Saving Lives with Effective Data Visualization: Evaluating the Effectiveness of Indiana's Driver Education Curriculum

Amratansh Sharma, Aditi Vatse, Matthew A. Lanham
Purdue University Krannert School of Management
sharm321@purdue.edu; avatse@purdue.edu; lanhamm@purdue.edu

Abstract

In the 21st century which is all about information, there is lots of data and less information. Studying trends and deriving information is important, but what matters the most is how that information can be presented and visualized to add value to the business. In this project, we collaborated with the Indiana Bureau of Motor Vehicles (BMV) to help them understand and evaluate the current effectiveness of their driver education curriculum being taught to various drivers. Specifically, we analyze the effect of driver education on getting the operator licenses, what type and number of citations are seen from different age groups with or without driver education, and the impact of driver education on fatalities. We also developed a logistic regression model to provide odds-ratios of the effects of certain curriculum coverage, but only used these parameter estimates as support for what was clearer to the stakeholders – effective data visualizations. We developed the BMV Data Diagnostic Tool, designed in Tableau, so that they can have a better understanding of the impact that Driver Education has on the community and they can take further steps to take appropriate actions with it.

Introduction

Driver education is one of the important resource for new drivers. It works towards preparing citizens for a better driving experience. A driver training program basically is made up of 30 hours of classroom training and six hours of behind the wheel training with a BMV licensed driver training school. There are various schools which provides the training. Online training is also available in alternative to in-classroom training, however the six hours behind the vehicle training is mandatory.

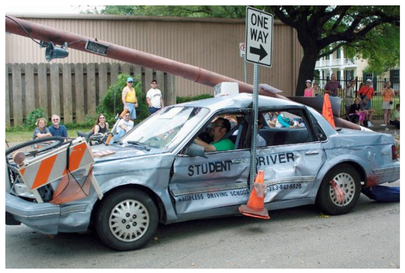

Figure 1. Possibilities in case of No Driver Education

In this project, our focus is to determine the impact of driver education. As well as provide answers to the following research questions.

Research Objectives:

- ❑ What is the percentage of individuals there were issued operator licenses completed a driver education program?
- ❑ What is the number and type of citations issued to individuals by age group with and without driver education including the need to file a certificate of compliance?
- ❑ What is the number of fatalities by age group with and without driver education?

Further, we have designed and evaluated a logistic regression model to determine the effect of a particular citation on the driver education received. This will help derive necessary changes to the curriculum.

Literature Review

After understanding the business problem and research objectives, we researched studies on driver education importance and techniques that were used in those studies to improve driving outcomes. Below are the findings:

Paper	Learning	Linear Regression	Logistic Regression
The impact of driver education on self-reported collisions among young drivers with a graduated license	Impact of driver education dependent on stage of driver learning in which it occurs.		✓
The Impact of Driving Knowledge on Motor Vehicle Fatalities	Written test help produce better novice drivers	✓	✓
Do driver training programs reduce crashes and traffic violations? — A critical examination of the literature	Changes attitudinal and maturational factors underlying risky and dangerous behavior through classroom and on road training		✓

Table 1. Literature review summary and methods used in those studies

Methodology

Figure 2 outlines our development steps, starting from file processing, Tableau dashboarding and statistical model development to create the BMV Diagnostic Tool.

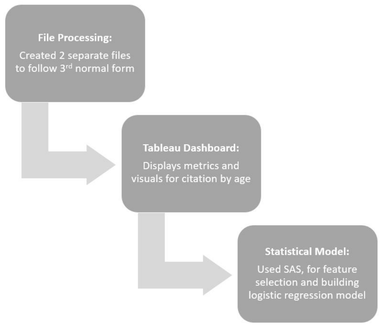

Figure 2. Development Steps

Data
The data used provides details of number of individuals provided driver license with and without driver education. It also provides the details of age groups, type and number for citations by various driving schools, and if a group passed or failed the driving test.

Data Cleaning & Pre-Processing
EDA was used to identify the outliers and trends. Missing values were treated and incorrect formats were standardized.

Model Design and Methodology Selection
To evaluate the effects and influence of citations, we developed a logistic regression model so we could make causal inferences of the effects. We observed that there were 321 'X' independent variables, such as citations, age, etc., which were too many in number for interpretation. Although the model was accurate (cross-validated accuracy = 90%), many variables were not statistically significant drivers of having had diver education.

Data Dimensionality Methods Investigated:

1) Principal component analysis (PCA)
2) Forward Selection
3) Backward Elimination
4) Subset Selection

We reduced the number of independent variables by using PCA. Once it was implemented, 8 principal component explained 45% of the variance in the data. While this explained the effect of between 100-150 variables, PCA was not as clear to the decision-makers of the effect of each citation. We performed forward selection, backward elimination and stepwise selection in SAS. After doing the whole process, we came across the 7 most important variables which provided easier interpretation to our partners.

Model Evaluation / Statistical & Business Performance Measures

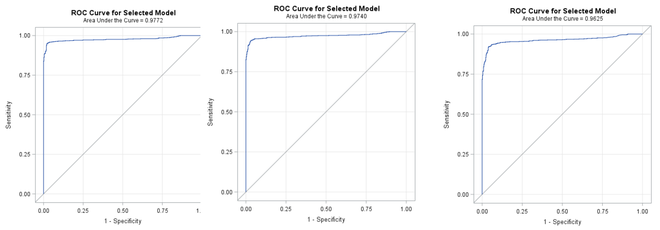

Figure 3. ROC Curves for Forward, Backward and Stepwise Selection (left to right)

We applied 5-fold cross-validation and achieved an accuracy of 88.6%. Further, we also utilized a grid search algorithm to validate to improve the accuracy to 93.97%.

Results

The first part of the solution was to allow a one stop view of the overall effectiveness of driver education. Figure 4 shows dynamic visualizations which provide the details as following:

- It provides percentage of people who took driver education and passed it successfully.
- It provides the number of Fatality citations issued. This number is classified by status of driver education.
- To slice and dice the data, an interactive chart was created. This chart helped filter the data by status of driver education, if certificate of insurance was required and type to citations.
- A detailed data table was added to allow drilldown into details by selections from the chart. This detailed table provides details like number of individuals with particular citation in an age group.

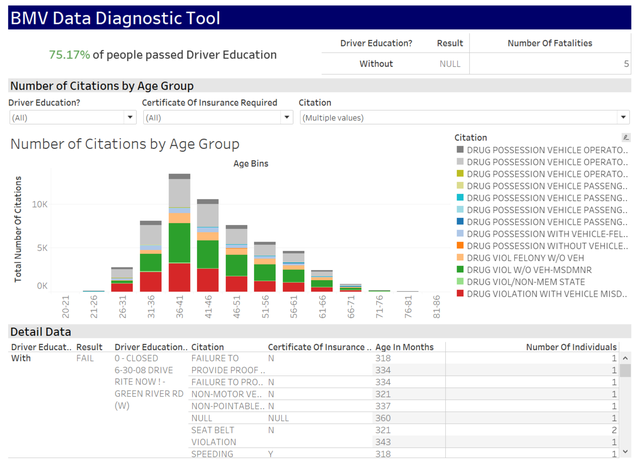

Figure 4. BMV Diagnostic dashboard

The second part of the solution was to recommend the areas of improvement in the curriculum. Based on our logistic regression model, we concluded that citations occur the most in absence of driver citation.

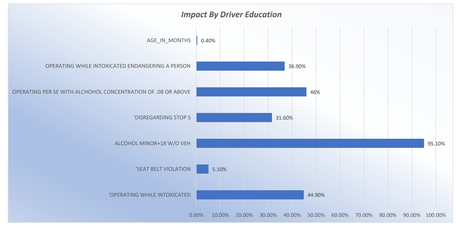

Figure 5. Impact of Driver Education on Citations

Conclusions

In this project we developed the BMV Diagnostic Tool, which provides effective data visualization supported by model estimates that helped the Indiana BMV make strategic decisions to improve their driving education curriculum and training. We show that getting the community in driver education programs can lead to a significant decrease in drivers being cited for intoxicated driving, driving thru stop signs, and seat belt violations, all of which lead to loss of life within our communities.

Acknowledgements

We thank Professor Matthew Lanham for constant guidance on this project. We also thank the Indiana BMV for providing us an opportunity to make a positive impact on our community and the great State of Indiana.