Using Supervised Learning Methods to Predict the Probability of a Pipeline Having an Incident Involving a Release of Substance

Team Members: Chandra Keerthi, Herteg Kohar, Xiying Qin, Ziqi Li
Wilfrid Laurier University

Advisor: Dr. Xu (Sunny) Wang

INTRODUCTION

The Canada Energy Regulator (CER) has a mandate to protect people and the environment during the construction, operation, and abandonment of oil and gas pipelines and associated facilities. Despite its best efforts, in the past 12 years, there have been 723 pipeline incidents that involved the release of oil or natural gas.

OBJECTIVES

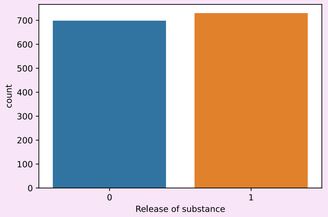
- Predict the probability of incidents related to the release of substances from 2008 to 2020.
- Identify the most important geographical and meteorological features can be identified and used to create a predictive model.
- Evaluate the performance of our models to choose the best one.

METHODS

Exploratory Data Analysis

- Comprehensive incidents dataset was augmented.
- Dependent variable – "Release of Substance" is encoded with 1 indicating release of substance and 0 as no release of substance.

Figure 1: Number of observations showing release of substance and no release of substance

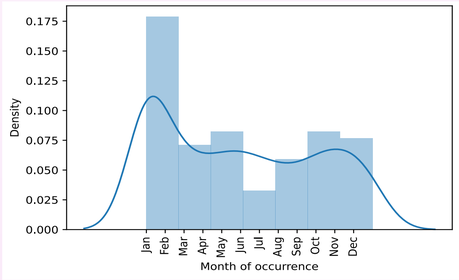


Feature Engineering

- One-hot encoding for 'Incident Types", "Province", "Land Use" to make the categories dummy variables.
- All geographical and meteorological factors related to dependent variable are binary.
- Longitude and Latitude values were augmented as show below to preserve its value during **preprocessing via standard scaler in Python**.

$$Longitude = cos(Latitude) * cos(Longitude)$$
$$Latitude = cos(Latitude) * sin(Longitude)$$

Figure 2: Distribution of Release of Substance Based on Month Occurred



Feature Selection

- LASSO Logistic Regression was used for feature selection, each feature was scored from least to most important.
- After the regression coefficients were shrunk, whichever features had a non-zero importance were selected to be used in our models.

Figure 3: Top 5 features selected from LASSO Logistic Regression

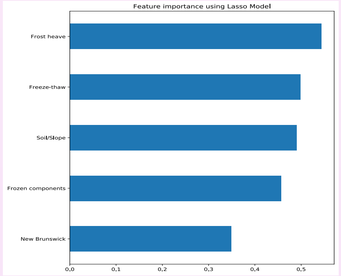
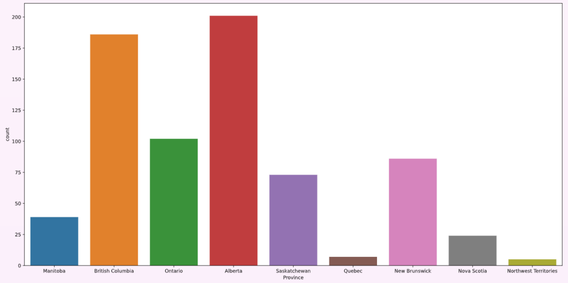


Figure 4: The number of incidents per Province in Canada



MODELS

Grid Search with 5 – Fold Cross Validation to pick the best parameters used for each model:

Models were evaluated using mean absolute error:

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

Random Forests

- Maximum features is set to automatic, after feature selection we have 43 variables
- Number of estimators is set to 500
- Used Gini Index as a criterion

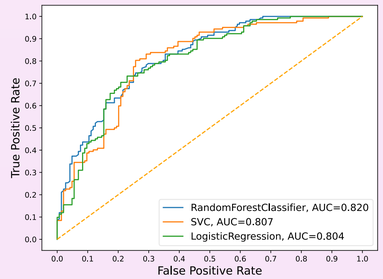
Support Vector Classifier

- Used Radial Basis Kernel
- Penalty parameter(C) = 1.2

Logistic Regression

- LASSO penalty

Figure 5: Receiver Operating Characteristic Curve of all the models with the Area under Curve Score



METRICS

Methods	Training Accuracy	Test Accuracy	Cross Validation Accuracy (5-fold CV)	Mean Absolute Error(MAE)	Area under Curve (AUC)
Logistic Regression	72.70%	73.43%	69.56%	0.2657	0.804
Random Forests	99.73%	74.47%	72.70%	0.2552	0.820
Support Vector Classifier (SVC)	81.10%	77.27%	72.53%	0.2272	0.807

CONCLUSION

The **Support Vector Classifier** has the **lowest MAE score** and **one of the highest AUC scores**. Our dataset is mostly sparse, which is why Random Forest overfitted. The most important features are frost heave, freeze-thaw, and other frozen components. We believe that our analysis prompts the possibility of using Support Vector Classifier to predict the release of substances at a given location.

REFERENCES

Government of Canada, C. E. R. (2020, November 7). Canada Energy Regulator / Régie de l'énergie du Canada. CER. https://www.cer-rec.gc.ca/en/safety-environment/industry-performance/interactive-pipeline/.

Government of Canada, C. E. R. (2021, April 15). Canada Energy Regulator / Régie de l'énergie du Canada. CER. https://www.cer-rec.gc.ca/en/safety-environment/industry-performance/interactive-pipeline/incident-data.html.

Acknowledgement: Funding support from NSERC