Proposal for a Credit Risk Prediction System

Group conceptual project

Group 27—Sana Sawant and Vedika Khandelwal

Table of Contents

Business Understanding

As a fin-tech start-up, we are developing a solution to reduce loan default risks using data-driven predictive analytics for financial institutions to benefit and integrate in their processes. As a Platform as a Service (PaaS) company, we are focused on providing our clients and lenders with actionable insights based on borrower data. This platform proposal will assist institutions evaluate risk more accurately and restructure their lending frameworks accordingly to reduce their default rate by identifying low-risk applications, hence improving financial security, profitability and compliance with credit risk management regulations.

The business objective is building a predictive model to estimate the likelihood of loan defaults, and categorize borrowers into risky or safe classifications, and identifying the significant variables impacting default probability, to enable proactive risk management. The project is framed as a classification task, with loan status as the target variable. Predictor variables include demographics, loan characteristics, credit history, and employment details. The scope covers the complete ETL process, including exploratory data analysis (EDA), data cleaning and transformation, model development, evaluation, deployment on the model into clients' loan approval systems, and feedback and monitoring.

The key stakeholders involved in the project are our data scientists, project managers, and business analysts, as well as the client's IT and risk management teams. We have made the assumption that our sample set fairly represents the broader trends and that borrowing patterns remain stable across time and demographics. Risks, such as missing data, overfitting, and misaligned goals, will be addressed through robust data validation and regular stakeholder reviews. This report is based on a small sample dataset for beta testing, the first MVP product development has been estimated with a budget of $40,000 covering data acquisition, personnel, software, and other costs.

Success will be measured by the model’s performance (AUC > 0.80), a 10% reduction in default rates within eighteen months, smooth integration of the model into operations, and positive feedback from stakeholders.

Data Understanding

The dataset provides a well-distributed and varied set of variables for predicting loan default, as it includes a mix of demographic (income, employment length, credit history), financial, and credit-related attributes (interest rate and loan amount) relevant to assessing borrower risk making it a robust and holistic framework for prediction.

However, there are some limitations to the dataset with respect to data quality. As discussed in further detail in the Data Preparation section, there are inconsistencies, anomalies, and missing values in the dataset. For example, columns such as emp length contain significant missing values, which if not resolved would lead to lower learning and model accuracy. Additionally, the dataset is imbalanced, with more non-default cases than default cases, potentially biasing predictions. These issues can be resolved in the Data Preparations section through data cleaning and EDA.

Certain attributes are likely more influential in predicting loan defaults. For example, loan-to-income ratio, credit history length, interest rate, and prior default records are directly tied to repayment behavior and are expected to carry significant weight. Others, like home ownership status, may have a more indirect impact as a predictive value, which proves helpful when used in combination with other variables. Prioritizing the most relevant data ensures better predictive accuracy and efficient resource allocation during data preparation.

While the cost of preparing this dataset is manageable, it could rise if additional external data, like macroeconomic indicators or credit bureau records, are included. To remain cost-efficient, efforts should focus on the attributes most aligned with the business problem. By prioritizing relevant variables, unnecessary expenses can be avoided while maintaining high predictive performance.

The dataset includes quantitative variables such as age, income, loan amount, interest rate, and credit history length for quantitative precision and qualitative variables, such as home ownership, loan intent, and loan grade for contextual analysis.

By carefully considering the strengths, limitations, and key characteristics of the dataset, this analysis is well-positioned to predict loan defaults accurately while staying aligned with the project's goals.

Data Preparation

Phase 1: Data Integration

In the first phase of data preprocessing, we focused on collecting, selecting, and integrating relevant data from the provided dataset. The dataset, which contained 32,581 rows and 12 columns, was loaded into a Pandas dataframe. The numerical and categorical variables included a mix of demographic information (e.g., person_age, person_income, person_home_ownership), loan-related details (e.g., loan_amnt, loan_int_rate, loan_status, loan_intent), and employment history (e.g., person_emp_length, cb_person_cred_hist_length). Only the relevant columns were retained, ensuring unnecessary data that could complicate the analysis was excluded.

Phase 2: Data Cleaning

The second phase focused on data cleaning, which was done to improve the quality of the data, eliminate inconsistencies and improve the data’s reliability. We conducted anomaly detection using visualizations to detect outliers and statistical techniques to eliminate them. The box plots highlighted the range of the dataset and the extreme values in each of the visualized variables such as person_age, person_income, and loan_amnt. To ensure the dataset's integrity, outliers were removed using the Interquartile Range (IQR) method, reducing the dataset size from 32,581 rows to 25,929 rows. Now that the outliers had been eliminated and the dataset held valid and reliable values, we then focused our attention to missing values identified in the following columns: person_emp_length (895 missing values) and loan_int_rate (3,116 missing values). We conducted a straightforward mean imputation method to preserve the dataset's overall distribution.

Phase 3: Data Transformation

We conducted feature engineering to develop two new features that would be strong predictors of credit risk default and improve the model performance. The first variable created is **Loan-to-Employment Length Ratio**, this ratio reflects the relationship between employment stability and loan amount, and the **Interest Rate-to-Loan Amount Ratio**, which captures the interest burden relative to the loan size. This adds further dimensions to the data proving useful for our predictive model. Categorical variables, person_home_ownership, loan_intent, and loan_grade were encoded to numerical values to allow for consistency and compatibility with the Random Forest.

Phase 4: Data Reduction

The final phase of preprocessing was partially achieved during the cleaning phase, where 6,652 rows were eliminated through anomaly analysis. A correlation analysis was performed to identify multicollinearity and redundant variables, however, all none were identified and all the variables were retained as valid potential predictors. In future steps, imbalances in the objective function can be tackled through selective sampling techniques like oversampling or undersampling can be applied. The data is now streamlined, reliable and holds valid predictive capabilities making it suitable for machine learning.

Descriptive Statistics

Dataset after pre-processing: 25,929 rows and 14 columns (including engineered features).

- person_age: Mean = 27.7 years, standard deviation = 6.3 years.
- person_income: Mean = $66,074.85, standard deviation = $61,983.12.
- loan_amnt: Mean = $9,589.37, standard deviation = $6,322.08.

Modeling

Model Selection: Random Forest for Credit Risk Prediction

When building a predictive model for credit risk, it is imperative to evaluate and compare the various machine learning models and identify the best fit for the dataset. In this case, Random Forest was chosen over simpler models like Logistic Regression, Linear Regression, and standalone Decision Trees, primarily due to the **skewed nature of the dataset** and its superior handling of non-linearity and relationship complexities.

Why Random Forest Is the Best Choice

1. Robustness to Skewed Data

As an ensemble learning method, Random Forest combines multiple Decision Trees, making it particularly immune to imbalanced datasets. Its process of subsetting and aggregating results allows it to tackle imbalance, missing values and outliers. Alternate models such as logistic regression use the log-odds of the target variable, making them vulnerable to an imbalanced class, and require explicit preprocessing to manage missing data and outliers. Decision Trees often overfit skewed data by focusing excessively on minimizing impurity without accounting for imbalance.

2. Non-Linear Relationships and Feature Interactions

The Credit risk dataset has complex, non-linear relationships and interactions between its different features, such as income, employment length, and loan amount. Random Forest captures these complexities by learning unique aspects of the data, and modeling the nonlinear interactions. Linear Regression on the other hand, produces overly simplistic outcomes resulting in low performance. Random Forest amplifies the capabilities of decision trees and overcomes its structural limitations to deliver comprehensive insights.

3. Overfitting Prevention

Random Forest incorporates such as bootstrap aggregation to tackle overfitting. Each tree is trained on a random subset of the data, and random subsets of features are used to split nodes. These techniques introduce randomness, ensuring that the model generalizes well to new, unseen data. In contrast, a single Decision Tree is highly prone to overfitting as it seeks to perfectly classify the whole training data. Logistic Regression, while less prone to overfitting, lacks the versatility to capture complex, non-linear patterns in data.

4. Feature Importance

Random Forest reliably develops a feature importance measure to identify the most impactful predictors. This capability is especially valuable in credit risk applications, where interpretability is essential for meeting regulatory and business requirements. Logistic Regression and Decision Trees indicate variable importance but struggle with non-linearity and overfitting, while Random Forest overcomes these obstacles to provide a more reliable and insightful feature importance score.

Model Design

The dataset was split into **70% training data** and **30% test data** to ensure a robust evaluation of the model on unseen data. **Model Choice:** A **Random Forest Classifier** was chosen with 100 estimators (n_estimators=100). By averaging the predictions of 100 trees, the model minimizes the impact of noise and outliers, resulting in stable and accurate performance. This combination of randomness and aggregation ensures reliable generalization to new data.

Evaluation

The Random Forest model for credit risk prediction demonstrated strong performance, achieving an **accuracy of 91.99%**, indicating reliable overall predictions. To evaluate its ability to distinguish between defaulters and non-defaulters, a **ROC-AUC score of 0.8148** was calculated, highlighting its robustness in classification. While these metrics reflect strong performance, a deeper analysis through a confusion matrix provided additional insights.

The confusion matrix revealed that the model correctly identified **880 defaulters** (true positives) and **5,414 non-defaulters** (true negatives). However, it also misclassified **497 defaulters** as non-defaulters (false negatives) and **51 non-defaulters** as defaulters (false positives). The model's low false positive rate ensures reliable identification of non-defaulters, but reducing false negatives remains crucial to mitigate financial risks associated with undetected defaulters.

The classification report showed a **precision of 0.95** for defaulters, meaning the model was highly accurate in identifying defaulters among predicted cases. However, the **recall for defaulters was 0.64**, indicating it identified only 64% of actual defaulters. For non-defaulters, the model achieved high precision (**0.92**) and recall (**0.99**), leading to an F1-score of **0.95** for this class.

From a business perspective, the model aligns well with the goal of reducing financial risk by reliably identifying, and categorizing applicants into defaulters and non-defaults. The insights provided by the confusion matrix and classification report also build stakeholder confidence in the model's decisions. To further enhance performance, recall for defaulters needs to be improved through hyperparameter tuning or balancing the dataset.

Deployment

Deploying the model on client systems, will require focus on transitioning it into production and ensuring its ongoing reliability. The model will be exported using tools like joblib or pickle and integrated with APIs for real-time loan application scoring. Containerization with tools like Docker can be assessed to ensure performance consistency across environments. Post-deployment, continuous metric monitoring is essential, while also retraining the model regularly to incorporate new data and address changes in borrower behavior or economic conditions. Automation tools such as Airflow streamline retraining schedules, can be integrated to ensure that the model remains current and effective without manual intervention and feedback loop with business teams should be defined to ensure sustainable operations.

Bibliography

1. Tse, L. (2020). *Credit Risk Dataset*. Kaggle.com. https://www.kaggle.com/datasets/laotse/credit-risk-dataset?resource=download
2. Huh, K. (2021, February 13). *Surviving in a random forest with imbalanced datasets*. Medium. https://medium.com/sfu-cspmp/surviving-in-a-random-forest-with-imbalanced-datasets-b98b963d52eb
3. Kreiger, JR. "Evaluating a Random Forest Model." *Medium*, Analytics Vidhya, 10 Sept. 2021, medium.com/analytics-vidhya/evaluating-a-random-forest-model-9d165595ad56.
4. sklearn.ensemble.RandomForestClassifier — scikit-learn 0.20.3 documentation. (2018). Scikit-Learn.org.https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
5. Ibm. (2024, August 9). Containerization. What is containerization? https://www.ibm.com/topics/containerization
6. Srivastava, T. (2024, October 21). *12 Important model evaluation Metrics for Machine Learning Everyone should know (Updated 2024)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/
7. W3Schools.com. (n.d.). https://www.w3schools.com/python/pandas/pandas_cleaning_empty_cells.asp

Appendix

1. Credit risk Prediction Model Sample System

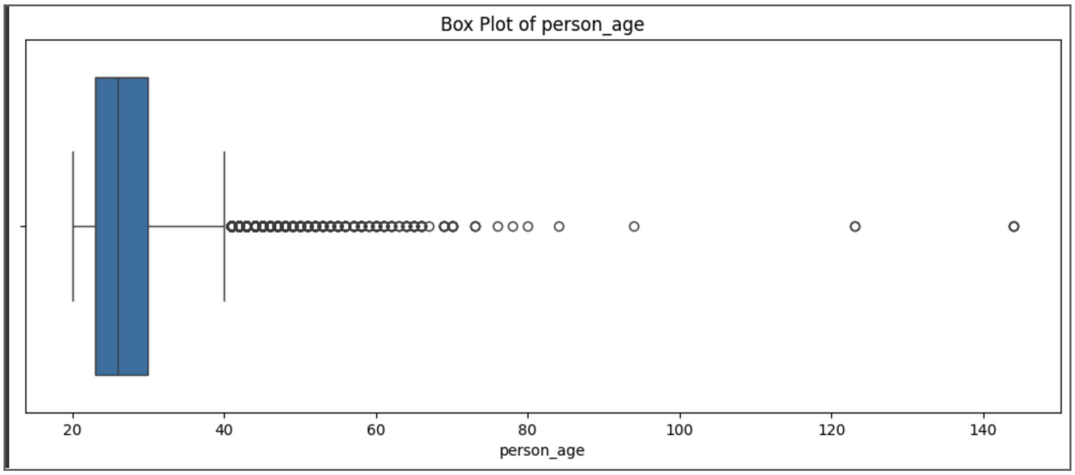
   https://colab.research.google.com/drive/1e7z69XH9djLoBGNhTRy06JVUIZHpTpS6?usp=sharing

2. Dataset: credit_risk_dataset (Upload this csv on the .ipynb file above and run the file to see a prototype of the model system).

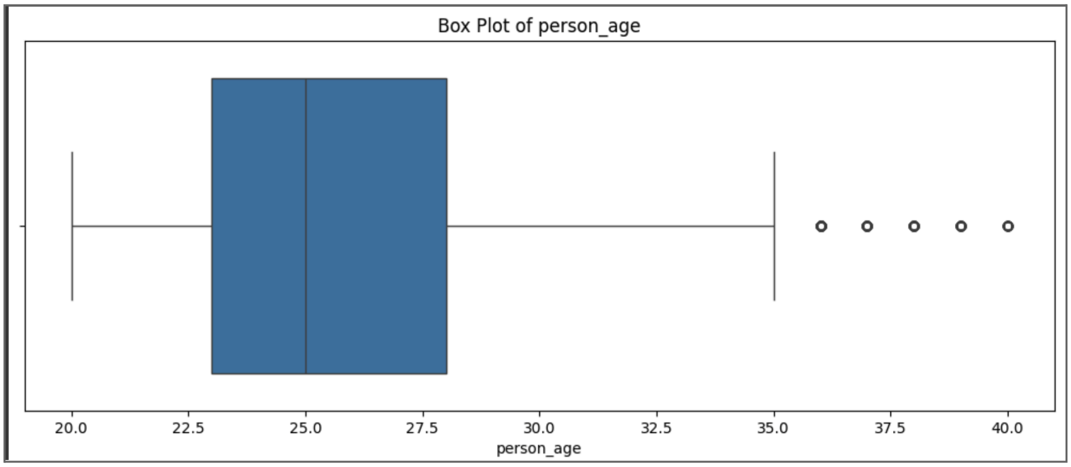
   https://docs.google.com/spreadsheets/d/1wX3_9uzGUZcm2MoreYluiXFS9Z4zPNQpUMEtAXCY4NQ/edit?usp=sharin

3. The Box Plot for each variable is shared in the code file above, some of the more significant box plots have been displayed below.

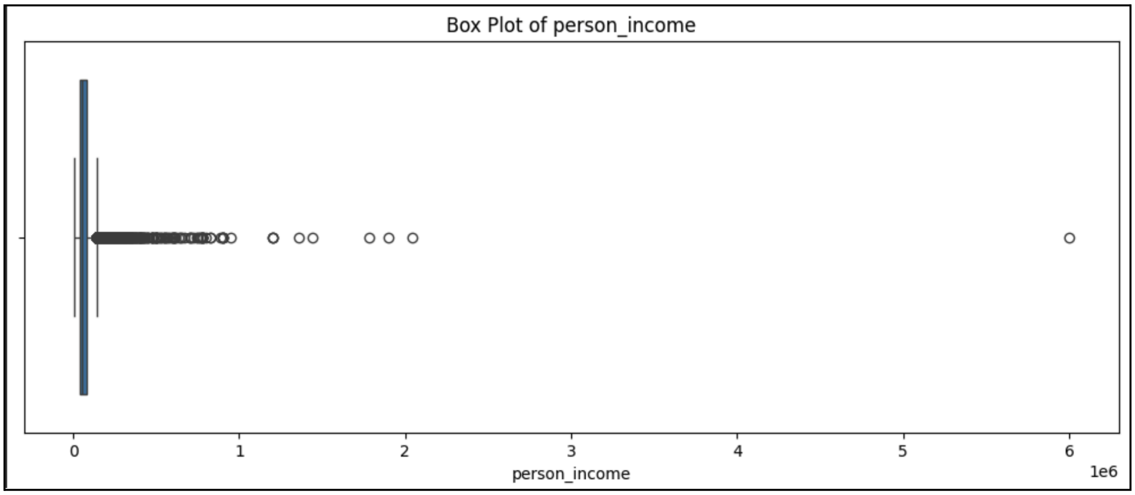
Person Age before Cleaning



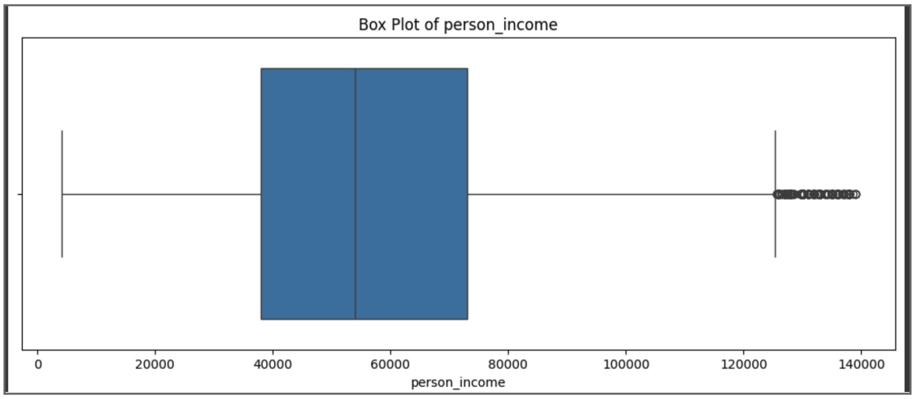
Box Plot for Person Age after Cleaning



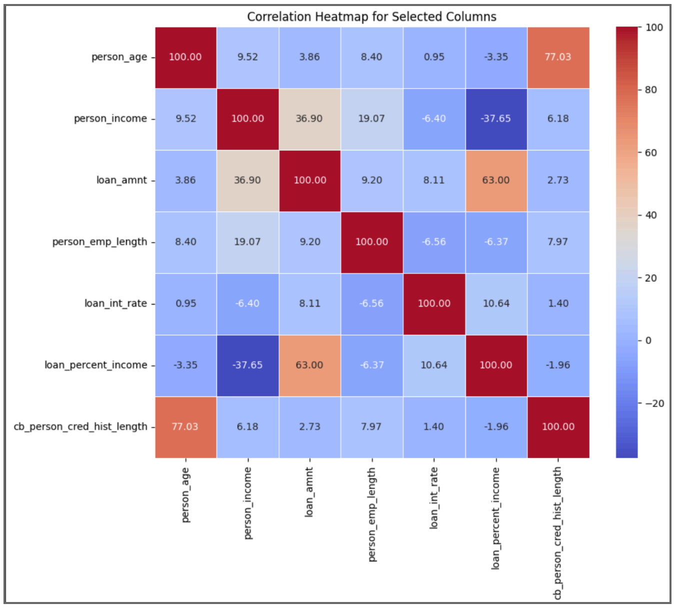
Box plot of personal income before cleaning



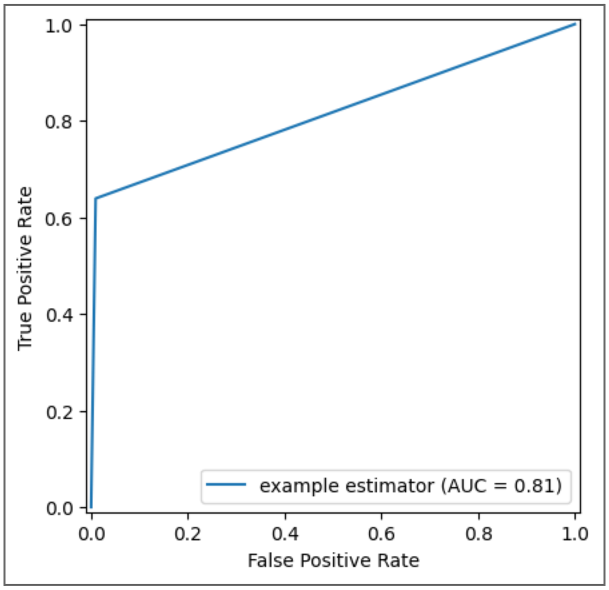
Box plot of personal income after cleaning



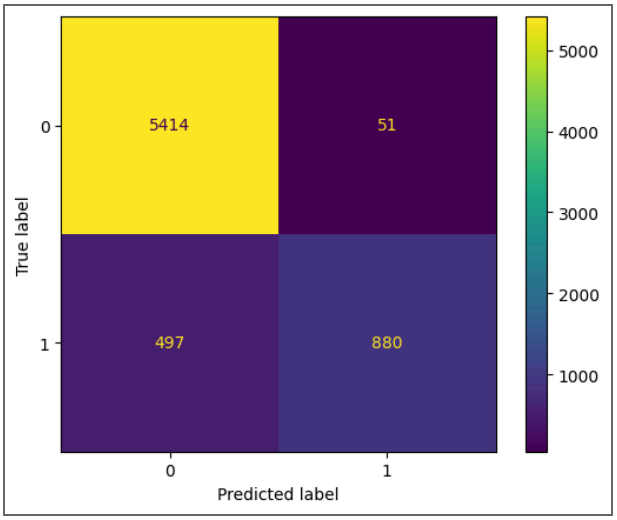
4. Correlation matrix



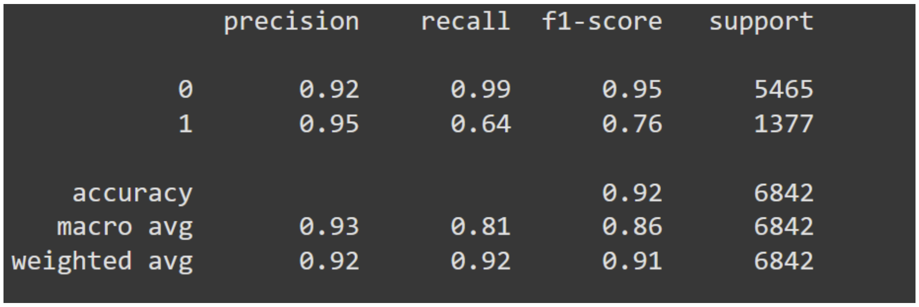
5. AUC-ROC Curve



6. Confusion Matrix



7. Classification and Performance Report



```
              precision    recall  f1-score   support

           0       0.92      0.99      0.95      5465
           1       0.95      0.64      0.76      1377

    accuracy                           0.92      6842
   macro avg       0.93      0.81      0.86      6842
weighted avg       0.92      0.92      0.91      6842
```