Predicting Incident Severity at Seattle, WA

IBM Applied Data Science Capstone

Road Traffic Costs Lives and Money

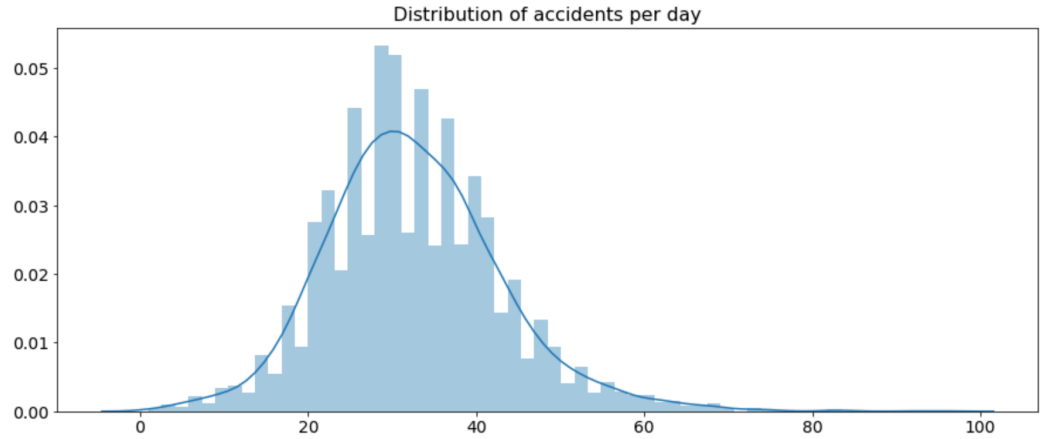
- In 2019 alone, 36,120 lives were taken away due to all sorts of causes towards road incidents — driver misconduct, adverse weather, etc.
- Spendings on property recovery, medical costs, legal bills, and loss of earnings sums up towards billions of dollars on an annual basis.
- Identifying factors that contribute towards these incidents and their severity may be insightful for local authorities and drivers who wish to maintain a safe driving environment.
- This project attempts to model the road condition in order to predict whether it is safe and smooth for drivers to go onto road trips in Seattle, WA.

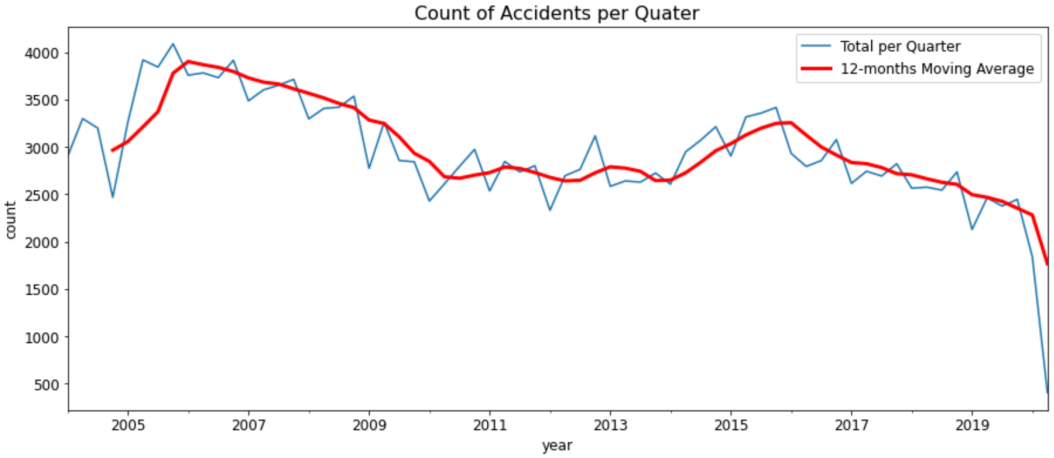
Data

- Data extracted from local authority
- Incidents in Seattle from 2004 and 2020 are available, with details such as location, place, time, conditions, fatalities and other are provided
- Total of 194,673 entries and 38 attributes are available

Dependent Variable	
SEVERITYCODE	A code that corresponds to the severity of the collision
Independent Variables	
INATTENTIONIND	Whether or not the collision was due to inattention
UNDERINFL	Whether or not a driver was under influences of drugs or alcohol
SPEEDING	Whether or not speeding was a factor in the collision
WEATHER	Description of weather conditions during the time of collisions
ROADCOND	Condition of road during the time of collision
LIGHTCOND	Condition of lighting during the time of collision

Exploratory Analysis

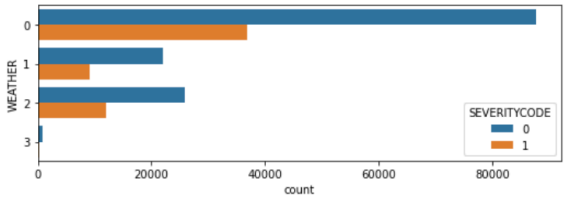


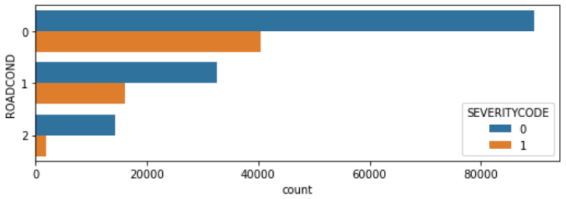


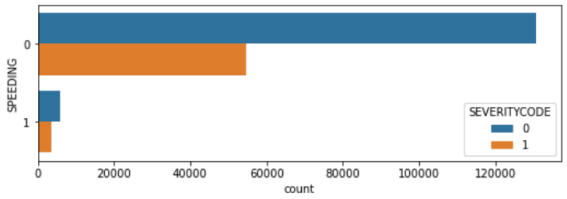
- Maximum around 30 incidents per day
- Decreasing trend of incident count in the past 16 years, with frequency highs occurring in Q2 and Q3 of the year on a regular trend

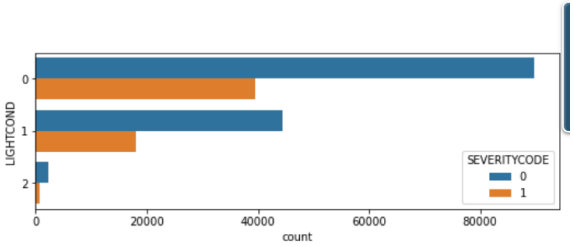
Exploratory Analysis - cont.

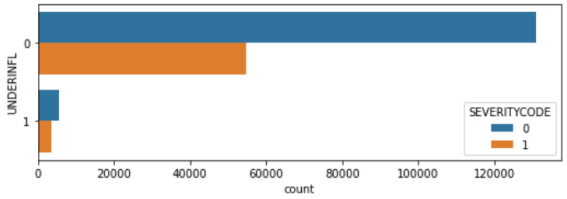
- Entries were replaced with scores to represent relative significance
- Unknown entries were replaced with scores in according to the proportion of filled scores
- Unbalanced dataset with code 1 incidents being the majority
- Use of SMOTE to balance the training data

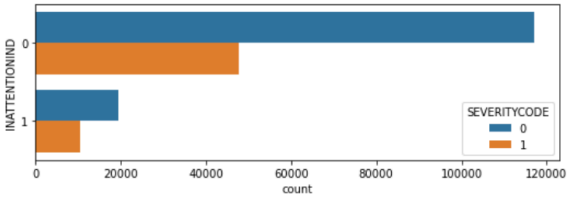






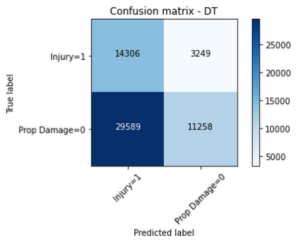




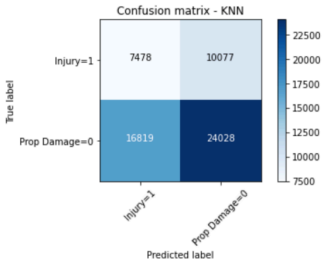


Results

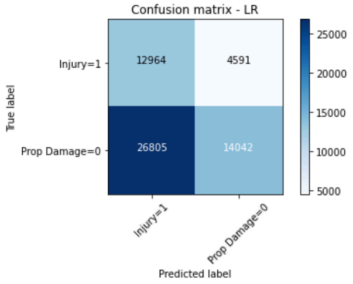
	precision	recall	f1-score	support
0	0.78	0.28	0.41	40847
1	0.33	0.81	0.47	17555
accuracy			0.44	58402
macro avg	0.55	0.55	0.44	58402
weighted avg	0.64	0.44	0.42	58402



	precision	recall	f1-score	support
0	0.70	0.59	0.64	40847
1	0.31	0.43	0.36	17555
accuracy			0.54	58402
macro avg	0.51	0.51	0.50	58402
weighted avg	0.59	0.54	0.56	58402



	precision	recall	f1-score	support
0	0.75	0.34	0.47	40847
1	0.33	0.74	0.45	17555
accuracy			0.46	58402
macro avg	0.54	0.54	0.46	58402
weighted avg	0.63	0.46	0.47	58402



Left upper: Decision Tree
Left lower: Logistic Regression
Right upper: k-Nearest Neighbour (optimum: k=4)

Results - cont.

Model	f1 Score	Accuracy Score
Decision Tree	0.42	0.438
Logistic Regression	0.47	0.462
k-Nearest Neighbour	0.56	0.539

Conclusion and Recommendation

- Built models to predict incident severity based on 6 independent variables that changes randomly
- Accuracy of model has rooms for improvements
- Better performances may be benefitted from the followings:
  - More balanced dataset
  - Less BLANK (i.e. np.nan) cells where attributes are unknown
  - More factors taken into account (i.e. date and address of collision)