2025

KMUTT Python: Final Project

DATA ANALYSE - ML

PORCHER LENNY, PUPAT LUCAS, MOUSSAVI CAMBYSE

Table of contents

I. Problem Definition

The objective of this project is to build a predictive model capable of determining whether a bank loan will be approved or not, based on the applicant's profile. Accurately predicting loan approval helps financial institutions minimize credit risk and optimize their decision-making processes when evaluating loan applications.

II. Dataset Used

Title: BCR - Loaners Data for Solvency Prediction

Source: Kaggle Dataset

This dataset provides anonymized data on past loan applicants, along with their respective loan approval outcomes. It serves as a solid foundation for training and evaluating machine learning models for credit risk assessment and solvency prediction.

Choosing this dataset was a challenge for us because Kaggle provided no definitions for the columns and no example code. We had to explore the dataset in depth without knowing whether the available data would be sufficient to train our future model.

A) Dataset Description

The dataset includes multiple records, each representing a unique loan application, along with attributes that describe the financial and demographic characteristics of the applicant. The target variable indicates whether the loan was granted or not.

B) Features Overview

Below is a description of each 42 columns included in the dataset:

Column Name	Description
Unnamed: 0	Index column from the original file, not relevant for analysis.
SK_ID_CURR	Unique identifier for each loan applicant.
NAME_CONTRACT_TYPE	Type of loan contract requested (Cash loans, etc.).
CODE_GENDER	Gender of the applicant (M = Male, F = Female).
FLAG_OWN_CAR	Indicates if the applicant owns a car (Y = Yes, N = No).
FLAG_OWN_REALTY	Indicates if the applicant owns real estate property (Y = Yes, N = No).
CNT_CHILDREN	Number of children or dependents.
AMT_INCOME_TOTAL	Applicant's total income.
AMT_CREDIT	Total credit amount of the loan requested.

AMT_ANNUITY	Loan annuity amount — the expected periodic (e.g., monthly) payment.
AMT_GOODS_PRICE	Price of the goods for which the loan is given.
NUM_ANNUITY	Number of annuity payments.
CREDIT_SHARE	Ratio of the loan to the income or other credit-related metric.
LOANER_AGE	Age of the applicant.
YEARS_EMPLOYED	Number of years the applicant has been employed.
REVENUE_CLASS	Income class (encoded numerically).
NAME_TYPE_SUITE	Who was accompanying the applicant when applying (e.g., Unaccompanied, Family, etc.).
NAME_INCOME_TYPE	Source of income (Working, Pensioner, Businessman, etc.).
NAME_EDUCATION_TYPE	Highest education level attained.
NAME_FAMILY_STATUS	Marital/family status (Married, Single, Widow, etc.).
NAME_HOUSING_TYPE	Type of housing (House / apartment, With parents, Rented apartment, etc.).
REGION_POPULATION_RELATIVE	Represents the relative population density of the applicant's region compared to the most populated region in the dataset.
DAYS_REGISTRATION	Number of days since the applicant registered in the system (negative value indicates days ago).
DAYS_ID_PUBLISH	Days since the applicant’s ID was last changed or updated.
FLAG_MOBIL	Whether the applicant provided a mobile number (1 = Yes, 0 = No).
FLAG_EMP_PHONE	Whether the applicant provided a work phone number.
FLAG_WORK_PHONE	Indicates if the applicant has a work phone.
FLAG_CONT_MOBILE	Whether the mobile phone number is reachable.
FLAG_PHONE	Whether the applicant provided a home phone number.
FLAG_EMAIL	Whether the applicant provided an email address.
CNT_FAM_MEMBERS	Number of family members.
REGION_RATING_CLIENT	Rating of the region where the client lives (by external source).
REGION_RATING_CLIENT_W_CITY	Regional rating taking into account the city.

WEEKDAY_APPR_PROCESS_START	Day of the week when the application process started.
HOUR_APPR_PROCESS_START	Hour of the day when the application process started.
REG_REGION_NOT_LIVE_REGION	Whether the applicant's registered region differs from the living region.
REG_REGION_NOT_WORK_REGION	Whether the applicant's registered region differs from the working region.
LIVE_REGION_NOT_WORK_REGION	Whether the living region differs from the working region.
REG_CITY_NOT_LIVE_CITY	Whether the registered city is different from the living city.
REG_CITY_NOT_WORK_CITY	Whether the registered city is different from the working city.
LIVE_CITY_NOT_WORK_CITY	Whether the living city is different from the working city.
TARGET	Target variable: 0 if the loan was approved, 1 if it was rejected (i.e., default or risk detected).

III. Dataset Analysis

A) Preprocessing

After importing the CSV file containing 307,493 rows and 42 columns, our first step was to become familiar with the dataset and understand the nature of the data it contains.

We began by identifying the purpose of each column. To do this, we determined the type of data (qualitative, quantitative, nominal, discrete, etc.) and analyzed the relationship between the column names and their corresponding data types. A few column interpretations were assisted by ChatGPT.

Next, we addressed the issue of missing data using several strategies to reduce the presence of null values:

- Row deletion: For instance, we removed rows that lacked a loan approval outcome or had more than one undefined value.
- Median imputation: For continuous numerical features, such as the property price associated with the loan, we replaced missing values with the median (instead of the mean) to minimize the influence of outliers.
- Mode imputation: For nominal features, we filled in missing values using the most frequent (mode) value.

After this initial preprocessing, we obtained a clean and well-understood dataset with 307,493 rows and 42 columns, free of any missing values.

B) Data Visualization

The main goal of the *Data Visualization* section was to explore and understand the relationships between various variables in the loan application dataset to identify factors that influence loan approval or rejection. Through a series of graphs and statistical analyses, several key insights emerged.

First, the distribution of loan decisions revealed a strong imbalance in the dataset: approximately 92% of applications are approved, indicating a potential bias that must be considered during modeling. Demographic analysis showed that women have slightly higher approval rates than men, and that individuals under 40 years old are more likely to be approved, likely due to life events (housing, vehicle) requiring financial support.

Variables such as family status and education level also have a notable impact. For example, married individuals and those with higher education levels show higher approval rates. Stable income sources, such as salaried employment, are also associated with better chances of approval, although income alone does not always differentiate outcomes. Regarding housing, those living in stable conditions (in a house or with parents) are more frequently approved.

From a financial perspective, applicants with moderate incomes (between 50,000 and 200,000) and loan amounts in the mid-range (300,000 to 700,000) tend to be approved more often. A low credit-to-income ratio (CREDIT_SHARE) is also a good indicator of creditworthiness. Annuities between 20,000 and 40,000 are the most common among approved loans. Conversely, very low incomes are often associated with rejections.

Regional analysis showed that population density and geographic stability (same residence, work, and registration location) positively influence loan approval. Mobility scores were calculated to summarize this, and the city-based mobility score was found to be more informative than the region-based one.

Finally, contract type (revolving credit vs. cash loans) and whether the applicant was accompanied during the application process have a subtle but real impact: revolving loans, which typically involve smaller amounts, are more likely to be approved, and applicants accompanied by a family member or trusted person showed slightly higher approval rates.

In conclusion, this section clarified which variables are most relevant to retain for the next modeling steps. Several columns deemed redundant or non-informative were dropped to simplify the dataset. The analysis provided a strong foundation for feature selection and prepared the ground for data preprocessing and machine learning model development.

C) Deleting Unnecessary Columns

This section marked the conclusion of our exploratory analysis. After identifying the most relevant features, we proceeded to remove columns that were either redundant or uninformative for our prediction task, such as CNT_FAM_MEMBERS, FLAG_OWN_CAR, and FLAG_OWN_REALTY. Additionally, we eliminated variables like FLAG_PHONE, FLAG_EMAIL, and DAYS_REGISTRATION, which either contained universally shared information or held little to no predictive value.

As a result of this refinement process, we narrowed the dataset down to 22 meaningful and well-selected variables, ready to be used for the machine learning section.

D) Formatting Data for Machine Learning

At this stage, we identified 7 categorical (string/object) columns that needed to be transformed using one-hot encoding to make them suitable for machine learning algorithms. This step ensures that our model can correctly interpret categorical data without assuming any ordinal relationship between categories.

In addition, we had several continuous numerical columns that required normalization to improve model performance and convergence. We chose to apply logarithmic normalization, as it helps reduce the impact of outliers and skewed distributions, making the data more consistent and suitable for training.

With these transformations completed, our dataset is now fully prepared for the machine learning phase.

E) Machine learning

To predict loan approval outcomes, we tackled the problem as a binary classification task, where the target variable indicates whether a loan was approved or rejected. We started by testing several standard machine learning algorithms using imbalanced data, as around 92% of the applications were approved, creating a significant class imbalance. Despite adjusting class weights, initial results, particularly with the SGDClassifier, were unsatisfactory, showing a low F1 score and poor recall for the minority class. This highlighted the limitations of applying standard models directly to skewed datasets, as they tend to favor the dominant class.

To address this, we implemented SMOTE (Synthetic Minority Over-sampling Technique), which generates synthetic examples for the minority class by interpolating between existing instances. This method allowed us to balance the dataset without simply duplicating records, resulting in a significant performance boost. Our SGDClassifier improved notably, achieving higher precision and F1 scores. However, we observed a slight drop in recall, indicating that the model still misclassified some true positives.

To further enhance generalization and reduce computational cost, we applied a hybrid resampling strategy that combines under sampling of the majority class and SMOTE oversampling of the minority class. This approach reduced the dataset size while maintaining balance, enabling us to test more complex models such as K Nearest Neighbors, Random Forest, Logistic Regression, and Support Vector Classifier. Among these, KNN yielded the best F1 score (~74%), closely followed by Random Forest, while Logistic Regression and SVC performed less effectively.

These experiments demonstrated that handling class imbalance is crucial for reliable predictions and that resampling strategies, when properly applied, can significantly improve model performance. However, we also recognized that excessive under sampling might reduce data diversity and limit model learning. Overall, the combination of resampling techniques and model comparison helped us identify the most suitable approaches for this financial classification task.

F) Deep learning

Our curiosity led us to experiment with a deep learning model, driven by the ambition to improve upon the models previously tested. For this comparison, we applied over-sampling to our dataset in order to fairly evaluate our results against the best-performing model so far: the SGDClassifier. To build the neural network, we relied on concepts studied in class, such as the sigmoid activation function, the Adam optimizer, EarlyStopping callback and loss curves.

Overall, we are satisfied with the outcome, as the model appears to generalize well. While the F1-score has improved, the results are more nuanced: the low recall suggests that the model misses a significant number of positive cases. This indicates instability in the model's predictions. Once again, this may be due to the lack of real rejected loan samples in the dataset, which affects the model's ability to learn rare cases properly.

However, this model could potentially be improved by introducing additional callbacks or by applying regularization techniques, such as dropout, to make the network more robust.

IV. Reflection on Responsible AI

Developing a predictive model in the banking sector, as explored in this project, involves not only technical considerations but also critical ethical challenges. In line with responsible AI practices covered in our course, several key aspects must be addressed to ensure fairness, transparency, and long-term reliability.

First, the model is susceptible to reproducing historical and representation biases, particularly if certain demographic groups (such as women), low-income individuals, or large families are underrepresented or have been treated unfairly in the training data.

These biases must be carefully identified and mitigated to avoid discriminatory outcomes. Moreover, it is important to recognize that a model's capabilities are inherently limited by the data it learns from. It cannot account for individual circumstances or proprietary bank scoring systems, which highlights the need for transparent communication of its limitations to stakeholders.

Additionally, evaluating the model through a single metric like the F1-score is insufficient. A broader set of fairness metrics (such as demographic parity), equal opportunity, or precision by group should be used to uncover and assess potential disparate impacts on different user segments. Another essential aspect is data privacy, especially when handling sensitive financial information. Incorporating mechanisms like differential privacy can help preserve client confidentiality without compromising predictive performance.

Finally, the model's effectiveness cannot be static. Because user behavior and institutional policies evolve, it is crucial to continuously monitor and update the system. This ensures the model remains accurate, fair, and aligns with real-world conditions, while minimizing the risk of deployment bias over time.

V. Conclusion

This project aimed to predict whether a loan application would be approved or rejected based on socio-economic and financial data. After thorough data cleaning and exploratory analysis, we formulated the task as a binary classification problem. Due to a strong class imbalance, we tested various resampling strategies, including SMOTE and a hybrid approach, to enhance model performance. Several algorithms were evaluated, with SGDClassifier delivering the best results. Finally, we incorporated a reflection on bias, transparency, and data protection, in line with Responsible AI principles, to ensure the model is used in an ethical and reliable manner.