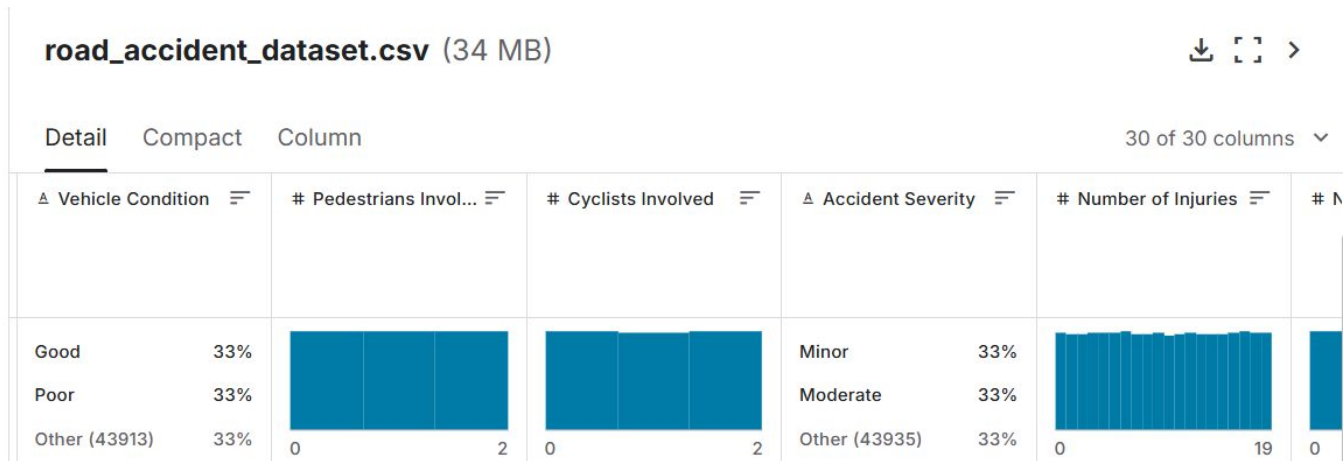


A Multivariate Analysis of Road Accident Severity: Predictive Modeling and Variable Importance

By: Marcellus Mwangi

INTRO TO INITIAL DATASET

- Synthetic dataset led to inaccurate predictions



ALTERNATIVE DATASET

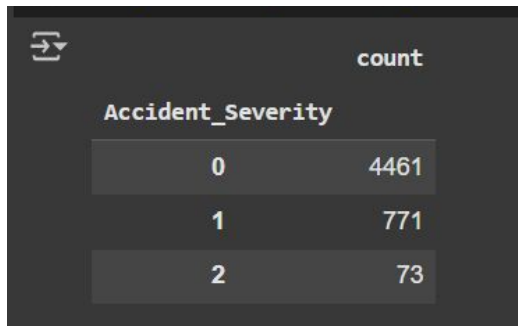
- **Size:** 5305 rows, 23 columns
- **Outcome Variable:** Accident Severity
 - 0:Slight , 1: Severe , 2:Fatal
- **Exploratory Data Analysis:** Multicollinearity, insignificant variables with too many categories, Standardizing numeric variables, One-hot encoding categorical variables.
- **Size after encoding:** 74 columns and 5305 rows



Variable	Description	Variable Type
Road Type	The type of road the accident occurred on	Categorical
Time	The time the accident happened	Numerical
Speed Limit	The speed limit of the area the accident happened	Numerical
Year	The year the accident happened	Numerical
Urban or Rural Area	Was the area urban or rural	Categorical
Number of Vehicles	The number of vehicles involved in the accident	Numerical
Day of week	The day of the week the accident happened	Categorical
Road Surface conditions	The condition of the road the accident occurred on	Categorical
Latitude	The latitude of the accident location	Numerical
Weather conditions	What the weather condition was when it occurred	Categorical

Imbalance of the Dataset

- **Data Imbalance:** occurs when the distribution of classes in a dataset is skewed, with one class significantly outnumbering the others
- Accident data is generally imbalanced because there are very few fatal accidents that happen due to improved vehicle safety features
- May cause very biased predictions to the over represented class




Accident_Severity	count
0	4461
1	771
2	73

Dealing with the Imbalance


- **Synthetic Minority Oversampling Technique(SMOTE):** a machine learning technique used to address class imbalance in datasets
- How SMOTE works:
 - Identify the minority class
 - Find nearest neighbors: For each instance in the minority class, SMOTE finds its k-nearest neighbors within the minority class.
 - Generate synthetic samples
 - Balanced dataset is the result
- Only applied SMOTE to training data



APPLICATION OF SMOTE



	count
Accident_Severity	
0	4461
1	771
2	73



	count
Accident_Severity	
1	3116
0	3116
2	3116

dtype: int64

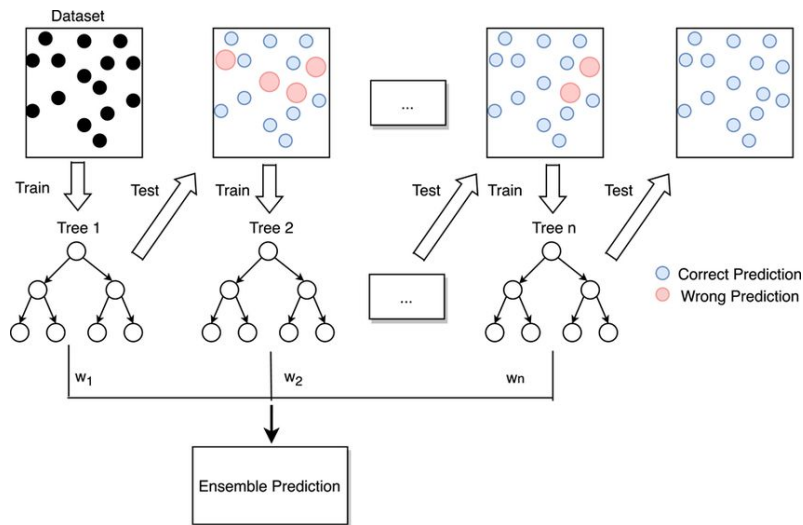
Models

Model	slight	Severe	Fatal	Overall Accuracy
Random Forest	0.88%	0.12%	0%	76.1%
Multinomial logistic regression	0.91%	0.09%	0%	78.8%
Ordinal Logistic Regression	0.87%	0.19%	0%	75.2%
Neural Network	0.87%	0.17%	0%	76.44%
Gradient Boosting Machine	0.84%	0.15%	0.13%	77.8%

- I used a 5-fold cross validation to find the best parameters for random forest, neural network and GBM

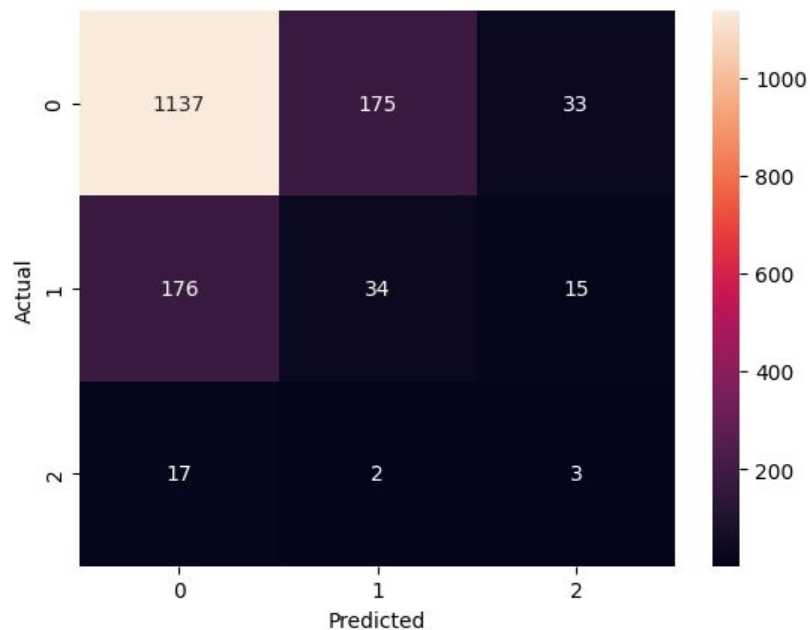
Gradient Boosting Machine

- Gradient Boosting Machines (GBMs) are powerful machine learning algorithms that build models sequentially, with each new model focusing on correcting the errors of the previous models

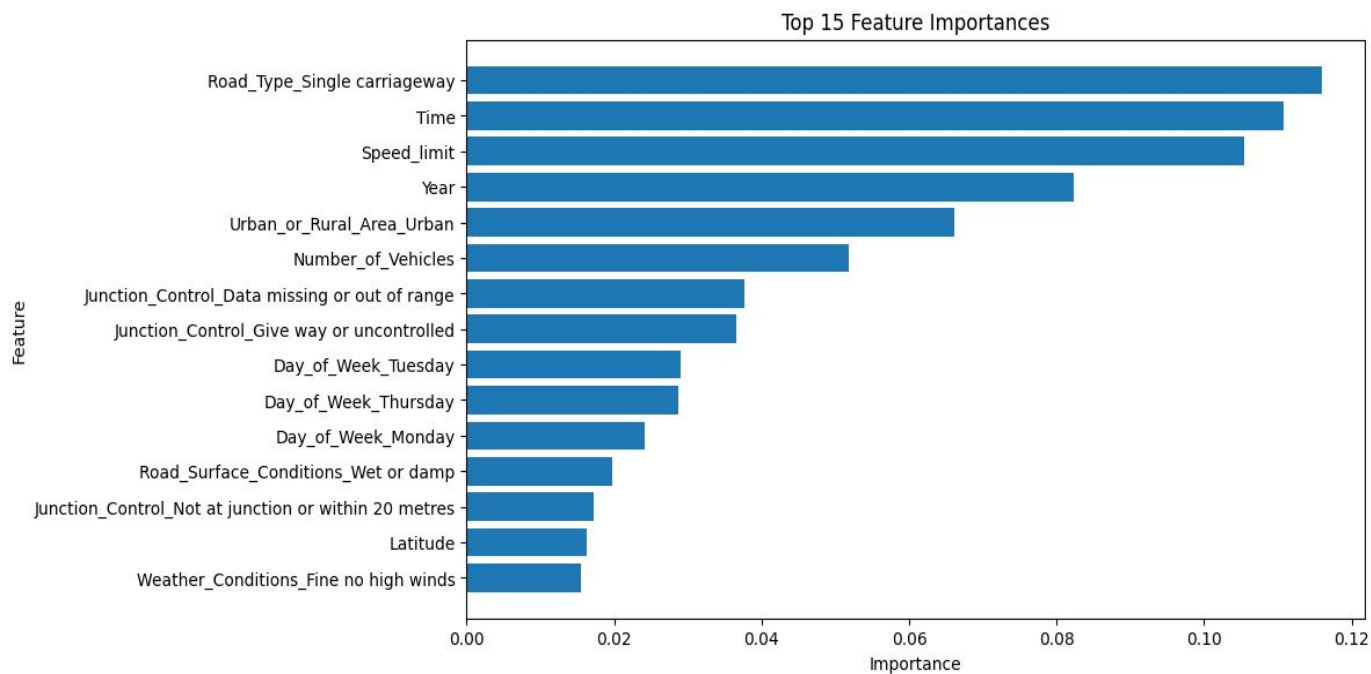


GBM Results

- Model accuracy: 73%
- There is still imbalance in the predictions of the model



Feature Importance

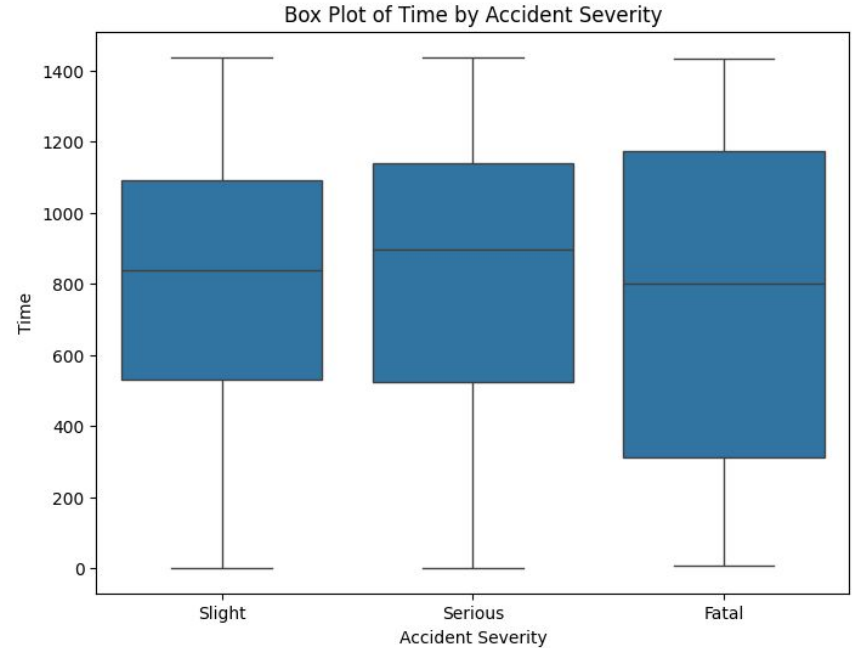
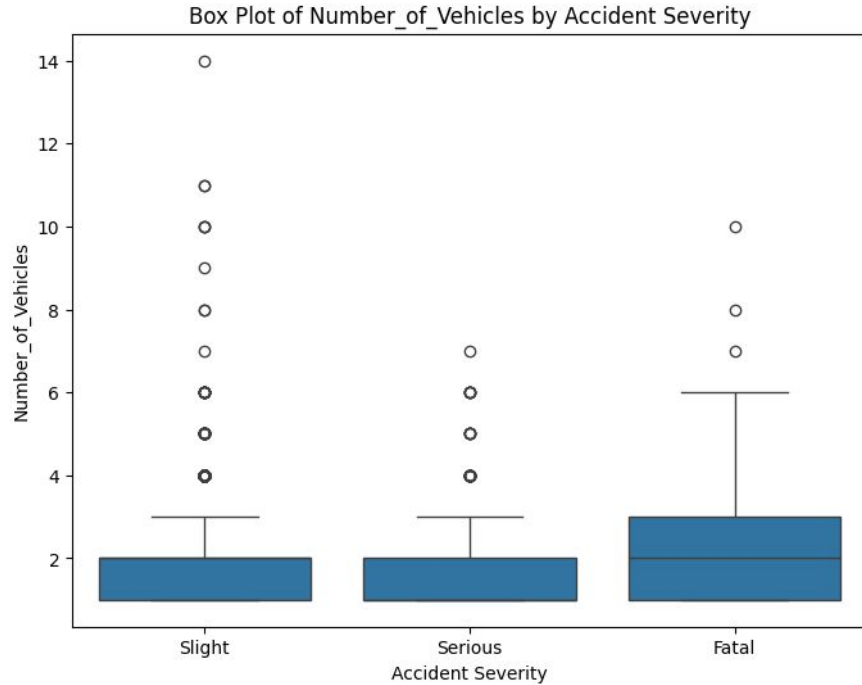


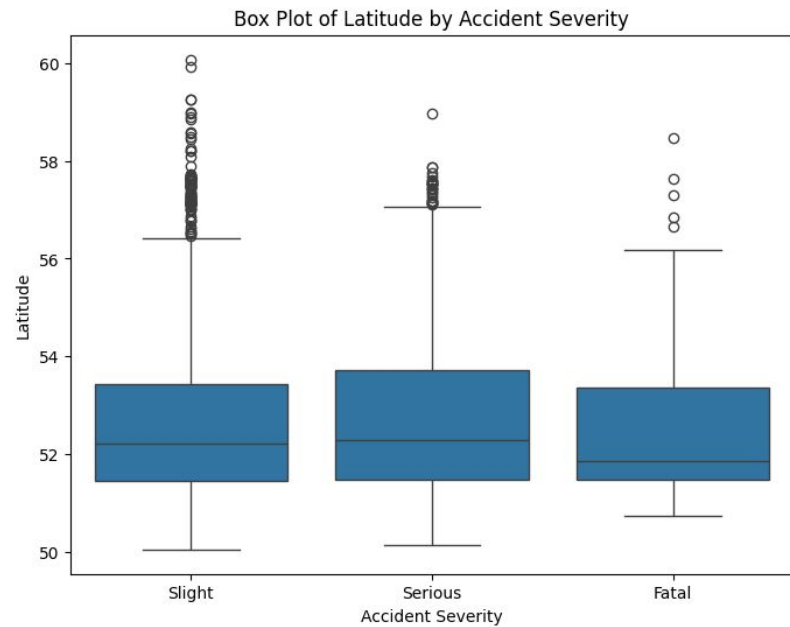
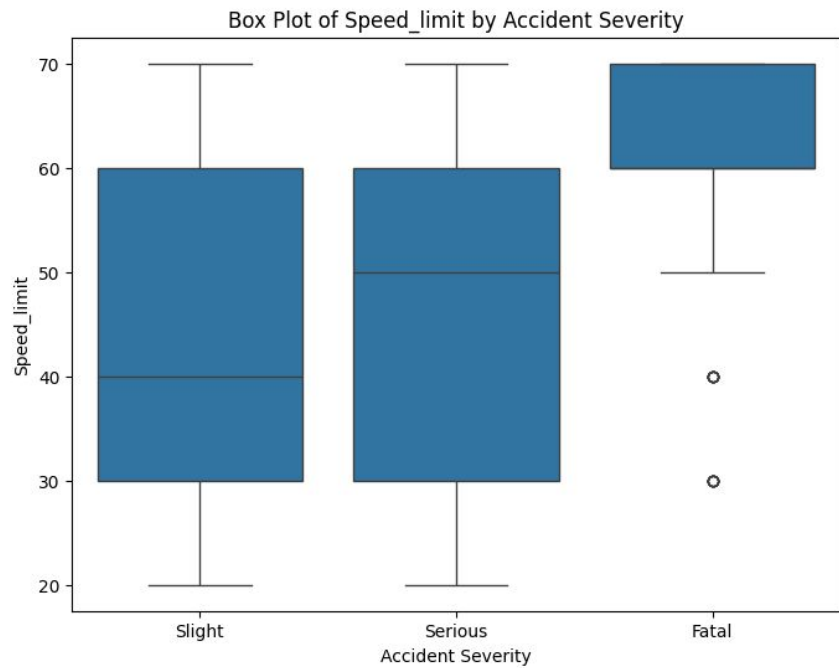
Limitations of prediction with High Imbalance

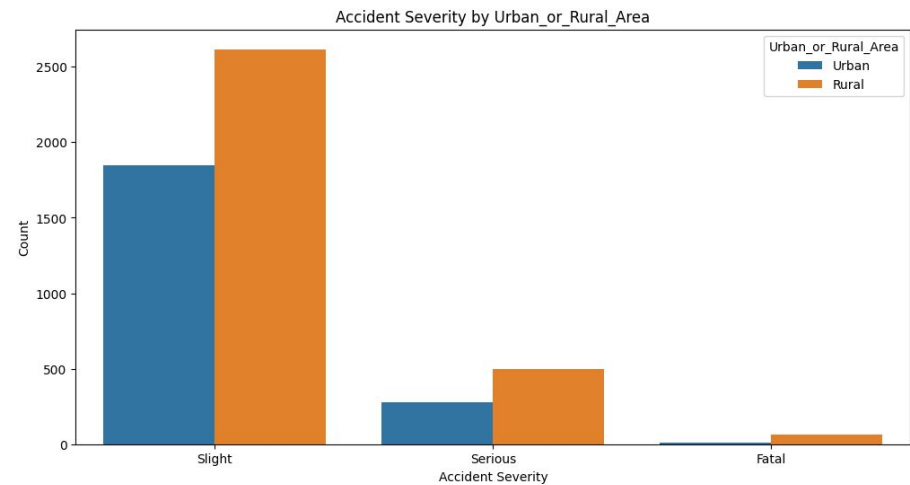
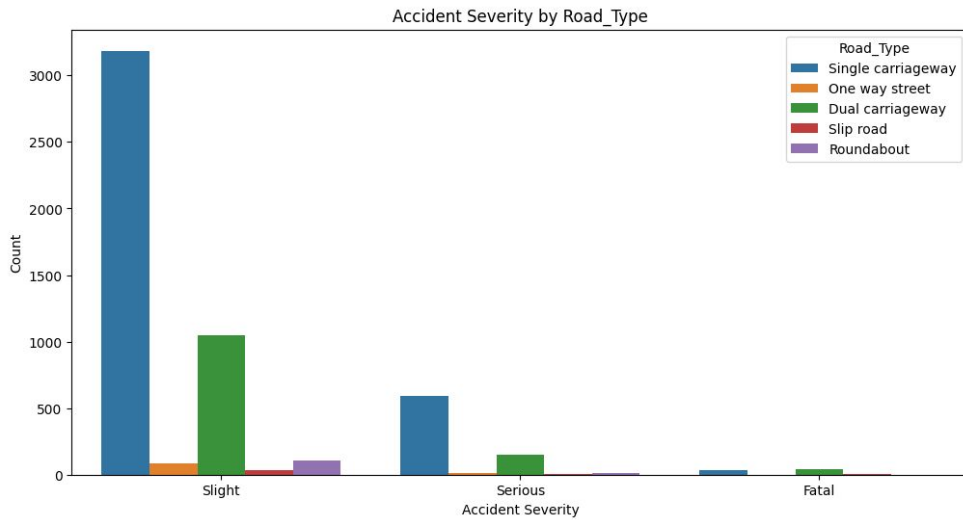
- The high imbalance makes it hard to rely on the results of the models for predictive analysis
- I used the GBM model for exploratory analysis rather than predictions

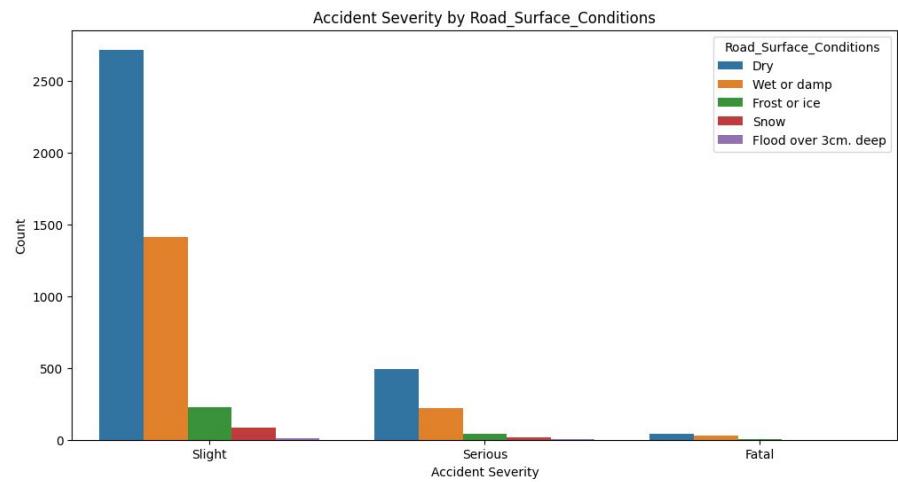
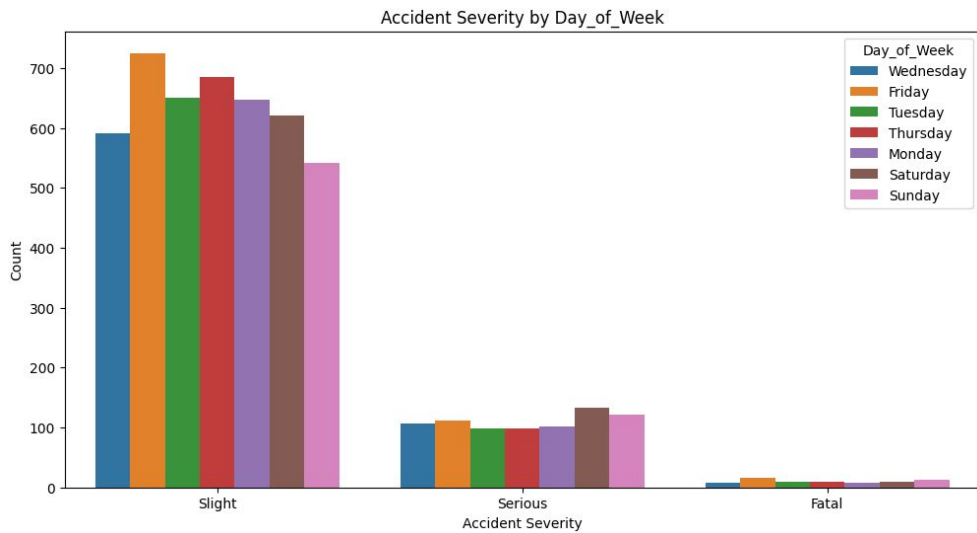


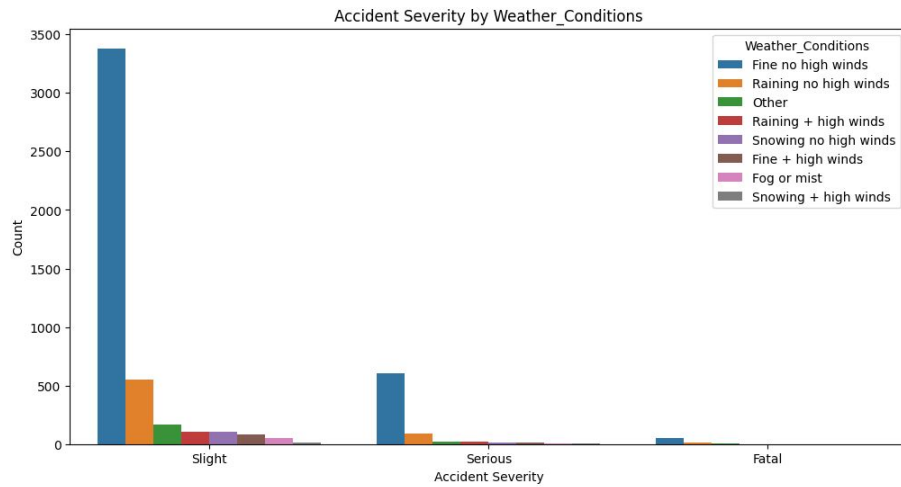
Variable plots











Conclusion

- Accident severity is very imbalanced
- Fitting a prediction model that is accurate for all classes is a challenge
- SMOTE was used to mitigate the issue but did not completely solve it
- Accuracy for each class is important not just overall
- Training model with SMOTE, if model accuracy is still imbalanced model should not be used for predictions



References

- *An introduction to statistical learning.* (n.d.). An Introduction to Statistical Learning. <https://www.statlearning.com/>
- *How to Deal with Imbalanced Datasets with SMOTE algorithm.* (2022, June 10). <https://www.turing.com/kb/smote-for-an-imbalanced-dataset>
- Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7. <https://doi.org/10.3389/fnbot.2013.00021>
- Rennie, J. D. M., Massachusetts Institute of Technology, Srebro, N., & University of Toronto. (n.d.). Loss Functions for Preference Levels: Regression with Discrete Ordered Labels. In *Massachusetts Institute of Technology* [Journal-article].
<https://home.ttic.edu/~nati/Publications/RennieSrebroIJCAI05.pdf>
- Sugatagh. (n.d.). *GitHub - sugatagh/Road-Traffic-Accident-Severity-Classification: The aim of the project is to build prediction models to classify severity of road traffic accidents (slight injury, serious injury or fatal injury) based on various relevant information regarding the involved vehicles, drivers, casualties and surrounding conditions.* GitHub. <https://github.com/sugatagh/Road-Traffic-Accident-Severity-Classification>
- *Variable Importance Plots—An Introduction to the vip Package.* (n.d.). <https://koalaverse.github.io/vip/articles/vip.html>