Data Discovery Document for



BMIS 2.0

Prepared by Tiger Analytics

Sep 09, 2016

**Table of Contents**

[1 Objective 3](#_Toc461211542)

[2 Data Sources 3](#_Toc461211543)

[3 Analyses 4](#_Toc461211544)

[3.1 Sensors 4](#_Toc461211545)

[3.2 Clamps 5](#_Toc461211546)

[3.3 Stop Case 6](#_Toc461211547)

[4 Model Structure 7](#_Toc461211548)

[4.1 Model for Sensor Failure 7](#_Toc461211549)

[4.2 Model for Clamp Failure 7](#_Toc461211550)

[4.3 Model for Stop Cases 8](#_Toc461211551)

[5 Phase II Recommendation 8](#_Toc461211552)

[6 Solution Approach 9](#_Toc461211553)

# Objective

The objective of the of the Data Discovery phase is to understand the data that is available from MidTronics’ data sources and evaluate the sufficiency of data for building models for :

1. Predicting the Likelihood of Sensor Failures
2. Predicting the Likelihood of Clamp Failures
3. Predicting the Likelihood of Stop Case

# Data Sources

The following data was used for the analysis:

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **File Name** | **# Records** | **# Variables** |
| 1 | BMISFPV\_Aug10.txt | 21,299,436 | 85 |
| 2 | BMISFPV\_Aug10\_PackTest.txt | 4,382 | 32 |
| 3 | BMISFPV\_Aug10\_TestPack.txt | 7,143 | 108 |
| 4 | BMISFPV\_Aug10\_TestQC.txt | 2,556,083 | 55 |
| 5 | BMISFPV\_Aug10\_Tests\_ShopInfo.txt | 10,949 | 15 |
| 6 | BMISFPV\_Aug10\_Tests\_TestTypegroups.txt | 24 | 3 |
| 7 | BMISFPV\_Aug10\_Tests\_UserScreens.txt | 95,583 | 4 |
| 8 | BMISFPV\_Aug10\_TestsCharging.txt | 761,524 | 11 |
| 9 | BMISFPV\_Aug10\_TestsPartNumber.txt | 2,971,577 | 10 |
| 10 | BMISFPV\_Aug10\_TestSystem.txt | 698,027 | 28 |
| 11 | BMISFPV\_Aug10\_TestTotal.txt | 21,565,692 | 30 |
| 12 | BMISFPV\_Aug10TestsError.txt | 7,855,172 | 16 |
| 13 | BMISFPV\_Aug10TestsGraphPoint.txt | 421,928 | 55 |
| 14 | Test tables.xls(Tab : Tests\_Coupon) | 59 | 10 |
| 15 | Test tables.xls(Tab : Tests\_CableDrop) | 523 | 56 |
| 16 | Test tables.xls(Tab : TestLocationGroups) | 12 | 3 |
| 17 | ToolTypes.xlsx | 14 | 3 |
| 18 | Battery Decisions.xlsx | 1,701 | 6 |

Of this the ‘core’ table most relevant to the analysis had the following records:

|  |  |  |
| --- | --- | --- |
| **Core Table(BMISFPV\_Aug10.txt)** | | |
| **Metric** | **Count** | **%** |
| # Clients | 39 |  |
| # Records | 21,299,436 |  |
| # Records in 2012 | 199,536 | 1% |
| # Records in 2013 | 6,452,893 | 30% |
| # Records in 2014 | 6,713,763 | 32% |
| # Records in 2015 | 6,454,823 | 30% |
| # Records in 2016 | 1,471,932 | 7% |
| Others/Outliers | 6,489 | 0.03% |

For all the analyses the information in the core table was used. In addition, the information in the Charging table was used for the analyses on Stop Cases.

# Analyses

In each of the three cases, we defined the objective variable to be predicted based on the inputs received. We then subsequently observed the distribution of the values in the historical data in order to see and if there was sufficient variation across the likely predictive variables.

## Sensors

The following ranges were used to define ‘Failure’, ‘Warning’ and ‘Normal’.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Temperature Range** | **Records** | **Unique Tools** |
| **Failure** | **<= -60F** | 9 | 3 |
| **Warning** | **-60 F to - 40 F** | 293,080 | 510 |
| **Normal** | **-40F to -140F** | 20,556,114 | 11,596 |
| **Warning** | **140F to 160F** | 54,054 | 2,677 |
| **Failure** | **>=160F** | 395,892 | 1,412 |
|  |  | **21,299,149** | **11,631** |

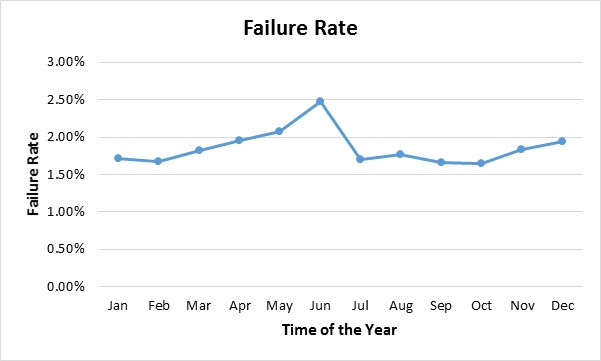
The tools here are both chargers as well as testers. At a tool level the objective is to predict the likelihood the next transaction of that tool will have a likely failure of the temperature sensor, given the history thus far.

The failures are distributed well by location, timeline (history), there is all notable variability by month of the year. The combination of location and month of the year, is a good proxy for weather, which we expect to be an influential variable.

**Failure Rate by Location**

*Regions were defined based on contiguous groupings of geographies.*

**Failure Rate by Month**



**Note:** The conclusions about data adequacy are very much dependent on the exact definition of the ‘Failure’. So, if the temperature ranges that define failure were to be modified, it will likely impact the predictive power of the model.

## Clamps

The data for this is available only for one client: Client 12.  
A clamp is said to have failed if the tool (may be a charger or a tester) registered a positive or negative resistance >=500 micro-ohms.

We observed the following distribution of observations

|  |  |  |  |
| --- | --- | --- | --- |
| **Observations** | | | |
| **Clamp non-Failure** | Below 500 micro-ohms in both Positive and Negative Resistance | 210,015 | 99.72% |
| **Clamp Failure** | '>=500 mico-ohms in either Positive or Negative Resistance | 594 | 0.28% |
|  |  | **210,609** |  |
|  |  |  |  |
| **Unique Tools** | | |  |
| Never experienced clamp failure |  | 33 |  |
| At least one instance of clamp failure |  | 40 |  |
|  |  | **73** |  |

As the model is to be built at tool level, it is necessary to see sufficient number of tools, variability across tools and sufficient number of observations per tool.

The **594** observations of failure are observed in 40 Tools.

Almost all the observations are from Illinois and California.

**We found some variation by time and location**

**Note**: The conclusions about data adequacy are very much dependent on the exact definition of the ‘Clamp Failure’. So, if the resistance ranges that define failure were to be modified, it will likely impact the predictive power of the model.

## Stop Case

Here the Stop Case of interest is Stop Case **18**.

The objective is to predict the likelihood that the next charging transaction of the battery will return a StopCase 18   
The distribution was observed to be as follows:

|  |  |  |
| --- | --- | --- |
| **Total Count of Stop Cases** | **Count** | **%** |
| Stop Case 18 | 9,216 | 1.3% |
| Stop Case 0 | 332,462 | 45.4% |
| Rest of the Stop Cases | 390,019 | 54.3% |
|  | 731,697 | 100% |

As a proxy for the battery we used the vehicle identification number.

The number of unique Vin numbers that have experienced stop case 18 are 5918

Here we calculated the average number of times we can expect to see a battery per year, in the data:

* Only 17% of the Vins have appeared more than once
* The average no. of days between two charging sessions for the same battery is around 150 days.

There is insufficient enough history at a battery (Vin Number) level.

# Model Structure

In the case of both sensors and clamps, the prediction is at the tool level. i.e. based on the history of that particular tool the likelihood of the failure of the next observation of that tool.

## Model for Sensor Failure

* Model output = a score that is proportional to the likelihood of Sensor Failure (i.e. Temperature falling in the range of either **<= -60 F** or **>= 160F**)
* Model Type = A logistic regression model (or a Random Forest Model, to be chosen based on what model type will be most predictive and stable)
* Input Variables: Historical (up until the current week’s) values of Temperature readings, voltage, CCA, State of Charge, Location, Temperature etc
* Note: the model takes into account all information available to predict the imminent failure of the sensor in the next week
* Score Threshold
  + a pre-decided threshold score over which the sensor’s next transaction will likely fail.
  + Threshold is decided based on the acceptable trade-off between detection of failures and false-positives. This can be revised in production with no change to the underlying model.

## Model for Clamp Failure

* Model output = a score that is proportional to the likelihood of Clamp Failure (i.e. Positive of Negative Resistance >= 500 micro-ohms)
* Model Type = A logistic regression model (or a Random Forest Model to be chosen based on what is most predictive and stable)
* Input Variables: Historical (up until the current week’s) values of Temperature readings, voltage, CCA, State of Charge, Location, Temperature etc
* Note: the model takes into account all information available to predict the imminent failure of the sensor in the next week
* Score Threshold
  + a pre-decided threshold score over which the sensor’s next transaction will likely fail.
  + Threshold is decided based on the acceptable trade-off between detection of failures and false-positives. This can be revised in production with no change to the underlying model

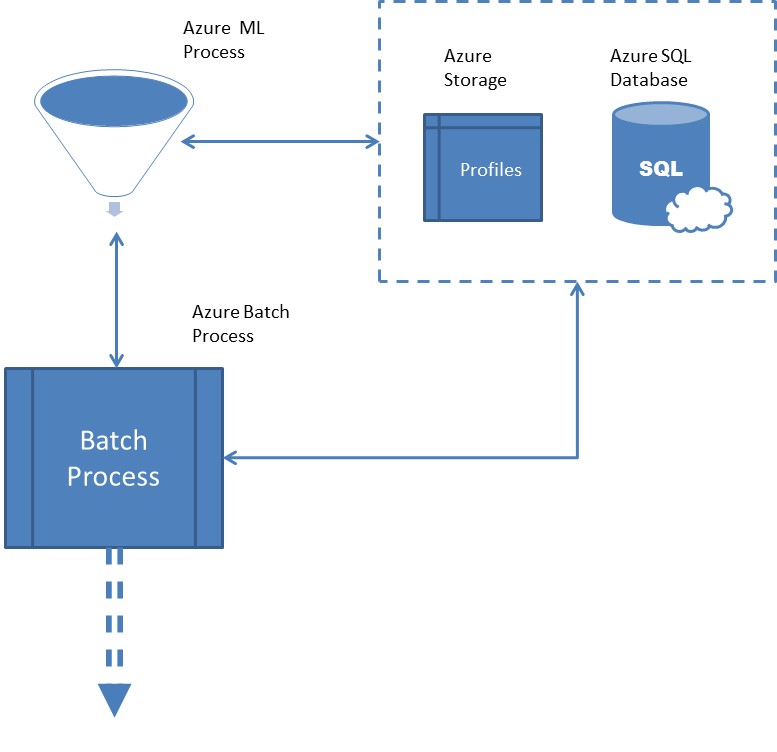
## Model for Stop Cases

This is a model that cannot be built at a tool level but would need to be built at a battery (proxied by vin number) level. However, as we do not see a battery itself to have many charges per year, the ‘trigger’ for an updated calculation in the score will not be a charge in that specific battery itself. We will need to update the score periodically based on updates on other charging transactions of other batteries.  
  
However such a model would be sufficiently predictive, only if there were some inherent patterns by which batteries (proxied by vins) can be grouped, such that they had similar patterns of occurrences of Stop Case 18 consistently over time.  
  
We created clusters of the charging observations by vin number, adding in other variables related to the charging event, location, time etc. Unfortunately there was not much clear differentiation across clusters. This points to the fact that the likelihood of a stop-case is substantially influenced only by that very battery and not to any generalizable characteristics that are common to other batteries.

# Phase II Recommendation

* We recommend that we proceed to Phase II for predicting Temperature Sensor Failure.
* With the data that will be available in production we can build a reasonable model for Clamp Failure
  + However we expect it may not be as strongly predictive as the Temp. Sensor Failure Moderl
  + If we are able to collect data of Clamp Failure across multiple clients, to make it as rich as the data of Temperature Sensors, then we will be able to build a good Clamp Failure Model.
* We do not see a scope for a good predictive model to predict the Stop Cases.
  + The limitation is not the availability in data as much as it is the structure of the problem itself – the causes of Stop Case seem particular to the battery itself and not sufficiently generalizable.

# Solution Approach



* The model equations are first developed and then implemented in Azure ML framework
* The variables that will be used to score the model in production will come from
  + Direct variables collected in the Azure SQL DB
  + Features/Profiles calculated from the variables, that will in turn feed the model
* Based on all charging and testing transactions in the recent history including that week, the values of the Features/Profiles will be calculated
* These will feed into the model equations and a score for each tool will be generated.
  + i.e. one score for likelihood of Sensor Failure (as scored by the Sensor Failure Model)
  + one score for the likelihood of Clamp Failure (as scored by the Clamp Failure Model)
* The threshold above which a flag out to be raised would be pre-decided based on analyses
  + This threshold value will exist alongside the features and will be automatically called in after model scoring to raise the flag
  + The threshold value should also be mutable, based on any downstream analyses, without having to make any change in the model
* The output of the batch-process would finally be the list of tools where
  + The flag for ‘likely impending Sensor failure’ is raised
  + The flag for ‘likely impending Clamp failure’ is raised