Predicting rust test results for Lubrizol engine oil additives

Samuel Hardy
35245562
Lancaster University
s.hardy3@lancaster.ac.uk

Geyi Liu
35552231
Lancaster University
g.liu12@lancaster.ac.uk

Eugene Magee
35680875
Lancaster University
e.a.magee@lancaster.ac.uk

Alex Meehan
35793868
Lancaster University
a.meehan1@lancaster.ac.uk

Kieran Molloy
35762970
Lancaster University
k.molloy@lancaster.ac.uk

Thomas Harrison
35584689
Lancaster University
t.l.harrison@lancaster.ac.uk

ACM Reference Format:
Samuel Hardy, Geyi Liu, Eugene Magee, Alex Meehan, Kieran Molloy, and Thomas Harrison. 2021. Predicting rust test results for Lubrizol engine oil additives. In *Proceedings of* . ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

CONTRIBUTIONS

0.1 Samuel

- Ridge regression model
- Lasso regression model
- Models and Results (report)

0.2 Thomas

- Elastic Net model
- Logistic regression model
- Models and Results (report)

0.3 Geyi

- Linear regression model
- Random forest classification model
- Random forest regression model
- Models and Results (report)

0.4 Eugene

- Presentation writing and presenter
- Exploratory analysis
- Pre-processing (including report)
- Conclusion (report)

0.5 Alex

- Presentation writing
- Exploratory analysis
- Feature selection (including report)
- Research strategy, validity and potential biases (report)

0.6 Kieran

- Exploratory analysis
- Alternative Pre-Processing Strategy
- Stochastic Gradient Descent Model
- Gradient Boosting Model
- Extra Trees Model
- Feature Selection - Information Gain, Correlation
- Attribute Importance Ranking
- Hyper-Parameter optimisation, KFold Validation
- Robust models - RANSAC, Theil Sen and Huber
- Models and Results (report)

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Lubrizol is a leading global specialty molecular chemicals company, supplying businesses in the transportation, industrial and consumer markets. Lubrizol's Additives division makes advanced products to improve engine oil performance and this is the focus of our research. Additives are chemical compounds that improve the lubricant performance of a base oil, making up 10 to 30% of the final engine oil. The additives perform a variety of different roles, including preserving the base oil's main friction and wear properties, detergent cleaning, and preventing rust in the engine [4] [2]. The base oil is designed and produced by the oil company, Lubrizol provide the oil company with the additives, and the oil company then sell the final product to the vehicle manufacturer, repair shop or consumer.

Lubrizol have gathered rust prevention test results for 6,360 different combinations of additives and base oils, which formed the data set for this project. Physical laboratory tests can be both time consuming and expensive to perform, so an effective modelling process would make the product development process faster and lower cost. Each combination of base oil and additive interacts in a unique way, so both need to be considered together. Lubrizol have also found it to be more broadly applicable to consider the chemical properties of the additives and base oils, rather than the ingredients.

This paper considers several different approaches to modelling rust test results, given the chemical properties of different combinations of engine oil additives and base oils. Exploratory analysis is performed to understand the data, relevant data pre-processing techniques are utilised to maximise model effectiveness, and multiple regression machine learning techniques are investigated.

The purpose of this research is to investigate the features of oil that have the most impact on rust prevention, specifically:

- What chemistry groups are influential on a test?
- Within chemistry groups, which predictor variables best describe the effect?
- Are there any outliers in the data and how do they influence results?
- Build a robust model which predicts how chemistry performs for new formulations.

The primary objective of this research is to develop a robust model which predicts the rust test results for a given set of chemical properties for the additives and base oil. The robustness of the models will be assessed using the following metrics:

(1) Adjusted R^2
(2) Mean Standard Error
(3) Mean Absolute Error
(4) The number of variables required for the model
(5) Any compute time considerations, particularly for predicting
(6) A qualitative assessment on interpretability of the model

We will be assessing the most influential groups by first finding the most influential predictor variables within the groups and then aggregating back up to the groups. This will be done by finding the variables that have the most statistically significant contributions within the model. From this we will be able to identify the chemistry groups that contain the most significant variables, helping us to identify which ones are most influential.

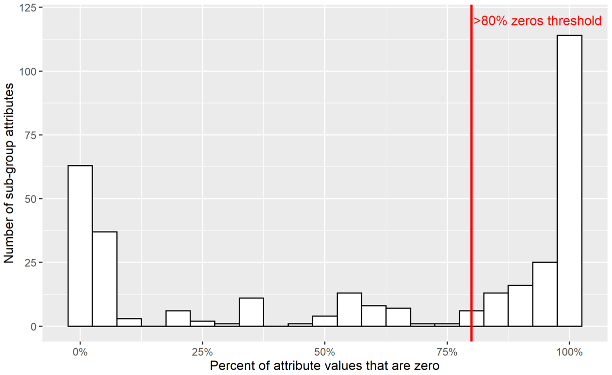


Figure 1: Histogram of % of attribute values that are zero

Our results show that Elastic Net is the most appropriate regression model for predicting rust test results, based on 15 chemical properties of the engine oil additives and base oil, balancing accuracy with simplicity and explainability.

2 METHODOLOGY

2.1 Feature selection

The most distinguishing feature of the Lubrizol dataset is that it is highly dimensional: The raw data has 810 attributes for 6,360 different tests. The 810 attributes break down as follows:

- 476 ingredient attributes
- 267 chemical sub-groups aggregated into 12 Groups relating to the engine oil additives
- A further 65 chemical sub-groups describing the base oils
- The lab the tests were conducted at: One of three (A/B/C)
- The response variable, the percentage rust in the test

The breadth of the raw data means that dimension reduction will be critical for successful modelling.

The data has been anonymised for commercial reasons, so it is not possible to apply domain knowledge at any stage in the process. The Lubrizol briefing advised us to base the modelling on the chemical groups, rather than ingredients, as this was simpler and more useful.

The base oil is determined by the oil manufacturer, so these 65 sub-groups are important but cannot be varied by Lubrizol. This means that we should model them, but consider their impact separately to the other 12 chemical groups for the engine oil additives.

Several of the chemical group attributes contained a large proportion of zero values, as shown in Figure 1. The chart shows a particular peak of 133 attributes that have >95% zero values. Chemical sub-groups with >80% zero values were excluded from the modelling, as they were unlikely to have a significant impact on overall results and may not have sufficient sample sizes to be representative. This was confirmed by re-running a simple linear regression model with a >90% threshold and achieving the same outcomes. This meant that 169 attributes were excluded from modelling because >80% values were zero (163 Group 1-12 attributes + 6 Group 13 attributes)

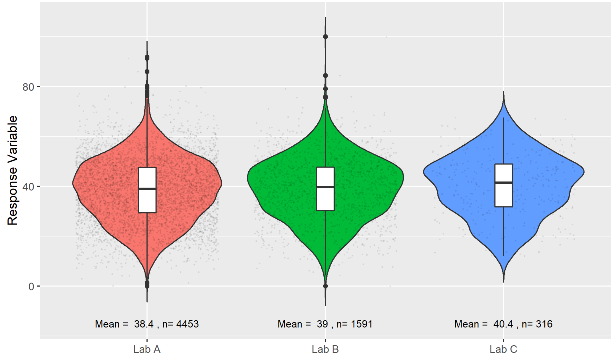


Figure 2: Violin Plot of the Response Variable for each Lab

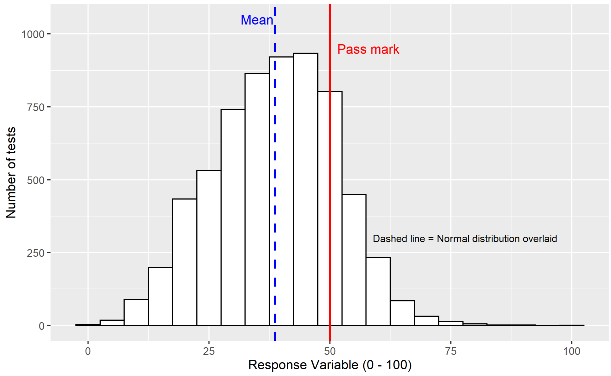


Figure 3: Histogram of the Response Variable

The test results were compared across the three different labs, shown in Figure 2. The test response scores were similar across all 3 labs, with a range of 2 percentage points, which means that we do not need to either correct the data for this or include the Lab attribute in the modelling.

The net result of these feature selection steps is that the original 810 attributes are reduced to 163, with limited impact on accuracy.

2.2 Pre-processing

Principle component analysis was explored as a potential avenue for transforming the data and reducing the dimensionality. A preliminary PCA showed us that we would require 22 variables to explain 90% of the variance in the original data. On balance, we decided against PCA transformation, since the reduction in variables is not enough to offset the variables losing interpretability, making it harder to discover the most significant groups and sub-groups.

One important consideration was that the variables had vastly different ranges with some only going from 0 to 0.5 and others going above 1000. This made it clear that standardisation or normalisation had to be done. We decided on doing both of these since they achieve similar goals of bringing variables into comparable ranges but some models may function better with standardisation and others my rely on normalisation to get the best results. Standardisation sets the mean to 0 and the standard deviation to 1 for each feature whereas the normalisation we performed sets the minimum value to 0 and maximum value to 1 for each feature.

We also investigated outliers and found that, using a threshold of 6 standard deviations, a small percentage of the values were considered outliers and these equated to around 0.24% of all the values in the data table. The number of outliers was reduced to 0.13% of the values once feature selection was done, however the number of observations containing an outlier was around 10% of the total number of observations. We decided to leave these outliers in since it felt wrong to eliminate an observation based on a single outlier value out of over 160 features. We will investigate whether or not this will have a significant effect on the accuracy of the models. Investigating this can be done by creating and testing models both with and without outliers removed so that we can compare how these models differ.

The effect of these outliers was tested on the similar linear model discussed later. We found that, keeping outliers in, the adjusted R^2 of the model was 0.6235. Once we removed the outliers and ran the same model with the same variables we arrived at an adjusted R^2 of 0.6237 showing little change as a result of keeping them in. It should also be noted that a small number of variables became less significant as a result but the significant variables remain highly significant and so we can say that the inclusion of outliers is unlikely to significantly effect our conclusions or the accuracy of models.

2.3 Research strategy

The distributions of the 170 model attributes were reviewed via an R Shiny app, which showed that they are mainly exponentially distributed, with some normal distributions. This means that models not assuming a normal distribution are likely to perform better.

The distribution of the output variable was also reviewed and can be seen in Figure 3. It is approximately normally distributed, with 19% of test results considered a "pass" (>50). The main research question could be approached as a classification problem (>50), which is Lubrizol's primary priority. However, it would be even better to also predict the response output via regression, as this provides a more detailed evaluation of the additive (e.g. did it just fail or fail by a long way?) and this additional information could be useful in developing products further (e.g. make a small adjustment or try something very different).

There are two other salient characteristics of the data set that are relevant to our research strategy. The first characteristic is that a large number of attributes are correlated with the response variable: Of the 163 features selected in pre-processing, 5 attributes have an absolute Pearson Correlation Coefficient >0.5, 65 attributes >0.25 and 93 attributes >0.1. The second characteristic is the high level of collinearity across attributes, which is demonstrated by the large proportion of the selected features that have a high Variance Inflation Factor, shown in Figure 4.

These two characteristics suggest that further feature selection steps will be required in the modelling process, which will need to be sophisticated enough to accommodate the large number of attributes are correlated with the response variable and the high levels of collinearity. They also suggest that there are still likely to be a large number of variables involved at the end of modelling

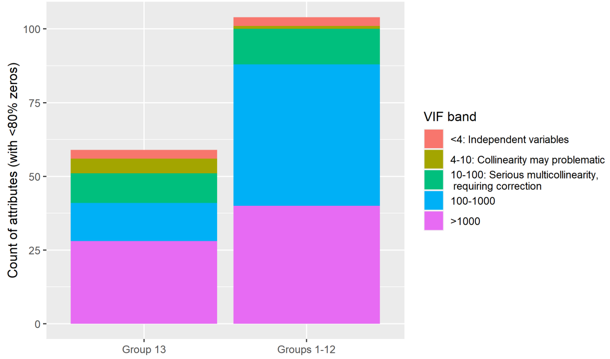


Figure 4: Number of attributes within each Variance Inflation Factor (VIF) band

process, so our approach will be to model at the sub-group level and then aggregate results back to the overall group level at the end of the process.

To enable analysis of modelling results, the data was split into 70% records for training the model (4,452 records) and 30% for testing (1,908 records).

2.4 Models considered

The first model considered was a full linear regression. Linear regression works by treating the response variable as a linear combination of the explanatory or predictor variables. This was selected to act as a very simple initial model designed to give us a preliminary look at the relationships between the variables and the response. It was then possible to do a form of feature selection on the variables in this model by performing two different types of selection: backwards selection and stepwise selection. Both of these forms of selection work by removing the least statistically significant variables and reducing the number of variables in the model while maintaining similar accuracy. In the case of backwards we simply remove the least significant variables whereas stepwise involves taking multiple steps to remove variables based on their significance and variance inflation factor, VIF.

Next we investigated ridge regression, which was selected since it is able to account for multicollinearity [10]. The large amount of multicollinearity within the data can impose limitations on the accuracy of a standard linear regression. Problems can occur because collinearity causes standard errors on the regression coefficients to be large, which can create inaccurate regression coefficients [1]. Inaccurate coefficients will lead to inaccurate predictions from the model, so limiting this should be a priority.

Least Absolute Shrinkage and Selection Operator (or LASSO) regression was also investigated as it can perform a form of variable selection on the data, further reducing the number of variables required in a model and making the model easier to interpret [6]. This is another technique that can be used to provide us with a definitive list of the most influential variables and, as a result, the most influential chemical groups. LASSO regression can also help to increase the accuracy of the model predictions [9], potentially making it a useful tool in creating a robust predictive model.

Elastic Net regression combines the benefits of both ridge regression and LASSO regression by creating a model that accounts for multicollinearity while also selecting variables and regularising the data. This allows it to overcome some of the limitations of the two previous models and so it was a desirable model type for us to consider [5]. Elastic Net should also allow us to find the most influential variables due to the variable selection it can perform.

We are able to classify data using a random forest, a type of ensemble classifier. Random forest is a supervised learning algorithm and an extension of decision tree [8]. Random forests work by training many random decision trees and predicts the label of an observation by combining the results of all trees. They are an incredibly powerful classification tool which typically has high accuracy, making it desirable. It is not sensitive to collinearity, and the results are robust to missing and unbalanced data [3]. Furthermore, it can effectively deal with highly dimensional data sets, and does not need dimension reduction. The most important useful feature for this project is that they are able to measure the relative importance of variables, which allows us to identify which ones are significant. The trade off is that the training time is long for a large number of trees.

3 RESULTS

3.1 Linear models

The variables from the linear regression model are shown in figure 5, and Table 1 records some metrics related to the performance of the linear model. All variables in the model are statistically significant, and there is no collinearity problem in this model. The linear model also allowed us to find coefficients for each variable. These coefficients mean the effect of each variable on the Response variable. For instance, Group1_6 has a negative effect on the Response variable which means that as the value of Group1_6 increases, the value of Response decreases. We can see that Group1_11 had a significant impact on the model and Group13 had the most variables in this model.

3.2 Ridge

Table 1 shows that ridge regression offers statistically insignificant improvements in $adjusted\ R^2$, MSE, and MAE. However, the ridge regression model does utilise three fewer variables while providing similar performance metrics, which highlights the reduction in model complexity mentioned previously. From the thirteen variables used in the ridge regression model, seven belonged to Group1, including two of the four largest coefficients.

Since variable selection is the aim of these regression models the commonalities between the selected attributes of each model should be noted. Due to the similarity between ridge and linear regression it is unsurprising that 30.76% of attributes that appeared in the ridge regression model are also present in the linear regression model, namely Group1_11, Group1_20, Group13_5, and Group13_6 and this can be seen in figure 5. In contrast to the linear model the results of ridge regression determine Group1 to be the most significant chemical group instead of Group13, which is the second most frequent and influential group in the ridge regression model.

3.3 LASSO

Fourteen variables are utilised by the LASSO regression model, 28.57% of these variables were also present in the ridge regression model, where as 50% were also found in the linear model. These results further evidence the similarity between all three of these approaches that is shown via the performance data in table 1. Only two attributes were present in the linear, ridge, and LASSO models, these were Group1_11 and Group13_6, with Group1_11 being the most significantly weighted feature in all models, this furthers the impact of Group1_11 as suggested in section 3.6.

The features selected by LASSO share much in common with those chosen via linear regression as previously stated, this includes the ordering of the most impactful chemical groups. While Group1_11 is the single most influential variable, Group13 has the most occurrences in both models, followed closely by Group2. This provides an interesting insight into the types of chemicals which effect Response, especially since the data is corroborated across multiple models.

3.4 Elastic Net

Elastic Net regression combines feature elimination from Lasso regression and feature coefficient reduction from Ridge regression which makes it an ideal choice of model. This model uses a five fold cross validation method to find the values of the parameters alpha (mixing parameter) and lambda (regularisation parameter) that best optimise the model.

Selecting the top fifteen attributes showed that Group1, Group2 and Group13 were the most influential on the model with Group1_11, Group1_9, Group2_18 having the strongest positive contribution and Group13_50, Group13_6 having the most negative contribution. After building a new dataset composed of these top fifteen most influential values, Elastic Net was used as a classification algorithm by taking a 30:70 split of test/train data and converting the prediction response output into a binary format based on the prerequisite of what classifies as a 'pass' and 'fail'.

As shown in table 2, taking the positive class as '0' (corresponds to a fail), the accuracy of the model was found to be 0.935 with a 95% confidence interval of (0.930, 0.944) and a balanced accuracy of 0.768 which shows that the model predicted the correct response primarily in the 'fail' cases and was more inaccurate in predicting the 'pass' which will be due to the class imbalance. Since there is a large class imbalance where the negative class is the minority. To account for the class imbalance, the weighted classification metrics were considered. The lowest classification metric was G-mean which was recorded as 0.7491, this measures the balance between the class majority and class minority. All other metrics were calculated to be upward of 0.93 which indicates this model predicted accurately overall.

Elastic Net took 6.96 ± 0.11 minutes to perform the regression model and 6.48 ± 0.5012 seconds to predict the response variable in the classification method.

3.5 Logistic regression

Figure 5 shows all the sub-groups that were used in the final logistic model, Group1_11 and Group2_9 have the strongest contribution towards the model predicting a pass since they had a significantly larger estimate value compared to the other positive estimates where as Group13_9 contributes the most towards a fail.

Logistic regression was used as a classification model by taking a 30:70 test/train split of the original dataset to train the logistic model and create predicted values for the response variable using the test data. Implementing a classification matrix allows a variety of different classification metrics to be reviewed to determine the performance of the model; the accuracy of the model was found to be 0.934 with a 95% confidence interval of (0.93, 0.94) indicating that the model predicted the majority of the response variables correctly. Table 2 shows the different weighted classification metric values. Table 2 shows that both the balanced accuracy and G-mean values are notably lower in comparison to the rest of the values, this could a result of the class imbalance between pass and fail. Since 19% of the values in the response variable relate to a pass then there could be a lack of training data for the model to accurately create a model which can detect the regularities.

The R-squared that is used in ordinary least squares regression can not be applied to logistic regression, this is because logistic regression does not aim to minimise variance but calculates estimates of maximum likelihood[7]. To evaluate the goodness of fit for a logistic regression model there are pseudo R-squared statistics however, the is no universally accepted best version. Each version of pseudo R-squared statistics are calculated slightly differently and give their own interpretation making it difficult to chose which statistic best suits the model. The pseudo R-squared values calculated for this model was McKelvey Zavoina which produced a pseudo R-squared value of 0.8114772 which is an indication that the model is of a good fit.

3.6 Random Forest regression model

One of the beneficial characteristics of random forest model is that we can list the variables that have important influence on Response variable. Here we choose the top 15 variables that are most important to Response in the random forest model. % IncMSE and Incnodepurity are two indexes to judge the importance of predictive variables. The larger their values are, the more important their corresponding variables are.

Comparing the two indexes, we can find that Group1_11 and Group1_9 are the most important variables in regression model. The top 15 variables of importance are concentrated in Group1, Group2 and Group13. Because random forest is not a linear algorithm, its regression model results will be different from the linear regression results. By comparing the results of the two models, we can find that Group1_11 is of great importance to Response in both models. In order to verify the importance of variables, we build a new regression model with only the top 15 variables of importance. Here we discover that the performance of the model constructed by the first 15 variables based on % IncMSE is better than that based on IncNodePurity. The model performance metrics based on IncMSE are shown in Table 1. Compared with the original model, the % Var explained value of the new model only decreased by 0.31. When the number of variables is greatly reduced, the value of % Var explained does not decrease much. This proves that these 15 variables are very important in the model.

Table 1: Regression Performance Metrics

Model	Variables	Adjusted R^2	MSE	MAE
Linear	16	0.6235	60.3025	5.9741
Ridge	13	0.6267	59.6350	5.9939
LASSO	14	0.6215	60.4573	5.9942
Elastic Net	15	0.6708	61.7424	5.8377
Random Forest	15	0.6894*	50.0376**	-

*: % Var explained (similar with R-squared, a measure of goodness of fit).
**: Mean of squared residuals.

Table 2: Weighted classification metrics for the predicted response value using the optimal subgroups.

Model	Elastic Net	Logistic Reg.	Random Forest
Accuracy	0.9373	0.9352	0.9099
Balanced accuracy	0.7681	0.8231	0.8154
Precision	0.9373	0.9351	0.9098
Recall	0.9373	0.9351	0.9098
F1-score	0.9373	0.9351	0.9098
G-mean	0.7491	0.8154	0.8099

G-mean = sqrt(sensitivity * specificity)

3.7 Random Forest classification model

A random forest can also be used to build classification model. The random forest also achieved good results in classification. The relevant classification metrics for the random forest classifier are listed in Table 2. It should be noted that the results of the random forest model will change with different seeds, but the change in the prediction accuracy will be small.

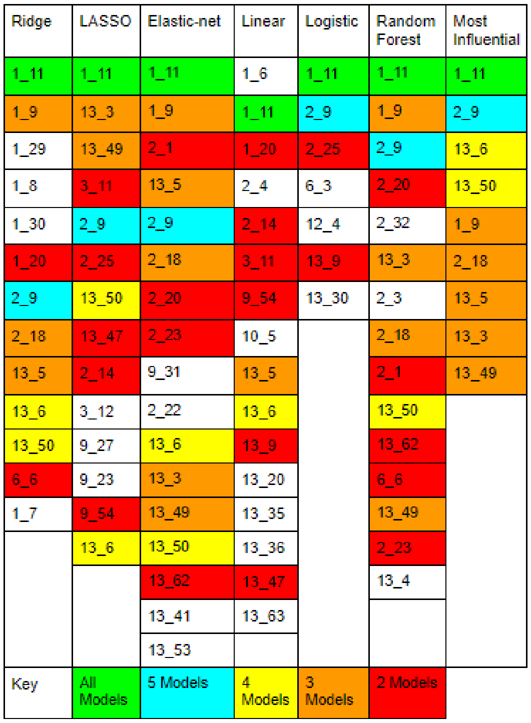
In this classification model, we also find 15 variables that are most important to classification. There are two indexes to measure the importance of variables. Mean Decrease Accuracy shows how much our model accuracy decreases if we leave out that variable. Mean Decrease Gini is a measure of variable importance based on the Gini impurity index used for the calculating the splits in trees. Comparing the two indexes, we can find that Group1_11, Group1_9 and Group1_2 are the most important variables that affect whether the product passes the test. And the top 15 variables of importance are concentrated in the Group1 and Group2.

3.8 Overall results comparison

The accuracy of each regression model is assessed in Table 1 (Linear/ Ridge/ LASSO/ Elastic Net), which shows that Elastic Net performed best across all metrics, with an *adjusted* R^2 of 0.67 for 15 variables. Ridge, LASSO, and Elastic Net are similar in regression approach and thus in performance, with only 0.0197 difference in *adjusted* R^2 between Elastic Net and it's nearest competition (Ridge). This result may be reflective of Elastic Net regression aiming to combine the best of Linear, Ridge and LASSO modelling.

The performance of the Elastic Net regression model was then compared to two classification models in this and Table 2 (Random Forest / Logistic Regression). Both classifiers performed better than Elastic Net for classification metrics, but they provide less informative output data (pass/fail) compared to regression approaches (continuous prediction values).

The most significant attributes for each model are shown and compared in Figure 5. Three chemical groups reoccurred consistently in both the final models and during development: Group1, Group2, and Group13. There were 9 sub-groups that were present in three or more of the six models, shown on the right-hand side of the grid. Some of the inconsistency in attributes selected by each model may be explained by the high levels of multicollinearity between chemical sub-groups, as they are interchangeable from a mathematical perspective, as the effect of the attribute may be linearly related to and thus predictable by another attribute in the same chemical group.



Ridge	LASSO	Elastic-net	Linear	Logistic	Random Forest	Most Influential
1_11	1_11	1_11	1_6	1_11	1_11	1_11
1_9	13_3	1_9	1_11	2_9	1_9	2_9
1_29	13_49	2_1	1_20	2_25	2_9	13_6
1_8	3_11	13_5	2_4	6_3	2_20	13_50
1_30	2_9	2_9	2_14	12_4	2_32	1_9
1_20	2_25	2_18	3_11	13_9	13_3	2_18
2_9	13_50	2_20	9_54	13_30	2_3	13_5
2_18	13_47	2_23	10_5		2_18	13_3
13_5	2_14	9_31	13_5		2_1	13_49
13_6	3_12	2_22	13_6		13_50	
13_50	9_27	13_6	13_9		13_62	
6_6	9_23	13_3	13_20		6_6	
1_7	9_54	13_49	13_35		13_49	
	13_6	13_50	13_36		2_23	
		13_62	13_47		13_4	
		13_41	13_63			
		13_53				
Key	All Models	5 Models	4 Models	3 Models	2 Models	

Figure 5: Grid showing the occurrence of subgroups

3.9 Exploration of other approaches

To understand how far the results could be improved with more complex approaches and many more variables, a number of alternatives were explored.

A different pre-processing approach was developed, using the full dataset (ingredients, group attributes and lab) and then ranking attributes using the feature importances from a gradient boosting regression model. Further to this, using a correlation feature selection and a mutual information gain model, the model can be reduced further without accuracy penalties and attaining lower computation times.

Robust regression was used to fit a regression model in the presence of corrupt data: either outliers, or error in the model. RANSAC (RANdom SAmple Consensus) fits a model from random subsets of inliers from the complete data set. RANSAC is a non-deterministic algorithm producing only a reasonable result with a certain probability, which is dependent on the number of iterations - this does not give good results in the current implementation but could be improved with parameter tuning. The Theil Sen Regression model uses a generalisation of the median in multiple dimensions. It is thus robust to multivariate outliers, but the robustness of the estimator decreases quickly with the dimensionality of the problem. This had an impact as feature reduction methods reduce the dimensionality of the problem. The Huber Regression model applies a linear loss to samples that are classified as outliers. It differs from Theil Sen and RANSAC because it does not ignore the effect of the outliers but gives a lesser weight to them. This performs strongest out of the robust methods.

The strongest results are found from extremely randomised trees with kfold crossover and grid search to optimise parameters, with an R^2 of 0.6995, MSE of 45.2316 and MAE of 5.1230. Mutual information feature reduction is unable to decrease computation time, where as the Correlation method does - by 6 seconds. Whilst the Attribute Information Ranking method is able to reduce the computation by a factor of 6 whilst only losing 0.5% accuracy.

3.10 Research validity and potential biases

The construct validity of this research is dependent on the assumption that the chemical process of rust can be analytically modelled using data on the properties of the additives and base engine oil. Whilst this research team do not have the domain knowledge or information on what the specific properties are, it is known that chemical processes are not necessarily a scalar function and can often require specific combinations for a given outcome. Some of these characteristics may be challenging to reflect in even the best of regression models. In addition, this analysis is dependent on having data for all of the relevant properties for the additives and oil. Whilst there are 332 properties in the data, it cannot be certain that includes all relevant properties. Another consideration is that the physical rust test results may not entirely accurately represent real-world corrosion challenges in an engine.

Even the most accurate models have potential conclusion validity issues, as at least 20% of error or variance unexplained, which could be connected with the construct validity issues just described.

Perhaps the most significant consideration is that our approach focuses on minimising the overall error for the data set of test results, which will inherently bias the models to the more common tests. New Product Development is likely to explore less common or even novel areas, where a single overall model may not be accurate. The opportunity cost for Lubrizol could be very significant if a good innovation were stopped incorrectly, so one potential development could be to use domain knowledge to develop models for each technology area, however Lubrizol scientists choose to define those. This would ensure that no technology area is under-represented and the best commercial outcomes are achieved.

Results are not generalised in this research, but an unanswered external validity consideration is whether a similar modelling approach could also be successful in predicting the test results for other common engine oil tests, or even, more broadly, other chemical tests.

4 CONCLUSIONS

In this project we investigated chemical data provided by Lubrizol in order to identify the most influential chemical groups and variables on rust prevention in engine oils. We used this data to create a number of models and identified the best of these models, based on accuracy measures, simplicity and explainability. We found that the elastic-net was the best regression model, allowing for an accurate prediction of the rust response variable. We also found that logistic regression was the most robust model for classification if Lubrizol wished to determine simply whether an additive will pass or fail the rust test. Across all models, groups 1, 2, and 13 were identified as the most influential chemical groups, although it should be noted that Group13 is determined by the given base oils and not the additives. It also appeared from our results that the most influential sub-group variables were Group1_11 and Group2_9.

We found a small number of outliers in the data, equating to around 0.13% of the values in the reduced data set. We tested a simple linear model with outliers in and with outliers removed and found that it had no significant impact on the model and, as a result, we kept the outliers in for modelling so as to preserve the data.

Our approach to model creation provided us with a number of useful models for predicting the rust performance of engine oil additives however there are other ways this could have been approached. For example, we could have started the project by identifying the most influential groups and variables and then use these to create more advanced blackbox models. An example would be the use of Neural Networks which are extremely powerful tools but they are not particularly easy to interpret meaning it could be difficult to establish influential groups from them. Finding the influential groups before using a Neural Network may allow us to circumvent this issue. With this method we may also be able to perform a principle component analysis, helping us to force independence of the features which would overcome the multicollinearity issues.

The exploration into other approaches gave some promising results with models that appeared to perform well both in terms of machine learning metrics and computation time. Further investigation could be done to determine the suitability of these more complex models for Lubrizol and the project brief.

This modelling approach could be developed further by using Lubrizol domain knowledge to develop specific models for each technology area, to ensure the model is effective in research areas outside of the most common historical tests. It would be interesting to investigate, with the help of additional data from Lubrizol, whether our modelling techniques could also be applied to other oil evaluation tests such as wear, friction, or viscosity.

REFERENCES

[1] Aylin Alin. 2010. Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics* 2, 3 (2010), 370–374.

[2] M Gulzar, HH Masjuki, MA Kalam, M Varman, NWM Zulkifli, RA Mufti, and Rehan Zahid. 2016. Tribological performance of nanoparticles as lubricating oil additives. *Journal of Nanoparticle Research* 18, 8 (2016), 1–25.

[3] Biau Gérard and Scornet Erwan. 2016. A random forest guided tour. *TEST* 2, 25 (2016), 197–227. https://doi.org/10.1007/s11749-016-0481-7

[4] GS Kapur, A Chopra, AS Sarpal, SSV Ramakumar, and SK Jain. 1999. Studies on competitive interactions and blending order of engine oil additives by variable temperature 31P-NMR and IR spectroscopy. *Tribology transactions* 42, 4 (1999), 807–812.

[5] Joseph O. Ogutu, Torben Schulz-Streeck, and Hans-Peter Piepho. 2012. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC proceedings* 6 Suppl 2, Suppl 2 (21 May 2012), 1–6.

[6] J. Ranstam and J. A. Cook. 2018. LASSO regression. *British Journal of Surgery* 105, 10 (Sept. 2018), 1348–1348. https://doi.org/10.1002/bjs.10895

[7] Thomas J Smith and Cornelius M McKenna. 2013. A comparison of logistic regression pseudo R2 indices. *Multiple Linear Regression Viewpoints* 39, 2 (2013), 17–26.

[8] Vladimir Svetnik, Andy Liaw, J. Christopher Culberson Christopher Tong, Robert P. Sheridan, and Bradley P. Feuston. 2003. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences* 6, 43 (2003), 1947–1958. https://doi.org/10.1021/ci034160g

[9] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 1 (1996), 267–288.

[10] John Zhang and Mahmud Ibrahim. 2005. A simulation study on SPSS ridge regression and ordinary least squares regression procedures for multicollinearity data. *Journal of Applied Statistics* 32, 6 (2005), 571–588.