Insert
Insert Subtitle

Toomas Roosma
Student ID: 1381691
Github repo with commit

August 22, 2022

1   Introduction

In New York City, the yellow taxicabs are a widely recognised symbols of the city. With the arrival of ridesharing services as Uber and Lyft in 2011 [1] and 2014 [2] , the Taxi and Limousine Commission (TLC) has been forced to adapt to the new gig economy [XXXsome source]. With the introduction of e-hailing apps and green taxis, TLC has already made steps to adapt. This report assumes the perspective of a taxi or rideshare company with the objective of forecasting demand for pickups in various locations around NYC. [method1] and [method2] will be used to predict demand, based on data provided by the TLC from 2017 to 2019. This project aims to be also beneficial for taxi drivers, giving them insights, when to work and when not, giving better control over their work time. Yellow and green taxi data has been used, and they will be collectively referred to as taxi services.

TODO vist siia tuleb kirjutada kõike eeldused: - ainult sõidu alustamisel olnud ilm on oluline - kogu nyc ilm on sama - öelda miks ei kasuta 2020+, sest koroona ja muudab, eeldan et maailm muutub tagasi normi

- introduction 1 mark point: I used 2018data to predict x using x.

2   Preprocessing, Analysis, and Geospatial Visualisation

2.1   Datasets

This report uses mainly the taxi trip data provided by the New York City Taxi and Limousine Commission (NYCTLC) [XXXsource]. Files detailing the outlines of different taxi zones were also provided by the NYCTLC and used for visualisation purposes.

Weather data from LaGuardia Airport recorded by the United States National Center for Environmental Information's (NCEI) Integrated Global Surface dataset [3] was combined with the previously mentioned dataset. This work makes the assumption, that the weather at the moment when the decision to hail a taxi is done, has significant importance, whether or not the consumer will hail a taxi or not. It is also assumed, that the weather at the time of dropoff does not matter, because taxis are mainly used to circulate from door to door.

2.2 Preprocessing

2.2.1 Dataset range selection

We are working on the assumption that people are more pepared for bad weather during winter months, so only data from summer months (XXX which ones) was used. Since 2017, the ride sharing companies Uber and Lyft have been required to share and publish drop-off date/time and locations [4]. Both these companies have undeniably have aquired significant importance in the ride service providing scene in NYC, so data from 2017 and onwards was used. (XXXtodo visualise green, yellow, fhv usage in NYC).

On the other end, data from 2020 and onwards was not used due to the lockdown and subsequent restriction put in place by NYC authorities. This means, that summer 2019 the last year included in this study. The world has since slowly gone back to normal, but as seen on model performance (XXX panna subsection 2022 discussion) of 2022, this return has not been completed. (XXXrääkida 2022). After all the preprocessing, we were left with 271934805 unique taxi trip entries.

2.2.2 Feature selection

Although the all the yearly datasets provided by the NYCTLC are said to follow one standard, they are not consistent with each other. There are also some outliers who do not follow the given structure.

The weather data recorded at the LaGuardia airport by the NCEI [5] contains many irrelevant fields for this analysis, like monthly averages and quality control codes. There are also datafields that are too local or not applicable for taxi services at the selected period, like wave height and snow accumulations. The type of daily present weather being reported was discarded due to conveying to little differential information about each hour. Cloud coverage was repurposed from a ordinal variable to discrete, as cloud coverage of missing hours was imputed from neighboring hours. Other weather attributes were imputed in a similar fashion.(XXXtodo teha teised ka nii)

2.2.3 Eliminating outliers

(XXTODO siia võiks leida ilusasti numbrid)

Trips with invalid pickup and drop-off location ID were discarded from the dataset, as the shapefiles defined only location IDs in the range 1-263.

Trips with pickup date outside the defined range were discarded. Drop-offs outside the selected range were left in, because we are interested in predicting pickups.

Trips with negative duration were detected and discarded.

Trips with distances over 6 hours were discarded from the dataset. According to google maps, a trip between Totenville and Wakefield (extremities of area of study) would take only 2h during rush hours.

Trips with distances over 150 miles were discarded from the dataset, as the the average trip length was 3 miles with a standard deviation of 7 miles. The same trip mentioned previously is 46 miles one way.

Trips with no passenger data or zero passengers were removed.

The following data was retained from the taxi dataset:

- Month, day and hour

- Weekday

- Pickup location ID

And the following from the weather dataset:

- Wind speed

- Visibility

- Air temperature

- Dew point

- Atmospheric pressure

- Cloud coverage

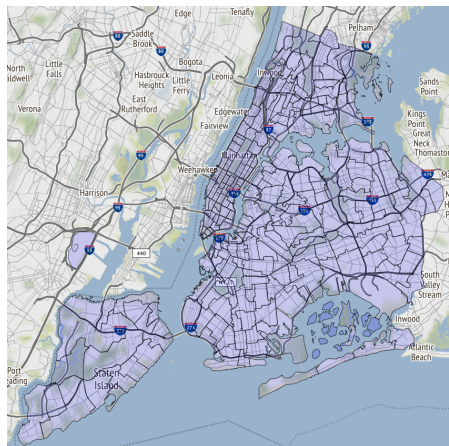[XXTODO] - vähemalt 1 good geospacial Visualisation

Example code for figures:

Figure 1: Some caption

3   Modelling

TODO: - võib olla ka random forrest, ei pea olema mathematical (linear vms) - esimene hinne: tuleb justify miks võtad. (nt linear model falsefully classifieb, ja seejärel teed penalised ja läheb halvemaks, siis teist ei kasuta) - tuleb 2 strictly erinevat, kui üks lin mod c teine, siis ei loe

- tuleb reference provideda??? okei kui see pole prereq jooksul kaetud, ft küsida - conclusion: nt airport areas are special

- categorical data peab olema onehot encoded (weather, locationIDs)

Example of a maths equation:

$$Y = X\beta + \epsilon \tag{1}$$

Example of an aligned equation (& denotes the symbol to align):

$$E[\mathbf{y}] = X\beta + E[0]$$
$$= X\beta$$

Example of an in-line equation $\epsilon \sim N(0,1)$

4   Recommendations

TODO: -ei ole summary of entire paper -

5   Conclusion

6   TODOs

- the reason snow was not included is because so few days had them. column AJ1(kuni an1) + visualise it over every year

References

[1] Lisa Eadicicco. *Uber just went public — here's what the app looked like when it first launched in New York City in 2011.* `https://www.businessinsider.com/uber-old-app-new-york-city-ipo-2019-5`. Accessed: 2022-08-21.

[2] Christine Lagorio-Chafkin. *Lyft in New York City: Let's Try This One More Time.* `https://www.inc.com/christine-lagorio/lyft-another-nyc-launch-attempt.html`. Accessed: 2022-08-21.

[3] National Centers for Environmental Information. *Integrated Surface Dataset (Global).* `https://www.ncei.noaa.gov/access/search/data-search/global-hourly`. Accessed: 2022-08-20.

[4] New York City Taxi and Limousine Commission. *TLC Trip Records User Guide.* `https://www1.nyc.gov/assets/tlc/downloads/pdf/trip_record_user_guide.pdf`. Accessed: 2022-08-20.

[5] National Centers for Environmental Information, US Air Force - 14th Weather Squadron. *FEDERAL CLIMATE COMPLEX DATA DOCUMENTATION FOR INTEGRATED SURFACE DATA (ISD).* `https://www.ncei.noaa.gov/data/global-hourly/doc/isd-format-document.pdf`. Accessed: 2022-08-20.