Using Machine Learning to Predict Collision Severity

Studying Collision Data from 2004 to 2020 for the Seattle Dept. of Transportation

Sam Steffes | September 4, 2020

Introduction

In the Spring of 2015, the Seattle Department of Transportation (SDOT) released a 10-year Strategic Vision for providing safe and sustainable transportation infrastructure for the city. This plan highlighted the adoption of a *Vision Zero* goal to eliminate serious and fatal crashes by the year 2030. However, as demonstrated in the figure below, while the annual volume of total collisions has decreased 27.6% since 2015, the number of severe collisions (those where injuries or fatalities occur) has decreased by only 18.4%, and has accounted for a larger portion of all incidents year-over-year since the report was released. This suggests that the initiatives and actions implemented to reduce collisions, while successful, are not adequately targeting the conditions or locations that result in severe collisions.

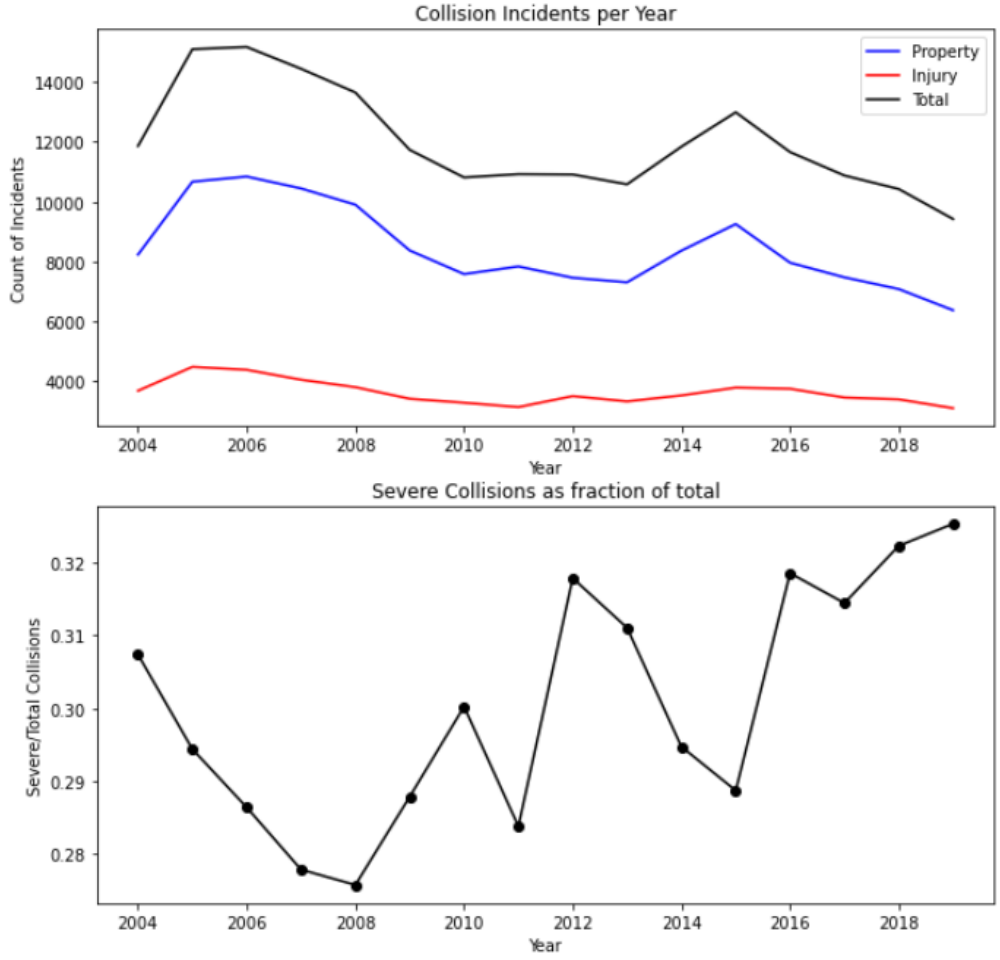

Figure 1 – Trend in collision results from 2004-2019.

Developing programs and strategies to accelerate the elimination of severe collisions first requires a better understanding of why they occur. Since the severity of collisions likely depends on many conditions, obvious correlation with discrete parameters would be difficult to identify. Therefore, a machine learning model that can evaluate a diverse feature-set and predict outcomes is a more effective analytical approach. A model that accurately predicts the severity of collisions offers several benefits for the SDOT's efforts to improve roadway safety, such as:

- identifying the factors that most contribute to severe collisions allowing more targeted improvements,
- providing a way to quantify (evaluate) and select improvement options that have the most benefit (i.e. reduction in severe incidents), and
- supporting the development and deployment of decision-support tools that optimize emergency response services and dispatch them more effectively.

Data

Details

Models to predict severity were trained using historic collision records collected by the Seattle Police Department and maintained by the SDOT Traffic Management Division. The dataset includes collision incidents from between January 2004 and May 2020. Records are assigned to two groups – 'property damage only' and 'injury' collisions – which made it ideal for supervised learning classification algorithms; the SEVERITYCODE feature that identifies the severity outcome was used as the target variable. There are approximately 195,000 samples in the dataset; however, there are 2.35 'property' collisions for every 1 'injury' collision, so data had to be balanced in order to avoid skewing the model according to historic volumes.

```
Type 1 collisions (property damage) =  136485
Type 2 collisions (with injury) =  58188
Ratio of Collisions with property damage only to those with injuries: 2.35
```

Figure 2: count of collisions by classification, demonstrating the need to balance the data for training

The dataset describes each collision using 36 different features, which provided a robust set of independent variables as possible parameters for training the model. The features identify:

- the date, time, position and location details of the collision;
- codes used by the state and DOT to categorize the collision;
- characteristics of the incident, such as type of collision and the number of people, pedestrians, bicycles and vehicles involved;
- environmental factors like weather and road condition; and
- the existence of driver behaviors including speeding, inattentiveness and the presence of drugs or alcohol

Much of the dataset is discrete information or Boolean fields that divide the samples into different categories, groups or states. The COLLISION type field, for example, classifies each incident according to ten common types of collisions. Several fields – including INATTENTIONIND, UNDERINFL and SPEEDING – identify the presence or absence of specific conditions; however, this is done using both Y/N and 0/1 variables, which required data cleansing. There are several continuous data fields, including the date

and time values, as well as the numeric counts of people, cars, etc. Some samples have one or more blank or incomplete fields which required cleansing or processing. Occasionally, the INCDTTM field - which captures the date and time of the incident – contains only the date portion of the timestamp.

Initial Feature Selection

The table below lists 12 features from the dataset that were initially identified as possible candidates for training the model, along with high-level assumptions about their potential significance. Data exploration steps described in the Methodology section demonstrate how these features were examined to determine possible correlation with collision severity and final selection for inclusion in the training features.

Column Name	Description	Possible Application/Value/Impact to model
COLLISIONTYPE	A keyword describing the collision, eg 'head-on', 'angled', 'cycles', etc.	The orientation or nature of the collision could be valuable to the likelihood of injury to people involved.
PERSONCOUNT	Total number of people involved in the collision.	More people involved in the collision means more chances for injuries... people could be seated in the vehicle in spots that were closer to the impact, etc.
PEDCOUNT	The number of pedestrians involved in the collision.	Pedestrians are at greater risk of injury when struck by vehicles because they are not protected by another vehicle frame, seatbelt, airbag, etc.
PEDCYLCOUNT	The number of bicycles involved.	As with pedestrians, bicyclists are also at greater risk of injury when struck by vehicles.
VEHCOUNT	The number of vehicles involved in the collision.	More vehicles are more chances for injury, and might suggest more extensive impact, damage or severity.
INCDATE/INCDTTM	Date and time recordings for the incident records.	This is not as clear; however, the date or time that an incident occurs may be a proxy for driving behaviours that result in more severe incidents. For example, the are higher traffic volumes during weekday rushhours, and people are in a hurry or distracted.
INATTENTIONIND	If collision was due to inattention.	Distracted driving certainly increases the probabibily of getting in an accident, and may affect severity as well.
UNDERINFL	If driver was under influence of drugs/alchohol.	Someone under the influence might be driving recklessly, causing the collision to be more severe, and may also not be wearing their seatbelt, increasing their own injury risk.
WEATHER	Weather at the time of incident.	Weather could have contributed to the nature of the collision, or the control of the vehicles before and after the collision.
ROADCOND	Condition of road at the time of incident.	This is probably closely aligned with weather, but could also capture the maintenance condition of the roadway.
LIGHTCOND	Light conditions at the time of incident.	Poor visibility may have contributed to the collision, especially where there are already blind or hidden turns, difficult or dangerous maneuvers are required, or acceptable speeds are already high.
SPEEDING	If speeding was a factor in the collision.	Speeding or reckless driving could result in more forceful collisions.

Table 1: candidate model features from the SDOT collision dataset

Methodology

The following sub-sections describe the steps taken to examine and process the dataset, before training and evaluating several classification models.

Exploratory Analysis

Each of the 12 features identified in Table 1 was explored for possible impact on collision severity. This was done using various plotting methods, in order to visualize potential relationships between the candidate variable and severity target. Often this was accomplished using count or histogram plots that showed the distribution of property and injury collisions for different characteristics. In some cases, the data was manipulated to show the rate of severe collisions, taken as the proportion of injury type collisions as part of the total number of incidents. Calculating the rate of injury collisions was particularly useful when the volume of samples was low in groupings for certain characteristics. The goal of the exploratory analysis was to identify features that showed variation in the distribution or proportion of injury collisions for different values or conditions. Such features would better help the model to better distinguish between classifications.

Impact of Date and Time

The date and time of incidents was explored to determine whether there was any relationship between severity and the day-of-the-week or hour-of-the-day when the collision occurred. The initial assumptions where that severity rates might be higher on weekdays and during rush-hour periods when more vehicles are on the road, creating more opportunities for collisions. Table 2 shows collision volumes and severity rates, grouped by day-of-the-week. Higher injury rates are observed for Monday-Thursday compared to Friday-Sunday; however, these differences are muted.

	Weekday	Property	Injury	Total	inj_rate
0	0	18365	7973	26338	0.302719
1	1	19825	8731	28556	0.305750
2	2	20021	8757	28778	0.304295
3	3	20306	9018	29324	0.307530
4	4	22774	9559	32333	0.295642
5	5	19342	8047	27389	0.293804
6	6	15852	6103	21955	0.277978

Table 2: collision volumes by type, and injury rates grouped by days of the week (Monday is 0)

Alternatively, as shown in Figure 2, there were noticeable peaks in injury rate during the rush-hour periods between 5-8AM and 3-7PM. This is especially pronounced when separating weekdays from weekends, seen by comparing the red and blue lines in the figure. While the specific hour does not directly cause collisions to be more severe, it likely serves as a proxy for behaviors or conditions (not captured in the dataset) that result in more severe collisions. For example, drivers may exhibit more risky practices during rush-hour.

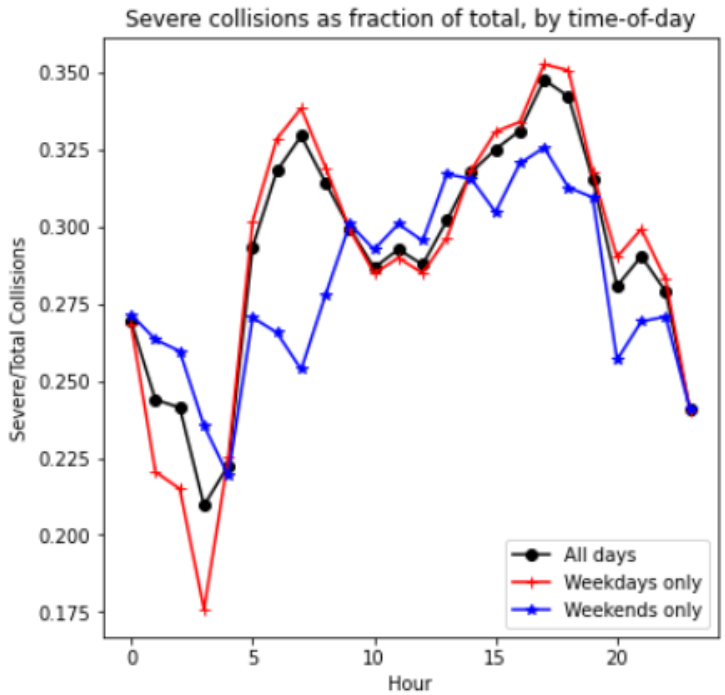

Figure 2: severe collision rate by hour of the day

Impact of Collision Type

The type of collision was also analyzed to determine the relative impact on collision severity, with predictable results. A significant number of collisions involving pedestrians and bicycles result in injury, compared to other types that describe impacts between two or more vehicles. Collisions with parked cars, where speeds are likely lower and only one car is occupied, had the lowest rate of injuries.

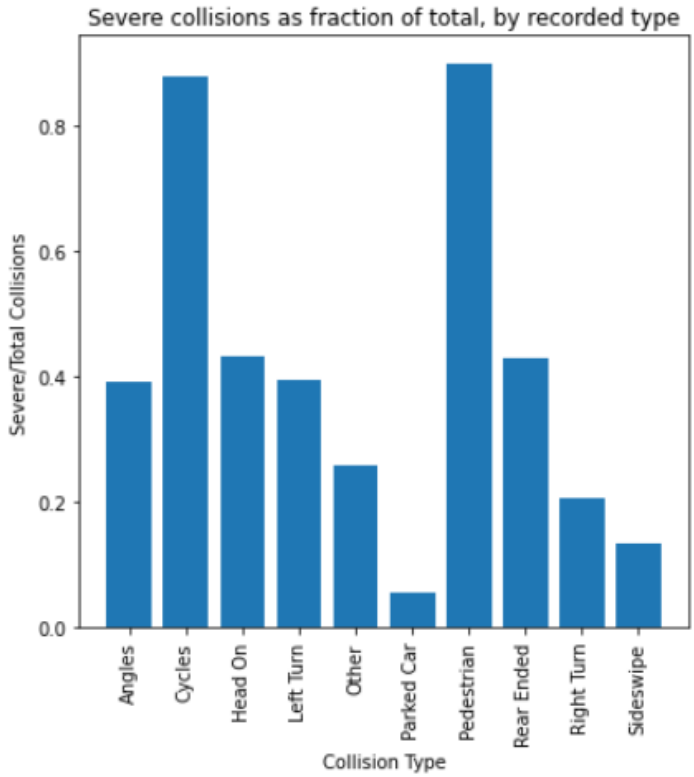

Figure 3: severe collision rate by collision type

Person, Vehicle and other Counts

Relationships between severity and the number of people, pedestrians, bicycles and vehicles involved were also studied. The distribution of collision types for different counts can be seen in Figure 4, below. It is clear that more injuries occur for collisions involving 1 or more pedestrians or bicycles, which confirmed what had already been observed when exploring severity by collision type. Interestingly, the rate of injury is higher for collisions involving 3 or more people, or something other than 2 vehicles (evidenced by the more balanced distributions). This can be explained for collisions involving 1 vehicle, which largely reflect those involving passengers and bicycles. Higher injury rates for collisions involving 3 or more people or cars could reflect incidents where the damage was more extensive (3+ vehicles) or where passengers riding in a vehicle were seated in locations more exposed to the force of the collision.

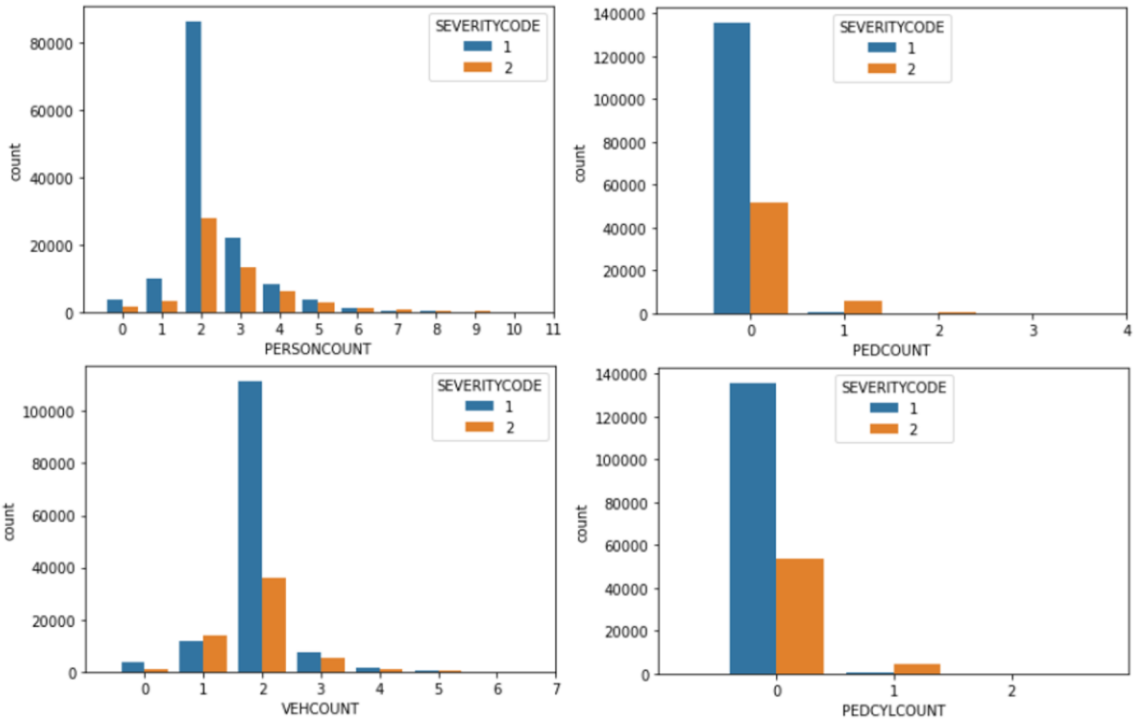

Figure 4: distribution of collision types for different counts of people, pedestrians, bicycles and vehicles

Environmental Factors and Driver Distractions

Lastly, several features describing environmental conditions and recorded driver influences were explored. Surprisingly, poor weather or road conditions did not result in disproportionately more injury collisions. In fact, for some characteristics such as snowy weather or icy road conditions, the proportion of collisions resulting in injuries was less than for "default" (sample majority) conditions. This could be explained by drivers observing extra caution during episodes of inclement weather (e.g. by reducing speed), such that collisions that do occur are less dangerous. The volume of samples where the most severe weather and road conditions occur is also quite low. Therefore, the impact of weather was further explored by sorting samples into majority (clear/cloudy/overcast) and minority (rain/snow/inclement) cases and computing the injury rate. The results, in Table 3, show a negligible difference.

	Weather	Property	Injury	Total	inj_rate
0	0	94266	44588	138854	0.321114
1	1	23231	11584	34815	0.332730
2	2	14991	932	15923	0.058532

Table 3: injury rate for different weather conditions (0=clear/overcast, 1=rain/inclement, 2=other)

Separately, collisions involving inattentive driving, drug or alcohol influence, or speeding resulted in higher injury rates than when these characteristics were not present.

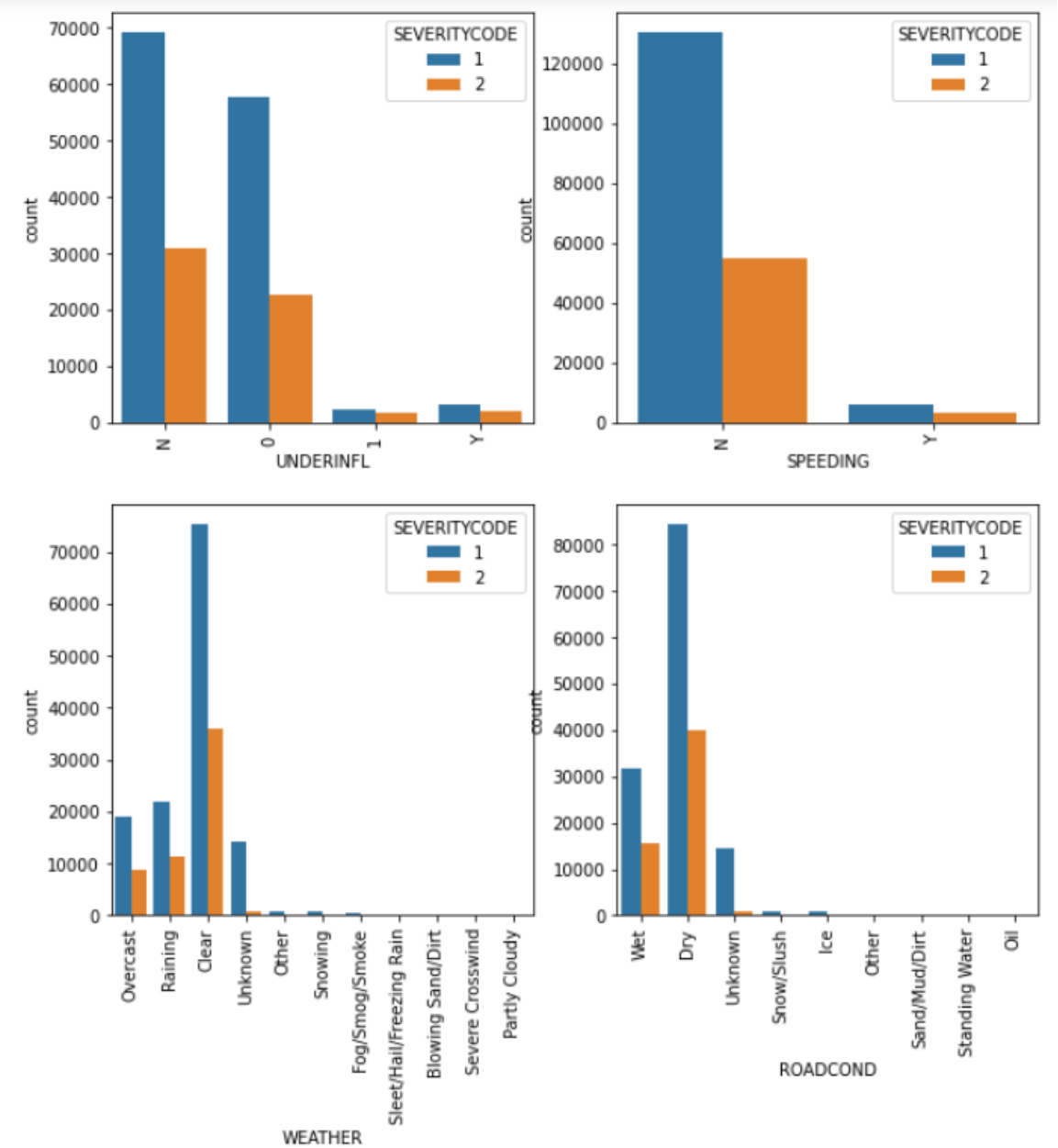

Figure 5: countplots showing distribution of collision types for different features categories

Final Feature Selection

After exploring these 12 features, those with the most significant impact on collision severity were collated for use in training the model. A summary of the findings for each feature is provided in Table 4, along with comments on any required processing or cleansing for the included parameters. Most notably, the INCDTTM, PERSONCOUNT and VEHCOUNT columns required transformation to one-hot encoding features that identified each sample with respect to a particular threshold. The remaining selected features were primarily Boolean fields that easily mapped to one-hot encoding.

Column Name	Description	Findings	Included & How
COLLISIONTYPE	A keyword describing the collision, eg 'head-on', 'angled', 'cycles', etc.	There are differences in the injury rate between different types of collisions, including significant increases for Pedestrian and Cycle types.	**Yes**: Collision type will be included using one-hot encoding.
PERSONCOUNT	Total number of people involved in the collision.	The proportion of injury collisions is higher when there are >3 people involved. May indicate more cars which could be more severe incidents, or more people in the vehicles, seated in locations that absorbed more force.	**Yes**: a simplified one-hot feature identifying if there are fewer (0) or 3 or more (1) people involved.
PEDCOUNT	The number of pedestrians involved in the collision.	Injuries occur much more frequently when there are 1 or more pedestrians involved. However, this relationship is already captured by the 'Pedestrian' collision type.	*No*: This would make a redundant feature and does not need to be included.
PEDCYLCOUNT	The number of bicycles involved.	Injuries occur much more frequently where bicycles are involved. However, again this is captured by the collision type impact.	*No*: This would be redundant.
VEHCOUNT	The number of vehicles involved in the collision.	Countplot shows that a higher portion of collisions involving something other than 2 cars result in injury. Further analysis showed that many collisions for 1 vehicle involve pedestrians or bikes, so the higher injury rate makes sense. Collisions involving 3 or more cars may have been more dangerous or severe, and also include more opportunities for injury.	**Yes**: a simplified one-hot feature identifying the number of vehicles as 2 (1) or not (0).
INCDATE/INCDTTM	Date and time recordings for the incident records.	The volume of total collisions and rate of injury collisions does not fluctuate much for the different days of the week, however, there are clear peaks in the rate of injury collisions during the typical rushhour periods.	**Yes**: datetime value will be used to create a one-hot encoding feature that captures whether the collision occured during rushhour (1) or not (0).
INATTENTIONIND	If collision was due to inattention.	There is a slightly higher rate of injury collisions among samples where inattentive driving was identified.	**Yes**:this is already a Y/N feature, which can be converted to 0,1 coding.
UNDERINFL	If driver was under influence of drugs/alchohol.	There is a higher rate of injury collisions among samples where drugs or alcohol were involved.	**Yes**:this is a Y/N feature which will be converted to 0,1 coding. Some cleansing will be required since currently both Y/N and 0/1 formats are used. Also, there are ~5000 blank records that can be dropped.
WEATHER	Weather at the time of incident.	More severe weather conditions did not show injury rates out of proportion with the total population.	*No*
ROADCOND	Condition of road at the time of incident.	More dangerous or road conditions did not show injury rates significantly out of proportion with the total population.	*No*
LIGHTCOND	Light conditions at the time of incident.	The proportion of injury collisions did not stand out for any of the different lighting categories with meaningful volume.	*No*
SPEEDING	If speeding was a factor in the collision.	The proportion of injury collisions was 1.28 times higher when speeding was involved.	**Yes**:this is a Y/N feature which will be converted to 0,1 coding.

Table 4: identification and discussion of selected model features

Data Cleansing and Processing

After selecting and compiling the desired modeling features, several actions were taken to cleanse the dataset of incomplete records. Figure 6, below, shows that the blank values were isolated to the COLLISIONTYPE and UNDERINFL columns. These blanks overlapped for all but one sample, which limited the impact of removing these records.

```
In [7]: #Lets check how many blank values there are in each column:
        df.isnull().sum(axis=0)

Out[7]: DATETIME            0
        COLLISIONTYPE    4904
        PERSONCOUNT         0
        VEHCOUNT            0
        INATTENTIONIND      0
        UNDERINFL        4884
        SPEEDING            0
        SEVERITYCODE        0
        dtype: int64
```

Figure 6: code snippet and output showing count of blank values by feature (column)

Additionally, exploration of the PERSONCOUNT feature identified over 5500 cases with a recorded count of 0. All but 5 of these samples had a VEHCOUNT of at least 1, further suggesting this value as a

recording error. Since PERSONCOUNT was selected as a model feature, these records were also removed to avoid any patterns that might be introduced by incorrect samples.

```
#looking at the distribution for PERSONCOUNT,
df['PERSONCOUNT'].value_counts()
```

```
2     111386
3      35138
4      14445
1      11727
5       6584
0       5541
6       2702
7       1131
8        533
9        216
```

```
#what is the VEHCOUNT distribution when the PERSONCOUNT = 0
df[df['PERSONCOUNT']==0]['VEHCOUNT'].value_counts()
```

```
2     4267
1      771
3      390
4       80
5       19
6        5
0        5
7        2
11       1
9        1
```

Figure 7: code snippets showing the number of questionable samples with a PERSONCOUNT of 0

Lastly, over 25,000 samples lacked any recorded time component, resulting in a default timestamp of '00:00:00.' Again, since incident time was selected for the model (transformed as rush-hour Y/N) these samples were removed to avoid undue influence. Figure 9, which shows the distribution of these removed samples by year, suggests reporting or data management practices were improved starting in 2010.

	DATETIME	COLLISIONTYPE	PERSONCOUNT	VEHCOUNT	INATTENTIONIND
0	2013-03-27 14:54:00	Angles	2	2	0
1	2006-12-20 18:55:00	Sideswipe	2	2	0
2	2004-11-18 10:20:00	Parked Car	4	3	0
3	2013-03-29 09:26:00	Other	3	3	0
4	2004-01-28 08:04:00	Angles	2	2	0
5	2019-04-20 17:42:00	Angles	2	2	0
6	2008-12-09 00:00:00	Angles	2	2	0

Table 5: data snapshot with row 6 showing example of missing time value (defaults to '00:00:00')

```
#now lets create a new df with the samples that fall between a time of 00:00:00 and 00:00:01
#this represents midnight and one second after midnight, and since the timestamp values only have hours and minutes
#it should only capture those samples where not time originally existed and the timestamp defaulted to 00:00:00
df2 = df.between_time('00:00:00','00:00:01')
print(df2.shape)
df2.head()
```

```
(25526, 10)
```

Figure 8: code snippet showing count of records where timestamp is '00:00:00'

```
#lets look at the distribution of the removed samples by year...
df2.groupby(df2.DATETIME.dt.year).count()['SEVERITYCODE']
```

```
DATETIME
2004    3583
2005    4439
2006    4598
2007    4335
2008    4073
2009    3399
2010     126
2011     626
2012      24
2013      65
2014      54
2015      38
2016      62
2017      31
2018      45
2019      23
2020       5
Name: SEVERITYCODE, dtype: int64
```

Figure 9: code snippet showing distribution of samples with missing time values by year

After cleansing the dataset, several columns were transformed to create more simplified features for the model. Person and vehicle counts were manipulated to create one-hot encoding features that identified samples according to specific numeric thresholds. The timestamp was used to identify whether each incident occur during or outside of the AM or PM rush-hour periods. Collision type was also converted to one-hot encoding. The final dataset, after cleansing and processing, included 16 features, 1 target variable and 158,701 samples.

```
#add columns: PCOUNT_OVER2, VCOUNT_IS2 to track whether there are 3 or more people and whether there are 2 vehicles or other
df['PCOUNT_OVER2'] = [1 if b>=3 else 0 for b in df['PERSONCOUNT']]
df['VCOUNT_IS2'] = [1 if b==2 else 0 for b in df['VEHCOUNT']]
df.head()
```

Figure 10: code snippet showing creation of one-hot encoding features for people and vehicle counts

Data Balancing

The last step to prepare the data for model training was balancing the number of samples from both the majority and minority classes. This was necessary in order to avoid biasing the model towards predicting all samples to the majority class. Since the volume of the minority class (injury collisions) was large, down-sampling was used on the majority class to create the final balanced dataset.

```
#lets observe how unbalanced the dataset is...
features_df['SEVERITYCODE'].value_counts()
```

```
1    110649
2     48052
Name: SEVERITYCODE, dtype: int64
```

Figure 11: value counts for each severity classification prior to balancing the data

	RUSHHOUR	PCOUNT_OVER2	VCOUNT_IS2	INATTENTIONIND	UNDERINFL	SPEEDING	Angles	Cycles	Head On	Left Turn	Other	Parked Car	Pedestrian	Rea Ende
0	1	0	1	0	0	0	0	0	0	0	0	1	0	
1	0	0	1	0	0	0	0	0	0	0	0	1	0	
2	1	0	0	0	0	0	0	0	0	0	0	0	1	
3	0	0	1	1	0	0	1	0	0	0	0	0	0	
4	0	0	1	0	0	0	1	0	0	0	0	0	0	

Table 6: snapshot of final feature-set showing one-hot encoding format

Modeling

Prediction models were built using four different classification algorithms, specifically 'k-nearest neighbors (KNN), decision tree, support vector machines (SVM) and logistic regression (LR). Classification was the best approach because of the nature of the dataset and the availability of a clear target variable with significant historic, labeled results. Prior to any modeling, the balanced dataset was split into train and test sub-sets, with the test set accounting for 25% of the data. Where parameter tuning was required, model performance was evaluated using simple accuracy or log-loss metrics. In addition to classification predictions, probabilities were also predicted for the test set using each algorithm to facilitate evaluation using the log-loss and area under the ROC curve (AUROC).

```
#using the test_train method, create test and train subsets of X and y
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=4)
print ('Train set:', X_train.shape,  y_train.shape)
print ('Test set:', X_test.shape,  y_test.shape)
```

```
Train set: (72078, 16) (72078,)
Test set: (24026, 16) (24026,)
```

Figure 12: code snippet showing preparation and size of train and test sub-sets

K-Nearest Neighbors

The optimum KNN model was found by iterating through different 'k' values and plotting the accuracy of the model in predicting the test set. Initially, 'k' values between 0 and 300, in increments of 15, were used to observe the model performance over a large range. Once the optimum range was identified, a second loop was completed to iterate through a local range of values. Despite this approach, an optimum k-value was difficult to locate. As shown in Figures 13 and 14, model performance was quite choppy, even for adjacent values of k. Ultimately, a k value of 250 was selected.

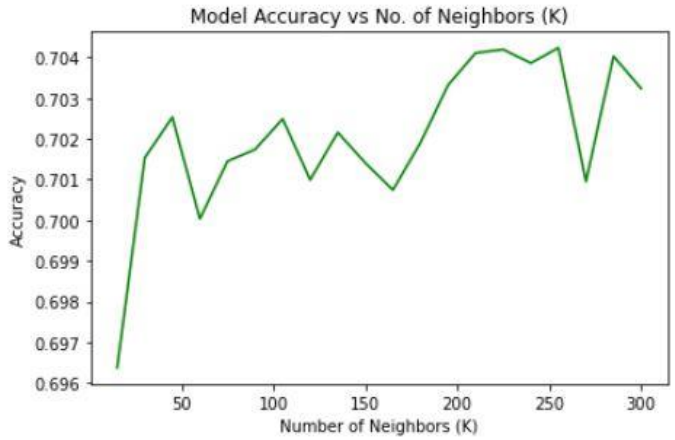

Figure 13: model accuracy for k-values between 15 and 300 (in steps of 15) showing optimum near 250

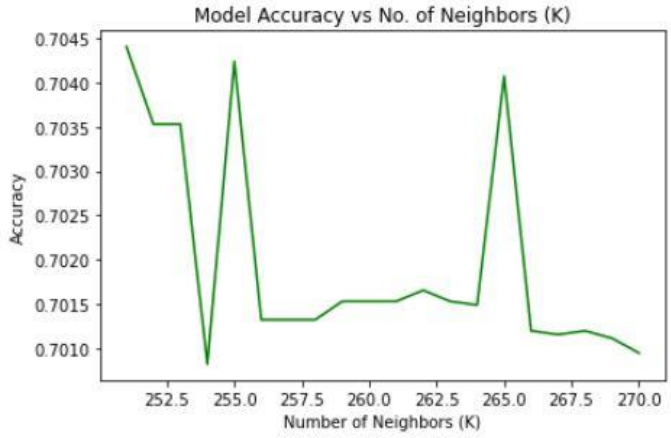

Figure 14: model accuracy for k-values between 251 and 270 showing choppiness between iterations

Decision Tree

A decision tree model was built using the RandonForest method from the Python Scikit-Learn package. The parameters of the model were automatically generated by the classifier tool and not pre-set.

```
#train the model on the train subsets
RFDT = RandomForestClassifier().fit(X_train,y_train)
DT_pred = RFDT.predict(X_test)
print('Accuracy for Decision Tree: ',metrics.accuracy_score(y_test,DT_pred))
```

```
Accuracy for Decision Tree:  0.7050695080329643
```

Figure 15: code snippet showing fit and testing of RandomForest Decision Tree.

Support Vector Machines

Four different SVM models were trained using linear, radial-based (rbf), polynomial and sigmoid kernel functions. The linear model was identified as the best performing model, as it combined the highest accuracy with the most balanced predictions between the two classes. The linear, ‘rbf’ and polynomial models achieve accuracies slightly above 0.70, but were biased towards predicting injury type collisions. The sigmoid model was more balanced, but only achieved an accuracy of 0.6299.

Logistic Regression

Logistic regression models were trained using 'liblinear', 'newton-cg' and 'sag' solver functions, and a regularization parameter, C, of 0.01. After selecting the 'liblinear' solver method, the C parameter was optimized by iterating through 50 values between 0.01 and 0.50, and plotting the log-loss performance of the model. This resulted in a C parameter of 0.14; however, the overall prediction accuracy was lower for the model with C of 0.14 versus 0.01, so the original parameter was kept.

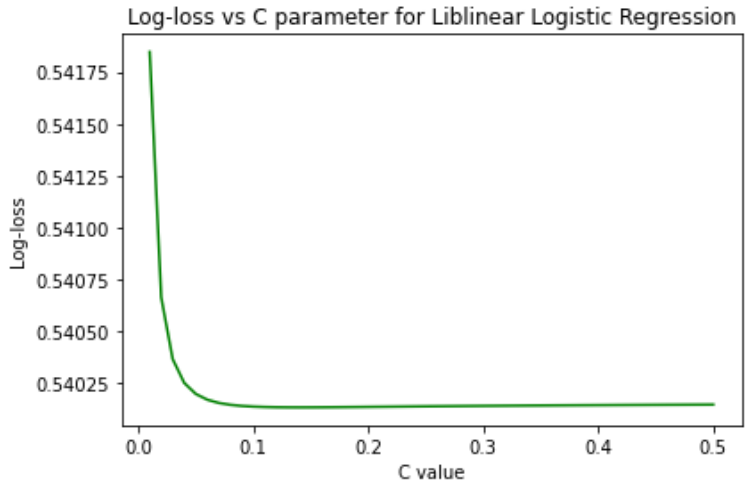

Figure 16: log-loss of 'liblinear' Logistic Regression models for different values of C (regularization)

Results

After tuning the parameters of each classification method, the optimized models were evaluated using several metrics, including simple accuracy score, Jaccard similarity, F1-score, log-loss and area under the ROC curve (AUROC). The Logistic Regression model performed best in 3 of the 5 metrics: accuracy, F1-score and AUROC. The KNN algorithm performed best in the other two: Jaccard and Log-loss. Each of the models achieved similar prediction accuracy on both the train and test datasets, indicating that the models were not too overfit to the training data.

	Algorithm	Model feature	Accuracy	Jaccard	F1-score	Log-loss	AUROC
0	KNN	k=250	0.704612	0.507870	0.541453	0.540904	0.786865
1	Decision Tree	Random Forest	0.705070	0.503503	0.541335	0.547774	0.788573
2	SVM	linear kernel	0.702572	0.506798	0.539207	0.605143	0.750791
3	Logistic Regression	liblinear, C=0.01	0.705985	0.500636	0.541839	0.541847	0.790964

Table 7: final results of optimized models for each classification algorithm

Discussion

The precise similarity in the evaluation metrics for the four different models was unexpected. This is likely a reflection of several things, the impact of specific features selected, and the size of the dataset versus the narrowness of the target classifications (24,000 samples predicted into only 2 categories). There must be some remaining disorder in the selected features that a prediction accuracy of around 70% is the maximum that can be achieved. The similarity is particularly interesting because there are

clearly some differences between the patterns and relationships that each of the models recognized. While all were more biased towards predicting injury type collisions, each did so with varying degrees of sensitivity. The KNN and SVM models were skewed by around 4500 predictions, while the Decision Tree and Logistic Regression models both predicted more than 5000 injury collisions in excess of property collisions. The best performing model by most metrics, Linear Regression, was in fact the most skewed, as shown in Figure 17.

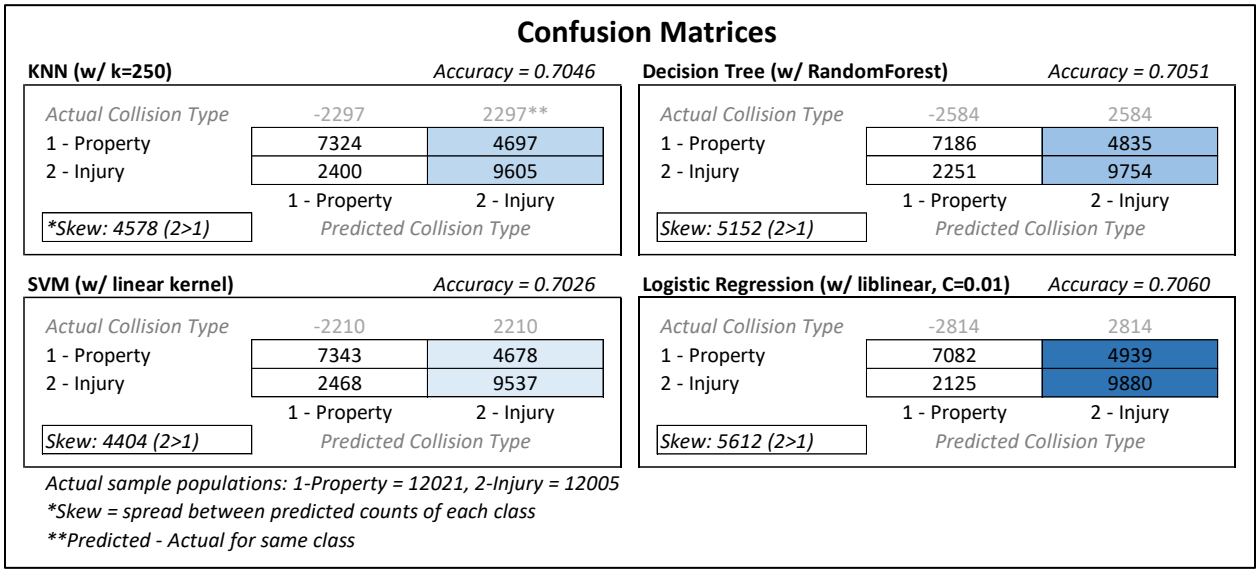

Figure 17: confusion matrix for optimized classification models

The stability of the metrics across each model despite these fluctuations in sensitivity suggest that each of the models provide a respectable fit; a model that simply predicted more samples to the injury class might have suffered a more significant decrease in accuracy. Indeed, the precision of the AUROC values – which evaluates the probability of a prediction rather than just the outcome – indicates that the models were relatively aligned on predictions for any given sample, and the exacty sensitivity pushed different amounts of more neutral cases into the injury classification.

Depending on the intended application or deployment of the model, such liberal prediction of severe collisions may not be undesirable. For example, if the goal is to predict severe collisions based only on preliminary incident details to choose the appropriate level of emergency response, false negatives (incorrectly predicted injury collisions) would be preferred to false positives (incorrect property prediction) in terms of the possible impact on human life. Alternatively, if the goal is to optimize more limited response resources, or to predict the proportion of severe collisions from modeled outcomes of a new traffic arrangement, then a more balanced model could be preferred.

One last consideration might be the relative speed of the models. Exact time consumption was not recorded; however, from anecdotal observation, the KNN, SVM and Decision Tree models took several minutes to fit, while the Logistic Regression models completed in only seconds. As more and more data is used to train the model, compute time will become a greater concern.

Conclusions

This study looked at developing a machine learning model to predict collision severity using data from the Seattle DOT on more than 190,000 incidents between January 2004 and May 2020. Sixteen meaningful features were extracted from the dataset after exploring their impact on collision severity. The features were then cleansed of incomplete or inaccurate records, and processed to create a matrix of one-hot encoded variables. A final balanced dataset of 48,000 samples from both target variable groups was used to train prediction models using four popular classification algorithms. Each of the models achieved prediction accuracies of around 0.70; however, they were biased towards predicting injury type (more severe) collisions. The models could be used by the Seattle DOT to better understand what factors cause more severe collisions to occur, or facilitate the development of calculations to quantitatively evaluate what initiatives will have the greatest impact on improving roadway safety by reducing severe collisions.

While the models were able to predict outcomes with relatively high accuracy, the bias towards injury type collisions and extreme similarity between the model performance metrics are both curious. Further exploration of the dataset could be conducted to determine if more relevant parameters exist, or if the selected parameters have improperly influenced the model. It would be interesting to apply this model to data from other sources or DOTs to evaluate the usability and effectiveness. Additional collision characteristics – not available in the applied dataset – could also be explored, such as vehicle make and model, posted speed limit versus actual vehicle speeds and whether or not airbags were released. There are many opportunities to continue exploring the prediction of collision severity.