Debt Portfolio Capstone

Author: Mark Yingling

Wednesday, 17 October 2018

[1 Problem Statement 3](#_Toc527543242)

[2 Gather Data 4](#_Toc527543243)

[3 Exploratory Data Analysis (EDA) 5](#_Toc527543244)

[3.1 Data Dictionary 5](#_Toc527543245)

[3.2 Data interpretation and the workflow 6](#_Toc527543246)

[3.3 Feature Engineering 6](#_Toc527543247)

[4 Model Pre-processing 7](#_Toc527543248)

[4.1 Scaling Data 7](#_Toc527543249)

[4.2 Leave-One-Out Cross Validation 7](#_Toc527543250)

[4.3 Stratified K-fold Cross Validation 7](#_Toc527543251)

[4.4 SMOTE (Synthetic Minority Over-sampling Technique) 7](#_Toc527543252)

[4.5 GridSearch 7](#_Toc527543253)

[5 Model the data 8](#_Toc527543254)

[5.1 Logistic Regression 8](#_Toc527543255)

[**5.1.1** Confusion Matrix Evaluation Metrics 8](#_Toc527543256)

[**5.1.2** ROC Curve 8](#_Toc527543257)

[**5.1.3** Cross Validation Score 8](#_Toc527543258)

[**5.1.4** Stratified K-fold Cross Validation 8](#_Toc527543259)

[**5.1.5** IBM Watson Studio Model Evaluation 8](#_Toc527543260)

[**5.1.6** Overall Evaluation 8](#_Toc527543261)

[5.2 DecisionTreeClassifier 9](#_Toc527543262)

[**5.2.1** Confusion Matrix Evaluation Metrics 9](#_Toc527543263)

[**5.2.2** ROC Curve 9](#_Toc527543264)

[**5.2.3** Cross Validation Score 9](#_Toc527543265)

[**5.2.4** Stratified K-fold Cross Validation 9](#_Toc527543266)

[**5.2.5** Overall Evaluation 9](#_Toc527543267)

[5.3 RandomForestClassifier 9](#_Toc527543268)

[**5.3.1** Confusion Matrix Evaluation Metrics 9](#_Toc527543269)

[**5.3.2** ROC Curve 10](#_Toc527543270)

[**5.3.3** Cross Validation Score 10](#_Toc527543271)

[**5.3.4** Stratified K-fold Cross Validation 10](#_Toc527543272)

[**5.3.5** Overall Evaluation 10](#_Toc527543273)

[**5.3.6** IBM Watson Studio Model Evaluation 10](#_Toc527543274)

[5.4 GradientBoostingClassifier 10](#_Toc527543275)

[**5.4.1** Confusion Matrix Evaluation Metrics 10](#_Toc527543276)

[**5.4.2** ROC Curve 11](#_Toc527543277)

[**5.4.3** Cross Validation Score 11](#_Toc527543278)

[**5.4.4** Stratified K-fold Cross Validation 11](#_Toc527543279)

[**5.4.5** Overall Evaluation 11](#_Toc527543280)

[**5.4.6** IBM Watson Studio Model Evaluation 11](#_Toc527543281)

[5.5 KNeighbors 11](#_Toc527543282)

[**5.5.1** Confusion Matrix Evaluation Metrics 11](#_Toc527543283)

[**5.5.2** ROC Curve 12](#_Toc527543284)

[**5.5.3** Cross Validation Score 12](#_Toc527543285)

[**5.5.4** Stratified K-fold Cross Validation 12](#_Toc527543286)

[**5.5.5** Overall Evaluation 12](#_Toc527543287)

[6 Evaluate the Model 13](#_Toc527543288)

[7 Risks and Limitations 14](#_Toc527543289)

[8 Answer the Problem 15](#_Toc527543290)

# Problem Statement

Debt is a large problem in the United States. Many people have fallen into large amounts of debt much of which have a staggering interest rates attached to it. A significant number of these people have chosen not to repay this debt. This can be attributed to any number of reasons (loss of employment, other costs taking precedence, with the interest being so high they don't see any reduction in the principle amount, etc…) After a period of time trying to collect on the debt creditors will write this debt off and sell it in hopes of recouping some of what is owed. The process by which debt is sold comes in the form of debt portfolios i.e., a number of different debtors are bound together and sold as a percentage of the total amount owed. This could me pennies on the dollar, but the amount paid varies on the amount of information given to the buyer about the debtors included in the portfolio. For instance, all statements and correspondence each debtor has had with the selling creditor. The more data provided the more the portfolio will cost, however the more data you have the better you will be able to predict if someone will eventually repay what is owed.

Getting people who have a demonstrated a propensity of ignoring their debts to pay is the key to success when purchasing portfolio s of debt. How then do we do this? It is our belief that people in this position would like a way out. Either to just have the load of the debt off their shoulders, or to get their credit back so they my buy a house, car, etc… This was implemented successfully with this portfolio by forgiving the interest and fees currently associated with the loan and reducing the interest rate to prime plus 3 for debtors meeting the following criteria:

* Pass a bankruptcy check
* Lives in the Denver metro local
* Remaining balance < 15k
* Didn’t look at interest rates of 8%
* Married and own their home

Can we accurately identify if the portfolio is good one to purchase by determining the Likelihood of someone repaying their debt?

* What were your findings?
* What risks/limitations/assumptions affect these findings?

# Gather Data

We have a portfolio of debt that was purchased on October 24, 2008. This portfolio is comprised of 42 debtors with a total debt amount of $182,222.36. The cost of the portfolio was $6,000.

The data came in the form of two paper spreadsheets. I used Abbyy FineReader, which is an optical character recognition (OCR) piece of software that allowed me to read the paper documents into a .csv file. From there I needed to separate several columns that the software merged together because of several reasons; some of the cells were highlighted, there were several hand written notes, and the pages were taped together so that the rows could be seen with all the columns. I also needed to fix several of the date, name, and address fields for the same reasons as above.

# Exploratory Data Analysis (EDA)

Upon the initial look at the two spreadsheets it became clear that they were the same group of accounts. When the portfolio was purchased two people were doing the analysis. I’m guessing that the two spreadsheets were each individual’s work product. The first data set had less missing data so we chose to work with it.

## Data Dictionary

There are 42 rows of data with 28 columns

|  |  |  |
| --- | --- | --- |
| Column Name | Data Type | Description |
| Account Number | int64 | Issuers account number |
| PortID | object | Issuer id |
| Remaining Balance | float64 | The remaining balance owed |
| Interests Fees | float64 | Interest and fees associated with the account |
| Date Opened | datetime64[ns] | Opening date for the card |
| Charge Off Date | datetime64[ns] | When account was written off by the issuer |
| Last Pay Date | datetime64[ns] | Last date a payment was made on the loan |
| Last Pay Amount | float64 | The amount last paid |
| Last Activity Date | datetime64[ns] | The date of the last payment |
| Interest Rate % | float64 | Interest rate of the loan |
| Issuer | object | The issuer of the credit card i.e., Chase, or some other bank |
| Merchant | object | This is the company that offers the card to its customers i.e., United Airlines Mileage Plus or the bank itself |
| FCFRA Date | datetime64[ns] | Fair Credit Reporting Act – This is the last time the issuer reported to the credit bureau. |
| OOS Date | datetime64[ns] | Out of Statute – Six years after the last payment made by the consumer |
| Account Type | object | Type of debt portfolio |
| Last Name | object | Debtors last name |
| First Name | object | Debtors first name |
| Middle Name | object | Debtors middle name |
| Address | object | Debtors address |
| City | object | Debtors city of residence |
| State | object | Debtors home state |
| Zip | object | Debtors zip code |
| County | object | Debtors home county |
| SSN | float64 | Debtors social security number |
| Home Phone | float64 | Debtors home phone number |
| Employer Phone | float64 | Debtors employers phone number |
| Cycle | object | How “fresh the account is i.e., how many times did the issuer send it out to a collection agency prior to sale |
| Paid | int64 | Was the debt paid |

## Data interpretation and the workflow

In doing the manual analysis my SME took the following into account:

* Bankruptcy check
* Denver metro local
* Home owner weighted higher then renters
* Remaining balance < 15k
* Didn’t look at interest rates of 8%
* Look to see if they were married
* Looked at the cycle the loan was in
* Calls were also made to each of the debtors and questions were asked to determine their current financial situation.

## Feature Engineering

* Because the Out of Statute date is an important date I subtracted the date the portfolio was purchased from it to determine how many days remained until the loan was out of statute, and then added a column to the dataframe.
* The Cycle is also an important column in determining if someone will pay back the debt. I created dummy values for this column which generated three new columns; 'Cycle\_QUATERNARY', 'Cycle\_SECONDARY', 'Cycle\_TERTIARY'

# Model Pre-processing

Prior to fitting any models some pre-processing of the data will help build better models.

## Scaling Data

Feature scaling can vary your results a lot while using certain algorithms and have a minimal or no effect in others. Because many of the models I will be building are based on the distance between two data points. Not having scaled data will cause problems.

## Leave-One-Out Cross Validation

Because my data set is very small I incorporated the leave-on-out cross validation. Leave-one-out cross validation is K-fold cross validation taken to its logical extreme, with K equal to N, the number of data points in the set. That means that N separate times, the function approximator is trained on all the data except for one point and a prediction is made for that point.

## Stratified K-fold Cross Validation

Stratification is the process of rearranging the data so as to ensure that each fold is a good representative of the whole. For example, in a binary classification problem where each class comprises of 50% of the data, it is best to arrange the data such that in every fold, each class comprises of about half the instances.

Because the data set is small this will Randomly split your entire dataset into k”folds” with each class comprising 50% of the data.

## SMOTE (Synthetic Minority Over-sampling Technique)

Because there is not a lot of data we need to over-sample the data to get more samples of the minority class. SMOTE is an oversampling method which creates “synthetic” example rather than oversampling by replacements. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen.

## GridSearch

Prior to fitting any models, I ran a grid search looking to fine the best hyper-parameters.

# Model the data

## Logistic Regression

The first model I tried was a logistic regression.

### Confusion Matrix Evaluation Metrics

I evaluated the model using model evaluation metrics such as accuracy, precision, and recall.

* Accuracy: Classification rate of for the training set was 90.41%. The test set scored 100%.
* Precision: How accurate is your model. In other words, you can say, when a model makes a prediction, how often it is correct. In this case, the training set scored 89.19% and the test was correct 100% of the time.
* Recall: If there are debtors who have paid in the training set this model can identify them 91.67% of the time. In the test set it can identify them 100% of the time

### ROC Curve

Receiver Operating Characteristic (ROC) curve is a plot of the true positive rate against the false positive rate. It shows the tradeoff between sensitivity and specificity.

* AUC score for the case was 0.972. AUC score 1 represents perfect classifier.

### Cross Validation Score

The cross-validation score was 86.49%

### Stratified K-fold Cross Validation

The Stratified K-fold cross-validation score was 86.49%

### IBM Watson Studio Model Evaluation

I wanted to see how Watson Studio’s automated model builder would do with this data. It’s version of a Logistic Regression model gave a AUC score of .7, however the overall evaluation was “poor”. The automated model builder doesn’t allow for the user to see the code that creates the model. That being said I was not able to tune it.

### Overall Evaluation

Based on all the scores I think that this might be a good model, even though the test set was 100%. The observations from the test set might come from the fact that there is only one observations in that set. However, the cross-validation and Stratified KFold scores were very good.

## DecisionTreeClassifier

The next model I tried was a single decision tree classifier. I found that there were three important features for this model; “Interests Fees”, “Cycle\_Quanternary”, and “OOS Days Left”.

### Confusion Matrix Evaluation Metrics

I evaluated the model using model evaluation metrics such as accuracy, precision, and recall.

* Accuracy: Classification rate of for the training set was 95.15%. The test set scored 100%.
* Precision: How accurate is your model. In other words, you can say, when a model makes a prediction, how often it is correct. In this case, the training set scored 87.80% and the test was correct 100% of the time.
* Recall: If there are debtors who have paid in the training set this model can identify them 100% of the time. In the test set it can identify them 100% of the time.

### ROC Curve

Receiver Operating Characteristic (ROC) curve is a plot of the true positive rate against the false positive rate. It shows the tradeoff between sensitivity and specificity.

* AUC score for the case is 0.947. AUC score 1 represents perfect classifier.

### Cross Validation Score

The cross-validation score was 86.49%

### Stratified K-fold Cross Validation

The cross-validation score was 77.03%

### Overall Evaluation

Based on all the scores I think that this might be a ok model, even though the test set was 100%. The observations from the test set might come from the fact that there is only one observations in that set. However, the cross-validation and Stratified KFold scores did not match very well.

## RandomForestClassifier

Our next model was a random forest. I found that there were eight important features for this model; “Interest Fees”, “Last Pay Amount”, “OOS Days Left”, “Remaining Balance”, “Cycle\_Quanternary”, “Cycle\_Tertiary”, “Interest Rate”, and “Cycle\_Secondary”.

### Confusion Matrix Evaluation Metrics

I evaluated the model using model evaluation metrics such as accuracy, precision, and recall.

* Accuracy: Classification rate of for the training set was 91.78%. The test set scored 100%.
* Precision: How accurate is your model. In other words, you can say, when a model makes a prediction, how often it is correct. In this case, the training set scored 85.71% and the test was correct 100% of the time.
* Recall: If there are debtors who have paid in the training set this model can identify them 100% of the time. In the test set it can identify them 100% of the time.

### ROC Curve

Receiver Operating Characteristic (ROC) curve is a plot of the true positive rate against the false positive rate. It shows the tradeoff between sensitivity and specificity.

* AUC score for the case is 0.947. AUC score 1 represents perfect classifier.

### Cross Validation Score

The cross-validation score was 94.59%

### Stratified K-fold Cross Validation

The cross-validation score was 90.54%

### Overall Evaluation

Based on all the scores I think that this might be a good model, even though the test set was 100%. The observations from the test set might come from the fact that there is only one observation in that set. However, the cross-validation and Stratified KFold scores were very good.

### IBM Watson Studio Model Evaluation

I wanted to see how Watson Studio’s automated model builder would do with this data. It’s version of a Random Forest model gave a AUC score of .5, however the overall evaluation was “poor”. The automated model builder doesn’t allow for the user to see the code that creates the model. That being said I was not able to tune it.

## GradientBoostingClassifier

The next model was a gradient boost. I found that there were four important features for this model; “Interest Rate”, “Cycle\_Quanternary”, “Last Pay Amount”, and “OOS Days Left”.

### Confusion Matrix Evaluation Metrics

I evaluated the model using model evaluation metrics such as accuracy, precision, and recall.

* Accuracy: Classification rate of for the training set was 95.89%. The test set scored 100%.
* Precision: How accurate is your model. In other words, you can say, when a model makes a prediction, how often it is correct. In this case, the training set scored 92.31% and the test was correct 100% of the time.
* Recall: If there are debtors who have paid in the training set this model can identify them 100% of the time. In the test set it can identify them 100% of the time.

### ROC Curve

Receiver Operating Characteristic (ROC) curve is a plot of the true positive rate against the false positive rate. It shows the tradeoff between sensitivity and specificity.

* AUC score for the case is 0.979. AUC score 1 represents perfect classifier.

### Cross Validation Score

The cross-validation score was 91.89%

### Stratified K-fold Cross Validation

The cross-validation score was 91.89%

### Overall Evaluation

Based on all the scores I think that this might be a good model, even though the test set was 100%. The observations from the test set might come from the fact that there is only one observations in that set. However the cross-validation and Stratified KFold scores were very good.

### IBM Watson Studio Model Evaluation

I wanted to see how Watson Studio’s automated model builder would do with this data. It’s version of a Random Forest model gave a AUC score of .833, also the overall evaluation was “good”. The automated model builder doesn’t allow for the user to see the code that creates the model. That being said I was not able to tune it.

## KNeighbors

The final model used was K Nearest Neighbors. I used the following as features; “Cycle\_Quanternary”, “Cycle\_Tertiary”, and “Cycle\_Secondary.

### Confusion Matrix Evaluation Metrics

I evaluated the model using model evaluation metrics such as accuracy, precision, and recall.

* Accuracy: Classification rate of for the training set was 58.9%. The test set scored 100%.
* Precision: How accurate is your model. In other words, you can say, when a model makes a prediction, how often it is correct. In this case, both the training set the test was correct 100% of the time.
* Recall: If there are debtors who have paid in the training set this model can identify them 16.67% of the time. In the test set it can identify them 100% of the time.

### ROC Curve

Receiver Operating Characteristic (ROC) curve is a plot of the true positive rate against the false positive rate. It shows the tradeoff between sensitivity and specificity.

* AUC score for the case is 0.595. AUC score 1 represents perfect classifier, and 0.595 represents a worthless classifier.

### Cross Validation Score

The cross-validation score was 59.46%

### Stratified K-fold Cross Validation

The cross-validation score was 86.49%

### Overall Evaluation

Based on all the scores I think that this is a bad model, even though the test set was 100%. The observations from the test set might come from the fact that there is only one observation in that set. However, the cross-validation and Stratified KFold scores did not match very well. Also, the low accuracy and recall scores on the test set make this model not better then chance.

# Evaluate the Model

Because the data set is small evaluation is a bit tricky, however based on the Cross-Validation and Stratified K-fold scores combined with the ROC curves I believe that the following models could answer the problem statement in descending from best to worst;

1. Gradient Boosting – Based on the high model evaluation metrics and cross validation scores.
2. Random Forest – Slightly lower than the Gradient Boosting model.
3. Logistic Regression – Slightly lower than the Random Forest model.

The following models would be poor performers in descending order from least to worst;

1. Decision Tree – Because there was such a disparity in the cross-validation scores.
2. K Nearest Neighbors – Because there was such a disparity in the cross-validation scores, and the Recall score on the training set was very low

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy | | Precision | | Recall | | Cross-Validation | Stratified K-fold |
| Training | Test | Training | Test | Training | Test |
| Logistic Regression | 90.41 | 100 | 89.19 | 100 | 91.67 | 100 | 86.49 | 86.49 |
| Decision Tree | 93.15 | 100 | 87.80 | 100 | 100 | 100 | 86.49 | 77.03 |
| Random Forest | 91.78 | 100 | 85.71 | 100 | 100 | 100 | 94.59 | 90.54 |
| GradientBoosting | 95.89 | 100 | 92.31 | 100 | 100 | 100 | 91.89 | 91.89 |
| K Nearest Neighbors | 58.9 | 100 | 100 | 100 | 16.67 | 100 | 59.46 | 86.49 |

# Risks and Limitations

Would like to create a confidence column to weight the debtor based on questioning by the SME. The line of questioning would determine if the debtor has improved their financial situation and are in a better position to repay the loan. Not having this information for this project limited the models accuracy. Going forward, if this model were to be deployed in a production environment having this additional column added as a feature to the model would reduce the risk of investment losses.

# Answer the Problem

Going back to the original problem statement; Can we accurately identify if the portfolio is good one to purchase? I believe that the model can give a very good indication as to whether a portfolio of debt will have enough debtors that will repay the debt to produce a good return on the investment.

End of document