Literature Review

Loan Default Prediction

Name: Annas Imtiaz

Student Number: 501203680

Supervisor's Name: Ceni Babaoglu

Submission Date: June 05, 2023

Ryerson University

Table of Contents

Abstract

Credit is the central revenue-generating activity for the banking / credit industry and comes in two primary forms. In Unsecured credit, there is no security / collateral for the bank to fall back on in case a user defaults. Secured credit, on the other hand, has some form of collateral for the bank to fall back on. Therefore, in the case of unsecured credit, it becomes even more important for the issuer to improve its decision-making accuracy.

Over the past few decades, a lot of academic research and application has been focused towards improving prediction accuracy of whether a new loan applicant will default or not, and issuers are always striving to improve this accuracy. Many machine learning algorithms including neural networks have been used to classify new applicants as potential good or bad borrowers.

Usually, an issuer's portfolio, or a publicly sourced dataset, would be imbalanced. Imbalance here means that the number of defaulted cases is less than the number of non-defaulted cases. Default is usually a binary classification, with 1 representing default and 0 representing non-default. As per Canadian Bankers Association (CBA), more than 70% of Canadians pay their credit card balance in full each month. This is itself reflective of the skewed nature of the problem, with a greater proportion of non-default instances. This makes training a model for default cases a bit difficult, since the model is unable to learn the characteristics of defaulted customers, due to low volume / high volatility.

A lot of academic research has tried to balance this imbalance, using either Undersampling, or Oversampling. Undersampling means that non-defaulted observations are deleted to make their count as close to the defaulted cases as possible. Oversampling, on the other hand, refers to adding sample defaulted observations, which have similar characteristics as actual defaulted observations, to have a sufficient volume for the algorithm to learn.

In some areas, academic research falls short of what is required by industry. One such instance is default prediction. Academic research is based primarily on predicting whether a new applicant is likely to default or not. However, a yes / no decision is not what banks / credit issuers are looking for. Credit issuers are more interested in a rank-order of borrowers based on their probability of default / credit score. They can then use this rank-order to define what level of risk they are willing to take, based on their risk appetite and overall economic forecasts. This allows them to set a benchmark through which they can make a risk – return trade-off decision, i.e., what level of risk are they willing to take if it offers a specific level of return. This benchmark / cut-off is obviously dynamic, and as economic conditions improve, issuers generally are more likely to take on risk, while risk aversion increases when economic uncertainty looms.

Research Questions:

- Which is the classification algorithm that produces the best results in terms of Accuracy, F-Score and AUROC?
- Are any of the independent variables correlated, leading to potential over-fitting if one of them is not excluded from analysis?
- Do the chosen algorithms have the same level of accuracy on both the train and test dataset?
- Do balanced datasets deliver better results as compared to imbalanced datasets?

Scope of research:

Based on the shortcoming of majority of academic research as noted earlier, this research will take traditional default prediction a step forward. In the first phase, the research will predict whether a new customer will default, with the outcome being a binary 0/1. However, in the next stage, the research will proceed to calculate probability of default, and eventually credit scores of these applicants (if deemed fit, as the data set already has FICO Scores). This would thus allow issuers to determine what level of risk they wish to undertake and set a benchmark cut-off based on each issuer's own risk appetite.

To answer our research questions, we will be using Decision Tree, Random Forest, Logistic Regression, KNN, ADA Boost and XGBoost. We will then perform Cross Validation on the top 2 models. Finally, ROC curves will be plotted for the top two performing models and their AUC Scores will be compared. This should answer our first research question. The same process will be repeated for both the original as well as balanced datasets to answer our fourth research question. For questions two and three, they would be answered during the pre-processing phase.

Data source:

Data used for this research is an open source dataset from Kaggle (Lending Club Loan Data Analysis | Kaggle). It has a total of 13 features (12 independent variables and 1 dependent variable) with 9578 observations. The data has a relatively good representation of both cases (default: 1533, non-default: 8045).

GitHub Repository:

The research files and codes will be uploaded to this public repository here: https://github.com/AnnasImtiaz/Loan-Default-Prediction

Limitations of research:

This dataset considers only factors specific to the borrower; however, credit decisions by issuers are influenced by external factors as well, including economic pulse, risk appetite and central bank measures to name a few. Therefore, the research may need to be built-upon with these external factors to give a comprehensive solution to lenders.

Literature Review

The first paper studied was 'An Investigation of Credit Card Default Prediction in the Imbalanced Datasets'[1]. The hypothesis was whether models developed using different ML techniques are significantly different from each other, and whether sampling techniques would improve the model's performance. The study used three different datasets and compared performance across each.

Three imbalanced datasets were used. Once data was preprocessed and resampled, Gradient Boosted Decision Tree (GBDT) model, an ensemble-based learning method, was used for modeling and its results were compared with traditional machine learning models (Random Forest, Bagging, KNN, Logistic Regression, Ada Boost and Stacking).

Next, a few resampling methods were used to undersample and oversample the data. A Gradient Boosted Decision automatically selects significant features during the modelling phase. It also gives the relative importance, also known as relative influence, of the independent features. To control overfitting of the Gradient Boosted decision Tree, a regularization shrinkage was also added.

To evaluate the models, Accuracy, Precision, Recall, F-Measure and AUROC. Additionally, Geometric Mean was also used to address imbalanced classes of the original datasets.

Results: For the initial datasets, GBDT outperformed all the other algorithms. In terms of Undersampling methods, Cluster Centroids has outperformed Random Undersampling as well as Near Miss methods. Amongst all the Oversampling techniques, K-means SMOTE outperformed Random oversampling, ADAYSN, SMOTE, Borderline-SMOTE and SMOTETomek. Across application of the algorithms on oversampled dataset, Random Forest had the best results when using Random Oversampling, while GBDT had the best results across all the other oversampling methods. The research concluded that balanced datasets had better accuracy as compared to imbalanced datasets, and that oversampling techniques produced better results compared to under sampling techniques.

The ANOVA test rejected the hypothesis and showed that the proposed methods using imbalanced datasets have significantly improved performance from the baseline model. Credit amount, marital status and education level were found to be significant features.

Authors in 'A hybrid interpretable credit card users default prediction model based on RIPPER'[2] focus more on interpretability of default prediction models rather than their accuracy. Specifically, the study focused on operation time and stability. By models being difficult to interpret, researchers here refer to domain experts being unable to understand the mechanics behind a model, and thus not able to improve on it. Therefore, for this research, interpretability is defined as the model can be understood, the rules are completely understandable, and the number of rules is reasonable. Researchers are of the view that in some jurisdictions, it is a legal obligation to justify a refused credit decision, which is only possible if the model is interpretable.

After data normalization, RELIEF Algorithm was used for feature selection. This is a feature weighted algorithm, which assigns weights to features based on correlation with other features. Then, features with weight lower than a certain threshold are dropped.

Since the dataset being used is imbalanced, SMOTE is used to add some sample data points to create equality between both classes. RIPPER is customized 2-loop algorithm that the authors recommend.

In order to measure performance, F-Value and AUROC are used. The authors conclude that their proposed refined Relief algorithm boasts of better performance as compared to traditional algorithms.

Authors of 'Credit Card Default Prediction using Machine Learning Techniques'[3] aim to find the correlation and predictive power of factors contributing to credit card default. For data pre-processing, some variables were converted from numeric to factor, while Correlation based Feature Selection was used to reduce dimensionality. For this paper, researchers used Logistic Regression, Decision Trees and Random Forest to predict default.

Based on a comparison of true positive vs false positive, Random Forest displayed the best accuracy and the highest area under the curve.

In their research paper 'Comparison of Different Ensemble Methods in Credit Card Default Prediction'[4], the authors try to apply ensemble machine learning methods on both originally imbalanced as well as after applying some balancing techniques to the dataset. In both cases, min-max scalar has been used to scale the data. The study primarily aims to answer:

- How well ensemble methods work on credit card default prediction?
- Are ensemble methods better than other machine learning methods when used on skewed datasets?
- Does balancing the dataset have an impact on relative performance gain of ensemble methods?

The ensemble techniques used in this research are Bagging, Boosting (AdaBoosting and XGBoosting), Voting, and random forests (RF). Authors have pertinently identified that accuracy on positive decisions is relatively more critical in this case, as an incorrect credit decision of issuing a credit card to a customer who will default will result in greater loss for the issuer. Some of the metrics used in this paper are Accuracy, Precision, Recall, ROC and AUC – however, F1 score is used to assess ensemble methods.

For imbalanced dataset, F1 was used as the evaluation metric. Stacking had the best F1 score, followed by Neural Networks. In terms of AUROC, Stacking was followed by XGBoost. Dataset balancing was done by down sampling it. Once the data is balanced and reshuffled, XGBoost comes out to have the highest accuracy, as well as AUROC.

The paper 'Research on Default Prediction for Credit Card Users Based on XGBoost-LSTM Model'[5] uses XGBoost, which is widely used in financial classification models, and Long-Short Term Memory (LSTM), which is widely used in time-series information. In this research, default prediction is primarily based on account, credit bureau, and transaction flow data.

Machine learning algorithms and deep learning algorithms are used on the same data set to construct default prediction models, and the prediction accuracy and modeling workload are compared, ultimately revealing that the deep learning model has high prediction accuracy. One interesting approach used by researchers

here was calculating ‘desire of funds’ which is based on the number of days the credit card was issued and it’s first use.

Performance evaluation metrics include accuracy, precision, recall and AUC. KNN and SVM algorithm for default prediction appear to be less effective, and the decision tree, random forest, AdaBoost, and XGBoost algorithms are better; of these, the XGBoost algorithm is the best. LSTM is a form of Recurrent Neural Network (RNN). The results suggest that a fusion of XGBoost and LSTM yields significantly better results than a simple XGBoost method.

The last paper referenced here is ‘Research on Credit Card Default Prediction Based on k-Means SMOTE and BP Neural Network’[6]. This paper proposes a prediction model based on k-means SMOTE and BP neural network. K means SMOTE is used to change the data distribution. Next, random forest is used to calculate importance of the features and this importance is then substituted into the initial weights of BP neural networks. The paper uses KNN, logistic regression, SVM, random forest and decision trees and then compares the results of these 6 methods.

The 23 features are converted into 89 input variables. Next, PCA is used to reduce these 89 to 27 input variables. The results suggest that the proposed algorithm leads to an increased performance compared to all the other algorithms.

Descriptive Statistics of dataset

There are no null values in the dataset:

```
1 loan_data.isnull().sum()
```

```
credit.policy        0
purpose              0
int.rate             0
installment          0
log.annual.inc       0
dti                  0
fico                 0
days.with.cr.line    0
revol.bal            0
revol.util           0
inq.last.6mths       0
delinq.2yrs          0
pub.rec              0
not.fully.paid       0
dtype: int64
```

There is a mix of categorical and numeric variables in the dataset:

```
1 loan_data.head()
```

it.policy	purpose	int.rate	installment	log.annual.inc	dti	fico	days.with.cr.line	revol.bal	revol.util	inq.last.6mths	delinq.2yrs	pub.rec	not.fully.paid
1	debt_consolidation	0.1189	829.10	11.350407	19.48	737	5639.958333	28854	52.1	0	0	0	0
1	credit_card	0.1071	228.22	11.082143	14.29	707	2760.000000	33623	76.7	0	0	0	0
1	debt_consolidation	0.1357	366.86	10.373491	11.63	682	4710.000000	3511	25.6	1	0	0	0
1	debt_consolidation	0.1008	162.34	11.350407	8.10	712	2699.958333	33667	73.2	1	0	0	0
1	credit_card	0.1426	102.92	11.299732	14.97	667	4066.000000	4740	39.5	0	1	0	0

Mix of numerical and categorical features.

- Categorical Features:
  - purpose
  - credit.policy
  - inq.last.6.mths
  - delinq.2yrs
  - pub.rec
  - not.fully.paid
- Numerical Features:
  - int.rate
  - installment
  - log.annual.inc
  - dti
  - fico
  - days.with.cr.line
  - revol.bal
  - revol.util

Some features appear to have outlier values, and may need to be treated:

```
1 loan_data.describe().T
```

	count	mean	std	min	25%	50%	75%	max
credit.policy	9578.0	0.804970	0.396245	0.000000	1.000000	1.000000	1.000000	1.000000e+00
int.rate	9578.0	0.122640	0.026847	0.060000	0.103900	0.122100	0.140700	2.164000e-01
installment	9578.0	319.089413	207.071301	15.670000	163.770000	268.950000	432.762500	9.401400e+02
log.annual.inc	9578.0	10.932117	0.614813	7.547502	10.558414	10.928884	11.291293	1.452835e+01
dti	9578.0	12.606679	6.883970	0.000000	7.212500	12.665000	17.950000	2.996000e+01
fico	9578.0	710.846314	37.970537	612.000000	682.000000	707.000000	737.000000	8.270000e+02
days.with.cr.line	9578.0	4560.767197	2496.930377	178.958333	2820.000000	4139.958333	5730.000000	1.763996e+04
revol.bal	9578.0	16913.963876	33756.189557	0.000000	3187.000000	8596.000000	18249.500000	1.207359e+06
revol.util	9578.0	46.799236	29.014417	0.000000	22.600000	46.300000	70.900000	1.190000e+02
inq.last.6mths	9578.0	1.577469	2.200245	0.000000	0.000000	1.000000	2.000000	3.300000e+01
delinq.2yrs	9578.0	0.163708	0.546215	0.000000	0.000000	0.000000	0.000000	1.300000e+01
pub.rec	9578.0	0.062122	0.262126	0.000000	0.000000	0.000000	0.000000	5.000000e+00
not.fully.paid	9578.0	0.160054	0.366676	0.000000	0.000000	0.000000	0.000000	1.000000e+00

References

1. Alam, T. M., Shaukat, K., Hameed, I. A., & Luo, S. (2020). An investigation of credit card default prediction in the imbalanced datasets Institute of Electrical and Electronics Engineers. doi:10.1109/ACCESS.2020.3033784 https://eq6sp2kj9f.search.serialssolutions.com/?ctx_ver=Z39.88-2004&ctx_enc=info%3Aofi%2Fenc%3AUTF-8&rfr_id=info%3Asid%2Fsummon.serialssolutions.com&rft_val_fmt=info%3Aofi%2Ffmt%3Akev%3Amtx%3Ajournal&rft.genre=article&rft.atitle=An+Investigation+of+Credit+Card+Default+Prediction+in+the+Imbalanced+Datasets&rft.jtitle=IEEE+access&rft.au=Alam%2C+Talha+Mahboob&rft.au=Shaukat%2C+Kamran&rft.au=Hameed%2C+Ibrahim+A.&rft.au=Luo%2C+Suhuai&rft.date=2020-01-01&rft.pub=IEEE&rft.eissn=2169-3536&rft.volume=8&rft.spage=201173&rft.epage=201198&rft_id=info:doi/10.1109%2FACCESS.2020.3033784&rft.externalDocID=6287639&paramdict=en-US
2. Xu P, Ding Z, Pan M. (2018, January 16) A hybrid interpretable credit card users default prediction model based on RIPPER. Concurrency Computat Pract Exper. 2018;30:e4445. https://doi.org/10.1002/cpe.4445
3. Y. Sayjadah, I. A. T. Hashem, F. Alotaibi and K. A. Kasmiran, "Credit Card Default Prediction using Machine Learning Techniques," 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA), Subang Jaya, Malaysia, 2018, pp. 1-4, doi: 10.1109/ICACCAF.2018.8776802. https://ieeexplore-ieee-org.ezproxy.lib.torontomu.ca/document/8776802
4. Azhi Abdalmohammed Faraj, Didam Ahmed Mahmud and Bilal Najmaddin Rashid (2021, November 05) Comparison of Different Ensemble Methods in Credit Card Default Prediction https://journals.uhd.edu.iq/index.php/uhdjst/article/view/806/640

5. Gao, J., Sun, W., & Sui, X. (2021). Research on Default Prediction for Credit Card Users Based on XGBoost-LSTM Model. Discrete Dynamics in Nature and Society, 2021https://doi.org/10.1155/2021/5080472

   https://www.proquest.com/docview/2613966865/fulltextPDF/37DACFBE1A1148A0PQ/1?accountid=13631
6. Chen, Y., & Zhang, R. (2021). Research on Credit Card Default Prediction Based on k-Means SMOTE and BP Neural Network. Complexity, 2021https://doi.org/10.1155/2021/6618841

   https://www.proquest.com/docview/2503352687?parentSessionId=bEFXwIo8L%2B05%2FxWutfHVd7jlNg6%2F7F78bhorMkI1lMc%3D&pq-origsite=summon&accountid=13631