Final Project

Advanced Data Analytics

By: Gina C. Miller

Data Analytics Strategy to Address Credit/Financing Defaults for GE

**The Data Source and Proposed Business Value**

The data set is 1,000 past credit applicants based on 31 variables and was created in response to the financial crisis of 2008-2009 (DAT-650, 2020). The goal of this dataset is to determine the likelihood of default for new credit based on the variables (DAT-650, 2020). The impact of the dollar loss to the company on a default loan by a customer is 150% of the remaining balance (DAT-650, 2020). The calculation for the cost-benefit of this model is as follows: The current default rate is 30%. The average loan balance is $3,271.26. The cost of the default of the loans at 150% of the loan amount averages $4,906.88. The cost is $981,377.40 over 1,000 loans. If we could lower the default rate to 25%, we would save $245,344.50 . This is a significant amount.

According to Shearer, the CRISP-DM model (CRoss-Industry Standard Process for Data Mining) encourages best practices and offers organizations a structure needed to realize better and faster results from its data mining (Shearer, 2000). The first two phases are Business Understanding and Data Understanding and arrows go between both (Shearer, 2000). Shearer describes Phase 1, Business Understanding, as possibly being the most important phase (Shearer, 2000). It is important for the data analysts to work with the project team and those with specific domain business knowledge, so we are all clear on what the objective is, and everyone knows how to address it. There are four steps to be followed during Phase 1: determine the business objective, assess the situation, determine data mining goals, and produce a project plan (Shearer, 2000). In this situation, we know the problem is defaults on credit loans and we are trying to determine if the existing dataset will aid in creating a predictive analytic model to detect credit-risk applicants.

Phase 2, Data Understanding is where we look at the data and identify any problems, missing data, gain initial insights and identify if the data is of good quality (Shearer, 2000). As we gain more insight into the data, we can increase our business understanding and our knowledge as to whether or not this data is complete enough to accomplish the task at hand. Tasks include a data visualization analysis where we could identify specific patterns. One of the first things a data analyst will do, during this phase, is initial Exploratory Data Analysis with some basic statistical summaries as well as visualizations such as bar graphs and line charts.

Predictive analytics, as the name implies, is using a variety of techniques such as data mining, statistical modeling, and machine learning algorithms to try and predict what will happen in the future. By looking at the data patterns and trends in the historical data, we can try these techniques to see if we can predict if a person will default.

Logistic regression, linear programming, and discriminant analysis are the current industry standard predictive analytic tools used to identify credit-risk applicants (Oskardottir, 2018). With the advent of machine learning techniques, neural networks, support vector machines, and random forests have been added to the mix (Oskardottir, 2018). The traditional demographics for credit-risk scoring are between 8 and 15 variables and consist of demographics (age, time at residence, time at job, postal code), existing relationship (time at bank, number and type of products, payment performance, previous claims), credit bureau (inquiries, trades, delinquency, public records) and real estate data (Siddiqi, 2006). Also the use of social media data is being used more often instead of only traditional scores (Siddiqi, 2006). Rosella Machine Intelligence and Data’s website mentions the use of decision trees, random forests, and logistic regression as common predictive analytic techniques for credit risk analysis (roselladb.com, n.d.). Parasarathy, a data scientist, states that random forest is a particular machine learning technique that is used to identify if someone will default on a loan (Parasarathy, n.d.). Random forest in Rattle (R statistical programming software) is a predictive analytic model where you build multiple decision trees into a “forest” of trees into a single model (Williams, 2011). According to Williams, the random forest algorithm “tends to produce quite accurate models” because it reduces the instability observed in a single decision tree (Williams, 2011). It is more robust to changes in the data, more robust to noise (variables that have little relationship to the target variable) and works well with large datasets (Williams, 2011). The idea is rather than just one decision tree, now we have a forest of decision trees and “many models working together are better than one model doing it all”, as Williams states (Williams, 2011). For this reason, I believe a random forest model will work well with the dataset provided.

There are a couple more data fields that would be helpful in this model. The first is income, which is not included at all in the dataset. Another helpful field of data would be “cash-flow” data (doddfrankupdate.com, 2019). A study by the FinRegLab.com shows that cash-flow data was a more effective predictor of creditworthiness than traditional credit scores (doddfrankupdate.com, 2019).

Phase 3 of CRISP-DM is data preparation and this data set is fairly “clean”. If we gain new data fields they can be added in and certain fields were removed for ethical considerations.

Phase 4, modeling, is what we have done. A random forest model was chosen and run.

Phase 5 is evaluation and as you will see, we performed many different evaluation metrics for the model.

Finally, Phase 6 is deployment of the model, after the project is complete. This is done on new data and leads to more business understanding and actually starts the CRISP-DM process again.

**Ethical Concerns and Implications**

There are some ethical concerns with a few of the fields in the dataset. While many countries do allow the use of age and gender in credit scorecards, this should not be a regular practice as it could lead to discriminatory lending practices (Baer, 2017). Ethnicity, gender, age, religious affiliation, and marital status should not be used in a dataset at all, and this extends to using those variables from social media also. In this particular data set, education could be discriminatory, and I suggest not using that. A high education may imply a higher paying job, however, as the FinRegLab.com research study shows, looking at cash-flow data is much better than some traditional fields used in consumer credit reports (doddfrankupdate.com, 2019). For obvious reasons, the variables MALE\_DIV, MALE\_SINGLE, MALE\_MAR\_WID should be discarded (this includes the gender of the applicant and his marital status which could lead to discrimination). I would also exclude age, it is possibly discriminatory, and as with education, I would exclude the JOB variable which is the nature of the job such as management, unskilled, or a highly qualified employee. I refer back to the study done by FinRegLab.com in that it is more useful to see the cash flow data rather than assume somebody that is “highly qualified” is more creditworthy than an unskilled laborer (doddfrankupdate.com, 2019). I would also remove the FOREIGN variable which indicates whether or not the applicant is a foreign worker. I believe that indicating someone as foreign is discriminatory because it indicates ethnicity. It is especially important to disregard those variables that could lead to discriminatory lending practices that are not only illegal but would ruin many organizations reputation and increase reputational risk (Baer, 2017).

**Model Creation**

In a study of predictive models for loan default risk assessments, Coser compared various classification models such as Light Gradient Boosted Model, Extreme Gradient Boosted Model, Logistic Regression, and Random Forest (Coser, 2019). Using public Lending Club data, they discovered they achieved the best results using the random forest model (Coser, 2019). Another study, by Zhu, also using Lending Club data, found that the random forest model, compared with decision tree, logistic regression, and support vector machine, performed best (Zhu, 2019). The random forest model had a 98% accuracy rate compared to the decision tree 95% accuracy rate (Zhu, 2019). Zhu notes that random forest models have significant advantages such as it runs efficiently on large databases, fixes well with errors in class population unbalanced datasets, and it has an effective method for estimating missing data and still maintaining accuracy (Zhu, 2019). Wu performed a study comparing random forest, decision tree, neural networks, and support vector machines (Wu, 2016). They used random forest to evaluate the financial variables and construct a credit rating prediction model (Wu, 2016). Random forest is able to select critical variables by analyzing credit rating indicators that are highly accurate (Wu, 2016). In the study, random forest was the best performing model. (Wu, 2016). The worst was support vector machines with only 67.3% accuracy (Wu, 2016). As noted, just decreasing the default rate by 5% is a substantial savings to the organization and I recommend a random forest model for this dataset.

**Model Value for Organization**

There has been current research in using machine learning methods for predicting loan defaults and credit risk assessments rather than just the FICO credit score and other scores determined by a few credit reporting agencies (Zhu, 2019). This is due to an increased need for a better understanding of customer behavior and profile (Coser, 2019). With an increased interest in data exploration and analysis in the financial/banking field, the ability to discriminate between clients as “good” or “bad” borrowers is important for banks and other financial institutions (Coser, 2019). Even a slight improvement in the accuracy of prediction may result in a considerable increase in profitability (Coser, 2019). Early identification of customers who have a significant risk of falling into default may help financial institutions prevent bad loans (Coser, 2019). In the third quarter of 2018, the delinquency rate on consumer loans was 2.28% (Coser, 2019). The overall default rate for consumer loans in the third quarter of 2009 reached an historical high of 10.1% (Coser, 2019). Currently, with the Coronavirus pandemic, the unemployment rate as of June 2020 is 11.1% (bls.gov, 2020). In April 2020, Forbes reported a 12% consumer loan impairment (Kauflin, Gara, 2020). The Mortgage Bankers Association reported a 4.36% default on mortgage loans in the first quarter of 2020 (mba.org, 2020). It cannot be overstated that the assessment of consumer risk of default is particularly important to a financial institutions profitability.

Since banks do make money on lending money, the fundamental problem becomes the assessment of a client’s creditworthiness (Coser, 2019). Based on this research, I used a random forest model, in the statistical software package R and Rattle, as the pilot plan/model.

**Model Creation**

The random forest model was tested with the marital status, gender, job, and ethnicity attributes removed. The out-of-bag, OOB, estimate of error rate was 24.57%. This means that when the model is applied to new datasets, the answers will be in error 24.57 % of the time (Williams, 2009). The confusion matrix shows the disagreement between the model’s predictions and actual outcomes (Williams, 2009). Below is the confusion matrix:

OOB estimate of error rate: 24.57%

Confusion matrix:

0 1 class.error

0 427 63 0.1285714

1 109 101 0.5190476

Interpreting this is not that difficult. The model and training dataset agree that for 427 of the observations, the applicant will not default and 101 will default (Williams, 2009). However, there are 109 applicants that the model predicted will not default but did and 63 predicted will default but did not. The model predicts someone will default, when they will not, 52% of the time. However, only about 13% of the time does it predict someone will not default and they do.

The area under the curve is 0.6762 which is not great, but not bad either, considering our dataset is imbalanced. There are more “0s” than the “1s” we wish to predict. The error matrix is:

Error matrix for the Random Forest model on Credit Data.csv [validate] (counts):

Predicted

Actual 0 1 Error

0 104 6 5.5

1 21 19 52.5

Error matrix for the Random Forest model on Credit Data.csv [validate] (proportions):

Predicted

Actual 0 1 Error

0 69.3 4.0 5.5

1 14.0 12.7 52.5

Overall error: 18%, Averaged class error: 29%

In this case, 52.5% of the time, the model predicted one would default when one did not. Only 5.5% of the time did the model predict an applicant would not default and they did. In that regard, it may be more important to accurately predict the defaulters then those that would not.

**Pilot Plan Results and Evaluation**

I used a lift curve which assesses how well the model is performing in identifying the positive (a person will default, 1’s) instances:

A map of a person

Description automatically generated

The maximum lift is 4. At 10% of the population have the highest probability of defaulting 4 times more than average.

Lift is calculated as the ratio of positives (1s) divided by the ratio of 1s in the entire dataset (Zornova, 2020). The maximum lift is 4. At 10% of the observations in the dataset, 40% of the observations are more positive than average. That is, 10% of the population have the highest probability of defaulting 4 times more than the average. At 50%, about 1 ½ of the observations have the highest probability of defaulting, 1.5 times more than the average.

Next I evaluated the model with a Precision/Recall curve:

A close up of a map

Description automatically generated

The Precision/Recall chart is a good evaluation for a dataset that is imbalanced (Brownlee, 2018). The reason is that when calculating precision and recall, it does not make use of the true negatives (Brownlee, 2018). It is only concerned with the correct prediction of the positive minority class (Brownlee, 2018). In this dataset, 700 observations do not default compared to 300 that do. The important area of this chart bows at the (1.0,1.0) point. While this is not the perfect precision/recall chart, it is not terrible considering how much money can be saved by accurately predicting just a small percentage of defaulters.

Another good evaluation chart is the Performance Risk Chart:

A close up of a map

Description automatically generated

According to this chart, 27% of the observations are of interest (Williams, 2009). The diagonal line in the plot is a random 50% caseload and indicates a 50% performance (only half the observations of interest will be found) (Williams, 2009). The dashed green line is the performance achieved when using the model to predict defaults (Williams, 2009). At 50% of the observations, we catch 82% of defaulters. The blue line indicates the lift of this model (Williams, 2009). The lift is about 2.5 times or twice as likely to predict a default.

The last chart is the ROC or “receiver operating curve”. This curve is a favorite in evaluating the performance in a model. A random model has an area under the curve of 0.5. A perfect model has an area under the curve of 1.0. The following chart is the ROC and AUC for our random forest model:

A close up of a map

Description automatically generated

The AUC is close to 80%. This is an indicator of a strong classifier, however, Brownlee and Takaya suggest that the ROC for an imbalanced dataset can be “deceiving” (Brownlee, 2018; Takaya, 2015). For that reason, I placed this chart last and focused on the error matrix, lift curve, performance risk chart, and precision/recall chart.

The variables of importance used in this model were: checking account, duration of the loan, amount of the loan, credit history, guarantor, installment rate, savings account, age, number of credits, and employment. While age did show up as an important variable, it should not be used in the final model since it could be discriminatory.

**Project Proposal Full Implementation**

Data?

What data is available?

We have 1,000 credit loans with a 30% default rate and 31 variables.

Value Carrier

What decision will be augmented?

The dataset has a predictive signal and can help predict who will default on a loan.

Value Proposition?

What value will this model give to the organization?

Current default rate: 30%

Average loan balance: $3271.26

Cost of default: $4906.88

Cost over 1,000 loans: $981,337.50

Savings at 25% default rate:$245,344.50

Data?

What data do we need?

Salary and income data would be helpful.

Metrics?

How will we measure the model’s performance?

OOB error rate, Precision/Recall Curve, Lift Curve, Risk Curve, and ROC curve.

Models?

What model can be derived from the data?

A random forest model did indicate a predictive value.

Stakeholders?

GE, management, and employees are key stakeholders because of the substantial savings. Customers can be stakeholders, benefitting from the savings on defaulted loans.

Costs?

What costs will be incurred?

This will be determined by the project management team however, with a savings of $245,344.50 over 1,000 loans or $2,453,445,000 over 10,000 loans

No predictive model is 100% perfect. But I have demonstrated what the savings to GE will be by decreasing the default rate only 5%. This is a substantial number that saves both the company and customers money because GE will not have to pass on these costs to the customer. The dataset available is enough to show a predictive signal. If we get additional data, in the form of salary or social media data, we can add that too. Using the CRISP-DM process, we will work fully with the project team in understanding the business problem and then using the data we have available to create models that will solve the problem. Additional data will be prepared, additional models will be created, and, ultimately, we will deploy this model for use on new data. As it is used, the CRISP-DM process starts over again as we gain more business understanding, as data changes, and we tune our model to continually improve the model performance.

**References**

Shearer, Colin. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing.* Vol. 5. No. 4. Retrieved from mineracaodedados.files.wordpress.com

Siddiqi, Naeem. (n.d.). Intelligent Credit Scoring 2nd Edition. Wiley and Sons. Retrieved from support.sas.com

Zhu, Lin. Qiu, Dafeng. Et.al. (2019). A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science.* Vol 162. <https://doi.org/10.1016/j.proes.2019.12.017>

Coser, Alexandru. (2019). Predictive Models for Loan Default Risk Assessment. *Economics Computation and Economic Cybernetics Studies and Research.* Vol 53. Iss 2. Retrieved from <https://doi.org/1024818/18423264/53.2.19.09>

Wu, Hsu-Che. Wu, Yu-Ting. (2016). Evaluating credit rating prediction by using the KMV Model and R. *Kybernetes.* Vol 45. Iss. 10. Retrieved from www-emerald-com.exproxy.snhu.edu/insights/10.1108/k-12-2014-0285-2016

(Sayad, n.d.) Model Evaluation. Retrieved from web.archive.org/

Williams, Graham. (2009). Data Mining with Rattle and R. Springer Science+Business Media LLC. New York, New York.

Kauflin, Jeff. Gara, Antoine. (2020). Exclusive: The Coming Default Crisis with Online Loans. Retrieved from forbes.com

Bureau of Labor Statistics. (2020). Current Unemployment Rate. Retrieved from bls.gov

Mortgage Banking Association. (2020). Current Mortgage Loan Default Rate. Retrieved from mba.org

Brownlee, Jason. (2018). How to Use ROC Curves and Precision Recall Curves for Classification in Python. Retrieved from machinelearningmastery.com

Takaya, Saito. (2015). The Precision-Recall Plot is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS One.* Vol. 10. Issue 3. Retrieved from ncbi.nlm.nih.gov

Zornova, Jaime. (2020). The Lift Curve: Unveiled. Retrieved from towardsdatascience.com

AI Pilot Canvas. Retrieved and adapted from [www.scaleoutsystems.com](http://www.scaleoutsystems.com)

Baer, Tobias. (2017). Social Media and Credit Scoring: Hype or the Future? Retrieved from linkedin.com

Author Unknown. (2019). Applying predictive analytics to credit scoring, beyond. Retrieved from doddfrankupdate.com

Rosella (n.d.). Rosella Machine Intelligence and Data Mining. Retrieved from roselladb.com

Parasarathy, Sriram. (n.d.) Top 5 Predictive Analytic Models and Algorithms. Retrieved from logianalytics.com

SNHU. (2020). DAT-650 Use Cases Document. *Southern New Hampshire University.*

Manchester, New Hampshire.