greatlearning

PGP BDML

Capstone Project – Recommender Systems

Interim Report

1. Proposed Solution

We are working on building a Recommendation engine for the capstone project, product recommendation is a filtering system that seeks to predict and show the items that a user would most likely purchase.

For now, we have covered the Popularity based recommendation that list the top 10 popular items purchased by the customers. Below are the steps taken to build the model -

1. Importing the necessary packages
2. Loading the Retail dataset [https://archive.ics.uci.edu/ml/datasets/online+retail].
3. Understand the data to get insights on the features
4. Perform Exploratory Data Analysis(EDA)
5. Sanitize the data
6. Visualize the data to find hidden patterns
7. Build a Popularity based recommendation system which lists top 10 popular items for a particular country. Using the popularity, we will also observe the purchase trends of the popular items across the regions.

2. Evaluation metrics

Popularity based recommender systems offers very primitive results, there is not really an evaluation metrics that we can consider for our project as we do not have real time data feed.
We have instead, split the dataset into training and test sets and evaluated the purchase power of the popular items over the test set.

The data is split based on the timeframe into 75% and 25% i.e.
Training Data - Dec 01 2010 to Sep 01 2011
Test Data - Sep 02 2011 to Dec 09 2011

Going forward as we build more models, the models will be evaluated based on the intuitional sense the recommendation system is making.

For popularity based recommendation system, the model recommends items to the users based on how popular those items are among users.
The stock code for the item, for which the highest number of quantity sold is considered as the most popular item.

Below is a snapshot of the popular items purchased in France.

```
top_10 = popular_items(input())
top_10
```

```
France
```

StockCode	Description
21086	SET/6 RED SPOTTY PAPER CUPS
21094	SET/6 RED SPOTTY PAPER PLATES
21212	PACK OF 72 RETROSPOT CAKE CASES
21731	RED TOADSTOOL LED NIGHT LIGHT
22492	MINI PAINT SET VINTAGE
22551	PLASTERS IN TIN SPACEBOY
22554	PLASTERS IN TIN WOODLAND ANIMALS
22556	PLASTERS IN TIN CIRCUS PARADE
23084	RABBIT NIGHT LIGHT
84879	ASSORTED COLOUR BIRD ORNAMENT

3. Exploratory Data Analysis

1. Total number of records is 541909
2. Analysis on the data and its types, the dataset contains 8 features:
   - InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
   - StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
   - Description: Product (item) name. Nominal.
   - Quantity: The quantities of each product (item) per transaction. Numeric.
   - InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.
   - UnitPrice: Unit price. Numeric, Product price per unit in sterling.
   - CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
   - Country: Country name. Nominal, the name of the country where each customer resides.
3. There are 38 countries data in the retail dataset and includes one country tagged Unspecified, we need more information on this country.
4. 'Unspecified' country has contributed ONLY for purchases (2667 units of currency) and no return or cancelled transactions.
5. Description and Customer ID have null and empty values. There are 1454 empty records for Description and 135080 for Customer ID.
6. Percentage of missing customer IDs is pretty significant (25%) and that could impact the results.
7. There are negative quantity records, the negative quantity related to returned or cancelled items, the invoice No for return transactions start with C.
8. Dataset contains Sale and Return/Cancelled orders. Sale transactions are 22064 transactions and return/cancelled transactions are 3836.
9. The data contains a year of transactional information from 2010-12-01 to 2011-12-09.

10. There are about 15% of return or cancel transactions, we can do some study on the return transactions to identify items that have patterns in these transactions and help prevent future cancellations.
11. The data is biased towards United Kingdom, contributes to about 90% of dataset.

```
retail_dataframe.Country.value_counts()
```

```
United Kingdom          495478
Germany                   9495
France                    8557
EIRE                      8196
Spain                     2533
Netherlands               2371
Belgium                   2069
Switzerland               2002
Portugal                  1519
Australia                 1259
```

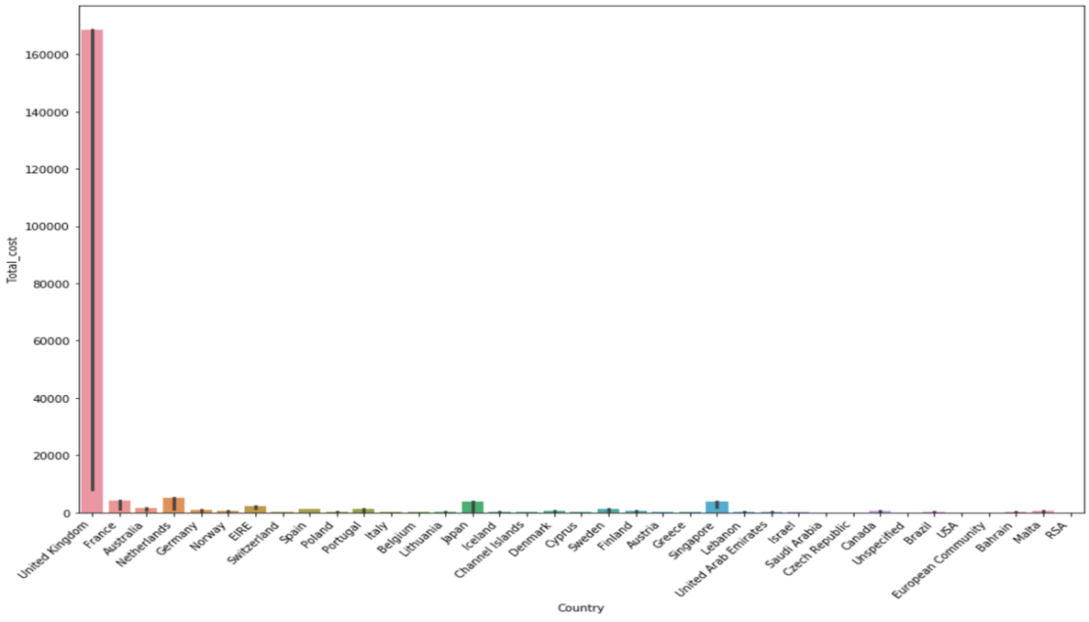
Transactional data summary -

- *Total number of records: 541909*
- *Total number of Transactions: 25900*
- *Number of Sale Transactions: 22064*
- *Number of Cancel or Return Transactions: 3836*
- *Percentage of Cancel or Return Transactions: 15 %*
- *Total number of distinct stock codes purchased or cancelled: 4070*
- *Total number Customers purchased or cancelled transactions with the retail chain: 4372*
- *Total number of countries: 38*
- *Total number of customer records missing 135080*
- *Percentage of customer records missing 25.0 %*

4. Exploratory Visualization

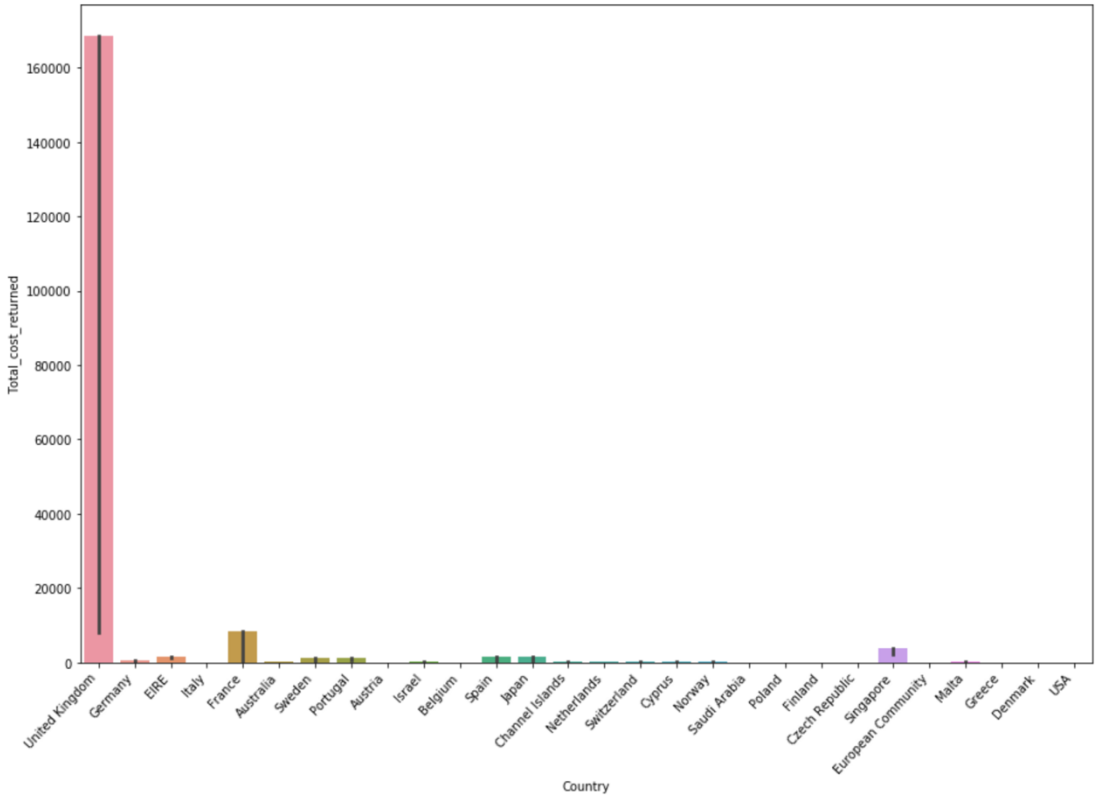
Purchase transactions:

1. United Kingdom and Netherlands have contributed as the top two revenue generators, while Saudi Arabia and Bahrain have contributed the least



Return/Cancel transactions:

1. United Kingdom, EIRE and France have the highest rate of returns or cancelled transaction amounts, while European Community and Saudi Arabia have the least cancelled transaction amounts.



Country	Quantity	Total_cost_returned_abs
United Kingdom	-260939	540518.16
EIRE	-4196	15260.68
France	-1624	12311.21
Singapore	-7	12158.90
Germany	-1815	7168.93

2. An interesting observation is Singapore had only 7 items each of quantity 1 returned and has the 4th highest return amount. These products have StockCode 'M' with description 'Manual', there seems to be something wrong here as the product codes aren't available. Will need to examine these invoices more in detail.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Total_cost_returned_abs
144830	C548830	M	Manual	-1	2011-04-04 13:08:00	162.60	12744.0	Singapore	162.60
144831	C548830	M	Manual	-1	2011-04-04 13:08:00	2382.92	12744.0	Singapore	2382.92
144832	C548830	M	Manual	-1	2011-04-04 13:08:00	239.30	12744.0	Singapore	239.30
144833	C548830	M	Manual	-1	2011-04-04 13:08:00	1252.95	12744.0	Singapore	1252.95
144834	C548834	M	Manual	-1	2011-04-04 13:09:00	2053.07	12744.0	Singapore	2053.07
406404	C571750	M	Manual	-1	2011-10-19 11:16:00	3949.32	12744.0	Singapore	3949.32
406405	C571750	M	Manual	-1	2011-10-19 11:16:00	2118.74	12744.0	Singapore	2118.74

5. Summary of Initial Findings

Revenue contributors

Increase in revenue (Purchasing power)

Country	Quantity	Total_cost
United Kingdom	4269472	7.308392e+06
Netherlands	200937	2.854463e+05
EIRE	140525	2.655459e+05
Germany	119263	2.288671e+05
France	111472	2.090240e+05

United Kingdom and Netherlands have contributed as the top two revenue generators

Country	Quantity	Total_cost
Brazil	356	1143.60
RSA	352	1002.31
Czech Republic	671	826.74
Bahrain	260	548.40
Saudi Arabia	80	145.92

Saudi Arabia and Bahrain have contributed the least

Decrease in revenue (Cancelled transactions)

Country	Quantity	Total_cost_returned_abs
United Kingdom	-260939	540518.16
EIRE	-4196	15260.68
France	-1624	12311.21
Singapore	-7	12158.90
Germany	-1815	7168.93

United Kingdom, EIRE and France have the highest rate of returns or cancelled transaction amounts

Country	Quantity	Total_cost_returned_abs
Czech Republic	-79	119.02
Greece	-1	50.00
Austria	-54	44.36
Saudi Arabia	-5	14.75
European Community	-2	8.50

European Community and Saudi Arabia have the least cancelled transaction amounts.

Popular items (specific to a country)

```
top_10 = popular_items(input())
top_10
```

```
France
```

StockCode	Description
21086	SET/6 RED SPOTTY PAPER CUPS
21094	SET/6 RED SPOTTY PAPER PLATES
21212	PACK OF 72 RETROSPOT CAKE CASES
21731	RED TOADSTOOL LED NIGHT LIGHT
22492	MINI PAINT SET VINTAGE
22551	PLASTERS IN TIN SPACEBOY
22554	PLASTERS IN TIN WOODLAND ANIMALS
22556	PLASTERS IN TIN CIRCUS PARADE
23084	RABBIT NIGHT LIGHT
84879	ASSORTED COLOUR BIRD ORNAMENT

Purchase trends

The data is skewed to the European market and it can be seen that France and United Kingdom have a good purchase trend of items purchased during the 1st 3 quarters of 2011 and the last quarter of 2011. They have 7 and 5 items common in the two segments of purchase history.

There are lot of countries (Japan, Portugal, Greece, Israel, USA, Brazil, RSA, Austria, Hong Kong, Canada, Saudi Arabia) where there are no common items purchased during two segments of purchase history.

6. Challenges

Below are some of the challenges -

- the dataset is not clean, has a lot of missing values for Customer ID and product description. Besides, multiple products (StockCode) have more than one description.
- the dataset is biased, has lot of data related to European market and more so towards United Kingdom
- no performance metrics to validate the authenticity of the algorithm
- the huge dataset could pose challenges with respect to data processing when working with other recommendation engine models
- there is no real time data feed to evaluate the model, we are working on the static data and no incremental transactions are coming to validate the findings

7. NextSteps

We have completed the primitive popularity based recommendation. Taking the foundation based on our finding in this Algorithm, we will be exploring other recommendation engine models (Collaborative filtering, Content-Based Filtering, Hybrid Recommendation Systems).