Classification for Credit Defaulting

Executive Summary

I decided to do a classification predicting payment defaulting for university and grad student with the intent to use the information gained to understand which predictors to weight in more when predicting payment defaulting in Central and East Africa.

The data I had chosen included 26 factors and 30,000 observations. After some data cleaning, I transformed it into a time series format to strengthen its structure. This modification changed the data dimension from 30,000 to 180,000 observations and 26 to 13 predictors. I, then, filtered the data to include only university and graduate students from age 22 to 45 because they were a good proxy for my group of interest, which consisted of small entrepreneurs from Central and East Africa. University and grad school students were the best proxy because they shared similar fluctuating income streams and inconsistent repayment behaviors as observed in this region of Africa. To predict loan default for next month (*default_nm*), I ran three classifications models (logistic regression, randomforest, boosting), and the best model came from stochastic gradient boosting, which achieved 84.4 % accuracy, followed by 83.7 % from randomforest and 77% from logistic regression.

As I ran each regression model, it's worth to note the following:

- *Limit_bal*, *pay_amt*, and *bill_amt* were consistently among the most important predictors of payment defaulting while *bill_period*, *pay_period* and *repay_period* were the least important.
- While increasing the number of trees did not necessarily improve accuracy using randomforest, it certainly improved accuracy using stochastic gradient boosting.
- The more reliable models were randomforest and boosting since they obtained higher accuracy measure and came with other metrics that provided reliable insights in understanding which factors were most important in improving accuracy (predicting defaulting). The logistic had a short coming because it showed that sex was not a significant factor in loan defaulting. This was at best vacuously true, because it was only true for this dataset. Given that I wanted to use these results for market insight, I dismissed that as valuable piece of information since there were relatively few women in business and the work place (in general) in Central and East Africa; thus, they were at a higher risk of default.

In conclusion, overall, I was happy with the performance of each model particularly randomForest and stochastic gradient boosting. I achieved a respectable level of accuracy given how small this dataset was. More importantly, I learned which predictors were important in predicting loan defaulting. These results affirmed what I had hypothesized prior to this study as important predictors payment defaulting (only the logistic regression failed to attest that *sex* was also a key predictor along *limit_bal, pay_amt and bill_amt*).

Introduction

Consumer credit makes up a sizeable portion of the US credit market. Since the introduction of banks, consumers have used banks to finance their business and personal needs in the form of credit. In such a consumption driven economy, consumers' spending has been so vital that credit access has been made available to increase the buying capacity of consumers. Virtually everyone in the United States has relied on credit (personal loans, mortgages, etc.) to buy goods and make some investments. In general, credit-borrowing is encouraged by the financial institutions, and it is healthy for the economy. There's, however, a concern for credit defaulting that puts banks and the economy at risk of a recession. Mortgage defaulting in the housing market was the leading cause of what has come to be known as the great session. My interest, however, lied in predicting the probability of default for small entrepreneurs in Central, East Africa who have used their ability to access credit to finance their business. This market is relatively new since it has only begun to spring lately since some countries in this region are experiencing high economic growth. There are foreign investors who would love to commit their capital in this high growth, high risk market but are skeptical of the borrower's ability to respect their financial obligations. While this data did not reflect the economic standing of my target market, students were a potential proxy. The reason was that students had unstable stream of income and repayment behaviors that sort of aligned with the credit market for small entrepreneurs in that region. A word of caution, I was certainly not using students as a proxy to make a strong extrapolation to my market of interest but, rather, used the results of this proxy to test against my hypothesis of which factors were indicative of the likelihood of default in this region.

Data Aggregation

The data for this project was obtained from UCI archives. I attempted three classification to predict credit defaulting for next month payments: logistic regression, randomforest and boosting. Also, I transformed the dataset into a time series format from its original format.

Variable Description

The following variables were obtained from aggregating the original dataset:

- *Limit_bal: amount of the given credit*
- *Sex: male and female*
- *Educ: grad school, university, high school, and others*
- *Marriage: married, single, others*
- *Age: year*
- *Repay_period: monthly repayments from April to September 2005.*
- *Status: categorical repayment measurement scale*
- *Default_nm: (current or default) tracks next month payments.*
- *Pay-period: tracks when payments are supposed to be made*
- *Pay-amt: how much payment was made*
- *Bill_period: tracks when the monthly bill statements was sent*
- *Bill-amt: how much was the outstanding monthly bill*

Model Building

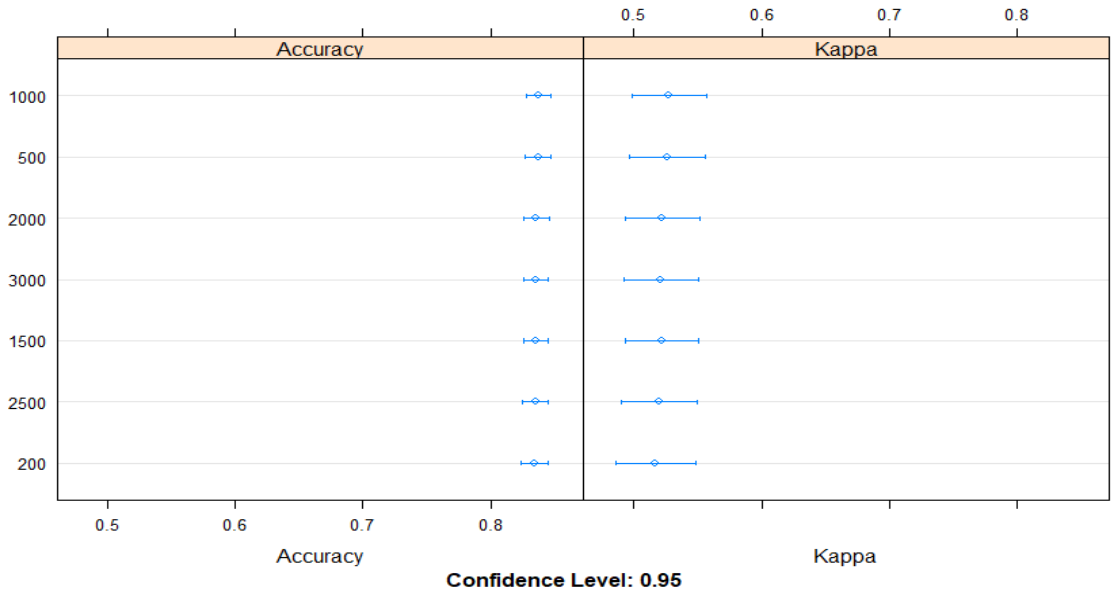
For this prediction project, I created four datasets. However, I only used the set that was created from filtering university and graduate students, which, as described earlier, was used as a proxy for learning purposes. This dataset had a little over 1000 observations. Hence, I used cross validation to improve prediction accuracy since it was the recommended sampling techniques as opposed to creating a training and a testing set. The other three datasets were large sets for intellectual exploration. They included a training set of 100,000 observations, a validation set of 40,000 observations and a test set of 40,000 observations. However, they were not included in this report.

Model 1

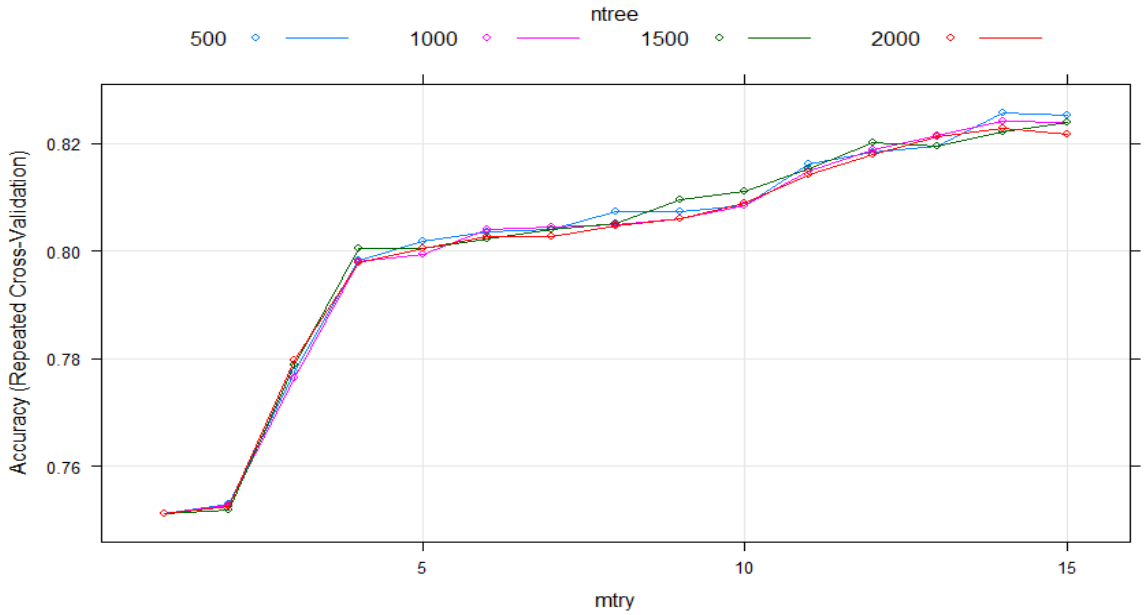
I ran a series of randomForest to attempt to predict credit default for students. I achieved about 83.7% accuracy using cross validation. I used manual search to reach this level of accuracy. The manual search peaked with the number of trees at 1000.

```
Models: 1000, 1500, 2000, 2500, 3000, 200, 500
Number of resamples: 30

Accuracy
          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
1000 0.7822581 0.8211382 0.8373984 0.8373830 0.8536585 0.9032258    0
1500 0.7822581 0.8211382 0.8360656 0.8354749 0.8536585 0.9112903    0
2000 0.7822581 0.8211382 0.8367320 0.8357570 0.8519375 0.9112903    0
2500 0.7822581 0.8211382 0.8360656 0.8352061 0.8536585 0.9112903    0
3000 0.7822581 0.8211382 0.8292683 0.8357526 0.8536585 0.9112903    0
200  0.7822581 0.8133851 0.8360656 0.8346729 0.8536585 0.9032258    0
500  0.7886179 0.8200387 0.8326669 0.8371054 0.8548387 0.9112903    0
```

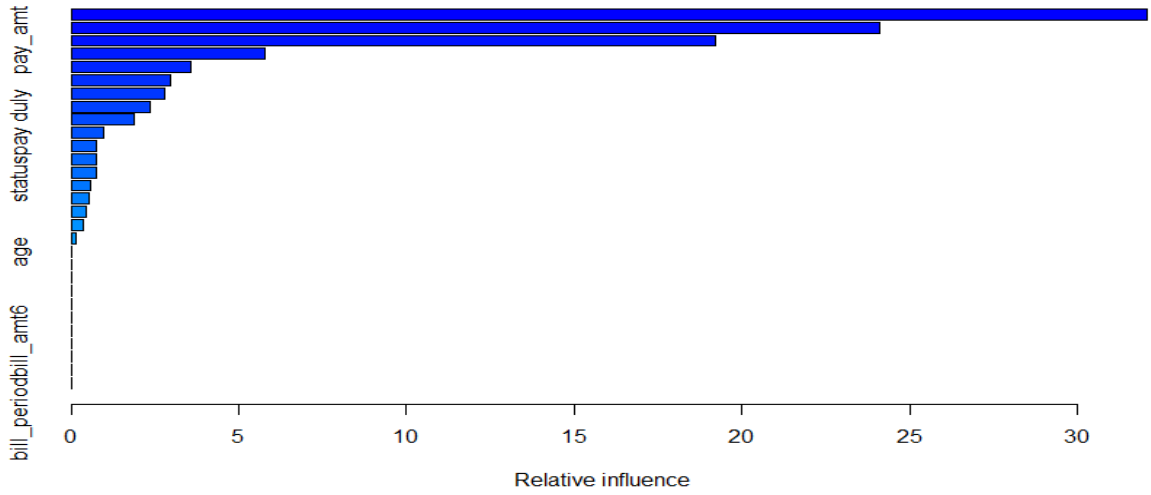


While the above model achieved its highest accuracy with 1000 trees, other models obtained from tuning paramters suggested that further increase in the number of trees did not necessarily improve accuracy. There was a diminishing effect past 1000 trees.



Model 2

I performed a series of stochastic gradient boosting algorithm. The highest accuracy came from tuning inte action. depth, n. tree and shrinkage while n. minosbsinnode was held constant. This model achieved about 84.4% accuracy, which was a slight improvement compared to the 83.7% achieve through randomforest. The three most important variables in predicting default were *bill_amt, limit_bal, and pay_amt.*



Unlike randomforest, boosting suggested the higher number of trees in the algorithm provided enough evidence in improvement in accuracy.

Model 3

The last model that I used was the logistic regression. It offered room to make prediction, but it allows the possibility of an inference based for generalization purposes. It's a good model with ease of interpretability. The best model achieved an accuracy of 77% accuracy because I transformed some variables such that they had a non-linear relationship to *default_nm* compared to the original model, which assumed that all predictors had a linear relation with the predicted *default_nm*.

```
Generalized Linear Model

1230 samples
  11 predictor
   2 classes: 'current', 'default'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 5 times)
Summary of sample sizes: 1107, 1107, 1106, 1107, 1106, 1107, ...
Resampling results:

  Accuracy   Kappa
  0.7692688  0.2511681
```

```
Coefficients: (11 not defined because of singularities)
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -4.590e-01  5.809e-01  -0.790 0.429463
limit_bal                 9.932e-12  1.561e-12   6.364 1.97e-10 ***
sexmale                   1.677e-01  1.593e-01   1.053 0.292520
educuniversity           -4.006e-01  2.510e-01  -1.596 0.110427
marriageothers            2.119e+00  4.132e-01   5.128 2.93e-07 ***
marriagesingle            5.731e-01  2.437e-01   2.352 0.018684 *
age                              NA         NA      NA       NA
repay_periodrepay_0505    4.283e-02  2.469e-01   0.173 0.862291
repay_periodrepay_0605   -2.578e-02  2.478e-01  -0.104 0.917119
repay_periodrepay_0705   -9.347e-02  2.488e-01  -0.376 0.707186
repay_periodrepay_0805   -1.733e-01  2.509e-01  -0.691 0.489835
repay_periodrepay_0905   -1.982e-01  2.626e-01  -0.755 0.450279
`status2M-delay`          6.011e-01  5.154e-01   1.166 0.243473
`status3M-delay`          4.384e-01  7.292e-01   0.601 0.547759
`status4M-delay`         -1.437e-01  1.345e+00  -0.107 0.914908
`statusno credit-use`    -2.147e+00  5.970e-01  -3.597 0.000322 ***
`statuspay duly`         -1.098e+00  5.138e-01  -2.136 0.032650 *
`statusrevolving credit` -1.201e+00  4.929e-01  -2.437 0.014806 *
pay_periodpay_amt2               NA         NA      NA       NA
pay_periodpay_amt3               NA         NA      NA       NA
pay_periodpay_amt4               NA         NA      NA       NA
pay_periodpay_amt5               NA         NA      NA       NA
pay_periodpay_amt6               NA         NA      NA       NA
pay_amt                   1.414e-11  3.347e-11   0.422 0.672724
bill_periodbill_amt2             NA         NA      NA       NA
bill_periodbill_amt3             NA         NA      NA       NA
bill_periodbill_amt4             NA         NA      NA       NA
bill_periodbill_amt5             NA         NA      NA       NA
bill_periodbill_amt6             NA         NA      NA       NA
bill_amt                 -9.088e-12  4.395e-12  -2.068 0.038668 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1380.0  on 1229  degrees of freedom
Residual deviance: 1201.5  on 1211  degrees of freedom
AIC: 1239.5

Number of Fisher Scoring iterations: 4
```

The above table seemed to support the other algorithm in identifying which variables were most important in determining the accuracy of the model. As we see, *bill_amt, pay_amt* and *limit_bal* consistently appeared important. In regression jargon, they were significant while *pay_period, bill_period* and *repay_period* were not significant. In addition, we see that the logistic regression gave us away to interpret the log odds of default. That is, each variable coefficient represented some increase or decrease in the percentage fluctuation of *default_nm.* However, since I was purely interested in accuracy, then the logistic regression yielded the least accurate measure.

Conclusion

I decided to do a prediction analysis on payment default for university and grad student because in theory they were the best proxy for predicting payment defaulting of small entrepreneurs in Central and East Africa since these two groups had unreliable streams of income and inconsistent repayment behaviors. I performed three regressions models that yielded accuracy measurement above 76% on average with stochastic gradient boosting achieving the highest accuracy of 84.4%. All the algorithm reliably showed that limit_bal, pay_amt, and bill_amt were among the top predictors for predicting payment default while other predictors such as with *pay_period, bill_period* and *repay_period* were the least important. Though randomForest and stochastic gradient boosting obtained higher accuracy measure for load defaulting, the logistic regression yielded lower measurement in comparison. However, it gave a better since of interpretability since it brought with it an inferential component as well determined by the coefficient of each predictor. Nevertheless, since my objective was prediction accuracy, boosting was my best model in terms of accuracy measure. To finish, I would say that the model results suggested that the size of the loan, the amount billed to the customer and paid by the customer are probably the most important factors in predicting payment defaulting. Thus, a moderate extrapolation of these results would say that *limit_bal, pay_amt and bill_amt* are the most important variables to consider when deciding which small entrepreneurs were likely to default. However, the logistic regression had some short comings. It suggested that *sex* of the student was a not significant predictor, which at best was vacuously true since gender would be an important factor in Central and East Africa given the limited number of females in business. Overall, the accuracy level and other performance metrics for both randomForest and stochastic gradient boosting were reliable enough that I could learn which predictors would call for caution when determining the probability of payment default in this region.

Citation:

- Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients
- https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients