**Module 7: Capstone Project Final Report**

Michael Stephens

Colorado State University Global

MIS581 – Capstone: Business Intelligence and Data Analytics

Professor Morad

August 2, 2022

**Table of Contents**

Abstract3

Introduction/Background4

Problem/Purpose Statement5

Objectives5

Overview of Study6

Variables7

Preliminary Data Exploration9

Research Question and Hypotheses10

Literature Review11

Research Design14

Methodology14

Methods14

Limitations16

Ethical Considerations16

Findings17

Descriptive Analysis17

Linear Regression Model20

Logistic Regression Model22

Conclusion24

Recommendations25

References26

**Abstract**

Unplanned stoppages caused by machine breakdowns are a significant problem in the manufacturing industry. Traditional maintenance strategies in which maintenance activities are planned at fixed intervals tend to be inefficient and costly. Recent technological developments support the adoption of predictive maintenance (PdM) in factories. Machine failures rarely occur instantaneously; some level of degradation or change in condition occurs as equipment shifts from operational to non-operational. In recent years, intelligent sensors have become a cost-effective option for monitoring the state of industrial equipment, including speed, temperature, and pressure. This paper proposes the design of two machine learning models to predict remaining useful life (RUL) and classify equipment nearing failure using condition-based sensor readings. The two methods chosen for this research study are linear and logistic regression.

The models in this project were developed using the NASA C-MAPSS dataset, which contains sensor readings from 100 turbofan engines. Both models show a statistically significant relationship between condition-based sensor readings and turbofan engine health. Predictions of remaining useful life (measured in cycles) from the linear regression model had an MAE of 34.1 and RMSE of 43.95. The logistic regression model was 97.55% accurate in classifying healthy engines (over 20 cycles from failure) and unhealthy engines (within 20 cycles of failure). These models show promise for guiding predictive maintenance decisions in manufacturing facilities. Data-driven decision-making will lead to fewer breakdowns, as maintenance is proactively determined by the real-time condition of the equipment rather than a predetermined schedule.

**Module 7: Capstone Project Final Report**

The manufacturing industry is rapidly evolving with new innovative technologies and process automation. However, equipment downtime remains a significant issue among manufacturers. Factories are estimated to lose 5% to 20% of productivity from unplanned machine breakdowns (Kane et al., 2022). When equipment fails unexpectedly, production is interrupted until maintenance is carried out to return the system to an operational state. These unplanned stoppages have many undesirable effects, including higher maintenance costs, lost output, late deliveries, wasted materials, and safety risks. The annual cost of repair and maintenance is $55.2 billion in the United States (Census Bureau, 2021).

It is the role of the maintenance team to prevent breakdowns by detecting possible problems with the system before a failure occurs. Unfortunately, some threats can be very difficult to identify through routine inspections without specialized tools (Kiangala & Wang, 2018). Modern technology enables the condition of equipment to be checked with minimal intrusion. In today’s manufacturing environment, various cost-effective sensors are available to monitor and record data concerning critical pieces of equipment (Pech et al., 2021). These sensors are capable of providing real-time measurements from the machine, such as temperature, force, vibration, pressure, flow, and noise (Daniyan et al., 2022). Recent research suggests that machine learning models trained with time-series sensor data may be effective at predicting equipment failure in production plants (Leukel et al., 2021).

According to Marr (2015), successful companies do not make decisions based on opinion but instead use data-driven facts. Predictive maintenance (PdM), also known as condition-based maintenance, is a concept that uses data analytics to manage the care of machinery based on future predictions. PdM is focused on detecting signs of machine failure to guide maintenance tasks for maximizing equipment service life and reliability. This research paper will study the significance of current and historical machine data in forecasting future equipment failure and its potential for minimizing or eliminating unplanned downtime in the factory by identifying unhealthy equipment before it fails.

**Problem/Purpose Statement**

Develop meaningful predictive models that classify industrial equipment as healthy or unhealthy and predict remaining useful life (RUL) using machine sensor readings as independent variables. These models will be designed to assist a manufacturing company’s maintenance decisions, reduce production costs, and improve operational efficiency. This research project aims to demonstrate the value of data and machine learning in the manufacturing industry.

**Objectives**

This paper will present and discuss two machine learning models. The first model is a classification model that will identify when equipment is nearing failure based on sensor readings. The second model is a prediction model that uses the same sensor readings to estimate a machine’s remaining useful life (RUL) or the time before expected failure. These predictive models are to be used by maintenance technicians to identify, investigate, and prevent machine failures in the factory. This project aims to minimize maintenance costs and maximize net profit for the manufacturer. The following goals and objectives will be targeted.

1. Review literature regarding predictive maintenance and discuss past findings on the subject. Provide an overview of the predictive maintenance process, benefits, and opportunities.
2. Collect historical machine data from sensors that measure conditional factors such as temperature, speed, and pressure. Prepare and clean a dataset for modeling and analysis.
3. Discuss the relevant considerations for the safety, privacy, and security of the data stored from factory sensors.
4. Conduct correlation analysis and present descriptive statistics to identify which sensors may have predictive value in forecasting machine health.
5. Using machine learning and inferential statistics, determine which sensors (if any) are statistically significant in predicting remaining useful life. Predictor variables with a p-value less than the significance level of 5% will be considered significant.
   1. Which current conditions from sensor readings are statistically significant in identifying unhealthy equipment?
   2. Which current conditions from sensor readings are statistically significant in predicting the remaining useful life (RUL) of equipment?
6. Develop a classification model that reliably classifies a machine as “healthy” or “unhealthy” with an accuracy greater than 75%. The definition of “unhealthy” is equipment nearing failure and will be further refined after exploratory data analysis.
7. Develop a predictive model that reliably forecasts remain useful life of equipment. Model performance will be measured using root mean squared error (RMSE).

**Overview of Study**

Released by NASA in 2008 for predictive maintenance research, the C-MAPSS FD001 dataset contains simulated data for 100 gas turbine aircraft engines. C-MAPSS is an acronym for Commercial Modular Aero-Propulsion System Simulation and is a program for generating realistic large commercial turbofan engine data (Saxena et al., 2008). According to researchers at NASA, the FD001 dataset includes a time series of system responses from 21 machine sensors under three operational settings (Saxena et al., 2008). Each observation in the dataset represents a snapshot taken during one engine cycle (flight), and each machine begins with a different level of initial wear. Observations were recorded from the collection time until engine failure. This dataset was chosen for its potential to predict remaining useful life (RUL) and diagnose unhealthy equipment using real-world sensor readings. The early detection of an aircraft engine fault is paramount for safety and reliability.

**Variables**

The C-MAPSS FD001 dataset consists of 28 unique variables. To understand these variables, it is important to discuss the basic components of a turbofan engine. As seen in Figure 1, a turbofan engine has five rotating parts – an intake fan, low-pressure compressor (LPC), high-pressure compressor (HPC), low-pressure turbine (LPT), and high-pressure turbine (HPT). The sensor readings in the dataset provide system status information from these components, such as temperature, speed, and pressure during each flight. Patterns in the time-series data from these sensors may be predictive of engine failure.

**Figure 1**

*Simple Diagram of Turbofan Engine*

Chart, diagram

Description automatically generated

*Note*. This figure shows a simple diagram of a turbofan engine. From “User’s Guide for the Commercial Modular Aero-Propulsion System Simulation (C-MAPSS),” by J. S. Litt, D. K. Frederick, and J. A. DeCastro, 2007, p. 9 (https://ntrs.nasa.gov/citations/20070034949).

The data dictionary in Table 1 shows the name, description, unit, and quantitative data type of each variable in the C-MAPSS dataset (Saxena et al., 2008). This dataset contains no qualitative variables. The full dataset is available for download on the Kaggle website at https://www.kaggle.com/datasets/behrad3d/nasa-cmaps.

**Table 1**

*C-MAPSS Data Dictionary*

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Variable Name** | **Description** | **Data Type** |
| 1 | Engine\_ID | Unique engine identifier | ID |
| 2 | Cycle | Number of flights | Discrete |
| 3 | OpSetting1 | Undisclosed flight operational setting | Interval |
| 4 | OpSetting2 | Undisclosed flight operational setting | Interval |
| 5 | OpSetting3 | Undisclosed flight operational setting | Interval |
| 6 | T2 | Total temperature at fan inlet (°R) | Ratio |
| 7 | T24 | Total temperature at LPC outlet (°R) | Ratio |
| 8 | T30 | Total temperature at HPC outlet (°R) | Ratio |
| 9 | T50 | Total temperature at LPT outlet (°R) | Ratio |
| 10 | P2 | Pressure at fan inlet (psia) | Ratio |
| 11 | P15 | Total pressure in bypass-duct (psia) | Ratio |
| 12 | P30 | Total pressure at HPC outlet (psia) | Ratio |
| 13 | Nf | Physical fan speed (rpm) | Ratio |
| 14 | Nc | Physical core speed (rpm) | Ratio |
| 15 | epr | Engine pressure ratio (discharge pressure ÷ P2) | Ratio |
| 16 | Ps30 | Static pressure at HPC outlet (psia) | Ratio |
| 17 | Phi | Ratio of fuel flow to Ps30 | Ratio |
| 18 | NRf | Corrected fan speed | Ratio |
| 19 | NRc | Corrected core speed | Ratio |
| 20 | BPR | Bypass ratio | Ratio |
| 21 | farB | Burner fuel-air ratio | Ratio |
| 22 | htBleed | Bleed enthalpy | Ratio |
| 23 | Nf\_dmd | Demanded fan speed | Ratio |
| 24 | PCNfR\_dmd | Demanded corrected fan speed | Ratio |
| 25 | W31 | HPT coolant bleed (lbm/s) | Ratio |
| 26 | W32 | LPT coolant bleed (lbm/s) | Ratio |
| 27 | RUL | Remaining useful life of engine (in cycles) *\*calculated variable (Max cycles – Cycles)* | Ratio (Target) |
| 28 | Near\_Failure | Indicates whether an engine is within 20 cycles of failure *\*calculated variable (if RUL < 21, then Near\_Failure = 1, else Near\_Falure = 0)* | Binary  (Target) |

This project aims to predict a machine’s remaining useful life (RUL) using real-time sensor readings and classify a machine as a near failure if it is close to breakdown. The NASA turbofan engine dataset is a good fit for these objectives because it contains 21 sensor readings from the beginning of data collection to the failure of the observed equipment. Thus, we can calculate the true RUL of each engine by subtracting the current cycle from the max cycle. Using data analytics and machine learning, we can test which of the 21 sensors (if any) have a statistically significant relationship with RUL. The resulting model can be used to make future predictions of RUL with new sensor data.

Additionally, this dataset provides enough information to create a binary target (Near\_Failure), which identifies observations that are within 20 cycles of failure. This binary target will be utilized to solve a classification problem. Maintenance technicians would be interested in classifying aircraft engines that are near failure (unhealthy), so they can be repaired or decommissioned before a catastrophic event. The concepts and objectives from this study are applicable to manufacturers looking to deploy similar models to predict the failure of factory equipment. An accurate model for classifying equipment nearing failure will ensure the appropriate maintenance tasks are performed in the factory to avoid unplanned downtime.

**Preliminary Data Exploration**

From the summary statistics in Table 2, we can see that six sensors have a standard deviation of 0 and thus do not change throughout the time series (T2, P2, epr, farB, Nf\_dmd, PCNfR\_dmd). These sensors will be removed from the dataset as they do not have any predictive value. Operating settings 1, 2, and 3 will also be dropped from the list of features because they are outside the scope of this study. The summary statistics show that the dataset contains a sample of 20,631 observations, sufficient for hypothesis testing, and no variables are missing data. Further analysis of individual sensors is required to determine their usefulness in predicting RUL and classifying near failure.

**Table 2**

*Summary Statistics (All Variables From C-MAPSS)*

Table

Description automatically generated

**Research Question and Hypotheses**

As shown in Table 3, this project will focus on answering whether a predictive relationship exists between machine sensor readings and equipment health. Hypothesis testing analyzes two mutually exclusive assertions about a population to decide which assertion is best supported by sample data (Davila, 2016). Two different hypotheses will be tested using turbofan engine data from NASA’s C-MAPSS dataset. The first hypothesis focuses on the relationship between sensor readings and an engine’s remaining useful life (RUL). The second hypothesis suggests that sensor readings influence the probability that a machine is within 20 cycles of failure. The predictor (input) candidates are sensors T24, T30, T50, P15, P30, Nf, Nc, Ps30, Phi, NRf, NRc, BPR, htBleed, W31, and W32. The planned output variables are remaining useful life (measured in cycles) and *Near\_Failure* (binary variable where 1 classifies an engine within 20 cycles of failure and 0 classifies an engine as more than 20 cycles from failure).

**Table 3**

*Research Question and Hypotheses*

|  |  |
| --- | --- |
| **Research Question** | |
| Is there a predictive relationship between engine health and component sensor readings? | |
| **Hypothesis 1** | |
| Ho (null) | There is no significant relationship between a turbofan engine’s remaining useful life (RUL) and its sensor readings for T24, T30, T50, P15, P30, Nf, Nc, Ps30, Phi, NRf, NRc, BPR, htBleed, W31, or W32. |
| Ha (alternative) | There is a significant relationship between a turbofan engine’s remaining useful life (RUL) and one or more of its sensor readings for T24, T30, T50, P15, P30, Nf, Nc, Ps30, Phi, NRf, NRc, BPR, htBleed, W31, or W32. |
| **Hypothesis 2** | |
| Ho (null) | None of the sensor predictors (T24, T30, T50, P15, P30, Nf, Nc, Ps30, Phi, NRf, NRc, BPR, htBleed, W31, or W32) have a statistically significant relationship with the binary response variable, *Near\_Failure*. |
| Ha (alternative) | At least one of the sensor predictors (T24, T30, T50, P15, P30, Nf, Nc, Ps30, Phi, NRf, NRc, BPR, htBleed, W31, or W32) has a statistically significant relationship with the binary response variable, *Near\_Failure*. |

**Literature Review**

There is a growing interest in the application of machine learning for predictive maintenance. The rapid advancement of IoT technology presents manufacturers with a cost-effective opportunity to invest in machine-embedded sensors for data collection and machine condition assessment. Several recent studies have examined the predictive value of intelligent sensor readings in estimating the remaining useful life of industrial equipment using machine learning.

Wu et al. (2017) created a random forest-based prognostic method to predict tool wear in milling operations using measurement data from modern sensors attached to CNC machines. Researchers found that a random forest trained with a sample of sensor data for cutting force, vibration, and acoustic emission could accurately predict the life expectancy of milling tools (Wu et al., 2017). Models trained with sensing technology were significantly more accurate in estimating tool wear than traditional methods that only use process parameters, cutting tool geometry, and material properties. At a training size of 70%, the resulting random forest model returned high accuracy metrics, with an R-squared of .990 and MSE of 10.156. This research is valuable to milling operations looking to optimize their maintenance decisions, such as when to change machine tooling.

Li et al. (2019) studied the degradation of rolling element bearings in rotating machines and how machine learning could be used to predict their remaining useful life. Bearings can cause failure or breakdown of the entire system if degradation is left unattended. Using a time series of vibration signals, researchers developed an exponential regression model to predict the total hours until bearing failure. The fitted exponential regression returned an adjusted R-squared of 0.957 and MSE of 14.57 (Li et al., 2019), suggesting that vibration signals have an exponential relationship with run time.

Researchers then improved their prediction results by building an artificial neural network (ANN), which is a supervised machine learning model that captures complex relationships between predictors and target variables. Neural networks are iterative learners, meaning that training records are passed through the network one at a time, and weights are updated to minimize standard error through a process called backpropagation (Abbott, 2014). After training was complete, the neural network’s performance was analyzed using a validation dataset. The performance of the ANN was better than the exponential regression, with a significantly lower MSE of 6.78. Therefore, researchers concluded that the neural network is more accurate at predicting bearing failure than exponential regression (Li et al., 2019).

Vibration signals were also the subject of a recent predictive maintenance study at a bottling factory. Kiangala & Wang (2018) analyzed real-time vibration speed data and its ability to predict early motor threats. Researchers found that motor vibration speed increases until it reaches an unacceptable state and forces unplanned downtime. Based on these results, a new predictive maintenance strategy was proposed to reduce equipment breakdowns. Per the new strategy, motors with vibration signals that reach an unsatisfactory threshold must be immediately scheduled for repair. A cloud-based dashboard report was created for continuous monitoring of vibration severity with automatic notification when new maintenance is required (Kiangala & Wang, 2018).

Liu et al. (2021) sought to deploy predictive maintenance to solve a multi-classification problem. Researchers in this study were interested in developing a model that could accurately classify factory equipment in four states of health – good, watching, warning, and fault. The predicted machine states could then be used to prioritize maintenance decisions and reduce the risk of future machine faults. Machine learning models were trained using operation data from various intelligent sensors installed on manufacturing resources. Researchers developed an improved deep adversarial learning (LSTM-GAN) model with an average classification accuracy of 98.87% (Liu et al., 2021).

Predictive maintenance research is trending towards the development of data-driven techniques for decision models (Bousdekis et al., 2019). Accurate real-time predictions of machine health offer several benefits, including maintenance cost reduction, increased lifespan of fixed assets, and fewer breakdowns interrupting production (Li et al., 2019). Advancements in sensor and big data technologies continue to support improvements to predictive maintenance algorithms. Thus, the manufacturing industry would benefit from further research into sensor-driven predictions of machine failure in order to deliver proactive maintenance recommendations.

**Research Design**

**Methodology**

This study will deploy a quantitative methodology that seeks to test the significance of machine sensors from the NASA turbofan engine dataset in predicting and diagnosing engine health. The C-MAPSS dataset supports quantitative analysis because it contains a large sample of numerical readings from 21 engine sensors with associated values for remaining useful life. From this dataset, we can generate inferential statistics that measure the significance of the relationship between machine sensors and turbofan engine health.

**Methods**

Several tools and techniques will be used to achieve project objectives. Data pre-processing and descriptive analysis will be conducted using Jupyter Notebook, which is an open-source application for creating and sharing data science projects online. Jupyter Notebook supports many programming languages, including Python (McKinney, 2017). Python has several libraries that will be useful for this project – Pandas, NumPy, and Matplotlib. Pandas will be used to create a data frame from the C-MAPSS text file, add variable names, insert target variables, and save the pre-processed data to a CSV file. NumPy will be used to create a correlation matrix to evaluate input variables and their relationship with the target variable RUL. The set of features may be narrowed based on each sensor’s correlation with the target variable. The toolset provided by Matplotlib will be utilized to create a visualization of scatterplots showing the trend in the time series of each predictor variable as the engine goes from healthy to failure. These trends will be helpful in determining which sensors have patterns that may be predictive of engine failure. Data will be split into training and validation sets to ensure an unbiased and objective evaluation of predictive performance.

The first hypothesis will be tested using multiple linear regression, a popular method for modeling the linear relationship between a response variable and multiple predictors. Linear regression provides a test statistic, F, which measures the overall model performance against a model with no independent variables. The F-value and its associated p-value will determine whether a statistically significant relationship exists between at least one engine sensor and remaining useful life. If the p-value from the F-test is less than the significance level (0.05), then we can conclude that enough evidence exists to reject the first null hypothesis. The t-value and p-value for each coefficient will give additional insight into which sensors are individually significant in predicting RUL.

Due to the binary nature of the second target variable, logistic regression will be used to test the second hypothesis. Logistic regression provides a test statistic, chi-square, to evaluate the significance of the relationship between multiple independent variables and a categorical (binary) dependent variable. In this case, we are interested in whether machine sensor readings are significant in classifying healthy turbofan engines (over 20 cycles from failure) versus unhealthy turbofan engines (within 20 cycles of failure). Like the F-test, if the chi-square test returns a p-value less than the significance level (0.05), the second null hypothesis will be rejected, and the predictive model will be deemed significant. A confusion matrix will be created to evaluate the model’s accuracy in classifying engines within 20 cycles of failure.

The necessary linear and logistic regression models will be developed using R, a widely used software program for statistical evaluation and data visualization (Kent State, 2022). Maintenance technicians would be interested in classifying aircraft engines that are near failure (unhealthy), so they can be repaired or decommissioned before a catastrophic event. The concepts and objectives from this study are applicable to manufacturers looking to deploy similar models to predict the failure of factory equipment. An accurate model for classifying equipment nearing failure will ensure the appropriate maintenance tasks are performed in the factory to avoid unplanned downtime.

**Limitations**

Ideally, a piece of factory equipment would be chosen for study. However, due to the proprietary nature of industrial machine sensor data, this analysis is limited to observations from 100 turbofan engines in the C-MAPSS dataset. Research of predictive models for the maintenance of aircraft engines will still provide value to manufacturers in determining the best strategy for model development and application.

This project has an 8-week time constraint which must be considered when defining its scope. Therefore, the decision was made to limit the research to two types of predictive models – logistic and linear regression. Alternative machine learning models may produce more accurate results and will be discussed for potential future research. Additionally, sensor readings are susceptible to noise and interference, and input variables may benefit from transformation techniques prior to modeling.

**Ethical Considerations**

Although machine data does not contain personal information, it may be proprietary to a business or organization. For this reason, the data used in this project will not be shared without express permission from the owners. In this case, NASA has released the C-MAPSS dataset for public use and research. No personally identifiable information will be shared without consent. If necessary, the dataset will be de-identified to avoid any conflict with sharing proprietary information. De-identification is the process of removing or masking variables so that the remaining data is anonymous.

The C-MAPSS dataset will only be used for the purpose and methods stated in this paper. This data will not be used for any activity outside this project’s defined scope, nor will it be unlawfully shared with third parties. The results of this proposed study are for the sole purpose of researching predictive maintenance and the value of real-time sensors in predicting machine failure. This study will be fully transparent about the source of data being used, the process for designing the algorithms for prediction, and the hypothesis test results. All parties involved will be treated with honesty and respect.

**Findings**

**Descriptive Analysis**

First, descriptive statistics were generated to review the basic features of the C-MAPSS dataset and its variables. Histograms are displayed in Figure 2, which visualize the distributions of sensor readings. The sensor for bypass-duct pressure, P15, appears to contain only two unique values with a near zero standard deviation. Figure 3 shows a matrix of scatterplots, visualizing each sensor’s relationship with RUL. Temperature sensor readings (T24, T30, T50), rotational speeds (Nf, NRf, Nc, NRc), static pressure (Ps30), bypass ratio (BPR), and bleed enthalpy (htBleed) appear to increase as remaining useful life decreases. Total pressure (P30), fuel-pressure ratio (Phi), and coolant bleeds (W31, W32) show a decreasing trend as useful life approaches zero.

From the correlation matrix in Figure 4, we can evaluate the strength and direction of the linear relationship between variables. The correlation matrix returns Pearson’s correlation coefficients, which range between -1 and 1. A coefficient of +1 indicates a perfect positive linear correlation, and -1 represents a perfect negative linear correlation. Zero indicates no correlation. Nine sensors (T24, T50, P30, Ps30, Phi, BPR, hrBleed, W31, W32) have a strong correlation with RUL (coefficient > 0.6). Five sensors (T30, Nf, Nc, NRf, NRc) have a moderate correlation with RUL (coefficient between 0.3 and 0.6). One sensor, P15, weakly correlates with RUL (coefficient < 0.3). Sensor P15 will be removed from the list of features as it does not appear to provide enough information for prediction.

**Figure 2**

*Distributions of Sensor Variables*

Diagram

Description automatically generated with low confidence

**Figure 3**

*Matrix of Scatterplots (RUL vs. Sensor Readings)*

Polygon

Description automatically generated with medium confidence

**Figure 4**

*C-MAPSS Correlation Matrix*

Chart, timeline

Description automatically generated

**Linear Regression Model**

Before running a linear regression, data was split randomly into a training (80%) and validation (20%) dataset. Fourteen predictors (T24, T30, T50, P30, Nf, Nc, Ps30, Phi, NRf, NRc, BPR, htBleed, W31, and W32) were selected as continuous variables to include in the model, with RUL as the target variable. A summary of the regression model built with the training dataset is shown in Figure 5. This model returned an F-value of 1,610. The p-value associated with this F-value is less than 0.0001, indicating that the model is statistically significant. Thus, the null hypothesis that no relationship exists between sensor readings and RUL is rejected. The significance of each independent (predictor) variable can be evaluated using the t-values and p-values listed in the coefficients table. Sensors with statistically significant coefficients (p < 0.05) are marked with asterisks.

**Figure 5**

*Linear Regression Results*

Text

Description automatically generated

R-squared, also known as the coefficient of determination, quantitatively measures a regression model’s goodness of fit. As seen in Figure 5, the linear regression returned an R-squared value of 0.5776, which means approximately 57.76% of the total variation in engine life is collectively explained by the model’s independent sensor variables. In Figure 6, a bar chart is displayed, showing which sensors contribute most to the prediction of RUL. The variables with the highest relative importance are Ps30, T50, and Phi.

**Figure 6**

*Bar Chart of Relative Importance of the Inputs Features*

Chart, bar chart

Description automatically generated

With the linear model coefficients, predictions can be made on new data. Predictions from the validation data returned an RMSE of 43.95 and MAE of 34.1, as seen in Figure 7. RMSE represents the square root of the average squared differences between predictions and actual observations, while MAE is the average absolute error in the prediction of RUL. These measures can be used to compare performance against other machine learning models.

**Figure 7**

*RMSE and MAE Calculations*

Graphical user interface, text, application

Description automatically generated

**Logistic Regression Model**

As previously discussed, a logistic regression model was developed to classify engines as healthy or unhealthy. The logistic regression, built using the training dataset, is displayed in Figure 8. The test statistic, chi-square, returned a value of 8,831.8 and a p-value of < 0.0001. These test results provide sufficient evidence to reject the second null hypothesis that no relationship exists between sensor readings and the binary target variable, *Near\_Failure*. Thus, the coefficients indicated by the regression model did not occur purely through chance. At least one or more of the chosen predictors influence the probability of engine failure. Like the linear regression, sensor coefficients with p-values less than 0.05 are considered individually significant.

**Figure 8**

*Logistic Regression Results*

Text

Description automatically generated with medium confidence

Text, letter

Description automatically generated

Logistic regression coefficients form an equation to calculate a new observation’s log(odds) or propensity for belonging to the unhealthy engine class. The accuracy of the logistic regression model’s predictions can be evaluated using a confusion matrix, as seen in Figure 9. Sensor values from the validation dataset were used to classify engines as healthy (0) or unhealthy (1), then compared with the actual values stored in the target variable, *Near\_Failure*. The model has an overall accuracy of 97.55%, a sensitivity of 85.10%, and a specificity of 98.95%. These results exceed the initial objective of developing a predictive model with an accuracy of at least 75%. Maintenance technicians can use this model to diagnose unhealthy engines from sensor readings.

**Figure 9**

*Confusion Matrix*

Text

Description automatically generated

The lift chart in Figure 10 gives a visual representation of how the logistic regression model performs against a naïve model. Performance appears to be quite good, with a large area between the lift curve and the baseline. The higher the lift, the more efficiently the model identifies positive class members (unhealthy engines).

**Figure 10**

*Lift Chart*

Chart, line chart

Description automatically generated

**Conclusion**

The evidence from the test results supports the theory that sensor readings for temperature, pressure, and rotational speed are predictive of remaining useful life and near failure of turbofan engines. The linear and logistic regression models were statistically significant, with p-values less than 0.001. Both models show potential for guiding predictive maintenance decisions. The resulting linear regression model can be used to forecast the remaining useful life of turbofan engines using real-time sensor readings. Alternatively, the logistic regression model can be used to identify machines within 20 cycles of failure. Predictions of RUL yielded a mean absolute error of 34.1 and root mean squared error of 43.95, while *Near\_Failure* classifications were 97.55% accurate. Manufacturers can apply these techniques by collecting sensor data to build predictive models for their own in-house equipment.

**Recommendations**

Modifying the selection of predictor variables for linear and logistic regression may improve model performance, as not all independent variables in this analysis turned out to be statistically significant. Additionally, the model coefficients may be distorted by collinearity between predictors. Removing one or more of the redundant variables could improve the coefficient estimates and standard errors.

Due to the extreme net loss associated with misclassifying an unhealthy engine as healthy, further adjustments are recommended to increase the sensitivity of the logistic regression predictions. Airlines should consider decreasing the cutoff value for the classification of unhealthy engines to minimize cost and maximize safety. The cutoff value could be automatically optimized using pre-determined decision values for each type of misclassification based on industry knowledge. Other machine learning models capable of binary classification should be considered, including but not limited to neural networks, decision trees, and random forests. Future research into these models may result in higher performance and accuracy.

**References**

Abbott, D. (2014). *Applied predictive analytics: Principles and techniques for the professional data analyst*. John Wiley & Sons.

Bousdekis, A., Lepenioti, K., Apostolou, D., & Mentzas, G. (2019). Decision making in predictive maintenance: Literature review and research agenda for industry 4.0. *IFAC-PapersOnLine, 52*(13), 607-612.

Census Bureau. (2021, December). *Annual survey of manufacturers*. https://data.census.gov/cedsci/table?q=AM1831BASIC&tid=ASMAREA2017.AM1831BASIC01

Daniyan, I., Daniyan, L., & Mpofu, K. (2022). The use of sensor based technology for enhancing maintenance operations. In I. A. Daniyan (Ed.), *Advances in manufacturing technologies and production engineering*, (pp. 89-100). Bentham Science Publishers. https://doi.org/10.2174/97898150397711220101

Davila, E. (2016, December 19). *Hypothesis testing* [Video]. LinkedIn Learning. https://www.linkedin.com/learning/statistics-foundations-3/hypothesis-testing-2

Kane, A. P., Kore, A. S., Khandale, A. N., Nigade, S. S., & Joshi, P. P. (2022). *Predictive maintenance using machine learning*. https://arxiv.org/abs/2205.09402

Kent State. (2022, May 12). *Statistical & qualitative data analysis software: About R and RStudio*. https://libguides.library.kent.edu/statconsulting/r

Kiangala, K. S., & Wang, Z. (2018). Initiating predictive maintenance for a conveyor motor in a bottling plant using industry 4.0 concepts. *The International Journal of Advanced Manufacturing Technology, 97*(9–12), 3251. https://doi.org/10.1007/s00170-018-2093-8

Leukel, J., González, J., & Riekert, M. (2021). Adoption of machine learning technology for failure prediction in industrial maintenance: A systematic review. *Journal of Manufacturing Systems, 61*, 87–96. https://doi.org/10.1016/j.jmsy.2021.08.012

Li, X., Elasha, F., Shanbr, S., & Mba, D. (2019). Remaining useful life prediction of rolling element bearings using supervised machine learning. *Energies, 12*(14), 2705. https://doi.org/10.3390/en12142705

Litt, J. S., Frederick, D. K., & DeCastro, J. A. (2007). *User’s guide for the commercial modular aero-propulsion system simulation (C-MAPSS)*. https://ntrs.nasa.gov/citations/20070034949

Liu, C., Tang, D., Zhu, H., & Nie, Q. (2021). A novel predictive maintenance method based on deep adversarial learning in the intelligent manufacturing system. *IEEE Access, 9*, 49557-49575.

Marr, B. (2015). *Big data: Using SMART big data, analytics, and metrics to make better decisions and improve performance*. Wiley.

McKinney, W. (2017). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython* (2nd ed.). O’Reilly Media.

Pech, M., Vrchota, J., & Bednář, J. (2021). Predictive maintenance and intelligent sensors in smart factory: Review. *Sensors, 21*(4), 1470. MDPI AG. http://dx.doi.org/10.3390/s21041470

Saxena, A., Simon, D., Goebel, K., & Eklund, N. (2008). Damage propagation modeling for aircraft engine prognostics. *International Conference on Prognostics and Health Management (PHM)*. https://ntrs.nasa.gov/citations/20090029214

Wu, D., Jennings, C., Terpenny, J., Gao, R. X., & Kumara, S. (2017). A comparative study on machine learning algorithms for smart manufacturing: tool wear prediction using random forests. *Journal of Manufacturing Science and Engineering, 139*(7). https://doi.org/10.1115/1.4036350