Homework 10

First steps of the project

Herman Klas Ratas, Karl Marti Toots

30. Nov 2019

Identifying business goals

Background

Traffic constitutes a major part of our societal routine. Traffic is governed and regulated by traffic laws for travel safety and efficiency for everyone. But with the determination of rules also come the events of breaking them. Violations and accidents in traffic are commonplace events in high population density areas like towns and cities. The violations are usually associated with bringing danger to other traffic users. It is the task of the law enforcement to catch and react to these violations in order to ensure the safety of traffic users. Tartu is the second largest city in Estonia and traffic violations occur there every day. Knowing more about where, when and in what conditions the violations occur the most could prove useful information to its citizens.

The conditions and places in which traffic violations could occur constitute to the probability of them taking place. The relevant conditions include time relations like the time of day and week and the season, location e.g. the city center or a rural area, weather at the time and so on. Information about the violations and their conditions is readily available on the internet for the curious of us to examine. When examined, some patterns associated with the conditions and events could uncover hidden truths about violations at certain locations and times. It would provide great foresight for the law enforcement and traffic users to have description, backed up with correlation in data, for relatively higher event densities related to some weather conditions, times and places.

Business goals

The main goal is to provide insight and foresight for the law enforcement and traffic users to have description, backed up with correlation in data, for relatively higher event densities related to some weather conditions, times and places.

Business success criteria

The business goal will be successful by the amount of citizens of Tartu, traffic users and law enforcement persons who will uncover new knowledge from the project results.

Assessing the situation

Inventory of resources

The list of resources is the two people involved, Karl Marti Toots and Herman Klas Ratas, and their time contribution. Another resource is the knowledge of the course and lab instructors, which could be utilized in need. In addition, previous labs resources can be put to use in the project. The software used is the python language and Jupyter notebook.

Requirements, assumptions, and constraints

The main requirement is following the schedule for completion in order to meet the deadline. Some other requirements are verifying that all required references to the data sources are included and all the tasks agreed upon in the project are fulfilled to the best of the team abilities.

Risks and contingencies

In case of an internet outage, relocation to the university institute is an option for both team members. In case of shortage of time, the business goals could be reviewed and reevaluated to assess what could be produced in the given timeframe. In case of lack of experience or knowledge, the resource of the lab instructors can be also utilized.

Terminology

In our project, there are no special domain knowledge terms at this time.

Costs and benefits

In order to direct towards working on tasks agreed upon and understanding the team mates done work more efficiently, weekly meetings have been organized to discuss and understand the results and success of the previous work and set out the tasks to complete for the next week. This project management model will in the end save time resources, ensure the involvement of the whole team and improve the results quality.

Defining the data-mining goals

Data-mining goals

The goal of this project is to assess which are the most relevant parts of the chosen data, create visualizations of traffic violations in Tartu and construct correlative descriptive models including providing validated results. One of the goals is to discover a list of places and times that correlatively have most traffic violations. In the project, some effort will also go towards accomplishing creating visually appealing graphs and heatmaps in order to make grasping of the traffic concerned information intuitive and exciting for the future user of the results.

Data-mining success criteria

The project success will be measured by the amount of uncovered hidden truths about these violations in Tartu, the number of places and times with specific weather conditions correlated to the amount of traffic violations. Other success criteria include the quality of the descriptive models and their evaluation values. In addition, the visualizations of the results should be intuitive and appealing to the reviewer.

Gathering data

As the aim of our project is to map the density of traffic violations around the city of Tartu and traffic behaviors dependence on weather, we need 2 types of data: information on traffic violations and weather. For the data to be useful, a few criteria must be met. Firstly, all data must have information of date and time so we can link traffic violations with weather at that specific moment.

Secondly, we need information about where the violation took place to be able to map the traffic violations around the city.

Data used in this project will be gathered from two public sources - 1) traffic violations are found as open data on the website of Estonian Police and Border Guard Board (referred to as Police), which is updated weekly, 2) weather data is gathered from observations made with the equipment of University of Tartu, Institute of Physics, Laboratory of Environmental Physics (data stored at meteo.physic.ut.ee, referred to as UT). Data is dated back to the year 1999 and can be called out for a predetermined period (smallest period is a day) with only desired information like temperature and precipitation.

Describing and exploring data

The open data of the Police is taken from their information system where they register the event and most important information, they can gather at that time. This means that the data only contains information available at the time of the event fixation. During this fixation, a maximum of 26 fields of information can be entered, although some fields may be unfilled. This calls for filtering of the data as only five fields (date, time, city, coordinate x, coordinate y) are necessary for our project's purposes. Of those attributes extra processing is needed for coordinates, as both the x and the y component has been given with an interval, which gives us an 500x500 m area. To represent it better, mean values of both coordinates will be taken.

UT is collecting weather data with 6 attributes, of which only 3 are necessary for this project (plus date and time): air temperature, precipitation and radiation flux. All these attributes are the easiest ones to form an understanding of the current conditions in the city - with a lower temperature, there might be ice and with higher temperature people tend to be outside and move more. When it is raining, roads are more slippery and lower radiation flux tells us that there are more clouds which means less daylight and lower visibility.

Verifying data quality

During first inspections on both datasets, the quality of the data seems very good, everything is structured and easily readable, with columns having few or no unfilled values. However, one potential problem has been found. UT's equipment gathers weather data in five second intervals

which means that every day around 17 000 rows of data is collected. As the time period observed is around 4 years, the amount of data needed to be processed is unreasonably high. To optimize the data gathering and processing it may be needed to call out data from the database for each occurring event separately, which would lower the amount of data processed with each request and as a whole.

Planning the project

Description of the task and subtasks	Time estimation (per team member)
Homework 10	4 h
Research & Retrieving data - Research datasets and finding the project idea - Acquiring a list of places to retrieve relevant data from - Research for how to implement making the required heatmaps in visualization - Find previous work done on the matter	7 h
Data cleaning - Set unimportant data aside (in a copy of the whole dataset) - Deal with invalid values / default values / other incorrect input - the coordinates need to be converted to a usable form - date and time data format, other items data format - filter the data for Tartu city related data only	5 h
Modelling & Evaluation - Find correlative insight, how weather/time/other conditions affect traffic behavior - Models - regression models for amount of traffic violations depending on conditions - regression models for other things that come to mind - Set up an efficient way to train and evaluate the models	8 h
Visualization - Producing heatmaps for the most interesting knowledge gained - Producing graphs describing evaluation and validation of the results and visualizing found relations	**8 h**
Poster + the poster session - 6 h + 3 h - Choose sections for the poster and the most important aspects of the project results - Make it visually appealing, providing the created heatmaps and graphs - Insert surprising gained knowledge for marketing	6 h + 3 h