Car Insurance Implications of Accident Severity by Age and Sex

APPLIED DATA SCIENCE CAPSTONE COURSE:
IBM DATA SCIENCE PROFESSIONAL CERTIFICATION ON COURSERA

by Jamie Boschan

September 27, 2020

I. Introduction

A. Background

Car insurance companies use many factors in their models to segment their customers into price categories. Some of the factors they consider include where customers live (metro area, rural vs suburban), car model, marital status, income, credit history, and annual mileage. Additionally, car insurance companies underwrite policies based on how risky they believe each driver will be to insure. Historically, car insurance companies have charged higher premiums to young (16 to 25) and senior drivers and to men based on statistically higher propensity to get into severe car crashes. According to fact sheets on Carinsurance.com and thezebra.com, average car insurance rates nationwide are highest for 16-year-old drivers and decrease annually until a driver turns 25, at which point they stabilize for many years. Then for senior drivers, rates begin to increase again around age 65. For drivers 16 to 25, male drivers pay higher rates than female drivers in every age category. Once drivers reach 25, the gender difference disappears.

B. Business Problem

The audience for this project is car insurance companies. To continue to accurately underwrite car insurance policies, insurance companies need to know if it is still true that age and gender affect the severity of car crashes. In this project, I will analyze vehicle crash data to see if the historical patterns still hold, and therefore, current underwriting practices are still appropriate. My hypothesis is that it is still true that younger drivers (under 20) and older drivers (over 75) get into more crashes that cause injuries or deaths.

II. Data Acquisition and Preparation

A. Data Sources

Since I live in Pennsylvania, I decided to utilize data from my own state and region. The Pennsylvania Department of Transportation (PennDOT), collects vehicle crash data from every county in Pennsylvania every year. This crash data is published, shared with the national Department of Transportation, and can be used by the state and municipal governments to make new laws to improve safety or by car insurance companies to revise their underwriting tables for insurance. In this project, I focused on vehicle crash data from Philadelphia County, in the year 2018. I used the data from 2018, the most recent year available, because my objective is to see if longstanding assumptions are still true. The data

was downloaded from the PennDot Crash Download Map, then uploaded into IBM Cloud Storage for this project. The full data dictionary for the PA crash data can be found here.

The PennDot zip file contained eight (8) .csv files, each for a different table, with a Crash Record Number (CRN) as the index key in all tables:

- COMMVEH_2018_Philadelphia.csv
- CRASH_2018_Philadelphia.csv
- CYCLE_2018_Philadelphia.csv
- FLAG_2018_Philadelphia.csv
- PERSON_2018_Philadelpia.csv
- ROADWAY_2018_Philadelphia.csv
- TRAILVEH_2018_Philadelphia.csv
- VEHICLE_2018_Philadelphia.csv

B. Feature Selection

Since there were hundreds of columns contained in the eight tables I downloaded from PENNDOT, I studied the data dictionary to identify the most useful tables and the relevant columns from each one. The research question at hand is whether drivers who are under 25, over 65, or male get into more severe accidents than drivers who are not, I started with the PERSON_2018_Philadelphia.csv table. I dropped all rows in the PERSON table who were not identified as drivers, as drivers are the focus of the study. I then identified relevant columns in the CRASH_2018_Philadelphia.csv and FLAG_2018_Philadelphia.csv tables that I thought car insurance companies might consider in their models – most but not all variations of age or sex/gender. Here are the columns I initially selected:

PERSON	CRASH	FLAG
• **CRN** • PERSON_TYPE • INJ_SEVERITY • AGE • SEX	• **CRN** • COLLISION_TYPE • FATAL_COUNT • INJURY_COUNT • DRIVER_COUNT_16YR • DRIVER_COUNT_17YR • DRIVER_COUNT_18YR • DRIVER_COUNT_19YR • DRIVER_COUNT_20YR • DRIVER_COUNT_50_64YR • DRIVER_COUNT_65_74YR • DRIVER_COUNT_75PLUS • MAX_SEVERITY_LEVEL'	• **CRN** • PROPERTY_DAMAGE_ONLY • INJURY_OR_FATAL • INJURY • FATAL • ALCOHOL_RELATED • DRINKING_DRIVER • UNDERAGE_DRNK_DRV • CELL_PHONE • SPEEDING • SPEEDING_RELATED • AGGRESSIVE_DRIVING • DRIVER_16YR • DRIVER_17YR • DRIVER_18YR • DRIVER_19YR • DRIVER_20YR • DRIVER_50_64YR • DRIVER_65_74YR • DRIVER_75PLUS

C. Data Cleaning and Preparation

First, I uploaded all eight .csv files directly into my project on IBM Watson Studio, and then imported the three tables of interest (PERSON, FLAG, & CRASH) into my project notebook. I read each one into a Pandas data frame (python terminology for a table), then dropped all rows in the PERSON table that were not coded as drivers. To create my final data frame (or table) for the project, I extracted the columns I needed into new data frames for each table, then joined my three smaller data frames using the CRN (Crash Record Number) as the index on which to match the rows. This data frame contained 19054 rows and 35 columns.

Once I had constructed my data frame, additional data cleaning was required before analysis could begin. First, I had to identify and handle missing data. I ran a loop over the data frame to identify missing data in each column and found that the SEX variable had 39 rows with missing data. I dropped from the table all rows where SEX was missing a value.

Next, I had to check data types for each column, and convert any String variables to a numeric type. All the binary Yes/No variables in the original tables were imported as String types, and SEX was also originally a string variable containing "M," "F," and "U" values. I converted these columns to numeric variables, with 1 and 0 replacing Yes/No for the binary flag variables and 0, 1, and 2 replacing the letters in the SEX variable. I then dropped all rows where SEX=2 (unknown) to remove ambiguous rows that could confuse the model. I also had to rename two of the binary categories, which imported with typos in the column names.

Finally, I was not sure that any of the existing variables in the dataset were quite right to use as a target/predictor variable representing accident severity, so I computed three new ones. First, I created "SEVERITY CODE" and "SEVERITY DESC" (one numeric, one string) to mimic the ones in the example dataset from the course. Both variables indicated either "Property Damage Only" or an "Injury collision." Second, I recoded "MAX_SEVERITY_LEVEL" into a new variable called "SEVERITY_LEVEL," reducing nine categories (0 – Not injured, 1 – Killed, 2 – Suspected Serious injury, 3 – Suspected Minor injury, 4 – Possible Injury, 8 – Injury/ Unknown Severity, 9 – Unknown if Injured) down to three (0 – No Injuries, 1 – Injuries, 2 – Fatal). I dropped all rows where MAX_SEVERITY_LEVEL was "9 – Unknown if Injured" as they could confuse the model.

III. Methodology

A. Exploratory Data Analysis

1. Charting relationships between AGE and/or SEX and TARGET variables

Many of the variables in this dataset are categorical in nature, even if they are coded numerically. Therefore, I used different types of charts – scatterplots, box plots, and bar charts – to get a sense of the relationships between the variables. Below is a selection of the charts I created, with descriptions.

a) Target Candidate: SEVERITY_CODE (1=Property Damage Only, 2=Injuries)

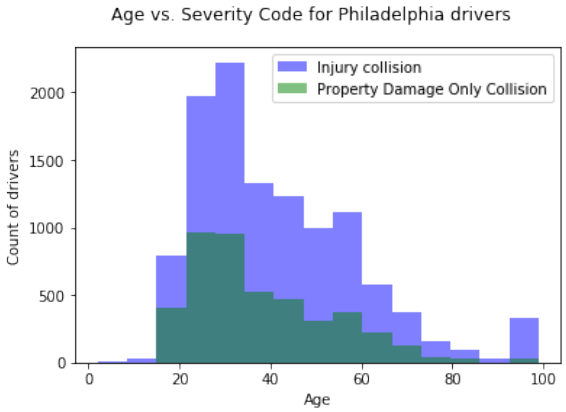


Figure 1: Histogram of Age vs Severity Code

First, I constructed a histogram (figure 1) to see the number of drivers by age who had injury collisions vs. property damage only collisions. The chart shows that at every age group, more drivers had injury collisions than property damage-only ones, but the relative volume differs.

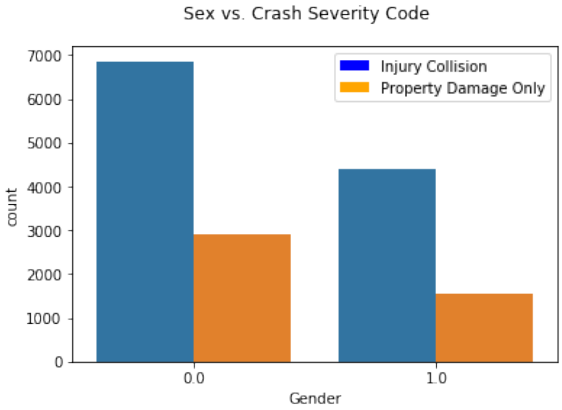


Figure 2: Gender vs. Severity Code

Next, I constructed a bar chart to compare injury vs. property damage-only collisions by gender. This chart quite clearly indicates that men (value 0) had more collisions of both types than women (value 1). However, for both men and women injury collisions appear to be more common in this chart.

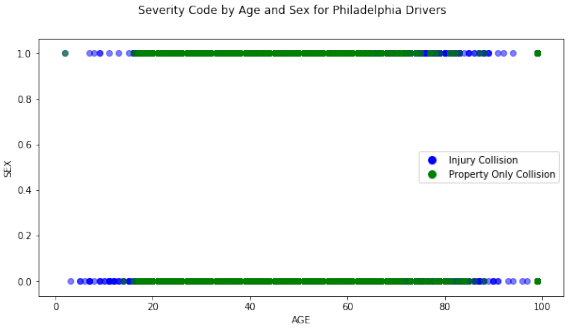


Figure 3: Age vs. Gender vs. Severity Code

Finally, I constructed a scatterplot comparing Severity Code by both Age and Gender. For both men and women, we see that injury and property-only collisions overlap substantially between about age 20 and 85, but at the lower and upper ends of the age range, there are more injury collisions for both men and women.

b) Target Candidate: INJURY_OR_FATAL (0=No injuries, 1=Injury or Fatal)

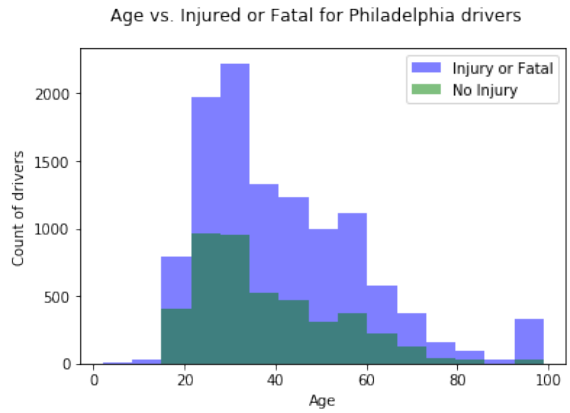


Figure 4 shows another histogram, this time of the number of drivers who had injury/fatal or no-injury collisions by age. This histogram is like the one above in the patterns it reveals, with more crashes having injuries than not across the age range.

Figure 4: Age vs. Injury_or_Fatal

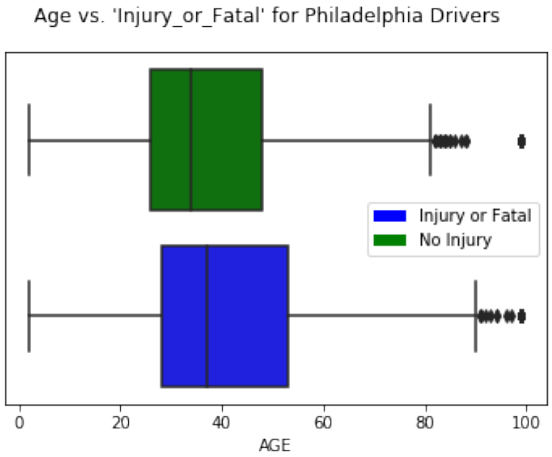
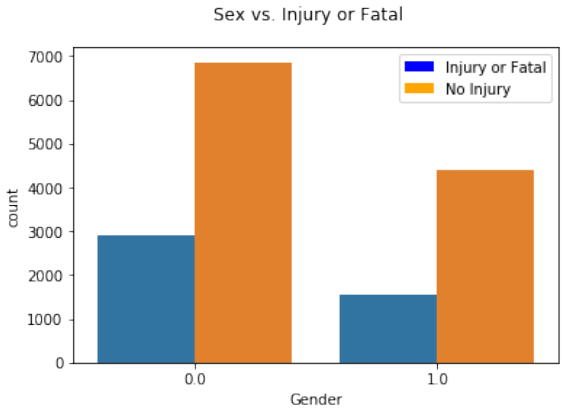


Figure 5, like figure 4, compares injury/fatal vs non-injury collisions by age. However, this chart is a horizontal box plot, rather than a histogram. We can see that the range of non-injury vs. injury/fatal collisions are similar, with outliers starting around 80 for non-injury collisions and around 90 for injury/fatal collisions.

Figure 5: Age vs. Injury_or_Fatal



Finally, figure 6 depicts a bar chart comparing injury/fatal to non-injury collisions by gender. Here, there is a more pronounced pattern where for both genders, most crashes do not result in injuries, but men (coded 0) have more crashes of both types than women.

Figure 6: Sex vs. Injury_or_fatal

c) Target Candidate: SEVERITY_LEVEL (0=No injuries, 1=Injuries, 2=Fatal)

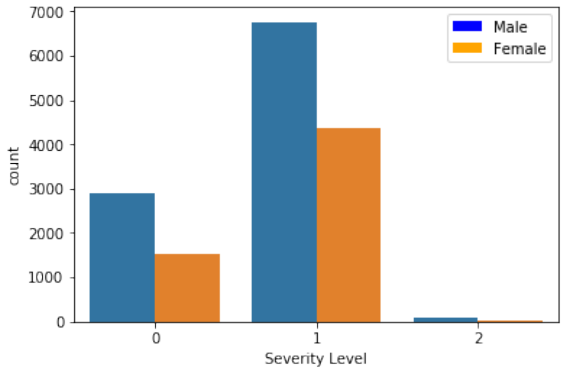


Figure 7: Severity Level vs. Gender

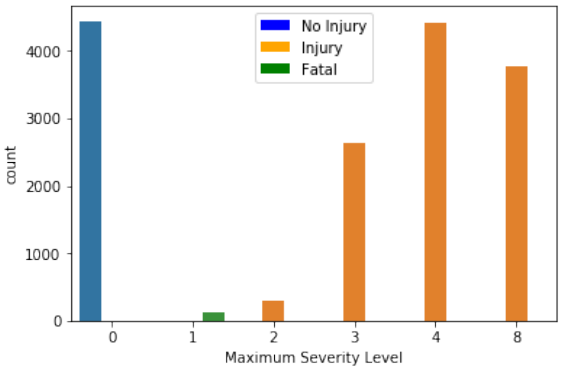


Figure 8: Severity Level vs. Max Severity Level

Finally, for Severity_Level, I only created two charts. Figure 7 shows Severity Level by Gender and Figure 8 compares Severity Level to Max_Severity_Level, the variable from which I derived it. In Figure 8, we can see that the recoding was done correctly, and in both charts, we can see that the number of crashes with fatalities is exceedingly small relative to the number of crashes with or without injuries. Figure 7, as above, shows that male drivers have more accidents in all three categories than female drivers.

2. Choosing a target variable

Initially, I believed that property damage and injuries or fatalities could both theoretically factor into a variable representing accident severity. There were no variables in the original data files that factored in both types of damage, so I constructed "SEVERITY_CODE" and "SEVERITY_DESC" variables as described above. I also considered several injury/fatality-focused variables as possible targets: "INJURY_OR_FATAL," "INJURY," "FATAL," "MAX_SEVERITY_LEVEL," and my own "SEVERITY_LEVEL" variable. Based on the charts shown above, and others not shown, "SEVERITY_LEVEL" was the best predictor variable and was used in all models.

B. Modeling

The research question in this project was to assess whether car insurance companies should continue to incorporate age and sex into their risk models when pricing car insurance. The chosen target variable, SEVERITY_LEVEL, is a categorical variable (even though it is coded numerically). Therefore, we must use a model that works with categorical targets. The Data Science courses in this certification program covered quite a few classification and segmentation models which fit this criterion: Decision Trees, K Nearest Neighbors, Logistic Regression, K-means, and SVM, among others. Of all these, logistic regression is the only one with which it is easy to not only see how well the model makes predictions, but to also analyze the relative import of each factor in the model. Therefore, I used logistic regression exclusively in modeling this problem. I built three logistic regression models, dropping predictor variables that seemed unhelpful as I went along. To build these models, I used SKLearn to split the data frame into training and testing sets, then built the model twice – once using SKLearn and once using Stats models. I built each model twice because Stats models turned out to have more statistical reporting capabilities, which I needed to assess the statistical significance of the predictor variables.

1. Model #1

The first logistic regression model I built utilized 20 predictor variables: SEVERITY_LEVEL as a function of 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.

2. Model #2

The second logistic regression model I built utilized 11 predictor variables: SEVERITY_LEVEL as a function of AGE, SEX, ALCOHOL_RELATED, CELL_PHONE, SPEEDING, DRIVER_16YR, DRIVER_17YR, DRIVER_18YR, DRIVER_19YR, DRIVER_20YR, and DRIVER_75PLUS.

From Model #1 to Model #2, I removed variables that were correlated or closely related to other predictor variables in the model, because these can cause multicollinearity, which reduces the statistical validity of the model. Specifically, I removed SPEEDING_RELATED for being related to SPEEDING and _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 for being correlated with DRIVER_16YR, DRIVER_17YR, DRIVER_18YR, DRIVER_19YR, DRIVER_20YR, and DRIVER_75PLUS respectively.

3. Model #3

The third and final logistic regression model I built utilized 8 predictor variables: SEVERITY_LEVEL as a function of AGE, SEX, DRIVER_16YR, DRIVER_17YR, DRIVER_18YR, DRIVER_19YR, DRIVER_20YR, and DRIVER_75PLUS.

In this model, I removed all remaining variables not related to the question regarding the impact of Age and Sex on car crash severity – namely ALCOHOL_RELATED, CELL_PHONE, and SPEEDING.

C. Model Evaluation: Predictive Accuracy and Inferential Statistics

All three models were evaluated two ways: 1) predictive accuracy, and 2) inferential statistics.

1. Predictive accuracy for all three models

Predictive accuracy measures indicate how well the model can predict which category a new item belongs to in the dataset. I used all three covered in the course: Jaccard, F1 score, and Log Loss.

Metric	Model #1	Model #2	Model #3
Jaccard	0.7087935429056924	0.7085811384876806	0.7085811384876806
F1	0.59	0.59	0.59
Log Loss	0.6401573637769501	0.6476053070669625	0.6480436298207022

2. Statistical validity and Inferential statistics

For each of the three models, I checked the statistical validity of the model by checking errors and correlation tables. The correlation tables validated the choice to remove predictor variables from models #2 and #3.

Here are the statistical summaries for each model:

Metric	Model #1	Model #2	Model #3
Statistical summary	explained variance: -0.0038 mean squared log error: 0.137 r2: -0.3432 MAE: 0.2912 MSE: 0.2912 RMSE: 0.5396	explained variance: 0.0 mean squared log error: 0.1371 r2: -0.3441 MAE: 0.2914 MSE: 0.2914 RMSE: 0.5398	explained variance: 0.0 mean squared log error: 0.1371 r2: -0.3441 MAE: 0.2914 MSE: 0.2914 RMSE: 0.5398
Significant Predictors, SEVERITY_LEVEL=1 (Injuries); p<.05	Age, Sex, plus others	Age, Sex	Age, Sex
Significant Predictors, SEVERITY_LEVEL=2 (Fatalities); p<.05	Age, Sex, plus others	Age, Sex, Alcohol-related, Speeding, Driver_20YR	Age, Sex, Driver_20YR

IV. Results

The predictive accuracy scores were nearly identical for all three models, which means all three can predict the likelihood of severe crashes of new drivers well. The accuracy measures do not really address the business problem here, however.

The regression summary statistics tell us that the models I built do not have particularly strong explanatory power. The first model is not statistically valid because of multicollinearity between several of the predictors and should be ignored. Models #2 and #3, however, show meaningful results. Both models indicate that Age and Sex are statistically significant variables in relationship to accidents that had injuries, and both models indicate that Age, Sex, and DRIVER_20YR are statistically significant variables in relationship to accidents that had fatalities. I will base my discussion and conclusions on Model #3, which was the most relevant to the business problem.

V. Discussion

All three models that I built had strong predictive accuracy, but weak explanatory power. The models were sufficient – particularly model #3 – in answering the business question at hand, which was whether car insurance companies should continue to factor age and sex into their driver risk models when pricing car insurance, as they have historically done. Model #3 showed that age and sex both have statistically significant relationships with accident severity, so it is valid for car insurance companies to factor them into their models.

The relationship between age and accident severity appeared to be weak in all my models. This apparent weakness was probably because the relationship between age and accident severity is not linear. Younger drivers may indeed be riskier drivers because they misperceive danger and may be more likely to drive under the influence of alcohol or drugs. Similarly, older drivers may be riskier if they start to have trouble seeing and/or may simply be at higher risk of injury. Drivers who are in between these extremes are generally less risky drivers. My model could probably be improved by applying a polynomial transformation to the age variable or breaking up drivers into age groups so the risk can be more accurately measured.

The relationship between sex and accident severity was more evident, both in the exploratory data analysis and in the resulting models. However, thezebra.com says that Pennsylvania is one of six states that do not use gender in assigning car insurance premiums. PA is now a state that allows transgender and gender non-conforming drivers to use an X gender marker on their drivers licenses, so equity concerns might be why the state does not utilize gender as a predictor, even though the data suggests they could do so.

VI. Conclusion

Historically, car insurance companies have assigned higher car insurance rates to young drivers (under 25), older drivers (over 65), and men due to statistically higher incidence of severe car accidents. I set out to determine whether these practices are still valid using vehicle crash data from Philadelphia, PA for the year 2018. I combined key variables from three tables, PERSON, CRASH, and FLAG to create the data frame for analysis. I computed a new target variable, SEVERITY_LEVEL, which reduced the original MAX_SEVERITY_LEVEL variable to just three categories: 0 – No injuries, 1 – Injuries, 2 – Fatalities. I built three logistic regression models, removing unnecessary or over-correlated predictor variables each time. All three models had strong, and nearly identical, predictive accuracy. Model #1 was not statistically valid, because predictor variables were correlated with each other. The final model I used was Model #3. Model #3 had weak statistical explanatory power, but inference testing confirmed that Age and Sex both had statistically significant relationships to accident severity. Therefore, it is still valid for car insurance companies to use these variables in their risk models and pricing structures.