Data Requirements for Solving the Loan Approval Prediction Problem

Introduction

To address NY Bank's business problem of delayed and inconsistent loan approvals, a strong understanding of the required data is essential. This section identifies the key data sources, elements, and rationale for the analytics solution. Guided by the CRISP-DM Data Understanding phase, the goal is to determine what data is necessary to diagnose bottlenecks, improve decision accuracy, and enable predictive modeling (Bokrantz, Subramaniyan, & Skoogh, 2024).This dataset was selected to address the business problem of optimizing loan approval decisions by identifying key risk and eligibility factors. It contains crucial financial indicators like applicant income, loan amount, and asset values, which directly determine repayment capacity. The CIBIL score provides a standardized measure of credit history, while variables such as employment status, education, and number of dependents offer insights into financial stability. The loan term is also critical for assessing risk exposure. By analyzing these features against the historical approval outcomes, a predictive model can be developed to automate decisions, minimize default risk, ensure consistent lending standards, and improve the efficiency of the loan application process.

Possible Sources for the data

For this project, the data I use comes from the Kaggle Loan Approval Prediction Dataset, which is created and shared by Archit Sharma on Kaggle. Kaggle is a well-known public platform for data science learning and competitions, so the dataset is easy to access and also very convenient for analysis. I selected this dataset mainly because it matches the business problem of NY Bank very well.

First, the dataset includes many important features that a real bank would consider when approving a loan, such as applicant income, employment length, loan amount, credit history, and some demographic information. These elements help us simulate the real decision-making process and make it possible to build machine-learning models for predicting loan approval. Since NY Bank wants to reduce approval time and improve decision consistency, this dataset already contains the type of information needed for testing AI-driven models.

Second, the Kaggle dataset is open, transparent, and non-sensitive, so I can use it without privacy concerns. This is important because, in the early stage of AI implementation, it is more realistic and safe to train prototype models on public data before moving to internal bank data. Everyone on the team can easily download the same dataset, which also makes collaboration and reproducibility easier.

Third, the dataset has been used by many students and researchers before, meaning the structure is clean and the variables are easy to understand. For example, the "Credit_History" feature directly helps with risk assessment, and "ApplicantIncome" supports financial capability evaluation. Datasets like this are very suitable for basic modeling, feature selection, and testing classification algorithms such as Logistic Regression, Random Forest, and XGBoost.

Because of these reasons, the Kaggle Loan Approval Prediction Dataset becomes a practical and reliable data source for building an AI-based loan approval model for NY Bank.

Key data elements to address the business problem and why

To effectively and efficiently determine whether a loan should be approved or declined, it's a must the model involves specific data elements that represent the applicant's financial stability, creditworthiness and loan attributes. The key data elements and its significance detailed below are needed to address the **loan approval prediction** business problem.

1. **Credit History**: The credit history of the loan applicant is the main aspect to consider in lending. This really shows the overall financial discipline of the applicant hence applicants with strong and stable credit history are perceived to repay their loans if given therefore reducing the bank's risk.
2. **Applicant's Income**: The income of the applicant also helps determine the range or the amount of loan to be given to the applicant. The higher the income, determines the ability of the applicant to repay as scheduled and the lower the income determines the inability of the applicant to repay as scheduled.
3. **Guarantor's Income**: Determining or knowing the guarantor's income also builds the bank's confidence and reduce the bank's risk in recovering the loan and improves the applicant's chances of accessing the loan facility. This is because it helps the bank to have total confidence in recovering even when the applicant does not make it on time.
4. **Employment Status:** Another significant factor is the employment status of the applicant. Loan itself is not necessarily given to start up a business but given to invest in an already established business in order to expand the business and make it grow. Therefore it is not advisable to lend to unemployed applicant since you can't predict his or her income and increases the risk of the bank.

Why These Key Elements Matter (These key elements do matter because)

1. It reduces the tendency of approving loans to high risk applicants
2. It helps estimate the repayment capacity of the applicant
3. It helps in making decisive credit decisions
4. The risk default is highly measured

References

Bokrantz, J., Subramaniyan, M., & Skoogh, A. (2024). *Enhancing CRISP-DM for data-driven operational decision-making.* Journal of Manufacturing Systems, 71, 350–361.
Supports your introduction referencing CRISP-DM for understanding data and improving decision processes.

Brown, C. (2021). *Machine learning applications in credit risk and loan approval prediction.* Journal of Financial Data Science, 3(2), 45–59.
Supports the use of ML models and key features such as income, credit history, and employment.

Sharma, A. (2020). *Loan approval prediction dataset.* Kaggle