DSCI 552: Machine Learning for Data Science

Problem Set 1

Instructor: Dr. Marcin Abram – email: mjabram@usc.edu

Deadline: Thursday, February 4, 2020 at 10 am PDT

You can submit the solutions (pdfs) on Blackboard. The deadline to submit solutions is always the second Thursday at 10 am (PDT). As long as the problem set is open, you will be able to send multiple answers (the last submitted answer will matter).

Task (20 point)

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 price of a used car. 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 as `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 a report, describing your main findings, on my desk, on Thursday, February 4, at 10 am."

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-items long. She would like to learn just about some "general trnds". To give ou 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, that dynamics is not constant. Value of a younger cars decreases faster then the value of an old car. For example, the value of cars that are less then 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 must be easily interpretable. It means, we are restricted only to simple models, like linear regression, Ridge regression, LASSO, or Elastic Net. Additionally, we needs to know, how accurate the model is. You must chose the best model and report its root means square error. Describe everything in your report. 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 validatoin set to determine the best model; then you will use the test set to estimate the model accuracy. In your report you should describe how you trained the models, how you selected the best one, and how you tested it's performance at the end."

The senior developer took you aside and said: "My task is to deploy your model to production. But I can not deploy a paper-report. I need your code. However, remember that I'm not a Data Scientist like 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 a "typical run". The code should demonstrate your approach end-to-end. People should just specify the path to the dataset, run it, and see the final results. Other 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.

Learning Objective: You will be able to design a machine learning pipeline. You will be able to create a technical report. You will be able to communicate your findings with several archetypal people.

Data

You can find the dataset, as `used_car_dataset.csv` file, at `https://www.kaggle.com/c/usc-dsci552-32415d-spring2021/data`.

Report

To help you, I prepared a template for you, see `https://www.overleaf.com/read/qcbcsdscvcdb`. You are encouraged to use the template, but if you prefer to use another editor – or to made some modifications – this is

acceptable. Just remember to always export your report to the pdf format at the end.

Report Submission

Submit your report (a pdf file) using Blackboard (check the "Assignments" section on Blackboard page of our class).

Code Submission

We created for you a GitHub Classroom, where you can create a private repositories, see `https://classroom.github.com/classrooms/77759207-usc-dsci552-32415d-spring2021-classroom`. In the following days we will add you to the GitHub Classroom. We will communicate the details of that process using Blackboard announcements.

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 were 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 the high-level stuff; She will probably read only the conclusions and look at the main figure). *(4 points for the report)*
- Your manager (he would like to see a detailed report; he might also look at some parts of the code). *(6 points for the report and 2 points for the code)*
- A senior developer (she would like to see the code; she might not read the report at all). *(8 points for the code)*

Your final score is the sum of scores given by each person. Cumulatively, there are 10 points for the report and 10 points for the code.

Each person (the CEO, the manager, the developer) will use the following grade rubrics.

Grade Component	Meets Expectations	Approaches Expectations	Needs Improvement
Completeness			
Clarity and Support			
Validity			

By "Completeness" we mean, that all parts of the question are addressed. By "Clarity" we mean, that the text and the code is written in accessible way. By "Support", we mean that you provide sufficient evidences to back-up your statements. You must cite any source you use (even if you happened to copy or adapt a snippet of a code from the internet – you should still treat it as a citation. You should clearly mark how large is that snipped and provide adequate references).

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 tasked to write. We are dedicated to help you do the best work – and while we keep high standards for you (and for us) – in the same time, we acknowledge, that you have limited time and limited resources to complete your task. This report doesn't have to be perfect to give you 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 questions using either Blackboard Discussion Forum, or (which seems to be the proffered mode of communication for most of you) Piazza.

Optional Challenge

We also created an optional challenge for you. There is no additional credits for participating in it[1]. However, we encourage you to try. We created a special class competition on Kaggle. Go to `https://www.kaggle.com/t/54b90f9a951b4d03805f400b5fa0be46`. You will find there a special file, called `used_car_dataset_PREDICT.csv` – it is a special test dataset, where I removed the "price" column. Train a model (you are not restricted to linear models anymore – you can use whatever you want) and make your predictions. On Thursday, February 4, during our lecture, I will present the leaderboard.

Have fun!

[1] Grading students based on the in-class competition is a very bad educational technique! We try to avoid it in this class. However, creating a small competition, and making the participation optional – this could be fun!