Milestone 2: Data Selection and Project Proposal

# 1. Choose a Business Problem

## Define the Business Problem or Question:

I plan to explore the problem of loan default risk in the banking industry. The goal is to predict whether a customer is likely to default on a loan based on their personal characteristics, financial background, and loan details. Banks use predictive analytics to identify customers at risk of default and make informed decisions on whether to approve or reject loans. Rather than simply rejecting high-risk applications, I aim to go a step further: developing a model that adjusts the interest rate or fee structure based on the predicted probability of default. This strategy allows banks to extend loans even to higher-risk customers while mitigating their financial exposure by offering adjusted terms.

Why is the Problem Important?  
Predicting loan default is crucial for the financial health of banks and lending institutions. A high default rate leads to significant losses, but rejecting too many customers can also result in lost revenue opportunities. By predicting default risk and tailoring loan terms to account for that risk, banks can make more profitable decisions. Offering higher interest rates or one-off fees to higher-risk borrowers could turn a potentially lost opportunity into a revenue-generating one, creating a win-win situation for both the bank and the customer.

# 2. Identify the Dataset

Key Characteristics:  
I will use a loan default dataset that includes information on borrower characteristics (such as salary, marital status, employment history), loan characteristics (loan amount, interest rate, duration), and whether the borrower defaulted on the loan. Currently I have chosen 2 datasets from Kaggle with over 250 thousand unique rows of information.

*Lending Club Loan Data*. (2021, June 17). Kaggle. https://www.kaggle.com/datasets/adarshsng/lending-club-loan-data-csv

*Loan Default Prediction Dataset*. (2023, September 11). Kaggle. https://www.kaggle.com/datasets/nikhil1e9/loan-default

Why is the Dataset Relevant?  
These datasets are essential for building a model that can predict loan default risk. By analyzing past data of customers who either defaulted or paid off their loans, I can identify the factors that contribute to default and use these insights to assess future applicants. Additionally, understanding how specific borrower characteristics relate to default probability will enable the bank to offer personalized interest rates or fees.

# 3. Plan for the Model(s)

Types of Models You Plan to Use:  
For predicting loan default, I will start with logistic regression, a standard method for binary classification. This model will allow me to predict the probability that a given borrower will default. Depending on the initial results, I may also explore more complex models like random forests or gradient boosting to capture non-linear relationships and improve predictive performance.

Evaluation Metrics:  
To evaluate the model, I will focus on metrics such as:

* Precision: To ensure that when we default is predicted, it is likely to happen.
* Recall: To minimize the number of false negatives (i.e., failing to predict a default when one is likely).
* F1-score: To balance precision and recall, especially in a case where false negatives or false positives have significant financial consequences.
* AUC-ROC: To assess the overall discriminatory ability of the model, especially for varying thresholds of risk tolerance.

# 4. What You Hope to Learn

Desired Insights:  
I expect to identify the key factors that contribute to loan default risk. These could include low income, high debt-to-income ratios, poor credit scores, or other personal and financial characteristics. The goal is to build a model that can predict default risk for future customers and offer recommendations for either rejection or approval with adjusted interest rates.

Broader Business Value:  
The insights gained from this analysis will allow the bank to extend loans to a broader customer base by adjusting terms for higher-risk customers rather than outright rejecting them. This could increase revenue streams while balancing risk. Moreover, the model will help improve decision-making, enhance customer satisfaction by offering more personalized loan terms, and create competitive differentiation for the bank.

# 5. Risks or Ethical Concerns

Potential Risks:  
A major risk in this project is the possibility of biased data. If certain demographics are over- or under-represented in the training data, the model could unfairly favor or penalize specific groups, leading to biased lending decisions. Additionally, the data could contain outliers or missing information, which could affect the accuracy of the model.

Ethical Concerns:  
It is critical to ensure that the model does not discriminate based on race, gender, or age, which are sensitive characteristics in loan approval processes. Ensuring fairness in the model and complying with regulatory requirements, such as the Equal Credit Opportunity Act (ECOA), is essential.

# 6. Contingency Plan

If the Original Plan Doesn't Work:  
If the dataset does not contain sufficient variability, or if model performance is inadequate, I will explore additional features that could improve predictions, such as incorporating external data like credit scores or market interest rates. Additionally, if logistic regression and tree-based models do not yield satisfactory results, I will explore more advanced methods to capture more complex patterns in the data.

# 7. Additional Considerations

Execution Plan:  
I will use Python for data analysis and modeling, with libraries like pandas, scikit-learn, and matplotlib. Regular model performance checks will help ensure that the project is progressing smoothly.