DSCI 552: MACHINE LEARNING FOR DATA SCIENCE

PROBLEM SET 1

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

Deadline: Thursday, February 11, 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 were approached by three people, your CEO, your direct technical manager, and a senior developer. They explained that your first task is to explore a certain dataset, propose and fit a predictive model, evaluate the model performance, and interpret the results.

Your CEO said: "The dataset describes conditions of various used cars and their current prices. I would like to learn what drives prices of used cars. Look at the dataset and find the main factors that affect the value of a car – and then explain it to me. Additionally, assess the impact of some special modifications (denoted by F1, F2, F3 and F4 in your dataset) on the price. This would help us to understand, if we should make the modifications before selling a car or not. I would like to see the report, describing your main findings, on my desk, on Thursday, February 11, 2021 at 10 A.M. "

Hint: You are asked to find general trends in the data. Report whatever you think is the most important. Your CEO doesn't want to see a list that is 20-times long. She would like to learn just about some general trends. To give you an example, one general trend could be "The price decrease with the age of the car. Holding all other factors constant, with each year, the price of a car decreases by $570. However, these dynamics are not constant. Value of younger cars decreases faster than the value of an old car. For example, the value of cars that are less than 5 years old, decreases nearly $2,500 per year." (This is just an example; your numbers might be different). Your second task you have to check both, the impact and the statistical significance of the F1-F4 attributes for making the price predictions.

Your Technical Manager said: "I would like you to propose a predictive model, that can be used to determine price of a used car. The problem is that the state-law demands that this model be easily interpretable. It means that we are restricted to use simple methods like Linear Regression, Ridge Regression, LASSO and Elastic Net. Additionally, we need to know how

Linear Regression
Random Forest?

accurate the model is. You must choose the best model and report its root mean square error. Describe everything in your report and I will study it carefully”.

Hint: In the most typical approach, you need to build three datasets: a training set, a validation set and a test set. You will use validation set to determine the best model; the test set to estimate model accuracy. In your report you should describe how you trained the models, how you selected the best one and how you tested its performance at the end.

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 `used_car_dataset.csv` on the Assignments section of Blackboard.

Report

https://www.overleaf.com/read/vnvhqxkpdhbk ✓

To help you, I prepared a template. See https://www.overleaf.com/project/600e2b977f2c[illegible]646f599. 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-ps1). Link to participate in the competition: https://www.kaggle.com/t/1f4f39dbdf17407895ef5811bf575807. 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!