Electronics & ICT Academy
Indian Institute of Technology, Guwahati

Post Graduate Certification in
Data Science

Finding Repeatable Customers in E-Commerce

Industry Grade Project – III: PG Certification in Data Science by E&ICT Academy

Table of Contents

Aim of the Project

The objective of this project is to build a data-driven practical solution for Pink Clover to find the repeatable customers and understand the relationship between such customers.

Background

Pink Clover is proud to have been one of the most successful companies in the field of fashion eCommerce with its wide range of products on offer for more than 7 years. Irrespective of the size of the challenges, they have logistics, power, and technology to fulfill the ever-rising demands. The comprehensive range of products offered by them is top-notch in the market. From core fashion styles to premium denim product for every season and trend is manufactured by the designers of Pink Clover.

The designers of Pink Clover work with a vision to design unique products for everyone, every community, and eco-friendly around the globe. In recent years, many new competitors have emerged in the fashion industry offering cheap and affordable products with wide range of styles. The monthly sales of Pink Clover have dropped significantly as the competitors are offering the products at much lower rates. Pink Clover has built its reputation based on the highest quality of products and services to its customers. The decline in the sales is affecting the internal operations of the company as they are struggling with a loss in revenue.

Analytics team at Pink Clover analyzed the situation in every way possible. The recent reports from the analytics team focuses on customer retention. According to the recent report from the analytics team, the probability of selling a product to a new customer is around 15-30 percent as compared to a repeat customer.

The report also mentioned the famous Pareto principle. The 80/20 principle or the Pareto principle was the discovery of an Italian economist named Vilfredo Pareto. The Pareto principle states that roughly 80% of the effects come from 20% causes. Happy, repeat customers constituting around 20% of the customer base generate 80% of profits.

To retain the customer the analytics team suggested running campaigns and offers to attract new and to retain the old customers. The company ran promotional campaigns for several products to retain the customers. To boost the retention of customers, the best way is to find the loyal customers who purchase the product even after redeeming the offer (Repeatable Customers).

For example, if a customer is offered a coupon to pick up one extra pair of shoes on buying one. Of 100 customers who redeem this coupon, let us assume 30 of them purchase the pair of shoes again. Then the target is to find those 30 loyal customers out of those 100.

Let us assume, that Mr.X, who is a regular customer of Pink Clover stores, was offered a coupon to buy StarWars shirt on 24th April 2013. Then the complete transactional history of Mr.X from 1st Jan 2012 to 23rd April 2013 has been recorded.

Process Flow

As mentioned in the above use cases, the final objective of this project is to build Machine Learning models that can predict whether a customer will buy a product after redeeming the offer or not. To achieve that objective, you need to abide by the following three phases of Data Science:

1. ETL System

The company needs to have a system where all the data is centrally available in one place. As a Data Scientist, you will have to create a Datawarehouse that holds all the information in one place.

2. Predictive Model

Create a predictive model that predicts if a customer will buy a product after redeeming the offer or not. You will have to explore the data beforehand, and get familiar with the data, deal with the issues present in the data size and generate all the required features to train a better model.

3. Relationship between the customers and offers

Once the features are built, use them for clustering and find similar offers based on the clusters generated.

Tasks to be performed

Following are the guidelines to perform various tasks related to each of the three phases:

Using the datasets provided, and your analytical knowledge, perform the tasks mentioned below. Each task has been broken down into multiple sub-tasks for the benefit of the Data Scientists.

1. ETL System

With the growing business, the company needs to have a system where all the data is available in one place. As a Data Scientist, the company wants you to create a Datawarehouse by using SQLAlchemy. Use the following datasets to develop a schema for a Datawarehouse:

- offers.csv.gz: Contains details on offers available
- train_history.csv.gz: Contains details on products offered to each customer from 1st March 2013 to 3rd April 2013

- test_history.csv.gz: Contains details on products offered to each customer from 4th April 2013 to 30th April 2013

Sub-Tasks:

- Identify the dimensions and facts from the given data dictionaries
- Identify the fact table and dimension tables for the schema
- Identify the Star schema
- Using SQLite as OLAP store, write the ETL for the dimension and fact tables

2. Predictive Model

As highlighted in the problem statement, the company wants to discover repeatable customers.

Sub-Task 1: Basic Understanding of the Problem

Even though the problem statement is thoroughly explained, the sub-task has been designed to make sure you comprehend the main problem statement and how solving the problem helps the business.

Sub-Task 2: Basic Exploratory Data Analysis

The following points should stand out in the analysis:

- Understand the basic notion of a product here, i.e., a product is nothing but a combination of a brand, company, and category

- Customers from test files are different from those listed in the train file, and there is always one offer per customer
- Test offer dates are quite different from train offer dates
- Offer in the test file might be different from that in the train file. You cannot build one model per offer
- A customer may not have bought the product offered to them in the past
- Customer may return some products, and, in those cases, the product size, quantity or amount might be negative

Sub-Task 3: Handle the Transactions File Issue

- The transaction file is 2GB in compressed form, and when uncompressed, it expands as much as 11GB. You should be able to figure out how to solve the size problem
- Hint: The transaction file is sorted by customer id. You can break the transactions record into one file per customer
- To save more storage, you can save the file in Parquet File Format
- Performing all these tasks will result in around 2GB of transactions data in uncompressed form
- Now, whenever you need any data for processing, you can call these files in buffer memory and perform the analysis

Sub-Task 4: Feature Engineering

- Now, to build a model, we need some presentation of features
- Neither do we have product demographics, nor customer demographics. All we have is historical transactions per customer
- We know that the basic notion of a product is its brand, company, and category (the department is first 2 digits of category)
- Using the following combinations of the brand, category, and company, a transactional similarity can be made with each transaction and the product. For instance, how many times a customer has bought products from the brand or category or company. However, there are 11 possible combinations, as listed below:

(brand), (company), (category), (department), (brand, company), (brand, category), (brand, company, category), (department), (brand, department), (company, department), (brand, company, department)

- Now, we can have time lags on this as well, for instance, how many times the customer bought a product in last 10 days, or 20 days, 30 days, 60 days and so on till 360 days
- We have 4 transactional facts, i.e., how many times a product was bought and its aggregated size, aggregated quantity, and aggregated amount
- Note that some customers may return the product/commodity. Hence, all the metrics above apply to the returned products/commodities as well
- There are many chances the customer may have never bought the product on offer ever. So, some never bought features, for example, never bought brand (0,1), or never bought category X, etc.

- Including all these with time-lapse of (10,20,30,45,60,120,240,360) days, 960 features should be available

Sub-Task 5: Modeling

- The train - validation divide is not so simple. It should be done properly to represent the held-out set
- Now, if we say 80% of data in the training set, it means we should find a date from 1st March 2013 to 3rd April 2013, where the % of offers are ~80% in training
- After the train-validation divide, you can choose from a wide range of models like simple logistic regression to decision trees or bagging methods like random forests or boosting methods like Gradient Boosting machine or XGboost
- The validation evaluation can be done by AUC scoring, and then, you should report this number

Sub-Task 6: Evaluation

- You are supposed to submit the predictions in test data as their probabilities (as defined in submissions.csv.gz)
- Calculate the AUC score and it should be at least be >0.50, i.e., the model should be able to correctly classify at least 50 percent of the records in test data

3. Relationship between the Customers and Offers

To find the relationships between the customers and offer, you will have to apply clustering methods.

Sub-Tasks:

- There will be some categorical columns and some numerical columns; in such case, the clustering method cannot be applied directly. You can use **Gower's distance** to handle the mixed data
- Find a similar offer based on the clustering method applied

Dataset Description

To solve the problem, the company provides a Point of Sale (POS) for about 300k of its customers across various regions, markets, and stores, with their years' worth of transactions, until the offer was made to the customer.

For the training data, you are provided with data such as when the product was offered to the customer, and if the customer made a repeat purchase.

For test data, you are only provided with the offer data and offer id. You need to predict if the customer will make a repeated purchase of the product or not.

- Customers are equally divided between train and test
- Only one offer per customer is included in either of the two
- All brand, customer, company, and category information is masked with ids
- The first two digits of category represent a department

Complete List of Files

File	Description
train_history.csv.gz	Training data in compressed
test_hisotry.csv.gz	Test data in compressed form
transactions.csv.gz	Transaction history of each customer in compressed form
offers.csv.gz	Information on the offers in compressed form

You are provided with the following files:

- **train_history.csv.gz:** Contains information on when the offer was made to the customer and if the customer became a repeat customer. All the offers that were made from 1st March 2013 to 3rdth April 2013.

Column Name	Data Type	Description
id	UUID	Unique identification of the customer
chain	UUID	Unique identification of the store
offer	UUID	Unique identification of the offer

market	UUID	Unique identification of the market
repeattrips	Numeric	Number of times the customer bought the same offered product after he redeemed the offer.
repeater	String	T if repeattrips>0 else f. (this is the response variable)
offerdate	Date(YYYY-MM-DD)	Date on which the offer was made

- test_hisotry.csv.gz: Contains information on when the offer was made to the customer. The response variable is not given. All the offers that were made from 4th April 2013 to 30th April 2013.

Column Name	Data Type	Description
id	UUID	Unique identification of the customer
chain	UUID	Unique identification of the store
offer	UUID	Unique identification of the offer

market	UUID	Unique identification of the market
offerdate	Date(YYYY-MM-DD)	Date on which the offer was made

- **transactions.csv.gz:** Contains a year's transaction history of each customer until they redeemed the offer. All POS data for each customer from 1st Jan 2012 till purchase date.

Column Name	Data Type	Description
id	UUID	Unique identification of the customer
chain	UUID	Unique identification of the store
dept	UUID	Unique identification of the product department
category	UUID	Unique identification of the product category
company	UUID	Unique identification of the product company
brand	UUID	Unique identification of the product brand
date	Date(YYYY-MM-DD)	The date of purchase

productsize	Numeric	Volume of the unit product purchases
productmeasure	Numeric	Unit of measure
purchasequantity	Numeric	Number of units purchased
purchaseamount	Numeric	Net amount in dollars of the purchase

- offers.csv.gz: Contains information on category, company, brand, the offer, and the net utility value (in dollars) to the customer.

Column Name	Data Type	Description
offer	UUID	Unique identification of the offer
company	UUID	Unique identification of the product company in offer
brand	UUID	Unique identification of the product brand in offer
quantity	Numeric	Minimum number of units of product to be purchased to get the discount
offervalue	Numeric	Net utility value (in dollars) to the customer

How to Start with the Project?

1. Log in to the **Google Co-lab**, load the notebook to the environment. Go to **Runtime** and choose the **Change runtime type**.

2. Import all the necessary Python packages: NumPy and pandas for numerical processing, data importing, preprocessing, etc., Matplotlib for plotting pose joints and showing images, sci-kit-learn for splitting datasets, and PySpark for performing machine learning on Big Data.

3. From here, you can take over to the project and start building the machine learning model and ETL system.

How to submit your project?

- Share your project solution for each process different jupyter notebooks(.ipynb) files and submission.csv.gz to support@edureka.co
- The .ipynb file should contain the details of each step in the markdown
- Add a description of the approach taken for each of the three phases

- Add answers to any subjective decisions made throughout the process in the markdown
- You can even upload your code into your GitHub repository and share the link of your repository with us

Marks Allocation

1. ETL Sytem [30 Marks]
2. Predictive Model [50 Marks] (Marks based on the number of datasets combined to build the features)
3. Relationship between the customers and offers [20 Marks]
4. A detailed explanation of steps in the markdown is important to allocate marks