Predictive Analysis for Financial Loan Approval

Danni Zhang
Fatima Mahmood
Romet Vislapuu

Link to GitHub repository: https://github.com/fatima97mahmood/h1_IDS

Business Understanding

Background

The financial sector is one of the most important parts of our modern economy. The world's governments and companies run on debt and so do modern households. With every loan, the risk of the debtor defaulting on the loan varies. This project aims to analyze synthetic data inspired by real-world credit and financial datasets to assess loan approval likelihood, focusing on loans granted to individuals.

Business goals

The business goal of this project is to be able to predict binary loan approval outcomes based on the Kaggle dataset to inform stakeholders of key predictors influencing loan decisions.

Business Success Criteria

For this project, we are aiming to predict binary loan approval outcomes with an accuracy above 90%.

Inventory of Resources

For our project, we are using the Loan Approval Classification Dataset on Kaggle. This dataset has 45,000 records and 14 features. These and other details will be further expanded upon in the Data Understanding section.

We will be using Jupyter Notebook with Python and various visualization libraries. Hardware wise we have three laptops.

Our team is made up of three highly motivated students who are all currently taking the Introduction to Data Science LTAT.02.002 course where we have learned about Data preprocessing, machine learning, visualization, and other techniques related to data science. Additionally, one of our group members has domain knowledge in economics.

Requirements, Assumptions, and Constraints

For the models we will produce to work in real-life scenarios, we must have our dataset represent realistic outcomes of the lending business.

We are constrained by our limited computational resources and the tight timeline presented by the course.

Risks and Contingencies

1. Missing the submission date due to technical difficulties.

– ***Mitigation***: Having an internal deadline before the submission one, to have time to fix any unforeseen issues that might have arisen.

The following issues were noted while exploring the data. See the Data Understanding chapter for further details.

2. Outliers or data quality issues.

– ***Mitigation****:* Cleaning and robust imputation methods.

3. Potential overfitting in models.

– ***Mitigation****:* Use of cross-validation and regularization.

Terminology

- ***Binary Classification****:* Loan approval prediction. If the person was approved or denied the loan.
- **Decision tree:** Supervised learning algorithm styled like a tree.
- **Random forest:** Combining multiple decision trees that is more accurate than a single decision tree.
- **Neural Networks**: Machine learning algorithm inspired by the structure of a brain.
- **XGBoost:** Machine learning algorithm popular on Kaggle.
- **Logistic regression:** Supervised machine learning algorithm that produces a probability value between 0 and 1.
- **Support vector machine**: Also known as SVM. Supervised machine learning algorithm.
- ***R^2*, R-Squared**: Statistical measure used to represent a regression model's goodness of fit.
- ***Hyperparameter tuning***: Hyperparameters are configuration variables that are set before the training process of a model begins. Tuning means finding the best hyperparameters for the desired result.

Costs and Benefits

As the requirements of the course constrain the size and complexity of the project, the costs are minor.

- **Costs**: Time for data preparation and modeling. Computational resources.
- **Benefits**: Practical insights for loan processing efficiency.

Data-Mining Goals

Creation of six classification models for loan status using the following six algorithms, optimized with hyperparameter tuning:

1. Decision tree
2. Random forest
3. Neural networks
4. XGBoost
5. Logistic regression
6. Support vector machine

Additionally, we hope to produce an analysis of feature importance in decision-making.

Data-Mining Success Criteria:

We measure our success by the two following measures:

- Predictive accuracy for binary classification above 90%.
- Regression model performance with R^2 above 0.75.

Data Understanding

Gathering Data

This dataset is a synthetic version inspired by the original Credit Risk dataset on Kaggle, enriched with additional variables derived from Financial Risk for Loan Approval data. It serves as a valuable resource for predicting the *loan status* of potential applicants (approved or not approved), encompassing demographic, financial, and credit-related features. With 45,000 observations and 13 features, all complete and without missing values, the dataset provides a solid foundation for detailed analysis and modeling.

Describing Data

The dataset incorporates a variety of numerical features, such as age, annual

income, years of employment experience, loan amount requested, loan interest rate, the proportion of income allocated to the loan, the length of credit history in years, and credit score. These quantitative attributes provide critical insights into an individual's financial behavior and creditworthiness. Additionally, categorical features like gender, education level, home ownership status, loan intent, and previous loan defaults contribute to understanding demographic and behavioral factors influencing credit risk.

Exploring Data and Verifying Data Quality

An examination of specific features reveals interesting patterns. In terms of education, a significant portion of applicants hold at least an undergraduate degree, with Bachelor's and Associate degrees being the most common qualifications. Gender distribution is slightly skewed, with male applicants accounting for 55.2% of the dataset. Regarding home ownership, most applicants are renters, while only a small fraction own homes outright. The dataset also reflects diverse motivations for loan applications, with education being the most common purpose, followed by medical expenses and entrepreneurial ventures. An examination of previous loan defaults reveals that the dataset captures a nearly even split between applicants with and without default histories, making it a valuable predictor for risk modeling.

Despite its comprehensiveness, the dataset includes certain inconsistencies and challenges. For example, unrealistic values such as an age of 144 years or 125 years of employment experience highlight the need for data cleaning. Additionally, the dataset is imbalanced, a factor that can affect model training. Techniques like Synthetic Minority Oversampling (SMOTE) may be required to address this imbalance and ensure fair model performance.

An initial exploration of the data indicates that most numerical features are not normally distributed. For these skewed variables, a combination of logistic transformations and the Interquartile Range (IQR) method was employed to manage outliers effectively. In contrast, for the normally distributed feature *loan interest rate*, outliers were identified and removed using the Z-score method with a threshold of three standard deviations. These preprocessing steps ensure the dataset is primed

for robust and reliable model development.

Planning the Project

This project focuses on data understanding, data preparation, feature engineering, and machine learning model development for loan approval prediction. The key steps include:

- **Data Preparation and Feature Engineering:** We will handle the outliers in variables such as age and experience through capping or transformation techniques. The skewed features will be addressed using log transformations where necessary. Further, we will encode the categorical variables using one-hot and label encoding.
- **Model Development**: After splitting the data into train and test sets, we will evaluate multiple machine learning models by assigning specific tasks to each team member.
- **Evaluation**: Finally, we will compare the trained models using metrics like AUC and ROC curves, F1-score, and Precision/Recall to account for imbalanced data. We will also account for training and validation losses to monitor for overfitting. We will refine the results based on the insights obtained and return to earlier steps if needed.

Machine Learning Models:

We will implement the following models with hyperparameter tuning:

1. **Decision Trees and Random Forest** (Romet):

Tune parameters like max_depth, min_samples_split, and n_estimators. The goal will be to focus on explainability and feature importance.

2. **Logistic Regression and SVM** (Danni):

Optimize regularization strength (c) in Logistic Regression and kernels (linear, rbf) in SVM for robust predictions.

3. **Neural Networks and XGBoost** (Fatima):

We will tune hidden layers, neurons, and learning rates for Neural Networks. Optimize learning_rate, max_depth, and n_estimators for XGBoost.

Feature Importance: We will identify critical predictors by employing feature importance scores from Random Forests and XGBoost. We will also use Recursive Feature Elimination (RFE) to rank variables by their impact on performance.

Task	**Hours (per member)**
Exploratory Data Analysis (EDA)	7
Feature Engineering	1
Model Training	12
Model Evaluation	6
Reporting and Presentation	4
Total	**30**