Input:

- NYC divided into small areas, taxi demand during each time segment

Output:

- Predict demand in the “future” time for each area in NYC

Baseline?

- Previous day at the same time in the same area
- Average over ??? (same days previous years/over all days/all weekdays….) in the same area

Model?

- ARIMA
- Bayesian network
- Linear Least-Squares Regression
- Support Vector Regression
- Decision Tree Regression
- Neural networks
- …Many more

Suggestions on questions to answer about models:

- Pros & Cons
- What input/feature set they tried?
- What results did they get?
- Will it be hard/easy to implement? Is there some library for that?
- Is it suitable for our taxi problem? Why?

Evaluation?

- Just make list of used metrics, Mina is processing it on her own

TODO:

- Discuss anomaly detection with the rest
- Mention lack of Trello and Slack usage
- Google Drive - Document folder?

Review :

- Simple model : with the grid separation pre-processing, do a mean over the eight neighbours, mean over year/month/week over 1 cluster, mean over weekdays in 1 cluster, do mean just over particular daytimes/whole days
- Linear least-squares regression and decision tree reg as simple models

---

- Random forest
- Poisson model and weighted poisson model
- ARIMA as a complex model
- ANN & bayesian net

Bayesian networks

Prediction of urban human mobility using large-scale taxi traces and its applications
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.438.4632&rep=rep1&type=pdf

- Not very successful, working worse than baseline
- Input: only previous time segment's demand + index of predicted segment
- If we are going to try this, we need to choose some different features
- Results:
    - Dependent on the length of time segment (1/2/3/6/12 hrs)
    - sMAPE: 35-65 (Baseline: 10-30)
    - NMAE: 40-118 (Baseline: 20-40)

General notes:

- Wide spread method, brief googling looks like there are many libraries for that

ARIMA

Prediction of urban human mobility using large-scale taxi traces and its applications
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.438.4632&rep=rep1&type=pdf

- Input v1: Pick-up in the nearest past few time segments
- Results v1:
    - Similar to baseline, not really better
    - Dependent on the length of time segment (1/2/3/6/12 hrs)
    - sMAPE: 20-30 (Baseline: 10-30)
    - NMAE: 20-40 (Baseline: 20-40)
- Input v2: Input v1 with ARIMA + Do some **custom math** with data from the same time from the previous day
- Results v2:
    - Better than baseline
    - sMAPE: 5-15 (Baseline: 10-30)
    - NMAE: 10-20 (Baseline: 20-40)
- Math does not seem too intuitive in v2 approach, but the method appears to be doable

General notes:

- Can be used for time series analysis

Predicting Taxi-Passenger Demand using Streaming Data
http://www.inescporto.pt/~jgama/KDUSPub/Journals/16.pdf

- Features: random errors, model weights
- Result:
    - sMAPE: 24%
- General notes: “ARIMA combines the most recent samples from the series to produce a forecast and to update itself to changes in the model” (useful for real time forecast)

A Predictive Model for the Passenger Demand on a Taxi Network
http://ieeexplore.ieee.org.zorac.aub.aau.dk/stamp/stamp.jsp?arnumber=6338680&tag=1

- Features: random errors, model weights
- Result: sMAPE: 24%
- Usage is the same as in the upper paper

Linear least-squares regression

Predicting Taxi Pickups in New York City
http://www.vivekchoksi.com/papers/taxi_pickups.pdf

- Features v1: Zone, Hour of day, Day of week, Hourly rainfall
- Results v1:
    - RMSD: 138.05 (Baseline 145.78)
    - R^2: 0.7595 (Baseline: 0.7318)
- Features v2: Zone, Hour of day, Day of week, Zone * Hour of day, Zone * Day of week * Hour of day
- Results v2:
    - **Best**
    - RMSD: 40.07 (Baseline 145.78)
    - R^2: 0.9797 (Baseline: 0.7318)
- Features v3: Zone, Hour of day, Day of week, Zone * Hour of day, Zone * Day of week * Hour of day, Zone * Day of week * Hourly rainfall
- Results v1:
    - RMSD: 40.74 (Baseline 145.78)
    - R^2: 0.9791 (Baseline: 0.7318)
- “The model converged to an optimum R2 value of about 0.98 using 8000 iterations of stochastic gradient descent and parameter values n_0 = 0.2, and p = 0.4”

Epsilon support vector regression

Predicting Taxi Pickups in New York City
http://www.vivekchoksi.com/papers/taxi_pickups.pdf

- Probably long computation
- Trained only on 50 000 randomly selected training examples
- Features: Zone, Hour of day, Day of week, Zone * Hour of day, Zone * Day of week * Hour of day
- Results:
    - RMSD: 79.77 (Baseline 145.78)
    - R^2: 0.9197 (Baseline: 0.7318)
- **On large sets computationally expensive**
- “For reference, training the support vector regression using the full training set did not complete in even 8 hours of running on a Stanford Barley machine using 4 cores”
- “Our model performed best with a high C value of 1 x 107, indicating that lower values of C underfit the data and resulted in too few support vectors”

Decision Tree Regression

Predicting Taxi Pickups in New York City
http://www.vivekchoksi.com/papers/taxi_pickups.pdf

- Features: Zone, Hour of day, Day of week, Hourly rainfall
- Results:
    - RMSD: 33.47 (Baseline 145.78)
    - R^2: 0.9858 (Baseline: 0.7318)
- “Of the values we swept, our model performed best with a minimum of 2 examples per leaf and a maximum tree-depth of 100. With greater tree-depths, the model achieved the same performance on the test set, suggesting that tree-depths greater than 100 contribute to overfitting.”

Sliding Window Ensemble Framework

Predicting Taxi-Passenger Demand using Streaming Data
http://www.inescporto.pt/~jgama/KDUSPub/Journals/16.pdf

- Features: other (low level) models and their predictions for the same time
- Result:
    - sMAPE: 24%
- General notes: this model could be useful when combining other models into one; in the paper three models were combined Two Poisson models and ARIMA
- Mentions half million row dataset

A Predictive Model for the Passenger Demand on a Taxi Network
http://ieeexplore.ieee.org.zorac.aub.aau.dk/stamp/stamp.jsp?arnumber=6338680&tag=1

- Features: predictions of all the “lower models”, forecasting accuracy of all models
- Result: sMAPE: 24%
- Used as a model which combines more “low level models”

Time Varying Poisson Model

Predicting Taxi-Passenger Demand using Streaming Data
http://www.inescporto.pt/~jgama/KDUSPub/Journals/16.pdf

- Features: average rate of Poisson process over a full week, relative change for the weekday (depends on day), relative change for the period in the day (depends on day and time in the day {splitted in intervals})
- Result:
    - sMAPE:
        - ~26% for Poisson Mean
        - ~26% for Weighted Poisson Mean
- General notes: this model could be useful if we consider splitting day into time intervals and predict demand for each of those; Based on this model is Weighted TIme Varying Poisson Model which takes “seasonal demands” into consideration;
- Mentions half million row data set

A Predictive Model for the Passenger Demand on a Taxi Network
http://ieeexplore.ieee.org.zorac.aub.aau.dk/stamp/stamp.jsp?arnumber=6338680&tag=1

- Features: average rate of Poisson process over a full week, relative change for the weekday (depends on day), relative change for the period in the day (depends on day and time in the day {splitted in intervals})
- Result: sMAPE:
        - ~26% for Poisson Mean
        - ~26% for Weighted Poisson Mean

- ***Weighted Poisson Mean*** -> increase the relevance of the demand pattern observed in the last week comparing to the patterns observed several weeks ago
    - Weight set is calculated using smoothing factor and number of historical periods considered in the initial average

Linear model

$$Y = \sum_{i=1}^{n} \beta_i X_i + \beta_0 + \varepsilon$$

Modeling Taxi Trip Demand by Time of Day in New York City
http://jessie-yang.github.io/TRR%2014-4217%20Manuscript.pdf

- Features: a bunch of demographics, time of the day, time needed for accessing the "transport" (we have nothing of these)
- Results: R^2: 0.29-0.63 (varying in day times)
- We do not have the features they used… but the model is super simple, so we might try that
- I can imagine, that demand will be linearly dependable on the demand at the same time from the previous days, but no other possible features come to my mind

List of evaluation methods in papers:

Prediction of urban human mobility using large-scale taxi traces and its applications
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.438.4632&rep=rep1&type=pdf

- sMAPE
- NMAE
- Custom math-heavy method comparing **real** distance and time spent by taxis compared to distance and time spent by taxis if they had used **suggested** routes (suggestions based on predictions) -- method was performed on historical data only, so it seems usable and statistics sounds quite cool, but realization is very hard to comprehend

Predicting Taxi Pickups in New York City
http://www.vivekchoksi.com/papers/taxi_pickups.pdf

- RMSD because it favors consistency and heavily penalizes predictions with a high deviation from the true number of pickups
- R^2 value (coefficient of determination) in order to evaluate how well the models perform relative to the variance of the data set

Our ideas :

- Making a means of all the previous dates
- Making a mean of the previous week
- Taking into account the demand of the neighbors (if the data is as pixels taking the mean or doing something with the 8 neighbors of a pixel)
- Random forest
- Having several weak models and combine them.
- Neural networks?...

Unused papers:

Prediction of urban human mobility using large-scale taxi traces and its applications

- overengineered maths
- more focus on the flow, taking into consideration the road capacity...

Developing a Large-scale Taxi Dispatching System for Urban Networks

- crazy math
- aiming for taxi dispatching system
- uses data we don't have (network average speed)

Developing a Large-scale Taxi Dispatching System for Urban Networks

- Old, just reviewing other papers...