Portfolio Project: Predicting Robotic Failure

MIS530-1

Cesy Zamarripa

**Introduction**

For this course, the portfolio project consists of performing predictive analytics on a chosen dataset. Utilizing the UCI Machine Learning repository, a dataset that consisted of data collected on robotics and automation was chosen for investigation. The business problem chosen was: can robotic failure be predicted by the robots’ electrical currents, temperatures, speed across joints and other monitoring information on the machines provided by the dataset? The target variable in this investigation would be the protective stops variable.

**The Dataset and it’s Variables**

The dataset chosen is a multivariate time-series dataset. It offers 23 variables that are all operational parameters such as temperatures, electrical currents, and protective stops from a UR3 cobot (a cop robot prototype). This dataset is perfectly primed for classification, clustering or even regressive analysis. With the business question at hand, both regression modeling and decision-tree modeling will be performed and compared using techniques learned over the course. As stated, the target variable will be the protective stops variable, and the time stamp variable will be rejected for the scope of this investigation. All other variables will be considered as inputs.

**Alternate and Null Hypothesis**

With the business question regarding predictive failure based on the variables in the dataset, the following hypothesis were formed.

H0: the electrical currents, temperatures and joint speeds are not statistically relevant for predicting protective stops.

HA: the electrical currents, temperatures and joint speeds are statistically relevant for predicting protective stops. This would mean that a model is rendered offering a prediction stronger than random guessing.

**Descriptive Statistics**

To initially look at the data, some descriptive analysis was performed. This allows us to gain a better understanding if the data has any entries that can skew any modeling analysis that will be performed. Searching for any extreme outliers was done by creating a histogram of the variables. Another important factor that can skew the results is missing values. The StatExplore node was utilized in Enterprise Miner to explore those missing values. Some of the output results are shown in figure 1.

**Figure 1**

*Missing Values*

A screenshot of a computer

Description automatically generated

Note: Date and time shown.

From the figure one can see that all the variables have the same number of missing values, which is 54 but you can also see that there are 7355 non-missing variables which is a great sample size. StatExplore was added to the diagram process to obtain a histogram. The diagram flow is shown in figure 2.

**Figure 2**

*Diagram Process*

A white background with black dots

Description automatically generated

Note: Graph Explore run completion confirmation shown.

After adding the StatExplore node, the input variables were all selected to be included in the histogram. As shown in figure 3.

**Figure 3**

*Histogram of all the Variables*

A screenshot of a computer

Description automatically generated

Note: Date and time shown.

From the histograms of all the input variables, it appears that variable 23 could be considered an outlier. It also appears that most all other variables are normally distributed with some variables having spikes for certain values, but not far out of the range of relative distributions. One can also see that variable 23 is the only binary variable among all the others.

From the Missing values figure, one can see other stats shared about each variable, that cross references with the histogram such as the range or the kurtosis values of the distribution relative to a normal distribution.

**Next Steps**

Now that missing values and outliers have been identified, the next step in the portfolio project will be cleaning the data to address those found discrepancies that might impose or skew the results of the model trying to be achieved to model robotic failure.

**Cleaning and Preparing the Data**

When cleaning and preparing a dataset, taking into consideration that there may be missing values that affect the model is essential. After initially performing some descriptive analysis, this dataset does indeed have missing values. This was addressed with imputing the dataset in SAS Enterprise miner using Impute node. Imputing allows for the missing data to be replaced with substituted values. The variable suggested to be an outlier, was just a binary variable, thus it was not filtered out of the dataset.

**Regression Model**

The first method of modeling utilized was a regression model. This technique was chosen because a regression model allows you to see which variables are most relevant to predicting the target variable. This was done using the regression node in Enterprise Miner. The results are shown in Figures 4-6.

**Figure 4**

*Regression Model Lift Results*

A screen shot of a graph

Description automatically generated

Note: Date and time shown.

From the lift chart, one can observe that the lift maintains above 2 for 40% of the data. This means that for 40% of the data the model is 2 times better at predicting protective stops than random guessing.

**Figure 5**

*Effect Number Results*

A screenshot of a computer

Description automatically generated

Note: Date and time shown.

The effects chart shows how relevant the coefficient of each variable is to the model. The blue signifies a positive relation and the red in a negative correlation.

**Figure 6**

*Fit Statistics of the Regression Model*

A screenshot of a computer

Description automatically generated

Note: Date and time shown.

From the fit statistics, the measure of interest to compare to other models is the Average Square Error (ASE). For the regression the ASE for the validation data is 0.035393 which isn’t extremely deviated from the training ASE. If that was the case, it would suggest the regression was overfitting.

**Decision Trees Modeling**

The second method chosen for modeling the data was through the use of decision trees. This was to be compared to the regression model. The results are shown in the following figures.

**Figure 7**

*Cumulative Lift*

A screen shot of a computer

Description automatically generated

Note: Date and time shown.

From this lift chart one can see that the lift is similar to the regression lift chart in that it maintains above 2 for approximately 40% of the data. One thing to note about this lift chart compared to that of the regression is that the lift is much higher for 10% of the data in the case of the decision tree model.

**Figure 8**

*Decision Tree*

A computer screen shot of a diagram

Description automatically generated

Note: Date and time shown.

The results of the tree display several things. The darker the node, the more relevant as well as the darker connecting line. Here variable 8 is the temperature of one of the joints, and variable 15 is the speed of a different joint, variable 5 is the current of the same joint as variable 15.

**Figure 9**

*Fit Statistics Results*

A screenshot of a computer

Description automatically generated

Note: Date and time shown.

From the fit statistics one can see the ASE for the model is 0.037364, which suggests the regression model is slightly better. This can be improved further with several options. Looking at the leaf statistics, one can choose if certain leaves need to be pruned from the model. This could be a possible next step in further analysis. Figure 10 shows the leaf statistic results.

**Figure 10**

*Leaf Statistic Results*

A screenshot of a computer

Description automatically generated

Note: Date and time shown.

From the leaf statistics, one could make a suggestion to prune the last 2 leaves. This would mean pruning one of the child nodes in the decision-making process to get a simpler tree.

**Further Model Comparison**

To better compare the models, a model comparison node was added to the flow diagram and shown in Figure 11. This allowed for a better side-by-side comparison of the two models.

**Figure 11**

*Final Diagram Flow*

A screenshot of a computer

Description automatically generated

Note: Date and time shown

**Figure 12**

*Lift Chart Comparison*

A screenshot of a computer

Description automatically generated

Note: Date and time shown.

From figure 12, one can see how the regression model in red compared to the decision tree model in blue. The Decision tree model has a higher lift under 20% but for up to 40% of the date, the regression model holds slightly higher.

**Figure 13**

*Fit Statistic Comparison Data*

A screenshot of a computer

Description automatically generated

Note: Date and time shown.

From the fit statistics, the average square error in the model comparison shows slightly lower for the regression model.

**Answering the Business Problem**

The business problem posed was, can the protective stop variable be predicted by the measures recorded for the temperatures, speed, and currents of the robot. After viewing the results of both models, and analysis of the lift charts, the Null Hypothesis can be rejected. Both models prove to show that for 40% of the data, the model is 2 times a better predictor than random guessing, which suggests a statistical significance in predicting robotic failure.

**Lessons Learned**

Throughout the entire course and the modules, learning what different modeling techniques could be applied was very insightful. Though only two types of modeling were presented in the project, there are a multitude of other methods that could have been performed and compared to reach the best predicting model. The use of neural networks could have been applied or even a boosting gradient technique or additionally performing other regression techniques such as partial least squares or Least Angle Regressions. Next steps in this project could include those forms of analysis as well as using a node that uses all the models superpositions to obtain the best-fitting model. This would allow us to most-accurately state how well the input variables are at predicting the target variable.

References

Degan Kettles. (2016, October 31). *Regressions with SAS Enterprise Miner* [Video]. YouTube. https://www.youtube.com/watch?v=wQ24e0LrmPg

deVille, B., Neville, P., & SAS Institute Inc. (2013). *Decision trees for analytics using SAS® Enterprise MinerTM*. SAS Institute Inc. https://analyticsconsultores.com.mx/wp-content/uploads/2019/03/Decision-Trees-for-Analytics-Using-SAS-Enterprise-Miner-B.-de-Ville-P.-Neville-SAS-2013.pdf

Dothang Truong. (2022, December 27). *SAS Enterprise Miner - Regression Modeling* [Video]. YouTube. https://www.youtube.com/watch?v=8vbGm6iuUoY

McCarthy, R. V., McCarthy, M. M., Ceccucci, W., & Halawi, L. (2019). *Applying predictive analytics: Finding Value in Data*. Springer.

Prashant Bhowmik. (2016, April 8). *SAS Enterprise Miner- Decision Trees* [Video]. YouTube. https://www.youtube.com/watch?v=pFBR6iXtOsc

*SAS Help Center*. (n.d.-a). https://documentation.sas.com/doc/en/emref/14.3/n1jqzz8cssr9m2n1ktx2iyv87q56.htm

*SAS Help Center*. (n.d.-b). https://documentation.sas.com/doc/en/emref/14.3/n1gvjknzxid2a2n12f7co56t1vsu.htm

Tyrovolas, M., Stylios, C., Aliev, K., & Antonelli, D. (2024). Leveraging Information Flow-Based fuzzy cognitive maps for interpretable fault diagnosis in industrial robotics. In *IFIP advances in information and communication technology* (pp. 98–110). https://doi.org/10.1007/978-3-031-63851-0\_6

*UCI Machine Learning Repository*. (n.d.). https://archive.ics.uci.edu/dataset/963/ur3+cobotops