Business understanding

Identifying business goals

Estonian Police and Border Guard Board has data of traffic violations made in the last 5 years. It contains every traffic violation that was recorded by the police from all over Estonia. Based on that data we can predict how likely a violation is made in a certain location or with a specific vehicle. The predicting model would use general details from which a person could not make an accurate prediction. The resulting machine learning model is a success if it is accurate enough to be used by the police or in insurance companies to help them make correct decisions and more accurate assessments.

Inventory of resources

Inventory of this project contains high-quality data from Estonian Police and Border Guard Board. The data has every recorded traffic violation made in Estonia in the last five years. Each entry is a violation with details about it. Details include but are not limited to: ‘vehicle type’, ‘location’, ‘driver’s age’, ‘date’. There are three novice data miners working on this project. Help from experts is available. Work is being done on basic computers using Jupyter Notebook, pandas, numpy, keras, etc.

Requirements, assumptions and constraints

By 16th December this project must be completed. Finished work is acceptable if it has a well enough predicting model. The data is public, available for everyone and free for usage. It has a Creative Commons 3.0 public copyright license.

Risks and contingencies

There aren’t any causes that could delay completion directly. Problems could arise only when we run into difficulties that come from lack of experience when training the model. Due to limited time, those can be hard to solve even with help from professionals.

Terminology

The general terminology of the dataset:

- **Vehicle type:** the general type of vehicle (truck, car, etc)
- **The country of the vehicle:** the registration country of the vehicle. In our dataset, it's either Estonia or just "other"
- **The brand of the vehicle:** the producer of the vehicle (such as "BMW", "Volvo", etc)
- **The year of the vehicle:** the production year of the vehicle.
- **Infraction:** a (traffic) crime that has been committed and recorded by the police
- **Road type:** in Estonia there's either maantee (highway), or street (tänav).
- **Infraction seriousness:** the seriousness of the infraction. Either misdemeanour (väärtegu) or felony (kuritegu).

Costs and benefits

Apart from the time spent on training the model, they are negligible. As for benefits, the model could be used by the police or insurance companies to determine risk, given the properties of a vehicle and its owner.

Data-mining goals

The goal of this project is to deliver a model that can make assumptions for a probability of a traffic violation given some details. Alongside there is a goal to make a presentation that explains this project and its results. Success criteria: a model that generalizes well. As far as we have seen, nobody has done this before based on Estonian data so we're kind of breaking new ground here.

Data understanding

Gathering data

The idea of our project came after finding the data, not that we had an idea and after that, we had to gather data. Despite the fact that there are definitely some requirements and selection criteria that the data has to follow. Firstly, the dataset has to be open-source and free to use furthermore it has to be big enough for decent analysis. Secondly, the features have to be described correctly, and errors should be able to be fixed with simple data cleaning. Lastly, the features should be relevant and give a big picture. What I mean by that is that the features should help us draw on some kind of conclusion. For example, location and time already give us the ability to calculate the percentage of an accident at a given location at the given time. Add more useful features like the violation type, and type of car, and we already have 4 features to create different combinations to predict from.

Describing data

Our dataset has a total of 26 features, that can be used.

- Properties about the location and time of the crime
    - Date of crime
    - Time of crime
    - Day of crime
    - County where the crime was committed
    - Municipality where the crime was committed
    - Name where the crime was committed(often same as a municipality)
    - Was it committed on a street or road(highway)
    - Name of street or road
    - On what KM of that road did the violation happen
    - Lest_X coordinate
    - Lest_Y coordinate

- Properties about the type of crime
    - Was it a traffic violation or something else
    - Paragraph of the violation (from the Traffic law)
    - Full description of paragraph that was violated
    - Which subParagraph was violated
    - Which point on the paragraph
    - Type of violated legal rule
- Properties about the vehicle
    - The country where the vehicle is registered at
    - Make of car (BMW, Volvo, etc)
    - Production date
- Properties about the driver
    - Age
    - Gender
    - Country of residence
    - Type of violation

As can be seen above, there are a lot of features to draw conclusions from. This also poses a problem, that there is too much data. To combat this, the features of the type of crime will be classified under common violations(DUI, no registration and so on) and only that will be used. And furthermore, depending on the goal we will use only the necessary features. From the location and date features, mainly the date and coordinates and general location might be used. From the violation type features, only the general classification will be used and from the features of vehicle and driver, most can be used for interesting analysis and predictions. Some features will be dropped, that we deem unnecessary like the id of case and the date of the violation.

Exploring data

The data comes from actual reports that officers fill out after finding a violation, so the data is quite accurate. We have a total of over 468000 rows of data, and this is data about the last five years. 1.3 million people in Estonia. Although there are some anomalies, that require data cleaning. For example, car makes can be written quite wrong - we have 11 different variations of

Mercedes-Benz and also quite often some reports have odd specifications or ways to write the make description, instead of Mercedez-Benz "Mersedes Bens" or instead of BMW, they've written the model number as 520. Furthermore, there are 75 different paragraphs represented, and some cars age have been written down as “OLD”. One fun insight: all recorded violations done with Ferraris are related to speeding.

Verifying data quality

The quality of data is mostly very good, since it has been gathered on the field, and invalid data can be disputed by the defendant. Some rows that have no values or missing useful values will be removed. Also, we would like to translate the data into English. We have checked the unique values of all fields, and anomalies mainly occur in the car names, so that we will fix in data cleaning. To make our lives simpler we will also group together the paragraphs under common labels like traffic_violation, no_registration, DUI and so on because often they are the same violation but different naming. Also, we will convert Lest_X and Lest_Y to usable coordinates by taking the middle of the given range to get the approximate location. And replace any other anomalies like car production date marked as “OLD”.

Planning the project

The tasks we need to perform:

- **Data exploration**: What kind of data are we dealing with here? - Just using jupyter notebook, pandas and matplotlib to visualize a bit
    - General data exploration: **Sander 2 h, Raul 0 h, Krister 0 h**
- **Cleaning the data**. The original dataset is very big, and very unclean. For example: Some instances have some rows missing, need to remove those, etc - just using pandas, numpy and jupyter notebooks
    - Clean the car brands. There are over 450 different car brands, some are misspellings etc. **Sander 3 h, Raul 3 h, Krister 0 h**
    - Classify the infractions. The original dataset has the infraction performed as paragraphs from the traffic law. Since this is very specific, we need to classify the 180 or so different paragraphs that appear in the dataset. **Sander 4 h, Raul 4 h, Krister 4 h**
    - Figure out the Location column: The location is encoded as a weird, Estonia specific coordinates. Need to figure out how to translate that into a more general system. **Sander 1 h, Raul 0 h, Krister 3 h**
    - Translate the data as needed: would like them to be in English. **Sander 1 h, Raul 1 h, Krister 1 h**
    - Other general cleaning tasks. **Sander 3 h, Raul 3 h, Krister 3 h**
- **Feature engineering & training the model** - using jupyter notebook, probably keras?
    - Figuring out what kind of features do we want for the risk quantifier. **Sander 3 h, Raul 3 h, Krister 3 h**
    - Figuring out which architecture do we want for the risk quantifier. **Sander 3 h, Raul 3 h, Krister 3 h**
    - Getting the risk quantifier to work. **Sander 7h, Raul 10h, Krister 10h**
- **Interpreting the results and poster design**: **Sander 3h, Raul 3h, Krister 3h**