LoanAnalytics Inc.™

A Portfolio Management Strategy

Case Study: Loan Default Prediction

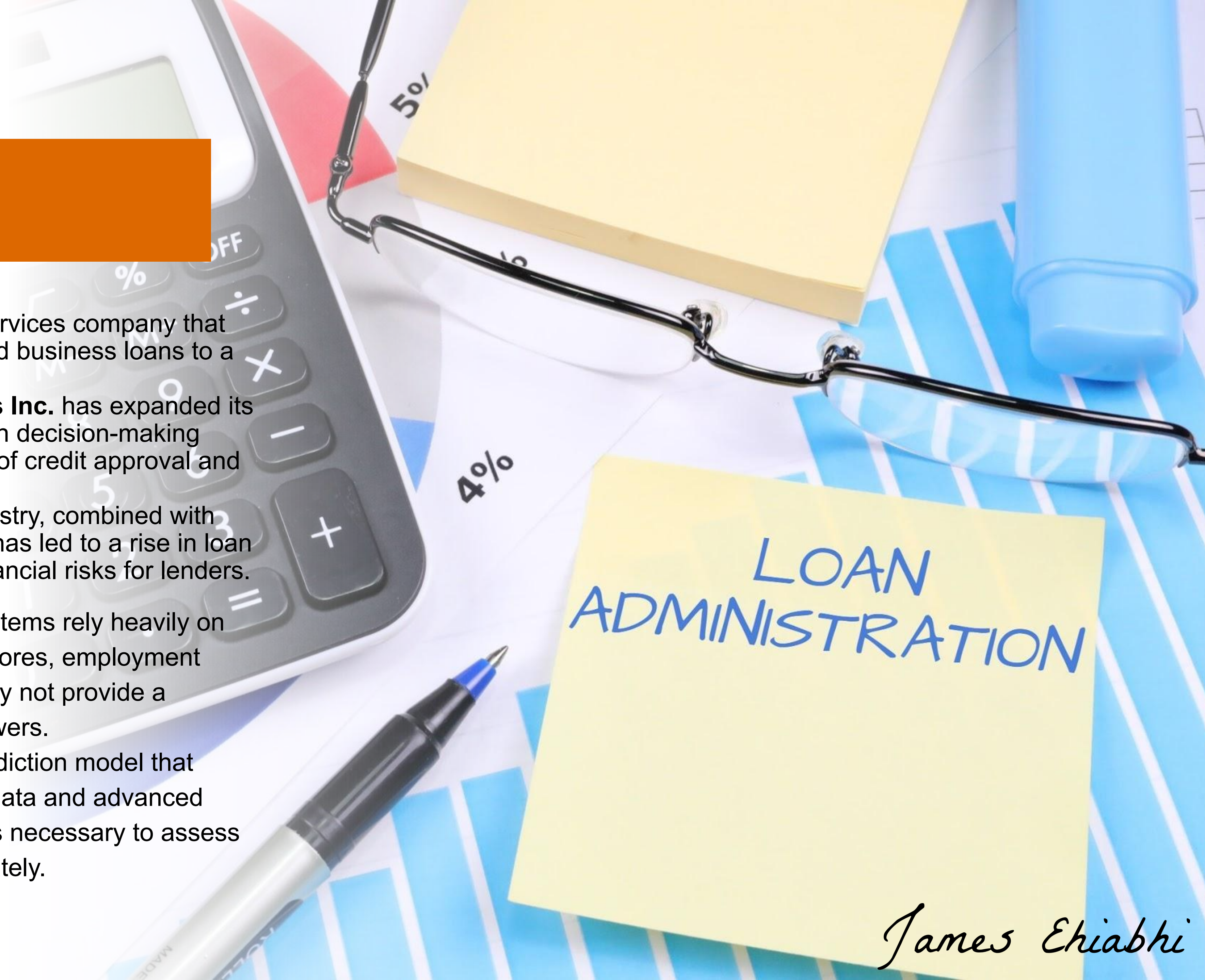
James Ehiabhi
Data Scientist | ML & AI Expert

kingjamesuwe@gmail.com
GitHub : JamesEhiabhi
in James Ehiabhi

Background

LoanAnalytics Inc. is a financial services company that specializes in providing personal and business loans to a broad range of clients.
Established in 2010, **LoanAnalytics Inc.** has expanded its client base by integrating data-driven decision-making processes, improving the efficiency of credit approval and risk management.
The rapid growth of the lending industry, combined with increasing economic uncertainties, has led to a rise in loan defaults, which poses significant financial risks for lenders.

- Traditional loan evaluation systems rely heavily on static information like credit scores, employment history, and income, which may not provide a complete risk profile for borrowers.
- An innovative loan default prediction model that integrates dynamic borrower data and advanced machine learning algorithms is necessary to assess the risk of default more accurately.



James Ehiabhi

Problem Statement

- LoanAnalytics Inc. seeks to implement an automated loan default prediction model that can identify high-risk borrowers early in the loan approval process. The model should help the company:

  a. Predict Borrower Default Risk:
     - ■ Develop an algorithm that can predict the likelihood of a borrower defaulting on their loan based on historical data.

  b. Improve Loan Approval Efficiency:
     - ■ Reduce the time it takes to process loan applications by incorporating predictive insights, thus reducing human intervention.

  c. Minimize Financial Losses:
     - ■ Enhance profitability by accurately identifying high-risk borrowers and taking preventive actions, such as offering adjusted interest rates or requiring collateral.

James Ehiabhi

Dataset

The dataset contains borrower information, financial metrics, and loan details, including:

- **LoanID:** Unique identifier for each loan.
- **Age:** Borrower's age.
- **Income:** Borrower's annual income.
- **LoanAmount:** Amount of the loan.
- **CreditScore:** Borrower's credit score.
- **MonthsEmployed:** Number of months the borrower has been employed.
- **NumCreditLines:** Number of credit lines the borrower has.
- **InterestRate:** Loan's interest rate.
- **LoanTerm:** Loan term in months.
- **DTIRatio:** Debt-to-income ratio.
- **Education:** Borrower's education level.
- **EmploymentType:** Employment status (e.g., Full-time, Unemployed).
- **MaritalStatus:** Marital status of the borrower.
- **HasMortgage:** Indicates if the borrower has a mortgage.
- **HasDependents:** Indicates if the borrower has dependents.
- **LoanPurpose:** Purpose of the loan (e.g., Auto, Business).
- **HasCoSigner:** Indicates if the borrower has a cosigner.
- **Default:** Whether the borrower defaulted on the loan (0 = No, 1 = Yes).

James Ehiabhi

Task

LoanAnalytics Inc. has provided a dataset of historical loans. Your task is to conduct an in-depth analysis of the dataset using the following steps:

1. **Univariate Analysis:**
    - Numerical Variables:
        - Analyze the distribution of key numerical features such as ApplicantIncome, LoanAmount, Age, and Loan_Amount_Term using histograms and summary statistics (mean, median, standard deviation).
        - Identify any outliers and skewed distributions that may need further treatment (e.g., log transformation).
    - Categorical Variables:
        - Examine the frequency distribution of categorical features such as Gender, Education, Married, and Property_Area using bar charts and count plots.
        - Explore any imbalance in the Default variable (target) to assess how many borrowers defaulted versus those who didn’t.
2. **Bivariate Analysis:**
    - Numerical vs Numerical:
        - Explore the relationships between pairs of numerical features, such as Income vs LoanAmount and Age vs Loan_Amount_Term, using scatter plots and correlation heatmaps to understand the strength and direction of relationships.
    - Numerical vs Categorical:
        - Compare how numerical variables like Income and LoanAmount vary across different categories (e.g., Education, Gender, and Loan_Status) using box plots or violin plots.
    - Categorical vs Categorical:
        - Investigate the relationships between categorical variables such as Married and Loan_Status, Education and Loan_Status, using stacked bar plots and chi-square tests to determine if these relationships are statistically significant.
3. **Multivariate Analysis:**
    - Numerical and Categorical:
        - Conduct a multivariate analysis to explore the combined influence of multiple variables. Use techniques such as:
            - Pair Plots: Visualize the pairwise relationships between numerical features (e.g., Income, LoanAmount, and Age) and group them by categorical variables such as Loan_Status.
            - Correlation Matrix: Analyze the correlation between all numerical variables and identify multicollinearity, which may affect predictive modeling.
    - Explore the interaction between categorical variables (e.g., Education and Self_Employed) and their combined influence on loan approval and default risk.
4. Feature Engineering:
    - Create new features based on domain knowledge to enhance predictive power:
        - Income-to-Loan Ratio: Generate a feature that measures the ratio of total income (Income + LoanAmount) to the loan amount.
        - Age Bucketing: Group Age into age ranges to analyze its effect on loan default.
        - One-hot encode categorical features such as Gender, Married, and Property_Area for machine learning models.

James Ehiabhi