Machine Learning using R

Veeral Patel

Regis University

April 12th, 2021

1. **Introduction**

For this project, we were expected to download R and R Studio in order to be able to open data files, and use R for data analysis. We were using a dataset that was included in one of the R packages. This dataset included species of plants and their petal measurements along with other measurements of different species of plants. We were to use R in order to perform mathematical operations on that dataset which helps us gather more insight on the data. We worked with many datasets for this homework. This was an introduction in Machine Learning where we were expected to identify the correct features that would predict of someone would default on their loans or not.

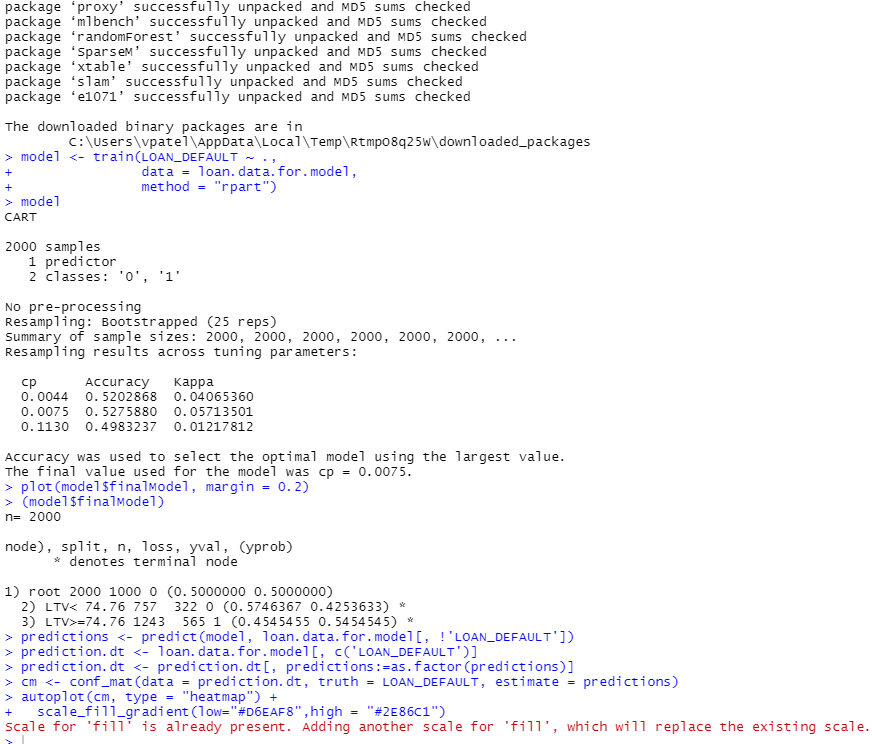
1. **Methods/Code/Screenshots**

I used a windows computer and I had to install R, R Studio (IDE) and all of the dependency modules that were needed to perform the data analysis. I have worked with R before but it was very brief and it was in a science class. I had only used R before to calculate averages of large sets of data that was in an .rda format which is like a really compressed file. This was a new adventure for me to work with R Studio with more data analysis. I watched a few quick tutorials on YouTube in order to assist me with getting everything installed on my computer. Most of the code in this assignment was supposed by the professor. I was in charge of coming up with a few calculations that included standard deviation, mean, box plot, histogram, and answering questions about the dataset. I also found it generally helpful to actually understand the data by just opening it up in R studio in a data frame before getting started with the assignment. I also used the head() function in order to get a quick glimpse of the data within R Studio.

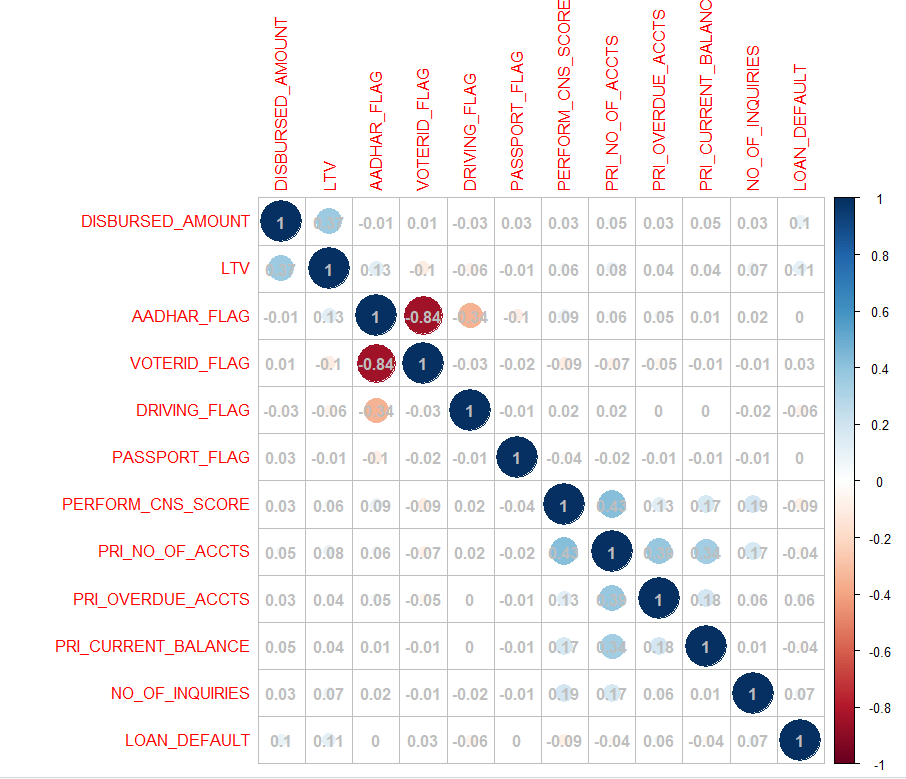
In this lab, I tried to predict whether or not a person would default on their loans. I used two different features to try to predict this outcome, LTV and PERFORM\_CNS\_SCORE. The LTV feature is a measure of the loan to the value of the thing being purchased. The PERFORM\_CNS\_SCORE is a measure of their credit score. Then, I used a correlation map, by doing so I quickly realized that the PERFORM\_CNS\_SCORE does not really have any correlation between that and the odds of defaulting on the loan. Also, with the same correlation map, I realized that the LTV had a correlation between doing the defaulting of their loan. Then, I used a model (rpart) to train my outcome and then I used that model to predict the odds of defaulting on the loan. I analyzed the result by heat map.

1. **Results/Analysis**

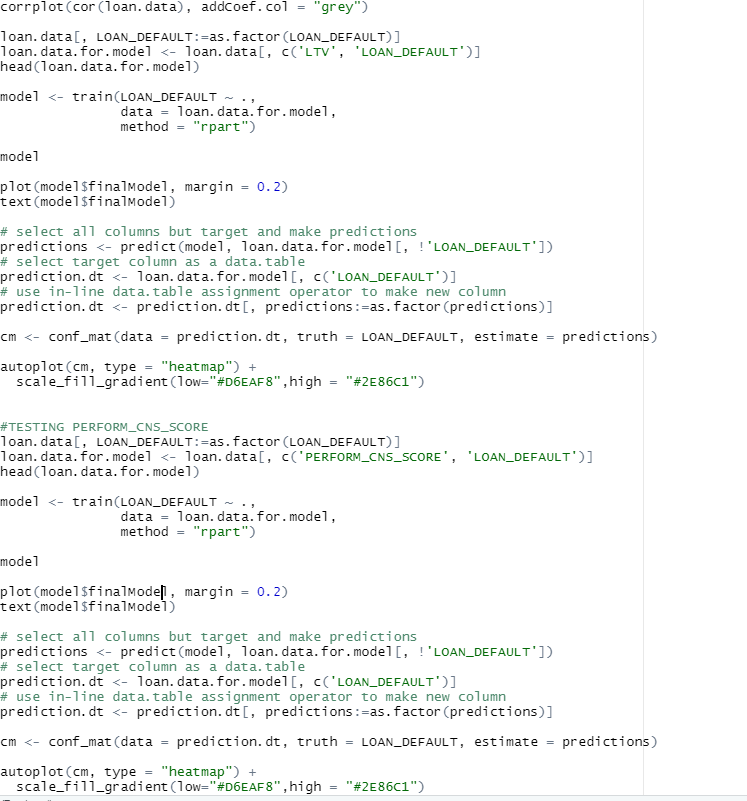
Workthrough of datasheets:



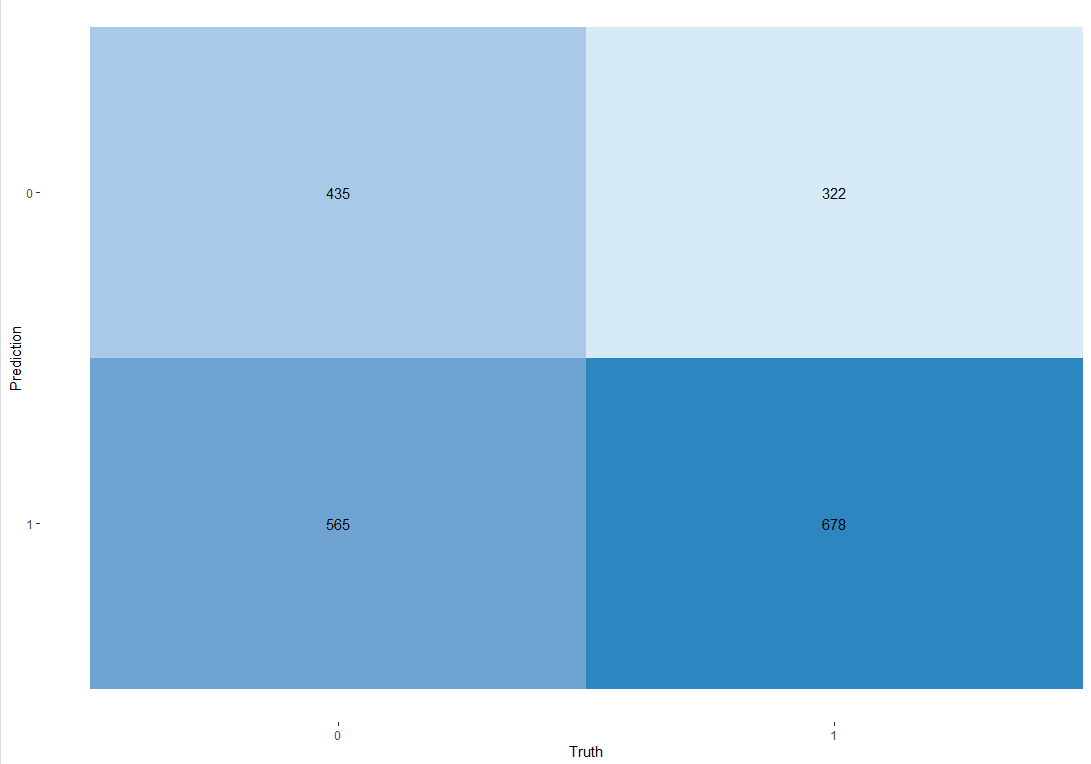
Correlation Plot of features with target:



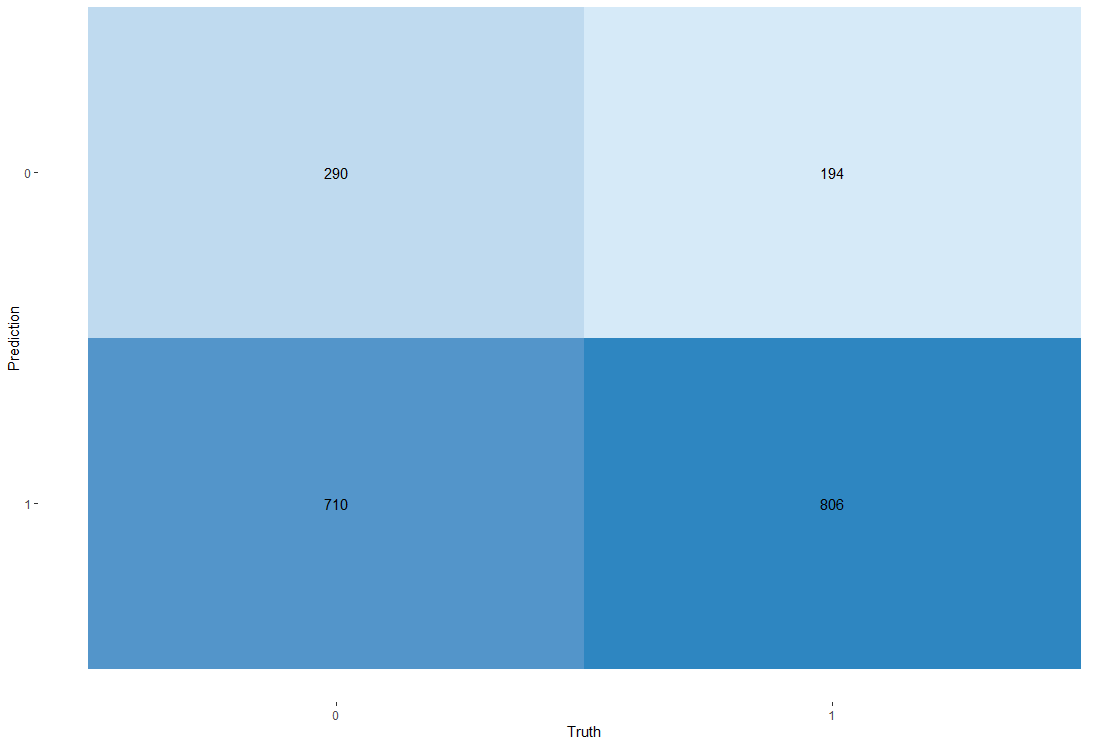
Code for two features used in model for machine learning:



Heat map of Data when using LTV as factor:



Heatmap of data when using PERFORM\_CNS\_SCORE:



1. **Analysis of Results**

In this lab, we had results from trying to predict whether or not an individual would default on their loans or not. I used two variables to try and predict this outcome, LTV and PERFORM\_CNS\_SCORE. The LTV feature is a measure of the loan to the value of the thing being purchased. The PERFORM\_CNS\_SCORE is a measure of their credit score. Then, I used a correlation map, by doing so I quickly realized that the PERFORM\_CNS\_SCORE does not really have any correlation between that and the odds of defaulting on the loan. Also, with the same correlation map, I realized that the LTV had a correlation between doing the defaulting of their loan.

The EDA was a large part of the analysis as well. By doing EDA, I used a correlation plot, which told us that there is a .11 correlation between LTV and LOAN\_DEFAULT. While, there is a -.9 correlation between PERFORM\_CNS\_SCORE and LOAN\_DEFAULT.

Then by looking at the strongest correlation, LTV, I chose that feature as part of the feature selection process. Then, I used the caret ML library in R to run a decision tree model, then I got the predictions and plotted them. I used a confusion matrix to plot them which kind of looks like a heatmap.

The heat map for LTV shows that in the 2000 times that the model ran, it was able to use LTV to predict whether or not the person would default or not correctly 1113 times.

The heat map for PERFORM\_CNS\_SCORE shows that in the 2000 times that the model ran, it was able to use PERFORM\_CNS\_SCORE to predict whether or not the person would default or not correctly 1096 times.

1. **Conclusion**

I thought that this was a really good introduction to machine learning. I think that it would have made more sense to start with regression or something like that instead of decision trees. I do not know if I did it right but I think that for a lab that my predictions from the model were not that accurate, maybe I did something incorrectly? I think that this model could have been adjusted to use logistical regression as well since it is a binary prediction that we are trying to make here. I think that this technology fits really well into the field of data science because we are actively using data to try to predict future outcomes. Something like this is really useful in any field out there because we are using past data to predict future data, I can’t really think of many instances where this would not be helpful at the least. I had a really good experience with this. It was really helpful that you provided to code to use the carat library to use decision trees. I wish that there was more of a mathematical explanation as to what was happening on the back end of the decision trees. I am still confused at how decision trees work.

1. **Reference:**

Hornik, K., Buchta, C., & Zeileis, A. (2009). Open-source machine learning: R meets Weka. *Computational Statistics*, *24*(2), 225-232.

Quinlan, J. R. (1987). Simplifying decision trees. *International journal of man-machine studies*, *27*(3), 221-234.