Assignment Part I: Involved reading an article that provided a case study of an actual decision sciences system and answers several questions about the application such as holistic…

Experiment Findings · March 2019

DOI: 10.13140/RG.2.2.24534.50247

1 author:

Sunil Mandhan
Hitachi, Ltd.

7 PUBLICATIONS **4** CITATIONS

1. PROBLEM/PURPOSE

Assignment Part I: Involved reading an article that provided a case study of an actual decision sciences system and answers several questions about the application such as holistic decision system, data quality, technical constraints, modeling approach etc.

Assignment Part 2: To analyze data associated with the services aspect of the Automobile manufacturing company that a company has purchased, and present some conclusions and recommendations to senior management.

3. ASSUMPTIONS

- The data values (those were out of range) were removed.
- Data 3 file was dropped while merging the files as the range of time to failure variable was very different from other files.

4. RESULTS

Using all the predictors, the performance was:
R-square=55.9, Adjusted R-square= 51.3, C-p= 8.0 and s=48.889.

By using the key predictors the values were:
R-square= 61.5, Adjusted R-square= 60.1, C-p= 5.0 and s=43.221.

The final model equation as a result of regressing analysis:
Time to Failure = - 485 - 74.1 Die Position - 0.0419 Vibration + 98.4 Impurities + 0.683 NM

5. CONCLUSIONS

This assignment helped in understanding the development of an Automated Decision System (ADS) and the application of Six Sigma principles namely DFSS (Design for Six Sigma) and DMAIC (Design, Measure, Analyze, Improve and Control) throughout its lifecycle. Some of the design trade-offs were addressed that came into play while designing the decision system. Also, some of the major data quality issues were addressed and solutions were recommended.

Table of Contents

List of Figures

List of Tables

1. PROBLEM STATEMENT

The problem consists of two parts and is described below.

Assignment Part I:

Involved reading an article that provided a case study of an actual decision sciences system and answers several questions about the application such as holistic decision system, data quality, technical constraints, modeling approach etc.

Assignment Part 2:

To analyze data associated with the services aspect of the Automobile manufacturing company that a company has purchased, and present some conclusions and recommendations to senior management.

2. RESULTS AND ANALYSIS

2.1 Assignment Part 1:

Q1. In what ways do the authors depict a holistic decision sciences system in their application, rather than simply a mathematical model? Provide as many specific examples as possible.

Answer:

The authors Patterson, Bonissone and Pavese present the development of an Automated Decision System (ADS) and the application of Six Sigma principles namely DFSS (Design for Six Sigma) and DMAIC (Design, Measure, Analyze, Improve and Control) throughout its lifecycle. Although from a technical perspective, designing an ADS is a big technical and mathematical challenge and a typical research paper may become encumbered by nice looking plots and technical details of approaches like Fuzzy-Logic (FL), Neural-Network (NN), Evolutionary Algorithm (EA) etc., yet, the paper follows a practical approach to satisfy realistic business goals and explores various challenges in designing, maintaining and updating such a system and their repercussions and remedial actions.

There are many examples throughout the paper, which clearly show that the authors have opted for a holistic analysis of the problem at hand over a complex mathematical model that breaks down in the face of realistic constraints. They are listed below:

1. The authors present not just a mathematical model for an Automated Decision Engine (ADE) but provide supporting practices and techniques to sustain the model throughout its lifecycle from definition to retirement. They discuss methods to build, use, and monitor, maintain and update the model so that it remains useful and up-to-date by adopting a DFSS+DMAIC approach.
2. The authors discuss the design trade-offs that come into play in the designing of the decision system:
    a. Accuracy versus Coverage
    b. Accuracy versus Interpretability
    c. Run-time efficiency versus configurable architecture

    Before formulating the model, they explore why a classifier must choose to maximize its scope or percentage of correct decisions. They point out that compliance/legal regulations may require the decisions to be transparent and justifiable and still not be overly simplistic. They also discuss the benefits of configuration driven architecture vis-à-vis hardcoded configuration to increase the speed of decision making.
3. The choice of modeling the input vector X as a combination of discrete, continuous and attribute variables and output vector Y as a discrete variable (bins) is motivated by the nature of the problem at hand which is an insurance underwriting decision system.

4. The choice of modeling techniques is driven by problem constraints such as non-linearity, transparency, business flexibility and openness to interpretations. Thus the authors justify their mathematical models.
5. The authors provide answers to the critical questions "What do I want to do?" and "How will I know I did it?" They dig out various CTQ's (Critical to Quality) for the model, Information Technology and maintainability requirements. Their CTQ's are based on dollar value in NPV (Net Present Value) and they propose to ensure maintainability by creating an activity checklist and homework problems to ascertain familiarity of maintenance team with the specific tools and processes.
6. The authors deal with data-quality and data-consolidation issues by designing a data gathering process and storing data in a temporary database. They compensated for under-represented categories of data by stratifying random sampling and finding performance metrics for each stratum. Most importantly, the authors created a dataset of Standard Reference Decisions (SRD's) which was used to test the validity of the data itself by brainstorming with subject experts and arriving at a consensus decision. Feedback was given to the data collection team to incorporate any missing/hidden input variables in the future.
7. The authors discuss technology selection wherein they discard Rule Based (RB) systems for not being scalable, Logistic Regression (LR) for low accuracy. It finally proposes to use the more accurate but opaque schemes of Neural Network (NN) and Multivariate Adaptive Regression Splines (MARS) for offline quality assurance and the lesser accurate but transparent Fuzzy Logic (FL) model for online underwriting decision making. This way they achieved a balance between accuracy and interpretability.
8. To account for the inability of business users to set optimal parameters for the FL model, the authors used an Evolutionary Algorithm (EA) to optimize the FL model parameters.
9. The authors validate their model against measures of coverage and global and relative accuracy, which are very practical and business critical parameters.
10. The authors used (Failure Modes Effects and Analysis) FMEA and root cause analysis to maintain and update the model continuously.

Q2. In what ways did the authors consider and address data quality? Do you feel these efforts were sufficient; why or why not?

Answer:

The authors use data to train the automated decision model and also to validate it. The authors had a resource of about 3000 cases of insurance cases which had been manually underwritten. The authors chose about 1200 of these for building the model and kept the rest as backup data. After an initial brainstorming session, a sufficient set of input parameters X's for the decision making process were decided on.

It was found that the historical data did not contain all these X's. These were acquired during the next data collection process. Also many of the parameters were not available in a digital form which could be queried and searched. Hence authors designed a data gathering process by taking a random sample of historical cases, and having the relevant data from those cases entered into a temporary database.

Authors also found that certain types of cases were under-represented in the data, especially those that, although within the ADE's scope, were placed into worst-risk categories. To mitigate this, authors took additional care in the next data collection round to collect random cases from the above category. This resulted in stratification of performance metrics for various categories.

The authors also tried to validate the data itself by deriving a dataset of SRD's from the consensus amongst subject experts on a set of insurance cases. This was used to do a Measurement System Analysis (MSA) of the original data.

The authors made significant efforts to identify data-quality issues and tried to mitigate them. As there were business limitations, such as the IT system which was being developed concurrently with the ADE, the authors had to face data quality and scarcity issues. But, the authors incorporated sufficient procedures to improve data quality.

Q3. There were some practical constraints in this case that limited the modeling approaches that could be used. Explain what these constraints were, and why they led the authors to the approach that they eventually used.

Answer:

There were indeed many constraints that limited the approaches that could be used to model the insurance underwriting problem at hand. They majorly concerned with the following four requirements:

1. Accuracy- imposes limits on both global and relative accuracy of the ADE
2. Coverage- the minimum scope of the ADE
3. Transparency: A legal/compliance requirement for decisions to be transparent
4. Business flexibility: should have a good balance between risk-tolerance for market competitiveness and risk-avoidance to prevent over-exposure to risk.

Due to the nature of problem at hand, some modeling approaches were better suited. Some examples of data related constraints are:

1. The underwriting mapping is highly non-linear, i.e. small incremental changes in one of the input components can cause large changes in the corresponding rate class.
2. There was no "true" decision for many insurance cases as underwriter's subjective judgment always led to variability in their decisions.

The input vector X was modeled as a combination of continuous, discrete and attributes variables while the output vector Y was a discrete variable that represented difference rate classes of insurance cases such as standard, worse- case etc.

There were many modeling technique s that could be used, namely, RB systems, LR, NN, MARS and FL. RB was rejected as an RB system already existed and was under patent. This was done to avoid patent infringement issues. The LR model could not provide the required level of accuracy owing to the strong non-linearity in the data. The NN and MARS models while providing good accuracy were not interpretable in the sense that their decisions were not transparent and could not be justified in logical sentences. This impaired transparency of insurance decisions to the customer and hence were sidelined for underwriting decision process. The FL model was found to sufficiently accurate and transparent to be used in the ADE. The NN and MARS models were used to assure offline decision quality.

Since the business users would not have been able to always select the optimal set of parameters for the FL model, the parameters were optimized through an EA and then fed to the FL model. Thus the business users could simply set up the basic structure and allow the EA to perform the final parameter selection and optimization.

Three parameters, coverage, relative accuracy and global accuracy were chosen to be the measurement criteria which were to be maximized by the model.

Q4. In "plain English", rather than technical jargon, explain the basic modeling approach used by the authors.

Answer.

Authors used the following modeling approaches:

1. fuzzy logic modeling for online production use
2. Evolutionary algorithms (EA) were used to tune and optimize the parameters used for decision engine built on for fuzzy logic modeling.

Each of the above modeling approaches are described in ample details below.

1. **Fuzzy model**

   This model was primarily chosen for online production use because of its accuracy, simplicity and interpretability it provides.

   The Fuzzy Logic Engine (FLE) uses rule sets to encode underwriting standards. Each rule set represents a set of fuzzy constraints defining the boundaries between rate classes.

   These constraints were first determined from the underwriting guidelines. They were then refined using knowledge engineering sessions with expert underwriters to identify factors such as blood pressure levels and cholesterol levels, which are critical in defining the applicant's risk and corresponding premium.

   The goal of the classifier is to assign an applicant to the most competitive rate class, providing that the applicant's vital data meet all of the constraints of that particular rate class to a minimum degree of satisfaction.

2. **Evolutionary Algorithms (EA)**

   These algorithms are used to tune the parameters for the decision engine based on fuzzy logic modeling.

   The design engine's design parameters must be tuned, monitored, and maintained to assure the classifier's optimal performance. EAs are chosen, an optimization paradigm based on the theory of evolution and natural selection.

   EA used in this system is composed of a population of individuals ('chromosomes'), each of which contains a vector of elements that represent distinct tunable parameters to configure the decision engine (classifier). A chromosome, the genotypic representation of an individual, defines a complete parametric configuration of the classifier. Thus, an instance of such a classifier can be initialized for each chromosome. Each chromosome of the population goes through a decoding process to allow them to initialize the classifier.

Each classifier is then tested on all the cases in the test set, assigning a rate class to each case. In this case, the test set was composed of a subset of the SRDs. The quality can be determined of the configuration encoded by the chromosome (the 'fitness' of the chromosome) by analyzing the results of the test. EA in this case uses mutation (randomly permuting parameters of a single chromosome) to produce new individuals in the population. The more fit chromosomes in generation t will be more likely to be selected for this and pass their genetic material to the next generation t + 1. Similarly, the less fit solutions will be culled from the population. At the conclusion of the EA's execution the best chromosome of the last generation determines the classifier's configuration.

Q5. How did the authors "close the loop" to ensure effective use of the final analytical system, and maintenance over time?

Answer.

"Closing the loop" ensures that project or system shall be useful over its life time for the business and it will be adaptable to the desired changes in future.

The ultimate success of any project can be judged by its acceptance by the user community, and its impact on business over a period of time. This success can be achieved by the following principles:

1. Integrating the outcome of the project (i.e., decision tool, analysis, report) to the everyday business process.
2. Ensuring a "control plan" to maintain and improve it over time.

More details on closing the loop are given below.

1. **Integrating the outcome of the project with the everyday business process:**

Authors have suggested using the financial analytical system for every case. This will ensures that the said system is the main entity handling the business requirements of analyzing the underwriting cases.
The system then decides if it can make the decision itself or human intervention is needed for completing the case.
This process clearly shows that the system is the primary processing unit and integrated into the business so well that all the cases are passed through it.

Also, the system is adaptable to the experiences based on new leanings and business rules and can be trained with little efforts.

2. **Maintenance and Control plan:**

The approach used for the maintenance is six sigma DMAIC process. By applying the rigor of the Six Sigma DMAIC roadmap during the remainder of the engine's lifecycle, business risk can be minimized and the return on the investment to build the automated decision system can be maximized. DMAIC activities pertaining to this project are described below.

Define (D):
All the CTQs that were taken care during the building of the system should be monitored. The two main CTQs for the discussed financial systems are:

1. Coverage:
   It provides what percentage of cases is handled by the automatic decision engine. If this is being affected then lots of resources and money will be wasted doing long manual work.

2. Accuracy
   Also the system should be accurate in results it provides otherwise serious financial losses can occur and confidence in the system may be lost.

Measure (M) and Analysis (A):
If some statistical evidence is know that there is some shift in the decision indicator, and then audit is raised for the system to know whether this shift is real or there is a problem with the decision engine. If the shift is due to the decision engine then maintenance staff is notified.

Improve (I) and Control (C)
The necessary changes to the decision engine rules are done in this phase based on the changed context and rules.

These changes are done with the help of update and maintenance tools provided with the financial system. Also these changes to the engine are tested with the help of applicable SRDs and new SRDs. After the successful testing new decision engine is used in the business process.

The system was built so that it can be changed and tuned easily. There were many issues faced when building the system as adaptable to the future changes.

The first step to solving this issue was to design the engine so that additions and changes could be made without altering the actual engine software itself. Therefore, a completely configuration-file-driven design was chosen.

No arbitrary limits were placed on the number of FL (Fuzzy Logic) rules or the type of data they could consider. Therefore, the maintenance of the knowledge base of the model was separated from the maintenance of the software. Business users could be in charge of the former, while the IT staff would be in charge of the latter.

The second step in planning for maintenance was to realize that business users would not always be able to select the optimal set of parameters for the FL (Fuzzy Logic) model. They may be able to provide the structure of the model (e.g. that an applicant's blood pressure should influence their rate class assignment), and an initial estimate of the parameters, but setting the detailed parameters would be a challenge.

To address this, a novel approach was taken, to couple the FL model with an evolutionary algorithm (EA) optimization, which could aid the business users in finding the correct set of model parameters to make the right decisions given a data set.

Thus, business users would not have to worry about setting exactly precise parameters within the engine, but could simply set up the basic structure and allow the EA to perform the final parameter search and optimization.

An additional benefit of this approach was that it could be applied right away to the SRD (Standard Reference Decision) data set, to ensure that the decision engine was making optimal decisions even prior to its being put into production.

2.2 Assignment Part 2:

Q1: Based on the data available, what appear to be the key drivers of transmission reliability?

Answer:

To find the key drivers, firstly all the four data files were analyzed individually. Each of the variables present in the file was plotted against the Time to failure. Looking at the plots generated some initial guess of the key drivers were made for each of the data file separately.

Die Positions in all the files appeared as one important variable to predict the Time of Failure. Vibrations also showed some correlation with the Time to Failure. Then Minitab was used to do best subset regression analysis with Die Positions, Vibrations as Predictor to be used for all the models and rest of the variables as Free Predictors. Later Impurities and Torque was also added to the list of Predictors for the model.

When the data from all the four files was merged the same predictors (Die Positions, Vibrations, Impurities and Torque) again came out to the key predictors. In this the Data 3 file was dropped as the range of the variable Time to failure in this file was very different from the other files.

Using all the predictors, the performance was:

R-square=55.9, Adjusted R-square= 51.3, C-p= 8.0 and s=48.889.

By using the key predictors the values were:

R-square= 61.5, Adjusted R-square= 60.1, C-p= 5.0 and s=43.221.

The graphs showing the plot of Time to Failure versus each key predictor for the final merged data are shown below:

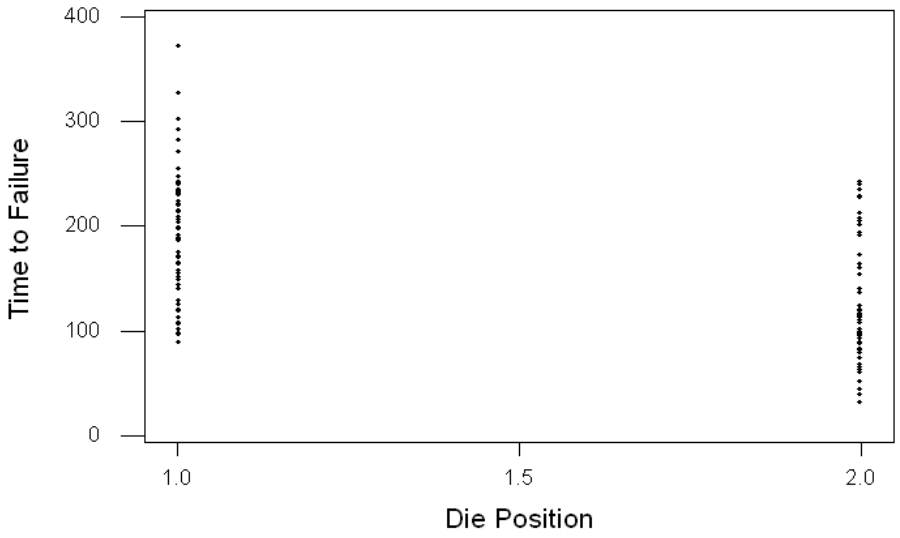


Figure 1 – Graph of Time to Failure versus Die Position

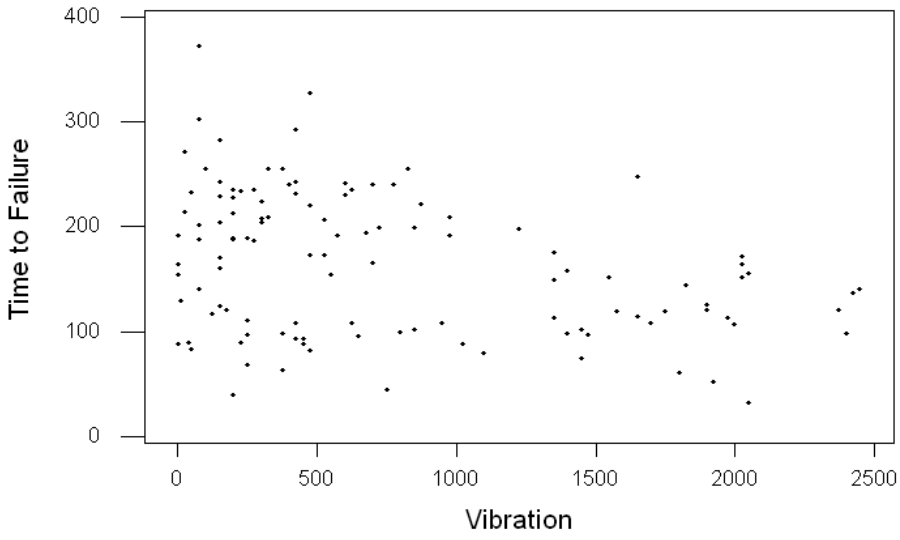


Figure 2 – Graph of Time to Failure versus Vibration

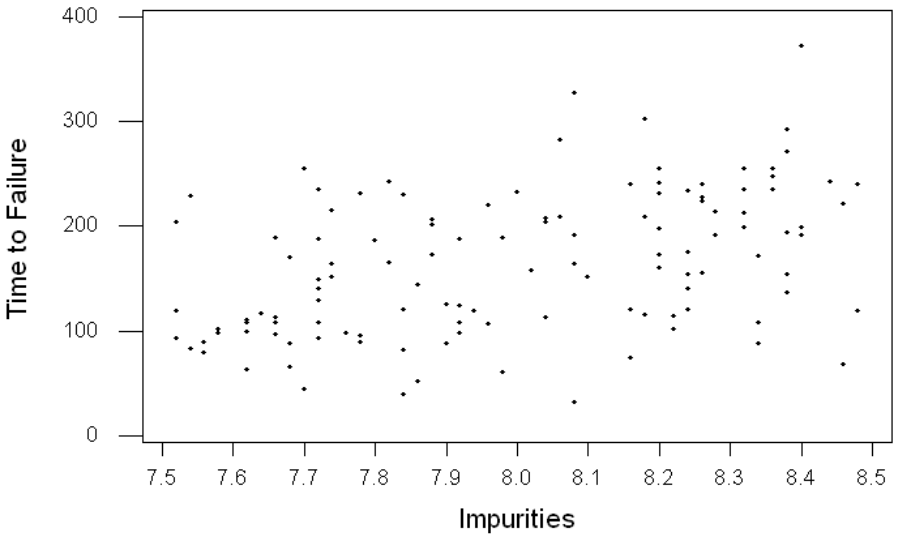


Figure 3 – Graph of Time to Failure versus Impurities

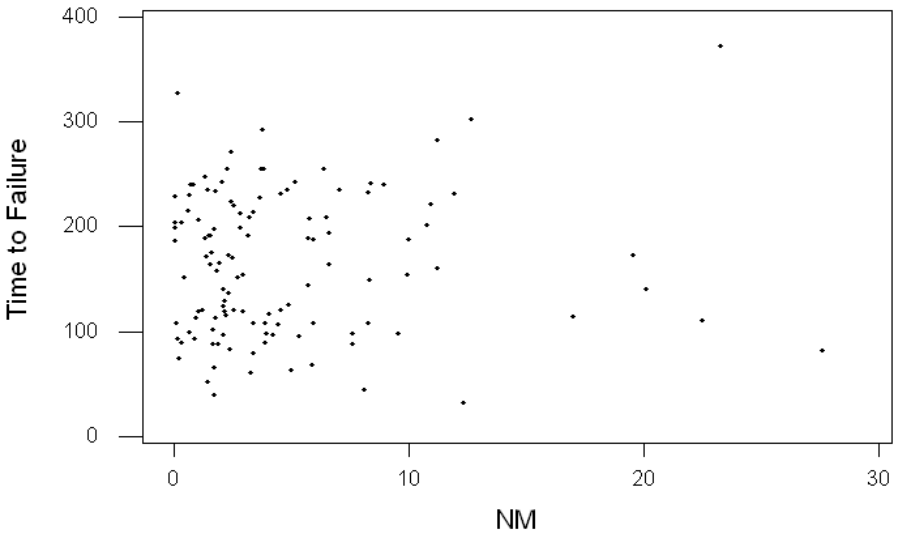


Figure 4 – Graph of Time to Failure versus Torque (NM)

After doing a Regression Analysis, the final model came out to be:

The regression equation is below:

Time to Failure = - 485 - 74.1 Die Position - 0.0419 Vibration + 98.4 Impurities + 0.683 NM

Note:

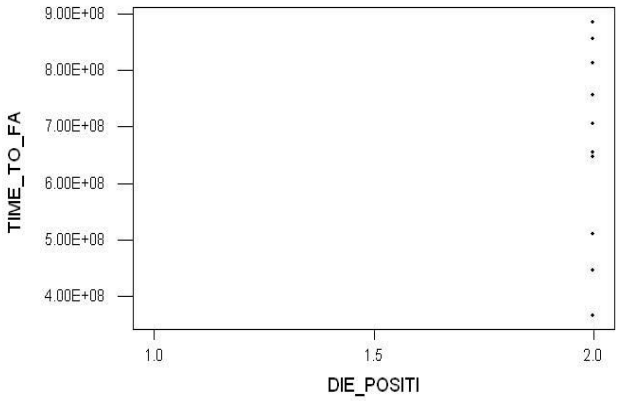
More detailed analysis and the plots for the other variables can be found in "Additional Analysis & Documentation Rev 1.pdf".

Q2: What issues have you found relative to data quality? What is your recommendation as to how these issues might be better addressed going forward?

Answer:

The major data quality issues faced is as follows:

1. The four enclosed data files did not contain the same set of variables.
2. The same variables were named differently. For example—NM, Torque Test refers to the same variable but are named differently. Some places missing value is left blank and at others it is denoted by "**#**".
3. In the third data file there are two Die Position values- 1 and 2. But the Time to Failure values is absent for all the rows where the Die Position value is 1. Hence the plot looks as follows:



4. In many of the Data file some of the values looked very much different from the usual values that the variable takes. A snapshot of Data 1 is shown below with the unusual values highlighted.

Table 1: Snapshot of Data 1

Serial #	Die Position	Design Type	Shift	Vibration	Impurities	Clearance	NM	Time to Failure
23	1	1	1	750	100000000	68.32	6.91	192
17	1	1000	2	425	7.64	68.53	1.81	146
44	2	1	2	1050	8.12	69.86	10.33	0
45	1	1	3	1350	8.46	69.16	5.6	0
46	2	#	1	925	8.2	1000	2.9	84
47	1	2	2	1300	999999999	70.63	3.62	849600
48	2	2	3	975	7.52	70.21	3.74	10800
49	1	2	1	1175	8.3	71.33	2.26	72000
50	2	#	1	1100	8.44	68.95	0.47	26100

5. The range of Time to Failure variable for the Data File 3 was very different from the usual range of the variable in the other three files.

Recommendation as to how these issues might be better addressed going forward:

1. The different names of the same variable can be figured out either by looking at the name itself like Torque and NM can be the same variable as the unit for Torque is NM. Additional validation can be done by plotting the two variables.
2. The unusual looking values should be dropped off as they can bias the model. For some of these values it would be very apparent that they are incorrect but for some values a decision will have to be taken whether it is a wrong value or whether it is due to some unusual operation of the machine.

Q3: What are your thoughts relative to the company's current approach to managing the data on transmission reliability? What is your recommendation as to how this might be handled better in the future?

Answer:

Company's current approach to managing the data on transmission reliability is leading to a lot of data quality issues. The same set of variables is not collected on different transmissions. A general convention of naming the variable is not followed. Some very absurd values are getting in. There is lot of Missing Values in the files.

Recommendation as to how this might be handled better in the future:

1. A common Convention for naming the variables should be followed.
2. Some checks on the range of values of the variables should be done before entering into the file.
3. Problems like very different values of a variable like Time to Failure in File 3 should be looked into.
4. There are some weird patterns in the file. For example The Time to Failure decreases with increase in impurity although it is generally believed that high impurities should cause transmission failure.

Q4: To what extent did data issues limit your ability to develop useful models?

Answer:

Some of the data issues did limit our ability to develop useful models. Few examples are as follows:

The Data 3 file could not be used at all for the final model as the range of the Y Variable (Time to Failure) was very different form that in the other file. If used this file could have biased the final model. Although another model for this file alone can be built. But there is lot of missing values in the file. There are two Die Position Values 1 and 2 but the y variable is absent for all the die positions with values 1.

Similarly in Data 1 file. There was lot of values for some of the variables that were very different from the values that the corresponding variable usually takes. Some of these values were very apparent to be wrong values. But for some of the values a decision had to be taken whether they are wrong values or just outliers. Some of the values for Time to Failure in Data 1 were very high and looked like wrong values. But since the Data 3 file also has high values for Time to Failure variable so there was confusion whether it is a wrong value or outliers. Although in order not to bias the data these values were dropped off.

Q5: Assuming good reliability models can be developed in the future, what are your recommendations as to "closing the loop" and ensuring that these models are effectively utilized and maintained over time?

Answer:

If a good reliability models are developed then in order to ensure that these models are effectively utilized and maintained over time the Six Sigma DMAIC should be used:

Define:
The CTQs that were taken care during the building of the system was Residuals. This should be monitored otherwise serious financial losses can occur and confidence in the system may be lost.

Measure and Analysis:
If some statistical evidence is know that there is some shift in any of the indicator metrics, and then audit should be raised for the system to know whether this shift is real or there is a problem with the reliability models. If the shift is due to the reliability models then maintenance staff should be notified.

Improve and Control:
The necessary changes to the reliability models should be done in this phase based on the changed context and rules. Also these changes to the models are tested. After the successful testing new reliability model should be used in the business process. The reliability models should be developed in such a way that it can be changed and tuned easily if needed.

3. DISCUSSION

This assignment helped in understanding the development of an Automated Decision System (ADS) and the application of Six Sigma principles namely DFSS (Design for Six Sigma) and DMAIC (Design, Measure, Analyze, Improve and Control) throughout its lifecycle. Some of the design trade-offs that came into play while designing the decision system were:

1. Accuracy versus Coverage
2. Accuracy versus Interpretability
3. Run-time efficiency versus configurable architecture.

Some of the major data quality issues were addressed and solutions were recommended:

Issues:

1. The four enclosed data files did not contain the same set of variables.
2. The same variables were named differently.
3. Some places missing value is left blank and at others it is denoted by "**#**".
4. In the third data file there are two Die Position values- 1 and 2. But the Time to Failure values is absent for all the rows where the Die Position value is 1.
5. Some of the values in the data files had different values from the usual values that the variable takes.

Recommendation:

1. The different names of the same variable can be figured out either by looking at the name itself like Torque and NM can be the same variable as the unit for Torque is NM. Additional validation can be done by plotting the two variables.
2. The unusual looking values should be dropped off as they can bias the model. For some of these values it would be very apparent that they are incorrect but for some values a decision will have to be taken whether it is a wrong value or it is due to some unusual operation of the machine.

Company's current approach to managing the data on transmission reliability is leading to a lot of data quality issues. Recommendation as to how this can be handled better in the future:

1. A common Convention for naming the variables should be followed.
2. Some checks on the range of values of the variables should be done before entering into the file.
3. Problems like very different values of a variable like Time to Failure in File 3 should be looked into.
4. There are some weird patterns in the file. For example The Time to Failure decreases with increase in impurity although it is generally believed that high impurities should cause transmission failure.