Machine Learning-Driven Credit Risk Analysis

Oct 31, 2023

Partheesh Marwah Practical Data Science
MS DATA SCIENCE
Seidenberg School Of Computer Science and Information Systems, Pace University

Agenda

- Executive summary
- Project plan recap
- Data
- Exploratory data analysis
- Modeling methods
- Findings
- Recommendations and next steps

Executive summary

Problem: The objective of this project is to develop a predictive model using machine learning to identify potential loan defaulters based on historical financial data. The aim is to support the decision-making process in lending by minimizing the risk of financial loss through more accurate and data-driven risk assessments.

Solution: Leveraging advanced machine learning algorithms, we have implemented a series of predictive models tailored to forecast credit risk with a high degree of accuracy. These models process vast amounts of historical financial data to discern patterns indicative of potential defaulters, effectively streamlining MoneyMe's loan approval process. The outcome is a robust risk mitigation strategy that enhances the security of transactions by efficiently pinpointing reliable borrowers, thus fostering prudent lending practices and financial stability for MoneyMe.

Project plan recap

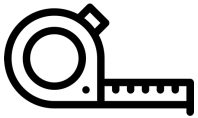
Deliverable	Details	Due Date	Status
Data & EDA	Create final assignment deck skeleton and fill in the deck up to the end of the EDA section	10/31/23	Complete
Methods, Findings, and Recommendations	Fill in the Methods, Findings, and Recommendations sections of the deck	11/14/23	In Progress
Final presentation	Send in final completed deck and present it	12/05/23	Not started

Data

DATA OVERVIEW

DATA SOURCE

Data was provided firsthand by the CTO of MoneyMe, ensuring a reliable and direct source.



SAMPLE SIZE

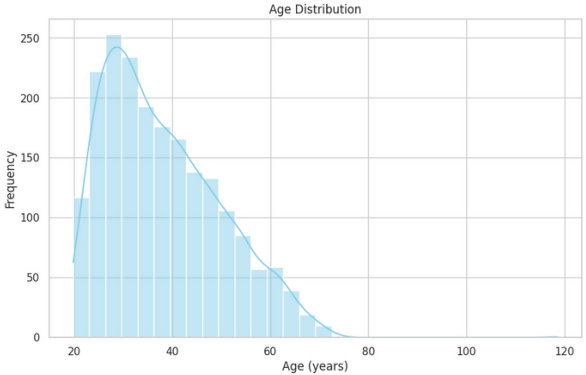
Our analysis is based on a comprehensive dataset consisting of 2,247 individual records, ensuring a robust and representative foundation for our insights and predictions.

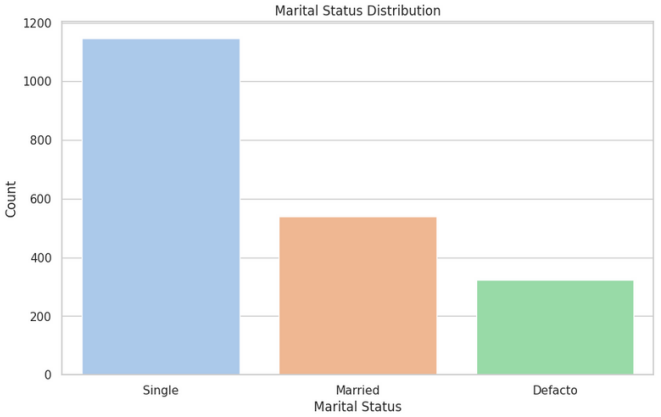
TIME PERIOD

Our dataset encapsulates the first quarter of 2020, a dynamic period that allows us to analyze financial behaviors and trends at the onset of global changes.

Exploratory Data Analysis

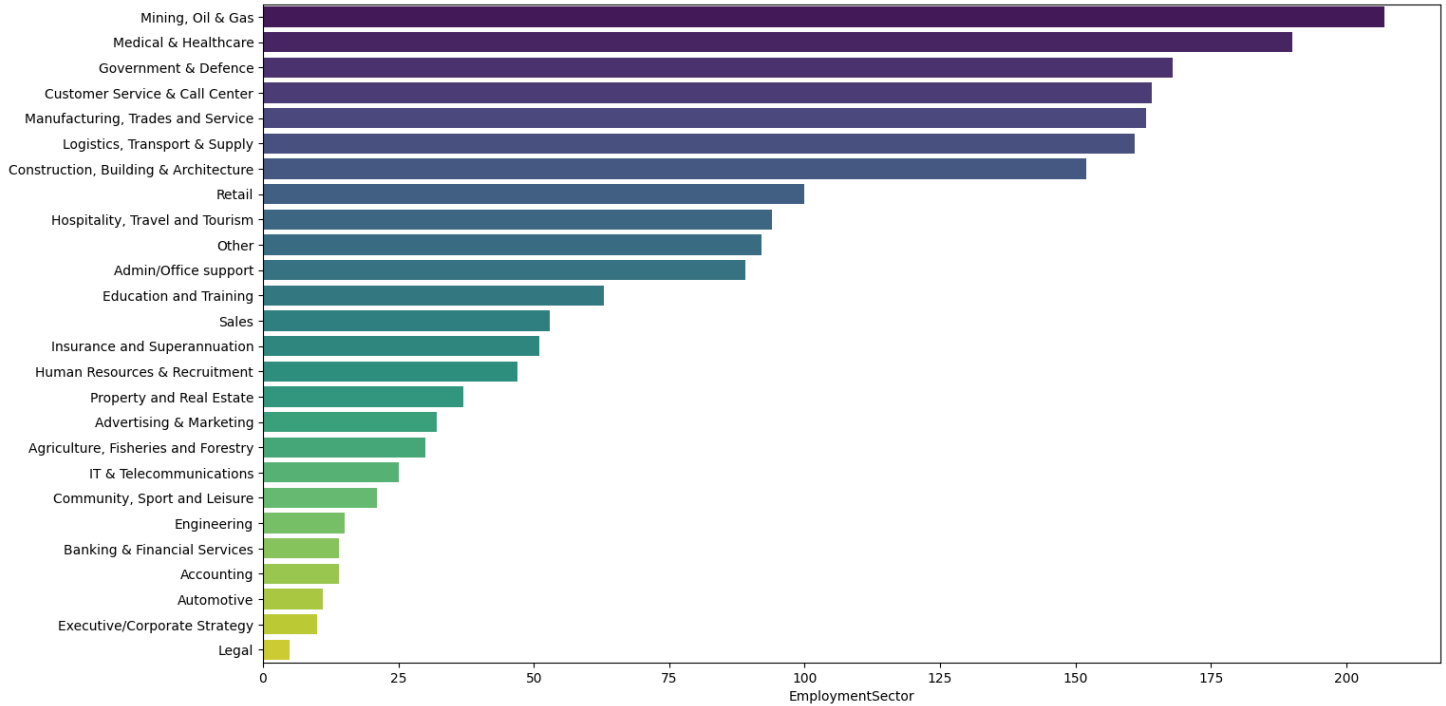
DEMOGRAPHICS





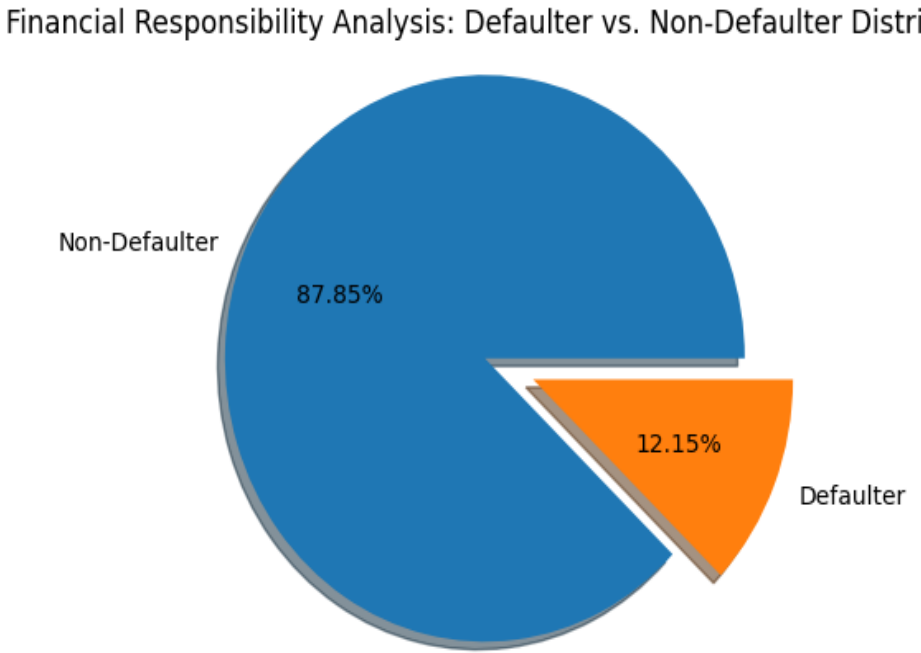
- GENDER BALANCE: THE DATA INDICATES A FAIR MIX OF BOTH "MALE" AND "FEMALE" PARTICIPANTS.
- Age Range: Participants vary widely in age, ensuring diverse insights.

Distribution of Employment Sectors Among Borrowers



THE "HOSPITALITY, TRAVEL AND TOURISM" AND "MANUFACTURING, TRADES AND SERVICE" SECTORS STAND OUT AS THE PRIMARY SEGMENTS WHERE PROFESSIONALS ARE SEEKING FINANCIAL ASSISTANCE, SUGGESTING THESE AREAS MAY HAVE THE HIGHEST DEMAND FOR LENDING SOLUTIONS.

DEFAULTER VS NON- FEAULTER DISTRIBUTION

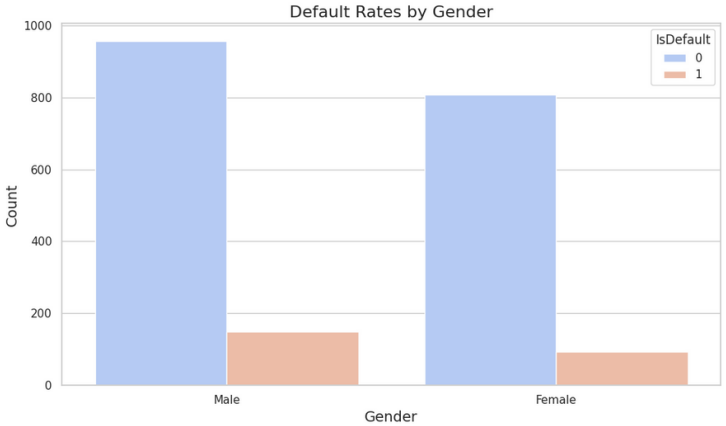


- MAJORITY NON-DEFAULTERS INDICATE STABLE CREDIT HEALTH; HOWEVER, DEFAULT PRESENCE SUGGESTS RISK MITIGATION POTENTIAL.

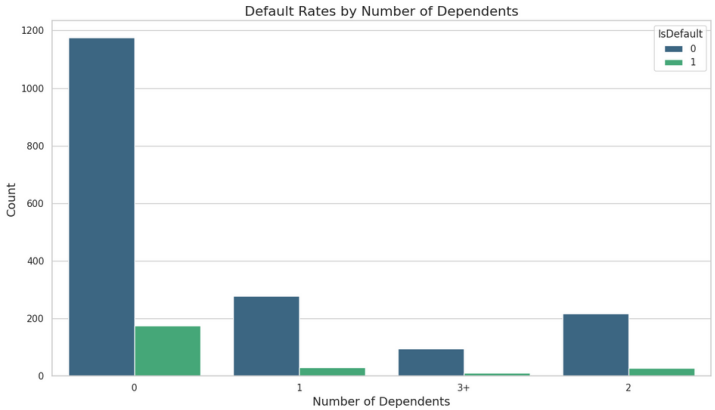
- Data underscores need for refined credit risk assessments to reduce defaults.

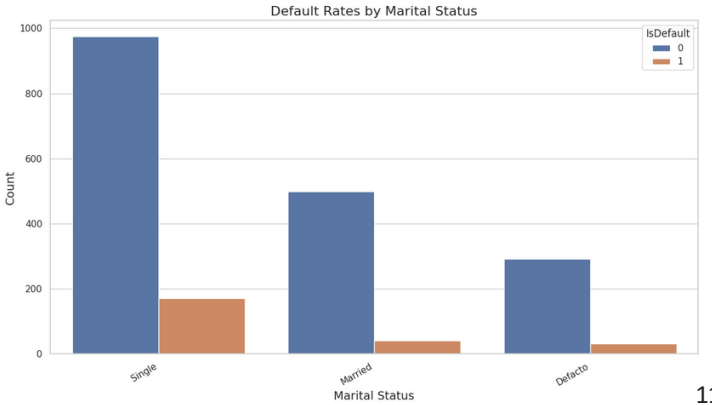
DEFAULT TRENDS BY DEMOGRAPHIC FEATURES

As we see, Gender plays no major role in determining defaulters based on the data.

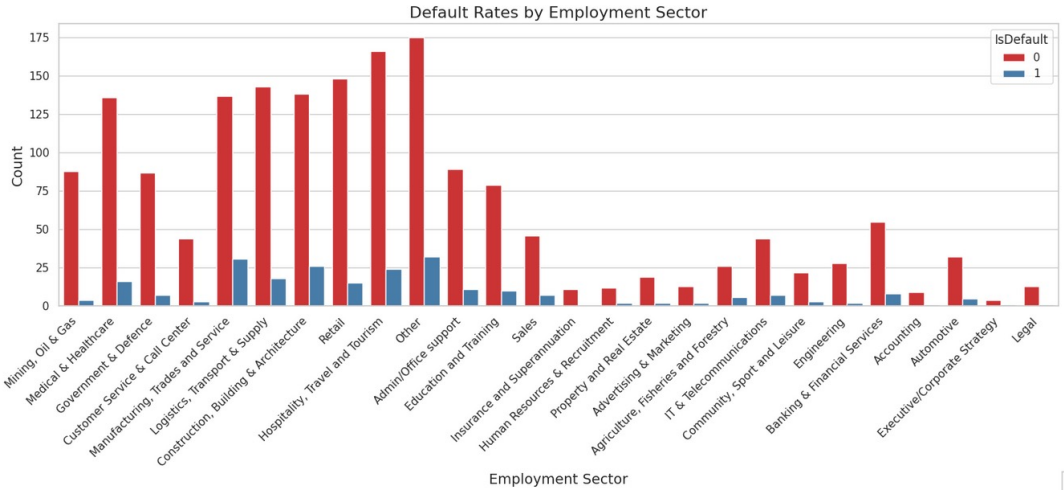


As per data, Single individuals with no dependents are more likely to default in compared to Married people with dependents.



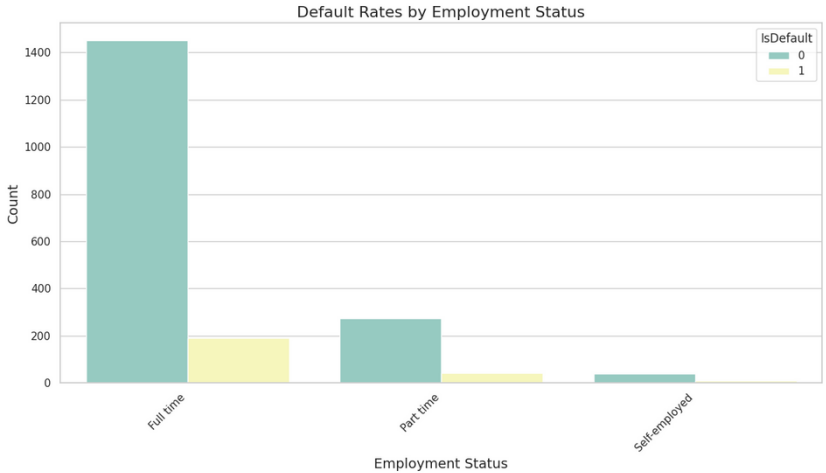


DEFAULT TRENDS BY EMPLOYEMENT FEATURES



- AS PER THE VISUALIZATION, THE SECTORS OF HOSPITALITY, TOURISM, AND MANUFACTURING EMERGED AS THE PREDOMINANT EMPLOYMENT SECTORS WITH THE HIGHEST INCIDENCES OF DEFAULTS.

- THE PART-TIME EMPLOYMENT STATUS TO DEFAULT RATIO COMES OUT AS THE HIGHEST.



Modeling methods

TARGET VARIABLE

"In my project, the output variable identifies an individual as either a 'Defaulter' or 'Non-defaulter,' serving as a key indicator of credit risk, which is essential for making informed lending decisions and managing financial risks effectively.

IMPORTANCE

1.Risk Mitigation: By accurately identifying defaulters, the model helps in minimizing the risk of bad debt, ensuring a healthier credit portfolio.
2.Strategic Decision Making: It aids in tailoring credit policies and underwriting standards, enabling more strategic decision-making in credit offerings and customer segmentation.

BUSINESS PROBLEM IT SOLVES:

By accurately predicting 'IsDefaulter', our model directly contributes to reducing financial losses from uncollected debts, enhances the efficiency of our credit allocation, and strengthens customer trust through more responsible lending practices

MODEL RELEVANCE

The 'IsDefaulter' target variable is pivotal in guiding the design, training, and evaluation of our predictive models, ensuring they are finely tuned to assess credit risk accurately and support informed decision-making

REAL WORLD APPLICATION

Predicting 'IsDefaulter' status helps financial institutions reduce loan defaults, improve profitability, and offer better terms to reliable customers, enhancing market stability.

KEY FEATURES USED IN THE MODEL

BREIF DESCRIPTION

The effectiveness of our predictive model heavily relies on the set of features we choose to include. These features are selected based on their relevance and impact on the target variable, 'IsDefaulter'

FEATURE RATIONALE

We selected features that provide comprehensive insights into an individual's financial behavior and credit history, as these are strong indicators of their likelihood to default.

FEATURE CATEGORIES

- Consumer Behavior: Features that track spending patterns, payment history, etc.
- Financial History: Includes credit score, previous loan history, etc.
- Demographic Information: Age, employment status, income level, etc.

SELECTED FEATURES

- **Financial History**: Indices 1, 2, 7, 11 (e.g., Credit Score, Debt Ratio)
- **Loan Details**: Indices 12, 13, 15, 28, 29, 30 (e.g., Loan Amount, Interest Rate)
- **Personal Information**: Indices 33, 69, 70, 71, 72 (e.g., Age, Employment Status)
- **Economic Indicators**: Indices 75, 76, 83, 88 (e.g., Market Trends, Employment Rate)
- **Behavioral Metrics**: Indices 95, 97, 98, 103, 104, 105 (e.g., Payment History, Spending Behavior)

UNDERSTANDING OUR DECISION TREE MODEL(NON-TECHNICAL)

Decision Tree in Everyday Terms

- "Imagine a tree with branches leading to different outcomes. Each branch represents a decision based on simple questions about the borrower's information.”
- "For example, one branch might ask, 'Is the borrower's income above a certain level?' Depending on the answer, we follow the branch to the next question, leading us closer to predicting if they might default."

WHY DECISION TREE?

- Direct and Practical: "The Decision Tree model stands out for its straightforward approach. It's like having a step-by-step guide to evaluate whether someone might default on their loan, making complex decisions more manageable.”
- Adaptable to Changing Data: "In the dynamic world of finance, conditions change rapidly. Decision Trees can be easily updated as we get new data, ensuring our predictions stay relevant and accurate.”
- "Just like a loan officer, the model examines various aspects of a borrower’s profile – income, credit history, etc. – to make an informed decision."

FURTHER READING

- For more technical details, please see our detailed explanation in the Appendix

UNDERSTANDING OUR DECISION TREE MODEL

TECHNICAL OVERVIEW

A non-parametric approach used for both classification and regression tasks. It segments the dataset into subsets based on feature value thresholds, forming a tree-like model of decisions.

Rationale for Choosing Decision Trees in Loan Defaulter Prediction

- "Optimal for our mixed data types (categorical and numerical), effectively handling the varied aspects of financial data."
- "Able to capture non-linear relationships and intricate patterns in borrower profiles, crucial for accurate defaulter prediction."

TECHNICAL ADVANTAGES

- Transparent decision-making process – easy to trace how each feature influences the prediction."
- "Efficient with large datasets, enabling us to process extensive borrower data swiftly and accurately."

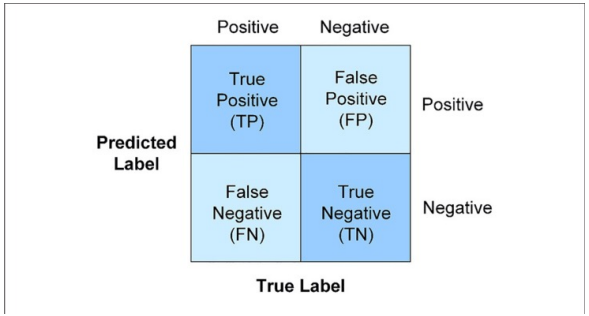
Implementation Insights

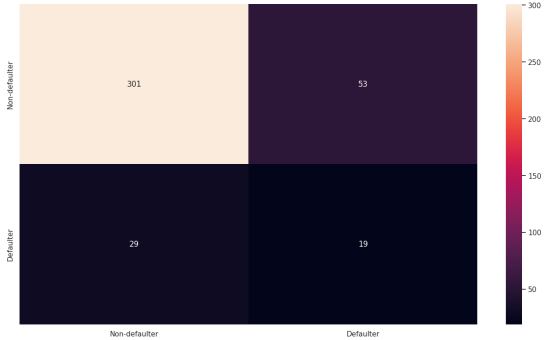
- "Implemented using entropy and information gain to determine the best feature splits, maximizing the homogeneity of each node."
- "Pruning methods applied to curb overfitting, enhancing the model's generalizability to unseen data."

Findings

Prioritizing Recall: Key to Identifying Loan Defaulters

Maximizing Recall: Ensuring We Identify Most Defaulters



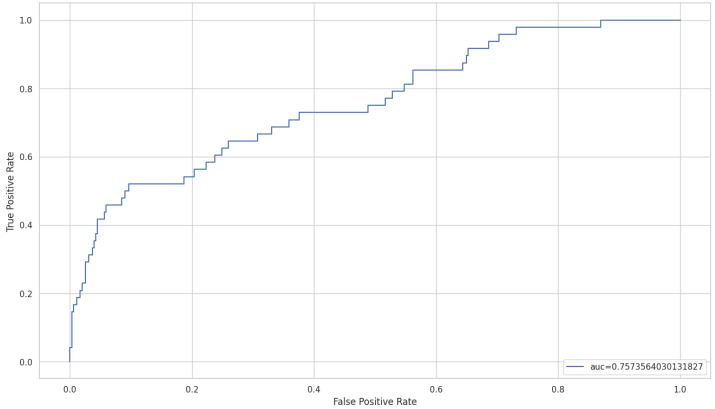


Recall indicates our model's ability to identify actual defaulters. A high recall means we're catching a large percentage of potential default risks.

With high recall, we minimize the risk of lending to defaulters, protecting our financial interests. It’s about not letting defaulters slip through the net.

•**Ideal Recall**: "For defaulter prediction models, aiming for a high recall, such as above 80%, is typically ideal."
•**Our Achievement**: "Through our modeling, we successfully achieved a recall of 86.3%, aligning closely with our target, ensuring we effectively identify the majority of potential defaulters."

Beyond Recall: A Comprehensive View of Model Metrics



•**Overall Accuracy**: Our model achieves an impressive accuracy of 86.3%, reflecting its effectiveness in predicting loan defaults accurately.
•**Cross-Validation Confidence**: With an average cross-validation score of 87.8%, our model demonstrates consistent performance across different subsets of data, ensuring reliability and robustness.
•**Precision in Predictions**:
- **Non-Defaulters**: High precision of 91% in predicting non-default cases, indicating a strong ability to identify reliable borrowers.
- **Defaulters**: A precision of 26% in predicting defaulters, pointing towards potential areas for model refinement to better identify high-risk loans.

•**Recall - Identifying Default Risks**:
- **Non-Defaulters**: Successfully captures 85% of actual non-default cases, reinforcing the model's effectiveness in recognizing secure loans.
- **Defaulters**: With a 40% recall rate for defaulters, the model shows a promising start in flagging potential default risks, with room for improvement.

•**F1-Score**:
- **Non-Defaulters**: An F1-score of 88% indicates a well-balanced model for non-default predictions.

•**Model Reliability and Future Directions**:
These metrics show a strong foundation in our predictive capabilities while highlighting areas for future enhancements, particularly in improving precision for identifying defaulters.

Recommendations & next steps

Targeted Strategies for Improved Risk Management

1. **Duration Matters:** Enhancing Engagement in Financial Interactions
2. **Actionable Insight**: Implement initiatives to extend and deepen customer interactions during financial consultations or loan application processes.
3. **Tactics**:
    1. **Agent Training**: Develop specialized training programs to equip loan officers with advanced communication skills that foster in-depth discussions with clients.
    2. **Process Refinement**: Optimize the loan application process to encourage thorough customer engagement and understanding.
    3. **Incentives for Depth**: Introduce rewards for loan officers who effectively engage clients in comprehensive financial discussions, potentially leading to more informed lending decisions.
4. **Seasonal Patterns:** Capitalizing on High-Opportunity Periods
5. **Actionable Insight**: Identify and leverage periods with historically lower default rates or better financial stability among borrowers.
6. **Strategic Marketing**: Intensify loan marketing efforts during these times, targeting demographics that show higher financial stability or lower default rates in these periods.
7. **Past Success Indicator: Prioritizing Reliable Borrowers**
8. **Actionable Insight**: Focus on clients with a positive history of repayment and financial stability.
9. **Tailored Approach**: Develop customized loan offers and terms for these low-risk clients, using their past financial behavior as a guiding factor for personalized lending solutions.

Forging Ahead: Future Directions for Our Project

Expanded Data Collection:

Currently, our analysis is based on a limited dataset from Q1. To enhance the model's efficiency and accuracy, we'll expand our dataset to include more comprehensive data across multiple quarters or even years. This broader dataset will provide more insights and improve the model's predictive power.

Integration of Additional Variables:

Explore the inclusion of more nuanced variables such as macroeconomic indicators, borrower's financial behavior over time, and other relevant financial metrics to provide a more detailed risk assessment."

Advanced Model Development:

Investigate and experiment with more advanced machine learning models, such as Random Forest or Gradient Boosting, to further improve our prediction accuracy, especially in terms of precision and recall.

Continuous Model Refinement

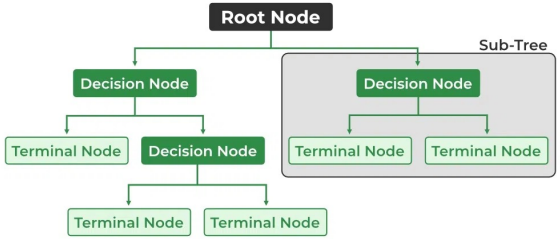
Regularly update and refine the model with new data and insights. This will help us stay ahead of changing financial trends and borrower behaviors.

Appendix

GITHUB

HERE'S THE GITHUB LINK TO THE CODE FOR THE PROJECT

DECISION TREE



A decision tree is a flowchart-like tree structure where each internal node denotes the feature, branches denote the rules and the leaf nodes denote the result of the algorithm. It is a versatile supervised machine-learning algorithm, which is used for both classification and regression problems. It is one of the very powerful algorithms. And it is also used in Random Forest to train on different subsets of training data, which makes random forest one of the most powerful algorithms in machine learning.

TERMINOLOGIES:

•**Root Node:** It is the topmost node in the tree, which represents the complete dataset. It is the starting point of the decision-making process.
•Decision/Internal Node: A node that symbolizes a choice regarding an input feature. Branching off of internal nodes connects them to leaf nodes or other internal nodes.
•**Leaf/Terminal Node:** A node without any child nodes that indicates a class label or a numerical value.
•**Splitting:** The process of splitting a node into two or more sub-nodes using a split criterion and a selected feature.
•**Branch/Sub-Tree:** A subsection of the decision tree starts at an internal node and ends at the leaf nodes.
•**Parent Node:** The node that divides into one or more child nodes.
•**Child Node:** The nodes that emerge when a parent node is split.
•**Impurity**: A measurement of the target variable's homogeneity in a subset of data. It refers to the degree of randomness or uncertainty in a set of examples. The **Gini index** and **entropy** are two commonly used impurity measurements in decision trees for classifications task
•**Variance**: Variance measures how much the predicted and the target variables vary in different samples of a dataset. It is used for regression problems in decision trees. **Mean squared error, Mean Absolute Error, friedman_mse, or Half Poisson deviance** are used to measure the variance for the regression tasks in the decision tree.
•**Information Gain:** Information gain is a measure of the reduction in impurity achieved by splitting a dataset on a particular feature in a decision tree. The splitting criterion is determined by the feature that offers the greatest information gain, It is used to determine the most informative feature to split on at each node of the tree, with the goal of creating pure subsets
•**Pruning**: The process of removing branches from the tree that do not provide any additional information or lead to overfitting.