COMP1801 - Machine Learning Coursework Report

Azhar Muhammed - 001364857
Word Count:

1. Executive Summary

The purpose of this report is to develop machine learning models to predict the lifespan of metal parts and determine whether they are defective based on a given dataset. The project addresses two key areas: regression to predict lifespan and classification to determine defect status. Regression was implemented using models like Linear Regression, Random Forest, and an Artificial Neural Network (ANN), with ANN ultimately chosen for its robustness and ability to capture complex relationships. The classification task involved Logistic Regression and ANN to label parts as defective or not based on a lifespan threshold of 1500 hours. Through careful feature crafting, preprocessing, and evaluation using metrics such as RMSE, R^2, Weighted F1-Score, and Recall, the study concludes that the ANN model is preferable for both regression and classification tasks due to its superior performance in handling complex patterns, class imbalance, and providing accurate predictions.

2. Data Exploration

The purpose of this data exploration is to understand the dataset, identify significant relationships among features, and determine the best predictors for modeling metal part lifespan. By employing visualizations such as scatter plots, regression lines, and polynomial feature transformations, this project aims to elucidate the factors influencing part longevity and guide model development to ensure robust, precise, and interpretable predictive models.

The dataset was loaded using Python's `pandas` library, and inconsistencies or missing values were addressed. The target variable, `Lifespan`, had 998 unique values out of 1000 data points, indicating an almost entirely unique dataset.

Initial exploratory data analysis (EDA) was conducted using scatter plots, histograms, and other visualizations to understand relationships between features and `Lifespan` and to identify linear and non-linear trends.

The analysis revealed that `coolingRate` had a very weak positive correlation with `Lifespan`, leading to its tentative inclusion in the model. ‘quenchTime‘ and ‘forgeTime‘ showed slight positive correlations, suggesting that extending these processes could enhance metal durability by improving internal structure.

Alloy composition analysis showed that **Nickel%** had a weak positive correlation, likely enhancing lifespan through corrosion resistance, while **Iron%** showed a slight negative correlation, possibly reducing durability due to brittleness. **Cobalt%** and **Chromium%** exhibited weak positive trends, consistent with their role in improving metal strength.

smallDefects required a polynomial transformation to capture its parabolic relationship with `Lifespan`. Initially, small defects had little impact, but beyond a threshold, they reduced durability. Based on these observations, `coolingRate` and `smallDefects` (with a quadratic term) were selected as key features for modeling lifespan. All features were thoroughly considered, as the intention is to implement a Neural Network in future tasks to capture all potential interactions and non-linearities.

The high number of unique `Lifespan` values (998 out of 1000) favored a regression approach for predicting continuous outcomes. Polynomial Regression was applied to capture non-linear relationships for `coolingRate` and `smallDefects`.

In summary, the data exploration provided valuable insights into feature relationships with `Lifespan`, informing both feature selection and modeling approaches for robust durability prediction.

3. Regression Implementation

This section focuses on regression as a method to predict the lifespan of metal parts using the given dataset. Regression analysis was used to develop a model capable of accurately predicting the lifespan based on features such as alloy composition, manufacturing conditions, and defects. In this context, an Artificial Neural Network (ANN) was chosen due to its robustness and capability to consider all features for predicting lifespan, whereas the Random Forest and XGBoost models focused specifically on the most important features. The following sections will detail the chosen regression models, the preprocessing and tuning methodologies, and an evaluation of their performances, aiming to establish the most suitable regression model for this task.

3.1. Methodology

Two regression models were chosen for this task: Linear Regression and Artificial Neural Network. Linear Regression was selected for its simplicity and as a baseline to compare against more complex methods, whereas Random Forest Regressor was selected due to its ability to handle non-linear relationships and interactions without requiring extensive feature engineering. During initial testing, tree-based models, particularly Random Forest and XGBoost, demonstrated superior performance compared to other models, such as Artificial Neural Networks (ANNs). However, in the Random Forest and XGBoost models where only focusing on high importance features. I have chose ANN due its robustness and considering all features for predicting lifespan. The dataset contained mixed data types, necessitating careful preprocessing to standardize and encode features appropriately. StandardScaler was applied to normalize numerical features, which improved model convergence and stability by scaling features to have a mean of 0 and standard deviation of 1. This choice of scaler was particularly suited due to the approximate normal distribution of several numerical features.

To address categorical variables, a combination of **One-Hot Encoding**, **Label Encoding** and a hybrid approach was employed. The hybrid approach ensured that high-cardinality categorical variables were appropriately represented, while the partType column were treated with Label Encoding to retain ordinal relationships where applicable. However, this approach shown low results while One-Hot encoding performed better. A consistent train-test split of 70To avoid overfitting and ensure that the model generalizes well, **cross-validation** was implemented along with a validation set. This allowed better use of the available data for both training and validation, thus enhancing the reliability of the model performance estimates. Hyperparameter tuning was performed for each model to optimize performance. For Linear Regression, regularization techniques such as Lasso and Ridge regression were considered to improve model stability and prevent overfitting. For the Artificial Neural Network, key hyperparameters including learning rate, number of units per layer, and dropout rates were tuned using RandomSearchCV. After conducting 150 trials, the Sequential Neural Network achieved an optimal validation loss of 12,098.92. The best architecture consisted of two layers, with the first layer containing 64 units and a dropout rate of 0.1, and the second layer containing 112 units with a dropout rate of 0.2. An optimal learning rate of 0.00566 was identified. The evaluation metrics for the Neural Network model on the test set, using a wrapper class, were as follows: RMSE = 113.29, $R^2 = 0.90$, and MAE = 90.19. This relatively simple architecture, combined with batch normalization and dropout, effectively balanced model complexity and performance. While a final validation loss of 61,680.13 was recorded in one trial, the configuration with the best observed performance can serve as a strong baseline for future work. The consistent use of batch normalization and moderate dropout rates provided effective regularization, enhancing the model's ability to capture complex interactions without overfitting.

3.2. Evaluation

The chosen regression models were trained on the training set and evaluated using the testing set to compare their effectiveness in predicting part lifespan. Each model's performance was quantified using several metrics such as R2, RMSE and MAE that are particularly well-suited for regression tasks involving continuous numerical outputs.

3.3. Critical Review

The metrics used to evaluate the models included Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 Score. RMSE was selected for its ability to penalize larger errors, which is crucial when predicting lifespans that may vary significantly. MAE provided a straightforward measure of the average error, and R^2 Score quantified the proportion of variance explained by the model. These metrics offered comprehensive insights into both prediction accuracy and error magnitude. The initial comparison of the models highlighted that the Random Forest Regressor outperformed Linear Regression in terms of RMSE and R^2 Score, capturing the non-linear relationships in the data more effectively. While Linear

Regression provided a simple baseline, it struggled with the complex dependencies inherent in the dataset. In contrast, the Random Forest Regressor demonstrated superior flexibility and performance. Despite the strong performance of Random Forest, a finely tuned **Artificial Neural Network (ANN)** was ultimately chosen due to its robustness in considering all features and its ability to model complex relationships comprehensively. The ANN achieved an RMSE of 113.29, an R^2 of 0.90, and an MAE of 90.19, indicating a high level of accuracy in lifespan prediction. In comparison, Random Forest and XGBoost models focused primarily on high-importance features but did not achieve the same level of accuracy across all metrics.

The final step involved integrating all processes into a single **pipeline**, which combined preprocessing, model training, and hyperparameter tuning. This approach ensured that the resulting model was both efficient to deploy and consistent in performance across various scenarios.

4. Classification Implementation

4.1. Feature Crafting

4.2. Methodology

4.3. Evaluation

4.4. Critical Review

5. Conclusions