Machine Learning for Predicting Severity of Collisions

Introduction

1. Car accidents is one of the leading causes of death in the world.
2. Despite all being damaging, the severity of an accident varies from inflecting just a property damage all the way up to fatality.
3. The severity of the collision depends on many factors such as weather and road conditions, the number of pedestrians or bicyclers involved, etc.

Problem

1. In this study, we will use machine learning techniques to predict the severity of collision based on several attributes.
2. This problem is interesting to the Departments of Transportation especially in the Seattle City. The study will reveal useful information about collisions that can be used to reduce both the collision severity and occurrences.

Objectives

This project aims at predicting the severity of car accidents using Seattle City data.

Data

1. The data source is the Seattle City Open Data Portal [1].
2. The data contains 40 features and 221141 records.

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.

Data continues . . .

Data Wrangling & Feature Engineering:

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).

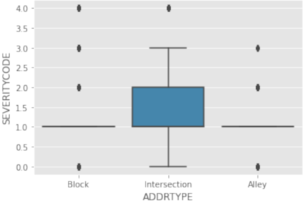
In total, the number of features is increased from 25 (after data cleaning) to 81 (after feature engineering).

Feature Selection

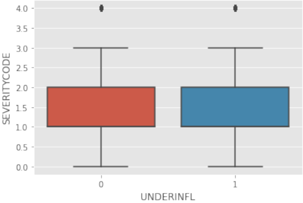
1. We use the **chi test** together with the correlation matrix to select the top features.
2. The feature selection is perform by considering the top positively and negatively correlated features using a correlation matrix.

Exploratory Data Analysis

Accidents occurring at intersections are more sever.

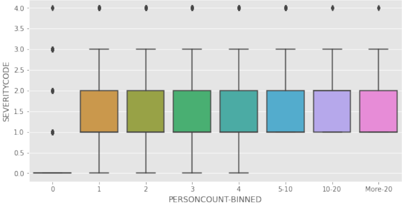


whether the driver was under the influence of alcohol or not, the severity of the accident remains the same!!

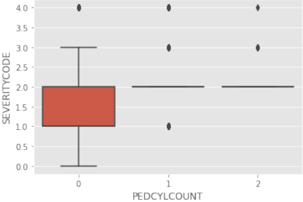


Exploratory Data Analysis continues . . .

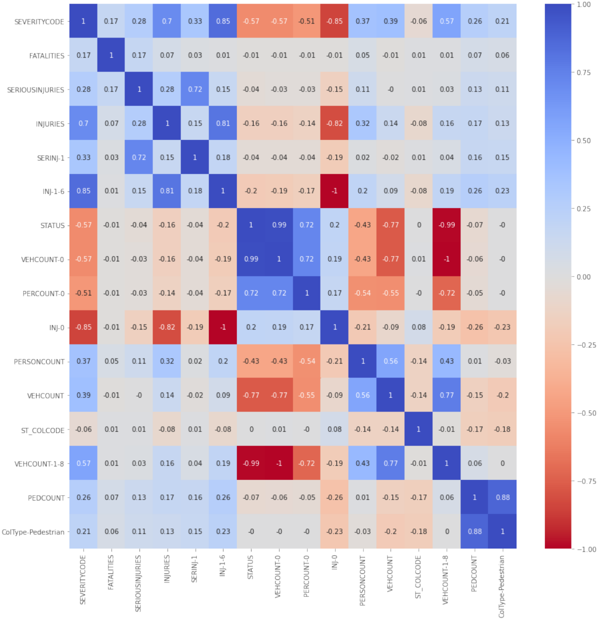
An accident involving pedestrians is more sever. If 5 ore more pedestrians are involved, the accident tends to be even more sever.



Accidents involving bicyclers are more sever.



The correlation matrix of the top 15 chosen features for the model.



	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
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

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.

Results & Discussion

Main Result

The predictive model achieves an accuracy of 99% and is able to predict all classes including the minority classes.

1. Surprisingly, the geographic data tends to be insignificant for this problem.
2. 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. Perhaps drivers tend to drive more carefully under those 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.

Conclusion

1. 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.
2. 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.
3. 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.