Table of Contents

Introduction 2

Problem statement 2

Target audience 2

Scope and main aim 2

Data 3

Source 3

Analysis plan 5

Methodology 5

Data cleaning 5

Missing data 5

EDA 7

Daily trend 8

Yearly trend 8

Monthly trend 9

Week day trend 9

Correlation plot 10

Model development 10

Target variable 10

Splitting training and test set and balancing data 11

Classification models 11

Results 11

Discussion 11

Conclusion 12

Introduction

Problem statement

The deaths and injuries from traffic accidents are now a world phenomenon. Many countries all over the world are greatly concerned about the growth of mortality and injury rate on the road. According to WHO, every year the lives of approximately 1.35 million people are cut short as a result of a road traffic crash. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury.

Road traffic injuries cause considerable economic losses to individuals, their families, and to nations as a whole. These losses arise from the cost of treatment as well as lost productivity for those killed or disabled by their injuries, and for family members who need to take time off work or school to care for the injured. Road traffic crashes cost most countries 3% of their gross domestic product.

Target audience

Motorists, people who use public transport, police and medical personnel are usually inconvenienced greatly when accidents occur and especially severe accidents. For, severe accidents it takes quite some time and resources for the police and medical personnel to clear the scene. This thus leads to many hours of wait. Students are likely to be late for their classes, people going for interviews get late, those expected in important meetings may not be able to attend, surgeons and doctors expected to save lives run late, just to name but a few. Although severe accidents are unexpected, it would be possible to build a machine learning algorithm to predict severe accidents using various factors that are determined to be related to a severe accident occurring. This would thus serve as an automated reference to help people make decisions where and when there is a high chance of a severe accident occurring and hence plan accordingly.

Scope and main aim

Factors influencing accident frequency may vary from the ones affecting the severity; hence, it is suggested that their analysis should be performed carefully. A set of explanatory variables, which could include: driver attributes (e.g., age and gender, whether under influence of alcohol), vehicle features (e.g., body type, vehicle age and number of vehicles involved in the accident), road characteristics (e.g., number of lanes, road surface conditions, intersection control and types of road), weather conditions, day of week, time of day, speed limit and accident characteristics (e.g., accident’s main cause) have been shown to be possible predictors of accident severity.

The **aim of this project** is therefore to develop a machine learning algorithm that would warn motorists and other people using public transport where and when there is a high chance that a severe accident would occur and hence help them plan accordingly to avoid delays and inconveniences.

Data

Source

To evaluate the objectives of this project, a dataset by SDOT Traffic Management Divisions, Traffic Records Group in Seattle was downloaded using [this link.](#) This data is updated weekly and captures all types of collisions since 2004 to present. A record of severity of the collision, “SEVERITYCODE” is provided alongside other 39 variables related to a given collision. The dataset has a record of 221266 collisions.

Below is a snapshot of the first 5 rows of the dataset and the number of rows and columns at the end.

```
In [2]: df_collisions.head()
```

Out[2]:

	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	LOCATION	...	ROADCOND	LIGHTCOND	PEDROWNOTGRNT	SDOTCOLNUM	SPEEDING
0	-122.386772	47.564720	1	326234	327734	E984735	Matched	Intersection	31893.0	CALIFORNIA AVE SW AND SW GENESEE ST	...	Dry	Daylight	Y	NaN	NaN
1	-122.341806	47.686934	2	326246	327746	E985430	Matched	Intersection	24228.0	STONE AVE N AND N 80TH ST	...	Wet	Dark - Street Lights On	NaN	NaN	NaN
2	-122.374899	47.668666	3	329254	330754	EA16720	Matched	Block	NaN	NW MARKET ST BETWEEN 14TH AVE NW AND 15TH AVE NW	...	Dry	Daylight	NaN	NaN	NaN
3	-122.300758	47.683047	4	21200	21200	1227970	Matched	Intersection	24661.0	25TH AVE NE AND NE 75TH ST	...	Wet	Dark - Street Lights On	NaN	4160038.0	NaN
4	-122.313053	47.567241	5	17000	17000	1793348	Unmatched	Block	NaN	S DAKOTA ST BETWEEN 15TH AVE S AND 16TH AVE S	...	NaN	NaN	NaN	4289025.0	NaN

5 rows × 40 columns

```
In [3]: df_collisions.shape
Out[3]: (221266, 40)
```

Figure 1: First 5 rows and shape of data

See below also a list of the various names of columns provided.

```
In [10]: print(list(df_collisions))
['X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO', 'STATUS', 'ADDRTYPE', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE', 'EXCEPTRSNDESC', 'SEVERITYCOD
E', 'SEVERITYDESC', 'COLLISIONTYPE', 'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INJURIES', 'SERIOUSINJURIES', 'FATALITIES', 'INCDATE',
'INCDTTM', 'JUNCTIONTYPE', 'SDOT_COLCODE', 'SDOT_COLDESC', 'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND', 'PEDROWNOTGRNT', 'SDOT
COLNUM', 'SPEEDING', 'ST_COLCODE', 'ST_COLDESC', 'SEGLANEKEY', 'CROSSWALKKEY', 'HITPARKEDCAR']
```

Figure 2: List of names of columns

Columns that are not providing much information on the model such as ID, KEY etc will be dropped. For the purpose of this project, we will explore the various variables and select features that we would use in the prediction based on their importance in predicting severity of an accident. Through feature engineering more variables will be obtained to ensure good prediction of our model.

The target variable is '**SEVERITYCODE',** which captures the various categories of severity of the accident as follows; 0 – Unknown, 1 – prop damage, 2 – injury, 2b – severe injury and 3 – fatality. See below the description of the number and percentage of collisions in each category where we observe significance imbalance in the data. For instance the number of collisions with code 1 (prop damage) were the highest (137485), contributing to 62% of all the collisions.

```
In [11]: #number of collisions in each severity category
         df_collisions.SEVERITYCODE.value_counts()

Out[11]: 1     137485
         2      58698
         0      21635
         2b      3098
         3        349
         Name: SEVERITYCODE, dtype: int64

In [8]: #percentage of collisions in each severity category
        df_collisions.SEVERITYCODE.value_counts() / len(df_collisions)

Out[8]: 1     0.621356
        2     0.265283
        0     0.097778
        2b    0.014001
        3     0.001577
        Name: SEVERITYCODE, dtype: float64
```

Figure 3: Categories of target variable

Below is also the number of missing values for each column which will guide us in the process of selection of variables. We observe that in our target variable "**SEVERITYCODE**" we have one missing value. We would drop this row going forward. For the other columns a decision will be made on whether to impute or drop the missing values.

```
In [9]: df_collisions.isnull().sum()

Out[9]: X                     7469
        Y                     7469
        OBJECTID                 0
        INCKEY                   0
        COLDETKEY                0
        REPORTNO                 0
        STATUS                   0
        ADDRTYPE              3712
        INTKEY              149443
        LOCATION              4586
        EXCEPTRSNCODE       120403
        EXCEPTRSNDESC       209491
        SEVERITYCODE             1
        SEVERITYDESC             0
        COLLISIONTYPE        26499
        PERSONCOUNT              0
        PEDCOUNT                 0
        PEDCYLCOUNT              0
        VEHCOUNT                 0
        INJURIES                 0
        SERIOUSINJURIES          0
        FATALITIES               0
        INCDATE                  0
        INCDTTM                  0
        JUNCTIONTYPE         11967
        SDOT_COLCODE             1
        SDOT_COLDESC             1
        INATTENTIONIND      191078
        UNDERINFL            26479
        WEATHER              26688
        ROADCOND             26608
        LIGHTCOND            26776
        PEDROWNOTGRNT       216078
        SDOTCOLNUM           94061
        SPEEDING            211353
        ST_COLCODE            9413
        ST_COLDESC           26499
        SEGLANEKEY               0
        CROSSWALKKEY             0
        HITPARKEDCAR             0
```

Figure 4: Missing values in each column

Analysis plan

After the data has been balanced, exploratory data analysis will be carried out to identify relationships between the various characteristics in the dataset with the target variable, severity of accident. Missing data will also be investigated and corrected. A set of predictions will also be identified to use in the prediction model. Since our data is labeled, supervised machine learning techniques will be used. The data will be split into training and test sets for training and validation. Since our target variable is likely to be binary after balancing the data, classification models will be evaluated to select the best classifier. The following classification models will be evaluated; K – nearest neighbor, Naïve Bayes, Random Forest, Logistic Regression, Gradient Boosting, XGBoost and Support Vector Machines. Various measures of accuracy such as Jaccard index, F1 score and log loss will be used to select the best classification model.

Methodology

Data cleaning

In this section, variables that were descriptions of other coded variables were dropped. Columns that also were used for identification of the collision were also dropped. Geometry columns, that is longitudes and latitudes were also dropped since the nature of the task is not related to mapping.

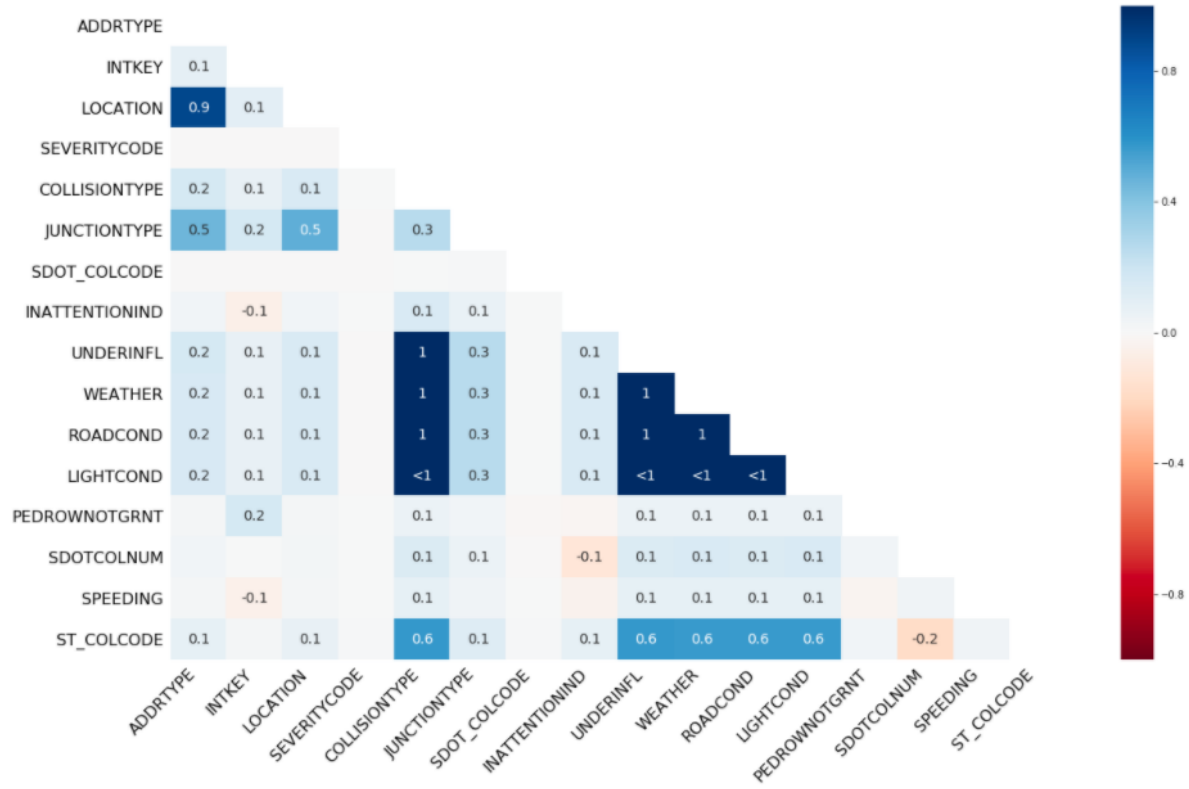
Missing data

To evaluate level of missingness, the number and proportions of missing values were obtained. See below the tables and a missing heat map. The variable with the highest missingness was whether or not the pedestrian right of way was not granted (**PEDROWNOTGRNT**) at 97.7% of data missing followed by **SPEEDING** with 95% of data missing.

```
df_cols.isnull().sum()/len(df_cols)
```

```
ADDRTYPE            0.016776
INTKEY              0.675400
LOCATION            0.020726
SEVERITYCODE        0.000005
COLLISIONTYPE       0.119761
PERSONCOUNT         0.000000
PEDCOUNT            0.000000
PEDCYLCOUNT         0.000000
VEHCOUNT            0.000000
INJURIES            0.000000
SERIOUSINJURIES     0.000000
FATALITIES          0.000000
INCDATE             0.000000
INCDTTM             0.000000
JUNCTIONTYPE        0.054084
SDOT_COLCODE        0.000005
INATTENTIONIND      0.863567
UNDERINFL           0.119670
WEATHER             0.120615
ROADCOND            0.120253
LIGHTCOND           0.121013
PEDROWNOTGRNT       0.976553
SDOTCOLNUM          0.425104
SPEEDING            0.955199
ST_COLCODE          0.042542
SEGLANEKEY          0.000000
CROSSWALKKEY        0.000000
HITPARKEDCAR        0.000000
dtype: float64
```

A missing heat map visualizes the missingness pattern as shown below where we see for instance high association in missingness between LOCATION and ADDRTYPE. Which means for these two variables the missingness is systematic. The darker the color the higher the association in missingness between two variables. We see that the highest missingness we observed for PEDROWNOTGRNT and SPEEDING is not associated with any other variable but it is independent and just caused by the variables only. This could be due to not entering this data.

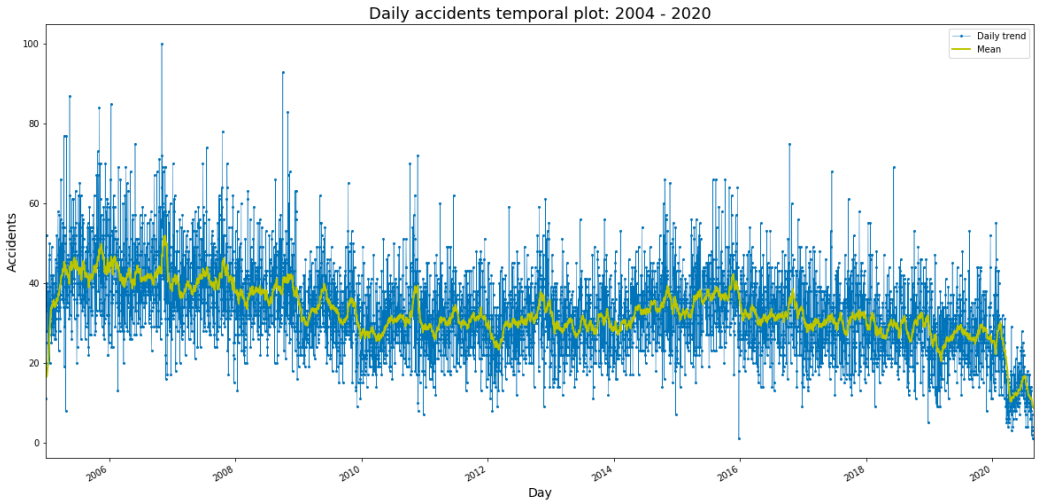


Variables with more than 40% of data were excluded to reduce the uncertainty in the imputation of the values. Usually a Multiple Imputation approach would have been used for imputation but is time intensive. Mean and most frequent values were used to impute for variables with less than 40% data missing.

EDA

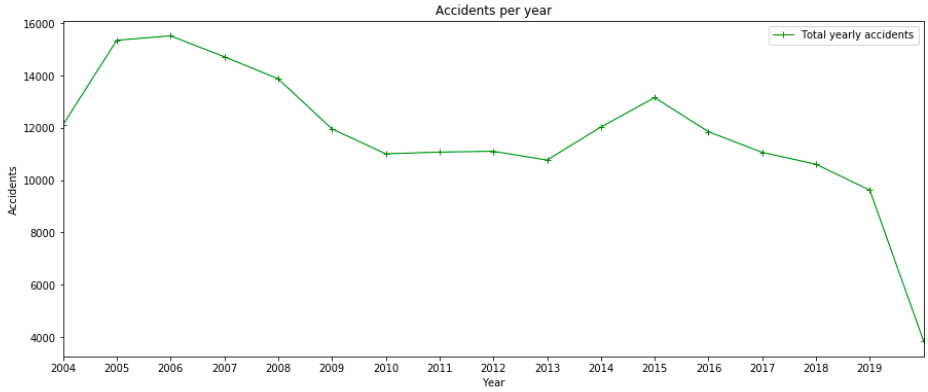
To understand the data better, temporal trends on the number of accidents since 2004 to present were obtained. This entailed making sure the date variables are in a date time format. The year, month and week day were extracted.

Daily trend



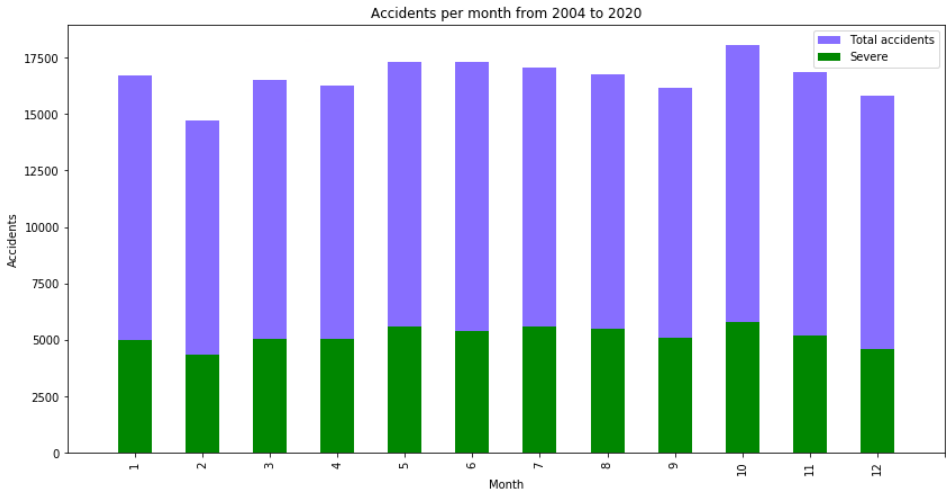
We see a steady trend overtime, however in 2020 there is a significant drop of daily accidents. This could be due to the restriction in movement during the COVID19.

Yearly trend



We see a slight decline after 2006 and a significant drop after 2019.

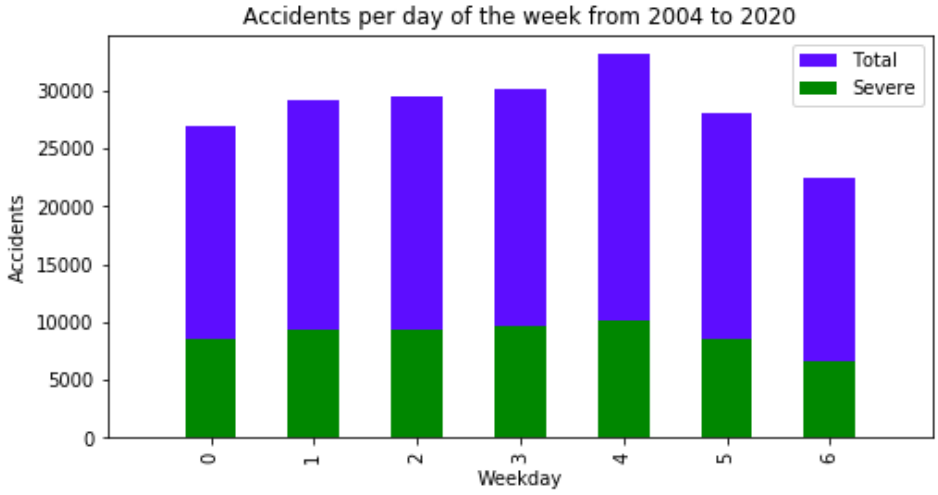
Monthly trend



There isn't a significant change monthly.

Week day trend

The next shows week day number of accidents.



Next, the relationship between the variables was obtained using a correlation plot as shown below.

Correlation plot

	SEVERITYCODE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	INJURIES	SERIOUSINJURIES	FATALITIES	SDOT_COLCODE	UNDERINFL	SEGLANEKEY	CROSSWAL
SEVERITYCODE	1	0.370575	0.260548	0.203206	0.38489	0.700391	0.280069	0.168462	0.311591	0.0966905	0.0974853	0.1
PERSONCOUNT	0.370575	1	0.0116041	-0.00916478	0.558807	0.319327	0.107266	0.0463413	0.008312	0.0538994	-0.00834905	-0.01
PEDCOUNT	0.260548	0.0116041	1	-0.0158227	-0.154356	0.167294	0.132543	0.0728061	0.260955	0.0319634	0.00126502	0.5
PEDCYLCOUNT	0.203206	-0.00916478	-0.0158227	1	-0.15074	0.122495	0.0620951	0.0110157	0.369044	-0.0126391	0.456218	0
VEHCOUNT	0.38489	0.558807	-0.154356	-0.15074	1	0.142673	-0.00306759	-0.0106176	-0.0785456	0.0507102	-0.0752901	-0.1
INJURIES	0.700391	0.319327	0.167294	0.122495	0.142673	1	0.279368	0.0671804	0.138528	0.0641258	0.0593988	0.1
SERIOUSINJURIES	0.280069	0.107266	0.132543	0.0620951	-0.00306759	0.279368	1	0.173007	0.0866685	0.0472939	0.0315772	0.05
FATALITIES	0.168462	0.0463413	0.0728061	0.0110157	-0.0106176	0.0671804	0.173007	1	0.0458338	0.0431549	0.00511154	0.03
SDOT_COLCODE	0.311591	0.008312	0.260955	0.369044	-0.0785456	0.138528	0.0866685	0.0458338	1	0.115384	0.202098	0.1
UNDERINFL	0.0966905	0.0538994	0.0319634	-0.0126391	0.0507102	0.0641258	0.0472939	0.0431549	0.115384	1	-0.00595666	-0.004
SEGLANEKEY	0.0974853	-0.00834905	0.00126502	0.456218	-0.0752901	0.0593988	0.0315772	0.00511154	0.202098	-0.00595666	1	-0.003
CROSSWALKKEY	0.167778	-0.0102738	0.553888	0.10378	-0.120686	0.100689	0.0559026	0.0318511	0.187266	-0.00407634	-0.00353797	
year	-0.0257858	-0.0673016	0.0214149	0.0276288	-0.108892	-0.00442378	-0.00503714	-0.00087498	-0.0846856	-0.0139701	0.0231274	0.05
month	-0.00444333	-0.00767659	0.00431671	0.00518652	-0.0104186	0.000967585	-0.00069574	0.00441976	0.00671416	0.0010763	0.00364075	0.006
weekday	0.000308069	0.0552597	-0.0176523	-0.0235545	0.017723	0.00840898	0.00351161	0.0045957	0.0162656	0.0732051	-0.0134219	-0.01
LOCATION1	-0.0270454	-0.0123105	-0.0361607	-0.0189904	0.00448877	-0.0273793	-0.00324572	0.000387719	0.0306835	0.0069568	-0.011494	-0.04
COLLISIONTYPE1	-0.0617565	0.0375634	0.102328	-0.215409	0.113219	-0.104717	-0.0236771	-0.00183166	0.00892772	0.00761879	-0.0989727	0.03
JUNCTIONTYPE1	-0.264412	-0.15276	-0.128913	-0.0939488	-0.0859161	-0.193834	-0.03328	-0.00752107	-0.15698	0.0225128	-0.0408399	-0.1
LIGHTCOND1	-0.123226	-0.0858907	-0.0532116	0.00502127	-0.0688841	-0.0751511	-0.0299769	-0.0177534	-0.174937	-0.225657	0.00108349	-0.02
ROADCOND1	0.0973057	0.0770512	0.0197047	-0.0361814	0.173137	-0.00178518	-0.00593351	-0.00732513	0.018023	0.00582948	-0.0163792	0.01
WEATHER1	0.0573449	0.0528063	0.00320965	-0.0393614	0.182701	-0.0490784	-0.0133418	-0.0081717	-0.0429878	-0.0226485	-0.0202497	0.003
HITPARKEDCAR1	-0.201678	-0.11906	-0.0437938	-0.0365627	-0.123492	-0.106276	-0.02004	-0.0090251	-0.158467	-0.00980709	-0.0183863	-0.03
Alley	-0.0225235	-0.0244881	0.0020348	-0.00694962	-0.0208688	-0.0227097	-0.00376006	-0.0023759	-0.0879492	0.0014004	-0.00506764	-0.008
Block	-0.193272	-0.0841446	-0.144518	-0.087084	1.69995e-05	-0.166582	-0.0341676	-0.00815805	-0.00487136	0.0335608	-0.0397057	
Intersection	0.196892	0.0876876	0.144696	0.0882867	0.0027786	0.170144	0.0347779	0.00850179	0.0166685	-0.0338531	0.0405085	0.1

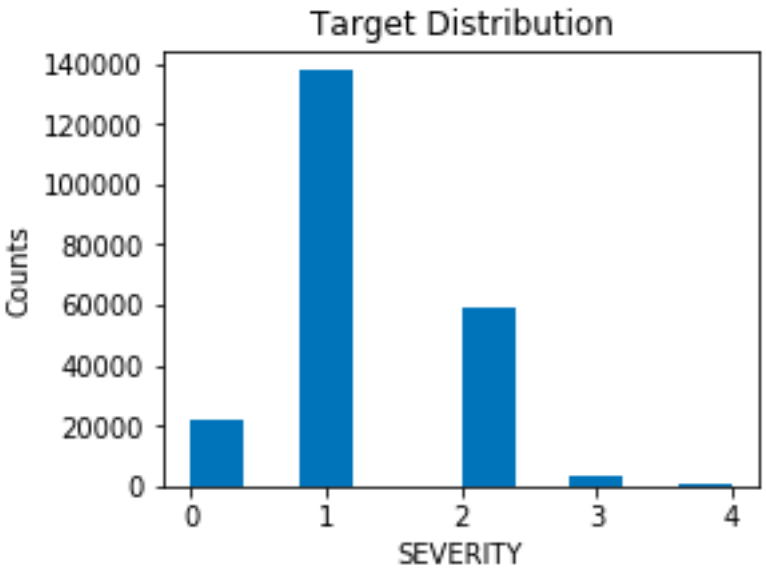
Features that were significantly associated with severity were taken forward for modelling. Clearly we see year, month, weekday, Location and Alley have a very small correlation with severity.

Model development

Target variable

The target variable has 5 categories. See below the distribution of these categories.

See below the distribution of the target.



Splitting training and test set and balancing data

To ensure balanced data, the data was stratified when splitting by the target variable. The training set included 80% of data while 20% was used as test set.

Classification models

Four classification models were tested to obtain the best classifier. These are;

i. KNN
ii. Decision Tree
iii. SVM
iv. Logistic Regression

Results

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.993153	0.993169	NA
Decision Tree	0.990374	0.990331	NA
SVM	0.991051	0.991137	NA
Logistic Regression	0.973154	0.973286	0.0914651

Discussion

i. In the results, it shows that among three machine learning methods, KNN excels other methods with only small difference in recall. Although KNN provide the best performance evaluation, parameter tuning in KNN is computationally exhaustive.
ii. In using decision tree, the maximum depth is 4 for the best accuracy performance on training data.
iii. Among all the ML methods, kNN is the safest approach
iv. To optimize the selection of features, using PCA is advisable to fasten the process of prominent feature selection.

Conclusion

i. The data from Seattle helps in identifying useful factors that help in building a predictive model
ii. The classification models predicted severity of the accidents accurately and KNN classifier is the best
iii. More features could be tested to assess how well they predict severity of accidents
iv. Missing data in some features could be corrected by ensuring these features are also captured