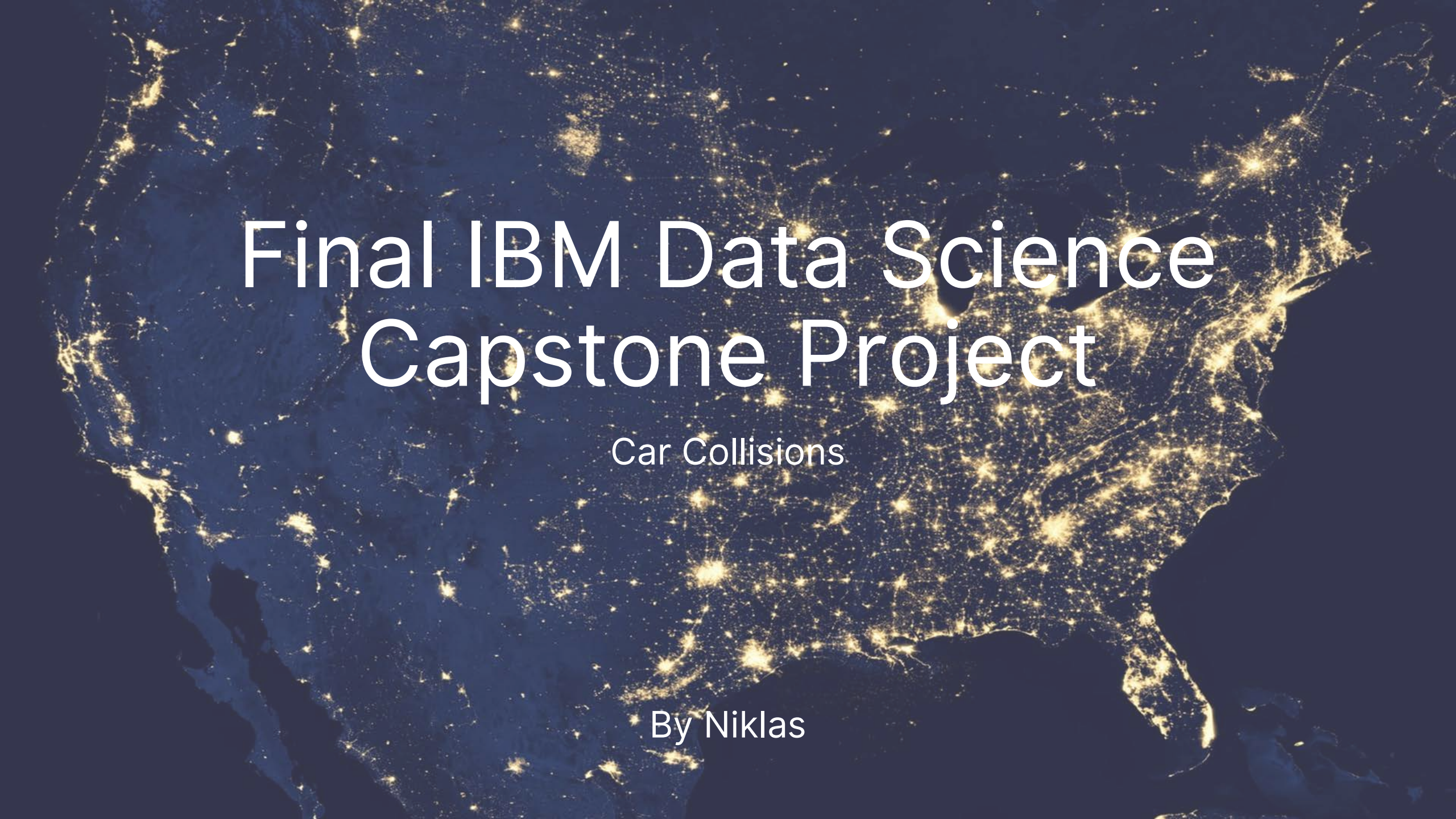


Final IBM Data Science Capstone Project

Car Collisions

By Niklas

Introduction

Vehicle accidents can have tremendous impacts on a person's life - both from a health perspective but also on a financial one. On those, multiple factors like traffic, rain, snow, wind, or road conditions can determine the likeliness and severity of an accident.

- *Can it be possible to determine and quantify the risk even before starting the engine by knowing certain conditions that can be expected during the ride?*

Lets find out!

Dataset

A Collision data set was used that included information about almost 200,000 accidents recorded in Seattle, WA, since the year 2004. Overall, the data set contained various variables with possible influence on the target variable, the **SEVERITY** of an accident. These predictor variables included the following:

- *WEATHER*
- *ROADCONDITION*
- *LIGHTCONDITION*
- *ADDRESS-TYPE*

Methodology

1. Data Analysis

2. Machine Learning Algorithms

2.1 KNN

2.2 Decision Tree

2.3 SVM

2.4 Linear Regression

Data Analysis – Part 1

Address

ADDRTYPE	Alley	Block	Intersection	All
Injury	0.892183	0.761363	0.568655	0.696664
Property Damage	0.107817	0.238637	0.431345	0.303336
Total Column	1.000000	1.000000	1.000000	1.000000

Percentage of Property damage collisions from Total

Percentage of Injury collisions from Total

Total

Weather

WEATHER	Blowing Sand/Dirt	Clear	Fog/Smog/Smoke	Other	Overcast	Partly Cloudy	Raining	Severe Crosswind	Sleet/Hail/Freezing Rain	Snowing	Unknown	All
Injury	0.734694	0.67673	0.669627	0.855696	0.683678	0.4	0.662179	0.72	0.758929	0.813616	0.943808	0.696664
Property Damage	0.265306	0.32327	0.330373	0.144304	0.316322	0.6	0.337821	0.28	0.241071	0.186384	0.056192	0.303336
Total Column	1.000000	1.00000	1.000000	1.000000	1.000000	1.0	1.000000	1.00	1.000000	1.000000	1.000000	1.000000

Data Analysis – Part 2

Road-Conditions

ROADCOND	Dry	Ice	Oil	Other	Sand/Mud/Dirt	Snow/Slush	Standing Water	Unknown	Wet	All
Injury	0.677531	0.77368	0.625	0.66129	0.69863	0.833669	0.738739	0.947891	0.667535	0.696664
Property Damage	0.322469	0.22632	0.375	0.33871	0.30137	0.166331	0.261261	0.052109	0.332465	0.303336
Total Column	1.000000	1.00000	1.000	1.00000	1.00000	1.000000	1.000000	1.000000	1.000000	1.000000

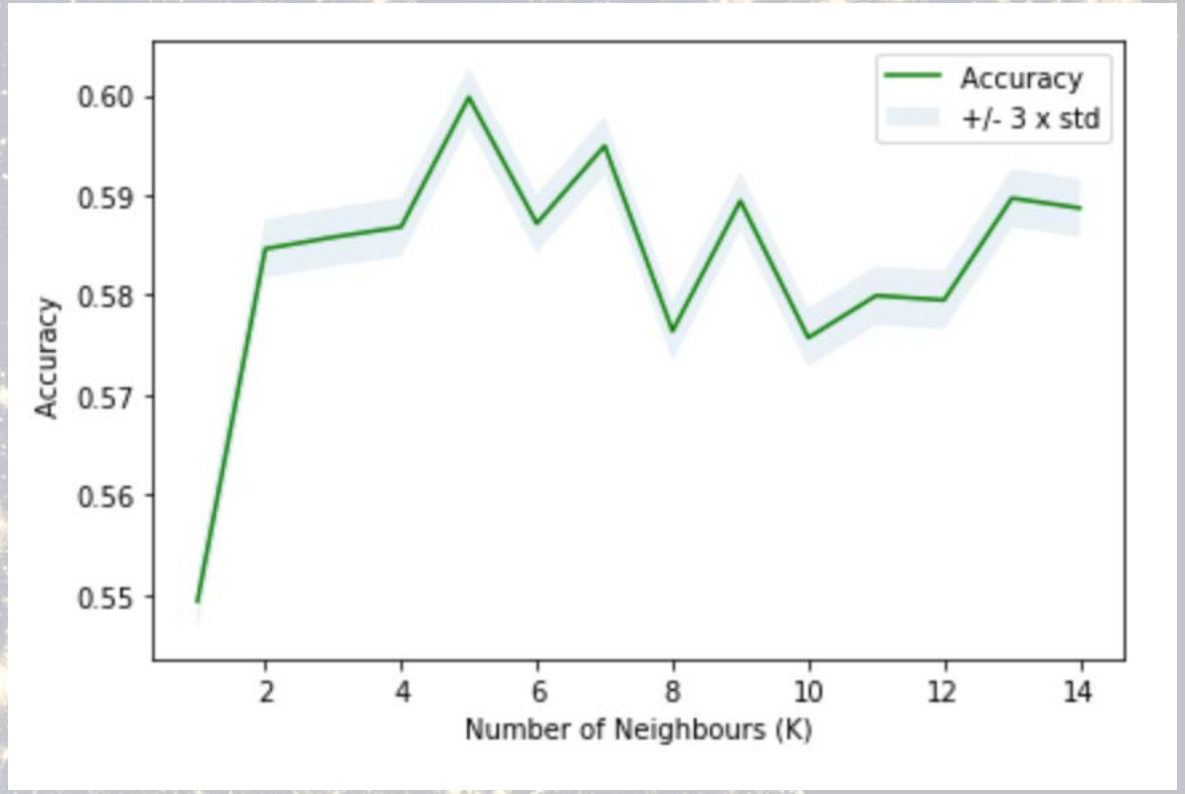
Light-Conditions

LIGHTCOND	Dark - No Street Lights	Dark - Street Lights Off	Dark - Street Lights On	Dark - Unknown Lighting	Dawn	Daylight	Dusk	Other	Unknown	All
Injury	0.781127	0.733953	0.701053	0.636364	0.669611	0.667181	0.668663	0.770925	0.95325	0.696664
Property Damage	0.218873	0.266047	0.298947	0.363636	0.330389	0.332819	0.331337	0.229075	0.04675	0.303336
Total Column	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.00000	1.000000

ML Algorithms Results - KNN

Graph for the identification of the optimal K-value for the KNN

Result: K=5



ML Algorithms Results - Overall

Evaluation of the four ML algorithms for the Jaccard and F1-Score (LogLoss for LR)

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.588707	0.580611	N/A
Decision Tree	0.605548	0.599037	N/A
SVM	0.605618	0.598745	N/A
Logistic Regression	0.604212	0.598791	0.661124

Discussion

This report has to be viewed under consideration of several potential weaknesses whose influence could be investigated going further:

- "Other" and "Unknown" values can be filtered out of the dataset
- Other variables could have a strong influence, such as alcohol and tiredness level of the driver
- Only two different severity types were investigated, potentially there are more
- From those investigated, the injuries can also have involved pedestrians or cyclists who where not considered

Conclusion – Part 1

At first, certain conditions increase or decrease the injury risk related to a collision:

Lower Injury Risk

- At Intersections
- Clear, rainy or foggy weather
- Dry, wet, and oily roads

Higher Injury Risk

- At Alleys, Blocks
- Storms, Hail, Snowing, Blowing Sand & Dirt
- Icy, snowy, and watered roads
- Dark light conditions

Conclusion – Part 2

For determining further severity of accidents, two algorithms are proposed to use to evaluate certain road, address, weather, and light conditions:

- Decision Tree
- Linear Regression