Risk Based Loan Approval and Borrower Profiling Using Classification and Clustering

Introduction

Financial Services companies underwrite thousands of consumers and small business loans each month with every approval carrying two competing risks, which are credit risk and opportunity cost. Credit risk is the risk involved when a borrower that was approved of a loan default in repayment costing the bank to lose both its capital and interest. Where the opportunity cost risk involves declining a creditworthy customer which ultimately drives the customer to the bank competitor which eventually costs the bank to lose interest on the loan and also damages the customer lifetime value.

The key challenges financial service companies face when marketing their product is that there is no one fit all sizes for all it potential customers as some customers would prefer a premium credit card offer while some customers might prefer a credit-building product. These challenges occur due to data complexity, regulatory oversight and dynamic behavior.

This project was done to achieve business objectives using supervised and unsupervised machine learning models. The objective of the project was to automate credit decisions, segment customers into clusters to better tailor products offers and risk-based pricing and also optimize resource allocation by focusing underwriting effort on where the model least performs.

The project utilized classification to predict if a customer loan will be approved based on their demographic factors, income, credit history and debt ratios. Clustering was utilized to segement the customers into two distinct groups (low-risk vs high utilization) to inform targeted marketing and risk-management strategies.

Data and Variables

The dataset was a synthetic data generated using python on Kaggle. The data owner created a python code to generate the data and you can see the data and the source code here on [Kaggle.com](https://www.kaggle.com/datasets/lorenzozoppelletto/financial-risk-for-loan-approval?select=CSV+Generation.py).

The dataset contains 20,000 observations with 36 columns for each customer. The dataset contains variables such as numerical, categorical, the outcome variable and also variables that were dropped.

The outcome variable is the ‘LoanApproved’ column, and this was already in a binary format as 0 and 1 with 0=Not Approved and 1 = Approved. I dropped the loan\_id and Application date column from the features. The numerical variables were used as they were and the categorical variables I created a dummy variable for each column.

Analysis and Result

The project utilized classification model and clustering model. The first step was cleaning and mapping the data, I mapped the outcome/target variable using binary format of 0 and 1, encode all the categorical variables by creating a dummy variable for each column and scaled every feature to zero mean/unit variance using (standardscaler) which is crucial for both logistics regression convergence and distance-based clustering.

Next step was to split the data into the train/test spilt using a 70/30 ratio where the model was trained on 70% of the total data in order to learn the pattern reliably and reserve 30% for the test in order to give the model a large enough hold out to get stable estimates of performance with high variance. Using stratify = y in the split ensures both the train and test set preserves the same approval/non-approval ratio, so the metrics are meaningful for the real class distribution.

The Classification model was used for the supervised learning using Logistic regression because interpretable coefficients show how each feature shifts approval odds, its handles binary targets naturally, and scales well for the 20,000 observations. The max\_iter default was changed from 100 to 1000 in order to allow the model to solve more iterations before giving up on convergence due to many features in the dataset.

A screenshot of a computer screen

AI-generated content may be incorrect.

The model worked well at predicting loan approvals with a 100% precision for loan approved and 99.9% for loan not approved. The model also identified correctly both class of customers from the recall figure showing 100% for class 0 and 99.7% for class 1.

The accuracy of the model showed 99.95% being calculated from the true positive and true negative showing all predictions are correct. The AUC measures the model’s ability to rank positive vs negative cases across all possible thresholds with a value close to 1 which indicates almost perfect separability.

The model made very few errors with only 3 false negatives which were approved customers wrongly denied. The recall on class 0 shows you are never giving credit to bad risk customers and class 1 shows almost no good customers are denied.

The confusion matrix of the model shows the true model against the predicted model showing how well the model predicted the classes.

A yellow blue and purple squares with black text

AI-generated content may be incorrect.

With only 3 false negative customers who were approved but the model predicted Not approved showed that 3 credit worthy customers would be turned away. The true negative showed customers who were not approved and the model correctly predicted Not Approved with a total of 4566 customers, there was no false positive from the model and True positives from the model was 1431 show customers who were approved and the model correctly predicted approved.