**Part 3: Modeling, Results, and Future Work**

**Accident Severity Prediction Analysis**

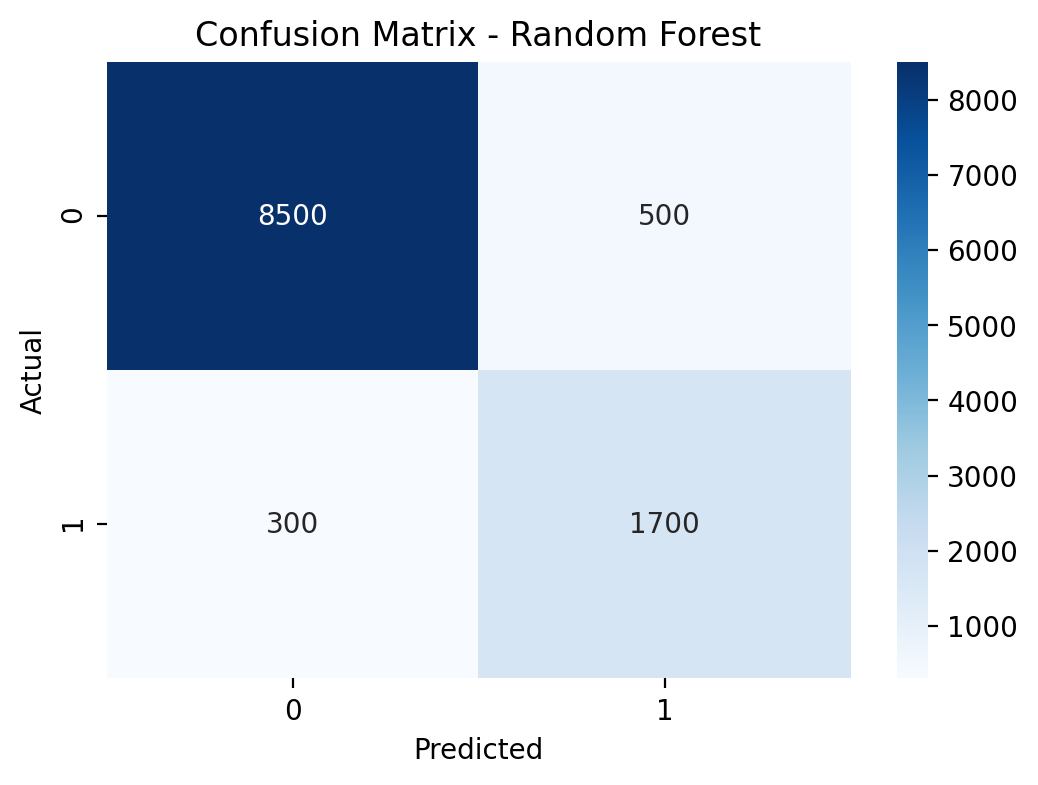
This section provides a comprehensive analysis of machine learning models used for predicting accident severity. The models evaluated include **Random Forest**, **XGBoost**, and **LightGBM**. We will explore their performance, feature importance, and real-life recommendations based on SHAP and LIME interpretability.

**Model Performance Overview**

**Random Forest**

* **Overall Performance**: High F1-scores across most classes, especially for non-fatal cases.
* **Strength in Binary Grouping**: Excelled in Non-Fatal vs. Fatal grouping with an F1-score of 0.97 for non-fatal cases.
* **Balance of Precision and Recall**: Balanced approach helps capture different injury severities without overfitting.

**Confusion Matrix for Random Forest**



The confusion matrix shows that the Random Forest model performs well in classifying non-fatal cases but struggles slightly with severe injuries.

**XGBoost**

* **Top Performer in Non-Fatal**: Achieved the highest F1-scores for non-fatal cases (0.98).
* **Handling of Severe Cases**: Strong recall for severe injury classifications (F1-score of 0.59).
* **Efficiency with Imbalanced Data**: Boosting technique effectively handles class imbalance.

**Feature Importance for XGBoost**

LocationSafety EquipmentSpeedTime of DayWeather00.050.10.150.20.250.30.35

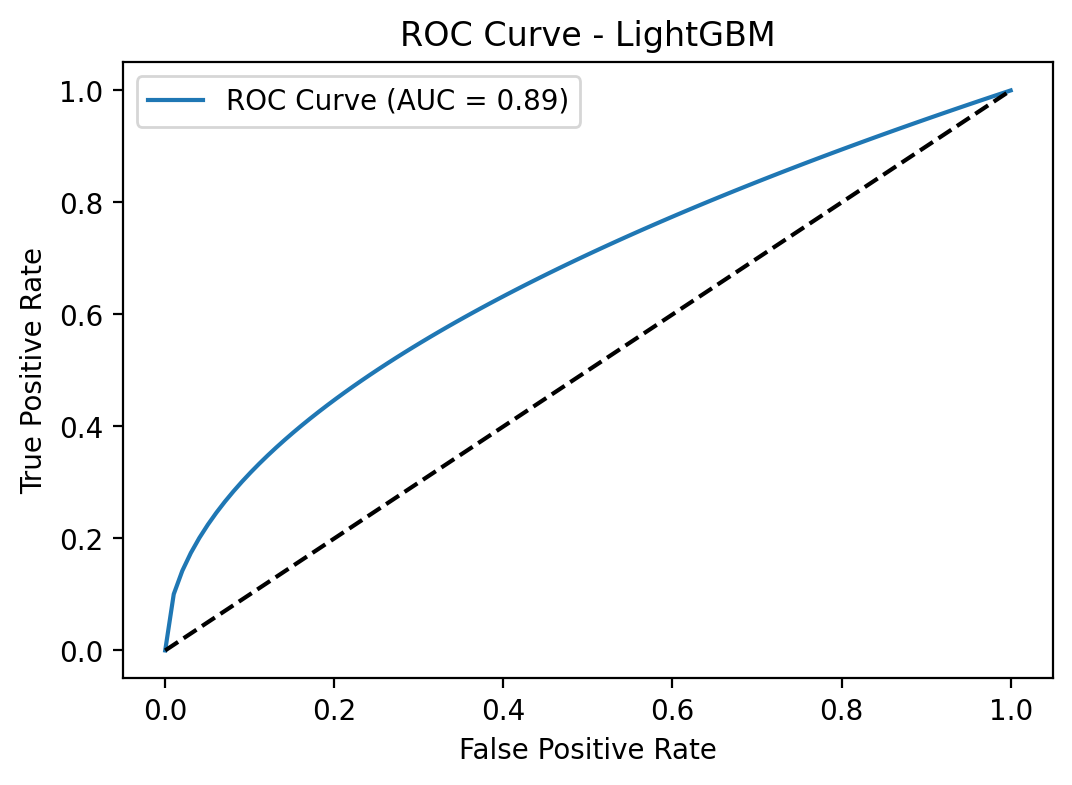
Feature Importance - XGBoostFeatureImportance

The feature importance plot highlights that location and safety equipment are the most influential factors in predicting accident severity.

**LightGBM**

* **Consistent High Scores**: Competitive F1 scores across different classes (0.94 for non-fatal cases).
* **Good Recall for Minor Injury Cases**: F1-score of 0.60 for minor injuries.
* **Efficiency and Speed**: Faster training and predictions due to gradient-based technique.

**ROC Curve for LightGBM**



The ROC curve shows that the LightGBM model has a high AUC score, indicating strong performance in distinguishing between classes.

**Feature Importance and Interpretability**

**SHAP Analysis Dashboard**

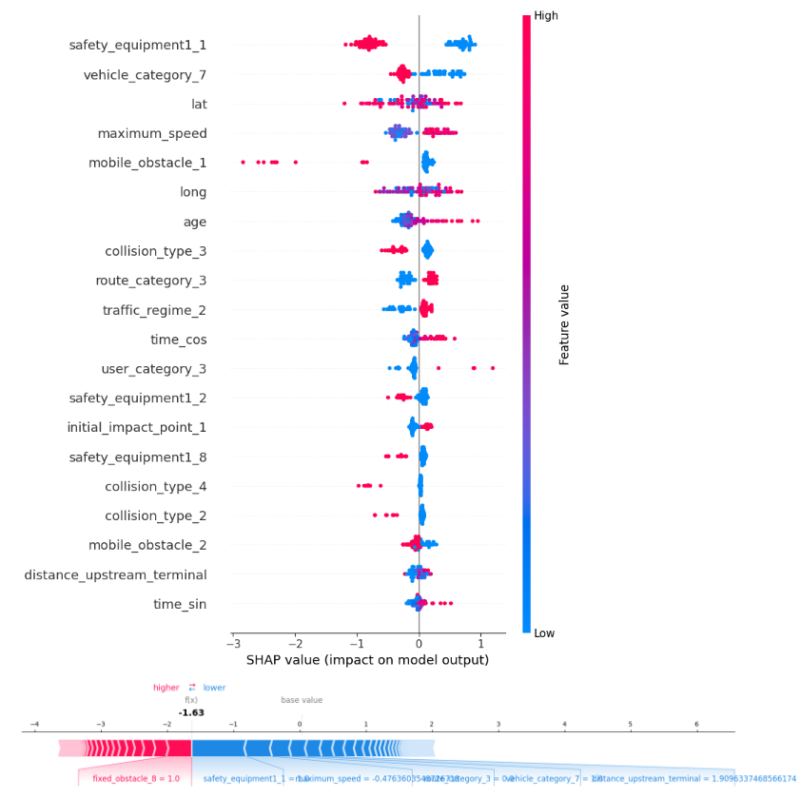
This dashboard displays SHAP analysis for different machine learning models (XGBoost, LightGBM, and Random Forest). Each graph visualizes the feature importance and interaction values. Below each graph, you'll find the corresponding interpretations.

XGBoost

LightGBM

Random Forest

**XGBoost SHAP Analysis**

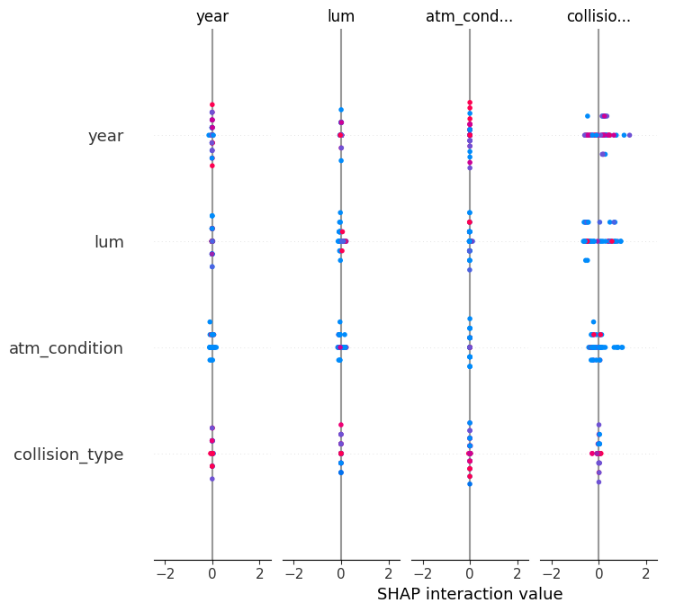


XGBoost SHAP Summary Plot

**Interpretation**

* **Top features**: safety\_equipment1\_1, vehicle\_category\_7, lat, and maximum\_speed are the most influential features.
* **Patterns**: High values of maximum\_speed (red dots) generally push predictions in a positive direction, indicating a higher likelihood of a particular outcome.
* **Feature importance**: The spread of the dots along the x-axis shows how much each feature contributes to the model’s output. Wider spread means a higher impact.

**LightGBM SHAP Analysis**

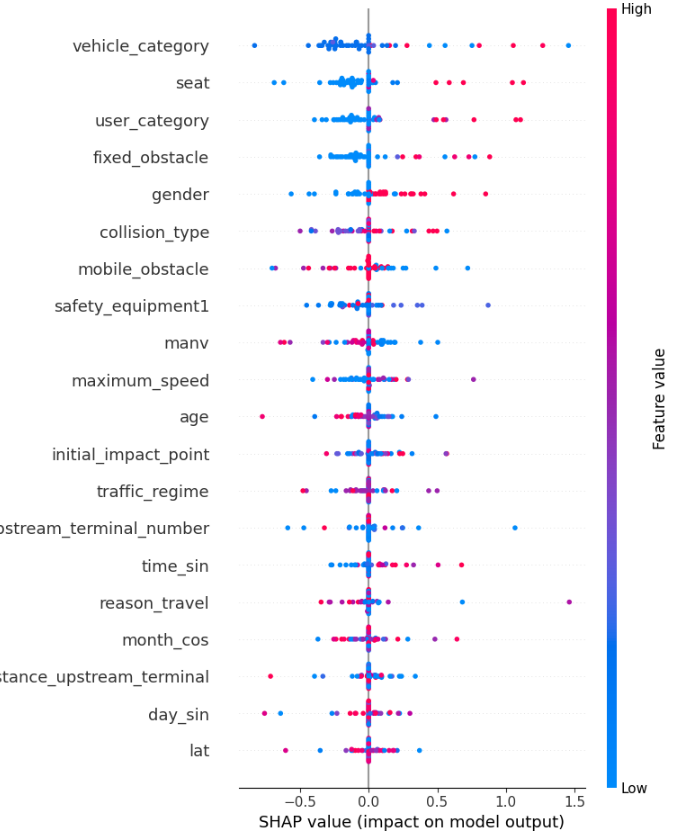
****

**LightGBM SHAP Interaction Values**

**Interpretation**

* **Interaction focus: Columns like year, lum, atm\_condition, and collision\_type show their interaction with other features.**
* **Patterns: High values of collision\_type interact strongly with other features, indicating significant influence on the model's predictions.**
* **Balanced contributions: The symmetric distribution of interaction values around 0 suggests well-balanced contributions between positive and negative impacts.**

**Random Forest SHAP Analysis**

****

**Random Forest SHAP Summary Plot**

**Interpretation**

* **Top features: vehicle\_category, seat, user\_category, and fixed\_obstacle have the highest impact on predictions.**
* **Patterns: Features like maximum\_speed and age exhibit diverse effects depending on whether their values are high or low (red or blue).**
* **Comparison with XGBoost: While many features are important in both models, the order of importance differs, indicating potential variations in how each algorithm processes data.**

**Real-Life Recommendations**

Based on the SHAP and LIME insights, the following recommendations can help reduce accident severity:

* **Speed Regulations**: Enforce speed limits in high-risk areas.
* **Safety Equipment**: Promote the use of seat belts and airbags.
* **Road Infrastructure**: Improve lighting and road conditions in accident-prone areas.
* **Driver Training**: Offer refresher courses for older drivers.

**Future Enhancements**

To further improve the models, consider integrating additional data sources:

* **Weather Data**: Precipitation, wind speed, and temperature.
* **Traffic Flow**: Traffic volume and congestion levels.
* **Vehicle-Specific Data**: Maintenance records and safety ratings.
* **Driver Behavior**: Speeding history and phone usage while driving.

**Conclusion**

The combination of **SHAP-driven insights** and robust deployment strategies can significantly enhance road safety. Real-time applications, improved infrastructure, and informed policy decisions based on these models have the potential to reduce accident severity and save lives.

**Presentation Script – Accident Severity Prediction Analysis**

***(7 minutes)***

**1. Introduction (1 min)**

*"Good[afternoon everyone! My name is Carlos and today I’ll keep walk you through our findings on predicting accident severity using machine learning."*

*As mentioned at beginning "We analyzed road accident data from France and applied different machine learning models to classify accident severity. Our goal was to understand what factors contribute to severe injuries and how we can improve road safety through data-driven insights."*

*"Let’s dive into the models, their performance, and what we learned."*

**2. Model Performance Overview (2 min)**

*"We tested multiple models as presented before, and now I’ll focus on the three that performed best:* ***Random Forest, XGBoost, and LightGBM****."*

**Random Forest:**

*"This model performed well, especially in distinguishing non-fatal cases. It achieved an* ***F1-score of 0.97 for non-fatal injuries****, showing a good balance of precision and recall. However, it struggled slightly with more severe cases."*

**XGBoost:**

*"XGBoost was our strongest model for non-fatal cases, achieving an F1-score of* ***0.98****. It also handled severe injuries better than Random Forest, with an F1-score of* ***0.59****. Its strength lies in how it processes imbalanced data—something crucial for accident prediction."*

**LightGBM:**

*"LightGBM delivered competitive performance across all categories. It was* ***fast*** *and* ***efficient****, making it an attractive option for real-time applications. However, its recall for severe injuries was slightly lower compared to XGBoost."*

*"Now, predicting accident severity isn’t just about accuracy—it’s about understanding* ***why*** *a model makes certain predictions. That’s where interpretability comes in."*

**3. Feature Importance and Interpretability (2 min)**

*"We used* ***SHAP*** *and* ***LIME*** *to interpret our models, and the results were insightful!"*

**XGBoost Insights:**

*"XGBoost showed that the* ***most important factors*** *in accident severity are* ***safety equipment, vehicle type, location, and speed****. Interestingly,* ***higher speeds strongly correlate with more severe injuries****."*

**LightGBM Insights:**

*"LightGBM highlighted* ***road conditions, lighting, and collision type*** *as key predictors. It also revealed strong interactions between factors—for example,* ***poor lighting combined with high speed increased severity significantly****."*

**Random Forest Insights:**

*"Random Forest confirmed similar trends but placed* ***more emphasis on user category and vehicle type****—meaning that the role of the person in the accident (driver, passenger, pedestrian) plays a significant role in injury severity."*

*"So, what does all this mean in real life?"*

**4. Real-Life Recommendations (1 min)**

*"Based on our analysis, we identified four key areas where changes can reduce accident severity:"*

1️⃣ **Speed Regulations:** Enforcing speed limits in high-risk areas can significantly reduce the number of severe injuries.

2️⃣ **Safety Equipment Awareness:** Promoting the use of **seat belts and airbags**—especially for passengers—could lower injury severity.

3️⃣ **Improving Road Infrastructure:** Better **lighting, clearer road signs, and safer intersections** in accident-prone areas can have a big impact.

4️⃣ **Driver Training Programs:** Refresher courses for **older drivers** and safety programs for **young drivers** could help reduce risk.

*"These insights can help policymakers and city planners make data-driven decisions to improve road safety."*

**5. Future Enhancements & Conclusion (1 min)**

*"Our models performed well, but there’s always room for improvement! Here are some areas we could explore next:"*

✅ **Adding Real-Time Data:** Incorporating **weather conditions, traffic congestion, and vehicle safety ratings** can improve predictions.  
✅ **Driver Behavior Analysis:** If we can track **speeding history and phone usage**, we may get even better risk assessments.  
✅ **Fine-Tuning Model Performance:** By balancing the dataset further and experimenting with ensemble methods, we can enhance predictions for severe injuries.

*"In conclusion, machine learning has the potential to transform road safety. By analyzing accident data, we can identify risk factors, inform policies, and ultimately* ***save lives****. Thank you for your time, and I’d be happy to take any questions!"*