# Development of a Predictive Model for Automobile Collision Severity Coursera Applied Data Science Capstone Project

September 1, 2020 Evan Franke

# Contents

Introduction

Data

Methodology

Results

Discussion

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**



SDOT Collision Code Matrix

	DIR.  Direction of trovel prior to collision  Directional cudes = 1	DIR.  Direction of trovel or impact  Directional codes =   One of trovel or impact  One of trove	COLLISION					
			STRIKING					
MOTOR VEHICLE (including TRAINS) and PEDALCYCLIST IN TRAFFIC			Histor Whicle In Operation	Driveriess Parke Veticie	Pedalcyclist in Traffic			
			10 13 13 14 14	30 31 33 33 34 34	50 31 42 33 34 34 44	STRUCK MOTOR VEHICLE	HEAD ON IN TRUNT END (oil head on) I FUNT END AT APPOLE IN EAST SUDE AT APPOLE IN EAST SUDE IN E	
			17 10 19 28 21 21 23 27	37 28 29 80 41 41 43	17 18 18 18 18 18 18 18	SIFUCE PEDALCYCLIST	HEAD-ON IN FRONT CIND IN ROOM SIND IN ROOM SIND IN REPT SIDE AT ANGLE IN REAL FIND IN ROOM SIDE SIDES OF ROOM IN ROOM IN ROOM SIDES OF ROOM IN ROOM IN ROOM SIDES OF ROOM IN ROOM SI	
			24 31 76	45 45	44 24 44	STRUCK PEDESTRIAN TRAIN OBJECT IN ROAD	ADWAY (Periodes curls, jersey barriers &	
			10	2	41	Anti Off SQuerror 14	O COLLISION IT FIXED CALECT - (includes conduits & condu	
			79	"	#V	ORDERORSO IN SCHOOL (1904-(1971-1941)		
			STRUCK		Note: 2nd Call Code Only			
			Hater Vehicle In Operation	Driveriess Minter Vehicle	Pedalcyclist in Traffic	Note: Zhi	d Cell Code Only	
			blank 01 84 84	62 06 A8 88	blask 70 88 84	STRUCK BY OTHER MOTOR VEHICLE  PROSESTIAN  TRANSCENSION HULLIN TRANSCENSION  TRANSCENSION		
PEDESTRIAN and PEDALCYCLIST NOT IN TRAFFIC	Direction of travel	Leg of intersection	STRUCK  OF OFF ROADWAY OF SECURES OF FAMILIES STREET  OF OFF ROADWAY OFF SECURES OF FAMILIES STREET  OF OFF SECURES OFF SECURES OFF SECURES STREET  OFF SECURES OF					
	0 = direction unk.		09	-		OJWN .		
	in directional codes	I thru 8 or 9 = non intersection 0 = middle of int.	STRUCK OF STRIKING  TO CROSSING - NOW INTERESTED THE S WASK  TO CROSSING - NOW INTERESTED THE STRIKE TO SHARE THE STRICT TO SHARE THE STRICT TH					
PEDESTRIAN only	applicable directional codes	applicable leg codes	STRUCI	BDADWAY WA	MIKING WITH THE AGAINST OF THE AGAIN	VEHICLE ON TEST THE 14	olari odali	

### **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
  - · Alley, Block or Intersection
- Weather
  - · Fog, Snow, Wind, Rain, etc.
- Road Conditions
  - Dry, Oil, Ice, Wet, Slush
- Light Conditions
  - Dawn, Daylight, Dusk, Dark
- Day of Week
  - Commuting and Traffic Congestion

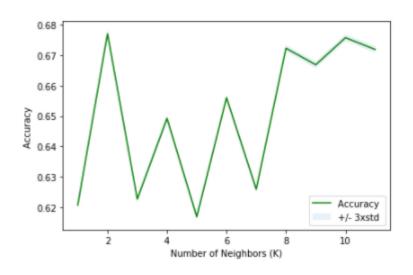
# **METHODOLOGY 2**

### **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 Neighbor
  - 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

Machine Learning Model	Test Set Accuracy
Support Vector Machine (SVM)	0.661
K-Nearest Neighbor (KNN)	0.676
<b>Logistic Regression</b>	0.701
Decision Tree	0.693

# **DISCUSSION**

<b>Prom</b>	isi	ng		
Resu	ts	Ok	otai	ned

Machine Learning able to predict an injury collision with 70% accuracy.

# **Tool May Reduce Accident Risk by:**

Monitoring subset of environmental factors from live sources

Supplying data to processing algorithm for risk calculation

Providing an alarm to driver when collision likelihood rises

# Further Possibilities for Improvement

Determine addition data fields for improved accuracy

Provide additional instructions for alternate safer routes

# CONCLUSION

# ACCURATE PREDICTIVE MODEL ACHIEVED

### **HIGHLIGHTS**

- Model developed from 200,000 collision records
- 70% injury prediction accuracy with Logistic Regression model
- Processing times fast enough to process live data feeds

### **NEXT STEPS**

- Identify live data source partners for software integration
- Identify hardware partners for application or hardware solution
- Improve public safety by integrating this solution