IBM Applied Data Science Capstone Report

Reducing The Severity of Car Accidents in Seattle

By Azizi Omar

September 2, 2020

1 Introduction

1.1 Background

Car accident, also called as a traffic collision, occurs when a vehicle collides with another vehicle of any size, animal, pedestrian, road debris or any other stationary objects, such as a tree, building or pole. Traffic collisions often result in property damage, injury, permanent disability, death and financial costs to all the individuals involved. In the United States, the yearly average number of car accident is 6 million [1]. 3 million people in the United States are injured every year in car accidents [1].

1.2 Business Problem

The severity of car accident is highly linked to the type and the number of involved parties. In general, the most severe car accidents are those which cause human injury. This is the aspect of car accident which we are to focus in this capstone project. Factors which contribute to this type of injury need to be determined. In this capstone project, after identifying the factors of car accident contributing to the human injury, we will investigate the methods to reduce the severity of car accidents in Seattle related to human injury.

1.3 Interest

Seattle Police Department (SPD) and the government of Seattle would be interested with the outcomes of this capstone. They would be able to use them to propose and implement adequate measures to reduce the human injury caused by car accident.

2 Data

2.1 Data Sources

For this project, we will use the shared data, which is the collision data provided by Seattle Police Department (SPD). The timeframe of the data based on the provided description is from 2004 to present. The dataset includes all types of collisions. For our purposes, we will only use collision data involving **vehicles with other parties**.

2.2 Data Description

A pandas-profiling[2] of the whole dataset has been generated. There are 38 attributes in this dataset, excluding the index. The attribute which is going to be our label or **target variable is the severity code ("SEVERITYCODE"),** as this variable will determine the severity of the car accident. The attribute of vehicles count ("VEHCOUNT") will help us to determine traffic collisions involving vehicles (cars). The rest of the attributes will be further investigated, in order to determine the important features which contribute to human injury caused by car accident.

There is a total of **189,588 observations** of traffic collisions involving minimum one vehicle (car). Percentage of missing cells is fairly high, 14.1% (1,042,065). The missing data and the outliers will be handled in the data preparation part of the next section.

3 Methodology

3.1 Data Preparation (Imputation & Transformation)

Based on the pandas-profiling generated previously, we could see that there is an attribute "SEVERITYCODE.1" which is an exact copy of another attribute "SEVERITYCODE", thus **this duplicate attribute** will be removed. The attribute "INCDATE" will be removed as it does not carry more information then the attribute "INCDTTM".

High cardinality attributes (with many unique values), such as "REPORTNO", "OBJECTID", "INCKEY", "COLDETKEY" , are bad choices of features. Because of this, they will be removed. Attributes "INTKEY","EXCEPTRSNCODE", "EXCEPTRSNDESC" and "SDOTCOLNUM" have a lot of **missing values**, at least 20% of the observations and they could not be imputed easily. Therefore, the attributes will be removed from disturbing the training process of the machine learning model. In this capstone project, **geospatial factors** will be ignored as they have different data treatment and different machine learning model. The involved attributes are as follow, "ADDRTYPE" and "LOCATION". Based on the given metadata (data reference sheet), there is a few attributes which have **little information**, almost

to none. They are “X”, “Y”, “STATUS”, “SEGLANEKEY”,”CROSSWALKKEY” and will be ignored as features. The attributes which **do not bring extra useful information** for model training part- the attributes to describe codes such as “SEVERITYDESC“, “SDOT_COLDESC“, “ST_COLDESC“ will be ignored as well. The attributes of the codes are able to replace these attributes. We will also ignore attribute “ST_COLCODE” as it is an unsupported pandas type column.

The attributes "INATTENTIONIND", "PEDROWNOTGRNT", "SPEEDING" and "UNDERINFL" have a high number of missing values. We noticed that these attributes have a single type of value “Y” representing a value “YES”, thus an **assumption of missing values being the value “N”** (No) is made from this observation. Therefore, we have **imputed** the missing values with a value “N”. For the attributes as follow, "COLLISIONTYPE", "JUNCTIONTYPE", "WEATHER", "ROADCOND" and "LIGHTCOND", also contain a great number of missing values. Based on the previous pandas-profiling, we could see that there is a value “Other” / “Unknown” to represent other unknown value or type, thus this can be used to replace the missing values.

Data integrity plays a big part in producing a good prediction model. We observed incoherent values in the attribute "UNDERINFL". The attribute contains other values than the expected values (“Y”, “N”). We replaced the value “1” with “Y” and the value “0” with “N”.

The **Boolean attributes** are then converted into “1” and “0”, for facilitating the training process of machine learning models, as most of models require a numerical value.

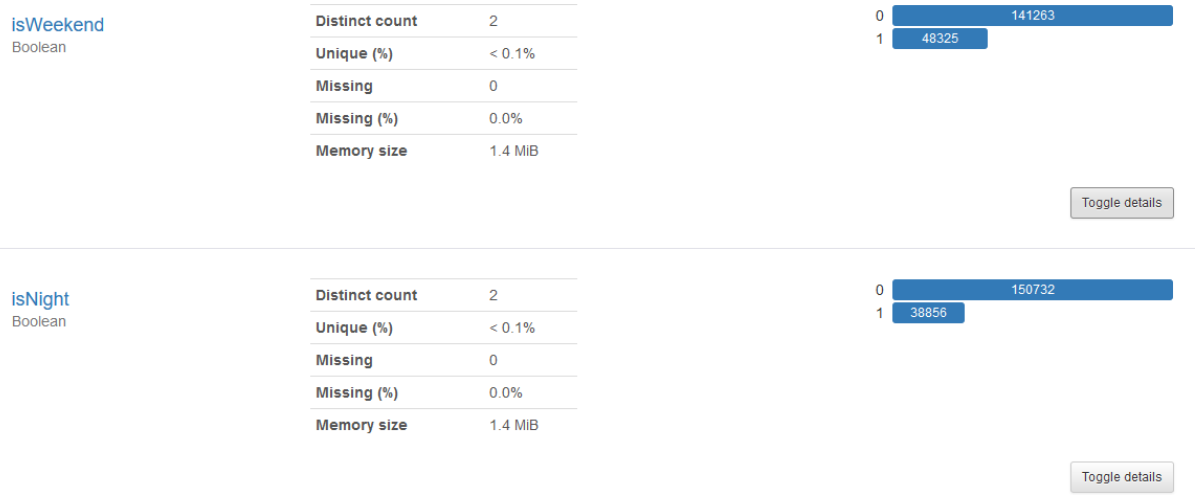
Transformation of severity code into “1” and ”0”. If the severity code equals to 1, referring to “Damaged property severity”, will be replaced with “0”, or else it equals to 2, referring to “Human injury severity” will be replaced with “1”. This new transformed attribute will indicate whether the severity of car accident related to human injury or not.

In order to **enrich and augment the dataset**, we add extra attributes “isWeekend” and “isNight”, which we obtained and transformed from the attribute of accident datetime “INCDTTM”. If the accident occurred during weekend, it will be flagged as “1” for the attribute “isWeekend”, whereas if the accident occurred after 6pm, it will be flagged as “1” for the attribute “isNight”.

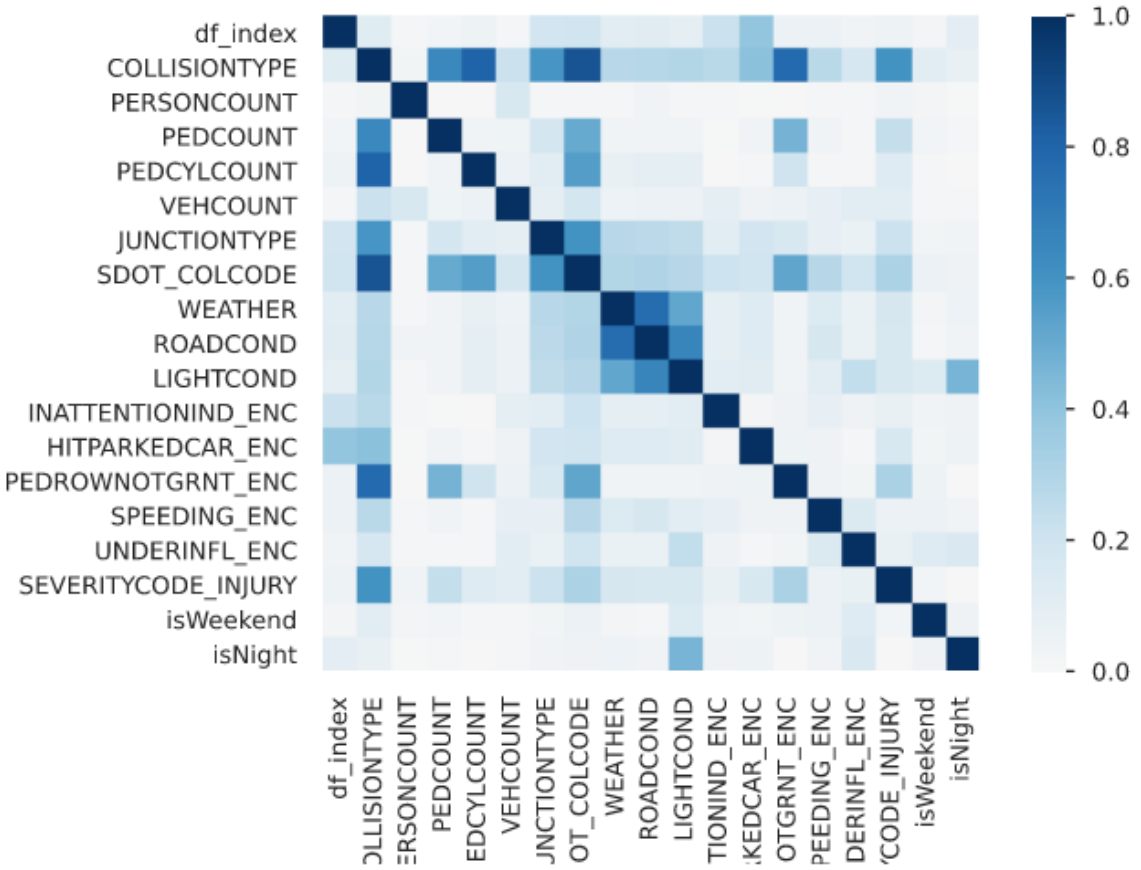
3.2 Exploratory Data Analysis (EDA) & Feature Selection

Another pandas-profiling[3] is generated for the newly prepared data. The data preparation part has reduced the number of attributes to 19 useful and important attributes. There is no more missing data in the dataset, neither duplicate columns. We have a total of 5 numerical attributes, and 14 categorical attributes.

We could see that the attributes shown in the above image have **a big number of value zero**, which may not be useful for the machine learning model.



Based on the horizontal bar graphs produced by the pandas-profiling, we could observe that there is a **little number** of car accidents **during weekend and during night** (after 6pm).



Based on the **correlation** provided by the pandas-profiling, using the algorithm "Phik", we could see that the target attribute of **"SEVERITYCODE_INJURY"** has high correlation with the attribute **"COLLISIONTYPE".** It also has slightly high correlation with the attributes as follow, **"SDOT_COLCODE", "PEDROWNOTGRNT_ENC", "HITPARKEDCAR_ENC", "JUNCTIONTYPE" and "PEDCOUNT".** These are the attributes to be fit into the model.

COLLISIONTYPE

Categorical

Common Values | Length | Unicode

Value	Count	Frequency (%)
Parked Car	47987	25.3%
Angles	34674	18.3%
Rear Ended	34090	18.0%
Other	23722	12.5%
Sideswipe	18609	9.8%
Left Turn	13703	7.2%
Pedestrian	6607	3.5%
Cycles	5216	2.8%
Right Turn	2956	1.6%
Head On	2024	1.1%

For the **highly correlated attribute of "COLLISIONTYPE",** the car accident involving parked car is the highest in number, followed by car accident involving angles, whereas the least being the head on type of car collision.

3.3 Machine Learning Model

Dataset has been split into **training and testing dataset** with a **ratio of 70:30**. For the model training part, we have adopted an **auto classification model developed by H2O.ai (AutoML)**. This is a robust model, where it will fit the dataset into all available classification models such as "StackedEnsemble"," XGBoost", "GBM", "Decision Tree", "Random Forest" and "KNN". From there, it will come out with **the best classification model** based on the metric that we have chosen, using "AUC" by default. We have specified the parameter "balance_classes" of H2O.ai in order to **upsampling** our dataset to avoid problems related to imbalanced dataset. Furthermore, we have included **cross-validation** in our model with 5 folds, in order to prevent overfitting. We will **exclude the model of "Deep Learning"** as it consumes a great deal of time to train the model.

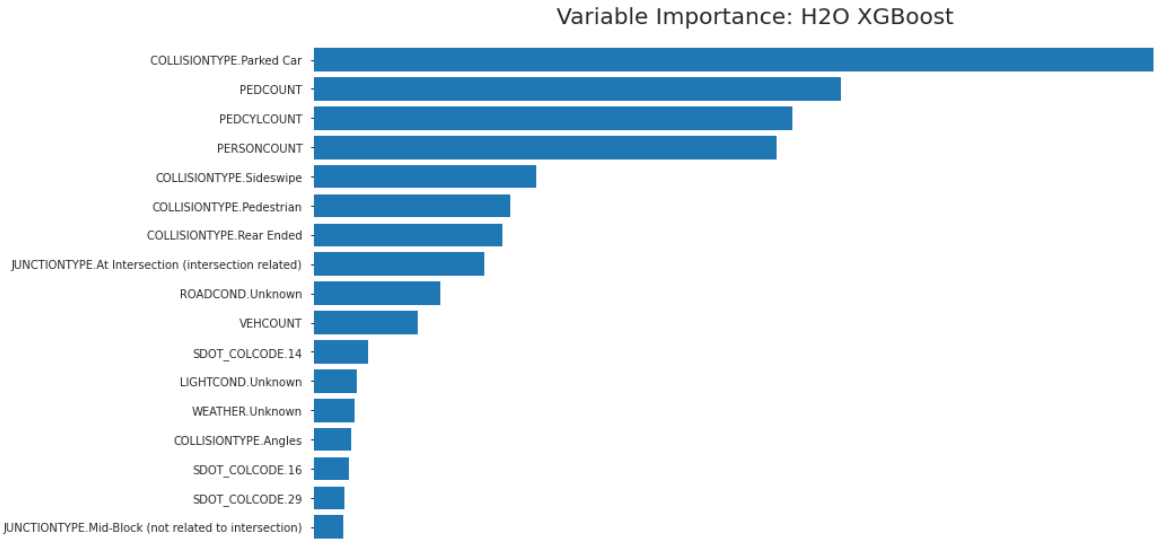
4 Results

model_id	auc	logloss	aucpr	mean_per_class_error	rmse	mse
StackedEnsemble_AllModels_AutoML_20200903_215827	0.803589	0.474433	0.64453	0.280415	0.396662	0.157341
StackedEnsemble_BestOfFamily_AutoML_20200903_215827	0.803243	0.474713	0.643306	0.280404	0.396777	0.157432
XGBoost_3_AutoML_20200903_215827	0.802843	0.46718	0.642304	0.280383	0.395467	0.156394
XGBoost_grid__1_AutoML_20200903_215827_model_2	0.802828	0.467266	0.642629	0.281156	0.395489	0.156411
GBM_grid__1_AutoML_20200903_215827_model_3	0.80274	0.467437	0.642457	0.280848	0.395579	0.156483

The total training time is **7min and 45s**. The result shows us that the **best model is "Stacked Ensemble All Models" with an AUC of 0.8036**, followed by "Stacked Ensemble Best of Family" and "XGBoost" with AUC of 0.8032 and 0.8028 respectively.

Looking into the best model, which is "Stacked Ensemble All Models, we could see that the **max accuracy of the model is 76.21%,** whereas the max precision is 90.99% and max recall is 100.00%.

We are also able to **retrieve important variables** from H2O.ai model of XGBoost which is our third best model.



We could see that the **attribute of collision type**, especially the value of **"Parked Car",** is the most **important variable** to determine whether the car accident contributes to severity related human injury. This is followed by the attribute pedestrian count and pedestrian / cyclist count.

5 Conclusion

Based on the obtained result, we can conclude that the most effective way of reducing severity of car accident, is by focusing on the **collision type of car/vehicles and the number of pedestrian / cyclist** involved in the car accidents. The Seattle Police Department (SPD) and Government of Seattle could use this results **by putting more surveillance camer**a near the pavement and the cycling lane where most of the pedestrian and cyclist are. Besides that, it is advisable to put surveillance camera in the parking area.

6 References

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

[2] https://zyron92.github.io/coursera/ibm_applied_ds_capstone_profiling_overall.html

[3] https://zyron92.github.io/coursera/ibm_applied_ds_capstone_profiling_final.html