Fatality Prediction of Road accidents Kenya


MDF
Innovators
• Mitch Mathiu
• Michael Omondi
• Deborah Okeyo
• Faith Wanjala

PROJECT OVERVIEW

Kenya experiences a high rate of traffic accidents, contributing to significant fatalities and injuries each year. Despite efforts to improve road safety, challenges persist due to factors such as poor infrastructure, reckless driving, and inadequate enforcement of traffic regulations. Understanding the dynamics behind these accidents is crucial for developing effective interventions.

BUSINESS

UNDERSTANDNING

Road traffic accidents are a significant public safety concern in Kenya, contributing to high fatality rates. Effective measures to reduce fatalities require identifying the factors that increase the likelihood of death in accidents. The goal of this project is to build a machine learning model that predicts the probability of fatal outcomes in road crashes using historical crash data. Insights from this model will help transportation agencies, public safety departments, and urban planners develop data-driven interventions to reduce fatalities.

Stakeholders for the project could include:

- Transportation Agencies eg SuperMetro sacco
- Private car owners
- Pedestrians, cyclists and motorists

Urban plan planners

DATA UNDERSTANDING

The project utilized crash data collected from Kenyan road accidents from 2012 to 2023,sourced from World Bank microdata platform.The data contains various features including:

- Crash date-The date the accident occured
- crash time-The time of the accident
- location-given in Longitudes and Latitudes
- Crash description keywords-keywords describing the nature of the accident-’pedestrian’,’motorcycle’,or ’fatality’

DATA PREPARATION

Data cleaning

- handle missing values
- handle duplicates
- Correct Data Types
- Remove Outliers
- Standardize Categotical values

Feature Engineering

- Data time feature engineering
- Seasonal environmental features etc

Training data=80%,tetsing data=20%

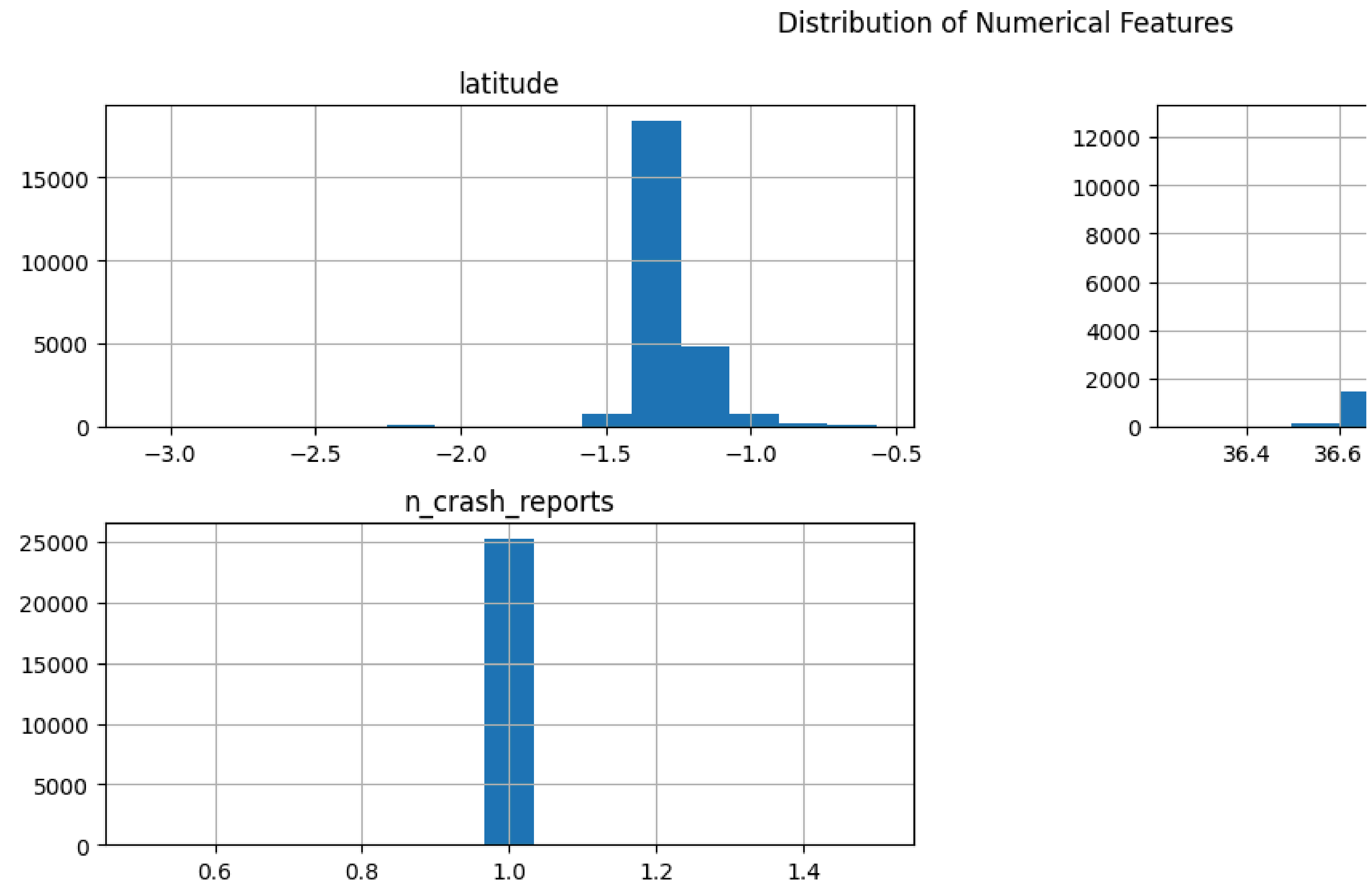
Data Splitting

Modeling

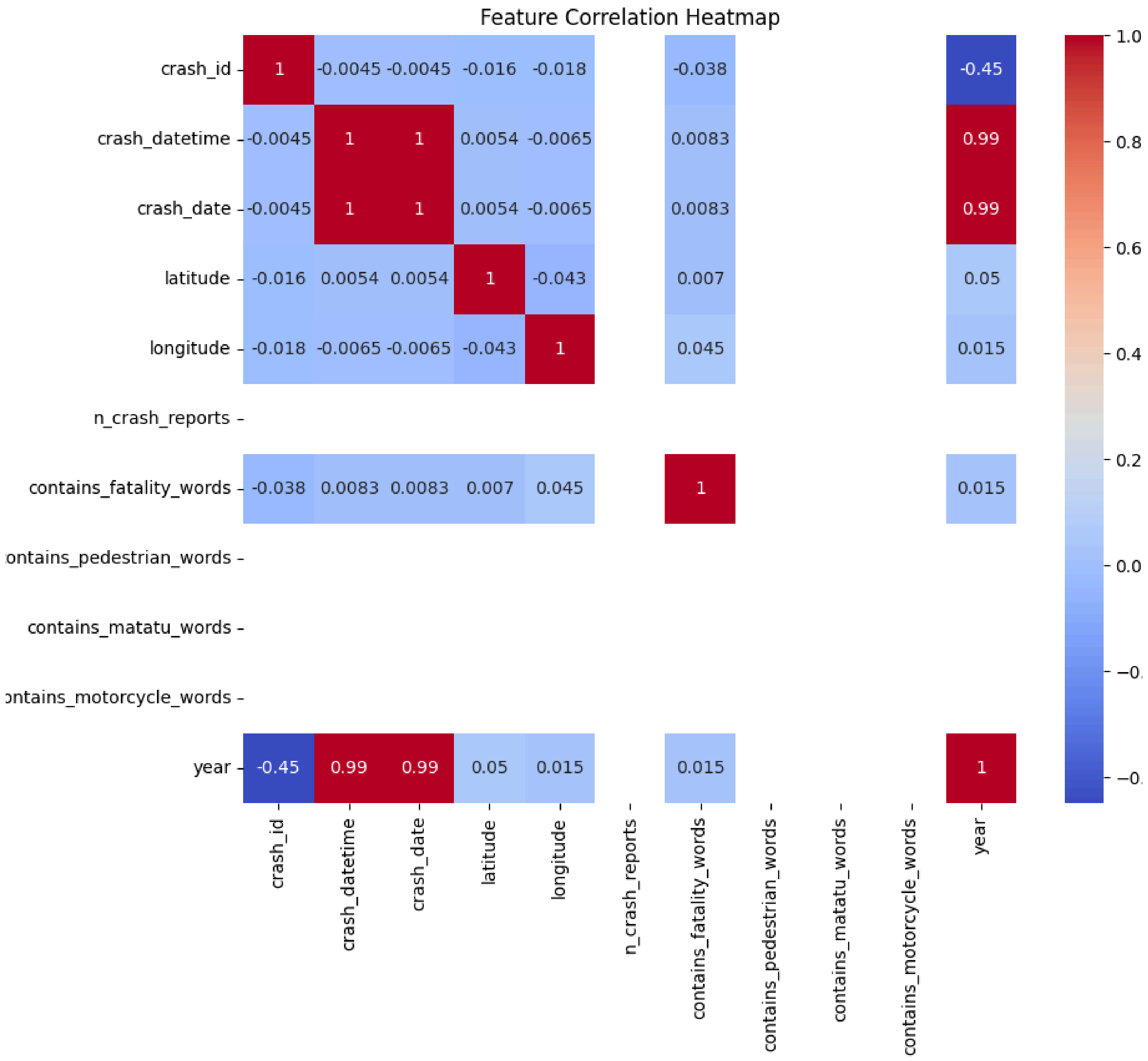
- Baseline model
- Logistic Regression
- Random Fores
- Support Vector Machine

ANALYSIS

Univariate Analysis



Bivariate Analysis



MODELING

1.Data Preprocessing
Feature/Target Split: Separate features (X) and target (y).
Handle Missing Data: Drop rows with missing values.
Datetime Handling: Convert datetime columns to Unix timestamps.
Train-Test Split: 80% for training, 20% for testing.
2.Feature Engineering
One-Hot Encoding: Convert categorical features to binary format.
Standardization: Scale numerical features to standardize them.
3.Class Imbalance Handling (SMOTE)
Use SMOTE to oversample the minority class and balance the dataset.
4. Multicollinearity Check
VIF Calculation: Drop features with high multicollinearity (VIF > 10).
5.Baseline Model
DummyClassifier (Most Frequent Strategy): Baseline accuracy = 50%.
6.Model Selection & Evaluation
Train multiple models (Logistic Regression, Random Forest, Gradient Boosting, SVM).
.Cross-Validation Scores:
Logistic Regression: 1.00
Random Forest: 1.00
Gradient Boosting: 1.00
Support Vector Machine: 0.53
7.Hyperparameter Tuning
GridSearchCV: Best parameters for Random Forest: n_estimators=50, max_depth=None, min_samples_split=2.
Best CV Score: 1.00
8.Model Performance Evaluation
Final Model (Random Forest):
Accuracy: 1.00
Precision: 1.00
Recall: 1.00
F1 Score: 1.00
ROC-AUC: 1.00
9.Ensemble Method (Optional)
Voting Classifier: Combined models achieve 1.00 test accuracy.

EVALUATION

The DummyClassifier provides a baseline accuracy of 94%. This high baseline likely reflects class imbalance in your data (e.g., more instances of "No Fatality" than "Fatality").

Model Selection and Cross-Validation:

We evaluated using four models: Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machine (SVM). Each achieved a cross-validation (CV) score of 100%, suggesting strong model performance during training, potentially due to the dataset or preprocessing steps.

Hyperparameter Tuning:

1. GridSearchCV optimized the Random Forest model with n_estimators=50, max_depth=None, and min_samples_split=2, achieving a best CV score of 100%. This model was then used as the primary model for further evaluation.

Model Evaluation:

2. The Random Forest model reached 100% on accuracy, precision, recall, F1 score, and ROC-AUC on the test set, indicating it accurately identifies both classes without errors. Additional metrics (e.g., Confusion Matrix, Classification Report, and ROC Curve) reaffirm this perfect performance.

Ensemble Model:

3. The Voting Classifier (ensemble of Logistic Regression, Random Forest, and Gradient Boosting) also achieved a 100% test score, suggesting robust generalization.

SUMMARY OF FINDINGS
Baseline Model:

Given these results, the Random Forest model (after hyperparameter tuning) is likely the best choice, as it combines high predictive power with interpretability through feature importance analysis. However, achieving 100% across all metrics suggests that further checks for overfitting might be worthwhile, especially if you plan to generalize this model beyond the dataset used here.

- Data Insights: The analysis revealed patterns in crash data, such as high-risk locations, peak times, and vehicle types involved in fatal crashes.
- Model Performance: The model demonstrated strong predictive accuracy, particularly in identifying high-risk scenarios, although certain conditions may require more data for improved prediction.

Recommendations

- Enhanced Feature Engineering: Incorporating additional variables such as vehicle type or traffic density at crash sites.
- Hyperparameter Tuning: Further tuning of model parameters to improve predictive accuracy.
- Real-Time Data Integration: Potential for integrating real-time data sources, like live traffic or weather updates, to improve prediction accuracy.

Real-World Context

Kenya has seen multiple tragic road incidents underscoring the need for enhanced road safety. Notable accident sites include:

- Kiambu Road: Frequent accidents, especially near Thindigua.
- Nairobi Expressway: Notable for high-speed incidents.
- Londiani, Thika, and Voi Road Crashes: Known areas with frequent fatalities, often due to high traffic and poor infrastructure.