Development of a Predictive Model for Automobile Collision Severity

**Applied Data Science Capstone by IBM/Coursera**

**BY -SHRAMAN JAIN**

**11th September,2020**

Contents

1. Introduction
2. Data
3. Methodology
4. Results
5. Discussion
6. Conclusion

INTRODUCTION

**CONVERTING DATA TO ACTION**

• Data only useful if it can be converted to actions

• Automobile data readily available for analysis and interpretation

• Empower improved decision making to save lives

**READY FOR INTEGRATION**

• Real time data can be readily interpreted into risk

• Exciting results based on first analysis attempts

• Many examples of how business partners can use this technology

DATA

COLLISION DATA

• Available from Seattle Department of Transportation (SDOT)

• Covers 2004 to 2020 • Nearly 200,000 Records

• Many Attributes Available

• Collision Details

• Road and Weather Conditions

• Geographic Information

• Environmental Variables

• Other Ancillary Information

**METHODOLOGY 1**

Collision data available, broken into injury vs. non-injury events. Nearly 40 associated attributes available for analysis, so it was necessary to determine what to use and what to omit

**Data to Exclude from Analysis**

* Database organization fields and keys
* Geographic location data
* Collision details (damages, event descriptions)

**Data to Include in Analysis**

* Address Type

• Weather

• Road Conditions

• Light Conditions

• Day of Week

**METHODOLOGY 2**

**DATA PREPARATION**

• One hot encoding translation for categorical variables

• Scaled results to prevent outsized influences on prediction

**SUPERVISED MACHINE LEARNING**

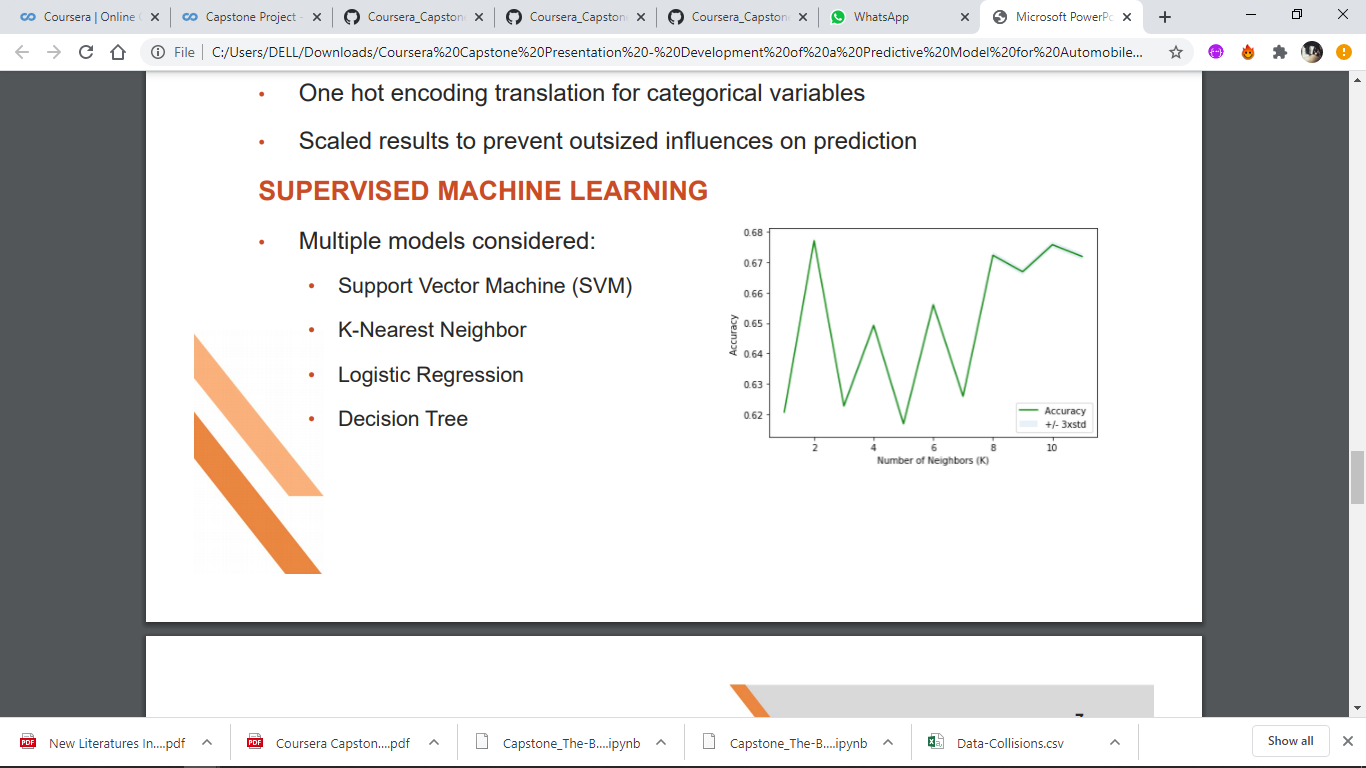
• Multiple models considered:

• Support Vector Machine (SVM)

• K-Nearest Neighbour

• Logistic Regression

• Decision Tree



**RESULTS**

**LIMITING TRAINING DATA SIZE**

* Utilizing entire data set led to long processing times
* Limiting to 10% provided ample training data while keeping processing times reasonable

**ACCURACY BY MACHINE LEARNING MODEL**

* All four models able to provide predictions
* Lowest accuracy with Support Vector Machine
* Highest accuracy with Logistic Regression

**Model vs Accuracy**

* **SVM – 0.661**
* **KNN – 0.676**
* **Logistic Regression - 0.701**
* **Decision Tree – 0.693**

**CONCLUSION**

• Model developed from 200,000 collision records

• 70% injury prediction accuracy with Logistic Regression model

• Processing times fast enough to process live data feeds