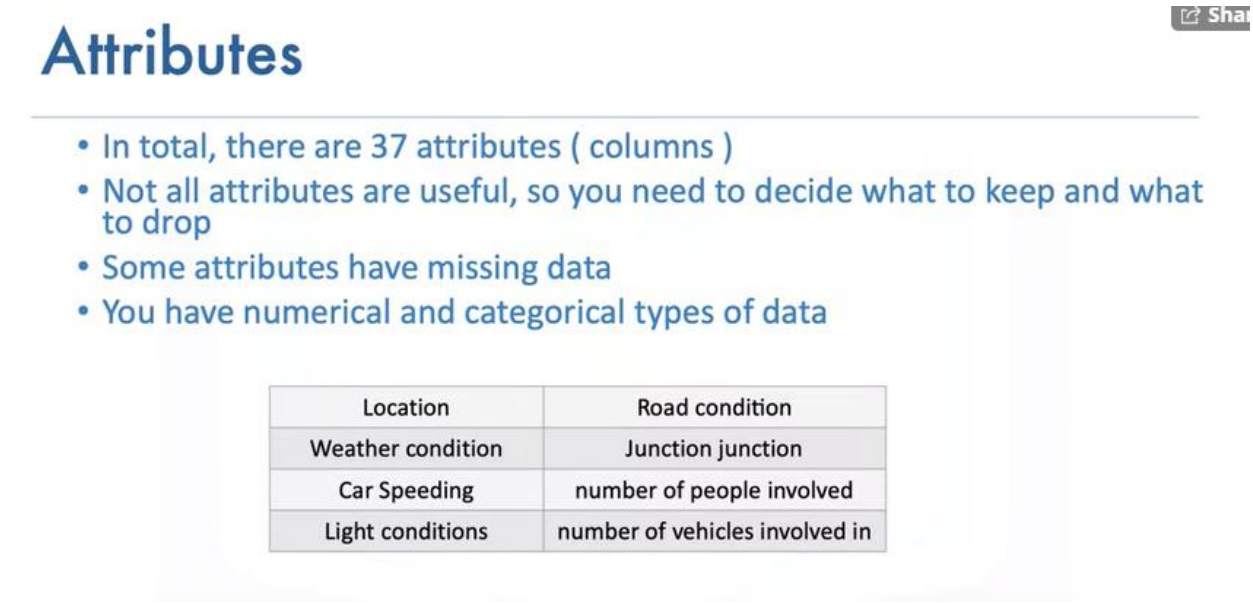
DATA science project –

Prediction of Accident Severity

By Narendra Jayaram
Dated 9/13/2020



Business Understanding:

The initial phase is to understand the project's objective from the business or application perspective. Then, you need to translate this knowledge into a machine learning problem with a preliminary plan to achieve the objectives.

Data understanding:

In this phase, you need to collect or extract the dataset from various sources such as csv file or SQL database. Then, you need to determine the attributes (columns) that you will use to train your machine learning model. Also, you will assess the condition of chosen attributes by looking for trends, certain patterns, skewed information, correlations, and so on.

Data Preparation:

The data preparation includes all the required activities to construct the final dataset which will be fed into the modeling tools. Data preparation can be performed multiple times and it includes balancing the labeled data, transformation, filling missing data, and cleaning the dataset.

Modeling:

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.

Evaluation:

Before proceeding to the deployment stage, the model needs to be evaluated thoroughly to ensure that the business or the applications' objectives are achieved. Certain metrics can be used for the model evaluation such as accuracy, recall, F1-score, precision, and others.

Prediction of Accident Severity

1. **Business Understanding:**

Seattle's plan to end traffic deaths and serious injuries on city streets by 2030.

What if each of us asked ourselves: what would be an acceptable number of traffic deaths for my family?

The answer is zero. You wouldn't even think twice about it. Now take that to the next level - what's an

acceptable number for your neighborhood? For our city?

While Seattle is one of the safest cities in the country, we still see more than 10,000 crashes a year, resulting in an average of 20 people losing their lives and over 150 people seriously injured. These are our friends, neighbors, and family members.

The thing is, traffic collisions aren't accidents - they're preventable through smarter street design, targeted enforcement, partnerships, and thoughtful public engagement. Together, we can make Seattle's streets safer for everyone.

Link

Problem Definition:

The City of Seattle and our partners are not satisfied with recent trends and we are prepared to take swift action to reduce collisions and save lives. The data is clear that our top priorities for the Vision Zero program moving forward must focus on the High Injury Network and pedestrian safety. The actions outlined in this document, along with our existing programs and maintenance efforts, will advance our safety goal. Vision Zero will always be about more than data. Each year, more than 180 lives were changed by traffic collisions. And the ripple effect is so much greater, as each person has a family, circle of friends, and larger community that continues to be impacted by loss, grief, and in many instances, long-term recovery.

1) To model the prediction of severity of an accident based on factors that are included in such an event.
2) This project will look at predicting the probability and severity of vehicular accidents based on weather and other characteristics, using historic collision data.
3) The thing is, traffic collisions aren't accidents - they're preventable through smarter street design, targeted enforcement, partnerships, and thoughtful public engagement. Together, we can make Seattle's streets safer for everyone

INTRODUCTION:

What's new?

- In February 2020, we released phase 2 of our **Bike and Pedestrian Safety Analysis**, to look at bicycle and pedestrian incident trends. This tool helps us proactively make safety enhancements across the city. This groundbreaking approach helps us prioritize locations, anticipate issues, and

make decisions informed by data.

- On December 10, Mayor Jenny A. Durkan **announced a series of steps to improve safety on City streets** and reaffirm the City's commitment to achieving the Vision Zero goal of ending traffic deaths and serious injuries by 2030. Mayor Durkan announced the City will **reduce speed limits to 25 miles per hour (mph) throughout the city**, **double the number of safety-enhanced traffic signals**, invest in engineering changes to create safer streets, create a new crash review task force, and launch additional traffic safety education and enforcement tactics. Read our **2019 update** to learn more.

  While we're excited to implement these steps, we also want to take a moment to remind everyone that we all have a role to play in improving safety. As you're traveling Seattle's streets, look out for yourself and for each other. Recognize that every intersection is unique, so stay alert. If you're driving, pay attention, slow down, and expect people are walking and biking in every part of the city at all times of the day.

Tasks Included in current notebook:

- Loading of data.
- Preprocessing.
- Visualizations.

++

2. Data understanding:

Data 2018 was one of the lowest years on record for transportation-related fatalities. Still, 14 people lost their lives last year and more than 170 people sustained serious injuries. As of December 2019, twenty-five people have been killed on our streets and another 153 people have been seriously injured. Impairment, distracted driving, speeding, and failure to stop for pedestrians continue to be the top contributing causes to collisions in Seattle. Pedestrians account for the majority of victims in fatal collisions in Seattle. Pedestrians make up just 4% of total collisions annually but more than 50% of fatalities. And we know that older pedestrians are more at risk. In 2019, the median age of pedestrians killed in collisions is 62 years old. More than 90% of severe collisions occur on our arterial streets.

Arterials carry the highest volumes of people walking, biking, driving, rolling, or riding transit and more than 80% of our arterial streets have speed limits higher than 30 miles per hour.

This project will look at predicting the probability and severity of vehicular accidents based on weather and other characteristics, using historic collision data.

Data Set Summary Data Set Basics Title Collisions—All Years Abstract All collisions provided by SPD and recorded by Traffic Records. Description This includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. Timeframe: 2004 to Present. Supplemental Information Update Frequency Weekly Keyword(s) SDOT, Seattle, Transportation, Accidents, Bicycle, Car, Collisions, Pedestrian, Traffic, Vehicle Contact Information Contact Organization SDOT Traffic Management Division, Traffic Records Group Contact Person SDOT GIS Analyst Contact Email DOT_IT_GIS@seattle.gov

Stakeholders:

- Public Development Authority of Seattle
- Transport authority

The dataset used for this project is based on car accidents which have taken place within the city of *Seattle, Washington* from the year *2004* to *2020*. This data is regarding the *severity of each car accidents* along with the time and conditions under which each accident occurred. Records are assigned to two groups – 'property damage only' and 'injury' collisions – which makes it ideal for a supervised learning classification model. There are approximately 195,000 samples in the dataset

The model aims to predict the severity of an accident, considering that, the variable of Severity Code was in the form of *1* (Property Damage Only) and *2* (Physical Injury)

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

These features identify: •

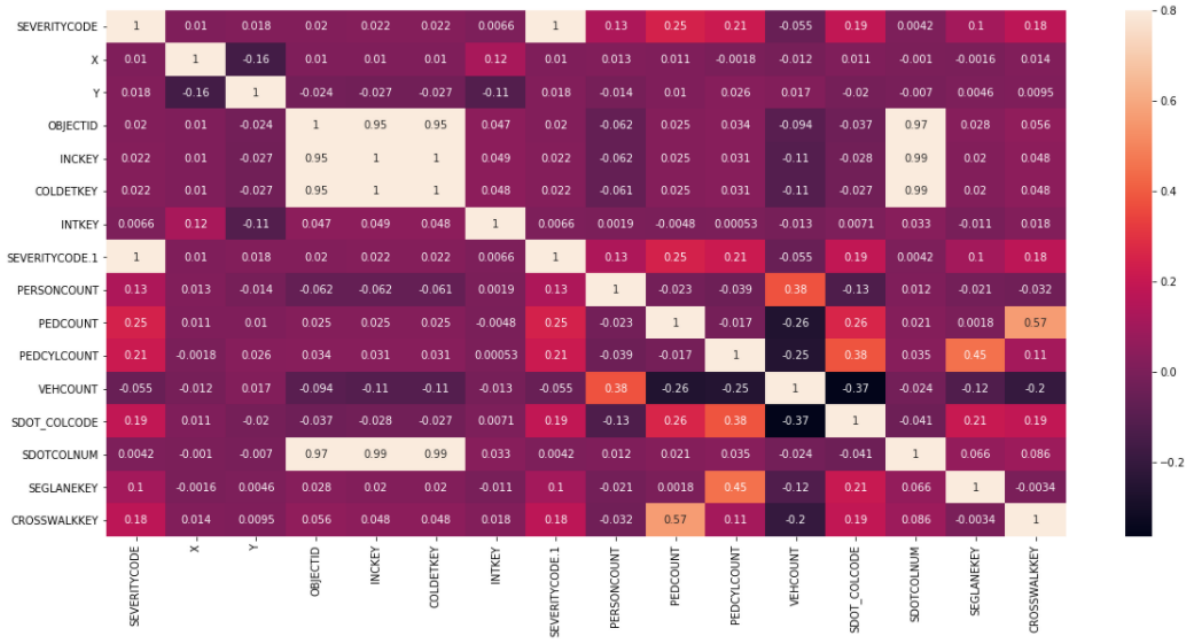
- the date, time, position and location details of the collision; •
- codes used by the state and DOT to categorize the collision; •
- characteristics of the incident, such as type of collision and the number of people, pedestrians, bicycles and vehicles involved; •
- environmental factors like weather and road condition; and •

- the existence of driver behaviors including speeding, inattentiveness and the presence of drugs or alcohol

A model for predicting severity will be built using historic collision records collected by the Seattle Police Department and maintained by the SDOT Traffic Management Division

```
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
Ratio of Collisions with injuries to those with property damage: 0.43
```

Correlation of data



Below are the count of null values in

```
PEDROWNOTGRNT     190006
EXCEPTRSNDESC     189035
SPEEDING          185340
INATTENTIONIND    164868
INTKEY            129603
EXCEPTRSNCODE     109862
SDOTCOLNUM         79737
JUNCTIONTYPE        6329
Y                   5334
X                   5334
LIGHTCOND           5170
WEATHER             5081
ROADCOND            5012
COLLISIONTYPE       4904
ST_COLDESC          4904
UNDERINFL           4884
LOCATION            2677
ADDRTYPE            1926
ST_COLCODE            18
dtype: int64
```

Categorical features

```
Index(['REPORTNO', 'STATUS', 'ADDRTYPE', 'LOCATION', 'EXCEPTRSNCODE',
       'EXCEPTRSNDESC', 'SEVERITYDESC', 'COLLISIONTYPE', 'INCDATE', 'INCDTTM',
       'JUNCTIONTYPE', 'SDOT_COLDESC', 'INATTENTIONIND', 'UNDERINFL',
       'WEATHER', 'ROADCOND', 'LIGHTCOND', 'PEDROWNOTGRNT', 'SPEEDING',
       'ST_COLCODE', 'ST_COLDESC', 'HITPARKEDCAR'],
      dtype='object')
```

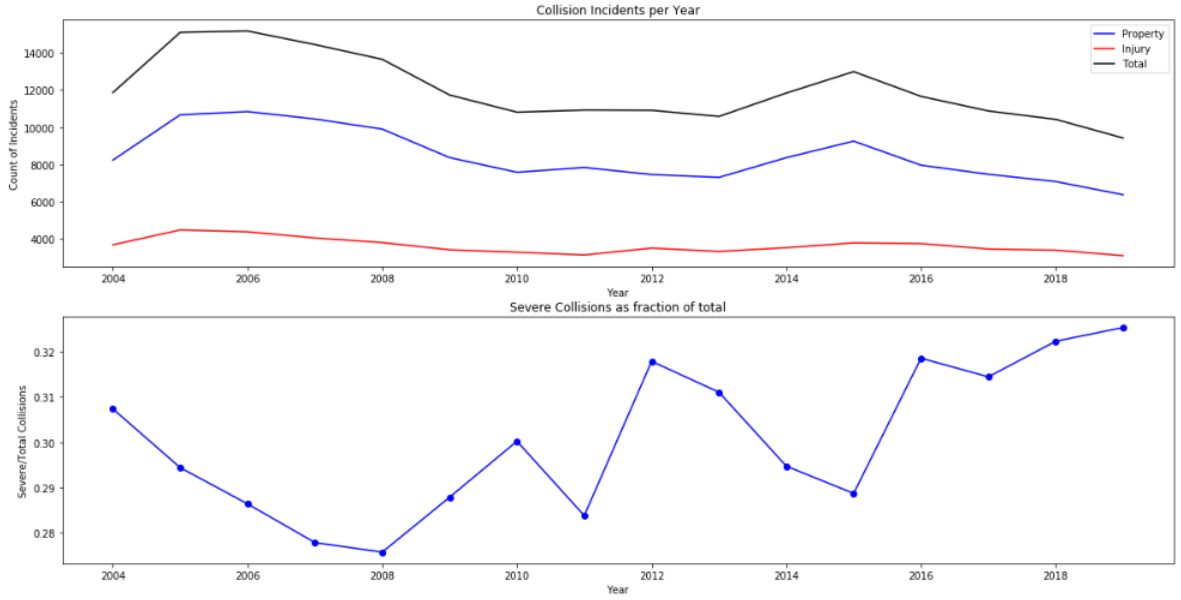
Numerical Features

```
Index(['SEVERITYCODE', 'X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'INTKEY',
       'SEVERITYCODE.1', 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT',
       'SDOT_COLCODE', 'SDOTCOLNUM', 'SEGLANEKEY', 'CROSSWALKKEY', 'DATETIME'],
      dtype='object')
```

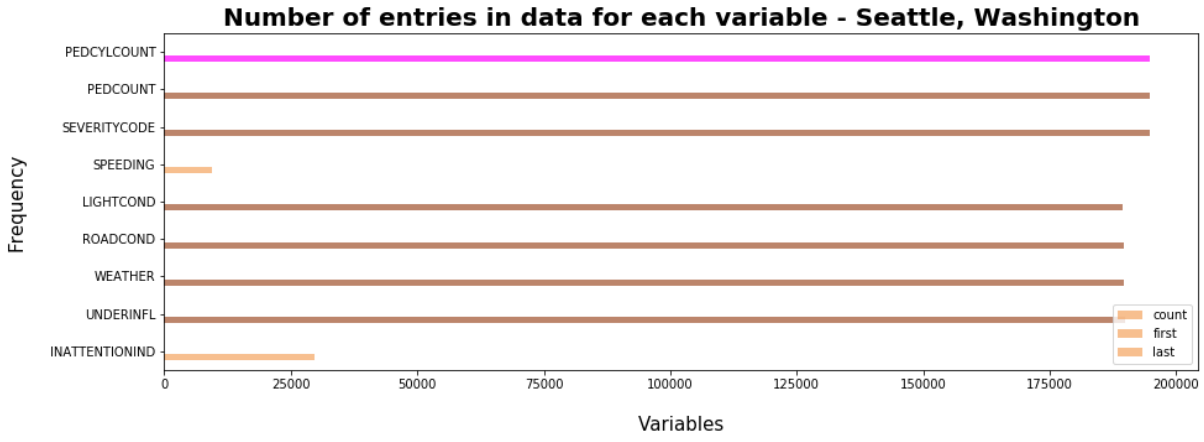
```
Numerical features : 16
Categorical features : 22
```

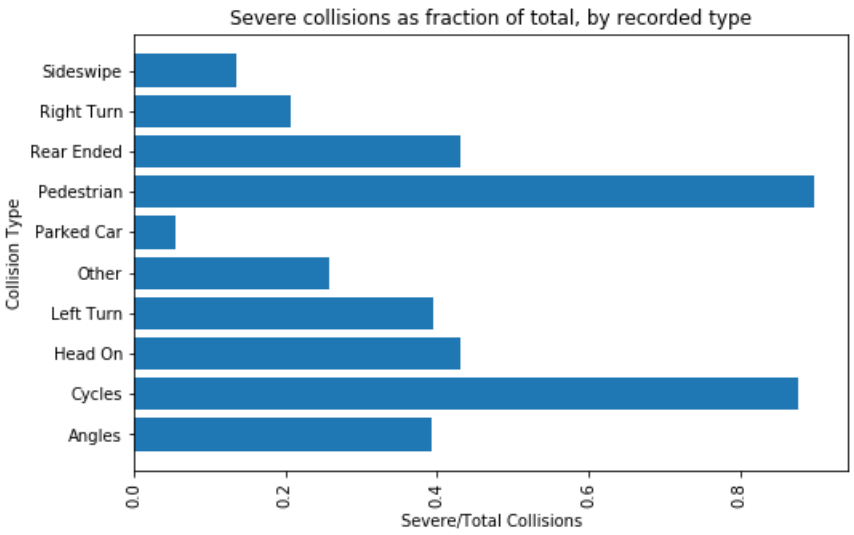
DATA CLEANING

3. EDA – Exploratory Data Analysis

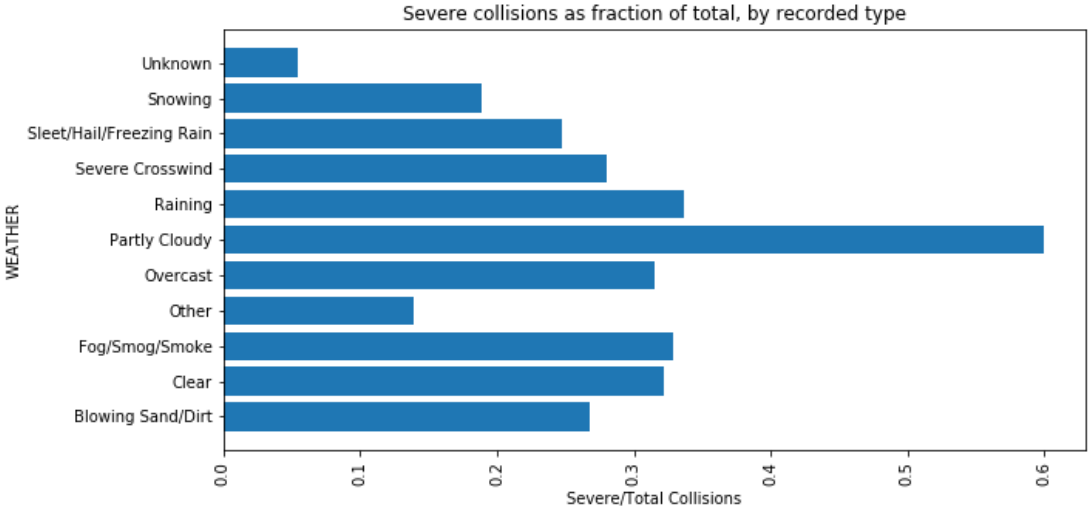


As we can see from the above chart, while the number of total incidents has decreased every year since 2015, the number of incidents where injuries occur has stayed mostly flat, accounting for a higher percentage of the total volume.¶

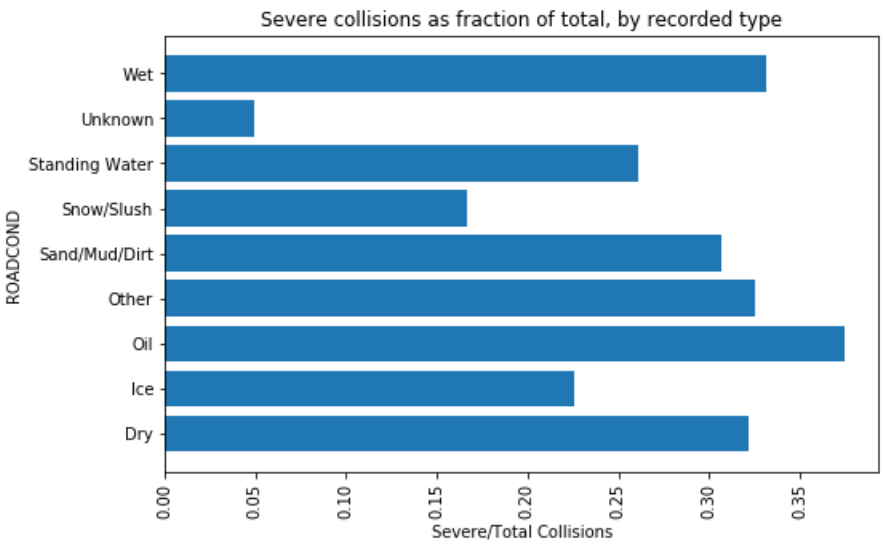




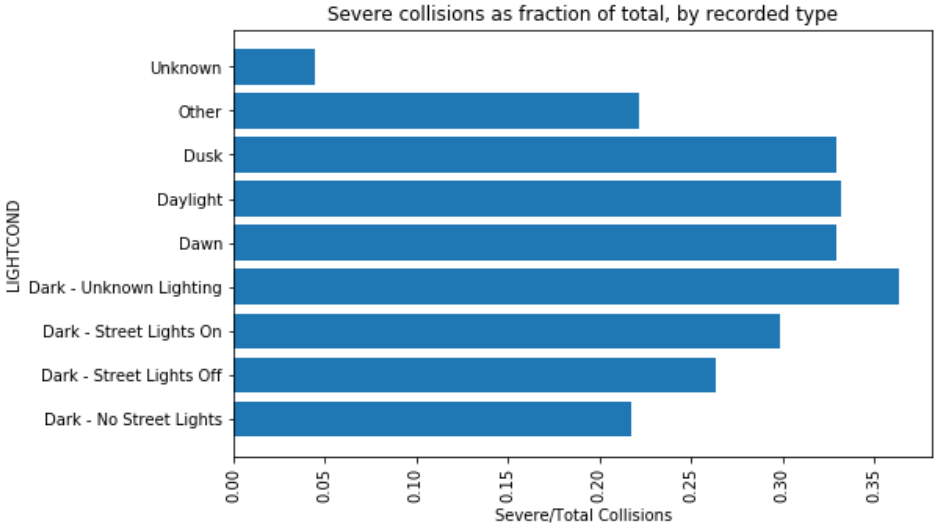
Pedestrian and Cycles have the high injury rate



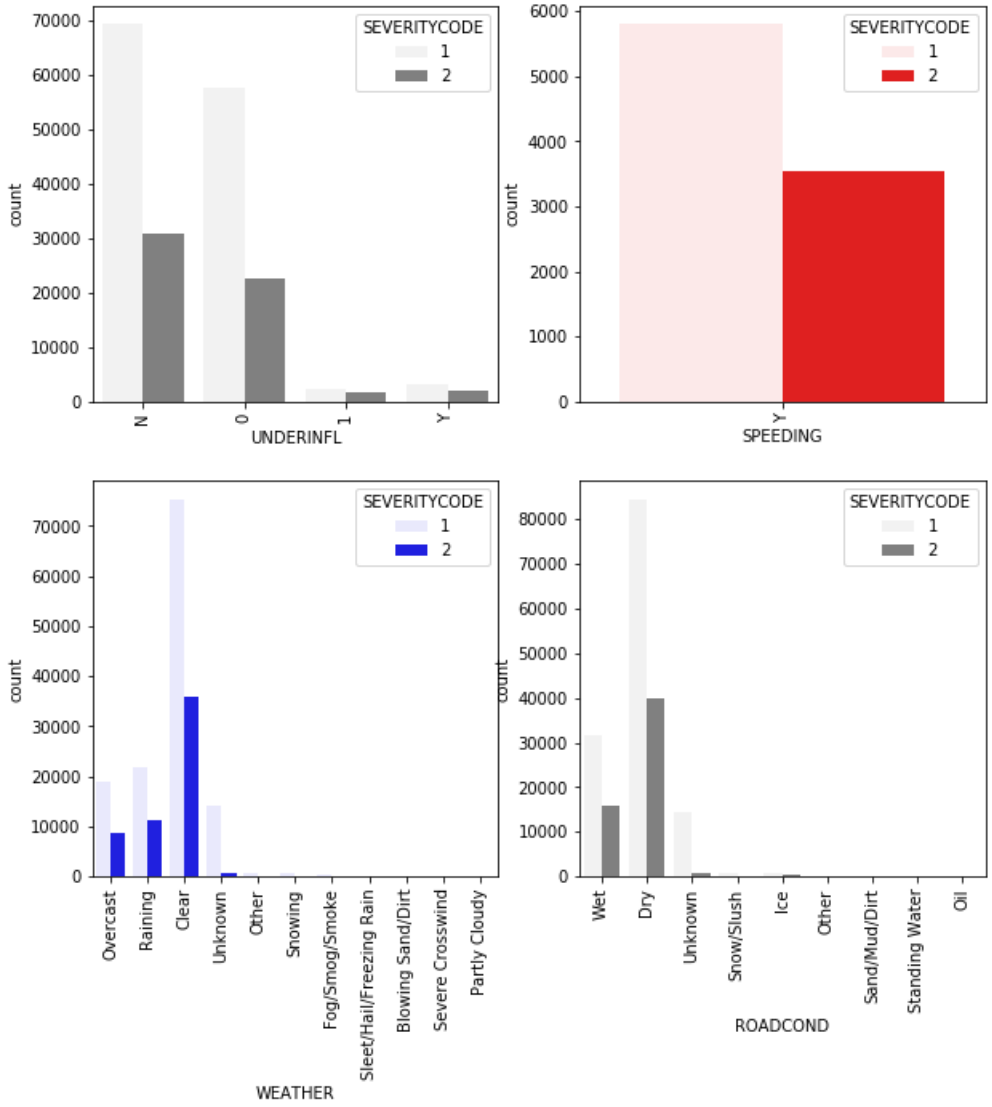
Partly cloudy has the high injury rate



oil and Wet Road has the high injury rate

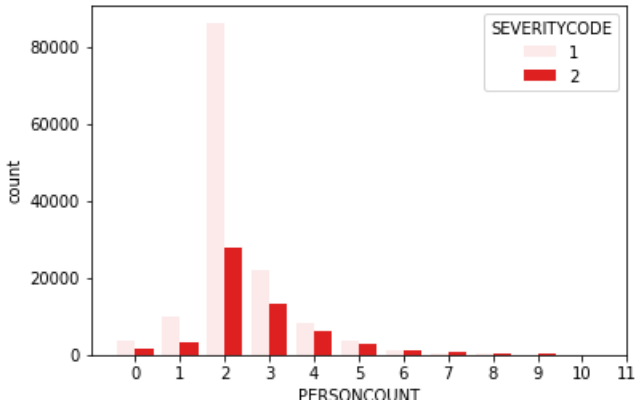


Dark-unknown lighting has the high injury rate

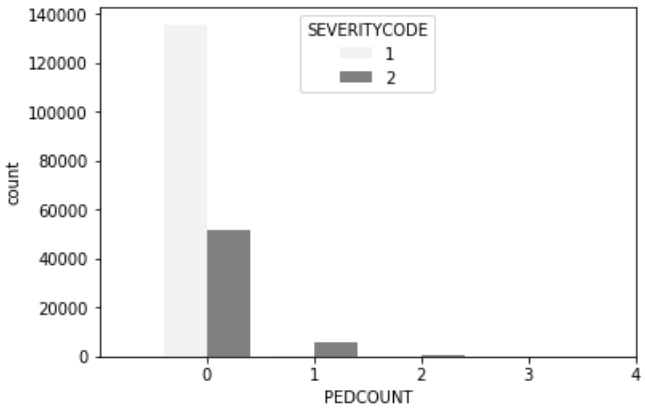

SEVERITYCODE
1
2
count
UNDERINFL
N
0
1
Y
SPEEDING
Y
WEATHER
Overcast
Raining
Clear
Unknown
Other
Snowing
Fog/Smog/Smoke
Sleet/Hail/Freezing Rain
Blowing Sand/Dirt
Severe Crosswind
Partly Cloudy
ROADCOND
Wet
Dry
Unknown
Snow/Slush
Ice
Other
Sand/Mud/Dirt
Standing Water
Oil

3.0 Data Cleansing and Preparation

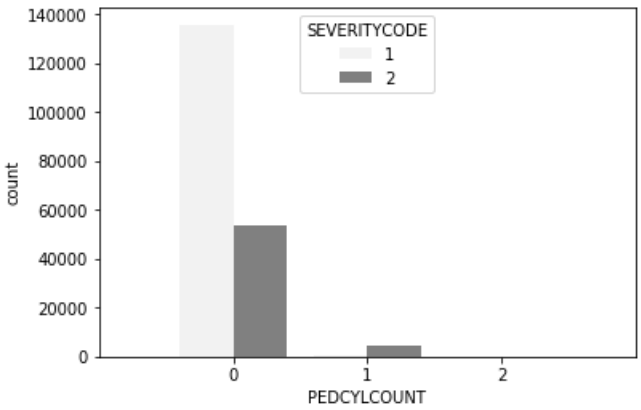
Based on the feature selection table above, the following data cleansing and prep will be completed:
drop PEDCOUNT / PEDCYLCOUNT / WEATHER / ROADCOND / LIGHTCOND
drop rows where COLLISIONTYPE, INATTENTIONIND, UNDERINFL or SPEEDING are blank
drop rows where PERSONCOUNT is 0
drop rows where the timestamp is 00:00:00
create column to identify if time is during either rushhour (5-8 or 15-18), then drop DATETIME
create column to identify if PERSONCOUNT is >=3, then drop PERSONCOUNT
create column to identify if VEHCOUNT is 2 or not, then drop VEHCOUNT
for INATTENTIONIND, convert N to 0 and Y to 1
for UNDERINFL, convert N to 0 and Y to 1
for SPEEDING, convert N to 0 and Y to 1
transform COLLISION type to one-hot encoding using get_dummies



>>First, clean the data to remove samples where the count of people is 0. Then, transform this feature to identify if there are more or less than 3 people involved (1 or 0, respectively) and use it in the model.¶

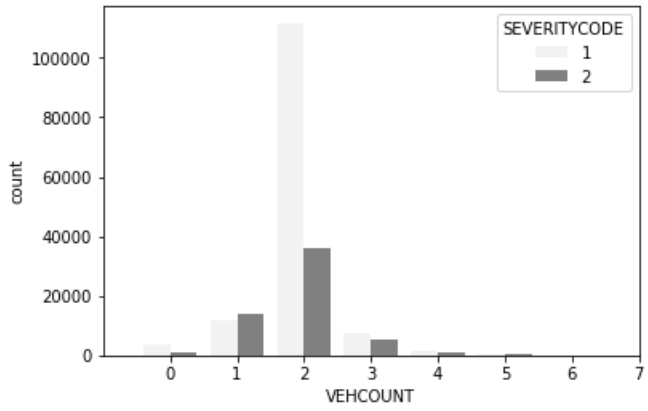


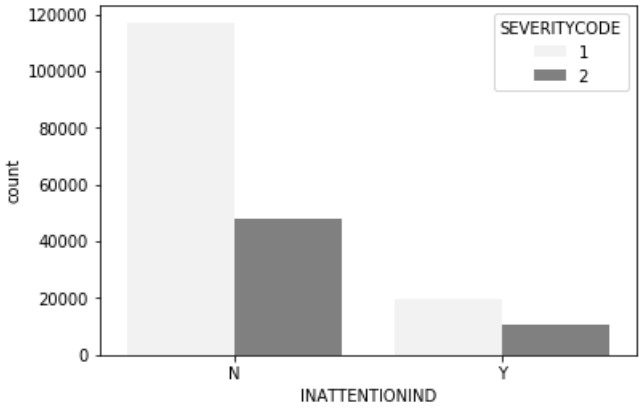
when the pedcount is >1 (at least one pedestrian is involved) injuries are more likely to occur. However, this is already captured by the 'Pedestrian' collision type, which is already in the feature set.¶



When bicycles are involved, injuries are more likely to occur.

> The 'Cycles' COLLISIONTYPE already introduces this relationship.

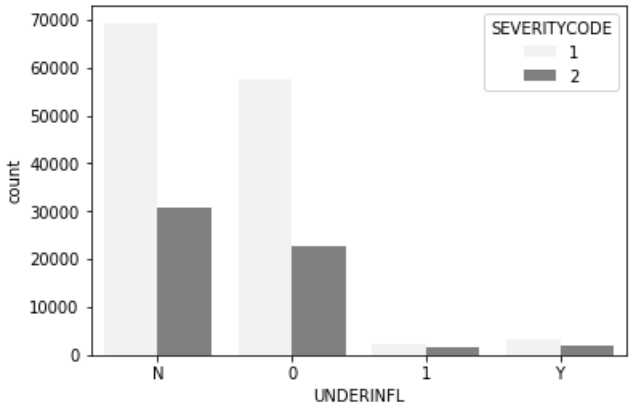




(INATTENTIONIND) Observation: collisions due to inattentive driving are slightly more likely to result in injuries than those where inattentiveness is not an issue; however, neither are more likely to result in injury than not.¶

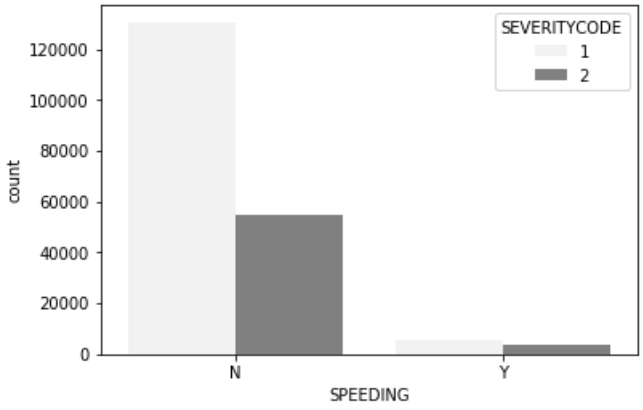
Injuries occur in a higher percentage of collisions where inattentive driving is present,

so lets use this as a feature. (INATTENTIONIND)¶



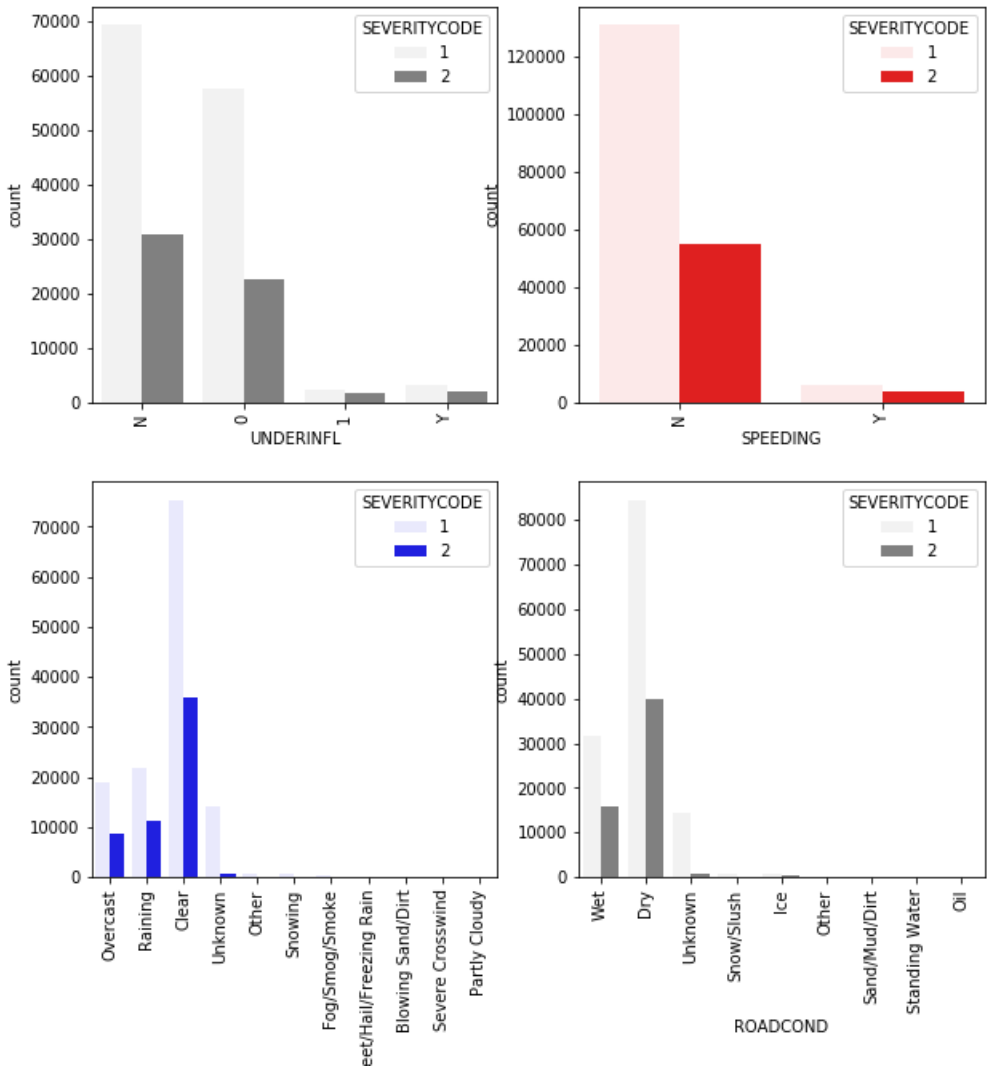
(UNDERINFL pt2) Observation: collisions where drugs or alcohol were involved resulted in injury at a higher rate than collisions where no drugs or alcohol were involved.¶

Lets include this as a feature. (UNDERINFL)



SPEEDING will be selected as a feature

pedestrian and bicycle count showed obvious associations with injury rate, but these would make redundant features because collision type already captures these scenarios



Summary

1. pedestrian and bicycle count showed obvious associations with injury rate, but these would make redundant features because collision type already captures these scenarios
2. meanwhile, contrary to expectation, Weather, Road and Light conditions didnt show remarkable patterns with respect to the rate of injury in collisions
3. featuress included - 'DATETIME', 'COLLISIONTYPE','PERSONCOUNT','VEHCOUNT','INATTENTIONIND','UNDERINFL','SPEEDING','SEVERITYCODE'

Observation: removing the samples where the PERSONCOUNT is recorded as 0 wont impact the proportionality of the target variable (SEVERITYCODE) and will still leave over 180k samples.. rather than have concern about the quality and impact of these records, lets remove them.¶

4. Balancing

Since there are more than 2x the number of property collisions (majority class) as there are injury collisions (minority class), the dataset should be balanced to avoid introducing bias to any models.

There are a number of ways to do this, including:

up-sampling the monority class
down-sampling the majority class
using other methods (Decision trees) or metrics (Area Under ROC Curve or Penalization Algorithms)
For this project, we will use a number of these methods.

The model will also be trained using a number of different methods, including a Decision Tree, in order to find the most effective model.
Area under the curve will also be used as an evaluation metric, in addition to accuracy score, jaccard index, f1 and logloss

5.0 Modeling¶

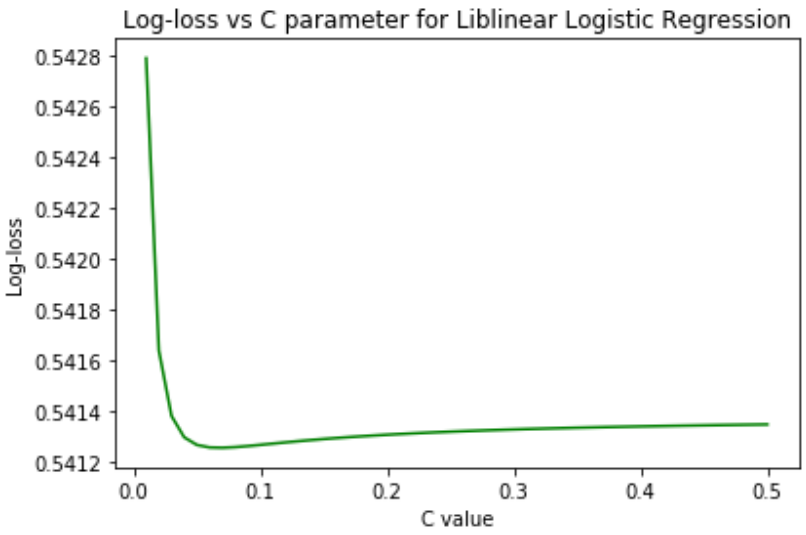
Now that the data is cleansed, processed and balanced, some Machine Learning models can be trained. Since this project is dealing with classification, several different methods will be used, including:

K Nearest Neighbor
Decision Tree
Support Vector Machines
Logistic Regression
First the data will be separated into train and test subsets. Some specific parameter tuning will be completed for each of the model types. Then the models will be evaluated using:

Simple accuracy score
Jaccard index
F1-score
Log-loss

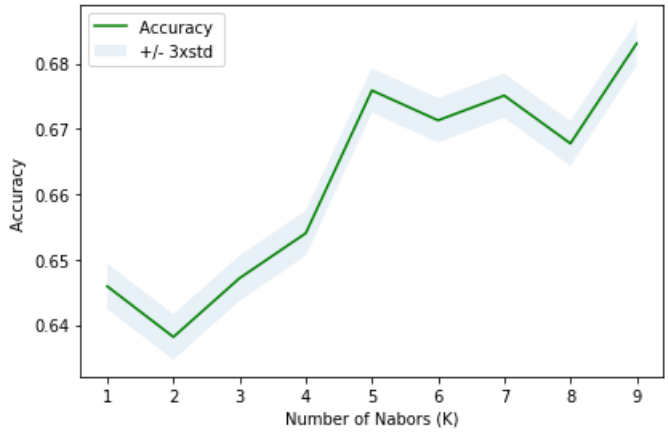
```
Jaccard similarity for liblinear Logistic Regression:  0.7042869034406215
Logloss for liblinear Logistic Regression:  0.5427897193198256
Test Accuracy for liblinear:  0.7042869034406215
[[ 8263  6121]
 [ 2405 12043]]
              precision    recall  f1-score   support

           1       0.77      0.57      0.66     14384
           2       0.66      0.83      0.74     14448

   micro avg       0.70      0.70      0.70     28832
   macro avg       0.72      0.70      0.70     28832
weighted avg       0.72      0.70      0.70     28832
```

Algorithm	Model feature	Accuracy	Jaccard	F1-score	Log-loss	AUROC	
0	Logistic Regression	liblinear, C=0.01	0.704287	0.492167	0.538934	0.54279	0.789685

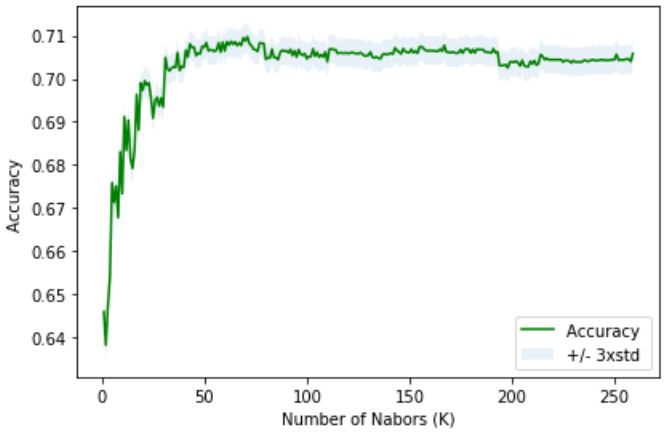



```
The best result is 0.5412554480316208 where C= 0.06999999999999999
```

Plot model accuracy for Different number of Neighbors



```
: array([0.65, 0.64, 0.65, 0.65, 0.68, 0.67, 0.68, 0.67, 0.68, 0.67, 0.69,
        0.68, 0.69, 0.68, 0.68, 0.68, 0.7 , 0.69, 0.7 , 0.7 , 0.7 , 0.7 ,
        0.7 , 0.7 , 0.69, 0.69, 0.7 , 0.69, 0.7 , 0.69, 0.7 , 0.7 , 0.7 ,
        0.7 , 0.7 , 0.7 , 0.71, 0.7 , 0.7 , 0.7 , 0.71, 0.71, 0.71, 0.71,
        0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71,
        0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71,
        0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71,
        0.71, 0.71, 0.7 , 0.7 , 0.7 , 0.71, 0.7 , 0.7 , 0.7 , 0.71, 0.71,
        0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71,
        0.7 , 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.7 ,
        0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71,
        0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71,
        0.7 , 0.71, 0.7 , 0.7 , 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71,
        0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71,
        0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71,
        0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.71,
        0.71, 0.71, 0.71, 0.71, 0.71, 0.71, 0.7 , 0.7 , 0.7 , 0.7 , 0.7 ,
        0.7 , 0.7 , 0.7 , 0.7 , 0.7 , 0.7 , 0.7 , 0.7 , 0.7 , 0.7 , 0.7 ,
        0.7 , 0.7 , 0.7 , 0.7 , 0.71, 0.71, 0.7 , 0.7 , 0.7 , 0.7 , 0.7 ,
        0.7 , 0.7 , 0.7 , 0.7 , 0.7 , 0.7 , 0.7 , 0.7 , 0.7 , 0.7 , 0.7 ,
        0.7 , 0.7 , 0.7 , 0.7 , 0.7 , 0.7 , 0.7 , 0.7 , 0.7 , 0.7 , 0.7 ,
        0.7 , 0.7 , 0.7 , 0.7 , 0.7 , 0.7 , 0.7 , 0.7 , 0.71, 0.7 , 0.7 ,
        0.7 , 0.7 , 0.7 , 0.7 , 0.7 , 0.71])
```




```
1 #confusion matrix & classification report
2 print(confusion_matrix(y_test,svm_pred1,labels=targetNames))
3 print(classification_report(y_test,svm_pred1))
```

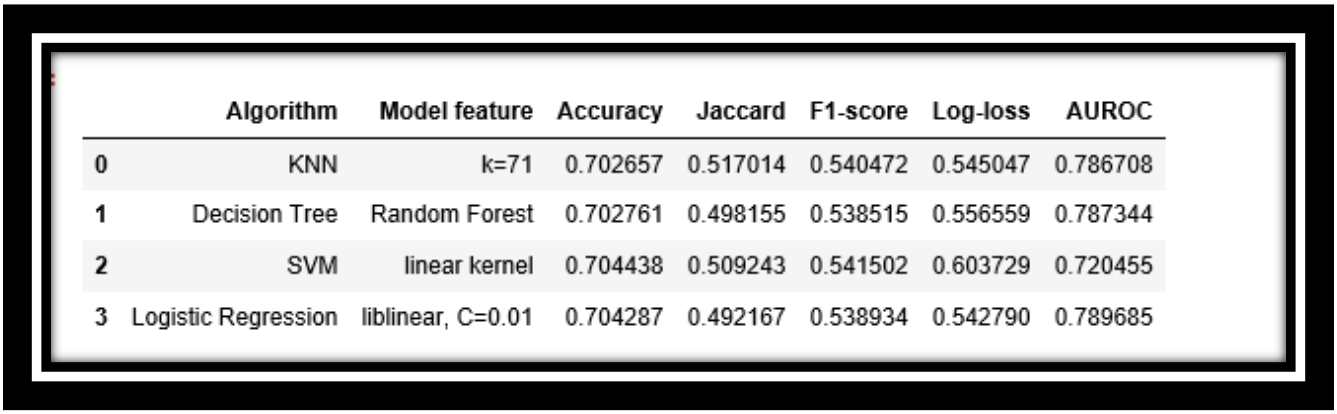
```
[[5895 3705]
 [1976 7645]]
              precision    recall  f1-score   support

           1       0.75      0.61      0.67      9600
           2       0.67      0.79      0.73      9621

   micro avg       0.70      0.70      0.70     19221
   macro avg       0.71      0.70      0.70     19221
weighted avg       0.71      0.70      0.70     19221
```

Algorithm	Model feature	Accuracy	Jaccard	F1-score	Log-loss	AUROC	
0	SVM	linear kernel	0.704438	0.509243	0.541502	0.603729	0.720455

6.0 Model Evaluation

In this section we will evaluate the optimized model of each classification method using:

Jaccard similarity F1-score Log-loss Area Under the ROC curve



	Algorithm	Model feature	Accuracy	Jaccard	F1-score	Log-loss	AUROC
0	KNN	k=71	0.702657	0.517014	0.540472	0.545047	0.786708
1	Decision Tree	Random Forest	0.702761	0.498155	0.538515	0.556559	0.787344
2	SVM	linear kernel	0.704438	0.509243	0.541502	0.603729	0.720455
3	Logistic Regression	liblinear, C=0.01	0.704287	0.492167	0.538934	0.542790	0.789685

Evaluation Summary:

The results are very similar for each of the different models.

The Logistic Regression approach performed best for 2 of the 5 metrics: simple accuracy,and AUROC F1 and log loss is SVM

while the KNN algorithm was the best according to Jacard.

the Logistic Regression model had the worst Jaccard similarity score of the 3 models.

Ultimately, the result are so similar that selecting a model might come down to its performance with respect to predicting against each classification. If being conservative - predicting fewer injury collisions is preferred -

then the SVM or KNN models might be best, whereas the Logistic Regression and Decision Tree model might be preferable if higher sensitivity is required for predicting injury type collisions.

Conclusions

- The thing is, traffic collisions aren't accidents - they're preventable through smarter street design, targeted enforcement, partnerships, and thoughtful public engagement. Together, we can make Seattle's streets safer for everyone
- Alternative features could be used from the same dataset
- Apply models to data from other sources or DOTs to evaluate effectiveness
- Explore impact of additional collision characteristics, such as vehicle make/model, posted vs actual vehicle speed, airbag release