Predicting Car Accident Severity

Machine Learning approaches to the problem

By Kauvin Lucas

Predicting car accident severity is challenging, but with the right model it is easier

- Car accidents are the number one problem in US transportation, and some of the main causes of fatality in the Washington state
- The city of Seattle is still far from reaching the goal of zero fatality rate by 2030
- Technological advancements may allow finding predictive patterns in the large amount of data whose variables are seemly uncorrelated
- The right prediction model may aid in the decision making processes of governments, insurance companies and healthcare institutions

Approaches to solve the problem

- No single factor is enough to explain the severity of the accident.
- By gathering several variables into a single model, it’s possible to find generalizable predictive patterns.
- This can only done by employing and evaluating machine learning models and data science methodology
- Unsupervised classification models that will be used: k-nearest neighbors, decision tree, support vector machine and logistic regression

Data collection and understanding

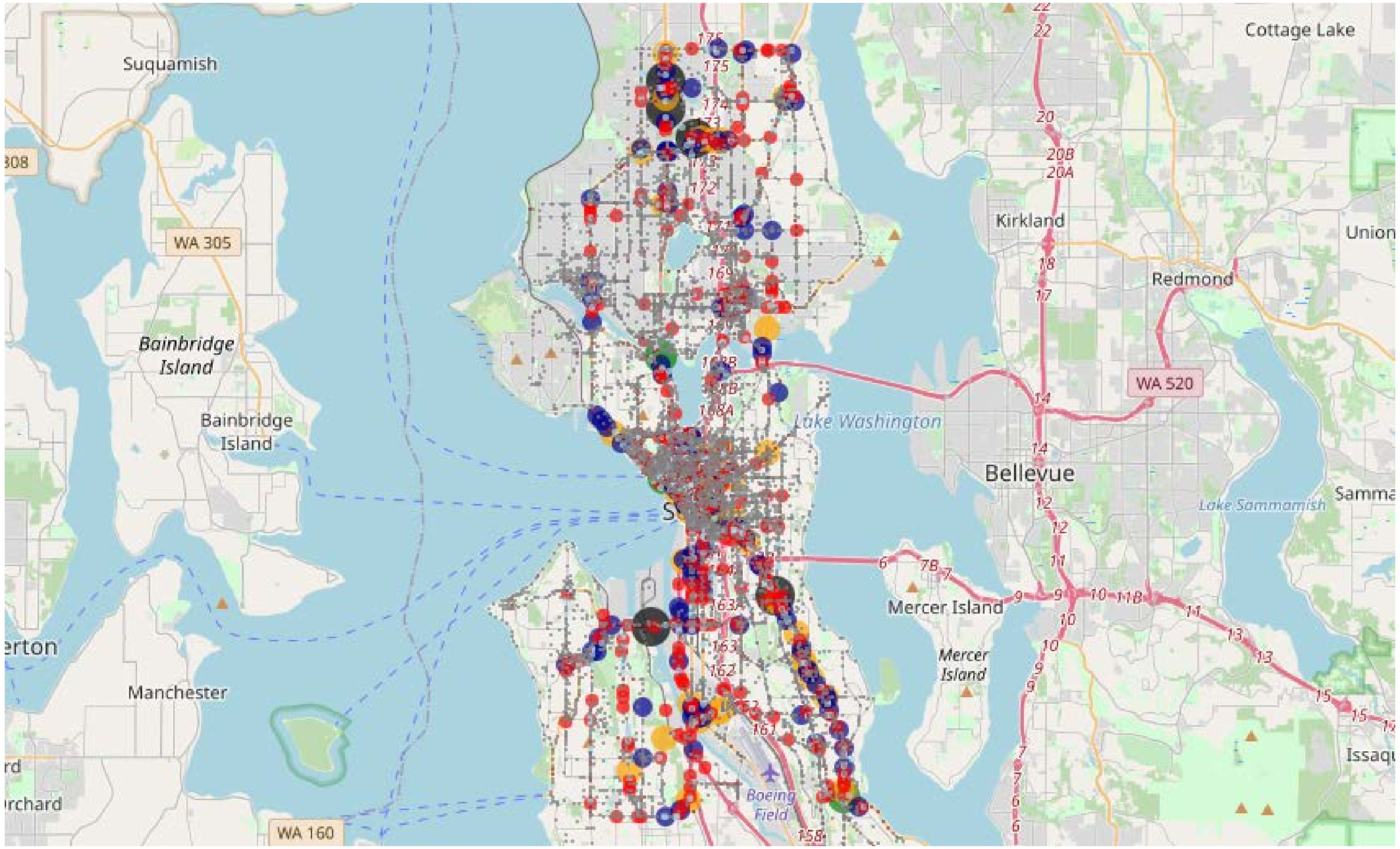
- The data used was provided by the Washington State Department of Transportation (WSDOT)
- The raw data has 37 columns and 194,673 rows
- 9 features were selected from the data
- A sample of 15,000 was taken from the dataset for modeling and evaluation purposes

Data collection and understanding (cont.)

- The output variable is the **severity code**, which can take values between 0 and 3, from least to most severe
- Only two different categories were recorded in the output variable: “property damage only” as 1 and “injury” as 2
- One more feature was created to be selected, which describes whether or not collision has occurred in holiday

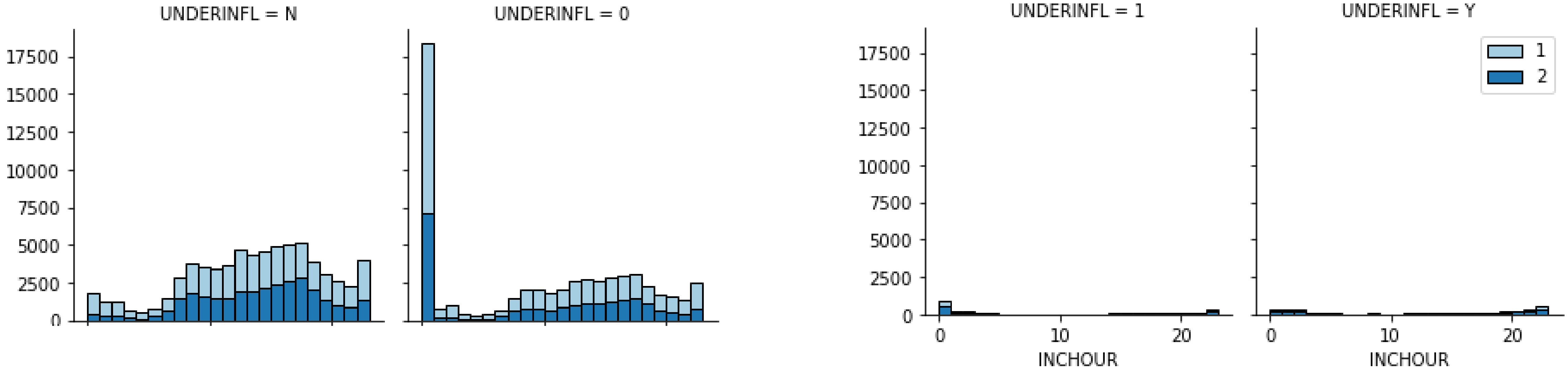
Data visualization

Map of Seattle showing several spots where the accidents have occurred



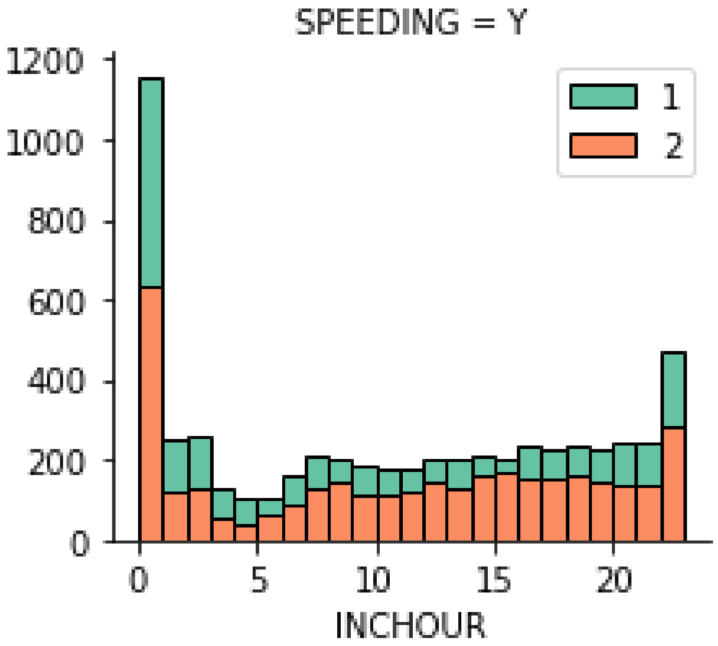
Data visualization

Accident frequencies comparison by hour of the day, severity code and whether the driver was drug/alcohol impaired



Data visualization

Time of the day may affect accident severity, as proportionally more severe accidents occur in afternoon than any other time of the day



Several other variables may influence the frequency of accidents. For example, more or less severe accidents may happen at intersection or around the block, or poorer streets light condition may cause the most severe accidents. A combination of all of these variables may be enough to make accurate predictions.

ML algorithms employed

- **k-nearest neighbors:** assumes that similar values exist in close proximity
- **Decision tree:** maps out the possible outcomes for each test attribute
- **Support vector machine:** intents to find a plane in a N-dimension that better separates two classes of data points
- **Logistic regression:** appropriate regression analysis when the dependent variable is dichotomous

Modelling

- Features extracted
- Train/test split and hyperparameter optimization employed
- Accuracy score calculated for each trained model:
    - **k-nearest neighbors:** accuracy score of 65.82% for the training set
    - **Decision tree:** accuracy of 65.93%
    - **Support vector machine:** accuracy of 66.23%
    - **Logistic regression:** accuracy of 61.93%

Evaluation

- The models were evaluated and reported by using Jaccard score, F1 score and log loss
- The results were the following:

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.65	0.61	NA
Decision Tree	0.65	0.54	NA
SVM	0.66	0.53	NA
LogisticRegression	0.66	0.61	0.63

Conclusion

- Some patterns in the data allowed for the models to make prediction with some decent accuracy
- By comparing each score, it is concluded that the logistic regression model is the most appropriate model, but the other scores were very similar
- The outcome variable can take 5 different categories, but it was treated as a binary variable since it had only 2 categories
- The feature data was randomly selected, and the scores may change lightly if the whole steps were retaken

Thanks!