Steven Brannum

BIA 6301 – ADM

Homework #3s

**Part A**

1. How is the analysis different than what you presented before? What changes, if any, did you have to make to the data?

This time I still removed the following columns:

* *atmospheric condition* – because nearly every observation was clear skies
* *fatalities ­*– because these were summarized across rows
* *person type* – because insurance covers every person in the accident no matter whether they’re a pedestrian, passenger, or driver.
* *race* – because I still believe it can be unethical or illegal to base business decisions on this
* all crash date information except for *crash date month* – because it was all from 2011 and the rest was very granular.

I decided to categorize the following columns:

* *states* into regions to follow the previous analysis.
* *Age* I categorized these into “Kid” ( < 15), “YoungAdult” (<30), “Adult” (<45), “OlderAdult” (<60), and “Retiring” ( > 59). I decided that younger than the driving age in most states was a good starting point and then using the same range categorized the rest of the ages. The splits were still fairly even except in the case of “Kid” having less than half the others and “YoungAdult“ having about 1.5x as many as the other three categories.
* *Alcohol Results* – I categorized into above the legal limit, at and below the legal limit and unknown.

I also standardized the values of “NA”, “Not Reported” and “Unknown” all to the value of “Unknown”.

1. Did this analysis give you any insights that you didn’t have before? If so, what are they? If not, why do you think that might be?

Since I left in more data this time, I believe there were more insights to gain. I noticed when a person had alcohol results above the legal limit there was a very high (86%) confidence that they were fatally injured with this occurring in 11% of our observations. Also, when the person’s alcohol levels were above the limit there was a strong (82%) confidence that they were male which we saw in 10.7% of the observations. The last valuable insight I see is that if you were above 59 years old there was a confidence of 62% that you would be fatally injured, and we can see this in 10% of the observations.

1. Does this new work support your previous recommendations? If so, how? If not, what recommendations would you make now? How confident are you in the results?

It doesn’t discount any of my previous recommendations. My recommendations last time were framed in the way that we should market more towards people that weren’t involved as much whereas the frequencies are showing what occurs the most and with the highest confidence and correlation between variables. Looking at this data I would now recommend to not insure people with DUIs or have clauses that don’t pay out for accidents while you’re legally intoxicated (is this legal?). I would still recommend that females are marketed towards more heavily then men due to pure numbers and lack of frequency rules against them. With such a small dataset it’s hard to be confident in anything. I would not be overly confident due to the highest confidence level of a “rule” happening is only 82% and being wrong nearly 1/5 of the time is a lot.

**Part B**

1. Consider/perform a variety of evaluation measures for the work you have already done. Is accuracy the best metric for performance for your question? If so, why? If not, why not? What measure would be better? How does your recommendation from Homework 1 change?

Since homework one was about who might respond to marketing material about a personal loan, I still believe accuracy is a good metric. We don’t really care about false positives (predict interested, but really not) in this case since casting a wider net is not necessarily bad for business. Although, you could look at it from the opposite side and want to eliminate false negatives (predict uninterested, but really are) because you want to make sure you’re marketing to all the people who might be interested.

1. What changes did you make to the original models? Why? What do you expect to see from these new specifications? What are the results?

This time I used k-fold cross validations on the naïve bayes model since the accuracy was lower than I liked. We did increase accuracy and the ability to filter out uninterested people, but we dropped the sensitivity to less than 10%. We almost always predicted they were not interested when they really would be interested which is the exact opposite of what we want. I also dropped the education column from the naïve bayes because it assumes independence between variables, but your salary has been shown to have a direct correlation with your education level. This actually did help slightly in prediction accuracy and doubled our prediction sensitivity.

1. Compare the entire set of models using your preferred criteria. Which performed best? Which would you recommend? Why?

Using a confusion matrix, I can see that the decision tree is still the most accurate and the naïve base and k-nearest neighbors have roughly the same accuracy. The decision tree also had the highest sensitivity accurately predicting 63% of the time when people were interested while the other two only predict correctly 31% correctly and 8% correctly. I would recommend using the decision tree to limit the false negatives (predicting uninterested when they were interested).

**Part C**

I plan on using a SQL table that contains user details from the website AshleyMadison.com. There are details about their attractions, location, physical attributes, etc. that can be categorized. Users of AshleyMadison.com must buy credits to send messages to other users on the platform. The business question would be, “Who is buying credits?”. Currently, I plan on using naïve bayes and frequency pattern analysis, probably the Apriori algorithm.