# Executive Summary

Machine failures are never a desired outcome when a customer is operating their piece of equipment, unfortunately it is something that is rarely able to be completely avoided. In this instance the machine failure is due to the failure of a specific part on the engine which will be referred to as the Widget. The Widget can be monitored using MEASUREY because when the Widget fails MEASUREY shifts downward. While a failure is never the desired outcome, there are large benefits for the customer if a failure can be predicted prior to the actual failure. Monitoring MEASUREY for a failure is valuable but it only notifies of a failure after the Widget has failed. Telematics data was utilized to predict the probability of failure prior to the actual failure of the Widget.

Telematics data for machines that have had a failure of the Widget and data for machines that have not had a failure were collected. A model was generated using this data that predicts the probability of failure for a given machine. The final model was able to correctly predict 60% of the failed machines but incorrectly predicted 26% of the non-failed machines. This means that 26% of the machines that did not have a failure were predicted to have a failure.

When customers are involved it is extremely important to not interrupt their operation unless there is an actual failure with their machine. In the case of the failed Widget there is no safety concern and the failure is not noticeable from the customer’s perspective. While the final model had some success in predicting failed machines, there were too many machines predicted as failed when they were actually not failed to justify using the model. It is suggested to simply monitor MEASUREY as opposed to relying solely on the model. The model can be used in tangent to monitoring MEASUREY to help identify high potential machines but should not be used to proactively repair machines.

Logistic Regression: Widget Failure

Christopher Spaude

March 20, 2018

# I. Introduction

John Deere is a company that “has delivered product and services to support those linked to the land”. John Deere is an original equipment manufacturer (OEM) that manufactures agriculture, construction, forestry, and turf care products. One division of John Deere is John Deere Power Systems (JDPS), whose main focus is producing diesel engines for John Deere products and other OEM businesses.

John Deere, with the consent of the customer, remotely collects machine data from customer vehicles on most of its large vehicle platforms. This typically includes large tractors, combines, excavators, crawlers, etc. This data is used to support customers and to help the engineering departments create better products. This remotely collected machine data is generically referred to as telematics data because the data is sent via cellular towers. This telematics data typically contains measurement data along with other vehicle information. Measurement data refers to data collected by an electronic control unit (ECU). This could be anything from what speed the engine is running to the ambient temperature. Due to privacy concerns all measurement data will be generically referred to as MEASURE1, MEASURE2,…,MEASUREn (where n is the number of measurements used).

One main benefit to the data collection system John Deere uses is the ability to look over a machine’s history. This allows engineers to look at data that led up to a specific event. This study is investigating the failure of a specific part on the engine that will be generically called Widget. The failure of the Widget can be seen in a telematics measurement that is collected, this specific measurement will be called MEASUREY. When the Widget fails there is a distinct shift in MEASUREY. In Figure 1, MEASUREY is plotted for the history of a single machine. Two vertical lines have been added to show when the Widget fails and when the Widget is replaced.

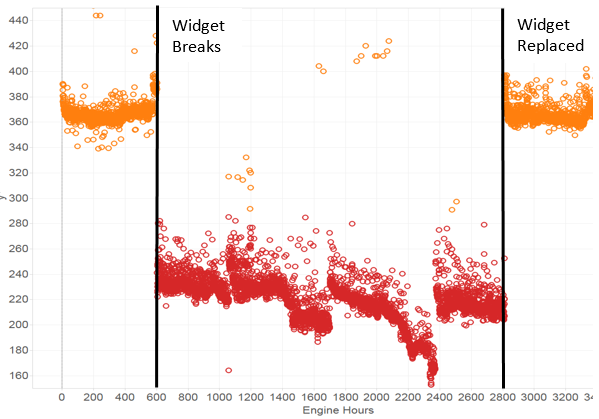


Figure 1: MEASUREY plotted versus engine hours

As seen in Figure 1, the shift in MEASUREY is large and downward when the Widget fails. When the Widget is replaced MEASUREY shifts equally as large upward. While monitoring machines for a shift in MEASUREY is important, it would be more valuable to detect the Widget failing before the failure occurs.

The goal of this study is to utilize the telematics data to estimate the probability the Widget is going to fail on a machine in the future. This will be done by collecting two populations: machines that had the Widget fail (this population will be referred to as the failed population) and machines that do not have a failed Widget (this population will be referred to as the non-failed population). The telematics data for each machine in these two populations will be modeled using logistic regression. The model will be assessed on how well it was able to categorize each machine into the failed and non-failed populations.

## Assumptions and limitations:

* The failure of the Widget is something that progresses over time. The failure of the Widget appears instantaneous when looking at MEASUREY but to be able to predict a future failure of the Widget it must fail due to damage over time. Sudden, random failures cannot be predicted prior to failure.
* The right measurements for predicting the failure of the Widget have been recorded. As mentioned previously, one benefit of the measurement collection system for John Deere is that historical data is stored, but one problem with the system is that the list of measurements being recorded for each machine is predefined when the machine is first manufactured. Not all measurements can be recorded for the entire life of a machine so the total list of possible measurements is reduced to a high importance list. This means that the right measurements may not have been recorded to predict the failure of the Widget.
* The non-failed population of machines will not fail the Widget in the near future. There is no way to determine if the non-failed machines will soon fail the Widget in the future so it is assumed that if the Widget has not failed, at the time the data is collected, that it will not fail in the future.
* The failed population all have similar failure modes. A part can fail many ways and for modeling to be successful it will be assumed that all the failures have the same root cause.

## Terminology

Below is a list of terms referred to throughout this report:

* Failed Machines: this is the population of machines that have failed the Widget.
* Non-Failed Machines: this is the population of machines that have not failed the Widget.
* Widget: this is a part on the engine that will be investigated.
* MEASUREY: this is the measurement that indicates the Widget has failed.
* Operating Time: this is the amount of time that a machine has been used. This is similar to the amount of miles driven for an on-road vehicle; off-road vehicles do not refer to amount of miles driven but instead the amount of hours the vehicle has been operated.

# II. Preparing and Exploring Data

As mentioned previously, JDPS provides engines to multiple different vehicle applications. Vehicle applications can have an influence on engine measurements even if the engine is the same. In this study the data set will be limited to a single type of vehicle. This will reduce variability in the measurements that is due to application differences.

## Populations: Failed vs non-Failed

Two different data sets need to be collected for this study, data on failed machines and non-failed machines. The list of failed machines was given by a group within John Deere focused on part failures. This list of machines was cross-validated using MEASUREY to confirm that a failure of the Widget had occurred. The list for non-failed machines was randomly selected from a database that stores all the telematics data. The list for non-failed machines was cross referenced with the list of failed machines to ensure a machine did appear in both categories.

For failed machines, only data prior to the failure was included. Since the goal is to predict the failure of the Widget prior to the actual failure, the data after the failure is irrelevant. For the non-failed machines all the data for the machine is included.

## Measurement List

In order to determine the measurement list that will be used for modeling, an interview was conducted with the lead engineer investigating this failure. During the interview, the lead engineer said that the failure mode is still not clearly defined for this issue. They have had some success repeating the failure but the root cause is still not clear.

The most success with repeating the failure has been to manipulate MEASUREX. MEASUREX is a histogram that records the amount of operating time each machine spends within each bin of the histogram. By manipulating the amount of time a machine spends in each bin, the engineers have had some success in repeating the failure of the Widget. The hypothesis of the lead engineer is that if a machine accumulates enough time in a particular bin of MEASUREX then the Widget will fail. To utilize MEASUREX in the logistic regression model each bin will be a separate measurement. The lead engineer also gave three categorical variable to include in the study as possible important factors. These will be referred to as CATEGORY1, CATEGORY2 and CATEGORY3.

## Exploring Data

Below is a summary of the raw data, STATUS is the indicator if the machine is in the failed population or the non-failed population:

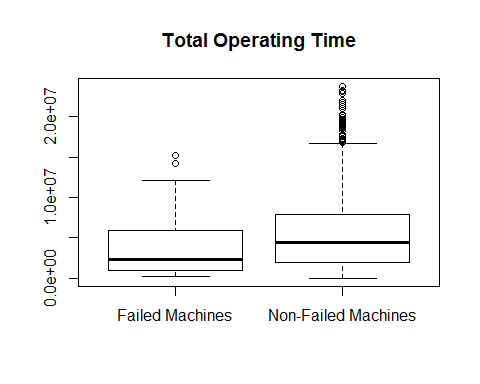
## STATUS CATEGORY1 CATEGORY2 CATEGORY3 MEASURE1   
## failed : 31 PE :1002 no :1030 no:1031 Min. : 7.0   
## non-failed:1000 NA's: 29 yes: 1 1st Qu.: 590.5   
## Median : 1126.0   
## Mean : 1409.0   
## 3rd Qu.: 1930.0   
## Max. :13753.0   
## MEASURE2 MEASURE3 MEASURE4 MEASURE5   
## Min. : 487 Min. : 4 Min. : 2 Min. : 372   
## 1st Qu.: 507324 1st Qu.: 1962 1st Qu.: 2814 1st Qu.: 28879   
## Median : 1155417 Median : 4517 Median : 5702 Median : 70664   
## Mean : 1587341 Mean : 9581 Mean : 10815 Mean : 106675   
## 3rd Qu.: 2264490 3rd Qu.: 8543 3rd Qu.: 10686 3rd Qu.: 141822   
## Max. :10392663 Max. :2400912 Max. :625003 Max. :1442866   
## MEASURE6 MEASURE7 MEASURE8 MEASURE9   
## Min. : 2 Min. : 2 Min. : 12 Min. : 16   
## 1st Qu.: 12443 1st Qu.: 16016 1st Qu.: 23455 1st Qu.: 36438   
## Median : 28890 Median : 39708 Median : 61195 Median : 99129   
## Mean : 43563 Mean : 63088 Mean : 104851 Mean : 163987   
## 3rd Qu.: 58718 3rd Qu.: 82852 3rd Qu.: 134339 3rd Qu.: 207109   
## Max. :399094 Max. :625050 Max. :1278998 Max. :1900046

## MEASURE10 MEASURE11 MEASURE12 MEASURE13   
## Min. : 23 Min. : 4 Min. : 6 Min. : 19   
## 1st Qu.: 112136 1st Qu.: 98735 1st Qu.: 121177 1st Qu.: 143085   
## Median : 330693 Median : 245111 Median : 299123 Median : 318836   
## Mean : 499414 Mean : 378279 Mean : 472174 Mean : 500372   
## 3rd Qu.: 665880 3rd Qu.: 491474 3rd Qu.: 643532 3rd Qu.: 682409   
## Max. :5592424 Max. :3319991 Max. :3871380 Max. :3515160   
## MEASURE14 MEASURE15 MEASURE16 MEASURE17   
## Min. : 6 Min. : 108 Min. : 1 Min. : 0.0   
## 1st Qu.: 135218 1st Qu.: 272866 1st Qu.: 97083 1st Qu.: 6.0   
## Median : 313871 Median : 620026 Median : 257650 Median : 48.0   
## Mean : 459823 Mean : 825223 Mean : 373306 Mean : 439.9   
## 3rd Qu.: 648202 3rd Qu.:1155878 3rd Qu.: 522404 3rd Qu.: 381.5   
## Max. :3280515 Max. :5900260 Max. :3514939 Max. :17088.0

After looking at the summary there are some parameters that can be removed because they will not be important to the model.

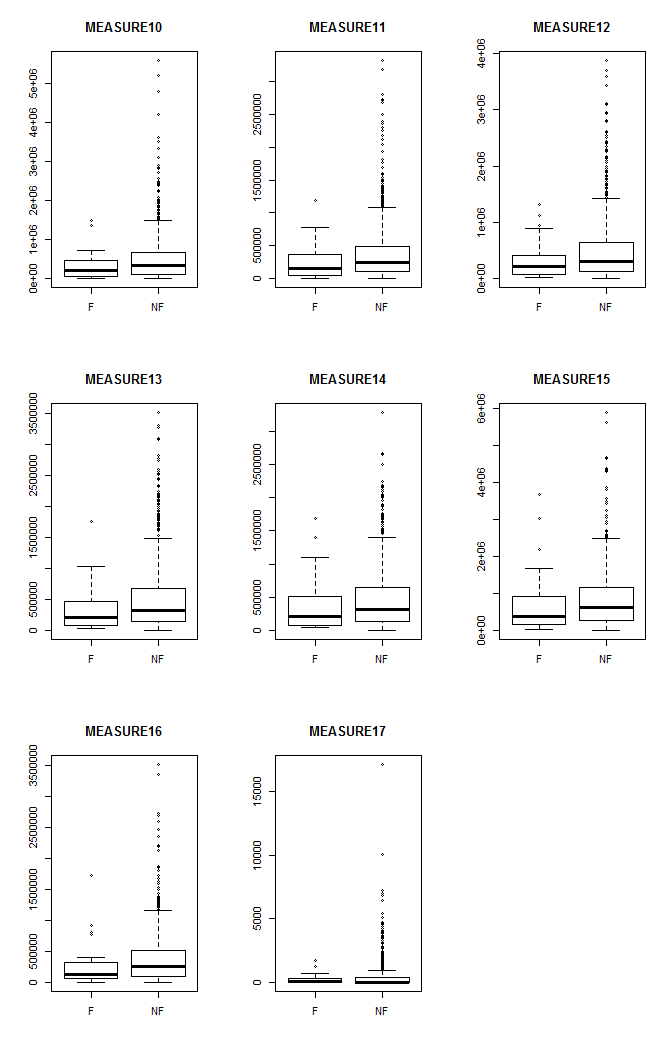
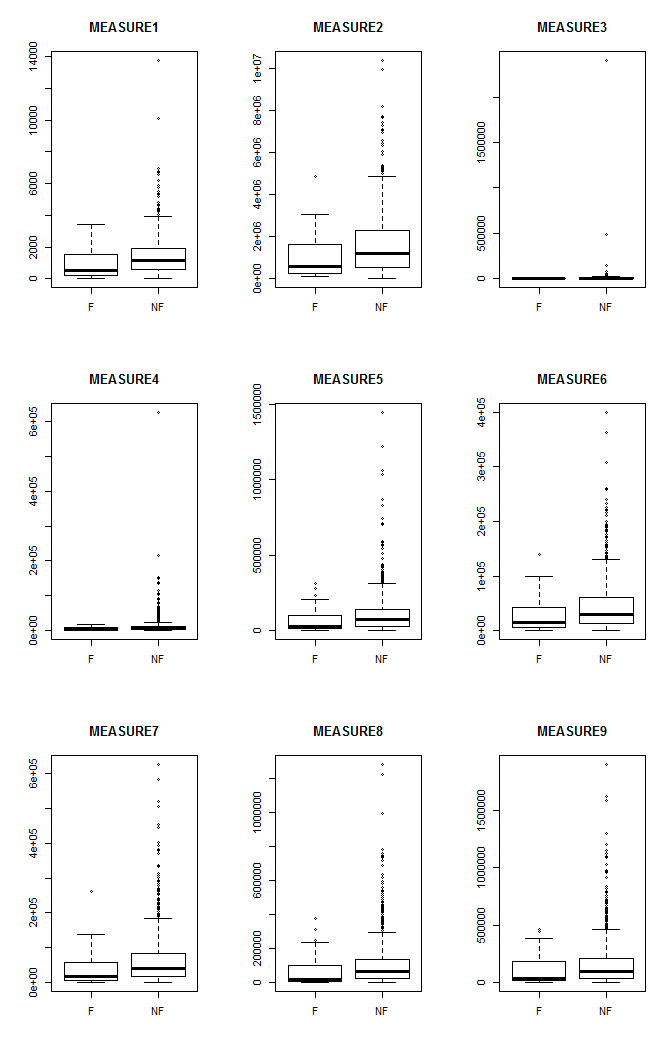
* CATEGORY1 can be removed. There is only one level, PE, and all other values are NA.
* CATEGORY2 can be removed. There are two levels for this parameter but there is only one vehicle that has ‘yes’ for CATEGORY2. Since there are 31 failed machines and 1000 non-failed machines it is assumed that this parameter will not be important.
* CATEGORY3 can be removed, there is only one level.

There is a very large distribution for each of the measurements (1-17), this is to be expected. As previously discussed, each measurement is a sum of time the machine operated in a particular bin of MEASUREX. If a machine has more total operating time it is reasonable to assume that there will be more time in each bin. Below is a box plot of total time in all bins for the failed and non-failed machines.



There is a very large number of outliers for the non-failed machines; these outliers are machines with a large amount of operating time. These machines will not be initially removed as this information may be valuable because these machines have operated for a long time without failing the Widget.

For a parameter to provide value it needs to be able to partial explain the difference between the failed and non-failed machines. In this specific study, parameters will be important if the failed population is larger than the non-failed population for that parameter. This would confirm the engineer’s hypothesis that the Widget failure is due to more time being spent in a particular bin of MEASUREX. To assess whether the failed machines have more time in a bin than the non-failed machines the medians will be compared. Below are boxplots for each measurement separated by failed and non-failed machines. Failed and non-failed machines will be labeled as F and NF respectively.



Looking through the boxplots is difficult due to the outliers and the heavy right skew. Based on the boxplots that are visible it appears that no measurement is greater for the failed machines then for the non-failed. This would go against what the engineer had suggested as a possible root cause. Below is the median for each measurement for failed and non-failed machines to help confirm.

## failed non-failed  
## MEASURE1 539.0 1140.5  
## MEASURE2 561366 1182387  
## MEASURE3 2510.0 4524.5  
## MEASURE4 2309 5768  
## MEASURE5 29856 71700  
## MEASURE6 15073 29155  
## MEASURE7 17924.0 40113.5  
## MEASURE8 21046 62947  
## MEASURE9 33395.0 100014.5  
## MEASURE10 205141 336572  
## MEASURE11 151153 248659  
## MEASURE12 212014 303548  
## MEASURE13 210883 323930  
## MEASURE14 205541.0 316324.5  
## MEASURE15 389940 626929  
## MEASURE16 129775 260118  
## MEASURE17 84 48

From the medians it would appear that only MEASURE17 may have importance for categorizing failed machines and non-failed machines. At this point it seems like the model will not be able to predict very well. One potential option may be to convert each bin from operating time in each bin to the proportion of the total time spent in each bin. This would also help reduce the outliers that have a large amount of operating time.

### Initial Model

Before any data manipulation is done, an initial model with the raw data will be built to use for comparison purposes later. All modeling will be assessed for accuracy using cross validation. To achieve this the data is separated into a test and training set of data. The training set of data will be used to build the model (contains 80% of the original data) and the test set will be used to validate the model. Below is a contingency table of the results from the model when applied to a test data set.

## Predicted Fail Predicted non-Failed  
## Failed Widget 0 10  
## non-Failed Widget 1 195

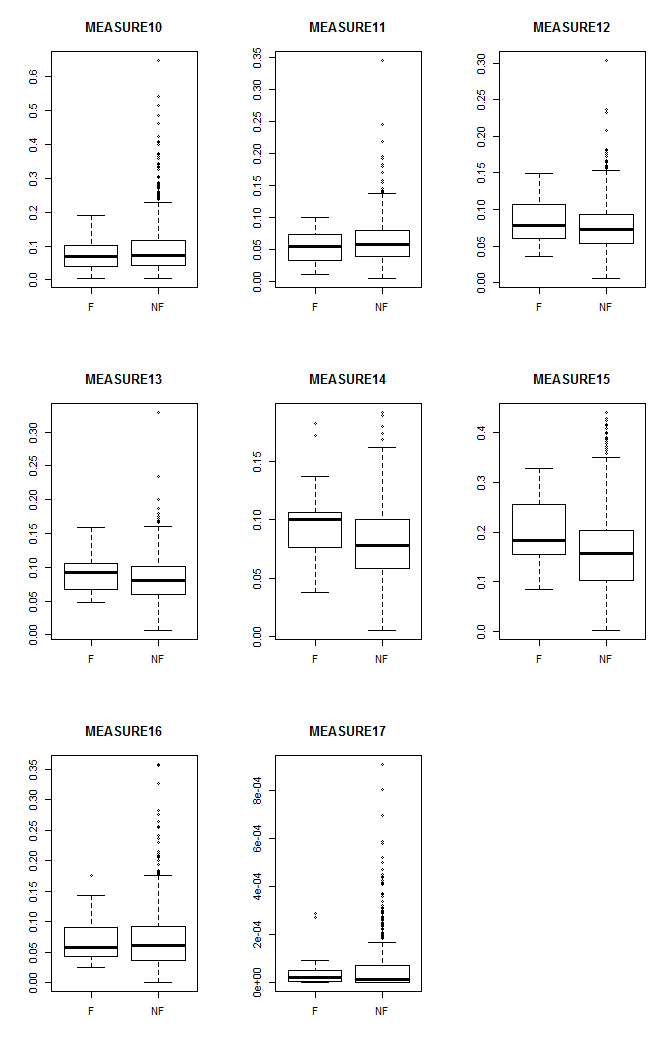
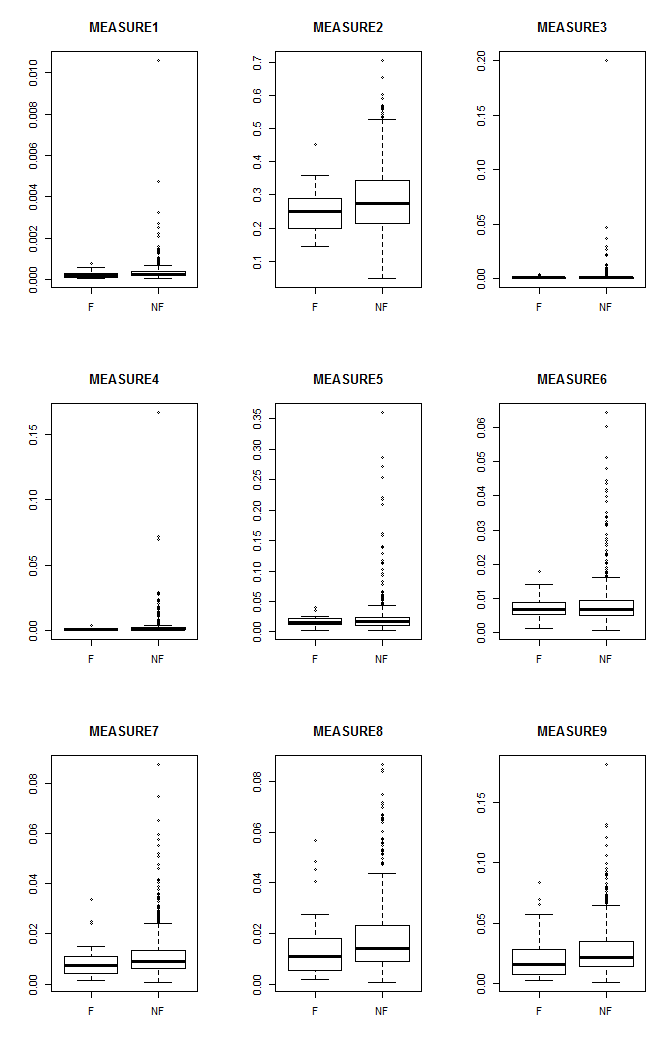
The percentage of correctly predicted outcomes is 95% and 0 of the machines with a failed Widget were actually predicted to be failed. This is not an effective model. The quantity of non-failed machines is vastly larger than the failed machines; this makes the model seem more effective than it is in real life. As long as the model categorizes a large portion of the test data as non-failed it will always appear to be accurate. This will be addressed in proceeding sections.

There were no significant coefficients from the logistic regression model. This seems to align with the initial assessment of the measurements. It is clear from the original model that some additional manipulation of the data needs to occur.

# III. Data Transformation

There are two main concerns with the data at this point. First, the predictor variables are heavily skewed with many outliers. Second, the failed population of machines is significantly smaller than the non-failed machines.

The predictor variables could be transformed to force a more normal distributed by taking the natural log of the predictor. This transformation does alleviate the skewness but there is still a fundamental issue with the data. From the initial look at the raw data it was observed that only one parameter had a larger median for the failed population than for the non-failed population. Transforming the predictor variables using the natural log will not alleviate this issue. Instead the predictor variables, which are operating time in bin, will be converted into the proportion of time in a bin. This is done by taking the time in a bin and dividing it by the total operating time in all bins. This transformation did not completely get rid of the data skewness but it did help as seen below:



The second concern is that the failed population is much smaller than the non-failed population. This issue is making the model appear more accurate than it is, even when cross validating. To help account for this, the failed machine population for the training set of data will be artificially increased using oversampling. New records will be added to the failed population by resampling from the original failed population. In this case 783 rows will be added so that the failed population is the same size as the non-failed population for the training set of data.

Below is a contingency table of the results from the model with the new data transformations:

## Predicted Fail Predicted non-Failed  
## Failed Widget 6 4  
## non-Failed Widget 51 145

The percentage of correctly predicted outcomes is 73% and 6 of the machines with a failed Widget were actually predicted to be failed. This is slightly better than the first model created but still not acceptable. The model was able to at least correctly predict some of the failed machines but the Type II error significantly increased. To improve the model’s accuracy statistically significant measurements will be determined.

# IV. Statistically Significant Measures

It is important to note here that previously the only significant measurements were the ones were the median for the failed population was greater than for the non-failed population. Now that the data has been transformed to a proportion of the total time this hypothesis no longer applies. Now a lower median for the non-failed population could be indicative of a different usage cycle of the machine. This changes the lead engineer’s hypothesis slightly but not significantly. John Deere vehicles are used differently by each customer so it is reasonable to believe that how the machine is being utilized could ultimately decide if the Widget will fail or not.

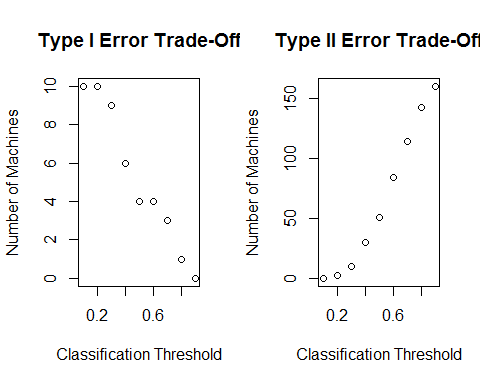
Using the stepwise method, the 17 original bins were able to be reduced to 9 statistically significant bins. The Variance Inflation Factor was also verified for each measurement, all were under two. A new model will be built using: MEASURE1, MEASURE2, MEASURE4, MEASURE11, MEASURE12, MEASURE13, MEASURE14, MEASURE15 and MEASURE16.

Below is a contingency table of the new model:

## Predicted Fail Predicted non-Failed  
## Failed Widget 6 4  
## non-Failed Widget 51 145

The percentage of correctly predicted outcomes is 73% and 6 of the machines with a failed Widget were actually predicted to be failed. There is no difference between this new model and the previous model. This new model with the reduced amount of parameters will be utilized as the best fitted model.

The classification threshold that has been used for determining if the Widget has failed or not is 0.5. This threshold can be adjusted in order to achieve more desirable results but there is a tradeoff. The tradeoff is between Type I and Type II errors. See the plots below to see the trade-offs between Type I and Type II error and the classification threshold.



For this particular case Type II errors are very costly which will be discussed in more detail later. For this reason the classification threshold should not be increased at all. The classification threshold could be reduced to help reduce the Type II error but then Type I error will increase and reduce the effectiveness of the model. A value of 0.5 appears to be a reasonable threshold.

# V. Results Summary

The final model created is the suggested model to utilize. This model utilizes the proportion of time instead of the raw time values and reduced the initial 17 parameters down to 10: MEASURE1, MEASURE2, MEASURE4, MEASURE11, MEASURE12, MEASURE13, MEASURE14, MEASURE15, MEASURE16 and MEASURE17. This model correctly predicted 60% of the failed machines.

When working with customers and customer’s data a company must be careful. In this particular case the failed Widget is not a potential safety concern and most customers will probably not notice the failed Widget. While it is good to replace broken Widgets, a major concern is disrupting a customer for a “non-issue”. This means that a model must be partly chosen on the amount of Type II errors which would cause a Widget to be replaced even when it is not broken.

Overall the model accuracy in general is not very high for predicting failed Widgets. With the model only being able to predict 60% of the failed Widgets and the fact that a failed Widget is not a safety concern, it may be beneficial to continue monitoring MEASUREY. The Widget can then be replaced if a shift in MEASUREY is seen. This model could be potentially used to create a list of high potential machines to monitor but it is not accurate enough to be the only tool to aid in fixing this failure.