Bati Bank Buy-Now-Pay-Later Credit Scoring: Exploratory Data Analysis (EDA) Interim Report

By Wolderufael Kassahun

Introduction:

This report details the initial findings from the Exploratory Data Analysis (EDA) stage of developing a credit scoring model for Bati Bank's new Buy-Now-Pay-Later (BNPL) service. We're leveraging data provided by the partnering eCommerce platform (Xente Challenge dataset) to understand customer behaviour and predict their creditworthiness.

Task 1 - Understanding Credit Risk

Credit risk is the financial risk that a borrower will default on their loan obligations, meaning they will fail to repay the principal and interest on time. This risk is a key concern for lenders, as it can lead to financial losses.

Factors Affecting Credit Risk:

- **Borrower's Financial Situation:** Income, debt-to-income ratio, credit history, and employment stability are crucial factors in assessing credit risk.
- **Economic Conditions:** Economic downturns can increase the likelihood of borrowers defaulting due to job losses or financial difficulties.
- **Loan Terms:** The interest rate, loan term, and collateral requirements can influence the level of credit risk.
- **Lender's Policies:** The lender's policies and procedures for evaluating creditworthiness and managing risk also play a role.

Managing Credit Risk:

- **Credit Scoring:** Lenders use credit scoring models to assess a borrower's creditworthiness based on various factors.
- **Collateral:** Requiring collateral (assets that can be seized in case of default) can mitigate credit risk.
- **Loan Terms:** Structuring loan terms with appropriate interest rates and repayment schedules can help manage risk.
- **Diversification:** Lenders can diversify their loan portfolio to reduce the impact of individual defaults.

Credit Risk Mitigation:

- **Credit Insurance:** Lenders can purchase credit insurance to protect themselves against losses due to borrower defaults.
- **Debt Collection:** Efficient debt collection practices can help recover funds from delinquent borrowers.

In conclusion, credit risk is a significant concern for lenders. By understanding the factors that affect credit risk and implementing effective risk management strategies, lenders can minimize their exposure to financial losses.

General Approach to Credit Risk Modelling and Scorecard Development:

The general steps involved in developing a credit risk model and scorecard:

1. Data Collection and Preparation:

- Gather relevant data on potential borrowers, including financial information, credit history, demographic details, and any other factors that might influence creditworthiness.
- Clean and preprocess the data to handle missing values, outliers, and inconsistencies.

2. Feature Engineering:

- Create new features or transform existing ones to improve model performance. This might involve creating ratios, calculating time-based metrics, or encoding categorical variables.

3. Model Selection and Training:

- Choose appropriate machine learning algorithms for credit risk modeling, such as logistic regression, decision trees, random forests, or gradient boosting machines.
- Split the data into training and testing sets.
- Train the models on the training data and evaluate their performance using metrics like accuracy, precision, recall, F1-score, and AUC-ROC.

4. Model Evaluation and Validation:

- Assess the model's performance on the testing set to ensure it generalizes well to unseen data.
- Consider techniques like cross-validation to improve model evaluation.

5. Credit Scorecard Development:

- Once a suitable model is selected, develop a credit scorecard that assigns a numerical score to each borrower based on their predicted probability of default.
- The scorecard should be calibrated to ensure that it aligns with the lender's risk appetite and regulatory requirements.

6. Implementation and Monitoring:

- Integrate the credit scoring model into the lender's decision-making process.
- Continuously monitor the model's performance and update it as needed to adapt to changing conditions and new data.

Key Considerations:

- **Regulatory Compliance:** Ensure that the credit risk model and scorecard comply with relevant regulations and industry standards.
- **Fair Lending:** Avoid discrimination based on protected characteristics when developing and using credit risk models.
- **Model Interpretability:** Consider using interpretable models or techniques to understand the factors driving credit risk predictions.

Literature review

The paper "RFMS Method for Credit Scoring Based on Bank Card Transaction Data" by Danyang Huang, Jing Zhou, and Hansheng Wang presents a new credit scoring model using the Recency, Frequency, Monetary (RFM) model. This model is based on bank card transaction data and aims to improve the accuracy of credit risk assessment.

The authors argue that traditional credit scoring models often rely heavily on demographic and financial information, which may not capture the dynamic nature of customer behavior. The RFM model, on the other hand, focuses on recent customer activity and spending patterns, which can provide valuable insights into creditworthiness.

The paper proposes a three-step approach to construct the RFM model:

1. **Data Preprocessing:** The authors preprocess the transaction data to handle missing values and outliers.
2. **RFM Score Calculation:** They calculate the recency, frequency, and monetary scores for each customer based on their transaction history.
3. **Credit Scoring Model Development:** They develop a credit scoring model using logistic regression, incorporating the RFM scores as predictor variables.

The paper evaluates the performance of the RFM model using a real-world dataset of bank card transactions. They compare the performance of the RFM model to a traditional credit scoring model based on demographic and financial information. The results show that the RFM model outperforms the traditional model in terms of accuracy, sensitivity, and specificity.

The authors conclude that the RFM model is a promising approach for credit scoring, particularly when dealing with dynamic customer behaviour and limited availability of traditional credit information. They suggest that future research could explore other machine learning techniques for credit scoring and investigate the impact of different RFM score calculation methods.

Overall, the paper provides a valuable contribution to the literature on credit scoring. It demonstrates the effectiveness of the RFM model in improving the accuracy of credit risk assessment based on bank card transaction data.

Task 2 - Exploratory Data Analysis (EDA)

Data Overview:

- I accessed the Xente Challenge dataset containing customer transaction information.
- Initial analysis reveals the following:
    - **Number of Rows:** 95,662
    - **Number of Columns:** 16
    - **Data Types:** A mix of object (categorical) and numerical types.

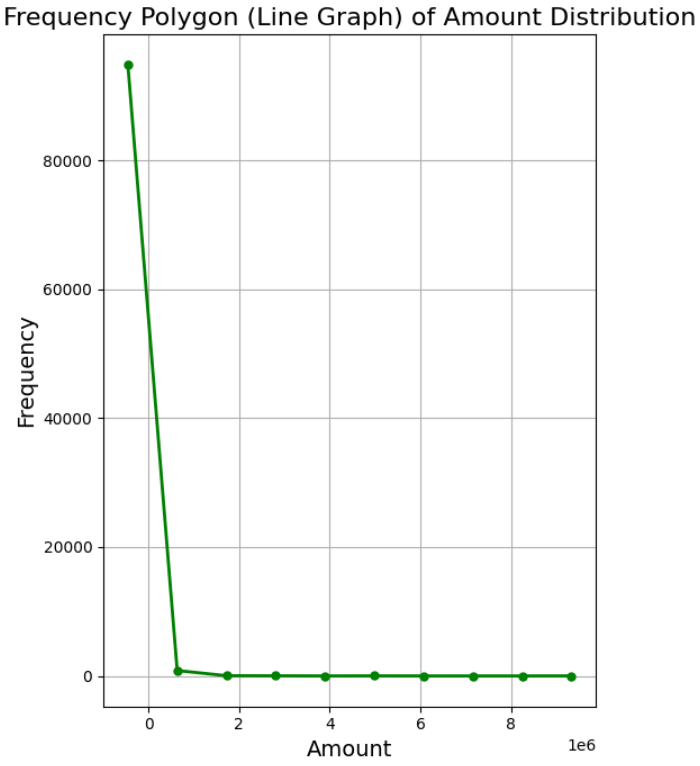
Key Features:

- **Transaction-level data:** Detailed information about individual transactions.
- **Customer information:** Details about customers involved in transactions.
- **Product information:** Details about the products purchased.
- **Transaction metadata:** Information about the context of the transaction (e.g., time, channel, pricing strategy).

Data Summary

- Both **Amount** and **Value** have a median of 1000, indicating that 50% of the transactions are below this amount.
- The mode for both columns is also 1000, suggesting that this is the most frequent transaction value.
- The standard deviation and variance for both columns are relatively high, indicating a wide spread of values.
- The skewness values for both columns are greater than 1, suggesting that the distributions are skewed to the right, with a longer tail on the right side.
- The kurtosis values for both columns are extremely high, indicating a heavy-tailed distribution with a significant number of outliers.
- These observations suggest that the transaction data for both Amount and Value are highly skewed and concentrated around smaller values, with a few large transactions (outliers) significantly affecting the overall distribution.
- Further analysis is needed to understand the underlying reasons for this skewed distribution and to explore potential implications for credit risk assessment and other modelling tasks.

Frequency Distribution



The frequency polygon provides a visual representation of the distribution of transaction amounts in the dataset.

Key Observations:

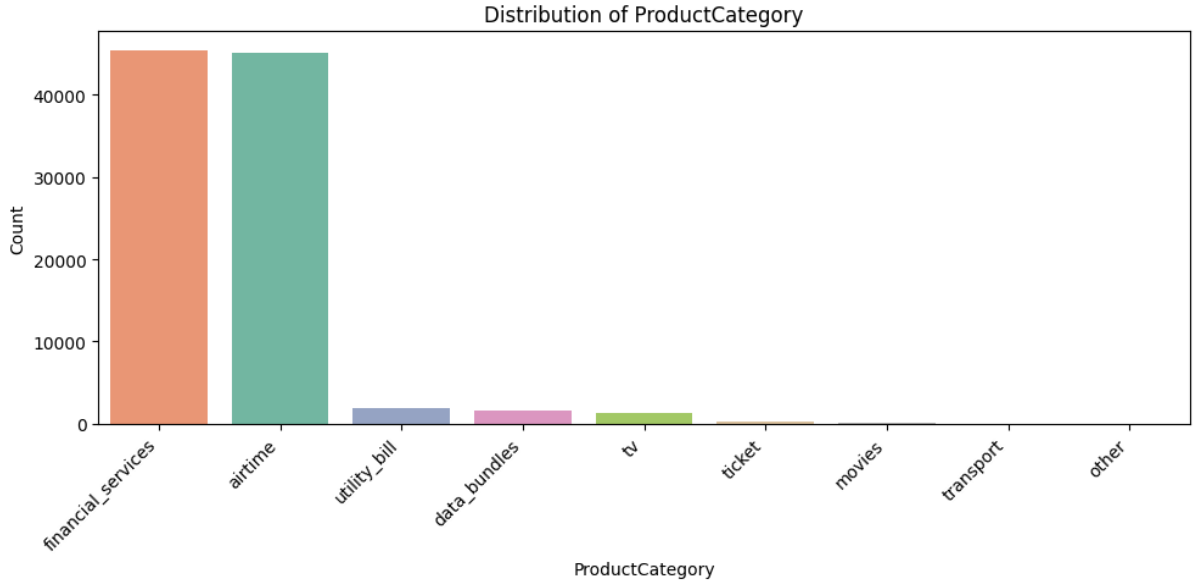
- **Right-Skewed Distribution:** The graph confirms the right-skewed nature of the `Amount` distribution, as observed in the summary statistics. This indicates that there are a few large transactions (outliers) that pull the mean to the right, while the majority of transactions are concentrated in the lower range.
- **Concentration at Lower Amounts:** The frequency polygon shows a sharp peak at lower amounts, suggesting that most transactions are relatively small.
- **Long Tail:** The presence of a long tail on the right side indicates the existence of infrequent but significantly large transactions.
- **Decreasing Frequency:** As the transaction amount increases, the frequency of transactions decreases, confirming the concentration at lower amounts.

These observations align with the findings from the summary statistics and further emphasize the skewed nature of the transaction data.

Potential Implications:

- **Outliers:** The presence of outliers (large transactions) might impact the overall analysis and modelling. It might be necessary to consider techniques like trimming to handle outliers appropriately.

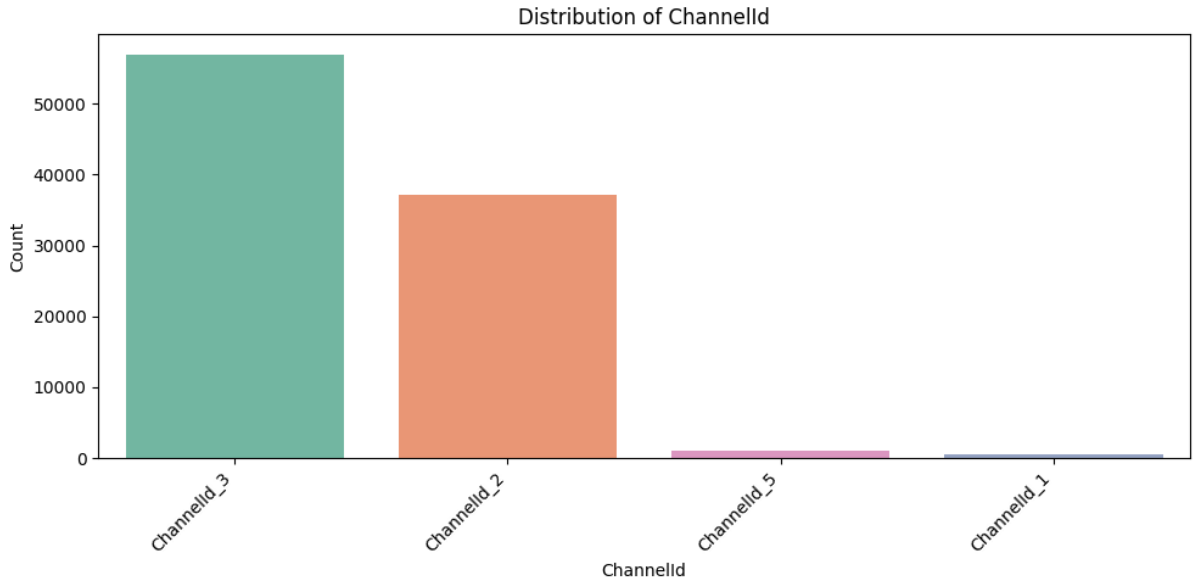
- **Credit Risk:** The concentration of transactions at lower amounts might suggest a relatively low-risk customer base. However, the presence of outliers could indicate potential risks associated with large transactions.
- **Customer Segmentation:** The skewed distribution might suggest the need for different customer segments based on transaction size.
- **Model Development:** The choice of modelling techniques should take into account the skewed distribution and potential outliers. Robust algorithms that can handle non-normal distributions might be more suitable.



The bar plot provides a visual representation of the distribution of product categories in the dataset.

Key Observations:

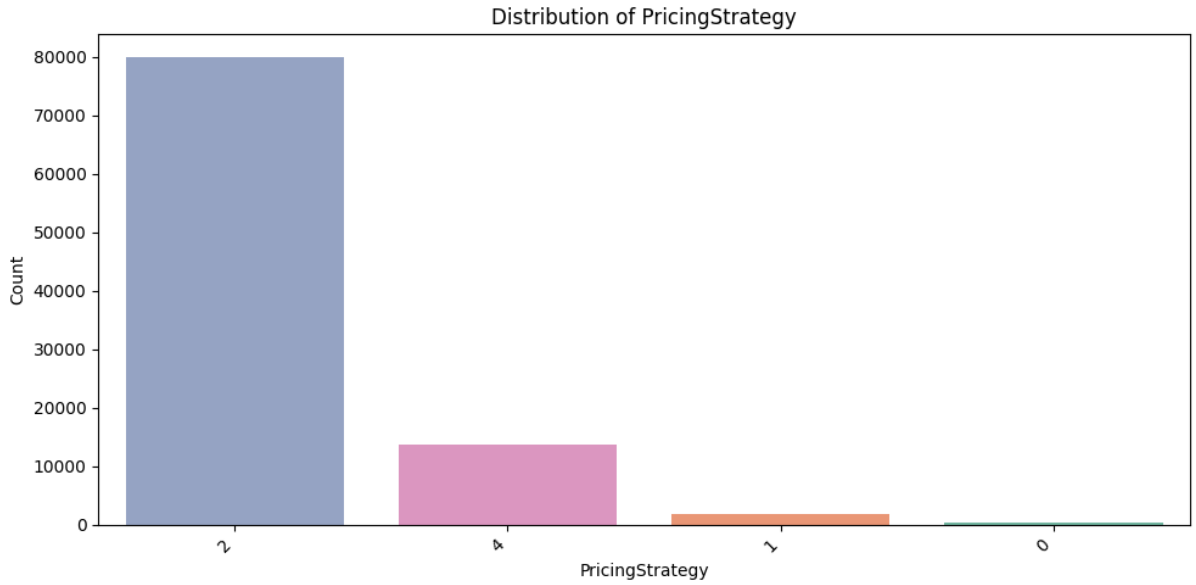
- **Dominant Categories:** The categories "financial services" and "airtime" have the highest counts, indicating that they are the most frequently purchased product categories.
- **Product Focus:** The dominance of "financial services" and "airtime" suggests that these product categories might be the primary focus of the business.
- **Customer Segmentation:** The skewed distribution might indicate different customer segments based on product preferences.
- **Risk Assessment:** The distribution of product categories might have implications for credit risk assessment, as different product categories could have varying levels of risk.



The bar plot provides a visual representation of the distribution of transaction channels in the dataset.

Key Observations:

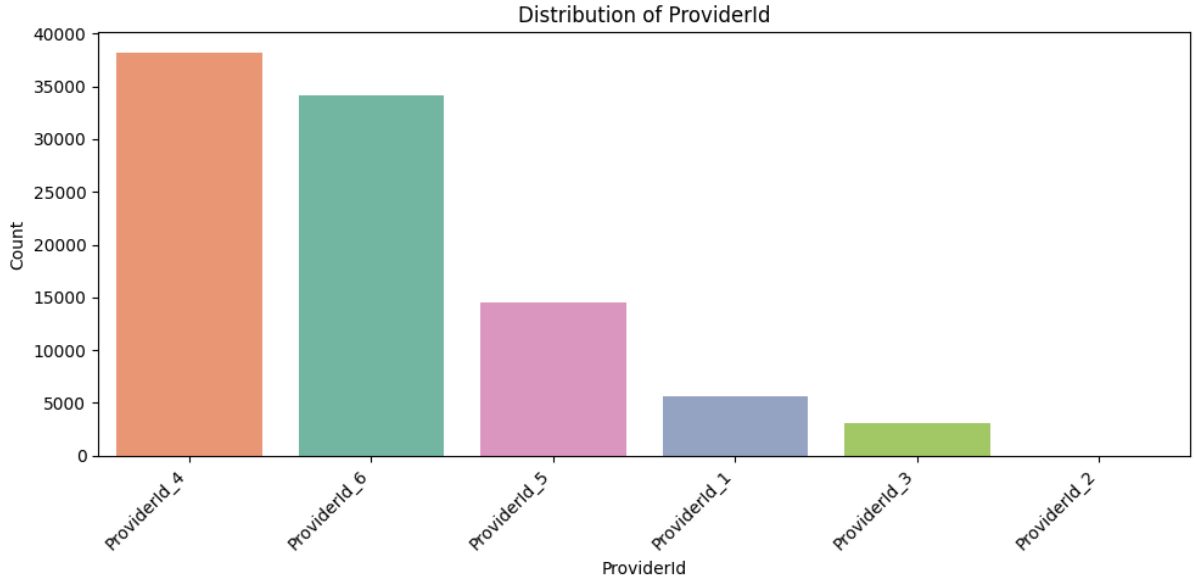
- **Dominant Channel:** "Channelld_3" is the most frequently used channel, accounting for a significant portion of the transactions.
- **Skewed Distribution:** The distribution is skewed to the right, with "Channelld_3" being the dominant channel.
- **Channel Focus:** The dominance of "Channelld_3" suggests that this channel might be the primary focus of the business or the most effective channel for acquiring customers.
- **Customer Segmentation:** The skewed distribution might indicate different customer segments based on channel preferences.
- **Risk Assessment:** The distribution of channels might have implications for credit risk assessment, as different channels could have varying levels of risk.



The bar plot provides a visual representation of the distribution of pricing strategies used in the dataset.

Key Observations:

- **Dominant Pricing Strategy:** The pricing strategy labeled "2" is the most frequently used, accounting for a significant portion of the transactions.
- **Pricing Focus:** The dominance of pricing strategy "2" suggests that this strategy might be the primary pricing approach used by the business.
- **Customer Segmentation:** The skewed distribution might indicate different customer segments based on pricing preferences.
- **Risk Assessment:** The distribution of pricing strategies might have implications for credit risk assessment, as different pricing strategies could be associated with varying levels of risk.

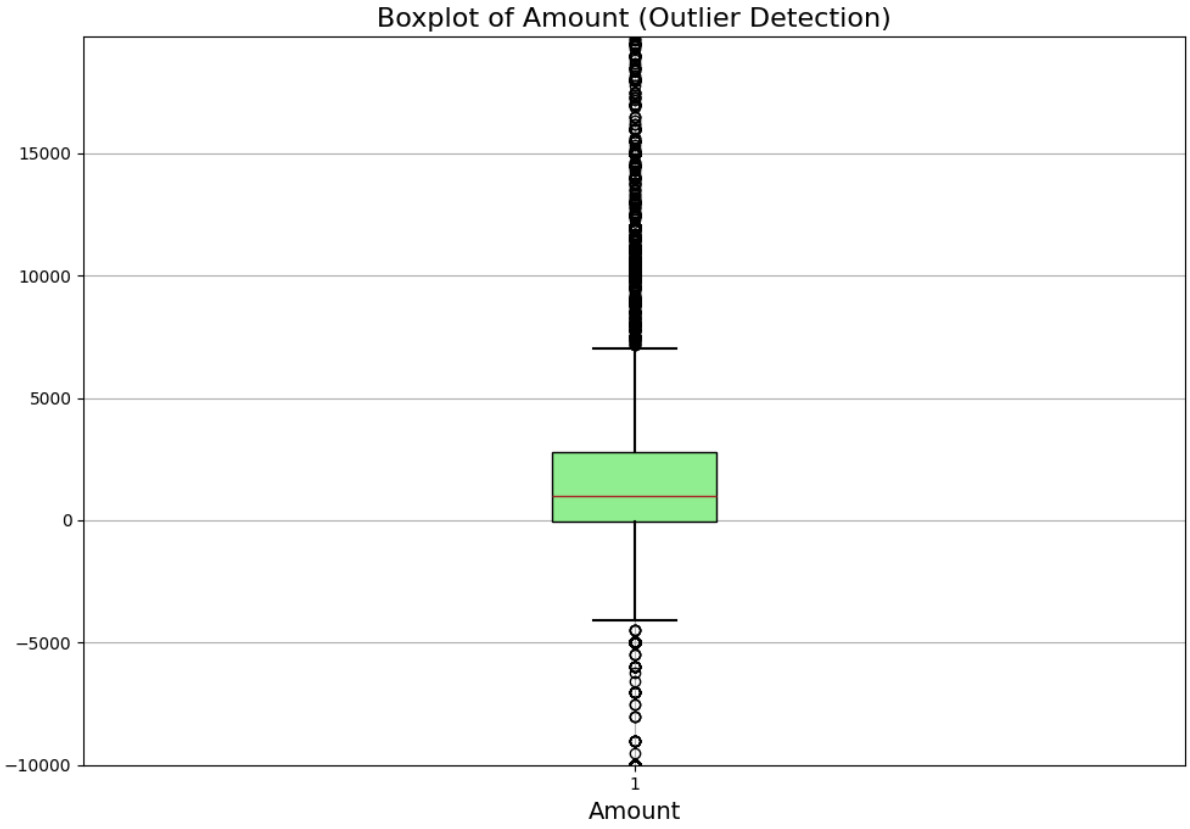


The bar plot provides a visual representation of the distribution of providers in the dataset.

Key Observations:

- **Dominant Provider:** "ProviderId_4" is the most frequently used provider, accounting for a significant portion of the transactions.
- **Provider Focus:** The dominance of "ProviderId_4" suggests that this provider might be the primary partner for the business or the most effective provider in terms of product availability or pricing.
- **Customer Segmentation:** The skewed distribution might indicate different customer segments based on provider preferences.

Outlier Detection



The boxplot provides a visual representation of the distribution of transaction amounts and helps identify potential outliers.

Key Observations:

- **Median:** The median transaction amount is around 2,000, indicating that 50% of the transactions are below this value.
- **IQR:** The interquartile range (IQR) is relatively small, suggesting that the majority of transactions are clustered within a narrow range.

- **Outliers:** The boxplot clearly shows a significant number of outliers, especially on the right side of the distribution. These outliers represent transactions with much larger amounts compared to the majority.
- **Skewness:** The presence of outliers confirms the right-skewed nature of the Amount distribution, as observed in previous analyses.

Implications:

- **Outlier Impact:** The outliers might have a significant impact on the overall analysis and modelling, as they can skew the results and make it difficult to accurately capture the underlying trends.
- **Data Cleaning:** It might be necessary to consider techniques like imputation or trimming to handle outliers appropriately and improve the data quality.
- **Model Robustness:** The chosen modelling techniques should be robust to outliers and able to handle skewed distributions.

Plan for next Task

Task 3: Feature Engineering

- **Complete the analysis mentioned in the interim report:** Analyze the distribution of numerical features, categorical features, and correlations.
- **Identify outliers and determine appropriate handling strategies:** Consider techniques like winsorization or trimming to address outliers.
- **Prepare the data for feature engineering:** Encode categorical variables, handle missing values, and normalize/standardize numerical features as needed.
- **Create aggregate features:** Calculate summary statistics for each customer (e.g., total transaction amount, average transaction amount).
- **Extract features:** Create new features from existing ones (e.g., transaction hour, day of week).
- **Feature selection:** Use techniques like correlation analysis or feature importance to select the most relevant features for modeling.

Task 4: Modeling

- **Split the data:** Divide the data into training and testing sets.
- **Choose and train models:** Select appropriate machine learning algorithms (e.g., logistic regression, random forest) and train them on the training data.
- **Hyperparameter tuning:** Fine-tune model parameters to improve performance.
- **Model evaluation:** Assess model performance using metrics like accuracy, precision, recall, F1-score, and AUC-ROC.
- **Compare models:** Compare the performance of different models to select the best one.

Task 5: Model Serving API Call

- **Choose a framework:** Select a suitable framework for building REST APIs (e.g., Flask, FastAPI).
- **Load the model:** Load the trained machine learning model.
- **Define API endpoints:** Create endpoints that accept input data and return predictions.
- **Handle requests:** Implement logic to receive input data, preprocess it, and make predictions using the model.
- **Return predictions:** Format the predictions and return them as a response.
- **Deployment:** Deploy the API to a web server or cloud platform.

Better to iterate on these steps as needed, refining the feature engineering process, experimenting with different models, and optimizing the API for performance.

References

[1]. https://github.com/wolderufael/Credit-Scoring-Model-for-eCommerce-Platform

[2].https://www3.stat.sinica.edu.tw/statistica/oldpdf/A28n535.pdf

[3].https://www.hkma.gov.hk/media/eng/doc/key-functions/financial-infrastructure/alternative_credit_scoring.pdf

[4].https://thedocs.worldbank.org/en/doc/935891585869698451-0130022020/original/CREDITSCORINGAPPROACHESGUIDELINESFINALWEB.pdf

[5].https://towardsdatascience.com/how-to-develop-a-credit-risk-model-and-scorecard-91335fc01f03

[6].https://corporatefinanceinstitute.com/resources/commercial-lending/credit-risk/

[7].https://www.risk-officer.com/Credit_Risk.htm