Project Documentation

Project Title:
Forecasting Road Accident Numbers in Saudi Arabia Using AI-Based Time-Series Models

Author: Fatimah Albaik
Email: AlbaikFatimah@gmail.com

---

1. Introduction

Traffic accidents are a leading cause of death and injury in Saudi Arabia, imposing significant human and economic costs. Predicting future traffic deaths and injuries allows policymakers to implement targeted interventions to improve road safety. This project applies AI-based time-series models to forecast monthly traffic deaths and injuries using historical government-reported data.

---

2. Problem Statement

Despite extensive reporting of traffic accident statistics in Saudi Arabia, there is currently no AI-based predictive system to forecast future accident outcomes. Developing such models can help authorities take proactive, data-driven decisions to reduce accident rates and their consequences.

---

3. Project Goal

To develop AI models that accurately forecast the monthly number of traffic deaths and injuries in Saudi Arabia using historical accident data, enabling safer roads through predictive insights.

---

4. Objectives

- Predict future monthly traffic deaths and injuries.
- Identify key factors influencing accident severity and trends.
- Evaluate and compare the performance of modern time-series regression models for this task.

---

5. Dataset Description

5.1 Source

The dataset was obtained from the **Saudi Open Data Platform** and includes the following files:

- Injured and Dead in Accidents 1437 H
- Injured and Dead in Accidents 1438 H
- Injured and Dead in Accidents 1439 H

5.2 Data Collection

Each file contained two separate sheets:

- **Deaths data sheet** with counts of deaths by region, month, gender, and age groups.
- **Injuries data sheet** with similar counts for injuries.

5.3 Data Integration and Size

- All files were combined into a single structured dataframe for modeling.
- **Total dataset size:** 1152 rows and 20 columns.

5.4 Key Features

- **Temporal features:** gregorian_year, gregorian_month, gregorian_day (engineered for time-series indexing).
- **Region:** region_number (region_name was dropped and encoded numerically).
- **Demographics:** male_count, female_count, age groups (under_18, 18_30, 30_40, 40_50, over_50), saudi_national_count, non_saudi_national_count.
- **Targets:** total_deaths, total_injuries.
- **Engineered Lag Features:** deaths_lag1, injuries_lag1 to capture previous month's trends for forecasting.

---

6. Data Preprocessing

6.1 Steps Taken

1. **Combining files:** Data from deaths and injuries sheets were rearranged and combined to form a unified dataframe.
2. **New Columns Created:**
   - total_deaths and total_injuries were added to track outcomes clearly across combined datasets.
   - gregorian_year was added since each original file represented a separate Hijri year.
3. **Column drops and replacements:**
   - Dropped region_name and replaced it with numeric region_number for modeling.
   - Dropped hijri_month_name and replaced it with numeric month_number for standardized temporal analysis.
4. **Date conversion:** Hijri dates were converted to Gregorian dates using Islamic calendar conversion to enable standard time-based indexing.
5. **Lag Features:** Created deaths_lag1 and injuries_lag1 representing previous month's deaths and injuries to capture autocorrelation and temporal dependencies crucial in time-series forecasting.

6.2 Data Integrity Checks

- Checked for null values and confirmed dataset completeness.
- Ensured numeric columns were correctly typed for modeling.

---

7. Exploratory Data Analysis (EDA)

7.1 Purpose

EDA was conducted to understand data distributions, trends, and potential feature relationships before model development.

7.2 Analyses Conducted

- **Dataset overview:** Verified structure, column types, and value ranges.
- **Summary statistics:** Reviewed mean, median, standard deviation, min, and max for all numeric features.

- **Correlation heatmaps:** Generated before and after feature engineering to identify linear relationships.
- **Temporal trends:**
  - Plotted yearly and monthly trends for deaths and injuries to detect seasonality and long-term changes.
- **Regional analysis:** Compared deaths and injuries across different regions to identify geographic disparities.
- **Demographic analysis:** Visualized age group distributions, gender counts, and nationality counts to understand accident victim profiles.
- **Animated visualizations:** Created region-wise animated plots to show changes in deaths and injuries over time dynamically.

7.3 Key EDA Findings

- Lag features showed strong correlations with targets, validating their inclusion.
- Region_number indicated clear differences in accident counts across regions.
- Monthly seasonality patterns were visible, with certain months showing higher accident rates.

---

8. Modeling Approach

8.1 Models Evaluated

1. **CatBoost Regressor**
   - Suitable for structured tabular data.
   - Handles categorical identifiers effectively without explicit encoding.
2. **LightGBM Regressor**
   - Integrated with time-series cross-validation for robust performance evaluation.
   - Demonstrated superior results in this project.

8.2 Strategy

- Separate models were trained for predicting total_deaths and total_injuries to specialize outcomes.
- TimeSeriesSplit was used to maintain temporal integrity and prevent data leakage.
- Lag features (previous month's deaths and injuries) were included to capture temporal dependencies.
- Hyperparameter tuning was applied to LightGBM to optimize performance.

---

9. Results

9.1 CatBoost

- **Deaths Prediction:**
  $R^2 = 0.76$ (with injuries_lag1), dropped to $R^2 = 0.38$ when injuries_lag1 was removed.
- **Injuries Prediction:**
  $R^2 = 0.77$ (with deaths_lag1), dropped to $R^2 = 0.47$ when deaths_lag1 was removed.

9.2 LightGBM

- **Deaths Prediction:**
  $R^2 = 0.95$ | MAE = 3.31 | RMSE = 6.32

- **Injuries Prediction:**
  $R^2 = 0.96$ | MAE = 12.09 | RMSE = 23.83

---

10. Model Phase Analysis

- **Lag features** were crucial, as previous month's deaths or injuries significantly improved accuracy by capturing temporal trends.
- **Region_number** consistently showed importance, revealing regional differences in accident outcomes.
- **LightGBM outperformed CatBoost**, delivering high accuracy and low errors in both targets.

---

11. Conclusion

AI-based time-series models, especially LightGBM, accurately forecasted monthly traffic deaths and injuries in Saudi Arabia. These models can support data-driven policies and proactive safety measures to reduce accidents and their consequences.

---

12. Future Work

- Integrate external data sources such as traffic volume, weather conditions, and enforcement activity for richer predictive models.
- Explore deep learning time-series models like Temporal Fusion Transformer for advanced multi-step forecasts.
- Deploy models into interactive dashboards for real-time use by policymakers.

---

13. Tools and Libraries

- **Programming Language:** Python
- **Data Processing:** pandas, numpy
- **Modeling:** scikit-learn, CatBoost, LightGBM
- **Visualization:** matplotlib, seaborn, plotly
- **Time-Series Deep Learning (tested):** PyTorch Forecasting
- **GPT-**4