POLI 175 Challenge 1

Due 11:59PM PT, Thursday February 11, 2021

General Instructions

In this assignment, we are providing you with a data set on criminal defendants. You'll work in teams of 3-4 to train supervised machine learning models that predict which defendants will recidivate (commit another crime) as a function of various characteristics.

Please submit this assignment by uploading your html (`.html`) and code (`.Rmd`) files onto **Canvas** (navigate to the Assignments tab on our Canvas course website) before the due time. **Only one person from each group should upload the files.**

Introduction

You are being provided with a historical data set of criminal defendants from Broward County, Florida. Your goal for this assignment is the following.

ASSIGNMENT

1. **Use the data to train and evaluate several predictive models that predict whether or not a criminal defendant will recidivate.**

2. **Choose a final model you consider would be best to deploy in the real world.**

Your team must decide which methods to employ, what processes and metrics to use for training your models, and what metrics and other considerations to take into account when choosing your final model. Some submission guidelines can be found at the end of this document, but this assignment is designed to be open-ended (though within the bounds of what we have learned in this class), similar to a real-world machine learning endeavor.

We will take your model and assess its performance on held-out test data that we are not sharing with you. Your model's test data performance will be a factor in your evaluation, though the bulk of your evaluation will be based on the extent to which you follow good practices in training, selecting, and assessing models; documenting your work and coding procedures; and justifying any key decisions you make along the way.

Data

Data Sets

Two data sets are provided.

1. `recidivism_data_sample.csv`: A sample of 6000 criminal defendants. You will use these data to build your predictive models, as you see fit.

2. `recidivism_data_pseudo_new.csv`: A sample of 3000 pseudo (fake) held-out observations. At the end of the assignment, you will use your final model to make predictions for these pseudo observations. **This is meant to verify that your final model can be used to predict onto (is compatible with) test data. You should not use these data at all in model training, nor to assess the predictive performance of your models.**

Variables

The data set `recidivism_data_sample.csv` contains the following variables:

- `id`: a unique identifier for each defendant.
- `recidivate`: a binary indicator of whether the defendant recidivated after the criminal charge.

  0: did not recidivate

  1: did recidivate

- `race`: a numeric indicator of race.

  1: White (Caucasian)

  2: Black (African American)

  3: Hispanic

  4: Asian

  5: Native American

  6: Other

  Note: Often times in data sets, an unordered categorical variable will be coded as numeric. In such cases, it is important that you take the appropriate steps to make sure any modeling procedures you apply to the data do not treat such variables as numeric variables!

- `sex`: a binary indicator.

  0: male

  1: female

- `age`: age in years.
- `juv_fel_count`: number of juvenile felony criminal charges in the past.
- `juv_misd_count`: number of juvenile misdemeanor criminal charges in the past.
- `priors_count`: number of non-juvenile criminal charges in the past.
- `charge_degree`: a binary indicator of the degree of the current charge.

  0: misdemeanor

  1: felony

- `charge_name`: a categorical variable describing the specific criminal charge.

The data set `recidivism_data_pseudo_new.csv` contains all of the same variables except for `id` and `recidivate`.

Guidelines

As described above, the assignment is the following.

1. **Use the data contained in `recidivism_data_sample.csv` to train and evaluate several predictive models that predict whether or not a criminal defendant will recidivate.**

2. **Choose a final model you consider would be best to deploy in the real world.**

In completing these tasks, you should follow the guidelines below.

- Use `R` for all coding procedures and complete the assignment using `R Markdown`.

- You must try at least three supervised learning methods we have covered so far in this class (e.g. logistic regression, LASSO, random forests, etc.).

- Follow the best practices and processes we've learned about for training predictive models to perform optimally on out-of-sample data. Use comments in your code to document the approaches and decisions you are making.

- From any of the methods you have tried, pick one final model that you think is "the best." This should be the model that you would choose to deploy in the real world. Justify that decision, discussing the metrics and other considerations you have focused on that have led you to pick that model over the other models. In other words, you must not only determine which is the best, but you must first define what you mean by "best" in this context.

- In your code, apply your final model to the pseudo data that are contained in `recidivism_data_pseudo_new.csv` to make predictions (predicted probabilities and classifications) for the observations in that pseudo data set. It is vital that you ensure that this process is performed successfully. This will ensure that your model is built properly and is compatible with out-of-sample data, which will allow us to assess your model on the real held-out data set that we (the instructors) have.

- Use detailed comments throughout your file to explain and document all of your decisions, assessments, and results. Your submission does not need to be structured like an essay or standard paper, but it does need to contain a sufficient amount of explanation in a sufficient amount of clarity to allow us to understand what you did and why.

Evaluation

Your grade for this challenge project will be determined by:

- Successful implementation of three different supervised learning methods that we have covered in class. (15%)
- Proper execution and usage of best practices in training your machine learning models. (30%)
- Clear and concise documentation of your work and coding procedures. (20%)
- Justification for any key decisions you make in training, selecting, and assessing models. (25%)
- Your final chosen model's performance on test data that we have held out. (10%)