Machine Learning as a tool to predict the severity of an accident: Case Study from Seattle (WA) 2004-2020 record

Dr. Daniel Rodelli
22th November, 2020

Introduction

- Car accidents are the leading cause of death among 5-29 year old people
- Huge direct and indirect economic impact
- Numerous **quantificable** factors affect the gravity of accidents
- **Machine Learning** is suited as this is a scientific approach for modelling and predicting the **severity** of an accident
- **Apply a machine learning to predict the severity of the accident using car collision data for the city of Seattle, USA.**

Business Plan

- A scenario where a witness gives information through an emergency call
- Which information is fundamental?
- Which information can a untrained witness observe and comunicate in a very short amount of time? → *Limited number of variables. Advantage: Quicker call and quicker response. Disadvantage: intrisically lower accuracy*
- How this information can be translated into a severity description of the accident

The final goal:

- **How emergency service can use this information to better manage assets (vehicles, personnel, hospitals)**

Methods

- Dataset exploration
- Dataset cleaning
- Algorithms:
  - K-Nearest Neighbors
  - Logistic Regression
  - Decision Tree
  - Random Forest
- Algorithms comparison

FEATURES:

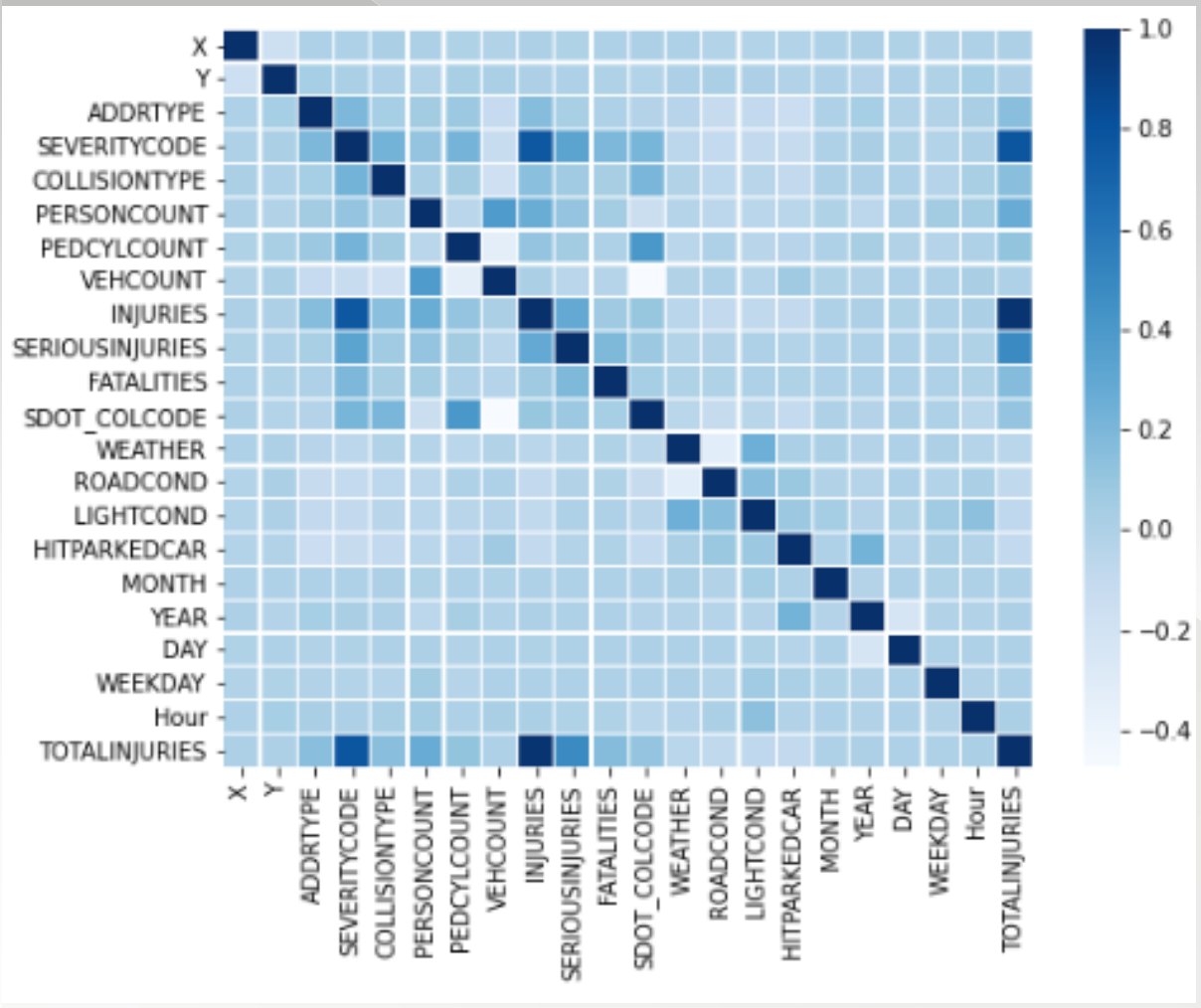
- Injuries - *fundamental*
- Vehicle Count
- Collision Code
- Collision Type
- X –Y coordinates - *fundamental*
- Weather
- Road Condition
- Light Condition
- Address Type
- Hour

TARGET:

Severity Code

CLASS	DESCRIPTION
1	**Accident with only property damage**
2	**Accident with injuries**
3	**Accident with severe injuries**
4	**Accident with fatalities**

Correlation Matrix

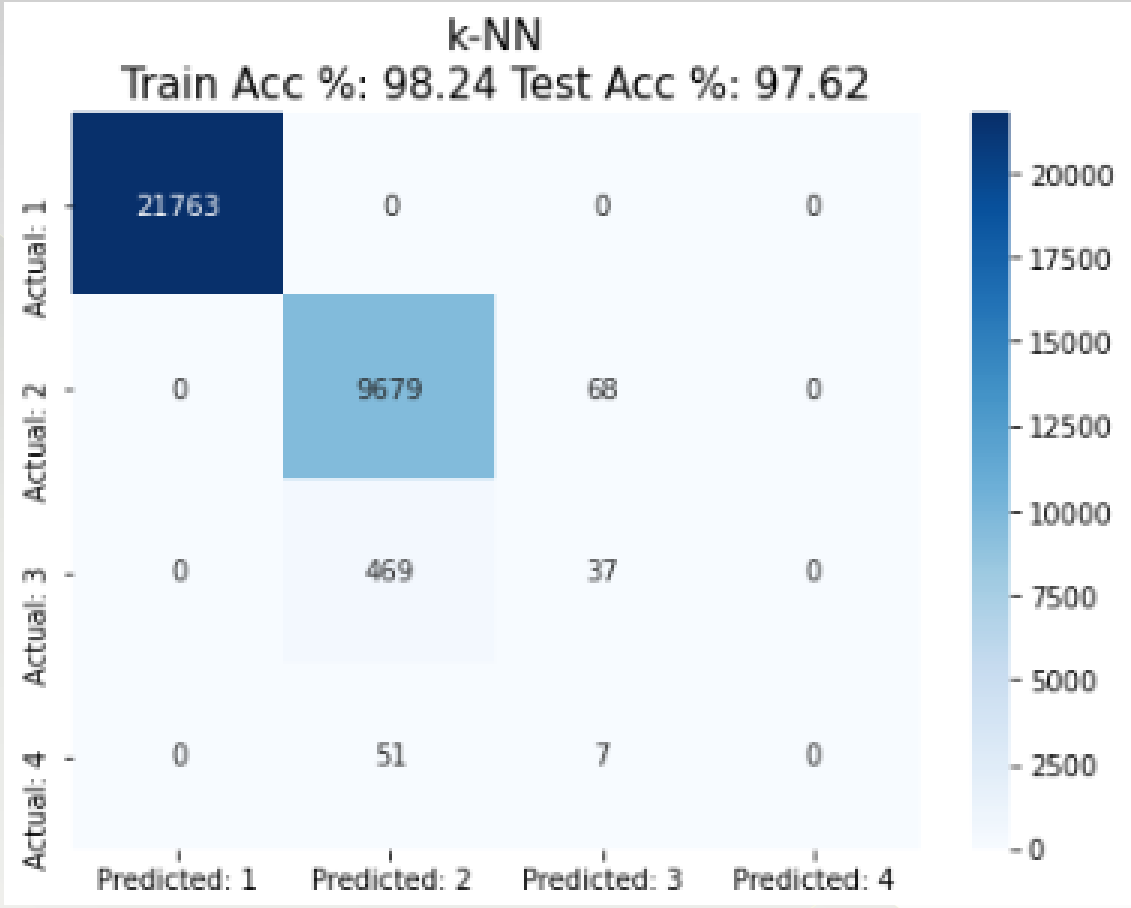


Very strong correlation between Severity Code and Injuries

MODELS

- **K-Nearest Neighbors**

The best accuracy (**0.9774**) was reached with a **K=5**



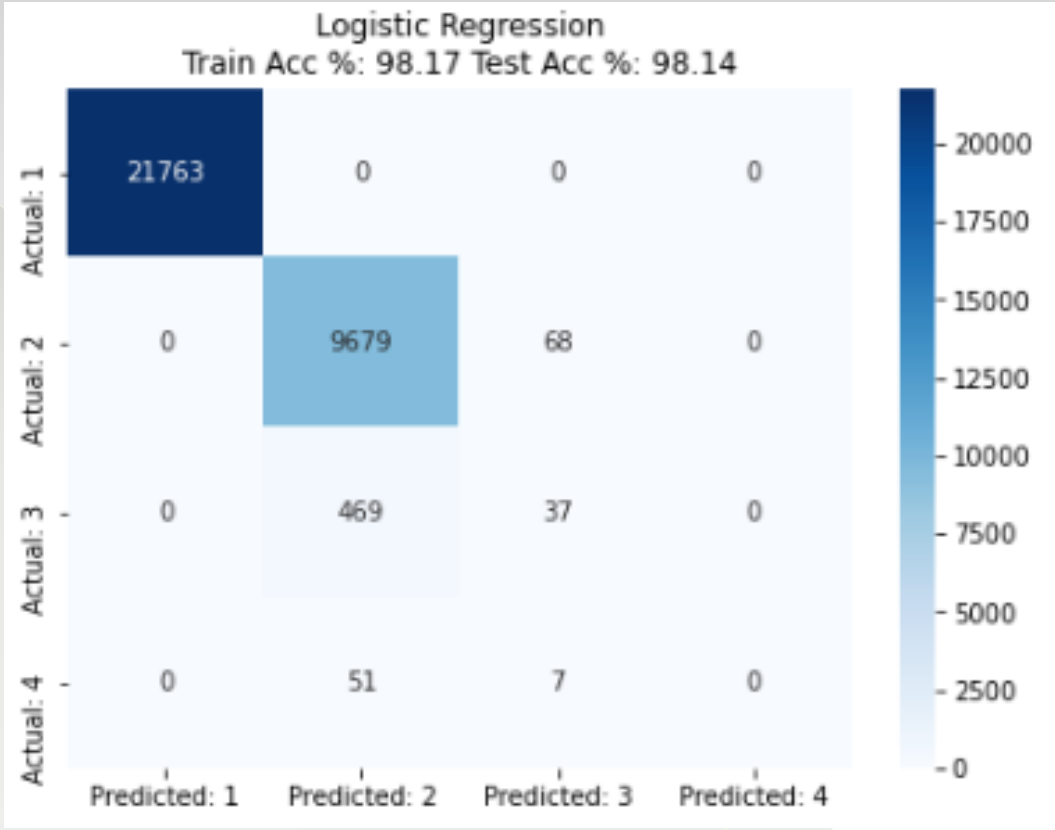
MODELS

- **Logistic Regression**

Best accuracy : **0.981**

C=**5**

Max iter**=1000**

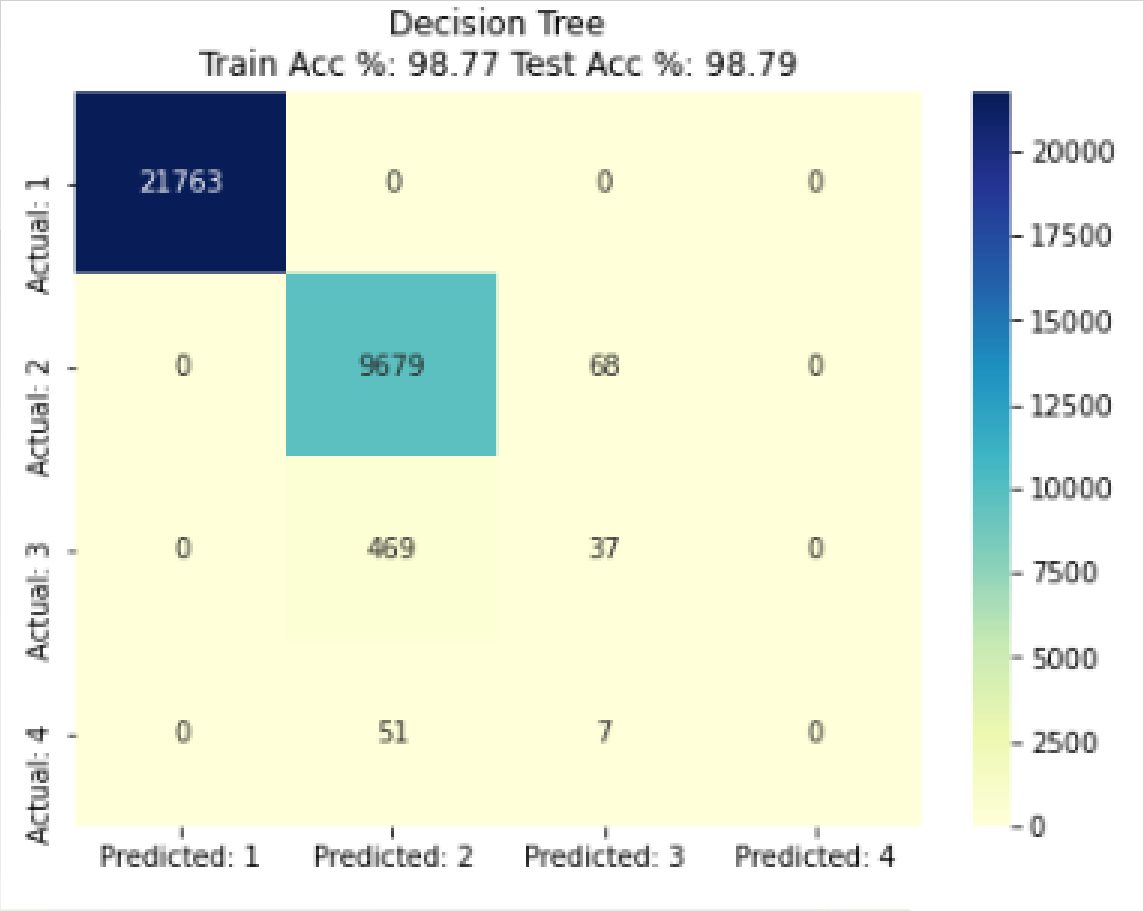


MODELS

- **Decision Tree**

Accuracy= **0.987**

Depth **= 5**

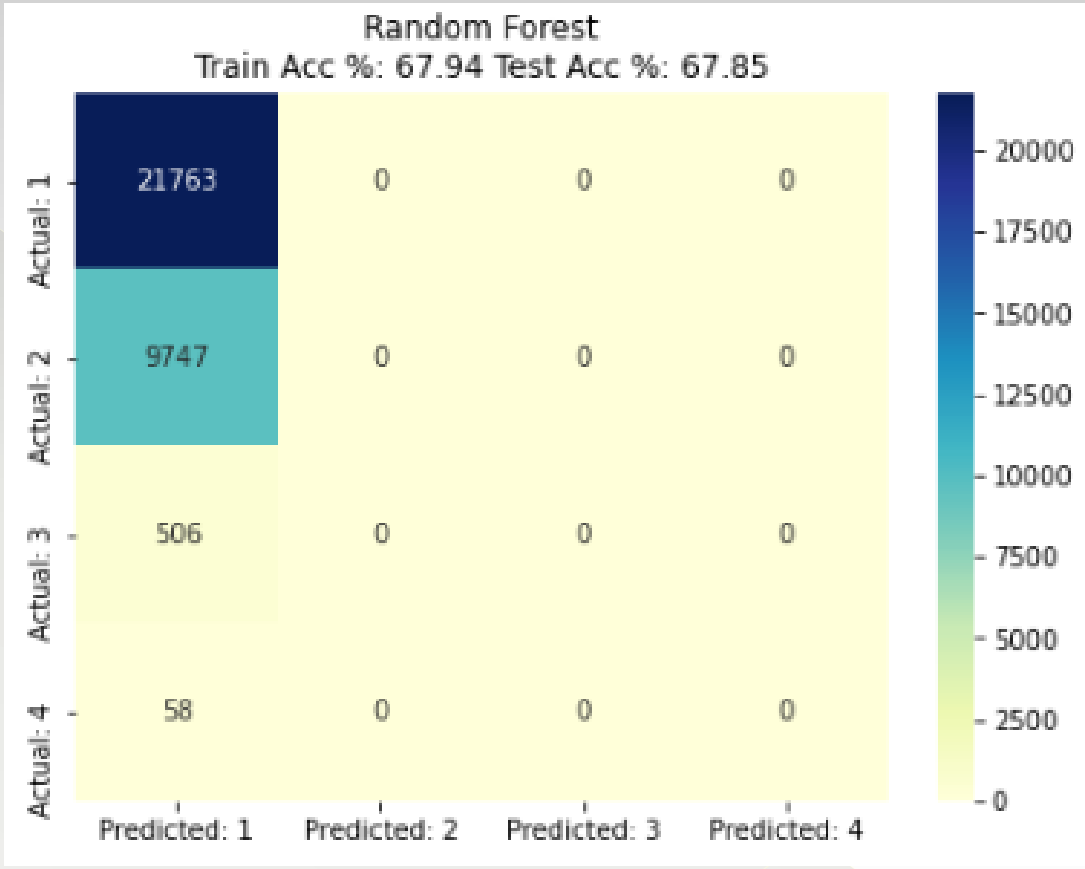


MODELS

- **Random Forest**

Accuracy = **0.679**

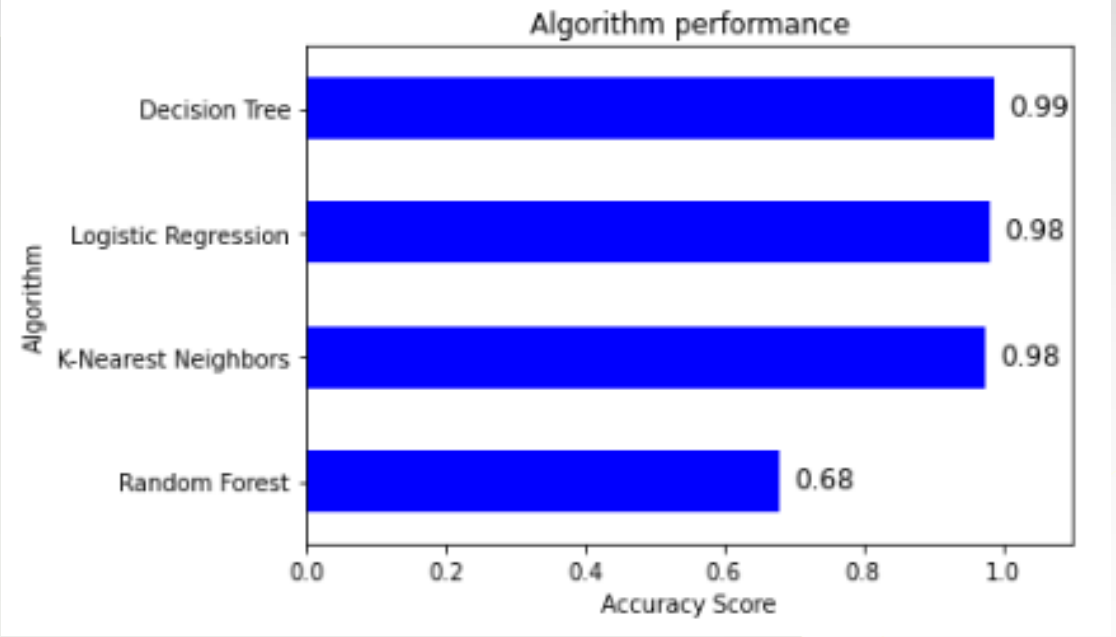
Accuracy too low
DISCARDED



MODELS COMPARISON

- **The model selected is Decision Tree**

	Algorithm	Accuracy_Score
3	Random Forest	0.678525
0	K-Nearest Neighbors	0.976242
1	Logistic Regression	0.981449
2	Decision Tree	0.987934



FEATURE IMPORTANCE

- **The model selected is Decision Tree**

The presence of **injuries** and the number of **vehicles** are the dominant features of the model

	FEATURE	IMPORTANCE
1	TOTALINJURIES	0.982
2	VEHCOUNT	0.015
3	ST_COLCODE	0.02
4	COLLISIONTYPE	0.01
5	X	0.00
6	Y	0.00
7	WEATHER	0.00
8	ROADCOND	0.00
9	LIGHTCOND	0.00
10	ADDRTYPE	0.00
11	HOUR	0.00

Discussing the results

- presence of **injuries** and the number of **vehicles** are the dominant features
- Location of an accident is also fundamental in a real case scenario
- 100% accuracy on Severity Code 1
- Slight underestimating the severty of the injuries
- →room to improvement with more features?

- **Low number of variables to inform → quicker call → quicker response → more lifes saved**

Conclusions

- Decision Tree model best predicts the severity of accidents . Accuracy: 98.7%
- Limiting by design the number of features affected the accuracy for injury severity
- Only 2 features are really necessary for the model (plus location, of course...)
- **LOW NUMBER OF VARIABLES MEANS QUICKER CALLS, QUICKER RESPONSE, MORE TIME TO SAVE LIVES.**