Predicting Type of Vehicle Incidents based on outside weather conditions

Dmitry Elsakov

September 26, 2020

1. Introduction

1.1. Background

The seaport city of Seattle is the largest city in the state of Washington, as well as the largest in the Pacific Northwest. As of the latest census, there were 713,700 people living in Seattle. Seattle residents get around by car, trolley, streetcar, public bus, bicycle, on foot, and by rail. With such bustling streets, it's no surprise that Seattle sees car accidents every day.

Therefore, it is advantageous for interested parties to accurately predict the type of incident based on outside conditions. And information from this study could be used by road department or police department to improve road conditions and reduce potential level of injuries on car incidents.

1.2. Problem

In 2015, a crash occurred in Washington every 4.5 minutes, while in 2017 – every 4 minutes. In 2015 Seattle recorded the highest number of car accidents in the state, at 14,508 (in second place was Tacoma with just 4,756). Although the city is taking steps to make the roadways safer for citizens, vehicle collisions are still a serious danger.

1.3. Interest

To reduce number of incidents by changing some of road conditions or traffic management, this study could be interested for Seattle Road department or even for Washington State Department of Transportation (SDOT) to know that changing road or light conditions could be helpful or not. Also, Seattle police department could use this data for improving safety transportation while outside conditions are not clear.

2. Data acquisition and cleaning

2.1. Data sources

Original dataset was provided by Washington State Department of Transportation (SDOT) and could be found on Kaggle here. The limited dataset with reduced number of rows for only few severity types could be found on IBM shared folder here. The metadata for the dataset with explanatory details of each column could be also found on IBM shared folder here. The both datasets contain details of car incidents from 2004 till present.

2.2. Data cleaning

Downloaded data contain many features and many of them are not provided in full or have a little non-Matched record. For this study we would be using cases in which only active vehicles are participated (i.e. at least one non-parked vehicle with non-disabled driver), therefore we will remove all cases where participants are (*we could use 'SDOT_COLCODE' code values and remove all codes except 1,3,4,10-29, which is only related to at least one Motor Vehicle in operation*):

- Pedestrians / Bicyclists only
- Pedestrians / Trains only
- Bicyclists / Parked Vehicles only
- Non-major vehicle with other non-vehicle cases

Also, I would like to specifically look at the human-related and human-unrelated cases for our data analysis, so for human-related cases we will remove all features except the following:

- Speeding data
- Driver Under Influence
- Driver Under Inattention

On the other hand, for human-unrelated cases, which are related to outside conditions only, we will remove all features except the following:

- Weather conditions
- Road Conditions
- Light Conditions

2.3. Feature selection

By looking at the provided data, we could see that our predicted output – Severity of the incident (*target or predictor*) would have only few values (*binary: 1 = property damage and 2 = injury*) in 'SEVERITYCODE' column (*when original data contain 5 different types, including fatalities*), so for that particular study we could predict only those two outputs to limit size of full dataset.

As input attributes, we would be using the following data columns: 'WEATHER', 'ROADCOND', 'LIGHTCOND' and I would like to use 'INCDTTM' for time of the day from DateTime format for human-unrelated cases and 'SPEEDING', 'INATTENTIONIND' and 'UNDERINFL' for human-related cases.

	SEVERITYCODE	INCDTTM	WEATHER	ROADCOND	LIGHTCOND
0	2	14	Overcast	Wet	Daylight
1	1	18	Raining	Wet	Dark - Street Lights On
2	1	10	Overcast	Dry	Daylight
3	1	9	Clear	Dry	Daylight
4	2	8	Raining	Wet	Daylight

Table 1: Example of human-unrelated features

	SEVERITY CODE	INCDT TM	INATTENTIONIND	UNDER INFL	WEATHER	ROAD COND	LIGHTCOND	SPEEDING
1	1	18	0	0	Raining	Wet	Dark - Street Lights On	0
2	1	10	0	0	Overcast	Dry	Daylight	0
4	2	8	0	0	Raining	Wet	Daylight	0
6	1	25	0	0	Raining	Wet	Daylight	0
8	1	13	0	0	Clear	Dry	Daylight	0

Table 2: Example of human-related features

3. Exploratory Data Analysis

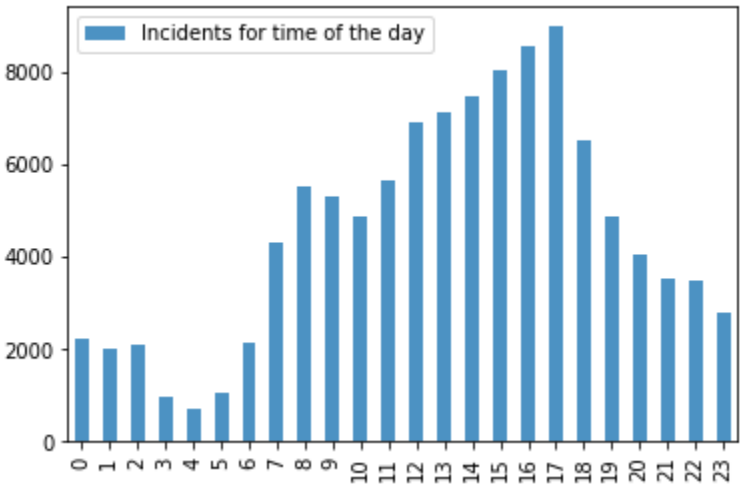
3.1. Target variable

As mentioned above, the Target variable in smaller dataset would have only few possible values:

- 1 = property damage
- 2 = injury

3.2. Relationship between incidents during time of the day and outside conditions

For the usage of histogram, we could clearly see what time is the most dangerous for driving as well as when is more safety conditions for specific weather, road or light conditions. Let's look at the visualization of that data:



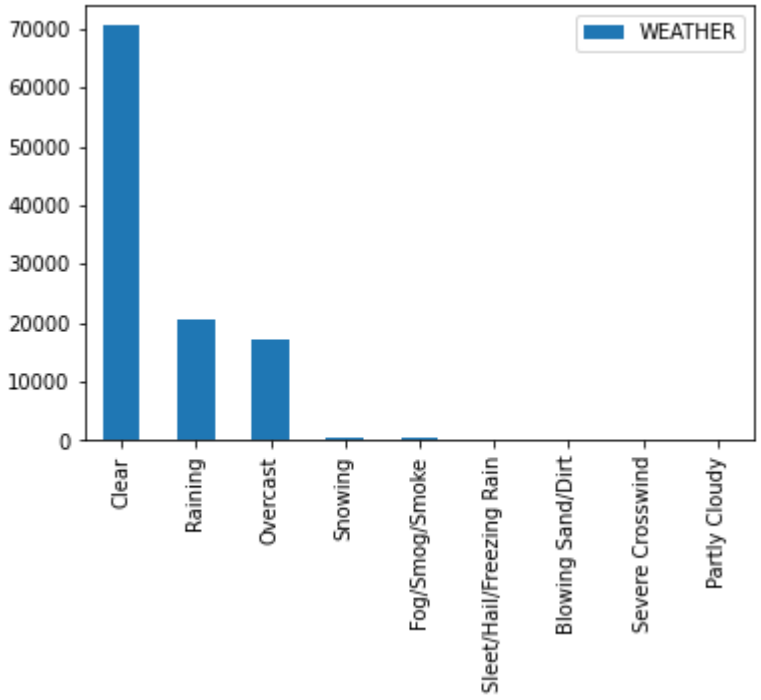
Picture 1: Histogram for all incidents during the day by hours

As we can see from the Picture 1 above, more incidents were happened within evening time, especially for the period of 2-5 PM.

Now, let's look at the specific cases in more details.

3.2.1. Relationship between incidents during time of the day and weather conditions

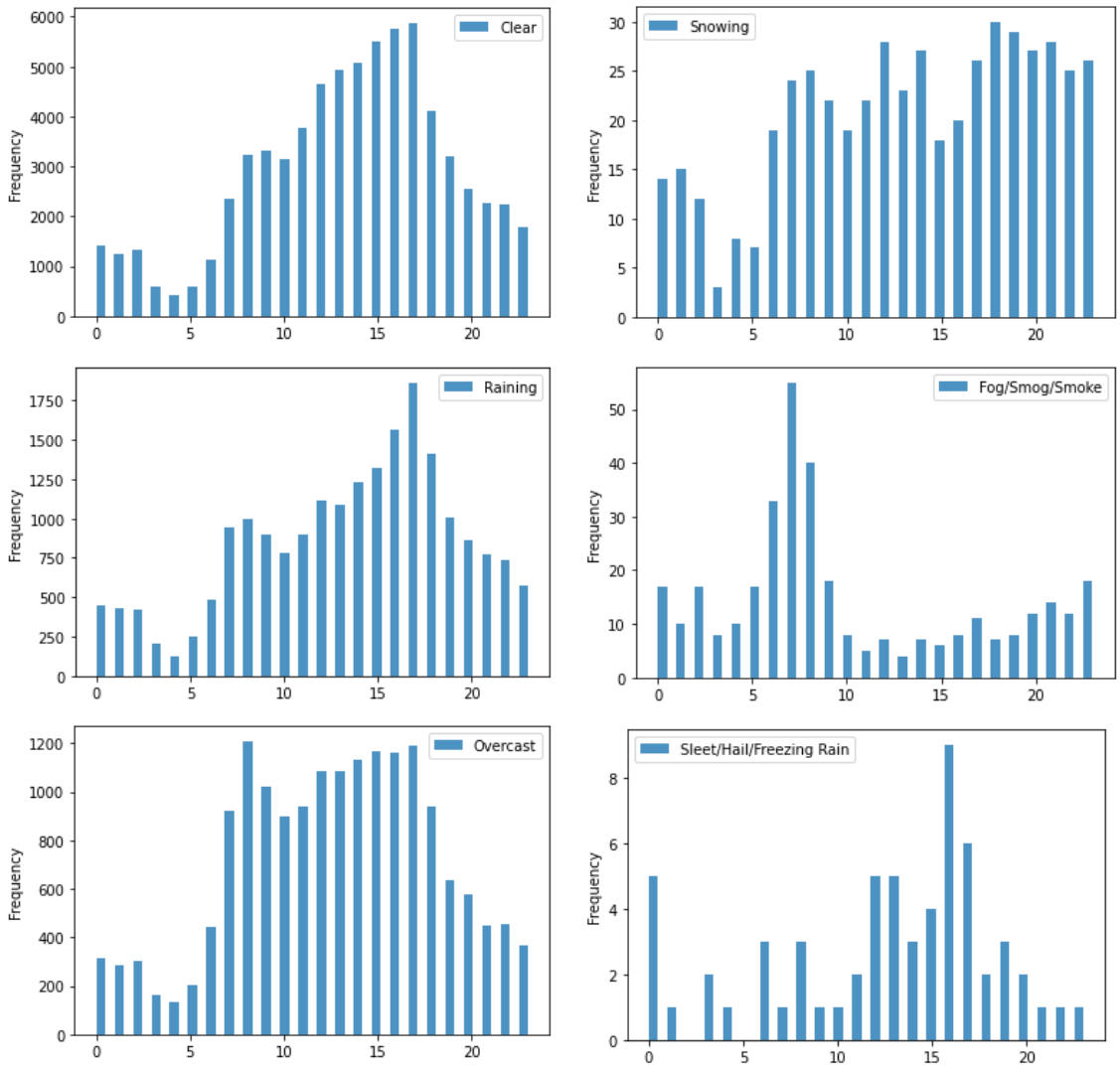
The different weather conditions listed on Picture 2 below noted that most of the incidents were happened in Clear weather, while Raining and Overcast weather conditions are on the second and third places respectively:



Picture 2: Incidents by Weather conditions

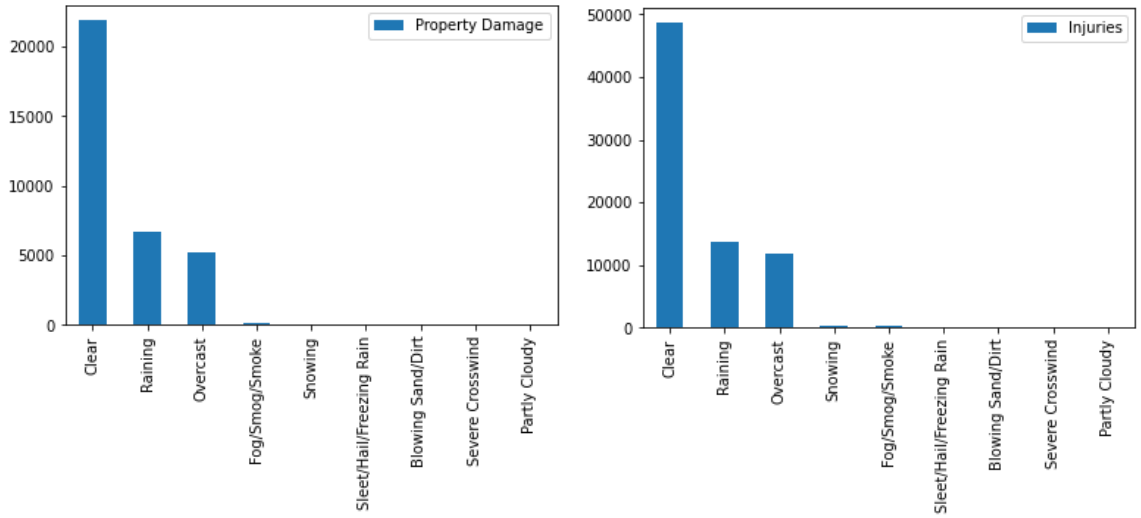
By looking on the data on the picture 3 below, we could reveal the following dependencies:

- While snowing, the incidents have similar probability during the day
- Overcast weather conditions have very similar to snow probability
- While Raining, most incidents are happened during evening time
- Fog / Smog / Smoke weather could be a reason of incidents for morning time
- Incidents in clear weather more likely happening on evening time (*may be when most of the people are commuting from work back to home*)



Picture 3: Incidents by Weather conditions hourly

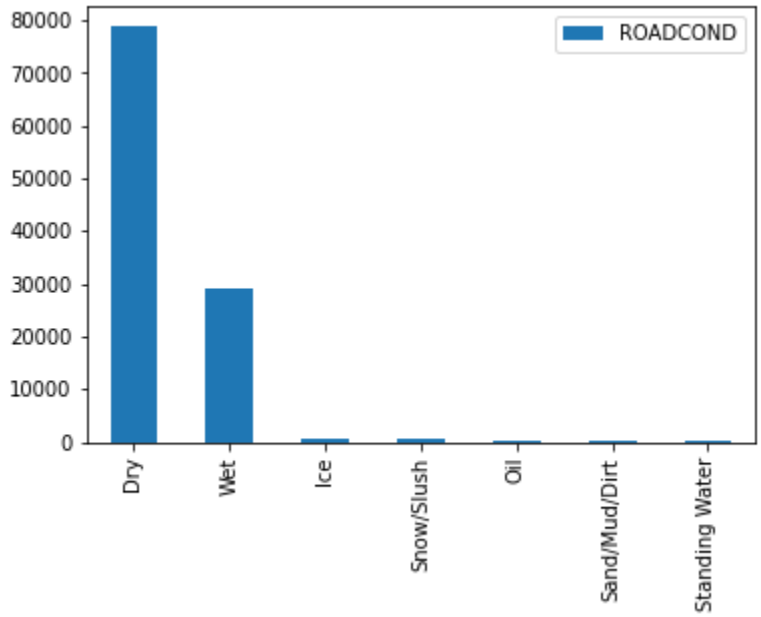
Finally, we could look at the incidents' severity distribution for different outside conditions. The picture 4 below displays the distribution for weather conditions, however Road and Light conditions would have similar views – the distribution of severity of the incident (*property damage vs injuries*) is very similar for any outside conditions.



Picture 4: Severity of Incidents by Weather conditions

3.2.2. Relationship between incidents during time of the day and road conditions

When we look at the Road conditions, we could find out that most of the incidents were happened on dry road, while wet taken the second most frequent place of incident probability, see picture 5 below:

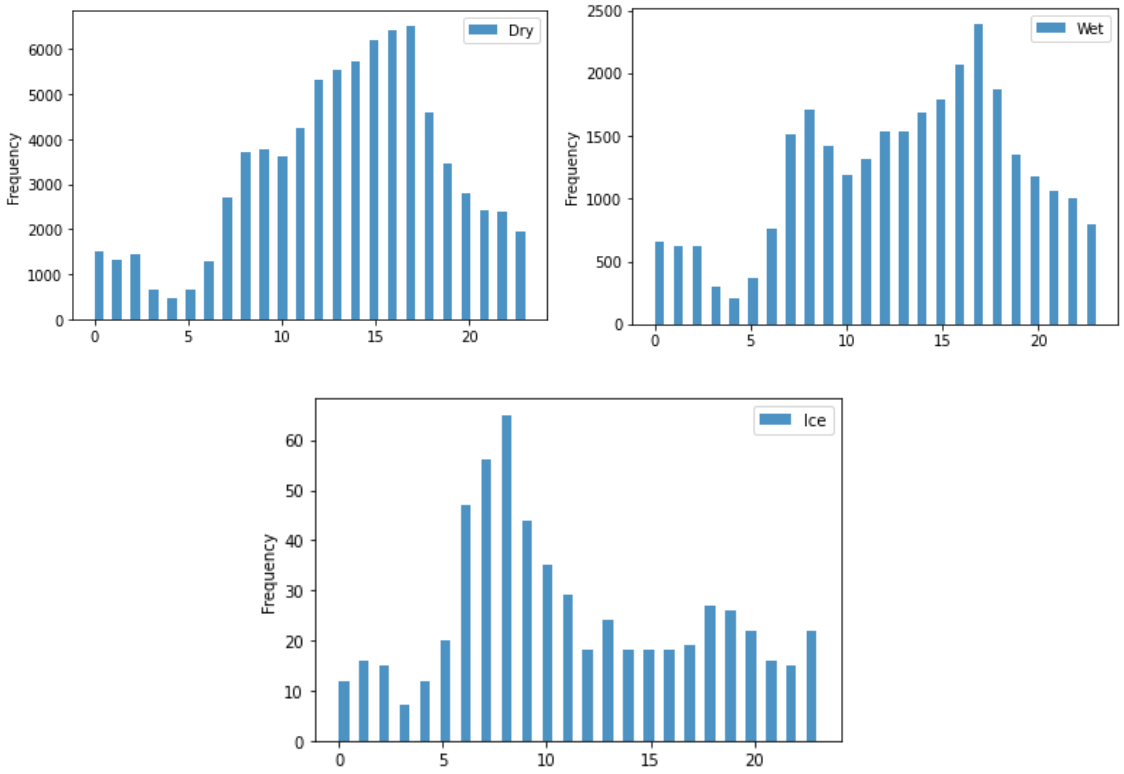


Picture 5: Incidents by Road conditions

For the detailed view of the data distribution for the time of the day of the first 3 most popular road conditions, we could see on picture 6 below the following:

- Dry road conditions are mostly correlated with general cases – evening time
- Wet road is more dangerous during the whole day

- Icy roads are most dangerous during morning hours (6-10 AM)

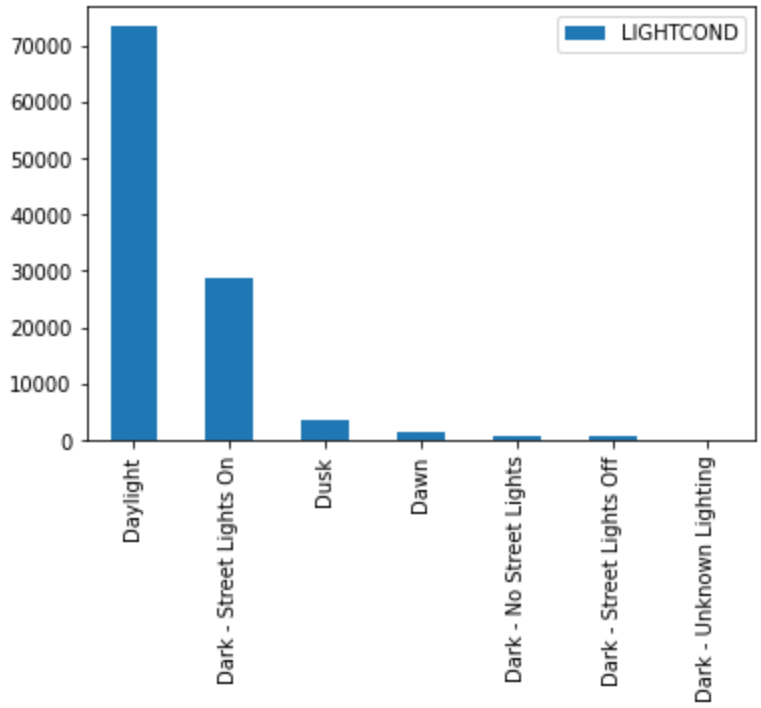


Picture 6: Incidents by Road conditions hourly

As mentioned above, distribution of incident severity by road conditions are almost same – there is no big difference between them.

3.2.3. Relationship between incidents during time of the day and light conditions

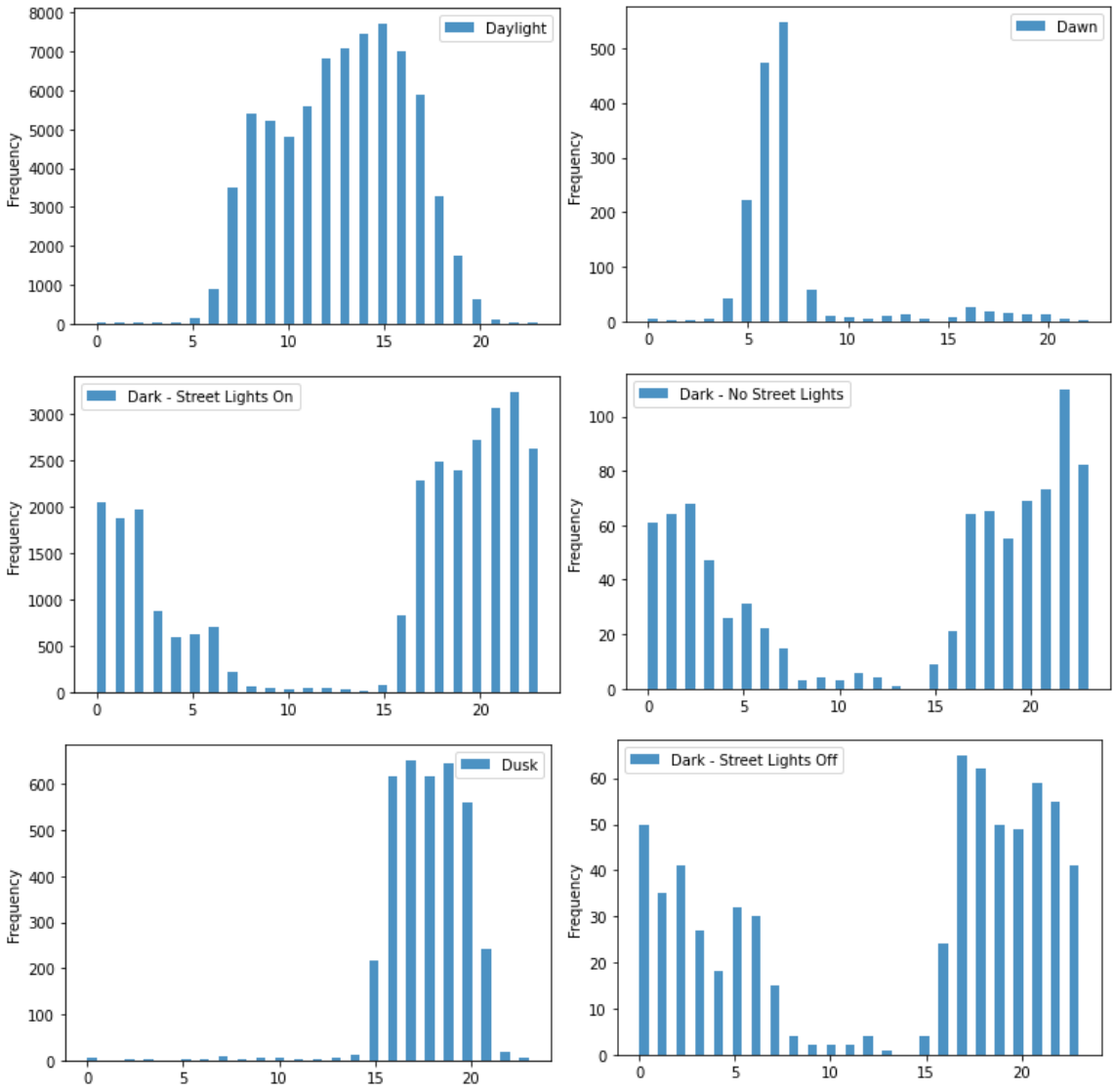
Also, we could look at the Light conditions and their relationship with number of incidents. Based on Picture 7 below, most of the incidents were happened on Daylight, while "Dark – Street Light On", Dusk and Dawn cases took second, third and fourth places:



Picture 7: Incidents by Light conditions

By looking on the data on the picture 8 below, we could reveal the following dependencies:

- During the Daylight more incidents are happened on evenings
- The quantity of incidents on Dusk 2+ times higher than Dawn cases
- There are some incidents while Dark – Street Lights Off (*which maybe could be avoided if Street lights would be switched On*)

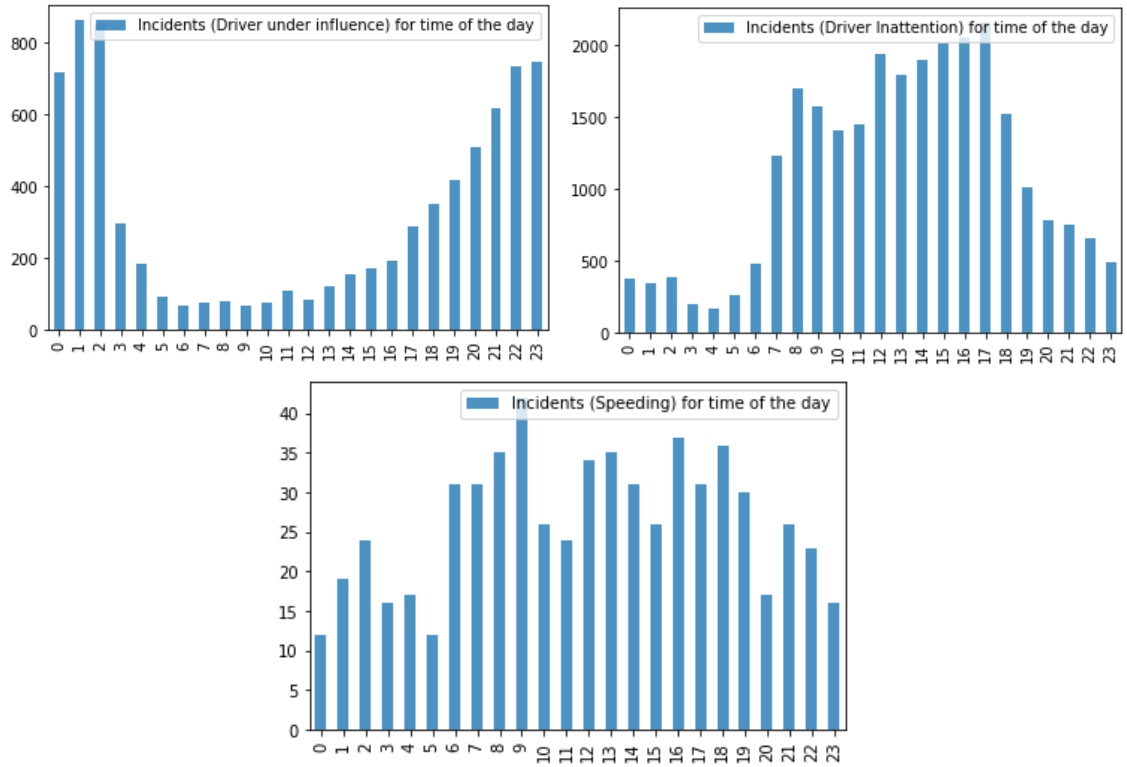


Picture 8: Incidents by Light conditions hourly

3.3. Relationship between incidents during time of the day and human-related incidents

Now we could look at the cases which were related to human behavior, i.e. when driver was under influence or inattention or while speeding.

For that means we would also use histogram of each cases during the day, so we could clearly see what time is the most dangerous for driving under specific driver conditions. Let's look at the visualization of that data:



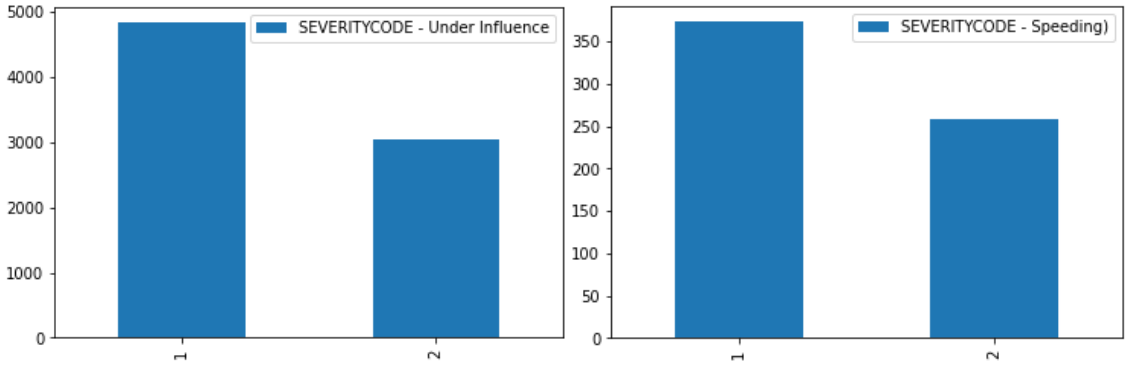
Picture 9: Incidents for Driver conditions by hour

So, based on the data above, we could make a following conclusions:

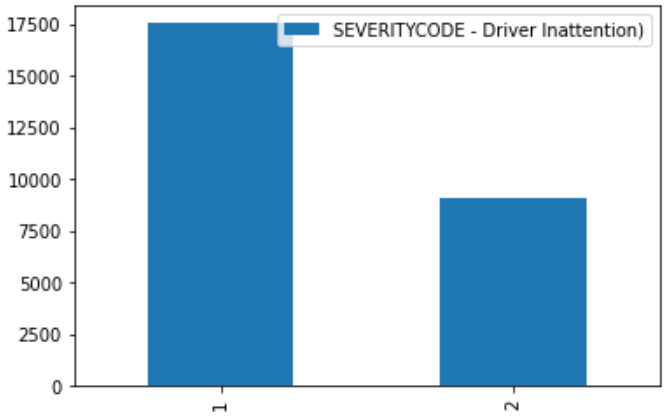
- The most cases related to driver under influence were taken place at night time
- Speeding related cases are distributed during the whole day
- Driver inattention cases very close to standard view (weather related)

3.4. Relationship between human-related incidents and incident severity

Based on the details above, let's look at the incident's distribution for all human-related cases and check the incident severity for each scenario:



Picture 10: Incident's severity for Driver conditions (under influence/speeding)



Picture 11: Incident's severity for Driver conditions (inattention)

As displayed on the pictures 10 and 11 above, the severity of the incidents in general are higher for human-related cases rather than outside conditions, i.e. as result of speeding or driver under influence there are ~2 times more cases with injuries rather than just property damage.

4. Results

4.1. Predictive Modeling

Generally, there are two types of predictive models – classification and regression. The difference is that while classification models could be helpful with prediction of output based on some inputs, they are not providing any additional information on the delta or amount of the changes of target feature.

In our case, we have only binary output feature, so we could use only classification models.

4.2. Classification models

To understand which model would be most relevant for our study, we could use the following models with evaluation:

- K-Nearest Neighbors (KNN)
- Decision Tree
- Logistic Regression (LR)
- Support Vector Machine (SVM)

After Data Cleaning we have ~75k rows for 'Property Damage' and ~34k for 'Injury' cases which means our dataset is Unbalanced. To fix that, we could Oversampling (Up-sampling) data of the minority class (*'injury' cases*), so we would have Balanced dataset (*~75k for each of the class*).

Unfortunately, not all data in our dataset is ready for Modeling - most of the columns contain 'object' type instead of numerical (int), therefore we have to convert them by label encoding to required data types (int).

4.3. Performances of different models

For evaluation performance of different models, we have to find out the best parameters for every model used. Let's look at the bast parameters calculated for each model in the table below.

MODEL	BEST PARAMETERS	ACCURACY
KNN	k = 21	**0.534**
DECISION TREE	criterion = 'entropy' max_depth = 50 max_features = 'auto' splitter = 'random'	**0.532**
LOGISTIC REGRESSION	C = 1 max_iter = 100 solver = 'lbfgs'	0.521
SVM	decision_function_shape = 'ovo' gamma = 'scale' kernel = 'linear'	0.527

*Table 3: Best parameters for each model. The best result highlighted by **bold***

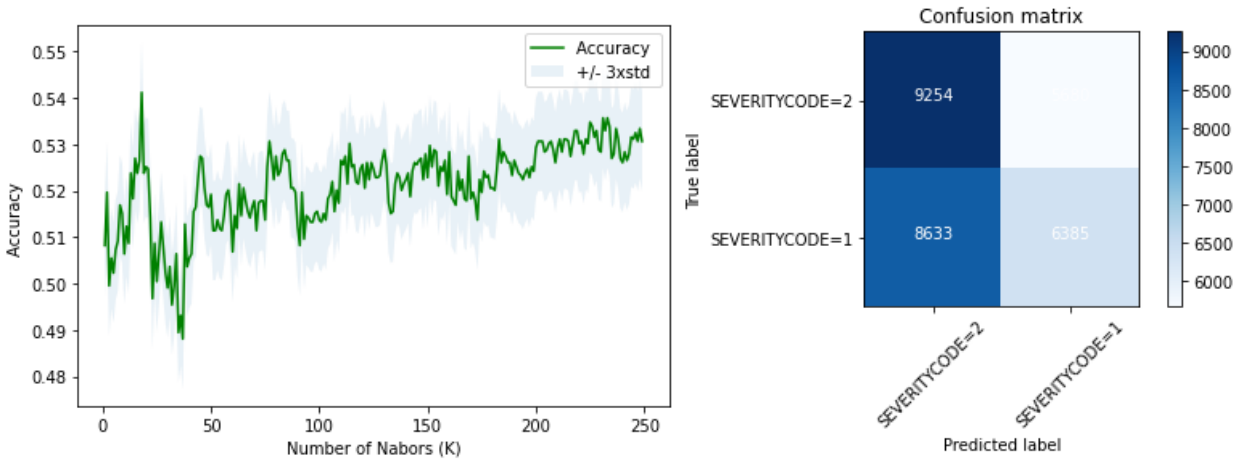
After found out the best parameters for each of the model, we could create them and calculate their performance. The details are provided in the table below.

MODEL	F1 SCORE	ACCURACY	JACCARD SCORE	LOG LOSS
KNN	0.533	0.534	0.429	N/A
DECISION TREE	0.525	0.532	0.305	N/A
LOGISTIC REGRESSION	0.518	0.521	0.284	0.695
SVM	0.521	0.527	0.302	N/A

Table 4: Performances of each model

As we can see from the table above, the best models are KNN and Decision Tree, while Logistic Regression displays the worst result. However, all models in overall displays the very close results.

Below the pictures with different K for KNN (as one of the best models) as well as Confusion Matrix for Logistic Regression (as one of the worst models).



Picture 12: Best K calculation (KNN) and Confusion matrix (LR)

5. Conclusions

Reviewing the details of the results in that particular study, we have analyzed relationship between severity of incidents (property damage vs injuries) based on both – outside conditions as well as human-related (driver condition) conditions. We were able to find out that more injuries rather than property damage cases related to driver conditions, while outside conditions (weather / road / light) is more related to property damage (higher probability). Also, some cases were related to Dark conditions, while Street Lights were Off, and they possibly could be mitigated if Street Lights would be turned On. In addition to that, on snowing roads the most of the cases were happened during morning time, so may be using some salt in the nights/evenings before snowing could help to reduce such cases.

Finally, we have created classification prediction models that based on outside conditions as well as driver conditions could predict (with some level of probability) the severity of incident, if it would happen.

That information could help drivers, road workers as well as police to take some measures for reducing incidents overall and injuries in particular case.

6. Future Directions

The current study was based on SDOT dataset data, limited by only few severity codes as well as only outside and driver conditions were taken in consideration. Both of these types of conditions could be also related to area / district of the city or age of the driver or season of the year (*i.e. winter vs summer*) and models, based on additional features could have better result of prediction probability of the incident severity.