A random forest Model for Road Speed Inference

# High-Level Description of Our Proposal

First, we formulate this problem as univariate time series prediction from our understanding of the problem definition and task requirements. But the task is not like traditional univariate time series task because we are not given the date of the prediction target and the information (observation data points) of both before and after prediction are provided, for example, the speed of early morning (0:00 am - 5:55am) and noon (11:00am - 3:55pm).

Three time series (speed from 00:00-6:00, 11:00-16:00, and 21:00-24:00) are set as features, and the direction of the speed also serves as another feature. The model outputs are time series of speed from 6:00-11:00 and 16:00-21:00. Random forest, known as ensemble model, is adopted to map the relation from what we know to the speed of time period we want to predict.

# Machine Learning Pipeline Overview

The team spent almost more than half of total time preprocessing the data. The critical steps in our **pipeline** includes:

1. extract historical speed of target road from given raw GPS files.
2. missing value imputation.
3. handle outliers.
4. training and validation datasets preparation.

The splitting of training and validation datasets in historical data is to find the best model/hyperparameters before we move on making final predictions on real datasets. Since we want our model to perform well on any given dates. Seven days of speed data are extracted individually and set as validation sets, which leaves the remaining training sets. After these manipulations, we have 7 folds of validation/training data. The seven validation sets are corresponded to and randomly sampled from each day of seven days of a week. The idea here is to develop seven models on these seven folds and get seven different performance of results, the final validation score is the average of them. So the candidate with best validation score is supposed to be the model performing the best on all seven days averagely.

1. Random forest implementation

# Discussion

The main advantages of our model are:

1. Fast to train, easy to deploy.

2. Not like traditional time series prediction models only consider the information beforehand, our model could utilize the information after the time period of predictions, which renders more accurate results.

3. the direction information is taken into account which helps model utilize speed history of opposite direction.

The limitation is that we do not search for the optimal hyperparameter of random forest due to time constraint, instead, we use default values. In the future, the model could be optimized by hyperparameter search.