Predicting car accident severity

Javier Lago – November 2020

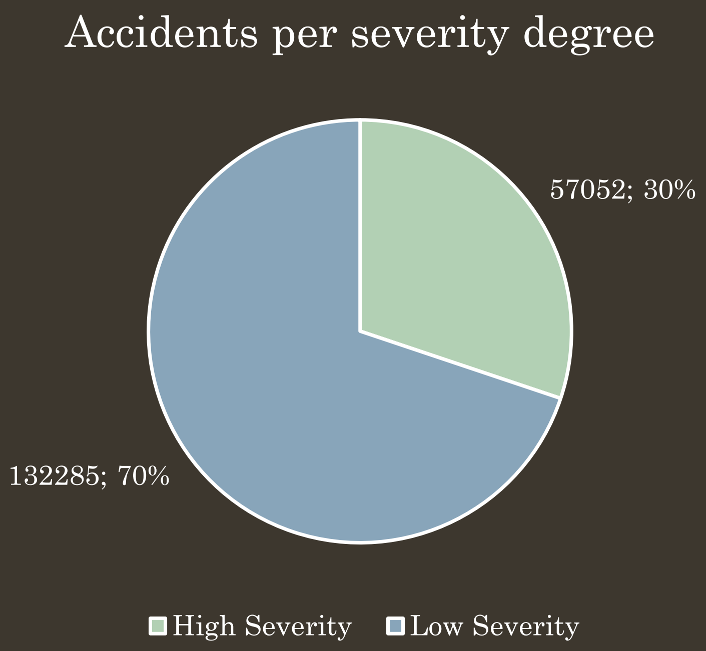
Introduction

- We would like to know if it is possible to predict the severity degree of a car accident depending on different parameters of the environment, and build a model that is able to perform this prediction and with which degree of accuracy.
- City authorities, traffic department and emergency services could use the output of the algorithm to implement preventive measures and allocate resources to minimize the risk of having high severity accidents and maximize response capacity.
- Emergency services could predict severity of an accident when receiving the first communication, being able to send the required resources and do an optimal coordination from the beginning, minimizing response time.

Data acquisition and cleaning

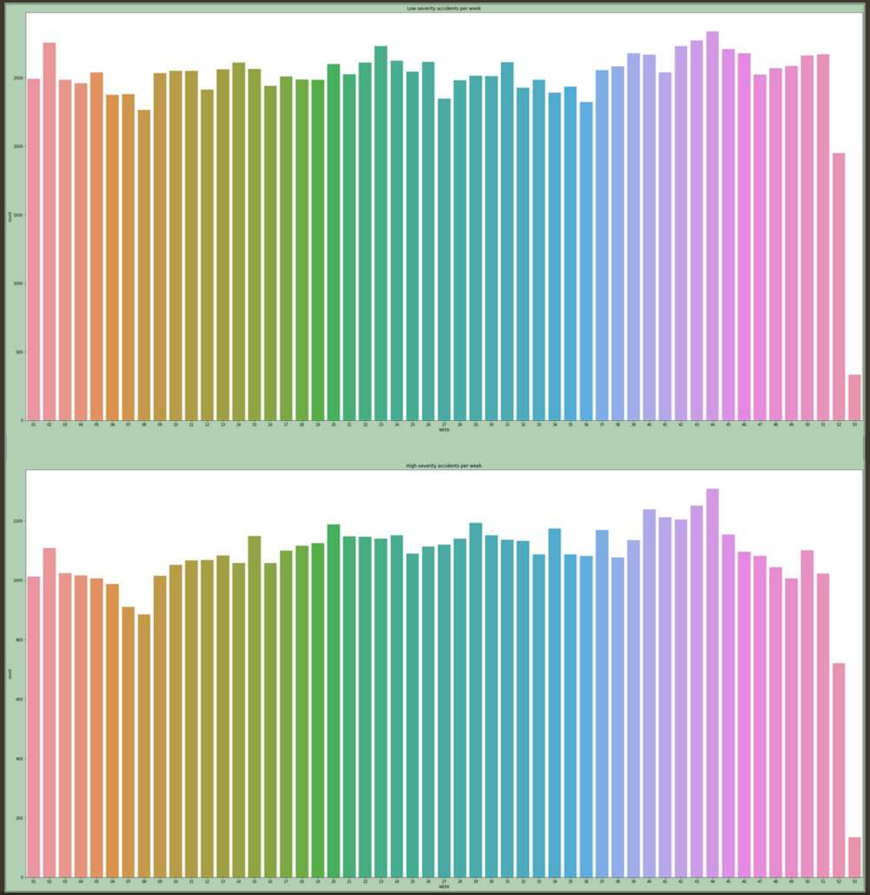
- Dataset used for the analysis is the Collisions Report provided by the Seattle Police Department (available at the Seattle Open Data Portal *https://data.seattle.gov/*).
- Original dataset contains 194673 records and 38 data fields.
- The data field containing the timestamp of the accident is transformed in three time fields (weekday, week and hour of the day), from which only week and weekday will be used. The other features used are light, road and weather condition, which we convert in binary through a hot encoding technique.
- Rows with feature values “Unknown” and “Other” or with a non relevant presence are eliminated. Rows with any empty feature are also eliminated.
- After the transformations, the final dataset contains 111686 records and 14 data fields (the target and 13 features, 11 of them binary).

Target: severity degree of a car accident



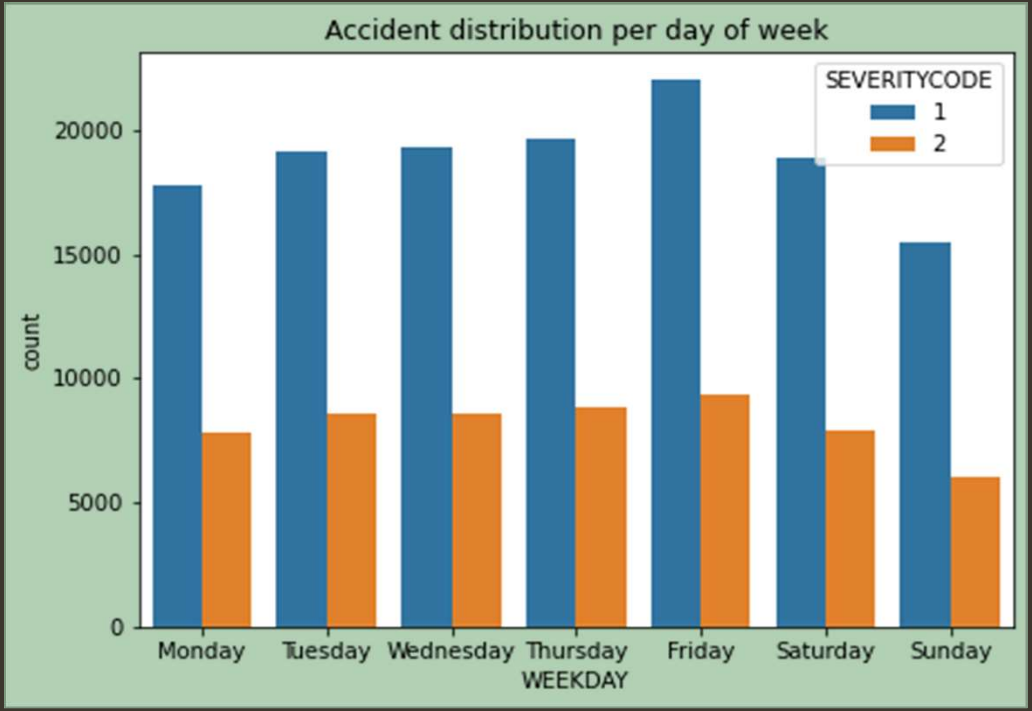
- Accidents are classified in two categories depending on its severity degree: high (2) or low (1).
- The dataset is strongly biased towards the low severity category, that contains approximately 70% of the cases.
- High severity is the category most interesting to us, as it requires urgent response and specialized emergency resources, and involves people suffering injuries.

Week Feature



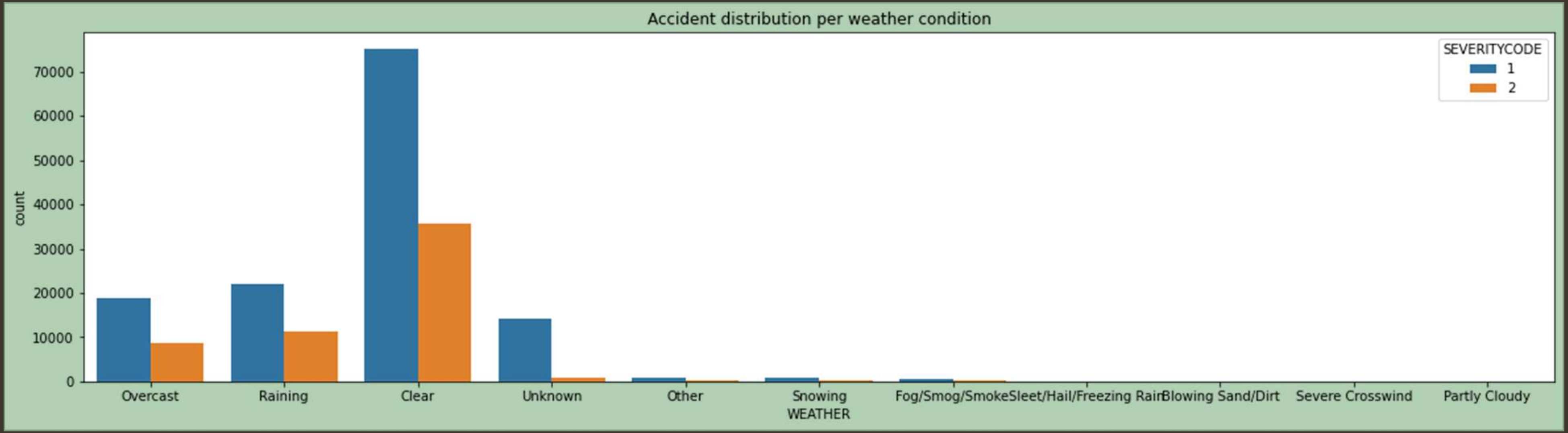
- Graphically, a similar distribution of accidents along the weeks of the year is observed both on low severity accidents (top image) and high severity accidents (bottom image).
- Average of the accident distribution per week and severity is 69.92% for low severity and 30.08% for high severity, with a standard deviation of 1.36, which is the same result for the total distribution among categories.
- It seems that this feature does not provide a clear way to classify the accidents between both categories.

Weekday Feature



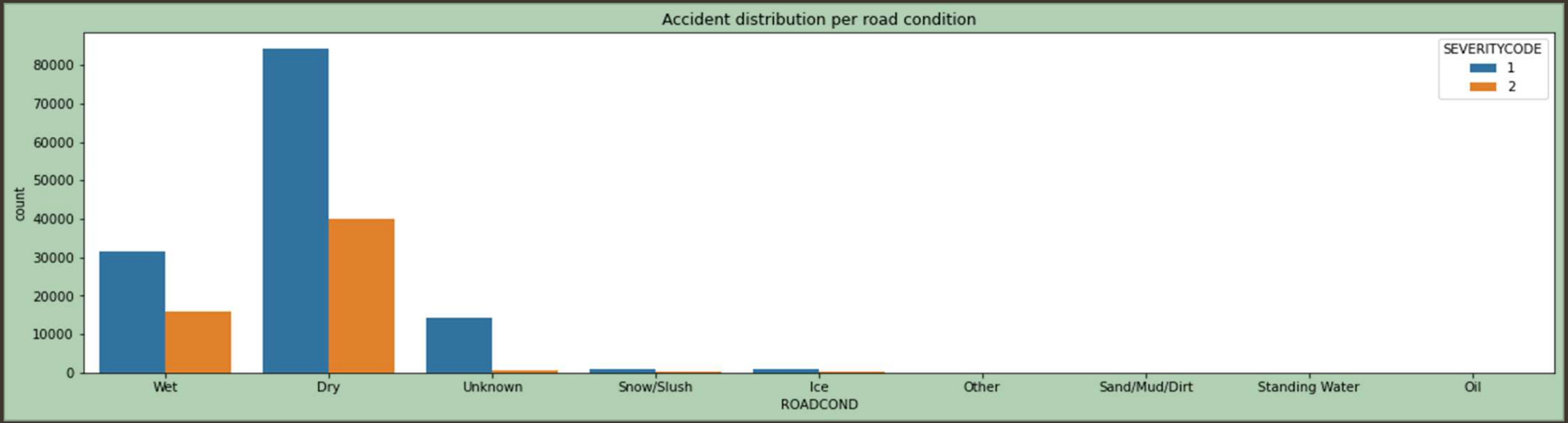
- If we study how accidents are distributed along the days of the week, we might expect to see some differences between workdays and weekends, or even between different workdays (for example Fridays).
- If we observe the distribution, the ratio between high and low severity accidents remains almost constant across the week, with average 69.94% to 30.06% and standard deviation of 1.12.

Weather Condition Feature



- If we observe accidents distribution per weather condition, we observe that there are two values that do not provide information (Unknown and Other). Also, there are only three “significant” values (Overcast, Raining and Clear), we will only keep those for the analysis.
- Proportion between both categories in these values is around 72% - 28%.

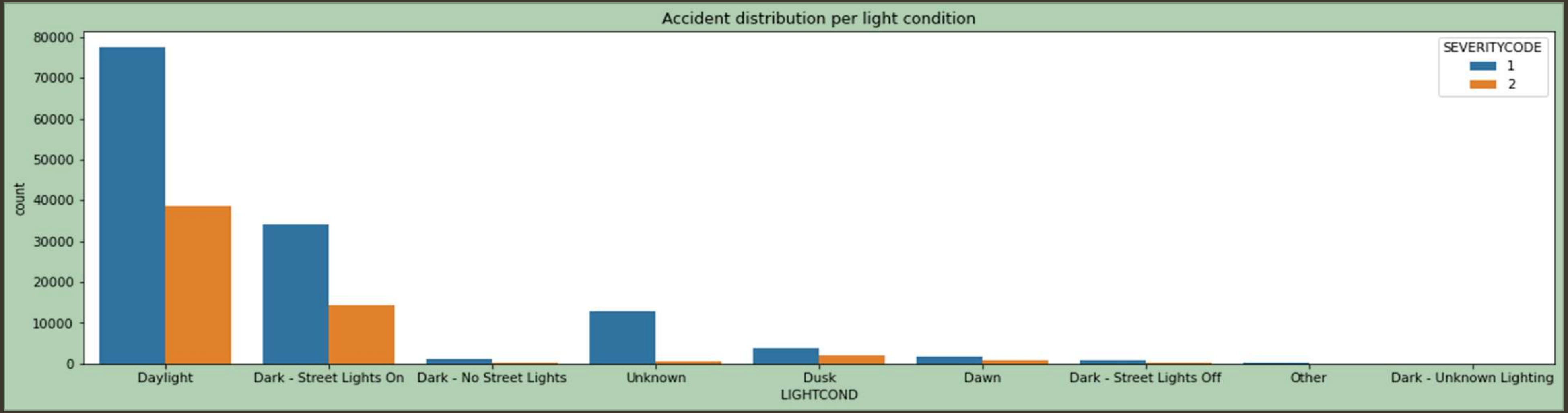
Road Condition Feature



- If we observe accidents distribution per road condition, we observe that there are two values that do not provide information (Unknown and Other). Also, there are only three "significant" values (Wet, Dry and the joint of Snow & Ice), we will only keep those for the analysis.

- Proportion between both categories in these values is also around 73% - 26%.

Ligth Condition Feature



- If we observe accidents distribution per light condition, we observe that there are three values that do not provide information (Unknown, Dark - Unknown and Other). We will keep the rest for the analysis, considering that we will take the joint of two values (Dark – No Street Lights and Dark – Street Lights Off) as a single value.
- Proportion between both categories in these values is also around 73% - 26%.

Having a biased dataset

Biased	Maximum	Optimal	Minimum
Parameter 1	-	Criteria = entropy	-
Parameter 2	-	Max. depth = 26	-
Recall	3.10%	3.10%	0.00%
Precision	42.86%	33.29%	0.00%
Accuracy	67.09%	66.06%	66.05%
F1	5.67%	5.67%	0.00%
F2	3.79%	3.79%	0.00%

Unbiased	Maximum	Optimal	Minimum
Parameter 1	-	Criteria = entropy	-
Parameter 2	-	Max. depth = 6	-
Recall	72.19%	58.27%	47.53%
Precision	51.31%	51.26%	50.18%
Accuracy	52.35%	52.35%	51.03%
F1	59.20%	54.54%	49.35%
F2	66.37%	56.72%	48.25%

- Using a biased dataset to train the model causes an overfitting problem (the algorithm has an accuracy around 66%, but a recall of only a 3%): low severity accidents are properly detected, but high severity accidents are almost no detected.
- Accuracy is not a good quality measure in this case.

- Applying an under sampling technique to the records of the category with stronger presence, in such a way that both categories are balanced, has better overall results.
- At its maximum level, accuracy is reduced in a 22%, but precision is increased in a 19%, and recall is more than 20 times higher.

Performance of the tested models

	KNN	Decision Tree	SVM	Logistic Regression	Random Forest	Gradient Tree	Voting (hard)	Voting (soft)
Parameter 1	K = 5	Criteria = entropy	Kernel = rbf	Solver = newton-cg	-	Estimators = 2	-	-
Parameter 2	-	Max. depth = 6	-	-	-	Learning rate = 0.2	-	-
Max. Recall	51.45%	72.19%	73.60%	54.12%	53.97%	68.50%	-	-
Recall	50.77%	58.27%	62.84%	54.12%	53.70%	56.52%	54.66%	54.77%
Min. Recall	26.12%	47.53%	49.39%	54.07%	49.76%	47.98%	-	-
Max. Precision	49.66%	51.31%	51.14%	51.30%	50.32%	51.37%	-	-
Precision	49.66%	51.26%	51.14%	51.30%	50.27%	51.37%	53.27%	51.96%
Min. Precision	49.08%	50.18%	49.41%	51.29%	49.79%	49.86%	-	-
Max. Accuracy	50.71%	52.35%	52.33%	52.30%	51.29%	52.43%	-	-
Accuracy	50.61%	52.35%	52.33%	52.30%	51.24%	52.43%	52.46%	51.15%
Min. Accuracy	50.01%	51.03%	50.38%	52.29%	50.74%	50.81%	-	-
F1	50.21%	54.54%	56.39%	52.67%	51.93%	53.82%	53.96%	53.33%
F2	50.54%	56.72%	60.09%	53.53%	52.97%	55.41%	50.45%	47.93%

- After testing and fine tuning different classification methods, both “simple” and “ensembled”, three of them are chosen by its best overall performance: SVM, Decision Tree and Gradient Tree Boost.

- Choosing a balanced configuration we obtain a recall value between 54.12% and 62.84%, a precision around 51% and an accuracy around 52.35%. Unfortunately, these values are not as good as desired, as we roughly predict properly half of the events.

Validation of the chosen models

	SVM	Decision Tree	Gradient Tree Boost
Recall	61.08%	54.56%	55.41%
Precision	34.67%	34.38%	34.98%
Accuracy	49.43%	50.22%	51.54%
F1	44.23%	43.99%	42.88%
F2	53.01%	52.17%	49.61%
Fitting Time	623.49 s	0.18 s	4.14 s
Execution Time	396.10 s	0.25 s	0.15 s
TP	33723	32880	30591
FN	21485	22328	24617
FP	63550	61383	56867
TN	49407	51574	56090

- To perform a final validation, we take the chosen methods with the fine tuned parameters and we train them with the whole unbiased dataset that we generated by under sampling the category that has the majority of cases.

- Once trained, we test them using the whole (biased) dataset. We can observe that there is an effect over some of the metrics.
  - Recall is reduced slightly (between 1 and 4 points), similarly to accuracy (between 1 and 2 points). We can still detect properly half of the high severity cases and half of the total cases.
  - Precision is strongly reduced (around a 33%, 17 points). This is due to the fact that we are wrongly classifying a lot more low severity accidents as high severity ones (low severity accidents are more frequent, and we only classify properly half of them).

Conclusions and potential future developments

- If we use the original dataset (biased) for training, we will get model overfitting.
- As the input of the model is strongly biased towards low severity, detection performance of high severity (the most interesting for us) suffers some degradation.
- In this case we are not able to use accuracy as the main metric to maximize when tuning the classification model as all categories are not equally important.
- The prediction of an accident severity degree based on the chosen set of features has a low performance, it seems that dependency with these factors is either low or null.
- Additional features shall be studied to look for an increase of model performance.
- A study regarding which are the factors more related to accident severity shall be done, to know if it is more dependent on other “environment” conditions, different than the ones considered, or in specific characteristics of the accidents.