
Car Accident Severity
Seattle City, USA

Overview

Introduction

Background and Business Problem

Data

Collection

Cleaning

Features Selection

Methodology

Further preprocessing

Exploratory Analysis

Model Building & Evaluation

Conclusion

Introduction (Background and Business Problem)

- Seattle has recently been experiencing an increase in the number of road accidents.
- As such, to align with long-term development vision of the City, key stakeholders within the Seattle Department of Transport (SDOT) want detailed insights on the car accident severity considering key factors that might have contributed to the accident.
- They have tasked a data scientist team to solve this problem.
- 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.

Data

- The readily available shared ‘Data-Collision.csv’ dataset (2004-2019) from SDOT will be used.
- The dataset has 38 attributes and 194,673 observations.
- Features comprising 9 independent numeric and categorial attributes were selected for the analysis.

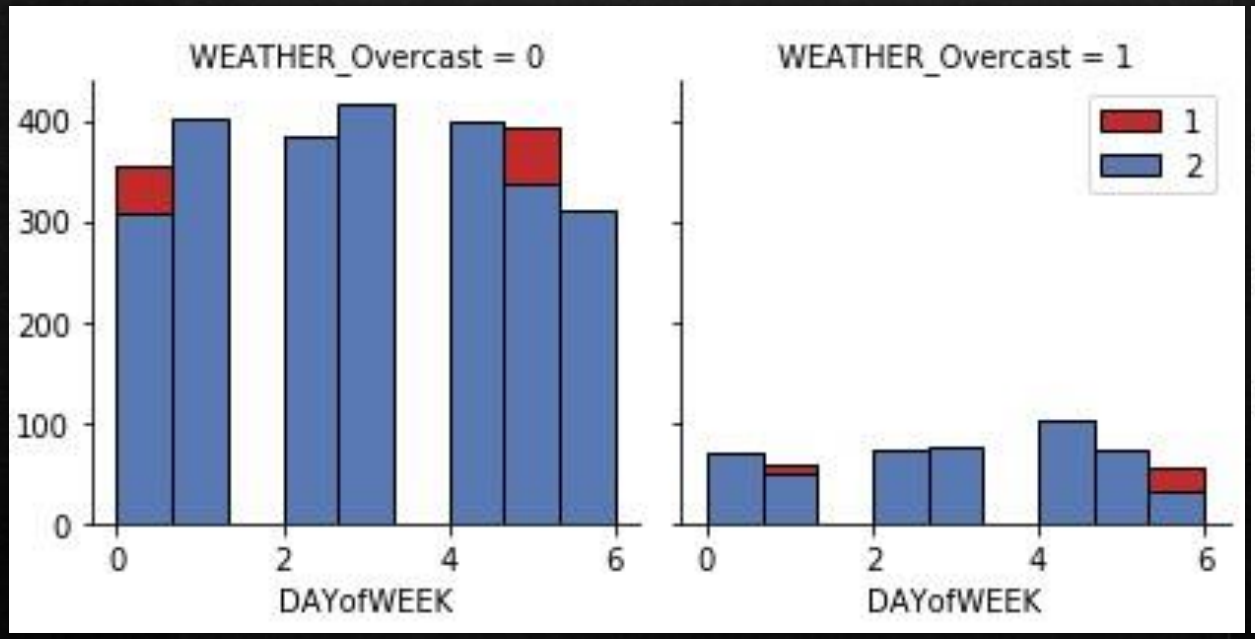
Methodology: Further - Pre-processing

- Handling missing data.
- Handling imbalanced classes (down-sample majority class).
- Handling categorial variables with many-levels.
- Descriptive statistical analysis.
- The cleaned data contained 20 features (including dummy variables).

Methodology: Exploratory Data Analysis

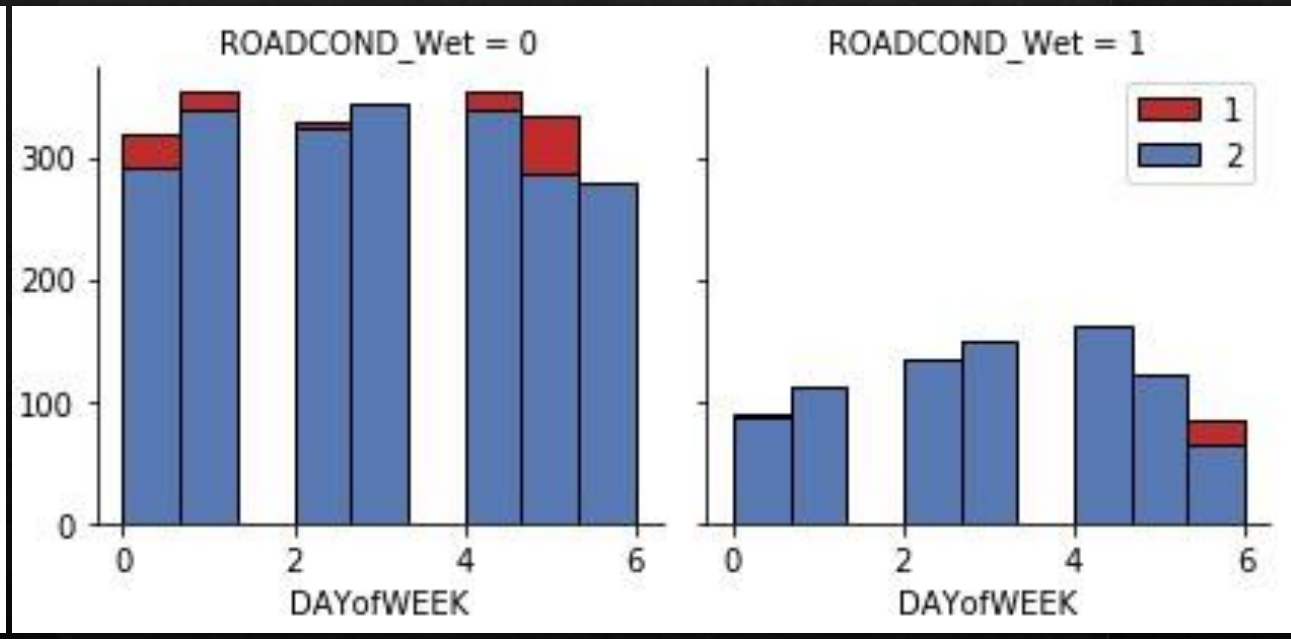
Weather Condition and Severity

(overcast, code 1 occurs more on weekends)



Road Condition and Severity

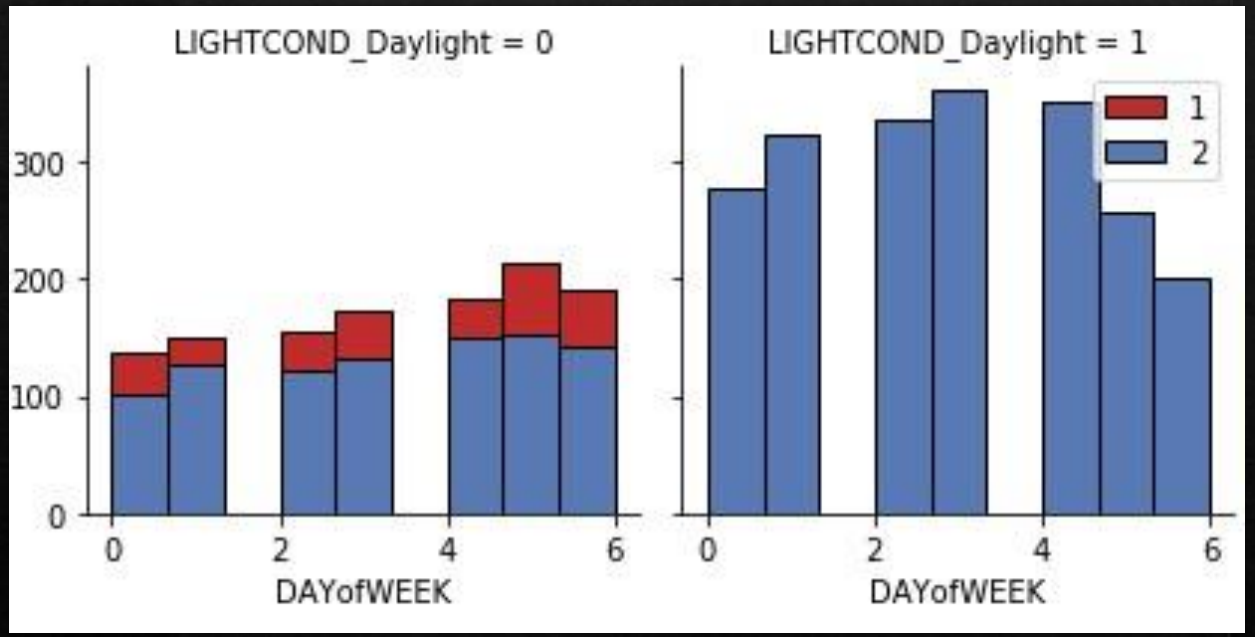
(wet, code 2 occurs more than code 1)



Methodology: Exploratory Data Analysis

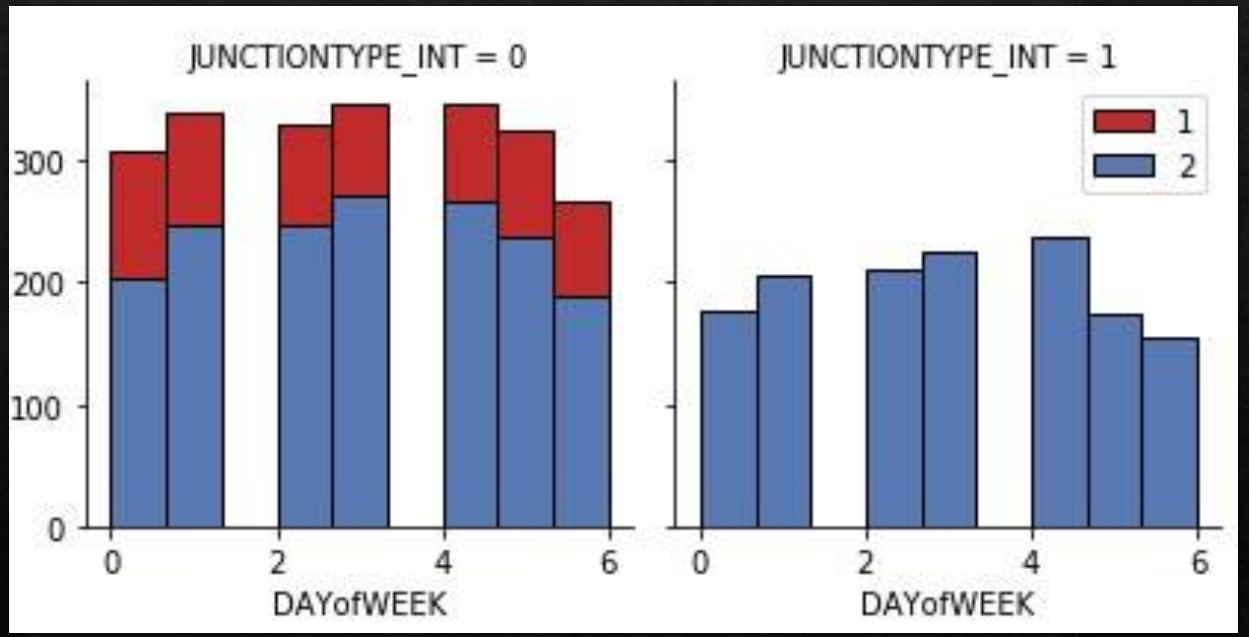
Light Condition and Severity

(daylight, lots of code 2)



Junction Type and Severity

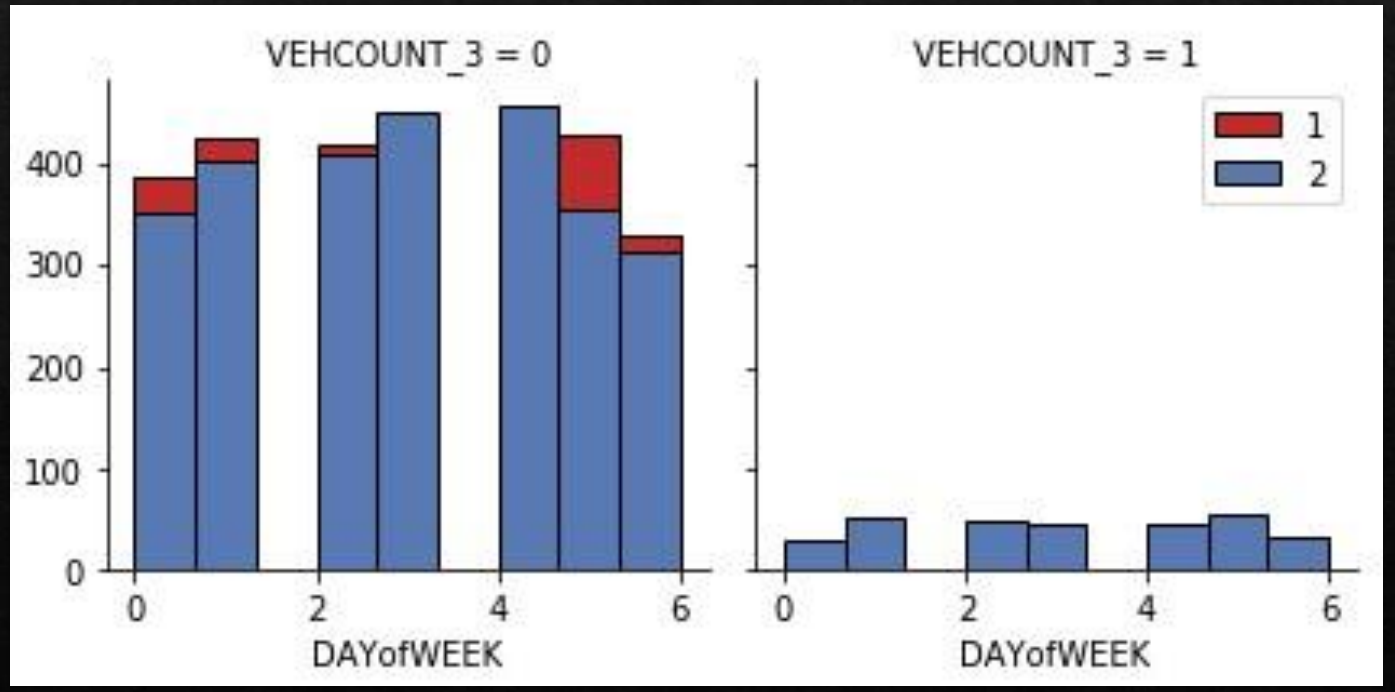
(intersection, lots of code 2)



Methodology: Exploratory Data Analysis

Vehicle Count and Severity

(3 vehicles, less impact but more of code 2)



Methodology: Build Model & Evaluation

- Classification Algorithms:
  - Support Vector Machine (SVM): kernel – radial basis function (rbf)
  - Decision Tree: max-depth – 4
  - K Nearest Neighbor (KNN) – best k = 5
  - Logistic Regression

Methodology: Model Evaluation

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

Conclusion

- Analysed the car accident severity based on the SDOT data-collisions dataset and considering key factors that might have contributed to the accident.
- Performed exploratory data analysis to gain insights.
- Built 4 models to predict the accident severity (1-property damage, 2-injury).
- Evaluated the accuracy of the models using unseen test data.
- Future works to study the impact of tuning hyper-parameters, other features, different algorithms and consequently improving accuracy.