Applied Data Science Capstone Project

Car accident severity

Author: Sergio Esposito
September 8th, 2020

Index

Introduction

In this project I intend to build a machine learning model capable of predicting the severity of a car accident based on some variables like weather, road conditions and other relevant variables.

My project could result interesting to:

- Insurance companies
- Transport/road traffic government agencies
- Car drivers

Data

I have used a dataset which compiles car accidents in the city of Seattle. The dataset has been provided by Coursera. Follow this link to see the metadata of the dataset https://drive.google.com/file/d/1uPOlUo2qaF-WYOYq5UljADLEqw1GGfwM/view?usp=sharing

The variable to predict is SEVERITYCODE

After analyzing the original dataset, I have decided not to use al the attributes, but the following:
ADDRTYPE
COLLISIONTYPE
PERSONCOUNT
VEHCOUNT
JUNCTIONTYPE
WEATHER
ROADCOND
LIGHTCOND
ST_COLCODE

I have applied data improvement techniques: clean rows with NaN values, convert object attributes to numeric values, convert the variable to predict to descriptive values, and balance the dataset.

Methodology

As the variable to predict has a discrete set of values, I have chosen to work with classification techniques. So, I have tested three classification techniques(SVM, Logistic Regression and Decision Tree with a subset of the data (20,000 rows).

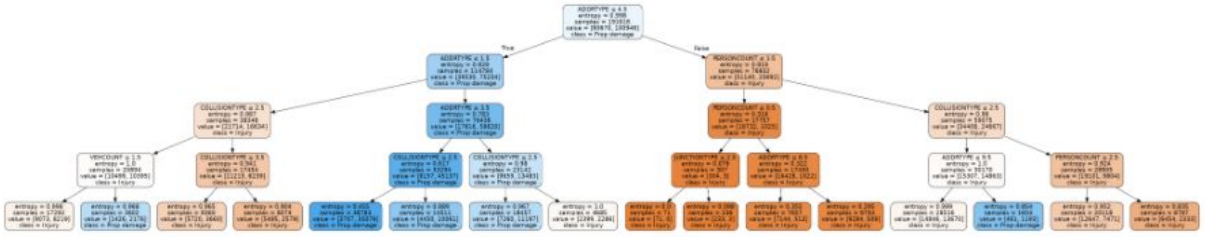
I have run the three techniques and compared the results:

- Logistic regression has provided jaccard_similarity_score=0.68625
- SVM has provided jaccard_similarity_score=0.73475
- Decision Tree has provided accuracy_score=0.7465
- SVM performance is worse than decision tree performance (around 10 times slower)

Having seen above premises I have chosen Decision Tree as my model

Results

My model is a Decision Tree whose visual representation can be seen below. More details of this representation can be seen in my Jupyter Notebook. Having run the model with the whole balanced dataset (almost 240,000 rows) it has been capable of predicting the severity of a car accident with a 69.3% of accuracy



Discussion

Whoever is interested in continuing this line of research could try other classification techniques like Random Forest and Gradient Boost. Also, Voting Ensemble could be used to combine the results of different models.

Conclusion

I have built a machine learning model, which using as input an accurate dataset with car accidents data can predict the severity of car accidents. This model can be useful to save

lives, improve transportation planning, and smart use of financial and infrastructure resources