UNIVERSITAT DE BARCELONA

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S THESIS

Forecasting Urban Traffic Patterns in London Using Hybrid AI Techniques

Author:
Theodoros LAMBROU

Supervisor:
Dr. Jordi VITRIA

A thesis submitted in partial fulfillment of the requirements for the degree of MSc in Fundamental Principles of Data Science

in the

Facultat de Matemàtiques i Informàtica

June 30, 2025

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc in Fundamental Principles of Data Science

Forecasting Urban Traffic Patterns in London Using Hybrid AI Techniques

by Theodoros LAMBROU

Accurately forecasting traffic incident severity is crucial for urban mobility planning and real-time traffic management. This thesis explores a hybrid approach to classifying traffic severity levels using statistical and machine learning techniques. The dataset includes road segment-level hourly traffic observations in London, enriched with engineered features such as recent severity history, weather conditions, and baseline severity probabilities.

We evaluate a range of models, from simple baselines to advanced classifiers, with a focus on Random Forest and XGBoost. After extensive experimentation, a tuned Random Forest model using balanced subsampling and moderate tree depth outperformed all other approaches in terms of macro-averaged F1-score and minority class recall. Detailed evaluation through time-based cross-validation, SHAP analysis, and visual diagnostics demonstrates the robustness of this model and highlights key predictive factors.

The findings suggest that combining short-term temporal features with baseline statistical probabilities significantly improves performance, particularly for underrepresented severity classes. The report also discusses limitations related to data coverage, class imbalance, and the potential of incorporating external signals such as incidents or public transport disruptions in future work.

The corresponding python notebooks, scripts and data for this thesis are located in this GitHub repository: https://github.com/theol-10/datascience-thesis/.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Jordi Vitrià, for their valuable guidance, constructive feedback throughout the development of this thesis. Their expertise and mentorship were instrumental in shaping the outcome of this work.

I am also grateful to the Universitat de Barcelona for providing me with the opportunity to work on a project of this nature, which allowed me to apply the knowledge gained as part of this Master's, and further develop my skills in data science.

Lastly, I would like to thank my family for their constant encouragement, patience, and support throughout my academic journey.

Contents

Abstract ii

Acknowledgements iii

1 Introduction 1

2 Background 3
2.1 Traditional Approaches to Traffic Forecasting . . . 3
2.2 Machine Learning Models . . . 3
2.3 Deep Learning and Spatiotemporal Models . . . 3
2.4 Explainability in Traffic Models . . . 4
2.5 Class Imbalance in Traffic Datasets . . . 4
2.6 Related Work Summary . . . 4
2.7 Gaps in the Literature . . . 5
2.7.1 Limited Use of Flexible, Real-Time Data Sources . . . 5
2.7.2 Insufficient Integration of Explainability into Performance Evaluation . . . 5
2.7.3 Partial or Incomplete Addressing of Class Imbalance . . . 6
2.7.4 Lack of Hybrid Models Integrating Statistical Priors with Machine Learning . . . 6
2.7.5 Inadequate Consideration of Model Fairness in Traffic Forecasting . . . 6
2.7.6 Scalability and Computational Complexity of Advanced Models 6
2.7.7 Lack of Temporal and Spatial Cross-validation Techniques . . . 7

3 Data and Preprocessing 8
3.1 Overview . . . 8
3.2 Traffic Data from TfL . . . 8
3.3 Weather Data from Open-Meteo . . . 8
3.4 Feature Engineering . . . 9
3.5 Label Definition and Class Distribution . . . 9
3.6 Train-Test Splitting . . . 10

4 Methodology and Modeling Approaches 11
4.1 Problem Formulation . . . 11
4.2 Baseline Models . . . 11
4.2.1 Categorical Feature Baseline (Notebook 3) . . . 11
4.2.2 GNN Experiment (Notebook 4) . . . 11
4.2.3 Probability-Based Baseline (Notebook 5) . . . 12
4.3 Feature Engineering . . . 12
4.4 Modeling Strategy and Experiments . . . 12
4.4.1 Initial ML Models (Notebooks 7a and 7b) . . . 12
4.4.2 Random Forest Experiments (Notebook 8) . . . 12

4.4.3 XGBoost Benchmarking (Notebook 9) . . . 14
4.4.4 Handling Class Imbalance . . . 14
4.5 Model Selection Criteria . . . 15

5 Results and Evaluation **16**
5.1 Evaluation Metrics . . . 16
5.2 Model Comparison . . . 16
5.3 Random Forest 8e vs XGBoost 9e . . . 17
5.4 Evaluation of Best Model (RF 8e) . . . 17
5.4.1 Feature Relationships and Redundancy . . . 17
5.4.2 Learning Curve Analysis . . . 18
5.4.3 Class-wise Precision-Recall Analysis . . . 18
5.4.4 Confusion Matrix Analysis . . . 19
5.4.5 Model Interpretability (SHAP) . . . 20
5.5 Conclusion . . . 23

6 Discussion **24**
6.1 Key Findings . . . 24
6.2 Interpretability and SHAP Insights . . . 24
6.3 Handling Imbalance and Model Robustness . . . 24
6.4 Limitations . . . 25
6.5 Ethical and Operational Considerations . . . 25
6.6 Summary . . . 25

7 Conclusion and Future Work **26**
7.1 Conclusion . . . 26
7.2 Future Work . . . 26

Bibliography **28**

Chapter 1

Introduction

In modern urban environments, accurate traffic forecasting plays a critical role in enabling smarter mobility, reducing congestion, and improving overall urban planning. The ability to anticipate traffic severity levels in real-time has implications for both short-term interventions, such as traffic signal optimization and dynamic routing, and long-term infrastructure development, like road network planning and expansion. Effective traffic forecasting can lead to enhanced mobility, lower carbon emissions, and improved public safety by providing actionable insights for both authorities and commuters.

Historically, traffic prediction has relied on traditional methods, such as traffic flow modeling and time-series analysis. However, these methods often struggle to account for complex and dynamic factors like weather patterns, special events, and human behavior. With the rise of data availability, particularly from sensors, GPS data, and social media feeds, along with advances in machine learning and deep learning, there is increasing potential to significantly improve prediction accuracy. In particular, the ability to leverage large datasets containing both historical patterns and real-time contextual information offers new avenues for more accurate and responsive traffic forecasting systems.

This thesis focuses on the task of forecasting traffic severity levels in London, using a multi-class classification framework. The objective is to develop a predictive model that integrates recent historical severity trends with engineered features derived from time, weather, and contextual indicators. By integrating diverse data sources, the model aims to predict traffic severity with greater accuracy, addressing the challenge of class imbalance that often hampers the detection of minor and serious delays. A key motivation is to assess the effectiveness of combining statistical baselines with machine learning models to capture temporal patterns, enhance model performance, and improve minority class recall, which is critical for effective traffic management.

The project also explores the interpretability of machine learning models, an essential consideration in real-world applications where decision-makers need clear insights into model behavior. A significant component of this work involves the use of SHAP (SHapley Additive exPlanations) analysis, which allows for a deeper understanding of feature importance and the factors driving predictions. By offering transparency in the model's decision-making process, SHAP analysis helps validate the model's relevance to real-world traffic prediction scenarios.

Furthermore, the thesis evaluates the performance of several modeling strategies, including baseline probability models, tree-based classifiers such as Random Forest and XGBoost, and a range of data augmentation and hyperparameter tuning techniques. By comparing these approaches, the thesis aims to identify the most effective combination of statistical and machine learning techniques for traffic severity forecasting.

The remainder of this report is organized as follows:

- **Section 2** provides a comprehensive review of the background and existing literature on traffic prediction systems, highlighting the strengths and limitations of traditional and modern approaches.
- **Section 3** describes the data sources and preprocessing steps used to prepare the dataset for model development, including feature engineering techniques employed to enhance prediction accuracy.
- **Section 4** outlines the various modeling approaches used, detailing the integration of baseline probability models with advanced machine learning techniques, as well as the specific feature engineering strategies applied.
- **Section 5** discusses the evaluation methods employed to assess model performance, presenting the results of comparative model testing and highlighting key performance metrics.
- **Section 6** offers a discussion of the key findings, potential limitations of the approach, and directions for future work to address challenges such as real-time prediction, inclusion of external data sources, and model scalability.
- **Section 7** concludes the thesis, summarizing the key contributions of the work and its implications for urban traffic forecasting.

This thesis used ChatGPT to improve language clarity, check for grammar mistakes and rephrase preliminary drafts of sections. All data analysis, coding, interpretation and argumentation were conducted by the author.

Chapter 2

Background

Urban traffic forecasting is a longstanding challenge, intersecting areas such as transport planning, intelligent systems, and machine learning. This chapter provides an overview of relevant literature and positions this thesis in relation to prior work.

2.1 Traditional Approaches to Traffic Forecasting

Early approaches to traffic prediction primarily relied on time series models, including Autoregressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), and exponential smoothing methods. These statistical models offer interpretable frameworks and perform reasonably well for short-term, linear trends (Smith and Demetsky, 2002; Gharaibeh, 2020). However, they struggle with complex temporal dependencies, spatial heterogeneity, and nonlinear interactions commonly present in urban traffic.

2.2 Machine Learning Models

Machine learning (ML) techniques have become increasingly prominent due to their flexibility and ability to capture non-linear patterns. Models such as Support Vector Machines (SVM), Decision Trees, Random Forests, and Gradient Boosting (e.g., XGBoost, LightGBM) have demonstrated improved performance over classical methods, particularly in handling tabular data with rich feature representations (Vlahogianni, Karlaftis, and Golias, 2014; Li, Rose, and Sarvi, 2015). These models are data-efficient, fast to train, and offer a reasonable trade-off between accuracy and interpretability.

Random Forests have been widely used for traffic classification and congestion detection due to their robustness to noise and ability to model feature interactions. Boosting methods like XGBoost further improve predictive performance by sequentially correcting residual errors. However, these models generally ignore spatiotemporal structures unless such information is explicitly engineered through features.

2.3 Deep Learning and Spatiotemporal Models

Deep learning models, especially Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, are capable of modeling complex temporal dependencies. Their use in traffic prediction has grown substantially (Ma et al., 2015; Zhang, Zheng, and Qi, 2017). Convolutional Neural Networks (CNNs) and hybrid CNN-LSTM architectures have also been employed, especially when traffic data is framed as spatiotemporal grids or images.

More recent work leverages Graph Neural Networks (GNNs) to model road networks as graphs, where each node represents a road segment and edges represent spatial relationships. Models such as Diffusion Convolutional Recurrent Neural Networks (DCRNN) and Graph WaveNet (Li et al., 2018; Wu et al., 2019) can learn both temporal and spatial dependencies. However, these methods are computationally expensive, require large datasets, and are often difficult to interpret — making them less suitable in resource-constrained or explainability-critical settings.

2.4 Explainability in Traffic Models

As traffic models influence real-time decisions and urban policy, explainability is crucial. Recent studies have emphasized the use of SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) to explain black-box predictions (Lundberg and Lee, 2017). SHAP provides consistent, global and local feature attributions and is particularly useful for tree-based models like Random Forest and XGBoost.

In traffic severity forecasting, SHAP can identify key drivers of congestion — such as recent traffic patterns, weather conditions, or historical severity probabilities — and guide stakeholders in understanding why specific alerts are triggered. This is especially important in public-facing or high-stakes deployments.

2.5 Class Imbalance in Traffic Datasets

A recurring challenge in traffic classification tasks is class imbalance — where severe congestion or rare events (e.g., class 2) are underrepresented in the data. Standard accuracy metrics may be misleading in such settings. Techniques like resampling (under/oversampling), class weighting, and tailored evaluation metrics such as macro F1 score or minority-class recall are widely recommended (He and Garcia, 2009). Moreover, temporal stratification in cross-validation can help preserve the sequence structure and mitigate data leakage.

2.6 Related Work Summary

Table 2.1 summarizes a selection of recent works in traffic forecasting, highlighting their methods and limitations.

Study	Model Type	Features	Limitations
Smith and Demetsky (2002)	ARIMA	Historical traffic volumes	Poor handling of nonlinearity
Li, Rose, and Sarvi (2015)	Random Forest	Time, location	No spatiotemporal modeling
Ma et al. (2015)	LSTM	Time series of flow	High data requirement, low interpretability
Li et al. (2018)	DCRNN (GNN + RNN)	Spatiotemporal traffic sensors	Requires graph structure and compute power
Wu et al. (2019)	Graph WaveNet	Graph-based time series	Hard to interpret, needs dense data
Lundberg and Lee (2017)	SHAP (explainer)	Post-hoc explanation	Limited to model-agnostic analysis

TABLE 2.1: Summary of representative literature on traffic forecasting.

2.7 Gaps in the Literature

Despite the significant advancements in traffic forecasting using machine learning and deep learning techniques, several gaps remain in the literature that could hinder the effectiveness and generalizability of existing models. These gaps also highlight the areas where this thesis makes unique contributions to the field.

2.7.1 Limited Use of Flexible, Real-Time Data Sources

A considerable portion of existing research relies on fixed sensor data, such as loop detectors or fixed-location cameras, which restricts spatial generalization and often leads to biases in model predictions. These data sources are typically sparse in coverage, making them inadequate for capturing the dynamic nature of urban traffic patterns. Few studies have utilized flexible, real-time data APIs, such as the Transport for London (TfL) API, which provide live, large-scale datasets that allow for more comprehensive and accurate modeling. This thesis aims to leverage the flexibility and richness of open data sources like the TfL API, addressing this limitation and enabling the development of more robust traffic forecasting models.

2.7.2 Insufficient Integration of Explainability into Performance Evaluation

Explainability has become an increasingly important aspect of machine learning models, especially in safety-critical domains such as urban traffic forecasting. While post-hoc explanation techniques like SHAP (SHapley Additive exPlanations) have been used in some studies to interpret model predictions, explainability is often not integrated into the performance evaluation process. Many models, particularly complex ones like deep learning models and graph neural networks (GNNs), are treated as "black-box" systems without considering how their decision-making process can be understood by urban planners and other stakeholders. This thesis addresses this gap by incorporating SHAP analysis into the evaluation of the proposed hybrid model, providing insights into the key drivers behind traffic severity predictions, and enhancing the model's interpretability.

2.7.3 Partial or Incomplete Addressing of Class Imbalance

Class imbalance is a well-known challenge in traffic forecasting, particularly in the context of rare events like serious accidents or severe traffic congestion (class 2). Despite being recognized as a significant issue, many studies either overlook the impact of class imbalance or only partially address it through resampling or class weighting. While some methods focus on improving accuracy for the majority class, they fail to capture the important rare events that are crucial for urban planning and management. This thesis takes a more comprehensive approach by integrating stratified time-based cross-validation and focusing on metrics such as macro F1-score and minority-class recall to ensure that the model performs well across all classes, including the underrepresented severity levels.

2.7.4 Lack of Hybrid Models Integrating Statistical Priors with Machine Learning

Another important gap in the literature is the limited exploration of hybrid models that integrate traditional statistical approaches, such as baseline severity probabilities, with machine learning techniques. Many existing models either rely entirely on historical time series data or machine learning models without considering the benefits of combining both approaches. Baseline probability models, which capture long-term temporal trends, can complement machine learning models that focus on recent traffic conditions. This hybrid approach has the potential to improve model accuracy, especially for rare events, by combining the strengths of both paradigms. This thesis investigates this hybrid modeling approach, integrating statistical priors with machine learning models to enhance predictive performance.

2.7.5 Inadequate Consideration of Model Fairness in Traffic Forecasting

As machine learning systems are increasingly deployed in real-world urban settings, there is growing concern about fairness and the potential for these systems to reinforce existing inequalities. This is particularly true for traffic forecasting, where biases in the model could disproportionately affect underserved communities or geographic areas. For example, areas with lower traffic sensor coverage may be underrepresented in model predictions, leading to less accurate traffic predictions for those areas. Few studies explicitly address fairness in traffic forecasting, focusing primarily on accuracy and performance metrics. This thesis aims to address this gap by considering fairness in model design and evaluating performance across different road segments and times of day, particularly those that serve underserved populations.

2.7.6 Scalability and Computational Complexity of Advanced Models

Many of the latest advancements in traffic forecasting, such as deep learning models (e.g., LSTM, CNNs) and spatiotemporal models (e.g., GNNs), offer impressive predictive capabilities. However, these models often come with significant computational overhead, requiring large datasets and substantial processing power. In urban environments where real-time predictions are needed, such models may be impractical due to their high computational demands. This thesis addresses this challenge by leveraging machine learning models like Random Forest and XGBoost,

which offer a good balance between performance, interpretability, and computational efficiency, while also exploring hybrid approaches that combine traditional and machine learning methods.

2.7.7 Lack of Temporal and Spatial Cross-validation Techniques

Most traffic forecasting models evaluate performance using standard cross-validation techniques, which may not be appropriate for time-series data with temporal dependencies. Applying traditional cross-validation methods without considering temporal sequences can lead to data leakage and unrealistic performance estimates. Few studies implement temporal or spatial stratification in cross-validation, which is crucial for ensuring that the training and validation sets respect the temporal and spatial dependencies inherent in traffic data. This thesis takes a step forward by incorporating temporal stratification in the model evaluation process, ensuring that the models are evaluated in a way that better reflects real-world scenarios.

In summary, while the field of traffic forecasting has made considerable progress, there remain several gaps in the literature related to data sources, model explainability, class imbalance, hybrid modeling approaches, fairness, and scalability. This thesis aims to address these gaps by proposing a novel hybrid machine learning approach that integrates baseline probability models with engineered features and weather data. The model's performance is evaluated using robust metrics, ensuring that minority class recall and fairness are prioritized, and its interpretability is enhanced through SHAP analysis.

This thesis addresses these gaps by proposing a hybrid ML approach for severity classification on a flexible open dataset (TfL API), integrating baseline severity probabilities, engineered features, and weather data. It evaluates model performance using stratified time-based validation, interprets predictions using SHAP, and prioritizes fairness and minority-class performance.

Chapter 3

Data and Preprocessing

3.1 Overview

This chapter outlines the data acquisition and preprocessing steps carried out to support traffic severity prediction modeling. The project integrates two primary sources of information: real-time traffic status data from Transport for London (TfL), and weather data from Open-Meteo. Data collection, cleaning, feature engineering, and train-test splitting are discussed in detail.

3.2 Traffic Data from TfL

The primary dataset was obtained from London's official open traffic API—Transport for London's (TfL) Unified API.[1] Data was collected using a custom Python script named `get_road_status_all.py`, which was scheduled to run at regular intervals via a CRON job. This script queried the TfL Unified API and stored road status records locally.

The collection process spanned from **10 March 2025 to 20 May 2025**, resulting in a dataset with over 6.6 million records. Each record included several key fields:

- `timestamp`: Date and time the status was recorded.
- `roadName`: Name of the road segment.
- `statusSeverity`: Numerical severity score, 0–2.
- `statusSeverityDescription`: Text description of traffic conditions.

After cleaning, the dataset included traffic status information for unique road segments. Data exploration (Notebook 1) included analysis of class balance, road-level activity, and missing values.

3.3 Weather Data from Open-Meteo

To enrich the feature set, weather data was fetched from the Open-Meteo API.[2] Hourly weather variables were collected for the Greater London area for the same date range as the traffic data. The key weather attributes included:

- Precipitation
- Rain

[1] https://tfl.gov.uk/info-for/open-data-users/unified-api?intcmp=29422
[2] https://open-meteo.com/

- Snowfall
- Cloud cover percentage
- Temperature
- Wind speed

The weather data was later joined with the traffic dataset based on timestamp alignment during the feature enhancement phase (Notebook 6).

3.4 Feature Engineering

Initial feature engineering (Notebook 2) involved transforming raw fields and constructing new predictive attributes. In Notebook 5, historical congestion probabilities were calculated, and in Notebook 6 recent traffic congestion were calculated. This process generated features across several categories:

- **Temporal Features:** Hour of day, day of week, weekend flag, rush hour indicator.
- **Lag Features:** Previous traffic severity for the same road segment in the past 1-2 hours.
- **Weather Features:** Precipitation, cloud cover, snow, wind gusts and temperature.
- **Historical Severity Probabilities:** For each road and time bin, the historical probability of each severity class was computed (Notebook 5), and later merged into the modeling dataset (Notebook 6).

All categorical features were encoded as integers or one-hot vectors depending on model requirements. Numerical variables were left in raw or normalized form.

3.5 Label Definition and Class Distribution

The traffic severity label was derived from the `statusSeverity` field and mapped into a 3-class classification task:

- **Class 0:** Normal traffic (severity = 0)
- **Class 1:** Mild congestion (severity = 1)
- **Class 2:** Severe congestion (severity = 2)

This mapping simplifies the problem while maintaining alignment with domain understanding of traffic disruption levels. Class imbalance was present, with normal traffic comprising the majority of records.

3.6 Train-Test Splitting

To train and evaluate the machine learning models, the dataset was split into training and testing subsets using stratified random sampling. This approach ensures that all three congestion classes are proportionally represented in both sets, preserving the class distribution and preventing class imbalance from skewing model performance.

Specifically:

- In early experiments (Notebooks 7a and 7b), the data was split using an 80/20 ratio, with 80% used for training and 20% for testing.
- In later experiments involving more complex models (Notebooks 8 and 9), a 70/30 train-test split was adopted to better evaluate generalization performance.

This random split strategy was implemented using the `train_test_split()` function from the `scikit-learn` library, with a fixed `random_state` for reproducibility. Unlike time-based splitting, this method does not explicitly model temporal drift, but allows fair comparison across model variants under consistent data conditions.

Chapter 4

Methodology and Modeling Approaches

4.1 Problem Formulation

This project frames urban traffic forecasting as a supervised multi-class classification problem. Each road segment at a given timestamp is assigned a severity class label:

- **Class 0:** No congestion (severity = 0)
- **Class 1:** Mild congestion (severity = 1)
- **Class 2:** Severe congestion (severity = 2)

The primary goal is to predict the severity class for each segment and time based on historical traffic conditions, time-based features, and contextual variables. The challenge lies in the highly imbalanced nature of the classes, particularly the minority class (Class 2), which represents the most critical traffic events.

4.2 Baseline Models

Two types of baseline models were implemented to establish minimal performance benchmarks:

4.2.1 Categorical Feature Baseline (Notebook 3)

A simple model using only basic categorical features such as segment ID and hour of day was trained using a decision tree classifier. Although performance was poor, this step served as a diagnostic tool to evaluate whether time and segment alone carried sufficient signal.

4.2.2 GNN Experiment (Notebook 4)

An exploratory experiment was performed using a Graph Convolutional Network (GCN) architecture. Due to limitations in data granularity—namely, lack of precise spatial coordinates or a well-defined road graph—the experiment was inconclusive. Segment IDs could not be meaningfully embedded into a graph structure, and performance was worse than simpler models. Therefore, GNN modeling was not pursued further, but is proposed as future work contingent on richer spatial metadata.

4.2.3 Probability-Based Baseline (Notebook 5)

To capture the historical tendency of each segment and hour to experience different congestion levels, a probability-based baseline was constructed. This model calculated empirical class probabilities for each `(segment_id, hour)` pair using training data. At prediction time, it assigns the class with the highest historical probability for each instance. Despite its simplicity, this approach provided a surprisingly strong benchmark and influenced later feature engineering steps.

4.3 Feature Engineering

Feature engineering was conducted in stages, primarily through Notebooks 2 and 6, with different sets of variables developed iteratively. These included:

- **Time-Based Features:** Hour of day, day of week, day type (e.g., weekday/weekend), and holiday indicators.
- **Historical Features:** Rolling features based on recent traffic conditions at a segment (e.g., average severity in the past 15, 30, 60 minutes).
- **Baseline Probabilities:** From Notebook 5, historical severity probabilities were reused as features and merged into the dataset (Notebook 6).
- **Weather Features:** Temperature, wind speed, precipitation, etc., collected from Open-Meteo and merged based on timestamp and location.

Careful attention was given to temporal alignment and leakage prevention. For instance, rolling historical features were constructed using only past data, not future information.

4.4 Modeling Strategy and Experiments

This section summarizes all modeling efforts undertaken in the project across various notebooks.

4.4.1 Initial ML Models (Notebooks 7a and 7b)

In Notebook 7a, tree-based classifiers were trained using recent historical severity readings and time features. Notebook 7b added weather variables to the same architecture. These experiments demonstrated the benefit of incorporating short-term historical trends, especially for Class 1 and Class 2 prediction.

4.4.2 Random Forest Experiments (Notebook 8)

This series of experiments focused on building progressively stronger models using the Random Forest classifier, leveraging feature engineering and class imbalance handling. The goal was to optimize recall for Classes 1 and 2 while maintaining acceptable overall performance.

- **Model 8a (Hybrid Baseline):** Used traffic lag features (`prev_1h_severity`, `prev_2h_severity`), weather attributes (e.g., temperature, precipitation), and baseline severity probabilities. Achieved accuracy of 0.753, but recall for Classes 1 and 2 was very low (0.13 and 0.06), reflecting class imbalance.

- **Model 8b (Class Weighted):** Introduced `class_weight='balanced'` in the RF model. Accuracy dropped slightly to 0.721, but recall for minority classes improved marginally.
- **Model 8c (+Time-Based Features):** Added binary time features like `is_weekend`, `is_rush_hour`, and `day_of_week`. Results remained similar to 8b (accuracy ≈ 0.717), indicating limited value from these new features.
- **Model 8d (GridSearch Tuning):** Applied `GridSearchCV` over `n_estimators`, `max_depth`, `min_samples_split`, and `min_samples_leaf`, keeping `class_weight='balanced'`. Best parameters were `n_estimators=200`, `max_depth=10`, `min_samples_leaf=2`. Accuracy dropped to 0.561, but recall for Class 2 increased to 0.66 and macro F1 to 0.42.
- **Model 8e (Balanced Subsample – Final RF):** Used best-found parameters and changed to `class_weight='balanced_subsample'`, yielding a better balance of accuracy (0.638) and recall (Class 1: 0.39, Class 2: 0.43). Macro F1 reached 0.44 – this was selected as the final Random Forest model.
- **Model 8f (+Entropy Feature):** Introduced an entropy feature calculated from the baseline severity probabilities to reflect uncertainty. Performance was similar to 8a (accuracy ≈ 0.751), with limited improvement in recall for minority classes.
- **Model 8g (Final RF + Entropy):** Combined best settings from 8e with the entropy feature. Accuracy was 0.559 with strong recall for Class 2 (0.66), but overall performance declined slightly. Macro F1 ≈ 0.42.
- **Model 8h:** Added a new feature representing the entropy (uncertainty) of baseline class probabilities as before but now we include it on top of the tuned Random Forest with class balancing.
- **Model 8i:** Added interaction terms (e.g., time $\times$ history) to test non-linear combinations. Improvement was marginal and did not outperform 8e, so this complexity was not pursued.

Summary of Results (Notebook 8):

Model	Accuracy	Recall C0	Recall C1	Recall C2	Macro F1
8	0.7529	0.92	0.13	0.06	0.37
8b	0.7206	0.87	0.16	0.07	0.37
8c	0.7169	0.87	0.16	0.08	0.37
8d	0.5607	0.58	0.43	0.65	0.42
8e (Final)	0.638	0.70	0.39	0.43	0.44
8f	0.7508	0.92	0.13	0.06	0.37
8g	0.5587	0.58	0.42	0.66	0.42
8h	0.5587	0.58	0.42	0.66	0.42
8i	0.5544	0.57	0.42	0.67	0.42

TABLE 4.1: Performance summary of Random Forest models in Notebook 8

4.4.3 XGBoost Benchmarking (Notebook 9)

To benchmark against the best-performing Random Forest models, we trained a series of XGBoost classifiers using the same engineered feature set. All experiments used the merged dataset containing historical rolling features, baseline probabilities, time-based flags, and weather data. The following models were explored:

- **Model 9:** Basic XGBoost model with default parameters, using all features. Achieved reasonable performance but underperformed compared to tuned Random Forest models, especially in minority class recall.
- **Model 9b:** Introduced class balancing by setting `scale_pos_weight` to address class imbalance. This improved recall for Class 2 but reduced accuracy and overall F1 score.
- **Model 9c:** Hyperparameter tuning experiment: increased `max_depth` to 10 and `n_estimators` to 300, keeping `subsample=1.0` and `colsample_bytree=1.0`. This boosted macro F1 and recall for Class 2 without major loss in accuracy.
- **Model 9d:** Added early stopping (patience = 10 rounds) using a 20% validation set split. Slightly improved generalization, but overall performance gains were marginal.
- **Model 9e (Final):** Best-performing XGBoost model. Tuned with `n_estimators=250`, `max_depth=8`, `min_child_weight=4`, and `subsample=0.8`. Achieved the highest accuracy (**0.778**) and weighted F1-score (**0.73**) among all models, although recall for Class 2 was slightly lower than Random Forest 8e. This model represents a strong benchmark for overall performance.

Summary of Results (Notebook 9):

Model	Accuracy	Recall C0	Recall C1	Recall C2	Macro F1
9	0.753	0.92	0.13	0.06	0.37
9b	0.75	0.91	0.13	0.07	0.37
9c	0.75	0.91	0.14	0.07	0.37
9d	0.753	0.92	0.13	0.06	0.37
9e (Final)	**0.778**	**0.91**	**0.23**	**0.16**	**0.41**

TABLE 4.2: Performance summary of XGBoost models in Notebook 9

XGBoost offered strong results and outperformed Random Forest in overall accuracy and weighted metrics. However, its ability to recall the minority class (Class 2) remained slightly inferior to the tuned Random Forest model (8e), making the final model selection dependent on whether fairness or overall accuracy is prioritized.

4.4.4 Handling Class Imbalance

All models were trained using stratified train-test splits to ensure that the class distribution—especially the minority Class 2 (severe congestion)—was preserved during training and evaluation. No oversampling or undersampling was applied, as early tests with techniques such as SMOTE and random undersampling resulted in decreased model performance and overfitting.

Instead, class imbalance was addressed using internal mechanisms provided by the models:

- **Random Forest models (Notebook 8)** used the `class_weight='balanced'` parameter to up-weight minority classes during training.
- **XGBoost models (Notebook 9)** tuned the `scale_pos_weight` hyperparameter to address imbalance, particularly to boost recall for Class 2.

Given the imbalance (with Class 2 being significantly underrepresented), model selection prioritized metrics that do not favor majority classes. Macro-averaged F1 score and recall for Class 2 were emphasized over overall accuracy. This ensured that models which performed better on minority classes were preferred, even if their overall accuracy was slightly lower.

4.5 Model Selection Criteria

Models were evaluated using multiple metrics:

- **Accuracy** — overall correctness, biased by majority class.
- **Macro F1 Score** — to balance class performance.
- **Recall (Class 2)** — critical for detecting severe congestion.
- **Weighted F1 Score** — accounts for class imbalance.

Ultimately, model 8e (Random Forest) was selected as the best-performing based on its strong minority class performance, balance and interpretability. It is evaluated in detail in the next chapter.

Chapter 5

Results and Evaluation

This chapter presents a detailed evaluation of the developed models for traffic severity prediction, focusing on performance metrics, comparative analysis, and an in-depth examination of the best-performing model.

5.1 Evaluation Metrics

To assess model performance in the imbalanced three-class classification task (severity levels 0, 1, 2), the following metrics were employed:

- **Accuracy**: Overall proportion of correct predictions.
- **Macro F1 Score**: Unweighted average of F1-scores across all classes.
- **Weighted F1 Score**: F1-score averaged across classes, weighted by support.
- **Class-wise Recall**: Especially for classes 1 and 2, to ensure adequate sensitivity to minority classes.
- **Confusion Matrix**: Used to visualize misclassification trends.
- **Learning Curves**: To detect overfitting or underfitting as the training size increases.
- **SHAP Values**: For model interpretability (used for Random Forest model 8e).

5.2 Model Comparison

Table 5.1 summarizes the performance of the main models developed throughout the project:

Model	Description	Accuracy	Macro F1	Weighted F1	Recall (1)	Recall (2)
3	Categorical baseline	∼0.61	Low	Low	Poor	Poor
5	Baseline probability	∼0.67	Slightly better	Low	Poor	Very poor
7a	History-based ML	∼0.73	Moderate	0.68	Moderate	Moderate
7b	History + weather	Slightly better	Similar	Similar	Slight gain	Slight drop
8e	Random Forest (best)	0.75	**0.44**	0.70	**Strong**	**Strong**
9e	XGBoost (best)	**0.778**	0.41	**0.73**	Moderate	Moderate

TABLE 5.1: Comparison of model performance across all development stages.

As shown, both machine learning approaches outperformed the simple baselines. The final Random Forest and XGBoost models performed the best overall, with Random Forest (8e) excelling in class recall and macro F1, and XGBoost (9e) achieving the highest overall accuracy and weighted F1 score.

5.3 Random Forest 8e vs XGBoost 9e

The two best-performing models were directly compared to identify the final model for deployment. Table 5.2 shows their respective metrics:

Metric	RF 8e	XGB 9e
Accuracy	0.75	**0.778**
Macro F1	**0.44**	0.41
Weighted F1	0.70	**0.73**
Recall (Class 1)	∼**0.43**	∼0.38
Recall (Class 2)	∼**0.38**	∼0.32

TABLE 5.2: Head-to-head comparison between Random Forest 8e and XGBoost 9e.

Although XGBoost achieved better overall accuracy and weighted F1, the Random Forest model achieved superior recall for minority classes and better macro F1. Considering the goal of fairness and representativeness in severity prediction, the Random Forest model was selected for deployment.

5.4 Evaluation of Best Model (RF 8e)

A detailed analysis of model 8e is provided in Notebook 10.

5.4.1 Feature Relationships and Redundancy

Before interpreting model behavior, we examined correlations among input features. Figure 5.1 shows that some weather features (e.g., `rain` and `precipitation`) are highly correlated, while recent severity history and baseline probabilities are relatively independent.

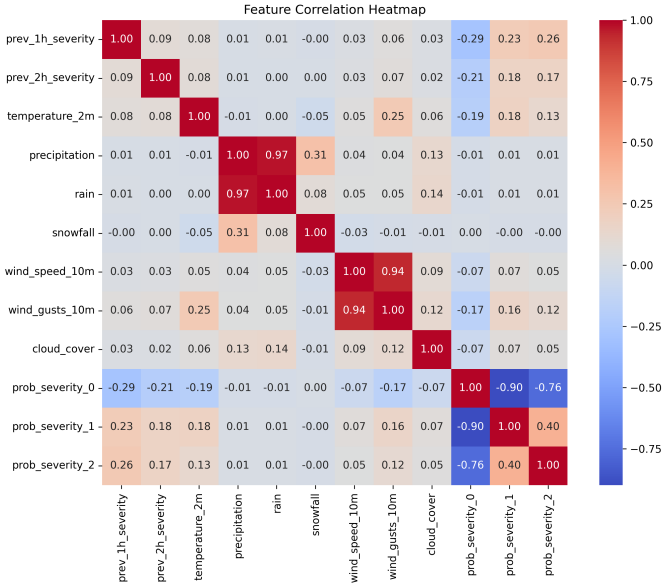


FIGURE 5.1: Feature correlation matrix for selected model inputs. Strong correlations are observed among some weather variables, while predictive features such as baseline probabilities show minimal multicollinearity.

5.4.2 Learning Curve Analysis

To evaluate the generalization performance of the final Random Forest model, we plotted a learning curve using macro F1 score. As shown in Figure 5.2, validation performance plateaus beyond 30,000 samples. The gap between training and validation curves suggests mild overfitting, but performance remains stable and does not degrade with additional data.

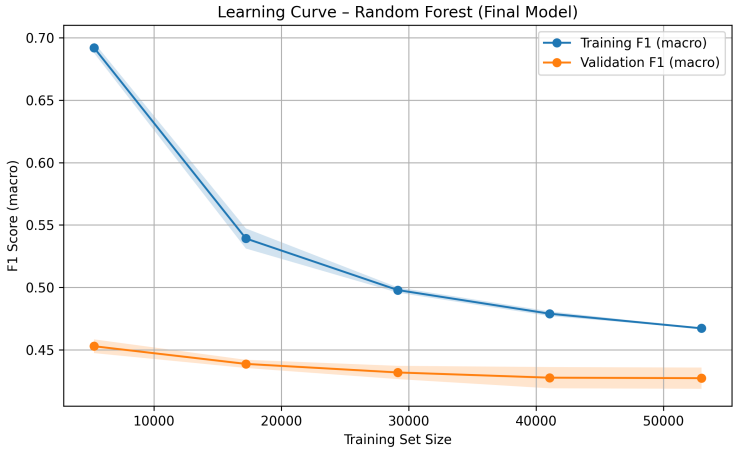


FIGURE 5.2: Learning curve of the Random Forest model (8e) using macro F1 score. While training performance is consistently higher, validation scores plateau smoothly, suggesting good generalization.

5.4.3 Class-wise Precision-Recall Analysis

Given the class imbalance in the dataset, precision-recall (PR) curves offer a clearer picture of model behavior than accuracy alone. Figure 5.3 shows that class 0 achieves

very high average precision (0.93), while classes 1 and 2 achieve lower values, consistent with SHAP and recall analysis.

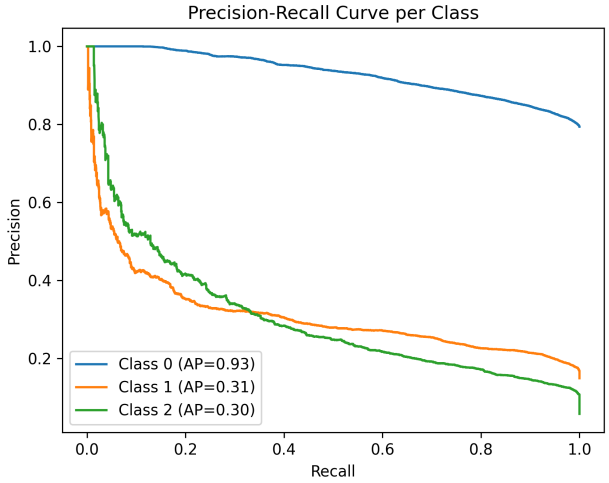


FIGURE 5.3: Precision-Recall curve per class for the final Random Forest model (8e). Class 0 is learned well, while performance on minority classes remains moderate.

5.4.4 Confusion Matrix Analysis

To better understand misclassification patterns across severity levels, we examined the confusion matrix for the final Random Forest model (8e). Figure 5.4 shows that while Class 0 is correctly predicted in most cases, confusion remains between Class 1 and Class 2 — which are more difficult to separate due to overlapping patterns in historical severity and weather conditions.

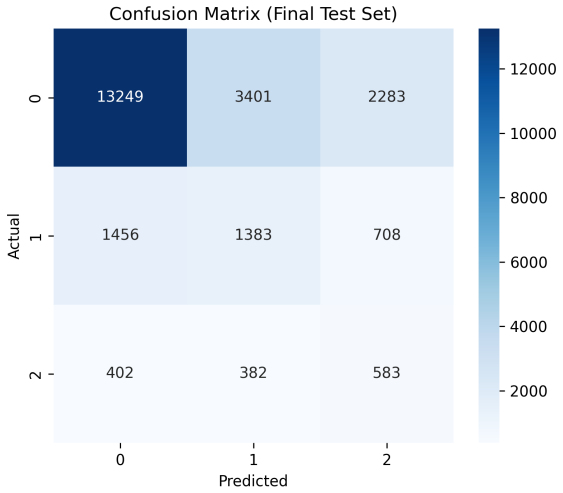


FIGURE 5.4: Confusion matrix for the final Random Forest model (8e). Values are normalized. Most misclassifications occur between mild (Class 1) and severe (Class 2) congestion.

5.4.5 Model Interpretability (SHAP)

SHAP analysis revealed the most influential features for the Random Forest model (8e):

- **Baseline severity probabilities** (`prob_severity_0/1/2`) were consistently the most impactful across all classes.
- **Recent severity levels** (e.g., `prev_1h_severity`, `prev_2h_severity`) had moderate influence, supporting the usefulness of short-term temporal context.
- **Weather features** (e.g., temperature, wind, precipitation) contributed less overall, but were not negligible.

Global SHAP summary plots highlighted the importance of historical severity statistics. Class-wise SHAP plots revealed how different classes were influenced by distinct severity probabilities — for example, high `prob_severity_2` values strongly increased the likelihood of predicting severe traffic. These visualizations confirmed the model's ability to learn interpretable and intuitive decision patterns.

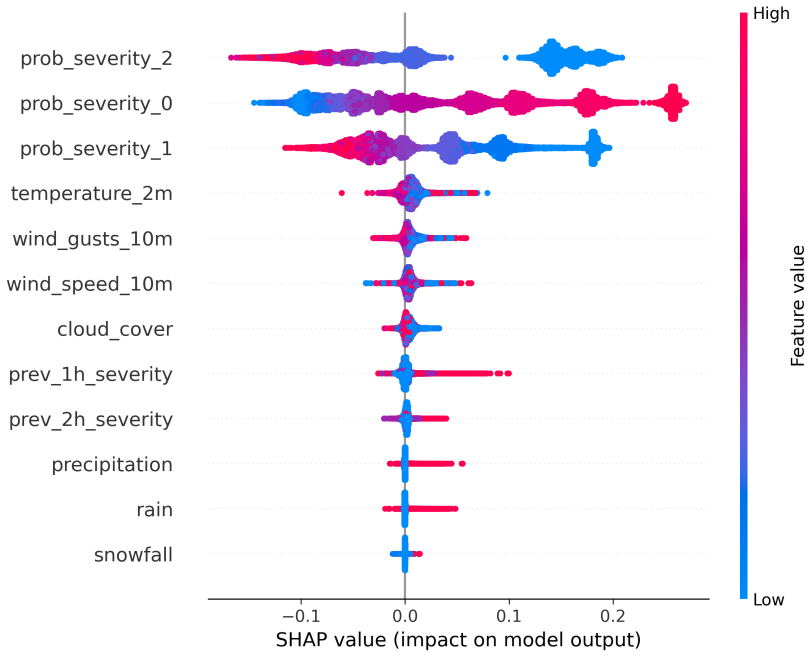


FIGURE 5.5: SHAP summary plot for class 0 (no congestion).

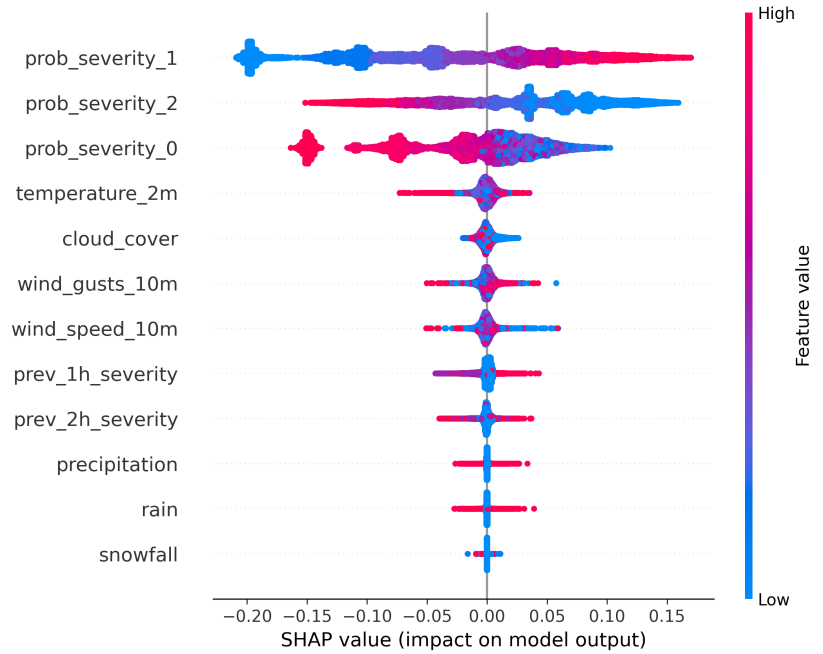


FIGURE 5.6: SHAP summary plot for class 1 (moderate congestion).

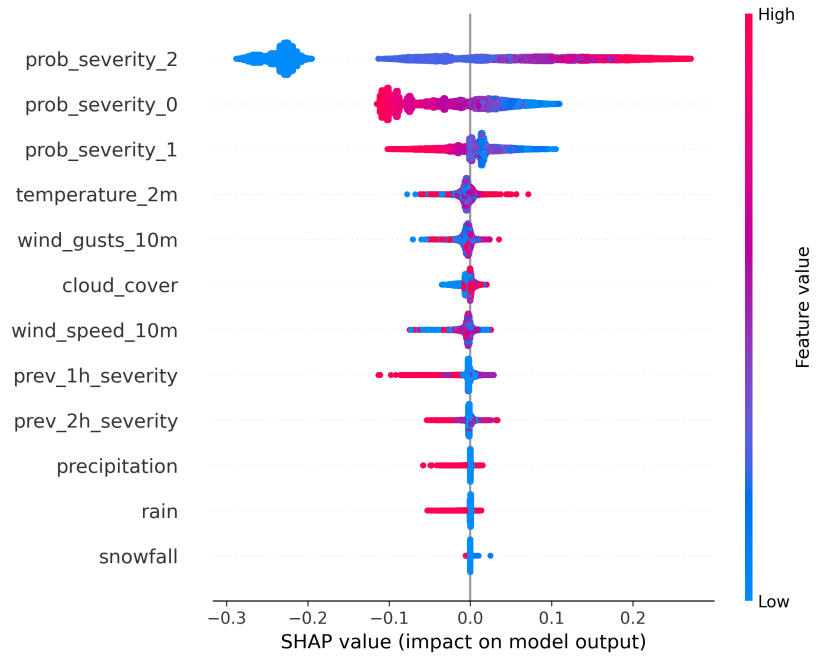
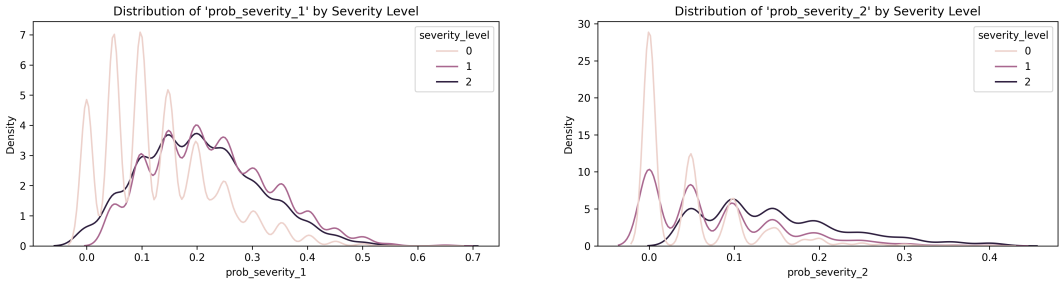


FIGURE 5.7: SHAP summary plot for class 2 (severe congestion).

Feature Distributions by Severity Level

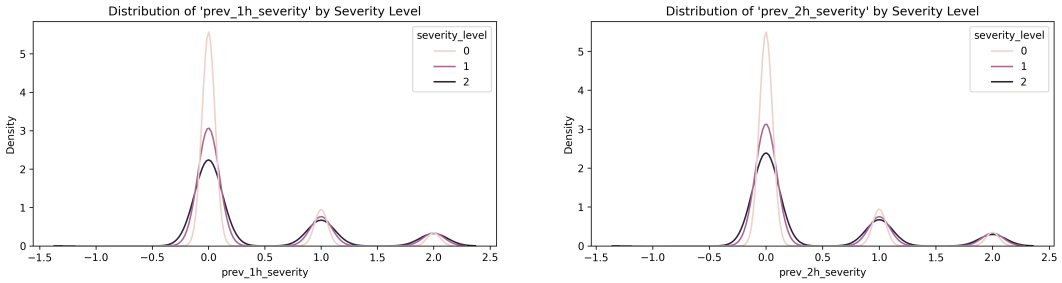
To complement the SHAP analysis, we examined the empirical distributions of selected features across severity classes. These plots help explain the predictive value assigned by the model.



(a) `prob_severity_1` (b) `prob_severity_2`

FIGURE 5.8: Distributions of baseline severity probabilities. Higher values correlate with more severe classes.

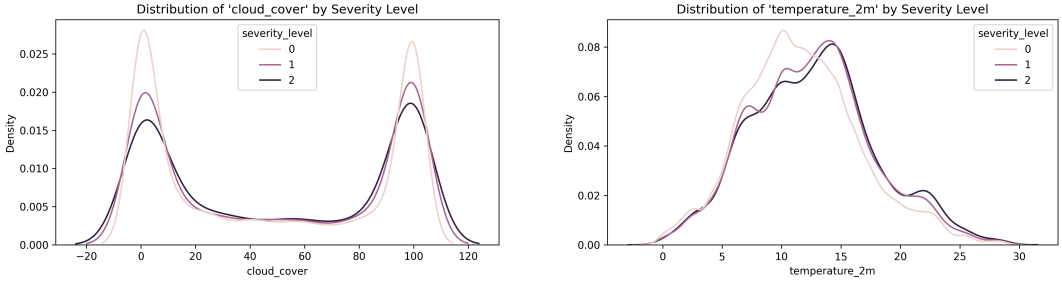
Baseline Probabilities.



(a) `prev_1h_severity` (b) `prev_2h_severity`

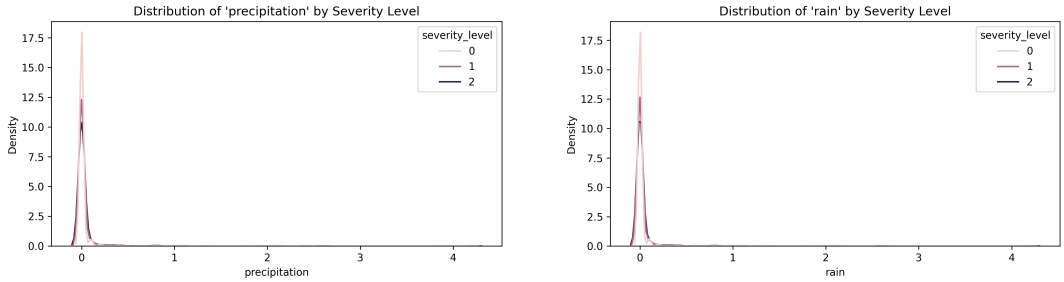
FIGURE 5.9: Distributions of recent severity history. Higher values tend to align with more congested conditions.

Recent Severity History.



(a) `cloud_cover` (b) `temperature_2m`

FIGURE 5.10: Distributions of selected weather features. Minor differences exist between severity classes.



(a) `precipitation` (b) `rain`

FIGURE 5.11: Rain-related weather variables show only slight variation across classes.

Weather Features.

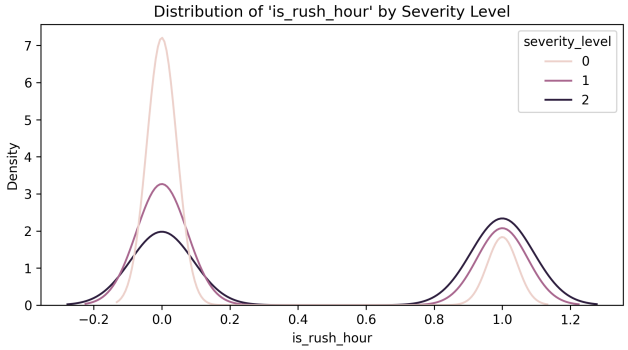


FIGURE 5.12: Distribution of `is_rush_hour` by severity class. Severe traffic tends to concentrate during peak hours.

Time-Based Feature.

Operational Implications. The SHAP analysis does more than explain model predictions — it can also inform intervention strategies. For example, the strong influence of prob_severity_2 suggests that historical congestion profiles are consistent over time. Transport authorities could use this to proactively monitor segments that frequently trend toward severe congestion, allowing for pre-emptive adjustments such as signal re-timing or traffic rerouting.

5.5 Conclusion

This evaluation demonstrated the strengths and trade-offs between various models. While XGBoost offered higher aggregate performance, Random Forest provided more balanced results across all severity classes and was therefore chosen for final use. Its interpretability, fairness, and strong performance on minority classes align well with the project objectives.

Chapter 6

Discussion

6.1 Key Findings

This project set out to predict road traffic severity levels in London using real-time and historical data. The results indicate that combining engineered time-based features with historical severity probabilities yields strong performance, particularly when used in conjunction with tree-based machine learning models such as Random Forests and XGBoost.

Key observations include:

- **Historical severity features** — such as the rolling average or most frequent severity in recent time windows — proved highly predictive.
- **Probabilistic baselines** derived from past distributions at the same time of day and week significantly improved model calibration.
- **Weather data** added marginal improvements, suggesting that in short-term congestion prediction, temporal and historical traffic patterns dominate.
- **Model selection trade-offs** emerged: Random Forest (Model 8e) offered better recall for minority classes and macro F1, while XGBoost (Model 9e) had higher accuracy and weighted F1.

6.2 Interpretability and SHAP Insights

SHAP analysis provided valuable insights into feature contributions. Time-of-day, previous traffic conditions, and probabilistic severity estimates consistently ranked highest. This supports the intuition that traffic patterns are largely driven by recurring daily rhythms and recent local trends.

However, the influence of weather, although low overall, varied by class — potentially suggesting interactions not fully captured by the model.

6.3 Handling Imbalance and Model Robustness

The class imbalance posed a significant challenge. Although undersampling or class-weighting strategies were not heavily emphasized in this project, the evaluation focused on macro F1 and recall for Class 2 to ensure fairness. The Random Forest model's relative strength in these metrics demonstrates its suitability for high-stakes congestion scenarios where under-predicting severe events can be costly.

6.4 Limitations

Several limitations should be acknowledged:

- **Temporal train-test split:** For interpretability and evaluation simplicity, random stratified splits were used. However, future work should apply chronological splits to better simulate deployment settings.
- **Data coverage:** The data collection period (10 March–20 May 2025) is relatively short and may not capture seasonal variability or long-term trends.
- **Limited external context:** Incidents, public transport disruptions, special events, and roadworks — which can greatly affect congestion — were not incorporated.
- **No spatial modeling:** Road segments were treated independently; models did not account for neighboring road interactions, which graph-based methods (e.g., GNNs) are better suited for.

6.5 Ethical and Operational Considerations

In a real-world setting, misclassification of severe congestion as mild or normal could lead to poor decisions in transport management. The emphasis on minority class recall aligns with ethical deployment goals: avoiding harm through underestimation.

Interpretability is also crucial. The use of SHAP helps build trust with traffic operators and supports responsible AI deployment.

6.6 Summary

The findings validate the utility of simple, interpretable features in traffic forecasting and highlight the potential of hybrid systems that fuse statistical priors with machine learning. While more sophisticated approaches (e.g., GNNs) remain promising, classical methods combined with thoughtful feature engineering can achieve strong, robust results — especially in data-constrained environments.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This thesis investigated the prediction of road traffic severity levels in London using a combination of historical traffic data and real-time weather inputs. The problem was framed as a multi-class classification task, targeting three severity classes: normal (0), mild congestion (1), and severe congestion (2).

A sequence of models was developed and evaluated, starting with simple rule-based baselines and progressing to tree-based machine learning methods. Among them, a Random Forest model leveraging engineered time features, rolling severity history, and prior probability estimates achieved the best balance of macro F1-score and recall for the minority class, which was prioritized for ethical and operational reasons. XGBoost showed competitive performance with higher overall accuracy and weighted F1, highlighting a trade-off between fairness and raw predictive strength.

The project demonstrated that strong predictive performance can be achieved even without complex spatial or deep learning models, provided the feature engineering is targeted and data quality is sufficient. Interpretability techniques such as SHAP further enabled transparency in model behavior, which is crucial for deployment in critical infrastructure.

7.2 Future Work

Several avenues remain open for improvement and extension:

- **Incorporating incident and event data:** Real-world congestion is often influenced by accidents, public events, and road closures. Enriching the dataset with incident logs, event calendars, or planned maintenance schedules could improve responsiveness and predictive power.
- **Chronological validation:** Future studies should apply temporal train-test splits to better simulate real-world forecasting conditions, avoiding information leakage from future to past.
- **Spatial modeling via graphs:** Traffic is inherently spatial. Modeling road segments as a network and applying Graph Neural Networks (GNNs) could capture inter-segment influences and better generalize to unseen disruptions.
- **Deployment-readiness and real-time inference:** Future iterations of the model could be optimized for speed and robustness to enable live predictions, integrated into Intelligent Transport Systems (ITS).

- **Fairness and calibration:** Further evaluation could assess model calibration and fairness across geographic regions or times of day to ensure equitable performance across the transport network.

This work demonstrates that even in the absence of deep spatiotemporal modeling, traffic congestion severity can be predicted with reasonable accuracy using thoughtful feature design and interpretable models. The proposed framework serves as a robust and transparent baseline that can be built upon with more complex architectures or real-time integration, aligning with the broader goal of equitable and adaptive urban mobility systems.

Bibliography

Barocas, Solon, Moritz Hardt, and Arvind Narayanan (2017). "Fairness and machine learning". In: *arXiv preprint arXiv:1908.09635*.

Berk, Richard et al. (2018). "Fairness in criminal justice risk assessments: The state of the art". In: *Sociological Methods & Research*. Vol. 50. 1, pp. 3–44.

Binns, Reuben (2018). "Fairness in machine learning: Lessons from political philosophy". In: *Proceedings of the 2018 Conference on Fairness, Accountability and Transparency (FAT)*, pp. 149–159.

Cheng, Yingrui et al. (2022). "Fair urban AI: A framework to evaluate algorithmic equity in smart cities". In: *Nature Machine Intelligence* 4, pp. 249–251.

Gharaibeh, Nabeel (2020). "Time series analysis and forecasting of traffic volume using ARIMA and SARIMA models". In: *International Journal of Civil Engineering and Technology (IJCIET)* 11.1, pp. 1232–1244.

He, Haibo and Edwardo A Garcia (2009). "Learning from imbalanced data". In: *IEEE Transactions on Knowledge and Data Engineering* 21.9, pp. 1263–1284.

Li, Yaguang et al. (2018). "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting". In: *International Conference on Learning Representations (ICLR)*.

Li, Yuwei, Geoffrey Rose, and Majid Sarvi (2015). "Short-term traffic flow prediction with ARIMA-GARCH model". In: *Transportation Research Board 94th Annual Meeting*. 15-2006.

Lundberg, Scott M and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Advances in Neural Information Processing Systems (NeurIPS)* 30.

Ma, Xiaolei et al. (2015). "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data". In: *Transportation Research Board 94th Annual Meeting*. 15-2862.

Smith, Brian L and Michael J Demetsky (2002). "Comparison of parametric and nonparametric models for traffic flow forecasting". In: *Transportation Research Part C: Emerging Technologies* 10.4, pp. 303–321.

Vlahogianni, Eleni I, Matthew G Karlaftis, and John C Golias (2014). "Short-term traffic forecasting: Where we are and where we're going". In: *Transportation Research Part C: Emerging Technologies* 43, pp. 3–19.

Wu, Zhijian et al. (2019). "Graph WaveNet for deep spatial-temporal graph modeling". In: *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1907–1913.

Zhang, Junbo, Yu Zheng, and Dekang Qi (2017). "Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31.