Springboard DSF Capstone Proposal

Guy Brodsky

May 10, 2017

## Capstone Project Proposal:

#### Analysing Data on Energy Pipeline Incidents to Predict Future Incidents

My idea for the capstone project component of the foundations of data science workshop is to create predictive models to see specifically which pipelines, or which type of pipeline system made with which type of materials are more likely to leak its respective oil, natural gas, or liquified natural gas (LNG/fracking). There have been thousands of incidents over the past few decades in the US, and as per the Pipeline and Hazardous Materials Safety Authority regulations, the operators of these pipelines have all filed reports on each of these incidents, ranging from the very detailed to somewhat detailed.

##### **The Problem**

These incidents usually have a harmful impact on the surrounding ecosystem, water resources, and stakeholders, specifically when these incidents occurr near or in a residential community, or near a body of water which supplies a local population. My main goal with this report is to analyse which specific factors, if any, have a significant effect on the chance of an accident happening. Gathering this information through analysing the incident reports will help to facilitate a better effort in preventing and reducing the number of accidents.

##### **The Intended Audience and Possible Solutions**

I hope that by making the report publicly available it might prove to be useful info for people involved in the construction, maintenance, and operation of energy pipeline systems in the US and elsewhere, whether they be gas, lng or oil pipelines. This report could also be an especially valuable resource for policymakers which create and implement the regulation of pipelines, and will hopefully influence improvement to the current regulation and governance of energy pipeline systems. Problematic materials, or ways of operating the systems might be avoided, once identified using statistical analysis.

##### **The Data**

To embark on this project, I'll be using [datasets publicly available](https://phmsa.dot.gov/pipeline/library/data-stats/distribution-transmission-and-gathering-lng-and-liquid-accident-and-incident-data%20/%22downloadables%22) through the PHMSA website. The website features datasets separated by type of pipeline system (lng, gas, oil), and type of data. The main datasets that I'll be using are those which contain all incident reports filed between the 1970s to the current 2010s, barring lng incidents which only include data from the past decade. These data sets feature small changes in column information, which I will clean up and render uniform for the sake of analysis and comparison. It also features dozens of categories (column values), and thousands of observations (rows), and so to make the project feasible I will restrict the analysis to significant data only, removing some of the less relevant data, such as the name of the person who filed the report.

##### **Deliverables**

To mark completion of the analysis, I will compartmentalize the project into several parts:  
*1.the code for initial data wrangling, exploratory analysis of relationships, and predictive modeling*  
*2.a written report which thoroughly explains my findings, accompanied with relevant graphics and examples of code used to arrive at conclusions.*  
*3.a slide deck used for presentations of the project*  
*4.an executive summary which includes the integral conclusions and some suggested solutions*  
*5.possibly also a letter - open or private - to an executive of the PHMSA in order to elicit some thoughts on my findings from the regulatory authority of the matter.*