Capstone Project – Predict Car Accident Severity

Jenny Wong

26 Oct, 2020

1 Introduction

1.1 Background

Life is the most valuable and irreplaceable thing in the world. I believe no one wants to have car accidents while travelling on road. This project is to predict the severity of a car accident by using various data of historical car accidents like weather, date, road conditions, etc. that would be able to let people drive more carefully or even change the route or date of the journey. So that not only saving the time to get involve into a car accident but saving people life.

1.2 Data Description

The dataset uses for the project is the collisions in Seattle provided by the Seattle Policy Department (SPD) and recorded by the Seattle Department of Transportation (SDOT) Traffic Management Division, Traffic Records Group. It includes all types of collisions from 2004 to present. You can download it from here and the metadata can be found from here.

This project uses the dataset to predict the severity of collision occurs in specific address type (i.e. alley, block and intersection) by given conditions during the collision like weather, road, light, speeding or not, etc. Since this is a labelled dataset, a supervised machine learning model is implemented for the project.

The dataset has unbalanced labels and noise which are required to perform data cleaning to make sure it becomes balance and clean. Afterwards, the dataset will be divided into training set and test set. Different models like KNN, SVM, Regression will be built by using the training data set. The model evaluation will be done by using the test set to report the accuracy of the model and come up a conclusion for the analysis.

This dataset is about traffic collisions in the City of Seattle. The **Data_Collisions.csv** data set includes 38 columns and 194673 rows which are provided by SPD and recorded by Traffic Records for the last 10 years. The dataset can be downloaded from here and the metadata from here.

The dataset includes all types of collisions. The fields of the data as below:

Field	Description
SEVERITYCODE	The severity of the collision.
X	The x-coordinate of the collision address.
Y	The y-coordinate of the collision address.
OBJECTID	ESRI Object ID Field.
INCKEY	A unique key for the incident.
COLDETKEY	A key that corresponds to the collision's detail.
REPORTNO	The report number for the collision.
STATUS	The status tells the information of the record matched or unmatched.
ADDRTYPE	The address type for the collision - Alley, Block or Intersection.
INTKEY	A key corresponding to the intersection to which the collision is associated.
LOCATION	A general location description for the collision.
EXCEPTRSNCODE	Unknown.
EXCEPTRSNDESC	Unknown.
SEVERITYCODE	Codifies the severity of the collision based on the fatality and disabling injury counts as well as pre-existing state severity codes.
SEVERITYDESC	A general description of the severity of the collision.
COLLISIONTYPE	A description of the type of collision that is represented.
PERSONCOUNT	The number of people involved in the collision.
PEDCOUNT	The number of pedestrians involved in the collision.
PEDCYLCOUNT	The number of cyclists involved in the collision.
VEHCOUNT	The number of vehicles involved in the collision.
INCDATE	The date of the collision.
INCDTTM	The date and the time of the collision if an exact time is known.
JUNCTIONTYPE	The type of junction at which the collision occurred.
SDOT_COLCODE	A code for the collision determined by the Seattle Department of Transportation.
SDOT_COLDESC	The human-readable description of the code given in the SDOT_COLCODE field.
INATTENTIONIND	Whether or not the collision was due to inattention of one or more of the involved parties.
UNDERINFL	Whether or not collision involved someone that was under the influence of drugs or alcohol. '1' if so, '0' if not, and "Null" if unknown.
WEATHER	The weather conditions at the time of the collision.
ROADCOND	The conditions of the road during the time of the collision.
LIGHTCOND	The light conditions during the accident.
PEDROWNOTGRNT	Whether or not the pedestrian involved in the collision was granted the right-of-way.
SDOTCOLNUM	Unknown.
SPEEDING	Whether or not speeding was a factor in the collision.
ST_COLCODE	A state code for the type of collision.
ST_COLDESC	A description of the state code for the type of collision.
SEGLANEKEY	Unknown.
CROSSWALKKEY	Unknown.
HITPARKEDCAR	Whether or not the collision included hitting a parked car.

Table 1

2 Data Visualization and Preprocessing

2.1 Understanding the data and perform data cleaning

We need to remove duplicated and unuseful columns and rows before forming a model. Since we are going to make use of the conditions of the road and the external factors of the environment to predict the severity of a car accident, all the unrelated data likes number of pedestrians involved, whether hitting a parked car or not will be removed from the source of the dataset. Also, those undefined well enough records will be removed as well to avoid data being miss-trained.

- Only the columns ‘SEVERITYCODE', ‘INCDTTM’, 'JUNCTIONTYPE', ‘ADDRTYPE’, 'WEATHER', 'ROADCOND', 'LIGHTCOND' are kept.
- Further analysis each field and the relationship of it against those related.

2.1.1 Further analysis on the field SEVERITYCODE

There are total 194673 records in which Severity = 1 contains 70.1%; Severity = 2 contains 29.9% where ‘1’ means property damage; ‘2’ means injury.

2.1.2 Further analysis on the field INTCDTTM

This is a field including the collision date and time from 2004 to 2020. We break it down into date, day, hour and year for further analysis.

2.1.2.1 Day

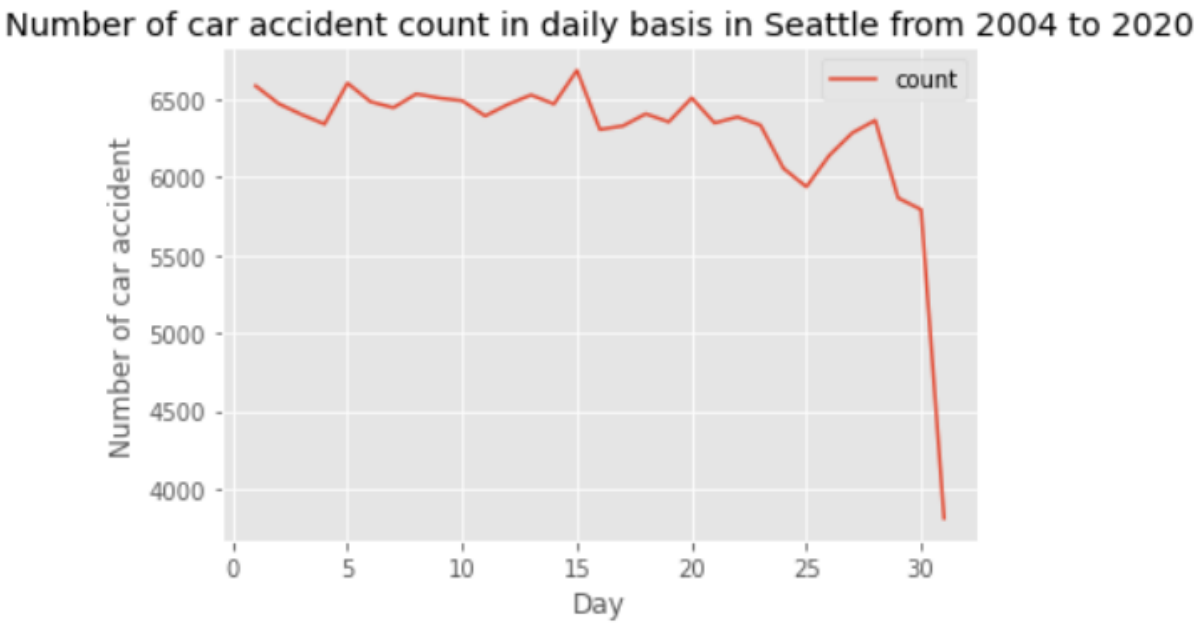

Figure 1

From the above figure "Number of car accident count in daily basis in Seattle from 2004 to 2020", we can see that all the day with the similar possibility of collisions only Day 31 drops half of number of collisions, it should because there are only half of the months out of a year has Day 31. Therefore, we can conclude that the field Day does not have significant effect on the severity of a car accident.

2.1.2.2 Month

Month	1	2	3	4	5	6	7	8	9	10	11	12
Percentage %	8.4	7.4	8.3	8.2	8.6	8.5	8.4	8.4	8.1	9.1	8.5	8.0

Figure 2

From the above figure "The percentage of collision occurred per month from 2004 to 2020 in Seattle", we can see that the number of collisions evenly distribute in a whole year. Therefore, we can conclude that the field Month does not have significant effect on the severity of a car accident.

2.1.2.3 Hour

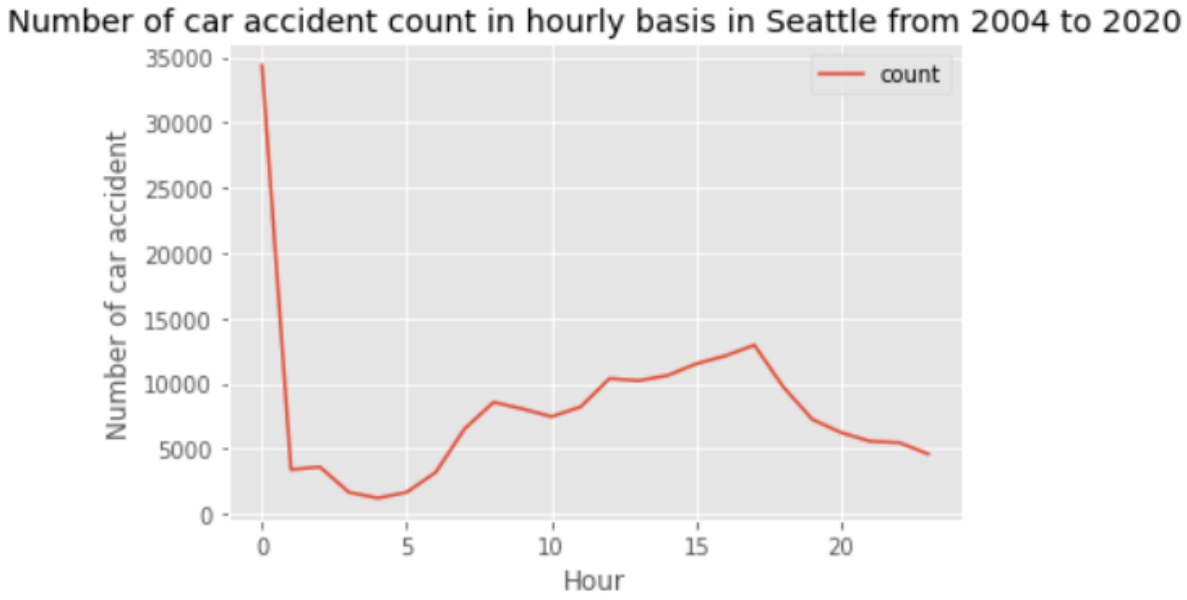

Figure 3

Hour	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Percentage %	17.7%	1.8%	1.9%	0.9%	0.6%	0.9%	1.6%	3.4%	4.4%	4.1%	3.8%	4.2%	5.3%	5.2%	5.5%	5.9%	6.2%	6.7%	5.0%	3.7%	3.2%	2.9%	2.8%	2.4%

Figure 4

From the above figure “Number of car accident count in hourly basis in Seattle from 2004 to 2020” and the figure “The percentage of collision occurred per hour from 2004 to 2020 in Seattle”, we can see that the number of collisions were extremely high from 00:00 to 01:00 of a day and the distribution of hour is uneven per day. Therefore, we can conclude that the field Hour has significant effect on the severity of a car accident.

2.1.3 Further analysis on the fields JUNCTIONTYPE and ADDRTYPE

	Counts	Percentage %
Mid-Block (not related to intersection)	89800	47.7%
At Intersection (intersection related)	62810	33.3%
Mid-Block (but intersection related)	22790	12.1%
Driveway Junction	10671	5.7%
At Intersection (but not related to intersection)	2098	1.1%
Ramp Junction	166	0.1%
Unknown	9	0.0%

Figure 5

	Counts	Percentage %
Block	126926	65.9%
Intersection	65070	33.8%
Alley	751	0.4%

Figure 6

JUNCTIONTYPE	ADDRTYPE	
At Intersection (but not related to intersection)	Intersection	0.999046
	Alley	0.000477
	Block	0.000477
At Intersection (intersection related)	Intersection	0.999936
	Block	0.000064
Driveway Junction	Block	0.994470
	Alley	0.005530
Mid-Block (but intersection related)	Block	0.999956
	Intersection	0.000044
Mid-Block (not related to intersection)	Block	0.997911
	Alley	0.001966
	Intersection	0.000123
Ramp Junction	Block	0.810976
	Intersection	0.189024
Unknown	Block	0.666667
	Intersection	0.333333

Figure 7

ADDRTYPE	JUNCTIONTYPE	
Alley	Mid-Block (not related to intersection)	0.745763
	Driveway Junction	0.250000
	At Intersection (but not related to intersection)	0.004237
Block	Mid-Block (not related to intersection)	0.727105
	Mid-Block (but intersection related)	0.185370
	Driveway Junction	0.086369
	Ramp Junction	0.001083
	At Intersection (intersection related)	0.000033
	Unknown	0.000033
	At Intersection (but not related to intersection)	0.000008
Intersection	At Intersection (intersection related)	0.967052
	At Intersection (but not related to intersection)	0.032255
	Ramp Junction	0.000478
	Mid-Block (not related to intersection)	0.000169
	Unknown	0.000031
	Mid-Block (but intersection related)	0.000015

Figure 8

From the above figure 5 “The breakdown of JUNCTIONTYPE of collision occurred in Seattle from 2004 to 2020”, the figure 6 “The breakdown of ADDRTYPE of collision occurred in Seattle from 2004 to 2020”, the figure 7 “The breakdown of JUNCTIONTYPE group with the breakdown of ADDRTYPE” and the figure 8 “The breakdown of ADDRTYPE group with the breakdown of JUNCTIONTYPE”, we can conclude that the field JUNCTIONTYPE can be simply defined to Block, Intersection and Alley. And since 'Alley' only dominate 0.4% out of the entire dataset, therefore, we replace all Alley to Block as well (as 75% of Alley related to block as well, 25% of Alley related to Driveway Junction which also can map to Block'). Thus, we can now drop the field JUNCTIONTYPE and replace all Alley to Block in the entire dataset.

2.1.4 Further analysis on the field LIGHTCOND

	Counts	Percentage %
Daylight	116137	61.3%
Dark - Street Lights On	48507	25.6%
Unknown	13473	7.1%
Dusk	5902	3.1%
Dawn	2502	1.3%
Dark - No Street Lights	1537	0.8%
Dark - Street Lights Off	1199	0.6%
Other	235	0.1%
Dark - Unknown Lighting	11	0.0%

Figure 9

From the above figure 9 “The breakdown of LIGHTCOND of collision occurred in Seattle from 2004 to 2020”, we can see that Daylight and Dark almost dominate the entire dataset, therefore, all Dark related fields (i.e. Dark – Street lights on, etc.) are group into a single field ‘Dark’ and ‘Dusk’, ‘Dawn’ and ‘Other’ are group into ‘Unknown”.

2.1.5 Further analysis on the fields WEATHER and ROADCOND

	Counts	Percentage %
Clear	111135	58.6%
Raining	33145	17.5%
Overcast	27714	14.6%
Unknown	15091	8.0%
Snowing	907	0.5%
Other	832	0.4%
Fog/Smog/Smoke	569	0.3%
Sleet/Hail/Freezing Rain	113	0.1%
Blowing Sand/Dirt	56	0.0%
Severe Crosswind	25	0.0%
Partly Cloudy	5	0.0%

Figure 10

	Counts	Percentage %
Dry	124510	65.6%
Wet	47474	25.0%
Unknown	15078	7.9%
Ice	1209	0.6%
Snow/Slush	1004	0.5%
Other	132	0.1%
Standing Water	115	0.1%
Sand/Mud/Dirt	75	0.0%
Oil	64	0.0%

Figure 11

From the above figure 10 “The breakdown of WEATHER of collision occurred in Seattle from 2004 to 2020”, we can see except Clear, Raining, Overcast and Unknown dominate the dataset. Therefore, all the other types of weather are grouped to Unknown.

From the above figure 11, “The breakdown of ROADCOND occurred in Seattle from 2004 to 2020”, we can see Dry and Wet dominate the dataset. Therefore, all the other types of road condition are grouped to Unknown.

WEATHER	ROADCOND	
Clear	Dry	0.955272
	Wet	0.032390
	Unknown	0.012338
Overcast	Dry	0.584109
	Wet	0.388059
	Unknown	0.027832
Raining	Wet	0.971968
	Dry	0.019342
	Unknown	0.008690
Unknown	Unknown	0.866185
	Dry	0.083258
	Wet	0.050557

Figure 12

ROADCOND	WEATHER	
Dry	Clear	0.853044
	Overcast	0.130039
	Unknown	0.011765
	Raining	0.005151
Unknown	Unknown	0.862409
	Clear	0.077629
	Overcast	0.043656
	Raining	0.016307
Wet	Raining	0.678862
	Overcast	0.226554
	Clear	0.075848
	Unknown	0.018736

Figure 13

From the above figure 12 “The breakdown of WEATHER group with the breakdown of ROADCOND” and the figure 8 “The breakdown of ROADCOND group with the breakdown of WEATHER”, we can see that whenever ROADCOND = 'Unknown', WEATHER has 86.2% = 'Unknown' as well and vice versa. Since, both fields are important to the prediction of severity of an accident, therefore, to prevent mis-leading the prediction, we will drop those records with ROADCOND or WEATHER = 'Unknown'.

2.1.6 Feature selection

After cleaning up the data, we use One Hot Encode to convert categorical features (“Weather”) to numerical values and all the values of the other features become “0” or “1”. Only 7 features are selected for building models, they are 'JUNCTIONTYPE', ‘ADDRTYPE’, 'WEATHER', 'ROADCOND', 'LIGHTCOND' and ‘hour’ to predict ‘SEVERITYCODE'.

3 Predictive Modeling with Evaluation

We use classification to predict the severity of a car accident with the selected features. The following algorithms are used:

a) K Nearest Neighbor(KNN)
b) Decision Tree
c) Support Vector Machine
d) Logistic Regression

3.1 Classification models with accuracy

We divide the sample dataset into two sets – training set and testing set. 80% of sample data is used for training the models and the remaining 20% of sample data is used to test the predictive models. After forming predictive models, we then use 3 different evaluation metrics to calculate the accuracy of the models, they are Jaccard index, F1-score, and Log Loss. For Jaccard index and F1-score, the best accuracy mark is 1. However, for Log Loss, the best accuracy mark is 0.

Algorithm	Jaccard Index	F1-score	Log Loss
KNN	0.63	0.61	N.A.
Decision Tree	0.67	0.54	N.A.
SVM	0.67	0.54	N.A.
Logistic Regression	0.67	0.54	0.62

Table 2

Table 2 summaries the accuracy of all predictive models. We can see the model KNN has 0.63 and 0.61 on Jaccard Index and F1-score respectively. However, for the models Decision Tree, SVM and Logistic Regression have the same scores on Jaccard Index and F1-score which are 0.67 and 0.54 respectively.

4 Conclusion

In this study, we analyze the traffic collisions in the city of Seattle from 2004 to 2020. We would like to use the historical car accidents data to predict the severity of car accident in the future. We use the date like weather, road conditions (i.e. dark or light, dry or wet), junction type and time to build predictive models and used ~135000 records to train up the models and ~34000 to test the model. This is a very useful exercise to let people know what the difference of road conditions and time would cause the result. Since from the analyze, we can see more accidents

are occurred when the weather and the road conditions are good (i.e. on a day with enough light, dry road or clear road). I guess this is because people pay extra attention when the weather and road conditions are not good but much relaxing when everything looks good. One interesting point is most of the collisions occurred at 12am to 1am and during the normal off-duty hours. I guess this is because people rush to go home so more accidents are caused. In conclusion, we should always pay attention when driving especially when the weather and road conditions are good.