Mcnulty Project – Kaggle Instacart Dataset
2/21
By: Michael Du

Server Hosting:
For running Jupyter Notebook on, I used EC2.
To transfer the necessary CSV files, I used Filezilla.

Things to note:

- For feature engineering, there were two issues that I ran into.
    - One was running into memory errors because the files were taking up too much space when I was joining them. I managed this by downcasting the column types, mainly from int64 to anything from int8 (Boolean columns) to int32. This was something Mike(the other one) showed me so for the first half of my project, I had to deal with a lot of kernel crashes
    - Second was dealing with the NaN values for Days Since Prior Orders, because products that were purchased for the first time would end up with an NaN. I could not simply discard these rows because NaN itself was indicative of first time purchase, which is relevant in predictions. I dealt with this two alternative ways. The first method was to create separate bins, which was used for Logistic Regression. This created 7 extra dummy variable columns. The second method was to convert the NaN values to -1, which I used for tree based methods.
- Because the dataset was so large (8+ million after grouping), I took out a 5% subset to run train/validation on. This 5% subset was assigned a 70/30 train/validate split, based on user_id
- There was significant imbalance, which reflected in my baseline logistic regression run, so I applied an even-weight oversampling.
- I ran logistic regression on individual features to gauge f1 scores for each one. In retrospect, and per Debbie's advice, I should've done Lasso regression or use XGBoost feature importance to identify important features.
- For my business approach, my intent was to use F1 as a more well-balanced metric, and utilize recall solely to calculate the dollar amount of revenue from reorder we'd be able to predict. In this case our goal was NOT to maximize recall. I was simply using the recall (0.3) to determine the percent of ALL reorders (10% of all orders are reorders) that was predictable. In this case, our model with it's 0.3 rate will allow us to predict the equivalent of 3% of all revenue.

Other thoughts:
- I didn't get around to using Tableau or D3 because while I understand the importance of it, these are things I learn on my own later, especially when I return to my former job as a web developer (at least part of my role will still involve web-dev) I wanted to take advantage of my time here to really make sure I understand the fundamentals of ML concepts we're learning. If I had more time though, I'd definitely have gotten to it. It's just that until the last minute, I was still experimenting and better understanding the classification models I used in this project.