2025

  TAN, James Anthroi A. - 240350922

  TAN, Xiuhao - 240244326

  ZHUANG, Beibei - 240383013

IT114116 - DSAI

2025/7/10



Credit Card Customer in DSA Bank

Table of Content

[Introduction 2](#_Toc202813805)

[Data 3](#_Toc202813806)

[Setting and Utilizing GitHub 4](#_Toc202813807)

[Roles 4](#_Toc202813808)

[Product Backlog items 4](#_Toc202813809)

[Sprint Backlog items 4](#_Toc202813810)

[Evaluation 5](#_Toc202813811)

[Conclusion 7](#_Toc202813812)

# Introduction

This project was initiated by the DSAI bank, and the project is aimed at consulting the credit card utilization circumstance on customers and provide management suggestions or strategy to seek characteristics of potential customers and have innovative insight for strengthening the market share. The suggestion and strategy are based on the following considerations:

1. They like to know the expected income of customers who are not stated their income information.
2. Overview of Existing and Attrition Customer circumstance
3. The pattern of customers' credit card utilization behaviours
4. New customer labelling and credit limit prediction

# Data

The only one dataset is from DSA bank and stored in an open dataset platform Kaggle. The dataset includes customer information and transaction records. The following is the metadata table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Field | Type | Description | range | Sample |
| CLIENTNUM | String | Customer Identification | - | 768805383 |
| Customer\_Age | Integer | Customer's Age in Years | 0 to 120 | 45 |
| Dependent\_count | Integer | Number of dependents of the customer's family | - | 4 |
| Credit\_Limit | Integer | Credit Limit on the Credit Card | - | 21341 |
| Total\_Revolving\_Bal | Integer | The money has not been paid off in full by the end of the billing cycle. | - | 233 |
| Avg\_Open\_To\_Buy | Integer | the difference between the total credit limit on the card and the current balance | - | 15412 |
| Total\_Trans\_Amt | Integer | Total Transaction Amount (Last 12 months) | - | 25 |
| Total\_Trans\_Ct | Integer | Total Transaction Count (Last 12 months) | - | 5 |
| Avg\_Utilization\_Ratio | Double | The % of credit consumed in terms of credit limit | 0 to 1 | 0.1547 |

# Setting and Utilizing GitHub

Provide the link of your GitHub project and some screen capture, including treeviews and document changing capture.

**GitHub Link**: <https://github.com/XiuhaoTan/TAN_4475Project.git>

# Roles

|  |  |  |
| --- | --- | --- |
| Name | Role | Responsibilities |
| TAN James Anthroi | Serving Product Owner |  |
| TAN Xiuhao | Product Owner |  |
| ZHUANG Beibei | Development Team |  |

# Product Backlog items

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sprint** | **Task** | **Category** | **priority** | **status** |
| 1 | Uploading data to R | DATA | LOW | Under Review |
| 1 | Data Cleansing | DATA | HIGH | Under Review |
| 2 | Relationship between price and bedroom numbers | Modelling | Middle | Committed |

# Sprint Backlog items

Use the template provided and make a screen capture for all sprint backlog item record

# Evaluation

1. Provide one insight of the customer’s information with respect to different education background

We first tabulated counts and average credit metrics by education level. We found that 3,128 customers hold graduate degrees, 2,013 have only high-school diplomas, and 1,487 report no formal education. When we joined that to average credit-limit and utilization data, graduate educated customers had 12% higher of average credit limits and 20% lower utilization ratios versus the uneducated group. That suggests higher-educated users enjoy better credit access and manage it more conservatively likely due to greater financial literacy and job stability.

For the bank, this means:

* Target premium products and loyalty rewards at graduates, who both carry higher limits and present lower default risk.
* Offer tailored financial-literacy tools or simplified credit-management features for lower-education segments to reduce over-utilization and late payments.

This is useful for the bank because those people with higher education may have more stable jobs and income, so they could be potential high-value customers. While people with lower education might need different credit strategies or support.

1. Guess the Income of Unknow Income Category customers

DSAI Bank encountered a group of customers who didn’t provide their income information, making it difficult to categorize them properly. To estimate their likely income range, we used their **average open-to-buy amount**, which reflects how much of their credit remains unused. This metric serves as a useful indirect indicator of financial capacity, as higher-income customers generally maintain higher available credit.

The average open-to-buy for customers marked “Unknown” is **$8,401.52**. When compared to known income brackets:

* Customers earning **$40K–$60K** have an average open-to-buy of **$4,290.27**
* Those in the **$60K–$80K** group average **$9,603.78**
* Higher-income customers (above $80K) show even larger values
* Lower-income groups fall much further below

The value for “Unknown” customers is much closer to the **$60K–$80K** bracket and far above lower-income groups. This makes it reasonable to assume that these individuals likely fall within the **middle-income range** of $60K–$80K. Their credit behavior supports this conclusion, allowing the bank to assign them fairly without needing direct income data.

1. Observe relation between transaction amount and it counts and provide one insight regards to customer consumption pattern

We plotted each customer’s total transaction amount against their total transaction count and fitted a linear regression (R²=0.65). The scatter shows a clear, positive correlation—customers who spend more tend to transact more often—but also three visible clusters:

* Low-count/low-spend: occasional users making few, small purchases
* Mid-count/mid-spend: steady users with moderate transaction habits
* High-count/high-spend: “power users” whose frequent small purchases drive large totals

We think that the transaction frequency is the primary driver of total spend. In practical terms, encouraging customers to increase the number of transactions (even small ones) will boost revenue more effectively than pushing up individual purchase sizes. Tailoring rewards—e.g., points per swipe, merchant-specific micro-offers—can tap into habitual buying behaviors and lift overall card usage.

1. Observe one distinguish utilization behavior in the customer dataset and raise one insight regards to customer behavior.

Analysis of credit data reveals that customers who keep their revolving balances very low (under $305) consistently achieve excellent credit ratings because they likely pay off their balances monthly, avoiding debt entirely. However, when balances exceed $1,173, risk surges dramatically—even among high-income earners ($80K+). For example, 59% of customers with balances over $1,173 receive "Poor" ratings, showing that high utilization (maxing out cards or missing payments) overpowers income advantages. This identifies a critical financial threshold: disciplined behavior (low balances) is the strongest predictor of credit health, while balances over $1,173 signal severe financial stress regardless of salary. Banks should prioritize helping customers stay below this danger zone to reduce default risk.

1. Extract and name the four observable types of customers and introduce the characteristics of each customer

After performing a four-cluster Gaussian mixture model using each customer’s transaction count and age, we examined the age statistics to understand who these customers really are. The results show four clear segments, each with unique behavior and life stage characteristics.

**1. Steady Mid-Age Spenders (Cluster 1)**

Customers in this group have an average age of 51, ranging from 39 to 65 years old. They tend to have well-established financial routines and moderate spending habits. Many likely use their credit cards for consistent, planned payments such as household expenses, bills, or healthcare. While not heavy swipers, they maintain regular usage patterns and are considered reliable clients.

**2. Dynamic Young Professionals (Cluster 2)**

Averaging about 40 years old and ranging from 26 to 46, these customers are younger adults with rising financial responsibilities. Their behavior points to more frequent transactions—digital payments, lifestyle purchases, and possibly family-related spending. This group would respond well to flexible rewards and mobile-based offers that fit their active lifestyles.

**3. Tech-Savvy Cautious Users (Cluster 3)**

Also aged 26 to 47, with an average near 39, this cluster resembles Cluster 2 but may be slightly more conservative. These users are comfortable with credit yet show measured habits. They might prefer essentials-based purchases and would benefit from educational offers like cashback tips or balance management tools.

**4. Senior Occasional Users (Cluster 4)**

This oldest segment has an age range of 48 to 73, with an average age near 54. They tend to transact less frequently and only for larger or scheduled purchases—travel, medical needs, or one-off insurance payments. They value stability and clarity. Premium services and concierge-style support would suit their cautious, long-term banking approach.

By combining age profiles with transaction behavior, this clustering approach allows the bank to define its customers not just by how they spend, but by who they are in life. These insights support more thoughtful product design, targeted outreach, and stronger customer relationships.

1. By comparing the similar cases with existing customers, evaluate the quality of the classification model, and you may suggest some advice(s) the improvement of the model quality

To evaluate the quality of the Naive Bayes classification model, we compared its predicted utilization rankings (Util\_Rank) for new customers with the actual rankings assigned to existing customers who share similar profiles. After applying the model, we found that most predictions for new customers fell under either the “Excellent” or “Poor” categories, while “Fair” and “Good” were rarely predicted. This trend closely mirrors the original training data, where the majority of existing customers were also labeled “Excellent” or “Good,” and only a small number were assigned to the “Fair” class. This suggests that the model has learned to focus on the most common patterns in the data, but struggles to recognize less frequent categories. When we examined specific individual cases—such as a customer aged 53, graduate level education, married, and earning between $80K and $120K—the predicted class was “Excellent.” We then filtered similar customers from the training set and observed that many of them also had an “Excellent” Util\_Rank, reinforcing that the model can make accurate predictions when the input is clearly aligned with known patterns.

However, when analyzing predictions classified as “Poor,” there were more inconsistencies. For several customers with moderate income and education levels, the model assigned “Poor” even when comparable individuals in the training data had better utilization ranks. This inconsistency may indicate that Naive Bayes—being a relatively simple model—doesn’t fully capture the relationships between variables, especially when features interact in complex ways. Additionally, since it assumes independence between predictors, it may overlook nuanced behavior that would require deeper analysis. To improve the model’s performance, we recommend balancing the dataset by ensuring that each class is represented more equally during training. Cleaning or refining categorical inputs—such as simplifying education levels or income brackets—may also help the model better distinguish between customer groups. Finally, considering a more advanced classification algorithm, like a decision tree or random forest, could allow the model to capture patterns that Naive Bayes typically overlooks, ultimately leading to more reliable predictions across all categories.

1. By comparing the similar cases with existing customers, evaluate the quality of the Prediction model, and you may suggest some advice(s) to improve the model quality.

To assess the prediction model’s reliability, we compared the credit limit forecast for a new customer with the actual credit limits of existing customers who had similar profiles. The selected new customer is 53 years old, male, with one dependent, holds a “Graduate” education level, is married, earns between **$80K and $120K**, has been with the bank for 47 months, and has a relationship count of 3. The model predicted a credit limit of **$6,686.13**. To evaluate this, we reviewed actual credit limits of comparable customers in the historical dataset, which ranged from **$1,438 to $15,987**, with a median of **$2,826** and an average of **$3,728**. While the predicted credit limit is above average, it remains within the observed range, particularly for customers in higher income brackets. This suggests the model appropriately captures customer behavior, though it may slightly overestimate limits for certain demographics. To improve model precision, we recommend refining income encoding (e.g. using numeric ranges), examining the impact of relationship duration, and evaluating alternate algorithms such as Gradient Boosting Machines for stronger predictive accuracy.

# Conclusion

DSAI Bank’s customer analysis has uncovered several meaningful insights that can help improve how the bank serves its customers and grows its business. One of the strongest patterns comes from examining customer education levels. A large number of the bank’s customers—over 3,000 individuals—have graduate degrees, and about 2,000 more have completed college. These highly educated customers typically show stronger financial discipline. On average, they receive higher credit limits and maintain lower utilization ratios, meaning they don’t use up all their available credit and are less likely to miss payments. This makes them ideal candidates for premium financial products, loyalty programs, and long-term relationship offerings. On the other hand, customers with only high school education or no formal education show more frequent signs of financial stress. They tend to carry higher balances, have smaller credit limits, and use a larger portion of their available credit. These customers may benefit from extra financial guidance, such as budgeting tools, simplified apps, or more frequent communication from relationship managers.

Another insight comes from the group of customers who listed their income category as “Unknown.” While at first this group may seem difficult to understand, a closer analysis helps us guess where they likely fit. By comparing their average open-to-buy amount—about $8,400—to other income groups, we found that they behave very similarly to those who report income between $60K and $80K. Their spending and credit usage patterns align closely with middle-income customers, so it would be reasonable to treat them as such when planning marketing offers or assigning credit limits. Doing this avoids accidentally undervaluing them or offering the wrong products, which helps build trust and improve customer satisfaction.

The analysis also revealed a strong pattern in how customers spend. When we looked at transaction count and total spending for each customer, there was a clear connection: people who spent more were also the ones making more transactions. In other words, frequent use of the card—even for small purchases—adds up to higher total spending over time. We noticed three groups: low-activity customers who make only a few transactions, steady users who swipe regularly but don’t overspend, and high-activity users who swipe often and contribute the most revenue. This finding suggests that it’s not just big purchases that matter. Encouraging customers to make frequent small purchases—through things like rewards per transaction or daily spending challenges—could be a strong way to increase revenue and engagement.

In credit behavior, one standout insight is about how customers manage their balance. Those who keep their balances very low—especially under $305—are much more likely to receive an “Excellent” credit rating. These users likely pay off their balance each month or stay within a comfortable spending range. On the other hand, when balances go above $1,173, the number of “Poor” credit ratings rises sharply, even among those earning high incomes. This suggests that high balances are more dangerous than low income when it comes to credit health. It also shows the bank should pay close attention to customers with large balances, even if their salary is high. Proactive tools like alerts, limit reminders, or helpful budgeting nudges could prevent these customers from becoming high-risk cases.

Finally, by grouping customers using clustering methods based on age and transaction count, we discovered four main types of users. First, there are **Steady Mid-Age Spenders**, usually aged 39 to 65, who spend moderately and stick to routines—great for stable, long-term plans. Next, **Dynamic Young Professionals** in their late 20s to mid-40s tend to spend actively and respond well to digital offers and flexible services. The third group, **Tech-Savvy Cautious Users**, is similar in age but spends more carefully; they may value educational tools or spending control features. Last are **Senior Occasional Users**, aged 48 to 73, who use their credit cards less frequently and mostly for large, one-off expenses. They may prefer personalized support, clear statements, and services that prioritize peace of mind. Understanding these groups allows the bank to better design campaigns, services, and rewards that match each group’s lifestyle and needs.

Altogether, these findings help DSAI Bank not only understand how its customers behave, but also how to better serve them. With more thoughtful credit strategies, better data use, and customer-centered product development, the bank can increase satisfaction, lower risk, and grow its long-term value.