Overall model building

Consider average overall models, time region models (day and period of week) and locational models in final blending

A model for each region where the regions are overlapping but every point is predicted using a unique region

A model using all the location data

A model for days of week

A model for (periods of) hours of week

- Blend of all

Build models using a relatively small subset of the data e.g. train 1-10. This is unlikely to hurt us given the amount of available data.

Check ins often happen in blobs – For all places? -> Use likelihood of surrounding checkins?

Validation strategy

Use last section of data as the validation part since this is most relevant to the predictive setting!

That way we can detect the best way to extrapolate recency of the regions.

Have a TRUE holdout sample!

General strategy – learn from the best

https://www.kaggle.com/c/facebook-recruiting-iv-human-or-bot/forums/t/14628/share-your-secret-sauce

Unbalanced data ideas

Undersampling majority class
Oversampling minority class
Neighborhood cleaning rule
Library(unbalanced)

Other

Consider problem as blocked classification rather than binary classification

What about candidate selection where mad x not >> mad y?? Not handled currently. It is hoped that the features madX/madY and relaxedMadX/relaxedMadY resolve these situations.