10-718 Data Analysis (Spring 2019)
Assignment 2: Taxi Travel Time Prediction
Version 1.0

Beat Kaggle baseline by 5pm on March 13, 2019.
Submit hard copy of report between 2pm and 5pm
on March 20, 2019 to GHC 8004.

1 Overview

Assignment 1 was focused on exploring the data and articulating a machine learning problem that solved the domain expert's goal of estimating taxi travel times in New York City. In this assignment, we are interested in implementing a preliminary pipeline based on our Assignment 1 discussion.

We will concentrate our attention on the initial set of variables that the domain expert provided. As such, for this assignment you are expected to use only the given data **without including any external information**. In particular, this means that any engineered features must be derived from this data or the maps/tables which were distributed with Assignment 1, and that you cannot augment the data with outside sources. The only exception is pre-processing the location IDs to convert them to latitudes and longitudes, which you may do.

In addition, you may not use the unlabelled validation/testing data as part of your pipeline (e.g., as unsupervised training data). You have access to it only for the purpose of labelling it once you have completed training.

2 Data

Just as in the previous assignment, we will be using data recorded by New York City taxicabs for the 2017 year. As before, the starting and ending locations are pre-processed by discretizing the region and reporting the index of the starting and ending sub-regions. Details about the discretization scheme can be found here.

2.1 Pre-processing

For this assignment, we have also performed some of the pre-processing steps discussed in class. In particular, we have:

- Added a travel time feature (in minutes).
- Removed trips before 2017.
- Removed trips whose travel time is less than a minute or more than 2 hours.
- Removed trips with zero passengers.
- Removed trips whose pick up or drop off zones are undefined or unknown.
- Removed trips whose payment type does not correspond to type 1 or type 2.

These same steps have been performed on the testing and validation data used for the Kaggle competition.

3 Loss Function

We were unable to come to on a unanimous consensus in class about which loss function best suited our purposes. Nonetheless, for this assignment, we will be using the root mean squared log error. More precisely, we will evaluate a set of predictions using the expression

$$RMSLE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\ln\frac{1+\hat{y}_i}{1+y_i}\right)^2}$$

where N is the number of trips evaluated, y_i is the actual travel time for trip i and $\hat{y}_i$ is the predicted travel time for the same trip. We use this loss because it has two properties that we consider desirable for our purposes: (1) it avoids over-penalizing outliers, and (2) it penalizes under-estimates more than over-estimates.

4 Kaggle Competition

As part of this assignment, you will be asked to beat a simple baseline model. You will not have access to the model, but you will have access to its performance (in terms of its $RMSLE$) on a set of validation data. The baseline is thus a submission in a Kaggle competition in which we expect you to participate. **You should create a Kaggle submission that beats the baseline's performance one week before the final report's deadline**.

The competition aspect of the assignment is meant to incentivize and motivate you. Your grade will not be negatively affected by your final ranking, but beating the baseline is part of the assignment grade. The competition, along with all the relevant data, can be found here.

5 Assignment 2

There are 3 components of this assignment, and your assignment report should contain a discussion for each of them.

1. **(6 points)** By 5pm on March 13, 2019, make a submission to Kaggle that beats the baseline. You may select up to two submissions for the public leaderboard; at least one of these two should beat the baseline. Note that you will only see performance on the validation set. You will be evaluated on a held-out test set, so do not over-fit to the validation set.

2. **(6 points)** Describe the pipeline used for your submission and present your results. Your description should be sufficiently detailed so someone could reproduce your exact pipeline. You must justify the choices you make so that a (non-ML expert) domain practitioner could understand why you chose your specific pipeline.

3. **(6 points)** Propose concrete and meaningful modifications or extensions to your solution. How would you improve upon your method? Justify your proposal (e.g., through a careful analysis of your current results, ablation studies, preliminary experiments, etc.). If you were to add external sources of information to your pipeline, what would you add and how would it address limitations of your current pipeline?

6 Timeline

The timing for this assignment is outlined below (please read carefully):

1. February 27 - Assignment 2 is released and the Kaggle competition begins. Students can make submissions and see their performance on a validation set.

2. March 13 - Deadline to beat the proposed baseline. Students are notified via email as to whether either of their two selected submissions beat the baseline on the held-out test set. Note that the competition remains open to give students the opportunity to refine their pipelines.
3. March 20 - Deadline for the final report. Kaggle competition ends.

You may use late days for the March 13 ("beat the baseline") or March 20 (final report) deadline. This will work the same way as using late days for any other assignment deadline. Using late days on one deadline does not shift the other deadline (e.g., if a student uses one late day for beating the baseline, they are still expected to submit the report on March 20 unless they plan to use an extra late day). Please notice that the Kaggle competition will end on March 20 whether or not a student uses late days for the final report deadline.