Predicting Car Accident Severity in Seattle

Ana Flávia Santos Souza

Introduction

Introduction

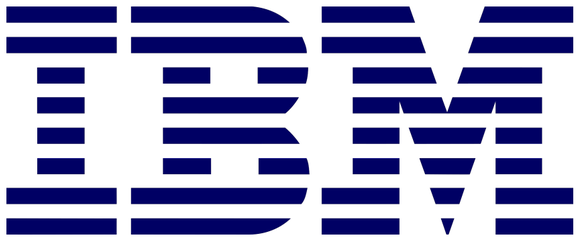
- In the US, there are around 6 million car accidents every year[1].
- Seattle, in contrast to other big cities in the country, is considered a safe place for driving[2].
- It is important to evaluate the accident rates to guarantee they remain low.
- If the authorities know beforehand how severe an accident was, they can better allocate assisting services.
- This project aims to predict car accident severity in Seattle, WA, using machine learning algorithms.

[1] https://www.driverknowledge.com/car-accident-statistics/
[2] https://quotewizard.com/news/posts/washington-most-distracted-driving-cities

Data collection

The data was collected from

Seattle Police Department



IBM

Methods

Data Cleaning

- I selected 13 attributes from the Seattle Police Department original data;
- Rows that had missing state collision code or events that happened after 2019, were excluded;
- Missing data was imputed with the most frequent value;
- The dataset was validated: columns had their errors fixed when needed.

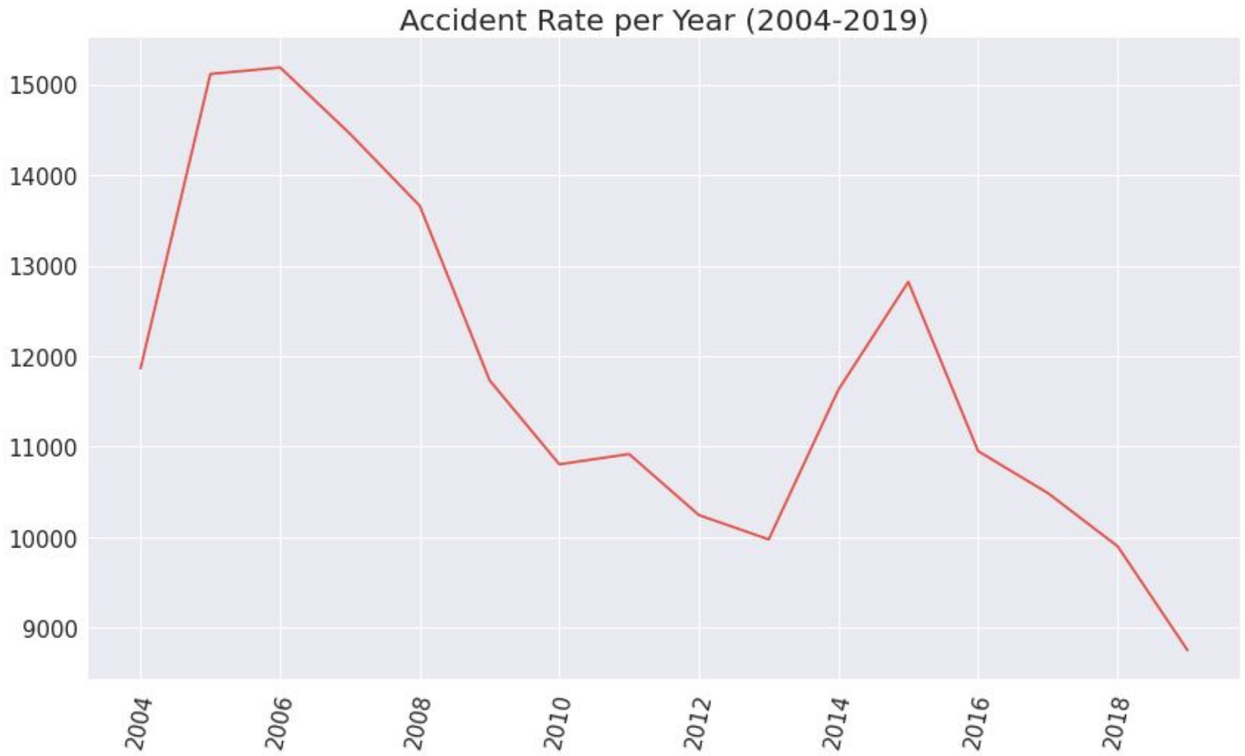
Data Analysis

- Analysis was done using the *pandas* library;
- Points of interest:
  - accident tendencies over the years, months and days of the week;
  - how much the type of collision, weather, road condition, light condition and the influence of drugs and alcohol affect the accident severity.

Modeling

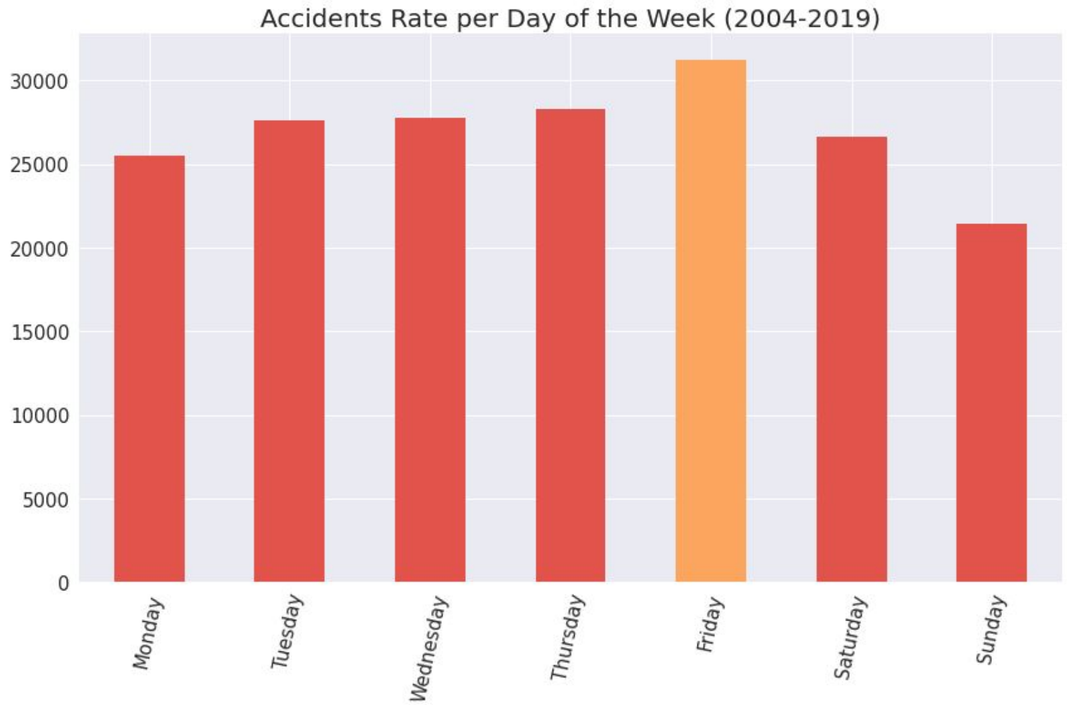
- Feature selection and modeling were done using *sklearn* and *imblearn* libraries;
- The dataset was split into training (70%) and testing (30%) sets;
- Imbalance problem on target variable was handled with under sampling;
- Models: Lasso regression and Random Forests;
- Metrics to evaluate models performances: F1-score, Area Under the Curve (AUC), recall, precision and accuracy;
- All the models had their probabilities calibrated and feature importance calculated.

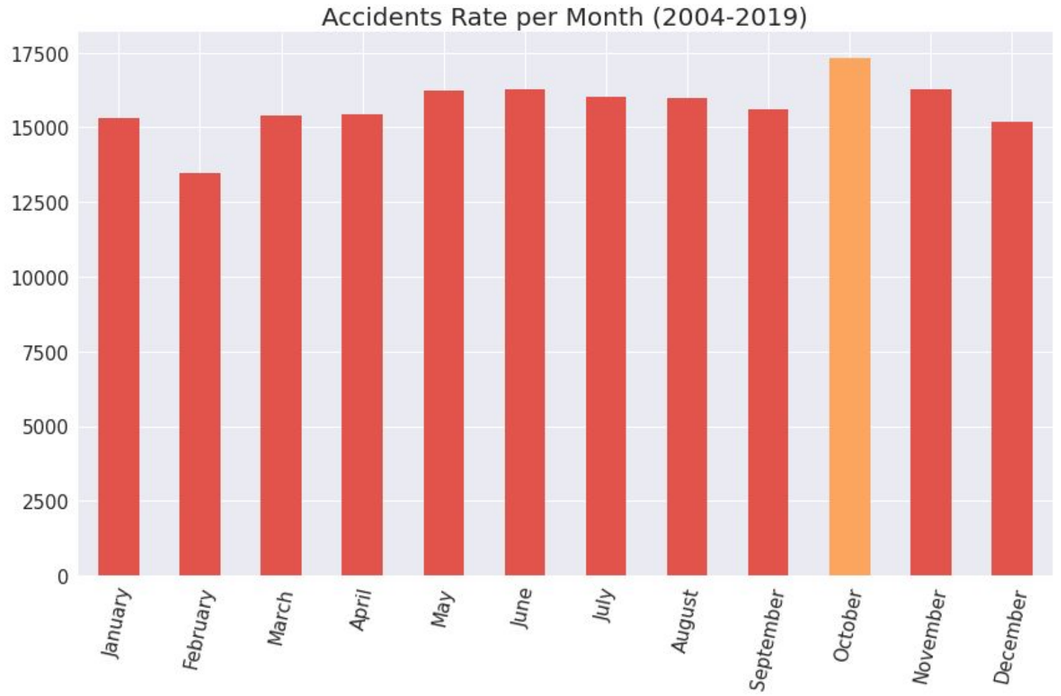
Results

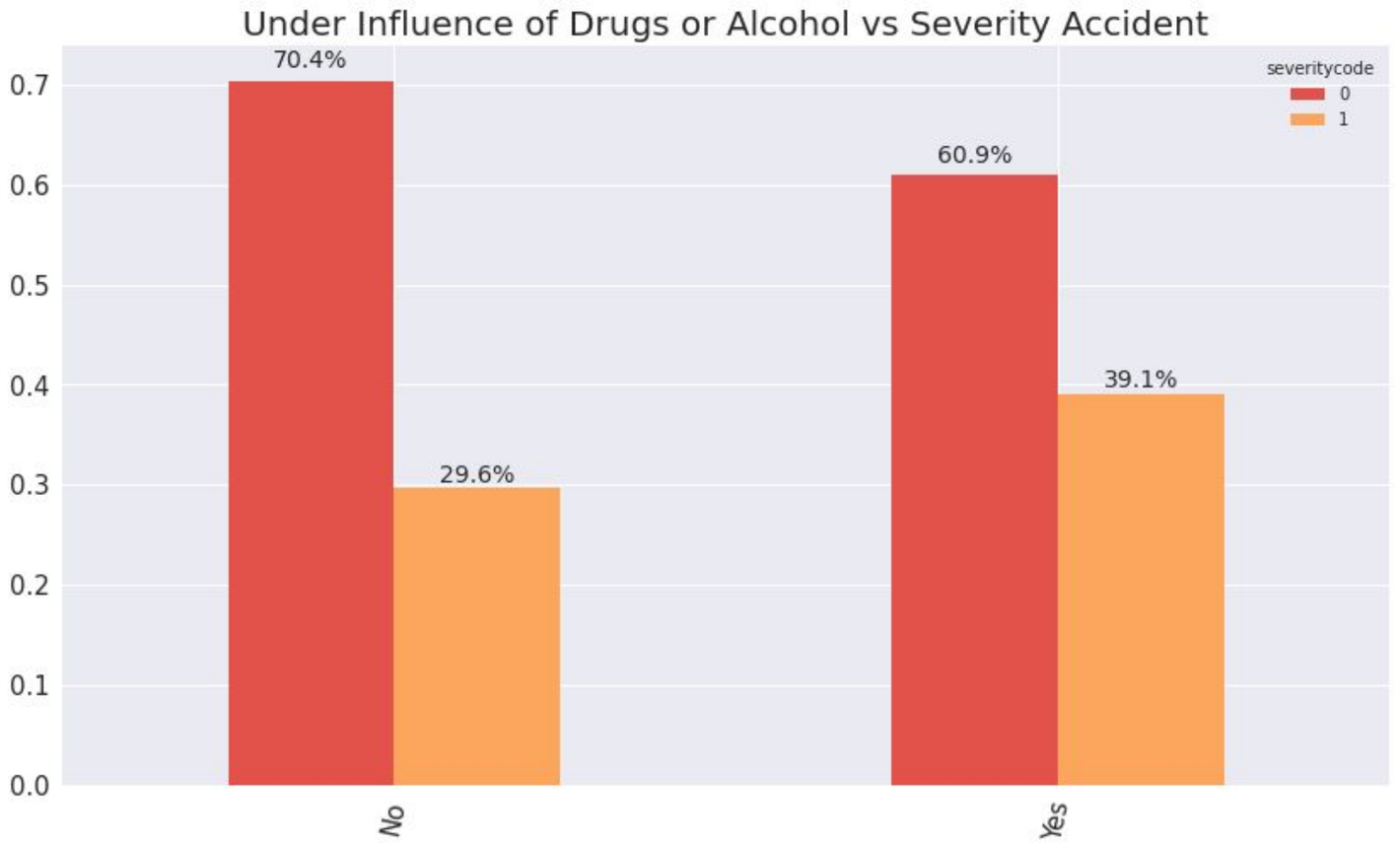


The accidents rate has been dropping for the past years, , except from 2004 to 2006 and 2013 to 2015. In 2019, there were less than 9,000 events.

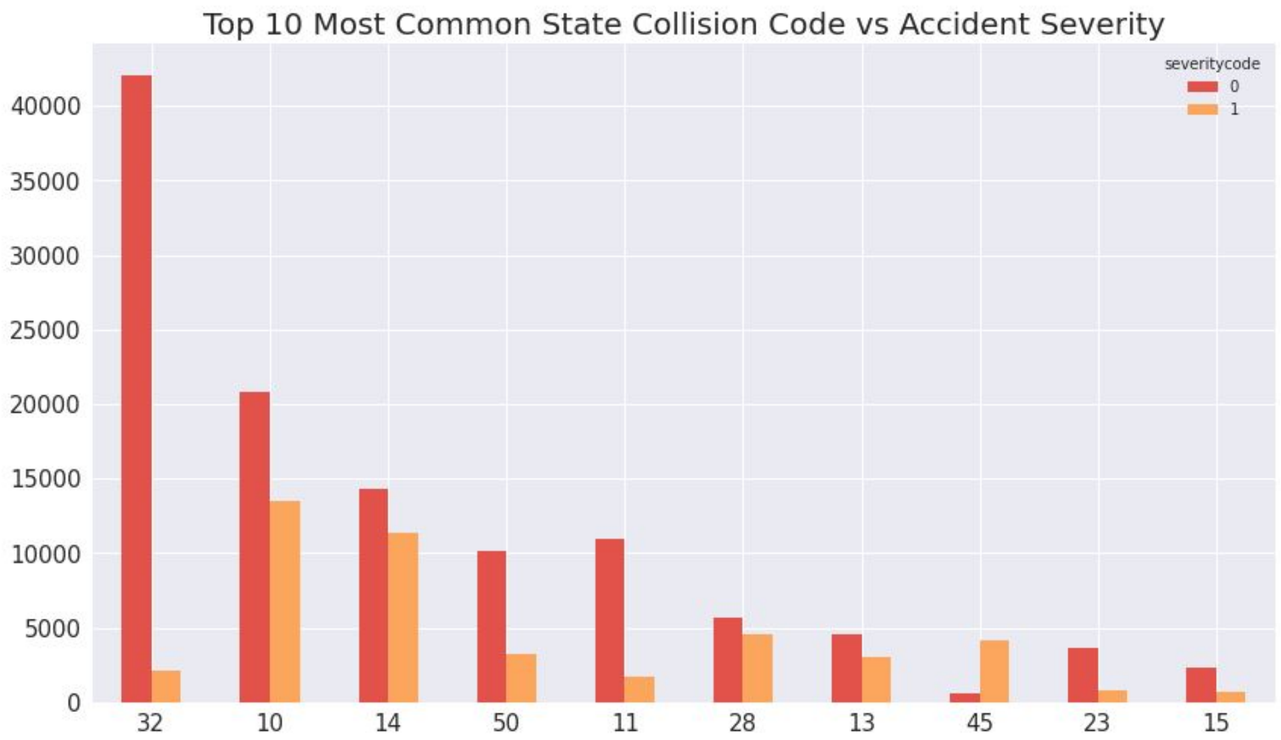
Accidents are more likely to occur on Fridays. In terms of months, October has been the month with the highest accident record from 2004 to 2019, with over 17,300 occurrences.





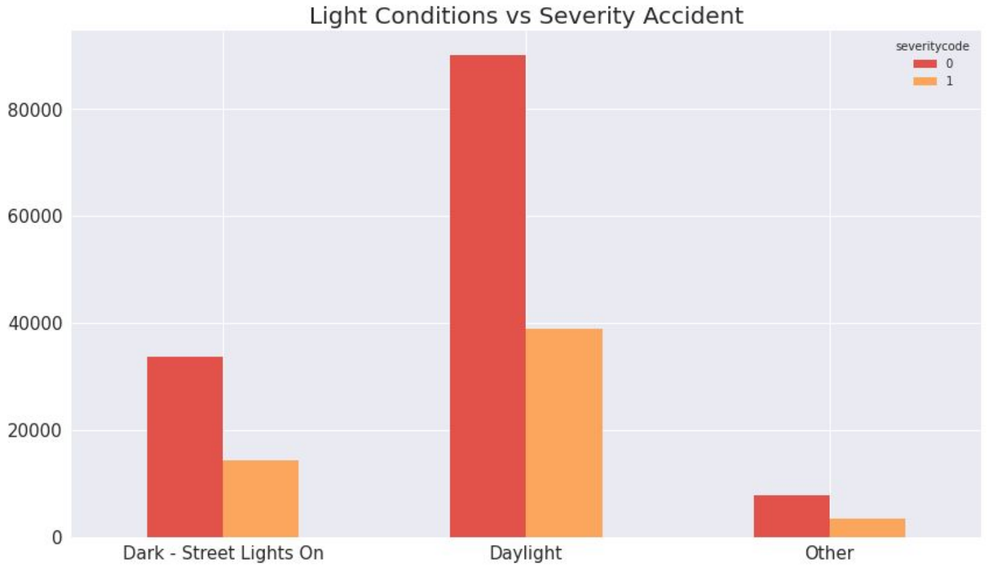


Impact of the use of drugs and alcohol on the accident severity, in percentage. When the driver was under influence of substances, the rate of severity 1 accidents was of 39.1%.

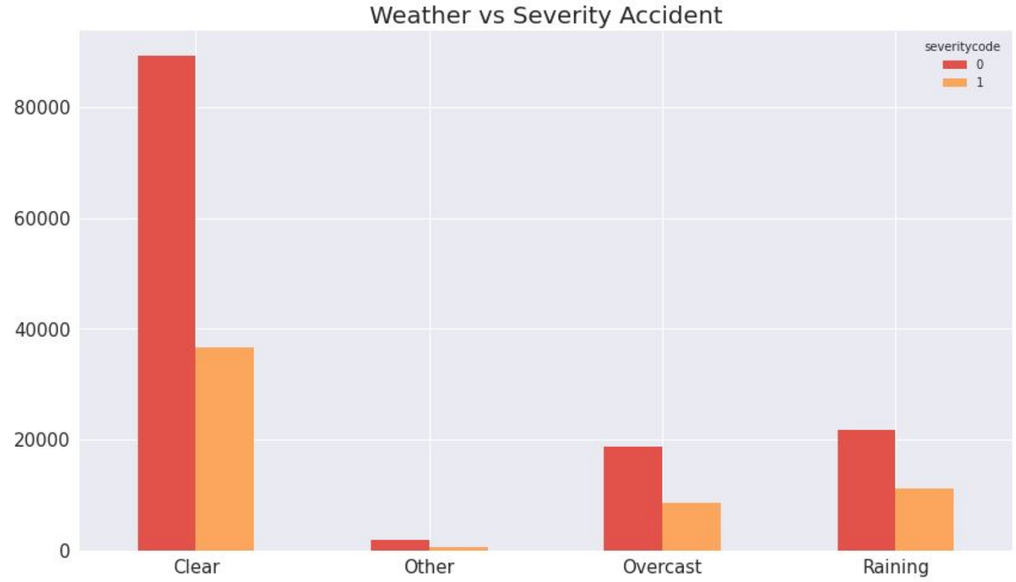


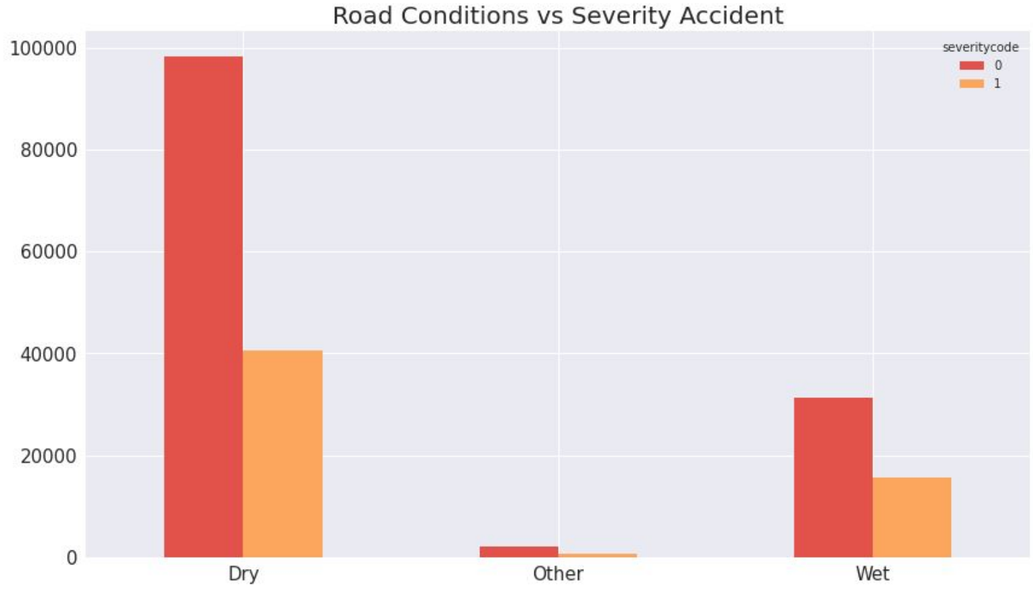
The 10 most common causes for an accident are responsible for 85% of all accidents. In accidents where a bicyclist was involved (45)[1], 87% of them had severity 1.

[1] The correlation between code and cause is available in the metadata file.

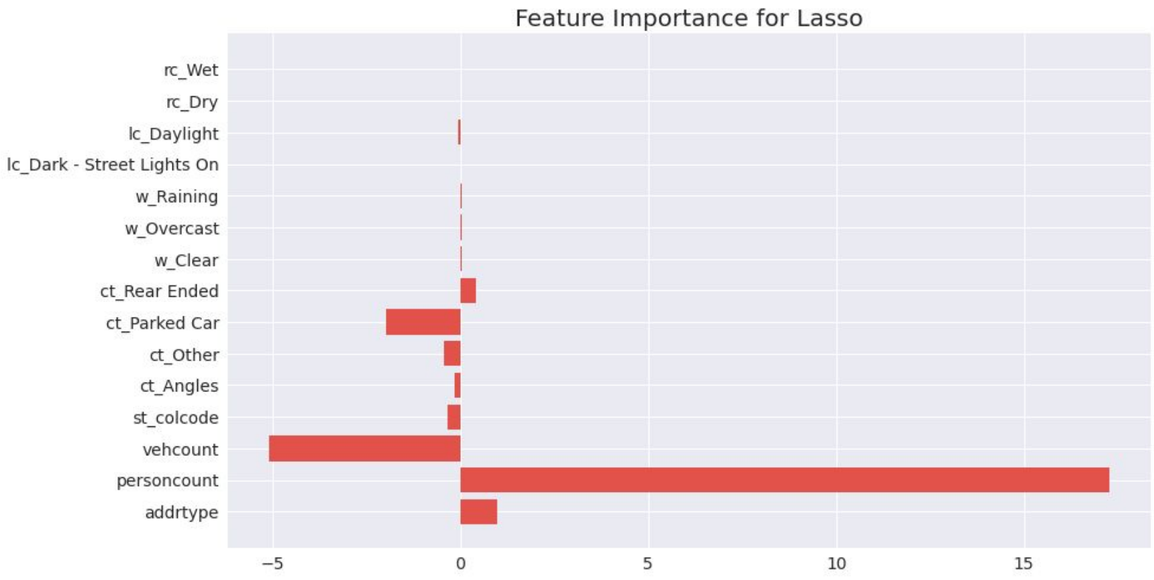


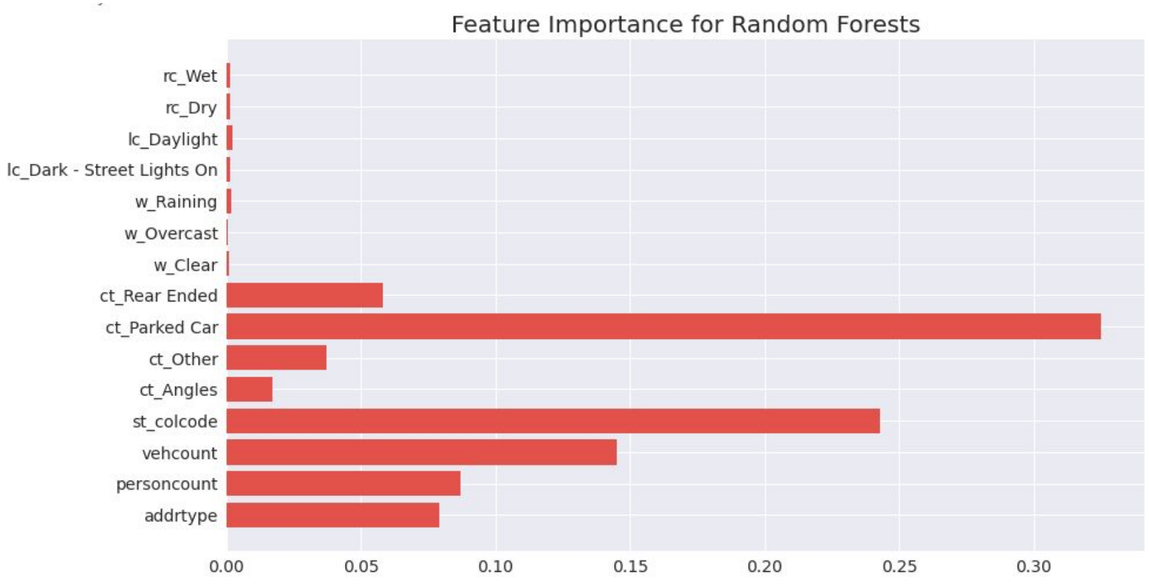
Most accidents occurred when the weather was clear (67%), the road was dry (74%), and during daylight (68%).





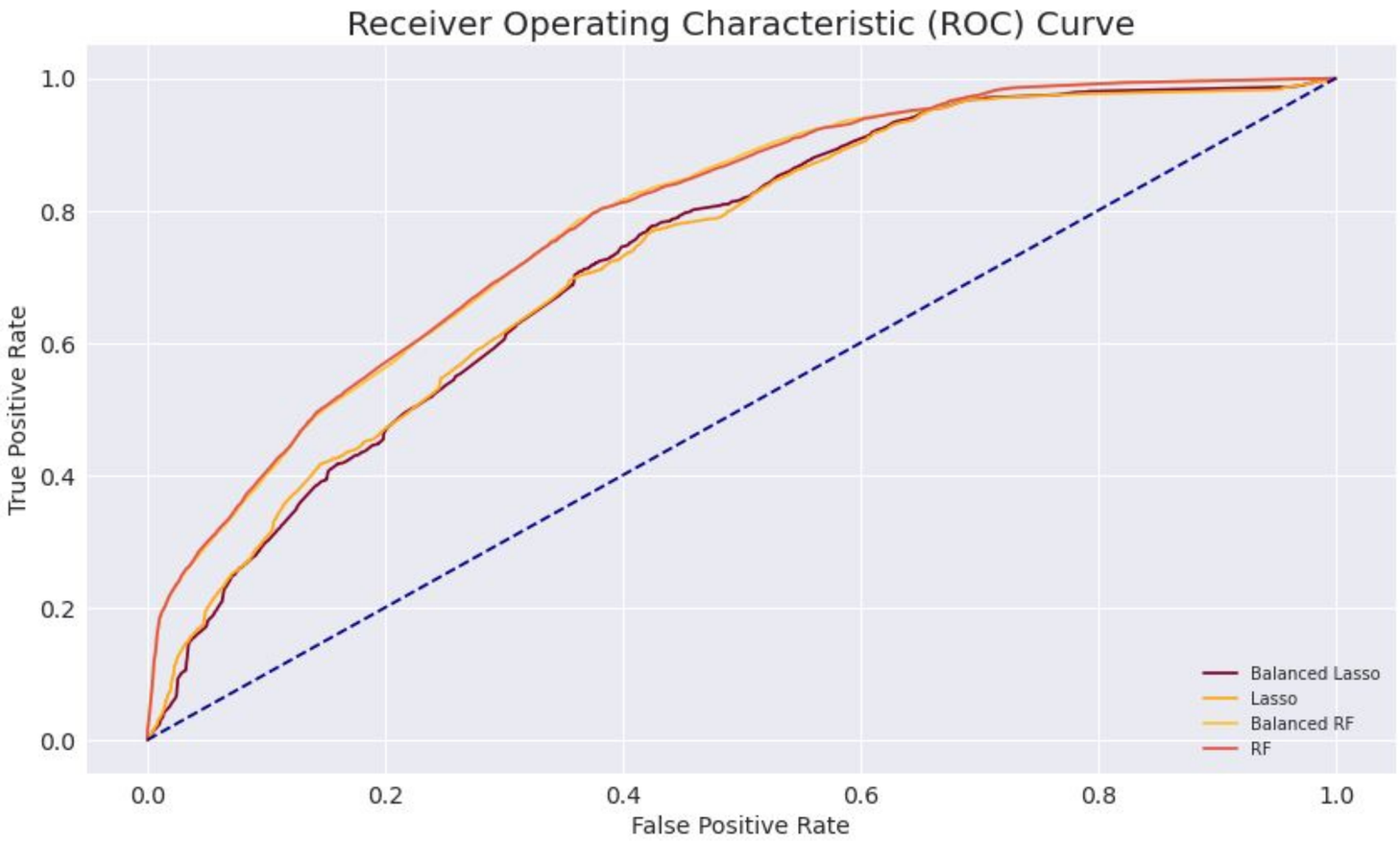
The variables that had the most impact on Lasso were *personcount*, *vehcount*, and *ct_Parked_Car* (collision on a parked car). For Random Forests, they were *ct_Parked_Car*, *st_colcode*, and *vehcount*.



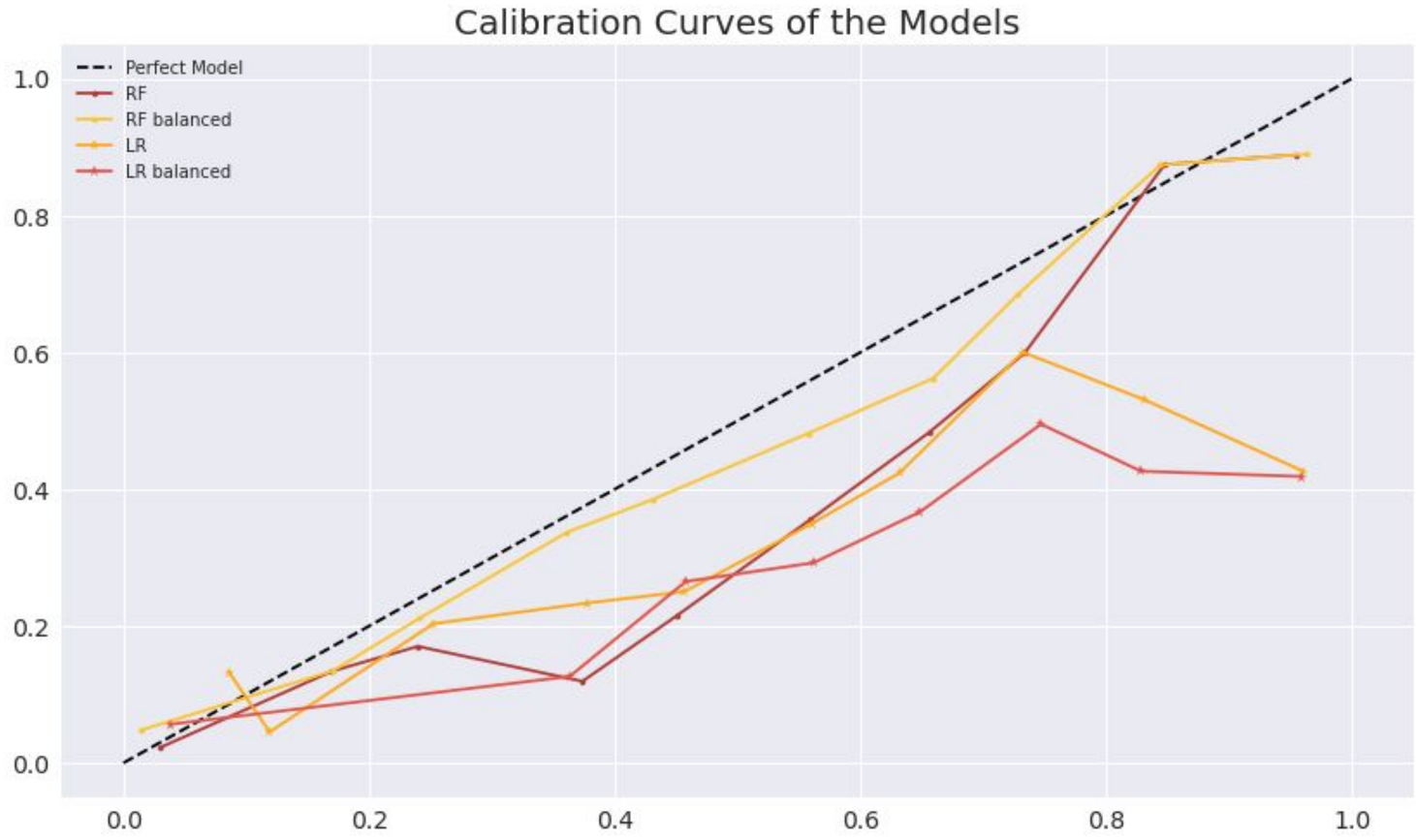


Models	Accuracy %	Rec0%	Rec1%	Prec0%	Prec1%	AUC	F1%
Lasso	0.72	0.94	0.23	0.74	0.61	0.73	0.68
Lasso balanced	0.60	0.52	0.81	0.86	0.42	0.73	0.62
RF	0.75	0.98	0.23	0.75	0.81	0.79	0.70
RF balanced	0.67	0.60	0.82	0.88	0.47	0.79	0.68

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



ROC curve of the models. The dotted line represents a model without prediction ability. Resampling didn't affect the models' Area Under the Curve.



Calibration curve of the models. The black line represents a model with perfectly calibrated probabilities. The balanced Random Forest was the model with the closest calibration curve to the ideal.

Conclusion

Conclusion

- The accident rate has been dropping in Seattle.
- Fridays are when most accidents occur, on average.
- When the weather conditions are clear, it's more likely to occur an accident.
- A driver under the influence of alcohol and drugs has more chances to cause a more serious accident
- 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.
- For future analysis:
  - include more variables in the analysis and the models to improve class 1 severity prediction precision.