CAPSTONE PROJECT

Car Accident Severity Report

Abstract

The purpose of this document is to present the report deliverable for the data science capstone project analyzing the car accident severity using dataset obtained for Seattle Department of Transport.

Kagiso Magowe

October 4 2020

Table of Contents

1. Introduction

Seattle being one of the fastest-growing major United States cities has recently been experiencing an increase in the number of road accidents. As such, this had caught an attention in key stakeholders within the Seattle Department of Transport (SDOT). Noting the City's long-term vision of making Seattle one of the most liveable cities in the world, the SDOT want detailed insights on the car accident severity (fatality) considering key factors that might have contributed to the accident. The findings will be very useful as they will enable stakeholders to deduce data-driven strategies to prevent future road accidents and thus, consequently saving lives.

2. Data

To solve this problem, the readily available shared "Data-Collisions.csv" dataset [1] from SDOT will be used. The dataset has 38 attributes (columns), with the first column being the target column labelled "SEVERITYCODE". The remaining 37 attributes are feature candidates to choose from when building the model. In total, the dataset contains 194 673 observations which are sufficient to train and test the machine learning model. The 37 attributes comprise categorial and numerical data. Key consideration will be taken when selecting feature columns, for instance, columns with very high numbers of missing values or are duplicates will be dropped e.g., SPEEDING (9333 non-null entries), EXCEPTRSNDESC (5638 non-null entries), PEDROWNOTGRNT (4667 non-null entries) and a duplicate column of SEVERITYCODE. On the other hand, feature columns including but not limited to weather condition (WEATHER), light conditions (LIGHTCOND), road condition (ROADCOND) junction type (JUNCTIONTYPE), number of people involved (PERSONCOUNT), number of vehicles involved (VEHCOUNT), and collision type (COLLISIONTYPE) will be viable candidates to consider when building the model. It is also worth noting that these columns have non-null values > 189 000. Finally, it is worth noting that the target label is very unbalanced ~ 7:3, hence the dataset needs to be balanced to avoid the effects of bias in the model. Table 1 shows a brief description of important attributes considered for model building.

Table 1. Description of important attributes.

Attribute (Column)	Description	Example (Column Levels)
WEATHER	Weather conditions	Clear, overcast, raining, other, unknown
ROADCOND	Road conditions	Wet, dry, other, unknown
LIGHTCOND	Light conditions	Dark-Street Lights On, daylight, dusk, unknown
UNDERINFL	Influence of drugs or alcohol status	Yes or No
JUNCTIONTYPE	Category of junction	Intersection, mid-block (intersection), mid-block (no intersection, driveway junction
VEHCOUNT	Number of vehicles involved in collision	0 to 12
COLLISIONTYPE	Collision type	Angles, parked car, left turn, rear ended, other
PERSONCOUNT	Total number of people involved in collision	Between 0 and 81
INCDTTM	Date and time of the incident	From 2004 to 2019
SEVERITYCODE	Severity code of the collision	1-property damage, 2-injury

It is worth mentioning that the "Data-Collisions.csv" dataset will be separated into training dataset (10000 observations) and testing dataset (1000 observations) files called "Data-Collisions_train.csv" and "Data-Collisions_test.csv", respectively. This will allow us to use the training dataset to build the prediction models and the testing dataset to evaluate the resulting models using unseen test data.

3. Methodology

In this section, we present the methodology adopted in addressing the business problem utilizing the provided datasets from SDOT. We analyse the dataset to summarize the main characteristics. We perform descriptive statistical analysis to describe our data using e.g., describe(), value_counts(), and info() methods. This allows us to gain more insights about the nature of the data including types, missing data, and whether target classes are balanced or not (to avoid biased model). We then pre-process the selected features including:

- Converting date and time to datetime object.
- Dropping missing data since they were very few relative to the entire dataset.
- Handling imbalanced classes by considering different techniques including 1) up-sample minority class, 2) down-sample majority class, 3) penalize algorithms and 4) tree-based algorithms. In this work, we adopted the down-sample approach since we had a very large dataset.
- Handling categorial variables with many levels. We used a two-step approach of 1) combining levels (to filter out rare levels) using frequency distribution i.e., we examined frequency distribution of each level and combined levels having frequency less than 5% of total observation (this threshold is standard and can be changed based on the distribution) and 2) dummy coding the resulting categorial variables.
- Data visualization of the resulting dataframe to examine the severity of the accidents at e.g., a particular day of the week considering a wide range of selected features (weather conditions, road conditions, light conditions, junction type).
- Finally, after data pre-processing stage, we used machine learning approaches to build a model that predicts (with the best accuracy) the accident severity and applied the following classification algorithms; Support Vector Machine (SVM), Decision Tree, K Nearest Neighbor (KNN), and Logistic Regression.

a. Exploratory Data Analysis

We examine some of the data visualization plots to look at the severity of accidents at a particular day of the week. That is, we highlight the relationship between the following selected one-hot encoded features; WEATHER_Overcast, ROADCOND_Wet, LIGHTCOND_Daylight, JUNCTIONTYPE_INT, and VEHCOUNT_3.

Figure 1 shows that when the weather condition is overcast, there are less accidents compared to when it is not. In addition, most accident collisions experience severity code 2 (injury) than severity code 1 (property damage). Collisions resulting in property damage appears to occur over the weekend or beginning of the week.

Similarly, Figure 2 shows that when the road condition is wet, there are less accidents compared to when it is not. Again, most accident collisions experience severity code 2 (injury) than severity code 1 (property damage). Collisions resulting in property damage appears to occur over the weekend or

beginning of the week. The injury type accidents tend to increase steadily between days 0 and 4 (weekdays).

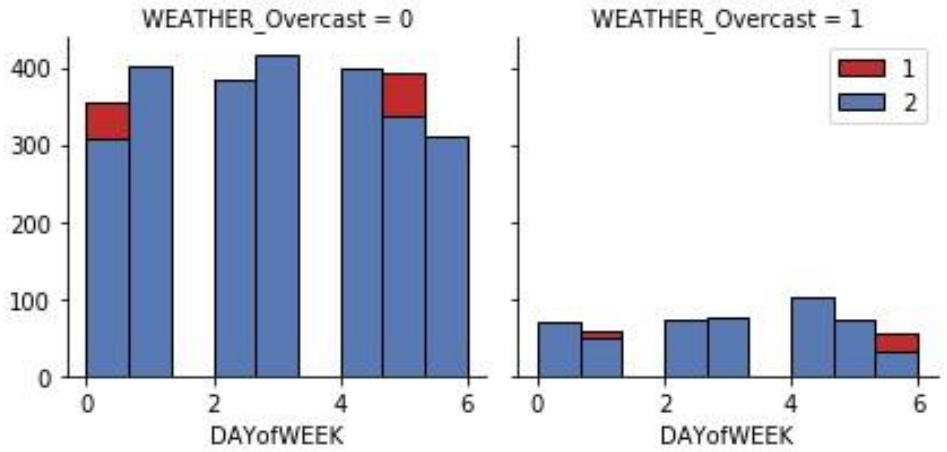

Figure 1: Relationship between weather condition (overcast) and accident severity.

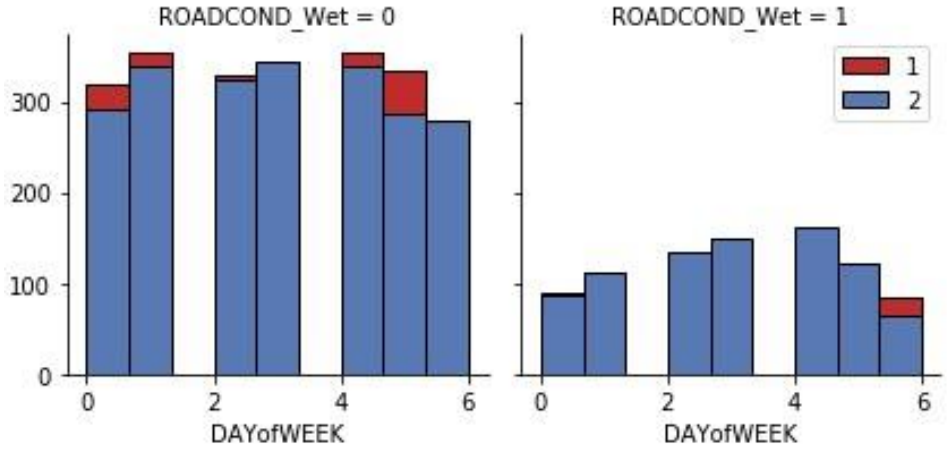

Figure 2: Relationship between road condition (wet) and accident severity.

Figure 3 shows that when the light condition is daylight, there are more accidents compared to when it is not. In addition, all of collisions under this condition seem to result in severity code 2 (injury).

Figure 4 shows that when the junction type is an intersection, there are less accidents compared to when it is not and all the accident collisions experience severity code 2 (injury).

Figure 5 shows that when the number of vehicles involved in a collision is 3, there are much fewer accidents compared to when it is not and all the accident collisions experience severity code 2 (injury).

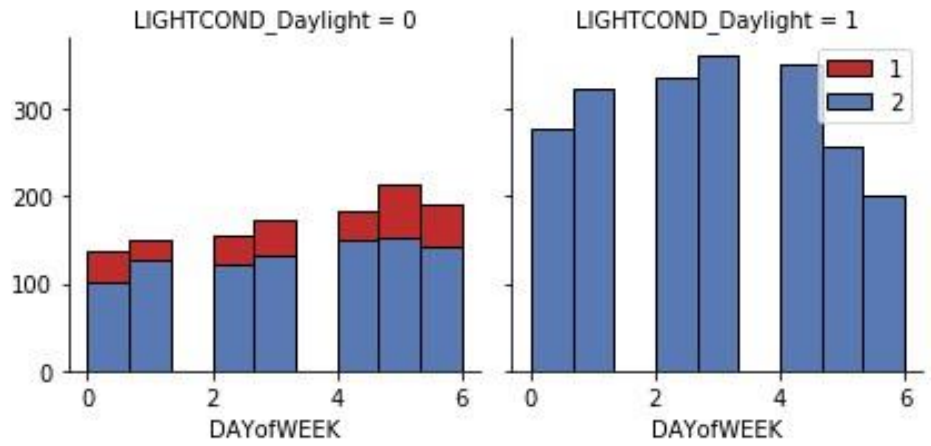

Figure 3: Relationship between light condition (daylight) and accident severity.

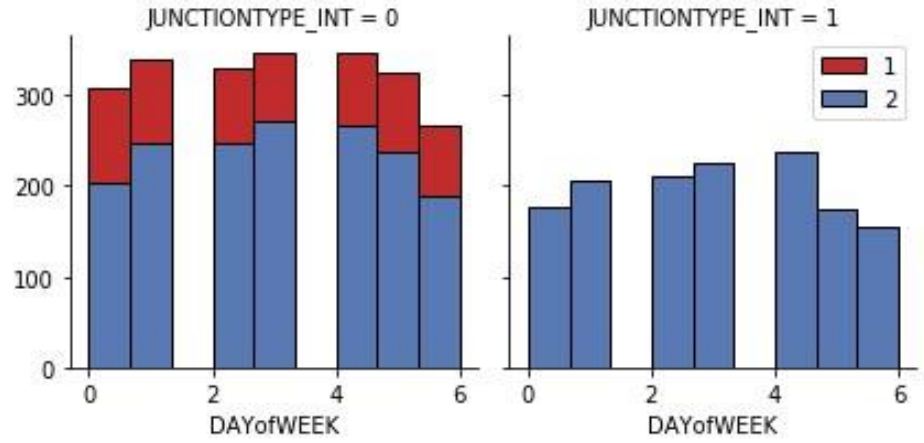

Figure 4: Relationship between junction type (intersection) and accident severity.

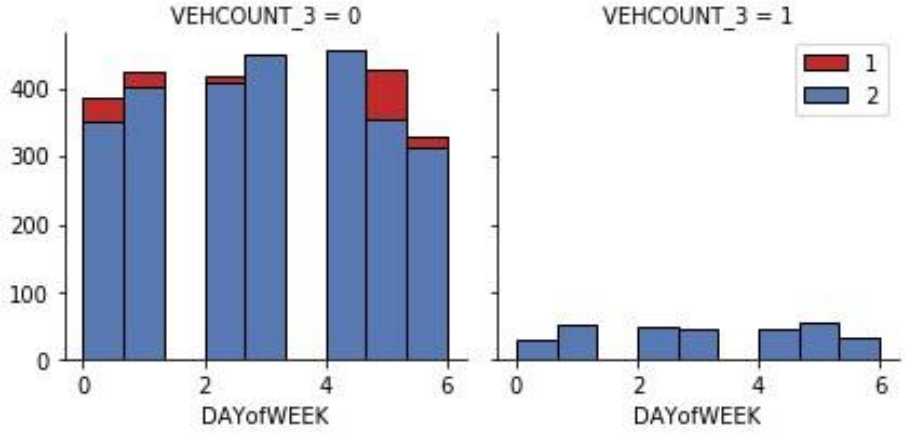

Figure 5: Relationship between vehicle count (3) and accident severity.

4. Results

This section presents the result of the machine learning classification algorithms considered namely:

- Support Vector Machine (SVM)
- Decision Tree
- K Nearest Neighbor (KNN)
- Logistic Regression

For accuracy evaluation, we considered Jaccard and Log-loss (where applicable) evaluation scoring metrics. Table 2 summarizes the accuracy of the developed models. It is worth noting that the algorithms were subjected to the same dataset (including random seeding) so that the comparison in the accuracy becomes fair. It is also worth mentioning that for the SVM, the selected kernel was the radial basis function (rbf); for the decision tree, the maximum depth was set to 4; and for the KNN, the best best k was firstly determined before building the model.

Table 2. Algorithms and evaluation metrics.

ALGORITHM	JACCARD	LOG-LOSS
SVM	0.6037	0.6580
Decision Tree	0.6058	NA
KNN	0.5832	NA
Logistic Regression	0.6188	0.6397

5. Discussion

As depicted in Table 2, the four algorithms perform similar, with KNN having a slightly lower Jaccard-similarity score of 0.5832. Logistic Regression had the best performance among all models when considering the Jaccard score.

It is worth noting that the as aforementioned in Section 2, the original dataset was imbalanced and as part of the pre-processing, down-sampling technique was applied to avoid biasing the model to a majority class. If the reader is not interested in performing such processing, it is recommended to use penalized SVM and Decision Tree algorithms which tend to be robust when dealing with imbalanced data.

6. Conclusion

In this capstone project, we analysed the car accident severity based on the SDOT data-collisions dataset and considering key factors that might have contributed to the accident. The selected features in this projected included weather conditions, road conditions, light conditions, junction type, number of vehicles involved in collision, and date & time of the incident. Data visualization was explored to give high-level insights onto how these factors contributed to the severity of the incident. In addition, predictive models based on well-known machine learning algorithms we developed to predict the severity of the accident. The built models were also subjected to unseen test data to evaluate their accuracy. Future works may include improving these models further by e.g., tuning and optimising hyper-parameters, varying features as well as applying other algorithms and different performance metrics.

7. References

[1] Coursera, [Online]. Available: https://www.coursera.org/learn/applied-data-science-capstone/supplement/Nh5uS/downloading-example-dataset. [Accessed October 2020].