Data Science Immersive: Capstone Project 1 Proposal

Prepared By Benjamin Olsen

### What Is The Problem You Want To Solve?

* The Myers-Briggs Type Indicator (MBTI) test, while long lauded by high school counselors and internet quizzes, has often found itself falling out of grace with academic and clinical psychologists due to its self-scoring and poor re-testability. Be that as it may, people everywhere still find validation in the self-administered test, and the MBTI has a loyal following of adults and teenagers alike. While the clinical application of the test has been discarded as unreliable and ultimately left to the wayside by clinicians, the test's usefulness has yet to be evaluated in the context of online behavior. This project will attempt to validate the relevance of the MBTI in an online social context based on an analysis of a person's online posting style. While MBTI is no longer considered to be theoretically validated construction of personality, a noticeable correlation in reported personality versus content of posts produced by this project could potentially point to a relevance in testing other, more clinically sound personality tests (such as the Big 5) as the information becomes more readily accessible.

### Who Is Your Client And Why Do They Care About This Problem?

* The primary application of this project is to understand whether MBTI has implications for targeted advertising. If there theoretically was a correlation between the content of a post and the personality (or self-perception) of its poster, an advertiser could potentially offer tailored, highly relevant targeted ads to subgroups of end-users based on the raw information contained in patterns within their posts. Combined with those users browsing habits, such information could help build an advertisement preference portfolio that evolves with the user (or group of users) based on their progressive behavior. While the MBTI has questionable usefulness in of itself, the fact that the results are self-reported could help indicate what *themes* those individuals strongly self-identify with, making the MBTI a decent starting point in terms of categorizing people based on what traits are exhibited in their posting history.

### What Data Are You Going To Use For This? How Will You Acquire This Data?

* The data in question comes from: https://www.kaggle.com/datasnaek/mbti-type
* This dataset contains over 8600 rows of data, on each row is a person’s:
  + Type (This person’s 4 letter MBTI code/type)
  + A section of each of the last 50 things they have posted (Each entry separated by "|||" (3 pipe characters))

### In Brief, Outline Your Approach To Solving This Problem.

* The first step would be to classify each comment according to the corresponding MBTI category, then analyze the text to create a baseline for which to compare additional comments.
* The next step would be use this baseline to try to accurately predict what classification of MBTI our test data will fall into (predicting what a commentator's personality type is based on a comment).

### What Are Your Deliverables?

* Original raw dataset .
* A wrangled dataset containing all comments split into individual entries and categorized by personality type reported.
* Codebase for parsing and visualization of data.
* Full textual analysis of the data, as well as laying out the architecture of the classification model.
* A summary of a testing of the accuracy of the model, and discussion of the results.
* Slide deck Summarizing findings.