Car Accident Severity Prediction

Introduction

- Road safety is a serious public concern, thus it is important to perform research on accident analysis and prediction.
- Goal is to predict accident severity based on time, traffic conditions, and other environmental factors
- Benefits
    - Provide relevant insights for the development of traffic rules and accident prevention
    - Provide practical information to the general public

Data Pre-processing

- Dataset containing collision data collected by the Seattle Department of Transportation from 2004 to 2020
- Missing values were either replaced with related values or dropped
- Most features contain categorical values, thus there were converted to numerical values
- Two classes were observed for the label (i.e. severity code): 1 - property damage, 2 - injury
- Class imbalance was fixed by under-sampling the majority class (severity = 1)
- Cleaned data contains 113,954 observations and 9 features (address type, collision type, junction type, weather, road condition, light condition, vehicle count, hour of day, and day of week)

Data Visualisation

Relationship between target variable and selected features

- It can be observed that most accidents occurred on Fridays and midnight, and that most features have noticeable impact on accident severity.

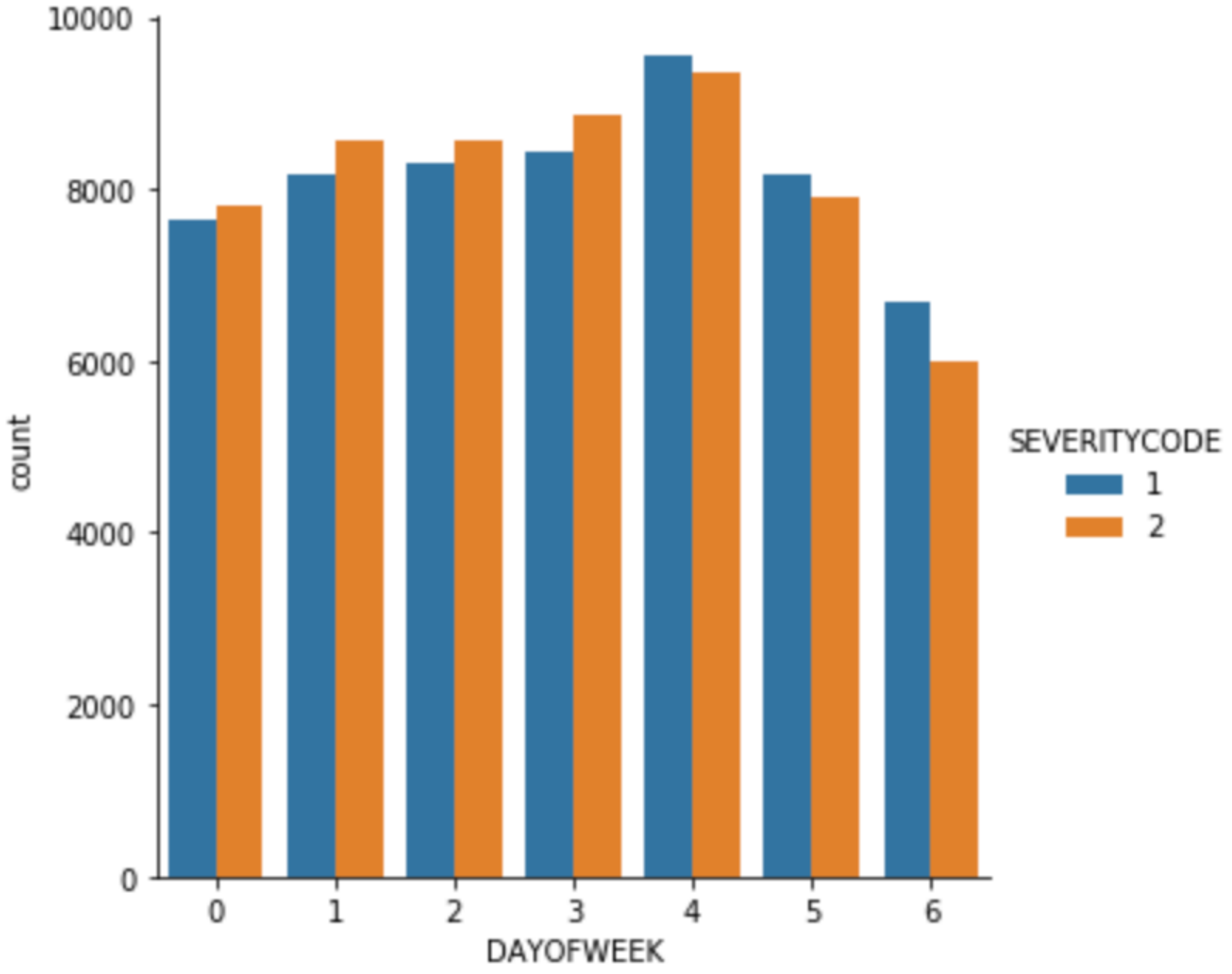


Figure 1. Distribution of severity per day

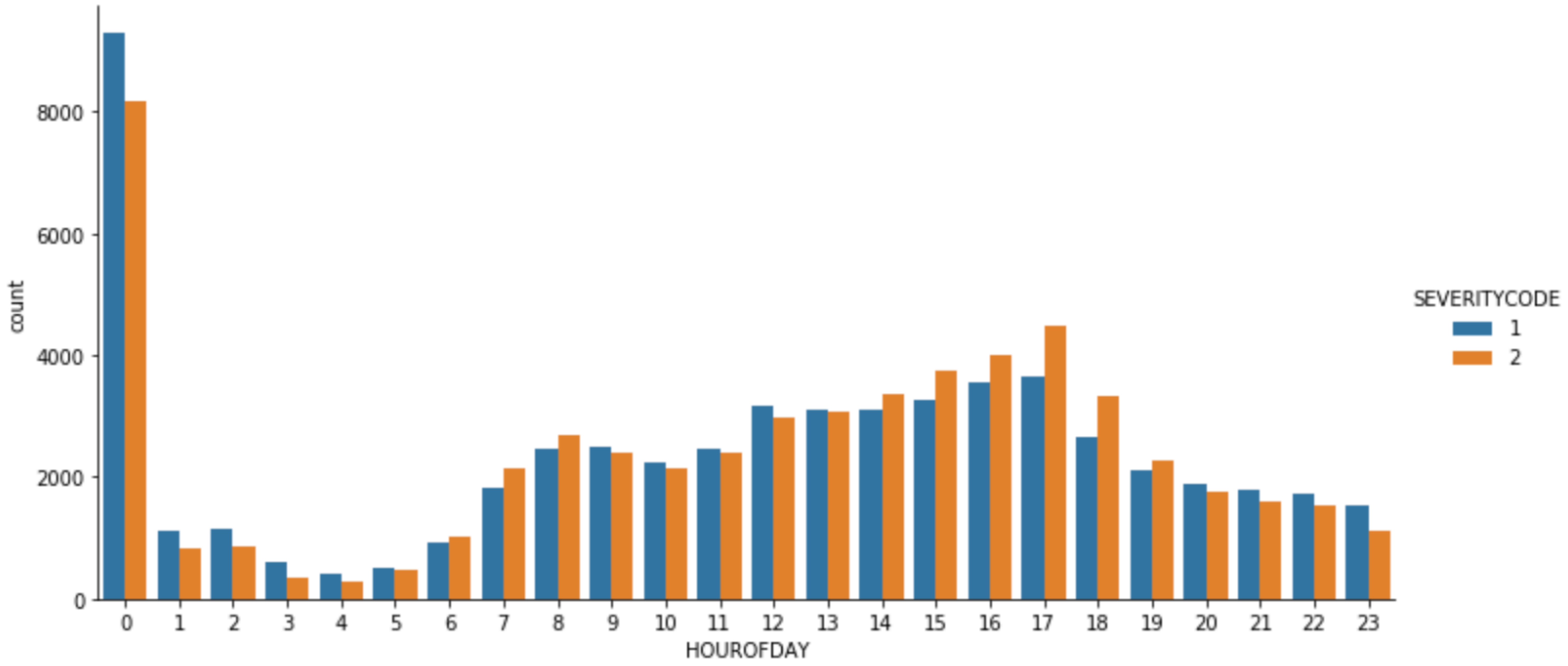


Figure 2. Distribution of severity per hour

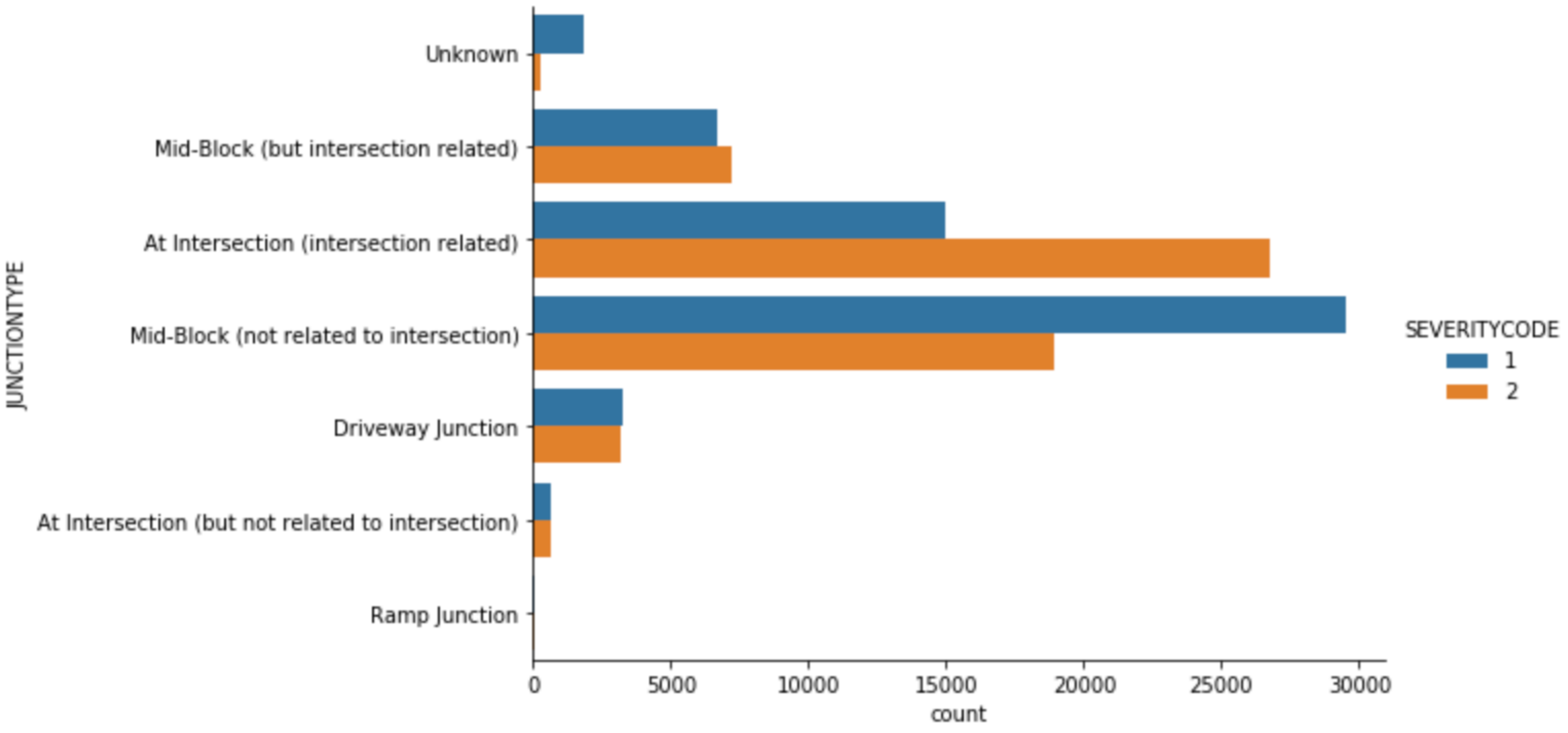


Figure 3. Impact of junction type on severity

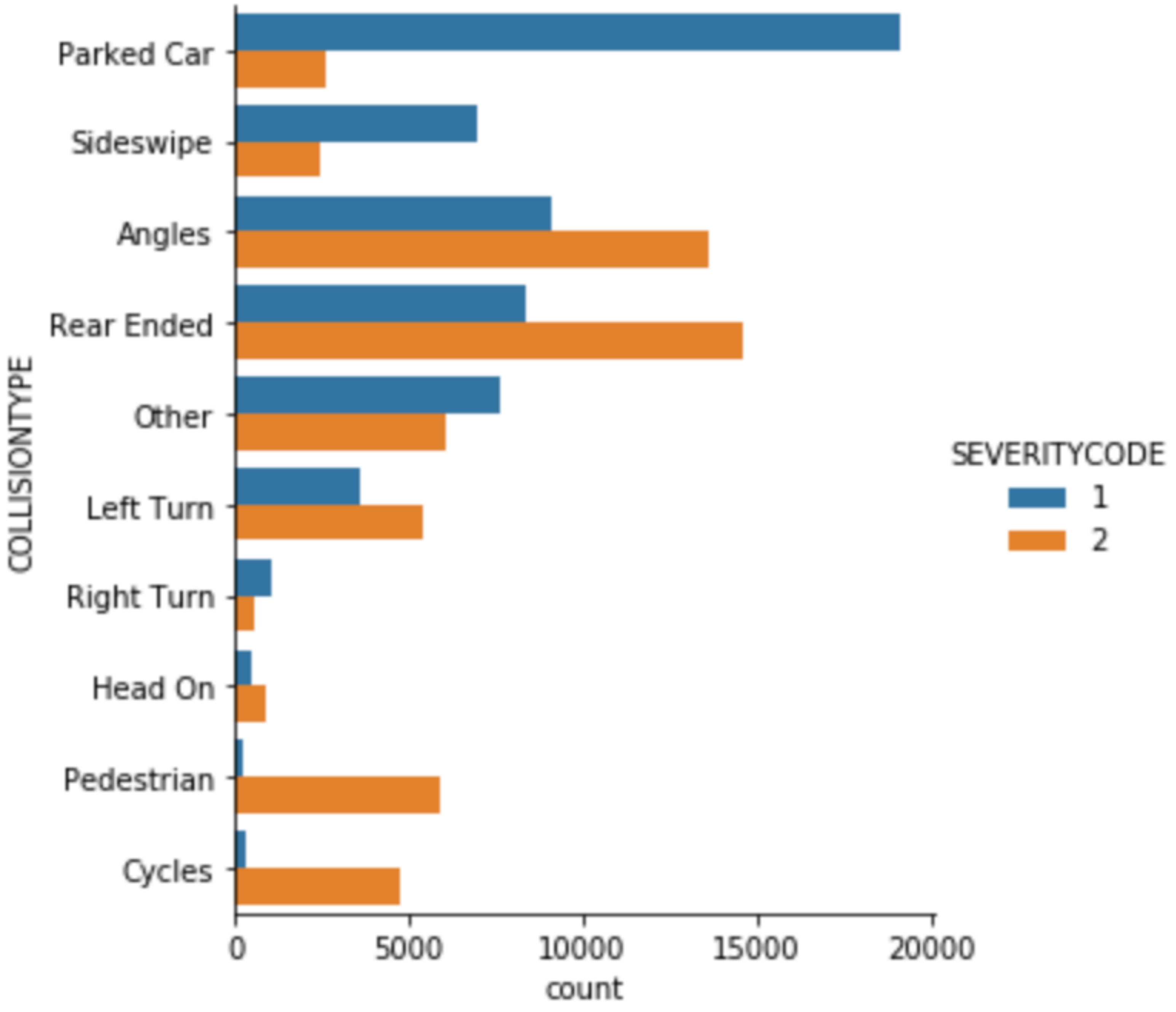
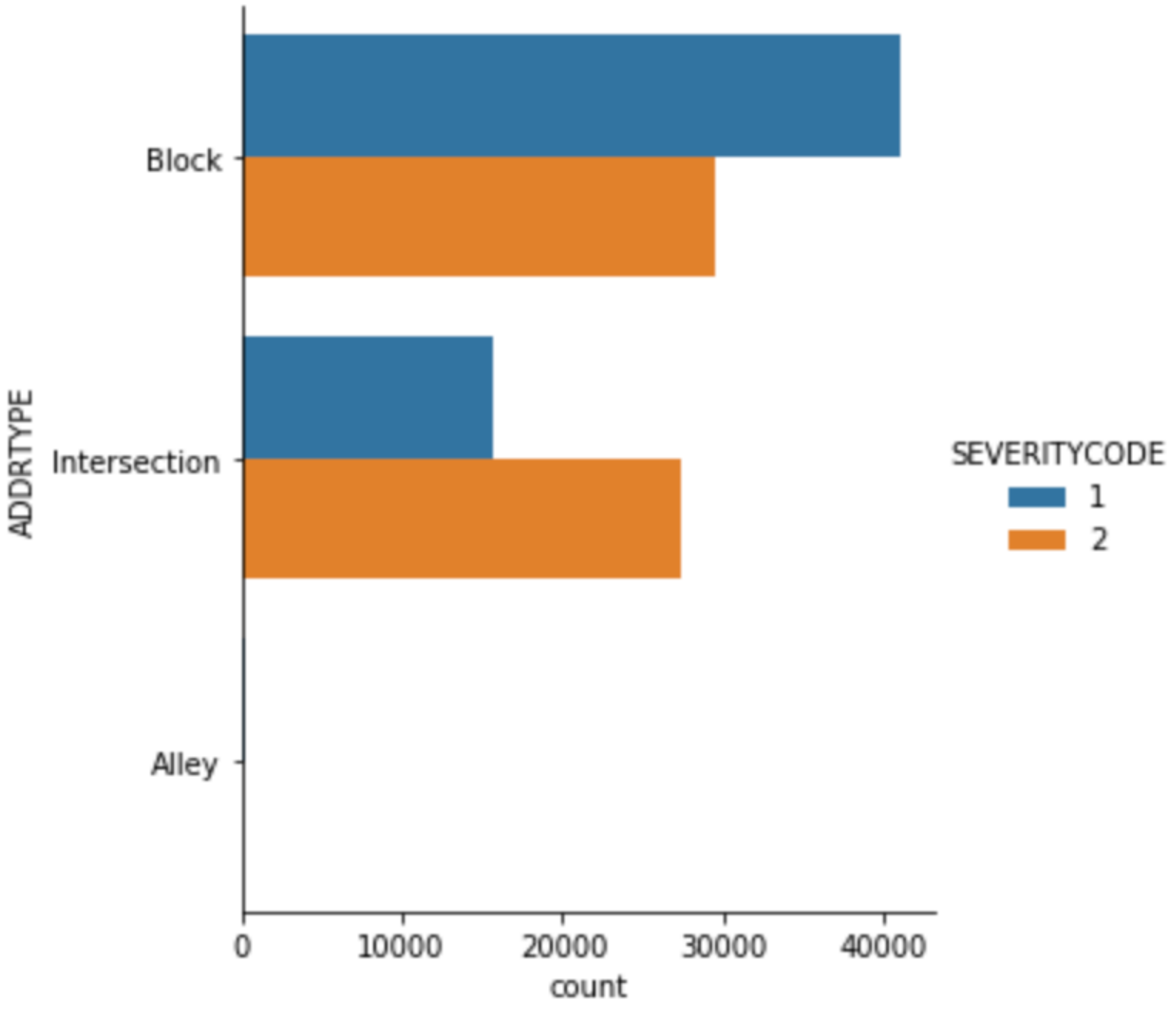


Figure 4. Impact of collision type on severity



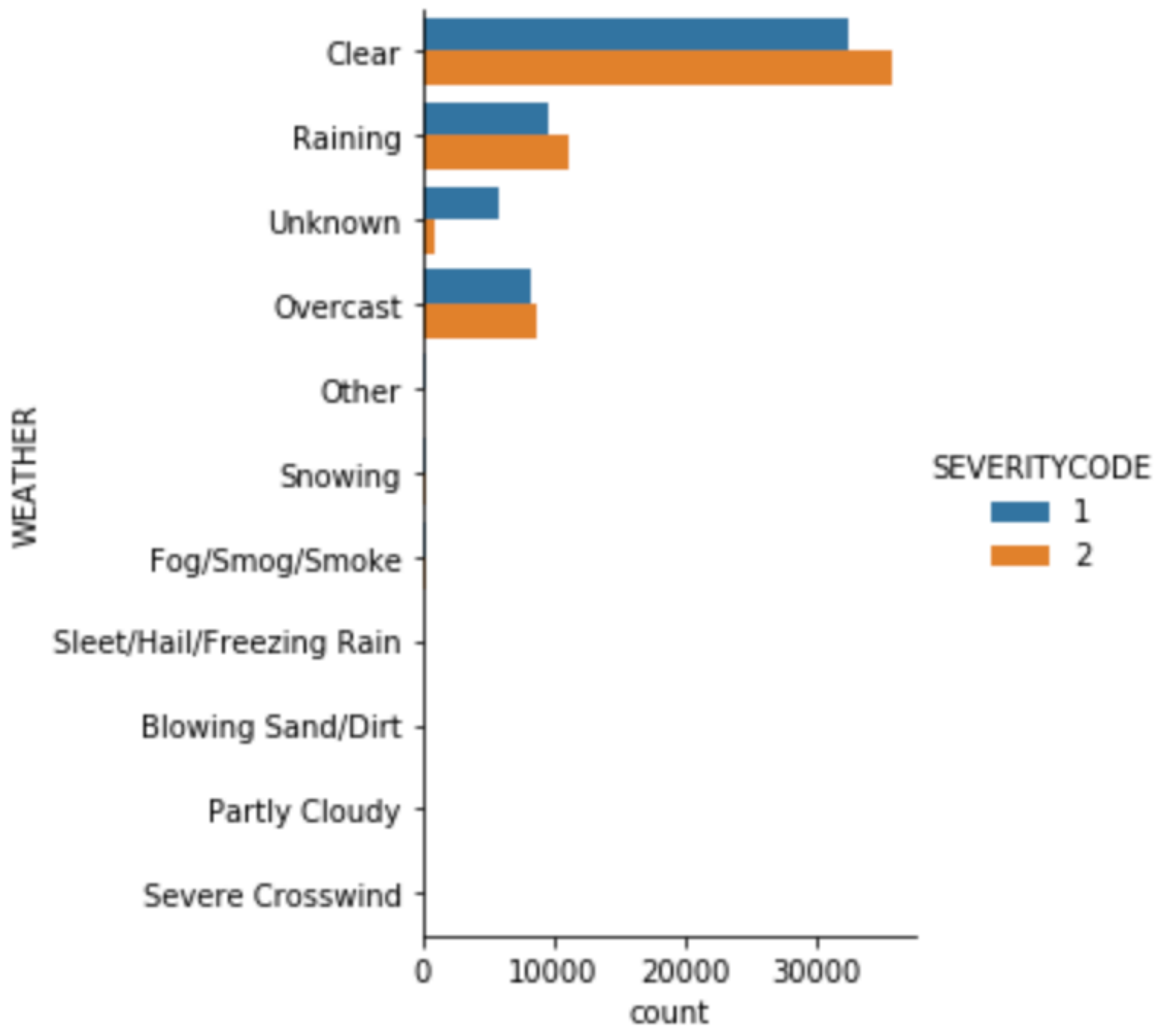


Figure 8. Impact of weather on severity

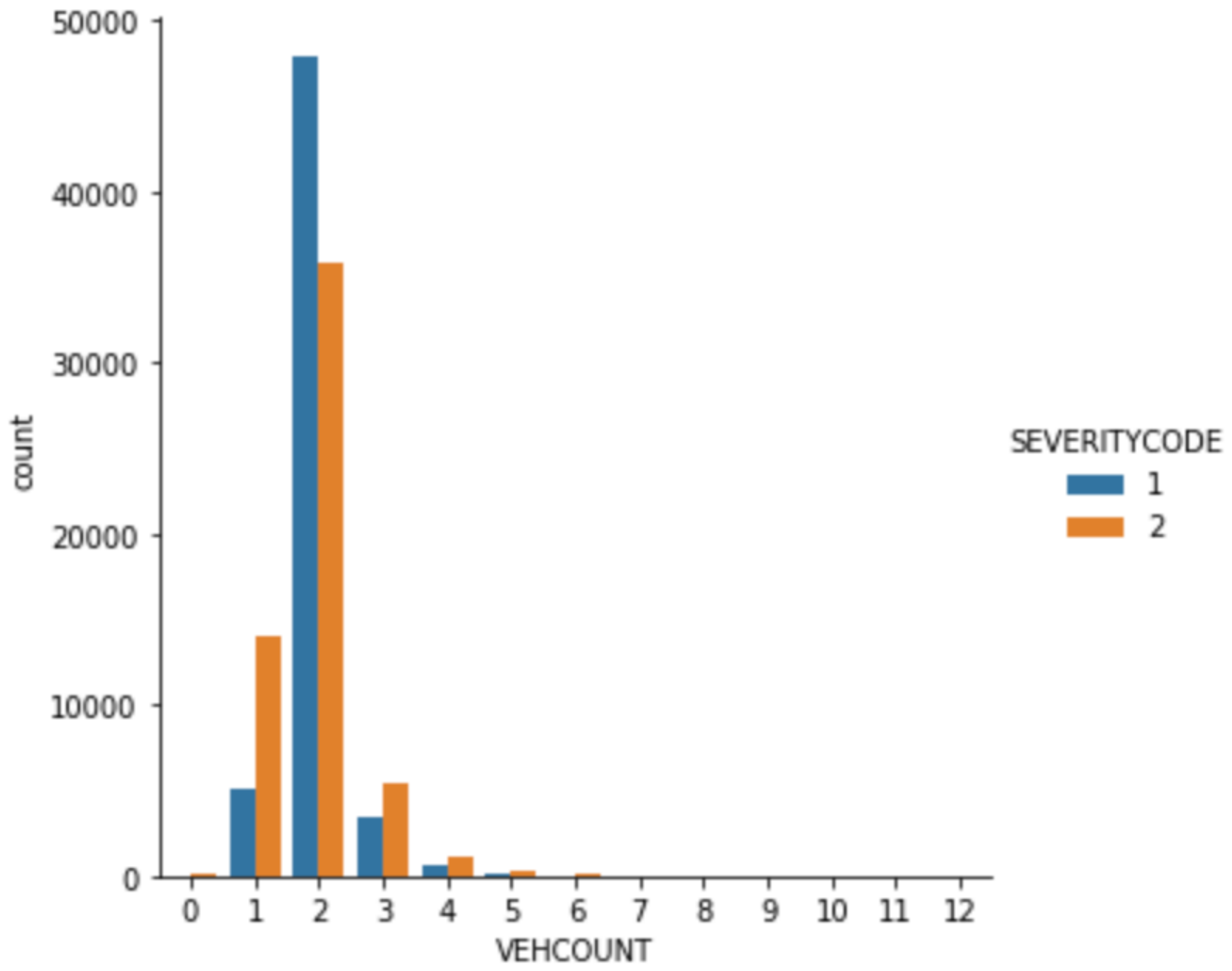


Figure 9. Impact of vehicle count on severity

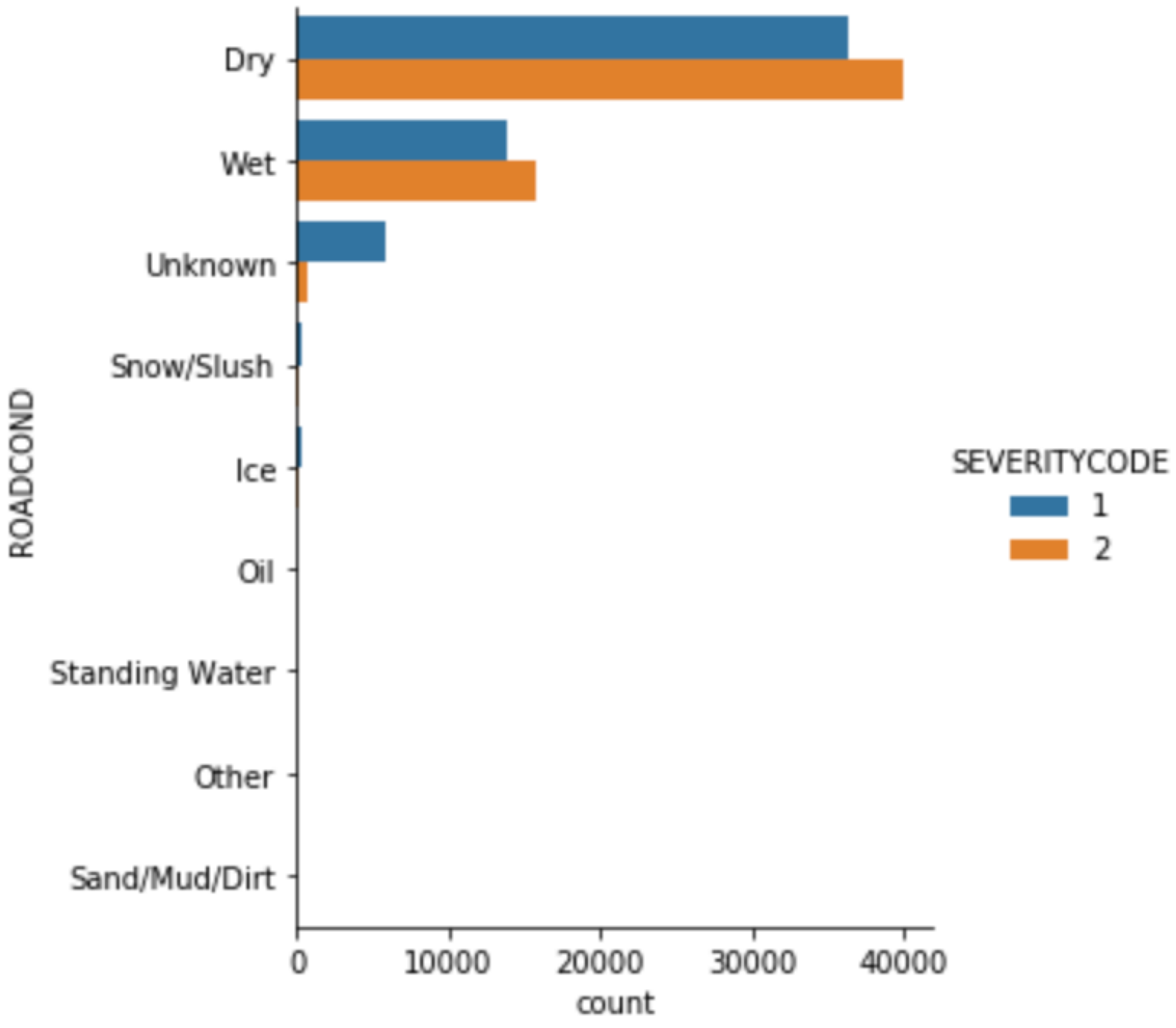


Figure 6. Impact of road condition on severity

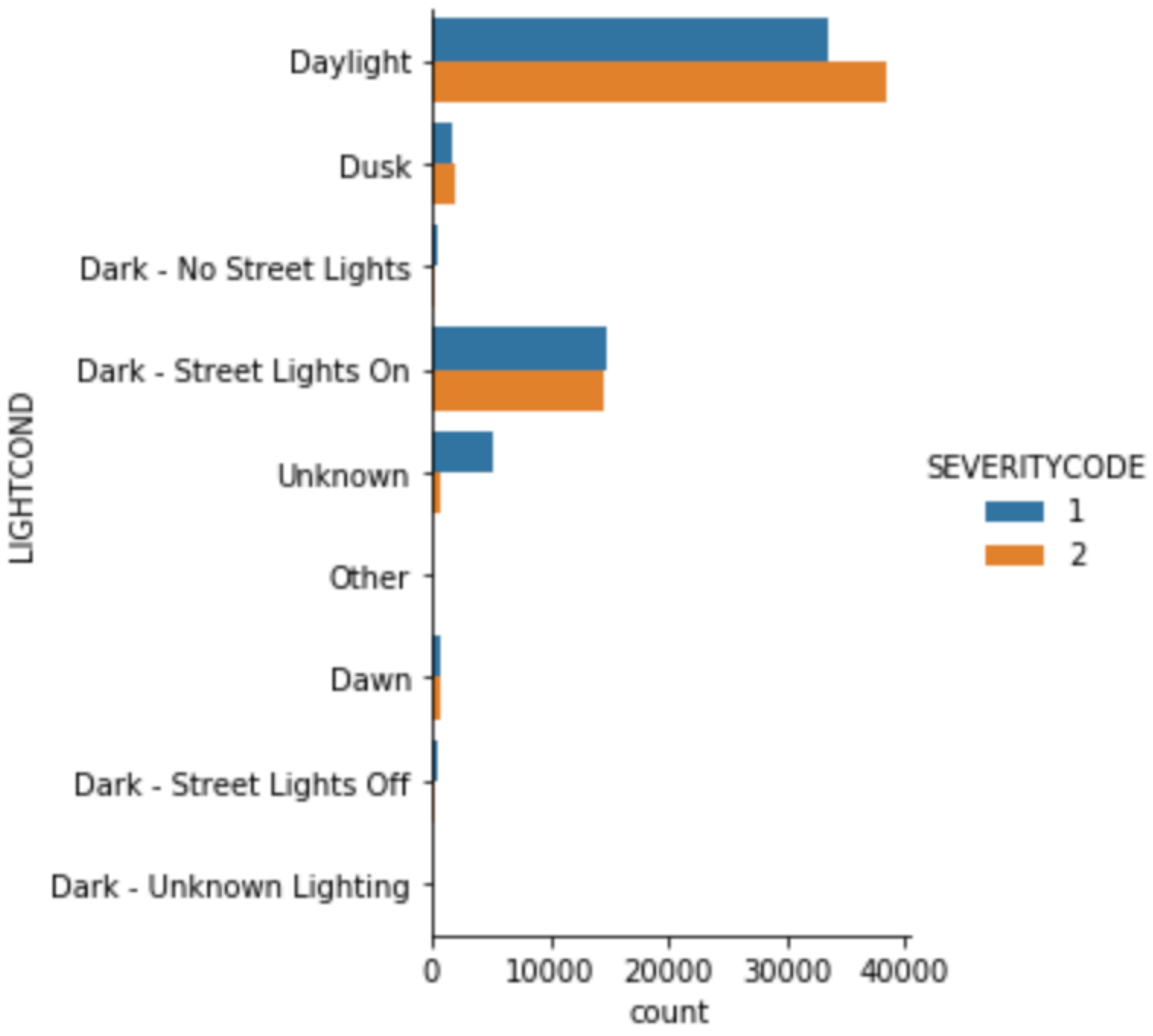
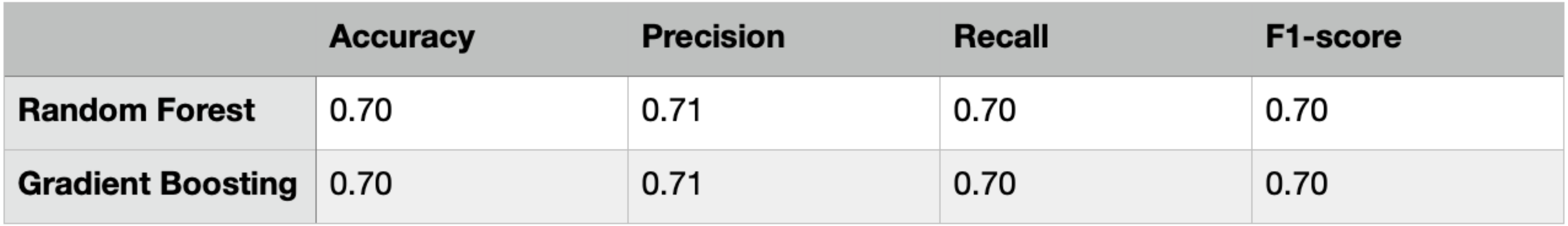


Figure 7. Impact of light condition on severity

Classification Models

Performance

- No significant difference between the performance of Random Forest and Gradient Boosting model

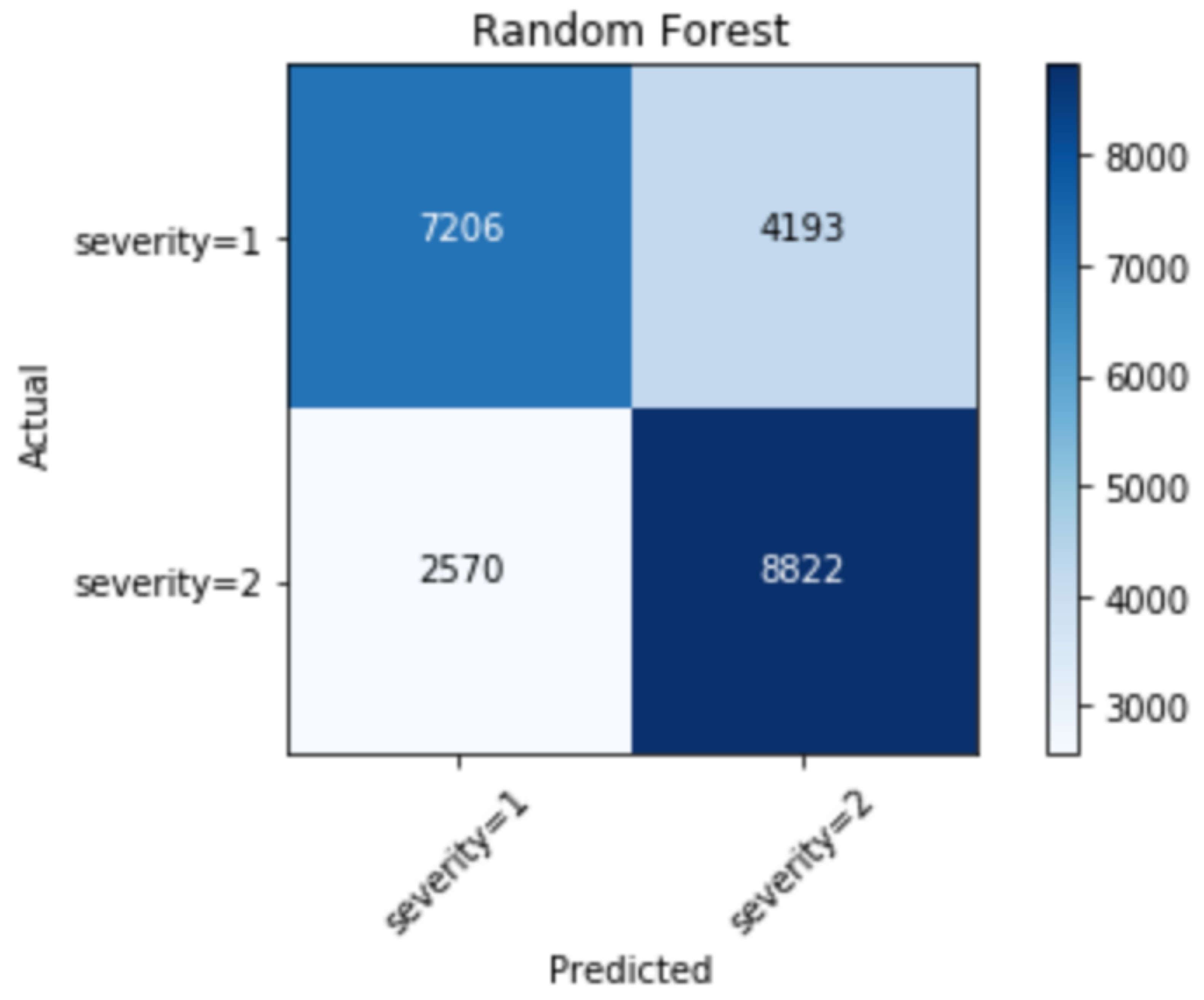


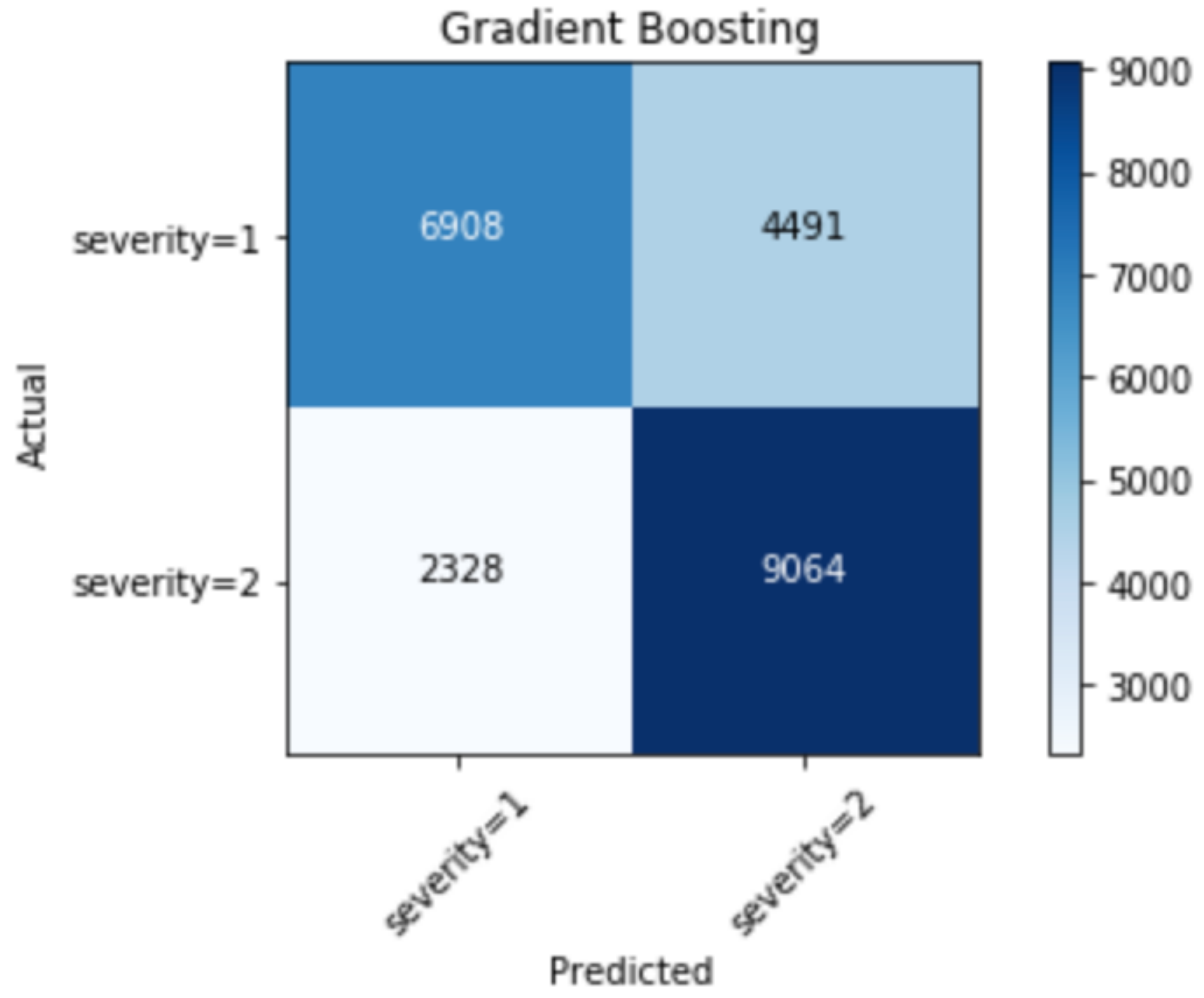
	Accuracy	Precision	Recall	F1-score
Random Forest	0.70	0.71	0.70	0.70
Gradient Boosting	0.70	0.71	0.70	0.70

Classification Models

Confusion Matrix

- Random Forest is better in predicting severity = 1
- Gradient Boosting is better in predicting severity = 2
- Overall, both models have equal balance between precision and recall

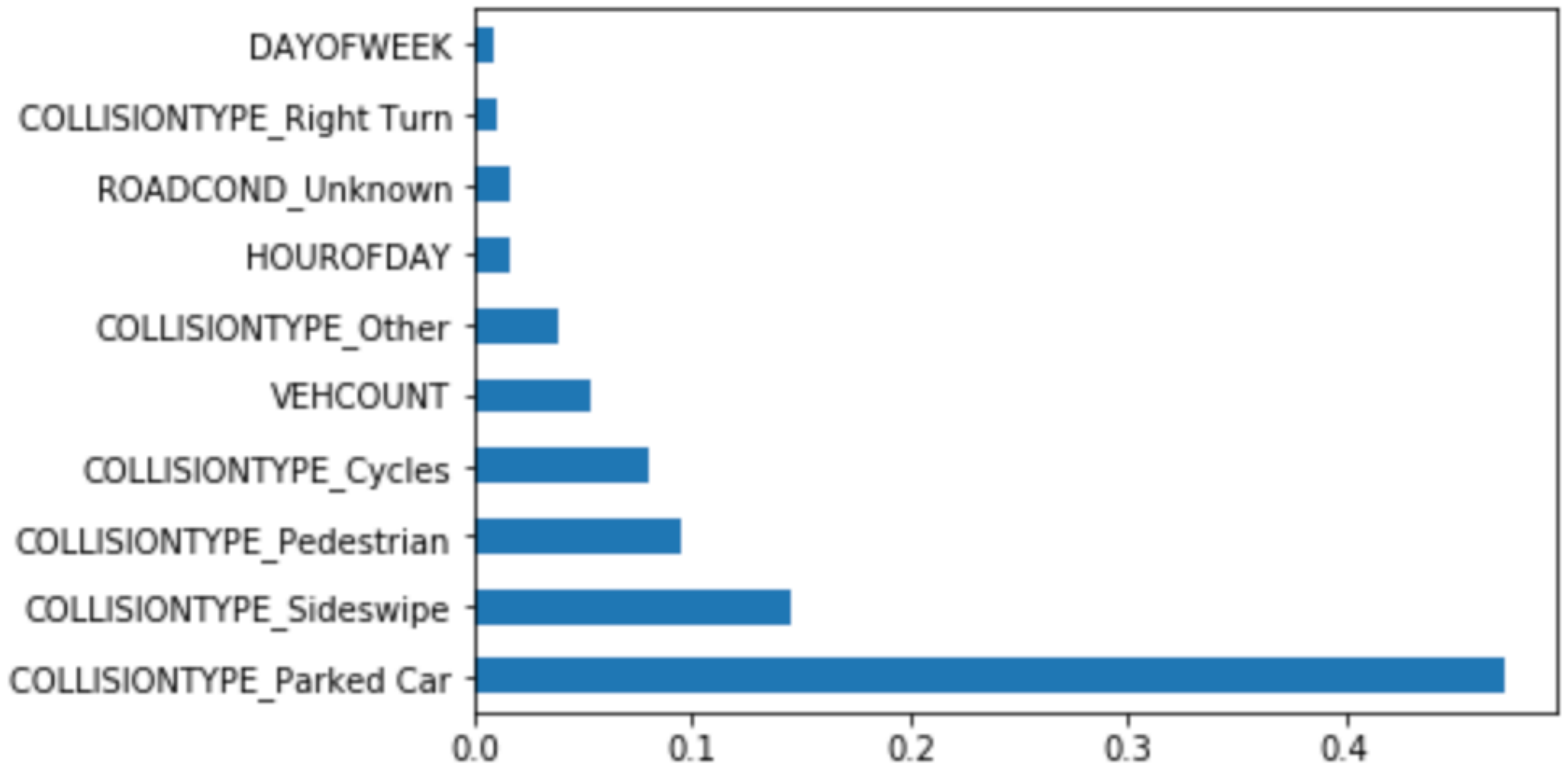




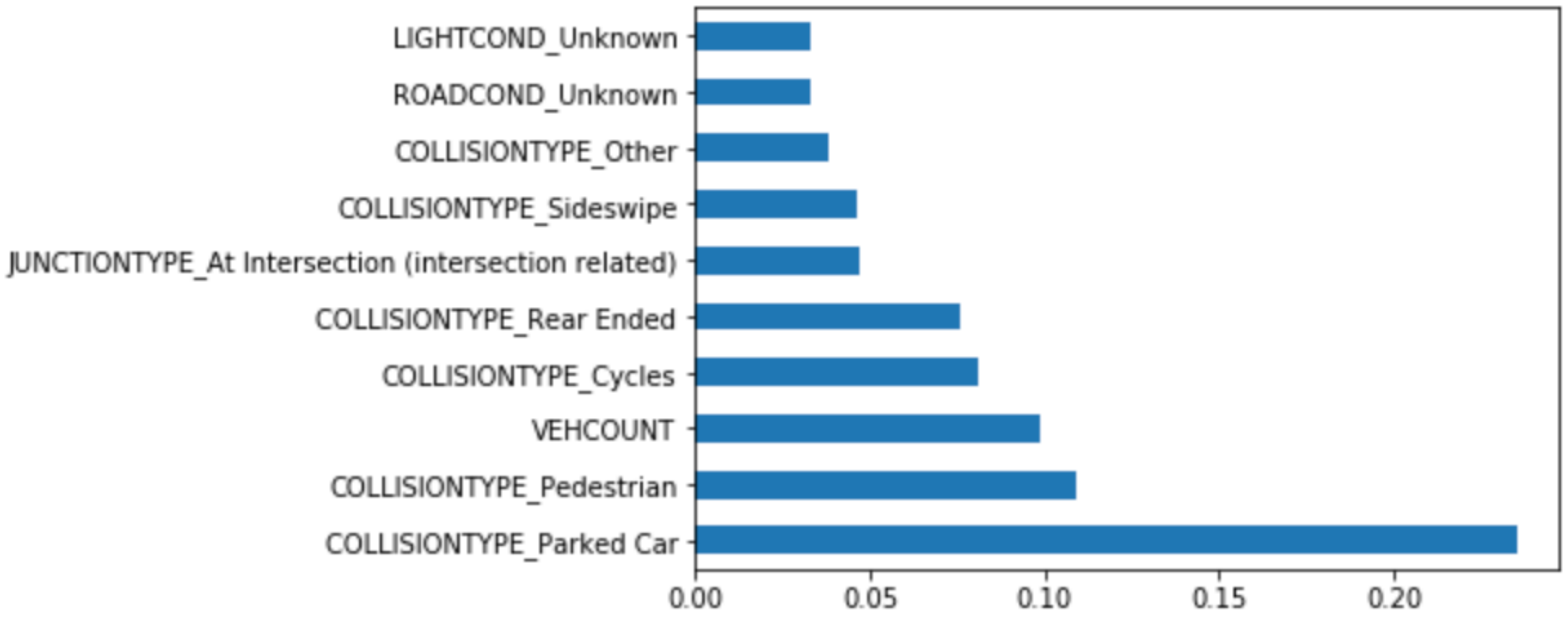
Classification Models

Feature Importance

- Both models ranked collision type and vehicle count as the most important features



Feature importance - Gradient Boosting



Feature Importance - Random Forest

Conclusions

- Results are inconclusive, as both models performed similarly
- Though models performed decently, accuracy can still be improved
- Use more advanced machine learning algorithm to improve performance
- Vehicle count and collision type were identified as the most important features in predicting accident severity