UNVEILING THE FUTURE...

Predicting Home Loan Approvals using Machine Learning

Pedro Azpurua, Akif Hasan, Henry Leighton, Graham Meadon

Agenda

- Project Objective
- Data Sources
- Technology Used
- Data Cleaning and Machine Learning
- Demonstration
- Conclusion
- Limitations and future development

Project Objective

Our Project Objective is to develop a comprehensive loan eligibility model that considers both historical loan data and key features (Applicant Income, Co-applicant Income, Loan Amount, Loan Amount Term, Credit History) as well as demographics (Gender, Dependents, Marital Status, Education).

The purpose of the model is to promote financial inclusion and reduce biases, ensuring fair access to loans for individuals from diverse backgrounds.

Data Sources

- Primary data source: De-identified Loan Data
- Additional data source: Geeks for Geeks

Technology Used

Data management and visualisations using these tools:

- Scikit-learn – for machine learning
- Python Pandas – for cleaning the data
- SQLite– converting the data into database
  - *Dataset with 598 lines of data*
- Python Flask-powered API - a resting API for data access
- Python Matplotlib – for data visualisations
- HTML/CSS/Bootstrap – for user interface

Data Cleaning and Machine Learning

The dataset contains 13 features

1	Loan	A unique id
2	Gender	Gender of the applicant Male/female
3	Married	Marital Status of the applicant, values will be Yes/ No
4	Dependents	It tells whether the applicant has any dependents or not.
5	Education	It will tell us whether the applicant is Graduated or not.
6	Self_Employed	This defines that the applicant is self-employed i.e. Yes/ No
7	ApplicantIncome	Applicant income
8	CoapplicantIncome	Co-applicant income
9	LoanAmount	Loan amount (in thousands)
10	Loan_Amount_Term	Terms of loan (in months)
11	Credit_History	Credit history of individual's repayment of their debts
12	Property_Area	Area of property i.e. Rural/Urban/Semi-urban
13	Loan_Status	Status of Loan Approved or not i.e. Y- Yes, N-No

ETL Process:

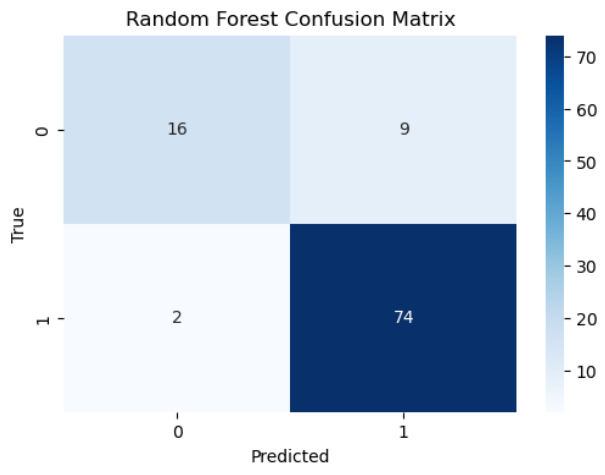
- Imported the data from SQL
- Dropped lines where we had empty values (NaN)
  - Reduced data sample from 598 to 505
- Dropped columns that were irrelevant
  - Loan_ID
- Converted categorical variables to one-hot encoding (0 or 1)
  - Gender, Married, Education, Self_Employed, Property_Area
- Exported to CSV for Machine Learning models
- The data was scaled before training and testing

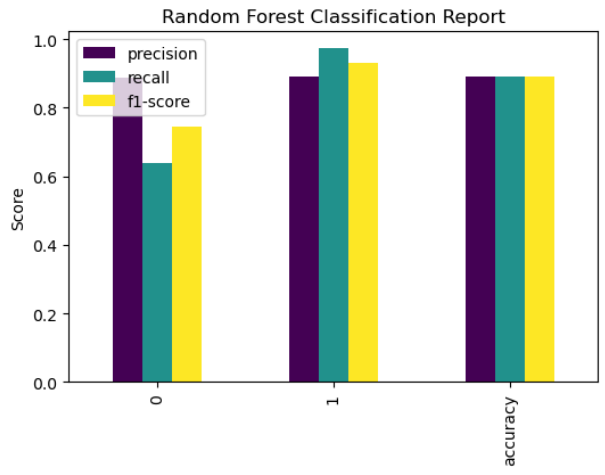
Data Cleaning and Machine Learning

We tested Supervised Learning models vs Neural Network for our Classification problem

Random Forest

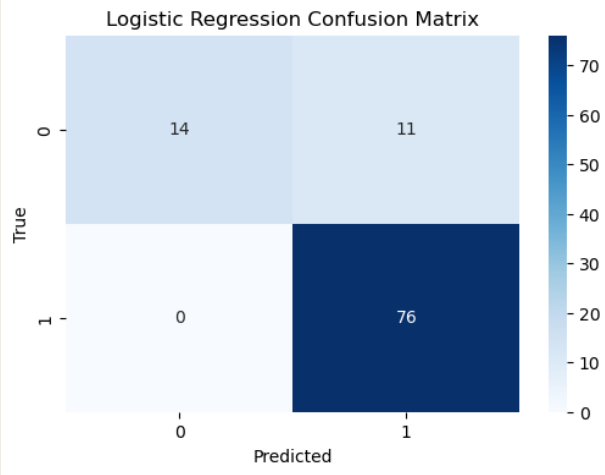
Model Accuracy: 89.11%

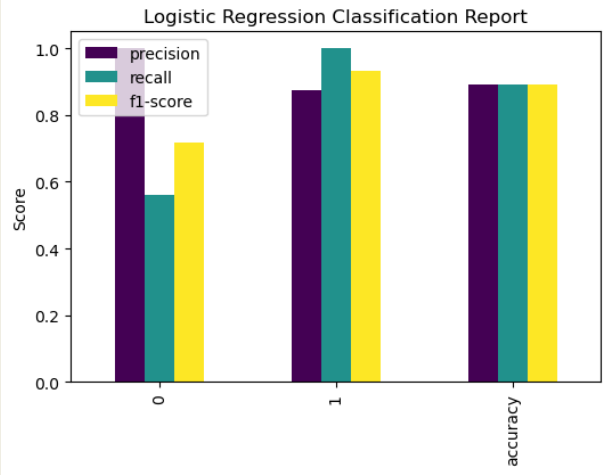




Logistic Regression

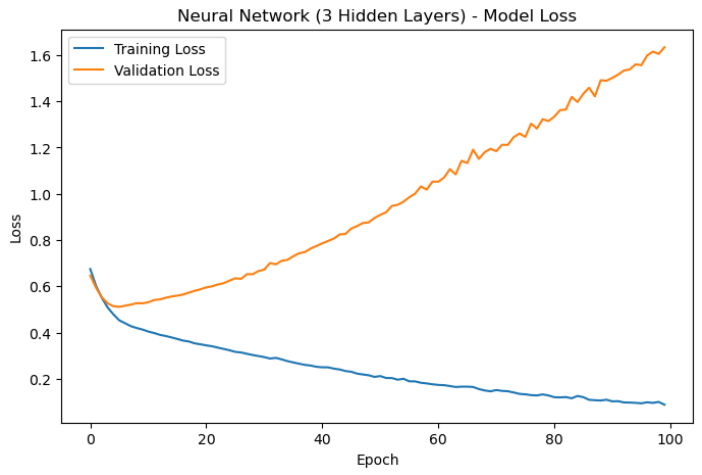
Model Accuracy: 89.11%

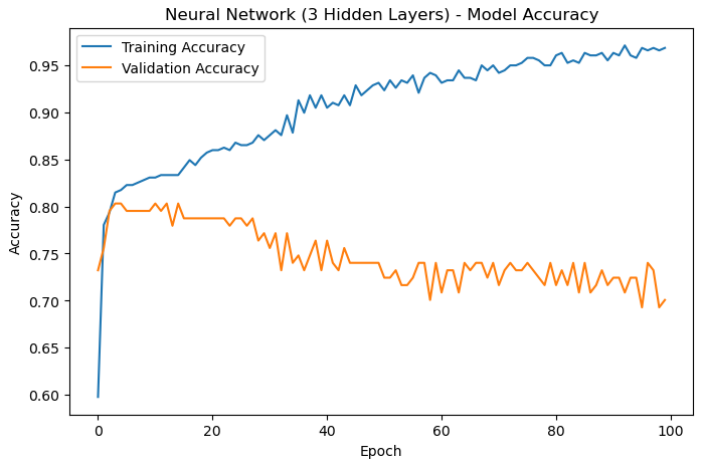




Neural Network

Model Accuracy: 70.08%





Data Cleaning and Machine Learning

Given the provided information and the trade-offs between the two models, we have selected the Random Forest model:

- **Balanced Performance:** The Random Forest model achieves balanced results for both classes, with precision and recall values more evenly distributed than in Logistic Regression. This suggests effective handling of both classes without excessive bias.

- **Improved Recall for class 0 (Denied)**: While the Random Forest model's recall for class 0 (denied) is slightly lower (0.64) than Logistic Regression, it maintains a reasonable level. This indicates successful capture of a significant portion of class 0 instances, contributing to overall balance.

- **Reduced Overfitting Risk**: Random Forest models incorporate strategies like multiple decision trees and random feature selection to mitigate overfitting. This enhances their ability to generalize to new data compared to Logistic Regression.

- **Versatility**: Random Forest models demonstrate adaptability by accommodating diverse data types and showing less sensitivity to outliers and feature scaling, unlike linear models such as Logistic Regression.

In conclusion, the Random Forest model offers balanced performance, lower overfitting risk, and versatility, making it a reliable option for achieving well-rounded and dependable classification outcomes.

DEMONSTRATION

Limitations and Future Development

Limitations:

- **Data Availability**: The accuracy of the model relies on the availability and quality of historical loan data.
- **Biases in Historical Data**: Biases in past lending practices may affect the fairness of predictions for underrepresented groups.
- **Complex Factors**: The model may not fully capture all factors influencing loan approval decisions.
- **Model Interpretability**: Some models lack transparency, making it difficult to explain predictions.

Future Development:

- **Inclusive Data Collection**: Collecting diverse and representative data to mitigate biases.
- **Fairness-aware Algorithms**: Developing algorithms to identify and address biases in predictions.
- **Explainable AI**: Integrating techniques to enhance the model's transparency.
- **Hybrid Models**: Combining different algorithms for more accurate predictions.
- **Real-time Data**: Utilizing real-time data for up-to-date loan approval decisions.
- **Ethical Frameworks**: Implementing ethical guidelines for responsible AI adoption.

Conclusion

Loan approval prediction using machine learning can improve the efficiency and accuracy of loan processing in financial institutions.

By building a robust predictive model that considers key features and demographics, banks can better assess the risk associated with loan applicants and make informed decisions about loan approvals.

This project demonstrates the potential of machine learning in the finance industry to streamline the loan approval process and promote financial inclusion.