DATA 403 Project 2

Fall 2023

Codebook:	Wednesday, November 15th
Presentation:	Friday, November 17th
Report:	Friday, November 17th

Overview

You are a data scientist at a bank working in the mortgages division. You have been tasked with building a model to predict whether an applicant will be able to repay their loan. Use the data from this Kaggle competition, specifically the data in application_train.csv, to build and evaluate your model.

The team with the best score on the Kaggle competition will get + 10 points of Extra Credit for each member. You may only submit **once** to the competition.

Data Prep and Feature Engineering

Decisions about how to prepare your data and how to create new variables are yours to make.

If you can think of a way to incorporate external data into your modeling process, that is very much allowed, but this is not a requirement of the project.

Model Implementation

You are welcome to use *Scikit-Learn* or *tidymodels* to prototype and to fit all models, but you should implement the **metrics** and the **cross-validation process** from scratch (you may use math operations such as those in *numpy* or *MASS* and data wrangling functions such as those in *pandas* or *dplyr*). You will then need to report the results from your implementation.

Model Selection

During your model selection process, you should try all of the following linear classification algorithms:

1. logistic regression, possibly with penalization
2. support vector machines
3. linear discriminant

You will need to decide what features to include and what values of hyperparameters to use.

One goal of this project is to compare the three classification algorithms. In order to make them comparable, you will need to consider the same options for features. How similar are their predictions? Can you explain why?

Test, Training, and Validation Sets

A big part of this project is how you split up the data into test, training, and/or validation set(s).

You should try, at a minimum:

1. A completely randomly sampled split
2. A stratified split
3. A split that is chosen in a non-random way, so that your test and/or validation sets can be considered to more accurately represent the data that will be seen when the system is deployed

How does changing the text/training split strategy impact performance of validation? Which approach makes you feel most confident about the model's eventual performance on the Kaggle holdout set?

Prediction Metrics

The Kaggle competition is judged on ROC-AUC, so you should certainly consider this metric in your analysis.

You should also consider at least **three** other possible metrics (more is even better!):

1. Accuracy
2. F1 Score
3. Another common metric of your choosing

Fairness

Find a best-performing model in terms of fairness, using one (or more) of the fairness metrics presented in DATA 401. Possible protected attributes to consider include gender and age. Analyze how well this model performs on the other prediction metrics you used, to see if there is a conflict between fairness and prediction performance.

Deliverables

Written

Turn in a 2-5 page report that discusses:
- how you chose the features and hyperparameters that you used
- how you decided between logistic regression / SVM / LDA
- an evaluation of the final model
- any ethical considerations with the use of the model

Coding

Turn in a Colab or R Markdown/Quarto notebook containing:

- Your implementations of the three classification algorithms
    - You will be partially graded on coding style. Your code does not have to be perfect, but it should not include any major inefficiencies, such as using a for loop when you could have manipulated vectors.
    - Your code should be well commented.
- A comparison of the predictions from the three algorithms.
    - How similar are they?
    - Can you explain why they agreed or disagreed in certain situations?
    - There are many ways to compare predictions from machine learning models; we will leave it to you to determine the best way to communicate your findings.
    - You should communicate your findings primarily through tables and graphs.
- An analysis of your “fair” model with respect to fairness and prediction performance

Presentation

On Friday, Nov. 17th, your group will give a 10 minute max business report presentation in class.