University of Tartu Institute of Computer Science

Estimating **Bolt** user penetration in Tartu

Team:
Anne Jääger
Hanna Maria Mägi
Milinda Tayashan

Tartu 2020

1. Introduction

Link to the repository:
https://github.com/HannaMariaMagi/BOLT-Estimating-Bolt-user-penetration-in-cities

The following report consists of three parts: Business understanding, Data understanding and Planning. The goal of this report is to plan how to estimate Bolt user penetration and its growth potential in Tartu.

2. Business understanding

2.1. Business goals

2.1.1. Background

Our project partner, Bolt, is looking for growth potential in their ride-hailing business line in Tartu. For this they need to get an overview of how big is the share of their current customers - riders - out of all residents per city district and estimate which regions have the highest potential to attract new riders.

2.1.2. Business goals

The goals of the project are to evaluate Bolt user penetration in Tartu city districts and estimate how many residents per district could be converted to new Bolt riders to increase the overall Bolt ride-hailing user penetration in Tartu.

2.1.3. Business success criteria

Business success criteria - number and share of Bolt riders per district has increased.

2.2. Situation assessment

2.2.1. Inventory of resources

We have a team of 3 students who will contribute 90 hours in total to the project. The analysis is based on user data from Bolt, demographic data from Statistics Estonia and map data from the

Tartu City government. Data will be analyzed in Jupyter Notebook, using Python (pandas, geopandas, matplotlib).

2.2.2. Requirements, assumptions, and constraints

Bolt data is protected with NDA which allows this data and results of the analysis to be used and presented only for the sake of this course. The analysis needs to be completed by December 14th and presented on December 17th.

2.2.3. Risks and contingencies

The most probable risk to jeopardize carrying out the analysis would be a team member getting infected with Covid-19 virus and not being able to contribute to completing the work on time. To prevent this situation we'll be ready to reorganize our tasks within the team if necessary.

2.2.4. Terminology

- Riders - Bolt customers who use ride-hailing service
- Rides - instances of provided ride-hailing service
- Pickup location - starting point of a ride
- Real destination - real finish point of a ride
- Home address - home address of a rider
- Work address - work address of a rider
- District - Tartu city district used for public administration
- ID - ride ID
- User ID - rider ID
- Time - date and time when a ride is created

2.2.5. Costs and benefits

No cost and benefits analysis will be conducted in this project.

2.3. Data-mining goals

2.3.1. Data-mining goals

1. Estimate Bolt user penetration in each Tartu city district:
    - create a dataset of rides which contains district data for all locations related to this ride (pickup, real destination, home address, work address);

    - develop a model to estimate the home address district for each rider currently missing this data (by pickup and real destination location and time data);
    - calculate the share of Bolt user penetration per district based on their real + estimated home address district;
2. Predict the number of people who could potentially be converted to Bolt riders per district, considering different aspects, e.g. distance from city centre, age, gender, etc.

2.3.2. Data-mining success criteria

Data-mining success criteria - model accuracy at least 90%.

3. Data understanding

3.1. Gathering data

3.1.1. Data requirements outline

The data needed to predict Bolt user penetration in Tartu would be:

- Data on Bolt rides in Tartu. When and where these took place in Tartu in the span of a month;
- Data on Tartu's population. From the same year as when the rides' data was gathered;
- Tartu city district map.

The data would need to be in the .csv and .shp (map geometry) formats.

3.1.2. Verify data availability

The data for the rides exists, since Bolt collects and saves information about every ride.

Data on Tartu's population can be found from the Tartu city's website. To be able to use this data, we will have to extract the information to a spreadsheet.

Tartu city district map is created by the Tartu City Government and we can ask permission to use it in our project.

3.1.3. Define selection criteria

Data about Bolt rides in Tartu was given to us by Bolt. We will only use data about the rides that were finished and include pickup and real destination location information.

Out of Tartu population data provided in Tartu city website we will use age and gender distribution information per district for our project (page 4).

The map containing geometry of 18 city districts of Tartu is provided by Tartu City Government and we will use it for data localization and visualization.

3.2. Describing data

3.2.1. Data given by Bolt

The data describes rides that were taken during the period of 31st March 2019 to 30th of April 2019. Some rows did not have the real_destination, home_address or work_address fields filled out.

The data that was given to us by Bolt contains 14 columns:

1. **ID *(int)*** - the id for the row.
2. **Pickup_lat *(float)***- the latitude of the pickup location.
3. **Pickup_lng *(float)*** - the longitude of the pickup location.
4. **Destination_lat *(float)*** - the latitude of the destination.
5. **Destination_lng *(float)*** - the longitude of the destination.
6. **Real_destination_lat *(float)*** - the latitude of the destination where the person was dropped off.
7. **Real_destination_lng *(float)*** - the longitude of the destination where the person was dropped off.
8. **Finished *(boolean)*** - whether or not the ride was finished.
9. **Created *(date)*** - when the ride was created.
10. **Home_address_lat *(float)*** - the latitude of the user's home address.
11. **Home_address_lng *(float)*** - the longitude of the user’s home address.
12. **Work_address_lat *(float)*** - the latitude of the user’s work address.
13. **Work_address_lng *(float)*** - the longitude of the user’s work address.
14. **New_id *(int)*** - the user’s id.

3.2.2. Tartu City Population Data description

Tartu consists of 18 districts. City data is represented in the website as the graphs for each year. 2019 data is selected for the project because it is the same year we have Bolt user data from. In order to use the data for our project we have to convert the data from graphs into a structured spreadsheet before analyzing or comparing it.

The graphs which represent the total number of population were selected. In these graphs the population is sorted according to gender - male and female, and age ranges of 0-6, 7-18, 19-64 and 65+ years.

The data that we extracted from Tartu city website includes 16 columns:

1. **Linnaosa** - District Name
2. **Pindala** (km2) - :Area
3. **Asustustihedus** (in/km2) - Population Density
4. **Mehed Naised Kokku** - Total population
5. **Naised(0-6)** - Female in age (0-6)
6. **Mehed(0-6)** - Male in age (0-6)
7. **Kokku(0-6)** - Total in age (0-6)
8. **Naised(7-18)** - Female in age (7-18)
9. **Mehed(7-18)** - Male in age (7-18)
10. **Kokku(7-18)** - Total in age (7-18)
11. **Naised(19-64)** - Female in age (19-64)
12. **Mehed(19-64)** - Male in age (19-64)
13. **Kokku(19-64)** - Total in age (19-64)
14. **Naised(65+)** - Female in age (65 +)
15. **Mehed(65+)** - Male in age (65+)
16. **Kokku(65+)** - Total in age (65+)

3.2.3. Tartu City map data description

Tartu City map data in .shp format contains 4 columns of 18 districts of Tartu:

1. **Nimi** - name of a district
2. **Shape_STAr**
3. **Shape_STLe**
4. **Geometry** - borders of a district in vector format

3.3. Exploring data

In the Bolt rides data we will remove the data where the rides were not finished. Out of the 65102 rides, that were given during the period of April 2019, 6568 of them were not finished. That is about 10% of the rides. There are 17447 unique users and 9016 of them have their home address filled out. That means that at least 51% of Bolt users in Tartu have their home address accessible for us to work with.

3.4. Verifying data quality

The data that Bolt gave is good enough to support our goals seeing as the data exists and we have access to it. However, the data that we gathered about Tartu's population is not as detailed as we would like it to be when it comes to describing the age of the population. For example, there are age gaps from 19-64 which makes us unable to separate 20 year olds from 40 year olds and so on.

4. Planning

4.1. Task list

Task	**Name**	**Time**
Planning and coordination meetings	All	10h
Collect data: Bolt rides data, Tartu city district map, Tartu city district demographic info	All	5h
Get to know Bolt data (Check how many users have a home and work address, Check how many recurrent users are there? Where are they from?, e.g.) See if we don't have to predict rows: ● How many unique users have home addresses ● Cleaning up data (remove data the we will not use)	Hanna	5h
Merge Bolt rides data and Tartu district map data to allocate district attributes to each ride by pickup, real destination, home address and work address location.	Anne	10h
Predict home address district for Bolt riders who currently miss this info by pickup destination and real destination location data and time using one of these methods: ● Decision Tree ● Random Forest ● Support Vector	Milinda	20h
Merge Bolt riders (home address district) data with Tartu demographic data to calculate and visualise Bolt user penetration per district by their real or predicted home address	Hanna	5h
Estimate the potential to convert non-users to users based on set criteria/assumptions, e.g. distance from city centre, demographics, or average predicted distance between home and work per district	All	15h
Visualization - which data to bring out to illustrate different tasks?	Anne	10h
Poster preparation	Hanna	5h
Presentation preparation	Anne	5h

4.2. Methods and tools

To achieve our goals we will be using Python with the pandas, geopandas, matplotlib and numpy extensions. For version control we will use GitHub. For data analysis we will use methods like: decision tree, random forest or support vector method, etc.