Progress Report - Team 10

Team members: Curtis Lin, Eddie Zhu, Kai Qi Lim, Youzhi Wu

## What’s done so far:

* Kick off project with team meetings, preliminary research on the Kaggle contest, and literature review. [all]
* Question formulation: Based on a set of features, would the advertisement be clicked? [all]
* Split full dataset for training and development (80% - 20%) in Spark [Curtis]
* Toy dataset establishment with about 1000 random samples from training set: read toy data into parquet format. We’ve also decided to use this file format for full dataset. [Curtis]
* Conduct EDA:
  + Data size, format, description [Curtis]
  + Histogram [Eddie]
  + Scatterplot [Eddie]
  + Null count [Eddie]
  + Heatmap to check trends in null rows [Kai Qi]
  + Heatmap to check similarity among different features [Kai Qi]
  + Heatmap to check correlation between numerical fields [Chloe]
  + Count number of unique categories in columns 14-39 [Chloe]

## Plan for next step:

* Preliminary feature engineering [Kai Qi]:
  + Logarithmic transformation and normalisation
  + Think about how to handle NA values, correlated features, etc.
* Preliminary algorithm implementation [Chloe]:
  + Implement logistic regression for toy data set in notebook
  + Implement logistic regression for toy data set in Spark
* Experiment with other models [Curtis]
* Experiment with other feature engineering techniques [Eddie]:
  + Feature selection: use Decision Tree / GBDT / Field-aware Factorization Machines
  + PCA
* Hyperparameter tuning [Kai Qi]
* Implement (optimal) algorithm in Spark for full dataset [Chloe]
* Finalize our report with analysis, application of course concepts, and conclusions. [all]