WESTERN GOVERNORS UNIVERSITY

Home Credit Default Risk Prediction and Policy Simulation

Capstone Project

Submitted to

Western Governors University College of IT

Erik Hnida

Erik Hnida
WGU D502

Abstract

This project presents a full data analysis and machine learning pipeline for predicting loan default probability using the publicly available Home Credit Default Risk dataset on Kaggle. The work progresses from raw data ingestion and exploratory data analysis to data preprocessing, feature engineering, model development, and evaluation. It includes a fully constructed GitHub repository that maintains all the notebooks, documentation, and source code used for the project for reproducibility and transparency. Additionally, it implements a GitHub actions workflow to demonstrate CI/CD principles. The project includes a final report documenting the analytical decisions, modeling tradeoffs, and conclusions drawn from the results. Additionally, a public Tableau story is included to provide visualizations for some of the key indicators of risk.

Erik Hnida
WGU D502

Default risk is one of the largest sources of loss for banking institutions across the United States. In this project Home Credit, a bank that specializes in helping individuals with little credit history and unbanked individuals get loans, is seeking to improve their default prediction capabilities so they can continue to serve their community without jeopardizing their ability to stay profitable. Using data analysis techniques and machine learning algorithms from the BSDA program, this project provides a full end-to-end data analysis pipeline that improves Home Credits current default risk calculations. If high-risk individuals are screened from the loaning process before being loaned too, Home Credit can save millions of dollars by preventing default before it can happen.

The dataset provided by Home Credit contains 10 artifacts that are used in determining default risk. In this project, 3 of these artifacts, “application_train.csv”, “bureau.csv”, and “HomeCredit_columns_description.csv” are used. The main table is “application_train.csv” and it contains loan and demographics information as well as whether the applicant defaulted (“TARGET”) for each application made to the bank. In this dataset, each row is equal to one loan. There is also an external reporting dataset, “bureau.csv” that was used to enhance the data in the application dataset. This bureau dataset contains as many rows as number of credits (loans) the applicant had before applying to Home Credit. Finally, the “HomeCredit_columns_description.csv” is a .csv file that is used to reference column meanings in all other datasets from the raw dataset.

For raw data ingestion and file structuring, the entire raw dataset was downloaded from Kaggle onto a structured local WSL Ubuntu environment intended to support reproducibility and separation of testing and production. This environment was connected to a GitHub repository for version control and final presentation. Once downloaded locally, the “ingestion.ipynb” notebook tested a reproducible process for opening “application_train.csv” using Pathlib and light exploratory analysis to determine default rate (8%), as well as shape and descriptions of the dataset.

Following this, the “preliminary_eda_01.ipynb” and “preprocessing_02” notebooks focused on determining the shape, investigating the default indication feature (“TARGET”), identifying columns with high missing data, identifying columns with NaN/Null or sentinel values, and investigating the middle statistics of the dataset to help direct further analysis and how these different values are handled. In this notebook, the default rate of the dataset was determined to be 8%, meaning that 8% of all applications defaulted eventually. Next, any columns with more than 50% missing data were dropped from the dataset, as imputation on this degree of missingness would heavily skew modeling algorithms. In the "DAYS_EMPLOYED" column, the sentinel value "365243" represents a null value in the dataset. To deal with this, these values were replaced with null values to avoid misrepresenting employment history. In the exploratory data analysis on middle statistics, outliers were not present, so fixes were not necessary. The final parts of the preprocessing step were to prepare the dataset for machine learning algorithms. To do this, the dataset was first split into four parts, between training and validation and between the X-variables and the y-variable, "TARGET". Then, numeric and categorical columns were identified. For any missing values in numeric columns, the data was imputed. To prepare categorical columns for the algorithms, they were One-hot encoded. After imputing and one-hot encoding were completed, the validation and training datasets were completed and ready to be passed to the two machine learning algorithms.

Following preprocessing, logistic regression and random forest classification were evaluated to predict loan default risk. Logistic regression was selected as the primary model due to its interpretability and robustness, while the random forest classifier served as a nonlinear comparison model. After testing, a decision threshold of 50% was determined to create the best performing models for this project. To measure the performance of the models, both algorithms were measured using their precision and recall scores as well as a capture and lift rate and for each model, the most important features for determining default risk were identified to help understand the profiling of the models.

The logistic regression model captures approximately 66.4% of all defaulters. To achieve this recall, it flags roughly 35% of the population as high risk, among whom the observed default rate is about 15.4%, nearly double the population baseline of 8.1%. Using capture and lift rate vs random, the model performed 3.71x better than the baseline when filtering for the top 5% of high-risk individuals and 3.10x better than the baseline when filtering for the top 10% of high-risk individuals. This is a meaningful improvement over the baseline. When observing the most impactful features for this modeling in determining risk, it heavily favored the loan-to-value relationship between “AMT_GOODS_PRICE” and “AMT_CREDIT”, education level, credit history, and work sector as the primary drivers of risk. This creates a profile where objective, controlled factors of an individual’s demographic features are being used as primary determinants in default risk.

The random forest classifier model captures approximately 52% of all defaulters. To achieve this recall, it flags roughly 22% of the population as high risk, among whom the observed default rate is about 19%, over double the population baseline of 8.1%. Using capture and lift rate vs random, the model performed 3.67x better than the baseline when filtering for the top 5% of high-risk individuals and 3.03x better than the baseline when filtering for the top 10% of high-risk individuals. The random forest classifier and logistic regression models different routes when determining default risk. Many of the most heavily weighted features in Random Forest focus on age, residence location, work tenure, and residence length. This seems to imply that some of the earliest trees in random forest are filtering wealthy, older, established individuals as low default risk early in the forest. Although the random forest classifier achieved higher precision, its conservative behavior and reliance on demographic stability features raised concerns regarding interpretability and potential sensitivity. Given its higher recall, transparent coefficients, and alignment with risk-averse lending strategies, logistic regression was selected as the champion model.

After the baseline modeling was completed for the project, I decided to see if the model could be enhanced by additional feature engineering and joining “bureau.csv” to the dataset using a SQLite3 database. After engineering the loan-to-value (LTV) feature and joining both tables, both models decreased in overall performance. This revealed two important insights: the baseline model already had external reporting data that was weighted extremely heavily in the “EXT_SOURCE_X” columns so the addition of “bureau.csv” data created additional noise; and the loan-to-value relationship on a linear model is already represented abstractly by their relationship with the “TARGET” variable, so the addition of the “LTV” column had little impact on model performance and was weighted very low on feature importance when determining risk. Although the additional work did not improve default risk calculation, it was an insightful experience that showed that the baseline “application_train.csv” model was the only one needed for the scope of this project.

Once all notebooks had been completed the dataset engineering and machine learning algorithm of the project was finalized, I transcribed the jupyter notebook code portions into source code python files that would be used to create a reproducible end-to-end pipeline that would ingest the raw dataset, process the data, train the logistic regression model, and deliver findings all in one script. This was done by creating 5 python modules containing functions to be called in a final python module, main. This source code is provided in the attached GitHub repository so that the script can be executed by any machine with the correct dependencies.

As a final reporting artifact, a scored dataset generated from the logistic regression model was used to create a Tableau story. The visualizations translate model outputs into interpretable risk indicators, highlighting demographic drivers, loan-to-value relationships, and predicted risk groupings to support business decision-making.

The predicted probability of default generated by the model represents an estimate of the likelihood that an applicant will fail to repay a loan under current conditions. Rather than serving as a strict approval or rejection rule, this probability is best interpreted as a relative measure of risk that can inform downstream decision-making. By ranking applicants from lowest to highest risk, the model enables lenders to prioritize manual review, adjust loan terms, or allocate capital more efficiently. Failing to identify a high-risk borrower (false negative) may result in direct financial loss through default, whereas incorrectly flagging a low-risk applicant (false positive) primarily represents an opportunity cost. The evaluated logistic regression model has high recall, meaning that it successfully identifies a substantial proportion of defaulters, albeit at the expense of increased false positives. This tradeoff minimizes default risk at the cost of total loan volume.

While the results of the project show the value of predictive modeling for loan default risk, some limitations must be pointed out. Loan defaults represent a relatively small proportion of all observations in the dataset, resulting in significant class imbalance. While modeling techniques and evaluation metrics were selected to account for this imbalance, it nonetheless complicates prediction and interpretation. Models optimized to capture defaulters may generate a higher number of false positives, which could lead to lost business opportunities if used without additional safeguards. The primary model selected for this project was logistic regression, however logistic regression assumes linear relationships between features and default risk, which may oversimplify more complex, nonlinear interactions present in the real-world. Additionally, the project focuses on predictive performance and does not address operational constraints such as regulatory compliance, fairness, or bias mitigation. In a real-world scenario these considerations are critical but were beyond the scope of the project.

This project demonstrated the development of an end-to-end data analysis and machine learning pipeline for predicting loan default risk, progressing from raw data ingestion to model evaluation and business-oriented visualization. The results show the logistic regression model can effectively rank applicants by default risk. While the current scope focused on a limited feature set and baseline modeling techniques, future work could incorporate richer credit bureau data, additional feature engineering, and more advanced modeling approaches to improve predictive performance. Further extensions may also include fairness analysis and regulatory considerations.

Erik Hnida
WGU D502
https://github.com/ehnidaWGU/D502-WGU-Capstone