Machine Learning for Predicting Severity of Collisions

1 Introduction

1.1 Background

Would it not be great if you could tell in advance how sever an accident will be if it occurs?! In this project, we will use a car-accident data of Seattle city in the Washington State, US from 2004 to 2013. The data can be download from [1].

The data contains many useful features such as weather and road conditions. The severity of the accident is classified into several classes ranging from property damage to fatality. This problem is interesting for the Departments of Transportation especially in the Seattle City. It can help the department linking weather, road and light conditions, certain locations and certain streets to very sever accidents.

1.2 Problem Statement

Car accidents is one of the leading causes of death in the US. In this project, we set out to predict the severity of car accidents using machine learning technique and real-world data.

It is expected that the machine learning classifier will be able to classify the severity of an accident based on its attributes. Furthermore, the exploratory data analysis will shed light on different hidden patterns in the data that can be useful for decision makers.

2 Data

2.1 Data Understanding

The data contains information about the car accidents occurred in the Seattle city in the US The data contains 221144 records and 40 attributes (features). The values of the target label (the severity of accidents) are described in Table 1.

Value	Description
0	Unknown
1	Property damage
2	Injury
2b	Serious injury
3	Fatality

Table 1: Target label.

2.2 Data Cleaning

1. Features with 50% or more missing values are dropped since imputing their missing values will alter the real distribution of the data significantly and therefore, can be misleading to the machine learning algorithm.
2. Missing values are replaced using a roulette-wheel replacement method that preserve the distribution of the data.

2.3 Data Wrangling & Feature Engineering

The following techniques are used to generate new, hopefully more informative, features out of the original features.

1. **Binning** is used to group values that are invariant to the target label to reduced the dimensions of the data.
2. **One-hot encoding** is used to convert categorical variables into binary columns.
3. **Binary encoding** is used when one-hot encoding is not practical (such as when there are two many levels for a categorical variable).

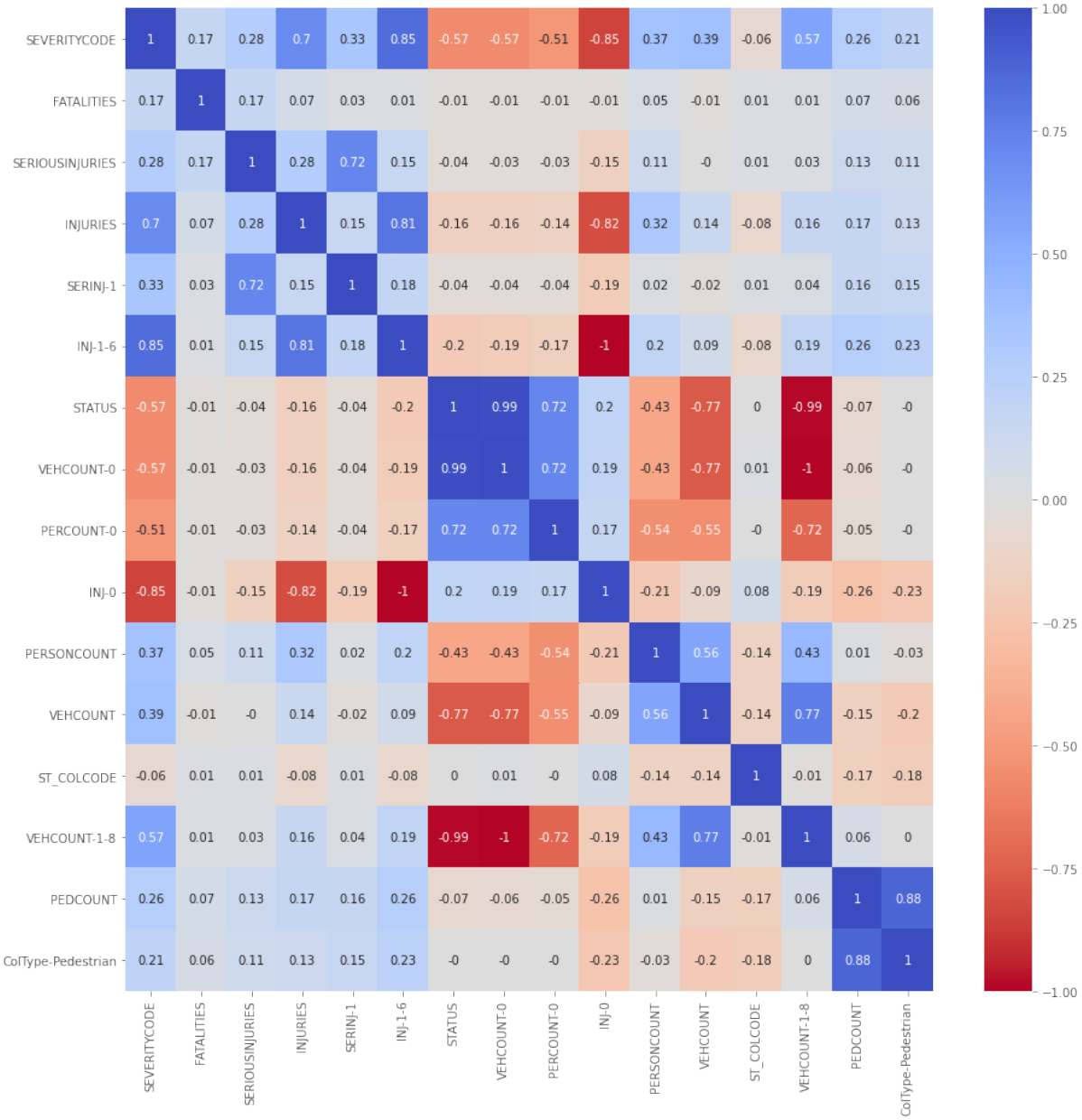
3 Methodology

3.1 Feature Selection

After the feature engineering step, we end up with 81 features that can be used in modeling. Feeding all these features to the model is likely to degrades the performance of the model since it will incur a very high number of parameters to be learned. Therefore, we perform feature selection as follows:

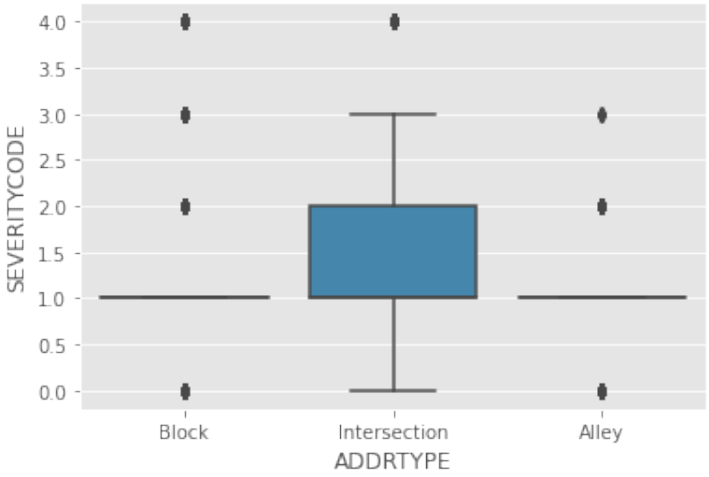
1. We use the **chi test** together with the correlation matrix to select the top features.
2. The feature selection is perform by considering the top positively and negatively correlated features using a correlation matrix.

The correlation matrix of the 15 features that are used in the model.

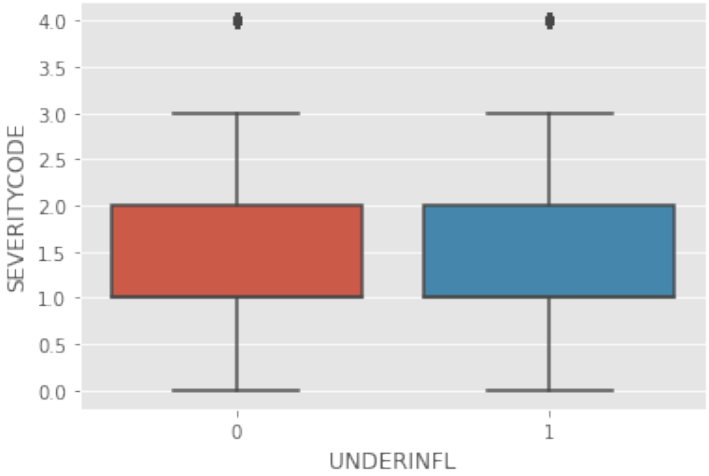

SEVERITYCODE 1 0.17 0.28 0.7 0.33 0.85 -0.57 -0.57 -0.51 -0.85 0.37 0.39 -0.06 0.57 0.26 0.21
FATALITIES 0.17 1 0.17 0.07 0.03 0.01 -0.01 -0.01 -0.01 -0.01 0.05 -0.01 0.01 0.01 0.07 0.06
SERIOUSINJURIES 0.28 0.17 1 0.28 0.72 0.15 -0.04 -0.03 -0.03 -0.15 0.11 -0 0.01 0.03 0.13 0.11
INJURIES 0.7 0.07 0.28 1 0.15 0.81 -0.16 -0.16 -0.14 -0.82 0.32 0.14 -0.08 0.16 0.17 0.13
SERINJ-1 0.33 0.03 0.72 0.15 1 0.18 -0.04 -0.04 -0.04 -0.19 0.02 -0.02 0.01 0.04 0.16 0.15
INJ-1-6 0.85 0.01 0.15 0.81 0.18 1 -0.2 -0.19 -0.17 -1 0.2 0.09 -0.08 0.19 0.26 0.23
STATUS -0.57 -0.01 -0.04 -0.16 -0.04 -0.2 1 0.99 0.72 0.2 -0.43 -0.77 0 -0.99 -0.07 -0
VEHCOUNT-0 -0.57 -0.01 -0.03 -0.16 -0.04 -0.19 0.99 1 0.72 0.19 -0.43 -0.77 0.01 -1 -0.06 -0
PERCOUNT-0 -0.51 -0.01 -0.03 -0.14 -0.04 -0.17 0.72 0.72 1 0.17 -0.54 -0.55 -0 -0.72 -0.05 -0
INJ-0 -0.85 -0.01 -0.15 -0.82 -0.19 -1 0.2 0.19 0.17 1 -0.21 -0.09 0.08 -0.19 -0.26 -0.23
PERSONCOUNT 0.37 0.05 0.11 0.32 0.02 0.2 -0.43 -0.43 -0.54 -0.21 1 0.56 -0.14 0.43 0.01 -0.03
VEHCOUNT 0.39 -0.01 -0 0.14 -0.02 0.09 -0.77 -0.77 -0.55 -0.09 0.56 1 -0.14 0.77 -0.15 -0.2
ST_COLCODE -0.06 0.01 0.01 -0.08 0.01 -0.08 0 0.01 -0 0.08 -0.14 -0.14 1 -0.01 -0.17 -0.18
VEHCOUNT-1-8 0.57 0.01 0.03 0.16 0.04 0.19 -0.99 -1 -0.72 -0.19 0.43 0.77 -0.01 1 0.06 0
PEDCOUNT 0.26 0.07 0.13 0.17 0.16 0.26 -0.07 -0.06 -0.05 -0.26 0.01 -0.15 -0.17 0.06 1 0.88
ColType-Pedestrian 0.21 0.06 0.11 0.13 0.15 0.23 -0 -0 -0 -0.23 -0.03 -0.2 -0.18 0 0.88 1
SEVERITYCODE
FATALITIES
SERIOUSINJURIES
INJURIES
SERINJ-1
INJ-1-6
STATUS
VEHCOUNT-0
PERCOUNT-0
INJ-0
PERSONCOUNT
VEHCOUNT
ST_COLCODE
VEHCOUNT-1-8
PEDCOUNT
ColType-Pedestrian
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

3.2 Exploratory Data Analysis

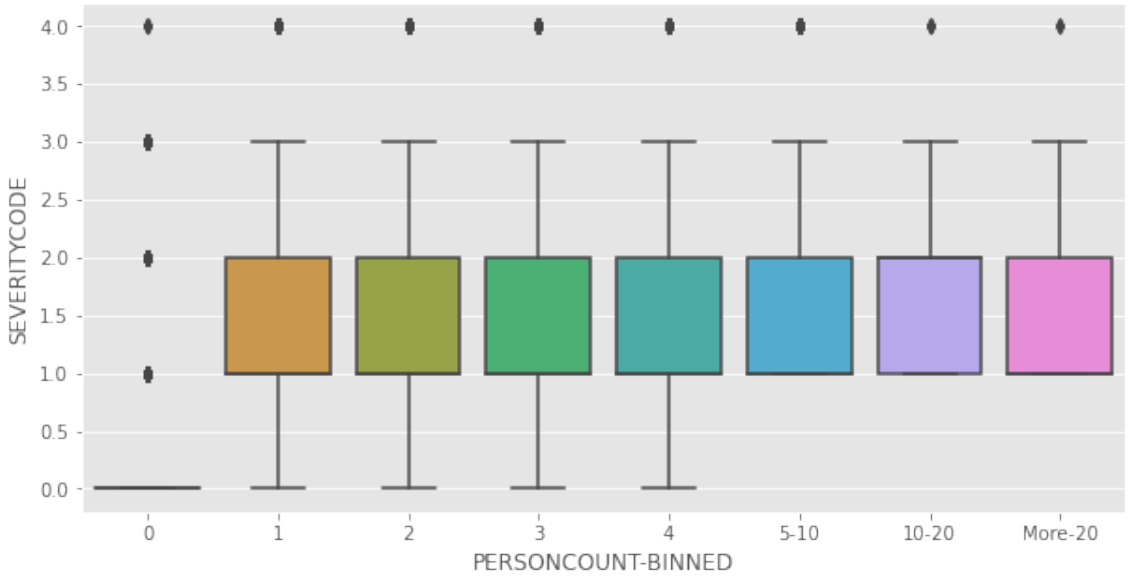
Accidents occurring at intersections are more sever.



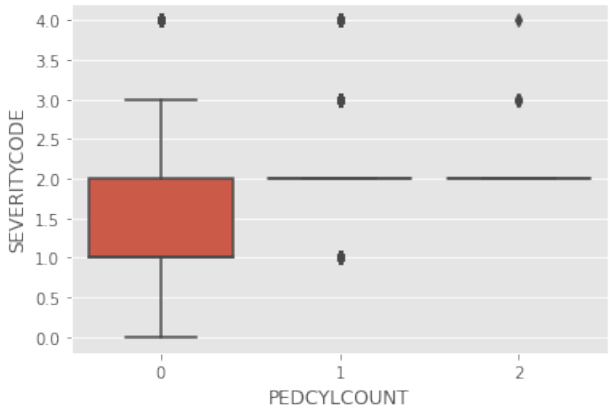
Whether the driver was under the influence of alcohol or not, the severity of the accident remains the same which counter our intuition. This could possibly be as a result of trying to drive more carefully when under alcohol influence. This is true for those who are not totally drunk.



An accident involving pedestrians is more sever. If 5 ore more pedestrians are involved, the accident tends to be even more sever.



Accidents involving bicyclers are more sever.



3.3 Modeling

The *random forest classifier* is used due its ability to deal with imbalanced classes without the need to *up-sampling* or *down-sampling* the original data. This is the case since the decision trees have hierarchical structures that enables them to effectively distinguished different classes regardless of the actual number of classes.

4 Result & Discussion

The predictive models that we built achieves an accuracy of 99% which is very high and is able to predict all five classes despite the fact that the problem is highly imbalanced. This ensures that the high accuracy of the model is not due to being very good at predicting the majority class. On contrary, the model

high predictive power is due to being able to distinguish all five classes. Random forest classifiers is used since its hierarchical structures enables decision trees to effectively distinguish different classes even if the data is imbalanced.

Important Observations:

1. Surprisingly, the geographic data tends to be insignificant for this problem. The target label is very weakly correlated with the locations at which the collisions occur.
2. The under-influence-of-alcohol state does not influence the severity of the accident which counters our intuition. In particular, the data reveals that the distribution of the severity of the accident remains invariant with the under-influence-of-alcohol state.
3. The poor weather and road conditions do not cause sever accidents. This could possibly be explained by the extra care exercised by the drivers under poor weather or road conditions.
4. Accidents at the intersection are more sever than those in alleys and blocks.

Recommendations:

1. The most sever accident involved pedestrians and bicyclers. Therefore, better policies to ensure the safety of pedestrians and bicyclers are needed.
2. Accidents at intersection are more sever than those in alleys and blocks. This needs to be investigated physically to see whether these intersections have working traffic lights. If not, perhaps, installing new ones will reduce the number of accidents.
3. Though very important, car speeding is mostly missing! This could be due to the difficulty of measuring the exact speed just before the accidents. However, we believe it is a very important predictive feature that should be better recorded.

5 Conclusion

The project aims at predicting the severity of car accident using data acquired from the Seattle City, US.The data is rich with 40 features and 221144 records.

The problem we dealt with is imbalanced as the number of records across classes. Therefore, we carefully chose the random forest classifiers owing to its power in distinguishing different imbalanced classes as this study showed. We ignored the resampling technique to balance the classes since the up-sampling will unacceptably increase the size of the data and the down-sampling will reduce it significantly.

The model is highly accurate (achieved an accuracy of 99%) and is able to predict all five classes which ensured that the high accuracy is not solely produced by predicting the majority class.

Reference

[1] Data Source: https://data.seattle.gov