Machine Learning Engineer Nanodegree

Capstone Proposal

James Wanderi Kinyanjui, June 17th 2019

Proposal

Segmentation of Credit Card Holders

Domain Background

One of the major revenue contributors of a financial institution such as a bank is cash advance to their customers in order to spend and pay them later at an interest. Credit card is a payment card issued by such an institution to their customers to enable the cardholder to pay a merchant for goods and services based on the customer's promise to the institution to pay them for the amounts plus other agreed charges [1].

In marketing and risk-avoidance strategies targeting the credit card holders, the banks should be able to know the holders' spending behaviours, affordability and credit risk in such a way that way that decisions can be made easily. Knowing if a card holder expends their credit card limit in the first week of the month might inform the bank to double the limit of the holder by updating the limit fortnightly on condition that holder can repay the credit advance.

This project shows a way of segmenting the card holders so that such decisions can be made easily by the bank in order to be adaptive in their product offerings, avoid risks, be effective and make profits.

Problem Statement

There is a rapid credit card adoption in most countries and this is challenging to the credit card issues in offering personalized solutions to the card holders.

If issuers are able to better predict consumer behaviour, they have a stronger chance of offering highly personalized solutions [2].

We would like to be able to derive groups of these cardholders that will be useful in informing the bank's decision in their product offerings. The goal of this project is to demonstrate how the grouping can be achieved and how these groups can be used in decision making.

Datasets and Inputs

The dataset is provided by Kaggle [3,4].
Only one file is provided with 8950 instances of customer credit card details.

Input Data Fields

- *CUST_ID* - Identification of credit card holder. Not the real identification.
- *BALANCE* - Available amount in the credit card holder to make purchases with.
- BALANCE_FREQUENCY - Frequency of the balance update. Value ranging from 0 to 1 with 0 being less frequently updated and 1 being high frequency of updating.
- PURCHASES - Amount of purchases.
- ONEOFF_PURCHASES - Maximum purchase amount paid in one instant.
- INSTALLMENTS_PURCHASES - Amount of purchases made in installments.
- CASH_ADVANCE - Cash advance by the cardholder.
- PURCHASES_FREQUENCY - Frequency of the purchases. Value ranging from 0 to 1 with 0 being less frequently and 1 being high frequency.
- ONEOFF_PURCHASES_FREQUENCY - Frequency of the one-off purchases.. Value ranging from 0 to 1 with 0 being less frequently and 1 being high frequency.
- PURCHASES_INSTALLMENTS_FREQUENCY - Frequency of the installment purchases. Value ranging from 0 to 1 with 0 being less frequently and 1 being high frequency.
- CASH_ADVANCE_FREQUENCY - Frequency of the cash advance by the card holder. Value ranging from 0 to 1 with 0 being less frequently and 1 being high frequency.
- CASH_ADVANCE_TRX - Number of transactions made when doing cash advance.
- PURCHASES_TRX - Number of purchase transaction made.
- CREDIT_LIMIT - Limit of credit card
- PAYMENTS - Amount of payment done by the card holder to the card issuer.
- MINIMUM_PAYMENTS - Minimum amount of payment done by the card holder to the card issuer.
- PRC_FULL_PAYMENT - Percent of full payment done by the card holder to the card issuer.
- TENURE - Tenure of credit card services to the card holder.

Solution Statement

Our goal is to segment the card holders and show the usefulness of segmentation. Therefore, the solution will comprise of the following:

1. Use 90% of data to derive useful clusters. The different clusters derived will then be assigned to every card holder.
   This data will then be split into 3 parts for training a predicting model.

The remaining data (10%) will be considered as new instances for prediction.

2. A Jupyter [5] notebook demonstrating the above. This notebook will be submitted to Kaggle.

Benchmark Model

A simple K-Means [6] with 2 clusters will be used to derive the useful clusters

Evaluation Metrics

We will use Silhouette score [7, 8] to evaluate the numbers of clusters derived.

Project Design

Step 1: Literature review

Read relevant materials on clustering using Machine Learning techniques and credit card issuance, marketing and usage within the banking sectors. This will prepare me to understand the problem to be solved and how to solve it.

Step 2: Data exploration

Explore the data through visualizations and code in order to understand how features are related to one another. We will observe statistical description of the dataset and consider the relevance of each feature.

Step 3: Data preprocessing

Through understanding of the data, we will create a better representation of data by performing feature scaling and outliers detection and removal.

Step 4: Cluster generation

We will use principal component analysis (PCA) to better understand and draw conclusions about the data. Through this process we will reduce the dimensionality of the data and use our benchmark model in deriving clusters.

We will then optimize our benchmark and/or use other clustering techniques in coming up with the best clusters based on the Silhouette score [7, 8].

We will visualize the data based on the derived clusters.

Step 5: Build predicting model

We will analyse, give meaning and give appropriate names to the clusters. We will then use this as labels to supervised learning.

Using this labelled data, we will build a prediction model that will be able to assign a label to new instance of data.

Step 6: Write final report and conclusions

As part of the final report of the project, clearly describe what we have done and its applications in the real world of credit card issuers.

References

[1] - O'Sullivan, Arthur; Steven M. Sheffrin (2003). Economics: Principles in action (Textbook). Upper Saddle River, New Jersey 07458: Pearson Prentice Hall. p. 261. ISBN 0-13-063085-3.

[2] - Mathur, Varun (2018). How Credit Card Issuers Can Gain from Machine Learning. https://www.publicissapient.com/news/How-credit-card-issuers-can-gain-from-machine-learning. Accessed on June 17, 2019.

[3] - Kaggle - (https://www.kaggle.com)

[4] - Bhasin, Arjun (2018). Credit Card Dataset for Clustering. https://www.kaggle.com/arjunbhasin2013/ccdata. Accessed on June 17, 2019

[5] - Jupyter - (https://jupyter.org/)

[6] - Wikipedia - (https://en.wikipedia.org/wiki/K-means_clustering)

[7] - Peter J. Rousseeuw (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics 20: 53-65. https://www.sciencedirect.com/science/article/pii/0377042787901257

[8] - Wikipedia - (https://en.wikipedia.org/wiki/Silhouette_(clustering))