Spam Text Case Study Rubric

DS 4002 – Fall 2024 - Jessica Li
Individual Assignment

General Description: You will build a predictive model to detect which words are associated with spam texts.

Why am I doing this? This assignment will give you a basic introduction to text analysis and classifying algorithms like Naives Bayes. Additionally, you will become familiar with GitHub, which is a useful tool in data science, computer science, and other tech fields. Finally, you will become more familiar with model evaluation metrics when looking at your final model.

What am I going to do? First, you will read and understand what your task is by reading the rubric and introduction document. Next, you will go to the GitHub repository to understand the dataset and the variables in it. Once you have done that, you will begin brainstorming how to clean the dataset and which model to use. After that, you will begin building your model. Afterwards, make sure to look at the evaluation metrics to determine how your model is doing and think of the reasons why behind its behavior.

Your final deliverables should include: A GitHub repository containing all materials used

Tips for success:

- Ask for help when you need it:
    - Your professor and TA are here to help you, so don't be afraid to ask for help.
- Research online:
    - Some basic code and resources are provided, but you're probably not going to find everything you need. Researching online is a great way to find what you need for this case study.
- Start general, then go granular:
    - First start with a general idea of what you want to do and what tools are appropriate. Also, make sure to look through the dataset and what the finished product should look like.
- One step at a time:
    - Don't overwhelm yourself by solely thinking of the final product. It helps to break things down into manageable pieces and go through them one at a time.

How will I know I have Succeeded? You will meet expectations on this case study when you follow the criteria in the rubric below:

[redacted]	[redacted]
Formatting	● GitHub repository (submitted via link in Canvas) ○ Should include: ■ README.md ■ Source code file ■ Dataset ■ REFERENCES.md ■ LICENSE.md
README.md	● Brief summary of what you've produced for the case study, this does not have to be highly detailed. It should describe the project context, model-building process, and the dataset you used.
Source Code File	● The code you used to build the model o Well-documented and commented
REFERENCES.md	● Cite any resources (journal articles, websites, etc.) used to create your model in IEEE documentation style.
LICENSE.md	● Terms if others want to use your work ● Use MIT format

Acknowledgements: Thank you Professor Alonzi for providing the rubric structure!