Hazardous Liquid Data Exploration and Visualization

Riordan Brennan

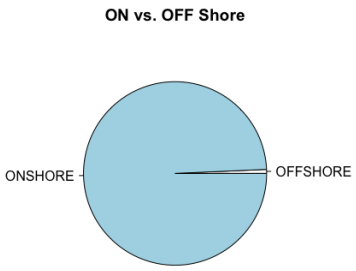
11/5/2020

Data Cleanup

I began my process of organizing my datasets by attempting to combine the datasets of hazardous liquid spills from 1986-2001, 2002-2009, and 2010-Present. However I soon realized this would be practically impossible as the variables for each dataset vary so widly that without a thourough understanding of this field, I could never confidently map the variables from one set to another. Because of this, I will just focus on the most recent dataset, 2010-Present, as I see no reason to create two models for the other datasets if they would just be outdated.

Once I got that sorted out, I began to remove variables. The dataset started with 606 fields, so I knew I had to get rid of a lot before doing any meaningful work. I began by removing any field that I considered a comment field, any non-numeric field where there was no repeated entries. Since this dataset comes from government forms, there are plenty of fields to leave a note or explain an answer, none of this is helpful to me so I removed them all. Next, I removed any identification fields, like names, phone numbers, addresses, etc. None of these data points would have broad enough categories to provide any insight.

My biggest pare down was removing all fields where over 50% of the respondents left the field blank. There were A TON of these fields in the dataset, since this is such a broad and comprhensive dataset, there are a lot of fields that only apply to a small subsection of the respondents, which is good but not very helpful to me. For example there was about 10-15 fields that only applied to offshore accidents, which was only a sliver of the responses



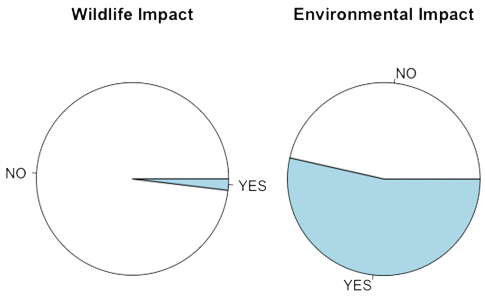
Next on the agenda was removing all fields that related to my two response variables, the estimated environmental cost and wildlife impact. This includes other estimated costs and any indicator variable put in by PHMSA on impact partly based on my response variables. This is simply because these variables wouldn't be known before the response variables in a real life setting, so it would kind of be cheating if I used them. I then finished by removed a few unnecessary quantitative fields that had a categorical counterpart

After this I converted some of my quantitative fields to categorical ones. Since I am going to be working with logit regresion model, it makes no sense to keep around quantitative fields, so I might as well get the most use out of them as I can. I started with the net loss of liquid, I decided, since I'm not expert in pipeline accidents, that I would just split it up into two categories: liquid lost and no liquid lost. So anything above 0.0 barrels lost would be fall into one category and the rest would be in the other. I then did the same with a much more important field, estimated environmental cost. I debated choosing some threshhold number to count as a 'significant cost', but with no real evidence to back it up, I chose to go with the same route as I did with liquid loss.

Now that I had all of the cleanup done, I could move my datasets into two sets, one for each response variable. The only difference between the two sets was that for each set, I removed the response variable from the other. I also removed some rows where there was no response for wildlife impact in that set.

Data Analysis

My first step in analysis was to compare the response variables and their splits. What I discovered was there were virtually no data points with a 'Yes' marked for wildlife impact. This came as a shock to me, I knew it wouldn't be that high of a percentage just from the data I had been glancing over, but I didn't realize there would be only *66* out 3,475 entries with a wildlife impact. Because this number was so low, I clearly didn't have enough data to create a robust model, so I unfortunately had to scrap the wildlife impact plan and just stick to environmental impact. As you can see in the pie charts, environmental impact has a pretty even split between yes and no, whereas wildlife impact has just as sliver of yes comparitively.



Now that I had it down to one response variable, I had to start looking for fields that could be the most helpful in a logit regression model. To do this I took all 55 fields and compared their categories with environmental impact. You can see the results in the charts below. Each bar is one category of one field, and it shows the percentage of responses with that category that had an environmental impact and the percentage that didn't. We're looking for anything that has a high or low percentage.

We seem to have gotten promising results for a good number of variables. I'm going to take a closer look at the commodity type, fatality indicator, ignition indicator, long term assessment, on and offshore, remidiation indicator, soil contamination, and water contamination.

We definitely have some promising variables. Soil contamination and the remediation indicator both look to have pretty good correlation to environmental impact. We also have some other individual categories within fields that look pretty good, like the yes on long term assesment or offshore accidents (although we knwo there weren't very many so that's maybe not a great indicator). The next step is to take these and put them into some models!