IEE 577: Data Science And Decision Systems

Portfolio Report

Kirity Ganga
MS in Computer Science
Arizona State University
kganga@asu.edu 1222577863

I. Introduction

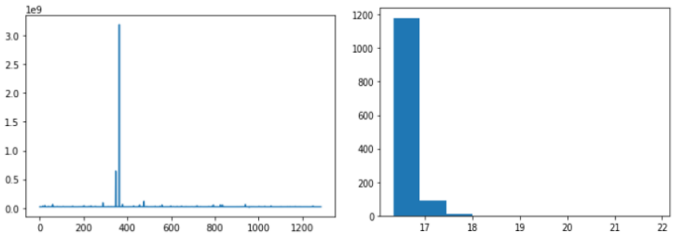
The occurrence of accidents due to Gas Leakage in factories may result in major economic and fatal losses. Therefore, it is necessary to accurately predict accidents. Several machine learning models have been applied to predict accidents under various time intervals and finally, an optimal model was selected by comparing the results of different models in terms of their accuracy, precision, recall, and F1 score. The results show that out of Bayesian networks and a few ensemble techniques, the Decision Tree and gradient boosting model exhibit the highest performance in terms of all the performance metrics. The root cause of the failure is predicted by using various embedding techniques such as TFIDF Vectorizer, Doc to Vector, BERT small, BERT large embedding, and for classification Naïve Bayes Classifier, fully connected Neural networks are used.

II. Description Of The Solution

Taking other factors into account as dependent variables in the data set, the goal is to predict the number of injuries/fatalities for a particular accident. The goal of this study is to predict the number of injuries/fatalities for a given accident by considering other factors as dependent variables in accidents.

Dataset **:** In the event of a pipeline accident or incident, pipeline operators must submit an incident report within 30 days to PHMSA. In-depth information about the location, facility and operation information, and the cause of the accident or incident are the key pieces of information collected. Specific information includes the time and location of the incident, number of any injuries and/or fatalities, commodity spilled/gas released, causes of failure, and evacuation procedures. Data sets are by following system types.[1] These reports published by the Department of Transportation (DOT) from "January 2010" are considered for our study.

Data Pre − processing **:** Before pre-processing the data the data had to cleaned by removing the null values from the dataset. Because the data has a lot of missing values we first limit our data to a certain number of columns that are useful for the prediction. Going through the external knowledge provided in reference[1] provided by the PHMSA itself, dimensions of the data were reduced by truncating few columns from the dataset that served no useful purpose in predicting the independent variables. After truncating the dataset, based on the assumptions null values were removed from the dataset. Then binning operation was performed where the cost data is segregated into bins i.e high cost and low cost. First, sum all the cost values in the dataset to measure the total damage cost.



It was observed that the binning of the high values and low-cost values was difficult as the appropriate threshold was difficult to measure. The total cost is thus set into the logarithmic filter to smooth the curve data(smoothing) and calculate the appropriate threshold. After choosing the appropriate threshold total cost is binned into two categories i.e high cost(1) and low cost(0) respectively. For classification of injuries and fatalities, if the count of the number of people injured and were fatal is more than zero, then the tuple is categorized as 1 else 0.

	Problem1	Problem2
Input	COMMODITY_RELEASED_TYPE,ACCIDENT_IDENTIFIER, FLOW_CONT_KEY_CRIT_IND, FLOW_CONT_MAIN_VALVE_IND, FLOW_CONT_SERVICE_VALVE_IND, LOW_CONT_METER_REG_IND, FLOW_CONT_EXCESS_FLOW_IND, FLOW_CONT_SQUEEZE_OFF_IND, FLOW_CONT_STOPPLE_FITNG_IND, IGNITE_IND, EXPLODE_IND, FEDERAL, LOCATION_TYPE, INCIDENT_AREA_TYPE, CROSSING, PIPE_FACILITY_TYPE, MATERIAL_INVOLVED, RELEASE_TYPE, EMPLOYEE_DRUG_TEST_IND, CONTRACTOR_DRUG_TEST_IND, INTERNAL_EXTERNAL, NATURAL_FORCE_TYPE, OUTSIDE_FORCE_TYPE	"Narrative" Column
Output/Target	Number of Fatalities/ Number of Injuries	"Cause" Column

III. Models Discovered:

Approach **1** From the visualization of the dataset, Since the dataset is unstructured and the dependent classes are skewed, linear algorithms were not used as it was clearly seen that the relationship between dependent and independent variables is not linear. It was also decided to choose distance invariant algorithms as the relation between the dependent and independent variables was uncertain and we have a large dataset. This is why Decision Tree algorithm was chosen as

Identify applicable funding agency here. If none, delete this.

base model. Ensemble models were inspired by decision tree algorithms that involve high interpretability(involves bagging and boosting) and finally, we also chose a probabilistic model as well to derive relations in data. As described earlier, SMOTE was applied to the dataset for training to overcome an imbalanced data issue and enhance the model performances:

Random Forest: Random Forest is based on the bagging algorithm and uses the Ensemble Learning technique. A subset of the data is used to create as many trees as possible and the output is combined. Using this method, decision trees are less likely to overfit, and the variance is also reduced, so accuracy is improved.

Decision Trees: A decision tree helps structure an algorithm in machine learning. Using a cost function, a decision tree algorithm will be used to split dataset features. Pruning is a process of pruning branches that may use irrelevant features from the decision tree before it is optimized.

Gradient Boosting: Gradient boosting is used to build a collection of predictors. In this approach, learners are taught sequentially by fitting simple models to the data and then analyzing the data for errors. At every stage, the goal of fitting successive trees (random sample) is to increase accuracy over the previous three.

```
9            Bayesian Network COST PREDICTION

100% |████████████████████████████████████████| 319/319 [1:04:15<00:00, 12.18s/it]
              precision    recall  f1-score   support

           0       0.63      0.70      0.66       179
           1       0.71      0.65      0.68       207

    accuracy                           0.67       386
   macro avg       0.67      0.68      0.67       386
weighted avg       0.68      0.67      0.67       386

Accuracy: 0.6735751295336787
[[125  54]
 [ 72 135]]
```

Bayesian Networks: A Bayesian Network falls under the category of Probabilistic Graphical Modelling (PGM) technique that is used to compute uncertainties by using the concept of probability. Popularly known as Belief Networks, Bayesian Networks are used to model uncertainties by using Directed Acyclic Graphs (DAG).[2] Using the Hill Climbing algorithm first it was asserted that the most optimized Bayesian network that can be architectured using the given data and then use that network to train the data and predict the output.

***Approach* 2 :** Few sate of art models for classification involving Natural Language Processing(NLP). For the NLP classification problem, word Embedding is the initial step followed by classification techniques. Word Embedding is the process of converting words into vectors by which the model can do further computations. In the present problem, dataset has sentences that need to be converted into vectors. So different embedding techniques were used to convert sentences into vectors.

The embedding vector size for Doc to Vector is 40 followed by fully connected Neural Network for classification resulting in a smaller number of trainable parameters i.e. 287.

Vectorizer Used	Classification	No trainable Parameters
TF-IDF	Multi Nominal naïve Bayes	-
Doc to Vec	Fully Connected Neural Network	287
BERT Large	Fully Connected Neural Network	5,383
BERT Small	Fully Connected Neural Network	903

It was observed that the trainable parameters for Bert large are the highest because Bert large gives the highest pooled output embedding layer with a vector of size 756 and a fully connected dense layer is used for classification. Hence the no of trainable parameters is 5,383 trainable parameters. Although a drop out of 0.3 is used to avoid overfitting since the number of trainable parameters is very high when compared with the data set size the model is overfitting. This resulted in low F1 scores.

IV. Hyperparameter Tuning

The model hyperparameters were the learning rate, max depth, max features, min samples leaf, min samples split, and n-estimators used for Decision trees and ensemble techniques, For Bayesian networks, the Hill-Climbing algorithm has been used to get the most optimized graph. To determine the model accuracy in the training process, a 10-fold cross-validation method was applied to the training dataset.

Target Variable	Optimal Model	Parameters	F1 Scores
BOOLEAN COST	Gradient Boosting algorithm	learning rate= 0.05 n_estimators= 500	The highest F1 score is 0.71
BOOLEAN INJURY	Decision tree algorithm	Criterion= 'Gini', Max_depth= 10, Min_samples_leaf= 1, Min_samples_split= 4	F1 Score for minor class is 0.44
BOOLEAN FATALITY	Gradient Boosting algorithm	Learning rate = 0.05 n_estimators = 500	F1 Score for a minor class is 0.1

As expected average accuracy of the models in the training phase were around 80-90%, except for the accuracy of the Bayesian network. This finding shows that the model performance is acceptable in terms of its accuracy.

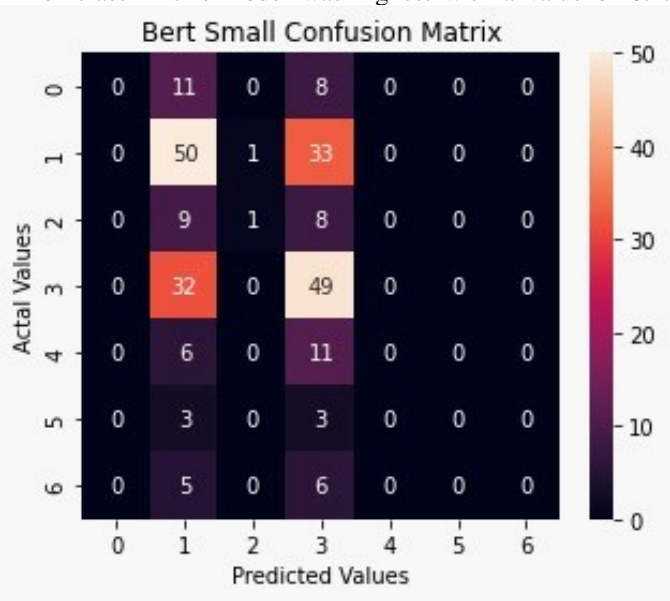
In the testing phase, the model performance was evaluated using the test data. The model performance indicators, specifically, the accuracy, precision, recall, and F1 score, were calculated.

NLP Model	Training Accuracy
Doc to Vec	58.70%
Large Bert	37.58%
Small Bert	41.30%

V. Results

It was concluded that the most optimistic model for predicting BOOLEAN COST is the gradient boosting algorithm with parameters learning_rate= 0.05 and n_estimators= 500. The F1 score in this model was highest with a value of 0.71. It was concluded that the most optimistic model for predicting BOOLEAN INJURY is the decision tree algorithm with parameters criterion= 'Gini', max_depth= 10, min_samples_leaf= 1, and min_samples_split= 4. The F1 score for minor class in this model was highest with a value of 0.44. It was concluded that the most optimistic model for predicting BOOLEAN

FATALITY is the gradient boosting algorithm with parameters learning_rate= 0.05 and n_estimators= 500. The F1 score for minor class in this model was highest with a value of 0.1.



It was referenced that the INVESTOR_OWNED column has the maximum importance in predicting BOOLEAN COST for a given tuple. It was referenced that the EXPLODE_IND column has the maximum importance in predicting BOOLEAN FATALITY for a given tuple. It was referenced that the IGNITE_IND column has the maximum importance in predicting BOOLEAN INJURY for a given tuple.

CONTRIBUTION

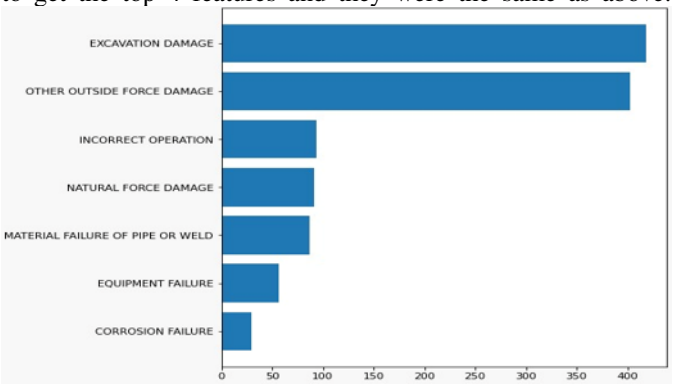
In order to assign responsibilities to my project partners, I first had to read over the issue description and fully comprehend it. After reviewing the data, I made sure I understood each attribute within the dataset and if any pre-processing or data cleaning was necessary for any of them. I used exploratory data analysis to uncover variables that are associated with a factory's dependent variables by comparing each column separately from pictorial representations.

After using a variety of visualization techniques, I modeled different Machine Learning models and performed hyper parameter tuning on them to improve their performance. I was assigned to model Approach 1 of the project where I had to model different machine algorithms and evaluate their performance. The project development was broken down into the following stages:

1. Exploratory Data Analysis:
2. Data Processing Predictive Modeling
3. Hyperparameter Tuning

LESSONS LEARNED

Starting with, as it was observed from all of the plots and analyses that the data is unstructured and highly skewed and thus it was over sampled and fed to machine learning models. We also had fitted a Random Forest on the dataset to get the top 4 features and they were the same as above.



As a result of this project and course, I learned how to tell a story using a dataset of numbers and words. There is a lot of information in the dataset, and thus I learned to feature engineer te models so that the prediction are done with highest accuracy. Different evaluation metrics were discovered and their significance was learnt. We chose F1 score as the evaluation metric at the end and improved it overtime.

LIST OF TEAM MEMBERS

This project had a three-person team, with

Kirity Ganga
Madhura Kolar Lakshmi
Venkata Sai Pramod Kuar Kasturi as team members.

REFERENCES

[1] https://catalog.data.gov/dataset/pipeline-accident-incident-reports
[2] https://www.edureka.co/blog/bayesian-networks/
[3] https://seaborn.pydata.org/
[4] https://numpy.org/
[5] https://scikit-learn.org/stable/