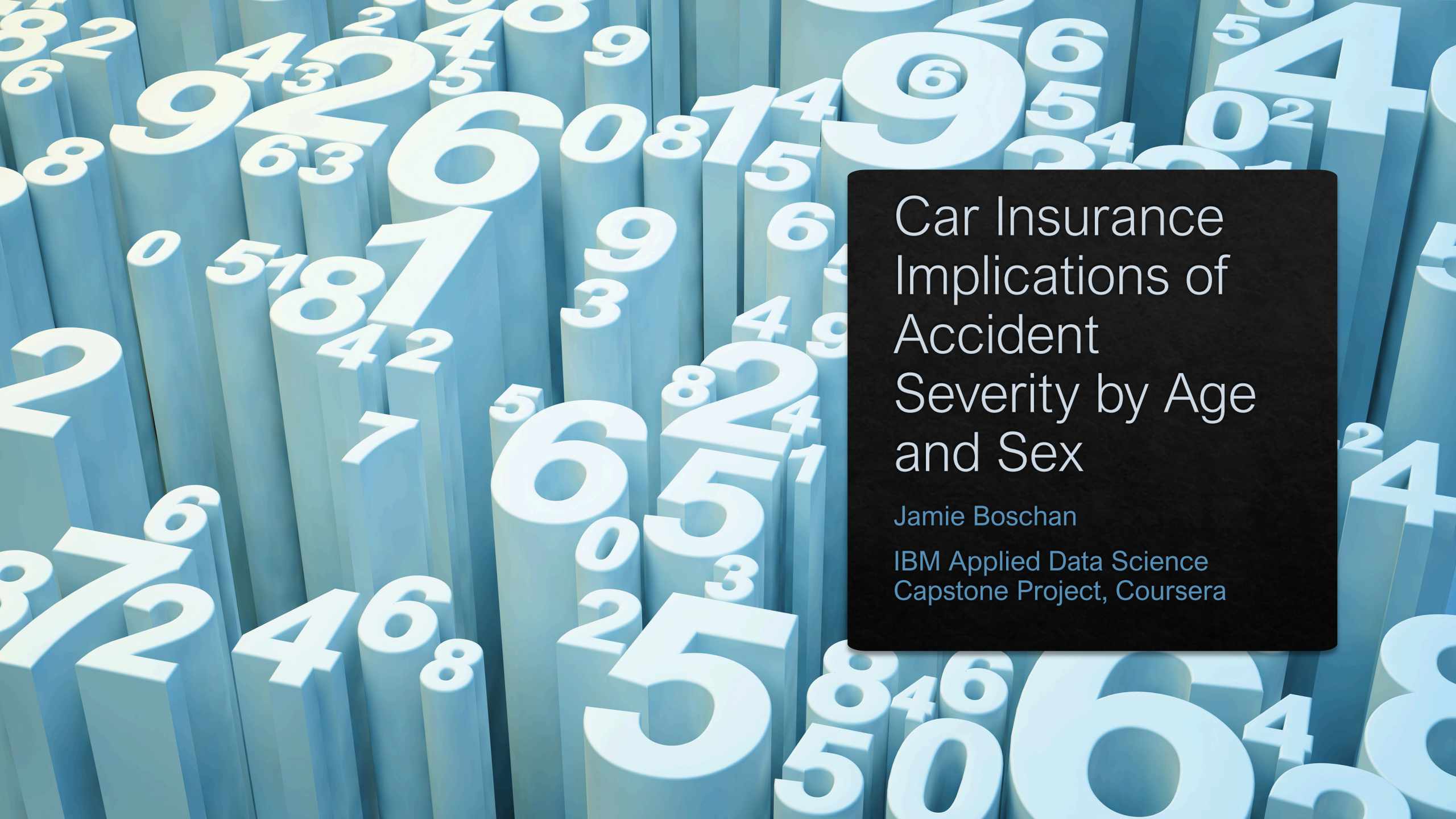


Car Insurance Implications of Accident Severity by Age and Sex

Jamie Boschan

IBM Applied Data Science Capstone Project, Coursera

Table of Contents

1. Project Overview

Business Problem

Should car insurance companies charge higher premiums based on age and sex of drivers?

Modeling

Logistic Regression

Predictive Accuracy

Inferential Statistics

Result

Yes, car insurance companies can charge higher premiums based on age and sex of drivers

2. Business Problem

Historically, younger (<25), older (>65) and male drivers were charged higher premiums because they got into more severe accidents

Should car insurance companies still factor age and gender into insurance prices?

3. Data – Overview

- Source: Pennsylvania Department of Transportation (PENNDOT)
- Content: crash data for Philadelphia county, PA in 2018
- Tables: PERSON, CRASH, FLAG

3. Data – Final Data Frame

	PERSON_TYPE	INJ_SEVERITY	AGE	SEX	PROPERTY_DAMAGE_ONLY	INJURY_OR_FATAL	INJURY	FATAL	ALCOHOL_RELATED	DRINKING
CRN										
2018000618	1	4	39	0.0	0	1	1	0	0	0
2018001306	1	0	55	0.0	0	1	1	0	0	0
2018001756	1	0	17	0.0	0	1	1	0	0	0
2018002147	1	0	18	0.0	1	0	0	0	0	0
2018002537	1	3	30	0.0	0	1	1	0	0	0

5 rows × 38 columns

4. Exploratory Data Analysis

Chart 1: Collision type vs Age vs Sex

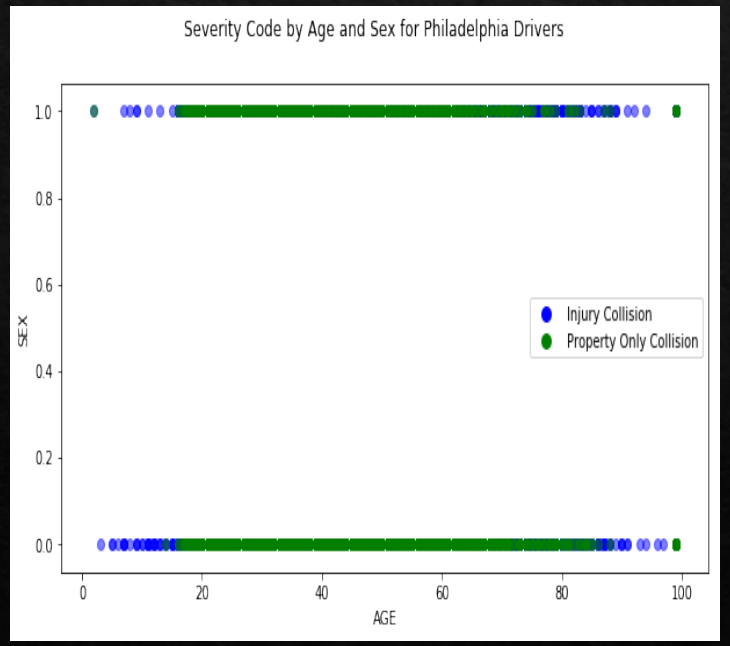


Chart 2: Gender vs. Severity_Level

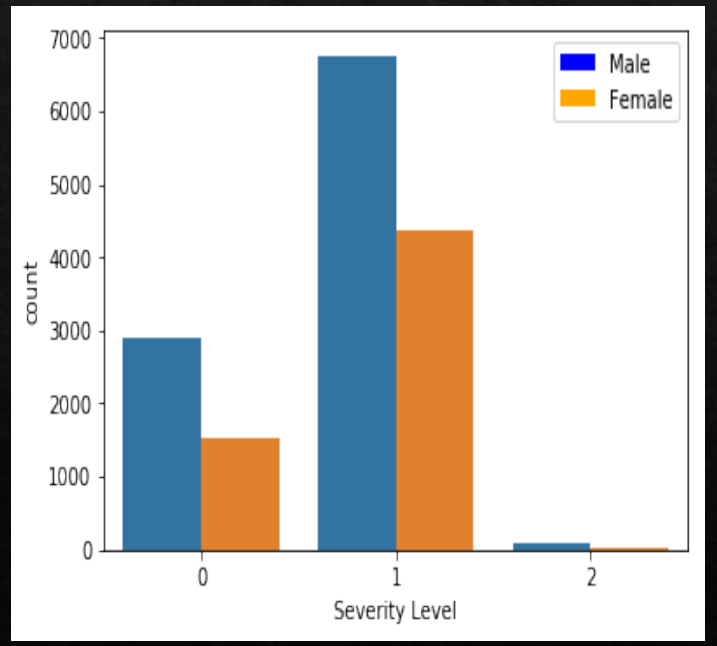
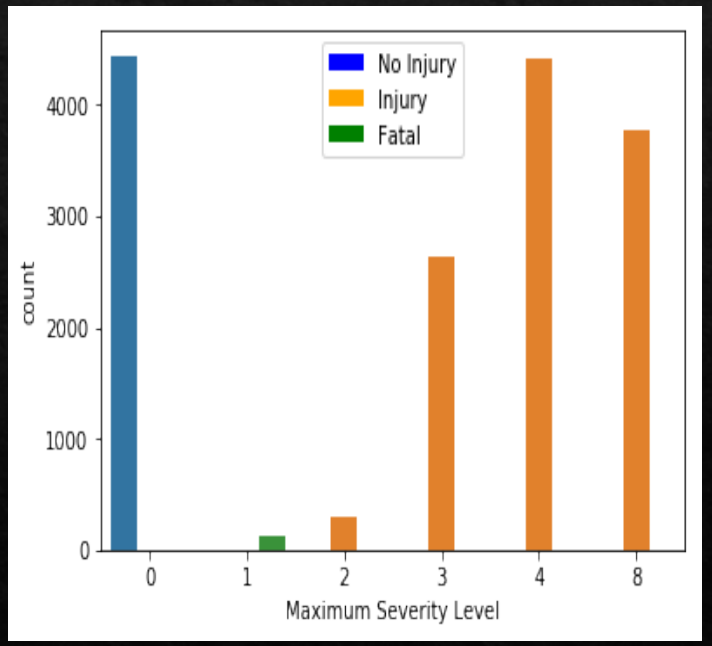


Chart 3: Severity_Level vs. Max_Severity_Level



Chosen target variable after Exploratory Analysis: SEVERITY_LEVEL

5. Modeling

Logistic Regression #1

Target: SEVERITY_LEVEL

Predictors (20):

AGE, SEX, ALCOHOL_RELATED, CELL_PHONE, SPEEDING, ~~SPEEDING_RELATED~~, DRIVER_16YR, DRIVER_17YR, DRIVER_18YR, DRIVER_19YR, DRIVER_20YR, DRIVER_75PLUS, ~~DRIVER_COUNT_16YR~~, ~~DRIVER_COUNT_17YR~~, ~~DRIVER_COUNT_18YR~~, ~~DRIVER_COUNT_19YR~~, ~~DRIVER_COUNT_20YR~~, ~~DRIVER_COUNT_50_64YR~~, ~~DRIVER_COUNT_65_74YR~~, and ~~DRIVER_COUNT_75PLUS~~

Logistic Regression #2

Target: SEVERITY_LEVEL

Predictors (11):

AGE, SEX, ~~ALCOHOL_RELATED~~, ~~CELL_PHONE, SPEEDING~~, DRIVER_16YR, DRIVER_17YR, DRIVER_18YR, DRIVER_19YR, DRIVER_20YR, and DRIVER_75PLUS.

Logistic Regression #3

Target: SEVERITY_LEVEL

Predictors (8):

AGE, SEX, DRIVER_16YR, DRIVER_17YR, DRIVER_18YR, DRIVER_19YR, DRIVER_20YR, and DRIVER_75PLUS

6. Model Evaluation: Model #3

Predictive Accuracy

Statistical Summary

Jaccard: 0.708	F1 Score: 0.59	Log Loss: 0.648	Explained Variance: 0	Mean squared log error: 0.1371	R^2: 0.1371	MAE: 0.1371	MSE: 0.2914	RMSE: 0.5398

Statistically significant predictors:
1) Crashes with injuries: Age, Sex; 2) Crashes with fatalities: Age, Sex, 20YR

7. Results & Conclusion

- Model #3 had strong predictive accuracy, but weak explanatory power. Nevertheless, model #3 showed that Age and Sex both have a statistically significant relationship with accident severity.
- This result answers the business question, which was whether these attributes should still be utilized by car insurance companies in setting insurance rates.
- The model was probably weak regarding age because the relationship isn't linear and needed a different transformation.
- PA does not currently use Sex in car insurance rates, probably for gender equity reasons, even though the data suggests it would be a valid choice.