Mariete Neitsov
Joosep Suuder

Analyzing traffic accidents in Estonia

Business understanding

Background

Road accidents are a big concern in Estonia, impacting people's safety and well-being. According to Transpordiamet in 2022, there were 1699 traffic accidents with human casualties. 1919 of those people were injured and for 50 people the accident was fatal. Between 2019 and 2021, an average of 55 people died in traffic accidents, and from 2020 to 2022, the average remained the same. The average number of fatalities over the three years had previously decreased, but now it has stabilized at the same level. This indicates that the reduction in the number of fatalities has been much slower than what is necessary to achieve the goals set in the traffic safety program.

Business goals

Our project could possibly benefit Estonia’s society because our goals include identifying high-risk zones for accidents and predicting the types of accidents that could happen. This gathered information could help to enhance road safety measures to reduce the number of injuries. If the project reveals specific demographic factors associated with higher accident rates, this information could also be used to design targeted awareness campaigns.

Business success criteria

The success of our project is based on its ability to predict the types of accidents that could occur. If the predictions are accurate and high-risk zones are found then it could indicate a valuable contribution to minimizing potential safety issues in traffic in specific areas with specific conditions and creating awareness campaigns towards different groups but this would have to be assessed by authorities who work in this field.

Inventory of resources

The main resource for this project is data about traffic accidents with human injury in Estonia from 2011 to 2021 and our knowledge gained from this data science course. We are also using Python as the primary programming language, along with various specialized packages.

Requirements, assumptions, and constraints

The project's requirements include accessing traffic accident data in Estonia, cleaning the data, and selecting appropriate predictive modeling techniques. We are assuming that the data is accurate and representative of the overall traffic conditions in Estonia. The constraints include privacy, so we must ensure that the data is anonymized and that no individuals can be identified from the data.

Risks and contingencies

Our time management skills and data science knowledge could be the cause of the delay of this project, but since this project has a strict deadline we are setting up a work schedule and consistently making efforts to reach our goals.

Terminology

Traffic accident - an accident involving at least one vehicle on a road open to public traffic in which at least one person is injured or killed
High-Risk Zones - Areas where the likelihood of traffic accidents is higher than average
Predictive modeling - a statistical technique using machine learning and data mining to predict and forecast likely future outcomes with the aid of historical and existing data
Costs and benefits
Costs and benefits analysis is not relevant for this project since it does not require any financial input.

Data-mining goals

1. Identifying relevant features for the predictive model.
2. Building a predictive model to forecast the likelihood of traffic injuries or to forecast accident type based on selected features.
3. Visualizing accidents on a map and Identifying high-risk zones to explore the relationship between accident density and various geographic features such as weather conditions and road type.
4. Compare traffic injuries across different demographic factors to possibly discover patterns and groups.

Data-mining success criteria

The data-mining would be successful if the predictive model should ideally have a high accuracy, somewhere around 85%. The visualizations of accidents on a map should clearly identify high-risk zones where injuries usually happen the most. The maps should be easily readable and understandable. Finally, the project would be even more successful if we are able to find patterns and groups among different demographic features such as age, first-time driving license owner, and vehicle type.

Data understanding

Gathering data

- *Data requirements*

To visualize traffic accidents on a map and to identify high-risk zones our data would need to contain information about where the accident took place. To identify patterns and groups we would require data on demographic factors, such as license type of participants, the use of safety equipment and the age range of the participants. To explore the relationship between environmental data and the rate of accidents we would need to have data containing the time of day, lighting at the place of the accident and information about the weather. Ideally all of our data would be in numerical format, and we should try to use one-hot encoding as much as possible.

- *Verify data availability*

The data we require is all available in the Estonian Transport Authorities' gathered data containing all traffic accidents from 2011 to 2021. This should be enough to achieve all of our stated goals.

- Define selection criteria

In our dataset we can disregard the case number column and all the PPA columns as these will not be needed for our goals. We will also need to clean up the dataset as some columns contain no value. In our chosen dataset we have coordinate data provided on accidents making it possible to map out the accidents on a map of Estonia. For identifying patterns and groups, the dataset contains numerical values of participating parties' vehicle type, use of safety equipment, whether or not participants had first-time driving licenses. There is also data for environmental factors at the time of the accident, for example: lighting, consisting of two columns, first of which can be one-hot encoded to specify whether it was dark or light time at the time of the accident, second specifying whether the area was illuminated given it was dark time. There is also data on the curvature of the road, road evenness and weather.

Describing the data

The traffic accident dataset comprises 15708 traffic incidents from 2011 to 2021. Data about the vehicle type, age of the participants, license type of participants in numerical binary format. Meaning the field contains 1 if the attribute applies to that specific case, 0 if otherwise. There are numerical columns for the number of casualties, number of injured persons, number of participants, and the

number of vehicles. There are many columns containing info about the street layout of where the accident occurred, accident specifics, like accident type. Also columns for the city, county and municipality. The dataset also contains environmental data like the time of day, lighting at the area of the incident, weather, the road. The dataset we have chosen should suit all of our needs for this project.

Exploring the Data

Isikuid (People Involved), Hukkunud (Fatalities), Vigastatuid (Injuries), Sõidukeid (Vehicles Involved): These columns provide counts of people involved, fatalities, injuries, and vehicles in each incident. The data shows a range of values, with the majority of incidents involving 1-2 people and 1-2 vehicles. Fatalities are rare, which is expected in traffic accident data. Other Columns: There are columns with binary values indicating the involvement of certain types of road users or conditions These might be useful for understanding specific risk factors. Toimumisaeg (Date and Time)**:** This is a key column but is currently in an object (string) format. It needs to be converted to a datetime format for analysis. Aadress (PPA) (Address), Maja nr (PPA) (House Number), Tänav (PPA) (Street), Ristuv tänav (PPA) (Cross Street): These address-related columns have a significant number of missing values, especially 'Ristuv tänav (PPA)' with over 82% missing values. Luckily for our project most of the PPA columns can be disregarded. GPS Coordinates: GPS X and Y coordinates are present, but some values are missing.

Verifying data quality

There are a couple of problems to solve before proceeding with the data, namely:

- Standardize the Toimumisaeg column.
- Remove rows with missing values.
- Check for any anomalies in the data, such as impossible values or extremely unusual records.
- One-hot encode the lighting values so that it is easier to work with.

Granted these steps are completed the data should be suitable for further work.

Project plan

Task	Hours
Data exploration and Analysis. Tools/Methods: Matplotlib, Seaborn	about two hours each.
Data cleaning and preprocessing Tools/Methods: Pandas,NumPy	about two hours each.
Feature selection and engineering Tools/Methods:one-hot encoding, date and time engineering	about two hours each.
Select an appropriate machine learning algorithm and train the model. Tools/Methods: Scikit-learn,multiclass classification.	about five hours each.
Evaluating the performance of the trained model. Tools/methods: Train-test split, cross-validation in Scikit-learn.	about three hours each.
Visualizing accidents on a map and trying different solutions Tools/methods: GeoPandas, Folium	about two hours each.
Model Optimization. Tools/methods: Hyperparameter tuning	about three hours each.
Identifying high-risk areas Tools/methods:	about two hours each
Conducting a final report and creating a poster Tools/methods:	about three hours each.

repository: https://github.com/MarieteNeitsov/IDS-project