

**NANYANG TECHNOLOGICAL UNIVERSITY**

**AY23/24 SEMESTER 1**

**BC2406: Analytics I: Visual & Predictive Techniques**

**Guardian Analytics: Comprehensive Pipeline Safety Solutions**

**Submitted by:**

|  |  |
| --- | --- |
| **Group Members** | **Matriculation Number** |
| AMIRUL HAZIQ BIN MUHAMMAD EFFENDY | U2210956K |
| ANG JO WEE | U2220795K |
| IAN KOH JIN | U2221932B |
| PANG BOSLYN | U2221298A |
| CHIANG QIN ZHI | U2221670K |

**Seminar Group Number:** S03

**Team Number:** T01

**Table of Contents**

[**Executive Summary 3**](#_5e0qlt30q52c)

[**1. Business Understanding 1**](#_mbra7aynk3mg)

[1.1 Business Problem 1](#_cxchhhcjuq7n)

[1.2 Existing Solutions Overview 1](#_aa5y7jhb27ca)

[1.3 Opportunity Statement 1](#_87cktyob0923)

[1.4 Methodology 2](#_xa5nveq0tonh)

[**2. Data Cleaning 3**](#_ars7srrm6tb5)

[**3. Data Exploration 5**](#_r230bzeb8omn)

[3.1 Univariate Analysis 5](#_6c5czfi2hblb)

[3.2 Bivariate Analysis 5](#_s39bke8bkfyh)

[**4. Modelling 7**](#_2xh5fh1iz7ef)

[4.1 Predicting Fatalities 7](#_p9reh6pvpagj)

[4.1.1 Logistic Regression Model 7](#_56jrd1oyebxk)

[4.1.2 Classification and Regression Trees (CART) 8](#_mb2hkepq2mkb)

[4.2 Predicting Injuries 8](#_o588acqbmj9u)

[4.2.1 Logistic Regression Model 9](#_fw5szrik7xpq)

[4.2.2 Classification and Regression Trees (CART) 9](#_d07sd31wryep)

[4.3 Predicting Percentile of Total Costs Incurred 10](#_1tuf1rttyart)

[4.3.1 Linear Regression Model 10](#_4tevmhd5zg5r)

[4.3.2 CART 11](#_v89ydqkvbted)

[4.4 Model Selection 12](#_qlklx8ow1cb)

[4.4.1 Categorical Variables (Fatalities & Injuries) 12](#_m07mpjwjdpgb)

[4.4.2 Continuous Variables (newcost\_Percentile) 13](#_izphd3lqsz51)

[**5. Business Implementation 15**](#_slly65b3lfpt)

[5.1 Recommendations 15](#_b823zpvvyl5u)

[5.1.1 Pre-Accident Solutions 15](#_637r0ptv069a)

[5.1.2 Post-Accident Solutions 16](#_se2ik7jan2om)

[5.2 Expected Outcomes 17](#_5k6fvqhk3ldg)

[5.3 Comparison with Current Solutions 18](#_wodf46f8z1dr)

[5.4 Limitations and Future Improvements 18](#_w9e7thealodh)

[**6. Conclusion 19**](#_lfyvlro1r0iv)

[**References 20**](#_rbe0ke61jf9y)

[**Appendices 22**](#_urw0158xlusl)

[Appendix A: Data Dictionary (Cleaned Data) 22](#_xux18cauyrk7)

[Appendix B: Data Exploration 24](#_7fni4il8vic9)

[Appendix C: Modelling 29](#_fm7j529hdfc7)

### 

### **Executive Summary**

Saudi Aramco, a global leader in the petroleum and natural gas industry, holds the world’s second-largest proven crude oil reserves and is the top oil producer daily. Despite having robust safety campaigns and systems in place, the company is dedicated to further enhancing workplace safety in the face of global increases in workplace accidents, which pose significant human and financial risks. To combat these challenges, Aramco has initiated various safety measures, including a multilingual safety campaign, a Safety Management System aligning with international standards, the SafeLife application and is exploring advanced technologies to ensure safer operations.

The project aims to leverage advanced data analytics to identify factors contributing to high accident rates in the oil pipeline sector, with the ultimate goal of predicting accident-prone scenarios to reduce accidents, costs, and enhance safety. The approach includes five stages: data cleaning, exploratory data analysis, variable selection, model building and presenting business solutions. The dataset used is “Oil Pipeline Accidents 2010 - Present” dataset on Kaggle and cleaned prior to analysis. The project's predictive models, which show high accuracy for injury and fatality predictions, suggest significant predictors including 'Cause Category', 'Liquid Type', and 'Pipeline Shutdown'.

From these factors, we propose several recommendations to reduce accidents and in turn reduce costs. These solutions complement Aramco’s existing safety management systems by offering pre- and post-accident strategies, such as infrastructure reinforcement, increased emergency drills, automation, targeted educational programs, and a Risk Assessment Matrix for rapid response.

These proactive strategies aim to provide significant improvement over current safety measures. However, the analysis is constrained by the limited public data and the inherent imbalanced data with regards to injuries and fatalities. Future improvements include securing more detailed data from Aramco to refine the predictive models further to the company’s specific operational context.

### 

### **1. Business Understanding**

#### **1.1 Business Problem**

Workplace safety plays a crucial role in a business organisation as it encompasses Standard Operating Procedures (SOPs), implementation of safety protocols, use of Personal Protective Equipment (PPE) and other policies that ensures a safe working environment.

The global oil industry, including giants like Aramco, has faced challenges related to workplace accidents, with repercussions ranging from loss of life to massive economic setbacks (Passwaters & Ugal, 2022).

Globally, there has been an increase in workplace accidents, resulting in fatalities and large economic costs for corporations. The International Labour Organisation (ILO) has reported that there are an estimated 340 million reported cases of workplace accidents annually and it is predicted that the actual number of accidents is significantly higher due to underreporting (ILO, 2011). From 2020 to 2021 alone, despite a 5% increase in work hours reported, the number of fatalities has increased by 42% (HSE Now, 2022).

Given the financial and reputational stakes, it becomes important for businesses, especially in high-risk sectors like oil, to invest not only in proactive safety measures, but also solutions that can be implemented after an accident occurs, so that economic losses are minimised and to prevent the safety of employees being compromised.

#### **1.2 Existing Solutions Overview**

A current solution implemented by Aramco is a multilingual safety campaign based on their Lifesaving Rules, aiming to mitigate over 90% of injury-causing incidents. Aramco's also has a Safety Management System (SMS) that is benchmarked with global industry standards with a strong safety culture. Aramco has also integrated digital safety solutions like the SafeLife mobile application that allows users to report hazards, near-misses and incidents, whilst also exploring innovative technologies like Auto Well Space Out initiative for safer operations (Aramco, 2021).

#### **1.3 Opportunity Statement**

Our project aims to analyse the causes and underlying factors linked to elevated accident rates to suggest safety measures that can be taken pre-accident and post-accident and predict the severity of accidents. This initiative would enhance safety protocols and minimise potential liabilities associated with workplace accidents. Our results will not only safeguard employee well-being but also significantly curtail operational disruptions, leading to considerable reductions in business expenditures and broader economic setbacks.

### 

#### **1.4 Methodology**

We will conduct our analysis in 4 different stages – data cleaning and EDA, selecting our Y variables, building our models through logistic regression, CART and linear regression, and finally, evaluating our models by analysing their performance metrics.

A diagram of a data analysis process

Description automatically generated***Figure 1.4:***  *Methodology Flowchart*

### **2. Data Cleaning**

Our dataset was found on [Kaggle](https://www.kaggle.com/datasets/usdot/pipeline-accidents), the “Oil Pipeline Accidents 2010- Present” Dataset (Department of Transportation, 2016). This database includes a record for each oil pipeline leak or spill since 2010, and includes important information like the number of injuries and fatalities, and associated costs . Since Kaggle is an open-source platform, its data has to be analysed, cleaned and processed carefully before proceeding to any kind of analysis.

***Table 2.1:*** *Cleaned Variables, Actions Taken and Rationale*

|  |  |  |
| --- | --- | --- |
| **Variables** | **Action** | **Reasoning** |
| Report Number/ Supplemental Number | Drop columns | Redundant as we can locate entries by the row number. |
| Operator ID/ Operator Name/ Pipeline/Facility Name/ Accident City/ Accident County/Accident State | Drop columns | Our analysis is not company-specific. |
| Accident Year | Data type changed to “factor” | Years should be a categorical variable. |
| Accident Date/Time | Data type changed to POSIXct format | POSIXct format will allow us to manipulate time-based operations easier. |
| Intentional Release (Barrels)/ Net Loss | Replace NA values with correct values, created a new column “Updated Net Loss (Barrels)” | 1. Intentional Release (Barrels) contains some N.A. values that are inaccurate. 2. Unintentional Release + Intentional Release - Liquid Recovery = Net Loss. 3. Assuming the other three variables are accurate, the rows with Intentional Release = NA are calculated and filled in respectively. 4. A new column called Updated Net Loss (Barrels) is then calculated with the prior relationship and cross-referenced with the original Net Loss (Barrels) to ensure that all values follow the prior relationship. |
| Liquid Name/ Liquid Subtype | Replace NA values with X and drop unused levels | There will only be Liquid Name if Liquid Subtype = ‘Other’ or ‘Other HVL’ and there will only be Liquid Subtype for certain Liquid Types.  We do not leave it as N.A. as the models omit rows with N.A. values during model training. This causes some levels to have no rows even after dropping unused levels during data cleaning, and causes an error. |
| Pipeline Shutdown | Replace NA and “” with “NO” | We can reference Shutdown Date/Time and Restart Date/Time variables and check if Pipeline Shutdowns occurred or not. |
| All Costs | Use 90% winsorization on `All Costs` (all data below 5th percentile above 95th percentile set accordingly) | Winsorization limits extreme values in the statistical data to reduce the effect of outliers. Due to the significant number of outliers in `All Costs`, and as winsorized estimators reduce outliers without deleting data, we decided to winsorize `All Costs` in a new column ‘newcost’. |
| Variables related to costs | Replace NA values with 0 | Upon inspection of the `All Costs` column, the values are accurate as it is the sum of all the Costs added up. |
| Restart Date/ Time / Shutdown Date/ Time | Replace missing values with NA, convert to POSIXct format, create new column ‘Shutdown Duration (Hours)’ | Missing values indicate that Shutdown / Restart did not occur, thus it is best to replace it with NA (Tierney, 2023). Analysing the duration of the shutdown is more important than the start and end time of the shutdown period. |
| Fatalities/ Injuries / Public Evacuations | Data type changed to categorical data type | We are interested in analysing if these occurred or not, not in the number as even a single occurrence is important. |
| All other variables concerning injuries/ public evacuations/ fatalities | Drop columns | These columns are too detailed and do not add value to this statistical study. |

### 

### **3. Data Exploration**

After the initial data cleaning is completed, data exploration is done to understand the dataset more effectively. The data dictionary of the dataset used can be found in [Appendix A](#_xux18cauyrk7). Key findings would be elaborated below and plots are found in [Appendix B](#_7fni4il8vic9).

#### **3.1 Univariate Analysis**

1. Accidents reported have been on an increasing trend, which is in line with ILO’s findings. This further supports the point that accidents in the oil and gas industry are increasingly prevalent and requires urgent attention. There was a lack of accident reports in 2017 in the dataset, as that was when the dataset was published.
2. The number of accidents starts to increase at 5:00 A.M. and peaks at around 8:00 A.M. This finding aligns with a study on workplace accidents, which reported that most workplace accidents happen 2 to 4 hours into the shift (US Bureau of Labor Statistics, 2016).
3. The cause category that had the greatest number of accidents reported is **MATERIAL/WELD/EQUIP FAILURE.** In this cause category, the highest number of accidents reported is the subcategory **PUMP OR PUMP-RELATED EQUIPMENT**. Equipment failures can cause disastrous accidents such as fires, explosions, and blowouts, threatening the safety of the employees (Kemmy Law Firm, 2022). Therefore, it is important for maintenance to be carried out regularly to ensure that equipment is safe-for-use, to reduce accidents caused by equipment failure.
4. The distribution of the variables Fatalities and Injuries are unbalanced as only 0.3% and 0.43% of the data had reported fatalities and injuries respectively. Thus, to handle the class imbalance, we introduced the use of class weights in our models to alleviate the skewed data.
5. The subcategory of cost that incurred the highest amount of cost is Environmental Remediation Cost. There is a large emphasis placed on environmental conservation, which is in line with Aramco’s aim to promote environmental awareness, as seen on the site.

#### **3.2 Bivariate Analysis**

1. The correlation matrix of the numerical variables in the dataset reported that the amount of liquid recovered and total costs incurred is positively correlated. It can be assumed that large-scale accidents would involve increased amounts of liquids released, thus the absolute amount of liquids recovered would most likely increase. The positive relationship suggests that recovery efforts are scaled based on the severity of the accident and the response plans deploy financial resources in proportion to the severity of the accident.
2. Furthermore, there is a positive correlation between liquid recovered (in barrels) and environmental remediation costs. This suggests that liquid recovery after an accident contributes to the high costs incurred to restore the environment.
3. The positive relationship between lost commodity costs and net liquid loss highlights the importance of preventing accidents and thus minimising the liquid lost in not only protecting the environment and public health, but also to reduce financial losses that are related to commodity loss.
4. The total costs incurred of accidents that caused a pipeline shutdown is significantly higher as compared to accidents that did not require a pipeline shutdown. Thus, it can be highlighted that pipeline shutdowns are extremely expensive and detrimental to the profits of the corporation.

### 

### **4. Modelling**

The objectives of modelling includes predicting the occurrence of fatalities and injuries when an accident occurs and the percentile of total costs, which would give insights to significant factors that affect these 3 variables and the severity of an accident after its occurrence. The same train-test datasets are used when comparing between models with the same response variable. We will analyse the models then evaluate the results through performance metrics.

#### **4.1 Predicting Fatalities**

We altered the data involving fatalities from one that presented how many fatalities were reported to binary categorical variables, which showed whether there were any fatalities or not. This is because fatalities involve the well-being of employees and we are less interested in predicting the amount, but rather if there were any fatalities and injuries when an accident is reported.

Since the distribution of the response variable “Fatalities” is imbalanced, (99.7% FALSE & 0.3% TRUE, [Appendix C](#_fm7j529hdfc7)) weights were used to handle the imbalance data when building the models. To reduce the risks of biased predictions, higher weights are assigned to the minority class (Fatalities = TRUE), allowing the models to pay more attention to its pattern (Singh, 2023).

To train the models, the dataset was randomly divided into train-test datasets with a ratio of 7:3. The train-test datasets have been observed to have similar distributions for the different variables, ensuring similarity in distribution between the two.

The predictor variables used to develop the models from the dataset are: Cause Category, Pipeline Location, Pipeline Type, Liquid Type, Liquid Subtype, Shutdown Duration (Hours), Liquid Ignition, Liquid Explosion, Evacuations, Injuries and Updated Net Loss (Barrels).

##### 4.1.1 Logistic Regression Model

The logistic regression model was developed using the training dataset and class weights were implemented. Through the logistic regression model we discovered that the statistically significant factors in predicting the occurrence of fatalities was whether the liquid subtype was **OTHER HVL** and whether the cause category was due to **OTHER OUTSIDE FORCE DAMAGE** ([Appendix C](#_fm7j529hdfc7)).

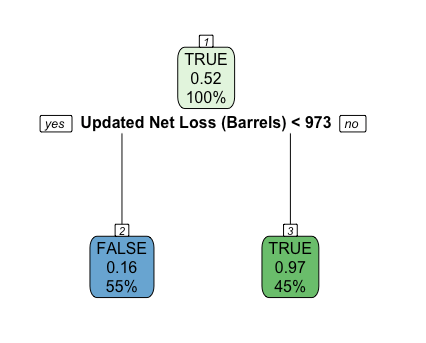
Testing the model on our test set, the confusion matrix can be seen in Table 4.1.1 which compares the actual and predicted values of Fatalities.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Predicted** | |
| *False* | *True* |
| **Actual** | *False* | 822 | 14 |
| *True* | 1 | 1 |

***Table 4.1.1:*** *Confusion Matrix For Test Set (Logistic Regression)*

##### 4.1.2 Classification and Regression Trees (CART)

To develop the CART, we first grew the tree to its maximum and found the appropriate Complexity Parameter (CP) to prune the maximal tree to its optimal tree. The appropriate CP computed was 0.3447 ([Appendix C](#_fm7j529hdfc7)) and it was used to prune the maximal tree. The CART below is the optimal tree for predicting Fatalities with 1 split and 2 child nodes.



***Fig 4.1.2:*** *Optimal CART To Predict Fatalities*

Table 4.1.2 presents the confusion matrix when the CART model was used to predict the testing dataset.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Predicted** | |
| *False* | *True* |
| **Actual** | *False* | 807 | 29 |
| *True* | 1 | 1 |

***Table 4.1.2:*** *Confusion Matrix For Test Set (CART)*

This CART may not be helpful in situations where an accident just occurred and predicting the occurrence of fatalities needs to be done swiftly to plan for the emergency response. This is because the most important predictor variable of the CART model is ‘Updated Net Loss (Barrels)’. Information about the Net Loss may not be readily available when the accident is still fresh.

#### **4.2 Predicting Injuries**

The treatment of the response variable “Injuries” is the same as “Fatalities” in Section 5.1. The variable shows the occurrence of injuries, and the distribution is imbalanced (99.6% FALSE & 0.4% TRUE, [Appendix C](#_fm7j529hdfc7)), thus class weights were implemented as well, to alleviate the bias towards the majority class. The models developed to predict Injuries were also trained and tested by randomly dividing the dataset into train and test data sets in the ratio of 7:3.

The predictor variables used to develop the models from the dataset are: Cause Category, Pipeline Location, Pipeline Type, Liquid Type, Liquid Subtype, Shutdown Duration (Hours), Liquid Ignition, Liquid Explosion, Evacuations, Fatalities, Updated Net Loss (Barrels).

##### 4.2.1 Logistic Regression Model

Table 4.2.1 is the resulting confusion matrix when the logistic regression model is used to predict the occurrence of injuries based on the testing set. It can be seen from the table that the model is able to accurately identify the occurrence of injuries in all the accidents that did have injuries reported.

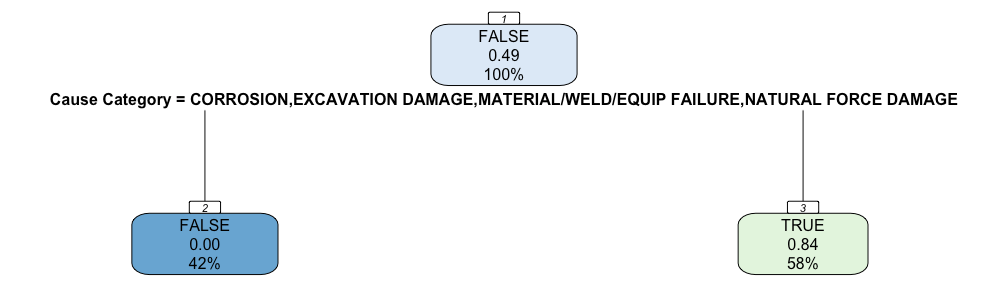
|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Predicted** | |
| *False* | *True* |
| **Actual** | *False* | 787 | 48 |
| *True* | 0 | 4 |

***Table 4.2.1:*** *Confusion Matrix For Test Set (Logistic Regression)*

Additionally, our logistic regression model gave insights to which factors are statistically significant in predicting the occurrence of injuries. The statistically significant factors are whether the cause of an accident is due to **INCORRECT OPERATION** of equipment and if the liquid type was **HVL OR OTHER FLAMMABLE OR TOXIC FLUID, GAS** ([Appendix C](#_fm7j529hdfc7)).

##### 4.2.2 Classification and Regression Trees (CART)

Using the train set used to develop the logistic regression in Fig 4.2.1, the CART was grown to its maximal tree and then pruned to its optimal using the appropriate CP value of 0.201 ([Appendix C](#_fm7j529hdfc7)). The optimal tree consists of 1 decision split and 2 child notes. Table 4.2.2 shows the comparison between actual and predicted values when the CART is used to predict the occurrence of injuries based on the test set.



***Fig 4.2.2:*** *Optimal CART to Predict Injuries*

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Predicted** | |
| *False* | *True* |
| **Actual** | *False* | 650 | 185 |
| *True* | 2 | 2 |

***Table 4.2.2:*** *Confusion Matrix For Test Set (CART)*

The variable with highest importance in the CART model is ‘Cause Category’ ([Appendix C](#_fm7j529hdfc7)). The optimal CART (Fig 4.2.2) is helpful in situations where predicting the occurrence of injuries need to be done in a quick manner and there is a lack of information regarding the accident.

#### **4.3 Predicting Percentile of Total Costs Incurred**

Initially, the response variable was the total cost of an accident. However, when developing the CART model, the optimal tree only had a single node (root node). This was due to increasing cross-validation error as the number of splits increases as well as a high standard error. Therefore, we decided to predict the percentile of the cost rather than the absolute cost. This resulted in better models, as converting it to percentile negated some of the skewness, similar to a logarithmic transformation. Furthermore, as the data for costs has an extremely large range with large amounts of outliers, we decided to use the winsorization to limit the extreme values, reducing the effects of outliers, as discussed in section 2.3.

Additionally, it has been reported that predicting percentiles may be more advantageous than predicting absolute values in the presence of outliers (Benoit & Van den Poel, 2009). Percentile values are able to provide insights on how the predicted cost incurred for a particular accident compares to other accidents. Furthermore, predicting the absolute value of the cost does not take into account inflation and changes in prices throughout the years.

The dataset was split into the training and testing dataset in the ratio of 7:3 and percentiles of the total costs (newcost\_Percentile) were added as an additional column for each dataset.

##### 4.3.1 Linear Regression Model

We started developing the linear regression model by using the majority of the variables (excluding newcost\_Percentile) as the predictor variables. The predictor variables used were Cause Category, Pipeline Location, Pipeline Type, Liquid Type, Liquid Subtype, Shutdown Duration (Hours), Liquid Ignition, Liquid Explosion, Evacuations, Injuries, Fatalities, Updated Net Loss (Barrels) and Pipeline Shutdown.

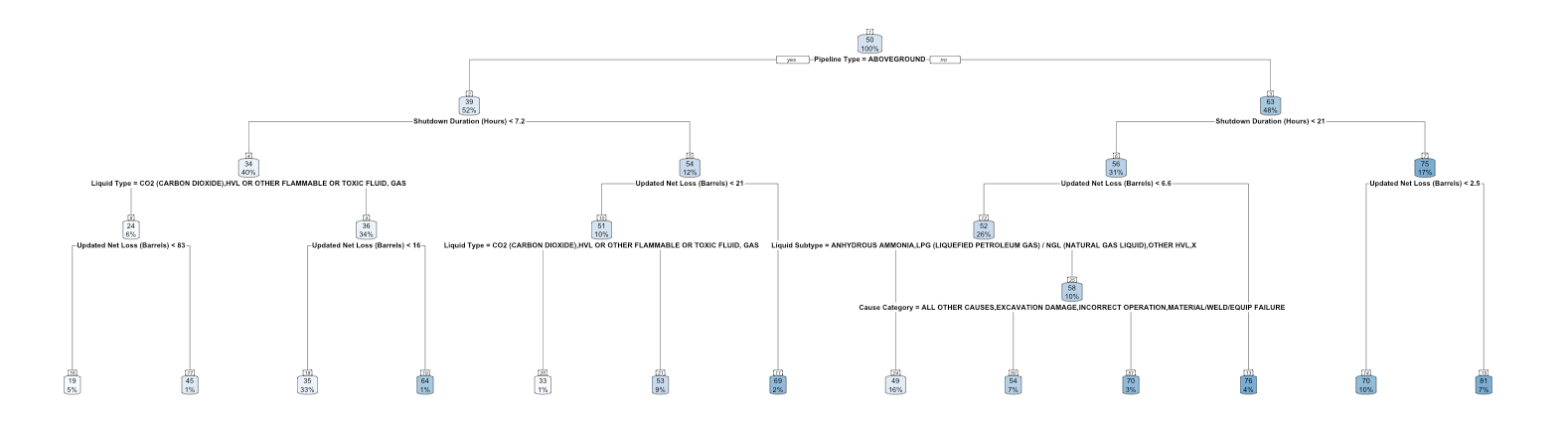
Then, using Akaike’s Information Criterion (AIC) we eliminated variables that did not help in the linear regression model. The next step was to observe the Variance Inflation Factor (VIF) of the remaining variables to remove variables that suffer from multicollinearity. However, no variables were removed from this step as the VIF values of all variables were below the threshold ([Appendix C](#_fm7j529hdfc7)).

The final linear regression model’s predictor variables were Cause Category, Pipeline Type, Liquid Type, Shutdown Duration (Hours), Liquid Explosion, Evacuations, Injuries, Updated Net Loss (Barrels) and Pipeline Shutdown.

##### 

##### 4.3.2 CART

Using the same train and test sets as the linear regression model above and X variables used in the initial linear regression (before AIC), we first grew the CART to the maximum and pruned the maximum tree subsequently to its optimal tree using the CP value 0.00499 ([Appendix C](#_fm7j529hdfc7)). The maximal tree had 55 splits and was pruned to the optimal tree which has 13 splits, to prevent overfitting of the model to the train set. Fig 4.3.2 presents the optimal tree to predict the percentiles of cost incurred. A higher definition picture of the optimal tree can be found in [(Appendix C)](#_fm7j529hdfc7).



***Fig 4.3.2:*** *Optimal Tree Predicting Percentile of Cost Incurred*

The optimal CART was then used to predict the percentiles of total cost incurred for the test set. Additionally, the model also provided insights to the important variables in predicting the percentile of cost. In this model, Pipeline Type is the most important variable ([Appendix C](#_fm7j529hdfc7)) in predicting the percentile of total costs incurred.

#### 

#### 

#### **4.4 Model Selection**

After developing the models for the response variable - Fatalities, Injuries and newcost (in percentile), we would be evaluating the models.

Table 4.4.1 contains the metrics that we would be using to evaluate our models.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Categorical Variables** | | | | | | | |
| **Predicted Variable** | **Model** | **Train Set** | | | **Test Set** | | |
| ***Accuracy*** | ***True +*** | ***True -*** | ***Accuracy*** | ***True +*** | ***True -*** |
| **Fatalities** | Logistic Regression | 98.9% | 100% | 99% | 98.2% | 50% | 98.3% |
| CART | 96.8% | 83.3% | 96.9% | 96.4% | 50% | 96.5% |
| **Injuries** | Logistic Regression | 95.9% | 100% | 95.8% | 94.3% | 100% | 94.2% |
| CART | 81.7% | 100% | 81.6% | 77.7% | 50% | 77.8% |
| **Continuous Variable** | | | | | | | |
| **Predicted Variable** | **Model** | **Train Set** | | | **Test Set** | | |
| ***RMSE*** | ***R2*** | | ***RMSE*** | | ***R2*** |
| **newcost (in percentile)** | Linear Regression | 24.52 | 0.275 | | 23.63 | | 0.337 |
| CART | 23.63 | 0.347 | | 22.51 | | 0.400 |

***Table 4.4.1:*** *Model Performance Metrics*

##### 4.4.1 Categorical Variables (Fatalities & Injuries)

From Table 4.4.1, it can be seen that in predicting categorical response variables, the logistic regression and CART models are highly accurate. However, only the logistic regression model that predicts the occurrence of injuries is highly sensitive. Sensitivity is measured using True Positive (TP) Rate. A highly sensitive model would be able to predict true positives. As for specificity, which is measured using the True Negative (TN) rate, all models have ratings above 75%. Therefore, the specificity of all models predicting the categorical variable is of a certain calibre.

Since the accuracy of all models when used to predict the test set is similar to the accuracy when the train set was used, we can be sure that the model does not suffer from underfitting or overfitting. Comparing the 3 metrics used to decide on the better performing model, the logistic regression models for predicting the occurrence of fatalities and injuries outperforms the CART model.

##### 4.4.2 Continuous Variables (newcost\_Percentile)

As for predicting newcost (in percentile), linear regression and CART were used. In addition to the 2 metrics used to evaluate the linear regression and CART models as seen in Table 4.4.1, model diagnostics are also carried out for the linear regression model (Table 4.4.2).

|  |  |
| --- | --- |
| Assumption that there is a linear association between the predictor and response variables hold. Additionally, since the red line is relatively constant at 0, the errors have a normal distribution with mean 0. | Standardised residuals are normally distributed as they fall on the 45-degree reference line. Thus, the assumption that errors are normally distributed are met. |
| As points are distributed with the same spread at each vertical slice, the assumption that errors are independent of the predictor variables and has a constant standard deviation holds. | No influential outliers. |

***Table 4.4.2:*** *Linear Regression Diagnostic Plots*

As seen in Table 4.4.1, the CART model performs better in all areas as compared to the linear regression model in predicting the percentile of the total costs incurred. The R2 value of the CART model when the test set was used is around 0.5, which suggests that half of the variability in the predicted data can be explained by the model. Therefore, the CART model is a better model as compared to the linear regression model as the R2 value of the linear regression model is lower than CART’s suggesting that a higher portion of the variability in the predicted data cannot be explained by the model. Additionally, the Root Mean Squared Error (RMSE) of the CART model for both the train and test sets are higher than the RMSE when linear regression was implemented.

### **5. Business Implementation**

#### **5.1 Recommendations**

We will split our solutions into two parts, pre-accident for solutions preventive in nature, and post-accident for solutions meant to reduce the severity of accidents that have already occurred.

##### **5.1.1 Pre-Accident Solutions**

Fatalities

The Logistic Regression model has illuminated critical variables that contribute to fatalities and injuries within the company's operations. The most statistically significant variables from our Logistic Regression model to predict Fatalities was when the categorical `Cause Category` was **OTHER OUTSIDE FORCE DAMAGE**.

Injuries

Meanwhile in our Logistic Regression Model for Injuries, the most statistically significant variables with the greatest magnitude of effect was when `Cause Category` was **INCORRECT OPERATION**, `Liquid Subtype` was **OTHER HVL**, and `Liquid Explosion` was also **Yes**.

|  |  |
| --- | --- |
| **Predicted**  **Variable** | **Significant Factors** |
| Fatalities | Cause Category: **OTHER OUTSIDE FORCE DAMAGE** |
| Injuries | Cause Category: **INCORRECT OPERATION**  Liquid Subtype: **OTHER HVL**  Liquid Explosion: **Yes** |

***Table 5.1.1:*** *Significant Factors for CART Predictions*

Recommended Solutions

According to the U.S. Department of Transportation, **OTHER OUTSIDE FORCE DAMAGE** to pipelines can include vandalism, sabotage or terrorism, vehicle or equipment contact not related to excavation, such as an automobile crash into a piece of pipeline equipment, or damage caused by accidents or fires from other businesses or industries that are nearby (U.S. Department of Transportation, 2014). To fortify against **OTHER OUTSIDE FORCE DAMAGE**, Aramco should reinforce its infrastructural resilience, particularly in areas where environmental and mechanical stressors are prevalent. This involves the integration of early detection systems equipped with sensors that can alert potential threats, such as heavy machinery encroaching on pipeline zones or abnormal seismic activities. Furthermore, the implementation of geofencing technology around critical infrastructure promises to trigger instant alerts, thereby averting potential damage by unauthorised activities.

Meanwhile for the significant factor **OTHER HVL**, HVL is an acronym for Highly Volatile Liquids. HVLs which are hazardous liquids which will form a vapour cloud when released into the air (Code of Federal Regulations, 2023). Thus, they necessitate specialised attention. The deployment of emergency response drills tailored to HVL-related incidents can be increased, bolstering the company’s readiness to respond to any accidents effectively. Furthermore, by increasing the use of robotics for automation to handle routine inspections and operations, especially in hazardous conditions, reduce human exposure to risk and also manpower costs.

To combat **INCORRECT OPERATION**, Aramco must design and implement targeted educational programs. These initiatives include simulation-based training and the application of human factors engineering to streamline operational processes, thereby reducing the likelihood of human error.

Liquid explosion risks can be countered with high-sensitivity detectors and automated shutdown systems, designed to react instantaneously to danger signs. Improved fire suppression technologies can be deployed to extinguish fires swiftly, minimising the risk to human life and infrastructure.

##### **5.1.2 Post-Accident Solutions**

In order to reduce the cost of accidents, we suggest greater planning of emergency response procedures for different tiers of risks. Utilising a Risk Assessment Matrix allows quick prediction of likely accident outcomes and allows Aramco to react to such accidents, thus mitigating further fatalities, injuries and costs. Aramco personnel can then use the results of the Risk Assessment Matrix to determine the appropriate actions to be taken according to the risk score.

Risk Assessment Matrix

When an accident occurs, Aramco can utilise a risk assessment matrix to predict the risk of an accident and determine the appropriate actions to take immediately.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **Predicted Severity** | | |
| No Harm: 1 | Injuries Only: 2 | Fatalities: 3 |
| **Predicted Cost (in Percentile)** | Cost < 25: 1 | *1* | *2* | *3* |
| 25 < Cost < 75: 2 | *2* | *4* | *6* |
| Cost > 75: 3 | *3* | *6* | *9* |

***Table 5.1.2:*** *Risk Assessment Matrix*

We use a scoring system to combine multiple factors into a single risk score that can be more easily categorised and visualised. This approach allows for a multi-faceted risk assessment while maintaining a format that is straightforward to interpret and act upon. The risk score is calculated by multiplying the likelihood score and the severity score. Risk scores below 3 are low risk and highlighted in green, risk scores equal to 3 or 4 are medium risk and highlighted in orange, and risk scores above 4 are high risk and highlighted in red.

To determine severity, Aramco personnel can use the CART and Logistic Regression models in Section 4.1 and Section 4.2 to determine if a fatality or injury will occur as a result of an incident. For example, when `Cause Category` is **CORROSION**, an injury is predicted to occur. To determine cost in percentile, CART model can be used to determine predicted cost of the accident in percentile. This risk matrix can be implemented as a digital system, where the Cost and Severity is automatically calculated using the models which are required for the risk score.

Tiered Emergency Responses

Aramco can utilise the results of the Risk Assessment Matrix to have a tiered emergency response system based on the risk score output. We propose a three-tiered system designed to effectively manage emergencies, encompassing low, medium, and high-risk accidents.

Low-risk accidents typically include minor incidents such as small leaks or equipment malfunctions These incidents, while not posing a significant threat, still require immediate attention to prevent further escalation. Such incidents require containment measures using available equipment such as the shutdown of certain machinery. Relevant personnel, such as the site manager, are promptly informed, and the incident is recorded in a log for future reference.

Medium-risk scenarios require a more robust response. Following the shutdown of the entire plant, the immediate activation of the on-site emergency response team is crucial. Local authorities such as the environmental agencies and fire department are then notified. In some cases, a precautionary evacuation of non-essential personnel may be necessary to ensure safety. Communication lines with local authorities and the company’s higher management are established to ensure a coordinated response.

High-risk incidents have the potential for widespread impact on the community, environment, and the company. Alongside the measures taken for medium-risk incidents, the response to high-risk emergencies also includes large-scale evacuations, establishing a command centre to handle the situation, coordinate response efforts, and notify regulatory bodies.

#### **5.2 Expected Outcomes**

Pre-accident Injury and Fatality Prevention

Whilst we are unable to know for sure how much our recommendations will reduce in reality, if we assume that there is a complete reduction in all the significant factors listed in Table 5.1.1 we can calculate the reduction in probabilities for injuries and fatalities. By using the logistic regression model coefficients, there is a reduction of 11.46% chance of occurrence for injuries and a reduction of 95.88% for fatalities assuming all other factors stay the same. However, in reality the other factors are unlikely to remain constant, and the reduction will likely be less drastic.

Post-accident Risk Assessment Matrix

A risk assessment matrix offers a real-time outlook by looking at current risk factors and the likely outcomes. It aids Aramco to take note of early warning signs or trigger events and prevent the accidents from getting worse (Lucidspark, n.d.). A risk assessment matrix also helps to get the entire team aligned through easy interpretability and visualisation. This allows for all types of staff, even non-technical personnel, to understand the potential risks and guide them to what specific actions are needed for them to take.

#### **5.3 Comparison with Current Solutions**

Currently, Aramco’s safety management system helps to identify and mitigate risks to improve safety performance and creates a safety culture that ensures employees have the necessary training, equipment, and protocols to complete tasks safely. Furthermore, Aramco conducts safety reviews and maintains an annual calendar for site visits by top management and deep dives into key process safety components (Aramco, n.d.). The usage of the risk assessment matrix can be a valuable addition to the current safety protocols and management system as it allows for a fast and easy use, leading to a rapid response from personnel during accidents. Furthermore, educational programmes to prevent incorrect operation of equipment are also in line with their current safety culture’s aims.

Aramco currently uses Unmanned Aerial Vehicles (UAV) to inspect high-rise equipment such as flare tips and prevents risk to human lives (Aramco, 2021). Similarly, our solution for handling Highly Volatile Liquids through the use of robotics and automation is also aligned with this overarching motivation. Aramco can expand its use of technology from UAVs to other types of robotics for inspecting Highly Volatile Liquids as well.

Aramco also tracks safety statistics and has developed Lifesaving Rules from them to cover the eight most hazardous activities at the company. The CART model’s variables used for splitting can be used to add on to these lifesaving rules, to further improve on the efficacy of these Lifesaving rules (Aramco, 2021).

However, one area that has not yet been touched on by Aramco are accidents caused by **OTHER OUTSIDE FORCE DAMAGE**. Thus, Aramco could place greater focus on this area to further reduce accidents through the solutions for **OTHER OUTSIDE FORCE DAMAGE** we have listed in **Section 5.1**.

#### **5.4 Limitations and Future Improvements**

As we could not find specific and detailed data for Aramco accidents, we had to resort to public data from United States pipeline accidents. Thus, this data might not be fully reflective of Aramco’s actual situation as there may be large differences in data across companies and geographical regions. There could also be more factors which are not found in our dataset that could affect accidents rates and severity.

Furthermore, our dataset is also highly imbalanced, with only a few observations of data having any injuries or fatalities. This could cause our models to be slightly inaccurate at predicting injuries and fatalities even through the use of weights for the model. The significant factors that influence the occurrence of fatalities and injuries could also be inaccurate due to the imbalanced data.

We could improve in the future by getting more specific data from Aramco’s accidents. This would allow our accidental model to be more accurate and specific with regards to Aramco’s oil pipelines.

### **6. Conclusion**

To sum up our project on workplace safety in the oil industry setting, more specifically Aramco, it is evidently crucial to implement proactive safety measures to combat the rise in workplace accidents and the related economic costs - amongst others.

Considering the fact that workplace safety is of utmost priority to Aramco, and how existing solutions - like safety campaigns, digital safety initiatives and benchmarked safety management systems - to address the challenges with regard to this area are lacking, we figured that we would delve deeper into it. Hence, we approach it by utilising data analytics to gain greater insight and ultimately devising strategies that could be more effective.

Through our data exploration, our findings underlined the aforementioned growing safety concerns in the oil industry and the root causes behind it. We also identified the top cost incurred subcategory and adjusted the skewness of the data by adding class weights and the use of percentiles. On top of that, we also explored the correlations between variables, and the cost-related impact of pipeline shutdowns in comparison to that in cases where it did not occur - the former being significantly higher than the latter.

In the next stage of our data analysis - that is the data visualisation and modelling phase - we used logistic regression and CART models to predict injuries and fatalities, as well as the costs incurred in the form of percentiles. We then compared the resultant varying prediction accuracy of the models before selecting the most appropriate one for our suggested approach.

In conclusion, combining these initial stages together, we suggest optimal measures to be taken in the future, before and after occurrence of such workplace accidents. Pre-accident solutions include reinforcing ARAMCO’s infrastructure resilience through the integration of early detection systems and the implementation of geofencing technology, increasing utilisation of robotics for automation, and targeted educational programs. Post-accident solutions consist of a more thorough planning of emergency response procedures with the aid of models like CART and a Risk Assessment Matrix. We believe that these solutions would definitely be valuable additions to ARAMCO’s current measures, while still being aligned to the company’s technology-driven initiatives and safety-centred culture.

### **References**

Aramco. (2021, July 8). *Building a culture of safety*. Aramco. Retrieved November 5, 2023, from <https://www.aramco.com/en/magazine/elements/2021/building-a-culture-of-safety>

Aramco. (n.d.). *Health and safety*. Aramco. Retrieved November 5, 2023, from <https://www.aramco.com/en/creating-value/sustainable-business-operations/health-and-safety>

Aramco. (2021). Our 2021 Sustainability Report. Retrieved 2023, from <https://americas.aramco.com/-/media/downloads/sustainability-report/2021-saudi-aramco-sustainability-report-safe-operations-en.pdf?la=en&amp;hash=D6A0F2119712661AAF6B56DD3BC2CFBDAD788DCE>

Benoit, D. F., & Van den Poel, D. (2009). Benefits of quantile regression for the analysis of customer lifetime value in a contractual setting: An application in financial services. *Expert Systems with Applications*, *36*(7), 10475–10484. <https://doi.org/10.1016/j.eswa.2009.01.031>

Code of Federal Regulations. (2023, October 31). *49 CFR Part 195 -- Transportation of Hazardous Liquids by Pipeline*. eCFR. Retrieved November 5, 2023, from <https://www.ecfr.gov/current/title-49/subtitle-B/chapter-I/subchapter-D/part-195>

Department of Transportation. (2016). *Oil Pipeline Accidents, 2010-Present*. Kaggle. Retrieved November 5, 2023, from <https://www.kaggle.com/datasets/usdot/pipeline-accidents>

HSE Now. (2022, June 20). *IOGP Releases 2021 Safety Data*. JPT. Retrieved November 5, 2023, from <https://jpt.spe.org/iogp-releases-2021-safety-data>

International Labour Organization. (2011, July 13). *The enormous burden of poor working conditions*.<https://www.ilo.org/moscow/areas-of-work/occupational-safety-and-health/WCMS_249278/lang--en/index.htm#:~:text=The%20enormous%20burden%20of%20poor%20working%20conditions&text=Worldwide%2C%20there%20are%20around%20340,of%20accidents%20and%20ill%20health>.

Kemmy Law Firm. (2022, March 29). *The most common causes of oil field explosions & accidents*. Kemmy Law Firm. <https://kemmylawfirm.com/blog/the-most-common-causes-of-oil-field-explosions-accidents/#:~:text=Equipment%20failures%20can%20cause%20fires,oil%20field%20workers%20may%20sustain>.

Loi, E. (2023, September 19). *22% rise in accidents at LTA Worksites in first 8 months of 2023: Chee Hong Tat*. The Straits Times. <https://www.straitstimes.com/singapore/22-rise-in-accidents-at-lta-sites-in-first-8-months-of-2023-chee-hong-tat-urges-vigilance>

Lucidspark. (n.d.). *Why you should use a risk assessment matrix. Why You Should Use a Risk Assessment Matrix*. Lucidspark. Retrieved November 4, 2023, from <https://lucidspark.com/blog/why-you-should-use-a-risk-assessment-matrix>

Passwaters, M., &amp; Ugal, N. (2022, March 25). Saudi Aramco’s oil storage facility in Jeddah hit by Houthi attack, fire erupts. Upstream Online | Latest oil and gas news. <https://www.upstreamonline.com/production/saudi-aramcos-oil-storage-facility-in-jeddah-hit-by-houthi-attack-fire-erupts/2-1-1191472>

Singh, K. (2023, July 6). *How to improve class imbalance using class weights in machine learning?*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2020/10/improve-class-imbalance-class-weights/>

Tierney, N. (2023, February 2). Replacing values with NA. Replacing values with na. <https://cran.r-project.org/web/packages/naniar/vignettes/replace-with-na.html#:~:text=When%20dealing%20with%20missing%20values,which%20values%20should%20be%20missing>

US Bureau of Labor Statistics. (2016). NONFATAL OCCUPATIONAL INJURIES AND ILLNESSES REQUIRING DAYS AWAY FROM WORK, 2015. <https://www.bls.gov/news.release/pdf/osh2.pdf>

U.S. Department of Transportation. (2014, July 29). *Fact Sheet: Other Outside Force*. Pipeline Risk Management Information System (PRIMIS). Retrieved November 4, 2023, from <https://primis.phmsa.dot.gov/comm/FactSheets/FSOtherOutsideForce.htm>

Wikimedia Foundation. (2023, October 23). Saudi Aramco. Wikipedia. <https://en.wikipedia.org/wiki/Saudi_Aramco>

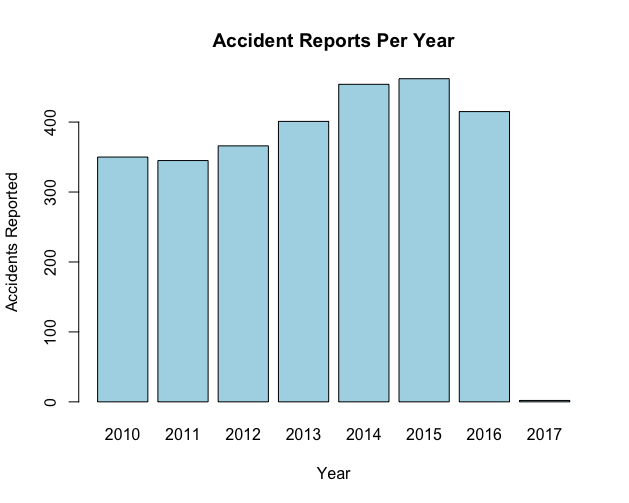
### 

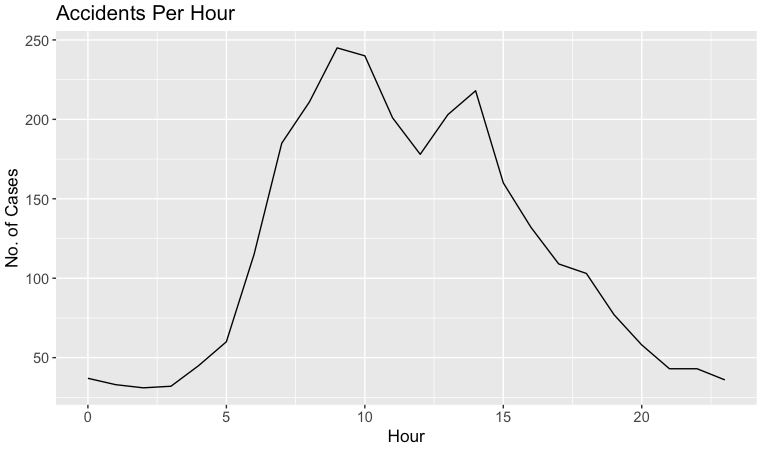
### **Appendices**

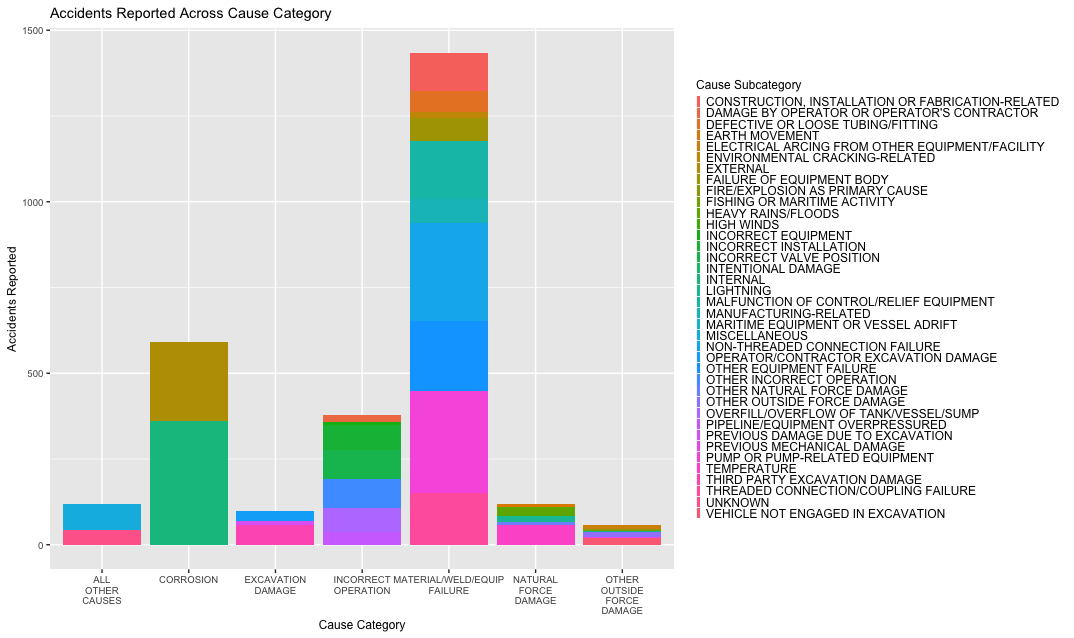
#### **Appendix A: Data Dictionary (Cleaned Data)**

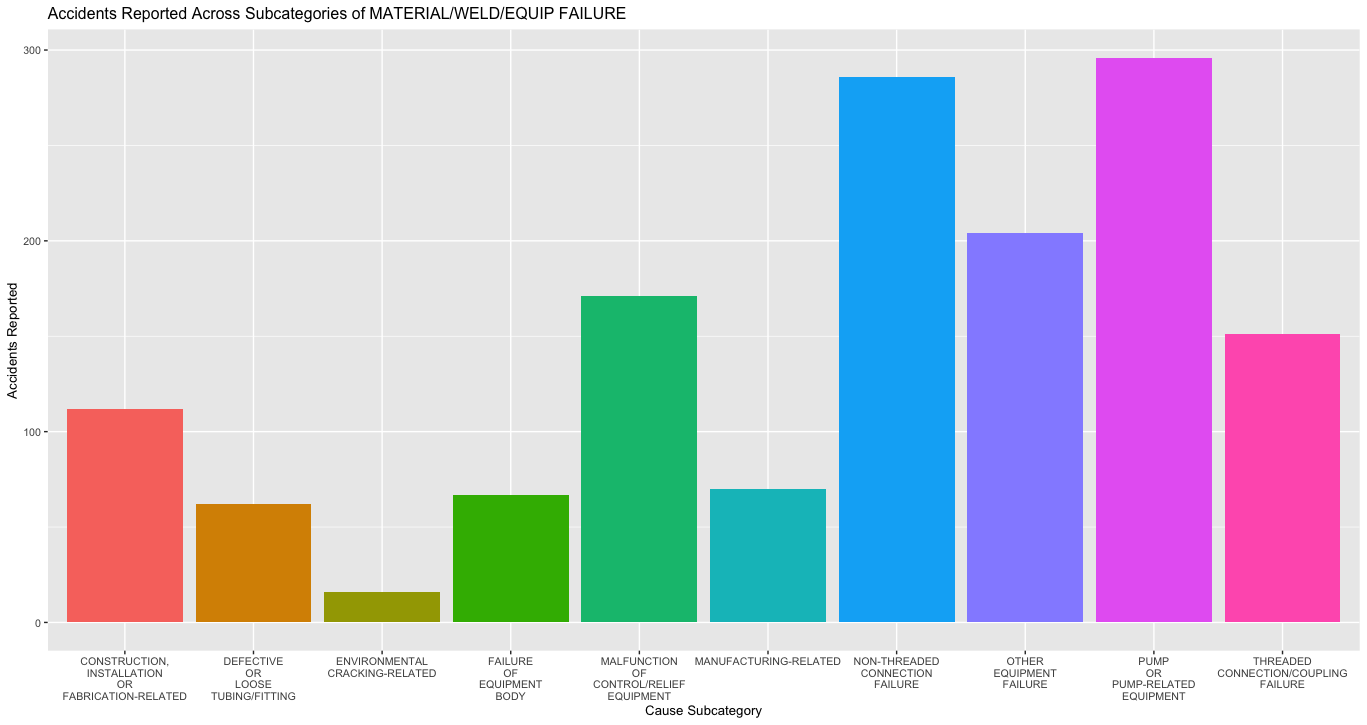
|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Type** | **Description** |
| Accident Year | Factor | Year of accident. Ranging from 2010 - 2018. |
| Accident Date / Time | Date: POSIXct  Time:  POSIXt | Date and time of accident in the format:  YYYY-MM-DD, Hours:Minutes:Seconds |
| Pipeline Location | Factor | Factor with 2 levels. Either Offshore or Onshore |
| Pipeline Type | Factor | Factor with 5 levels. Type of pipeline for onshore pipeline |
| Liquid Type | Factor | Factor with 5 levels. |
| Liquid Subtype | Factor | Factor with 8 levels. |
| Liquid Name | Factor | Factor with 68 levels. Only available when Liquid Subtype = OTHER HVL / OTHER. Otherwise the cell has NA. |
| Accident Latitude | Numeric | Measures the location’s distance north or south of the equator. |
| Accident Longitude | Numeric | Measures the location’s distance east or west of the equator. |
| Cause Category | Factor | Category of cause of accident |
| Cause Subcategory | Factor | Specific category of cause |
| Unintentional Release (Barrels) | Numeric | Amount of liquids released intentionally due to accident. Measured in barrels. |
| Intentional Release (Barrels) | Numeric | Amount of liquids released due to accident. Measured in barrels |
| Liquid Recovery (Barrels) | Numeric | Amount of liquid recovered in barrels |
| Updated Net Loss (Barrels) | Numeric | Corrected net loss measured in barrels.  Net Loss = Intentional Release + Unintentional Release - Liquid Recovery |
| Net Loss (Barrels) | Numeric | Uncorrected net loss measured in barrels from the dataset. |
| Liquid Ignition | Factor | Factor with 2 levels. |
| Liquid Explosion | Factor | Factor with 2 levels. |
| Pipeline Shutdown | Factor | Factor with 2 levels. |
| Shutdown Duration (Hours) | Numeric | Shutdown Duration = Restart Date/Time - Shutdown Date/Time.  Restart Date/Time and Shutdown Date/Time are from the original dataset. |
| Evacuations | Logical | Identifies if evacuation was carried out.  TRUE: Evacuation was carried out.  FALSE: Evacuation was not carried out. |
| Injuries | Logical | Identifies if there were injuries.  TRUE: Employees were injured.  FALSE: No employees were injured. |
| Fatalities | Logical | Identifies if there were fatalities.  TRUE: Accident was fatal.  FALSE: Accident was not fatal. |
| Property Damage Costs | Numeric | Costs related to property damage. |
| Lost Commodity Costs | Numeric | Costs related to raw materials lost. |
| Public / Private Property Damage Costs | Numeric | Costs related to damage caused to properties that do not belong to the corporation. |
| Emergency Response Costs | Numeric | Costs related to emergency response to the accident. |
| Environmental Remediation Costs | Numeric | Costs related to restoring the environment after the accident. |
| Other Costs | Numeric | Other costs. |
| All Costs | Numeric | Summation of all the other costs mentioned above. |
| newcost | Numeric | All Costs after winsorization. |

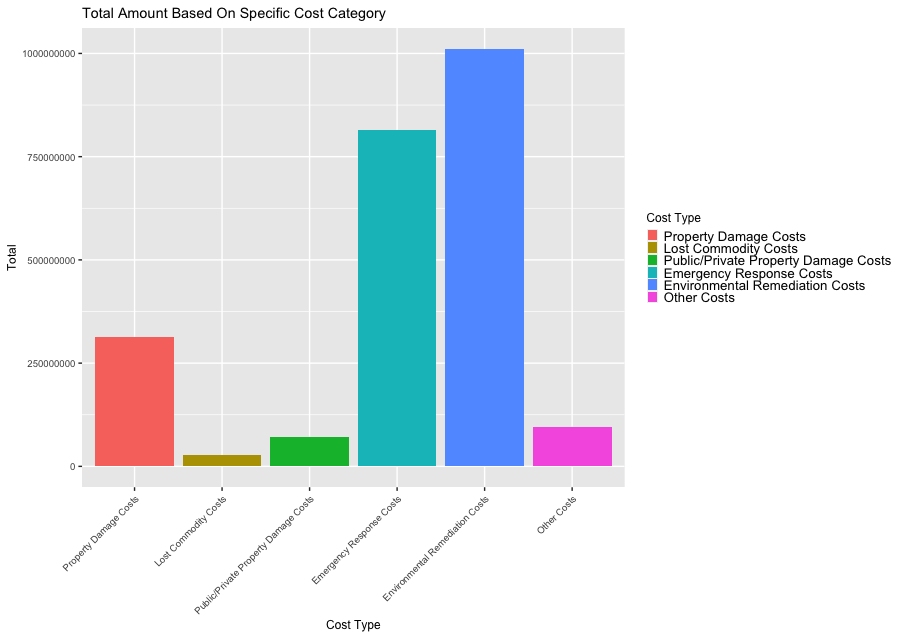
#### **Appendix B: Data Exploration**

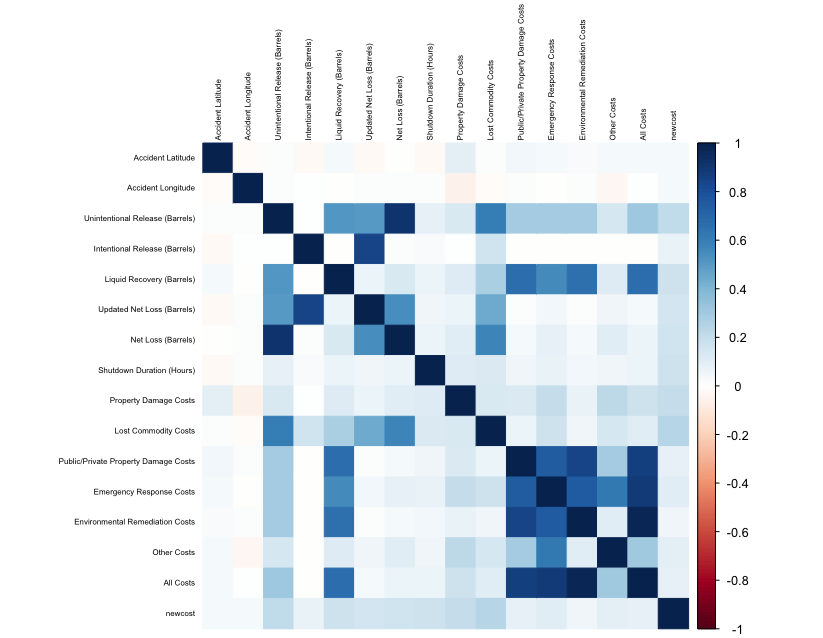




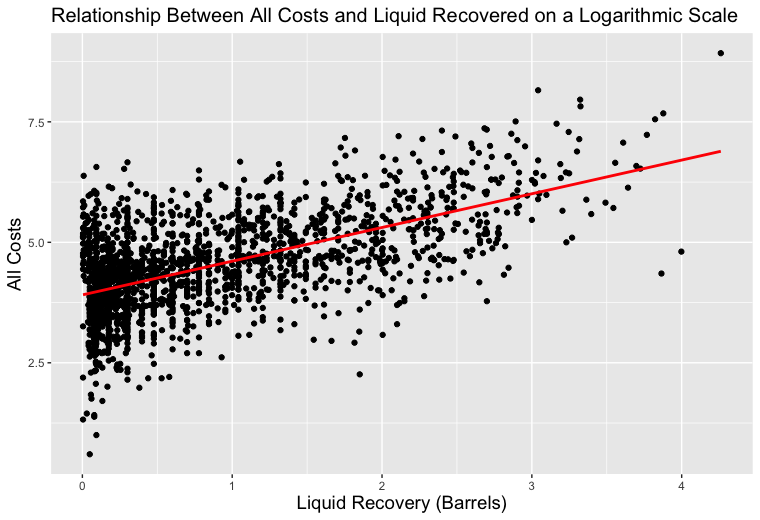


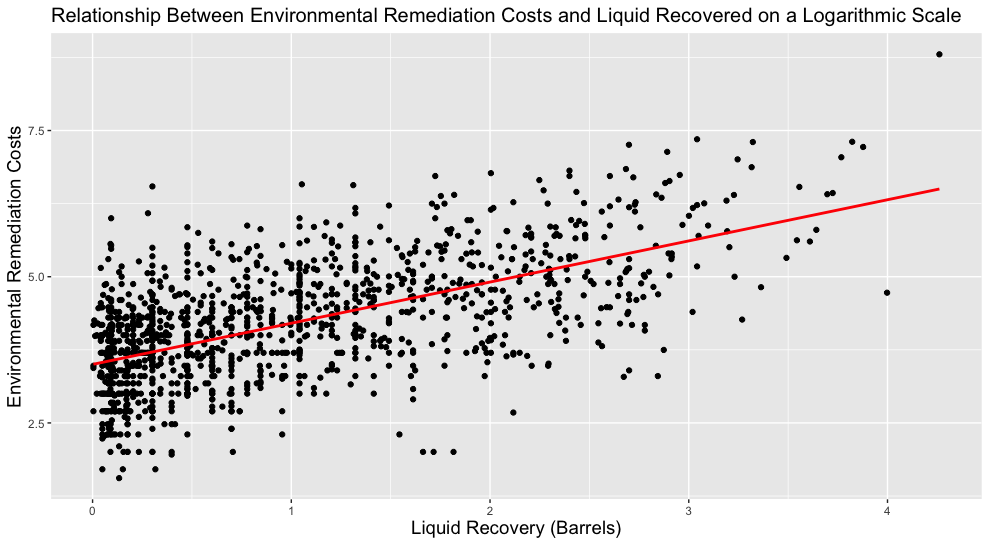


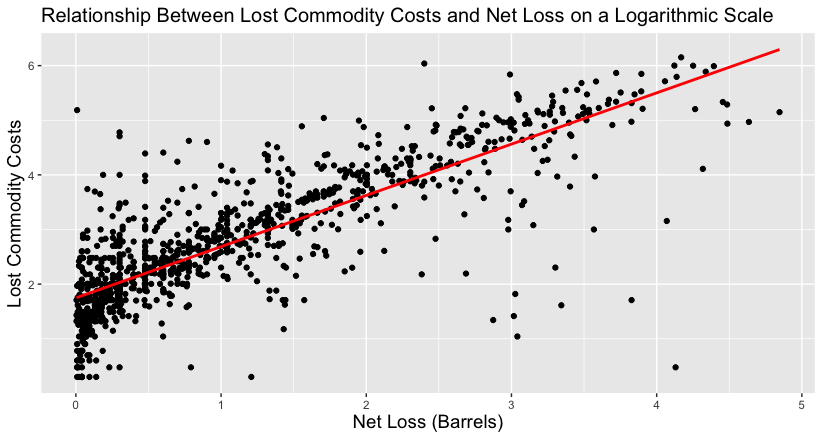


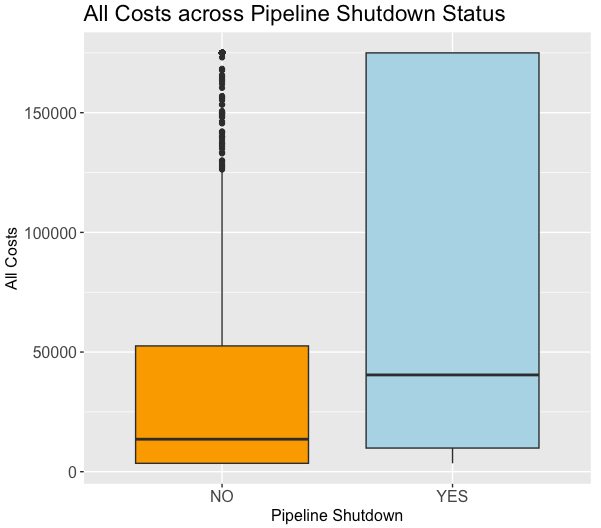


*Correlation Matrix of Numerical Variables*



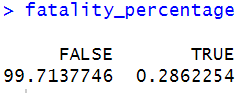






#### 

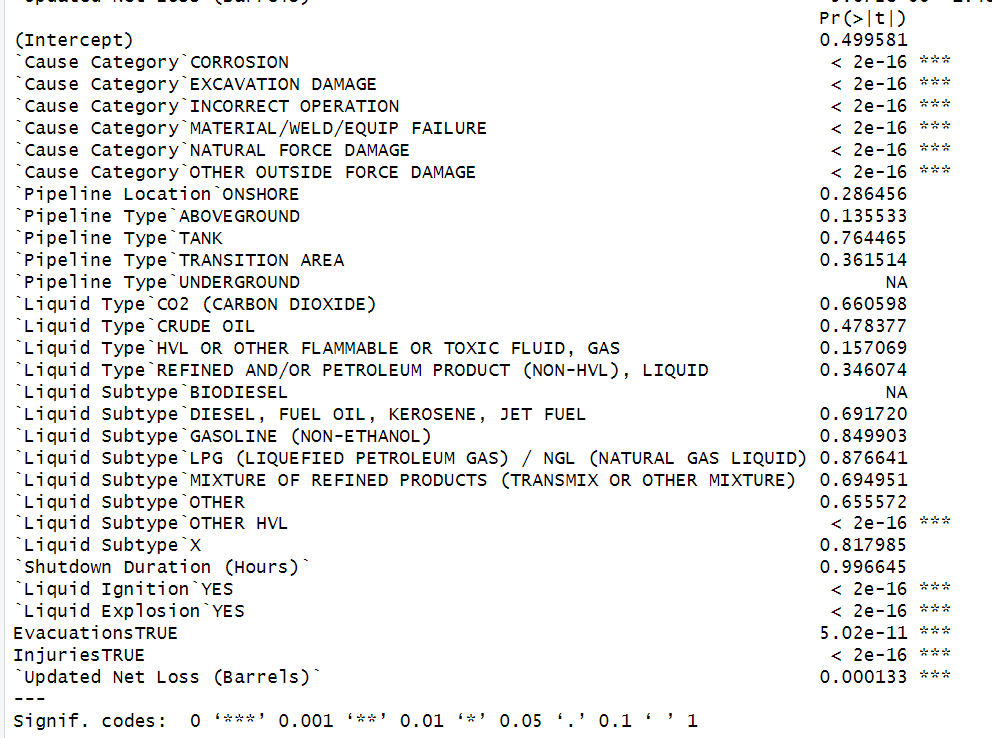
#### **Appendix C: Modelling**



*Imbalanced distribution of “Fatalities”*



*Prediction Accuracy of Logistic Regression Model in predicting “Fatalities” (Train Set)*

**

*Statistically significant factors in predicting the occurrence of fatalities (Logistic Regression)*

**

*Prediction Accuracy of Logistic Regression Model in predicting “Fatalities” (Test Set)*

**

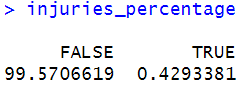
*Optimal CP for CART in predicting “Fatalities”*

**

*Prediction Accuracy of CART Model in predicting “Fatalities” (Train Set)*

**

*Prediction Accuracy of CART Model in predicting “Fatalities” (Test Set)*

**

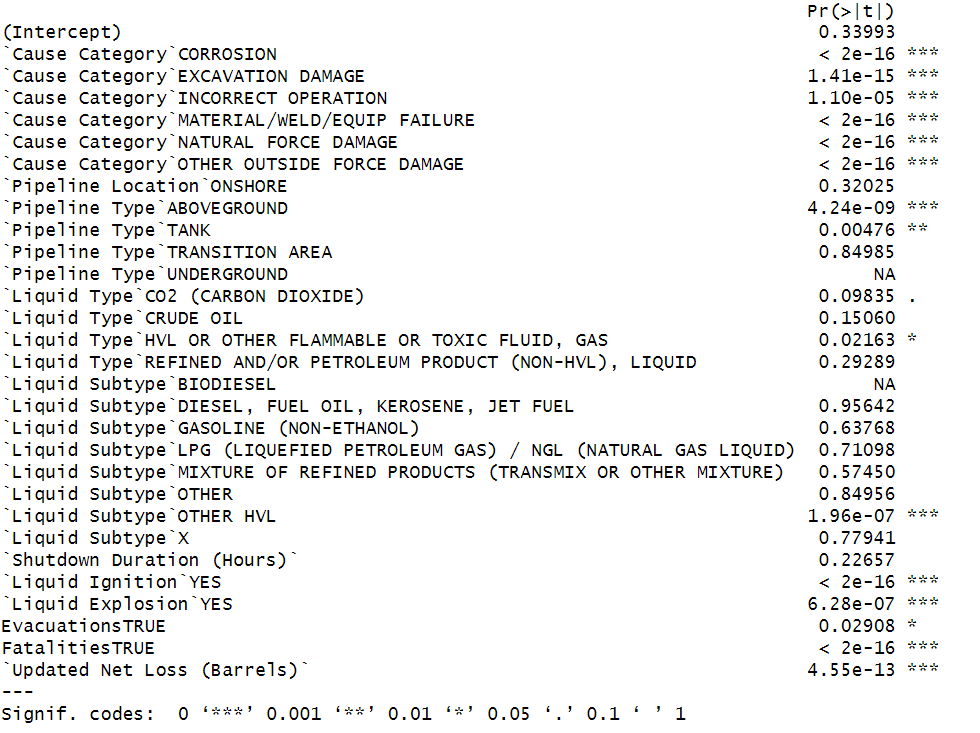
*Imbalanced distribution of “Injuries”*

**

*Prediction Accuracy of Logistic Regression Model in predicting “Injuries” (Train Set)*

**

*Prediction Accuracy of Logistic Regression Model in predicting “Injuries” (Test Set)*

**

*Statistically significant factors in predicting the occurrence of injuries(Logistic Regression)*

**

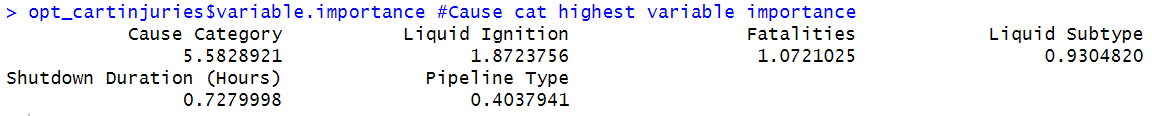
*Optimal CP for CART in predicting “Injuries”*

**

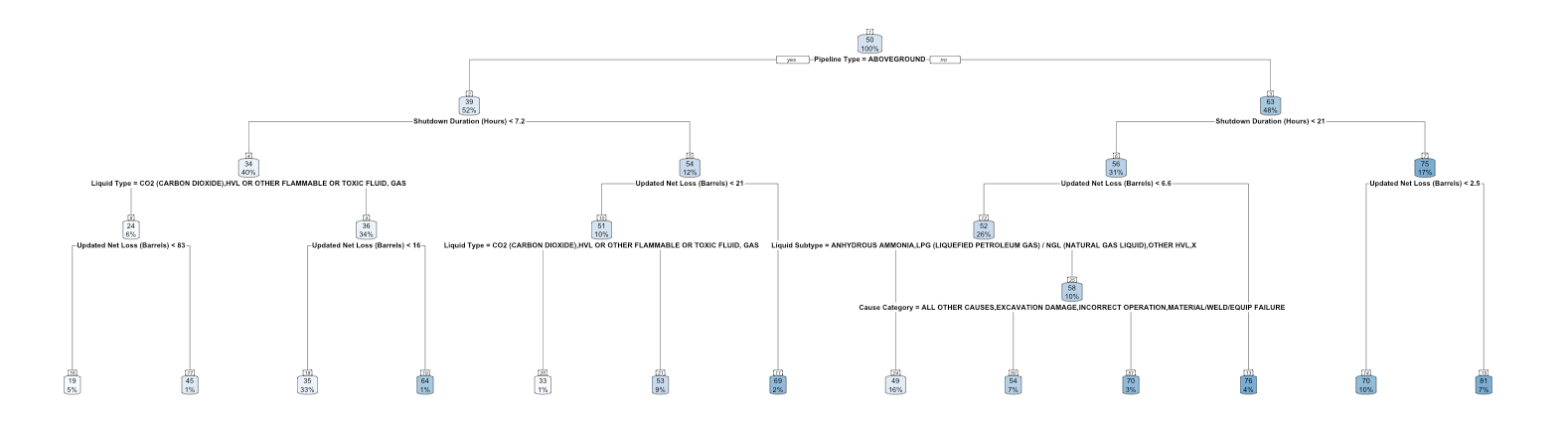
*Prediction Accuracy of CART Model in predicting “Injuries” (Train Set)*

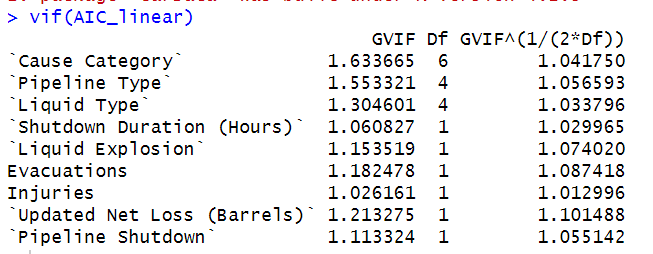
**

*Prediction Accuracy of CART Model in predicting “Injuries” (Test Set)*

**

*Variable importance of CART Model in predicting “Injuries”*

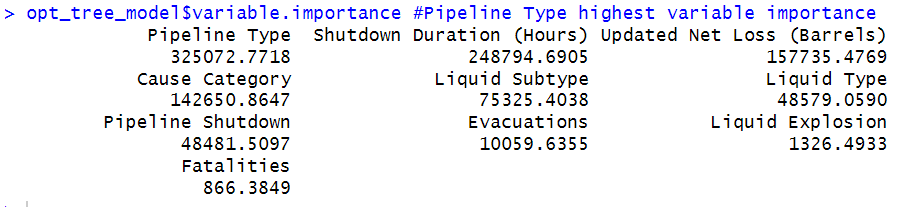


**

*VIF values below the threshold of 10*

**

*CP value for CART model in determining “New Cost”*

**

*Pipeline Type is the most important variable in determining “New Cost”*