**Data Analytics Boot Camp November, 2021 – May, 2022**

**Individual Self Assessment - Martha Richardson**

**Self Assessment**

My active roles included:

* Cleaning: programmed outliers logic and removed those rows, filled in total salary for base salary where it was missing.
* Preprocessing: coded binning for categorical data columns to preprocess for ML models, designed the original region binning which was based on US Census regions, and later coded an API call to retrieve lat/long data for cities in our data and binned each row based on its proximity to the 9 metro areas with largest concentrations of tech jobs. I coded a salary level field that created $25,000 bands so we could use logistic regression techniques to predict salaries.
* Cheryl did most of the first pass coding of the machine learning model but I worked extensively to refine them, testing subsets of the data in different ML models (by title, region, company, unsupervised learning class)
* I completed an unsupervised learning model which returned fairly coherent groups by salary, region, and experience.
* I copied Lucien’s code in the HTML and python app.py to create a fourth tab in our website which predicted salaries at the top 5 companies

After the first segment I was not very involved in the presentation, which was completed by Cheryl, and Lucien programmed the entire web/dashboard segment of the project from scratch by himself. As we wrapped up the project, I contributed by suggesting edits and proofreading the presentation and the website/dashboard. Throughout all phases of the project our whole team discussed and decided together on the direction and details of the project.

The most challenging thing for me was regrouping when our original expectation of a clear and accurate linear regression did not emerge from the data. For about a week it felt like our project was going to be an inaccurate mess. We overcame that issue by digging deeper into the data to find ways to improve accuracy using subsets of the data and unsupervised learning. By the end we had a deeper understanding of how our dataset worked, what we could reasonably deduce from it, and why it wasn’t what we expected.

**Team Assessment**

Communication protocol: We used Slack to communicate and worked together during class and office hours, with a couple of other zoom meetings early on. Since both of my teammates were working real world jobs while I was not, most of my work was completed during the workday. It is possible/probable that I sent too many messages out on Slack at times. I sometimes felt frustrated if responses to my questions took more than a day, and sometimes I had to just wait until we were all meeting again to get an answer. In another project, I would make more specific protocols about communication, including appropriate hours for sending messages and an expectation of how soon people ought to respond when a teammate asks a question.

Team strengths: Independence in our individual areas of the project, no one was a slacker and all contributed a lot to the project. There was no one who was “carried” as often happens in group projects. To a new cohort I would say, please know that your project will probably not be exactly what you envisioned at the outset. Things will progress and change as the project develops but it will all work out.

**Summary of Project**

We analyzed a Data Science and STEM salary dataset that we downloaded from Kaggle. We looked for trends in the data, using linear regression and a random forest supervised machine learning model to create salary prediction tools based on the smallest number of user inputs. Our resulting dashboard predicted salaries with user inputs (total years of experience, years of experience with current company, and region/company) with an accuracy between 40% for the random forest model (based on the whole dataset) to 50-60% with the linear regression models (based on the top 5 companies). The ranked feature importance analysis from the Random Forest machine learning model suggests that we used the most important features to predict on, and an unsupervised machine learning analysis supports the region groups we created.