Classification

Bryce Bowles
648 Business Data Analytics 901
11/06/2020

Assignment 4

1. Read Chapter 2.2 of Introduction to Statistical Learning in R and the article "The Foundations of Algorithmic Bias". Write a 1/2 to 1 page response (prose, in complete sentences). In your essay address the following:
    - Provide two definitions of bias.
    - How can we avoid some of the undesirable effects of each type of bias?
    - In addition to bias, what are other ethical challenges facing data scientists today? Are there other concerns that are more important than bias?

Use the Lending Club 2017 Q2 data for the remaining questions.

2. Build a logistic regression model and a classification tree model for predicting the final status of a loan based on variables available at the time at which a loan is awarded. Provide a confusion matrix and misclassification rate for each model for a test dataset. Which variables appear to be important for predicting outcome?

3. Build a logistic regression model and a classification tree model for predicting the final status of a loan based on the variables loan_amnt, funded_amnt_inv, term, int_rate, installment, grade, emp_length, home_ownership, annual_inc, verification_status, loan_status, purpose, title, dti, total_pymnt, delinq_2yrs, open_acc, pub_rec, last_pymnt_d, last_pymnt_amnt, application_type, revol_bal, revol_util, recoveries. Use the same training and testing observations as for question 2. Provide a confusion matrix and misclassification rate for each model for a test dataset. To format the date in last_pymnt_d properly, use the following code:

```
my_df$last_pymnt_d <- as.POSIXct(my_df$last_pymnt_d)
```

4. Plot the ROC curves for your four models. Which models perform best? To what do you attribute the differences in performance between the models in #2 and #3?

5. If you were considering investing in Lending Club loans, which model would you use to support your decision making?

Update 9/10/20: Create a report that contains a summary of the steps that you took and the results that you obtained. Do not include code in your report and avoid specific references to R - write the report as if you are reporting to someone unfamiliar with R (but you may assume familiarity with the analytics methods). Submit your code as a separate .R file so that the instructor can run it if necessary. Refer to the syllabus for more details.

Update 10/19/20: A link to the article "The Foundations of Algorithmic Bias" is in the Module Online Resources on the page Analytics Articles.

#1 - Essay

Techniques to evaluate models are provided throughout the readings to help assess model accuracy or decide which method produces the best results. Since no one Statistical Learning method dominates all others over all possible data sets, it is important to try more than one so that a model gets evaluated for optimal results. Quality of fit, bias-variance trade-off and classification setting all influence model accuracy.

Bias refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model. One way bias gets added to models is by "overfitting the data". This can be explained by how well a model's predictions actually match the observed data. The way we may measure this is to quantify the extent to which the predicted response value for a given observation is close to the true response value for that observation. In a regression setting, the most commonly used measure is the mean squared error (MSE). The MSE will be small if the predicted responses are very close to the true responses. Another way Bias can affect models is by the way the data are selected. I thought the "The Foundations of Algorithmic Bias" article mentioned an interesting and important point of when training a machine learning model, it is only going to learn on the training data's scenarios and other scenarios that may be in the test dataset will be left out. The data itself may be bias, depending on how parameters were selected and if individuals choosing those parameters contained any bias.

To avoid some of the undesirable effects of each type of bias, you will need to evaluate each way the data is being collected and used. For example, in linear regression, you may test the true f linear relationship between dependent variable and other variables in the data. If f is linear, linear regression will have no bias, making it very hard for a more flexible method to compete. In contrast, if the true f is highly non-linear and we have an ample number of training observations, then we may do better using a highly flexible approach. Another method is to use cross-validation, a way to estimate the test MSE using the training data. When collecting data, in general, the more data you use, the better off you will be.

Other ethical challenges facing data scientists today include MSE, variance, and squared bias. Variance refers to the amount by which f would change if we estimated it using a different training data set. As a general rule, using more flexible methods, the variance will increase and the bias will decrease. Low bias and low variance are best for predictions. Another ethical challenge may be getting a good test set. Good test set performance of a statistical learning method bias-variance requires low variance as well as low squared bias. The challenge lies in finding a method for which both the variance and the squared bias are low. In a real-life situation in which f is unobserved, it is generally not possible to explicitly compute the test MSE, bias, or variance for a statistical learning method.

Assessing model accuracy depends on quality of fit, bias-variance trade-off, classification setting and much more. Bias may be included in a predetermined dataset. After the readings, it is my recommendation that data scientists should be aware of this and continuously collect data to avoid predetermined bias's so that models adjust to changes made over time. In addition to continuously collecting data, continuously evaluating the performance of the model as the data changes will also be essential in obtaining optimal results.

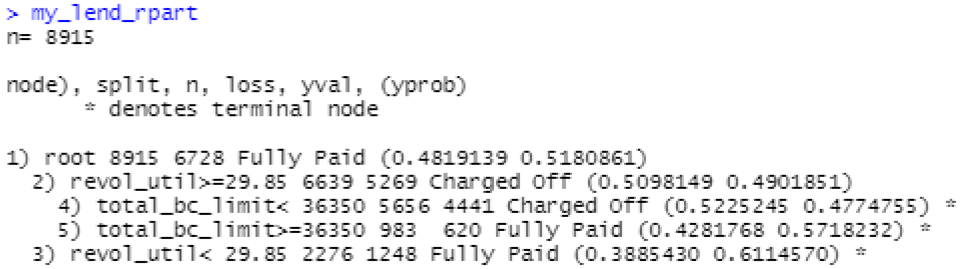
#2 - Logistic regression and classification tree model for predicting the final status of a loan based on variables available at the time at which a loan is awarded.

After reading in the data, columns were chosen to help predict the final status of a loan such as application_type, total_bal_ex_mort, total_bc_limit and revol_util, int_rate, emp_length, annual_inc, dti, fico_range_low, and fico_range_high. Columns where then filtered to exclude all status types except “Fully Paid” and “Charged Off” and to only include the final status’. Linear regression can only predict classification models in instances where you have two outcomes. The data was split to get a sample of 10% while choosing the same proportions of status types in both the training and test data. 1,000 samples will provide a good enough turnout while keeping computational speed fast. Next, the missing data (N/As) are replaced, using the median from training data to fill the test data to prevent leakage. Since the ratio is about 4 “charged off” to 1 “fully paid”, weights were added to indicate the significance of each type.

A logistic regression model was then built to predict the loan status. Probability for each loan that was fully paid was calculated and we used this to determine the prediction. If the probability was anything over 0.5 its prediction is “fully Paid” and anything under 0.5 is “Charged Off”. A confusion matrix was then used to compare the actual results verses the predicted. Of those that were Charged off, 8235 correctly classified and 6897 were not. Of those that were Fully Paid, 46362 were correctly classified and 18734 were not. The logistic regression model had a misclassification rate of 0.319477.

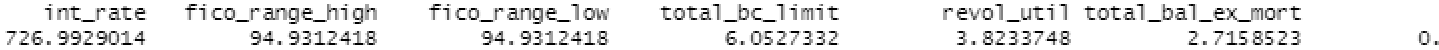
The Classification Tree model was then built to compare to the logistic regression model, still predicting the loan status. A confusion matrix was used was used to compare the actual results verses the predicted. Of those that were Charged off, 9972 were correctly classified and 5160 were not. Of those that were Fully Paid, 40291 were correctly classified and 24805 were not. This model had a misclassification rate of 0.373498. Below is how the tree is broken down as well as variable importance. The int_rate, fico_range_high and fico_range_low appear to be important for predicting outcome.

Tree:



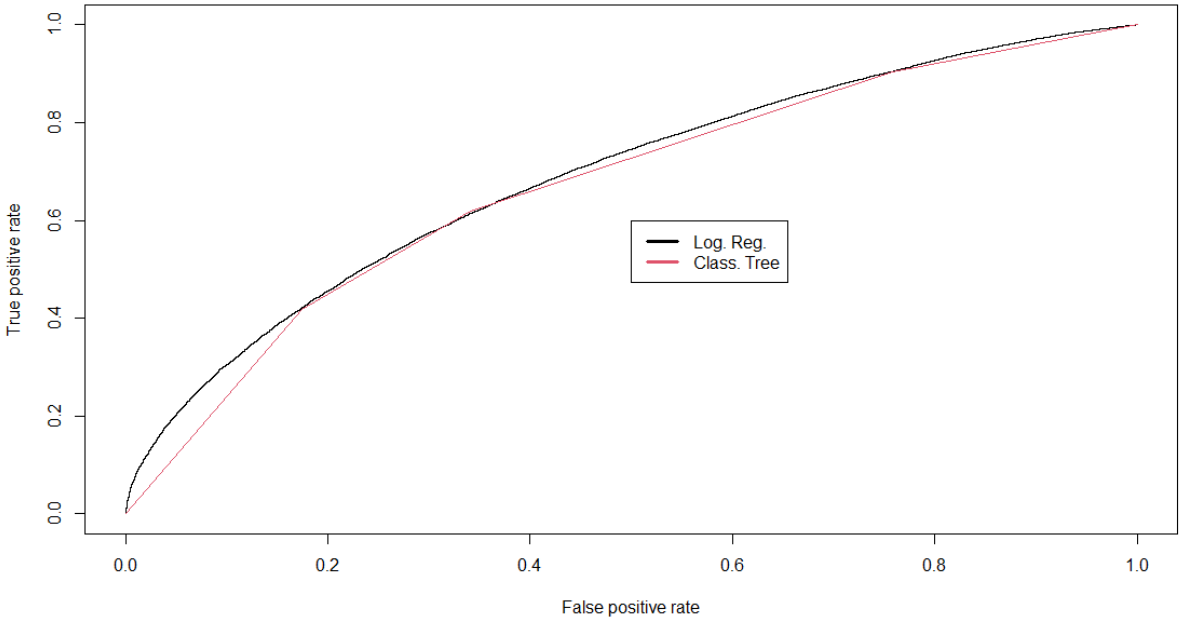
```
> my_lend_rpart
n= 8915

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 8915 6728 Fully Paid (0.4819139 0.5180861)
  2) revol_util>=29.85 6639 5269 Charged Off (0.5098149 0.4901851)
    4) total_bc_limit< 36350 5656 4441 Charged Off (0.5225245 0.4774755) *
    5) total_bc_limit>=36350 983  620 Fully Paid (0.4281768 0.5718232) *
  3) revol_util< 29.85 2276 1248 Fully Paid (0.3885430 0.6114570) *
```

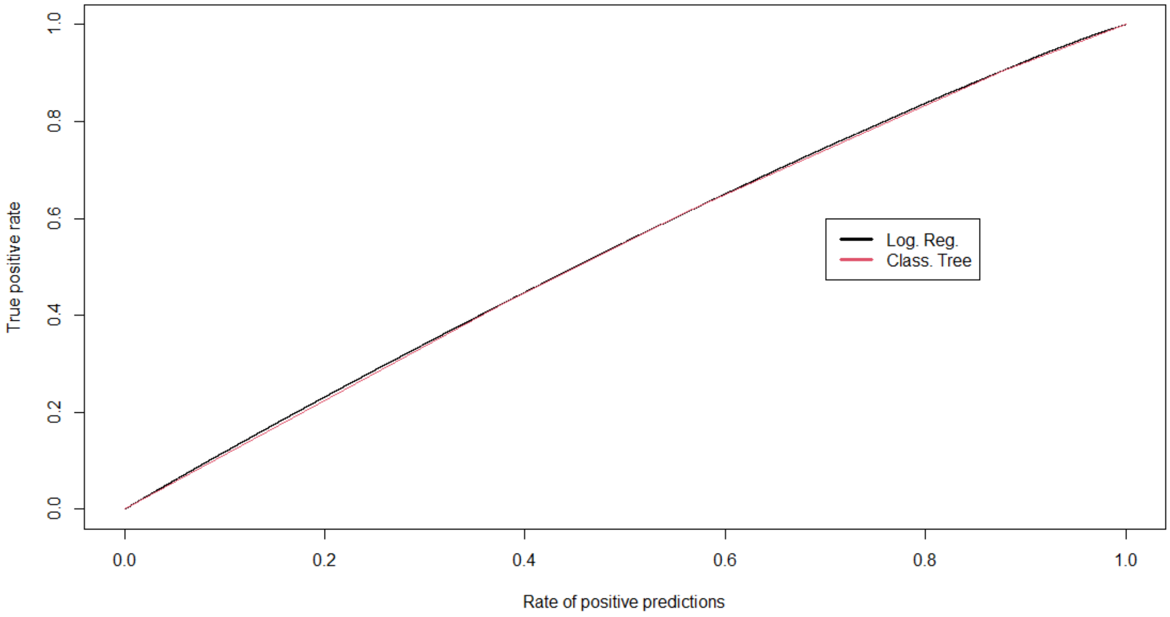
Variable Importance:



```
   int_rate   fico_range_high    fico_range_low    total_bc_limit        revol_util total_bal_ex_mort
726.9929014        94.9312418        94.9312418         6.0527332         3.8233748         2.7158523         0.
```

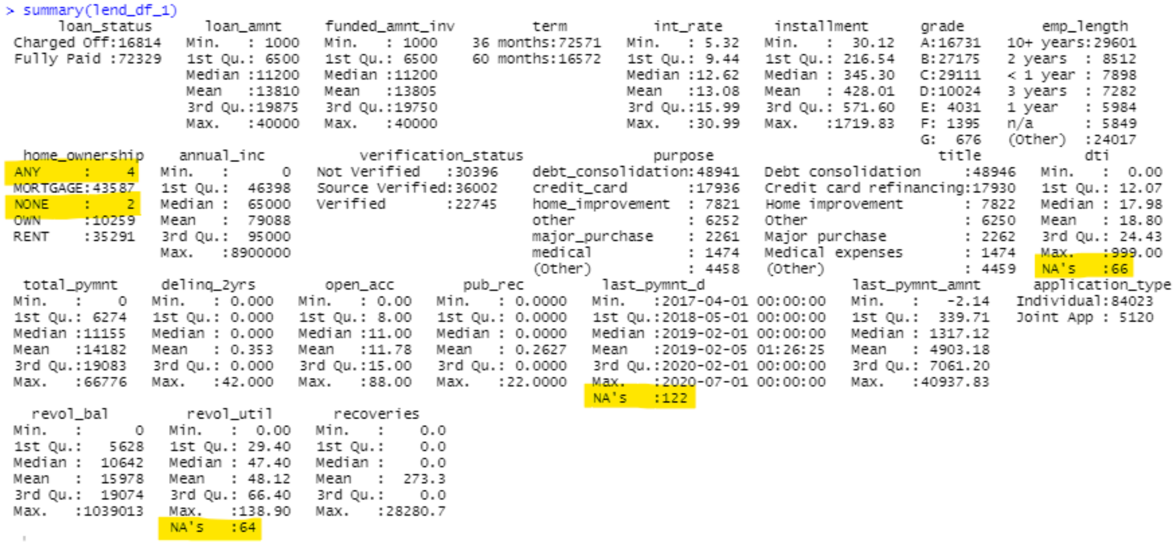
A ROC Curve was used to display the models. The Classification Tree model had an AUC (Area under the ROC Curve) value of 0.6710865 of and Logistic Regression had a 0.6883902 value, indicating the logistic regression is the best model for this scenario.





#3 - Logistic regression model and a classification tree model for predicting the final status of a loan based on the variables *loan_amnt, funded_amnt_inv, term, int_rate, installment, grade, emp_length, home_ownership, annual_inc, verification_status, loan_status, purpose, title, dti, total_pymnt, delinq_2yrs, open_acc, pub_rec, last_pymnt_d, last_pymnt_amnt, application_type, revol_bal, revol_util, recoveries*.

First, the variables were modified from the previous models. Columns where again filtered to exclude all status types except “Fully Paid” and “Charged Off” and to only include the final status’. Similar preprocessing steps were used to remove any NA’s and adjust the home_owndership variable categories. Once again, the categories “ANY” and “NONE” were grouped with RENT. The below summary was used to identify variable preprocessing needs. The data was split to get a sample of 10% while choosing the same proportions of status types in both the training and test data. Next, the missing data (N/As) are replaced, using the median from training data to fill the test data to prevent leakage. Since the ratio of loan_status is about 4 “charged off” to 1 “fully paid”, weights were added to indicate the significance of each type.

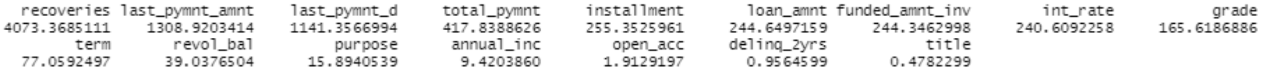


```
> summary(lend_df_1)
       loan_status      loan_amnt      funded_amnt_inv          term          int_rate       installment         grade         emp_length   
 Charged Off:16814   Min.   : 1000   Min.   : 1000    36 months:72571   Min.   : 5.32   Min.   :  30.12   A:16731   10+ years:29601  
 Fully Paid :72329   1st Qu.: 6500   1st Qu.: 6500    60 months:16572   1st Qu.: 9.44   1st Qu.: 216.54   B:27175   2 years  : 8512  
                     Median :11200   Median :11200                      Median :12.62   Median : 345.30   C:29111   < 1 year : 7898  
                     Mean   :13810   Mean   :13805                      Mean   :13.08   Mean   : 428.01   D:10024   3 years  : 7282  
                     3rd Qu.:19875   3rd Qu.:19750                      3rd Qu.:15.99   3rd Qu.: 571.60   E: 4031   1 year   : 5984  
                     Max.   :40000   Max.   :40000                      Max.   :30.99   Max.   :1719.83   F: 1395   n/a      : 5849  
                                                                                                          G:  676   (Other)  :24017  
  home_ownership    annual_inc          verification_status                purpose                         title                dti        
 ANY     :    4   Min.   :      0   Not Verified   :30396   debt_consolidation:48941   Debt consolidation     :48946   Min.   :  0.00  
 MORTGAGE:43587   1st Qu.:  46398   Source Verified:36002   credit_card       :17936   Credit card refinancing:17930   1st Qu.: 12.07  
 NONE    :    2   Median :  65000   Verified       :22745   home_improvement  : 7821   Home improvement       : 7822   Median : 17.98  
 OWN     :10259   Mean   :  79088                            other             : 6252   Other                  : 6250   Mean   : 18.80  
 RENT    :35291   3rd Qu.:  95000                            major_purchase    : 2261   Major purchase         : 2262   3rd Qu.: 24.43  
                  Max.   :8900000                            medical           : 1474   Medical expenses       : 1474   Max.   :999.00  
                                                             (Other)           : 4458   (Other)                : 4459   NA's   :66      
  total_pymnt     delinq_2yrs        open_acc        pub_rec          last_pymnt_d                   last_pymnt_amnt     application_type
 Min.   :    0   Min.   : 0.000   Min.   : 0.00   Min.   : 0.0000   Min.   :2017-04-01 00:00:00   Min.   :   -2.14   Individual:84023  
 1st Qu.: 6274   1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 0.0000   1st Qu.:2018-05-01 00:00:00   1st Qu.:  339.71   Joint App : 5120  
 Median :11155   Median : 0.000   Median :11.00   Median : 0.0000   Median :2019-02-01 00:00:00   Median : 1317.12                     
 Mean   :14182   Mean   : 0.353   Mean   :11.78   Mean   : 0.2627   Mean   :2019-02-05 01:26:25   Mean   : 4903.18                     
 3rd Qu.:19083   3rd Qu.: 0.000   3rd Qu.:15.00   3rd Qu.: 0.0000   3rd Qu.:2020-02-01 00:00:00   3rd Qu.: 7061.20                     
 Max.   :66776   Max.   :42.000   Max.   :88.00   Max.   :22.0000   Max.   :2020-07-01 00:00:00   Max.   :40937.83                     
                                                                    NA's   :122                                                        
   revol_bal         revol_util       recoveries     
 Min.   :      0   Min.   :  0.00   Min.   :    0.0  
 1st Qu.:   5628   1st Qu.: 29.40   1st Qu.:    0.0  
 Median :  10642   Median : 47.40   Median :    0.0  
 Mean   :  15978   Mean   : 48.12   Mean   :  273.3  
 3rd Qu.:  19074   3rd Qu.: 66.40   3rd Qu.:    0.0  
 Max.   :1039013   Max.   :138.90   Max.   :28280.7  
                   NA's   :64                        
```

When creating a logistic regression and classification tree models, the same steps in the above question 2 model were used.

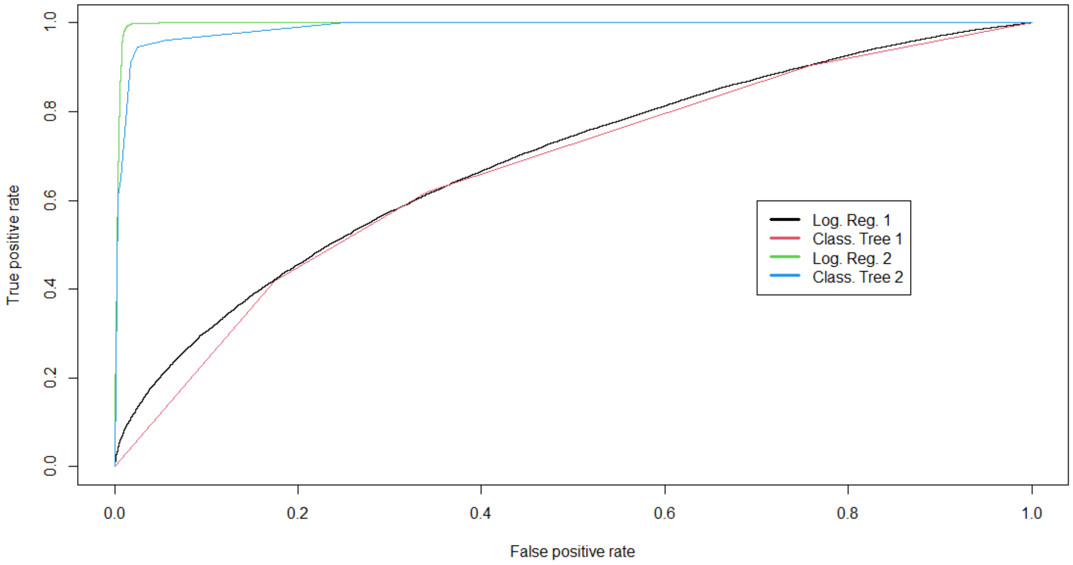
Logistic Regression model: Of those that were Charged off, 14876 correctly classified and 256 were not. Of those that were Fully Paid, 64867 were correctly classified and 229 were not. The logistic regression model had a misclassification rate of 0.006045271.

Classification Tree model: Of those that were Charged off, 14759 were correctly classified and 373 were not. Of those that were Fully Paid, 61500 were correctly classified and 3596 were not. This model had a misclassification rate of 0.04947151. The recoveries, last_pumnt_amnt, last_pymnt_d, total_pymnt etc… in that order appear to be important for predicting outcome. (the higher the value, the more important)



```
   recoveries last_pymnt_amnt    last_pymnt_d     total_pymnt     installment       loan_amnt funded_amnt_inv        int_rate           grade 
 4073.3685111    1308.9203414    1141.3566994     417.8388626     255.3525961     244.6497159     244.3462998     240.6092258     165.6186886 
         term       revol_bal         purpose      annual_inc        open_acc     delinq_2yrs           title 
   77.0592497      39.0376504      15.8940539       9.4203860       1.9129197       0.9564599       0.4782299 
```

#4 The Classification Tree model had an AUC (Area under the ROC Curve) value of 0.9885012 of and Logistic Regression had a 0.9967356 value, indicating the logistic regression is the best model for this scenario.



#5 The 2nd Logistic Regression created (green) performs the best. Other factors that come into play are overfitting and biases. If I were to consider investing in Lending Club loans, I'd use the first linear regression model (Black) to support my decision making. This model does not perform as well but it also suffers from less leakage and less overfitting of the model.