Predicting the severity of car accidents in Seattle for an Insurance Company

Juan Pablo Clavijo Acosta

September 07, 2020

1. Introduction

1.1 Background

Seattle is a seaport city on the West coast of the United States. Seattle is the largest city in the state of Washington and the Northwestern area of North America. There are 740300 people living in the city as of 2018 being the decade's fastest growing major U.S. city. Tourism is Seattle's 4th largest industry, which makes us understand that Seattle is a very busy city with a lot of movement. The city, like any other in the USA, possess different means of transportation, including and not limited to taxi cabs, bus service, rail/train transportation, and even ferries.

1.2 Problem

A growing Health Insurance business settled down in Seattle is concerned about the price and conditions of insurance policies related to car accidents. Stakeholders are having a hard time defining new policies and conditions related to coverage and cost of insurances related to car accidents. As no clear overview can be defined, more information is needed in order to make a decision. Increasing the cost of the insurance would highly impact the desire of customers to acquire the service, so stakeholders want to determine what changes can be made to the coverage of the health policy related to car accidents.

1.3 Interest

Our target audience will be the insurance company, and the customers that acquire the service. The company will highly benefit from the information as if they know and segment their policies depending on the classification of the car accident, they could deduce clauses that would help them reduce costs due to affected customers. Also, the customers acquiring the company service can be benefited as they'll be more aware of the type of accidents where they'll get coverage on, and how they could use the insurance service more effectively.

2. Data acquisition and Data Cleaning

2.1 Data Sources

The data we are working with comes from the SDOT Traffic management division and Traffic records group. All collisions have been provided by SPD and recorded by Traffic Records. The data set contains all type of collisions produced within the limits of the city of Seattle from year 2004 to the present year. IBM Coursera Capstone course provided the dataset on the following link.

2.2 Data Cleaning

The data downloaded is extensive with over 38 columns that represent the attributes related to the collisions or cases reported in the city of Seattle, and over almost 200,000 rows of information. The data set can be quite intimidating at first, but before cleaning the data frame of unnecessary information, we first needed to determine which attributes could help us predict the severity of the accident associated to each case.

But before selecting the attributes necessary to build our model, we started off by cleaning the dataset from missing values and repeating values in certain rows. We eliminated about 10,000 rows of repetitive content, and we did not drop the null rows as there were some columns that presented more than 100,000 null values.

Columns that represented inattention, involvement of pedestrians, and speeding has the most null values. These columns were categorized with yes and no answers, so I decided to create dummies of these columns and drop the originals in order to remove any null value presented in those columns.

The amount of null values was reduced drastically to 3000 – 6000 rows with null values in critical parameters like junction type and collision type. We proceeded to drop these rows in order to be able to work with this column attributes.

Finally, we defined if there was any incorrect type values in the current data set and found that the collision code was defined as object type rather than integer, so we proceed to change the type to the one that matches.

2.3 Feature set

As we stated before, the data set had over 38 different attribute parameters. I considered the severity index as the dependent variable of the whole set, and then determined the parameters that would be used as independent variables that shared some relationship with the severity index

As such I removed some parameters that I thought did not shared any value in determining the severity index in our current data set.

Kept attributes	Dropped attributes	Reason
ADDRTYPE	X, Y, LOCATION	Location of the accident was defined as unimportant. More important is where the accident took place
ST_COLCODE, ST_COLDESC	'OBJECTID','INCKEY','COLDETKEY', 'REPORTNO','STATUS','INTKEY'	Dropped all the id relevant columns as they are used just for indexing the collision
SEVERITYCODE	'SEVERITYDESC','SDOT_COLCODE', 'SDOT_COLDESC','SDOTCOLNUM', 'SEVERITYCODE.1'	Similar features, and more indexing

PEDCYCLCOUNT, INCDTTM, PEDROWNOTGRNT	'PERSONCOUNT','PEDCOUNT', 'VEHCOUNT','INCDATE',	Redundant date column, redundant involvement of pedestrians, and numbers of vehicles involved seemed irrelevant for our analysis

Table 1. Feature selection in our data cleaning process

After dropping these columns our data set still has 20 attribute columns to build our predictive model. We proceed with the Exploratory Data Analysis in order to determine which attributes are more relevant for the model.

3. Exploratory Data Analysis (EDA)

3.1 Relationship between accident frequency and weather

Analyzing the weather attribute of the data set, we determined that most of the collision produced in Seattle were under normal weather conditions or in clear days. Around 110,000 collisions were produced under clear sky. As such, I determined that weather is not a good factor to determine the severity of an accident as most of the accidents were produced under normal weather conditions.

3.2 Relationship between accident frequency and road condition

Again, most of the accidents were produced under normal dry road conditions. Over 120,000 accidents were recorded in which the condition of the road was dry, and as such I decided to not include road condition as a factor in determining the severity index as most of the cases registered were under dry road conditions.

3.3 Relationship between accident frequency and lightning condition

Counting values, I determined that most of the accidents were produced in daylight. More than 110,000 cases were produced in daylight, as such I decided not to include light conditions of the road as a factor to determine severity of the accidents.

3.4 Relationship between accident frequency and junction type

I determined that most of the cases were produced Mid-block and not necessary in a junction. Almost 90,000 cases were produced mid-block, with most of the other cases occurring at intersections. Still we don't have much information in how this parameter relates to the severity of the cases. As most of the accidents occurred at mid-block, I decided to analyze if the drivers where speeding next.

3.5 Relationship between accident frequency and speeding

After determining the frequency of speeding cases, I was amazed to encounter so few cases. Around 10,000 cases of person speeding were encountered. Still driving at excess speed is a variable that could let us determine if an accident is associated with a higher severity index, so it should have to be considered in the model.

3.6 Relationship between accident frequency and type of accident

Most of the big picture is drawn in this attribute, as this column indicates how the collision was produced. This column lists almost 62 different types of collisions and if the accident involved cyclist and/or pedestrians. Determining frequency of the cases, I

could see that the data is severely distributed. First, I determined the frequency of the type of collision before analyzing the type of incident associated with the collision. From the data I pulled the following pie chart describing the frequency of the accidents.

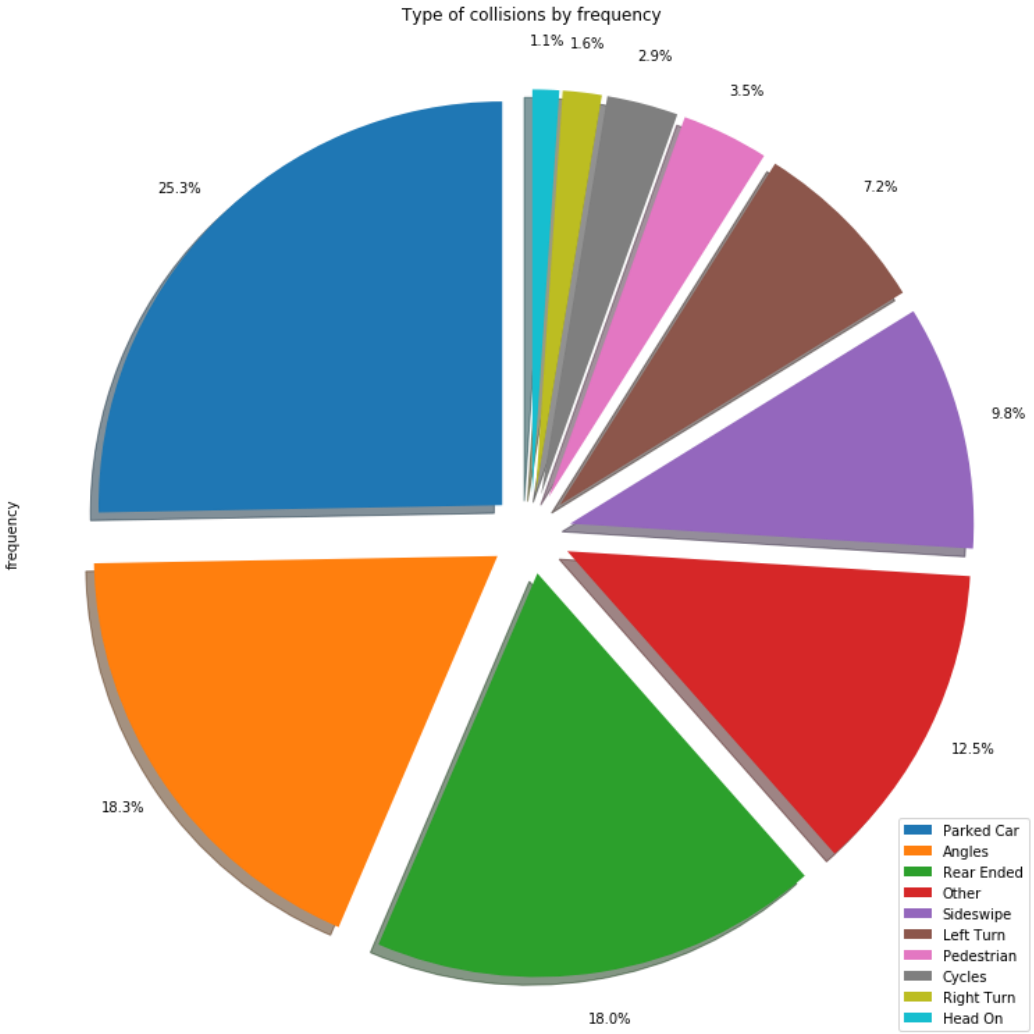

Fig. 1 Frequency of type of collisions registered

Most of the accidents have been registered as collisions involving parked cards. This type of collisions is generally not as severe as others and generally involve property damage. But we still have other cases to analyze that could give us a lead in determining which cases will be associated with a higher severity index.

3.7 Breaking down the severity index

As we encountered our best parameter to analyze if a case is associated with a higher severity index, I decided to determine how many cases involve only property damage, and what other cases involve injuries, serious injuries, and fatalities.

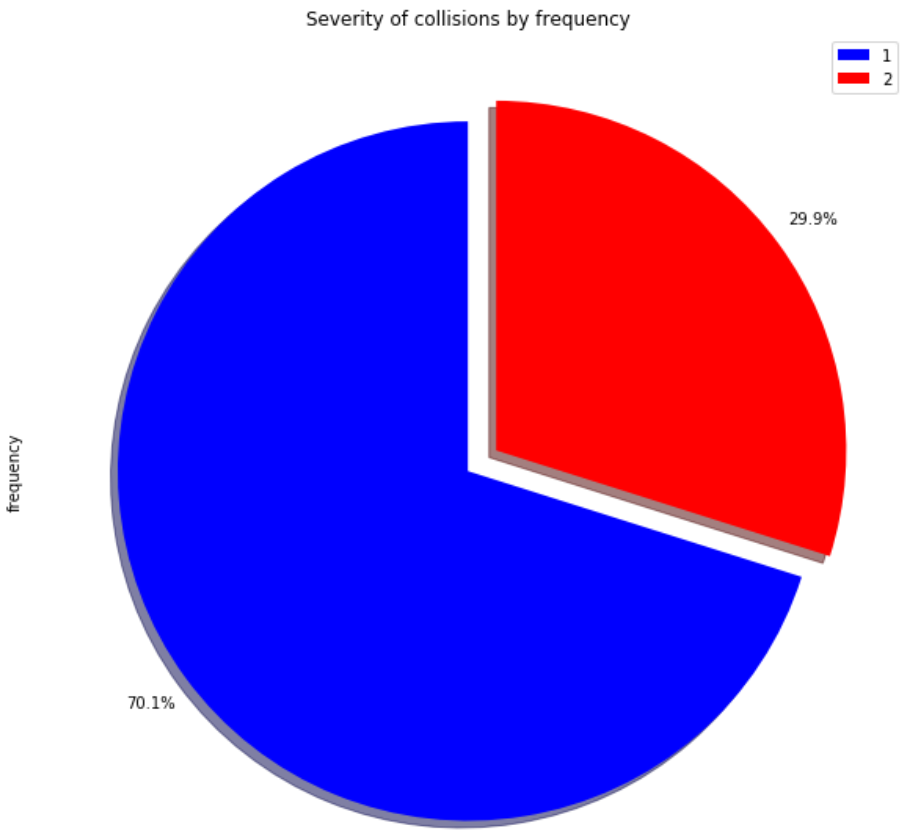

Fig. 2 Frequency of incidents associated to a severity index

Is good to see that we don't have any fatality or serious injury in all of the cases. As such, we will be breaking down our cases to analyze just these two types of severity. Type 1 involves only property damage, and type 2 involves incidents that involve small injuries in the collision. With this new information, we associate and group by the type of incident and the severity of the incident in order to determine a pattern.

3.8 Grouping severity and type of collision

In order to determine a trend, I grouped the attributes in a multi-index table to determine the frequency of the collision and how it associated to a certain type of severity. As a result, I obtained the following graph.

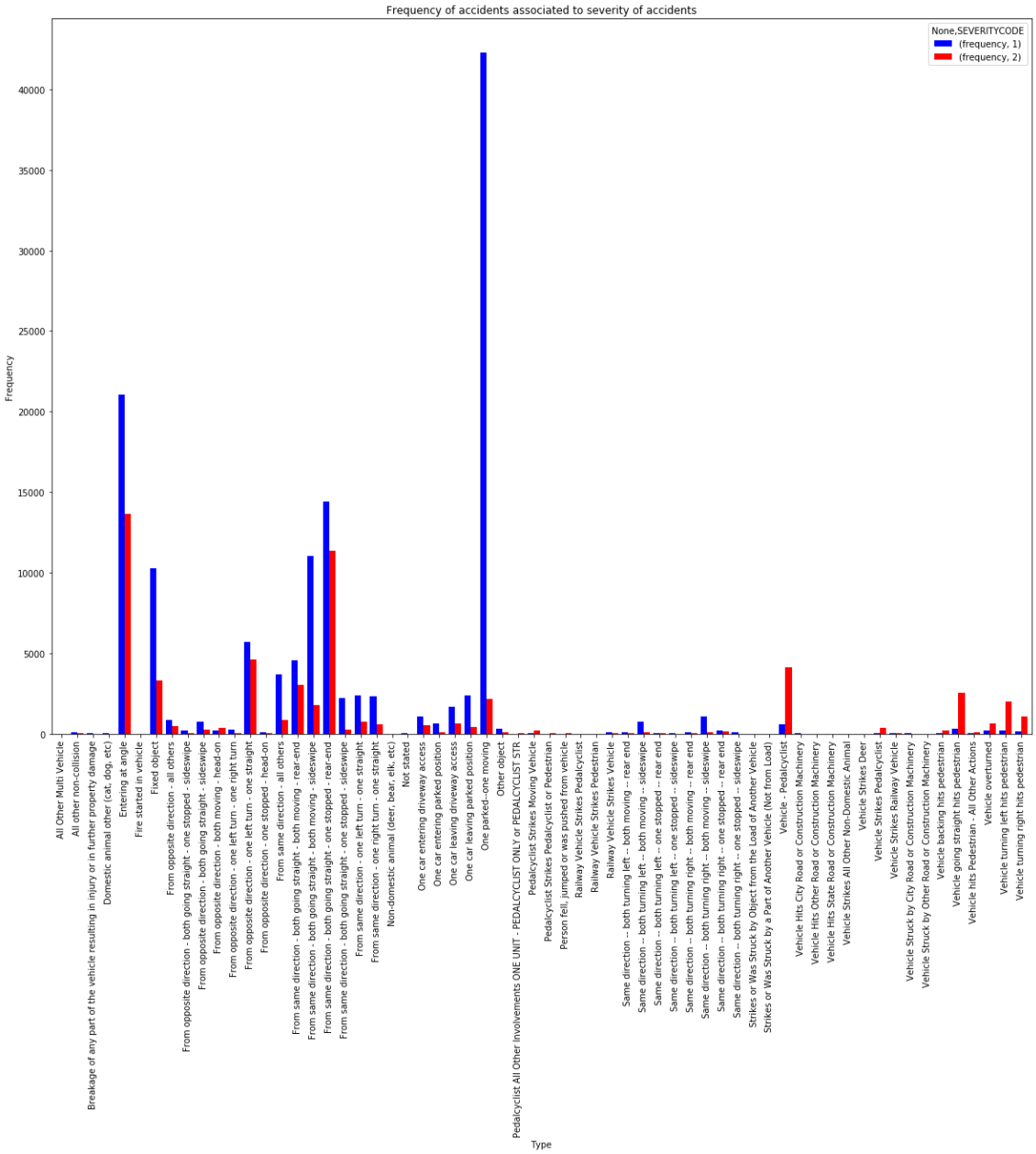

Fig. 3 Cases associated to a severity

As we can see, the most common type of collision is one parked and one moving. The severity associated to this type of collision relates to proprietary damage. On the other hand, we see that incidents produced at angles have the most cases related to injuries in term of severity, but we also see from the graph above that accidents involving pedestrians and cyclists will most probably end in injury. As such, we could use the type of collision in order to predict whether an incident is of class "1" or class "2". This will be our key independent variable. As such, we will add other factors to our clean data frame that will increase the accuracy of our prediction in terms of predicting the severity of the collision.

3.9 Attribute selection

The attribute selection for our final dataframe was based on frequency of the attribute and high correlation between the severity index and the attribute.

Attribute selected	Reason
Alcohol_No, Alcohol_Yes	High correlation with severity index
PEDCYCLCOUNT	High correlation with severity index, easily categorized as injury
ST_COLCODE	Main distribution of cases with a lot of variables to analyze
HitParkedYes, HitParkedNo	High correlation, it can be easily categorized as property damage
Inattention	High correlation with severity index
Speeding	High correlation with severity index, easily categorized as injury
Pedestrians	High correlation with severity index, easily categorized as injury

Table 2. Selection of attributes for the model

4. Predictive Modeling

As our model wants to determine the class or type of severity associated with the attributes selected, I decided to base the predictive model in a classification model, in which the result could be either 1 or 2. As such I decided to test the model using two different approaches, using K-Nearest Neighbors, and Logistic Regression.

4.1 Preparing the data

I started constructing the array selecting the specified attributes as our featured parameters and the severity code as our dependent variable. After that I normalized the set and trained it 60% for training and 40% for testing.

4.2 KNN approach

After testing the model, I found out that the best k to work with was 12 and obtained the following results for the accuracy of the model.

	Accuracy score	F1 score	Jacquard Similarity score
KNN	0.7407	0.6863	0.7407

4.3 Logistic Regression

I used Logistic regression as a different approach to corroborate the results of the KNN approach. I also decided to build a confusion matrix to determine how the values are being predicted and how accurate the model is. As such, I obtained the following results

	Acc Score	F1 score	Jacquard Similarity score	Log loss
Logistic Reg	0.7222	0.6536	0.7222	0.5625

As for the confusion matrix:

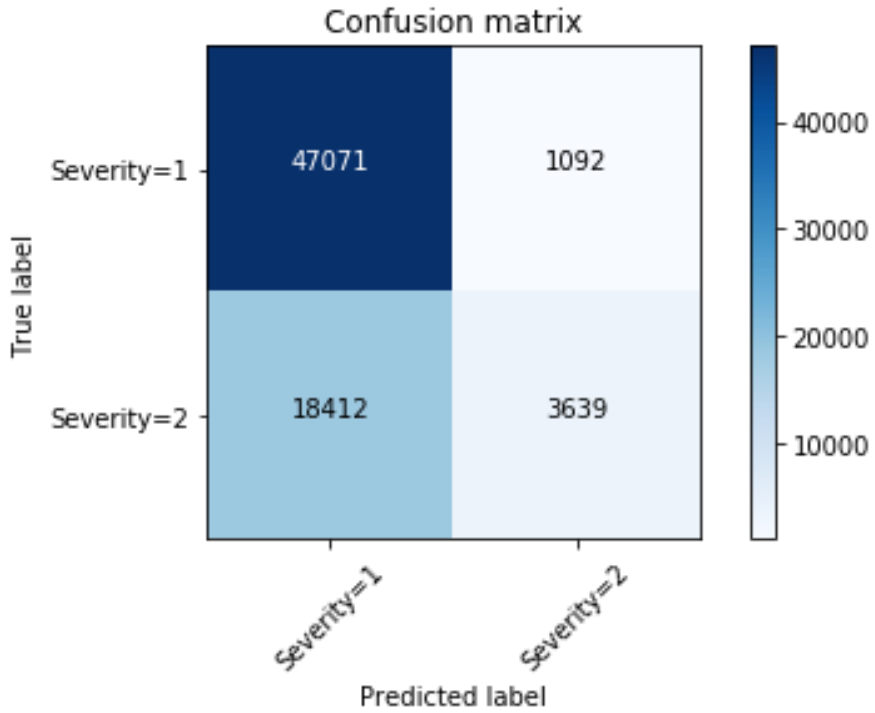

Fig. 4 Confusion Matrix

As we can see from the model we built, it can accurately predict accidents with a low severity code, but it is not that accurate predicting if an accident is associated to a severity code of 2 (injury related).

5. **Discussion of results**

From the model we built, it is clear that the data used to predict the severity of the accidents still suffers from uncertainty as we have a high accuracy rate for accidents related to property damage, but we have a low accuracy score when determining accidents that are related to injuries of the persons involved. Still the overall prediction score using KNN model and Logistic Regression has an accuracy rate of 73%, so we can still consider our model as accurate to determine the severity of an accident.

As the most common type of collisions are against parked cars, this parameter may have influenced the model to be more biased toward property damage as we determined that this type of collisions was associated to a lower severity index. To build a more accurate model we will need further information or attributes that could help us determine a better relationship between the severity index and the attribute selected.

6. **Conclusions**

From the recollection of the data, we have determined that the accidents that involve a higher severity index are those related to accidents involving pedestrians and cyclist with almost 100% certainty. So, we recommend to the insurance company that it should modify policies of car accidents that are related to hitting pedestrians or cyclists, and to some extent motorcycles. This way, the insurance company can decide if they reduce the coverage of the person that caused the accident and covering most of the medical expenses of the pedestrian or cyclist involved in the accident.