Masters of Engineering in Engineering Management

Data Analytics 2023-2024
(AL_ENGMF_9_1)

Continuous Assessment 1

February 2024

Student Name: Abhishek Kumar
Student Number: A00268732

Table of Contents

1. Business Understanding

1.1 Business Background

The local start-up team, including a software engineer and a manufacturing engineer, is beginning on a lucrative business opportunity to transform predictive maintenance procedures within the manufacturing business. The team views the project of modelling and categorising machine failure events according to current operating conditions as a promising commercial opportunity with a high potential for success.

The organisational structure demonstrates a dynamic and interdisciplinary approach, where the software engineer contributes skills in programming languages, while the manufacturing engineer provides an in-depth understanding of machine operations and failure modes. This collaborative synergy sets the foundation for creativity and problem-solving in predictive maintenance solutions.

Central to the team's objectives is the implementation of predictive analytics models on real-world data, enabling proactive detection of machine failure problems before they escalate into costly operational disruptions. Leveraging synthetic training data, the team strives to replicate varied failure scenarios and optimise predictive algorithms to enhance model accuracy and reliability.

The predominance of machines running without failure 95-99% of the time emphasises the vital requirement for powerful predictive analytics skills to effectively detect failure events throughout the remaining operational window. This urgency is emphasised by the high cost of downtime, underscoring the strategic relevance of building comprehensive predictive maintenance solutions capable of reducing operational disturbances and maximizing productivity.

Existing approaches to machine failure detection may lack the precision and speed required to meet emerging operational difficulties successfully. The start-up team intends to surpass existing methods by combining data-driven insights and predictive modelling tools to foresee and pre-emptively manage failure scenarios.

Collaboration between the software engineer and manufacturing engineer develops cross-functional synergy, assuring alignment with business objectives and operational reality. The team is poised to drive innovation and provide disruptive solutions that change predictive maintenance standards within the manufacturing sector.

1.2 Business Objective

- Predictive Modelling: The team aims to create strong predictive analytics models that can accurately differentiate between machine failure occurrences (fail = TRUE) and normal operating situations (fail = FALSE). The team tries to understand the complex connections between machine operational parameters and failure outcomes using synthetic training data and real-world observations.

- Business Viability: The successful modelling of machine failures is a tempting economic prospect due to the machinery's importance in many production processes. The team sees their predictive maintenance solutions as a key element of operational efficiency and cost-effectiveness, leading to significant returns on investment for industry stakeholders.

1.3 Success Criterion

The success criterion for the data mining project entails achieving a high predictive accuracy rate, reducing downtime and maintenance costs, improving Overall Equipment Effectiveness (OEE), enhancing predictive maintenance capabilities, and generating tangible business impact and ROI. Each criterion directly correlates with the specified business objectives, ensuring alignment with the project's strategic goals. Success will be measured by the project's ability to reliably classify machine failure events, optimize resource allocation, maximize equipment uptime, and deliver quantifiable cost savings and operational efficiencies. Continuous monitoring and evaluation will gauge the project's effectiveness in driving sustained business success within the manufacturing domain.

1.4 Estimation on Resources

- **Workforce:** Software Engineer is responsible for programming and agile methodologies. Manufacturing Engineer provides domain experience and insights into machine operations and failure modes.

- **Data**: Synthetic training data is available for simulating machinery operations and breakdown conditions.

- **Tools and Software**: Access to Data Analytics Platforms such as R language, Python, or specialist tools for predictive modelling and data analysis. Access to computational infrastructure for model training and validation.

- **Financial Support**: Allocation of funding for project implementation, including staff charges, software licencing, and data collecting fees.

- **Collaborative Networks**: Potential collaboration with external specialists or research institutions for additional insights and knowledge.

- **Infrastructure**: Access to data storage and management solutions for organizing and processing massive datasets.

- **Training and Development**: Opportunities for people training and skill development in data analytics and predictive modelling methodologies.

- **Project Management Support**: Provision of project management resources to oversee project milestones, schedules, and deliverables.

By leveraging these resources effectively, the project team may create and execute comprehensive predictive analytics solutions to address machine failure concerns inside the manufacturing environment. Continuous improvement and refining of resources will ensure the project's success in attaining its objectives and delivering tangible value to the start-up.

1.5 Risk Identification & Contingency Measure

- **Data Quality Issues**

Risk: Synthetic training data may not correctly reflect real-world settings, leading to model inaccurate.
Contingency: Implement data validation techniques to identify abnormalities and discrepancies. Augment synthetic data with real-world observations for greater model generalization.

- **Model Overfitting**

Risk: Models may overfit training data, resulting in poor generalization to unseen data.
Contingency: Apply regularisation techniques and cross-validation approaches to prevent overfitting. Evaluate model performance on validation datasets to ensure robustness.

- **Resource Constraints**

Risk: Limited computational resources may impair model training and validation operations.
Contingency: Optimize algorithms and employ cloud-based computing resources to scale computational power as needed. Prioritize tasks and distribute resources efficiently to enhance productivity.

- **Business Objective Misalignment**

Risk: Project outcomes may not fit with stakeholders' expectations or company objectives.
Contingency: Conduct regular stakeholder meetings to establish alignment between project objectives and company goals. Adjust project scope and deliverables as necessary to accommodate shifting business needs.

- **Integration Challenges**

Risk: Integrating predictive analytics models into existing operational procedures may provide technical obstacles.

Contingency: Collaborate closely with IT and operational departments to expedite integration processes. Develop user-friendly interfaces and documentation to promote seamless deployment of predictive models.

- **Model Interpretability**

Risk: Complex predictive models may lack interpretability, impeding stakeholders' comprehension and acceptability.
Contingency: Use interpretable model structures and visualization tools to explain model predictions. Provide training sessions and documentation to enhance stakeholders' knowledge of model outputs.

- **Data Privacy and Security**

Risk: Handling sensitive data may expose the firm to privacy breaches or security concerns.
Contingency: Implement effective data encryption and access control systems to preserve critical information. Comply with relevant data privacy legislation and standards to mitigate legal and reputational concerns.

- **Lack of Expertise**

Risk: Limited competence in data analytics and predictive modelling may delay project progress.
Contingency: Invest in training programs and external consultants to complement team skills. Foster a culture of knowledge sharing and collaboration to exploit varied skill sets efficiently.

By proactively identifying risks and executing contingency measures, the project team can prevent potential interruptions and boost the project's likelihood of success in delivering useful insights and solutions for predictive maintenance within manufacturing.

1.6 Data Mining Goals/Success Criterion

The main objective of data mining is to develop predictive analytics models that accurately and reliably distinguish between instances of machine failure (fail = TRUE) and normal operational conditions (fail = FALSE).

- **Evaluation of Model:**

Attain a high level of predicted accuracy, precision, and recall when categorising machine failure events using real-time operating information.
Develop models that accurately represent and analyse the intricate relationships between machine operational variables and failure results, utilising both synthetic training data and actual observations.

- **Business Impacts:**

Show concrete business benefits by implementing predictive maintenance solutions to save downtime, reduce operating disturbances, and optimise maintenance plans.
Assess the economic feasibility and potential ROI of accurately predicting machine breakdowns through predictive maintenance, highlighting its role in enhancing operational efficiency and cost-effectiveness for industry participants.

- **Validation and Generalisation:**

Validate generated models utilising robust evaluation approaches, assuring their effectiveness across varied operating scenarios and datasets.
Ensure model generalizability by testing against real-world observations and incorporating feedback from maintenance personnel to boost prediction accuracy and flexibility.

- **Integration and Deployment:**

Integrate validated models seamlessly into existing operational workflows and real-time monitoring systems, providing quick detection and response to probable machine failure scenarios.
Facilitate smooth deployment and user uptake through user-friendly interfaces and documentation, ensuring accessibility and usability across all stakeholder groups.

The project aims to provide actionable insights and transformative solutions in data mining to drive operational excellence, enhance business viability, and establish predictive maintenance as a key factor in efficiency and profitability in the manufacturing sector.

1.7 Tools for the Project and Resources

- **R Programming Language**: RStudio provides a comprehensive integrated development environment (IDE) for R programming, facilitating coding, debugging, and data analysis tasks.

- **Tidyverse Packages**: Building upon desctools, the Tidyverse suite (containing packages like dplyr and tidyr) provides robust tools for data manipulation and transformation, ensuring data is prepared properly for modelling.

- **Data Visualization Tools**:
  ggplot2: Developing insightful and customisable visualizations, allowing for analysis of correlations between machine operational parameters and failure consequences.
  plotly: Offers interactive plots and dashboards, enabling the presentation and exploration of complicated data patterns.

- **Machine Learning Libraries**:
  randomForest: Effective for generating ensemble learning models, for forecasting machine faults.
  glmnet: Provides tools for fitting generalized linear models with Lasso or Elastic-Net regularization, suited for handling high-dimensional data and feature selection.

- **Model Evaluation and Validation**:
  ROCR: Facilitates the evaluation of classification models with measures such as ROC curves and AUC scores.
  caret: Offers easy utilities for cross-validation, model selection, and performance assessment across different algorithms.

- **Cloud Computing Resources**: Consider employing cloud computing platforms like Amazon Web Services (AWS) for scalable computing resources and storage capabilities.

- **Version Control and Collaboration**: git integrated with RStudio enables version control and collaborative development, allowing team members to track changes, manage project iterations, and coordinate work effectively.

- **Documentation and Reporting**:
  rmarkdown: Provides a versatile framework for building reproducible reports and presentations, integrating R code, graphics, and narrative prose seamlessly.
  knitr: Works with rmarkdown to provide dynamic and interactive reports, boosting the communication of project findings and insights.

2. Data Understanding

2.1 Data Format and Data Quantity

- **Data Structure**: str(data) = The dataset is a data frame consisting of 10,000 observations (rows) and 6 variables (columns).

- **Variable Data Types**:
  - Air_temperature_.K., Process_temperature_.K., and Torque_.Nm. are numeric variables.
  - Rotational_speed_.rpm. and Tool_wear_.min. are integer variables.
  - Machine_failure is a logical (Boolean) variable signifying machine failure (TRUE) or regular operation (FALSE).

- **Variable Descriptions**:
  - Air_temperature_.K.: Represents the air temperature in Kelvin.
  - Process_temperature_.K.: Represents the process temperature in Kelvin.
  - Rotational_speed_.rpm.: Represents the rotational speed in revolutions per minute.

- Torque_.Nm.: Represents the torque in Newton-meters.
- Tool_wear_.min.: Represents the period of tool wear in minutes.
- Machine_failure: Indicates whether a machine failure happened (TRUE) or not (FALSE).

- **Data Integrity**: No missing values are obvious in the output, as shown by the absence of any 'NA' values in the data frame's structure.

- **Data Quantity**: The dataset's quantity of 10,000 rows and 6 columns provides a solid foundation for conducting comprehensive data analysis and modelling efforts aimed at understanding and predicting machine failure events within the manufacturing environment.

The **dataset appears to satisfy with business requirements** for predicting machine failures based on operational conditions. With 10,000 observations, it delivers sufficient data volume for modelling. Variables like as temperature, speed, torque, and tool wear provide significant insights into potential failure modes. The inclusion of the "Machine_failure" variable enables the replication of failure events. However, further examination of data quality and pre-processing is necessary to ensure reliability. Overall, the dataset is a great resource for building predictive maintenance solutions to boost operational efficiency and minimize downtime in manufacturing operations.

2.2 Interpretation and Exploratory Data Analysis

1. summary(data):
Range of Values:
For variables like Air_temperature_.K., Process_temperature_.K., Rotational_speed_.rpm., and Torque_.Nm., we see a broad range of values from the minimum to the maximum, showing variability in the measurements across different observations.
Air_temperature_.K. and Process_temperature_.K. range from about 295.3 to 304.5 Kelvin, while Rotational_speed_.rpm. ranges from 1168 to 2915 RPM, and Torque_.Nm. ranges from 3.20 to 76.60 Newton meters.

Central Tendency:
The mean values for Air_temperature_.K., Process_temperature_.K., Rotational_speed_.rpm., and Torque_.Nm. are generally close to the median values, indicating symmetric distributions with modest skewness.
The mean values for these variables provide an approximation of the central tendency or average value of the measurements.

Tool Wear Duration:
The Tool_wear_.min. variable spans from 0 to 253 minutes, with a median value of about 108 minutes. This shows that the period of tool wear varies greatly among observations, with some instances of protracted tool usage.

Machine Failure Incidence:
The Machine_failure variable is expressed by logical values (TRUE or FALSE), indicating if a machine failure occurred.
The summary reveals that out of 10,000 data, 339 cases are recognised as machine failures (TRUE), whereas the majority (9661 observations) indicate regular operation (FALSE).

Data Distribution:
The quartiles (1st, 2nd, and 3rd quartiles) provide insights into the distribution of values within each variable, helping to examine the dispersion and variability of the data.

2. **desc(data)**: see Appendix A

The blue part represents instances where machine failure is marked as FALSE, totalling 9,661 observations, which accounts for 96.6% of the dataset. The red part denotes instances where machine failure is marked as TRUE, with a frequency of 339 observations, equivalent to 3.4% of the dataset.

Confidence Intervals (CI): The 95% confidence intervals are calculated using the Wilson method. The lower bound (lci.95) for the proportion of FALSE instances lies at 96.2%, and the upper bound (uci.95) is at 96.9%. For TRUE instances, the lower bound is at 3.1%, and the upper bound is at 3.8%.

the range from 0.0 to 1.0 represents the counts of observations or instances within the dataset. This range indicates the proportion of occurrences of a particular category relative to the total number of observations.

3. ok_operation_row_indices <- which(!data$Machine_failure)

The ok_operation_row_indices variable is created using the which(!data$Machine_failure) expression. This expression selects the indices of rows where the Machine_failure column is False (i.e., successful operations). The ! operator negates the condition, so it essentially selects rows where Machine_failure is False.

4. fail_row_indices <- which(data$Machine_failure)

The fail_row_indices variable is created using the which(data$Machine_failure) expression. This expression selects the indices of rows where the Machine_failure column is True (i.e., failed operations).

Data partitioned and Suggestion:

Ok Operation Indices: Rows when machine operation is successful.
Failures Indices: Rows where machine failure occurred.

Here are suggestions:

Comparing Variables: Study how variables such as temperature, speed, torque, and tool wear differ between cases of machine failure and successful operation.
Identifying Patterns: Differences in distributions may suggest patterns or trends linked with machine failure.
Prediction Insights: Understanding these differences helps inform prediction algorithms. Variables with significant changes between failure and non-failure instances could be valuable indicators for future failure identification.
Maintenance Strategies: Insights gathered from these comparisons can aid in creating maintenance strategies.

5. summary(data[ok_operation_row_indices,])

It allows us to analyse the characteristics and distributions of variables specifically for instances where the machine operation was successful.

6. summary(data[fail_row_indices,])

It allows us to analyse the characteristics and distributions of variables specifically for instances where machine failure occurred.

7. **Desc(data[ok_operation_row_indices,])** : see Appendix B

It describes the partition of the data when the machine action was normal (FALSE). The blue part represents instances where machine failure is marked as FALSE, totalling 9,661 observations, which accounts for 100% of the partitioned dataset.

- Length: The data has 9,661 observations.
- n: All 9,661 observations are non-missing.
- NAs: There are no missing values in this data.

- Unique: There is just one unique value (FALSE) in the data, indicating that all observations represent regular machine operation.
- Frequency (freq): All 9,661 observations have the value FALSE, reflecting 100% of the data.
- Percentage (perc): 100% of the observations in the data represent regular machine operation.
- 95% Confidence Interval (95%-CI): The Wilson approach is used to determine the confidence interval, but in this case, it results in 100% for the lower and upper confidence intervals (lci.95 and uci.95), showing that the normal operation occurs with certainty within the dataset.

8. **Desc(data[fail_row_indices,])**: see Appendix C

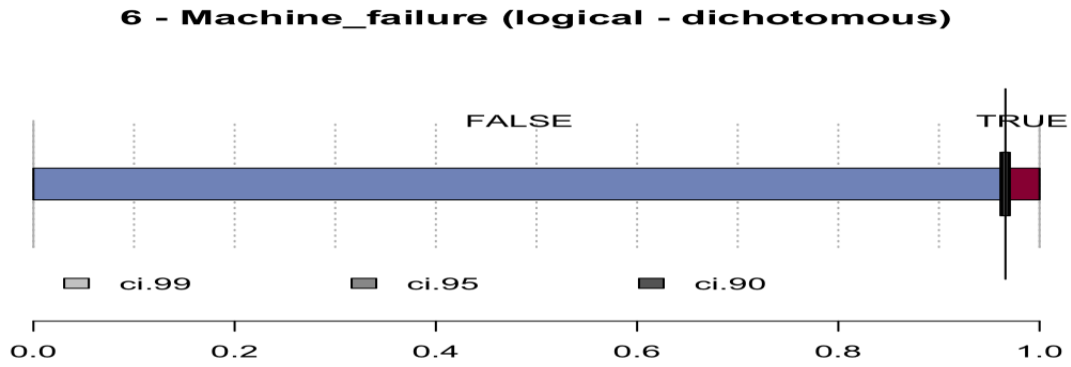
It describes the partitioned dataset where machine failure occurred, represented by the logical value TRUE. The blue part represents instances where machine failure is marked as TRUE, totalling 339 observations, which accounts for 100% of the partitioned dataset.

- Length: The data contains 339 observations where machine failure occurred.
- n: All 339 observations are non-missing.
- NAs: There are no missing values in this data.
- Unique: There is just one unique value (TRUE) in the data, suggesting that all observations represent machine failure.
- Frequency (freq): All 339 observations have the value TRUE, signifying 100% of the subgroup.
- Percentage (perc): 100% of the observations in the data represent machine failure.
- 95% Confidence Interval (95%-CI): The Wilson method is used to compute the confidence interval.

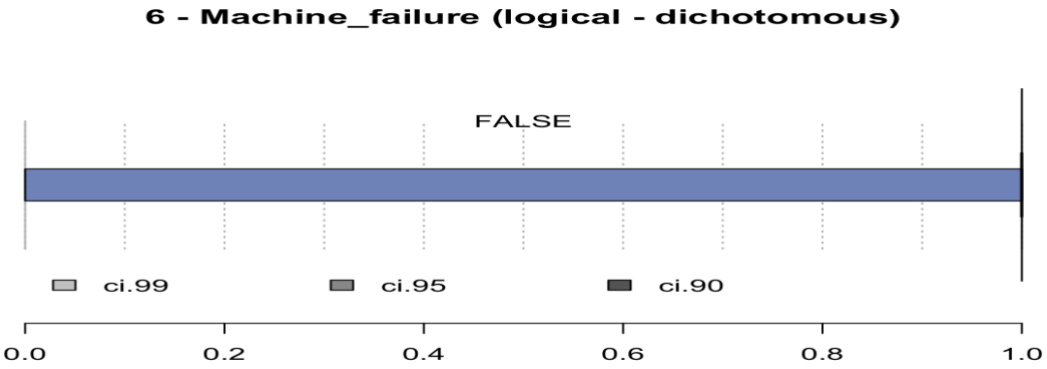
In this image, the lower confidence interval (lci.95) is 98.9%, and the upper confidence interval (uci.95) is 100.0%, suggesting with high certainty that machine failure is present within the dataset.

Appendix

Appendix A



Appendix B



Appendix C