Predicting the outcome of driving exams in Estonia

Team members: Vanessa Apuhtin, Lisette Pihor

Identifying your business goals

Background

The Estonian Transport Administration (ETA) is the government agency responsible for regulating and supervising the transport sector in Estonia. One of its functions is to conduct practical driving tests for people who want to obtain a driving licence. The ETA publishes open data on the driving exams that took place in Estonia from 2021 to 2023, including information on the examinees, examiners, exam locations, exam categories, exam results, and exam durations.

Business goals

Business goal is to improve the quality and efficiency of the practical driving tests, as well as to enhance the safety and satisfaction of the drivers and the public. To achieve these goals, the ETA needs to understand the factors that influence the outcome of the driving exams.

Business success criteria

- The accuracy and reliability of the predictive models developed by the project team, as evaluated by cross-validation and testing on unseen data.
- The interpretability and usefulness of the insights derived from the data analysis and the predictive models.

Assessing your situation

Inventory of resources

- People: The project team consists of two students (Vanessa ja Lisette) who learn data science and have essential skills in machine learning, and statistics.
- Data: The project team will use the open dataset from the ETA, consisting of driving exams that took place in Estonia from 2021 to 2023. The dataset has 13 attributes and about 130 000 records. The dataset is available in CSV format and can be downloaded from the Estonian Open Data Portal.

- Software: The project team will use various Python libraries and frameworks, such as pandas, numpy, scikit-learn, matplotlib, seaborn, and TensorFlow.

Requirements, assumptions, and constraints

- Requirements: The project team must complete the project by 10th of December, and deliver a presentation during the poster session. The project team must also follow the ethical and legal guidelines for data privacy.
- Assumptions: The project team assumes that the data provided by the ETA is accurate, complete, and representative of the population of interest. The project team also assumes that the data is sufficient and relevant for answering the research questions.
- Constraints: The project team faces some limitations in terms of the data availability and quality, such as missing values, outliers, and imbalanced classes.

Risks and contingencies

- Risks: The project team may encounter some difficulties because of the lack of the experience in obtaining, cleaning, and integrating the data, as well as in selecting, tuning, and evaluating the predictive models. The project team may also face some uncertainties in interpreting and communicating the results.
- Contingencies: The project team will mitigate the risks by conducting a thorough data exploration.

Terminology

- Examinee: A person who takes the practical driving test to obtain a driving licence.
- Examiner: A person who conducts and evaluates the practical driving test.
- Bureau: A location where the practical driving test is related to.
- Category: A type of driving licence that the examinee will obtain after passing. (additional information: driving licence categories)
- Additional information: A code that indicates some special conditions or requirements for the practical driving test or the driving licence. (additional information: appendix nr 18)
- Last driving school: The name of the driving school that the examinee studied in before taking the test.
- Driving teacher: If the driving teacher was also in the car (J-Yes/E-No)
- Result: The outcome, either passed or failed.
- Duration: The length of time that the practical driving test lasted, in minutes.
- Reason for aborting: The cause that led to the termination of the practical driving test before completion.
- Date: Exam year and month
- Mistakes: Exam mistakes
- Unsuccessful elements: An exam is divided into individual elements; this contains unsuccessful exam elements

Costs and benefits

- Costs: The project team will incur some costs in terms of time and effort for conducting the data analysis and modeling tasks, as well as for preparing and presenting the project deliverables.
- Benefits: The project team will gain some benefits in terms of knowledge, skills, and experience for applying data mining techniques and methods to a real-world problem.

Defining your data-mining goals

Data-mining goals

- Develop different models to predict whether an examinee will pass or fail the practical driving test, based on the available attributes in the dataset.
- Compare and evaluate the performance and robustness of the different models, using appropriate metrics and methods.
- Identify and analyze the most important and influential attributes that affect the outcome of the practical driving test, using various techniques and tools.
- Provide insights and explanations for the patterns and relationships discovered in the data, using visualizations and narratives.

Data-mining success criteria

- The accuracy, precision, recall, as well as the area under the ROC curve and the confusion matrix, for the binary classification task of predicting the exam result.
- The feature importance, correlation, and causation scores of the attributes, for the exploratory task of identifying the key factors that influence the exam outcome.

Data Understanding Report

Gathering data

Outline data requirements

- The data should contain information on the driving exams that took place in Estonia from 2021 to 2023, including the exam date, location, category, result, duration, and other relevant attributes.
- The data is structured and readable in CSV files, and have consistent and meaningful column names and values.
- The data should have a sufficient number of records and variables to support the goals and methods.

Verify data availability

The project team has verified that the required data is available and accessible from the Estonian Open Data Portal, which publishes open data on the driving exams in Estonia. The project team has obtained the ability to use and share the data for the project purposes.

Define selection criteria

- The project team will use the three CSV files, one for each year from 2021 to 2023, containing information on the driving exams that took place in Estonia during that period. The project team will concatenate the three files into one dataset, with a total of about 130,000 records.
- The project team will use all the attributes from the dataset.

Describing data

- The project team has used the pandas library in Python to read the CSV files and create a data frame object, which allows easy manipulation and analysis of the data.
- The project team has used the info() and describe() methods of the data frame object to get a summary of the data, such as the number of rows and columns, the data types, the missing values, and the basic statistics of the numerical variables.
- The project team has used the value_counts() and unique() methods of the data frame object to get the frequency and the number of unique values of the categorical variables.
- The project team has used the head() and tail() methods of the data frame object to get a glimpse of the first and the last rows of the data, and to check for any anomalies or outliers.

- The data has 13 columns and 129,217 rows.
- Some of the columns have missing or suspicious values, which we are going to fix later.

- The data has one numerical variable: duration. The duration variable has a mean of 39.5 minutes and a standard deviation of 18.0 minutes.
- Other variables are categorical. The bureau variable has 17 unique values, corresponding to the locations where the exams were related to. The category variable has 13 unique values, corresponding to the types of driving licences that the examinees obtained. The additional information variable has 8 unique values, corresponding to the codes that indicate some special conditions or requirements for the exams or the licences. The last driving school variable has 340 unique values, corresponding to the names of the driving schools that the examinees studied in. The driving teacher variable has two unique values, J and E, indicating whether the driving teacher was also in the car or not. The result variable has 4 unique values, indicating the outcome of the exams. The Reason for aborting variable has 7 unique values, showing reasons for finishing the exam earlier. The Unsuccessful elements value has 2064 unique values and the mistakes value has 376 unique values. The Examinee variable has 68303 unique values, and the Examiner has 56 unique values.
- The data is quite balanced: the result variable has 70 087 passed cases and 55 298 failed cases, which means that about 54% of the exams were passed and 42% were failed. There are also 2604 cases, where the examinee did not appear (2%) and 1228 cases, where the drive was discontinued (0.9%).

Corrections in the dataset

- We are adding Year and Month columns, by extracting information from the existing date column.
- After thorough investigation, our project team has decided to drop the following categories: "KATK_POHJUS" (Reason of aborting), "MITTEARVESTATUD" (Unsuccessful elements), and "VEAD" (Mistakes). This decision is based on the recognition that these values do not contribute meaningfully to predicting the outcome of the driving exam. Rather, they primarily indicate reasons for exam failure.
- The project team also eliminates rows where the “Result” is marked as “did not appear” or “drive was discontinued”. Additionally, we removed a row where the “Duration” was NaN, but the examinee had passed the exam. These changes ensure that exams indeed took place, with a minimum duration of at least 1 minute.
- On top of that, we change three rows, where the “Additional information” is marked as “101|78|78” to “101|78”, because the original values seem suspicious. The same goes with some duration values, which are abnormally large (such as 1024 min).
- Also, the project team has transformed the categorical values of “Result” and “Driving teacher” to binary format.
- Furthermore, we are changing the “Examinees” column to “Previous attempts”, where the value is the number of previous attempts. This makes sure that we don’t have any privacy concerns in our dataset.

Exploring data

- The project team has used the seaborn and matplotlib libraries in Python to create various visualizations of the data (like box and bar plots).

- “VARASEMAID KATSEID” (PREVIOUS ATTEMPTS): It is more common to pass the exam on the first or second attempt, after that it is more common to fail the test.
- “BYROO” (BUREAU): The result of the driving exam differs among different cities. Some cities have an outstanding passing rate (for example Tartu).
- “KATEGOORIA” (CATEGORY): The examinees who are doing a B-category exam (most common) have the lowest chance of passing the driving exam.
- “ERITINGIMUSED” (ADDITIONAL INFORMATION): The most common additional information is 101 (Primary driving licenses) and 101|78 (Primary automatic driving licenses).
- “VIIMANE AUTOKOOL” (LAST DRIVING SCHOOL): It didn’t give much information about the outcome. Some schools have better success rates than others but nothing unusual.
- “SÕIDUÕPETAJA KAASAS” (DRIVING TEACHER): Examinees without the driving teacher are more likely to pass the exam, but with the teacher the change is basically 50-50.
- “EKSAMINEERIJA” (EXAMINER): it's apparent that certain examiners demonstrate a significant variability in pass and fail rates.
- “SEISUND” (RESULT):
- “KESTUS” (DURATION): When the exam is shorter than the required time (below 35 or 45 min) then the outcome is negative. So the longer the exam the higher chance of passing.
- “AASTA” (YEAR): Did not influence the results.
- “KUU” (MONTH): The variable shows some seasonal patterns, with more exams taking place in the summer months than in the winter months. We also notice a slightly higher pass rate from April to September compared to the winter months.

Planning the project

DATE AND TIME	TEAM MEMBER	TASK	METHODS AND TOOLS	COMMENTS
28.11.23 / 3h	Vanessa	Business understanding text	Writing	Vanessa will write a text explaining the business problem, the objectives, and the expected outcomes of the project.
28.11.23 / 3h	Lisette	First look at the data	value_counts(), info(), describe(), unique(), head()	Lisette will use these methods to get a general overview of the data.
30.11.23 / 7h	Vanessa	Data understanding report & project plan	Writing	Vanessa will write a report summarizing the findings from the explorations in the data, highlighting any issues or insights.
30.11.23 / 7h	Lisette	Correcting dataset, deeper exploration	dropna(), groupby(), cumcount(), seaborn, matplotlib	Lisette will clean the data by handling missing values, outliers, and errors. She will also perform a deeper exploration of the data and visualization techniques.
05.12.23 / 8h	Lisette & Vanessa	Testing different models	decision trees, random forests, gradient boosting machines, One-hot encoding	Lisette and Vanessa will test different models such as decision tree, random forest. They will use scikit-learn to build and evaluate the models, pandas and numpy to manipulate the data, and train_test_split(), cross_val_score(), to split the data, perform cross-validation, and tune the hyperparameters.
06.12.23 / 7h	Lisette & Vanessa	Finding the best predicting models	Compering	Lisette and Vanessa will compare the performance of the models. We will select the best model based on the results.

07.12.23 / 5h	Lisette & Vanessa	Analyzing the outcome and making the poster	Writing	Lisette and Vanessa will analyze the outcome of the best model and explain how it answers the business problem. They will also make a poster to present their project, using PowerPoint or Canva to create a visually appealing design.