**Packaging Made Faster: BMW Automated Logistics Value Chain**

Kenny Barza

Karen Bassil

Emily El Hajj

Jason Eid

Presented to: Dr. Imad Bou-Hamad

**Abstract**

Problems in transportation and logistics had to be tackled long before computers became available to support decision making. Over the years, technology has evolved and problem decision, specifically in logistics and supply chain, has shifted from the hefty manual decision making to the more accurate and robust automated decision making. Like many other automotive companies, BMW Logistics Group wants to find the best way to fit certain vehicle products into certain packages which in return go into particular load units. In this paper, different classification and regression models are discussed and evaluated on the provided data to find ways of predicting packaging features from product features, and methods of predicting required load unit features from packaging features. Although our findings in predicting the numerical packaging features were limited due to many important outliers that should not be omitted from the data, we believe that in an ideal workspace where technical and logistics staff are present, many different ways can be approached to solve that issue. Finally, it is believed that the generated categorical packaging and load unit models are successful at predicting packaging variables and load unit features respectively.

**1. Introduction & Literature Review**

Along with the improvement in technology, it is no surprise that businesses are further focusing on implementing more advanced technological and analytical infrastructure into their daily operations. Looking deeply into a company’s organizational value chain, logistic faces numerous challenges as it requires superior planning in order to meticulously transport, store, and receive products into the company. Therefore, automating such process would enable BMW logistics group to optimize its operations systems and consequently reduce cost as well as complexity.

A research done by Knoll, Prüglmeier, & Reinhart (2016) shows that to be able to build a packaging model, we must first analyze existing combinations of material numbers. The latter constitute of the product’s weight as well as volume which in turn represents the three dimensions of x, y, and z. Another study explains that the quality index of a product can indicate the condition requirements of a package and thus, they concluded that the greater the index, the higher the likelihood that there is a superior packaging quality requirement (Knoll, Neumeier, Prüglmeier, & Reinhart, 2019).

Another way to predict numerous variables is by implementing the Random Forest approach. As explained by Ali et al. (2012), the random forest method has its advantages as it overcomes the issue of overfitting and can properly handle any outliers in the data as well as missing values.

Not that, random forest techniques represent an ensemble of trees used for classification and regression problems. A random forest takes an input and builds multiple decision trees. As output, the class/leaf nodes that have the highest mode or mean represent the prediction result.

**2. Problem Description**

BMW Logistics Group is facing a problem of fitting the product into the right package, and the package into the right load unit. In fact, the process of logistics planning is as follows; different products present distinct dimensions, weight, and requirements, and these products should be fitted into a specific packaging which in turn possesses certain dimensions, weight, and qualities. Consequently, the packages must also fit into load units where each unit presents different features. Traditionally, BMW Logistics Group used to check manually if a product fits certain packaging, and if a package fits certain load unit. Thus, doing this for each item is time and cost consuming. As a result, we decided to develop multiple regression and classification models that could potentially predict autonomously each features part and facilitate this tedious task.

**3. Data Description**

The dataset was provided by the Logistics Group of BMW Automotive in Germany. It contained data related to product, packaging, and load unit features. Numerical variables related to the dimension and weight of each product, packaging, and load unit was provided as well as the capacity of each packaging and load unit and the quality index of all products. Additionally, categorical variables were presented of whether a load unit and/or a packaging is one way or can be used several times, and/or whether it requires special wrapping or not. A more detailed metadata table can be found in the appendix under table x.

After examining our data, few anomalies were detected for which dimension values were equal to 0. Since this cannot be possible for the fact that our packages and load units are in rectangular shape, we decided to clean the data by omitting the rows that contained irregularities.

**4. Results and Discussion**

**Packaging X-Dimension, Packaging Y-Dimension, Packaging Z-Dimension**

To begin with, the variables to be predicted are packaging x, y, and z dimensions using linear regression, decision tree and random forest approach. In brief, linear model 1 contained all the numerical and categorical features. Linear model 2 included only the numerical variables. Linear model 3 was built by using log on the numerical variables. For x-dimensions, model 1 and model 2 present approximately the same RSME, whereas model3 which takes into consideration the linearity of the model by using log on the numerical variables has slightly lower RMSE than model 1 and 2. In terms of decision trees, to create model 4 the lowest x error was selected, and the outcome was a very complex tree with 304 splits, so we opted for another model 5 with a fair enough x error in order to simplify the tree. Both models 4 and 5 have lower RMSE than the linear regression models. However, if a decision was to be made in order to choose the best model, it would be the random forest which has the lowest RMSE of nearly 177cm.

For the packaging y dimension, the first three models are similar in terms of predictors with the ones presented in packaging x-dimension. Plus, another model numbered 4 was created and includes only the significant variables. For the linear regression models, the lowest RMSE belongs to the model 3 which accounts for linearity by using log on the predictors. Model 5 with 321 splits and model 6 with 22 splits both have RMSE lower than the linear regression models and are nearly equal, and as result, we would pick model 2.

Finally, for the packaging z dimension, it is prevalent that all three models in linear regression have approximately the same RMSE whereas the decision tree models 4 and 5 have lower RMSE. Model 4 constitutes of 301 splits because of the selection of the lowest x error whereas model 5 comprises of 36 splits with an RMSE almost equal to model 5. The random forest has the lowest RMSE in this case. However, if we consider the packaging z-dimension’s mean to be equal 393.8cm and compare it with 2(RMSE)~(355.54), then none of the models studied serve as an optimal predictor for this variable.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | |  | RMSE | Predictors/Splits/Trees | Mean |
| Linear Regression | | **Model 1** | 423.3578 | 7 predictors | 826.617 |
| **Model 2** | 424.1813 | 5 predictors |
| **Model 3** | 361.6160 | 5 predictors |
| Decision Tree | **Model 4** | | 226.8368 | 304 splits |
| **Model 5** | | 267.2891 | 25 splits |
| Random Forest | **Model 6** | | 177.5136 | 150 trees |

Table 1 – Accuracy measures for the regression models of packaging x dimension

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | |  | RMSE | Predictors/Splits/Trees | Mean |
| Linear Regression | | **Model 1** | 268.2202 | 7 predictors | 548.4564 |
| **Model 2** | 268.2817 | 6 predictors |
| **Model 3** | 268.4865 | 5 predictors |
| **Model 4** | 207.8145 | 7 predictors |
| Decision Tree | **Model 5** | | 154.3292 | 321 splits |
| **Model 6** | | 154.3459 | 22 splits |
| Random Forest | **Model 7** | | 177.7773 | 150 trees |

Table 2 – Accuracy measures for the regression models of packaging y dimension

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | |  | RMSE | Predictors/Splits/Trees | Mean |
| Linear Regression | | **Model 1** | 361.3197 | 7 predictors | 393.8238 |
| **Model 2** | 312.9610 | 5 predictors |
| **Model 3** | 330.1474 | 5 predictors |
| Decision Tree | **Model 4** | | 259.23030 | 301 splits |
| **Model 5** | | 258.5310 | 36 splits |

Table 3 – Accuracy measures for the regression models of packaging z dimension

**Packaging Weight**

This section tackles the weight of the package predictor using linear regression, decision tree and random forest approach. Model 1 and 2 have similar RMSE, but the difference between them is that model 1 includes all the predictors whereas model 2 contains only the 5 numeric predictors (product\_x\_dim, product\_y\_dim, product\_zdim, product\_weight, product\_quality\_index). Model 3 accounts for the linearity of the model since the residual vs. fitted graph did not demonstrate a horizontal line, therefore, log was used on the predictors but the RMSE seems to be higher than model 1 and model 2. The model 4 and model 5 belonging to the decision tree approach, both have same RMSE and lower than the previous models. The random forest method demonstrates the lowest RMSE amongst the rest of the models, however, this still does not make it the optimal model to be used, since the 2(RMSE)~(59) is higher than the mean which is equal to 41.1kg.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | |  | RMSE | Predictors/Splits | Mean |
| Linear Regression | | **Model 1** | 66.7674 | 7 predictors | 41.15612 |
| **Model 2** | 66.9032 | 5 predictors |
| **Model 3** | 73.4427 | 5 predictors |
| Decision Tree | **Model 4** | | 39.0807 | 193 splits |
| **Model 5** | | 39.2238 | 13 splits |
| Random Forest | **Model 6** | | 29.535 | 150 trees |

Table 4 – Accuracy measures for the regression models of packaging weight

**Product Per Packaging**

For the product per packaging, models 1 and 2 both belonging to the linear regression approach, both have the same RMSE, and the difference between them is that model 2 omitted all the insignificant variables making the number of predictors only 3. Model 3 has a slightly lower RMSE compared to model 1 and 2. Model 3 implemented the use of log on the predictors for the sake of linearity. The decision tree approach models have lower RMSE than the linear regression model. While model 4 has 93 splits and model 5 has 29 splits, both present same RMSE. The random forest approach offers the least RMSE amongst the other models, however the RMSE is approximately equal to the mean. As a result, none of the model are suited to predict accurately the outcome.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | |  | RMSE | Predictors/Splits | Mean |
| Linear Regression | | **Model 1** | 156.534 | 7 predictors | 101.6709 |
| **Model 2** | 156.758 | 3 predictors |
| **Model 3** | 145.898 | 5 predictors |
| Decision Tree | **Model 4** | | 121.584 | 93 splits |
| **Model 5** | | 120.804 | 29 splits |
| Random Forest | **Model 6** | | 102.9318 | 150 trees |

Table 5 – Accuracy measures for the regression models of product per packaging

**Packaging Load Capacity**

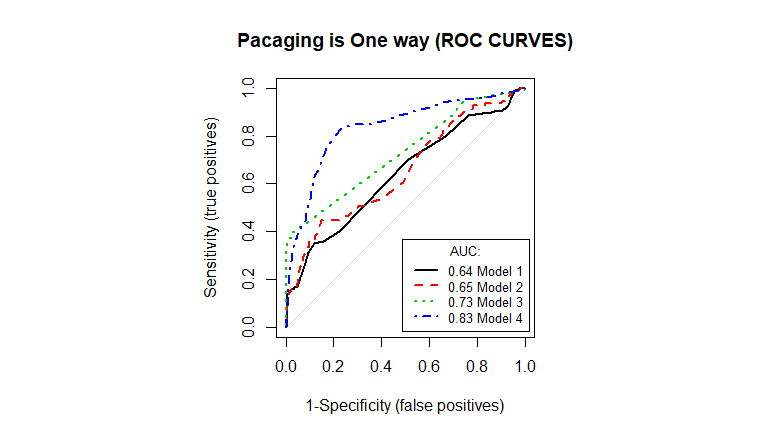
For the variable packaging load capacity, all of the models studied have a large RMSE in comparison with the mean, expect model 7. The latter could be deployed by the BMW logistics group to predict the outcome.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | |  | RMSE | Predictors/Splits | Mean |
| Linear Regression | | **Model 1** | 556.6301 | 7 predictors | 356.661 |
| **Model 2** | 556.5789 | 6 predictors |
| **Model 3** | 567.0656 | 5 predictors |
| **Model 4** | 498.8873 | 5 predictors |
| Decision Tree | **Model 5** | | 394.7114 | 291 splits |
| **Model 6** | | 420.7495 | 36 splits |
| Random Forest | **Model 7** | | 102.9318 | 150 trees |

Table 6 – Accuracy measures for the regression models of packaging load capacity

**Packaging is One Way**

In this section, we will be predicting the variable packaging\_is\_oneway using the logistic regression approach as well as the decision tree method. In brief, both models 1 and 2 included all the variables related to the product features, however, in model 2 we deployed the over-sampling and under-sampling technique simultaneously due to the low number of important class associated with the packaging is one-way variable that is presented in our data. In addition, models 3 and 4 were extracted from the decision tree model which also contained all the variables linked to product features. Model 3 was selected for the fact that it has the lowest validation error, and model 4 decision tree approach took into consideration the over and under sampling technique simultaneously.

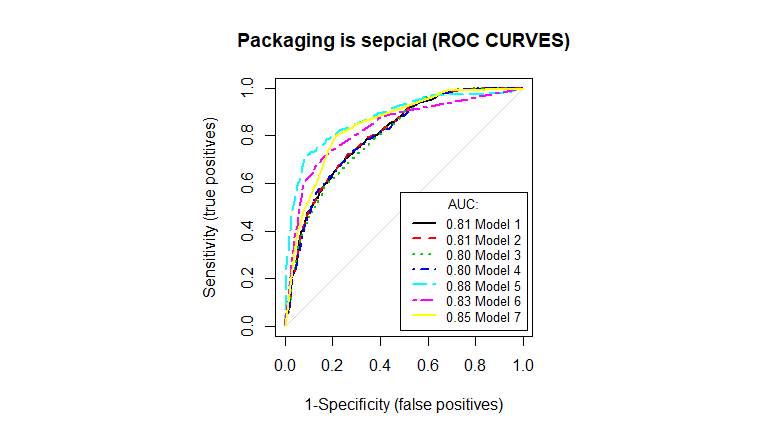
Table 7 summarizes the accuracy measures as well as the area under the curve scores for each model. To begin with, even though model 3 presents the highest overall accuracy, its sensitivity ratio is one of the lowest, so deploying such model would not be beneficial when predicting our variables since we mainly care about our important class which is to accurately predict if our packaging should be one way. Consequently, while model 4 presents the lowest overall accuracy, it also has the highest sensitivity ratio of approximately 75.2% as well as the highest AUC which is equal to 82.9%. The latter model results in better prediction related to the important class so deploying it might be considered the best approach. Figure 1 – Packaging is one-way ROC Curves

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  | Overall Accuracy | Sensitivity | Specificity | AUC | Predictors/Splits |
| Logistic Regression | | **Model 1** | 0.9674733 | 0.0000000 | 1.000000 | 0.6358085 | 7 predictors |
| **Model 2** | 0.8823836 | 0.3600000 | 0.8999462 | 0.6519742 | 7 predictors |
| Decision Tree | **Model 3** | | 0.9744991 | 0.3120000 | 0.9967725 | 0.7313729 | 22 splits |
| **Model 4** | | 0.826698 | 0.7520000 | 0.829209 | 0.8347574 | 29 splits |

Table 7 – Accuracy measures for the classification models of packaging is one way

**Packaging is Special**

Both the logistic regression method and the decision tree approach will be used to predict our variable packaging\_is\_special. To begin with, model 1 included all the variables related to the product features, however, in model 2 we included only the variables that were significant. In addition, model 3 was built using the same variables from model 1, the only difference is that we trained it using the three-sampling technique and decided to opt with the oversampling method. Similarly, model 4 included the variables of model 2 and considered the sampling technique that uses both the over-sampling and under-sampling method concurrently. We also generated decision trees to predict packaging\_is\_special where model 5 was chosen for the fact that it has the highest cross validation error, however, it turned out to be so complex with 149 splits. As a result, by limiting the number of splits to 30 and making the tree less complicated, model 6 was selected with the lowest cross validation error. Finally, model 7 was built with the same logic of model 6 but this time using the oversampling method.

Table 8 summarizes the accuracy measures as well as the area under the curve scores for each model. To begin with, among the logistic regression models, we can safely say that Model 4 is the best one since it has the highest accuracy and AUC. Also, it has a high specificity ratio despite it coming at third rank and has the second-best sensitivity, which is as well very close to model’s 3 ratio, and it is less complex. For the decision trees models, we can safely omit model 5 as it presents a very high number of splits. Looking at model 6 and 7 for which both offer similar numbers of splits, we can detect that model 6 provides us with a higher accuracy and specificity ratios while model 7 delivers a greater sensitivity ratio and Figure 2 – Packaging is special ROC Curves AUC score. Thus, we will be resorting to model 7 with better sensitivity as we are more interested in predicting the important class. Finally, between the best logistic model (model 3) and best decision tree model (model 7), we will be picking model 7 as it has the highest accuracy, sensitivity, and AUC score.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  | Overall Accuracy | Sensitivity | Specificity | AUC | Predictors/Splits |
| Logistic Regression | | **Model 1** | 0.7254749 | 0.129744 | 0.9745387 | 0.8058288 | 7 predictors |
| **Model 2** | 0.6969983 | 0.05030891 | 0.9833948 | 0.6969983 | 4 predictors |
| **Model 3** | 0.767109 | 0.4642542 | 0.8937269 | 0.7989076 | 7 predictors |
|  | | **Model 4** | 0.7756961 | 0.4561307 | 0.799631 | 0.8093062 | 4 predictors |
| Decision Tree | **Model 5** | | 0.8540203 | 0.7025596 | 0.9173432 | 0.8759955 | 149 splits |
| **Model 6** | | 0.8248764 | 0.6187114 | 0.9110701 | 0.832983 | 27 splits |
| **Model 7** | | 0.7887067 | 0.8014122 | 0.7833948 | 0.8470822 | 28 splits |

Table 8 – Accuracy measures for the classification models of packaging is special

**Load Unit X Dimension, Load Unit Y Dimension, Load Unit Z Dimension**

We also built different regression models to predict the numerical dependent variables load\_unit\_x\_dim, load\_unit\_y\_dim and load\_unit\_z\_dim. Plus, we calculated each model’s RMSE to evaluate the performance. In brief, linear model 1 contained all the numerical and categorical features of the load unit variable. Linear model 2 included only the significant variables. Linear model 3 was built by removing the categorical variable packaging raw material name which constitutes 20 levels. Linear Model 4 was built by removing all categorical variables. Furthermore, we assessed decision tree models: In the below tables (7, 8 & 9), model 5 represents a pruned tree with a certain number of splits, whereas model 6 represents a pruned tree with a lower number of splits to keep the model optimized and simple. However, to get good prediction results for load unit z dimension, a random forest model was deployed.

As shown in table 9 of load\_unit\_x\_dim, all models present a good overall RMSE, but it is obvious that decision trees represent better RMSE compared to the mean value of load\_unit\_x\_dimension. We must consider choosing between model 5 and model 6 since they have the best RMSEs. As a result, we would recommend BMW Logistics Group to deploy the decision tree model 6 which presents a good 2\*RMSE (~208) with 20 splits compared to the mean 1260 of load\_unit\_x\_dim. Likewise, as shown in table 10 of load\_unit\_y\_dim, we would recommend the Logistics group to deploy the decision tree model 6 which presents a good 2\*RMSE (~142) with 20 splits compared to the mean 832 of load\_unit\_y\_dim. Also, in table 11 of load\_unit\_z\_dim, random forest model 7 which presents a good 2\*RMSE (~110) with 150 trees compared to the mean 420 of load\_unit\_z\_dim is the advised model.

The below figure shows a visualization by a library called visreg that offers a convenient way to visualize the resulting fit and possibly gain some insight into the model. Based on the visualization, packaging\_x\_dim and packaging\_weight determine the result of load\_unit\_x\_dim where load\_unit\_x\_dim increases when packaging\_x\_dim increases and decreases when packaging\_weight increases, whereas packaging\_y\_dim and others do not determine the value of load\_unit\_x\_dim since almost a straight line is shown. To study the resulting fit of the remaining models of load\_unit\_y\_dim and load\_unit\_z\_dim, kindly refer to the rest of the figures in the appendix.

Chart, scatter chart

Description automatically generated

Figure 3 – Load unit is one way pruned decision tree

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | |  | RMSE | Predictors/Splits | Mean |
| Linear Regression | | Model 1 | 133.7996 | 26 Predictors | 1260.6 |
| Model 2 | 140.1001 | 7 Predictors |
| Model 3 | 139.4196 | 8 predictors |
|  | | Model 4 | 141.0983 | 6 predictors |
| Decision Tree | Model 5 | | 101.2579 | 32 splits |
| Model 6 | | 104.6099 | 20 splits |

Table 9 – Accuracy measures for the regression models of load unit x dimension

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | |  | RMSE | Predictors/Splits | Mean |
| Linear Regression | | **Model 1** | 87.77346 | 26 Predictors | 832.7194 |
| **Model 2** | 89.72729 | 7 Predictors |
| **Model 3** | 89.11194 | 8 predictors |
|  | | **Model 4** | 89.56966 | 6 predictors |
| Decision Tree | **Model 5** | | 69.5498 | 34 splits |
| **Model 6** | | 71.59569 | 20 splits |

Table 10 – Accuracy measures for the regression models of load unit y dimension

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | |  | RMSE | Predictors/Splits/trees | Mean |
| Linear Regression | | **Model 1** | 197.3013 | 26 Predictors | 420.8882 |
| **Model 2** | 222.1892 | 7 Predictors |
| **Model 3** | 218.6689 | 8 predictors |
|  | | **Model 4** | 218.6324 | 6 predictors |
| Decision Tree | **Model 5** | | 89.45925 | 67 splits |
| **Model 6** | | 92.92101 | 42 splits |
| Random Forest | **Model 7** | | 55.63437 | 150 trees |

Table 11 – Accuracy measures for the regression models of load unit z dimension

**Load Unit Weight**

For load\_unit\_weight, different regression models were assessed, the models’ predictors are similar to the ones in the previous section: 1) All predictors 2) Only significant variables 3) Without packaging\_raw\_material\_name 4) Without categorical variables 5) Pruned decision trees 6) Random Forest. As shown in table 12 of load\_unit\_weight, we would recommend BMW logistics group to deploy the random forest model 7 which presents a good 2\*RMSE (~16) with 155 trees compared to the mean 61 of load\_unit\_weight.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | |  | RMSE | Predictors/Splits/trees | Mean |
| Logistic Regression | | **Model 1** | 16.29229 | 26 Predictors | 61.62669 |
| **Model 2** | 17.02198 | 6 Predictors |
| **Model 3** | 16.94211 | 8 predictors |
|  | | **Model 4** | 18.01122 | 6 predictors |
| Decision Tree | **Model 5** | | 13.18962 | 60 splits |
| **Model 6** | | 14.55843 | 38 splits |
| Random forest | **Model 7** | | 8.381416 | 155 trees |

Table 12 – Accuracy measures for the regression models of load unit weight

**Packaging Per Load Unit**

For packagings\_per\_load\_unit, different regression models were assessed, the models’ predictors are like the ones in the previous section. However, outliers in the data were greatly affecting this variable’s prediction, so it was decided to keep the outliers since they represent an important part of our data and instead transform the numerical variables using logarithm (log).

As shown in the below table, decision tree model 5 presents 2\*RMSE (~0.90) with 40 splits compared to the logarithmic mean 2.1 of packagings per load unit: compared to other models, it is believed that this model works best.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | |  | RMSE | | Predictors/Splits/trees | | Mean | |
| Linear Regression | | **Model 1** | 0.6329259 | | 26 Predictors | | 2.147614 | |
| **Model 2** | 0.6623698 | | 8 Predictors | |
| **Model 3** | 0.7455134 | | 6 predictors | |
| Decision Tree | **Model 4** | | | 0.4558386 | | 52 splits | |  | |
| **Model 5** | | | 0.4594032 | | 40 splits | |

Table 13 – Accuracy measures for the regression models of log(packagings per load unit)

**Load Unit is One Way**

We also built four different logistic models to predict our categorical variable load\_unit\_is\_oneway. Plus, we calculated each model’s accuracy measures and visualized its AUC ROC curve to better evaluate the performance. In brief, model 1 contained all the numerical and categorical features of the packaging variable. Model 2 included only the numerical variables. Model 3 was built by choosing the variables that showed a strong relationship based on the box plots as well as the bar charts, and these included the three dimensions of the packaging (x, y, and z), and the load capacity of the packaging. Further, we assessed a model that includes all the packaging variables using the decision tree method.

Diagram, schematic

Description automatically generatedAs shown in table 14, all models present a very high overall accuracy ratio but, models 1, 2, and 3 provided us with a lower sensitivity ratio than of the decision tree model. For the fact that our aim is to predict the important class of 1’s which is if the packaging is one way, then we would recommend BMW logistics group to deploy the decision tree model which presents a very high accuracy and sensitivity ratios, equal to 0.99 and 0.94 each respectively, with only seven numbers of splits and a high AUC score of 0.991. Figure 4 – Load unit is one way pruned decision tree

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Overall Accuracy | Sensitivity | Specificity | AUC | Predictors/Splits |
| Logistic Regression | **Model 1** | 0.9786625 | 0.04705882 | 0.9997339 | 0.5233964 | 7 Predictors |
| **Model 2** | 0.9838668 | 0.6235941 | 0.9920170 | 0.9163103 | 5 Predictors |
| **Model 3** | 0.9815249 | 0.3411765 | 0.9960085 | 0.7762483 | 4 Predictors |
| Decision Tree | | 0.9914130 | 0.9411765 | 0.9925492 | 0.9910012 | 7 splits |

Table 14 – Accuracy measures for the classification models of load unit is one way

**Load Unit is Special**

Lastly, additional three models were built to predict our second categorical load\_unit\_is\_special variable. Each model’s accuracy measures and AUC rates were computed to compare different models’ performances. Model 1 contained all packaging variables. Model 2 included variables that showed significancy when tested in model 1, while model 3 included the significant variables of the latter. Additionally, a decision tree model was also tested to further compare different models’ performances.

From the below table, model 2 and model 3 show approximately same ratios and present the lowest sensitivities and thus, they can be eliminated for the fact that we are primally interested in predicting our important class of load unit is special. Consequently, we would recommend employing the pruned decision tree model which presents a sensitivity ratio of 0.986 and an accuracy equal to 0.996. Further, to reduce slightly the complexity, BMW could also consider model 1 which results in high accuracy and sensitivity ratios while considering 7 predictors which is less than of the decision tree model’s splits.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Overall Accuracy | Sensitivity | Specificity | AUC | Predictors/Splits |
| Logistic Regression | **Model 1** | 0.9942753 | 0.9381107 | 0.9991516 | 0.9977974 | 7 Predictors |
| **Model 2** | 0.9799636 | 0.7491857 | 1.0000000 | 0.9458101 | 4 Predictors |
| **Model 3** | 0.9794431 | 0.7426710 | 1.0000000 | 0.9488679 | 2 Predictors |
| Decision Tree | | 0.9960968 | 0.9869707 | 0.9968891 | 0.9967176 | 13 splits |

Table 15 – Accuracy measures for the classification models of load unit is special

**5. Conclusion and Recommendation**

The combination of data analysis and data modelling for prediction can greatly impact the performance of the supply chain department and logistical activities of a firm. Rather than having employees perform the tasks of fitting each product into packages and each packaging into a load unit, BMW Group Logistics Team will now be able to automate this process by applying linear and logistic regression as well as utilizing decision trees and the random forest approach. As a result, we would recommend BMW to implement the decision tree method for all the categorical variables such as packaging is special, packaging is one way, load unit is special, and load unit is one way, as well as for some numerical variables including packaging y-dimension, packaging per load unit, and load unit y-dimension since it lowered its RMSE significantly. In addition, they may use the random forest approach for the following variables: packaging x-dimension, load unit z-dimension, and load unit weight. Finally, for the last group of variables such as the packaging z-dimension, packaging weight, product per packaging, and load unit x-dimension, none of the trained models has shown an accurate prediction outcome and thus, we would highly encourage BMW to dig deeper into other methods of machine learning approaches and find the optimal algorithm to automate this process and decrease its operating expenses.

Another approach to implement for further greater prediction accuracy is to deal with the issue of outliers by working closely with logistics experts to better dictate which ones are important and which ones are not. Handling outliers is a critical process, and as Leys, C. et al. (2019) explain that there are three methods to manage the outliers, and they are as follows: keeping the outliers, removing the outliers, and recoding the outliers. The latter could be implemented by deploying the Winsorization approach as described by Leys, C. et al. (2019) where the outliers could be converted to a value at a certain percentile of the data.

# **References**

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random Forests and Decision Trees. *International Journal for Computer Science Issues, 9*(5).

Knoll, D., Neumeier, D., Prüglmeier, M., & Reinhart, G. (2019). An automated packaging planning approach using machine learning. *Procedia CIRP, 81*, 576-581.

Knoll, D., Prüglmeier, M., & Reinhart, G. (2016). Predicting Future Inbound Logistics Processes using Machine Learning. *Procedia CIRP, 52*, 145-150.

Leys, C., Declare, M., Mora, Y. L., Lakens, D., & Ley, C. (2019). How to Classify, Detect, and Manage Univariate and Multivariate Outliers, With Emphasis on Pre-Registration. *International Review of Social Pschychology, 32*(1).

**Appendices**

**Appendix A – Detailed table of data**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Column Name | Type | Description |
| Product Features | product\_generic\_family\_name | STRING | Product family name |
| product\_module\_number | STRING | Unique identifier for the Module to which the material belongs |
| product\_kogr\_number | STRING | Unique identifier for the Construction Group |
| product\_weight | DOUBLE | Weight of the product |
| product\_number | STRING | Product unique identifier as identified by BMW Group Logistics |
| product\_name | STRING | Product name, as identified by BMW Group Logistics |
| product\_is\_esp | BOOL | 1 if product requires electrostatic discharge protection, 0 if not |
| product\_is\_dangerous\_good | BOOL | 1 If product is dangerous good, 0 if not |
| product\_x\_dim | DOUBLE | Product height |
| product\_y\_dim | DOUBLE | Product width |
| product\_z\_dim | DOUBLE | Product depth |
| product\_quality\_index | DOUBLE | Quality score as identified by BMW Group Logistic |
| product\_supplier\_number | STRING | Unique identifier for product supplier |
| Packaging Features | packaging\_x\_dim | DOUBLE | Package height |
| packaging\_y\_dim | DOUBLE | Package width |
| packaging\_z\_dim | DOUBLE | Package depth |
| packaging\_weight | DOUBLE | Weight of the empty package |
| products\_per\_packaging | INT | Number of parts per package |
| packaging\_is\_special | BOOL | 1 if package is special, 0 if not |
| packaging\_is\_oneway | BOOL | 1 if package is designed for only one-time usage, 0 if not |
| packaging\_raw\_material\_name | STRING | Packaging material |
| packaging\_load\_capacity | DOUBLE | Maximum weight a package can carry |
| Load unit Features | load\_unit\_x\_dim | DOUBLE | Load unit height |
| load\_unit\_y\_dim | DOUBLE | Load unit width |
| load\_unit\_z\_dim | DOUBLE | Load unit depth |
| load\_unit\_weight | DOUBLE | Weight of the empty load unit |
| packagings\_per\_load\_unit | INT | Number of packages per load unit |
| load\_unit\_is\_special | BOOL | 1 if load unit is special, 0 if not |
| load\_unit\_is\_oneway | BOOL | 1 if load unit is designed for only one-time usage, 0 if not |

**Appendix B – Data Visualization**

1. **Bivariate**

**Chart, scatter chart

Description automatically generated**

**Chart, histogram, scatter chart

Description automatically generated**

**Chart

Description automatically generated**

**Chart

Description automatically generated**

**Chart, histogram

Description automatically generated**

**Chart, histogram

Description automatically generated**

**Chart, histogram, scatter chart

Description automatically generated**

**Chart, histogram

Description automatically generated**

**Chart, histogram, scatter chart

Description automatically generated**

**Chart, histogram

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Chart, histogram

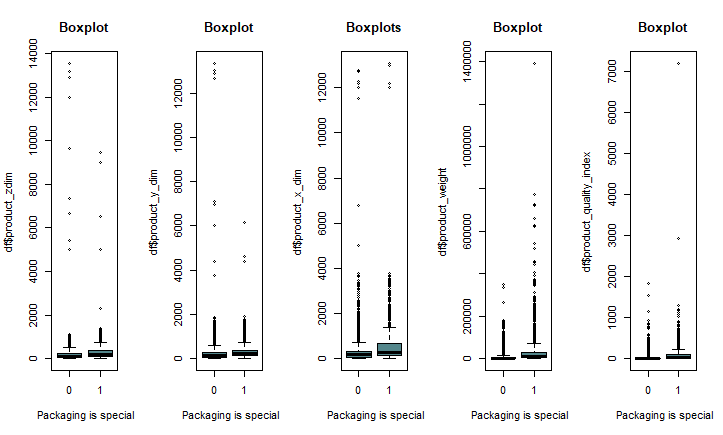
Description automatically generated**

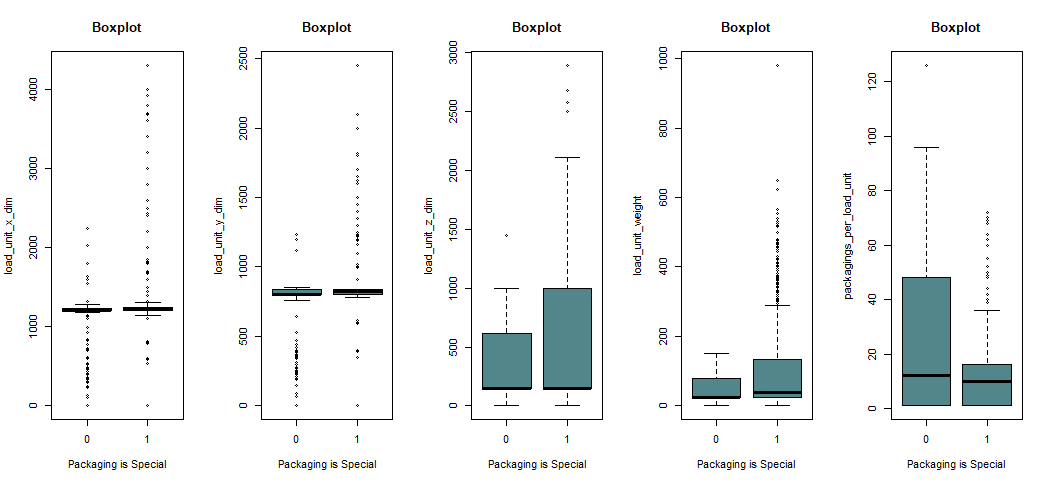
**Chart, histogram, box and whisker chart

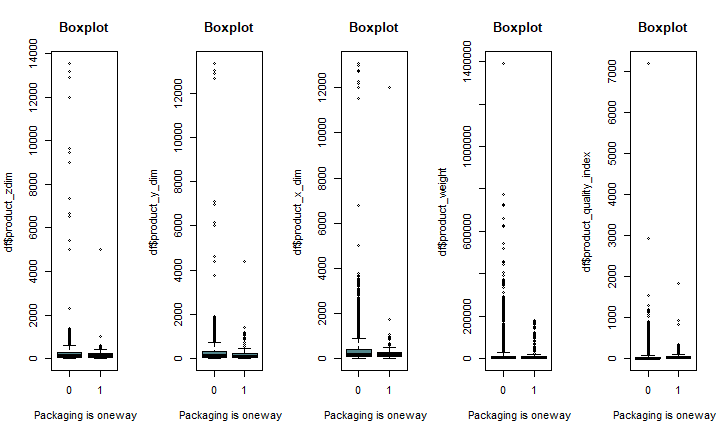
Description automatically generated**

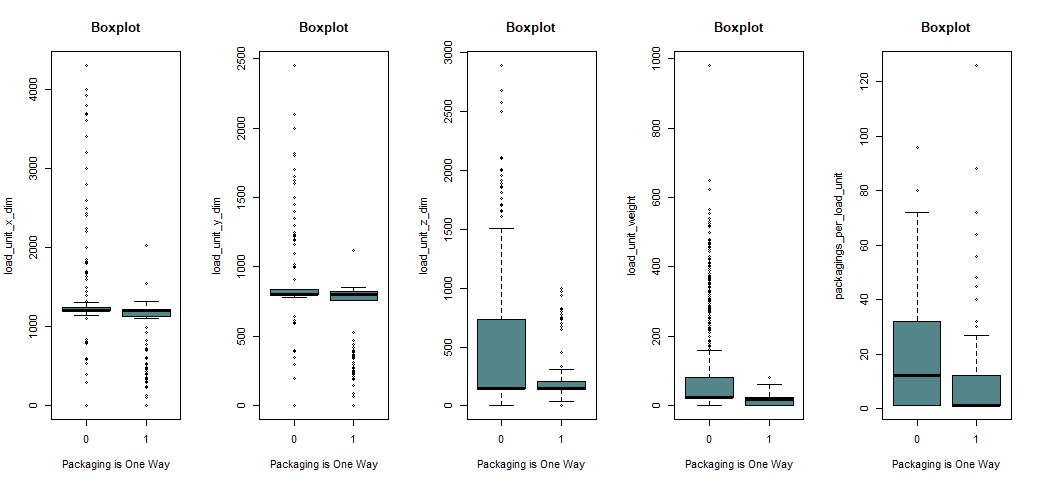
**Chart, box and whisker chart

Description automatically generated**







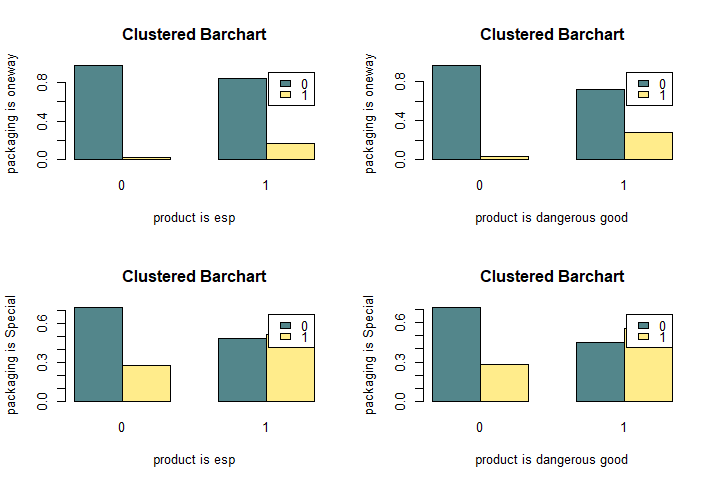


Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

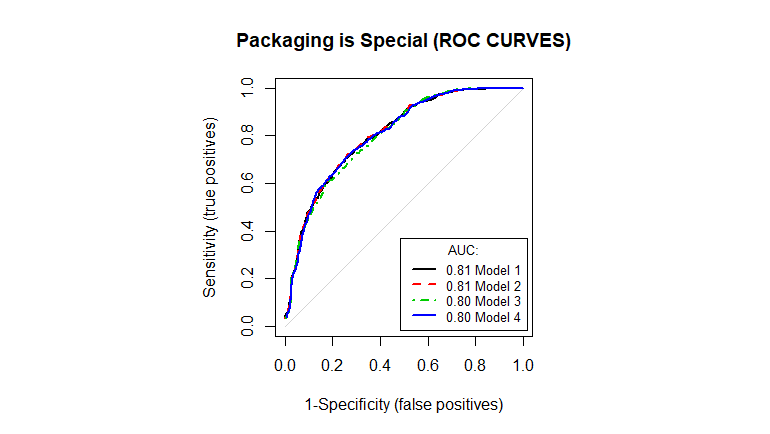
Description automatically generated

Chart, waterfall chart

Description automatically generated

**Appendix C – ROC Curves**

1. **Logistic Regression Models**

**Chart, line chart

Description automatically generated**

Chart

Description automatically generatedChart, line chart, histogram

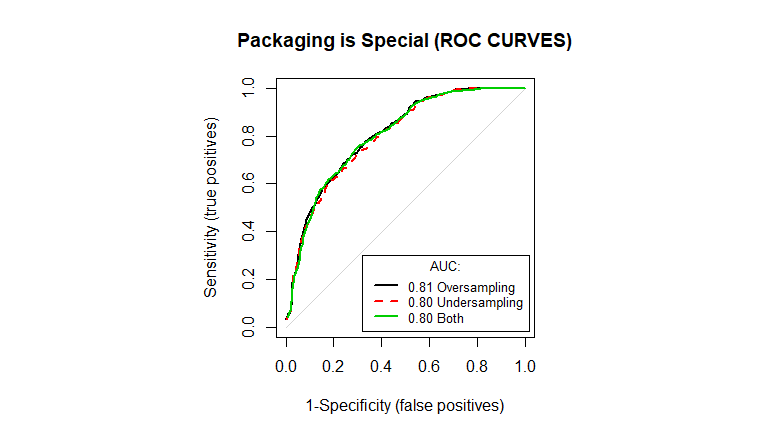
Description automatically generated

1. **Decision Tree Models**

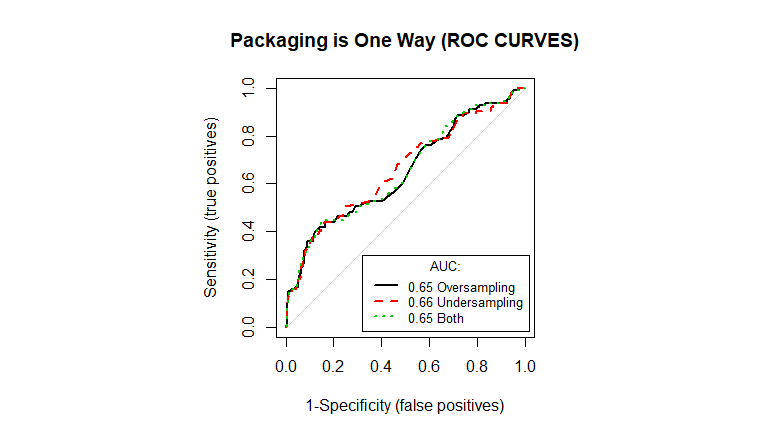
**Chart, histogram

Description automatically generatedChart, scatter chart

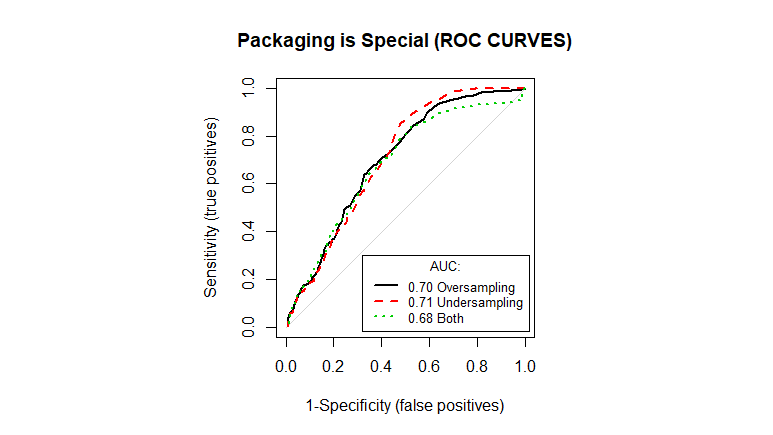
Description automatically generated**

1. **Different Sampling Techniques**
   1. **Logistic Regression Model**
      1. **Using 7 Variables**

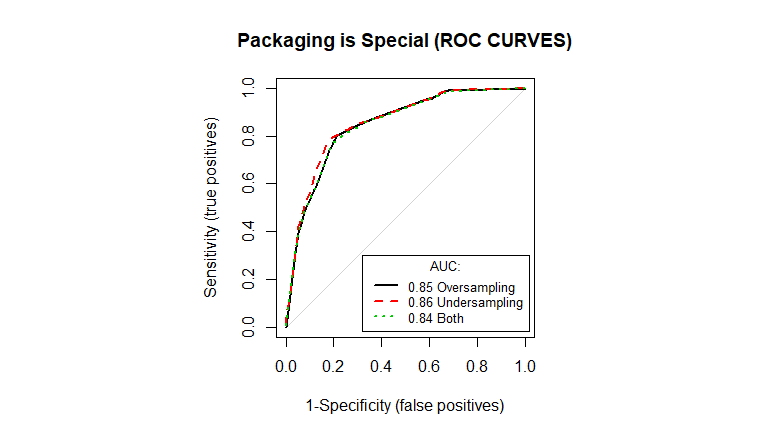
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Model* | *Accuracy* | *Sensitivity* | *Specificity* | *AUC* |
| Over Sampling | 0.775176 | 0.458959 | 0.90738 | 0.806833 |
| Under Sampling | 0.767109 | 0.462542 | 0.893727 | 0.798908 |
| Both | 0.767109 | 0.456311 | 0.897048 | 0.803956 |



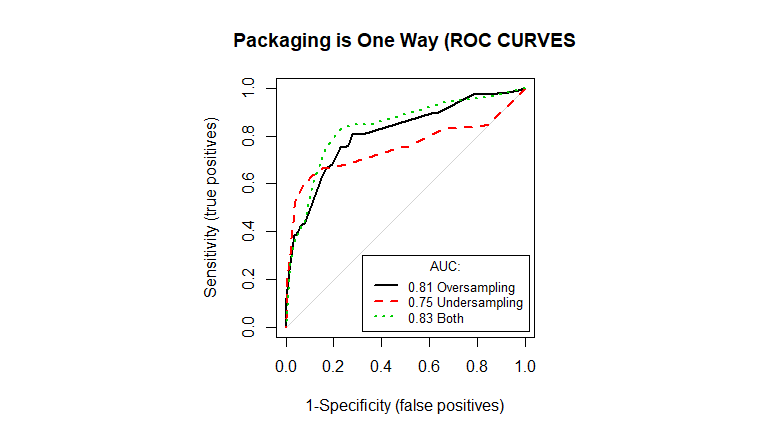
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Model* | *Accuracy* | *Sensitivity* | *Specificity* | *AUC* |
| Over Sampling | 0.891751 | 0.336 | 0.910436 | 0.6487 |
| Under Sampling | 0.880301 | 0.344 | 0.898332 | 0.6604 |
| Both | 0.882383 | 0.36 | 0.899946 | 0.6519 |

* + 1. **Using 4 Variables**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Model* | *Accuracy* | *Sensitivity* | *Specificity* | *AUC* |
| Over Sampling | 0.671871 | 0.380406 | 0.793727 | 0.7013 |
| Under Sampling | 0.675774 | 0.383937 | 0.797786 | 0.7083 |
| Both | 0.686127 | 0.414828 | 0.799631 | 0.6829 |

* 1. **Decision Tree Model**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Model* | *Accuracy* | *Sensitivity* | *Specificity* | *AUC* |
| Over Sampling | 0.788707 | 0.801422 | 0.783395 | 0.8470 |
| Under Sampling | 0.803799 | 0.790821 | 0.809225 | 0.8555 |
| Both | 0.794431 | 0.770521 | 0.804428 | 0.8439 |



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Model* | *Accuracy* | *Sensitivity* | *Specificity* | *AUC* |
| Over Sampling | 0.822566 | 0.664 | 0.831092 | 0.8470 |
| Under Sampling | 0.886027 | 0.632 | 0.89457 | 0.7512 |
| Both | 0.826698 | 0.752 | 0.829209 | 0.8292 |

**Appendix D – Model Visualization**

1. **Decision Tree**
   1. **Load Unit X Dimension**

**Chart, scatter chart

Description automatically generatedChart, box and whisker chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

* 1. **Load Unit Y-Dimension**

**Chart, scatter chart

Description automatically generatedChart, box and whisker chart

Description automatically generated**

**Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated**

* 1. **Packaging Per Load Unit**

**Chart, diagram, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated**

**Chart

Description automatically generatedChart, diagram, box and whisker chart

Description automatically generated**

1. **Random Forest**
   1. **Load Unit Z-Dimension**

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

* 1. **Load Unit Weight**

**Chart

Description automatically generated**

**Chart

Description automatically generated**

**Appendix E – Decision Trees**

Diagram, schematic

Description automatically generated