**Project Description**

**Overview**

Abay Bank S.C., inspired by the Great Abay River's strength, aims to promote growth by financing various sectors, generating employment, and fostering a stable financial system. This project focuses on optimizing credit risk assessment at Abay Bank using data analytics to develop a predictive model that identifies high-risk borrowers. The goal is to reduce loan defaults, improve the loan portfolio, and enhance the bank's financial stability and profitability.

**Organization**

Abay Bank S.C., established on July 14, 2010, and fully operational since November 3, 2010, is a key player in the Ethiopian banking sector. As of December 31, 2023, the bank has a paid-up capital of Birr 5.171 billion, 4,479 shareholders, and operates over 528 branches with more than 3 million account holders. The bank serves various sectors, including trade, agriculture, industry, and real estate.

**Business Problem**

Despite growth and technological advancements, Abay Bank struggles with accurately assessing credit risk, leading to higher loan defaults and financial losses. The current system cannot effectively leverage historical data to predict and mitigate risks, resulting in loans granted to high-risk borrowers.

**Objectives**

* **Develop a Predictive Model**: Create a machine learning model to predict loan default likelihood.
* **Reduce Loan Defaults**: Implement the model to minimize high-risk loans and improve loan portfolio quality.
* **Enhance Decision-Making**: Provide insights to the credit risk management team for informed lending decisions.
* **Improve Efficiency**: Streamline the credit assessment process, reducing evaluation time and resources.
* **Financial Stability**: Contribute to the bank's overall financial stability and profitability.

The provided dataset includes loan details, borrower demographics, repayment history, and account transactions, essential for building the predictive model. The project will involve data cleaning, exploratory data analysis (EDA), feature engineering, model development, and evaluation. The ultimate aim is to integrate the predictive model into Abay Bank's systems to automate credit risk assessment and enhance financial outcomes.

**Business Understanding**

**Organization and Industry**

Abay Bank S.C., established on July 14, 2010, is a leading financial institution in Ethiopia. Named after the Great Abay River, the bank symbolizes strength and growth potential. Since starting operations on November 3, 2010, Abay Bank has significantly contributed to the Ethiopian banking sector by financing various economic sectors. With a paid-up capital of Birr 5.171 billion and 4,479 shareholders as of December 31, 2023, the bank serves over 3 million account holders through 528 branches, offering services across trade, agriculture, industry, and real estate.

**Relevant Business Context**

The Ethiopian banking industry is rapidly evolving with technological advancements and increased regulatory oversight. Financial institutions, including Abay Bank, are adopting data analytics and machine learning to enhance efficiency, reduce risks, and provide personalized banking solutions. Despite this progress, Abay Bank struggles with accurate credit risk assessment, leading to higher loan defaults and financial losses.

**Key Stakeholders and Business Problem**

**Key Stakeholders**

1. **Credit Risk Management Team**: Uses the predictive model for informed lending decisions.
2. **Loan Officers**: Benefits from model insights to identify high-risk borrowers during the loan approval process.
3. **Senior Management**: Uses model results to enhance the bank's financial stability.
4. **IT Department**: Ensures seamless integration of the predictive model into the bank’s systems.
5. **Regulatory Bodies**: Ensures compliance with financial regulations, assisted by improved risk assessment processes.

**Business Problem**

Abay Bank's current credit risk assessment system fails to leverage historical data effectively, resulting in high loan defaults and financial losses. The inability to predict and mitigate risks accurately leads to granting loans to high-risk borrowers, adversely affecting the bank's profitability and increasing the rate of non-performing loans.

**Project Goals**

* **Develop a Predictive Model**: Create a robust machine learning model that accurately predicts the likelihood of loan defaults.
* **Reduce Loan Defaults**: Implement the predictive model to decrease the number of high-risk loans and improve the overall quality of the loan portfolio.
* **Enhance Decision-Making**: Provide actionable insights to the credit risk management team to make informed lending decisions.
* **Improve Efficiency**: Streamline the credit assessment process, reducing the time and resources required for loan evaluations.
* **Financial Stability**: Contribute to the overall financial stability and profitability of Abay Bank S.C.

**Expected KPIs or Metrics for Success**

* **Reduction in Loan Default Rate**: Measure the decrease in the percentage of defaulted loans after implementing the predictive model.
* **Model Accuracy**: Evaluate the predictive model's accuracy using metrics such as precision, recall
* **Operational Efficiency**: Assess the time and cost savings achieved in the loan assessment process.
* **Stakeholder Satisfaction**: Gather feedback from the credit risk management team and loan officers on the usability and effectiveness of the predictive model.
* **Financial Impact**: Analyze the improvement in the bank's profitability and reduction in non-performing loans.

**Data Acquisition**

**Identifying Relevant Data Sources**

To develop a robust credit risk assessment model, it's essential to gather comprehensive data from Abay Bank’s core banking system. This ensures the model has access to a diverse and detailed dataset to accurately predict loan defaults.

**Internal Data Sources**

All necessary data for this project will be sourced from Abay Bank's core banking system, which consolidates various types of information critical for credit risk assessment. Key data types include:

1. **Loan Data**:
   * Details such as account numbers, customer IDs, product codes, amounts financed, book dates, maturity dates, user-defined statuses, status descriptions, and component names.
2. **Borrower Demographics**:
   * Information including customer IDs, telephone numbers, mobile numbers, and addresses.
3. **Repayment History**:
   * Records of due dates, paid dates, amounts paid, total payment delays, and payment statuses.
4. **Additional Metrics**:
   * Calculated metrics such as loan tenure, total interest paid, penalty counts, and average principal to interest ratio.

**Process of Obtaining and Accessing the Data**

**Internal Data Collection**

As IT staff of Abay Bank, we have direct access to the core banking system, allowing us to extract the necessary data efficiently. The data collection process involves:

1. **Data Extraction**:
   * Use the bank's database querying tools to retrieve specific datasets from the core banking system. This involves running queries to collect detailed information on loans, borrower demographics, repayment history, and account transactions.
2. **Data Export**:
   * Export the extracted data into manageable formats such as CSV or Excel for initial analysis and preprocessing. This step ensures the data is in a suitable format for further processing and model development.

**External Data Collection** Currently, external data collection is not feasible for this project. However, if needed in the future, integration with external data sources such as credit bureaus and national statistics agencies could be considered to enhance the model's predictive capabilities.

By following these steps, Abay Bank S.C. ensures the availability of high-quality data necessary for building a predictive model that optimizes credit risk assessment. This comprehensive dataset will enable the bank to make informed decisions, reduce loan defaults, and enhance overall financial performance.

**Data Cleaning and Preprocessing**

**Data Cleaning Steps:**

**1. Handling Duplicates:**

* **Identify and Remove Exact Duplicates**:
  + Use the pandas function drop\_duplicates() to remove any exact duplicate rows from the dataset. This ensures that each transaction in the dataset is unique.
* **Handle Duplicate ACCOUNT\_NUMBER and CUSTOMER\_ID Appropriately**:
  + Since ACCOUNT\_NUMBER and CUSTOMER\_ID can be duplicated due to different repayment instances, do not remove these duplicates. Instead, handle them appropriately during aggregation steps to ensure accurate calculations.

**2. Missing Values:**

* **Check for Missing Values**:
  + Use the pandas function isnull().sum() to check for missing values in each column. This helps identify columns with missing data.
* **Ensure No Missing Values in Critical Columns**:
  + For critical columns like ACCOUNT\_NUMBER, CUSTOMER\_ID, AMOUNT\_FINANCED, and AMOUNT\_PAID, ensure there are no missing values. If missing values are found, decide on appropriate imputation strategies (e.g., filling with mean/median values) or discard those rows.

**3. Data Types:**

* **Convert Date Columns to Datetime Format**:
  + Convert the date columns (BOOK\_DATE, MATURITY\_DATE, DUE\_DATE, PAID\_DATE) to datetime format using the pandas to\_datetime() function for accurate date calculations.
* **Ensure Numeric Columns are in Appropriate Formats**:
  + Ensure that numeric columns like AMOUNT\_FINANCED and AMOUNT\_PAID are in appropriate numeric formats for further calculations.

**4. Correcting Inconsistent Entries:**

* **Standardize Categorical Values**:
  + Ensure consistent naming in categorical columns such as STATUSDESCRIPTION and USER\_DEFINED\_STATUS. This can be done by converting the text to lowercase or using a predefined mapping for standardization.

**Preprocessing Steps:**

1. **Feature Engineering:**
   * **Loan Tenure**: Calculate as the difference between MATURITY\_DATE and BOOK\_DATE.
   * **Payment Delays**: Calculate as the difference between PAID\_DATE and DUE\_DATE.
   * **Defaulted Loans**: Identify based on STATUSDESCRIPTION and payment delays.
2. **Aggregations**:
   * **Total Amount Financed**: Sum AMOUNT\_FINANCED for each CUSTOMER\_ID and ACCOUNT\_NUMBER.
   * **Total Amount Repaid**: Sum AMOUNT\_PAID for each CUSTOMER\_ID and ACCOUNT\_NUMBER.
   * **Number of Loans**: Count unique ACCOUNT\_NUMBER for each CUSTOMER\_ID.
   * **Total Payment Delays**: Sum delays in payments for each customer.
   * **Total Interest Paid**: Sum interest payments.
   * **Penalty Count**: Count penalties for each customer.

**Final Dataset Overview**

**Customer Loan Summary**:

* **CUSTOMER\_ID**: Unique identifier.
* **Total\_Amount\_Financed**: Total amount financed.
* **Total\_Amount\_Repaid**: Total amount repaid.
* **Number\_of\_Loans**: Number of unique loans.
* **Average\_Loan\_Tenure**: Average tenure of loans.
* **Total\_Payment\_Delays**: Total days of delayed payments.
* **Total\_Interest\_Paid**: Total interest paid.
* **Penalty\_Count**: Total penalties incurred.
* **Defaulted\_Loans**: Indicator of loan default status.

**Loan Repayment Details**:

* **CUSTOMER\_ID**: Unique identifier.
* **ACCOUNT\_NUMBER**: Loan account number.
* **PRODUCT\_CODE**: Loan product code.
* **STATUSDESCRIPTION**: Loan status description.
* **COMPONENT\_NAME**: Type of repayment (e.g., interest, principal).
* **Total\_Amount\_Paid**: Total amount paid per repayment type.
* **Due\_Date**: Repayment due date.
* **Paid\_Date**: Actual repayment date.

These steps ensure a clean and consistent dataset ready for effective loan risk assessment and predictive modeling.

**Exploratory Data Analysis**

**Highlight the Key Insights and Observations from the Data Analysis**

1. **Loan Amount Distribution:**
   * The dataset includes a wide range of loan amounts, from small loans to very large loans. This diversity indicates that the bank caters to a wide variety of financial needs.
   * Large loans, such as those exceeding millions, represent significant financial exposure and potentially higher risk.
2. **Repayment Behavior:**
   * The total amount repaid varies significantly across loans, with some loans showing full repayment while others have no repayment recorded. This variation can help identify high-risk customers and repayment patterns.
   * Loans with zero repayment might indicate non-performing loans or loans that have just been disbursed.
3. **Loan Tenure** 
   * The average loan tenure varies from very short durations (10 days) to multiple years, reflecting the variety of loan products offered by the bank.
4. **Payment Delays and Defaults:**
   * Significant payment delays are observed in the dataset. Identifying the reasons behind such delays can help improve risk assessment models.
   * The presence of multiple defaulted loans highlights the need for stricter credit evaluation processes.
5. **Interest vs. Principal Repayment:**
   * The average principal interest ratio shows substantial variation, with some loans having high ratios indicating more principal repayment compared to interest. This metric can help evaluate the repayment behavior and loan structuring.
6. **Product Code Analysis:**
   * The dominance of certain product codes indicates their popularity among customers. Analyzing product-specific trends and risks is crucial for the bank.

**Identify the Important Variables Relevant to the Business Problem**

**Important Variables Relevant to the Business Problem**

1. **CUSTOMER\_ID**: Unique identifier for tracking customer-specific data.
2. **ACCOUNT\_NUMBER**: Loan account number to distinguish different loans taken by the same customer.
3. **PRODUCT\_CODE**: Represents the type of loan product, essential for analyzing trends and risks.
4. **Total\_Amount\_Financed**: Key for understanding financial exposure and loan sizes.
5. **Total\_Amount\_Repaid**: Crucial for assessing repayment behavior and identifying potential defaults.
6. **Number\_of\_Loans**: Important for understanding customer engagement and loan portfolio size.
7. **Average\_Loan\_Tenure**: Relevant for analyzing loan durations and terms.
8. **Total\_Payment\_Delays**: Critical for assessing payment behavior and risk.
9. **PENALTY\_COUNT**: Indicates the frequency of penalties, relevant for understanding repayment issues.
10. **STATUSDESCRIPTION**: Provides context on loan status, crucial for risk assessment.
11. **COMPONENT\_NAME**: Specifies the component of repayment (e.g., interest, principal), useful for detailed repayment analysis.

**Modeling**

**Overview**

This section outlines the machine learning approach used for optimizing credit risk assessment at Abay Bank. It covers the models considered, the training and evaluation process, and the performance metrics of the selected model.

**Machine Learning Algorithms Considered**

To enhance the credit risk assessment process, various machine learning algorithms were evaluated:

1. **Logistic Regression**: A linear model used for binary classification that estimates the probability that a given input point belongs to a certain class.
2. **Decision Trees**: A non-linear model that splits the data into subsets based on feature values to make predictions.
3. **Random Forest**: An ensemble learning method that combines multiple decision trees to improve prediction accuracy and control over-fitting.

**Model Training and Evaluation Process**

The data was preprocessed to ensure consistency and reliability before training the models. Here is an overview of the steps taken:

1. **Data Preprocessing**:
   * Missing values were handled using SimpleImputer from sklearn. For numerical features, the median value was used, and for categorical features, the most frequent value was applied.
   * Numerical features were scaled using StandardScaler, and categorical features were one-hot encoded using OneHotEncoder.
2. **Model Training**:
   * The Random Forest Classifier was selected as the final model due to its robust performance and ability to handle a mix of numerical and categorical features.
3. **Model Evaluation**:
   * Performance was evaluated using several metrics: Accuracy, Precision, Recall, and F1-Score. These metrics provide a comprehensive understanding of the model's performance, particularly its ability to correctly identify both loan defaults (positive class) and non-defaults (negative class).
4. **Feature Importance**:
   * The importance of each feature was assessed to understand which variables were most influential in predicting loan defaults.

**Performance Metrics and Selected Model**

The Random Forest Classifier emerged as the best-performing model based on the evaluation metrics:

* **Accuracy**: 73.5% - This indicates that the model correctly classified 73.5% of the instances in the test set.
* **Precision**: 35.5% - This score indicates that when the model predicts a loan will default, it is correct 35.5% of the time.
* **Recall**: 37.3% - The model correctly identifies 37.3% of actual loan defaults.
* **F1 Score**: 17.5% - This balanced measure of precision and recall demonstrates that the model has a reasonable balance between correctly identifying defaults and not misclassifying non-defaults.

The Random Forest model's robustness and interpretability, combined with its ability to handle different types of data and provide feature importance, made it the preferred choice for credit risk assessment.

**Key Features**

The following features were identified as the most significant in predicting loan defaults:

1. **Total Payment Delays**: Indicates the cumulative number of days payments were delayed.
2. **PENALTY\_COUNT**: The number of penalties incurred, reflecting the borrower’s repayment issues.
3. **STATUSDESCRIPTION**: Categorizes the loan status (e.g., Loss, Substandard, Doubtful, Special Mention, Normal).
4. **COMPONENT\_NAME**: Specifies the component of repayment (e.g., interest, principal).

These features provide valuable insights into the borrower’s financial behavior and risk profile, enabling the bank to make more informed credit decisions.

**Communication of Results**

**Key Findings and Implications**

**Key Findings:**

1. **Model Performance**:
   * The Random Forest model achieved a precision of 35.5%, recall of 37.3%, and an F1 score of 17.5%, indicating a reasonable balance between correctly identifying loan defaults and not misclassifying non-defaults.
2. **Feature Importance**:
   * Significant features identified include:
     + **Total Payment Delays**: Reflects the cumulative number of days payments were delayed.
     + **PENALTY\_COUNT**: Indicates the number of penalties incurred, highlighting repayment issues.
     + **STATUSDESCRIPTION**: Categorizes the loan status (e.g., Loss, Substandard, Doubtful, Special Mention, and Normal).
     + **COMPONENT\_NAME**: Specifies the component of repayment (e.g., interest, principal).
   * These features provide valuable insights into borrower risk profiles, enabling more informed lending decisions.

**Implications for Business:**

1. **Enhanced Risk Assessment**:
   * Accurate predictions facilitate better lending decisions, reducing the number of non-performing loans and enhancing financial stability.
2. **Tailored Products**:
   * Insights from significant features allow the bank to create customized loan products and repayment plans tailored to different risk segments.
3. **Efficient Resource Allocation**:
   * Resources can be focused on monitoring high-risk loans while automating the approval process for low-risk loans.

**Limitations and Recommendations**

**Limitations:**

1. **Data Quality**:
   * The accuracy of the model is heavily dependent on the quality and completeness of the data. Inconsistent or missing data can adversely affect predictions.
2. **Model Generalization**:
   * The model’s performance may vary with new data, particularly if the new data differs significantly from the training data.
3. **Feature Dependency**:
   * The significance of features may change over time, necessitating regular updates to the model.

**Recommendations:**

1. **Data Enhancement**:
   * Improve data collection processes and integrate additional data sources, such as credit scores and employment history, to enrich the dataset.
2. **Regular Updates**:
   * Regularly evaluate and retrain the model to ensure it remains accurate and relevant.
3. **Feedback Loop**:
   * Establish a feedback loop with loan officers and risk managers to align model predictions with real-world observations and make necessary adjustments.
4. **Risk Mitigation**:
   * Develop strategies for managing high-risk loans, such as proactive customer engagement and restructuring repayment plans to mitigate potential defaults.
5. **Scalability and Integration**:
   * Integrate the model into existing systems to facilitate automated decision-making and scalability.

By addressing these limitations and implementing the recommendations, Abay Bank can enhance its credit risk assessment process, leading to improved decision-making and better financial performance.

**Conclusion**

**Summary of the Project**

This project aimed to optimize credit risk assessment at Abay Bank using machine learning techniques. The Random Forest model was identified as the most effective approach, achieving a precision of 35.5%, recall of 37.3%, and an F1 score of 17.5%. Key features influencing loan default predictions were identified, providing valuable insights into borrower risk profiles. The implementation of this model promises to enhance lending decisions, reduce non-performing loans, and improve resource allocation, ultimately strengthening the bank's financial stability.

**Value Delivered**

* **Enhanced Risk Assessment**: Accurate predictions of loan defaults lead to more informed and reliable lending decisions.
* **Tailored Products and Services**: Insights into borrower behavior allow for the development of customized loan products and repayment plans.
* **Operational Efficiency**: Automating the credit assessment process and focusing resources on high-risk loans improve overall efficiency.

**Lessons Learned**

1. **Data Quality is Crucial**: Clean, comprehensive data is essential for model performance. Inconsistent or incomplete data significantly impacts predictions.
2. **Model Maintenance**: Regular updates and retraining are necessary to ensure the model adapts to changing conditions and maintains accuracy.
3. **Collaboration**: Engaging with loan officers and risk managers provides valuable real-world insights, enhancing the model's relevance and applicability.
4. **Scalability**: Ensuring the model can handle large datasets and integrate seamlessly with existing systems is key to long-term success.

**Overall Experience**

The data analytics life cycle, from data preprocessing to model deployment, highlighted the importance of each stage in developing a robust predictive model. The iterative process of testing and refining models underscored the need for continuous improvement and adaptation. This project not only delivered immediate value to Abay Bank but also provided a framework for future data-driven decision-making initiatives.