Table of Contents

Seattle Car Accident Severity

Waleed Muhammad

October 26, 2020

1.Introduction

1.1 Background

Say you are driving to another city for work or to visit some friends. It is rainy and windy, and on the way, you come across a terrible traffic jam on the other side of the highway. Long lines of cars barely moving. As you keep driving, police cars start appearing from afar shutting down the highway. Oh, it is an accident and there's a helicopter transporting the ones involved in the crash to the nearest hospital. They must be in critical condition for all of this to be happening. Now, wouldn't it be great if there is something in place that could warn you, given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be, so that you would drive more carefully or even change your travel if you are able to. That is what this project will help uncover and solve.

1.2 Problem

I aim to find correlations in several variables which can lead to a car accident and how severe it is. This information can be used for accident prevention organizations, local governments, etc.

2. Understanding The Data

2.1 Data Cleaning

There are a handful of issues with the dataset. The variables are categorical and there is a lot of missing binary data. We can remove this or fill them. We will remove missing data for those large feature sets.

We are trying to predict the severity of a car accident, but there is a disparity in the amount of accidents that result in injuries and those that result in property damage. Because of this, an imbalanced dataset arises, which is an issue when predicting this model.

Lastly, there seems to be a correlation between certain properties. Light condition, road condition, and weather condition all have certain missing data points. Although, these are Missing Not Randomized (MNAR).

2.2 Feature Selection

The first step consists of 17 features. Many features have been removed because they contain missing values. Here are the features used to build the model:

Feature	Description
addr_type	Collision address type: • Alley • Block • Intersection
collision_type	Collision type
under_infl	Whether or not a driver involved was under the influence of drugs or alcohol.
weather	A description of the weather conditions during the time of the collision.
road_cond	The condition of the road during the collision.
light_cond	The light conditions during the collision.
hit_parked_car	Whether or not the collision involved hitting a parked car.
is_ped	Extracted from the original column named as PEDCOUNT. Whether or not a pedestrian was involved in the accident.
is_bike	Extracted from the original column named as PEDCYLCOUNT. Whether or not a bicycle was involved in the accident.
month	Extracted from the original column named as INCDATE. The month of the collision.
day_name	Extracted from the original column named as INCDATE. The day name of the collision.
hour	Extracted from the original column named as INCDTTM. The hour of the collision.
district	Districts determined as a result of the clustering of locations

3. Methodology

3.1 Exploratory Data Analysis

We will look at the distributions of variables and the relations of them with the target variable.

3.1.1 Target Variable

Type of Collisions (Dependent variable) has two unique values named *Property Damage* and *Injury*. There are many collisions that result in property damage, but the number of *injury* collisions is three times less. So we can assume the dataset is imbalanced.

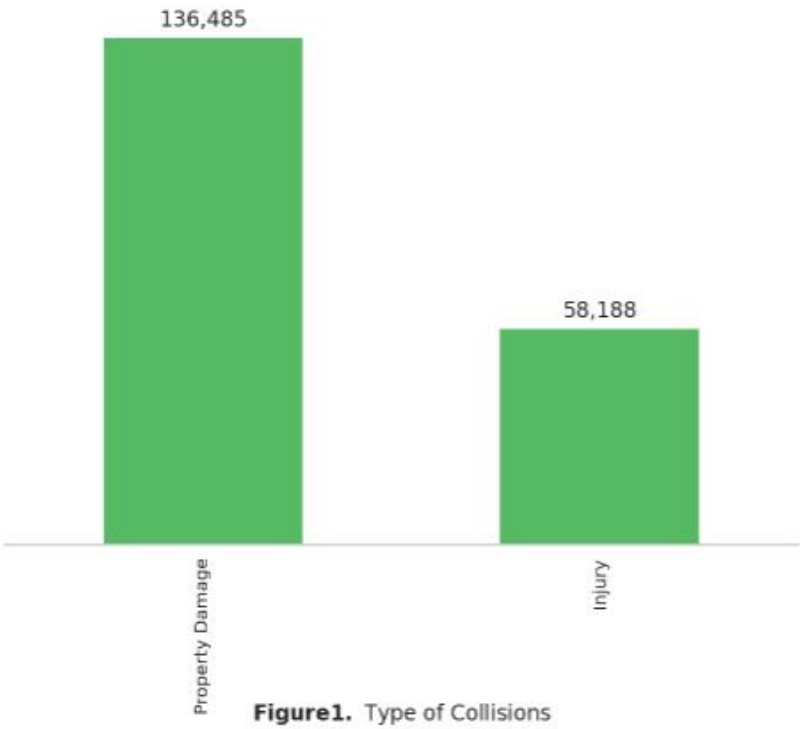


Figure1. Type of Collisions

3.1.2 Weather Conditions

Our hypothesis is that poor weather conditions correlate to car accident severity. But, when we look at the data, it is surprising to see that the most collisions occur in broad daylight. Most accidents occur in the Summer months such as June, July, and August. We can see that there is less rainfall in these months as well. Perhaps a new hypothesis is necessary such as "*people are more likely to go outside in clear weather*."

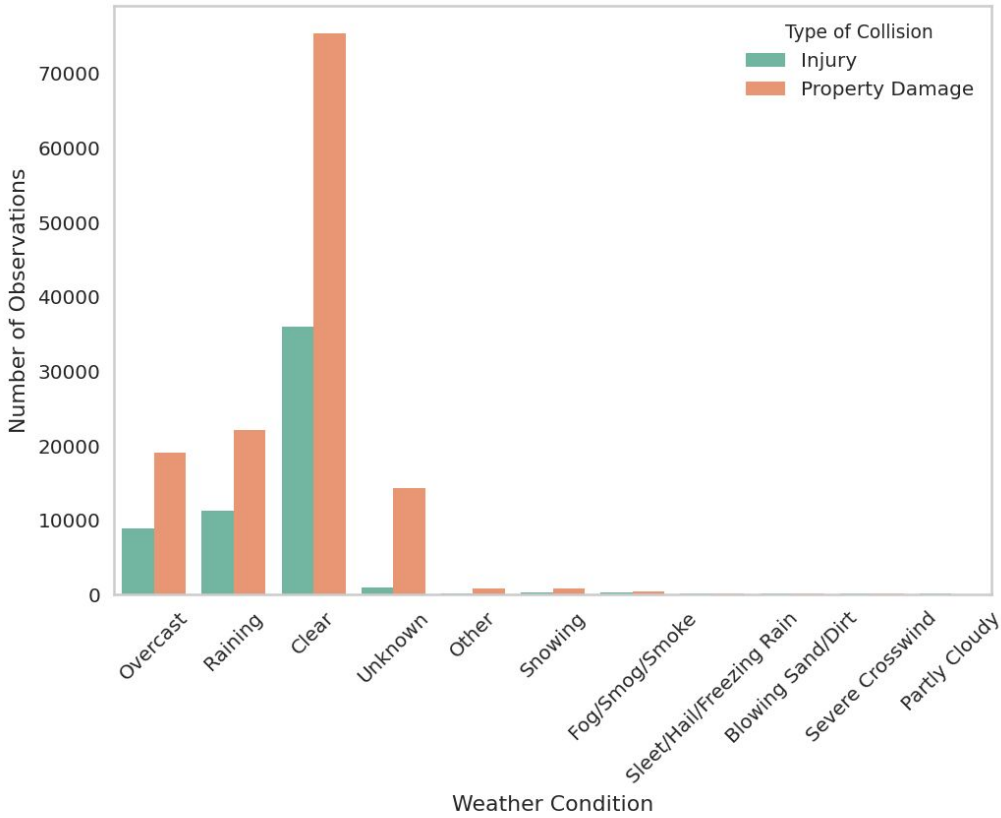


Figure2. Type of Weather Conditions

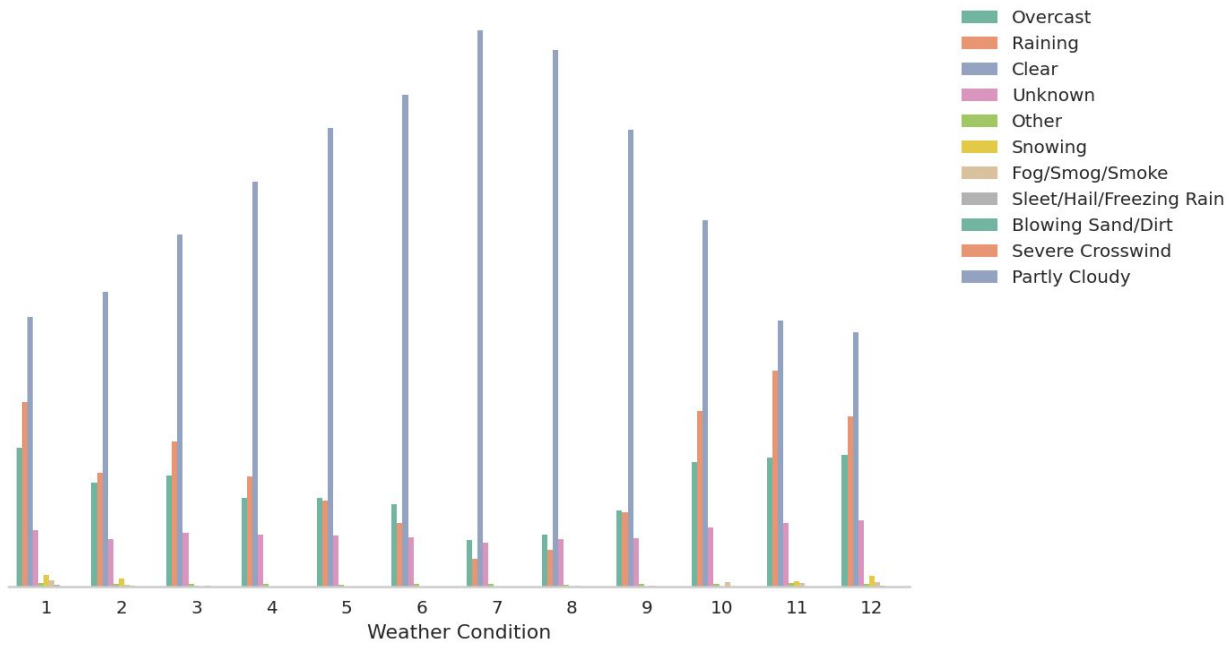


Figure3. Weather Conditions by Month

3.1.3 Road Conditions

For the most part, roads are dry during most accidents. But you can see that rainy days have lots of data as well.

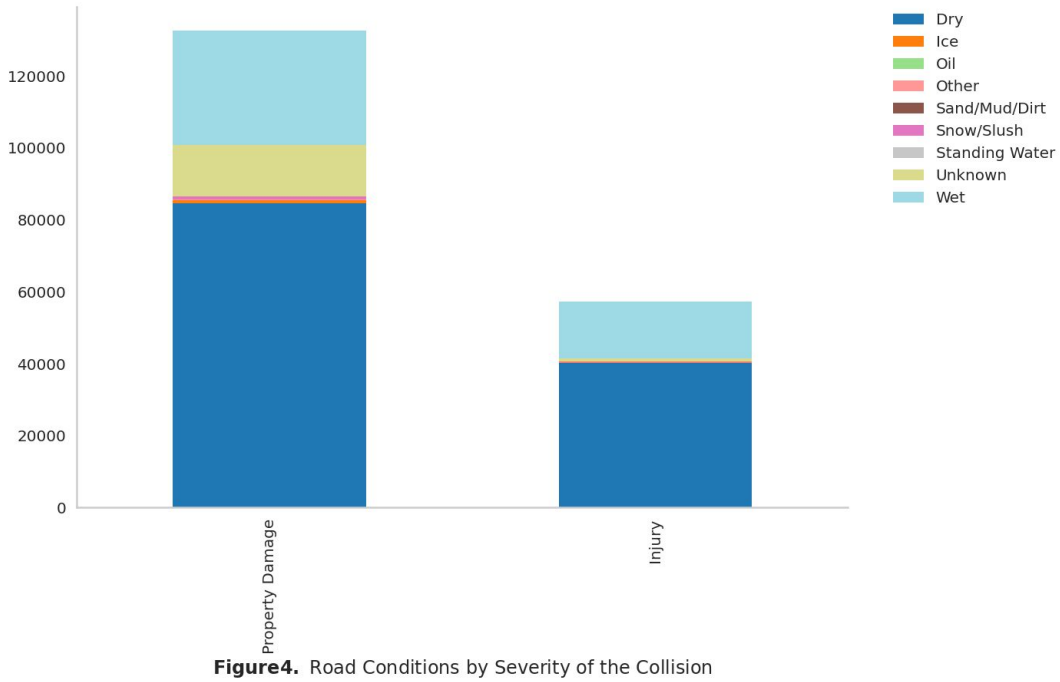
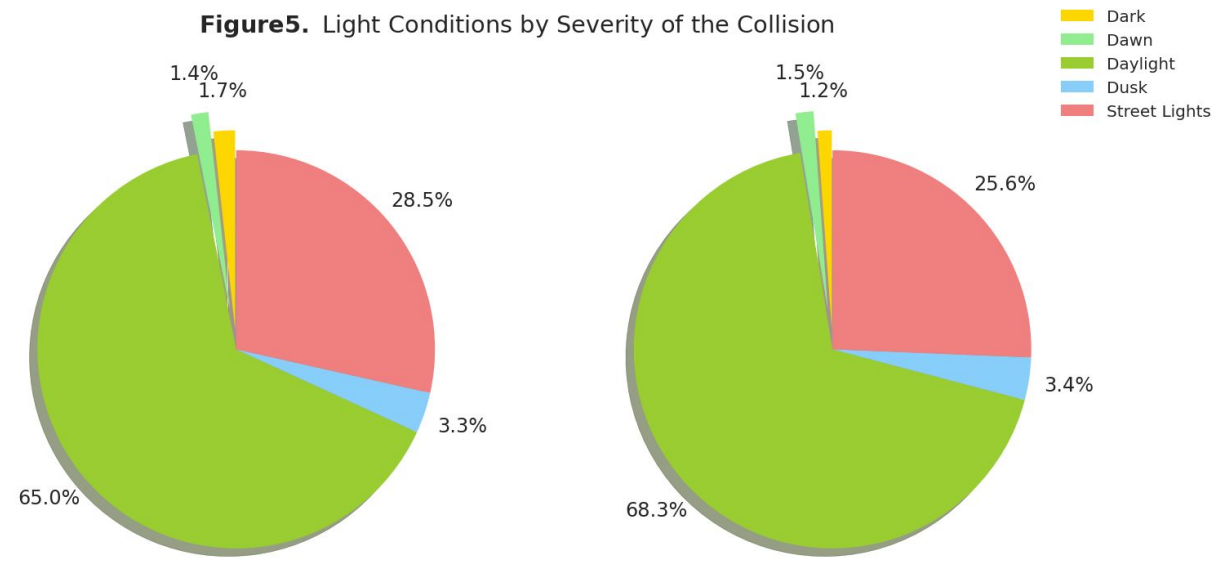


Figure4. Road Conditions by Severity of the Collision

3.1.4 Light Conditions

In Figure 5, we can see that most accidents happen in daylight and a significant amount under street lights. It is important to note that Seattle has great infrastructure because the least amount of accidents occur in the dark.

Figure5. Light Conditions by Severity of the Collision



Light Conditions for Property Damage Collisions

Light Conditions for Injury Collisions

3.1.5 Address Type ~ Collison Type

We can see that angle collisions occur most often at intersections. Also, parked car accidents are the most occurring accident type.

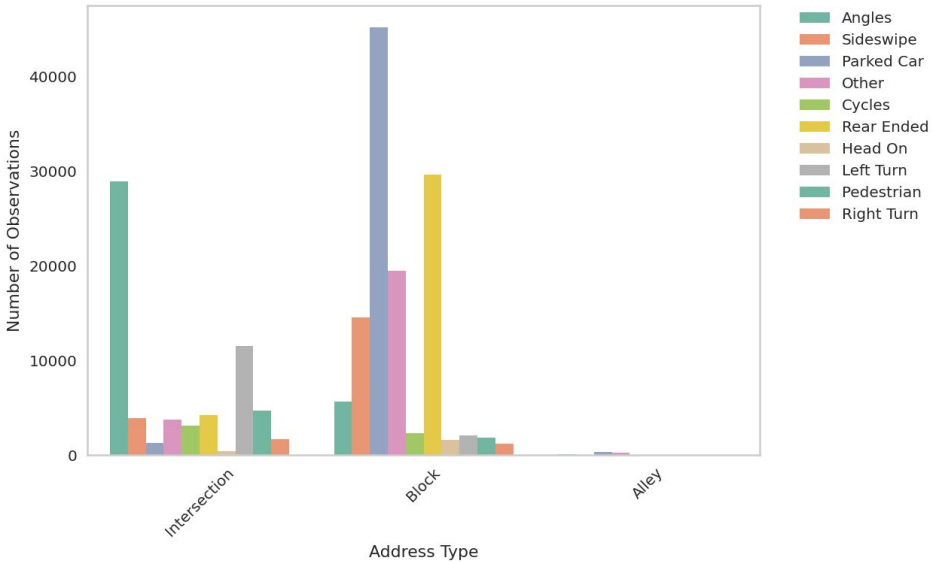
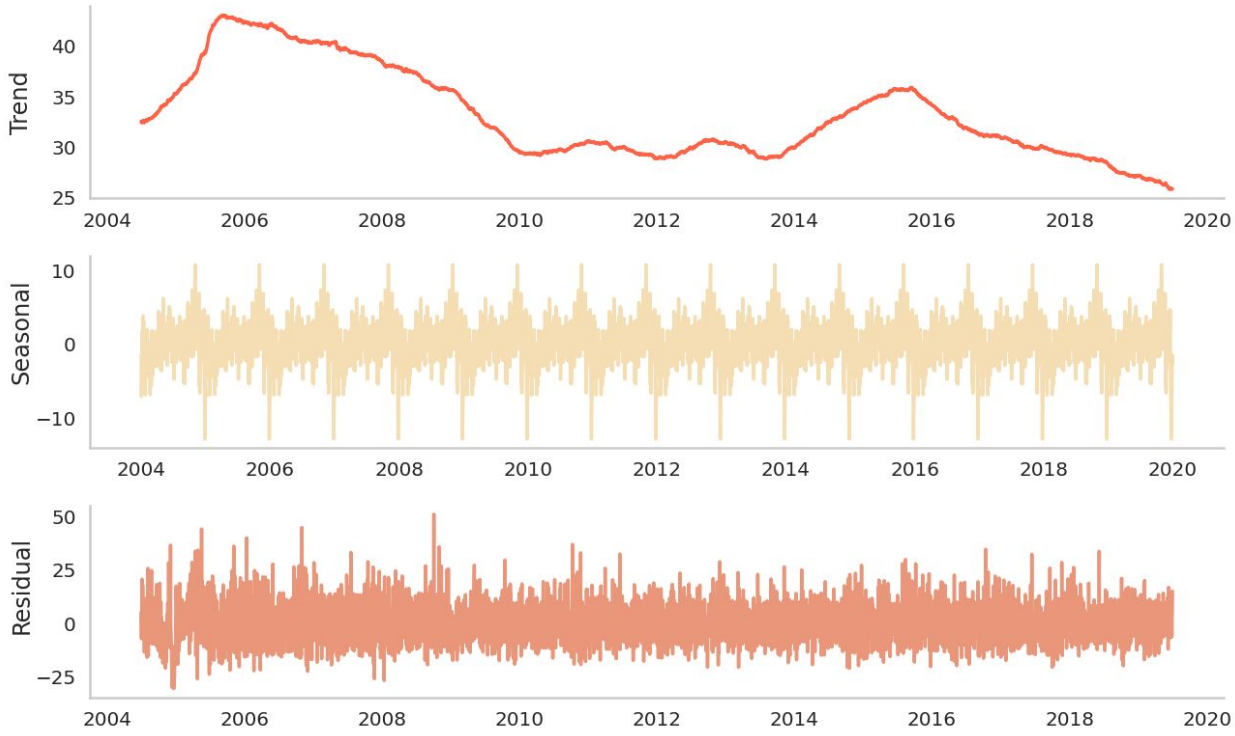


Figure6. Collision Types by Address Type

3.1.6 Seasonality

There is obviously seasonality in the data as we can see from the below graph. It is also important to consider that the reason why 5pm is the time where most accidents occur in a day is because it is rush hour.

Figure8. Seasonal Decompose



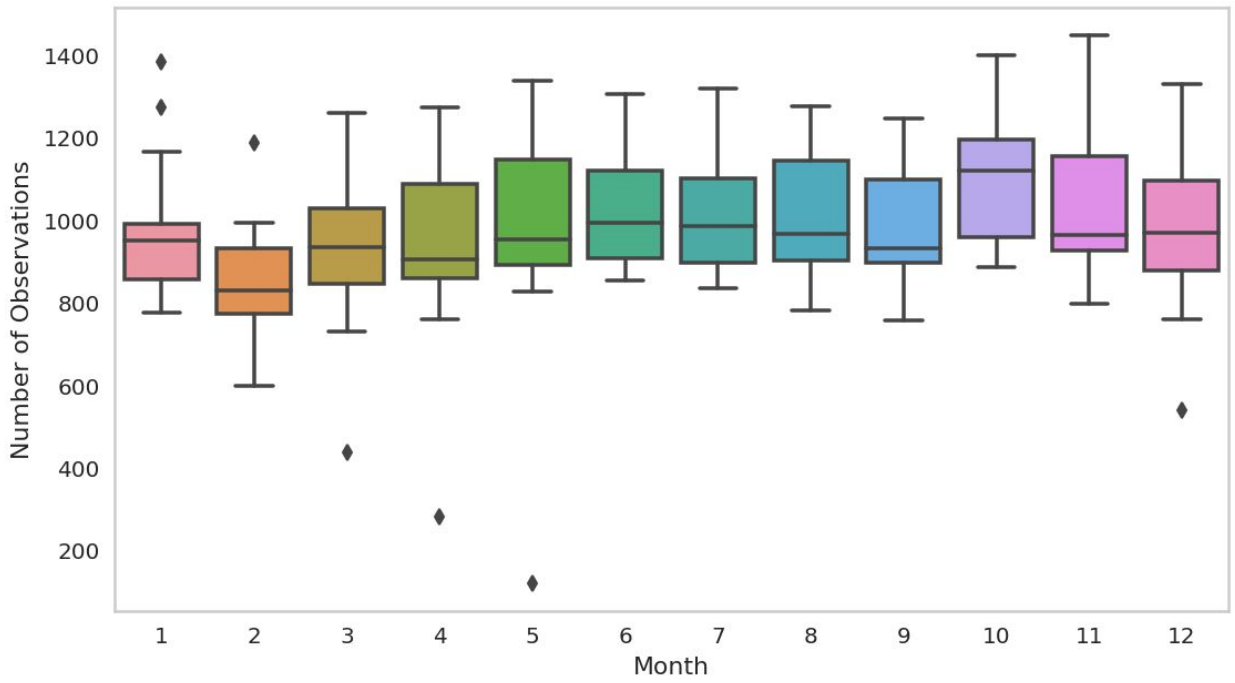


Figure9. Number of Collisions by Month

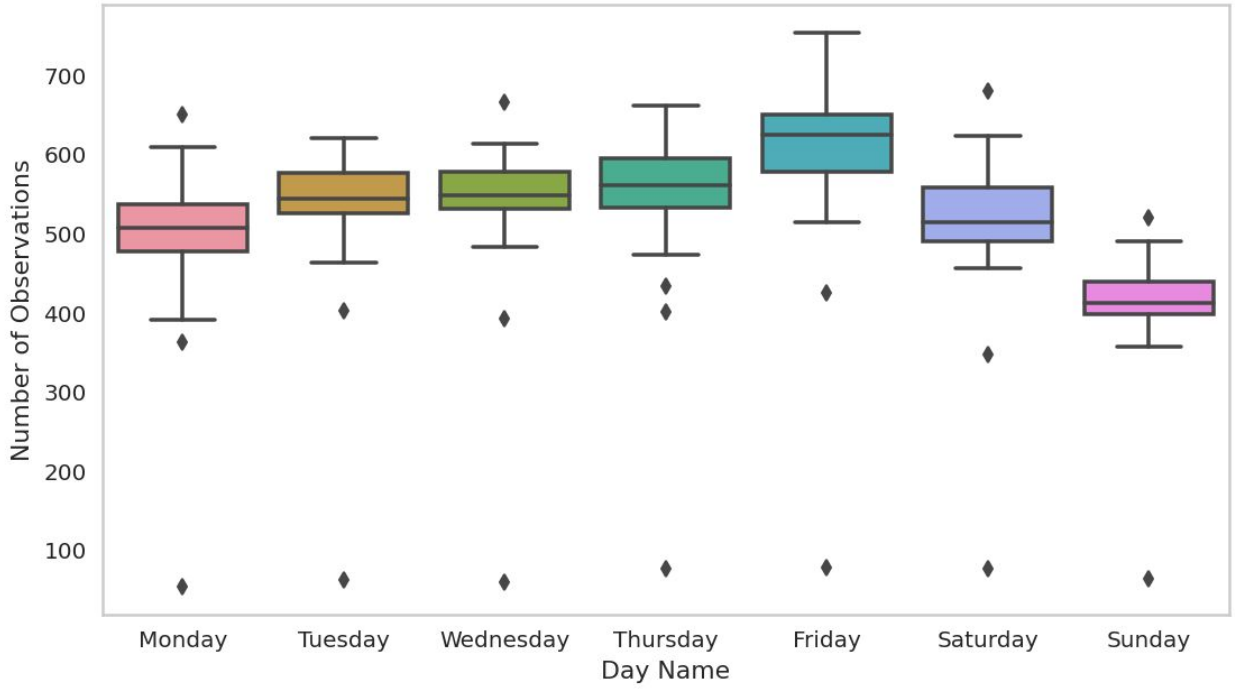


Figure10. Number of Collisions by Day Name

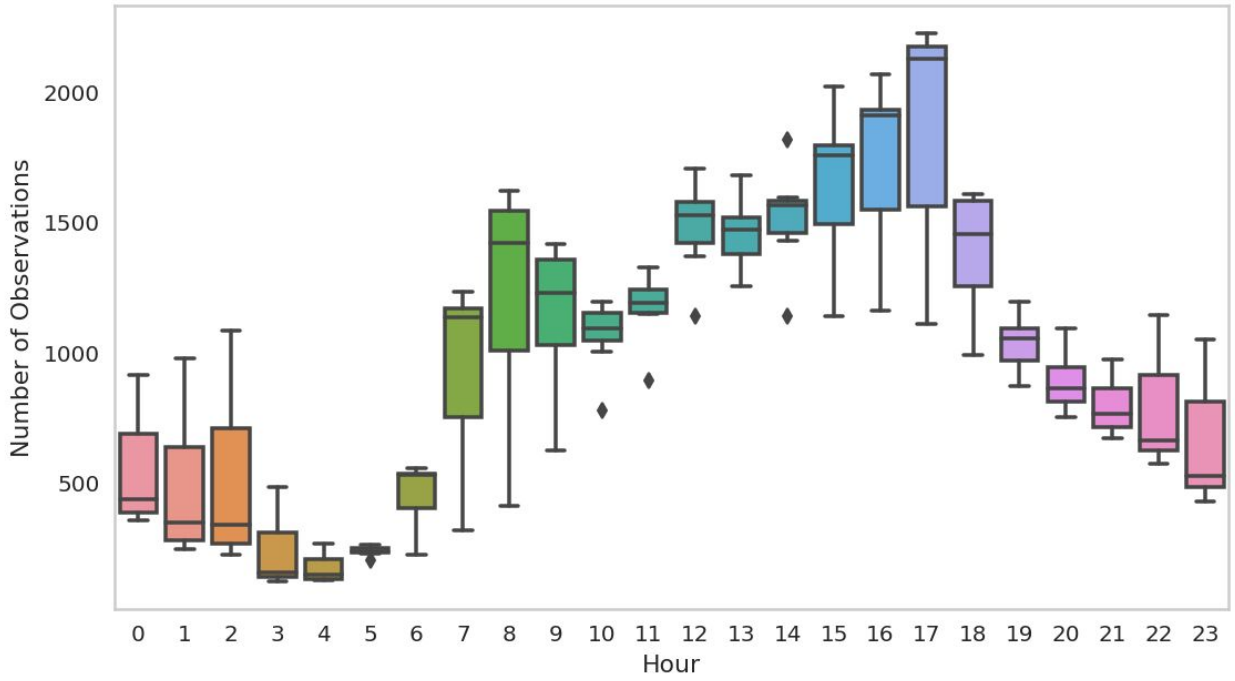


Figure11. Number of Collisions by Hour

3.2 Building a Model

Because the independent variable has two unique values, we can consider it as binary classification. This can be thought of as “Will an accident result in injury or not?” or “Will an accident result in property damage or not?” We will take a look at Binary Classification models such as Logistic Regression, Gradient Boosting Modifier, and Random Forests Classifier.

3.2.1 Logistic Regression

	Precision	Recall	F1 Score
Property Damage	.74	.98	.84

Injury	.80	.21	.34
Accuracy			.75
Macro Avg	.77	.59	.59
Weighted Avg	.76	.79	.69

Table 1. Logistic Regression results with class with class weight param

	Precision	**Recall**	**F1 Score**
Property Damage	.87	.59	.71
Injury	.46	.80	.58
Accuracy			.66
Macro Avg	.67	.70	.65
Weighted Avg	.75	.66	.67

Tabel 2. Logistic regression result with undersampling

	Precision	**Recall**	**F1 Score**
Property Damage	.56	.63	.72
Injury	.47	.76	.58

Accuracy			.67
Macro Avg	.66	.69	.65
Weighted Avg	.74	.67	.68

Table 3. Logistic regression results with undersampling and oversampling

3.2.2 Random Forests

Random Forests includes the class weight parameter like Logistic Regression. Additionally, the Random Forests classifier contains many different parameters which can be tuned. Here are the performance metrics of the Random Forests:

	Precision	**Recall**	**F1 Score**
Property Damage	.74	.98	.84
Injury	.80	.21	.34
Accuracy			.75
Macro Avg	.77	.59	.59
Weighted Avg	.76	.79	.69

Table 4. Random Forests results with class weight param

	Precision	Recall	F1 Score
Property Damage	.78	.80	.79
Injury	.50	.46	.48
Accuracy			.70
Macro Avg	.64	.63	.63
Weighted Avg	.69	.70	.70

Table 5. Random Forest results with undersampling (Tuned Model)

	Precision	Recall	F1 Score
Property Damage	.79	.74	.76
Injury	.47	.55	.51
Accuracy			.68
Macro Avg	.63	.64	.64
Weighted Avg	.70	.68	.69

Table 6. Random Forest results with undersampling and oversampling

3.2.3 Gradient boosting

Here are the performance metrics of the Gradient boosting

	Precision	Recall	F1 Score
Property Damage	.74	.99	.85
Injury	.85	.20	.33
Accuracy			.75
Macro Avg	.80	.59	.59
Weighted Avg	.78	.75	.69

Table 7. Gradient boosting results with undersampling

	Precision	Recall	F1 Score
Property Damage	.87	.62	.72
Injury	.47	.78	.58
Accuracy			.67
Macro Avg	.67	.70	.65
Weighted Avg	.75	.67	.68

Table 8. Gradient boosting with undersampling and oversampling.

4. Results and Discussion

When the training step is started without any balancing process between classes in the data set, the recall value for injuries caused by accidents seems quite low. At this point, it would not be wrong to say that the models predict that most accidents with injuries will result in property damage. The fact that the accidents resulting in injury cannot be predicted well, it means that the features that will differentiate the severity of the accident cannot be determined. Therefore, the suggestions to be made to prevent serious accidents cannot touch the right points.

In order to better predict the accidents that resulted in injury, the data set has been balanced and tested with classification algorithms. It is a well known fact that ensemble trees work in harmony with undersampling methods. After the necessary processes for balancing have been completed and the model has been builded, the recall value has increased for injury-related accidents, but there is a trade-off at this point: For accidents resulting in injury, while the recall value increases, the precision value decreases.

If a better solution cannot be found in such projects, the point to be waived should be chosen well. The basis of this decision is determining the goal, thus understanding the business. If causes of serious accidents are sought, better ways to predict serious accidents must be found and their causes investigated.

5. Conclusion

The goal of this project was to predict car accident severity and to determine the factors that affect severity. When examining the feature importance after the training step, it seems possible that there are several remarkable features. Whether the accident was hitting a parked car or not is extremely significant. It is also good to note that a pedestrian or cyclist being involved in the accident is significant as well.

Referring back to the exploratory data, we can see that accidents involving parked cars resulted in property damage.

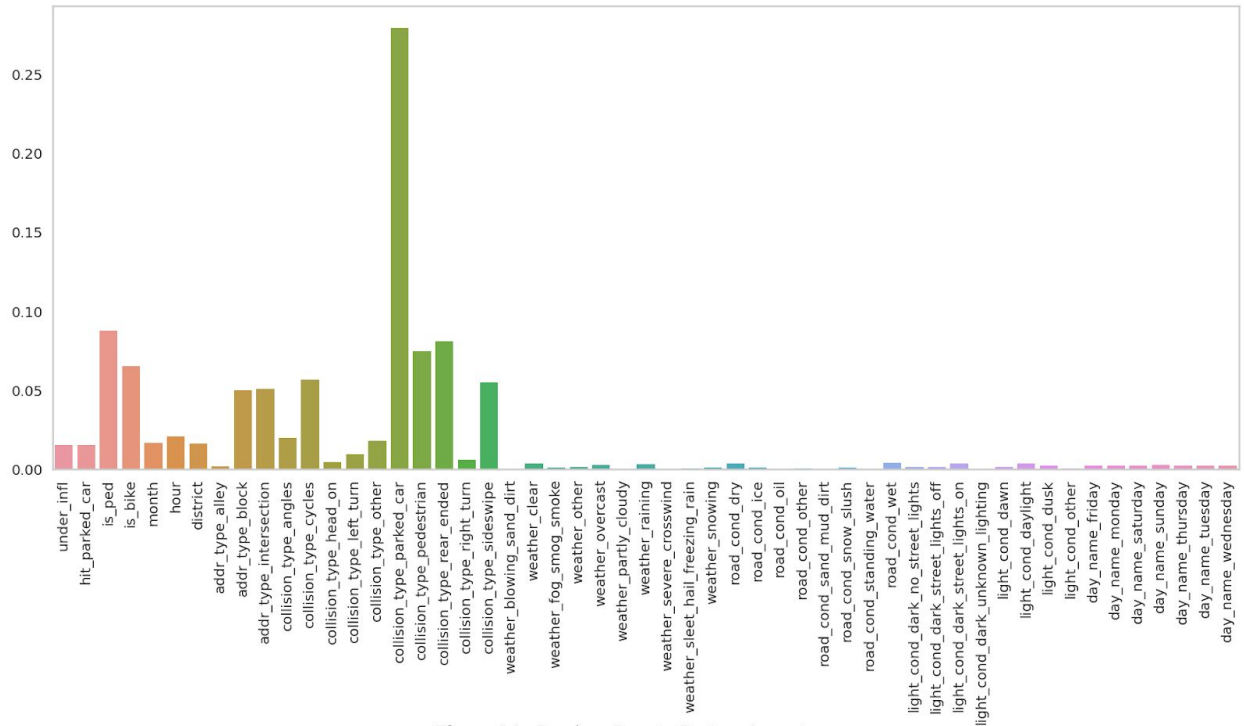


Figure14. Random Forests Feature Importances

Various points that can make suggestions based on all insights can be listed as follows,

- Drivers should be more careful where pedestrians and/or cyclists are concentrated. (is_ped and is_bike)
- Drivers should be more careful at intersections. (addr_type_intersection)
- For areas where parked cars are particularly concentrated, suggestions can be made to municipalities and/or relevant authorities to build parking lots or to increase security measures. (collision_type_parked_car)
- Extra measures can be taken to prevent driving under drugs and alcohol. (under_infl)
- Drivers should be more careful about sideswipes. (collision_type_sideswipe)
- Apart from the features that distinguish the two classes from each other, suggestions for situations where accidents occur frequently can be listed as follows:
- Drivers should be more careful on Fridays.
- At close hours to 5pm drivers should be more careful.
- Special precautions can be taken by the relevant authorities, especially since there are angles type collisions at intersections.