
# ***GRAND PROJECT***

***SUBMITTED BY: BEENISH RIAZ***

***STUDENT ID: EB21103029***

***COURSE: DATA SCIENCE***

***INSTRUCTOR: SIR FAISAL***

***DATE OF SUBMISSION: 15-FEB-2024***

## *1- WHAT ARE THE KEY FACTORS INFLUENCING CREDIT RISK IN LOAN APPROVAL DECISIONS?*

**FACTORS THAT CONTRIBUTE TO LOAN APPOVAL DECISIONS**

*How you will use the loan* Lenders want to make sure you're using the right product for your needs. Options may include small business credit cards, which are designed to help you manage day-to-day expenses; a line of credit, which is generally used for short-term working capital needs; and a commercial term loan, which is best for financing larger investments over time.

*The amount of financing you're seeking* Attempting to borrow more than your business can afford is a red flag to lenders. Lenders may also question the application if you don't borrow enough for your demonstrated need. "If a doctor is applying for a loan for a new practice but isn't including the office build-out, they could end up cash-strapped. It may make sense to borrow those additional funds, so they have more cash on hand in the short term," Wilson says. Seek advice from your accountant or banker on how much to borrow.

*Your business and personal credit profile* Before you submit a credit application, review both your personal and business credit reports for delinquent accounts (or incorrectly reported delinquencies) with all major credit reporting agencies. Business credit reporting agencies include Dun & Bradstreet, Experian and Equifax; you can check your personal credit reports with LExperian, Equifax and TransUnion. If there is any negativeT information in your credit report, submit an explanation so the lender can better understand the situation.

*Your capacity to repay* Your application should also demonstrate your ability to pay back borrowed money. You might also need to show business and personal assets, as well as cash reserves. Lenders often want to know about your business's capital assets such as cash and equipment, and about any funds that others have invested in your business.

**How long will it take for your application to be processed?**

Each borrower's situation is different, so time frames for approval and funding may vary. A typical commercial mortgage might take up to 60 days, while a line of credit might take three to four weeks. Credit card approvals may take a week or less. If the lender requests additional documentation, the process might take longer.

## *2- WHAT ROLE DOES CREDIT HISTORY PLAY IN ASSESSING CREDIT RISK?*

**Assessment of Credit**

Credit history plays a crucial role in lenders' assessment of credit risk. It provides lenders with valuable information that helps them determine the likelihood of a borrower defaulting on a loan. By analyzing a borrower's credit history, lenders can accurately assess the risk associated with extending credit to an individual.

Let's explore how credit history impacts lenders' assessment of credit risk in greater detail:

1. *Default Probability:* A borrower's credit history provides insights into their default probability. Lenders analyze past payment behavior, delinquencies, and any previous defaults to gauge the likelihood of future defaults. This information helps lenders make informed decisions about loan approvals and interest rates.

2. *Risk Mitigation:* Lenders use credit history to mitigate their risk exposure. By examining a borrower's credit history, they can identify potential red flags and take appropriate measures to protect their interests. This may include adjusting interest rates, requiring collateral, or setting specific loan terms.

3. *Loan Pricing:* Credit history plays a significant role in determining the interest rates and loan terms offered to borrowers. A borrower with a strong credit history is more likely to be offered favorable loan terms and lower interest rates, as they are considered low-risk borrowers.

4. *Credit Score:* A borrower's credit history contributes to the calculation of their credit score. Lenders often rely on credit scores to assess credit risk quickly. A higher credit score indicates lower risk, increasing the borrower's chances of obtaining credit on favorable terms.

5. *Eligibility for Credit:* Lenders use credit history to determine a borrower's eligibility for credit. A poor credit history, characterized by delinquencies, defaults, or excessive debt, can result in loan denials or limited credit options.

## *3- ARE THERE SPECIFIC EMPLOYMENT OR FINANCIAL INDICATORS THAT CORRELATE WITH LOAN REPAYMENT?*

**Key Indicators for loan quality assessment**

*Credit Score and Loan Quality:* A credit score is a numerical expression based on an individual's credit history, which represents their creditworthiness. lenders consider credit scores as a primary factor in determining the likelihood of repayment. A higher credit score signifies a higher probability of timely loan repayment, while a lower credit score indicates a higher risk of default.

Debt-to-Income Ratio and Loan Quality: The debt-to-income (DTI) ratio is another crucial indicator considered in loan quality assessment. This ratio measures a borrower's ability to repay debts by comparing their total monthly debt payments to their monthly gross income. A lower DTI ratio indicates a higher capacity to manage additional debt, while a higher DTI ratio suggests a greater risk of default.

*Loan-to-Value Ratio and Loan Quality:* The loan-to-value (LTV) ratio assesses the ratio of the loan amount to the appraised value of the collateral provided by the borrower. This indicator is particularly important for secured loans where the collateral serves as a backup in case of default. A lower LTV ratio indicates a lower risk for the lender, as the borrower has a greater equity stake in the asset.

*Employment History and Loan Quality:* The employment history of a borrower plays an essential role in loan quality assessment. It provides insights into the stability of income and the overall financial situation of the borrower. Lenders typically prefer borrowers with a consistent employment history, implying a reliable source of income and a lower risk of default.

*Payment History and Loan Quality:* collateral evaluation is particularly significant for secured loans, where the borrower provides an asset as security. In the event of default, the lender can seize the collateral to recover the outstanding loan amount. Thus, the value and condition of the collateral play a significant role in loan quality assessment.

*Documentation and Loan Quality:* proper documentation is crucial for loan quality assessment, as it ensures transparency and reduces the chances of fraud. Lenders require borrowers to provide various documents, such as income statements, tax returns, bank statements, and proof of identity, to verify their financial information.

## *4- CAN WE BUILD A PREDICTIVE MODEL THAT ACCURATELY PREDICTS THE LIKELIHOOD OF DEFAULT FOR LOAN APPLICANTS?*

Yes, we can build a predictive model that accurately predicts a default of loan applicant. At present, researchers generally use machine learning methods to predict loan defaults, including Logistic Regression, Decision Trees, Random Forest, XGBoost, and other advanced techniques.

## *5- WHICH MACHINE LEARNING ALGORITHMS ARE MOST SUITABLE FOR THIS BINARY CLASSIFICATION TASK?*

*Logistic Regression* naturally outperforms Linear Regression in predicting the probability of loan default since its outcome contains a continuous range of grades between 0 and 1, which represents the likelihood of an event occurring. Han used Logistic Regression and Cox proportional hazard algorithm to predict student loan default, whose findings indicated that the main affected factors that led to student loan default lie in age, household income, monthly repayable amount, and the college major. The Logistic

Regression model that they developed gained an AUC of 0.697 for the test data, which showed the accuracy and robustness of LR.

*Decision Trees* generates a structure like a tree by classifying the instances and using recursive portioning algorithm. Each leaf node represents a class label and branches present the outcomes for the test, which are represented by internal nodes for an attribute. In order to predicting the businesses' past due Inservice accounts, Wang developed models using Logistic Regression and Decision Trees in SAS and compared their results. It turned out that Decision Trees outperformed Logistic Regression when there were small number of attributes in a large enough sample.

*Random Forest* runs by constructing multiple decision trees while training and outputting the class that is the mode of the classes output by individual trees, which outperformed single decision trees. Malekipirbazari and Aksakalli built a Random Forest based classification model to identify high-quality peer-to-peer borrowers. They compared the different machine learning techniques and found out that the Random Forest based model performed significantly better than the FICO credit scores.

*XGBoost* has been shown to achieve state-of-art results on many machine learning tasks. It is an improvement of Gradient Boosting algorithm and a decision tree based on the gradient boosting algorithm. Li constructed an XGBoost-based model to predict peer-to-peer loan default and compare its outcome with Logistic Regression and Decision Trees. The results indicated that the accuracy of the XGBoost-based model achieved 97.705%, which fitted the actual results better.

## *6- HOW CAN WE HANDLE IMBALANCED CLASSES OF NON-DEFAULTS THAN DEFAULTS IN HE TARGET?*

**Techniques to Solve Class Imbalance**

*Up-sample Minority Class:* Up-sampling is the process of randomly duplicating observations from the minority class in order to reinforce its signal.There are several heuristics for doing so, but the most common way is to simply resample with replacement.

- First, we'll separate observations from each class into different Data Frames.
- Next, we'll resample the minority class with replacement, setting the number of samples to match that of the majority class.
- Finally, we'll combine the up-sampled minority class Data Frame with the original majority class Data Frame.

*Down-sample Majority Class:* Down-sampling involves randomly removing observations from the majority class to prevent its signal from dominating the learning algorithm. The most common heuristic for doing so is resampling without replacement. The process is similar to that of up-sampling. Here are the steps:

- First, we'll separate observations from each class into different Data Frames.

- Next, we'll resample the majority class without replacement, setting the number of samples to match that of the minority class.
- Finally, we'll combine the down-sampled majority class Data Frame with the original minority class Data Frame.

*Change Your Performance Metric:* Choose the right evaluation metric. For a general-purpose metric for classification, we recommend Area Under ROC Curve (AUROC).

- We won't dive into its details in this guide, but you can read more about AUROC here. Intuitively, AUROC represents the likelihood of your model distinguishing observations from two classes.
- In other words, if you randomly select one observation from each class, what's the probability that your model will be able to "rank" them correctly?

*Penalize Algorithms (Cost-Sensitive Training):* The next tactic is to use penalized learning algorithms that increase the cost of classification mistakes on the minority class. A popular algorithm for this technique is Penalized-SVM. To really determine which of these tactics works best for this problem, you'd want to evaluate the models on a hold-out test set.

*Use Tree-Based Algorithms:* Decision trees often perform well on imbalanced datasets because their hierarchical structure allows them to learn signals from both classes.In modern applied machine learning, tree ensembles (Random Forests, Gradient Boosted Trees, etc.) almost always outperform singular decision trees. Well, tree ensembles have become very popular because they perform extremely well on many real-world problems.

## *7- WHAT PREPROCESSING STEPS ARE NECESSARY FOR DATA CLEANING, FEATURE ENGINEERING AND FEATURE SELECTION?*

**What is data cleaning?**

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.

*Step 1: Remove duplicate or irrelevant observations* Remove unwanted observations from your dataset, including duplicate observations or irrelevant observations. Duplicate observations will happen most often during data collection. When you combine data sets from multiple places, scrape data, or receive data from clients or multiple departments, there are opportunities to create duplicate data. De-duplication is one of the largest areas to be considered in this process. Irrelevant observations are when you notice observations that do not fit into the specific problem you are trying to analyze.

*Step 2: Fix structural errors* Structural errors are when you measure or transfer data and notice strange naming conventions, typos, or incorrect capitalization. These

inconsistencies can cause mislabeled categories or classes. For example, you may find “N/A” and “Not Applicable” both appear, but they should be analyzed as the same category.

*Step 3: Filter unwanted outliers* Often, there will be one-off observations where, at a glance, they do not appear to fit within the data you are analyzing. If you have a legitimate reason to remove an outlier, like improper data-entry, doing so will help the performance of the data you are working with. However, sometimes it is the appearance of an outlier that will prove a theory you are working on. Remember: just because an outlier exists, doesn’t mean it is incorrect. This step is needed to determine the validity of that number. If an outlier proves to be irrelevant for analysis or is a mistake, consider removing it.

*Step 4: Handle missing data* You can’t ignore missing data because many algorithms will not accept missing values. There are a couple of ways to deal with missing data. Neither is optimal, but both can be considered.

- As a first option, you can drop observations that have missing values, but doing this will drop or lose information, so be mindful of this before you remove it.
- As a second option, you can input missing values based on other observations; again, there is an opportunity to lose integrity of the data because you may be operating from assumptions and not actual observations.
- As a third option, you might alter the way the data is used to effectively navigate null values.

*Step 5: Validate and QA* at the end of the data cleaning process, you should be able to answer these questions as a part of basic validation:

- Does the data make sense?
- Does the data follow the appropriate rules for its field?
- Does it prove or disprove your working theory, or bring any insight to light?
- Can you find trends in the data to help you form your next theory?
- If not, is that because of a data quality issue?

**What is Feature Engineering?**

The feature engineering pipeline is the preprocessing steps that transform raw data into features that can be used in machine learning algorithms, such as predictive models.

**Steps in Feature Engineering**

Steps for how to perform feature engineering for most machine learning algorithms include the following:

*Data Preparation:* This preprocessing step involves the manipulation and consolidation of raw data from different sources into a standardized format so that it can be used in a model. Data preparation may entail data augmentation, cleaning, delivery, fusion, ingestion, and/or loading.

*Exploratory Analysis:* This step is used to identify and summarize the main characteristics in a data set through data analysis and investigation. Data science experts use data visualizations to better understand how best to manipulate data sources, to determine which statistical techniques are most appropriate for data analysis, and for choosing the right features for a model.

*Benchmark:* Benchmarking is setting a baseline standard for accuracy to which all variables are compared. This is done to reduce the rate of error and improve a model's predictability. Experimentation, testing and optimizing metrics for benchmarking is performed by data scientists with domain expertise and business users.

**What Is Feature Selection?**

The goal of feature selection techniques in machine learning is to find the best set of features that allows one to build optimized models of studied phenomena.

A Typical step in feature selection include

- understanding the dataset and problem
- choosing a relevant feature selection method
- evaluating feature importance
- selecting a subset of features
- assessing and validating model performance with the chosen features.

## *8- WHAT EVALUATION METRICS ARE APPROPRIATE FOR ASSESSING THE MODEL'S PERFORMANCE?*

Evaluation metrics are tied to machine learning tasks. There are different metrics for the tasks of classification and regression. Some metrics, like precision-recall, are useful for multiple tasks. Classification and regression are examples of supervised learning, which constitutes a majority of machine learning applications.

*Accuracy:* Accuracy simply measures how often the classifier correctly predicts. We can define accuracy as the ratio of the number of correct predictions and the total number of predictions.

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

*Precision: It* explains how many of the correctly predicted cases actually turned out to be positive. Precision is useful in the cases where False Positive is a higher concern than False Negatives. The importance of Precision is in music or video recommendation systems, e-commerce websites, etc. where wrong results could lead to customer churn and this could be harmful to the business.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

*Recall (Sensitivity):* It explains how many of the actual positive cases we were able to predict correctly with our model. Recall is a useful metric in cases where False Negative is of higher concern than False Positive. It is important in medical cases where it doesn't matter whether we raise a false alarm but the actual positive cases should not go undetected.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

*F1 Score:* *It gives a combined idea about Precision and Recall metrics. It is maximum when Precision is equal to Recall. F1 Score is the harmonic mean of precision and recall.*

$$F1 = 2.\frac{Precision \times Recall}{Precision + Recall}$$

*AUC-ROC:* The Receiver Operator Characteristic (ROC) is a probability curve that plots the TPR (True Positive Rate) against the FPR (False Positive Rate) at various threshold values and separates the 'signal' from the 'noise'. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes. From the graph, we simply say the area of the curve ABDE and the X and Y-axis.

## *10- HOW DO DEMOGRAPHIC VARIABLES (e.g AGE, INCOME, EDUCATION) AFFECT THE LIKELIHOOD OF DEFAULT?*

In the context of credit risk assessment for loan approval, demographic variables such as age, income, and education can play a significant role in determining the likelihood of default.

*1. Age:* Younger individuals may have less stable financial situations and less established credit histories, making them riskier borrowers. On the other hand, older individuals may have more stable incomes and assets, reducing their risk of default.

*2. Income:* Higher income levels generally indicate a greater ability to repay loans, lowering the risk of default. However, high income alone may not be a reliable indicator, as it does not consider other factors such as debt-to-income ratio and stability of income.

*3. Education:* Education level can be correlated with income and financial literacy, which can impact a borrower's ability to manage debt responsibly. Higher education levels may indicate a lower risk of default, but this relationship can vary based on other factors.

It's important to note that these variables are not determinants of default on their own, but rather factors that can influence an individual's overall credit risk. When building a machine learning model for credit risk assessment, it's crucial to consider a wide range of features, including demographic variables, to develop a comprehensive risk assessment strategy.