Analysis of Seattle Collision Data

a presentation by Miranda Childs for Coursera's IBM Data Science Professional Certificate capstone project

Collision reduction starts with awareness

- In Seattle, WA, from 2004 to present there have been 194,673 collisions reported by the Seattle Police Department (SPD) to the Seattle Department of Transportation (SDOT). 58,188 of those collisions involved an injury.

- By determining and focusing on the factors that contribute most heavily to **severe collisions**, we can create meaningful strategies to reduce the number of accidents, especially those with injuries, in order to increase the wellbeing and longevity of our community.

- Vision Zero Network is "a collaborative campaign helping communities reach their goals of Vision Zero -- eliminating all traffic fatalities and severe injuries -- while increasing safe, healthy, equitable mobility for all." (https://visionzeronetwork.org). Through our thorough analysis we will make recommendations for the next campaigns and strategies that Vision Zero can execute in collaboration with SDOT.

Data Acquisition and Cleaning

- Collision data from 2004 to present, provided by the Traffic Records Group in conjunction with the Seattle Police Department and Seattle Department of Transportation. (collision data set, and metadata)
- 194,673 rows and 37 features in the raw dataset
- Duplicate columns and null values were dropped
- Features were transformed for consistency and accuracy
- Undersampling was utilized to balance the labeled data
- 4 features were chosen to predict the severity of collisions: Under The Influence, Weather, Road Condition, and Light Condition
- The cleaned data contains 111,840 rows and 4 features

Methodology aka look at all these cool visualizations!

- discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why
- In this phase, various algorithms and methods can be selected and applied to build the model including supervised machine learning techniques. You can select SVM, XGBoost, decision tree, or any other techniques. You can select a single or multiple machine learning models for the same data mining problem. At this phase, stepping back to the data preparation phase is often required.

Results

- discuss the results. “Certain metrics can be used for the model evaluation such as accuracy, recall, F1-score, precision, and others” — or should this be in discussion section???

Discussion

- section where you discuss any observations you noted and any recommendations you can make based on the results.

Conclusion and recommendations

- Collisions suck!
- A citywide campaign can emphasize the hazards of driving during low-light and poor weather conditions. Up to date digital signage can also be considered.
- Accuracy of the models can increase with better data collection.
- In future record keeping, create more comprehensive records with fewer null values. For example ‘Unknown’ or ‘Other’ should never be listed as weather conditions.