Default Detectives

BROWN

DEFAULT RISK

Introduction

This project is to predict how likely a borrower is to default or repay the loan. In order to achieve this we employ behavioural, demographic, and credit features of clients.

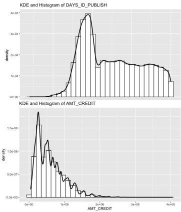
This is a binary classification task. The two classes of TARGET are:

- TARGET = 1: likely to default
- TARGET = 0: likely not to default

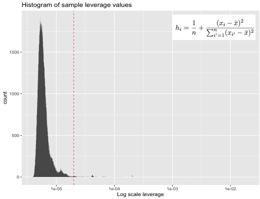
The target variable is imbalanced - originally 92% of the target values are 0's and only 8% are 1's.

EDA and Preprocessing

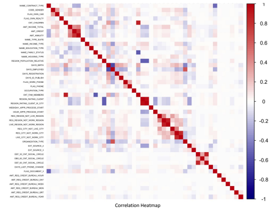
- EDA: KDE plots and the histograms for each numerical feature in the dataset



- Encode categorical features
- Imputation with MICE
- Remove Outliers (Leverage cutoff)

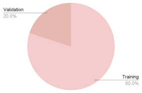


- Drop highly correlated variables
- Standardization



Methods

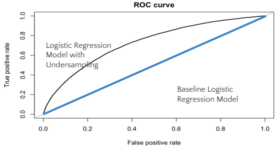
1. **Baseline** Vanilla logistic regression model
2. **Advanced logistic regression to address target imbalance**
   - **Stratified K-Fold Cross Validation:** Divided data into 10 stratified folds
   - **Undersampling**: Reduce the majority class (TARGET = 0)
3. **Random Forest Classification**
   - Train-test split
   - Undersampling
4. **Feature Importance**
   - Covariate magnitude - log odds with bootstrap
   - Wald-test for significance (p-value < 0.001)
   - Forward stepwise model selection on AIC



Results

We used the **AUROC** (Area under ROC) metric, which is insensitive to class imbalance.

The Stratified K-fold logistic regression has a very low AUROC (almost random) that the stratification did not get implemented as desired due to the lack of class=1 samples.



Model	AUROC
Baseline	0.5
Stratified K-Fold Logistic Regression	0.501
Undersampled Logistic Regression	0.726
Undersampled Random Forest	**0.732**

The undersampled random forest model performs best (AUROC=0.732), closely followed by the undersampled logistic regression (AUROC=0.726)

Since logistic regression is more interpretable, we explored its global feature importance below.

Notably, the odds to default the loan increases by 53% for male applicants compared to female counterparts.

Detectives On Mission

Nange Li
Preprocessing
Global Importance

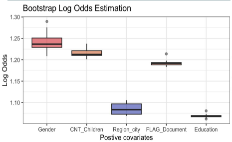
Radhika Mehrotra
Logistic Regression
Model AUROC

Sagarika Ramesh
Random Forest
Interpretation

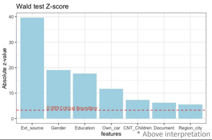
Ramya Harikrishnan
Imputation
Under Sampling

Global Feature Importance *

Covariate magnitude



Wald test z-score



Forward Stepwise

The order of selection

- EXT_SOURCE
- GENDER
- EDUCATION_TYPE
- DAYS_BIRTH
- FLAG_DOCUMENT
- FLAG_OWN_CAR
- DEFAULT_SURROUND

* Above interpretation is only for a subset of all the important covariates

Resources

GitHub

Dataset: https://www.kaggle.com/competitions/home-credit-default-risk/data

ISLRv2: https://www.statlearning.com/