Car Accident Severity

Applied Data Science Capstone

Sumeet Grewal

Table of Contents

Introduction

One of the most dangerous and complex activities we engage in on a daily basis is driving a vehicle. Increasing population density in major cities such as Seattle increases the propensity for traffic incidents and collisions.

According to a dataset on collisions in the city of Seattle, there have been over 190 000 incidents since 2004. Traffic collisions have an enormous impact, most notably in terms of physical injury and in property damage.

It is possible that certain environmental factors and driver behaviours increase the severity of collisions. Some of these conditions could be specific days and times, weather and road conditions, driver inattention, and speeding.

The key audience of this study is the City of Seattle as they would ideally undertake initiatives to reduce the number and severity of traffic collisions.

Project objective

Understand which factors have the greatest impact on the severity of collisions, and identify which mitigation tactics may be effective.

Data

The selected dataset contains all collisions provided by the Seattle Police Department (SPD) from 2004 - Present. It contains over 194000 records and 37 fields which are described in detail in the associated metadata file.

The objective is to develop a model to accurately predict the ‘severitycode’ label. This field currently either contains :

- a value of ‘1’ indicating property damage only or
- a value of ‘2’ indicating physical injury from the collision.

Data

The focus of this study will be on identifying the key contributors to the collision. This table illustrates some of the features that may be of interest

ATTRIBUTE	DATA TYPE	DESCRIPTION
WEATHER	Text	The weather conditions at the time of the collision
ROADCOND	Text	The conditions of the road at the time of the collision
LIGHTCOND	Text	The light conditions at the time of the collision
INATTENTIONIND	Text (Y/N)	Whether or not the collision was due to inattention by the driver
SPEEDING	Text (Y/N)	Whether or not the collision was due to speeding by the driver
PEDROWNOTGRNT	Text (Y/N)	Whether or not the pedestrian right of way was not granted by the driver
UNDERINFL	Text (Y/N)	Whether or not the driver was under the influence of drugs or alcohol

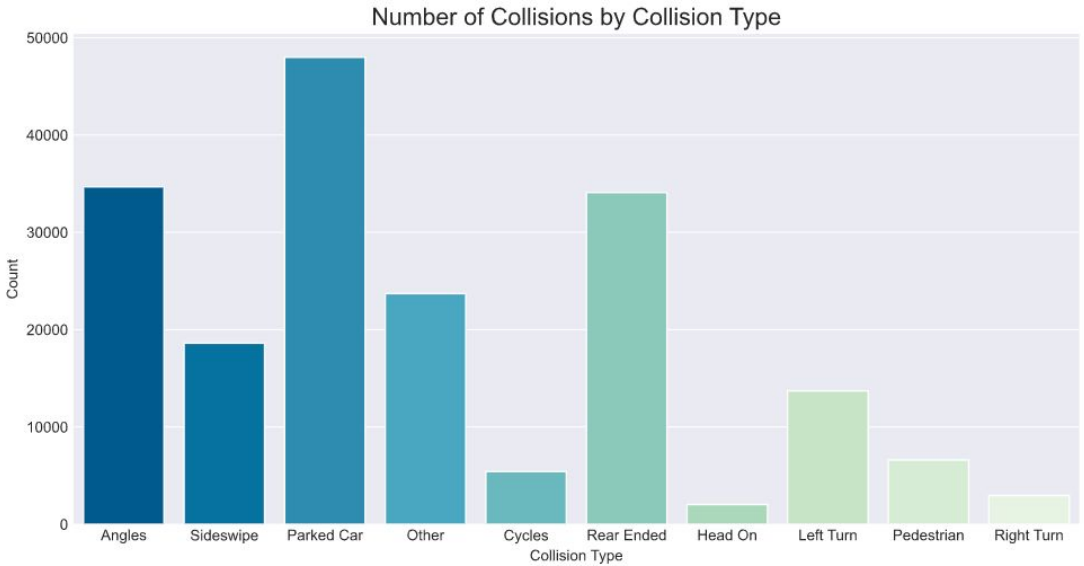
Methodology

02

Once unnecessary columns have been removed, the remaining columns are:

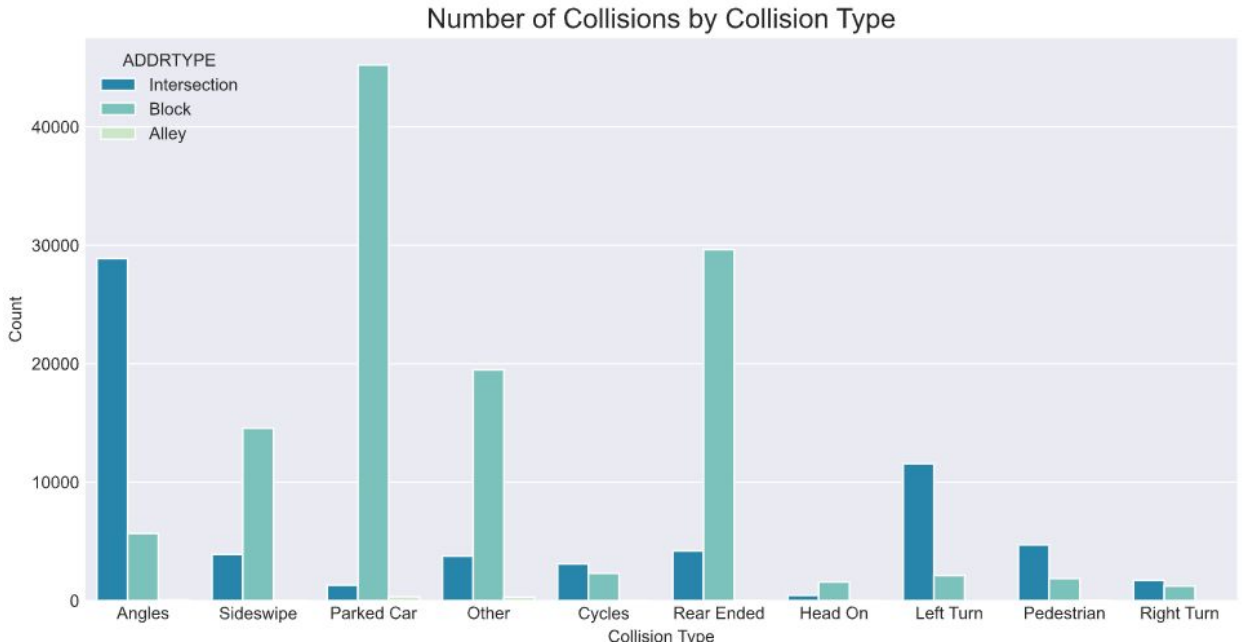
- SEVERITYCODE
- ADDRTYPE
- INTKEY
- COLLISIONTYPE
- INATTENTIONIND
- UNDERINFL
- WEATHER
- ROADCOND
- LIGHTCOND
- PEDROWNOTGRNT
- SPEEDING

A surprisingly significant number of collisions have been recorded where one of the vehicles involved is a parked car. The two next highest collision types are rear-end collisions and collisions at angles.



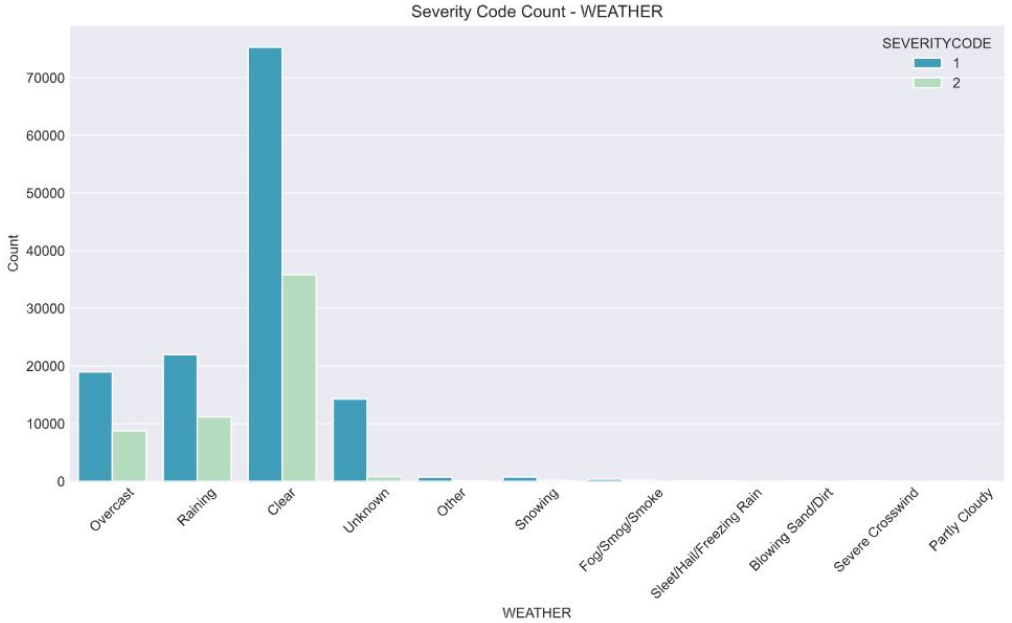
This graph illustrates the address type of each location so we can see which types of collisions are most common on ‘Blocks’ versus ‘Intersections’.

- Most intersection collisions involve angles and left turns.
- Pedestrians are more involved in intersection collisions.
- Block collisions are mostly rear-end, parked car, or sideswipe collisions.



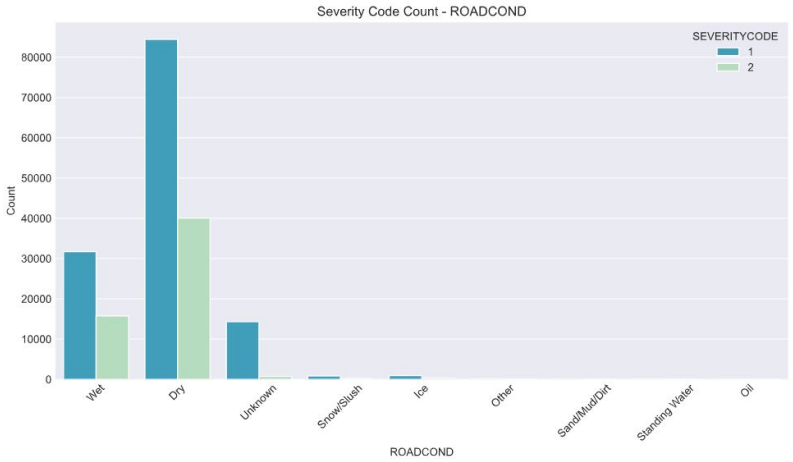
Looking at the weather conditions for the collisions, it is evident that most accidents occur in Clear conditions. An approximately equal proportion of the gross amount of collisions occur during Raining or Overcast collisions.

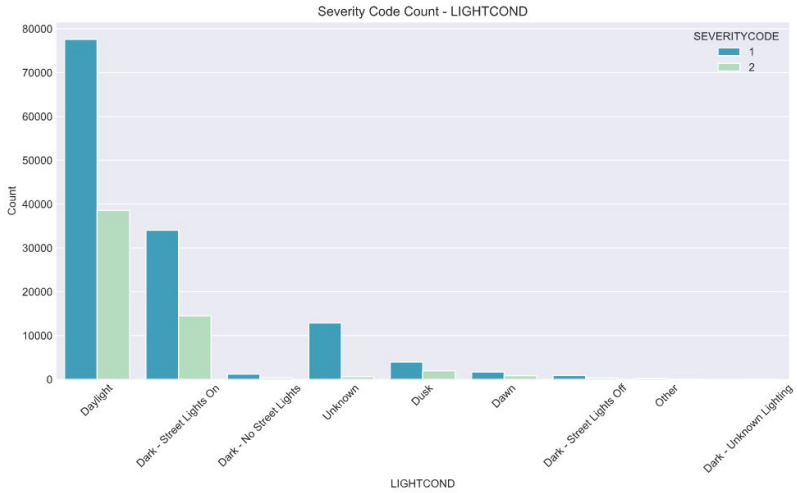
This distribution is likely skewed to reflect the overall incidence of each of these weather conditions. The data for the weather attribute should be normalized against the average distribution of weather conditions in Seattle.



A very similar trend is identifiable in the following graphs of Road Conditions and Light Conditions. Most accidents occur on dry roads and in daylight. One interesting feature of the Light Conditions graph is that more collisions occur at Dusk than at Dawn.

In both Raining and Clear conditions, a severity code value of 2 is approximately half as likely as a value of 1. This is also true for the comparison between Wet and Dry conditions, and the comparison between Daylight and Dark.





Data Cleaning

01

Drop Records

Drop records that have ‘Unknown’, ‘Other’, or empty in the various indicator columns

02

Normalization

Ensure that binary features are all encoded similarly - 1 for True and 0 for False

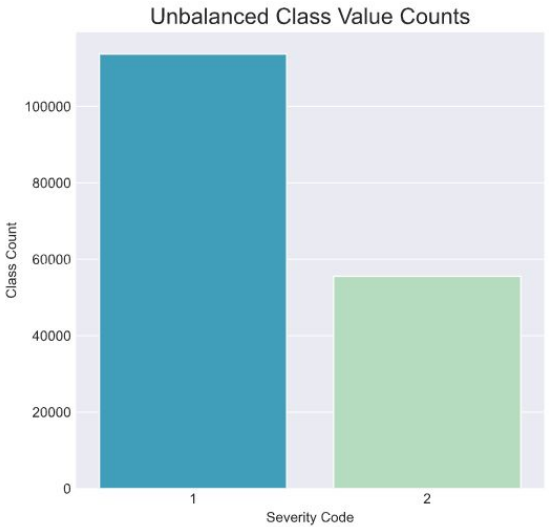
03

One Hot Encoding

Utilize One-Hot-Encoding to convert categorical variables such as Weather to a set of binary, integer indicator variables

Data Balancing

An undersampling technique is used to balance the dataset. An equal number of records with value 1 are randomly sampled from a shuffled dataframe. The sampled dataset is randomly combined with the records with a value of 2 to create a balanced dataframe.



Modeling

Data Selection	The first step is to define the dependent and independent variables. X will include all of the indicator columns that have been prepared and Y will be severity code.
Train/Test Split	The dataset is then divided into mutually exclusive training testing sets. This technique should improve the out-of-sample accuracy of the model. 20% of the records will be reserved for the test set.
Train Model	The logistic regression model is trained using the training dataset.
Evaluate & Tune	Various metrics are used to evaluate and tune the Logistic Regression model including F1-Score, Jaccard Score, and LogLoss metrics.To tune the model, we will first try different solvers to identify which one produces the maximum accuracy and then different values for regularization strength.

Results

From this classification report, we can see that the model is more accurate at predicting records with severity of 1 and not that great at labelling severity code 2. The weighted average f1-score indicates that the average accuracy of our model is approximately 0.52. From this we may conclude that a number of the predictor variables that we defined are likely inconsequential in predicting the severity code.

	precision	recall	f1-score	support
1	0.53	0.80	0.64	11083
2	0.60	0.31	0.41	11133
accuracy			0.55	22216
macro avg	0.57	0.55	0.52	22216
weighted avg	0.57	0.55	0.52	22216

Discussion

1. Based on the analysis in the report, intersection collisions are far more likely to be defined as being a result of angles, left turns, or pedestrians whereas ‘block’ collisions are more likely to involve a parked car or rear ending. The analysis indicates that the type of collision is not very correlated with the severity of an accident.

2. One key area where correlation does exist is between severity code 2 (persons injured) and the Pedestrian Right of Way Not Granted indicator. This makes logical sense as if a pedestrian is not given their right of way, they are more likely to be involved in the collision.

Discussion

3 One interesting difference is that more collisions occur at Dusk than at Dawn. However, most accidents in our dataset occurred in Dry Road Conditions and Daylight. This model doesn't take into account the various distributions of weather. An opportunity to improve the model is to normalize these values according to the actual average distribution of road, light, and weather conditions.

4 From a mitigation perspective, it would be unwise to apply resources uniformly across the city in order to address a specific issue. Mitigation efforts should be concentrated where they would have the greatest impact. It would thus be constructive to identify which intersections are the most dangerous. This way, the City of Seattle can optimally allocate their resources and mitigation tactics towards addressing key factors in key intersections.

Conclusion

A significant number of factors may determine whether a collision occurs and whether or not property is damaged or persons are injured. These include human behaviours such as distracted driving or speeding, and environmental factors such as weather, road, and light conditions.

The objective of this report was to identify if any such factors are accurately able to predict the severity of an accident. Severity was classified as a 1 if only property was damaged or as a 2 if injuries occurred as well.

Conclusion

After preparing, cleaning, and balancing the dataset, a logistic regression model was built to predict the severity code. The dataset was split into 80% for training and 20% for training, in order to improve the out of sample accuracy.

The model itself was able to somewhat accurately predict the class value of 1 but was not very effective at accurately predicting the class value of 2. The results of the analysis determined that the identified factors are not significant indicators of the severity code measurement.