**Optimizing Loan Decision-Making: A Comparative Analysis of Predictive Models**

**Sheetal Patangay**

School of Business Administration, The Pennsylvania State University

BUS 510: Business Analytics and Decision Modeling

Dinesh R. Pai, Ph.D.

|  |
| --- |
| **Table of Contents:** |
| 1. **Introduction** |
| 1. **Dataset Overview and Preprocessing** |
| 1. **Descriptive Statistics** |
| 1. **Visualization** |
| 1. **Multicollinearity** |
| 1. **Hypothesis Testing** |
| 1. **Various Analysis Models Considered**   ***6.1. Logistic Regression Model and Results***  ***6.2. Discriminant Analysis Model and Results***  ***6.3. Neural Network Analysis Model and Results*** |
| 1. **Comparison of Models** |
| 1. **Key Findings and Insights** |
| 1. **Conclusion and Lessons Learned** |

**INTRODUCTION:**

Access to financial resources is a cornerstone for individual and economic growth, and the process of loan approval is pivotal in facilitating this access. Traditional methods of evaluating loan applications are often time-consuming, subjective, and susceptible to human biases. Machine learning (ML) (IBM, 2023a) is a revolutionary field that comes under Artificial Intelligence, ML enables computers to learn from data and make intelligent decisions or predictions without explicit programming. It is based on the notion that systems are capable of autonomously learning from experience and becoming better at it, which enables them to change and adapt to new knowledge. The power of algorithms to extract patterns and insights from data lies at the heart of machine learning. These algorithms are able to generalize and forecast fresh, unknown data because they can find hidden structures, correlations, and trends within datasets.

**DATASET OVERVIEW AND PREPROCESSING:**

1. **Dataset:**

The dataset used for this study is obtained from Kaggle about L&T Financial services. The loan approval dataset is a compilation of financial information and related data that's used to assess a person's or an organization's eligibility for a loan from a lending institution. It takes into account several variables, including assets worth, loan status, loan length, loan amount, income, and employment status. This dataset is used to create models and algorithms that forecast the chance of loan acceptance based on the provided parameters.

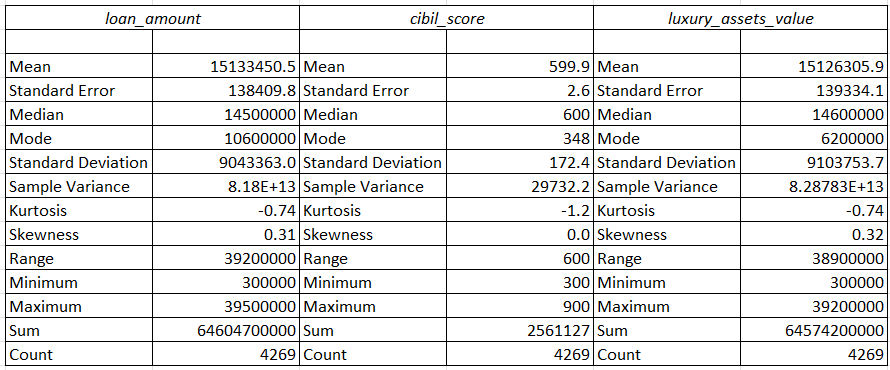
The dataset contains 4269 financial records, each of which has 13 columns, the last column ‘loan\_status’ is the target label that we want to predict. The possible values for this attribute are ‘Approved’ and ‘Not Approved’ hence this is a binary classification problem. The 13 columns are: loan\_id, no\_of\_dependents, education, self\_employed, income\_annum, loan\_amount, loan\_term, cibil\_score, residential\_assets\_value, commercial\_assets\_value, luxury\_assets\_value, bank\_asset\_value and loan\_status.

1. **Data preprocessing:**
2. First, we drop the ‘loan\_id’ column as it is not relevant to classification.
3. We check the dataset for any missing or NaN values, the dataset does not contain any missing values.
4. The columns: education, self\_employed and loan\_status have categorical values we encode these values as follows:

|  |  |  |
| --- | --- | --- |
| **Education** | **Graduated: 1** | **Not Graduated: 0** |
| **self\_employed** | Yes: 1 | No: 0 |
| **loan\_status** | Approved: 1 | Not Approved: 0 |

1. We split our dataset into training (70%) and test (30%) sets. We use the stratified split to ensure that both sets have sufficient data for both classes.

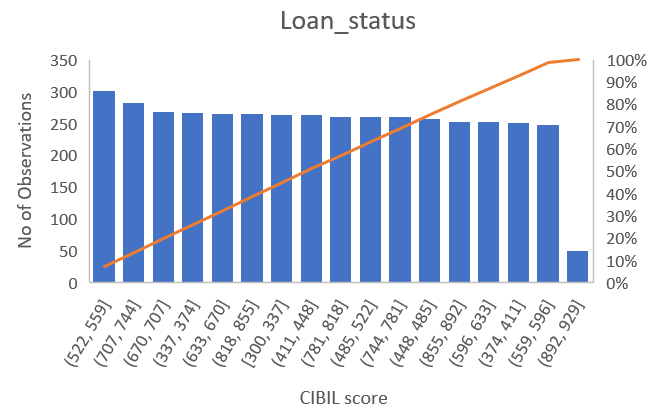
**DESCRIPTIVE STATISTICS:**



These are some of the variable’s descriptive analysis. Positive skewness in both loan amount and luxury assets value suggests a potential presence of higher values. The negative kurtosis indicates a flatter distribution compared to a normal distribution for all three variables. Further exploration, especially regarding potential outliers and inter-variable relationships, is warranted for a comprehensive understanding.

**VISUALISATION:**

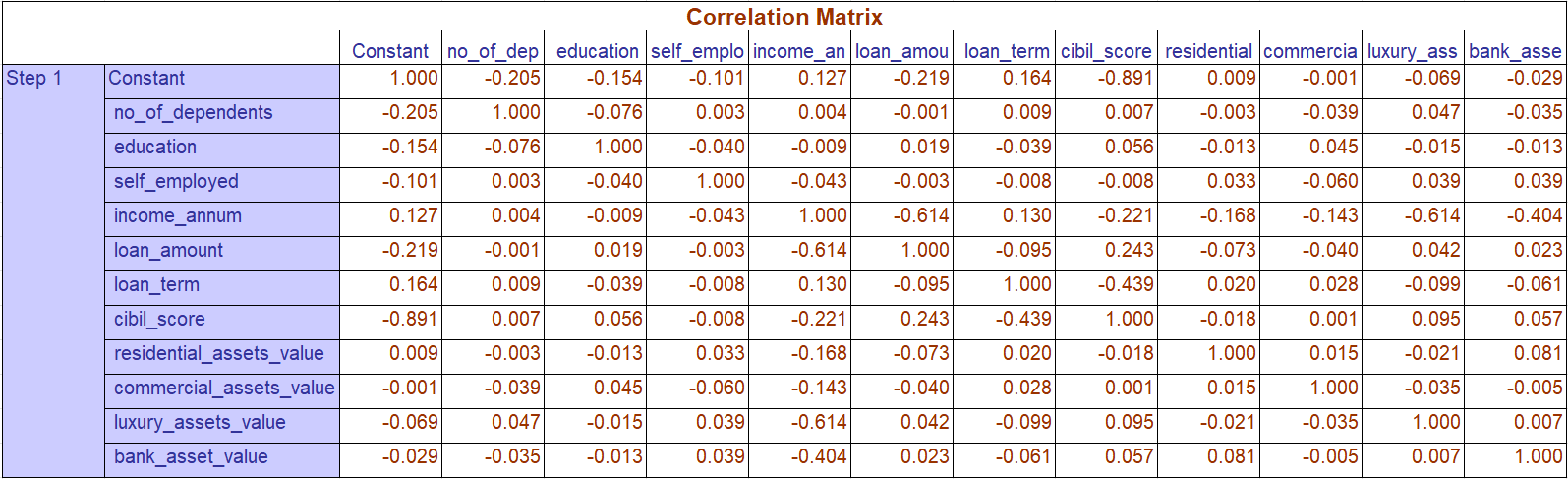
Histogram plot for Loan Status vs CIBIL score:



The chart shows the number of loan entities by their CIBIL score. The CIBIL score is a credit score that is used in India to assess the creditworthiness of an individual. The higher the CIBIL score, the better the creditworthiness.

The chart shows that many loan entities (52%) have a CIBIL score between 707 and 744. This is followed by entities with a CIBIL score between 670 and 707 (28%), and entities with a CIBIL score between 744 and 781 (12%). The remaining 10% of entities have a CIBIL score below 707 or above 781. The chart also shows that there is a slight decrease in the number of loan entities as the CIBIL score increases. This is likely because lenders are more likely to approve loans to borrowers with higher CIBIL scores.

**MULTICOLLINEARITY:**

**Scenario:** We considered a financial institution, L&T Financial Institution, using a dataset of historical loan applications to implement a neural network model for predicting loan approval or rejection.

**Objective:**

L&T Financial Institution wants to leverage neural networks, to automate and improve the accuracy of loan approval decisions based on historical data.

**HYPOTHESIS TESTING:** To conduct hypothesis testing for the data, we need to formulate a hypothesis. The goal is to assess if there is enough evidence in the sample data to reject the null hypothesis in favor of the alternative hypothesis.

**Null Hypothesis (Ho):** There is no significant difference in the mean values of the numerical features (e.g., income\_annum, loan\_amount, loan\_term, etc.) between approved (loan\_status=1) and rejected (loan\_status=0) loan applications.

**Alternative Hypothesis (Ha):** There is a significant difference in the mean values of the numerical features between approved and denied loan applications.

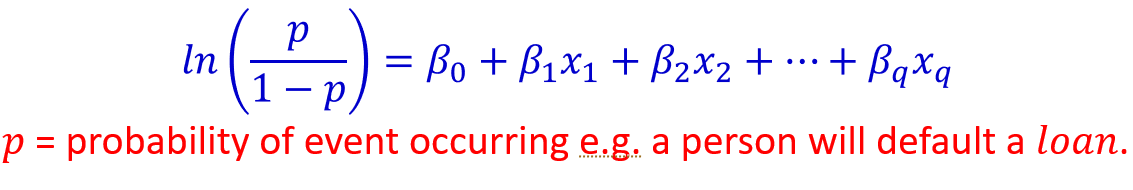
Since p-values are significant, we reject the null hypothesis in favor of the alternative hypothesis. This suggests that there is evidence to support the idea that there is a significant difference in the mean values of the numerical features between approved and denied loan applications.

**VARIOUS ANALYSIS MODELS CONSIDERED:**

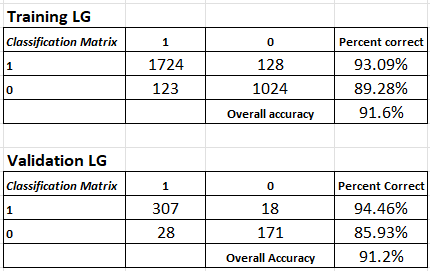
* **Logistic Regression:**

Logistic regression estimates the probability of an event occurring, such as voted or didn’t vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1.





In this logistic regression equation, z = ln(p/1-p) is the dependent or response variable and x is the independent variable. p/1-p is the odds ratio. The logistic regression model is trained on a historical dataset where loan outcomes are known. The model learns the relationships between the independent variables and the likelihood of loan approval. The performance of the model is then evaluated on a separate dataset using metrics such as accuracy, precision, recall, and the area under the ROC curve. Once the logistic regression model is trained and validated, it can be applied to new loan applications to predict the probability of approval. A screenshot of a calculator

Description automatically generated 

The logistic regression model exhibits promising performance, demonstrating strong predictive capabilities both during training and on unseen data in the validation set. The overall accuracy metrics for both training (91.6%) and validation (91.2%) suggest that the model is effectively capturing patterns in the data to make accurate predictions regarding loan approval. The classification matrices provide additional granularity, showing that the model is particularly adept at identifying instances with loan approval. However, there is a slight decrease in accuracy for instances with loan denial in the validation set. Continuous monitoring and potential model refinement could further enhance predictive accuracy and generalization to new data. Overall, the logistic regression model proves to be a reliable tool for making loan approval decisions based on the provided data.

* **Discriminant Analysis**

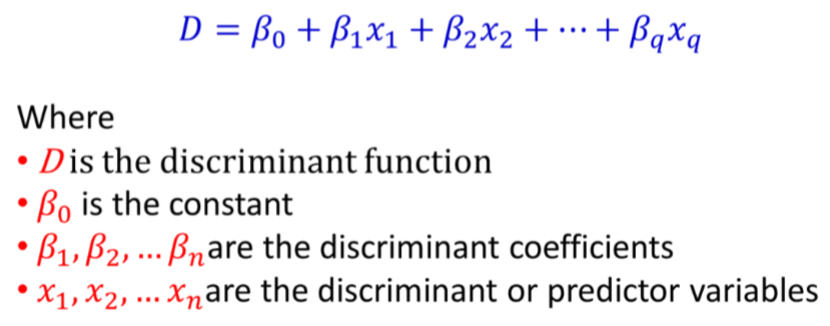
A discriminant analysis is used to compare effectiveness, understand profits and losses, and determine the best practices and recommendations to implement. Our goal is to improve the loan approval decision-making process in our scenario through an automated and data-driven approach. This analysis aims to improve loan processing efficiency, improve accuracy through identifying complex patterns, and ensure consistent and reliable decisions based on historical data by leveraging machine learning. Discriminant analysis can handle situations in which the dependent variable contains more than two classes. Logistic regression is one such method that has this limitation. Discriminant analysis is particularly useful for categorizing observations into predefined groups when the dependent variable is categorical. The method is good for problems that have more than two categories. The dependent variable in our data set has two categories: approved and rejected. With small datasets, discriminant analysis may perform better than more complex models like neural networks, which are more prone to overfitting.

**Model Architecture:**

For discriminant analysis, we divided the data set into training (70%) and validation (30%) datasets. We categorized the dependent variable "loan\_status" as 1's and 0's (1 for approved and 0 for rejected). With SPSS, we then conducted discriminant analysis on the training dataset. Using these coefficients, we calculated the discriminant function for the validation datasets of L (0) and L (1) as shown below.

A white sheet with black text and numbers

Description automatically generated



Using this discriminant function, we calculated the columns Prediction, Accuracy (1), Accuracy (0), Incorrect (1), Incorrect (0). Based on these columns, we calculated the classification matrix and determined its efficiency.

**Classification Matrix of DA:**

A screenshot of a computer

Description automatically generated

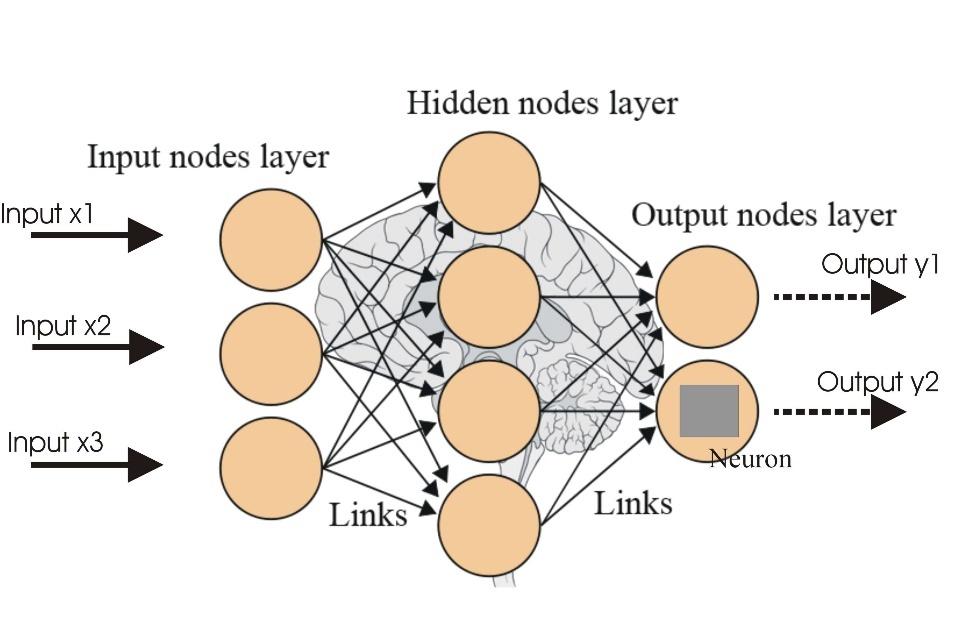
A screenshot of a computer

Description automatically generated

The model achieved relatively high accuracy on both training and validation datasets. An overall accuracy of 91.7% on the training dataset and 91.8% on the validation dataset suggests that your model performs well on the data.

* **Neural Network Analysis:**

Neural network analysis involves using artificial neural networks, inspired by the human brain, to analyze and interpret complex data patterns. In simple terms, it's like teaching a computer to recognize and understand information, much like how our brains process and learn from experiences. What makes neural network analysis unique is its ability to automatically identify intricate patterns and relationships within data, allowing it to make predictions or classifications without explicit programming. This method stands out from traditional analysis models because it excels at handling unstructured and large datasets, making it particularly powerful in tasks such as image and speech recognition, language processing, and complex decision-making, where conventional approaches may fall short. In essence, neural network analysis harnesses the power of artificial intelligence to unlock insights from data in a way that mirrors human cognitive processes.

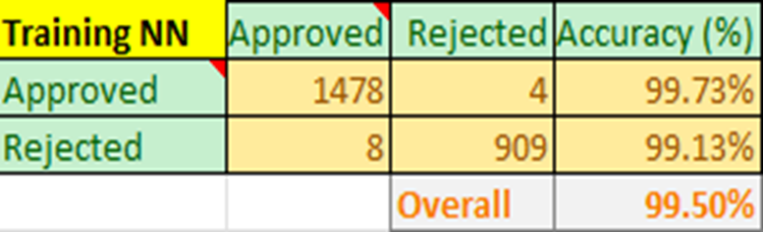


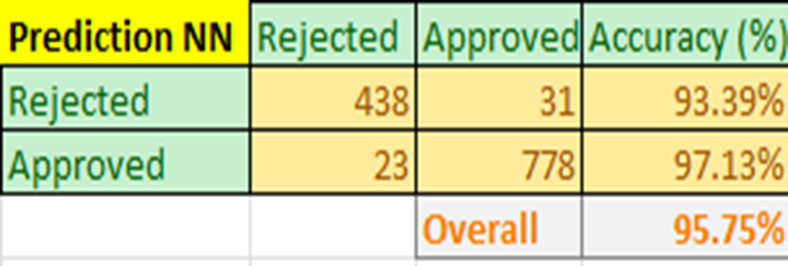
**Model Architecture:**

In our neural network analysis for loan approval, we initially divided the dataset into training (first 3000 rows) and validation sets. As we utilized neural network software in Excel, we did not explicitly encode categorical variables, relying on the neural network's ability to handle them implicitly. The resulting classification matrices and overall accuracy percentages were derived from this approach, optimizing the model for both training and validation datasets.

**Classification Matrix of NN:**

The classification matrices for the training and validation datasets provide a comprehensive view of the neural network's performance in predicting loan approval and rejection.





- ***Training Dataset*** ***Precision and Recall***: For the "Approved" class, precision is exceptionally high at 99.73%, indicating a low false positive rate. The recall for "Rejected" is 99.13%, demonstrating a low false negative rate.

- ***Validation Dataset*** ***Precision and Recall:*** Notably, the model exhibited a precision of 97.13% for the "Approved" class, signifying a low false positive rate. The "Rejected" class demonstrated a recall of 93.39%, indicating a low false negative rate.

**COMPARISON OF MODELS:**

|  |  |  |
| --- | --- | --- |
| Model | Training Accuracy | Validation Accuracy |
| Discriminant Analysis | 91.70% | 91.80% |
| Logistic Regression | 91.60% | 91.20% |
| Neural Network | 99.50% | 95.75% |

**Interpretation:**

Based on the overall accuracy matrix for the loan approval and rejection data analysis, the neural network model emerges as the most appropriate choice for the loan approval analysis. The training accuracy of the neural network is exceptionally high at 99.50%, indicating a robust ability to learn from the provided data. Additionally, the validation accuracy stands at 95.75%, demonstrating the model's effectiveness in generalizing to new data. In comparison, both logistic regression and discriminant analysis show slightly lower accuracies, with logistic regression at 91.60% and 91.20%, and discriminant analysis at 91.70% and 91.80% for training and validation, respectively. The significantly higher accuracy of the neural network, especially during training, suggests that it has successfully captured the underlying patterns in the data, making it the preferred model for making accurate and reliable loan approval predictions.

**KEY FINDING AND INSIGHTS:**

The neural network analysis for loan approval yielded several key insights that are essential for understanding the decision-making process and optimizing the lending system. Here are the summarized key insights:

**1. Dominance of CIBIL Score:** - The analysis reaffirmed the significant impact of the CIBIL score on loan approval decisions. A higher credit score consistently emerged as a strong predictor of loan approval, underlining the importance of creditworthiness in the model's decision-making.

**2. Income Annual and Loan Amount Interaction:** - An interesting finding was the nuanced relationship between annual income and the requested loan amount. While a higher income generally favored loan approval, there were instances where a proportionately large loan amount in relation to income influenced the model's decision, suggesting a careful consideration of the balance between income and loan size.

**3. Unexpected Influence of Luxury Assets:** - The presence of luxury assets was found to have a more pronounced impact than anticipated. Applicants with substantial luxury assets were often more favorably considered for loan approval, indicating a potential additional factor contributing to the decision-making process.

**4. Self-Employment Dynamics:** - The analysis revealed complexities in assessing the self-employment status of applicants. While being self-employed generally did not hinder loan approval, the nature and stability of self-employment emerged as critical factors, with certain self-employed individuals exhibiting strong creditworthiness.

**5. Education as a Decisive Factor:** - The educational background of applicants played a more decisive role than initially expected. Graduates consistently exhibited higher probabilities of loan approval, suggesting that education level is a relevant indicator of financial stability and reliability.

**6. Asset Values and Collateral Impact:** - Residential and commercial asset values were identified as strong indicators of financial stability and collateral. Higher values positively influenced loan approval decisions, emphasizing the importance of assessing an applicant's overall financial portfolio.

**CONCLUSION AND LESSONS LEARNED:**

In the world of financial services, the successful implementation of data mining techniques for loan approval prediction is a significant step towards enhancing efficiency, mitigating risks, and fostering financial inclusivity. This academic project explored the intricacies of building a predictive model that leverages historical data to assess and predict loan approval outcomes.

The potential impact of this project extends far beyond the confines of academic exploration. The deployment of the developed model into real-world scenarios has the capacity to streamline loan approval processes, reduce processing times, and, importantly, make lending decisions more equitable. The transparency and interpretability of the model provide insights into the factors influencing approval decisions, fostering trust among both lenders and loan applicants.

However, it is crucial to acknowledge the limitations of any predictive model and the dynamic nature of financial landscapes. Future improvements for this project can be the integration of new datasets into the model training and exploring more advanced techniques like deep learning.