TARTU SMART BIKE

Keit Järve, Elina Meier

Repo: https://github.com/3linameier/IDS2020

Group C22

1. Business understanding

Identifying your business goals

Background

Many smart bike users (including us) have encountered the problem that at specific times Tartu bikes are in specific locations. This means you might not get a bike from the location you wish or from any nearby docks in a 5 min walk radius.

Business goals

With this project we intend to make it easier for everyday bike users to plan their commuting. Our goal is to create a model to predict whether there are available bikes in the dock at certain times.

Business success criteria

People, who use Tartu smart bikes, can plan their time better, when they know if there is a chance of getting the bike from prefered dock or not and make their bike usage easier.

Assessing your situation

Inventory of resources

We have 2 csv files of data about bike usage from the period of 06.2019 - 07.2020. Also we are in contact with Ivar Oja who is the development manager transportation in Tartu city. For this project we are going to use python and jupyter notebook.

Requirements, assumptions, and constraints

We are planning to finish the project by the 13th of december and present it on the 17th. Requirements for acceptable finished work include having a model which shows the probabilities of a bike being in a certain dock on a certain time period. Also we would like that the model is more than 80% accurate. We have the access to the data and we have verified that it is suitable and appropriate for realizing our ideas.

Risks and contingencies

Our data has some rows with undetermined start station or end station names and that could mess with the number of bikes currently in the dock because the bike isn't removed from the dock or hasn't arrived in the dock.
In addition to that there might be some Tartu smart bike side errors in the data. For example in the unlocking times. We found some unlocking times after 01.00 at night, but the latest official time to unlock a bike is before 01.00. We will probably remove those rows.
Another risk might be the lack of time, as we only have two weeks to complete the project. Also we have many other subjects to focus on. The solution to this problem is to plan ahead and divide work into smaller portions.

Terminology

Dock - place where bikes are when they are not used

Costs and benefits

There are no monetary costs or benefits. The only cost is our time. Benefits include helping people in Tartu gain better control over their time and also us getting a chance to practise our newly-learned knowledge from the subject Introduction to data science.

Defining your data-mining goals

Data-mining goals

Our goal is to make a model which predicts whether there are bikes in the dock or not and where the bikes are most frequently at certain times.

Data-mining success criteria

Our model predicts the probability whether there are bikes in the dock or not with the accuracy of at least 80%.

2. Data understanding

Gathering data

Outline data requirements

The necessary types of data include: unlocking date and time, locking date and time, start dock, end dock.
We need the data in a csv file and the prefered time range is around one year to make sure we can analyse usage during all the seasons.

Verify data availability

We knew that the data exists because in Tartu smart bike homepage they have made some statistics on their own. There was also stated that anyone who has interest in analysing the data can ask for some data. We contacted them and got the needed data to create the model. Right now we possess the data we need and it is in a usable form.

Define selection criteria

We have 2 csv files (one from period 06.2019 - 12.2019 and one from period 12.2019 - 07.2020). Both files have the same format and same columns. Columns we have in our data are:
unlockedat - (yyyy / mm / dd) The date when the bike was unlocked from a specific dock
unlockedattime - (hh / mm / ss) The time when the bike was unlocked from a specific dock
lockedat - (yyyy / mm / dd) The date when the bike was locked to a specific dock
lockedattime - (hh / mm / ss) The time when the bike was locked to a specific dock
length - (kilometers) The distance how much bike user cycled from start dock until the end dock

startstationname - The name of the station where the bike was taken
endstationname - The name of the station where the bike was returned

Describing data

In the csv file from the time period 06.2019-12.2019 we have 734166 rows and in the csv file from period 12.2019-07.2020 we have 690332 rows. Both files have the same columns which are described in the previous task. From this data we can get the information when the bike arrived at a certain dock and when it left the dock. Using this data we can make another table where every row describes one dock and how many bikes it usually has at a given time.

Exploring data

First csv file has unlock and lock dates from the time period 01.06.2019 - 14.12.2019. First bike was unlocked on 01.06.2019 at 9.37.57 and the last bike was unlocked on 14.12.2019 at 15.26.26. There are 77 different start and end station names including ‘Määramata’, ‘Undetermined’ and ‘Warehouse’.

Second csv file has unlock and lock dates from the time period 14.12.2019 - 31.07.2020. First bike was unlocked on 14.12.2019 at 15.25.43 and the last bike was unlocked on 31.07.2020 at 00.58.43. There are 77 different start station names including ‘Määramata’, ‘Undetermined’ and ‘Warehouse’ and 78 different end station names. There is one station ‘Tour d’ÖÖ’ which is only an end station and not a start station in the second csv file.

Verifying data quality

We have quite decent data to achieve our goal to build the model. The biggest problem we found is that about 11% of the data has a start station name ‘Undetermined’. But we found that most of the time we can replace it with the real

start station name because mostly the time between unlocked and locked in these cases is very short (10-15 seconds) which means there is no possibility that the bike came from another dock. Also the distance in those cases is 0.00 or close to it which confirms that we can replace the value ‘Undetermined’ with the end station name.

3. Planning your project

Tasks

1. Replace ‘Undetermined’ start station names with correct station names E: 2h
2. Clean the data and make extra replacements where needed K: 2h
3. Make sure there are no unnecessary unlocking times during the night(01.00 - 5.00) E & K: 2h
4. Make a new table with station names and bikes at the dock at certain times (about 10-15 min time periods) K & E: 8h
5. Make predictions about each station K & E: 5h
6. Analyse the differences between school and summer time K & E: 3h
7. Analyse the differences between weekdays K & E: 2h
8. Make other interesting predictions with the data K & E: 3h
9. Make a 3 minute video about the project K & E: 1h
10. Make a poster about results K & E: 3h
11. Present the poster and received results K & E: 1,5h
12. Go around and explore other projects K & E: 1,5h

TOTAL:

- Keit (K): around 32h
- Elina (E): around 32h

Methods and tools

1. Jupyter Notebook
2. Python
3. Powerpoint
4. Pandas
5. Numpy
6. Matplotlib
7. Sklearn
8. OBS Studio
9. IMovie