VISHAL CHAUDHARY · DATA ANALYST · DUBLIN, IRELAND

Geospatial Analytics & Predictive Modelling · Personal Project

Road Collision Severity Analysis

XGBoost and Random Forest on 104,258 UK collision records to predict accident severity and map high-risk zones — personal project.

104,258	67%	#1 Predictor	50+
Collision Records	XGBoost Accuracy	Speed Limit	Risk Hotspots Mapped

TOOLS & TECHNOLOGIES

Python	XGBoost	Random Forest	Geopandas	Folium	SMOTE
Scikit-learn	Matplotlib				

Email vishal.ch1401@gmail.com	LinkedIn linkedin.com/in/vishal111	GitHub github.com/chaudhary521	Location Dublin, Ireland

PROBLEM STATEMENT

Road safety investment decisions are often made on political rather than data-driven grounds. This personal project asked which environmental and situational factors most strongly predict collision severity — and where on the UK road network do the highest-severity collisions cluster? The dual goals were a predictive model for triage and a geospatial hotspot map for infrastructure prioritisation.

DATASET

UK Department for Transport STATS19 road collision dataset: 104,258 collision records with 33 features per record. Variables include road type, posted speed limit, weather conditions, light conditions, junction detail, vehicle type, driver age band, and severity classification (slight / serious / fatal). Fatal collisions represented fewer than 2% of records — creating a significant class imbalance challenge.

APPROACH & METHODOLOGY

Applied the full CRISP-DM pipeline. Data cleaning addressed missing values in weather and road condition fields through mode imputation. SMOTE oversampling was applied specifically to fatal and serious collision classes to address extreme imbalance. XGBoost and Random Forest were benchmarked; XGBoost was selected as primary model. SHAP values generated feature importance rankings. Geospatial hotspot analysis used Geopandas for spatial joins and Folium for interactive map output.

KEY TECHNICAL HIGHLIGHTS

- XGBoost achieved 67% accuracy on three-class severity prediction (slight / serious / fatal).
- SHAP analysis confirmed speed limit as the single strongest predictor of collision severity.
- SMOTE applied to address extreme class imbalance — fatal collisions represented under 2% of data.
- Geospatial hotspot analysis mapped the 50 highest-risk road segments by collision density and severity score.
- Junction type, road surface condition, and light conditions ranked as secondary severity predictors.
- XGBoost outperformed Random Forest by 8 percentage points on minority class (fatal) recall specifically.

KEY INSIGHTS & RESULTS

Speed limits above 60 mph were associated with 4.2x higher fatal collision probability. Rural single-carriageways showed the highest severity index despite lower absolute traffic volume. Night-time collisions at unlit junctions had a 2.8x higher serious-injury rate. The hotspot map revealed that a disproportionate number of fatal collisions clustered on fewer than 12% of total road segments analysed.

BUSINESS IMPACT

The severity model gives transport authorities a data-driven tool for ranking road safety interventions by expected impact rather than political visibility. The geospatial hotspot maps directly support targeted speed limit

reviews, junction redesign, and lighting improvement programmes. The CRISP-DM pipeline is documented for annual retraining with updated STATS19 data releases.
