Predicting Car Accident Severity in Seattle

Ana Flávia Santos Souza

October 14, 2020

1. Introduction

In the US, there are around 6 million car accidents every year, and 90 people die because of that every day[1]. Great centers in the country, such as Los Angeles, New Orleans, and Baltimore, are usually the ones where most accidents occur[2]. Seattle, however, appears to be in the opposite direction. Even though it is one of the biggest cities in the country, when it comes to driving, Seattle is considered one of the safest places in the state of Washington[3]. That fact doesn't diminish the need to continually evaluate the city's accident rates and keep it track of how much damage each of these collisions causes, especially to guarantee that the rates will remain low.

Fortunately, data have become accessible as well as methods for visualization and analysis, making it easy to evaluate these rates. With the right data, it is possible to evaluate, identify, and predict how harmful an accident was or will be, using machine learning algorithms. This allows authorities to know how much services and efforts should be designated for this issue.

[1] https://www.driverknowledge.com/car-accident-statistics/

[2] https://quotewizard.com/news/posts/americas-most-car-accident-prone-cities

[3] https://quotewizard.com/news/posts/washington-most-distracted-driving-cities

1.1. Problem

The analysis of accident rates is important to understand how much damage is caused by a collision, and what can be done to decrease the harms. This project aims to use machine learning algorithms to predict car accident severity based on data of a period of fifteen years (2004-2019) from Seattle, WA.

1.2. Target Public

The key point of this project is to predict car accident severity based on the accident characteristics. This could be useful for the authorities responsible for assisting in these cases, such as police and medical services, to better understand how severe an accident is before getting into the location.

2. Data Collection

To examine car accident severity in Seattle, I used a CSV file provided by the Applied Data Science Capstone course on Coursera. The course also offered a metadata file with a description of the columns and their types. The original dataset was from the Seattle Police Department and Accident Traffic Records Department. It had 38 columns containing information about car accident occurrences from 2004 to 2020, including date and time of the accident, the accident's location, type of collision, number of people and vehicles involved, accident severity, and so on. With this data, I expected to get a perspective about the reasons that may lead to a car accident and predict the accident severity.

3. Methodology

3.1. Data Cleaning and Analysis

For cleaning and analyzing the data, I used the *Pandas* library. The dataset had about 194 thousand rows and 38 columns. The first thing I did was to check duplicated rows and delete them. There were none. Next, I selected what columns I would use for analysis and modeling, according to the metadata file. Thirteen columns total were selected: *severitycode, addrtype, collisiontype, personcount, pedcount, pedcylcount, vehcount, incdttm, weather, lightcond, roadcond, st_colcode,* and *underinfl.*

After, I fixed typos and errors on the columns entries. In the categorical variables, classes with low frequencies were summed together and converted into "Other" category. I also handled missing data. For the *st_colcode* column, rows where data was missing, were dropped. For the other variables, missing data was imputed according to the most frequent value. Finally, for the multiclass variables, dummies were created. The target variable *severitycode* was recoded, so class 1 became 0, and class 2 became 1. All the columns had their types corrected and validated.

Accidents that had occurred in 2020 were dropped since the year is still not over, and the data about it would be incomplete.

As the main point of interest, I observed what type of collision, weather, road condition, and light condition were most frequent during accidents. I also analyzed how these variables, plus the influence of drugs and alcohol, impact the accidents' severity. Lastly, I checked the number of people and vehicles involved in the accidents and accident frequency over time (accidents per year, month, and day of the week from 2004 to 2019).

3.2. Modeling

For feature selection, I removed predictors with near-zero variance. Fourteen of the initial variables and dummies were selected as predictors, and they were all normalized. The dataset was split into a training set (70%) for training the models, and a testing set (30%), for validating them. The target variable, *severitycode*, had a severe imbalance problem, so I treated that by under-sampling the majority class.

Since this is a classification problem, I used two classification models and compared their performances: Lasso Regression and Random Forests. The metrics chosen to measure performance were: F1-score, Area Under the Curve (AUC), recall, precision, and accuracy. All the models were tuned using cross-validation. Finally, I calculated feature importance and calibrated the probabilities of each model.

All the modeling process was done with the *sklearn* and *imblearn* libraries.

4. Results

4.1. Analysis

My analysis showed that in 15 years, Seattle's accident has been decreasing, except from 2004 to 2006 and 2013 to 2015, as shown in Figure 1.

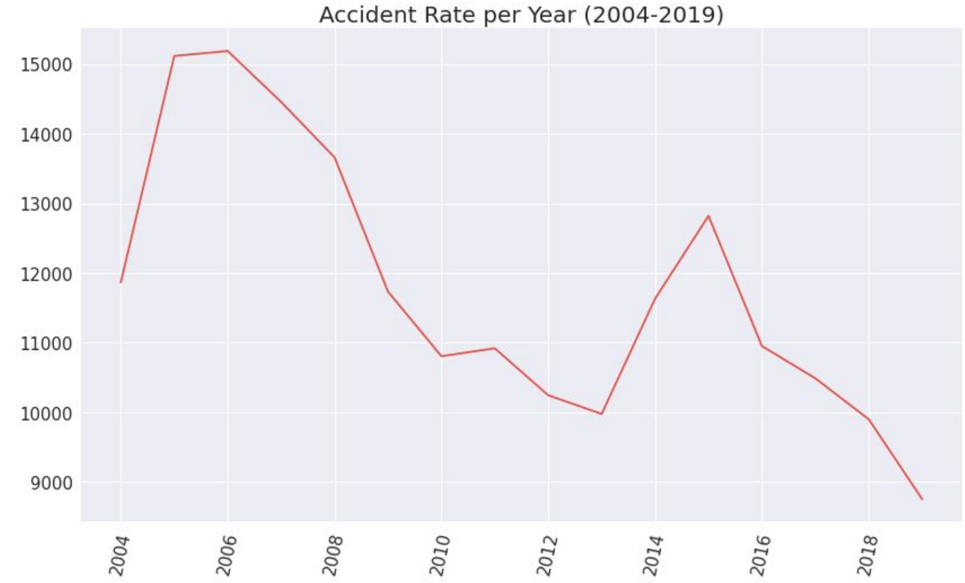

Figure 1: Accident rate per year, from 2004 to 2019.

I also found that most accidents have occurred on Fridays, especially around 5 pm. As for the month, October was the one with most occurrences, over 17,300 accidents from 2004 to 2019.

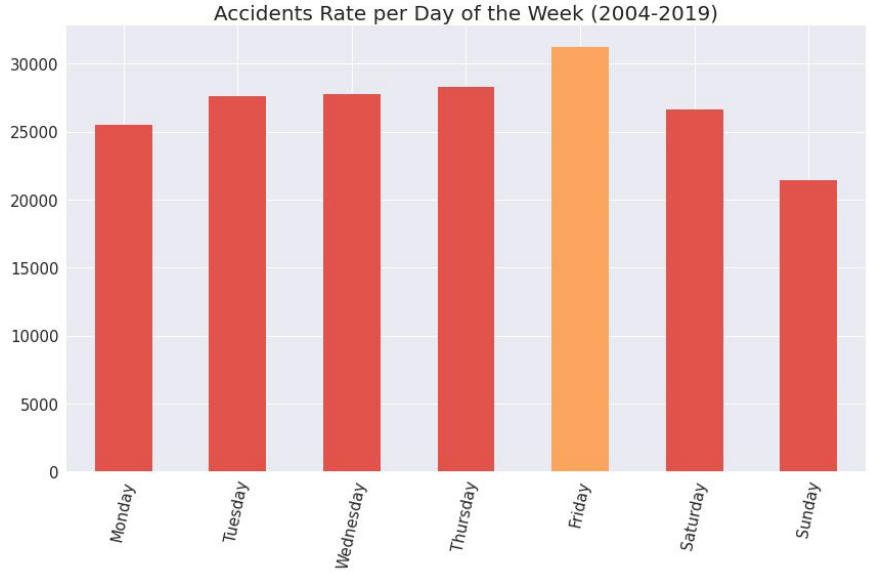

Figure 2: Accident rate per day of the week, from 2004 to 2019.

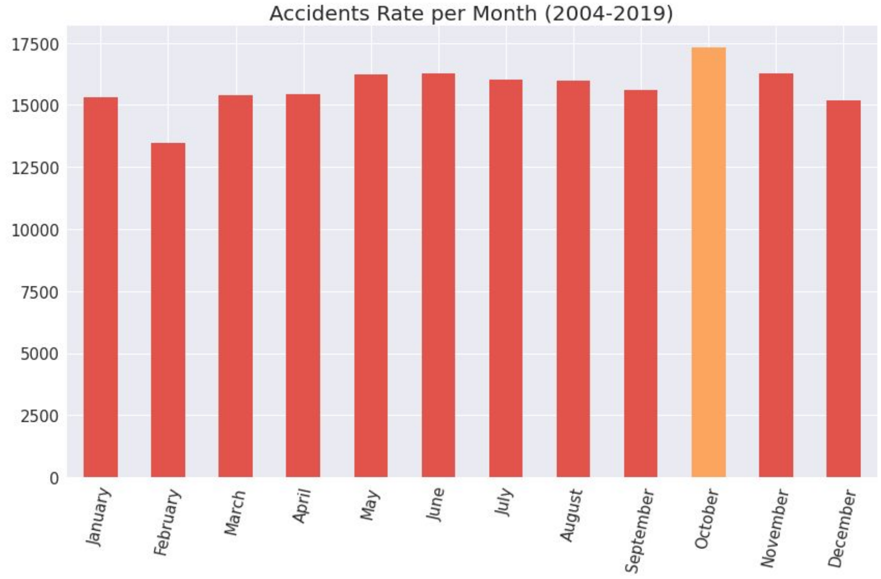

Figure 3: Accident rate per month, from 2004 to 2019.

The ten most common collision codes represented over 85% of all accident causes. For severity 0, the most common reason was one parked/one moving, while the most common cause for severity 1 was entering at an angle. In accidents involving bicycles, 87% of them presented severity 1. On average, the accidents involved two vehicles and two to three people, no matter how severe it was.

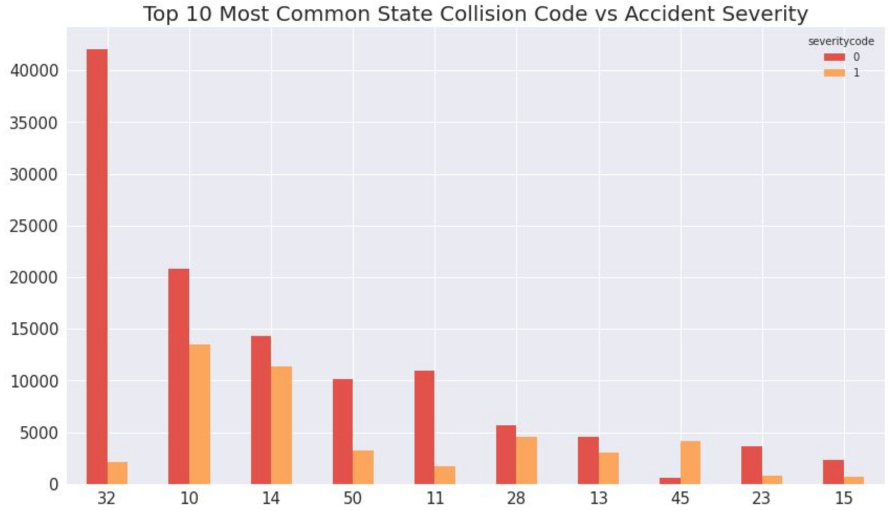

Figure 4: Impact of the most common state collision code on accident severity. The correlation between codes and causes is available in the metadata file.

In accidents where the driver was under the effect of drugs or alcohol, the severity 1 accidents represented around 39% of accidents. In contrast, when the driver was not under the influence of substances, they only represented 29.6% of the accidents.

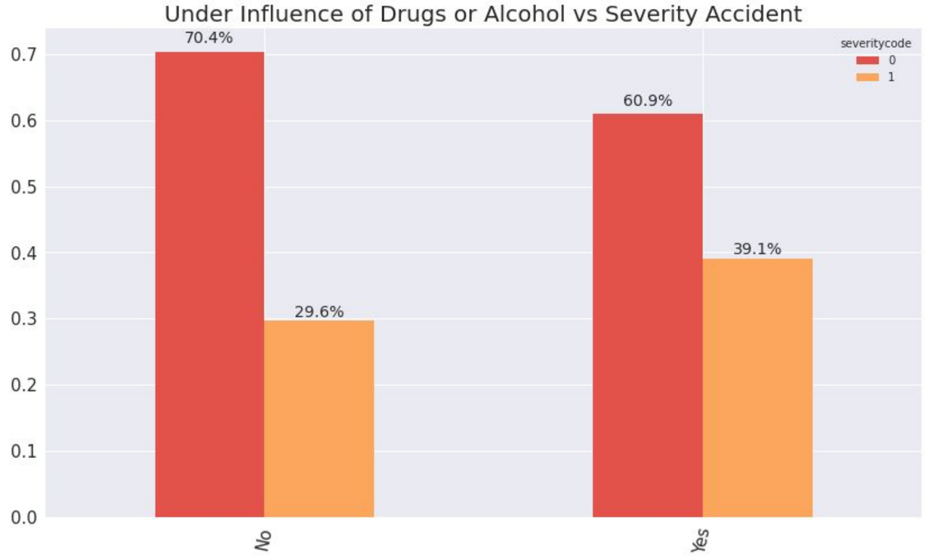

Figure 5: Impact of the use of drugs and alcohol on the accident severity, in percentage.

Finally, I found that most accidents occurred when the weather was clear (67%), the road was dry (74%), and during daylight (68%). These three conditions presented the biggest rates for both severity 0 accidents and severity 1 accidents.

4.2. Models

Before balancing, both models presented a similar performance. Lasso presented an accuracy of 0.72, while Random Forests' accuracy was of 0.75. The recall for class 1 was low. Both presented a value of 0.22.

After resampling, Lasso presented an F1-score of 0.62, recall for class 0 of 0.52, recall for class 1 of 0.81, precision for class 0 of 0.86, and precision for class 1 of 0.42. In Random Forests, these same metrics presented values of 0.67, 0.59, 0.81, 0.88 and 0.46, respectively. The AUCs for both models didn't change after resampling. Lasso presented an AUC of 0.73, and Random Forests presented a value of 0.79. Metrics from all models are displayed in the table below.

Models	Accuracy %	Rec0%	Rec1%	Prec0%	Prec1%	AUC	F1%
Lasso	0.72	0.94	0.23	0.74	0.61	0.73	0.68
Lasso balanced	0.60	0.52	0.81	0.86	0.42	0.73	0.62
RF	0.75	0.98	0.23	0.75	0.81	0.79	0.70
RF balanced	0.67	0.60	0.82	0.88	0.47	0.79	0.68

Table 1: Metrics of all 4 models for comparison: Lasso, Lasso balanced, Random Forest (RF) and Random Forest balanced.

As for feature importance, the variables that had the most impact on Lasso were *personcount*, *vehcount,* and *ct_Parked_Car* (collision on a parked car). For Random Forests, they were *ct_Parked_Car*, *st_colcode,* and *vehcount.* Variables related to road conditions, light conditions, and weather appeared to have a low impact on both models.

After calibration, the balanced Random Forests presented the best result according to the ideal model.

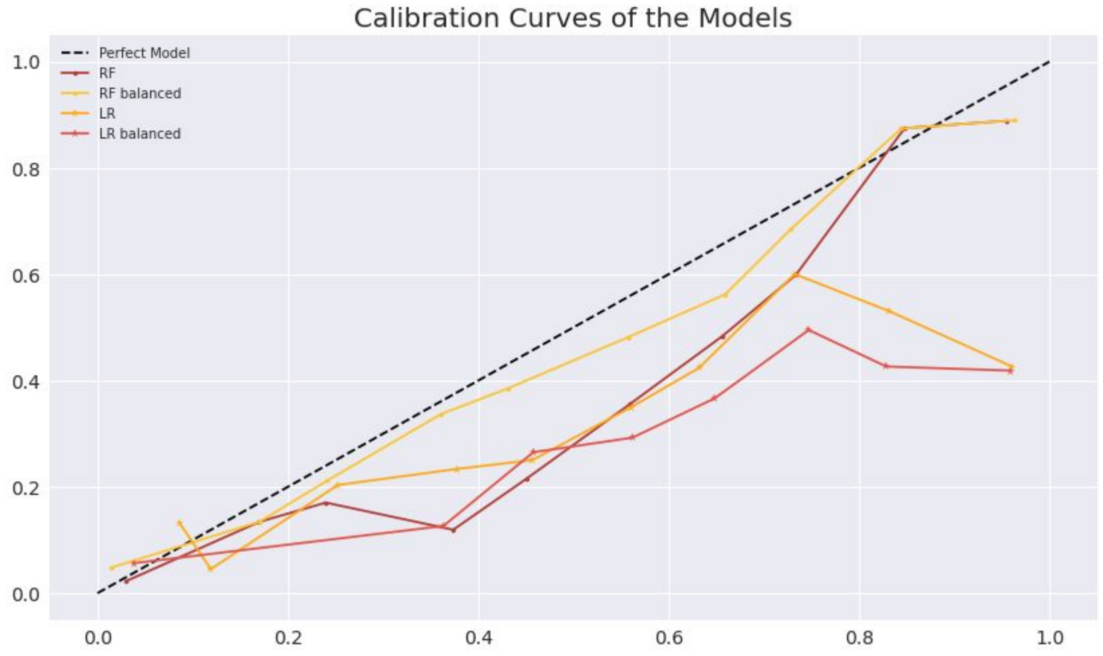

Figure 6: Calibration curve of the models. The black line represents a model with perfectly calibrated probabilities.

5. Discussion

The analysis suggested that Friday is when most accidents occur, which may have a connection to the fact that people get more tired and anxious to get home at the end of the week, and probably are less cautious with speed and traffic because of that. Sundays are when fewer accidents occur, which makes sense if we consider that people drive less on weekends.

The leading causes of accidents are related to the inattention of the driver. Since most of the accidents also occur during the daylight, on a dry road, and with clear weather, this suggests that when raining or snowing, people are more careful with their driving, what doesn't happen when the weather conditions seem harmless. The use of alcohol and drugs while driving also increases the chances of a collision being more severe.

The models initially presented a lousy performance when predicting class 1 severity because of imbalance. After resampling, Random Forest presented the best performance in all the selected metrics, which suggests this model is better for this accident prediction specifically. However, the model precision for class 1 was still lower than 0.5. Considering that the false positive for class 1 could mobilize an excess of police and medical services without its need, the model could use some improvement to classify better severity 1 accidents.

Some variables, like speeding and injuries-related, didn't appear on the dataset or had a significant proportion of missing values (over 40%). Yet, if available, they could improve models' prediction, since they are common indicators of how severe an accident was.

6. Conclusion

This project analyzed Seattle's accident rates from 2004 to 2019. Using data provided by the Seattle Police Department, it was possible to identify that the car accident rate has dropped over the years. Fridays, around 5 pm, are when accidents are more likely to occur, on average, and October is the month with most accidents.

When the weather conditions are clear, it's more likely to occur an accident. So people should be more careful when driving under these conditions. A driver under the influence of alcohol and drugs will probably cause a more serious accident, and bicyclists are more prone to get involved in a severity 1 accident.

With the use of machine learning models, it was possible to predict the severity of most accidents. This can help entities responsible for providing assistance in car collisions to allocate their forces better. One suggestion for further studies is to include more variables in the analysis and the models to improve class 1 severity prediction precision.