Interim Progress Report

Project title: Customer segmentation using unsupervised learning

Table of Contents

INTRODUCTION

The theoretical background of the project:

Over the years, increasing competition between companies and the availability of vast historical data has made data mining techniques widespread for finding critical and strategic information hidden within client's information.

The clustering technique is a more popular data mining technique in order to analyze customer segmentation. The main aim of data mining (DM) for Customer segmentation is that the past data can contain information that will be useful in the future. The main objective of data mining in customer segmentation is to identify patterns from historical data. (Thomas, June 23, 2020)

Customer segmentation is a famous and important data mining application among established customers. A segmentation project starts with identifying business goals and finishes with distinguishing marketing strategies for the segments. There are various types of segmentation according to specific criteria or attributes used for segmentation. For customer segmentation, data mining offers different clustering algorithms. (Adithyan, Nov 20, 2020) In order to identify behavioral data, analyze natural groupings of customers and suggest a solution based on data patterns, we can use clustering algorithms. So as long as data mining models are properly built, they can unfold groups with distinct profiles and features and lead to many segmentation schemes related to business relevance and value. One of the most popular clustering algorithms for customer segmentation is k-means clustering. (Eslamijam, December 28, 2020).

Statistical analysis is used to discover common patterns of consumer behaviour. Different interaction behaviours for distinct parts of the general population will result in distinct prediction patterns for various predictor variables, implying the need for segment-specific scorecards. (Kienow, April 26, 2018)

Overview of Project:

Today, credit cards are one of the most significantly used payment methods. These cards are used for services and products and are also sometimes used to borrow money. People use a credit card to make annual, weekly, monthly, and even daily acquisitions. This is a great opportunity for companies to research what people buy and when they buy it? They can be used to purchase high-quality products that otherwise could not be afforded, doing online transactions, as well as emergency transactions for unexpected expenses, and are more convenient than carrying cash, reducing the circulation of counterfeit money. Thus the use of credit cards is better than cash and debit cards. Customer segmentation is the process of classifying customers according to their industry, geography, revenue, account size, and staff count in order to determine where risk and opportunity exist within the portfolio. (Panek, 9 May 2019)

Credit card issuers traditionally target consumers using information about their behavior and demographics. In most departments, i.e., Risk department, Operations department, sales sector, and customer Service department, segmentation of credit card customers can be used. But marketing teams tend to benefit most of it.

In credit card marketing, customer segmentation can be used extensively for customer value and lifecycle management, including but not limited to:

- Increase your wallet participation and upgrade your primary card by identifying potential cardholders.
- Encourage banking products and cross-selling of secondary cards depending upon customer understanding.
- Update limits of cards and levels selectively for utmost return on risks.
- Retain the most valuable customers through accurate assessment of customer value and risk.
- Guide clients to use more profitable payment products and communication channels, also increase the profitability of payments.
- Increasing clients' appropriateness in communication and offers resulting in date settlement and - loyalty.

The **goal of this project** is the **segmentation of credit card users/holders**. For this document, we define and explain the necessary steps that can be taken to develop a model for grouping German credit card holders according to age, term/duration, and loan amount. I'll be using a data set that contains a number of behavioral attributes for credit card customers. The data set called "German Credit data" can be downloaded from the Kaggle website. This project aims to divide a bank's credit card users into segments that can help the bank understand their credit card users. We will segment our data using k-Means clustering, an algorithm for clustering unlabeled records for this project.

Tools used in this project:

In order to perform customer segmentation by k-means clustering, we used **python** as a data science tool.

Reason for choosing python:

Total of 8.2 million active Python users, "Up to 69% of learning developers and data scientists now use Python (compared to 24% of those using R). Python is popular with data scientists because of its simplicity. Although some data scientists have a computer science background or are familiar with other programming languages, many data scientists have experience or no experience in statistics, mathematics, or other technical fields. The grammar is easy to use and write, making this programming language easy to learn and easy to learn. Python is specifically popular among data scientists.

Deliverables

A context-level model that we used in our project is a Cross-industry standard process for data mining known as CRISP-DM. In our project, we used the CRISP-DM process model as system architecture. It consists of six phases, from business understanding to deployment. This process outlines all steps that can help us to perform data mining activities from business understanding to deployment. Since it is the most widely used data mining model & also easy to understand, we choose the CRISP approach for our project.

Phases of CRISP-DM

1. **BUSINESS UNDERSTANDING:** First, we need to understand the objectives of our project and requirements from a business perspective, then states these objectives in terms of data mining goals and prepare a preliminary plan for achieving the goals. Business understanding is explained briefly in the overview section and also explain in more detail in the final report.

2. **DATA COLLECTION**: We collected data about the customers of German bank from Kaggle.

3. **DATA PREPARATION:** The next step is to perform the data cleaning and transformation process. Data cleaning is done using different functions and libraries of python, which will explain in the final report.

4. **DATA MODELLING:** The next step is to select the modeling technique, test it, build the model and then assess it. For segmentation of bank users according to age, credit amount, and duration, we select k-means clustering algorithm.

5. **EVALUATION:** In this step, we thoroughly evaluate the model, review all the steps executed, and analyze the results

6. **DEPLOYMENT:** In this phase, the plan for deployment and for maintenance of the system must be made. In this project, the deployment phase is not included.

PROGRESS

Accomplishments

- ✓ Data loading into pandas data frame and understand the data set.
- ✓ Learn data type of every column in a data frame.
- ✓ Checking missing values in the data frame and also handle these null values.
- ✓ Checking duplicate rows in a data set.
- ✓ Perform and analyze univariate exploratory data analysis (EDA) of columns that have object datatype by plotting their count plots.
- ✓ Then analyze numerical columns using histogram and correlation matrix.
- ✓ Now perform bivariate analysis of both object types columns and numerical and write the observations.
- ✓ Before applying the k-means algorithm, we remove skewness of data as our age, duration, and credit columns are skewed.

- ✓ Finding optimal k-value by using different methods.

Problems

Finding optimal 'k' value for k-means:

For our dataset, finding the optimal k—value for the algorithm is difficult. First, I used the elbow method for identifying the value of k that gives good results. But Elbow methods fail for our dataset as the elbow is not clearly visible. Then I search for another method that is the silhouette score method, that is more efficient than the elbow method.

Planned work

Major tasks

Literature review

In this step, I review different works of literature regarding our business problem, i.e., customer segmentation using KMeans. As a result of this, we are able to understand our business goal.

Data set collection

We collected data from Kaggle website, which is data of German bank customers that holds credit cards.

Research different type of python libraries and algorithms

In order to perform segmentation in python, we study different python libraries, which we will discuss in the final report. Also, study k-means algorithms in depth from different online sources. This step helps in selecting the optimal 'k' value in K-means.

K-means algorithm

After all the data preprocessing steps, we plan to build the KMeans clustering algorithm, which leads to developing credit cardholders. This task is not completed yet.

Conclusion

Create different clusters and then interpretive these customers so that issuers understand their customers better.

Final report preparation:

Half report is done, which consists of theoretical things. The section that briefly explains coding steps and observation is remaining.

Bibliography

Adithyan, N., Nov 20, 2020. *medium.* [Online]
Available at: https://medium.com/codex/customer-segmentation-with-k-means-in-python-18336fb915be
[Accessed 25 June 2021].

Eslamijam, M., December 28, 2020. *bdtechtalks.* [Online]
Available at: https://bdtechtalks.com/2020/12/28/machine-learning-customer-segmentation/
[Accessed 25 June 2021].

Kienow, R., April 26, 2018 . *experian.* [Online]
Available at: https://www.experian.com/blogs/insights/2018/04/customer-segmentation-generation-potential-segments/
[Accessed 26 June 2021].

Panek, T., 9 May 2019. *dnb.* [Online]
Available at: https://www.dnb.com/ie/perspectives/finance-credit-risk/perfecting-your-portfolio-perspective.html
[Accessed 26 June 2021].

Thomas, S. R., June 23, 2020. *yieldify.* [Online]
Available at: https://www.yieldify.com/blog/types-of-market-segmentation/
[Accessed 25 June 2021].