Statistics 652: Project - Lending Club Loan History Challenge

Recently Siraj Raval, a famous AI Education YouTuber, made a video Intro to Statistics - Data Lit #2 and gave a data challenge to classify the *Loan Status* of the approved LendingClub loans.

The data can be downloaded from kaggle LendingClub.

The goal of the project is to apply all of the Machine Learning Algorithms we learn about in this class to the LendingClub data from 2012-2014. And to equal or improve on the accuracies already achieved, see the github for the competition LoanDefault-Prediction. (Note: It might have been more accurate to use Classification in the title of the github. The competition is to build a classifier not a numeric prediction model.)

For the project you will apply the 5 step learning process used in the Lantz book for each model in Chapter 11.7 Exericises Problem 6 (Naive Bayes is options because we did not cover this model in detail) in the Modern Data Science with R book to the LendingClub data. The data are in the *accepted_2007_2018q4.csv* data file.

Step 1 – collect data

Step 2 – exploring and preparing the data

Step 3 – training a model on the data

Step 4 – evaluating model performance

Step 5 – improving model performance

You should do your work in an R Project. There should be a **/data** directory in which your data files are downloaded to and merged together.

In your report, **Lastname_Firstname_Stat652_Project.Rmd** you should include an Abstract, Introduction, and Conclusions sections. It might be better to use separate .Rmd files for each models, if each becomes too large.

For the 5 steps you should do the following:

Step 1 – collect data

Download the data files from the kaggle LendingClub website and subset the data in the *accepted_2007_2018q4.csv* file for the years 2012-2014. Save an .Rds file and read in the .Rds file when needed. Use the **saveRDS()** and **readRDS()** R functions, here is a link readRDS. Or try using the *data.table* R package and the *fread()* fuction. Make a table of the variable names. Make a table of the number of rows and columns in the dataframe.

Step 2 – exploring and preparing the data

Summarize the important features of the data. Summarize some of the numeric and some of the categorical features, there maybe too many to summarize.

If you have time and are interested, see if you can get the trelliscope package to work for visualization.

Create Training and Testing datasets. Use a 75-25 split. (What was used for the accuracies on the github?)

Step 3 – training a model on the data

Try all of the main models we introduce in the Stat. 652 class. And if you are interests try some alternatives.

Step 4 – evaluating model performance

Compute the accuracy and, if appropriate, the area under the ROC curve (AUC) to rank the classification accuracy of each model.

If you have time try the h2O autoML function or the h2O DriverlessAI software for comparison.

Step 5 – improving model performance

Try to tune the parameters in each model to achieve best performance.

In the Conclusion section clearly state what you believe the best ML learning model is for classifying **Loan Status**, variable is *loan_status*.

Extra Credit: Once you have decided on the best model, refit it using all of the 2012-2014 data and then use your model to classify all of the 2015 data. Check the accuracy of your predictions.

If you have limited computer resources, use the model you settled on using the 2012-2014 training data and use that model to classify a 10% sample of the 2015 data.