**Facebook V: Predicting Check Ins**

“The goal of this competition is to predict which place a person would like to check in to. For the purposes of this competition, Facebook created an artificial world consisting of more than 100,000 places located in a 10 km by 10 km square. For a given set of coordinates, your task is to return a ranked list of the most likely places. Data was fabricated to resemble location signals coming from mobile devices, giving you a flavor of what it takes to work with real data complicated by inaccurate and noisy values. Inconsistent and erroneous location data can disrupt experience for services like Facebook Check In.”

Link: <https://www.kaggle.com/c/facebook-v-predicting-check-ins>

**Introduction**

This was a Kaggle Recruiting competition that occurred only a few months ago. Generally, competitions award either cash prizes, swag, or in this case recruiting opportunities. Facebook has been on Kaggle quite a few times already and this time they offered an extremely large data set that mimicks real-world location and time data.

Since this competition is already finished, you will find that everyone’s notebooks have been published and most of the data intricacies discovered. We even have the winner’s solutions available!

First place winner’s interview: <http://blog.kaggle.com/2016/08/16/facebook-v-predicting-check-ins-winners-interview-1st-place-tom-van-de-wiele/>

You are **NOT** expected to simply download the data and build a model without ever looking at Kaggle’s resources. Read them, all of them. You will find that many obstacles you face have been addressed already by someone else.

**Objective**

The evaluation metric is probably something completely new to you, called Mean Average Precision at 3 (MAP@3). This is often used for evaluating the quality of recommendations, such as the first 10 results on Google or the first dozen products returned from an Amazon search. Kaggle already has a FANTASTIC page on how the metric works:

<https://www.kaggle.com/wiki/MeanAveragePrecision>

Please read through it. At a high level, MAP quantifies quality between 0 and 1, zero being the recommendations are meaningless (e.g. none of them were clicked on) and one meaning all of your recommendations were effective (e.g. all of them were clicked on).

As before, in addition to the accuracy, you and your teammates will be drafting an ipython notebook detailing your data cleaning, exploratory analysis, and modeling methods.

**Logistics**

Project submissions (notebook and Kaggle score board submission) will be due by **Sunday 10/23 11:59pm.** Since this is a completed competition, your submission score will be scored twice – once for the public leaderboard (50% of the test data) and once for the private leaderboard (100% of the test data). Only your private leaderboard score will count for your grade. As before please submit your project notebooks on Gradescope.

**Notebook Guidelines**

1. Follow the general structure of notebooks you see on Kaggle:
   1. Basic data exploration with graphs, summary statistics, and outlier detections
   2. Data cleaning/Feature Engineering
   3. Model building
   4. Model evaluation
   5. Model interpretation
2. If you are in part of a team, you can split the work per section – but the best way to learn is to help do a little of everything. There are a lot of features and no person will produce the exact same analysis.
   1. Tip: The best model in terms of speed efficiency is probably KNN but, again, **explore the others**.Even though we won’t get to some of them that appear in the kernels (e.g. xgboost), it is a great way to have the team try out different things that others have used even if you don’t immediately understand what it does. Plus, it will boost your accuracy over everyone else ☺

**Grading**

We will set our own benchmark accuracy score and your model **MUST beat it to pass the class.** ☺

Points are distributed uniformly per team as follows:

70 points – Benchmark = (0.45836 Private Leaderboard Score)

.

.

100 points – Top team model.

Benchmark script: <https://www.kaggle.com/zfturbo/facebook-v-predicting-check-ins/mad-scripts-battle/code>

In fact, now that you are given the benchmark script, there should not be any reason for not being able to meet it. Another fact, this was (I think) the very first script published that gained a lot of traction for its simplicity since it used the very basic idea of counting instead of a standard machine learning model.

The notebook will also be pseudo-graded for effort and quality on a 0-50 scale (10 points X 5 sections in the notebook structure). The first two project notebooks **will not** count to your grade but your final two project notebooks will. We give a score so you can gauge what our expectations are.

For each section, point values are outlined below:

1. Data Exploration

2 points for graphs of any kind

2 points for statistics of any kind

3 points for thoroughness (did you analyze enough to get a good understanding of the data?)

3 points for quality/creativity

1. Data Cleaning & Feature Engineering

5 points for cleaning the data of the most apparent problems (e.g. missing values, outliers, dummy variables, etc.)

2 points for feature engineering any variable

3 points for “creative/unique” feature engineering

1. Model building

5 points for setting up the train and test sets correctly

2 points for successfully running a model

3 points for successfully running more than one model

1. Model evaluation

7 points for correct out-of-sample test accuracy to the public leaderboard accuracy (within a reasonable threshold)

3 points for demonstrating that your local test accuracy is correlated with the public leaderboard accuracy

1. Model interpretation

4 points for interpretation of any kind

3 points for thoroughness (did you fully grasp what your model is telling you about the data?)

3 points for comparing different models