CIS242 Final Project - Report For GCRoc Kentaro Tanaka

Assumptions Made:

In this memo to the CEO of GCRoc, I make a few assumptions.

1. We cannot have different auto insurance plans based on race, time of year (since plans are usually long term), or gender. (I keep these out of my analysis)
2. Fatal injuries cost more to cover and bring more trouble (in terms of lawsuits) than other types of injuries.

Dear CEO of GC Roc,

I hope you are doing well. I was able to complete a report on the analysis of the data you sent me on fatality accidents throughout the US.

The goal of this data analysis is to determine which region is the best one to expand our business to (Midwest, West, or South). For this reason, I have left out our current region of the Northeast from our analysis. I focused on the aspects of the data that would help us change and improve our offerings and left out characteristics that we can not make different offerings for such as race or gender.

As an insurance company what we would like to avoid the most are incidents involving death. They cost the most and put our company in danger of lawsuits. To start my analysis, I started looking at the different data that we have and how they could help us predict which type of people end up as fatalities in fatality accidents. I ran many tests on the different characteristics of individuals. The age range of the individual only had a 54.1% accuracy of predicting fatal injuries. The best indicator I could find was the type of person (driver, passenger, or pedestrian) and alcohol content in the system to predict if they were the ones to die during an accident. The type of person had an accuracy rate of 60.5% and alcohol content in the system had an accuracy rate of 74.2% of whether that person was going to face a fatal injury or not in a fatality-involving accident. I would like to point out, however, that these accuracy rates are actually quite low. During this analysis I was starting to realize that this set of data may not give us the correct results to make our business decision.

I decided to perform further analysis on the two factors that had the highest accuracy rates for predicting fatal injuries: alcohol level and person type.

I started with doing an analysis of alcohol in the system. Since this seemed like an obvious factor, I decided to include data from every region to do an overall analysis. In this analysis I excluded the data that didn’t report alcohol levels. As predicted, alcohol content has a strong relation to fatal injuries in fatal accidents. We can say with 84% accuracy that those with alcohol in their bodies are involved with fatal injuries 27.8% of the time in fatal accidents. Whether a person has alcohol content in their bodies is the largest indicator of whether they had a fatal injury in all of the data. This gives us enough evidence to tell us that no matter which region we decide to expand business in, we shouldn’t accept customers with a history of drunk driving. These types of customers are simply too much of a risk for us in terms of cost and legal cases. If you decide to insure those with a history of drunk driving, there should be a policy in place that charges them significantly more than a driver without a history of drunk driving.

After finding out that alcohol content is an obvious indicator regardless of region, I decided to study each region individually on its state and person type and how they relate to injury severity. I decided to filter the results based only on those that have over a rate of 60% accuracy. I will first share with you my findings, then explain why I think the current data does not provide information for us to make a decision of which region to expand to next.

In the Midwest, I was able to find two different associations with fatal injuries that have over a 60% accuracy. The first was if the individual were a pedestrian, there is a 91.7% accuracy that they have a fatal injury 3.8% of the time. The second one was that if they are in Iowa, and they are a driver of a motor vehicle, we can say with 63.3% accuracy that they have a fatal injury 3.3% of the time. However, this second association has relatively low accuracy and can be considered to not be very significant.

The South was the most promising with five different associations with fatal injuries that we can say with 60% accuracy. The first was once again if the individuals were pedestrians, where we can say that pedestrians have a fatal injury 5.9% of the time with 96.5% accuracy. The other associations were all in between 60-70% in accuracy which can be concluded as not being very significant. Even if there are five different associations here, it doesn’t help us very much if the majority of them we can only say with 70% accuracy.

Lastly, the West didn’t seem to give us any better results than the other two regions. Individuals who were pedestrians once again gave an accuracy over 90%, however there were no other associations that went over 70%.

After the analysis that I have made on the different factors and each of the regions, I must conclude that this dataset does not answer the question of which region is the best to expand our business to. The biggest associations I could find with fatal injuries was with alcohol presence and if they were pedestrians. These are both quite obvious facts. People with alcohol in their systems (especially drivers) are expected to have a higher rate of fatality then those that don’t. Individuals that are pedestrians should also have a higher rate of fatality, since they don’t have the protection of a car. If they were to be associated in the fatality accident the pedestrians would be the ones getting hit by the car.

Due to the explanations above and the relatively low association accuracy and significance percentages, I can conclude confidently that data on fatal accidents does not help us with making this decision at the moment. I wanted to ask if there possibly is a dataset based on all accidents in general (not only ones with fatalities), I think that this data could possibly give us further insight into which region to expand to next.

Thank you for your time and I hope that these findings do not come as a disappointment to you.

Sincerely yours,

Kentaro Tanaka

Data Scientist

GCRoc