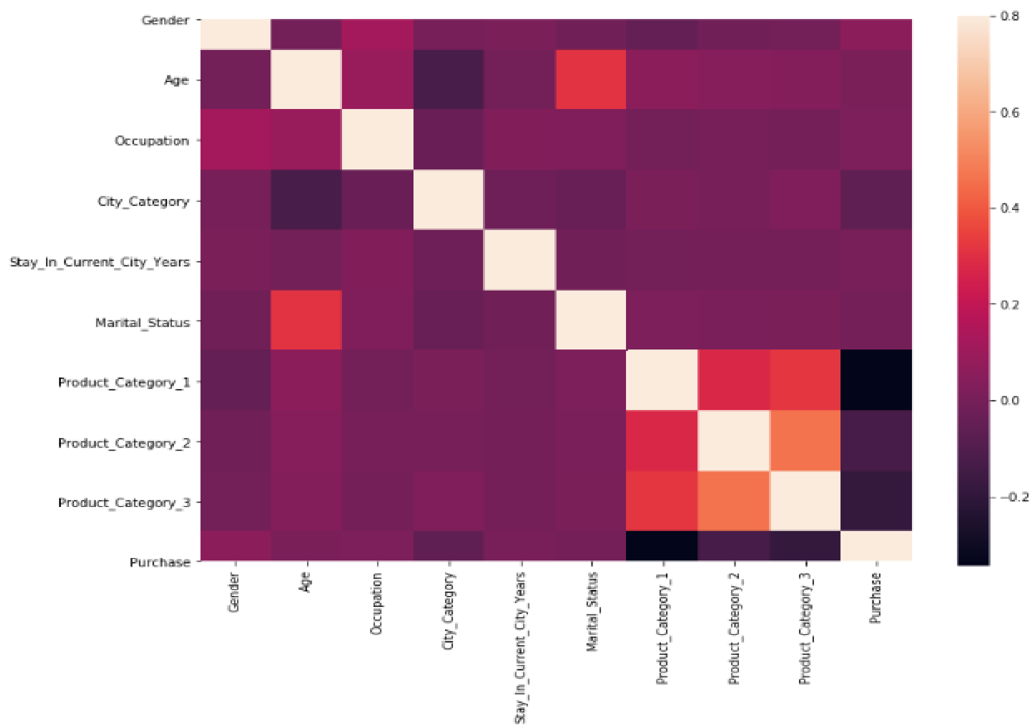
Data Cleaning

As we know that in order to use PySpark for analysis every variable should be in numeric form. We used dummy coding to converted categorical variables in numeric variables.

We performed the following steps to clean the data:

- Dropped irrelevant Variables - User_ID and Product_ID
- Replaced the null vales in Product Category 2 and Product Category 3 with mode (most frequent value)
- Removed '+' sign in Age and Stay_in _current_city_years variable.
- Converted Gender variable to numeric – M -> 1, F -> 0
- Converted Age bins into numeric.
- Converted City_category into numeric. (A -> 2, B -> 1, C -> 0)

Exploratory Data Analysis



Interpretations:

- Nothing is highly correlated with purchase variable
- Product_Category_1 has a negative correlation with Purchase.
- Maritial_Status and Age are strongly correlated as Expected
- Product_Category_3 has a strong correlation with Purchase.Maybe the products in this category were cheap.

Techniques

PySpark: We used Spark to do parallel computation on our huge dataset. Spark integrates well with Python. PySpark is the Python package that we used to make the magic happen.

In order to do exploratory data analysis and data cleaning we converted the Pyspark data frame into a pandas data frame and did the cleaning and analysis.

Machine Learning Pipeline

- At the core of the pyspark.ml module are the Transformer and Estimator classes
- We used a .transform() method that takes a DataFrame and returns a new DataFrame; usually the original one with a new column appended.
- We used a .fit() method. These methods also take a DataFrame, but instead of returning another DataFrame they return a model object.
- **Vector Assembler -** We combined all the columns containing our features into a single column. This must be done before modeling can take place because every Spark modeling routine expects the data to be in this form. We can do this by storing each of the values from a column as an entry in a vector.

```
+--------------------------+-----+
|features                  |label|
+--------------------------+-----+
|(9,[6,7,8],[1.0,2.0,15.0]) |11379|
|(9,[6,7,8],[2.0,4.0,9.0])  |9542 |
|(9,[6,7,8],[2.0,4.0,9.0])  |9761 |
|(9,[6,7,8],[2.0,4.0,14.0]) |12708|
|(9,[6,7,8],[2.0,4.0,16.0]) |13138|
|(9,[6,7,8],[2.0,4.0,16.0]) |13217|
|(9,[6,7,8],[3.0,4.0,5.0])  |7948 |
|(9,[6,7,8],[3.0,4.0,5.0])  |13270|
|(9,[6,7,8],[3.0,4.0,16.0]) |10821|
|(9,[6,7,8],[4.0,5.0,16.0]) |2061 |
|(9,[6,7,8],[4.0,5.0,16.0]) |2119 |
|(9,[6,7,8],[4.0,5.0,16.0]) |2770 |
|(9,[6,7,8],[4.0,5.0,16.0]) |2807 |
|(9,[6,7,8],[4.0,8.0,16.0]) |2099 |
|(9,[6,7,8],[5.0,8.0,16.0]) |3740 |
|(9,[6,7,8],[5.0,8.0,16.0]) |6888 |
|(9,[6,7,8],[5.0,11.0,16.0])|5365 |
|(9,[6,7,8],[5.0,14.0,16.0])|5205 |
|(9,[6,7,8],[5.0,14.0,16.0])|8817 |
|(9,[6,7,8],[5.0,14.0,16.0])|8831 |
+--------------------------+-----+
only showing top 20 rows
```

Splitting the Dataset: We split the data into test and train (80:20). Train dataset was used to train the model and test data was used to test the model.

Models

1. **Linear Regression-** We ran a linear regression model but a very poor Adjusted R^2 value of 13%
2. **Gradient Boosted Trees** – We ran a Gradient Boosted tree model and got an Adjusted R^2 value of 67%.

Results

features	label	prediction
(9,[6,7,8],[4.0,5.0,16.0])	2794	9254.764860117912
(9,[6,7,8],[8.0,8.0,16.0])	7915	7636.533777985297
(9,[4,6,7,8],[1.0,1.0,2.0,5.0])	15233	12190.110462702127
(9,[4,6,7,8],[1.0,1.0,15.0,16.0])	15759	10454.289513483309
(9,[4,6,7,8],[1.0,2.0,4.0,8.0])	15965	11317.020754045072
(9,[4,6,7,8],[1.0,2.0,4.0,9.0])	6438	11161.728120021817
(9,[4,6,7,8],[1.0,2.0,4.0,9.0])	16087	11161.728120021817
(9,[4,6,7,8],[1.0,3.0,4.0,16.0])	10973	9671.714332958354
(9,[4,6,7,8],[1.0,3.0,5.0,16.0])	7974	9669.591104115047
(9,[4,6,7,8],[1.0,4.0,5.0,12.0])	718	9887.796291307397
(9,[4,6,7,8],[1.0,4.0,5.0,16.0])	1450	9266.625755214372
(9,[4,6,7,8],[1.0,4.0,5.0,16.0])	2054	9266.625755214372
(9,[4,6,7,8],[1.0,5.0,8.0,16.0])	3584	8857.290719783778
(9,[4,6,7,8],[1.0,5.0,8.0,16.0])	5283	8857.290719783778
(9,[4,6,7,8],[1.0,5.0,8.0,16.0])	6875	8857.290719783778
(9,[4,6,7,8],[1.0,5.0,14.0,16.0])	6969	8844.551346723927
(9,[4,6,7,8],[1.0,5.0,14.0,16.0])	8766	8844.551346723927
(9,[4,6,7,8],[1.0,8.0,8.0,16.0])	5963	7648.394673081759
(9,[4,6,7,8],[1.0,8.0,8.0,16.0])	6178	7648.394673081759
(9,[4,6,7,8],[1.0,8.0,8.0,16.0])	9912	7648.394673081759

only showing top 20 rows

Model 1 – Linear Regression

features	label	prediction
(9,[6,7,8],[4.0,5.0,16.0])	2794	2905.806755479731
(9,[6,7,8],[8.0,8.0,16.0])	7915	8590.288033550183
(9,[4,6,7,8],[1.0,1.0,2.0,5.0])	15233	13508.645742561657
(9,[4,6,7,8],[1.0,1.0,15.0,16.0])	15759	14109.816904834077
(9,[4,6,7,8],[1.0,2.0,4.0,8.0])	15965	13400.21786301889
(9,[4,6,7,8],[1.0,2.0,4.0,9.0])	6438	11655.68027677495
(9,[4,6,7,8],[1.0,2.0,4.0,9.0])	16087	11655.68027677495
(9,[4,6,7,8],[1.0,3.0,4.0,16.0])	10973	9805.512814199854
(9,[4,6,7,8],[1.0,3.0,5.0,16.0])	7974	11306.573552734517
(9,[4,6,7,8],[1.0,4.0,5.0,12.0])	718	3087.8107187062105
(9,[4,6,7,8],[1.0,4.0,5.0,16.0])	1450	2210.8626279040486
(9,[4,6,7,8],[1.0,4.0,5.0,16.0])	2054	2210.8626279040486
(9,[4,6,7,8],[1.0,5.0,8.0,16.0])	3584	6461.110198812819
(9,[4,6,7,8],[1.0,5.0,8.0,16.0])	5283	6461.110198812819
(9,[4,6,7,8],[1.0,5.0,8.0,16.0])	6875	6461.110198812819
(9,[4,6,7,8],[1.0,5.0,14.0,16.0])	6969	6369.81529331221
(9,[4,6,7,8],[1.0,5.0,14.0,16.0])	8766	6369.81529331221
(9,[4,6,7,8],[1.0,8.0,8.0,16.0])	5963	7761.181776595798
(9,[4,6,7,8],[1.0,8.0,8.0,16.0])	6178	7761.181776595798
(9,[4,6,7,8],[1.0,8.0,8.0,16.0])	9912	7761.181776595798

only showing top 20 rows

Model 2- Gradient Boosted Trees

Our Final model (Model 2) was able to predict the target variable (Purchase amount) with an accuracy of 79% and an adjusted R2 value of 67%. **Gradient Boosted Trees** model was the better model.

Role of Team Members

1. **Gnana Teja Peddi:** Data cleaning, Exploratory Data analysis, report preparation
2. **Shubham Sharma:** Modelling, Inferences, report preparation, presentation preparation.