Rubric

Your instructional team will evaluate your project using the following criteria.

For Project 3 the evaluation categories are as follows:

The Data Science Process

- Problem Statement
- Data Collection
- Data Cleaning & EDA
- Preprocessing & Modeling
- Evaluation and Conceptual Understanding
- Conclusion and Recommendations

Organization and Professionalism

- Organization
- Visualizations
- Python Syntax and Control Flow
- Presentation

Scores will be out of 30 points based on the 10 categories in the rubric.

3 points per section

You must get at least a *1* in each section to pass the project.

Score	Interpretation
0	*Project fails to meet the minimum requirements for this item.*
1	*Project meets the minimum requirements for this item, but falls significantly short of portfolio-ready expectations.*
2	*Project exceeds the minimum requirements for this item, but falls short of portfolio-ready expectations.*

3	*Project meets or exceeds portfolio-ready expectations; demonstrates a thorough understanding of every outlined consideration.*

The Data Science Process

Problem Statement

- Is it clear what the goal of the project is?
- What type of model will be developed?
- How will success be evaluated?
- Is the scope of the project appropriate?
- Is it clear who cares about this or why this is important to investigate?
- Does the student consider the audience?

Score: 2

Comments: Great, specific problem statement. How will you determine which does better (i.e. what metric(s)). Who could benefit from this project?

Data Collection

- Was enough data gathered to generate a significant result?
- Was data collected that was useful and relevant to the project?
- Was thought given to the server receiving the requests such as considering number of requests per second?

Score: 3

Comments: Nice work. Very detailed. Consider other error handling to make it even better: other status codes, or a successful request containing no data, etc.

Data Cleaning and EDA

- Are missing values imputed/handled reasonably?
- Are distributions examined and described?
- Are outliers identified and addressed?
- Are appropriate summary statistics provided?
- Are steps taken during data cleaning and EDA framed appropriately?
- Does the student address whether they are likely to be able to answer their problem statement with the provided data given what they've discovered during EDA?

Score: 3

Comments: Thoughtful process. Null values are taken care of. Good that you noticed more cleaning was needed later and looped back. It's ok to include that in your earlier notebook after you find it. What about [removed] values in the selftext or title?

Preprocessing and Modeling

- Is text data successfully converted to a numeric representation?
- Are methods such as stop words, stemming, and lemmatization explored?
- Does the student properly split and/or sample the data for validation/training purposes?
- Does the student test and evaluate a variety of models to identify a production algorithm (AT MINIMUM:Bayes and one other model)?
- Does the student defend their choice of a final model relevant to the data at hand and the problem?
- Does the student explain how the model works and evaluate its performance successes/downfalls?

Score: 3

Comments: Very thorough. Since the target here is binary, regression doesn't really make sense. Interesting to look at the other features in isolation from the title and selftext. Would have been cool to see feature importances and some interpretation there. If you go down a path that doesn't lead anywhere, it is ok to leave it out of your project.

Evaluation and Conceptual Understanding

- Does the student accurately identify and explain the baseline score?
- Does the student select and use metrics relevant to the problem objective?
- Does the student interpret the results of their model for purposes of inference?
- Is domain knowledge demonstrated when interpreting results?
- Does the student provide appropriate interpretation with regards to any descriptive and inferential statistics?

Score: 2

Comments: Lots of metrics were calculated, but not really used. Maybe focus on a couple important ones to keep the message clear. If one subreddit isn't more important than the other then it makes the most sense to focus on basic accuracy.

Conclusion and Recommendations

- Does the student provide appropriate context to connect individual steps back to the overall project?
- Are the conclusions/recommendations clearly stated?
- Does the conclusion answer the original problem statement?
- Are future steps to move the project forward identified?

Score: 3

Comments: Great work.

Organization and Professionalism

Project Organization

- Are modules imported correctly (using appropriate aliases)?
- Are data imported/saved using relative paths?
- Does the README provide a good executive summary of the project?
- Is Markdown formatting used appropriately to structure notebooks?
- Is the process explained in comments/Markdown?
- Are files & directories organized?
- Do files and directories have well-structured, appropriate, consistent names?

Score: 3

Comments: Great job!

Visualizations

- Are sufficient visualizations provided?
- Do plots accurately demonstrate valid relationships?
- Are plots labeled properly?
- Are plots interpreted appropriately?
- Are plots formatted and scaled appropriately for inclusion in a notebook-based technical report?

Score: 2

Comments: Great job using visuals to help with cleaning. Make sure your plots have titles. Some of them could be a little larger. Would have been nice to see some comparisons between the subreddits (word counts, post lengths, etc.).

Python Syntax and Control Flow

- Is care taken to write human readable code?
- Is the code syntactically correct (no runtime errors)?
- Does the code generate desired results (logically correct)?
- Does the code follows general best practices and style guidelines?

Score: 3

Comments: Wow, amazing! Obvious command of python. But I might say be careful not to increase the complexity of your code or abstract too much away with too many functions depending on the purpose of the project. If the plan is to turn it into an app that's great, but if it is more exploratory in nature, it might be good to see more of the steps without jumping around looking for function definitions. Certainly not a problem, just maybe something to think about.

Presentation

- Is the problem statement clearly presented?
- Does a strong narrative run through the presentation building toward a final conclusion?
- Are the conclusions/recommendations clearly stated?
- Is the level of technicality appropriate for the intended audience?
- Does the student appropriately pace their presentation?
- Does the student deliver their message with clarity and volume?
- Are appropriate visualizations generated for the intended audience?
- Are visualizations necessary and useful for supporting conclusions/explaining findings?

Score: 3

Comments: Nice work! Maybe a little rushed and could use some more visualization.

Total Score: 27/30 (PASS)

Why did we choose this project?

This project covers three be concepts we cover in the course: Classification, Natural Language Processing, and Data Acquisition.