D10 - Crime reporting

Predicting crimes using machine learning and classifiers

LTAT.02.02 course project

Rasmus Saame, 6. rühm

Toomas Roosma, 6. rühm

Table of contents

1. Setting up

Our repository is hosted in GitHub and is shared with every instructor. Repository can be found at this link: https://github.com/ToomasRo/crime_predictor.

2. Business understanding

Identifying your business goals

NB! Since we chose this project simply out of interest we don't have a big company behind us that's depending on our work. Our background analysis is made up in one possible way what the business logic and requirements for the project might be. The closest thing to a business who might be interested is the Politsei ja Piirivalveamet.

Background

Every day multiple public order breaches are committed. So that the crimes wouldn't go unpunished, the police have to counter actions. One of the options is to punish criminals for their crimes. Another option is to prevent crimes by increasing the presence of the police officers in places where breach of the public order takes place most often.

Business goals

The Police expects to reduce the crime rates in the next 10 years by trying different methods. Police would like to reduce crimes by 5% each year and by 40% for the end of 10-year period.

Business success criteria

The success will be measured by counting the times and places where public order breaches take place. Crime rates change will be measured in each location separately and in Estonia total as well. The goal has been achieved when there are some locations where crime has dropped by previously mentioned amounts.

Assessing your situation

Inventory of resources

In our team we have 2 very dedicated data scientists Rasmus and Toomas who are going to work on the project.
We have data about public order breaches in the last 10 years (starting from 2012) from this source.
For software we are going to use Python, Jupyter Notebook and different Python libraries for data mining, such as pandas, numpy and sklearn, and Google.
For hardware we are going to use our HP laptops from Institute of computer science to do all of the calculations and presentations.

Requirements, assumptions and constraints

The data is distributed with Creative Commons 3.0 license and is available for sharing and modifying if properly referenced. At the end of the project we need to have a working model predicting what type of crime is most probable in certain location and statistics about public order breaches in the last 10 years.

Risks and contingencies

Possible risks include hardware failures, if our laptops break we can't work on the project. In this case we can find substitute computers and restore our progress from GitHub.

Terminology

Public order breach (avaliku korra rikkumine) – Action against the law that take place in public room.
Petty theft (pisivargus) – Small theft where stolen value is less than 40€.

Misdemeanor (väärtegu) – Smaller offense for which the usual punishment is fine or arrest.
Crime (kuritegu) – More significant offense for which the usual punishment is fine or imprisonment.

Costs and benefits

Possible costs include data scientists' time and effort. Different nutrients and drinks for developing also require some amount of money. Possible benefits include deeper knowledge about public order breaches, statistics about breaches, model predicting breaches, lower crime rates and better course project and grade.

Defining your data-mining goals

Data-mining goals

In the end of the project we'd like to have different models which predict the likeliness of a crime on a given time and location. In addition to that we'd like to have statistics and graphs about last 10-years offenses that happened in Estonia. Map plot about the likeliness of different offenses in Estonia.

Data-mining success criteria

Predicting crimes on a test set with a greater than 0.6 AUC. Statistical knowledge about different offenses that took place. Graphs that illustrate offenses in the last 10 years on an interactive map.

3. Data understanding

Gathering data

Outline data requirements

The data we need does not need to have a lot of features. We need to know the date and time, what type of public order breach happened and the location. Ideally the location would be precise GPS coordinates. The more data the better.

Verify data availability

The Estonian open data portal (https://avaandmed.eesti.ee/) provides us with data from 2012 to now.

Define selection criteria

We have found only one source of data, so we do not have much choice from where to get out data. The publisher of the data is trusted, so we can move forward with this data.

Describing data

We obtained three CSV files (one from 2012 to 2015, one from 2016 to 2020 and one from 2020 to now). The files are actually not comma separated but tab separated.
Every entry contains the beginning and ending date, time and weekday of the offense. It also contains the type of offense, the penalty paragraph, total monetary loss (how big the theft/damage was), the type of location where it occurred (on the street, parking lot, etc.). It also contains the name of the county and municipality where the event took place. For precise location description, it provides us with L-EST coordinates. Overall, it provides us with nearly 100 000 entries, so it should be enough. After dropping rows that contain null values, we are left with around 78000 entries, if we want to keep the total monetary loss. If we were to remove this row, we would have around 10000 rows more.

Exploring data

The most common type of offense is theft, and usually (for over 87%) the total monetary damage is in the range of 0 to 499€. A third of all crime is committed in Tallinn, so we might have problem that crime correlates directly to population density.

All of the data is in string format, so we need to reformat it. Also unfortunately, the L-EST coordinates give us the location with only 500-meter accuracy.

Looking at the frequency of crime per day, we can see a slow but steady decrease in crime per day. Every Saturday there is a small peak of crime and Mondays and Thursdays are lows.

An overwhelming majority of offenses are committed in Harjumaa, and they consisted mainly of (small)thefts. We might have to pivot our main goal from predicting crime in Estonia to predicting thefts in Tallinn (or Harjumaa).

Verifying data quality

The overall quality of the data is very good; we did not find any data entry error. Some columns were scarcely populated, the additional description (SyndmusTaiendavStatLiik) and total monetary loss columns were mostly empty.

There are three occurrences of missing data for counties and parishes that have been replaced with “!Sisestamata”.

Time during which offenses were committed did not have any outlier, so we may assume that there is no “default value” that the police puts on reports, when it does not know when it happened.

The type of offense column has usually many different types appended one to another, so we will have to choose one of them, either the first one (as they are ordered by accuracy) or split the same row into multiple rows, thus growing our dataset’s size. Doing so would give us around 15% more data to predict from.

4. Project plan

Plan

1. Coming up with a business (project) idea and making the initial plan. (2x 4h)
2. Investigate data – Purpose of this task is to get familiar with the data. (2x 3h)
3. Clean data – Purpose of this task is to select only relevant fields from the data and verify that the data is usable and make possible corrections. (2x 2h)
4. Decide on different approaches – In this task the target is to talk and discuss about different approaches for this project and choose models that we are going to try to implement. (2x 3h)
5. Implement different models – In this task we are going to implement different models chosen in previous task and test with parameters to find the best models. (2x 8h, but might take longer)
6. Analyze built models – Find how well the models behaved and calculate their efficiency. This is going to be done in a continuous feedback loop with steps 4 and 5. (2x 6h)
7. Build graphs (interactive) to visualize the results – This task's purpose is to make our results presentable. (2x 4h, but there is no real upper limit)
8. Present the project – This task consists of presenting the project at project session on December 16th. (2x 2h)

Since we had no preferences on which part we want to do, we decided that both of us do equal amounts of every part.

Tools and techniques

We are planning to use Python3 and different package made for manipulating data, such as numpy, pandas and sklearn. We plan to consider using different methods learned in this course such as KNN, RF and much more for best possible predictions.