Project Guidebook

Block 1D-2023-2024

AI Traffic Accident Prevention Breda

Authors:

Ionut-Alberto Boţoroga (230036)
Márk Molnár (232926)
Mirthe van der Horst (231763)

Date:

June 21, 2024

I. Introduction

Welcome to the guidebook for our project, focused on predicting the number of road accidents in Breda. This project aims to enhance road safety by leveraging advanced data analytics and machine learning techniques. Our goal is to provide actionable insights that can help reduce accidents and improve traffic management in the city.

II. Methodology

Data Collection

We utilized two primary datasets provided by ANWB and SWOV:

- **ANWB Safe Driving Dataset**: This dataset includes driving behavior data such as:
    - `road_name`: Name of the streets.
    - `event_start`: Date and time of the incidents.
    - `maxwaarde`: Maximum speed value or the g-force value for the incidents.
    - `category`: Type of incidents (e.g., speeding, harsh cornering, accelerating, braking).
    - `incident_severity`: Severity of the incidents.
- **SWOV Accident Dataset**: This dataset contains data related to road accidents, including:
    - Date of the accident.
    - Name of the area where the accident occurred.
    - Type of road involved.

Since these datasets contain sensitive information, we have adhered to strict data protection protocols to ensure confidentiality and data security.

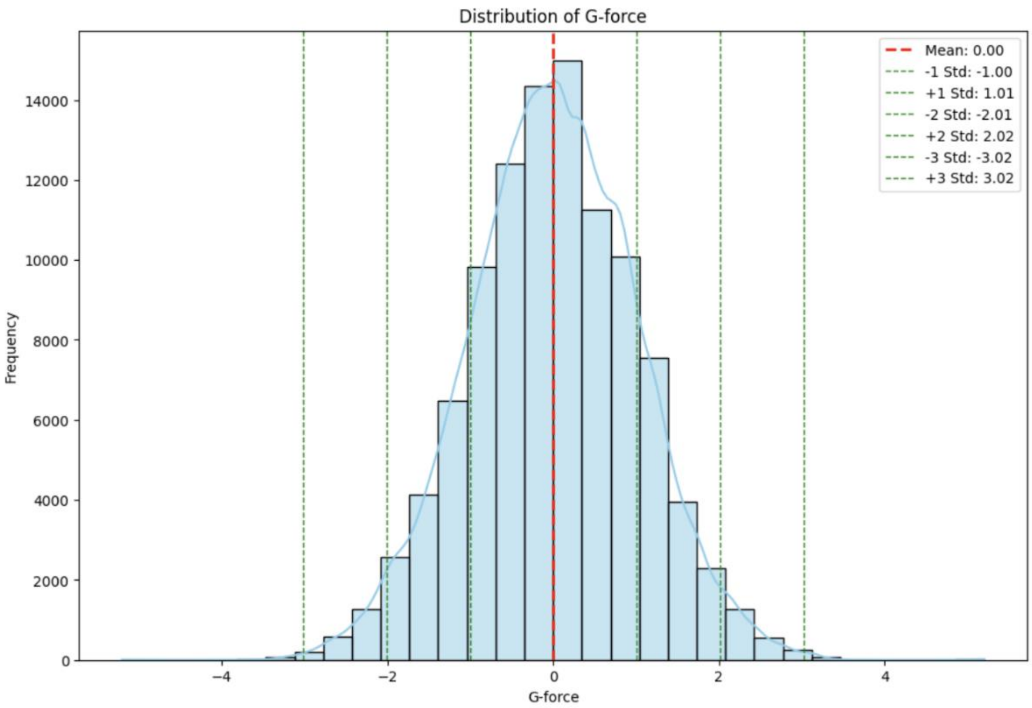
Data Pre-processing

We undertook several steps to prepare the data for analysis:

- **Cleaning**: Checked for zero or NaN values and removed any outliers.
- **Normalization**: Applied normalization to scale the data appropriately.
- **Logarithmic Transformation**: Identified that our data followed a logarithmic pattern, so we applied a logarithmic transformation to improve model performance.

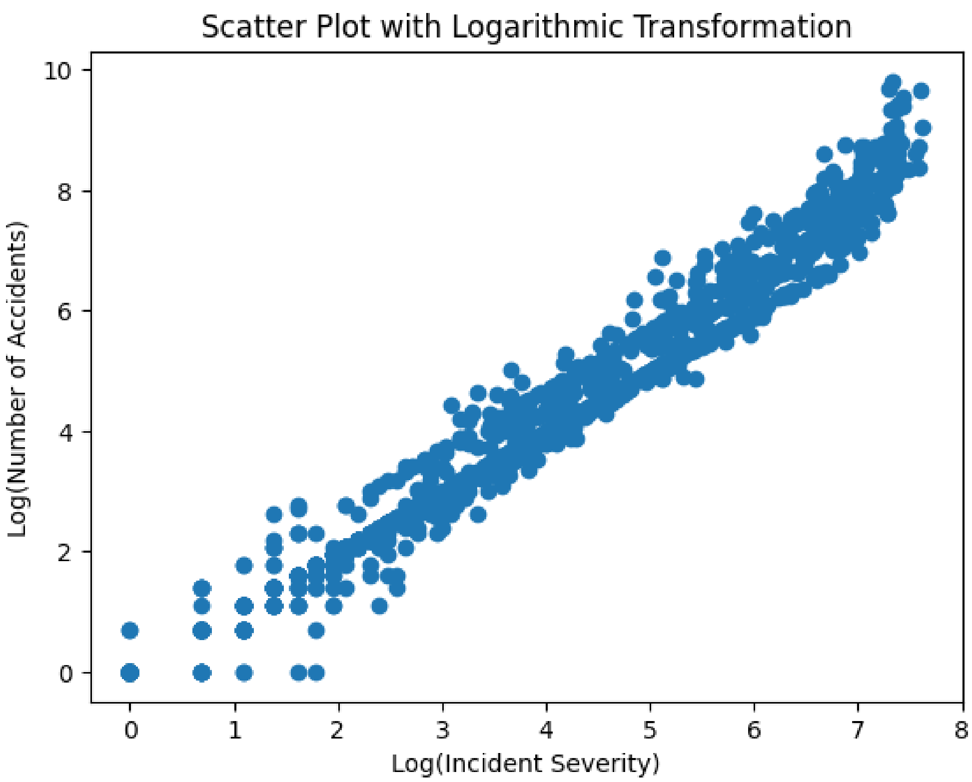
Detailed Pre-processing Steps:

1. **Missing Values**: Confirmed there were no missing values in the dataset.
2. **Outliers**: Used z-scores and Quantile Transformation to identify and remove outliers, keeping values within the -3 to 3 range.



1. Figure: Distribution of the normalized G-force values after Quantile Transformation.

3. **Transformation**: Applied logarithmic transformation after confirming the pattern through exploratory data analysis.



2. Figure: Scatter plot for the incident severity and the total number of accidents after applying Logarithmic Transformation

Model Selection

We explored several machine learning and deep learning techniques to identify the best model for our predictions:

- **Linear Regression**
- **XGBoost**
- **Random Forest Regression**
- **Neural Networks**: Both Deep Neural Networks (DNN) and Recurrent Neural Networks (RNN)

Training the Data

The data was split into training and test sets, with 80% for training and 20% for testing. We performed multiple iterations and hyperparameter tuning to optimize the models:

- **TensorFlow**: Used for building and training deep learning models.
- **Scikit-learn**: Used for machine learning algorithms.

Evaluation

We developed an evaluation function to assess the models using the following metrics:

- **Mean Absolute Error (MAE)**: Measures the average error magnitude.
- **Root Mean Squared Error (RMSE)**: Highlights larger errors.
- **R-squared (R^2)**: Indicates the proportion of variance explained by the model.

Together, these metrics provide a comprehensive evaluation of model accuracy and reliability.

Unit Testing

To ensure the final model operates as intended, we have added unit tests to most of its functions. We have considered several scenarios for seven different functions, almost entirely eliminating the possibility of the model performing differently than expected.

Deployment

For the stakeholders (the Municipality of Breda and ANWB), we developed an application using Streamlit. This application integrates the best-performing model and provides a user-friendly interface to display the risk factors for each street in Breda. Users can interact with the application to gain insights into high-risk areas and make informed decisions to enhance road safety. The application can also do forecasting for the next 7 days based on the previous 30 days of data.

III. User Manual for our Application

Welcome to the Risky Roads Application. This manual provides detailed instructions on how to use the application to assess road safety. The application uses historical, current, and forecasted data to predict risk factors for various streets.

What is the Risk Factor?

The Risk Factor is a calculated metric that helps predict the likelihood of traffic accidents on a particular street in Breda based on historical driving data. This model considers several key indicators of driving behavior and incident severity to estimate accident risk.

The features used in this calculation include:

- **Harsh Braking:** The number of instances where vehicles have braked sharply.
- **Harsh Cornering:** The number of instances where vehicles have taken sharp turns.
- **Harsh Accelerating:** The number of instances where vehicles have accelerated rapidly.
- **Speeding Incidents:** The number of instances where vehicles have exceeded the speed limit.
- **Average G-force:** The average gravitational force experienced during these incidents.
- **Incident Severity:** The average severity of these incidents.

Using these features, our model predicts the number of accidents with the following formula:

Number of Accidents = 0.0097 * Number Cornering + 0.0038 * Number Braking + 0.8612 * Incident Severity - 0.0319

How to Interpret the Weights

Each weight indicates the contribution of its corresponding feature to the risk of accidents. For example, a higher weight for Harsh Braking means that an increase in harsh braking incidents is strongly associated with a higher number of accidents.

Why is This Important?

Understanding the Risk Factor helps city planners, traffic authorities, and drivers identify high-risk areas and take proactive measures to improve road safety. By addressing the underlying causes highlighted by the model, we can work towards reducing the frequency and severity of accidents.

Installation

Ensure you have the required libraries installed:

```
pip install psycopg2 pandas joblib numpy scikit-learn streamlit plotly
```

And that you have the streamlit file and the weights of the Linear Regression model.

Application Features

1. Predict Current Risk

This feature provides an immediate assessment of the current risk levels of streets. It helps you understand immediate safety concerns and take necessary precautions.

1. **Predict for now**: Click the "Predict risk" button to get the current risk predictions.

2. **View Results**: A table with the current risk factor for each street will be displayed.

3. **Street Selection**: Select a street from the dropdown to view its specific risk factor and rank.

4. **Incident Map**: If incidents are found for the selected street, a map showing the locations of these incidents will be displayed.

2. Predict Risk for the Past 30 Days

This feature analyzes the risk levels of streets over the past month. It helps identify patterns and trends in road safety.

1. **Predict for past 30 days**: Click the "Predict 30 days in the past" button to get predictions for the past 30 days.
2. **View Results**: A table with the past risk factor for each street will be displayed.
3. **Street Selection**: Select a street from the dropdown to view its risk factor development over the past 30 days.

3. Predict Risk for the Next 7 Days

This feature forecasts the risk levels for the upcoming week using advanced predictive models and forecast data.

1. **Predict for next 7 days**: Click the "Predict next 7 days" button to get predictions for the next 7 days.
2. **View Results**: A table with the future risk factor for each street will be displayed.
3. **Street Selection**: Select a street from the dropdown to view its predicted risk factor over the next 7 days.

4. Concatenate Plots

This feature allows you to visualize the risk levels comprehensively by selecting a street after obtaining predictions for the past 30 days, present, and next 7 days.

1. **Generate Predictions**: Ensure you have generated predictions for the past 30 days, present, and next 7 days.
2. **Street Selection**: Select a street from the dropdown to concatenate plots.
3. **Concatenate Plots**: Click the "Concatenate Plots" button to view the combined risk factor predictions over time.

Behind the Scenes

Data Fetching

Data is fetched from a PostgreSQL database using `psycopg2`. Ensure the database credentials are correctly configured.

Normalization

Features are normalized using `MinMaxScaler` to ensure they are on a similar scale before feeding them into the model.

Predictions

The application uses a pre-trained linear regression model loaded using `joblib`. Predictions are made based on normalized features.

Visualization

Plots are generated using Plotly and displayed in the Streamlit interface for easy interpretation.

Functions

- `fetch_incident_data()`: Fetches incident data from the database.
- `normalize_features(df, features)`: Normalizes the specified features in the DataFrame.
- `predict_past()`: Predicts risk factors for the past 30 days.
- `predict_new_server_data()`: Predicts current risk factors based on the latest data.
- `predict_next_7_days()`: Predicts risk factors for the next 7 days using forecast data.

IV. Conclusion

Summary

Our project on predicting road accidents in Breda has shown significant promise in enhancing road safety through the application of Machine Learning and AI techniques. By leveraging detailed driving behavior data and historical accident records, we have created a predictive model that identifies high-risk roads and provides actionable insights for reducing accidents.

Key Findings

- **Accurate Predictions**: Our model has demonstrated strong predictive accuracy, with evaluation metrics indicating its reliability and effectiveness.
- **High-Risk Roads**: The insights gained have highlighted specific roads and conditions that are more prone to accidents, enabling targeted interventions.
- **User-Friendly Application**: The deployment of our model via a Streamlit application offers stakeholders an intuitive tool to access and utilize the predictions.
- **Quality Assurance**: Comprehensive unit testing was implemented throughout the development process to validate the functionality and reliability of our models and application components.

Implications

The successful implementation of this project can lead to several positive outcomes:

- **Improved Road Safety**: By identifying high-risk zones, the Municipality of Breda can prioritize road safety measures more effectively.
- **Informed Decision-Making**: Policymakers can use the data-driven insights to develop more informed strategies for traffic management and accident prevention.

Future Work

While our project has achieved its initial goals, there are several avenues for future research and improvement:

- **Data Expansion**: Incorporating additional data sources, such as real-time traffic updates and weather conditions, can further enhance the model's accuracy.
- **Model Refinement**: Continuous refinement and retraining of the model with new data will help maintain its relevance and accuracy over time.
- **Enhanced Predictive Models**: Our Recurrent Neural Network (RNN) model, which predicts the number of incidents for the next 7 days, has shown promising results. Further fine-tuning and validation of this model can improve short-term predictions and provide timely alerts for potential high-risk periods. Also, if the Safe Driving dataset from ANWB is updated every day, then we can retrain the RNN model weekly on new data.

- **Broader Deployment**: Expanding the deployment of the predictive tool to other municipalities can amplify its impact and contribute to nationwide road safety improvements.

Final Thoughts

The integration of AI and data science into traffic management represents a significant step forward in enhancing road safety. By continuing to innovate and collaborate, we can pave the way for safer roads and a better driving experience for everyone in Breda and beyond.