ETA Prediction Challenge

@DataHack

A taxi goes from Chinatown to Times Square. How long will it take to arrive?

The Team

Nir Malbin

An innovative ML expert and programmer, expert in feature engineering and selection techniques.
Kaggle Master

Gad Benram

Data Ninja, a pure professional in every data spect from gathering and exploring to modeling.
Kaggle Master

Seffi Cohen

CDS (Chief Data Scientist) for the Israeli Defense Forces, and pioneer in ML ensemble techniques.
Kaggle Master

Daniel Marcous

Data Wizard, expert in big data processing and production ready ML. Googler, Wazer and traffic analytics expert.

Taxi challenge by @Final

Data :

In this challenge, you are given data on taxi rides in New York, containing information on each ride such as the start and end points, date, time of day, distance, etc...

Goal :

Our purpose is to predict the travel time (in logarithmic scale) of a ride. The data is split to train and test sets, and we can use both general data of the ride with local data on similar rides from the train set.

Ride Information - **Given Dataset**

- From / To coordinates (lon, lat)
- Departure timestamp
- Trip distance (road distance)
- Vendor - Taxi company *(Found to be not important)*
- Passenger count *(Found to be not important)*

Machine Learning

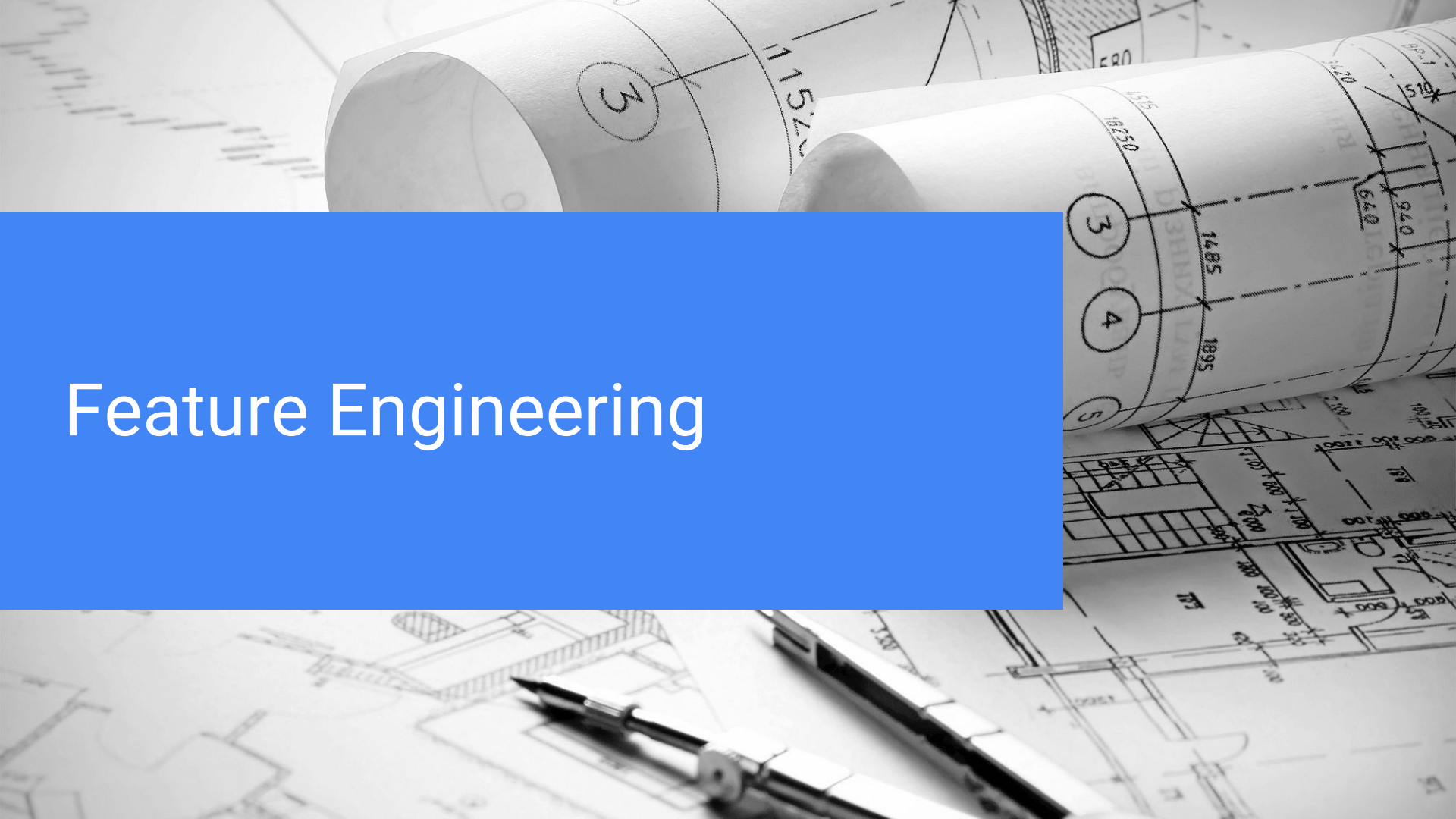
Predicting ETA Using :

1. Ride Information
2. Environment
3. Geography
4. Inferred States

Data Cleaning

- Box coordinates to NYC (remove 0.0 etc.)
- Remove very long / far rides (>2h/65km)
- Remove anomalies

Feature Engineering



Datetime based features

- Month start / end
- Day / Day of week / Hour / 15 Minute interval
- Is weekend / business day
- Is work hour (09:00-17:00)
- Is rush hour (morning / afternoon)
- Is holiday

City based features

- NYC Neighbourhood (pair crossing)
- Distance to points of interest (100X2)
  - Schools / Hospitals / Parks etc.

PCA ➔ 2

Weather based features

- Temperature
- Events - Rain / Snow etc.
- Humidity
- Wind
- Visibility
- Min / Max / Avg / std etc.

Inferred Traffic based features

- **Assumption** :
  our data is a representative sample of the NYC's - "driving population"

- Crowdedness

  - #rides in X radius

    - 100 / 500 / 1500 / 5000

    - Euclidean / Manhattan

PCA → 1

News based features

1. Crawling NYTimes
2. Topic Modeling
3. Finding topics correlated with ETA
4. Using top10 correlated topics as features
    a. Number of articles on a day for every topic

Modeling

Ensemble

- Random Forest
- Deep XGBoost
- Dropping the most important feature
- RNN – poor results

Extreme M.L.

5%

Independent thinking:

- *Adding members*
- *Feature engineering*
- *Data representations*

10%

- *Talking about the previous step*
- *Splitting to sub teams*

80%

- ***Data cleaning***
- ***Feature extraction***
- ***Feature selection***
- ***Modeling***
- ***Validation***

5%

Caveats

- Timeseries future mixing
- Crowdedness - assumes that data is a representative sample of the total car population
- Variance - taken from original validation dataset (constant)