Statistical Analysis of Rail Equipment Accident Data for Identifying Significant Variables in Rail Damage Cost Estimation

Alden Jettpace, Kitzel Padilla, Ramakrishna Reddy Mitta, Shruti Mukadam

1. Introduction

Rail-Freight is a vital part of the any modern logistics network in many countries, with the United States boasting a vast rail network before the invention of the car. However, rail-freight also comes with a range of risks and hazards, including accidents and incidents that can result in property damage, injuries, and fatalities. An accident can vary in severity from minor to disastrous, and the costs can be prohibitive. The Norfolk Southern derailment in East Palestine, Ohio is expected to cost the company upwards of $387 million, not including potential insurance claims on long term medical conditions from the carcinogens of the burned material (Funk, 2023). In this case, there is a clear example of the financial and human cost of a train accident.

The many firemen, brakemen, and engineers which run these trains are tasked with managing various systems and variables to ensure smooth operations, but such variables can be overwhelming, and this can cause uncertainty in ascertaining the risk of an accident at any given moment. Given records of previous accidents and derailments, could a subset of variables be selected which best explain the current costs and risk of a potential accident?

1.1 Problem Statement

Given a set of variables which represent the records of previous instances of rail accidents and equipment failures, there exists of subset of variables which are observable during operation of the train before a possible derailment which are statistically significant in predicting the total equipment-cost of an accident or derailment, should it occur in those measured circumstances.

- **Null Hypothesis**: No such subset exists, and all variables are either relevant or irrelevant to measuring accident costs
- **Alternative Hypothesis**: There is some subset of variables $Y < Z$ that are statistically significant to predicting accident costs, containing a majority of the explainable variance.

1.2 Explanation

Under the logic that a more severe accident will have on average a larger cost in damages, the summed cost of damage to the tracks and the train equipment was used as an indicator of the risk of a serious accident. Thus, the variables selected as significant are expected to be correlated with a risk of having a serious accident, and act as warning-indicators to train engineers when operating a train. Because of this focus on variables observable during operations, only variables recorded from the incident data which could be realistically observed before such accidents will be considered in the subset selection.

2. Previous Work

Because of the vastness of freight as a service in the US, various studies have been conducted to improve the safety of the system by identifying the most significant variables. Jing Chen et al utilized multiple data sources including census data, Federal Railroad Administration records of crash incidents and all rail-road crossings in the US, as well as fatality data, and utilized an alternative to the common best-subset selection, shrinkage, and non-linear-tree methods of variable selection. Using fatalities as the response, they compared the Lasso and Step methods, along with

Xgboost, then combined with variables from an arbitrary prediction model. The selected variables were then compared on a logistic-linear regression for predicting whether in those circumstances an accident would be fatal or not. Their work methodology, so much like our own, found that the mixture of the variables selected from each method had the highest accuracy (2021) Thus, we have reasons to suspect we will have similar results and will find a relevant subset of variables for predicting costs.

3. Aim

The primary objective of this project is to conduct a variable selection analysis on a dataset of rail equipment accidents to identify the most statistically significant variables for estimating the costs of damages incurred during such incidents. The study aims to employ rigorous statistical methods to select the best variables, which can provide a more accurate and reliable estimation of the costs involved in rail equipment accidents. The findings of this study are expected to contribute to the development of improved risk management strategies for rail equipment accidents and enhance the overall safety and efficiency of railway transportation systems.

4. Files

The process of cleaning, analyzing, and refining the data was spread across several files, with the raw data being refined down to our data for analysis. Below is a list of the files involved in the study and their purpose:

- **Rail_Equipment_Accident_Incident_Data.csv:** Curated FRA Safety data pertaining to Rail Equipment Accidents. Consists of over 200,000 rows and 160 columns. Cleaned for relevant and variables and non-empty instances in [Group_7_Dataclean.R]
- **Form54DataDict.xlsx**: File which explains the context of the variables/columns of the Rail Accident Data, indicating whether they were numeric or text
- **cleaned_train_data_1.csv**: Consists of 1269 rows and 47 columns, consists of variables in Rail Accident Data that did not have a missing value in variables selected which were considered measurable during operation (pre-accident). Output of [Group_7_Dataclean.R]
- **cleaned_rail_refined.csv:** Consists of 1269 rows and 41 columns, result of refining certain variables and combining others to improve correlation with summed_cost. Output of [Group_7_Data_Refined_Markdown.Rmd] file, analyzed for subset selection in [Group_7_Project_Subset.Rmd]
- **Group_7_Dataclean.R:** File where process of initial subset selection for variables only measurable before an accident where selected, and null-data-incidents removed
- **Group_7_Data_Refined_Markdown.Rmd:** Markdown explaining process of further data analysis and refinement to transform certain attributes such as Positive.Drug.Tests and Positive.Alcohol.Tests into Positive.DA.Test.
- **Group_7_Project_Subset.Rmd:** File where various subset selection methods were performed and compared, including Forward/Backward selection, Lasso-Regression, and RandomForest.

5. Methodology

5.1 Dataset Chosen

The data set that was chosen is the Rail Equipment Accident/ Incident Data set, downloaded from the U.S. Department of Transportation website (https://data.transportation.gov/Railroads/Rail-Equipment-Accident-Incident-Data/85tf-25kj). This dataset is being continuously updated and holds one hundred and sixty predictors and over two hundred thousand rows, representing newly reported accidents from the report form 54. The U.S Department of Transportation is one of the executive departments of the U.S. Federal government and this dataset is provided by the FRA Safety administration.

5.2 Dataset Exploration and Cleaning

Before cleaning the data and selecting dense-vector instances, we first had to do initial subset selection by choosing the variables which could be measured before an accident rather than recorded afterword, throwing away variables such as number of cars damaged, or number of cars derailed.

Following the initial subset selection, the data cleaning process encountered several problems. Most instances were missing almost all of their attribute data, with even missing Latitude/Longitude values reducing the data by nearly 90% of its instances. Many of the values were missing in other ways, being instead an empty string rather than an NA-value, and in the case of Track.Density, in the form of a string when its intended use is as a continuous value. The removal of so many instances from the missing NA combinations meant that many categorical variables lost some of their categories when converted into factors. This entire process, done in [Group_7_Dataclean.R], reduced the dataset down to 1269-instances out of the original 200,000. Finally, ultizing summed equipment cost (to do away with the variable total-cost which could include insurance claims and lawsuits), we created the response variable of summed_cost by summing the track and equipment damage costs.

The cleaned data set was further refined in the file [Group_7_Data_Refined_Markdown.Rmd] We first transformed the variables hours.conductors.on.duty and hours.engineers.on.duty into binary variables (signifying whether the employees had been on duty for over eight hours or not), grouping the variable Months into seasons, and fusing the less-frequent categories of rain/sleet/snow/cloudy in Weather.Condition into the new variable Un-Clear, making it a binary-categorical. These last changes were made because certain continuous variables might have different behavior after a certain cutoff point so categories for some of the variables were fused into larger groups to give the new larger group better representation in the data.

Within that same file, we also did initial data analysis and visualization of correlation with the response. The correlation and the anova score relative to summed_cost were visualized in Figure 1 and Figure 2.

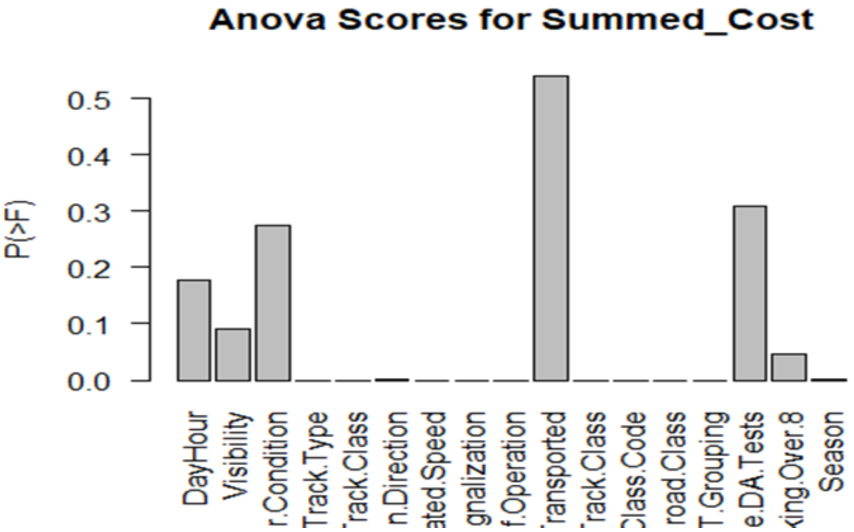

Figure 1: Bar chart of Anova P-value for categorical refined-variables to the response summed_cost.

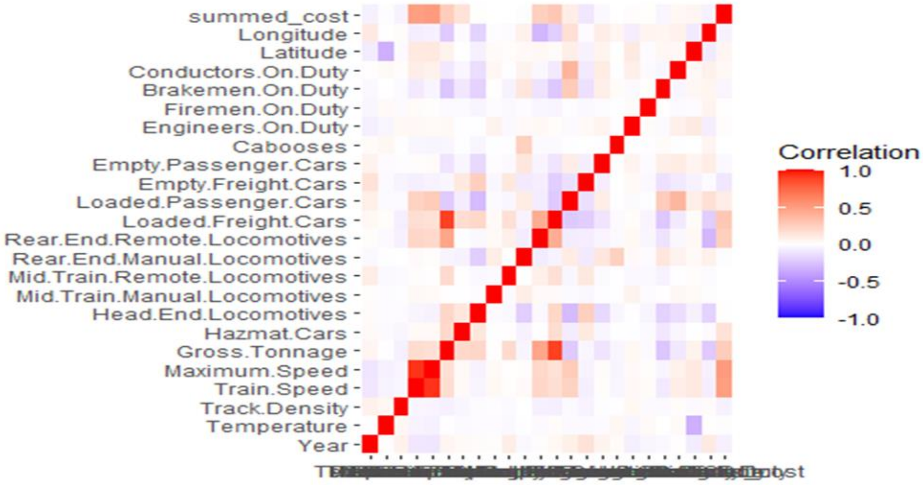

Figure 2: Correlation heatmap of categorical variables of refined data

5.3. Subset Selection:

To determine the optimal subset of variables for estimating the cost of railway damage, this study employed 3 statistical techniques: Forward/Backward subset-selection, Lasso coefficient reduction, and Random Forest. In the case of Lasso and forward and backward selection methods, categorical variables with multiple categories were retained in their entirety if they were considered significant in at least one of their categories. For example, if the DayHour variable was found to be relevant for the DayHour8 and DayHour16 categories, the entire DayHour variable would be considered significant when estimating the cost of railway damage. This approach ensured that all relevant information from categorical variables is included in the analysis and helped to improve the accuracy of cost estimation for railway accidents.

The variables deemed relevant by Forward/Backward selection and by Lasso reduction were recorded for later use. Following this, a Random Forest was constructed and repeatedly tested to optimal tuning for the data, before building a forest on those parameters. We then examined the importance-value of all the variables in the forest and selected the 22-most important variables by increased error on removal.

To perform Forward/Backward subset selection, the study used the regsubsets model with a maximum number of variables. The AIC, BIC, CP, Adj-R^2, and MSE values were plotted for each model, and the minimal value for AIC/BIC/CP or maximum value for Adj-R^2 will be used to select the optimal number of attributes. However, if the number of attributes selected by different criteria varied, the study utilized cross-validation to determine the minimum MSE attribute number. By using this approach, it helped to ensure that the model is not overfitting the data, and that the selected variables are truly the most relevant to make the estimation.

In the Lasso regression method, the study also used cross-validation to determine the best lambda value. Once the best lambda value was determined, the final Lasso regression model was fitted using this value, and the variables not reduced to zero coefficient were recorded.

In the Random Forest method, the study used cross-validation to determine the best tree parameters. Once the optimal parameters were identified, we built the Random Forest model using these parameters and examined the importance scores of the variables. Based on the importance scores, we select 22 variables that were deemed to be the most significant to make the estimation.

Finally, glm(generalized linear model) with cross-validation to estimate the summed cost by utilizing the variables from different subsets selected by the previous methods, and an lm(linear_model) to estimate to the adjusted R^2 score for each methods relevant subset, along with their unions and intersections. We then determined the best-subset by these scores.

6. Results

6.1 Forward Subset Selection

The number of attributes chosen for minimum BIC is 15, while for minimum CP and maximum adjusted R squared, it is 28. Since there is uncertainty in the number of attributes, we based our decision on cross-validation MSE. The minimum error was found at 28 out of 96 variables in regsubsets during cross-validation. This finding was supported by the maximum adjusted R squared in the previous examination and the minimum CP. Next, we examined the 28 variables within the forward selection. Since the 28 variables contain dummy variables of categorical variables, we considered the whole categorical variable significant, even if one dummy variable was significant. Consequently, we are left with 19 variables that were found to be significant in the forward subset selection process.

The variables are - DayHour, Visbility, Track.Type, Track.Class, Recorded.Estimated.Speed, Maximum.Speed, Method.of.Operation, Passengers.Transported, Hazmat.Cars, Rear.End.Remote.Locomotives, Loaded.Freight.Cars, Loaded.Passenger.Cars, Empty.Break.Cars, Brakemen.On.Duty, Latitude, Longitude, Joint.Track.Class, Rporting.Railroad.SMT, Positive.DA.Tests.

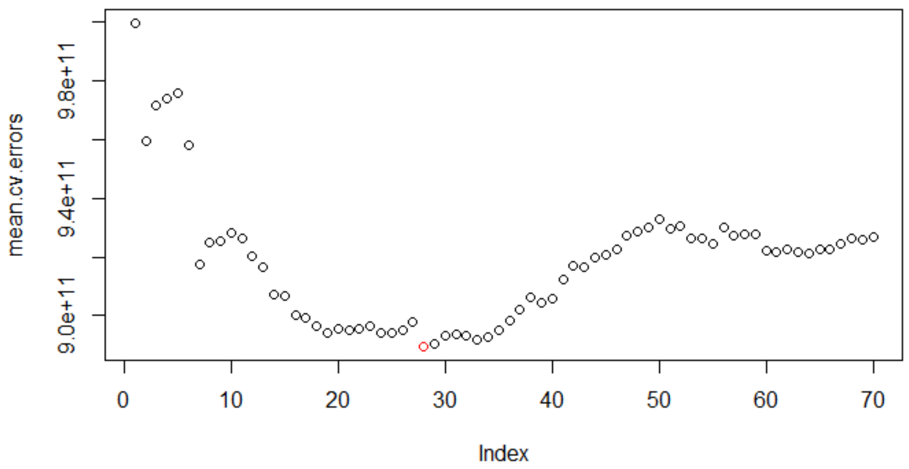

Figure 3: Forward Selection Cross-validation error score for number of attributes

6.2 Backward Subset Selection:

The number of attributes chosen for minimum BIC was 14, while for minimum CP was 28 and for maximum adjusted R squared was 34. Since there was uncertainty in the number of attributes, we based our decision on cross-validation MSE. The minimum error was found at 22 out of 96 variables in regsubsets during cross-validation. We then examined the 22 variables within the backward selection. Since the 22 variables contain dummy variables of categorical variables, we considered the whole categorical variable significant, even if one dummy variable was significant. Consequently, we were left with 16 variables that were found to be significant in the backward subset selection process.

The variables are - DayHour, Visibility, Track.Class, Recorded.Estimated.Speed,
Maximum.Speed, Passengers.Transported, Hazmat.Cars, Read.End.Remote.Locomotives,
Loaded.Freight.Cars, Loaded.Passenger.Cars, Brakemen.On.Duty, Latitude, Joint.Track.Class,
Reporting.Railroad.Smt.Grouping, Positive.DA.Tests, Season.

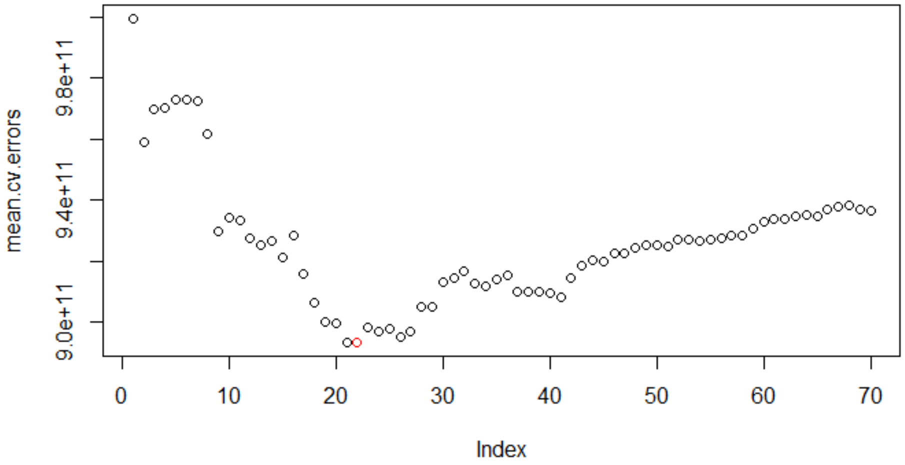

Figure 4: Backward Selection Cross-validation error score for number of attributes

6.3 Lasso Coefficient Shrinkage

We used a 10-fold glmnet to find the optimal lambda value, which was determined to be 22435.5 or exp(10.0184). Using this lambda value, we performed a lasso analysis on the entire dataset and found that 35 out of 96 variable coefficients were not reduced to zero. Since the 35 variables contain dummy variables of categorical variables, we considered the whole categorical variable significant, even if one dummy variable was significant. Consequently, we were left with 22 variables that were found to be significant in the Lasso Coefficient Shrinkage process.

The variables are - Train.Speed, Maximum.Speed, Hazmat.Cars, Rear.End.Remote.Locomotives, Loaded.Freight.Cars, Loaded.Passenger.Cars,Empty.Freight.Cars,Cabooses, Brakemen.On.Duty, Latitude, Longitude, DayHour, Visibility, Track.Type, Track.Class, Recorded.Estimaged.Speed, Signalization,Method.of.Operation,Passenegers.Transported,Joint.Track.Class, Reporting.Railroad.SMT.Grouping, Season.

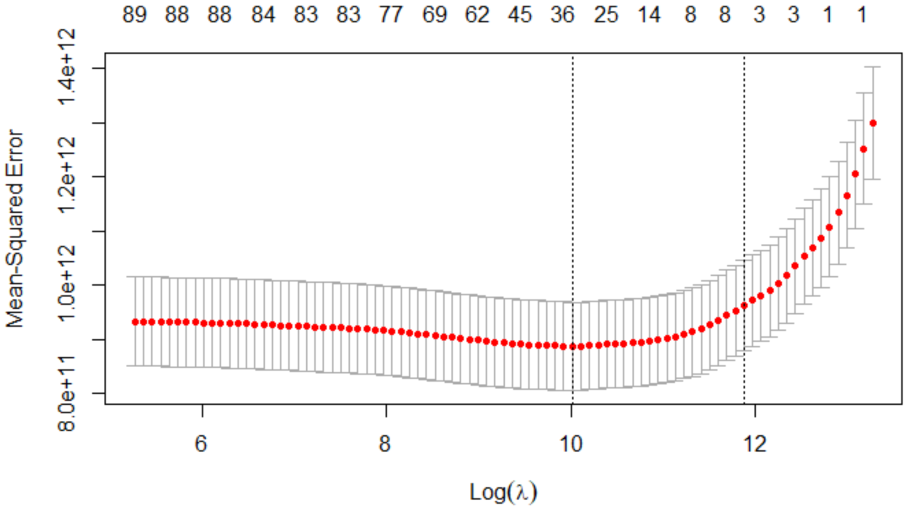

Figure 5: Lasso Regression Optimal lambda value

6.4 Random Forest Selection

After 24-loops with a repeated random-seed, we found that the best forest for this data would contain 200 trees and build a tree with at max 10 variables at a time. Using these parameters with the same seed number, we used the importance() function on the forest to find the importance-values for each variable. Sorting by this value, we found the most important variables by the value of increase in error if variable were removed, of these, the two 3 variables proved to have a far higher score compared to the other significant variables by an order of magnitudes.

The variables were – Maximum.Speed, Train.Speed, Gross.Tonnage, Loaded.Freight.Cars, Joint.Track.Class, Reporting.Railroad.SMT.Grouping, Latitude, Rear.End.Remote.Locomotives, Track.Class, Longitude, Brakemen.on.Duty, Train.Direction, Reporting.Railroad.Class, Working.Over.8, Loaded.Passenger.Cars, Recorded.Estimated.Speed, Empty.Freight.Cars, Passengers.Transported, Signalization.

6.5 Subset Comparison

Variable	Forward Selection	Backward Selection	Lasso Coefficient Reduction	Random Forest
DayHour	✓	✓	✓	
Visibility	✓	✓	✓	
Track.Class	✓	✓	✓	✓
Recorded.Estimated.Speed	✓	✓	✓	✓
Maximum.Speed	✓	✓	✓	✓
Passengers.Transported	✓	✓	✓	✓
Hazmat.Cars	✓	✓	✓	✓
Rear.End.Remote.Locomotives	✓	✓	✓	✓
Loaded.Freight.Cars	✓	✓	✓	✓
Loaded.Passengers.Cars	✓	✓	✓	✓
Brakemen.On.Duty	✓	✓	✓	✓
Latitude	✓	✓	✓	✓
Longitude	✓		✓	✓
Join.Track.Class	✓	✓	✓	✓
Reporting.Railroad.SMT.Grouping	✓	✓	✓	✓
Positve.DA.Tests	✓	✓		
Season		✓	✓	
Track.Type	✓		✓	
Method.of.Operation	✓		✓	✓

Train.Speed			✓	✓
Empty.Freight.Cars			✓	✓
Cabooses			✓	
Signalization			✓	✓
Gross.Tonnage				✓
Reporting.Railroad.Class				✓
Train.Direction				✓
Year				✓
Working.Over.8				✓
Track.Density				✓

Table 1

Table 1 presents the results of different feature selection methods used in the analysis of rail equipment accident data. The results show that each method selects a different subset of variables, and the performance metrics also vary between methods. Since each of these variables has been suggested as relevant by at least one subset selection method, the union and intersection of each subset is also examined.

Because of the varying results of each method, we compared their performance on 10-fold cross validation, along with the adjusted R^2 value trained on all the available data to determine the 'best subset'. And to ensure that at least some subset could exist that performs better than training on all the attributes, the results were compared against those of a model trained on all the data attributes.

Subset	**Cross Validation Error**	**Multiple R-squared**	**Adjusted R-squared**
Forward Selection	888254872871	0.4027201	0.3725235
Backward Selection	887501880873	0.4033223	0.3763625
Lasso Coefficient Reduction	896822361014	0.4081187	0.3755194
Random Forest	885813867025	0.3901162	0.3706081
Union	904362517918	0.4105676	0.3727047
Intersection	872783423019	0.3848176	0.3720120
All Variables	922930632921	0.4127649	0.3684647

Table 2

Table 2 presents the results of glm(generalized linear model) with cross-validation and lm(linear model) to estimate the summed cost by utilizing the variables from different subsets. Every linear model trained on a subset had a lower cross validation error than that of one trained on all the attributes and had a higher Adjusted R-Squared value. Even the largest subset, Union, contains 29 of the 40 variables, and by its R^2 score we can determine that the 11 excluded variables contain only 0.2% of the variance in summed_cost. The Intersection meanwhile consists only of 12 variables, yet it contains 38% of the total variance. This means that 28 of the variables only contain 3% of the variance of the response.

From the above, we have evidence to reject the null hypothesis and accept that alternative hypothesis: That there exists some subset of variables correlated with summed_cost which contains most of its information.

7. Discussion

While the backward selected subset had the larger adjusted R^2 value, it had a larger cross validation error compared to the intersection of all subsets. While we have reason to believe that at least one of the variables unique to the backward selection method is significant, they contain only 2% of the variance, so our group selected the intersection of all subsets as the 'best-subset'. Table 3 demonstrates the coefficients and P-value of the attributes of the intersection. If a value is categorical, it will only have the significant coefficients listed.

Variable	Coefficient	Standard Error	P-Value
Track.Class.7	-5857893.9	1225093.6	1.95e-06
Recorded.Estimated.Speed.Recorded	215490.2	54223.1	7.47e-05
Maximum.Speed	33174.6	2046.4	< 2e-16
Passengers.Transported.Yes	-1085215.1	280235.2	0.000113
Read.End.Remote.Locomotives	157708.3	52331.5	0.002634
Loaded.Freight.Cars	4224.2	675.9	5.63e-10
Loaded.Passenger.Cars	226130.2	46458.5	1.28e-06
Brakemen.On.Duty	182305.0	70517.9	0.009845
Latitude	-11570.7	5361.0	0.031095
Reporting.Railroad.SMT.GroupingSMT-7 - Commuter West	-928969.1	285355.1	0.001163
Hazmat.Cars	5760.3	1367.9	2.73e-05

Table 3: Coefficients and P-Values of significant intersection attributes

All the intersection variables except for Joint.Track.Class had at least one relevant categorical coefficient or continuous coefficient. As for why these unique relationships are found, there could be a variety of reasons which would require further data sources to examine.

Certain types of tracks could be more or less expensive to repair or prone to damage. That the Class 7 tracks have negative coefficient suggests that either they are harder to damage compared to Class 1 Tracks and thus less prone to damage or are simply less costly to repair. That having the speed be recorded rather than estimated means that damages are higher could imply that transports where the speed is especially monitored are by themselves usually in more hazardous conditions such as extreme weather and somehow absorbed the entire effect. That having passengers on board means less cost while at the same time having passenger cars having more cost could be explained by the presence of the cars being potential damage that could be suffered by there being more equipment that could be damaged. Meanwhile, passengers take up space and weight which is no longer allocated to the freight, and thus is no longer part of the cost.

8. Conclusion

In the future, it would be beneficial to base the analysis on more specific measures of accident severity, such as whether a train derailed or crashed, compared to less severe accidents or random accidents. This would allow for the identification of variables that are less directly correlated with the value in question, reducing the risk of subset noise interfering with the true shape of the data.

Besides the focus, in our data cleaning we should take greater steps to remove highly correlated data, rather than expect the subset selection methods to take care of this for us. The maximum speed and the Train Speed are highly correlated and thus could add unnecessary noise to the subset selection methods.

Additionally, efforts should be made to collect more complete and accurate data, as well as establishing better data standards for the Department of Transportation. This would help to ensure that the data used for analysis is reliable and comprehensive, allowing for more accurate identification of significant variables and improved cost estimation in the event of rail equipment accidents. Even without this, our data cleaning method could also have been improved. Rather than simply throwing away every instance of the data with the preselected variables that had an NA value in its rows, we instead should have simply found from the variables the instances which had the least missing, and within those instances where it was not missing fill in the missing values in the other attributes with either the mean or mode value, depending on if it was a continuous or categorical variable. While this would have added bias to the data, the additional instances would have most likely decreased it by even more.

Overall, while the selection methods selected were valid, more care should be taken the next time around in selecting data and finding additional sources to create a larger database to use for variable selection. Just as well, further care in pre-treatment and analysis of the data should be taken in future analysis.

References

Funk, J. (2023, April 26). *Norfolk Southern estimates Ohio derailment will cost $387M.* ABCNews.com. https://abcnews.go.com/Business/wireStory/norfolk-southern-estimates-ohio-derailment-cost-387m-98869616

Chen, J., Chen, X. & Yan, Y. (2021). An ensemble learning method for variable selection and its application on Railroad Fatal Accidents. *Journal of Physics: Conference Series,* 1955(1), 1-6. https://iopscience.iop.org/article/10.1088/1742-6596/1955/1/012065