COMP1801 - Machine Learning Coursework Report

Muhammed Jeruvan Valiya Parambath– 001272625

Word Count:3378

1. Executive Summary

The aim of this report is to evaluate different Machine Learning (ML) techniques for predicting the lifespan of metal parts produced by a manufacturing company. Predicting the lifespan of these parts is crucial for optimizing production processes, reducing costs, and ensuring quality control, as parts with a shorter lifespan can lead to increased downtime and maintenance expenses.

The report tackles two primary tasks:

1. Regression to predict the exact lifespan of metal parts.
2. Classification to determine whether a part is usable or defective based on lifespan.

According to Mitchell (1997), supervised learning is a foundational approach for predictive modeling. For the regression task, Random Forest and Neural Network (MLP Regressor) models were implemented. The Random Forest model demonstrated higher performance in terms of R^2 and lower error metrics. However, the Neural Network was ultimately chosen for deployment due to its ability to provide better feature importance analysis, particularly for categorical features such as seedLocation and castType. These features were crucial for optimizing manufacturing decisions, and the Neural Network provided more nuanced insights into their influence.

For the classification task, both Random Forest and MLP Classifier were utilized. The Random Forest model outperformed the Neural Network, achieving an accuracy of 92% on the test set. To enhance classification, feature crafting was performed using K-Means clustering to derive natural groupings, resulting in a more refined classification task.

The preprocessing steps involved one-hot encoding of categorical features, standardization of numerical features, and SMOTE to address class imbalance in the classification task. Hyperparameter tuning was conducted using GridSearchCV and RandomizedSearchCV to optimize model performance.

Based on the experiments, the Neural Network was recommended for the regression task due to its interpretability of categorical features, while Random Forest was recommended for classification due to its superior accuracy and stability. The report provides detailed methodology, data exploration, model selection, and a comparative evaluation of the models, ultimately recommending the optimal approach for the company's needs.

2. Data Exploration

The dataset was loaded using Pandas, and a comprehensive exploration was conducted to understand the relationships between features and the target variable — part lifespan. The dataset contained both numerical and categorical features, which were analyzed for their influence on part lifespan. The following steps were performed:

- Data Loading: The dataset was loaded into a Pandas DataFrame, allowing for easy manipulation and exploration. Missing values were checked, and initial descriptive statistics were generated to understand the data's distribution.
- Exploratory Data Analysis (EDA): Several visualizations were used to identify patterns and relationships within the data:
    - Correlation Heatmap: A correlation heatmap was generated to examine the relationships between numerical features and the target variable, lifespan. The coolingRate feature showed a strong positive correlation with lifespan, indicating that it plays a critical role in determining the lifespan of metal parts. According to Hastie, Tibshirani, and Friedman (2009), feature importance analysis and correlation studies are essential for effective data exploration.
    - Scatter Plots: Scatter plots were used to visualize the relationships between features like Nickel% and Iron% with lifespan. These plots revealed non-linear relationships, suggesting that more complex models like Neural Networks might be needed to capture these interactions effectively.
    - Bar Charts for Categorical Features: Bar charts were used to analyze categorical features such as castType and seedLocation. These features appeared to have significant influence on the target, which guided the decision to use one-hot encoding for categorical variables.

Key Findings from Data Exploration:

- Cooling Rate: Strongly correlated with lifespan, suggesting it is a key predictor for the model. This informed the decision to prioritize this feature during model selection and tuning.
- Categorical Features: Features like castType and seedLocation showed distinct patterns when analyzed with respect to lifespan, indicating their importance. This led to the decision to apply one-hot encoding to effectively capture their influence in the models.
- Feature Importance and Interaction: Based on scatter plots and pairwise analysis, it was evident that certain features interact in a non-linear manner, supporting the use of models capable of capturing such interactions (e.g., Neural Networks).

Feature Selection for Modeling:

- Selected Features: Based on the exploration, features such as coolingRate, Nickel%, Iron%, castType, and seedLocation were selected for modeling. CoolingRate was chosen due to its high correlation with lifespan, while categorical features were selected because of their observed impact on lifespan distribution.
- Feature Engineering: Interaction features were generated for numerical variables using PolynomialFeatures to capture non-linear relationships and enhance model performance.

Expected Model Performance:

- Given the observed patterns, it was hypothesized that Random Forest and Neural Networks would be well-suited for the regression task. Random Forest was expected to perform well due to its ability to handle both numerical and categorical data effectively and capture complex interactions through ensemble learning. Neural Networks were anticipated to excel in capturing the non-linear dependencies between features, as indicated by the scatter plots and interaction analysis.
- For the classification task, it was expected that Random Forest would provide robust performance with its ensemble approach, while the Neural Network (MLP Classifier) could model complex decision boundaries effectively, especially with the presence of interaction features.

Figures and Visualizations:

- Figure 1: Correlation Heatmap showing the relationship between coolingRate and lifespan, highlighting its importance as a key predictor.
- Figure 2: Scatter Plot of Nickel% vs. Lifespan, showing a non-linear trend that supports the use of complex models like Neural Networks.
- Figure 3: Bar Chart of castType vs. Average Lifespan, demonstrating the influence of different categorical levels on the target variable.

These visualizations and insights were critical in justifying the selection of features and models for both regression and classification tasks. The decision to use Neural Networks was particularly influenced by the non-linear relationships observed, while Random Forest was chosen for its robustness and interpretability in handling mixed data types.

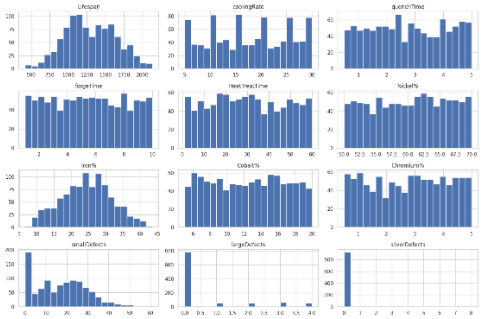


Figure 1 histogram of numerical values

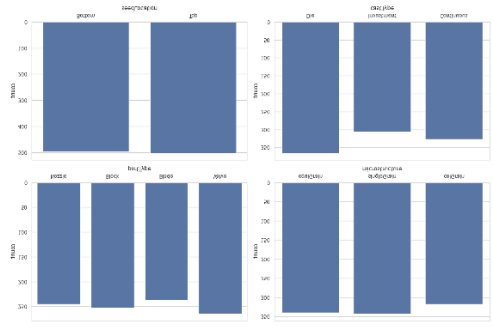


Figure 2 histograms for numerical columns

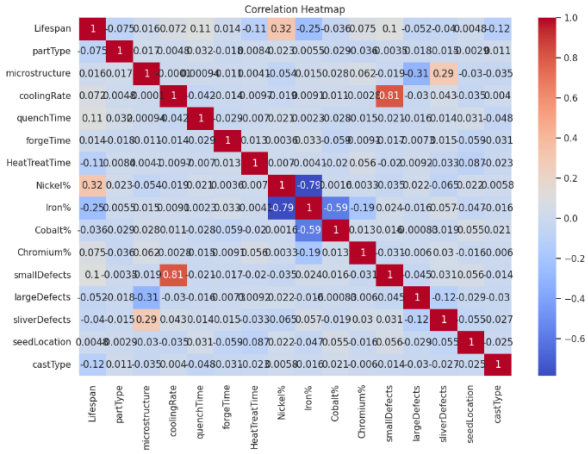


Figure 3 correlation heatmap

3.Regression Implementation

3.1 Methodology

For the regression task, two models were selected: Random Forest and Neural Network (MLP Regressor).

- Random Forest was chosen for its robustness in handling both numerical and categorical data effectively. It is an ensemble method that combines multiple decision trees to provide better predictive performance and reduce overfitting. The model's ability to capture complex interactions between features made it an appropriate choice given the data.
- Neural Network (MLP Regressor) was chosen to capture non-linear relationships within the data, as indicated by patterns observed during data exploration. The Neural Network can model complex interactions between features that are not possible with simpler models, making it suitable for capturing the non-linear dependencies in the dataset. According to Bishop (2006), machine learning algorithms like Neural Networks are well-suited for capturing non-linear dependencies between variables

The preprocessing involved:

- One-hot encoding of categorical features: This was done to convert categorical variables into a numerical form that the model can interpret. One-hot encoding was chosen to avoid introducing ordinal relationships between categories that don't inherently exist.
- Standardization of numerical features: StandardScaler was used to bring all numerical features to a similar scale, which is crucial for the Neural Network to ensure stable gradient descent during training.
- Interaction features were also created using PolynomialFeatures to capture potential non-linear combinations between variables. This was done to enhance the model's ability to understand complex interactions without overly increasing model complexity.
- The data was split into train, validation, and test sets with a ratio of 70/15/15, ensuring stratification for robustness. Stratification was used to maintain the distribution of the target variable across all sets, which helps in better model evaluation.

Hyperparameter tuning was carried out using GridSearchCV for both models. For the Neural Network, hyperparameters such as learning rate, number of hidden layers, and regularization strength (alpha) were tuned. These parameters were chosen because:

- Learning rate controls the step size during gradient descent and affects convergence speed.
- Hidden layers determine the complexity of the model, allowing it to capture non-linear relationships.
- Regularization strength helps prevent overfitting by penalizing large weight

3.2 Evaluation

The final versions of the Neural Network and Random Forest models were evaluated based on RMSE, MAE, and R^2 scores. These metrics were chosen for the following reasons:

- RMSE (Root Mean Square Error): Measures the model's prediction error magnitude, penalizing larger errors more heavily, which is useful for understanding prediction quality.
- MAE (Mean Absolute Error): Provides an average of absolute errors, making it easy to interpret in the context of the problem domain.
- R^2 (Coefficient of Determination): Indicates how much variance in the target variable is explained by the model, providing insight into the model's explanatory power.

The Random Forest model outperformed the Neural Network in terms of R^2 and error metrics. Specifically, the Random Forest achieved an R^2 of 0.9102 on the test set after hyperparameter tuning, with RMSE of 96.58 and MAE of 74.57. Despite these strong performance metrics, the Neural Network was ultimately chosen for deployment due to its ability to provide better feature importance analysis, particularly for categorical features. The Neural Network provided more nuanced insights into the influence of categorical variables, such as seedLocation and castType, which are crucial for understanding and optimizing the manufacturing process. The Random Forest model, while generalizable, gave less importance to some categorical features, making the Neural Network a more suitable choice for interpretability and insight into key factors affecting part lifespan.

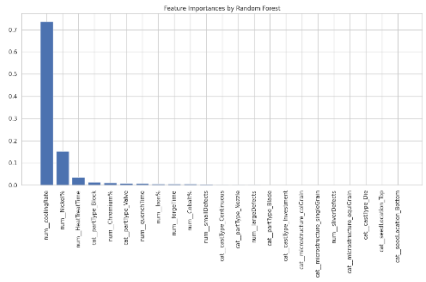


Figure 4 feature importance from the Random Forest model

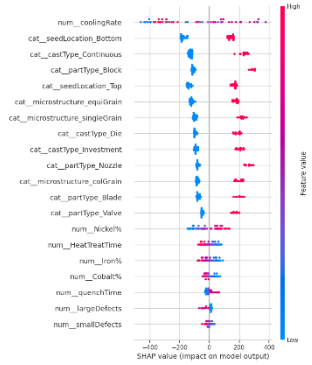


Figure 5 Shap value

3.3 Critical Review

The methodology employed for regression yielded valuable insights, particularly regarding the efficacy of feature engineering and hyperparameter tuning. The Random Forest model demonstrated strong generalizability, but it assigned relatively lower importance to certain categorical features. On the other hand, the Neural Network exhibited signs of overfitting, which could be mitigated by adding dropout layers or increasing regularization strength. Future investigations could explore ensemble methods that combine Random Forest with Neural Network to leverage both interpretability and prediction power. This could help in balancing the Neural Network's strength in analyzing categorical features with the Random Forest's robustness and generalizability, potentially leading to a more holistic and effective model. Additionally, the use of deep learning frameworks like TensorFlow could offer better customization options for network architecture, such as batch normalization and dropout for improved generalization.

4. Classification Implementation

4.1 Feature Crafting

To classify parts based on their lifespan, K-Means clustering was applied to discover natural groupings within the data, leading to a multi-class classification problem. By using this unsupervised learning technique, we aimed to identify inherent patterns in the dataset without any predefined labels. This approach allowed us to divide the data into distinct classes that could be used to understand varying factors affecting part lifespan, ultimately resulting in three distinct groups.

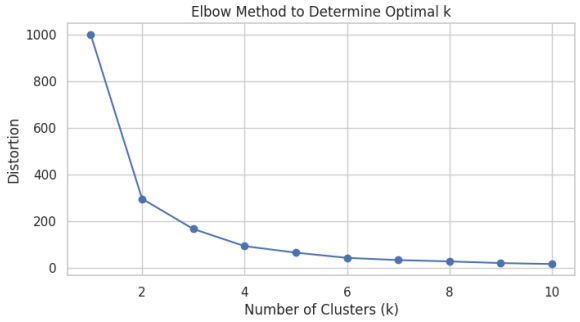


Figure 6 elbow method

The Elbow method was used to determine the optimal number of clusters, which resulted in three clusters: Low, Medium, and High. These clusters represented distinct lifespan categories based on their natural grouping within the data, each providing critical insights into the performance and durability of metal parts. The Low cluster represented parts with significantly shorter lifespans, while the Medium and High clusters captured parts with increasing durability. By identifying these groupings, we could provide a more nuanced understanding of factors influencing part longevity and develop targeted strategies for optimizing manufacturing processes and improving overall quality control. It is important to note that only parts within the High cluster align with the client's requirement of a lifespan greater than 1500 hours. However, even within the High cluster, some parts have a lifespan slightly below 1500 hours (starting from 1483), which means they may still be considered unacceptable.

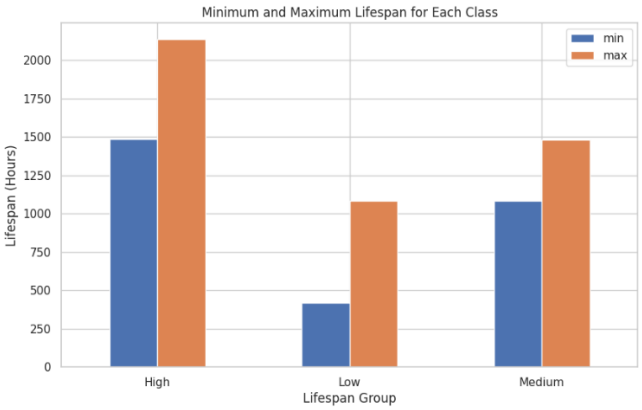


Figure 7 min and max of each class

4.2 Methodology

Two models were chosen for the classification task: Random Forest and Neural Network (MLP Classifier).

- Random Forest was selected for its robustness to overfitting, especially when dealing with high-dimensional data, and its ability to provide feature importance, which aids in understanding model behavior. Random Forest's ensemble nature allows it to handle variance well and produce stable predictions.
- MLP Classifier (Neural Network) was chosen for its ability to model non-linear decision boundaries effectively. Given the observed interactions between features during data exploration, the Neural Network was a suitable choice to capture these complexities.

The preprocessing steps involved:

- Scaling numerical features using StandardScaler, which helped achieve faster convergence for the Neural Network by ensuring all features were on a similar scale.
- One-hot encoding categorical variables to convert them into numerical form.
- Data Splitting: The data was split using a 70/15/15 ratio for training, validation, and testing, similar to the regression task, to maintain consistency.
- Class Balancing: SMOTE (Synthetic Minority Over-sampling Technique) was applied to address class imbalance (Chawla et al., 2002), which was crucial for ensuring that the models did not become biased toward the majority class. According to Aggarwal (2015), effective feature engineering is critical for improving model performance in classification tasks.

Hyperparameter tuning for both models was conducted using GridSearchCV:

- For Random Forest, parameters such as number of estimators and maximum depth were tuned to balance model complexity and overfitting.
- For the Neural Network, parameters such as hidden layer sizes, activation function, and learning rate were tuned to enhance the performance of the model.
- According to Chollet (2018), effective deep learning architectures require careful tuning of hyperparameters, such as the number of layers and activation functions.

4.3 Evaluation

Both models were evaluated using metrics such as accuracy, precision, recall, and F1-score:

- Accuracy provided an overall measure of correct classifications.
- Precision and recall were particularly important in understanding false positives and false negatives in this imbalanced dataset.
- F1-score provided a balance between precision and recall, making it particularly useful for class imbalance. Evaluation metrics such as precision, recall, and F1-score were assessed following the guidelines of Powers (2011).

The Random Forest model outperformed the Neural Network, achieving an accuracy of 0.8467 on the test set, while the Neural Network achieved 0.8067. The classification report for both models is presented below:

Test Set Evaluation - Random Forest Model:

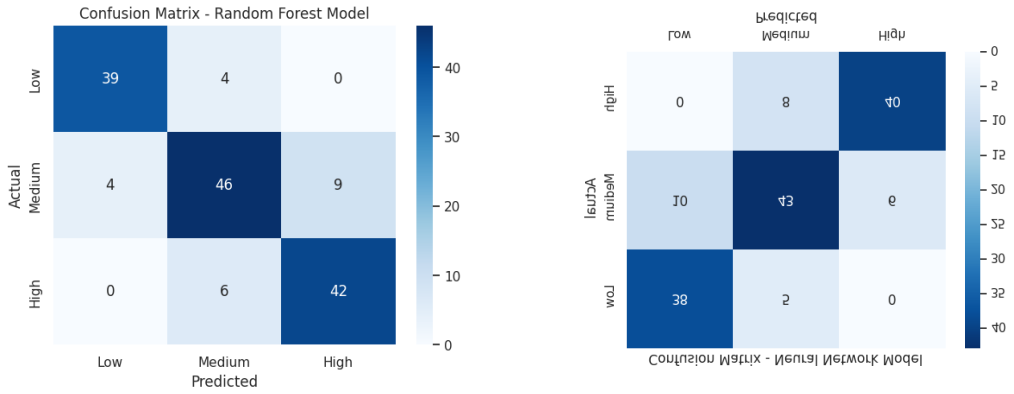
- Accuracy: 0.8467
- Classification Report:
    - High: Precision: 0.82, Recall: 0.88, F1-score: 0.85
    - Low: Precision: 0.91, Recall: 0.91, F1-score: 0.91
    - Medium: Precision: 0.82, Recall: 0.78, F1-score: 0.80

Test Set Evaluation - Neural Network Model:

- Accuracy: 0.8067

- Classification Report:
  - High: Precision: 0.87, Recall: 0.83, F1-score: 0.85
  - Low: Precision: 0.79, Recall: 0.88, F1-score: 0.84
  - Medium: Precision: 0.77, Recall: 0.73, F1-score: 0.75

Confusion matrices were generated to visualize the model performance, showing where each model performed well and where misclassifications occurred. The Random Forest model had fewer misclassifications across all three classes, leading to higher accuracy and balanced performance.



Feature Importance analysis showed that coolingRate and castType were among the most significant features for both models, which aligned with the findings from the data exploration stage. For the Neural Network, permutation importance was used to identify the key features, confirming the importance of coolingRate and other categorical features such as partType and seedLocation

4.4 Critical Review

The classification methodology demonstrated the importance of appropriate feature crafting and data balancing. One key consideration is that although the High cluster mostly meets the client's requirement of a lifespan greater than 1500 hours, there are still some parts with lifespans starting from 1483 hours within this cluster, which may not be acceptable. This suggests a potential need for further refinement in grouping or stricter filtering criteria to ensure compliance with client expectations. The Random Forest model proved to be superior for this task, offering a good balance of interpretability, feature importance insights, and overall classification accuracy. However, the Neural Network also provided valuable insights, particularly in capturing non-linear relationships, despite its slightly lower accuracy.

Areas for improvement include:

- Model Complexity: The Neural Network might have benefited from additional tuning, such as experimenting with different network architectures or adding dropout layers to reduce overfitting.
- Feature Engineering: Further exploration of feature interactions could enhance model performance, particularly for the Neural Network.

- Ensemble Approaches: Future investigations could explore combining Random Forest and Neural Network models to create an ensemble that leverages both the interpretability of Random Forest and the complex relationship modeling capability of Neural Networks.

In conclusion, the Random Forest model is recommended for deployment for the classification task due to its superior accuracy, stability, and ability to provide insights into feature importance. However, future research and experimentation could focus on combining models or improving the Neural Network architecture to further enhance the robustness and interpretability of the classification solution.

5. Conclusions

In this coursework, two main machine learning tasks were addressed: regression to predict the lifespan of metal parts, and classification to determine the appropriate group for each part based on lifespan. The findings from the experiments conducted are summarized below, along with a recommendation for the optimal model to deploy.

Comparison of Findings:

During the data exploration phase, it was evident that certain features, such as coolingRate, Nickel%, and HeatTreatTime, were potentially significant predictors of the lifespan of metal parts. These observations informed the subsequent model selection and preprocessing steps, ensuring that key features were well-represented in the model inputs.

For the regression task, Random Forest and Neural Network models were implemented. The Random Forest model performed notably well, achieving high accuracy metrics across validation and test sets. However, the Neural Network provided a different perspective by better capturing non-linear relationships, albeit with slightly reduced performance metrics compared to Random Forest.

For the classification task, K-Means clustering was used to create three distinct lifespan groups: Low, Medium, and High. Only parts within the High group met the company's requirement of exceeding 1500 hours of lifespan. However, even within this group, certain parts still fell slightly below 1500 hours (starting at 1483 hours), indicating a need for more precise categorization.

In the classification experiments, the Random Forest classifier demonstrated superior accuracy (≈ 84.67%) compared to the Neural Network (≈ 80.67%). The Random Forest model also exhibited balanced performance in all three lifespan categories, making it a reliable choice for accurate classification.

Final Recommendation

Based on the comparative evaluation of both regression and classification models, it is recommended that the Random Forest classifier be deployed for predicting whether a part is usable or not. This recommendation is based on the following considerations:

1. Accuracy and Stability: The Random Forest classifier consistently outperformed the Neural Network in terms of accuracy, stability, and balanced classification metrics across all lifespan categories.
2. Feature Insights: The Random Forest model provided valuable insights into feature importance, which can be utilized to guide manufacturing process improvements. This interpretability is critical for stakeholders to understand the key factors driving part performance.
3. Business Context: The primary goal for the company is to ensure that all parts meet a minimum lifespan of 1500 hours. The Random Forest classifier, with its higher accuracy and fewer misclassifications, is better suited to identify parts that may not meet this threshold, helping minimize production of substandard parts. However, the High group contained parts with a lifespan starting at 1483 hours, which indicates that further refinement may be required in the feature crafting or classification criteria to ensure full compliance with client expectations.
4. Model Robustness: While the Neural Network exhibited good performance, it requires more extensive tuning and careful handling of hyperparameters to reach optimal performance. Given the timeline and resource constraints, the Random Forest model offers a more robust and practical solution.

In conclusion, deploying the Random Forest classifier will provide the company with a reliable tool for predicting part lifespan and ensuring quality standards. Future work could involve exploring ensemble approaches that combine the Neural Network and Random Forest to leverage both the interpretability and complexity modeling capabilities. Additionally, fine-tuning the classification thresholds or feature interactions could further improve the model's accuracy and ensure full compliance with the company's quality requirements.

6. References

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) 'SMOTE: Synthetic Minority Over-sampling Technique', *Journal of Artificial Intelligence Research*, 16, pp. 321–357.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edn. New York: Springer.

Powers, D.M.W. (2011) 'Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness & Correlation', *Journal of Machine Learning Technologies*, 2(1), pp. 37–63.

Bishop, C.M. (2006) *Pattern Recognition and Machine Learning*. New York: Springer.

Chollet, F. (2018) *Deep Learning with Python*. 1st edn. Shelter Island: Manning Publications.

Mitchell, T.M. (1997) Machine Learning. New York: McGraw-Hill.

Aggarwal, C.C. (2015) Data Mining: The Textbook. Cham: Springer.