New Mexico Oil and Gas Field Spill Incidents Machine Learning

The New Mexico oil and gas field spills data collected by Oil Conversation Division (https://wwwapps.emnrd.state.nm.us/ocd/ocdpermitting/Data/Spills/Spills.aspx) is used to predict the missing incident severity through supervised classification machine learning. This document is submitted to my GitHub account together with Jupyter Notebook file "New_Mexico_Oil_and_Gas_Field_Spill_Incidents_Inferential_Machine_Learning.ipynb".

Step 1: Data Import and Wrangling

Import the table of spills containing oil and gas field spills data in New Mexico, and the columns are going to be analyzed are:

Facility: facility identifier if the incident happened in a facility
API: well identifier if the incident happened in a well
Operator Name, Severity, Incident Type, Material Spilled, Volume Spilled, Volume Recovered, Spill Cause, Spill Source, District, County, Waterway Affected, Ground Water Impact

Then forward and backward fill the NaNs in the table with the information of the same incident, and drop duplicated rows.

Step 2: Exploratory Data Analysis

From the EDA in the previous inferential statistics study, it's found that there are 4183 the missing values of incident severity out of 26454 rows.

And in the following section, I'm going to apply supervised machine learning to get the missing data of incident severity, and regenerate the correlation plots to compare with the plots in previous section.

Step 3: Machine Learning

To predict the missing incident severity (major or minor) based on the known incident severity and its features, first I convert the columns of data type from string categories to integer identifiers, and assume that the features are independent. Split the data based on known and missing severity data, and the known incident severity and features are used to train the model.

The KNeighborsClassifier is used to conduct the supervised classification machine learning. train_test_split function is used to split the train the train and test data with 20% test size, and use GridSearchCV to tune and cross validate the parameter `n_neighbors`.

Then apply the trained model to predict the missing incident severity data, and update the initial spills table, and the corrolations are re-evaluated among the monthly number of incidents,

volume spilled and recovered, severity, incident type, material spilled, cause, source, waterway and ground water impacted, location distribution, and operators.

Step 4: Conclusion

1. Assuming independent features of the incidents, the KNeighborsClassifier model is applied to conduct the supervised classification machine learning. After tuning and cross validating the parameter `n_neighbors`, I get 86.6% accuracy score for train data and 79.9% accuracy score for test data, with the best `n_neighbors`=5.
2. From the plots of the train data residual distribution and the test data residual distribution, we can see the residual distribution uniformly in 1 and -1 around 0, which further validates the model.
3. After applying the trained model to predict the missing incident severity data, the initial spills table is updated, and the corrolations are re-evaluated among the monthly number of incidents, volume spilled and recovered, severity, incident type, material spilled, cause, source, waterway and ground water impacted, location distribution, and operators.
4. After updating, there monthly number of incidents increases by 10 to 20/month after 2008, and maximum number of major and minor incidents is more than 120/month in 2016 with more minor incidents than major incidents.
5. The maximum monthly number of minor incidents is more than major incidents after updating the table.
6. The correlation between the number of major and minor incidents and incident type, material spilled, spill cause and spill source is adjusted slightly and reordered on each category.
7. The number of incidents in each district and county is adjusted slightly.
8. The total number of minor incidents caused by opertors increases after updating the data, for example, COG OPERATING LLC, ENTERPRISE PRODUCTS OPERATING LLC, BP AMERICA PRODUCTION COMPANY.