Predicting Credit Card Default

Mary Donovan Martello

Project 1: Milestone 4

Executive Summary

Defaults of credit card debt hurt both the lender company as well as the economy as a whole. Not only is the lending company hurt by not being repaid for its loan, but the borrowers and future borrowers are hurt by the company increasing interest rates and fees to make up for charge-off losses. Credit card companies must endeavor to predict if the status of customers' accounts indicate that the customers are on the verge of default so that they may mitigate future defaults. The goal of this project was to develop predictive classification models that will take as inputs features of customers' credit card accounts and classify the account as either (1) will default next month or (2) will not default next month.

Credit cards accounts include many features, including demographic data, credit limits applied, and multiple monthly values of how long the payments are overdue, debt outstanding, and amounts of previous payments. We examined these account features and selected the best combination of features for predicting a pending default. In exploring the credit account features, we also identified that some of the account features showed some distinction in the default cases but not in the cases that did not default in the next month.

The conclusions from this project are that the demographic features are not predictive of default, but the features of the length of time the due payments are late as well as the credit limit in relationship to the amount of debt incurred are predictive of pending default. Predictive models can be employed on actual credit accounts to predict pending defaults with a good degree of accuracy. The models built in the project will be valuable to credit card companies for detecting pending defaults. Further research should be performed to identify the statistical relationship between credit limit and payments due to help credit card companies make credit decisions in the future as well as to identify other key variables for predicting pending defaults.

Introduction/Background

Defaults of credit card debt are not only harmful to the lender company but also to the economy as a whole. This is demonstrated by the fact that the United States Federal Reserve tracks the delinquency rate and the charge-off rates of credit cards of commercial banks (Charge-Off and Delinquency Rates on Loans and Leases at Commercial Banks, 2021). There is $893 billion of outstanding commercial credit card debt in the United States and a 2.62% credit card charge-offs rate due to credit card defaults (15 Shocking Credit Card Debt Statistics, 2020) (Charge-Off and Delinquency Rates on Loans and Leases at Commercial Banks, 2021). Credit card companies must endeavor to predict if the status of customers' accounts indicate that the customers are on the verge of default so that they may mitigate future defaults.

The dataset used for this project includes actual credit card records at points in time in the past. Each record in the dataset also includes a label of whether the account holder defaulted the next month (default = 1) or did not default the next month (default = 0). The dataset is imbalanced with respect to the percent of records with default labels (22%) versus records with non-default labels (78%). This imbalance in the target variable must be addressed before or within the models in order to properly evaluate accuracy metrics from the models as the imbalance makes it much more likely to predict a non-default and be accurate regardless of the skill of the model (Alam, 2020) .

The goal of this project is to develop predictive classification models that will take as inputs features of credit card accounts and classify the accounts as either default or non-default.

Methods

The analysis is following the CRISP-DM stages for data science projects including Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. The Business Understanding analysis is summarized in the Introduction / Background section above and deployment is beyond the scope of this project. The Data Understanding, Data Preparation, and Modeling, stages are discussed in this Methods section. Evaluation is subsequently discussed below.

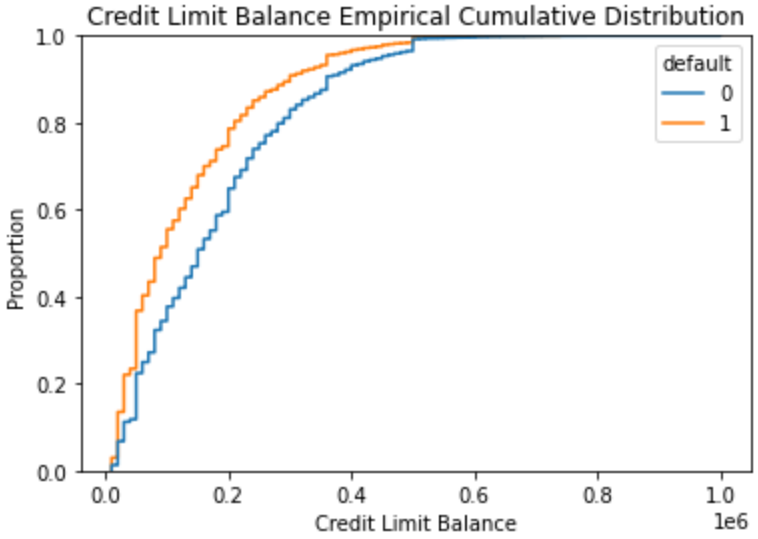
Data Understanding

The dataset used for this project includes actual credit card records at points in time in the past. The credit cards account features include demographic data of the account holder (age, gender, marital status and education), as well as the account holder's credit limit, and multiple monthly values of how long the payments are overdue, the debt outstanding, and the amounts of previous payments. Appendix A includes full details on the dataset.

Each record in the dataset also includes a label of whether the account holder defaulted the next month (default = 1) or did not default the next month (default = 0). The dataset is imbalanced with respect to the percent of records with default labels (22%) versus records with non-default labels (78%). The Data Understanding analysis utilized this target feature designation to examine differences in features when the target feature was designated as default vs non-default. We looked at differences in the data distributions and statistics for a subset of the default records versus a subset of the non-default records.

The Data Understanding analysis indicates that the demographic differences in the account holders are not distinctive in default cases versus non-default cases. However, the analysis did reveal that a few of the transaction-based features did repeatedly show distinction or significance with respect to the default target feature; namely the credit balance and monthly repayment status features.

The credit limit distributions and descriptive statistics showed distinctions between the default records and the non-default records. The histograms of the credit limit variables show that the default records have a greater proportion of lower credit limit accounts than the default subset. The below cumulative distribution plot also shows this distinction in credit limit values. It shows that across the entire distribution customers that defaulted had lower credit limits than customers that did not default. The defaulted account holders were likely given a lower account limit due to having a lower credit score at the time the account was opened.



The monthly repayment status distributions and descriptive statistics also showed distinctions between the default records and the non-default records. The distributions show that the subset of default records have higher counts in those variables that represent that the account holder was late on repaying the account balance than the subset of non-default records. In fact, even though there are 3.5 times as many non-default records as default records, the default subset has higher counts than the non-default subset in some instances. The descriptive statistics also show this distinction. The mean of the repayment status is higher for default records in every month, but the means of the default subset bill amounts, and previous payment variables are less than the means of the non-default subset in every month.

Point-biserial correlation is used for correlation between the binary target variable and the continuous predictive variables. None of the features have strong point-biserial correlations with the default target feature. However, the correlation of the most recent repayment status value is more closely correlated with the default target variable than any of the other features. Correlations between the other features were similarly weak. Although the different types of features are not strongly correlated, there is high correlation between the six different months of values for each of the repayment status, bill amount,

and previous payment amount features. Because these variables are so highly correlated (i.e., nearly the same value in six different variables), there is a multicollinearity problem.

Finally, as previously mentioned, the dataset is imbalanced with respect to the default target variable with only 22% of the records labeled as defaulting in the next month.

Data Preparation

The histograms show that of the six different months of payment information only one of those months show counts for payment status = 1 (payment is 1 month late) but the counts for all six months show over 2,000 counts for payment status = 2 (payment is 2 months late). This seems like a discrepancy because a payment cannot be two months late in August, for example, but not be one month late in July. Therefore, we assigned a value representing one month late in any record in which the next month's value represented that the payment was two months late.

Because there was a diverse range of numeric values in the quantitative variables, the quantitative variables were scaled. Because there was skew in some of the quantitative features, the quantitative features were log-transformed.

The goal of feature selection is to find a reduced subset of the input features in which maximum redundant and irrelevant information is eliminated (Alam, 2020). The number of highly correlated features resulting from six months of data for three different features were addressed by performing a Principal Component Analysis (PCA) on the quantitative features. The PCA analysis looks at linear relationships between the different types of features and combines variables to retain the most variance between the variables, thus addressing redundancy and multicollinearity (Field, 2012). The PCA model reduced the 20 quantitative features to 8 features.

Feature selection also includes determining which predictive features will be used as inputs in the model. Identifying the key features in a dataset that results in accurate predictions is a goal of the predictive model. The predictive models were tested with key features identified by an analysis of variance test and by the feature importance statistics from logistic regression, decision tree classifier, random forest classifier, and XGBoost classifier models that identified the most important features with

respect to the default target feature. The feature importance results from these models confirmed that the demographic variables are not important for the prediction and showed which of the PCA components were most important.

Modeling

Various classification techniques have been used for credit card default predictions, including Logistic Regression, Random Forest, K-Nearest Neighbor, Logistic Model Tree, Adaboost, Stacking, Gradient Boosted Decision Tree, XGBoost, and Neural Networks (Alam, 2020) (Rosen, 2020). The models used for this analysis are: (1) Logistic Regression, (2) LASSO (least absolute shrinkage and selection operator) Logistic Regression, (3) Gradient Boosting Model (GBM), (4) Random Forests, and (5) Artificial Neural Network (ANN).

The data has three categorical features. Because many statistical models can handle only numerical attributes, the categorical features were converted to numerical dummy features through one-hot encoding. Some research indicates that the categorical features should not be converted to dummy numerical features for models involving trees as these models do not involve calculation of distances (Comparative Analysis of Machine Learning Techniques for Detecting Insurance Claim Fraud, n.d.). Thus, the Random Forest model was run without one-hot encoding.

Logistic Regression / LASSO Logistic Regression / GBM / Random Forests

This section discusses the model evaluation and selection for the Logistic Regression, LASSO Logistic Regression, GBM and Random Forests models. The following section discusses the ANN model.

Phase One

Model evaluation includes assessing the general performance of a predictive algorithm and evaluating how well the model will work in the real world. Thus, the first phase of modeling included evaluating these four classification algorithms with baseline models. These models employed a grid search for tuning the four models' hyperparameters to find the best version of the model and by using

stratified k-fold cross-validation instead of the holdout method to split the data in order to address bias and variation issues with the holdout method. All of the input features were used in the models during this first phase of modeling. The accuracy scores from the baseline models indicate that they are all viable models. The log error of the models was also tracked because of the imbalance in the dataset.

Phase One CV Grid Models	Accuracy	Log Error
Logistic Regression	.7911	.4674
LASSO Logistic Regression	.7911	.4672
GBM	.7922	.4590
Random Forest	.7956	.4712

Phase Two

The second phase of modeling determined if the baseline models would be improved by using subsets of the select features identified in the feature selection analysis discussed above. Eight subsets of features were tested on a simple Logistic Regression model with an accuracy score. The best subset used six of the PCA variables and none of the categorical demographic variables. Using the best subset, these four classification models again employed a grid search for tuning the four models' hyperparameters and stratified k-fold cross-validation for splitting the data.

The performance of GBM and Random Forest models in the second phase improved over the baseline models, but the performance of the logistic regression models did not change to much degree.

Phase Two CV Grid Models	Accuracy	Log Error
Logistic Regression	.7904	.4690
LASSO Logistic Regression	.7904	.4690
GBM	.7992	.4524
Random Forest	.7972	.4694

We then used the best parameters from the grid search from these best subset models for the next model evaluation phase.

Phase Three

The next phase of our model evaluation is to address the imbalance in the dataset by selecting a threshold (other than the standard threshold of 0.5) for classifying the default target variable.

Classification predictive modeling involves predicting a class label for records, but models with imbalanced target variables can address the imbalance with the prediction of a probability of class membership (Brownlee, 2020) (Evgeniou & Zoumpoulis, n.d.) (Rosen, 2020).

We used the best parameters from the grid search from Phase Two to make predicted probabilities for positive outcome predictions of the target classification and used a Receiver Operating Characteristic (ROC) test to select the best classification threshold (Rosen, 2020). The probability predictions were then allocated between 1 and 0 (the binary classification values) based on the best threshold. Thus, we now have classification predictions with tuned parameters using an optimized classification threshold for an imbalanced target variable.

We evaluated these threshold optimized classification predictions with Recall and Log Loss Error evaluation metrics. A classification model is evaluated by Recall scores when, as in credit card default cases, there is a high cost associated with false negative predictions (Rosen, 2020). Recall scores calculate how many of the actual positive default cases are labeled as true positives by the model (Classification: Precision and Recall, n.d.) (Saxena, 2018). Because we are predicting probability of class membership, we use a probability metric, such as the Log Loss function, as it summarizes how well the predicted distribution of class membership matches the known class probability distribution (Brownlee, 2020) (Dembla, 2020). The best Recall score is 1.0 and best Log Loss Error is 0.0.

Phase Three Prob/Threshold	**Recall (overall)**	**Recall (Actual Defaults)**	**Log Loss Error**
Logistic Regression	.6850	.6067	.4703
LASSO Logistic Regression	.6850	.6062	.4704
GBM	.6950	.6243	.4545
Random Forest	.6850	.6022	.4605

Artificial Neural Network

This section discusses the model evaluation and selection for artificial neural networks (ANN). We also employed a deep learning ANN to see if it would improve on the above classification models.

We addressed the imbalance in the target variable in the neural networks by weighting each of the target's binary classifications.

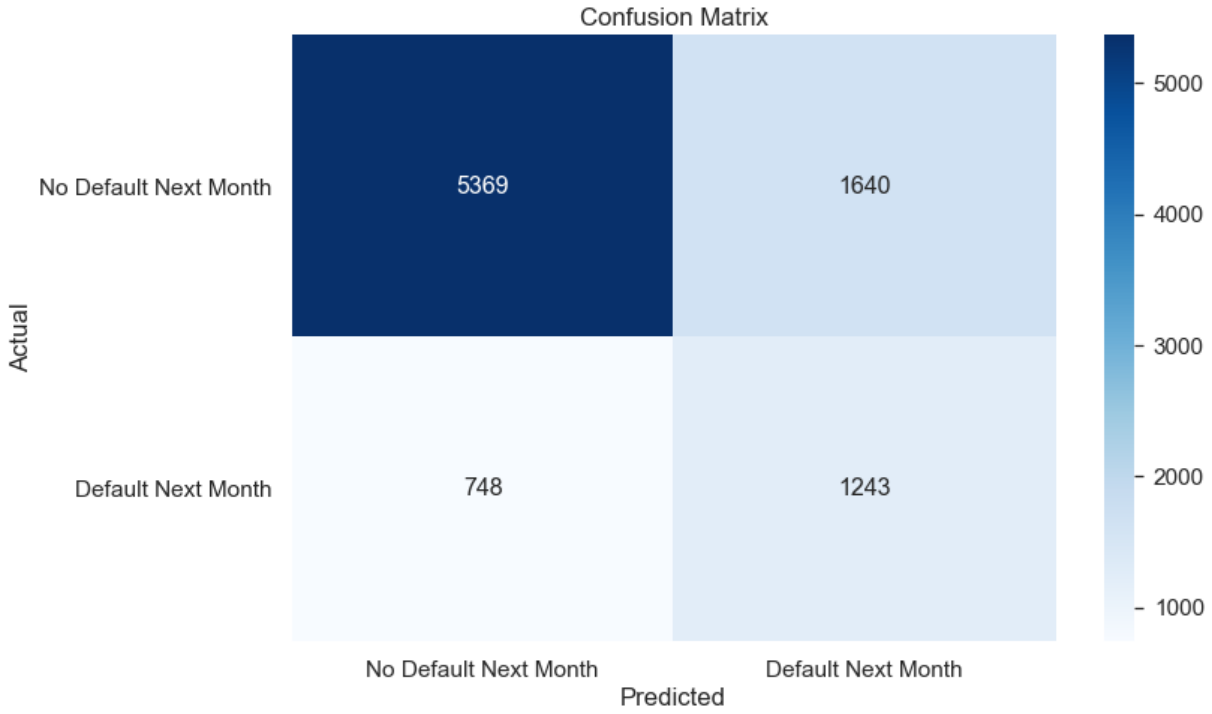
The first phase of modeling for the ANN was to run a baseline Sequential ANN model with a large number of epochs and small batch size and to employ validation data. With these results, we determined the optimal number of epochs to use to avoid overfitting. We then ran the model again with the adjusted epochs.

The next phase was to compare the above ANN model with a version with a smaller number of epochs but a larger batch size. We again tested the optimal number of epochs and ran another model with the adjusted number of epochs. The results improved with this second version of the ANN.

ANN Version	Accuracy	Recall (overall)	Loss
First baseline (overfitted)	.8074	.2677	.4523
First optimized	.6989	.6861	.5958
Second baseline (overfitted)	.7293	.6278	.5855
Second optimized	.7259	.6419	.6419

Results

Regarding the non-neural network models, we evaluated the Phase Three models, which used the best identified subset of input features, best parameters, and optimized classification threshold. The Gradient Boosting Classification (GBC) model had the best Recall and Log Loss Error scores. 62.43% of the actual default accounts were labeled as true positives by the Gradient Boosting model and overall, 69.50% of the default and non-default actual positives were identified correctly. This is a favorable result for default prediction cases because false negative predictions are costly to the credit card company. The GBC Log Loss Error loss score, which is indicative of how close the prediction probability is to the corresponding actual/true value, was also the best score among the four non-neural network models. The below confusion matrix of the GBC model also showed the best accuracy of predicting true positives of the default classification: 62.43% (1,243 of 1,991). This confusion matrix demonstrates the accuracy of the predictions made by comparing predicted classes and true classes.



The results of the artificial neural network model shows that the neural network outperformed three of the non-neural models with respect to Recall, with only the Gradient Boosting model performing better than the neural network model. The best neural network model has a Recall score of 68.61% and a binary crossentropy loss of 59.58%.

Discussion/Conclusion

The ANN model and the GBC model are viable and useful models for predicting credit card defaults. The Logistic Regression and Random Forests models performed nearly as well as the neural network and GBC models. The models are viable but we will need additional tuning of hyperparameters and perhaps a different approach to the imbalanced target variable.

In conclusion, we have determined that we have viable models for predicting credit card defaults. We have further identified the next project for credit card defaults to include (1) additional tuning of hyperparameters for the models created for this project, (2) testing other approaches to target variable

imbalance in the models, (3) employing new methodologies to address the primary contributing factors to default probability since the PCA components are not helping in analyzing individual factors, and (4) implementing a deployment strategy of making actual predictions and/or converting the probability results to a credit score that can be used in making lending decisions.

References

15 Shocking Credit Card Debt Statistics. (2020, August 10). Retrieved from CardRates.com: https://www.cardrates.com/advice/shocking-credit-card-debt-statistics/

Alam, T. M. (2020). An Investigation of Credit Card Default Prediction in the Imbalanced Datasets. *IEEE Access*, Vol. 8 201173-201198.

Brownlee, J. (2020, January 14). *A Gentle Introduction to Probability Metrics for Imbalanced Classification*. Retrieved from machinelearningmastery.com: https://machinelearningmastery.com/probability-metrics-for-imbalanced-classification/

Charge-Off and Delinquency Rates on Loans and Leases at Commercial Banks. (2021, February 21). Retrieved from federalreserve.gov: https://www.federalreserve.gov/releases/chargeoff/

Classification: Precision and Recall. (n.d.). Retrieved from Machine Learning Crash Course: https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall

Comparative Analysis of Machine Learning Techniques for Detecting Insurance Claim Fraud. (n.d.). Retrieved from wipro: https://www.wipro.com/en-US/analytics/comparative-analysis-of-machine-learning-techniques-for-detectin/

Dembla, G. (2020, November 17). *Intuition behind Log-loss score*. Retrieved from towardsdatascience.com: https://towardsdatascience.com/intuition-behind-log-loss-score-4e0c9979680a#:~:text=is%20dependent%20on.-,What%20does%20log%2Dloss%20conceptually%20mean%3F,is%20the%20log%2Dloss%20value.

Evgeniou, T., & Zoumpoulis, S. (n.d.). *Classification for Credit Card Default*. Retrieved from http://inseaddataanalytics.github.io/: http://inseaddataanalytics.github.io/INSEADAnalytics/CourseSessions/ClassificationProcessCreditCardDefault.html

Field, A. M. (2012). *Discovering Statistics Using R.* Los Angeles: Sage Publications Ltd.

Rosen, D. B. (2020, August 1). *How To Deal With Imbalanced Classification, Without Re-balancing the Data*. Retrieved from towardsdatascience.com: https://towardsdatascience.com/how-to-deal-with-imbalanced-classification-without-re-balancing-the-data-8a3c02353fe3

Saxena, S. (2018, May 11). *Precision vs Recall*. Retrieved from medium.com: https://medium.com/@shrutisaxena0617/precision-vs-recall-386cf9f89488#:~:text=While%20precision%20refers%20to%20the,correctly%20classified%20by%20your%20algorithm.&text=For%20problems%20where%20both%20precision,maximizes%20this%20F%2D1%20score.

Appendix A

This is the source of the data:

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Here is a data dictionary of the dataset:

- ID: ID of each client
- LIMIT_BAL: Amount of the given credit in NT (New Taiwan) dollars (includes individual and family/supplementary credit)
    - For reference, 1 US dollar = 28.57 NT dollars on 3/19/2021.
- SEX: Gender (1=male, 2=female)
- EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
- MARRIAGE: Marital status (1=married, 2=single, 3=others)
- AGE: Age in years
- PAY_0: Repayment status in September, 2005
- PAY_2: Repayment status in August, 2005 (scale same as above)
- PAY_3: Repayment status in July, 2005 (scale same as above)
- PAY_4: Repayment status in June, 2005 (scale same as above)
- PAY_5: Repayment status in May, 2005 (scale same as above)
- PAY_6: Repayment status in April, 2005 (scale same as above)
    - The above Pay variables have the following values: -2 means no account usage, -1 means balance paid in full, 0 means at least the minimum payment was made, 1 means payment delay for one month, 2 means payment delay for two months, etc. up to 9, which means payment delay for nine months and above.
- BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)
- BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)
- BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)
- BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)
- BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)
- BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)
- PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)
- PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)
- PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)
- PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)
- PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)
- PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)
- default.payment.next.month: Label for default payment (1=yes, 0=no)