Title:: Card Limit Analysis for Targeted Marketing

Submitted by Kirthi Chandra (Hero Vired)

Client:TwoDotSeven

TwoDotSeven Pte Ltd, 30 Cecil Street, #19-08 Prudential Tower, Singapore 049712,
contactus@twodotseven.com

Contents

1. Business Objective, Understanding and High-level Approach

The aim of this project is to create a card limit analysis using Python that can be help in targeted marketing efforts. This will analyze customers' card limits and provide valuable insights to optimize marketing campaigns and tailor offers to specific customer segments.

Business Objectives

This analysis can help the business minimize credit card default risk, increase revenue from interest and fees, and identify profitable customers for targeted marketing.

Problem Statement

By analyzing the factors that influence the setting of a card limit, the bank can benefit in several ways. The analysis can provide valuable insights into customer segments and their creditworthiness, allowing the bank to make informed decisions when determining credit limits. This can help minimize the risk of default by setting appropriate credit limits for different customers based on their income, credit rating, and other relevant factors. Additionally, the analysis can identify potential high-value customers who may be eligible for higher credit limits, enabling the bank to tailor marketing strategies to attract and retain profitable customers. Ultimately, this data-driven approach can lead to improved risk management, targeted marketing efforts, and increased profitability for the bank.

Task-1 -Data Cleaning & Visualization

The importance of data cleaning in the development of machine learning models cannot be overstated as it ensures that the model is trained on clean and unbiased data, resulting in more accurate predictions and decisions. By identifying and addressing errors and biases in the data, data cleaning plays a crucial role in enhancing the performance of machine learning models.

1. Join datasets of different formats using Python based on common fields.
2. Clean the data by removing duplicate, missing, and irrelevant values.
3. Analyze the data to understand its characteristics and trends.
4. Address any errors or inconsistencies in the data.
5. Ensure that the data is suitable for analysis

Task-2 -Exploratory Data Analysis (EDA)

is a technique that enables the examination and visualization of data to uncover patterns and trends that may not be obvious. It's an essential step in the machine learning model-building process as it helps to understand the data's characteristics and patterns. Through EDA, data scientists can discover relationships and trends in the data that can assist in building and training the model, and gain insights into the data that can inform the development of the model. In summary, EDA is a powerful tool for gaining a deeper understanding of the data and guiding the development of machine learning models.

1. Conduct comprehensive exploratory data analysis (EDA) to gain deep insights into the dataset.
2. Utilize descriptive statistics to understand the dataset's key features and distributions.
3. Create insightful visualizations, including histograms, box plots, and scatter plots, to identify trends and anomalies.
4. Analyze categorical features using pie charts and bar plots to understand their distributions.

5. Identify outliers and patterns in the data using appropriate statistical methods.
6. Perform hypothesis testing to validate assumptions and test correlations between variables.
7. Conduct correlation analysis to identify the relationships between numerical variables.
8. Utilize inferential statistical methods, such as t-tests to compare groups and analyze differences.
9. Apply chi-square tests to analyze associations between categorical variables.
10. Calculate measures of central tendency (mean, median, mode) and dispersion (variance, standard deviation) for relevant variables.
11. Apply data transformation techniques (e.g., logarithmic transformation) to normalize skewed variables.
12. Analyze the distribution of gender, ethnicity, education level, and marital status among the individuals.
13. Determine the percentage of students among the individuals and assess their credit ratings and credit limits.
14. Explore the relationship between income and credit limit using correlation and regression analysis.
15. Assess the impact of education level on credit limits by comparing different groups.
16. Calculate the average number of cards owned by individuals and analyze its relationship with credit limits.
17. Evaluate the distribution of account balances and identify individuals with high or low balances.

2. Methodology and Inferences

1. Importing and combining the Data based on customer ID using Outer Join
2. Data Inspection shows no null values and no Duplicates and all the data types are in their respective mode.

3. Data Cleaning & Visualization

3.1 New Columns Created:

Rating_Value columns is added based on CIBIL Score value categorising it in the form of poor (<550), average (<650), good (<750) and excellent.

3.2 Dropping existing Columns:

Droping Name and CustomerID as they are redundant in understanding the Credit Limit Analysis because of its unique variable nature.

3.3 Handling Outliers:

The maximum outliers for income column which do not fall in the wisker valleys are 9.33% but they are not showing any anomalies of spread, the same logic goes with Limit (2.75%), Rating (1.75%) and Cards (.92%) and hence no changes done for these outliers and would be retained. In the same case there are a handfull of 70 observations which shows outliers in all the three fields (Limit, Rating and Income) and these 70 observations are treated by eliminating the observations.

3.4 Feature Engineering:

There are certain features which are very few in numbers(<1%) like Divorced, Separated and Widow in marital status column and has been removed in this column to make the model more reliable. The feature where the balance equal zero covers 24.06% of the whole data set and is retained in order to understand the characteristics. This zero can also be the result of error like data imputation.

4. Address any errors or inconsistencies in the data

There is a high correlation between Balance, Limit and Rating and correlating these continuous variable with the rest categorical variable shows no sign of any abnormalities. But removing the balance equal zero is reducing the variance of the dataset.

The rest bivariate analysis shows the relevance of Caucasians, Female and Masters features in their predominant fields.

5. Measures of Central Tendency and Dispersion

The majority of users (maximum number) are 25 years old, but the average age across all users is 55. While the maximum number of users have a current account balance of 0, the average account balance is 446 with a standard deviation of 453.

```
The majority of users have around 3 cards each, as indicated by
the central measures.
```

```
The maximum number of users have an income of 10.7, but the aver
age income is 46.29 with a standard deviation of 36.9.
```

```
The limit values vary significantly, with a large variance of 47
,76,959.00. The mean limit is 4406, and the standard deviation i
s 2,185.63.
```

6. Inferential Statistical Analysis

Based on the inference, we can state that a statistical test indicates that the means of the continuous variables, namely Rating, Income, and Limit, are not equal to each other.

When comes to continuous variable there is a good association between the following combinations columns only 1. Gender and Education 2. Education and Ethnicity

7. Predictive Modelling

7.1 KMeans Clustering -- Continuous Variable

K-means clustering with 2 clusters shows that income can be divided into two segments with average incomes of 37.37 and 56.88, respectively. The corresponding average credit limits are 2,998.13 and 6,080.41, and average ratings are 237.32 and 431.00. However, the impact of these factors on credit card balance is not linear. Instead, the average balances are 161.64 and 785.31, respectively. This suggests that credit card balances are more extreme in cases where income, limit, and rating are all high or all low.

7.2 Decision Tree -- Continuous Variable

Rating appears to have a higher value in component analysis. This is indicated by the r2_score metric, which has a value of 0.417. An r2_score of 0.417 indicates that the algorithm is not performing very well, but it is still learning about the relationship between rating and the other elements in the dataset.

7.3 KMeans Clustering -- Continuous Variable (Rating)

The column named rating_cluster_3 has the following characteristics in other fields: Rating_cluster_3 with the number 2 represents a group with the highest average values for all aspects, including: 1. Income mean of 73 2. Credit limit of 6,486 3. Rating of 619 4. Number of cards of 3 5. Credit card balance of 824 Rating_cluster_3 with the number 2 also represents a group with the lowest average values for all aspects, including:

1. Income mean of 39
2. Credit limit of 2,429
3. Rating of 158
4. Number of cards of 2
5. Credit card balance of 59

7.4 Correlation Analysis -- Continuous Variables

There is a good association between the following combinations columns only

1. Gender and Education
2. Education and Ethnicity

8. Understanding the Relationship

8.1 Income and Credit Limit

```
    The Correlation is good from income to limit
```

8.2 Education Level on Credit Limit

The relation between the Education and Rating is same in terms of count and sum in terms of their hierarchy, except for one exception of Doctorate.

8.3 Average No. of Cards and its Limits

People on an average uses 3 cards

8.4 Understanding of Balance

The people with high credit balance are unmarried, female, Caucasian and either bachelor or masters education

9. A Few KPI Question which can be raised to check the dataset

9.1 Credit Card Default Risk

In summary, targeting senior citizens, monitoring customers with multiple credit cards, and considering the relationship between income and credit card balance can help in understanding credit card default risk. Senior citizens' higher ratings, the equal usage of credit cards by married and unmarried individuals, and the stability of credit card balances relative to income levels contribute to a better understanding of credit card default risk.

9.2 Revenue from Interest and Fees

Targeting senior citizens, particularly those who are Caucasian, and focusing on customers with higher credit card balances, credit card limits, and ratings can help credit card companies maximize revenue from interest and fees.

9.3 Profitable Customers for Targeted Marketing

The most profitable customers for targeted marketing efforts would likely be senior Caucasian individuals with a Master's in Education, who have higher ratings and potentially higher incomes. These customers possess characteristics that align with higher income levels and show a positive correlation between income and rating.

10. Collective Inferences

Understanding the credit card default risk can be approached as follows:

In summary, targeting senior citizens, monitoring customers with multiple credit cards, and considering the relationship between income and credit card balance can help in understanding credit card default risk. Senior citizens' higher ratings, the equal usage of credit cards by married and unmarried individuals, and the stability of credit card balances relative to income levels contribute to a better understanding of credit card default risk.

Revenue from Interest and Fees

Targeting senior citizens, particularly those who are Caucasian, and focusing on customers with higher credit card balances, credit card limits, and ratings can help credit card companies maximize revenue from interest and fees.

Profitable Customers for Targeted Marketing

The most profitable customers for targeted marketing efforts would likely be senior Caucasian individuals with a Master's in Education, who have higher ratings and potentially higher incomes. These customers possess characteristics that align with higher income levels and show a positive correlation between income and rating.

_________________THANK YOU_________________