**Project Description**

**Overview**

Ethiopia’s mighty river, the Great Abay, is a dramatic spectacle and a symbol of natural strength and grandeur. This tremendous natural strength similarly explains why we are named Abay Bank. We are here to foster growth and development by promoting and financing different sectors, thereby generating employment opportunities and accelerating capital formation, by ensuring a safe, stable, and sound financial system. This project focuses on optimizing credit risk assessment at Abay Bank S.C., a leading financial institution in Ethiopia. Leveraging data analytics, the primary objective is to develop a predictive model that accurately identifies high-risk borrowers. This initiative aims to reduce loan defaults, enhance the quality of the bank's loan portfolio, and improve the overall financial stability and profitability of the bank.

**Organization**

Abay Bank S.C. officially established on July 14, 2010, has fulfilled all the necessary requirements of the National Bank of Ethiopia. The bank started full-fledged banking operations on November 3, 2010, and has since grown into a major player in the Ethiopian banking sector. Currently, the paid-up capital of the bank is Birr 5.171 Billion, with 4,479 shareholders as of December 31, 2023. Abay Bank is poised to serve all economic sectors through its network of branches, extending its services to domestic trade and services, international trade, agriculture, industry, transportation, construction, and real estate sectors. As of December 31, 2023, the bank operates over 528 branches with more than 3,073,758 account holders.

**Business Problem**

Despite its growth and technological advancements, Abay Bank faces challenges in accurately assessing credit risk, leading to higher loan defaults and financial losses. Inaccurate risk assessments result in granting loans to high-risk borrowers, negatively impacting the bank's profitability and increasing non-performing loans. The current system lacks the capability to leverage historical data effectively to predict and mitigate these risks.

**Objectives**

* **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.

The dataset provided includes comprehensive details of loans, borrower demographics, repayment history, and account transactions, which are essential for building a robust predictive model. The project will follow a structured approach, including data cleaning and preprocessing, exploratory data analysis (EDA), feature engineering, model development, and evaluation. The ultimate goal is to integrate the predictive model into Abay Bank's existing systems to automate the credit risk assessment process and improve financial outcomes.

**Business Understanding**

**Organization and Industry**

Abay Bank S.C. is a prominent financial institution in Ethiopia, established on July 14, 2010. Named after the Great Abay River, the bank embodies strength and potential for growth. Since commencing full-fledged operations on November 3, 2010, Abay Bank has become a cornerstone in the Ethiopian banking sector, fostering economic development by financing various 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 its network of 528 branches. Abay Bank offers a wide range of services, including domestic and international trade, agriculture, industry, transportation, construction, and real estate.

**Relevant Business Context**

The Ethiopian banking industry is undergoing rapid transformation, driven by technological advancements and increasing regulatory oversight. Financial institutions are increasingly adopting data analytics and machine learning to enhance operational efficiency, mitigate risks, and offer personalized banking solutions. Abay Bank, with its commitment to leveraging state-of-the-art banking technology, is poised to lead the way in financial empowerment and economic development. However, the bank faces challenges in accurately assessing credit risk, which impacts its profitability and increases the rate of non-performing loans.

**Key Stakeholders and Business Problem**

**Key Stakeholders**

1. **Credit Risk Management Team**: Responsible for assessing and managing credit risk. They will use the predictive model to make informed lending decisions.
2. **Loan Officers**: Involved in the loan approval process and borrower evaluation. They will benefit from the model's insights to identify high-risk borrowers.
3. **Senior Management**: Overseeing the bank's financial performance and strategic direction. They will use the model's results to improve overall financial stability.
4. **IT Department**: Supporting data extraction, processing, and model implementation. They will ensure the seamless integration of the predictive model into the bank's systems.
5. **Regulatory Bodies**: Ensuring compliance with financial regulations and guidelines. The model will help the bank adhere to regulatory requirements by improving risk assessment processes.

**Business Problem**

Abay Bank faces challenges in accurately assessing credit risk, leading to higher loan defaults and financial losses. The current system lacks the capability to leverage historical data effectively to predict and mitigate these risks. This results in granting loans to high-risk borrowers, negatively impacting the bank's profitability and increasing 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, and ROC-AUC.
* **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**

For the successful implementation of the credit risk assessment model, it is crucial to gather comprehensive data from the bank’s core banking system. This ensures that the model has access to a diverse and detailed dataset to accurately predict loan defaults.

**Internal Data Sources**

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

1. **Loan Data**:
   * Details such as account numbers, customer IDs, product codes, amounts financed, book dates, maturity dates, user-defined statuses, and status descriptions.
2. **Borrower Demographics**:
   * Information including customer numbers, dates of birth, gender, national IDs, passport numbers, telephone numbers, mobile numbers, email addresses, and addresses.
3. **Repayment History**:
   * Records of due dates, paid dates, amounts paid, and payment statuses.
4. **Account Transactions**:
   * Transaction details including transaction reference numbers, account numbers, transaction codes, transaction amounts, transaction dates, and value dates.

**External Data Sources**

For this project, we are currently unable to obtain data from external sources such as the National Bank of Ethiopia. Therefore, we will rely solely on the internal data sources available within the bank’s core banking system.

**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. Here’s how the data collection process is handled:

1. **Database Queries**:
   * Utilize SQL queries to retrieve specific datasets from the core banking system. Example queries include:

-- Query to extract loan data

SELECT

ACCOUNT\_NUMBER, CUSTOMER\_ID, PRODUCT\_CODE, AMOUNT\_FINANCED,

BOOK\_DATE, MATURITY\_DATE, USER\_DEFINED\_STATUS, STATUSDESCRIPTION

FROM

CLTB\_ACCOUNT\_MASTER;

-- Query to extract borrower demographics

SELECT

CUSTOMER\_NO, DATE\_OF\_BIRTH, SEX, P\_NATIONAL\_ID, PASSPORT\_NO,

TELEPHONE, MOBILE\_NUMBER, E\_MAIL, D\_ADDRESS1, P\_ADDRESS1, P\_ADDRESS3

FROM

STTM\_CUSTOMER

JOIN

STTM\_CUST\_PERSONAL ON STTM\_CUSTOMER.CUSTOMER\_NO = STTM\_CUST\_PERSONAL.CUSTOMER\_NO;

-- Query to extract repayment history

SELECT

ACCOUNT\_NUMBER, DUE\_DATE, PAID\_DATE, AMOUNT\_PAID, PAID\_STATUS

FROM

CLTB\_REPAYMENT;

-- Query to extract account transactions

SELECT

TRN\_REF\_NO, AC\_NO, TRN\_CODE, FCY\_AMOUNT, TRN\_DT, VALUE\_DT

FROM

CLTB\_ACCOUNT\_TRANSACTIONS;

1. **Data Export**:

* Export the extracted data into CSV or Excel formats for initial analysis and preprocessing. This step ensures the data is in a manageable format for further processing.

**External Data Collection**

1. **API Integration**:
   * Integrate APIs from credit bureaus to fetch credit scores and detailed credit reports. Use secure connections to ensure data integrity.
   * Integrate APIs from national statistics agencies and the Central Bank of Ethiopia to automatically fetch the latest economic indicators.
2. **Manual Data Collection**:
   * In cases where API integration is not feasible, manually download reports from external sources and incorporate them into the dataset.

By following these steps, Abay Bank S.C. can ensure 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, DATE\_OF\_BIRTH) 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:**

* **Calculate Loan Tenure**:
  + Calculate the tenure of each loan as the difference between MATURITY\_DATE and BOOK\_DATE.
* **Calculate Age at Loan Booking**:
  + Calculate the age of customers at the time of loan booking from DATE\_OF\_BIRTH and BOOK\_DATE.

**2. Aggregations:**

* **Total Amount Financed**:
  + Sum AMOUNT\_FINANCED for unique combinations of CUSTOMER\_ID and ACCOUNT\_NUMBER.
* **Total Repayment**:
  + Sum AMOUNT\_PAID for unique combinations of CUSTOMER\_ID and ACCOUNT\_NUMBER.
* **Number of Loans**:
  + Count unique ACCOUNT\_NUMBER for each CUSTOMER\_ID.

**Final Dataset Overview**

**Description of Final Dataset:**

**Customer Loan Summary:**

* **CUSTOMER\_ID**: Unique identifier for customers.
* **Total\_Amount\_Financed**: Total amount financed for each customer by summing AMOUNT\_FINANCED across unique ACCOUNT\_NUMBER.
* **Total\_Amount\_Repaid**: Total amount repaid by each customer by summing AMOUNT\_PAID across unique ACCOUNT\_NUMBER.
* **Number\_of\_Loans**: Total number of unique loan accounts each customer has.
* **Average\_Loan\_Tenure**: Average loan tenure across all loans for each customer.
* **Age\_at\_Loan**: Age of the customer at the time of loan booking.

**Loan Repayment Details:**

* **CUSTOMER\_ID**: Unique identifier for customers.
* **ACCOUNT\_NUMBER**: Loan account number.
* **PRODUCT\_CODE**: Code representing the type of loan product.
* **COMPONENT\_NAME**: The type of repayment (e.g., interest, principal).
* **Total\_Amount\_Paid**: Sum of AMOUNT\_PAID for each repayment type per CUSTOMER\_ID and ACCOUNT\_NUMBER.
* **Due\_Date, Paid\_Date**: Specific dates related to each repayment.

By following these detailed data cleaning and preprocessing steps, the dataset will be prepared for further analysis and reporting, providing a comprehensive view of the customers' loan behaviors and repayment patterns, which is crucial for loan risk assessment.

///////////////////////////////////////////////////////////////////////////////////////

**Summary of the Loan Risk Assessment Process**

This script processes a dataset to generate a comprehensive summary for assessing loan risk. It includes key metrics like total amount financed, total amount repaid, number of loans, average loan tenure, age at loan booking, payment delays, and total interest paid, number of defaulted loans, and the ratio of principal to interest paid. These metrics provide a detailed view of each customer's loan behavior and repayment patterns, helping to assess loan risk effectively.

**Summary of the Report**

The generated report provides a detailed summary for each customer, including:

* **Total Amount Financed**: The sum of all loans taken by the customer.
* **Total Amount Repaid**: The total amount repaid by the customer.
* **Number of Loans**: The count of unique loans taken by the customer.
* **Average Loan Tenure**: The average duration of the loans.
* **Age at Loan**: The average age of the customer at the time of taking the loans.
* **Total Payment Delays**: The total number of days payments were delayed.
* **Total Interest Paid**: The total amount of interest paid by the customer.
* **Defaulted Loans**: The count of loans that were not fully repaid by the maturity date.
* **Average Principal to Interest Ratio**: The average ratio of the principal amount to the interest amount repaid.

**Exploratory Data Analysis (1-2 pages)**

**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 and Customer Age:**
   * The average loan tenure varies from very short durations (10 days) to multiple years, reflecting the variety of loan products offered by the bank.
   * The consistent age at loan booking around 44-45 years suggests that the bank's primary customers are middle-aged, possibly with established credit histories.
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**

1. CUSTOMER\_ID: Unique identifier for customers, crucial for tracking customer-specific data.
2. ACCOUNT\_NUMBER: Loan account number, important for distinguishing different loans taken by the same customer.
3. PRODUCT\_CODE: Code representing the type of loan product, relevant for analyzing product-specific trends and risks.
4. Total\_Amount\_Financed: Key for understanding the financial exposure and distribution of 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 the size of the loan portfolio.
7. Average\_Loan\_Tenure: Relevant for analyzing loan durations and terms.
8. Total\_Payment\_Delays: Critical for assessing payment behavior and risk.
9. Total\_Interest\_Paid: Important for understanding the interest burden on customers.
10. Defaulted\_Loans: Key for identifying high-risk customers.
11. Average\_Principal\_Interest\_Ratio: Relevant for understanding repayment patterns and customer behavior.
12. P\_ADDRESS1: The location of the borrower, useful for geographic analysis of loans.

**Modeling**

**Overview**

This section outlines the machine learning approach used for optimizing credit risk assessment at Abay Banks. 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. It 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.
4. **Gradient Boosting Machines (GBM)**: Another ensemble technique that builds models sequentially, with each new model attempting to correct the errors of the previous ones.

**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**:
   * **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**: 74% - This indicates that the model correctly classified 74% of the instances in the test set.
* **Precision**: 1.0 - This high precision score indicates that when the model predicts a loan will default, it is correct 100% of the time.
* **Recall**: 0.74 - The model correctly identifies 74% of actual loan defaults.
* **F1 Score**: 0.85 - This balanced measure of precision and recall demonstrates that the model has a good 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 Amount Financed**
2. **Total Amount Repaid**
3. **Number of Loans**
4. **Average Loan Tenure**
5. **Total Payment Delays**
6. **Total Interest Paid**
7. **Average Principal Interest Ratio**
8. **Product Code**

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 high precision (1.0) and good recall (0.74) and F1 Score (0.85), indicating strong accuracy in predicting loan defaults.
2. **Feature Importance**:
   * Significant features: Total Amount Financed, Total Amount Repaid, Number of Loans, Average Loan Tenure, Total Payment Delays, Total Interest Paid, Average Principal Interest Ratio, and Product Code.
   * These features offer valuable insights into borrower risk profiles.

**Implications for Business:**

1. **Enhanced Risk Assessment**:
   * Accurate predictions lead to better lending decisions, reducing non-performing loans and improving financial stability.
2. **Tailored Products**:
   * Insights allow the bank to create customized loan products and repayment plans for different risk segments.
3. **Efficient Resource Allocation**:
   * Focus on high-risk loans for monitoring, and automate low-risk loan approvals.

**Limitations and Recommendations**

**Limitations:**

1. **Data Quality**:
   * Model accuracy depends on data quality and completeness. Inconsistent or missing data can affect predictions.
2. **Model Generalization**:
   * Performance may vary with new data, especially if it differs from the training data.
3. **Feature Dependency**:
   * Features may change over time; regular model updates are necessary.

**Recommendations:**

1. **Data Enhancement**:
   * Improve data collection and integrate additional data sources like credit scores and employment history.
2. **Regular Updates**:
   * Regularly evaluate and retrain the model to maintain accuracy.
3. **Feedback Loop**:
   * Establish feedback from loan officers and risk managers to align model predictions with real-world observations.
4. **Risk Mitigation**:
   * Develop strategies for high-risk loans, such as proactive engagement and restructuring repayment plans.
5. **Scalability and Integration**:
   * Integrate the model into existing systems for automated decision-making and scalability.

By addressing these limitations and implementing the recommendations, Abay Banks can improve their credit risk assessment, leading to better decision-making and enhanced financial performance.

**Conclusion**

**Summary of the Project**

This project focused on optimizing credit risk assessment at Abay Banks using machine learning. The Random Forest model was identified as the most effective, delivering high precision (1.0) and balanced recall (0.74) and F1 Score (0.85). Key features influencing loan default predictions were identified, providing valuable insights into borrower risk profiles. The model's implementation promises improved lending decisions, reduced non-performing loans, and better resource allocation, ultimately enhancing 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**: The importance of clean, comprehensive data cannot be overstated. Inconsistent or incomplete data significantly impacts model performance.
2. **Model Maintenance**: Regular updates and retraining are necessary to ensure the model adapts to changing conditions and maintains its 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 its 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 Banks but also provided a framework for future data-driven decision-making initiatives.