UofT credit union limited  
Proposal for Services

Overview

## UofT Credit Union is a financial institution that provides financial services including checking and saving accounts, financial advice, mortgages and more.

## According to the latest reports, shareholders found that digital engagement between the company and customers is lower than the expected rate. We know that online services are the fastest and most efficient way to reach the customer to present profitable products.

## This data analysis project firstly will focus on existing customer data such as demography, financial status, and spending behaviors. (Payment types and amounts.). According to the result of the analysis, the model can be implemented for future customers for forecasting purposes.

## Machine learning techniques will be our main tool for grouping the customers. By the end of our analysis, we will try to identify the group of customers that will be eligible to receive online services engagement promotions.

## Questions we hope to answer with the data

## What is the online bank penetration in the different age ranges?

## Is there a consumer cluster more prone to use Online bank?

## In the cluster that is more prone to use Online bank, are there users that do not use or that could use it more often?

## What behaviours can be detected in customers in 2018 to predict trends and offer promotions and discounts to drive online engagement?

## The Objective (Financial Services)

* #1 Identifying ideal online engagement criteria
* #2 increasing online engagement by offering financial products
* #3 implementing the model for future customers, or potential future customers (Forecasting)

## The Objective (Data Analytics)

* # Do we have sufficient data sources?
* # Does Relational Database Service (RDS) have the raw data?
* # Does Relational Database Management system have the connection and work properly?
* # Data cleansing tools are properly working and suitable with data
* # Feature(s) and target(s) (applicable If Supervised ML) are identified according to our approach.
* # Proper Machine Learning technique identified and tested.
* # Results visualized, relation between outcomes

## Dataset, Tool,

* #1 Source Data: CSV
* #2 Database Services: Amazon Web Services – S3 (Cloud)
* #3 Database Management: PostgreSQL
* #4 Working Environments: Google Colab and GitHub
* #4 Programming Languages: Python, Spark, SQL
* #5 Related Libraries: Scikit-learn, Pandas, Mongo, Matplotlib, NumPy
* #6 Visualization: Tableau, Matplotlib

## Data Source

## For this project we will use csv files that the credit union send us. There are:

## Owners: have the data from the owners of the credit union

## ATM payments: have payments released at ATMs

## Online\_AutPay: have the payments in Online banking and Automatically payment

## Wallet: have all the transactions using mobile wallet as google pay, Samsung pay, apple pay and mobile pay

## Tap\_pos: have all the transactions using tap or POS (password)

This project will use data from 2018 for create the model and present for the directors the results

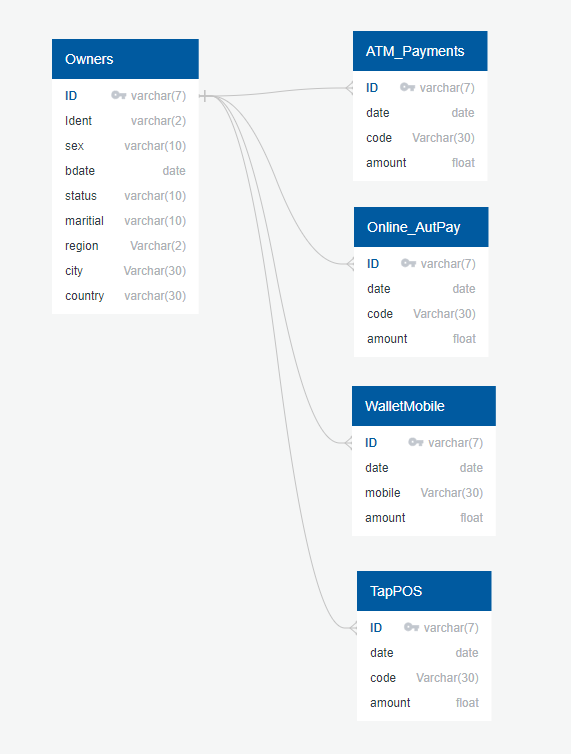
## Project Deliverables

Following is a complete list of all project deliverables:

|  |  |
| --- | --- |
| Deliverable | Description |
| Deliverable 1 | Extracting the data |
| Building the Amazon Web Services Bucket |
| Creating PostgreSQL Server, Connection with AWS and writing table queries |
| Deliverable #2 | Preprocessing data (Cleansing, Converting, Splitting) |
| Deliverable #3 | Identifying the Clusters of Preprocessed data |
| Converting cluster into class and merging class with clean data |
| 3D Visualization of the merged data frame |
| Deliverable #4 | Presentation of Overall Project. |

# Database Diagram

We will create 5 tables to import all the 5 csv files. The PK to link all Tables is ID that is a Varchar with 2 letters and 5 numbers.



## Machine Learning Model Mock Up

# Diagram Description automatically generated

We have chosen to use an unsupervised machine learning model, as we are hoping to find patterns and trends in the data. Unsupervised learning refers to training data with only input data and no corresponding output variables.

We will begin the model by importing all necessary packages/ modules. We will read and load the dataset into a dataframe. We will go over all the columns and remove any columns that are not necessary to the analysis. We will check the data type of each feature and transform as needed. Then, we will remove any rows where there is at least 1 null value and drop any duplicates. Once the data is cleaned and ready to be analyzed, we will remove the owner ID since it will not be used in the clustering algorithm. After, we will create variables for the text features and standardize the data. Once data is standardized, we will go through dimension reduction via principal component analysis. The Elbow Curve method is used to find Centroids (number of clusters) and the K-means Algorithm is initialized to define the clusters. We fit the model and predict the clusters. A new dataframe is created using the predicted clusters and features and merge the dataframe with the principal component analysis dataframe. The owner ID’s column is re-added into the clustered dataframe. The cluster results are then visualized with 2D and 3D plots.