IBM Applied Data Science Capstone Project

Predicting the severity of a car accident in the Seattle area

Rafael Mata M.

September 22 – 2020

Table of Contents

1. Introduction

1.1. Background

Nowadays the car is one of the most used medium to travel, to go to work, move between cities,etc. This phenomenon increases the number of cars in the streets and the possibilities of accidents between cars, cyclist or pedestrians.

Also conditions like location, weather, road staus, speed, light and others can influence the accident odds and this could result in injuries, car and property damage, fatalities, financial impact, medical bills, emotional impact or long-term consequences.

1.2. Business Problem

The question or problem to be answered with this project is knowing certain conditions (weather, location, day, road status, etc) what is the severity when you have a car accident.

1.3. Interest

Stakeholder groups that are affected for the car accidents are drivers, pedestrians, cyclist, local and regional authorities and others involved in the accidents.

This will help the stakeholders in different ways, for example:

- Avoiding to drive when there are some risky conditions
- Drive carefully due to certain conditions
- Reduce fatalities
- Reduce medical bills
- Minimizing fatal/injury car crash
- Identify locations with highest accidents rate and take actions

2. Data Understanding

2.1. Data source

In order to address the problem, there is a dataset called **Collision -All Years** which has data for the Seattle city car accidents. The file is in .csv format, and can be found in this link: DATASET, also a description for each file is HERE

The dataset contains **194673** observations and **38** features or columns, this is the list of features:

Table #1
Dataset fields

FIELD	FIELD	FIELD	FIELD	FIELD
SEVERITYCODE	ADDRTYPE	PERSONCOUNT	SDOT_COLDESC	SPEEDING
X	INTKEY	PEDCOUNT	INATTENTIONIND	ST_COLCODE
Y	LOCATION	PEDCYLCOUNT	UNDERINFL	ST_COLDESC
OBJECTID	EXCEPTRSNCODE	VEHCOUNT	WEATHER	SEGLANEKEY
INCKEY	EXCEPTRSNDESC	INCDATE	ROADCOND	CROSSWALKKEY
COLDETKEY	SEVERITYCODE.1	INCDTTM	LIGHTCOND	HITPARKEDCAR
REPORTNO	SEVERITYDESC	JUNCTIONTYPE	PEDROWNOTGRNT	
STATUS	COLLISIONTYPE	SDOT_COLCODE	SDOTCOLNUM	

2.2. Data cleaning and feature selection

Analyzing the dataset:

- 6 of the features have many missing values, for modeling those features that have more than 50% missing data will not be used for trainning or testing.

Table #2
Features with more NaN values

FIELD	%NaN	FIELD	%NaN	FIELD	%NaN
INTKEY	66.57%	EXCEPTRSNDESC	97.10%	INATTENTIONIND	84.69%
PEDROWNOTGRNT	97.60%	SPEEDING	95.20%.	EXCEPTRSNCODE	56.43%

- The target label is unbalance with 136485 for value 1 and 58188 for value 2 severity codes, so for trainning this columns must be balance
- The target is duplicate with the feature SEVERITYCODE.1 and SEVERITYDESC, so these features will not be considered
- There are features that are unique values that does not add significance to the model, so won´t be used as features for modeling.

Table #3
Features with unique values

FIELD	FIELD	FIELD	FIELD
OBJECTID	COLDEKEY	INTKEY	CROSSWALKKEY
INCKEY	REPORTNO	SEGLANEKEY	SDOTCOLNUM

- The feature status is totally umbalanced and is very similar to the target so won´t be used
- INCDATE and INCDTTM have the same date info so only INCOTTM will be used that have in addition the hour
- ADDRTYPE and JUNCTIONTYPE have the same information with JUNCTIONTYPE wiht more detail, so only this will be used
- FOR PEDCOUNT and PEDCYLCOUNT both are considered in the field COLLISIONTYPE so are not used
- SDOT_COLDESC and ST_COLDESC are descriptions of SDOT_COLCODE, ST_COLCODE are reduntdant and not used
- SDOT_COLCODE and ST_COLCODE describe similar conditions so only ST_COLCODE is used it has a better distribution
- HITPARKEDCAR is unbalanced and this characteristic is also included in the ST_COLCODE so is not used
- WEATHER and ROADCOND share similar characteristics and have a direct relation so only ROADCOND is used
- VEHCOUNT and PERSONCOUNT have many outliers so will need some cleaning to be used in the model

- UNDERINFL has letters (Y/N) mix with numbers (1/0) so needs to be fixed to be used in the model

After the analysis and data understanding, removing the features with excesive Nan values and the features with unique values, the list of features to be used in the model are:

Table #4
List of features to be used in the model

FEATURES	Description
X,Y	Geographic location of the accident
COLLISIONTYPE	Type of collision
VEHCOUNT	Number of vehicles involve in the accident
PERSONCOUNT	Number of persons involve in the accident
INCDTTM	Date and hour of the accident
JUNCTIONTYPE	Category of junction at which collision took place
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.
ROADCOND	Conditions of the road
LIGHTCOND	Light conditions
ST_COLCODE	A code provided by the state that describes the collision

Table #5
Target for the model

TARGET	Description
SEVERITY	Severity of the car accident

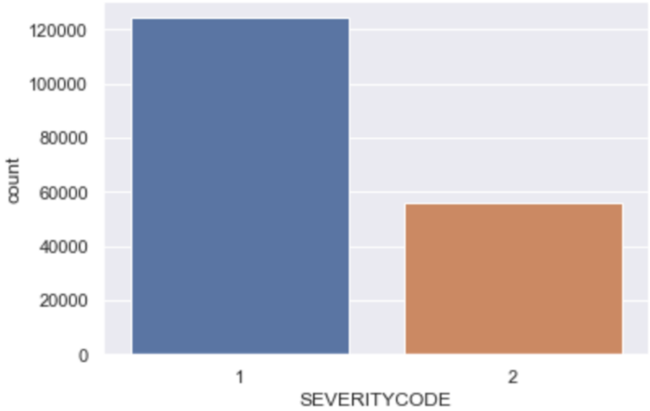
3. Exploratory data analysis methodology

3.1. The target feature

The target feature has two values, 1 and 2 that represent the severity of the car accident, but when are analyzed it shows a clearly unbalance between them.

Fig #1

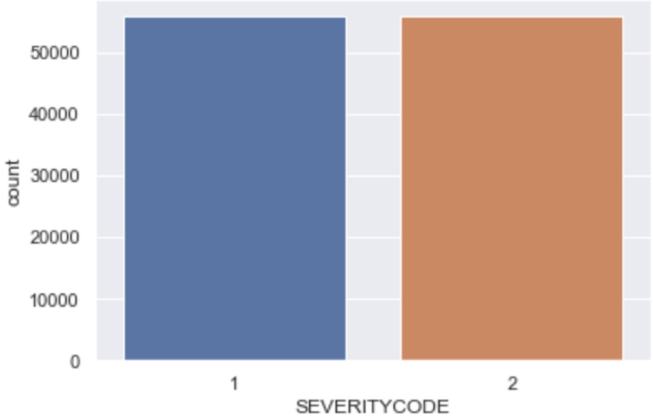
Target feature distribution



When modeling if the target feature of the dataset is unbalance this could affect the prediction and classification, so the dataset is downsample for the severity code 1 to balance the dataset, using a random sample with a size equal to severity 2 observations.

Fig #2

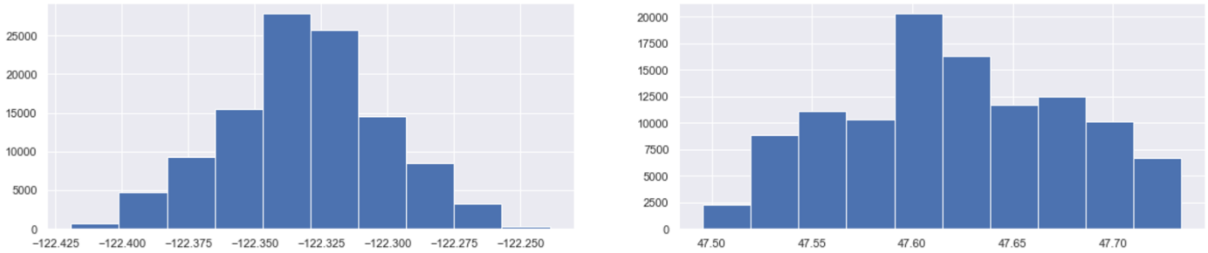
Target feature distribution after balance



3.2. Accidents severity and relation with the location

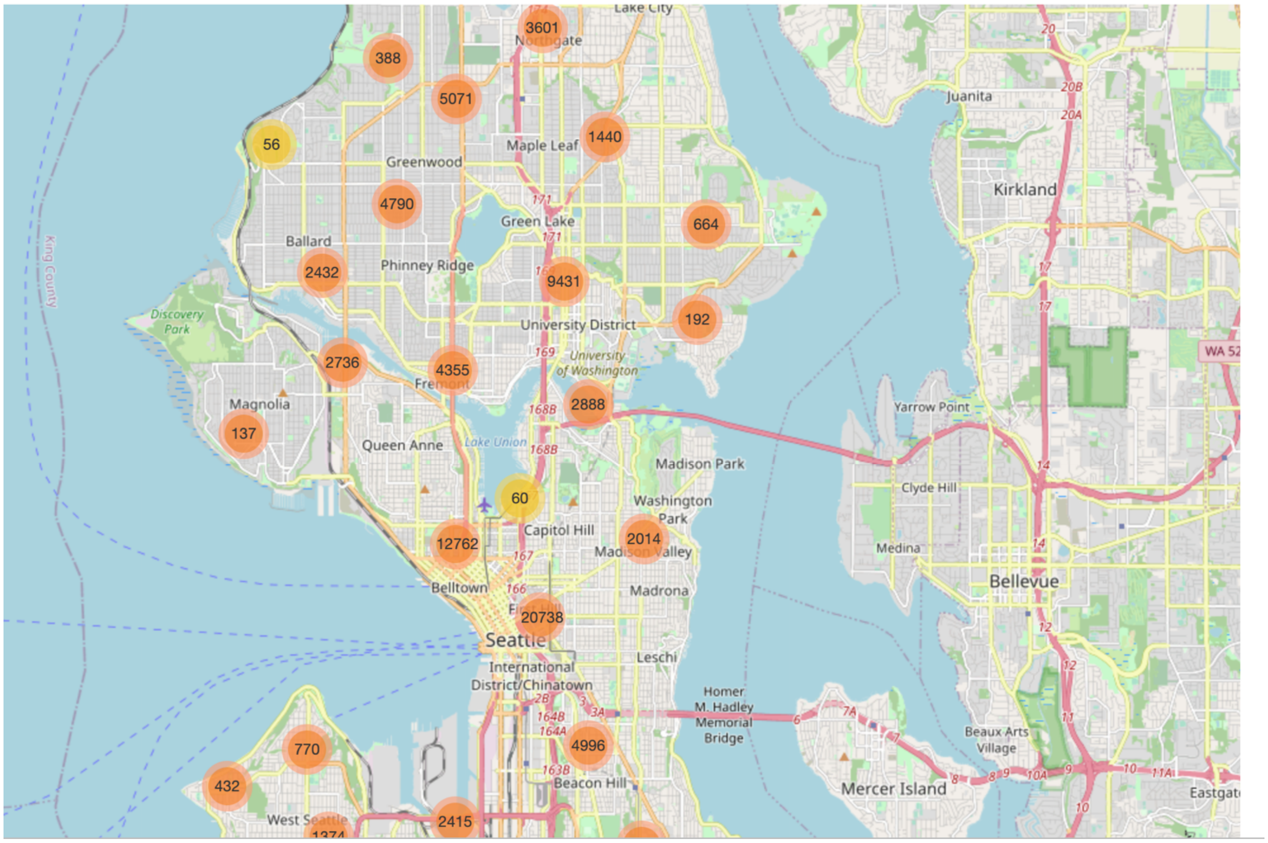
The X,Y fields are the longitude and latitude where the accident occurred, analyzing the distribution, some of them are concentrated in specific areas .

Fig #3
Accident coordinates distribution



And it can be seen in a map with the car accidents counts in the Seattle area.

Fig #4
Accidents count and distribution in the Seattle area



3.3. Relation between vehicles, persons involved in the accidents and severity

The persons and vehicle involved in the accidents are in the range of 1 to 4 in the majority of the observations but in a few cases it was found that 80 persons were involved in a car accident this is understanded as an outlier and those observations were removed from the dataset, when you explore the relation between accident severity, persons and vehicles there are no big difference.

Fig #5

Accidents vs people affected

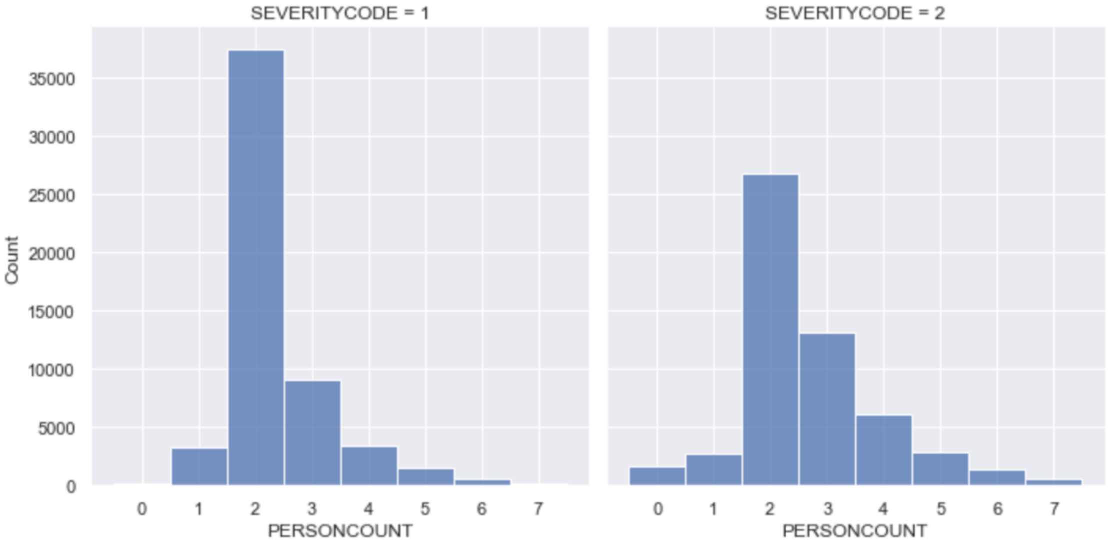
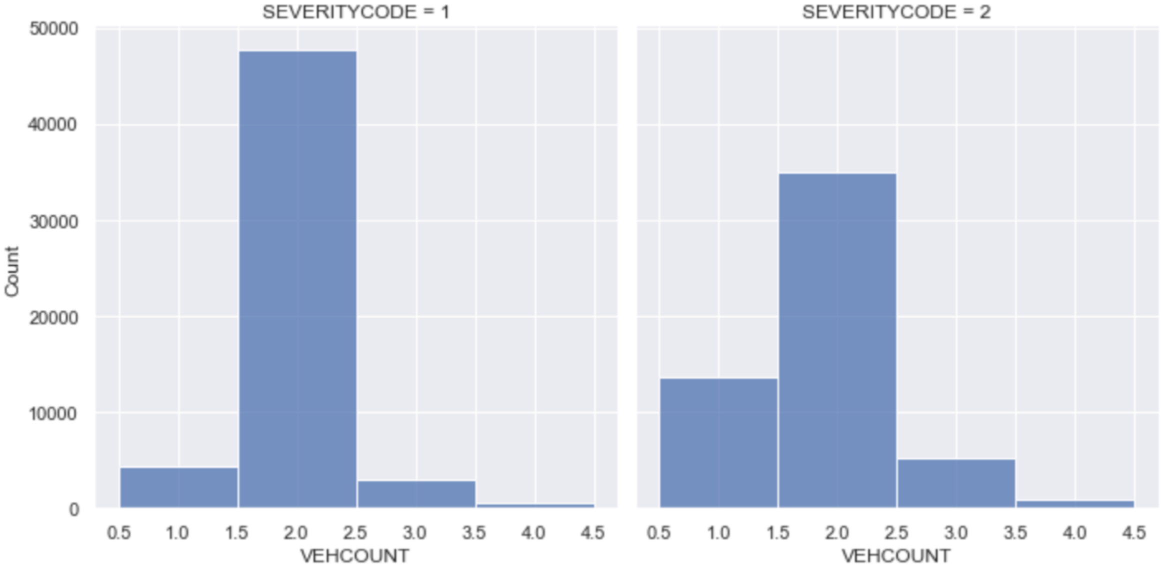


Fig #6

Accidents vs cars involved



3.4. Relation between collision type and severity

The collision type distribution is centered around two types of collisions, angles and reared card for both severities and Parked car collision for severity one.

Fig #7

Collisions types vs severity

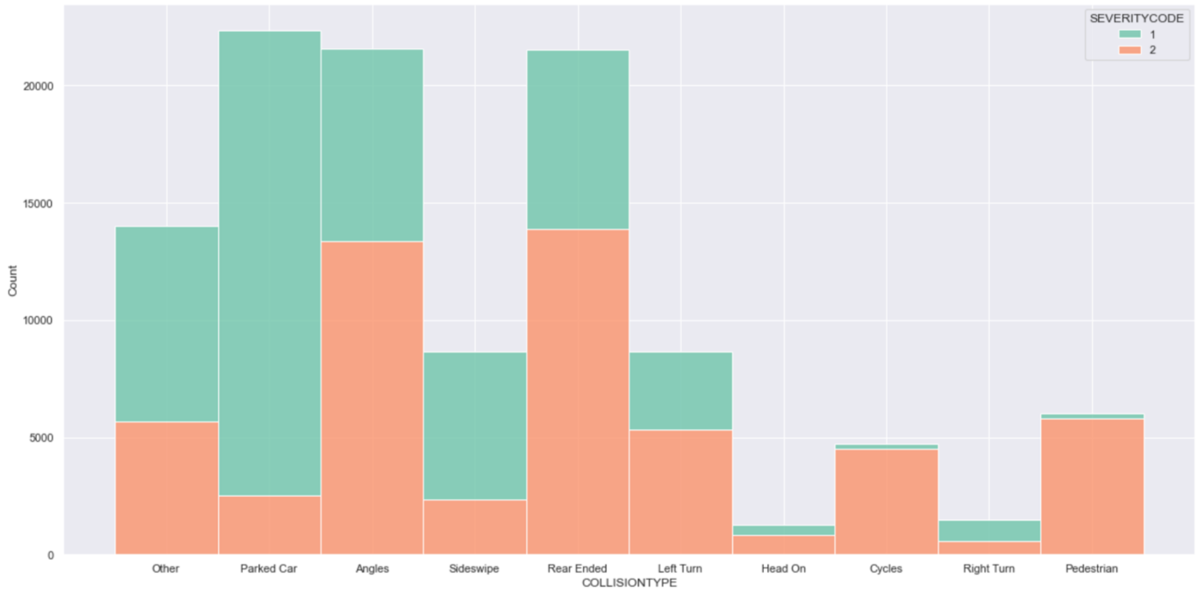
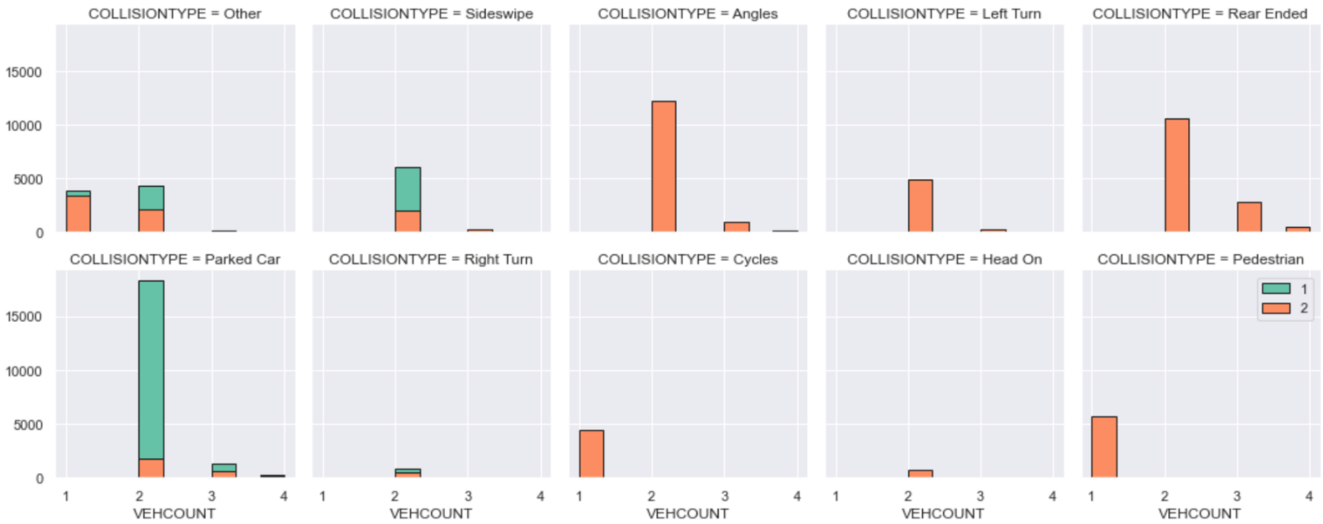


Fig #8

Collisions types vs severity and vehicles involved

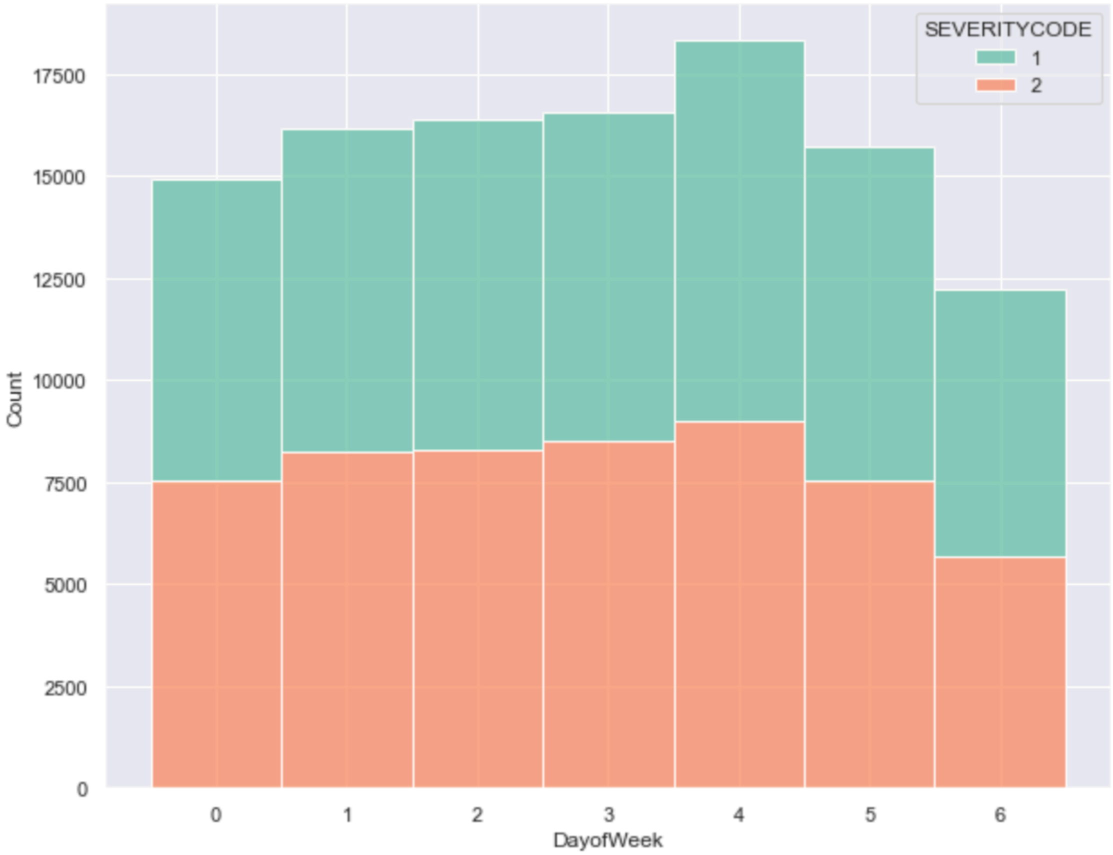


3.5. Relation between day of the week and severity

The field INCDTTM contains information about the date of the car accident, to find a relation this information is converted to day of the week to check if according the day there is more accidents or the severity distribution is related with the day, when the data is analyzed there is a little increased for the Fridays and decreased in Sundays, but the severity is balanced for all days of the week, so there is no clear relation between some specific days and the quantity of accidents or severity.

Fig #9

Relation between day of the week and accident severity



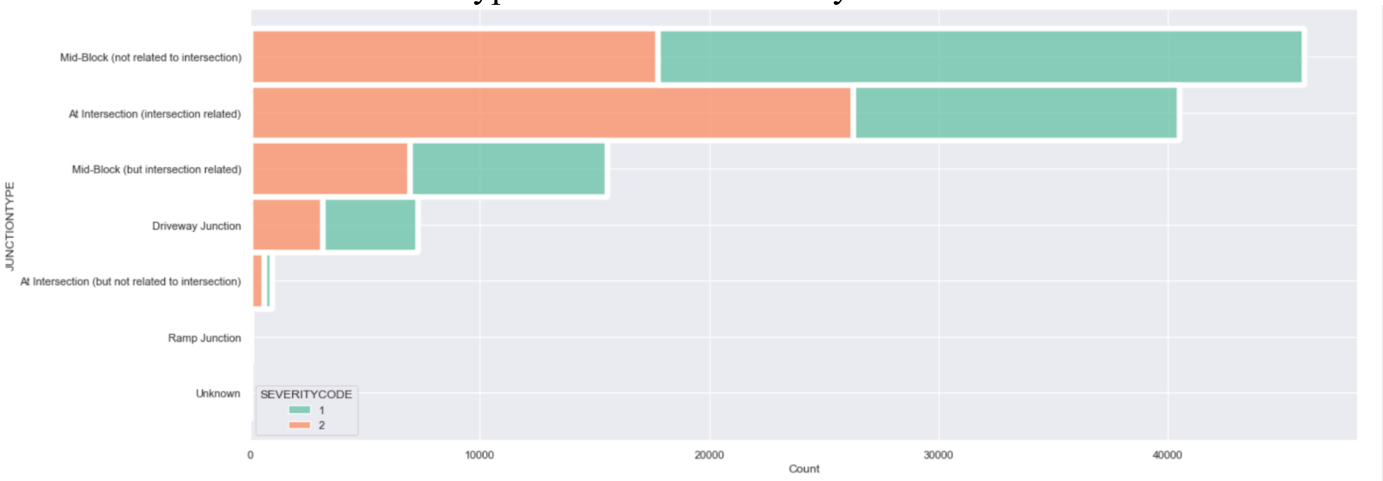
0: Monday

3.6. Relation between JUNCTION type and severity

The distribution of the Junction type and accidents is around two types at the Mid block and intersection, the rest of junction types have only a few accidents, also the severity is balanced between these two types of junctions.

Fig #10

Junction type and accident severity

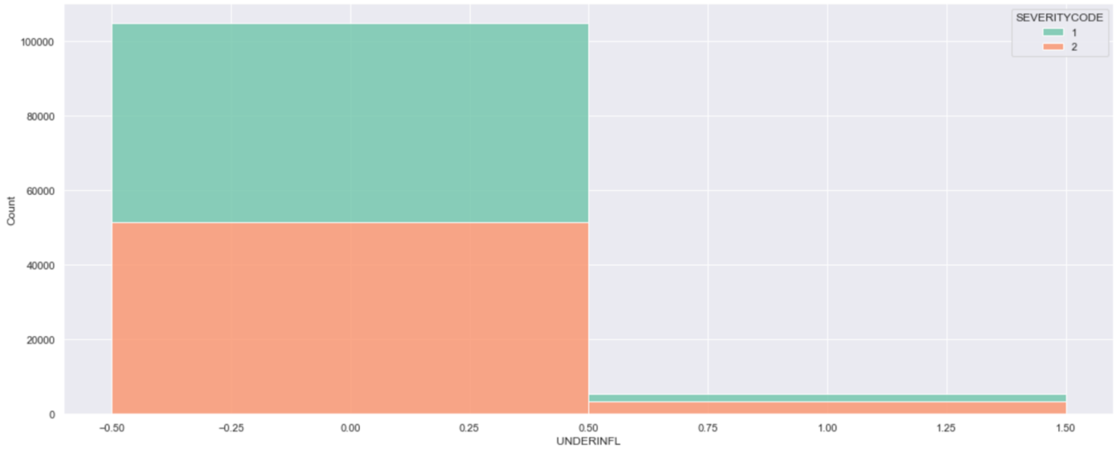


3.7. Relation between accident severity and drugs influence

Most of the observations occurred with no drugs influence, only a few samples are with the influence of substances but not represent a clear relation with the accidents quantity or severity.

Fig #11

Influence of drugs and accidents

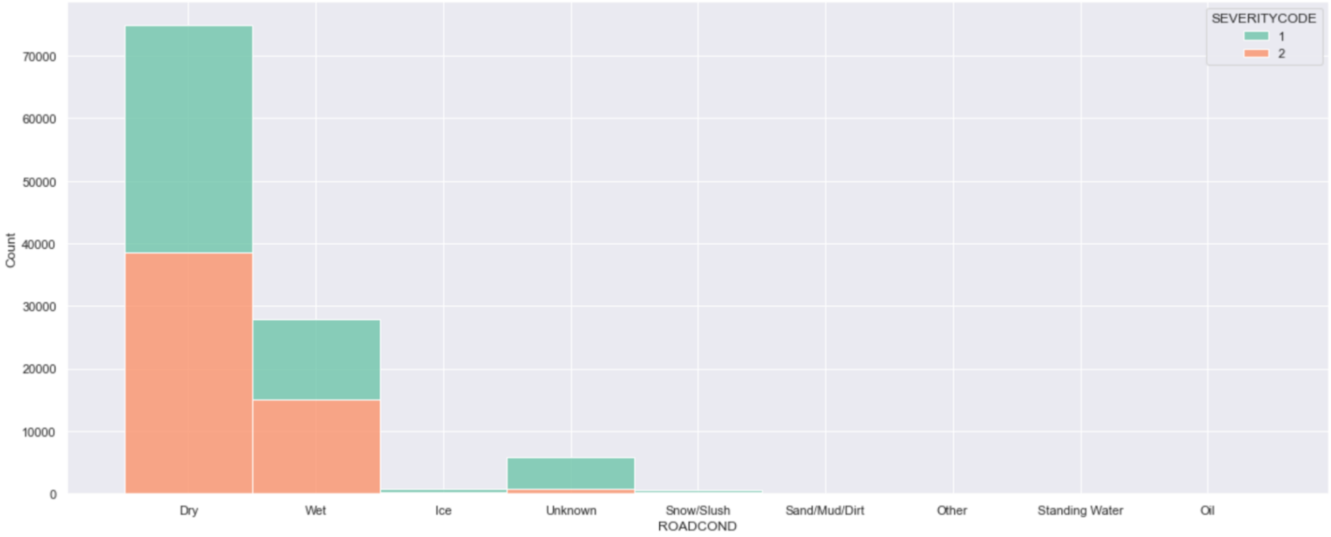


3.8. Road condition and accident severity relation

The road condition is an important factor that affects the car accidents, it is observed in the dataset that most of the accident occurred during dry conditions and in minor quantity during wet road conditions, I hypothesized that during wet or snowing conditions could occurred more accidents but according with the observations it is not the case, also the severity is balanced between all the different conditions .

Figure #12

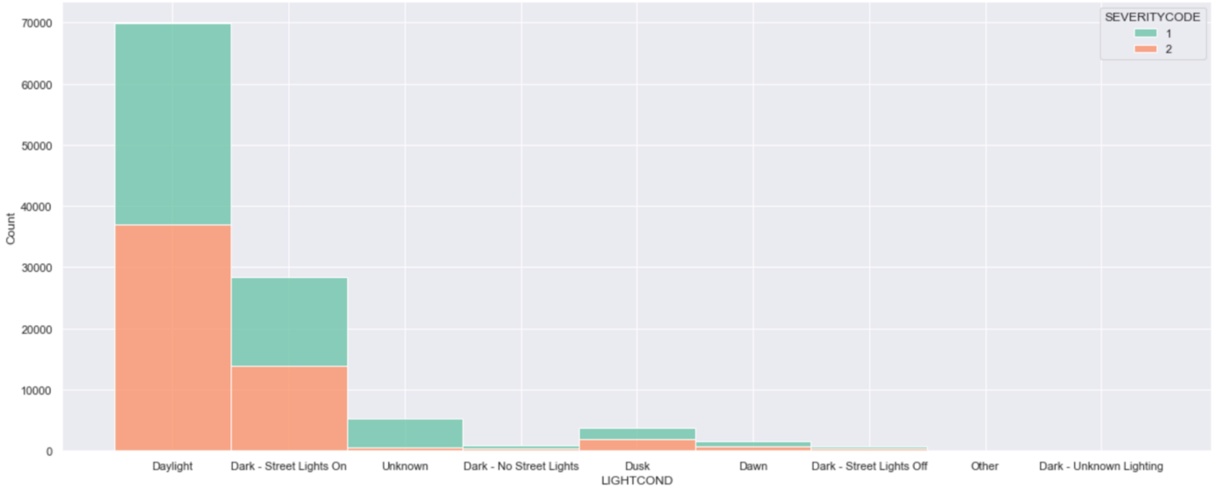
Road conditions and accident severity



3.9. Light condition and accident severity relation

The light conditions observations shows that the majority of the accidents occurred during day light and other important quantity during night but with street lights on, so the lights conditions does not have an important effect on the accidents, the severity of the accidents is balanced between all the different light conditions.

Figure #13
Light conditions and accident severity



3.10. ST_Colcode and accident severity relation

The ST_Colcode describe the collision type, when the dataset is analyzed it shows that the accidents with codes 10, 32 and 14 are the most frequent. The codes 10 is when the collision occurred when entering at angle, the code 32 when one car is parked and the other is moving and the code 14 when cars are going in the same direction straight.

Table #6
Street Collision codes summary

ST_COLCODE	Quantity
10	21560
32	20582
14	16732
28	6850

4. Predictive Modeling results

For this problem supervised learning models are used to predict the severity of the accidents, in this category there are different types of classification machine learning algorithms that can be used, three models are considered:

- KNN (K-nearest neighbors)
- SVM (Support vector machine)
- Logistic regression
-

The KNN and SVM are good to classify and in this case with two severities is a kind of binary classification problem, with the logistic regression also is possible to predict the probability of the severity.

To use the classification models all the information must be transformed to numeric values, in the dataset only the SEVERITY and location fields are in numeric format so all the other fields are coded in numeric values, also is very important for the models KNN and SVM that used distant that all the values are in the same range so the features are normalized in the range 0 – 1.

With all the features coded in numeric format, correlation is calculated to observe if there is relation between the features using a heat map with coefficients.
The figure 14 shows that there is no positive or negative relation between the features.

To train, evaluate and test the models the dataset is split in different groups in this way:

- 80% for training and evaluating
- 20% for testing and run the model on unseen data

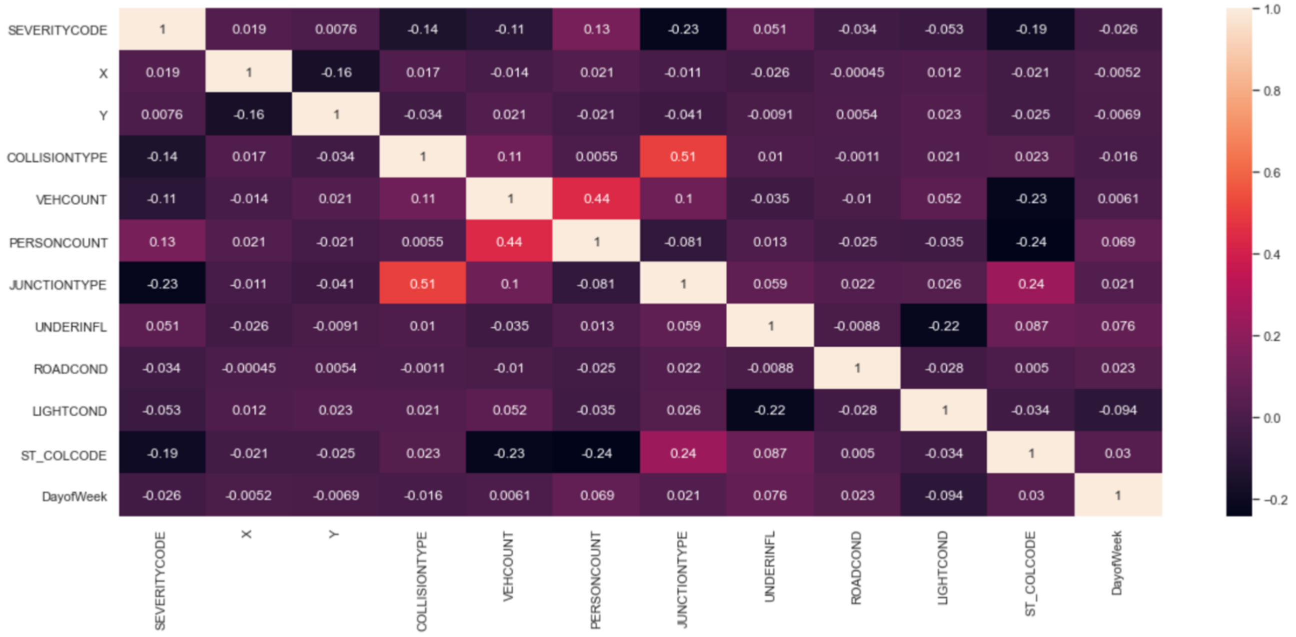
And the training part is also split in this form:

- 80% for training
- 20% for evaluating and calibrate the hyperparameter of each model
-

This schema prevents the overfitting and generalize well from training to testing information for the models.

Figure #14

Features correlation coefficient



	SEVERITYCODE	X	Y	COLLISIONTYPE	VEHCOUNT	PERSONCOUNT	JUNCTIONTYPE	UNDERINFL	ROADCOND	LIGHTCOND	ST_COLCODE	DayofWeek
SEVERITYCODE	1	0.019	0.0076	-0.14	-0.11	0.13	-0.23	0.051	-0.034	-0.053	-0.19	-0.026
X	0.019	1	-0.16	0.017	-0.014	0.021	-0.011	-0.026	-0.00045	0.012	-0.021	-0.0052
Y	0.0076	-0.16	1	-0.034	0.021	-0.021	-0.041	-0.0091	0.0054	0.023	-0.025	-0.0069
COLLISIONTYPE	-0.14	0.017	-0.034	1	0.11	0.0055	0.51	0.01	-0.0011	0.021	0.023	-0.016
VEHCOUNT	-0.11	-0.014	0.021	0.11	1	0.44	0.1	-0.035	-0.01	0.052	-0.23	0.0061
PERSONCOUNT	0.13	0.021	-0.021	0.0055	0.44	1	-0.081	0.013	-0.025	-0.035	-0.24	0.069
JUNCTIONTYPE	-0.23	-0.011	-0.041	0.51	0.1	-0.081	1	0.059	0.022	0.026	0.24	0.021
UNDERINFL	0.051	-0.026	-0.0091	0.01	-0.035	0.013	0.059	1	-0.0088	-0.22	0.087	0.076
ROADCOND	-0.034	-0.00045	0.0054	-0.0011	-0.01	-0.025	0.022	-0.0088	1	-0.028	0.005	0.023
LIGHTCOND	-0.053	0.012	0.023	0.021	0.052	-0.035	0.026	-0.22	-0.028	1	-0.034	-0.094
ST_COLCODE	-0.19	-0.021	-0.025	0.023	-0.23	-0.24	0.24	0.087	0.005	-0.034	1	0.03
DayofWeek	-0.026	-0.0052	-0.0069	-0.016	0.0061	0.069	0.021	0.076	0.023	-0.094	0.03	1

4.1.Model evaluation

Using the training and evaluation datasets each of the models are trained and their hyper parameters adjusted using grid search method to get the best performance in terms of the different metrics, the metrics used to evaluate the methods are:

- F1 score
- Recall score
- Accuracy score
- Precision score
- MCC Score

In the initial stages of the model evaluation the performance was below 50%, when analyzed for each model the reason was the unbalance datataset and the unnormalized data, once this was fix a 15% increase in performance was achieved, and to get the best results the grid search methodology was used to find the right hyper parameters for each model, this way an addition 3% in the performance was gotten.

The table #7 shows the final metrics results for each of the models, and the figure #15 shows the confusion matrix, after evaluating the results the SVM and KNN method gets the best results in predicting the accident severity in the area of Seattle and the Logistic regression method has the lowest performance.

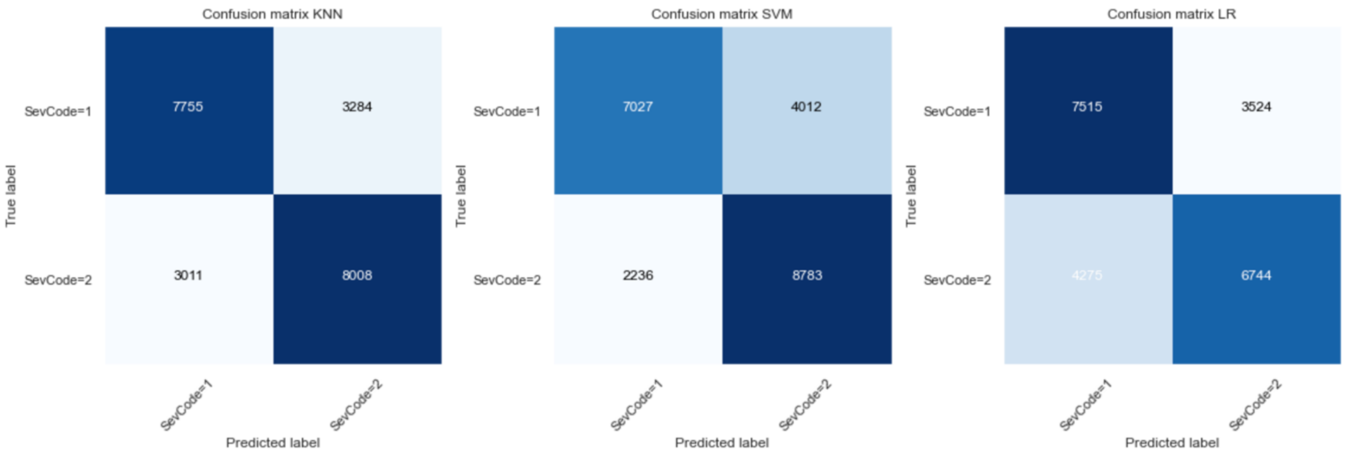
Table #7

Metrics results for each model

Algorithm	MCC	RECALL	Precision	F1-score	Accuracy	LogLoss
KNN	42.9%	70.3%	72.0%	71.5%	71.5%	NA
SVM	43.9%	63.7%	75.9%	71.5%	71.7%	NA
LogisticRegression	29.4%	68.1%	63.7%	64.6%	64.6%	63.5%

Figure #15

Confusion matrix for each model



Between the SVM and KNN the MCC (Mathew correlation coefficient) is used to find the best model, in this case SVM achieve the best performance in predicting the accident severity .

5. Discussion

Due to the nature of the problem three classification machine learning methods were used to classify the car accident severity, between them SVM has the best results, however a 75% accuracy could be considered low because we are dealing with accidents where people lives are involved, even though this is a good starting point to use the information a take preventive measurements to avoid the accidents or reduce their severity.

6. Conclussion

In this report a car accident dataset from the Seattle area was used to apply different machine learning models to predict the accidents severity, the CRISP-DM model was followed to get the data, cleaned, prepared, analyzed, build the model and evaluate it.

From the possible models to use, three classification models (KNN, SVM, LR) were developed, having the best predicting results the SVM model.

As starting point this model can be used to analyze the accidents in the Seattle area and predict the severity to take preventive measurements and reduce fatalities, financial lost or other associated costs or losses.

7. Future directions

As was mentioned before the models performance must be improved to have a better predicting accuracy at least 10% improve, this could be done using another features that are not in the dataset like cellular phone usage that have an important effect in the accidents rate or another type of model can be used and compare with the models implemented here.