Alternative Methods to Classifying Applicant Default Risk on Loans

Sarah Ellwein
sellwien@calpoly.edu

Jason Mulson
jmulson@calpoly.edu

Bella White
bwhite17@calpoly.edu

Abstract

Developing models to predict a credit applicant's ability to pay back their requested loan with high levels of accuracy has been a top priority for banks and credit groups globally. This paper outlines our approach to this objective. We describe our process of generating and testing linear classification models–logistic regression, support vector classifier, and linear discriminant analysis–with different splitting strategies–random, non-random, and stratified. We found logistic regression with non-random sampling had the best ROC-AUC score of all the models tested. Ethical considerations are of utmost priority in developing these models and will be covered and thoroughly discussed in this report.

Introduction

In August 2018, Home Credit Group published a dataset of anonymous loan applications, credit bureau data, and transaction history. They challenged Kaggle competitors to create a model that could consistently and accurately predict whether a loan applicant would default on their requested loan (*Home Credit Default Risk*, n.d.). The overarching goal of the project is to financially empower unbanked populations by using an alternative method of classifying an applicant's default risk. The provided data included information about each applicant's education, family, credit, financial, and working history, as well as information on the loan they requested. This dataset is historical, that is, it only includes information on approved applicants who have either defaulted or paid back their loans in full. From this data, we intend to predict whether or not an applicant will default on their loan if approved for one.

Data Preparation

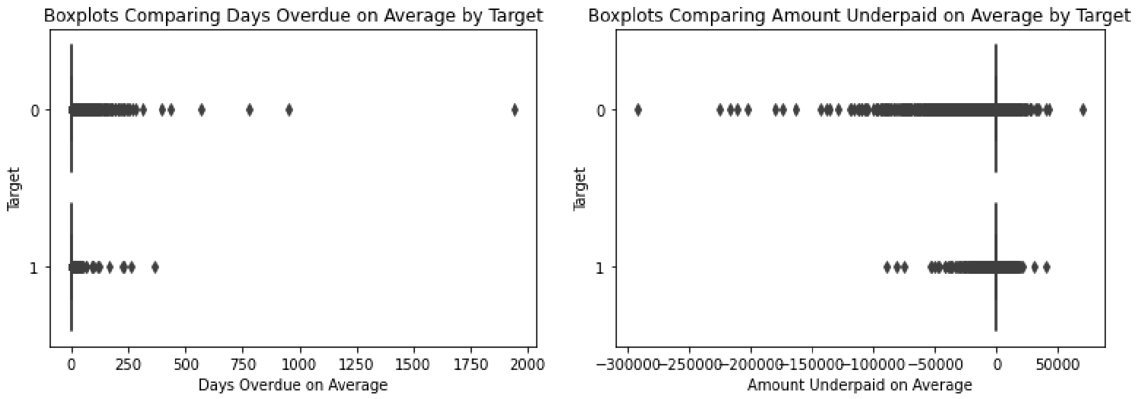
3.1 Data selection

It is important to consider our data in the context of approving credit for individuals. Since we are engaging with credit transactions, we must adhere to the Equal Credit Opportunity Act (ECOA) when building our model. The ECOA prohibits discrimination in any aspect of a credit transaction based on protected characteristics such as race, religion, nationality, sex, marital status, or age (*Equal Credit Opportunity Act*, n.d.). To ensure compliance avoid discrimination, we dropped several features from our dataset: *Coded_Gender* and *Days_Birth*.

We also checked features that could be used to infer any of the protected characteristics. Therefore, *Name_Family_Status* was dropped as it contained information about the individual's marital status. We dropped these features to ensure that our model didn't utilize any protected characteristics to make a decision on someone's ability to pay back a loan.

Our goal for determining the features to include in the model consisted of taking as many predictors as possible from the datasets provided. Our first step was deciding on certain features to remove from our initial dataset before introducing new features. The variables we dropped ended with the following words in their features: 'avg', 'medi', 'mode', 'house', 'day', 'week', 'change'. These features were not included in the model because they did not provide enough prediction power for classifying whether or not someone would repay their loan. Our next step was to grab predictors from the other datasets. We grabbed predictors that helped provide more information about each applicant. An example of the predictors we included in the dataset were if they received a loan from the company in the past and whether or not it was still active at the time. Next, we decided to do exploratory data analysis to understand more about the data.

Two numerical predictors we believed would help determine whether or not someone would repay their loan were the average amount of days overdue (*Days_Overdue*) and average amount underpaid (*Amt_Underpaid*) from previous loans. While there were applicants who did not have previous loans with the company, we still found some applicants who have averaged a significant amount of days past the due date of the loan and whether they over or under paid the loan on average. The figures of the two predictors compared to whether or not they repaid the loan are shown below:



The categorical features we found to have the highest association with the applicant's ability to repay their loan were *Occupation_Type*, *Organization_Type*, *Income_Type*, *Education_Type*, and *Housing_Type*. We found these five features to have the highest associations with the target variable using Cramer's V, which is a measure of association between two categorical variables. The table below shows the Cramer's V measurement for the five predictors with the highest association:

Predictor	Association with Target
Occupation Type	0.0801
Organization Type	0.0723
Income Type	0.0638
Education Type	0.0576
Housing Type	0.0370

We can see that all the measurements are smaller than 0.1 which is considered the minimum for a moderate measure of association between the two variables being compared. Although the predictors above do not have high Cramer's V measurements, keeping all of the predictors left in our final model would allow for the best predicting power.

3.2 Data cleaning

The initial dataset *application_train.csv* contained several missing values for the majority of features. Additionally, outliers were detected in certain features and are taken into consideration. This section will briefly outline techniques and strategies used to handle both missing data and anomalies.

Most rows with missing data were handled with simple imputation strategies. The following techniques were applied:

- Missing numerical data where features can be considered non-applicable is imputed with zero
  e.g. number of children, age of car, etc.
- Missing numerical data where missing values are different from having a zero value is imputed with zero and flagged in a separate column
  e.g. credit bureau scores, social circles, etc.
- Missing categorical data where the option isn't indicated for the applicant is imputed with either 'not applicable', 'unknown', or 'other' depending on the context
  e.g. Applicant's occupation type is indicated as 'other' assuming the occupation isn't found during the application process
- Missing categorical data where the common value is more than half the entire dataset is imputed with its mode
  e.g. Applicant's name type suite is imputed with 'Unaccompanied'

However some missing data required more exploration before we decided how to replace or impute data. An example of this is the *Days_Employed* feature which measures the continuous days an individual has been employed relative to to the date of their loan application. Due to this structure, the values should be negative or 0. However, when performing EDA on this feature to determine the distribution of days employed, we found that there were 32,123 rows with the same positive value of 365,243 days employed. Not only should this number be negative, but having 32,123 rows with the exact same outlier value was highly unusual.

Upon further investigation, we found this group had a lower rate of defaulting than the population average. Only 5.4% of these individuals defaulted compared to the population average of 8.1% defaulting. Although we didn't know what the 365,243 value meant, we concluded that it must be some kind of flag or indicator value. We created a new column called *Days_Employed_Outliers* to flag rows that contained the odd value, then replaced the *Days_Employed* value of these flagged individuals with the median value of the population. This ensured we didn't lose potentially significant insight while removing outlier values.

Data integration was an important part of this project due to the handful of datasets involved. We needed to find a way to aggregate previous clients' information in the final dataset for more insight into whether or not an applicant will repay their loan. Here are two examples where we used data aggregation to integrate information from the other data sources into our final dataset:

- *Num_Prev_Loans*: counting the number of previous loans the company has for each applicant and
- *Days_Overdue*: summing the number of days that previous clients were past due that was undeferred.

3.5 Data formatting

After imputing missing values and integrating additional data, the final dataset must be preprocessed for model use. The following techniques were applied:

- All numerical variables were standardized using z-scores
- All boolean variables were casted as integers (True/False becomes 1/0, respectively)
- All binary categorical variables with categories 'Yes'/'No' were casted as integers ('Yes'/'No' becomes 1/0, respectively)
- All categorical variables were one hot encoded

After preprocessing, the resulting dataset had 265 predictors. This excessive number of predictors slowed down the model during implementation, so predictors must be dropped to speed up the process. Since linear classifiers were used for this project, predictors with low linear correlation with the target (<0.007) were dropped. The resulting dataset has 151 predictors, which proved to be efficient for model implementation.

Data Modeling

4.1 Modeling technique

Three linear classification models were chosen for this project: logistic regression (LOG), support vector classifiers (SVC), and linear discriminant analysis (LDA)

LOG is implemented with gradient descent to determine the value of the predictors coefficients. Initial coefficients and learning rates were chosen based on the speed of convergence and lowest possible gradient value. The initial coefficients were the mean of each feature in the dataset with a learning rate of 2. Additionally, different ridge penalty coefficients were tested on LOG: 1, 5, 10. These penalties were tested using three fold cross validation; however, none of the values produced optimal results. An penalized ridge LOG model (with penalty 10) was implemented in addition with the non-penalized LOG model to provide comparisons that will be discussed in the model assessment section. Using LOG requires the assumption that the data has linear association with the logit odds of the target.

SVC is also implemented with gradient descent to determine the linear coefficients of an affine line separating the targets. Initial coefficients and learning rates were determined with the same criteria as LOG. The initial predictors were 0.000001 (any small value near 0 would work) and a learning rate of 0.000001. Using SVC doesn't require any assumptions about the data.

LDA is implemented using log-likelihood ratio. This only required the mean, standard deviation, and sample size of each target class. Although this procedure appears simple, matrix inversion created difficulties in computing linear coefficients for sparse data. This was resolved using truncated SVD, using 90% of total components to retain the most data while avoiding a singular matrix of data. Using LDA requires the assumption that the data for each target is multivariate normal.

4.2 Test design

We implemented three sampling methods for our training and holdout sets: random sampling, stratified sampling, and non-random sampling.

The random sampling selects 70% of the data at random to train the models on and reserves the remaining 30% of data for the holdout set. One might choose a random sampling in a balanced dataset, i.e. there are equal amounts of both the positive and negative target variable in the dataset.

Unlike the random sampling method, stratified sampling maintains the original distribution of target variables in the 7:3 training holdout split. This splitting method is optimal for handling unbalanced dataset like the HCDR dataset, which has a large majority class (92% of individuals didn't default on loan). The random sampling strategy does not guarantee capturing the minority class in the training set. (individual defaulted on loan). Stratified sampling ensures that both the training and dev set will have the same 92:8 distribution of target variables.

However stratified sampling doesn't resolve issues that come with an unbalanced dataset. The class we are trying to predict only represents 8% of the data, which could prove hard for a model to differentiate from the 92% of the majority class.Non-random sampling is implemented by dropping enough of the majority class to balance the distribution of target variables in the data. In the final dataset, we used non-random sampling to ensure that both the training and dev sets were 80% majority class and 20% minority class. The drawback of this method is the reduction of usable data, therefore, this method should be used in cases where the dataset is sufficiently "large" enough.

4.3 Model creation

The final dataset was tested on a combination of 4 different classification techniques with 3 different sampling strategies, hence testing 12 different models. The goal of all models produced is to maximize ROC-AUC score. In addition to optimizing this performance metric, each model should be robust to variance when classifying future data; therefore, three fold cross validation is applied to observe robustness.

After testing each model with different train/holdout split strategies, one model is chosen from each modeling technique based on ROC-AUC performance (excluding penalized LOG, which will be explained in model assessments). The tables below depict the results.

LOG, Non-random Sampling

Cross Validation Results

	Accuracy	Precision	Recall	F1 Score
Fold 1	80%	15%	48%	0.22
Fold 2	79%	15%	44%	0.22
Fold 3	79%	15%	45%	0.22

Train/Holdout Results

	Accuracy	Precision	Recall	F1 Score
Train	77%	25%	39%	0.22
Holdout	77%	26%	39%	0.31

Holdout ROC-AUC: 0.67

SVC, Non-random Sampling

Cross Validation Results

	Accuracy	Precision	Recall	F1 Score
Fold 1	49%	39%	17%	0.23
Fold 2	50%	41%	18%	0.25

Train/Holdout Results

	Accuracy	Precision	Recall	F1 Score
Train	50%	38%	17%	0.24
Holdout	51%	38%	17%	0.24

Fold 3	50%	40%	17%	0.24

Holdout ROC-AUC: 0.45

LDA, Stratified Sampling

Cross Validation Results

	Accuracy	Precision	Recall	F1 Score
Fold 1	8%	100%	8%	0.15
Fold 2	46%	87%	12%	0.21
Fold 3	48%	81%	13%	0.22

Train/Holdout Results

	Accuracy	Precision	Recall	F1 Score
Train	48%	86%	12%	0.21
Holdout	47%	86%	12%	0.21

Holdout ROC-AUC: 0.65

4.4 Model assessment

Each model had clear advantages and disadvantages for the task given. These traits may be due to the model's assumptions or its volatility to imbalanced data.

LOG proved to be the most robust model with consistent results when using a non-random sampling strategy, shown through cross validation. Additionally, it had the highest ROC-AUC score of all the models produced when tested on the holdout set. One of this model's weaknesses is its ability to identify applicants at risk of defaulting. Of the models shown above, LOG had the lowest recall score. Additionally, LOG doesn't perform robustly when using a random sampling strategy.

A penalized LOG model was also implemented; however, the performance was suboptimal overall. The ROC-AUC score was consistently 0.5 and never appeared to improve, regardless of sampling strategy. This may be due to additional bias that adding a penalty presents to a model using biased data. With this in mind, penalized LOG models were not included in the final results.

Unlike LOG, SVC had the worst results of all the models in general, regardless of sampling strategy. The only identifiable trait of this model is its robustness; however, the results still disappoint. SVC had the lowest ROC-AUC score, F1 score, and accuracy. SVC also had the longest runtime.

Finally, LDA with stratified sampling had promising results in the final train/holdout analysis. LDA had the second highest ROC-AUC score and the highest recall. Additionally, LDA runs the fastest of all the models shown. Although these results seem optimal, the cross validation results showed conflicting results. In the first fold, LDA had the worst accuracy and showed a tendency towards classifying all applicants as high-risk. In the second and third folds, the scores begin to show more balanced results. This showed evidence of high variance, which is concerning when attempting to predict future data. This lack of robustness is taken into consideration.

Of all the models produced, LOG with non-random sampling had an optimal ROC-AUC score with robust results and therefore was chosen as the final model. After submitting to Kaggle, it received the final ROC-AUC score of 0.60.

Evaluation

5.1 Results evaluation

Although the results from the final model were acceptable, it is still far from completion for commercial use. Several improvements should be made to mitigate bias and errors. One of the main challenges in producing the best model is lack of subject-matter expertise. Due to lack of knowledge, the model fails to use potentially helpful predictors that may improve model performance. Consultation from a loan counselor could provide insight for better predictors and guidance for error analysis.

5.2 Discussion

We had to consider the ethical and legal ramifications of each design decision made due to the sensitive nature of both the data and the model's predictions. This includes removing features directly or indirectly prohibited by the ECOA to ensure the model did not make decisions based on age, gender, or marital status.

An additional ethical consideration concerns the granularity of the dataset. Each observation of this public dataset includes private information of an individual. Although directly identifying information like names, addresses, and phone numbers were removed before publication, it is possible to identify an individual due to the high granularity of the dataset.

Another precaution is the representation of applications provided.The model developed in this project was trained on accepted applications, however, the acceptance criteria for these applications are unclear. This may perpetuate an unconscious bias when approving these loans since there are a lack of comparisons between the accepted and rejected population. Therefore the potential intrinsic bias in the data could be perpetuated through our model.

These issues potentially clash with the goal presented by Home Credit, as they expressed a desire to "broaden financial inclusion for the unbanked population" (*Home Credit Default Risk*, n.d.). Since the historical loan data does not include unbanked populations, it is challenging to build a model that will fairly evaluate an applicant's ability to pay back their loans. It may perpetuate the same historical discrimination or biases against the same population Home Credit hopes to financially empower.

Approving a loan can change someone's life: empowering them to buy their first home, purchase a reliable car to get to work, or pay for continued education. Using an algorithm to make this decision can bring forward ethical concerns from personal biases of those generating the dataset. It is important for design decisions and final models to be transparent to identify and combat potential biases towards groups or individuals. Publishing the ethical and legal justifications for design decisions at every step, being clear about how the training data was gathered, and identifying how the model is making decisions are all crucial steps in achieving transparency. With the exception of personal information and commercial secrets, the process of approving or denying a loan must be open to public scrutiny. Additionally, individuals of denied loan applications should be able to request investigation into why they were denied. The model is not infallible and could be perpetuating a bias which can only be caught by human review.

Our model is not the only entity that can misinterpret the presented information. We acknowledge that our model can be misinterpreted in a variety of ways by Home Credit and associates. Although the model can predict an individual's ability to pay back the requested loan with some accuracy, it is in no way a comprehensive evaluation of an individual's financial status as it has no way of evaluating unique circumstances. This model should be used in conjunction with human review and should not be utilized as the sole (or even primary) method of evaluating a loan application given these factors.

References

Equal Credit Opportunity Act. (n.d.). Federal Trade Commission. https://www.ftc.gov/legal-library/browse/statutes/equal-credit-opportunity-act

Home Credit Default Risk. (n.d.). Kaggle. https://www.kaggle.com/competitions/home-credit-default-risk