Identifying the causes and contributing factors of road collisions with injuries and fatalities in the City of Seattle, USA

Understanding the causes of serious road collisions leads to improvement

- Road collisions are the leading cause of death[1] in the USA for people aged <55
- Associated costs and productivity losses >$75 billion in 2017[2]
- By better understanding the causes, local Seattle Authorities can develop and implement policies and strategies to help reduce serious road collisions

[1] ASIRT [2] CDC

Data acquisition & cleaning

- Data obtained from the City of Seattle Open Data portal.
  - Downloaded on 6 September. Includes 221,266 records and 40 variables for road collisions 2004 to present, updated 5 September 2020.

- Unnecessary variables / keys as well as duplicate variables were discarded. Recorded with missing data were deleted. Some features were extracted.

- Cleaned data set consisted of 10 variables (DV & 9 IVs) with 174,452 observations.

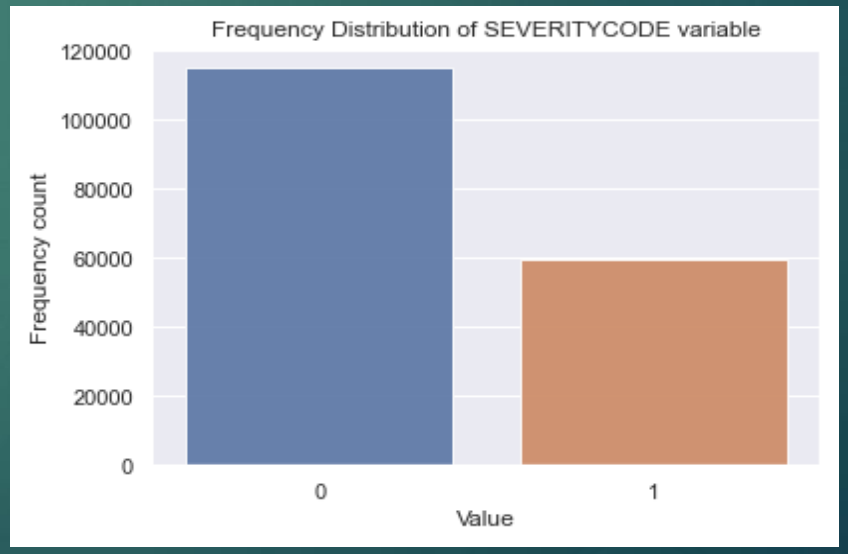
The target variable was processed to represent non-serious / serious collisions

- Originally consisting of 5 values, these were consolidated into two categories representing collisions resulting in 'property damage only' (0) and those resulting in 'injury or death' (1):

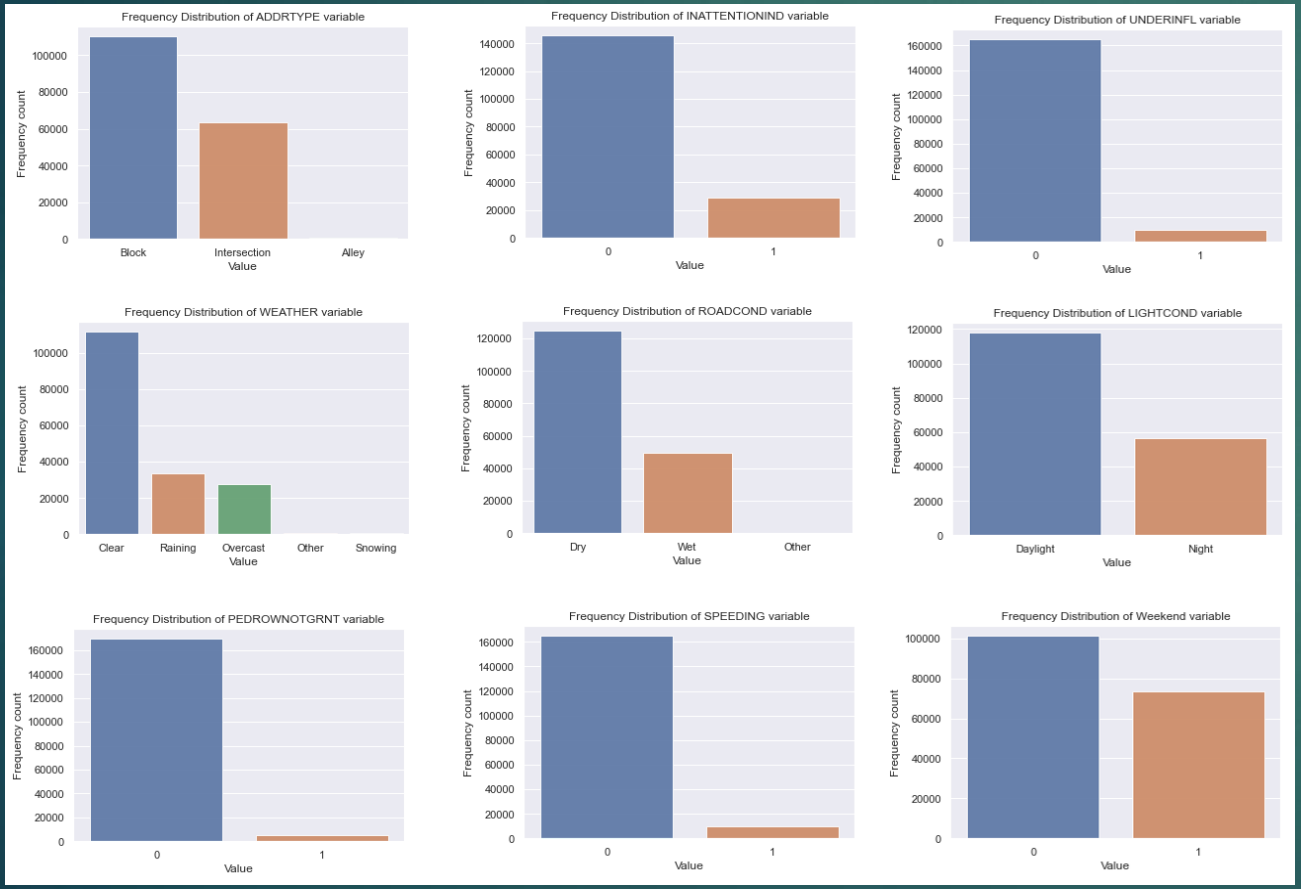
Before consolidation

0 - Unknown
1 - Property damage (only)
2 - Injury
2b - Serious Injury
3 - Fatality

After consolidation

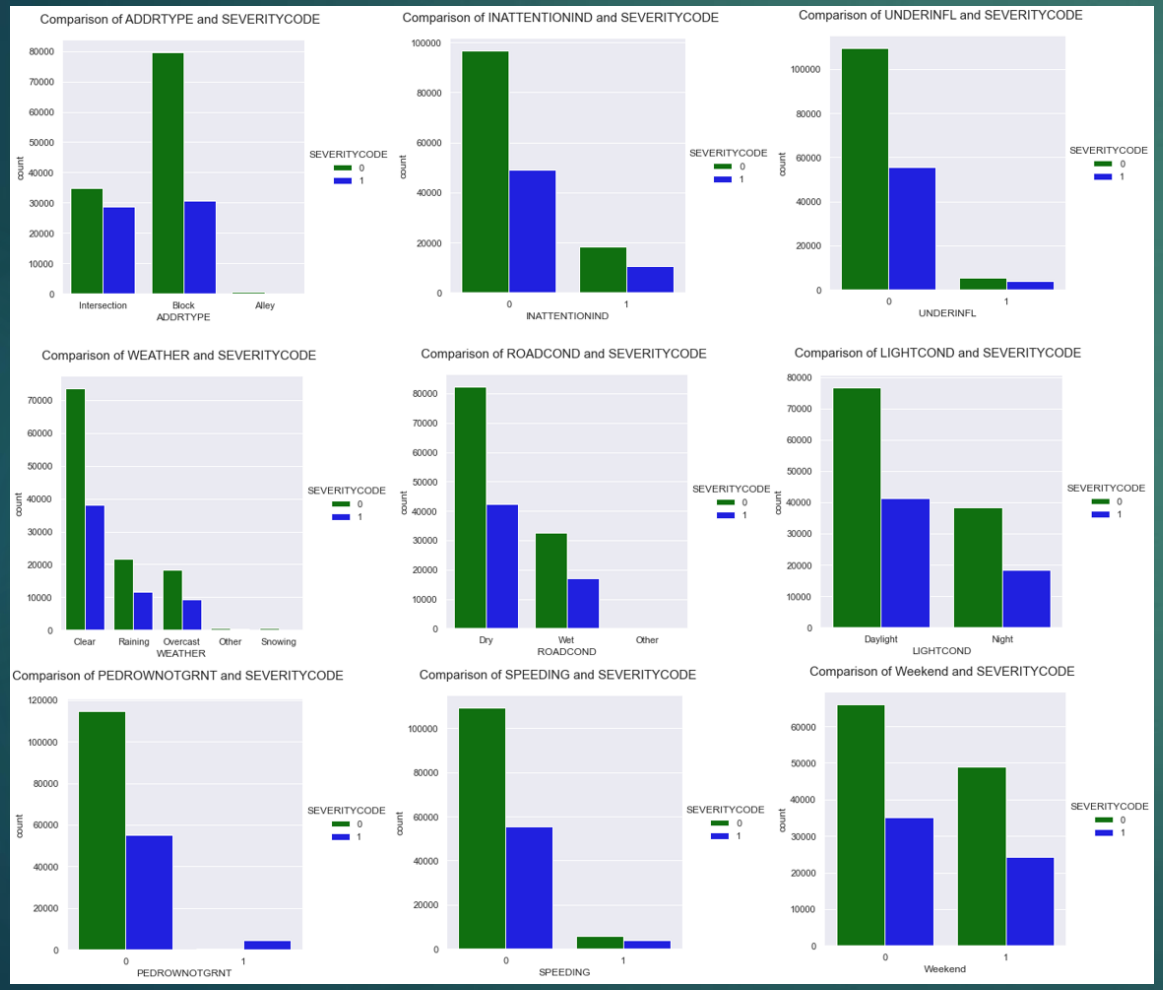


9 independent variables were used in the analysis



- These represented behaviours (e.g. inattention, under the influence) or conditions (weather, road and light conditions)
- Many variables had low levels of occurrence (i.e. few ‘1’ values)
- Converted to dummy variables prior to modelling

Several variables visually indicated a relationship with collision severity



- In particular, the following appeared (visually) to occur more frequently with serious collisions:
  - Intersections,
  - Driving under the Influence, and
  - Speeding

Modelling

The data was analysed using 4 sample balancing strategies and 4 prediction algorithms

Sample balancing strategies

- Original unbalanced sample
- Over-sampling minority class
- Under-sampling majority class
- Synthetic up-sampling of minority class using SMOTE

Prediction algorithms

- K Nearest Neighbour (KNN)
- Support Vector Machines (SVM)
- Decision Tree
- Logistic Regression

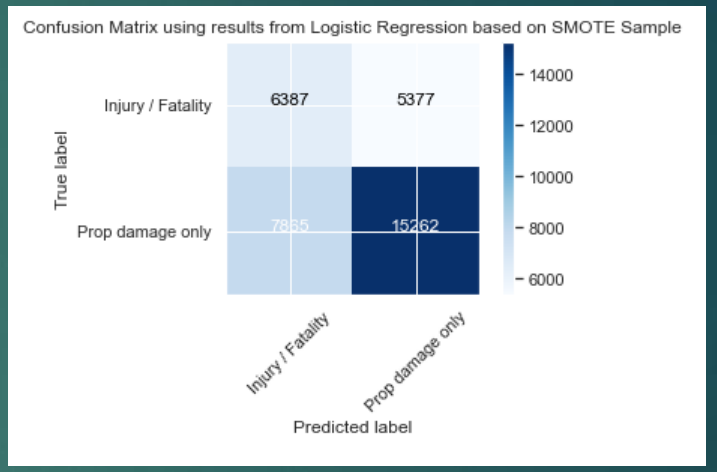
Logistic Regression using SMOTE sampling strategy was most accurate

F1 Scores by Sample & Algorithm	Original Imbalanced	Oversampled	Undersampled	SMOTE upsampled
KNN	0.622836	0.609429	0.575888	0.614389
SVM	0.594983	0.610955	0.624016	0.614423
Decision Tree	0.593774	0.599331	0.611951	0.624390
LogisticRegression	0.596860	0.625437	0.627413	0.627835

- While several accuracy metrics were calculated, the F1 score was used to compare all 16 combinations of sample balancing strategy and prediction algorithm.
- Logistic Regression on the SMOTE sample delivered the highest F1 score of 0.6278

Overall the best model performed moderately well

- There was a relatively high level of false positives and false negatives.
- This may indicate that not all contributing factors are captured in the model or dataset.
- Further investigation is required to establish whether other potential factors can be collected from collisions.



	precision	recall	f1-score	support
0	0.74	0.66	0.70	23127
1	0.45	0.54	0.49	11764
accuracy			0.62	34891
macro avg	0.59	0.60	0.59	34891
weighted avg	0.64	0.62	0.63	34891

Top 5 contributing factors for collisions causing injury / death

Contributing factor	Logistic Regression Coefficient
Pedestrian Right of Way Not Granted	2.314
Intersection location	0.602
Driver Under the Influence	0.529
Speeding	0.475
Inattention	0.303

- Examining the Logistic Regression coefficients highlights 5 top contributing factors
- 4 of these are 'human' factors, one is related to the location of collisions (Intersections)

Conclusions

- Built useful models to identify contributing factors leading to road collisions causing injury or death.
- However the accuracy of the model has potential for improvement
- Need for additional data to be collected, representing other possible causes / factors.
- Some examples could include:
  - Reckless driving behaviour
  - Condition of the vehicle / mechanical issues
  - Condition of the road (e.g. potholes, impaired view)