Credit Card Holder’s Risk

Group Number 7

1st David Huang 2nd Shraddha Waphare 3rd Yash Abhyankar 4th Venkateswararao Para

5024818 50442793 50442683 50442199

[dhuang34@buffalo.edu](mailto:dhuang34@buffalo.edu) [swaphare@buffalo.edu](mailto:swaphare@buffalo.edu) [yabhyank@buffalo.edu](mailto:yabhayank@buffalo.edu) [vpara@buffalo.edu](mailto:vpara@buffalo.edu)

Introduction

The interest rate on credit cards in the US is increasing due to high inflation and uncertainty of the financial market. In this model, we tried to determine the risks by analyzing the spending habits of credit card holders by segmenting the customers into different groups. Based on the result of our model, we can suggest financial management to reduce the risk of overspending and spread financial awareness. Overall, managing one’s credit card debt and controlling one’s spending habits is important to prevent going into debt. Currently, the credit card debt in the US is right now increasing at a rapid rate to 887 billion dollars in debt with inflation at a high rate of 8.2% each month which means many US citizens would need to curb their spending habits to ensure they don’t fall into debt which is becoming easier to do.

Methods/Results

Our dataset summarizes the usage behavior of about 9000 active credit cardholders during the last 6 months. The file is at a customer level with 18 behavioral variables like Balance, Purchase, Credit Limit, Payments, Cash Advance Frequency. The source of our data is Kaggle with link: [Data link](https://www.kaggle.com/datasets/jillanisofttech/market-segmentation-in-insurance-unsupervised?resource=download).

Our data is well structured and hence there was no need to rescale the data. We cleaned the data by removing null values and deleting the column Customer ID as it was just serving unique identification to the data. We also performed the EDA. By plotting the correlation matrix we checked the correlation of variables with each other. Before determining the cluster of Customers, we conducted PCA to reduce the dimensionality but maintain information as much as possible. We selected the first 10 PCs as they are covering 90% of data.

We used clustering methods to segment the customers.

1. K-means Clustering was used to determine how many clusters we need to divide the Customers into, which may represent their profile. From there we can determine what kind of offers we can provide to them. From the Elbow, the range of optimal k value is estimated to be between 4 and 7 (inclusive of both). For k = 7, there is an overlapping of clusters which is not suitable. Thus k=6 is estimated to be the choice for the number of model clusters.

Chart, scatter chart

Description automatically generated

Fig 1. K-means Clustering

1. Density Based Modeling has two main parameters that are epsilon and minimum points (Minpts). For each choice of Minpts, there is an optimal epsilon which determines the radius of the cluster. In our case the epsilon value is 0.15 for nearest neighborhood points equal to 20 which produced 4 different clusters.

Chart, scatter chart

Description automatically generated

Fig 2. Density Based Clustering

1. Distribution Based Modeling determines the top three of the predefined models based on the Bayesian Information Centric metric value. Preferably, the lowest value of BIC to maximize the maximum likelihood of the model prediction using the least amount of parameters. The corresponding algorithm is the expectation maximization. The selected models and respective number of allocated clusters are chosen to be any one of the best three based on the set seed value. One of the models showed 4 different clusters with some overlapping clusters.

Chart, scatter chart

Description automatically generated

Fig 3. Distribution Based Clustering

Conclusion

The final conclusion regarding this project is that there could be multiple different clusters ranging from 4 clusters from the distribution based clustering and density based clustering to 6 in the KNN clustering methods. This means that there would be approximately 4 to 6 different groups of risk levels with 4 to 6 different spending habits classifications.

Github Link: <https://github.com/david164b/EAS509Project_SDM>