Progress Report
CTR Prediction

As a team we have decided to implement a Field-Aware Factor Machine model to predict Click-Through-Rate for this final project.

Individual updates:

Erik:

Work performed:

- I performed the EDA for Section 3 of the report in which I calculated the distinct values and null values in each field discovering that the data will be highly sparse. The sparsity of the data poses a big challenge to us and will be used to guide our effort in feature engineering which I work with Noah to implement.

- I also collaborated with Connor on Section 2 giving him feedback on the content he wrote.

Future work:

- As a next step, I will work with Anu on implementing the home-grown large scale solution we will run on the assigned dataset in Section 4
- Make notes on the key points to include in Section 5.

Noah:
I've focused on EDA, hashing features and will be orchestrating runs on GCP. Similar to Eric, I looked at the type and distribution of data, but dove a little more into looking at distinct values and their specific counts in bins of 10, 100 and 1000 counts. We found that many of the distinct values in certain categories had values <10 and we're considering binning these into the same bin.

I also looked at 3 Idiots kaggle winning competition submission and compiled their C++ library to understand it a little better. It provided some insight into the integer values and how to handle hashing them.

Although not completed yet, I'll be looking to create a pipeline on GCP for transforming the data, building models and testing. We decided as a group that we'd use a method similar to HW5 and submit to the cloud rather than having jupyter notebooks live that we run through.

Connor:
After we agreed to implement a Field-Aware Factorization Machine model for our project, I took the lead in section 2 of the paper, explaining the algorithm and the math behind it, as well as creating a toy example. Moving forward, I will finalize this section, and then assist in the building

up our base model implementation to work with the larger dataset. I will also be working on section 5, discussing the course concepts we decide are the most relevant to our project and model implementation.

Anu:

My area of focus is section 4, the implementation of the algorithm, Created two implementations. First was just a logistic regression implemented in PySpark to make sure we have some baseline and can read and write the data and get familiar with the format. Also conducted EDA on the features. The second algorithm is a the FFM methodology from Algorithm 1 of the paper https://www.csie.ntu.edu.tw/~cjlin/papers/ffm.pdf. This implemented in pyspark and tested on a small data set to see if it runs, which it did.The algorithm uses Stochastic gradient descent and parallelizes this across multiple partitions/nodes using broadcast variables to send coefficients and running a parallel update to the coefficient matrix.