DS3000_FP4_SectionX_GroupY

December 9, 2019

DS 3000 - Fall 2019
DS Report
Project Title
Team Member Names

Executive Summary: Add your summary here (100-150 words)

Provide a brief summary of your project. After reading this executive summary, your readers should have a rough understanding of what you did in this project. You can think of this summary in terms of the four sections of the report and write 1-2 sentences describing each section.

0.1 Outline

1. INTRODUCTION
2. METHOD
3. RESULTS
4. DISCUSSION

0.2 1. INTRODUCTION

In this section, orient your readers to your project. You've already written some of these in previous deliverables. Based on your final analysis, revise your problem statement and write a concise introduction section. This section should touch upon the following points, but should be written in full paragraphs. Your writing should incorporate all of these points (and more if you like) in a coherent way. Remember that you are trying to convince your readers that this is an important problem to tackle.

Problem Statement * Describe the problem you would like to tackle. * What is the topic of your project? * What do you want to learn about it?

Significance of the Problem * Why is it important to tackle this problem in your project? * In what ways could the insights from this project be useful? * Has there been previous work on your topic? Do some research into your topic. Cite your sources appropriately. You can use the numbered reference format or APA (if you are more comfortable with it).

Questions/Hypothesis * End this section with a list of questions and hypotheses * You should tie these questions/hypotheses to the problem statement and its significance * e.g. Given the aforementioned problem and its importance, we set out to tackle the following questions:

Requirement: * You should have at least one question tapping into the comparison of various machine learning algorithms in predicting your target variable from your features variables.

0.3 2. METHOD

0.3.1 2.1. Data Acquisition

- Describe where you obtained your data. Provide a link to the original source.
- If you scraped your data, include your code as a script file.
- Your data should be stored in an online repository (e.g., GitHub) and your code should retrieve your data from that online resource. You can read csv files from the Web in the same way that you read files from local drive.
- Describe the dataset and variables. What do variables represent?

0.3.2 2.2. Variables

- If you are testing hypotheses, what are your IVs and DVs?
- For your predictive models, what are your features and target variables?

0.3.3 2.3. Data Analysis

- Specifically describe your predictive model. What outcome variable are you going to predict from what feature variables?
- Describe whether this is a supervised or unsupervised learning problem. Also identify the sub-category of the learning task (e.g. classification).
- What machine learning algorithms are you going to use? Why?

0.4 3. RESULTS

0.4.1 3.1. Data Wrangling

- Perform simple data cleaning (delete extra columns, deal with NA values, etc.)
- Perform data wrangling to extract your features and target values (e.g., grouping your dataframe by columns, applying functions to format dataframes, etc.)
- Preprocess your variables (e.g., scaling/transforming feature variables to normalize them)
- Feature extraction (dummy variables, new features from existing features, etc.)
- Use one feature selection technique to select a subset of your original features

0.4.2 3.2. Data Exploration

- Generate appropriate data visualizations for your key variables identified in the previous section
- You should have at least three visualizations (and at least two different visualization types)
- For each visualization provide an explanation regarding the variables involved and an interpretation of the graph.
- If you are using Plotly, insert your visualizations as images as well (upload the graph images to an online source, e.g. github, and link those in Jupyter Notebook)

0.4.3 3.3. Model Construction

- If you proposed hypotheses, conduct your hypothesis tests
- For your machine learning question(s), split data into training, validation, and testing sets (or use cross-validation)

- Apply machine learning algorithms (apply at least three algorithms)
- Train your algorithms

0.4.4   3.4. Model Evaluation

- Evaluate the performance of your algorithms on appropriate evaluation metrics, using your validation set
- Interpret your results from multiple models (and hypothesis tests, if any)

0.4.5   3.5. Model Optimization

- Tune your models using appropriate hyperparameters
- Explain why you are doing this (e.g., to avoid overfitting, etc.)

0.4.6   3.6. Model Testing

- Test your tuned algorithms using your testing set

0.5   4. DISCUSSION

- Provide a summary of the steps you took to analyze your data and test your predictive model

- Intepret your findings from 3.4., 3.5, and 3.6

    - Which algorithms did you compare?
    - Which algorithm(s) revealed best performance?
    - Which algorithm(s) should be used for your predictive model?

- If you tested hypotheses, interpret the results. What does it mean to have significant/non-significant differences with regards to your data?

- End this section with a conclusion paragraph containing some pointers for future work *(e.g., get more data, perform another analysis, etc.)

0.5.1   CONTRIBUTIONS

- Describe each team member's contributions to the report (who did what in each section)
- Remember this is a team effort!
- Each member of your team will provide peer evaluation of other team members. Your final grade on the project will be based on those peer evaluations.