DSCI 552: MACHINE LEARNING FOR DATA SCIENCE

PROBLEM SET 2

Instructors: Dr. Kristina Lerman (lerman@isi.edu) and Dr. Keith Burghardt (keithab@isi.edu)

Deadline: Thursday, February 25, 2021, 10 A.M PDT

You can submit the report on Blackboard and code on GitHub Classroom. As long as the problem set is open, you will be able to upload multiple answers (the last attempt will be graded).

TASK (20 points)

You are a new hire in a mid-size company. You have just completed your first task. Today morning, you handed your report in and now you circle around the office. You feel exhausted and a bit nervous. You don't really know what to expect next. After the lunch, you were approached by your technical manager. "Our boss wants to see you and we have to discuss the next task for you", he communicated. Following your manager, you enter the meeting room. The CEO and the senior developer already wait inside.

Your CEO said: "You did an excellent job! We are truly lucky, that you decided to join our firm. Our clients were very pleased with the model that you designed for them. The report was very helpful, they highlighted, that it was important for them to correctly understand the limitations of the model and to know the expected performance after it is deployed in a real world. However, now we have another task for you. We want to start a trial with a major hospital. I want you to prepare a proof-of-concept, so we can convince them, that a partnership with our firm can be beneficial for them. You will get a historical medical data. I want you to design a model, that can classify, if a certain treatment is recommended for the patient or not. Additionally, there are 6 additional features (denoted as **TestA, TestB, GeneC, GeneD, GeneE** and **GeneF** in your dataset) that we can use. However, they are really expensive and difficult to collect. I want you to assess, how useful they are. We meet with the hospital in two weeks. I want a detailed report describing your main findings, on my desk, on Thursday, February 25, at 10
A.M.

Hint: Your boss called your task "proof-of-concept", but in fact, the nature of that assignment is the same as the last time. You are asked to train a classification model and you must measure how good that model is. Additionally, you must give recommendations which features are important to collect. You should look

*at all variables, but at minimum, you should test the importance of **TestA, TestB, GeneC, GeneD, GeneE** and **GeneF**.*

Your Technical Manager said: "This time it really matters, that your model has a good performance. If we can show that our model makes less mistakes than a human doctor, it would be a big deal. Describe exactly how you tested your model. They are really going to look at that section. Additionally, similar to the last time, the interpretability of the model is very important. You should restrict yourself to logistic regression."

Hint: Remember, that accuracy alone, is not a good measure. We care both, about accuracy, precision and F1-score. Report also false positive and false negative. To choose a right model, you can use for example the AUC score. It is ok (it's even expected) that you will do some feature engineering. You can also try to add regularization to your logistic regression and test if it helps you or not. To show that the model can be interpreted, you can identify and explain the most important relations between the variables and the expected outcome (e.g., how the probability that the treatment is recommended changes with age? Or gender?).*

The Senior Developer took you aside and said: "My task is to deploy your model to production. But I cannot deploy a paper-report. I need your code. However, remember that I am not a Data Scientist list you. I have a different expertise. I will read your code, but you should make sure that I can follow and understand it – and that I know how to use it."

Hint: In the ideal case, people should be able to take your code, run it and recreate all your results. In a less ideal case, it should be a demonstration of typical run. The code should demonstrate your approach end-to-end. People should just specify the path to the dataset, run it and see final results. Another name for this is a technical demo. At your future work, you might be quite often asked to demo your results. People will expect you to present an end-to-end example where you read the raw data, train your model and evaluate the results of the predictions.

Data

You can find the dataset ps2_public.csv on the Assignments section of Blackboard.

Report

To help you, I prepared a template. See https://www.overleaf.com/read/vnvhqxkpdhbk. You are encouraged to use the template but you are free to use other editors or make modifications. Just ensure that the final submission of the report has to be in PDF. Submission of report has to be done on Blackboard.

Code Submission

We have created a GitHub Classroom where you can create private repositories. We will update the class on Piazza on how to go about uploading your solutions on this platform.

Grading Rules

In order to grade your work, we will role-play the following situation. We will assume, that you are a new-hire in our company. You are asked to provide a comprehensive technical report that illustrates your findings. We will evaluate it from the perspective of three people.

- Your CEO (she would like to hear high level stuff. She will probably only read the conclusions and look at main figure). (4 points for report).

- Your manager (he would like to see a detailed report; he might also look at some parts of the code). (6 points for report and 2 points for code).

- A senior developer (they would like to see the code and won't read the report at all). (8 points for code).

Your final score is: 10 points for report and 10 points for code.

Don't Panic

Don't panic. We understand that this is a large, open ended task. We also understand that this might be the very first technical report that you were asked to write. We are dedicated to help you do your best work all while keeping the standards high. We acknowledge that you have limited time and resources to complete the task. This report doesn't have to be perfect for 100% score.

If you don't know where to start read the Second Chapter of "Hands-On Machine Learning with Scikit-Learn, Keras and Tensorflow", 2nd Edition by Aurélien Géron. Check also the Appendix B. Machine Learning Project Checklist from that book.

If something is not clear, ask your questions on Piazza.

Note: Cite any source you use (even if you adopt/copy a snippet of code). Failure to do so would amount to plagiarism.

Optional Challenge

We also created an optional challenge for you. There is no additional credits for participating in it. However, we encourage you to give it a try. We created a special class competition on Kaggle (https://www.kaggle.com/c/usc-dsci552-section-32416d-spring-2021-ps2). Link to participate in the competition: https://www.kaggle.com/t/420b31050f1248c4a2c95ca73a25051a. You will find a special test dataset, where I removed the price column. Train a model (you are not restricted to linear models anymore) and make your predictions. Have fun!