Predicting Fatal Accidents in UK's Public Roads

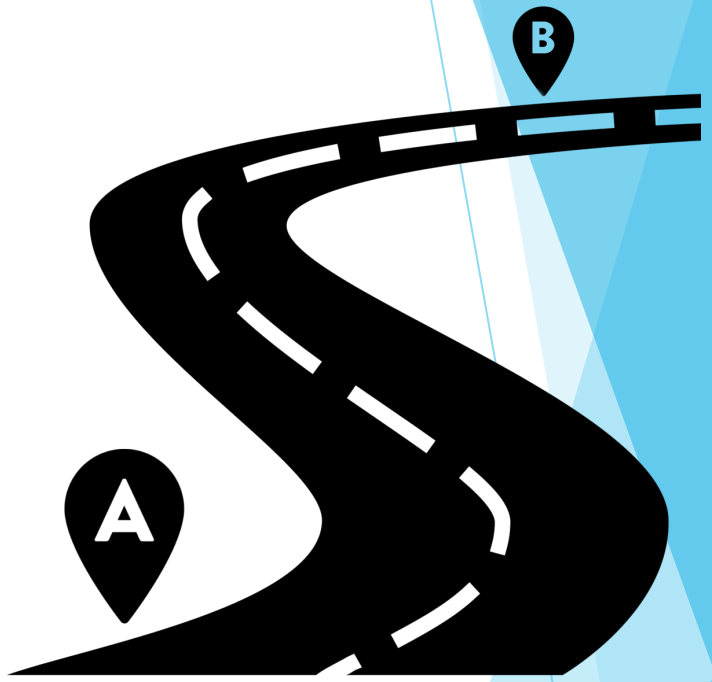
Achyut Kafle

October 1, 2018

Supervised Learning Capstone, Thinkful

Road Map

- Motivation
- Results
- Methods
  - Exploratory data analysis
  - Feature engineering and selection
  - Classification and evaluation
- Conclusions and Next Steps



Picture source

Motivation

- Road accidents claim thousands of lives and hundreds of thousands of slight to permanent injuries each year in UK.

- Serious costs associated with road accidents:
  - Human lives lost;
  - Healthcare costs of injured;
  - Economic and psychological costs of lost time due congestions following accidents

- Knowing what factors cause or lead to accidents and being able to predict future accidents based on observed data can help prevent or minimize the costs by implementing policies to address those factors.

Results

- Factors that contribute most to predicting fatal accidents are:
  - Number of vehicles involved, location (lat/longs), urban, etc.
- Support vector classifier performed the best on baseline hyperparameters.
- Random forest classifier improved the most after tuning hyperparameters.
- Boosting classifiers generally performed better than the bagging or stacking classifiers and extreme gradient boosting performed the best.

Methods: Data

Picture source

- Road safety data from 2016 on personal injury accidents on UK's public roads that are reported to the police and subsequently recorded from the UK Department for Transport and includes:
  - Accident severity (Slight, Serious, Fatal)
  - Number of vehicles
  - Number of casualties
  - Location (Latitude and Longitude; Northing and Easting)
  - Time (Hour, day and month)
  - Road surface conditions
  - Light conditions
  - Weather conditions
  - Road types and junction types

Picture source

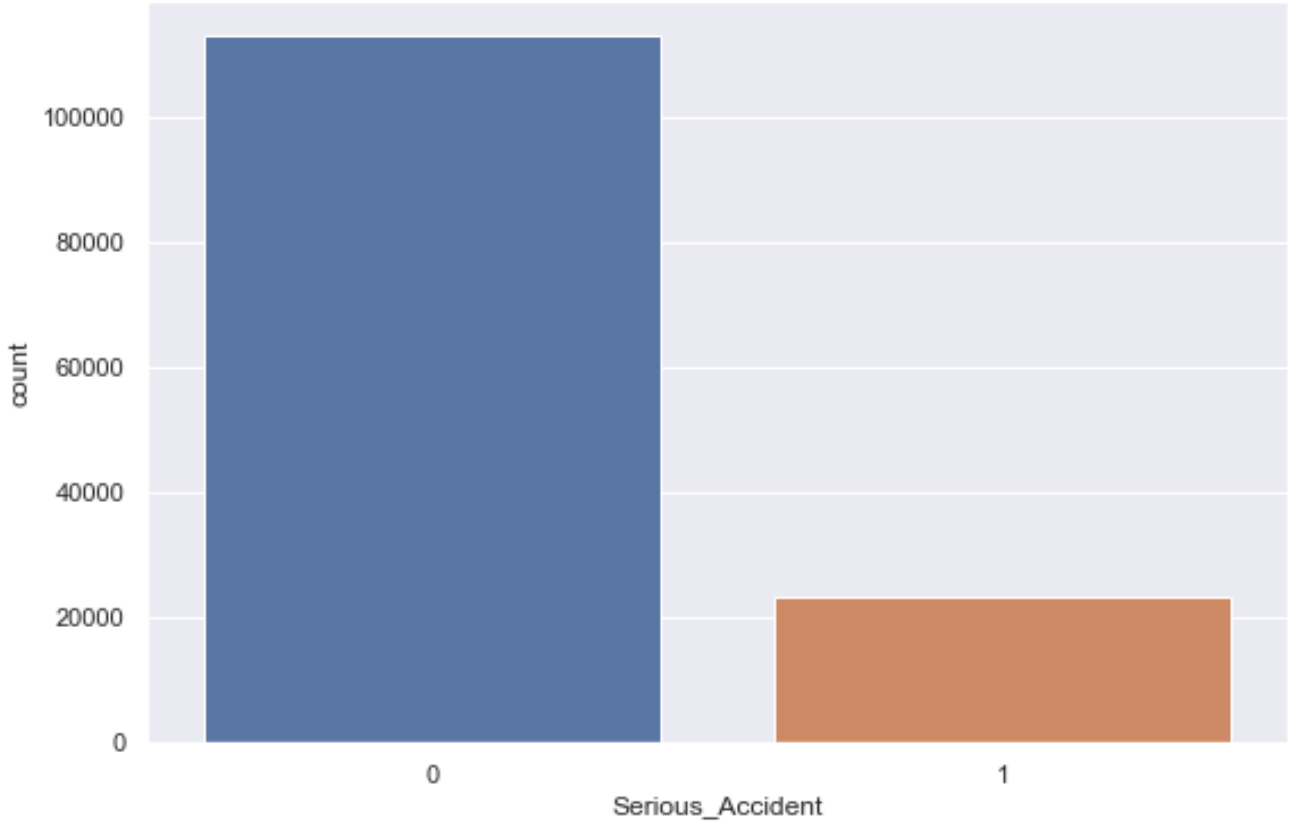
Picture source

Picture source

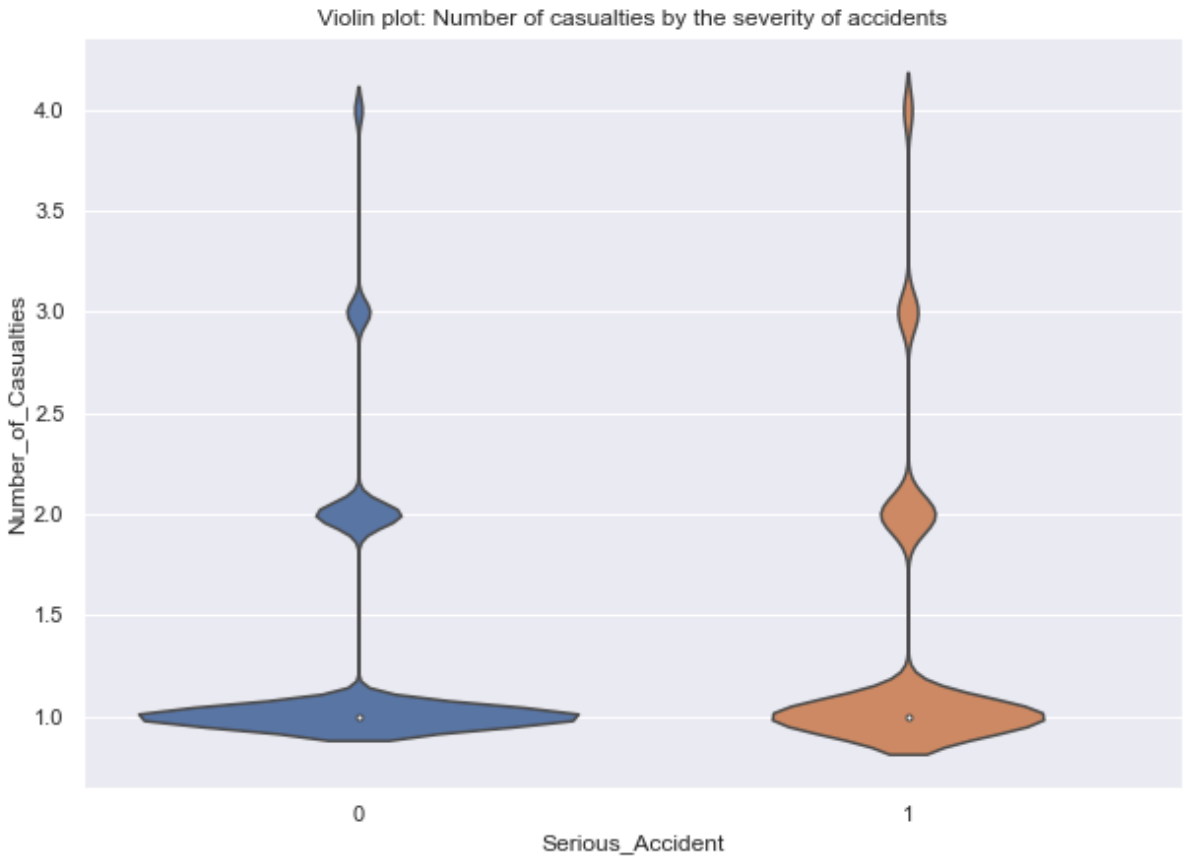
Picture source

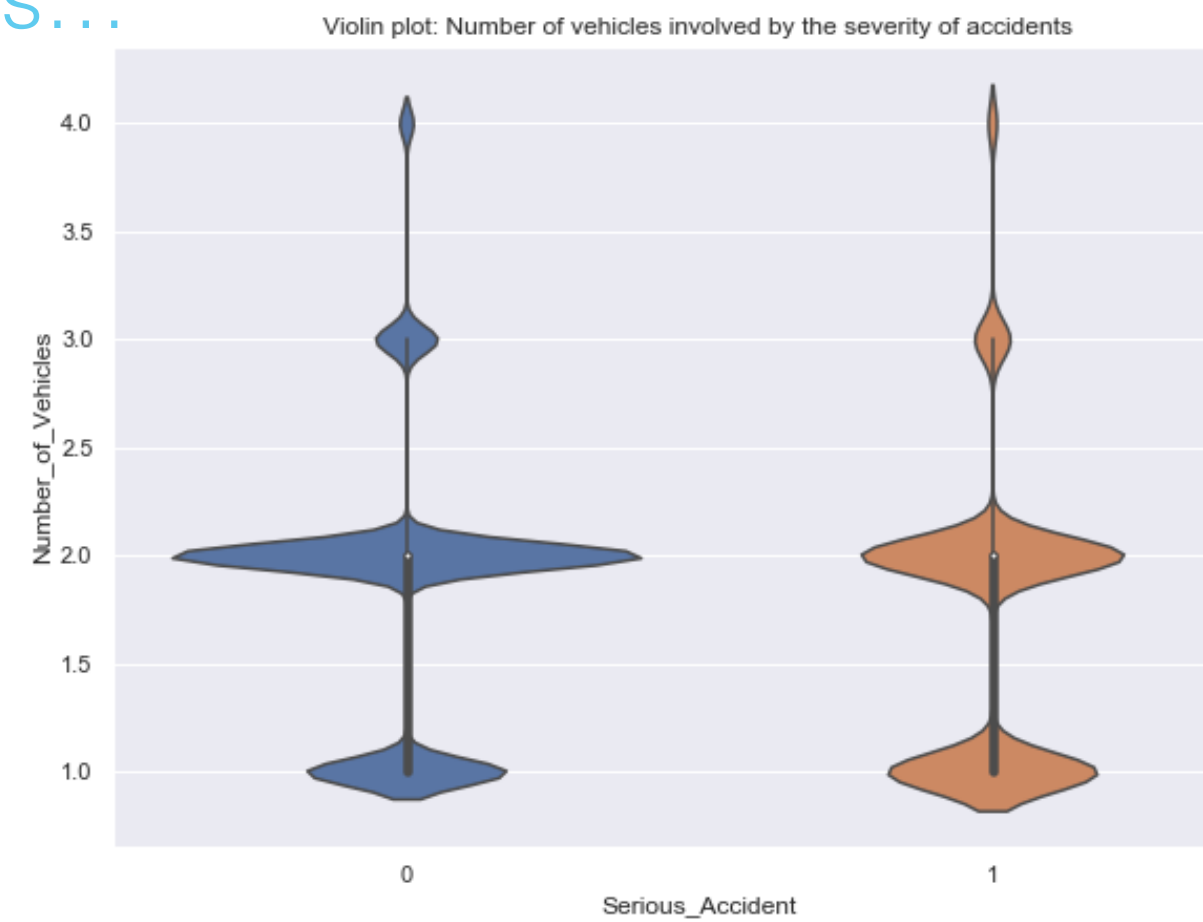
Methods: Exploratory data analysis

- **Missing data:**
  - Dropped few hundreds missing data points on few relevant features
- **Outcome variable of interest:**
  - Fatal/serious accidents
  - 133,000+ unique accidents of which 17% are fatal/serious



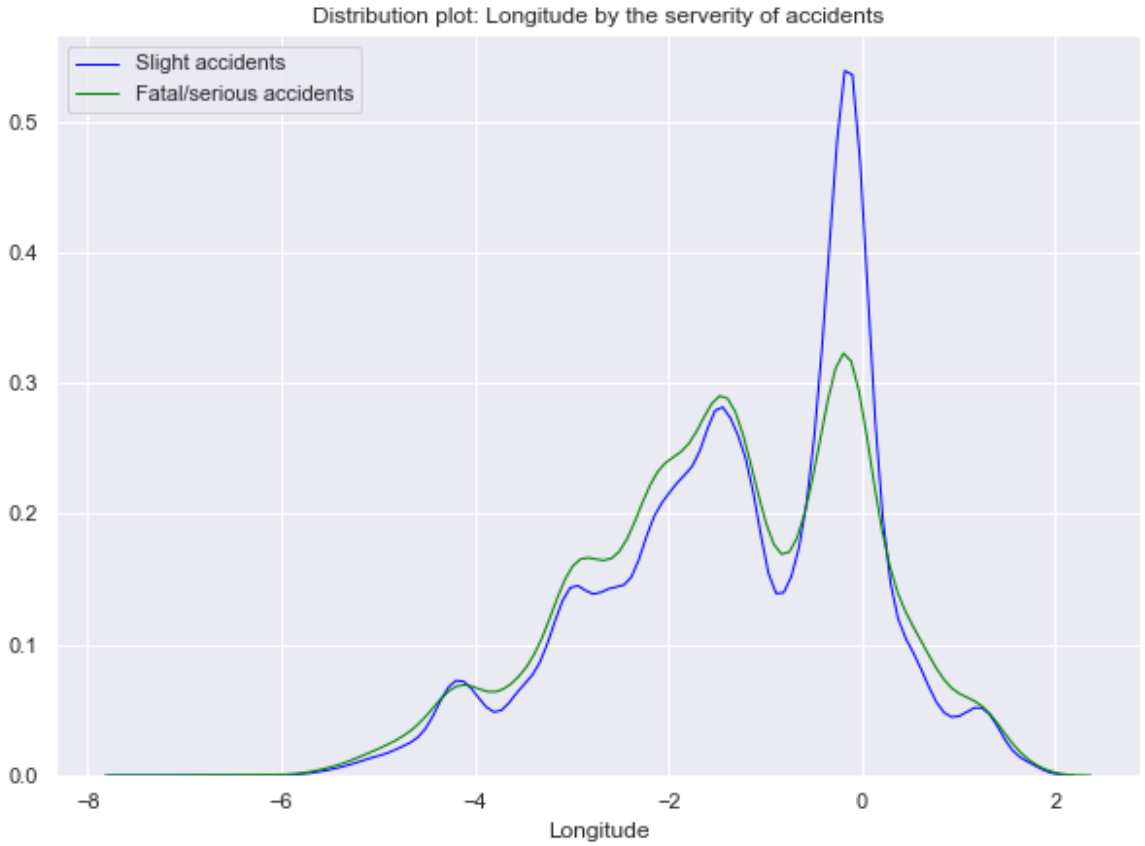
Methods: Exploratory data analysis…

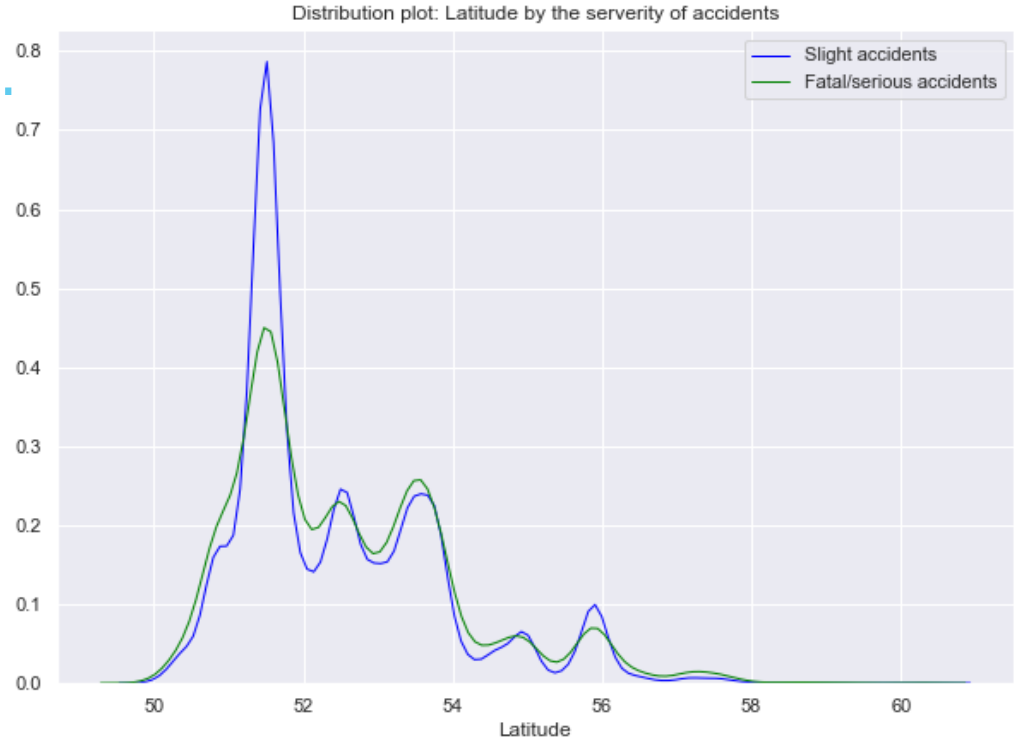




- **Numerical features:**
  - Number of vehicles;
  - Number of casualties

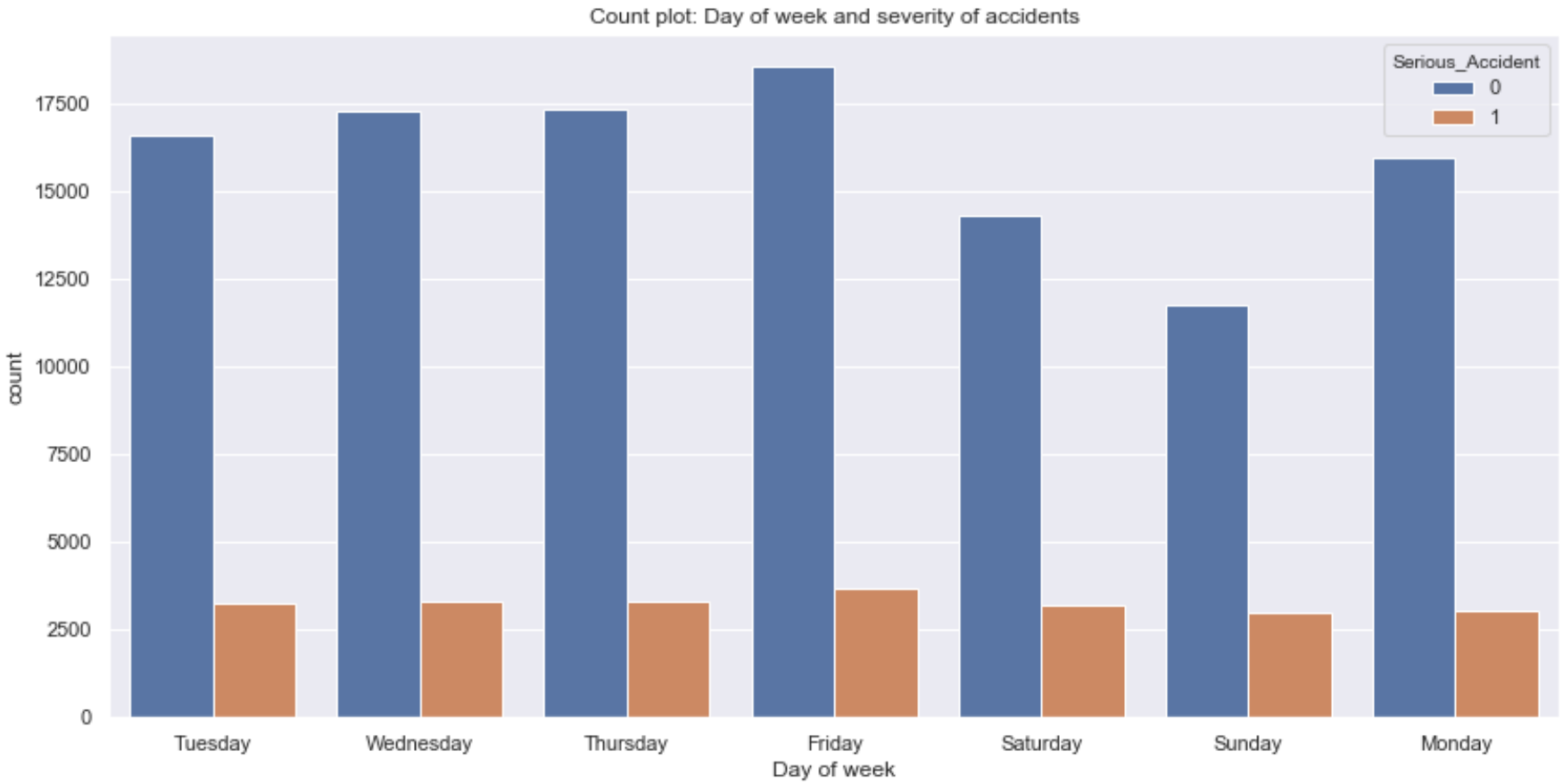
Methods: Exploratory data analysis…





- **Numerical features:**
  - Latitude
  - Longitude

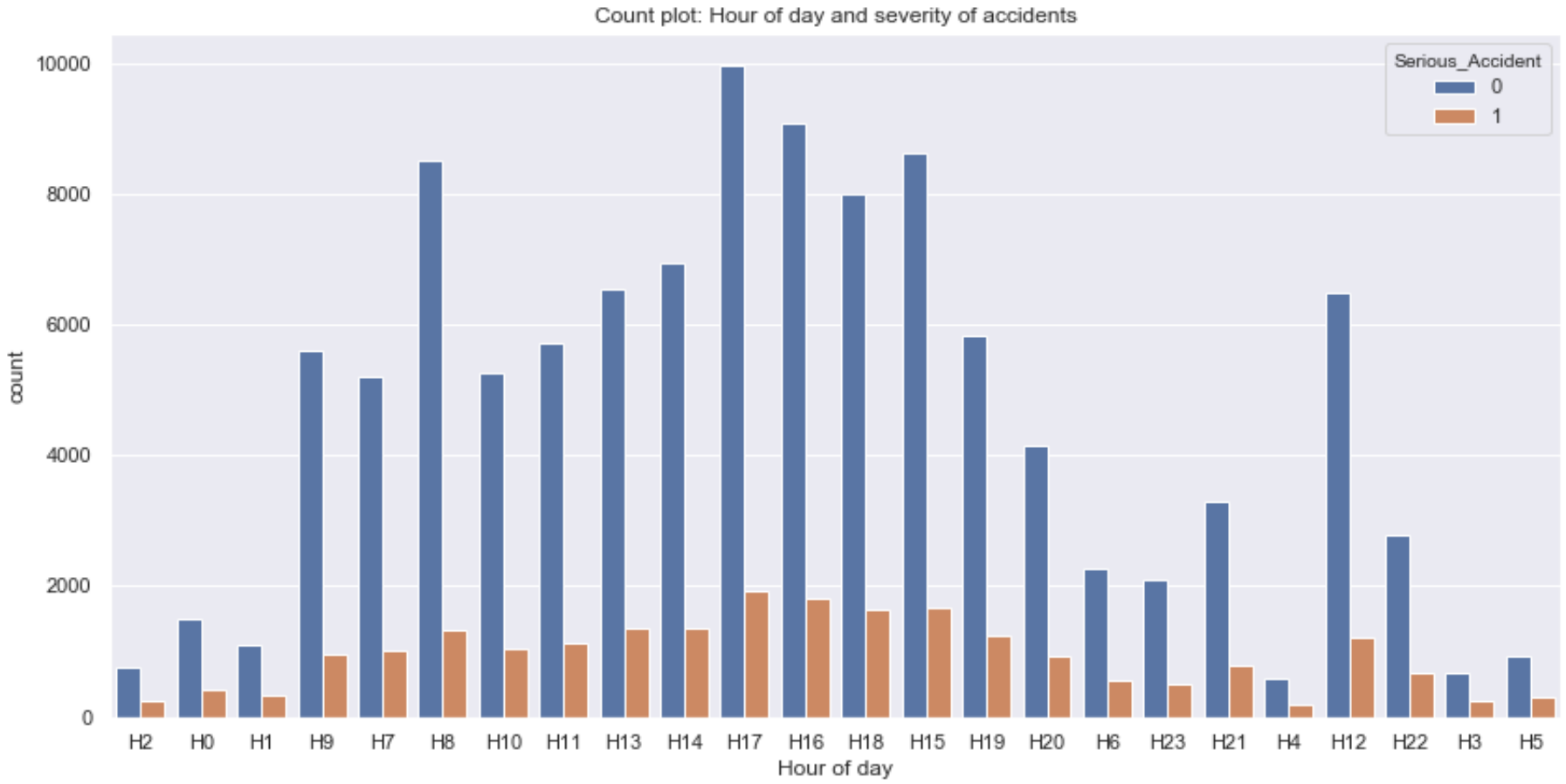
Methods: Exploratory data analysis…



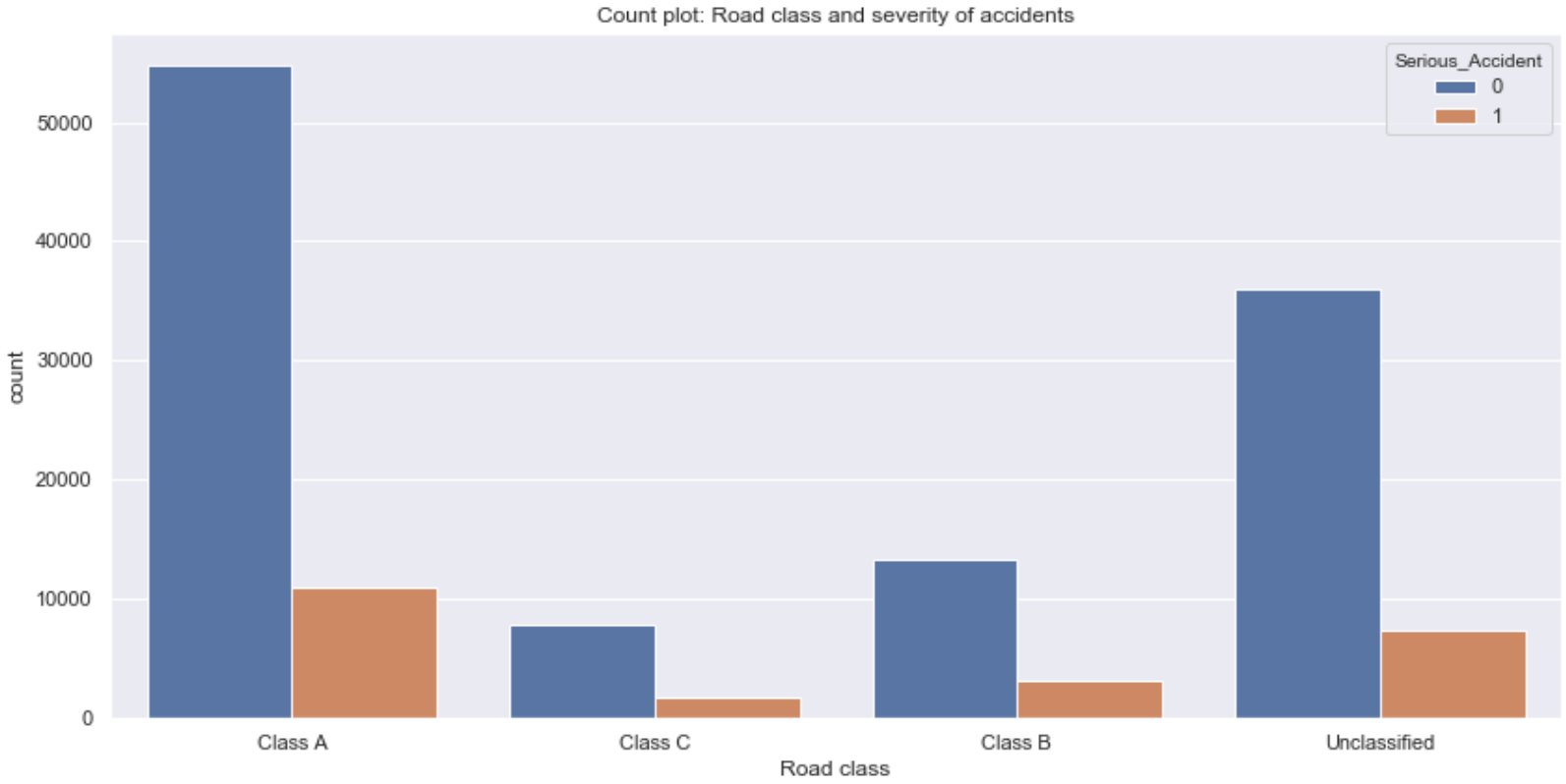
Least number of accidents on weekends and the most on Fridays

Methods: Exploratory data analysis…

- Accidents peak at multiple times: 3-6 pm and 8 am



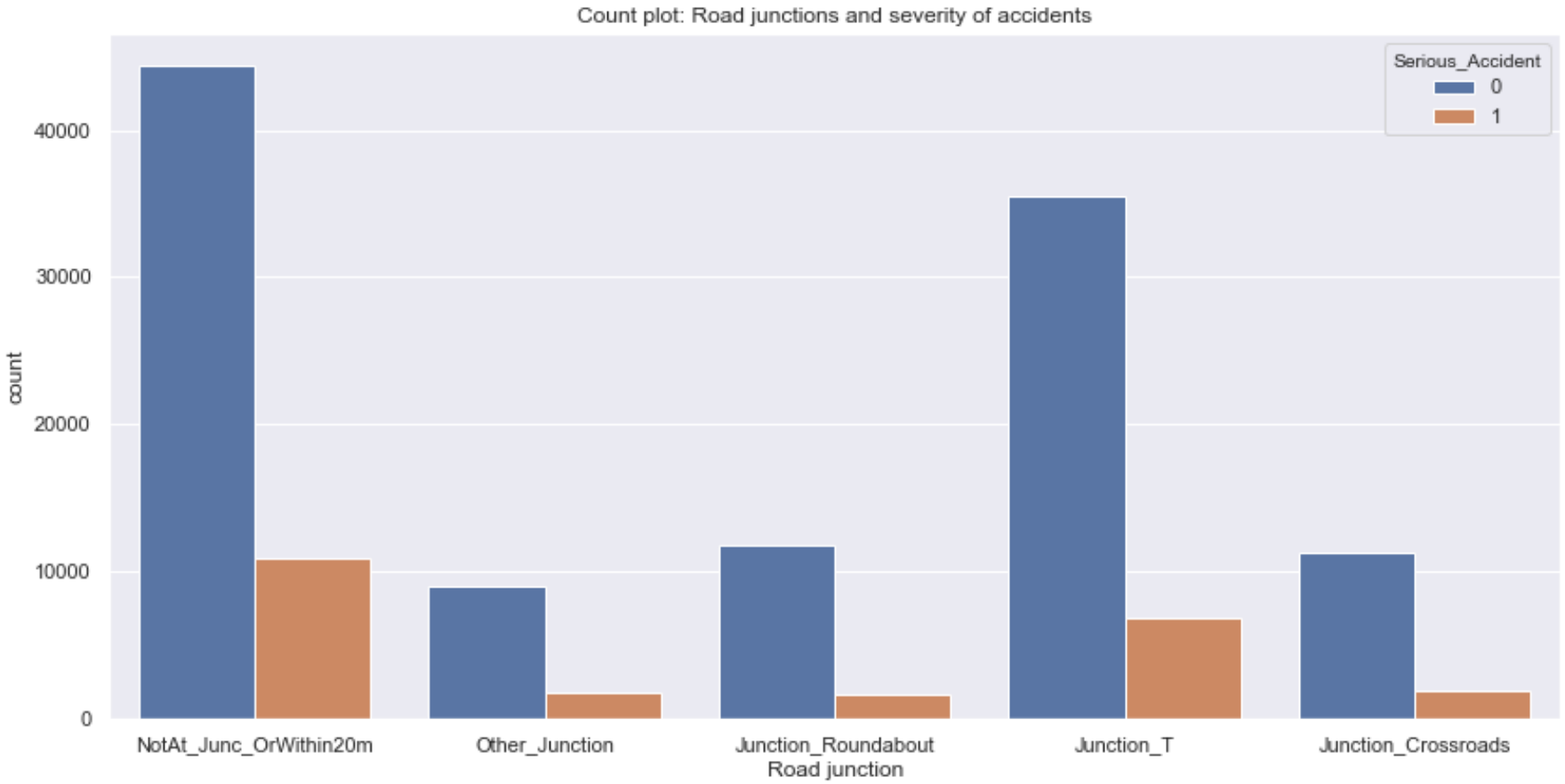
Methods: Exploratory data analysis…



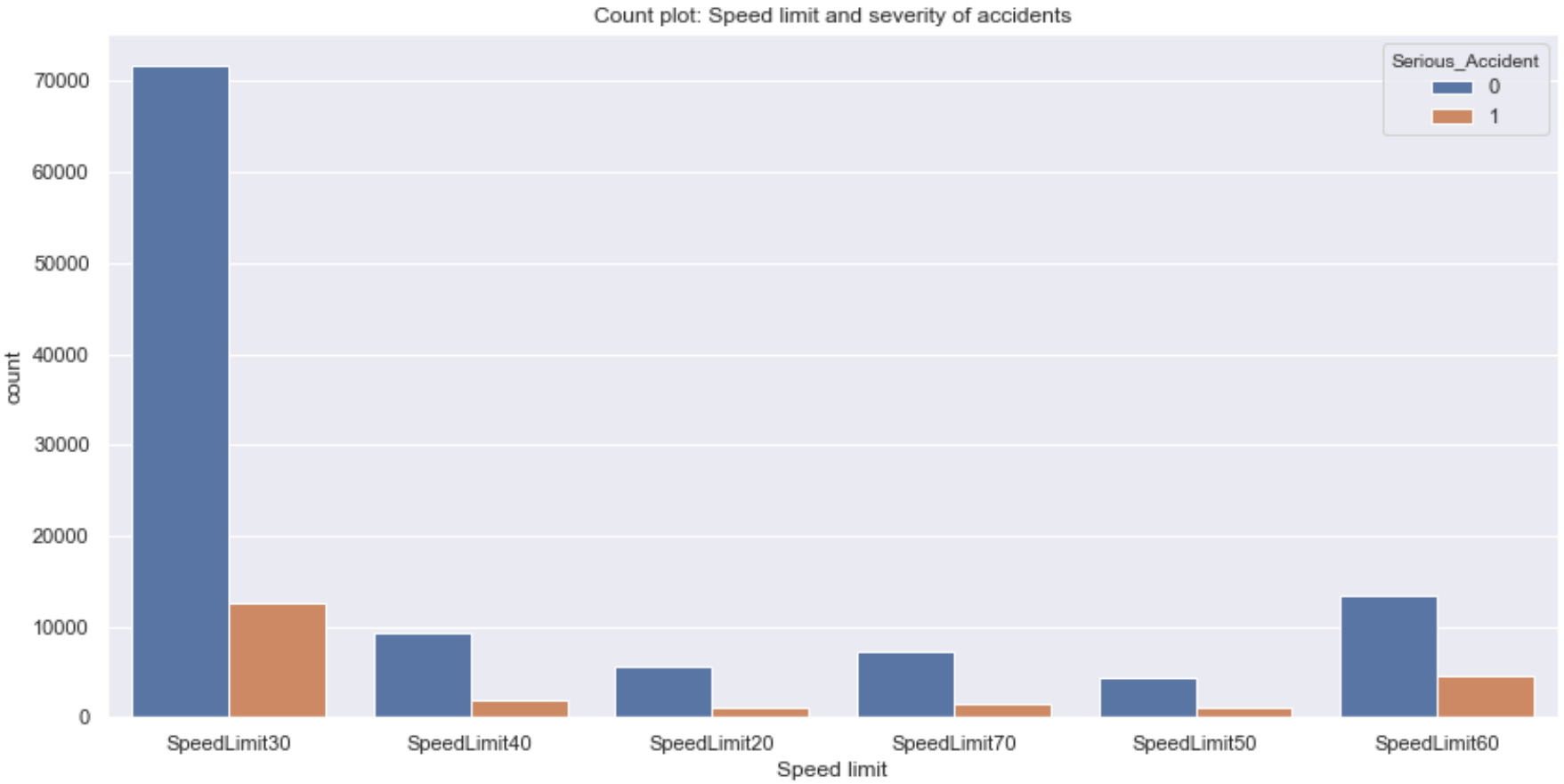
- More accidents happened on class A roads which are primary routes

Methods: Exploratory data analysis…

- More accidents happened not at junctions or within 20m of junctions followed by T junctions



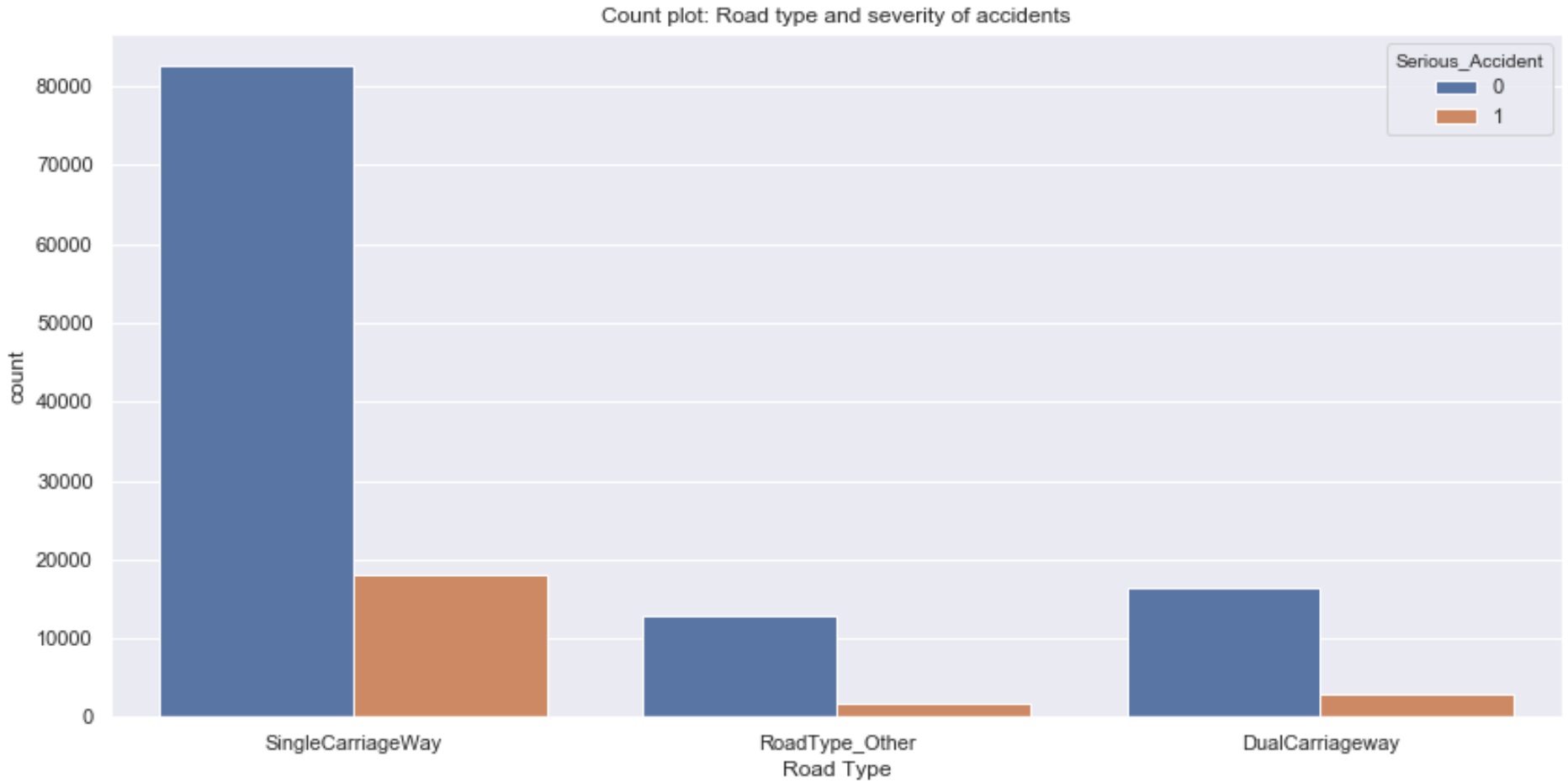
Methods: Exploratory data analysis…



A majority of accidents happened on roads with 30mph speed limit

Methods: Exploratory data analysis…

- A majority of accidents happened on single carriageway roads

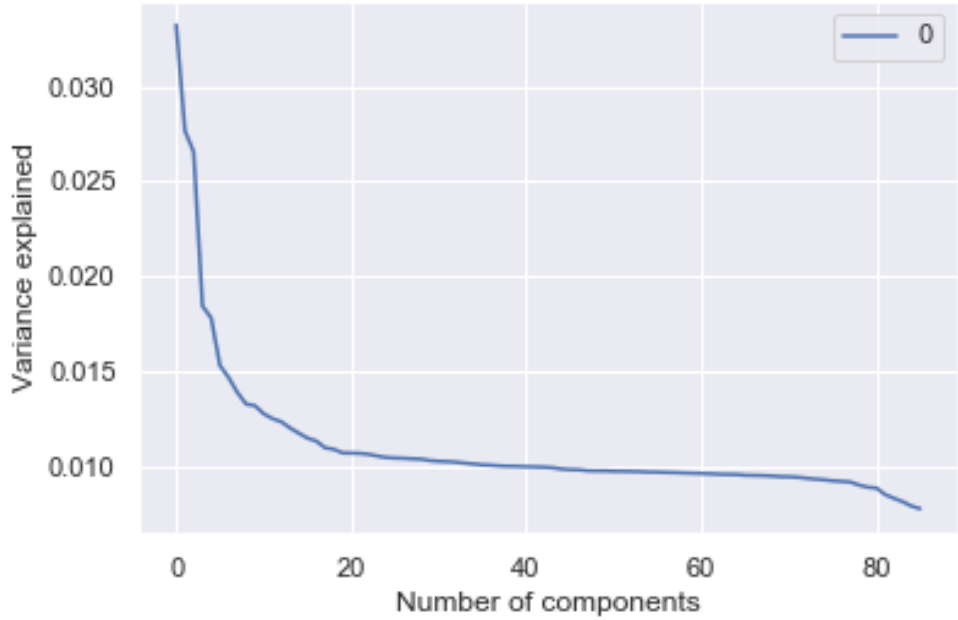


Methods: Feature engineering and selection

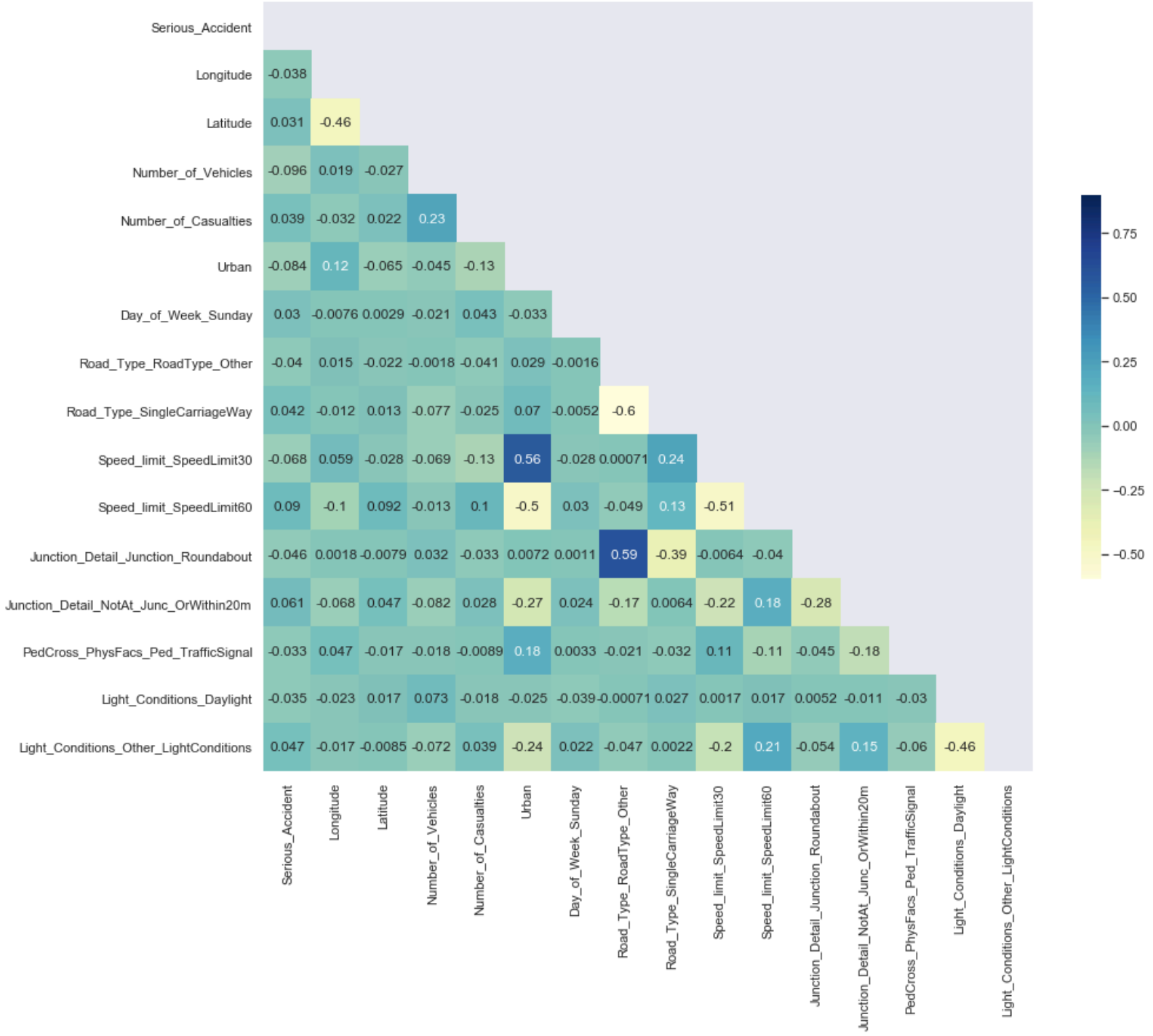
Picture source

- **Feature engineering**
  - Removed outliers e.g., Number of vehicles and casualties
  - Combined closely related categories into one category and then converted those into dummy or presence/absence features

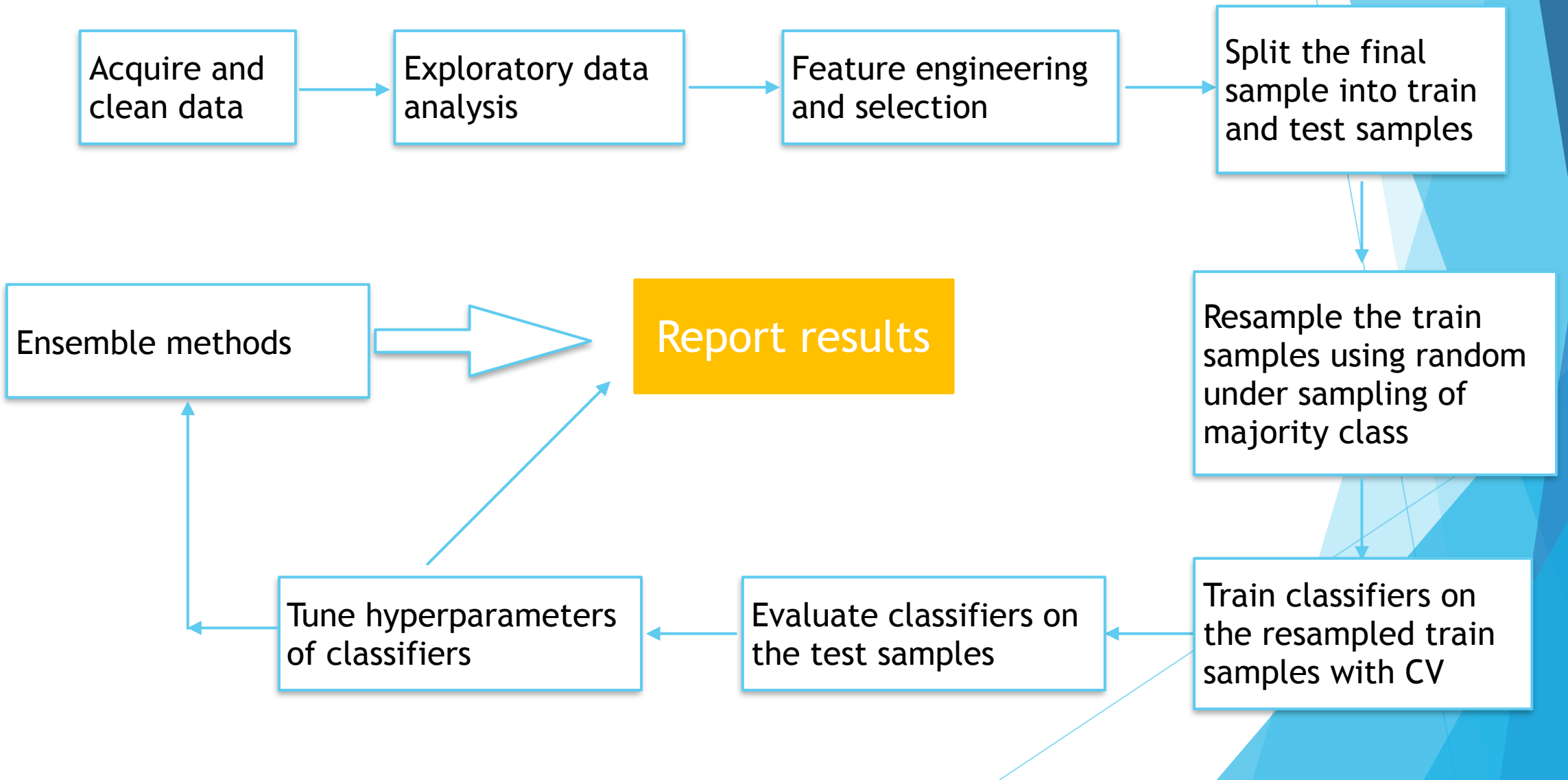
- **Feature selection**
  - Principal component analysis: 86 of 106 features needed to retain 95% of variance in features.
  - Used **15 k best** features: assessed the accuracy and area under ROC curve to select these features



Correlation matrix: Selected 15 k-best features



Classification and evaluation: Overall process

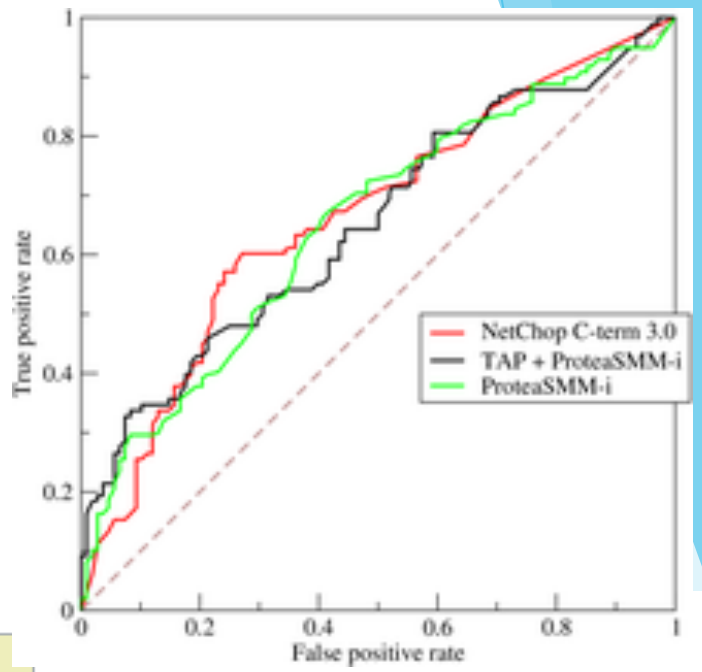


Classifiers

- **K-Nearest Neighbors (KNN)**
  - K nearest neighbors get to vote on classes.
- **Logistic Regression**
  - Likelihood of occurrence of positive outcome is used to determine class
- **Support Vector**
  - Data points that are closest to the decision boundary that best separates positive outcomes from the negative outcomes are support vectors. These closest points support or define the decision boundary.
- **Random Forest**
  - Multitude of decision trees at training and these decisions trees determine the final class.

Evaluation metrics

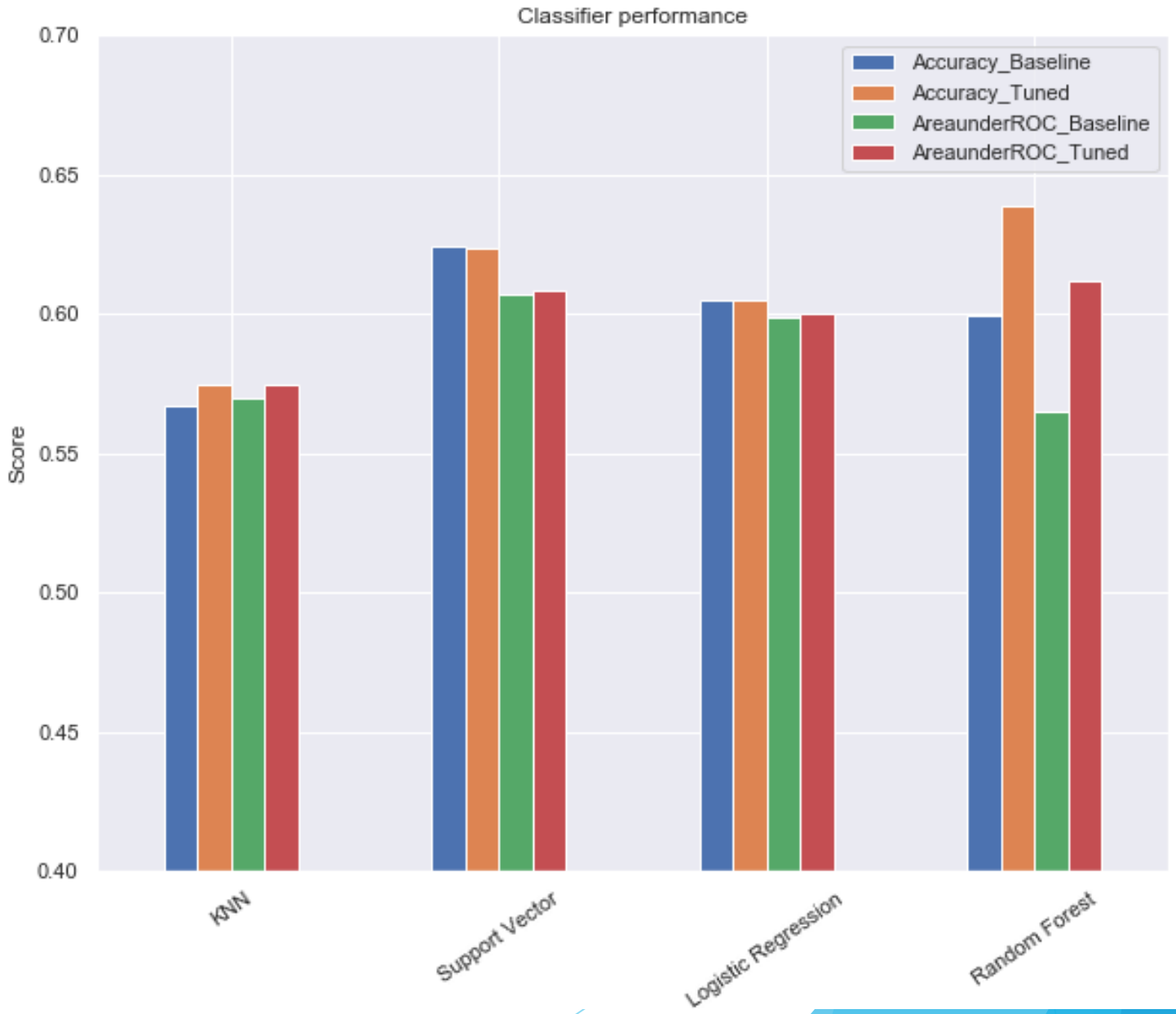
- Accuracy Score
- Area under Receiver Operating Characteristic (ROC) Curve



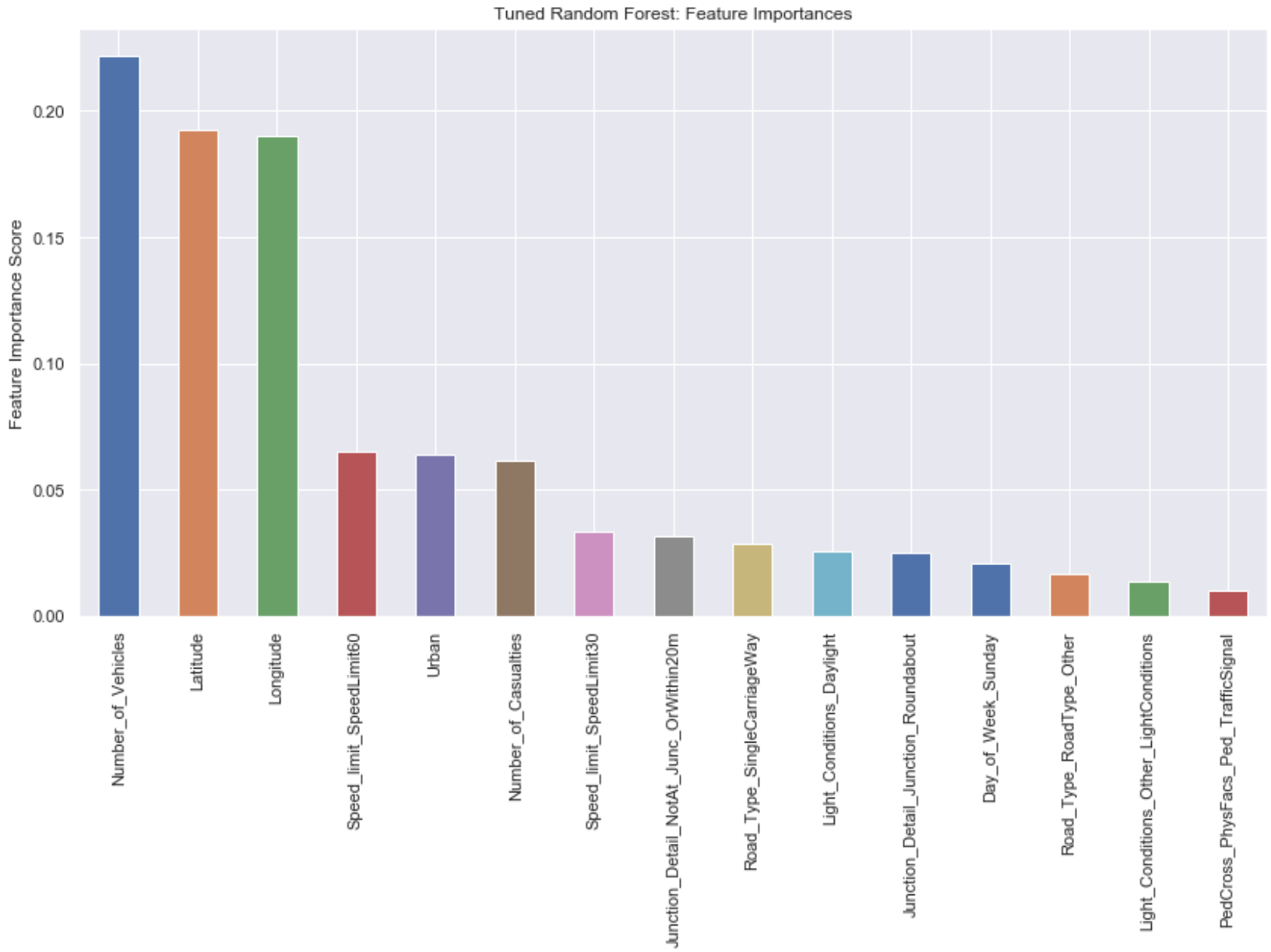
		True condition			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	F_1 score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$

Picture source

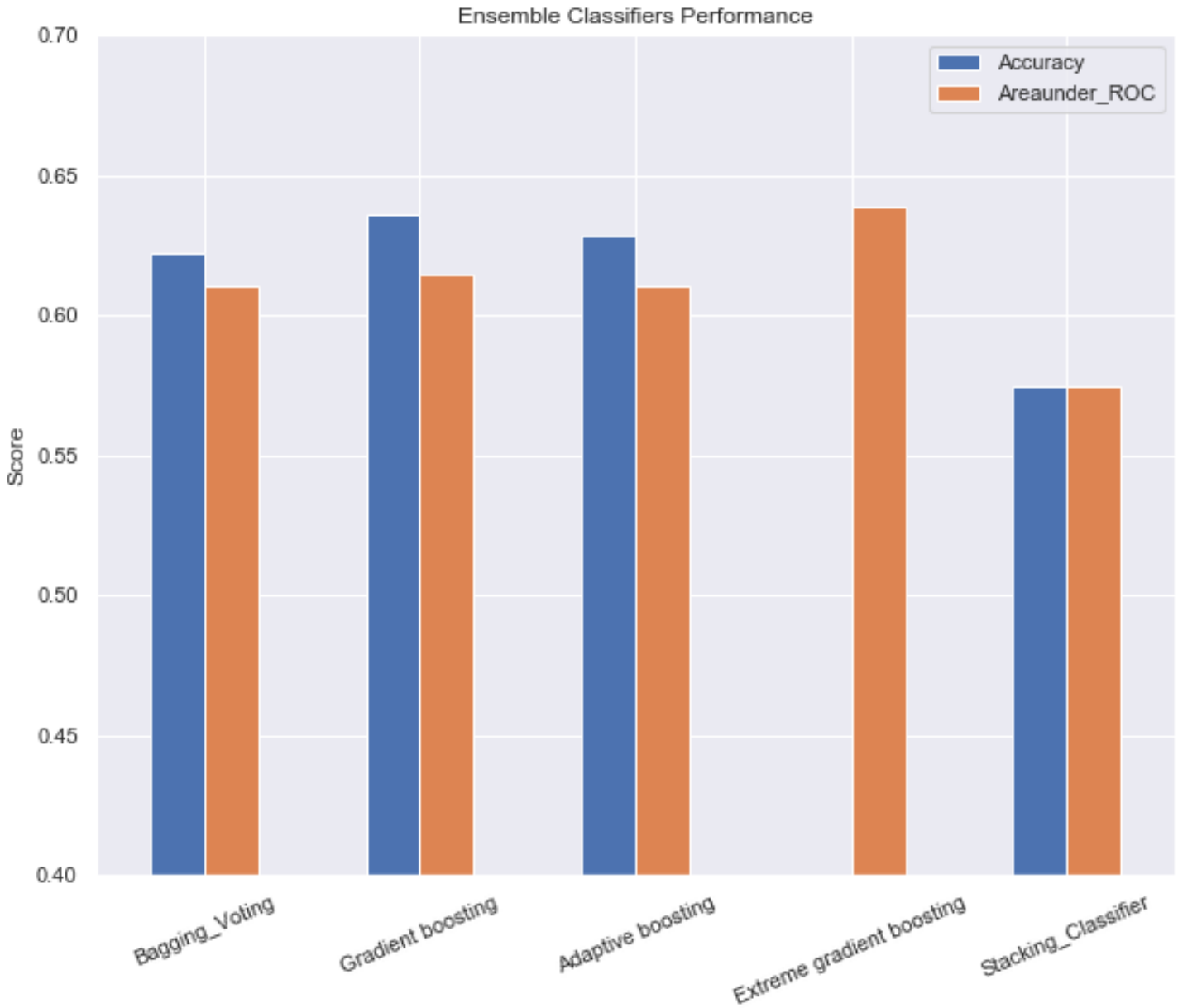
Performance before and after hyperparameters tuning



Feature importance: Tuned Random Forest



Performance of ensemble methods



Conclusions

- Support vector performed the best on baseline hyperparameters.
- Random forest improved the most after hyperparameters tuning.
- Boosting methods performed better than bagging or stacking ensemble methods.

Next steps

- Add vehicle and driver characteristics to the accidents data.
- Invest more time on tuning hyperparameters to improve performance.
- Try other resampling techniques.

Thank you!

Questions?