Predicting Severe Accidents in Seattle

By Heather Spero

Solving the Accident Problem

Seattle had 187 serious accidents in 2017

Many factors impact potential serious including weather, vehicle, light, time of year etc.

Important to predict the number of severe accidents

Emergency workers and hospitals could utilize this information to be prepared and efficient

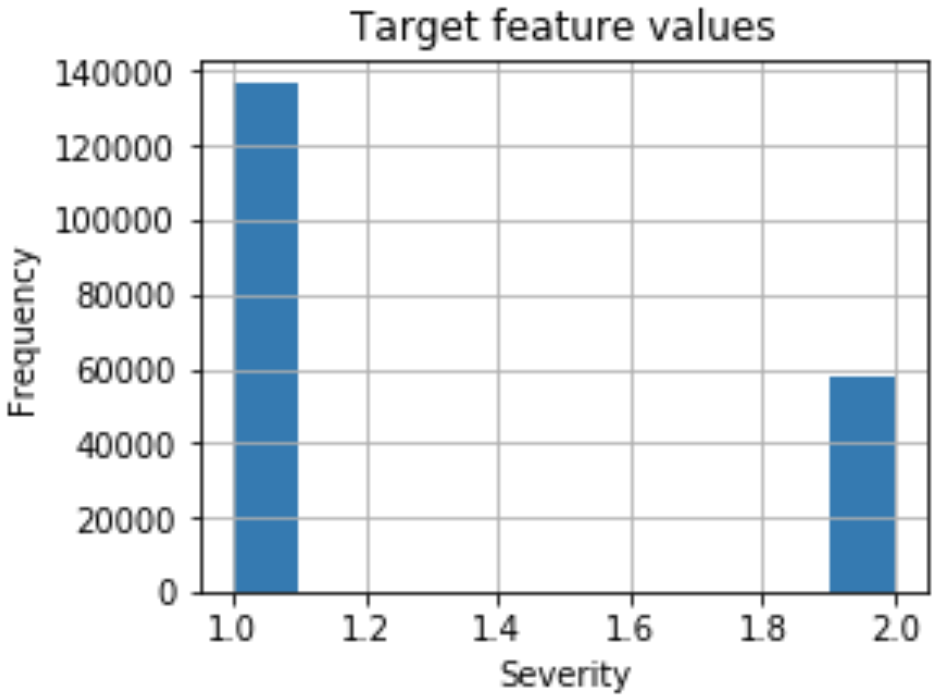
Data Acquisition

- Data was provided by SDP recorded by Traffic Records
- Target Value is SEVERITYCODE:
  - 1 – not severe accident
  - 2 – severe accident
- Primary variables utilized in analysis: Road Conditions, Light Conditions, month (time of year)

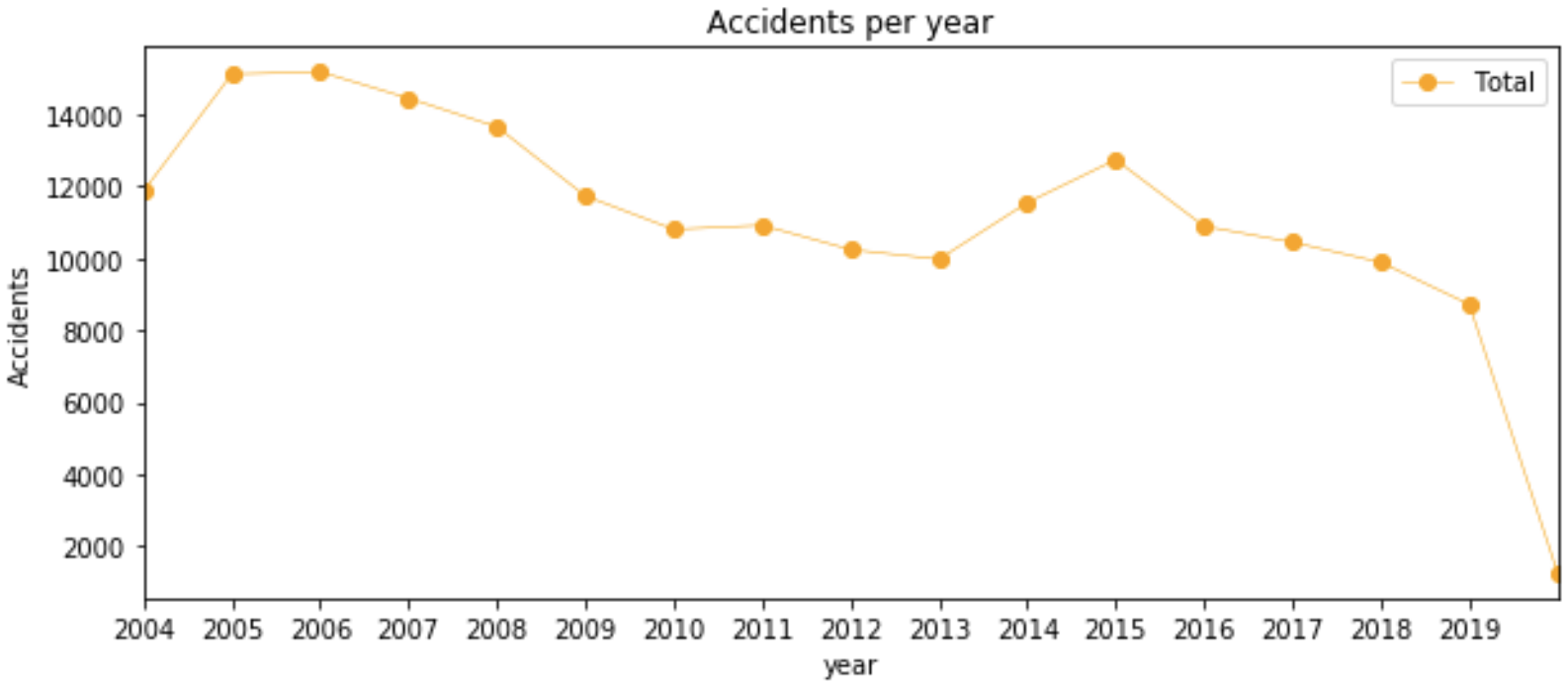
Data Cleaning

- Road Condition and Light Condition Data was categorical and had to be translated to numeric
- 2% of the records were missing Road Condition or Light Condition Data and were removed
- Unbalanced weighting of Target Variable
  - Over 2 times more Code 1's than Code 2's
  - Sample had to be rebalanced before creating the algorithm so Code 1 did not dominate
- Data was standardized prior to analysis

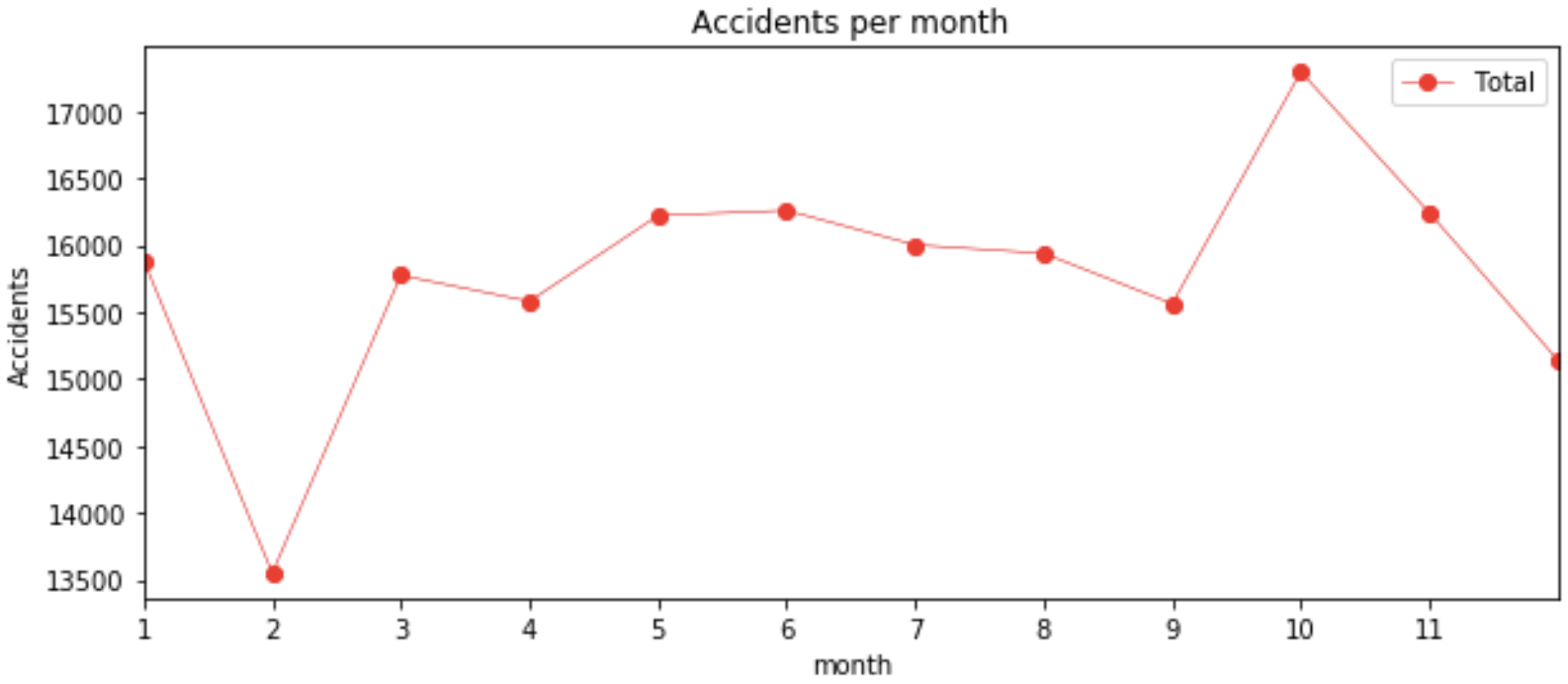
Target Distribution



Accidents Per Year



Accidents Per Month



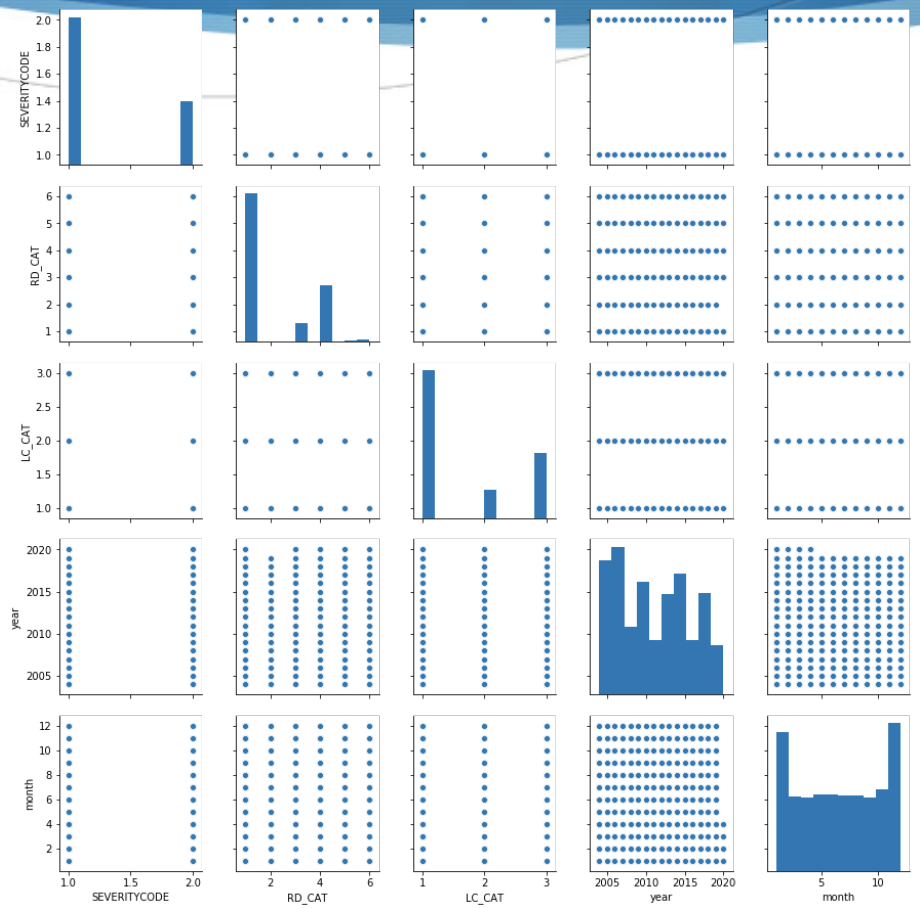
Correlation of Each Variable

	RD_CAT	LC_CAT	year	month
SEVERITYCODE	-.035	-.104	.023	.004

Review of Correlation

- There appears to be no significant correlation between the individual variables and the Target variable
- Most significant variable appears to be day light conditions however the value is still not significant
- The data will be analyzed through clustering and Logistic Regression to try and improve the predictive method

Comparative Distribution of the Variables



Predictive Methods for the Target(SeverityCode)

- K-Nearest Neighbor (KNN) KNN predicts by finding the most similar data point within k distance.
- Decision Tree model provides a layout of all possible outcomes
  - Random Forest model is also used to try and improve the accuracy.
- Logistic Regression model will only predict one of those two classes(binary) therefore appropriate for this instance

Best Model

- Only concerned with severe accidents so only looking for the model that predicts the most accurate for SEVERITYCODE 2
- The RECALL is the % of truly severe accidents that were properly predicted is an important statistic in this case
- The F1 Score is also important as it is calculated at the Target Value level
- The Jaccard Score is calculated on the accuracy of Code 1 and Code 2 combined so is less relevent for this study

Summary of Results for Severe Accidents

Algorithm	Jaccard	F1 Score	Precision	Recall
KNN	.53	0.50	0.53	0.47
Decision Tree	.56	0.63	0.54	0.76
Random Forest	.56	0.63	0.54	0.76
Logistic Regression	.53	.56	.53	.59

Conclusion

- The best model for the purpose of predicting the severe car accidents in the Seattle area is the Decision Tree (and Random Forest)
- The Decision Tree Recall score of 76% is far superior
  - KNN and Logistic Regression Recall scores was .53
- The Decision Tree F1 Score was slightly higher at .63
  - KNN and Logistic Regression F1 scores were .5 and .56

Discussion

- Beneficial to create a predictive model for severe accidents
- Utilizing the variables Road conditions, Light Conditions and month in Decision Tree model
- Multiple applications:
  - Police
  - Ambulances/Hospitals
  - Firefighters
  - Notification to all drivers for more caution