Project done by: S.R. Vidhyambika

REMAINING USEFUL LIFE PREDICTION FOR NASA TURBOFAN DATASET: A COMPREHENSIVE OVERVIEW

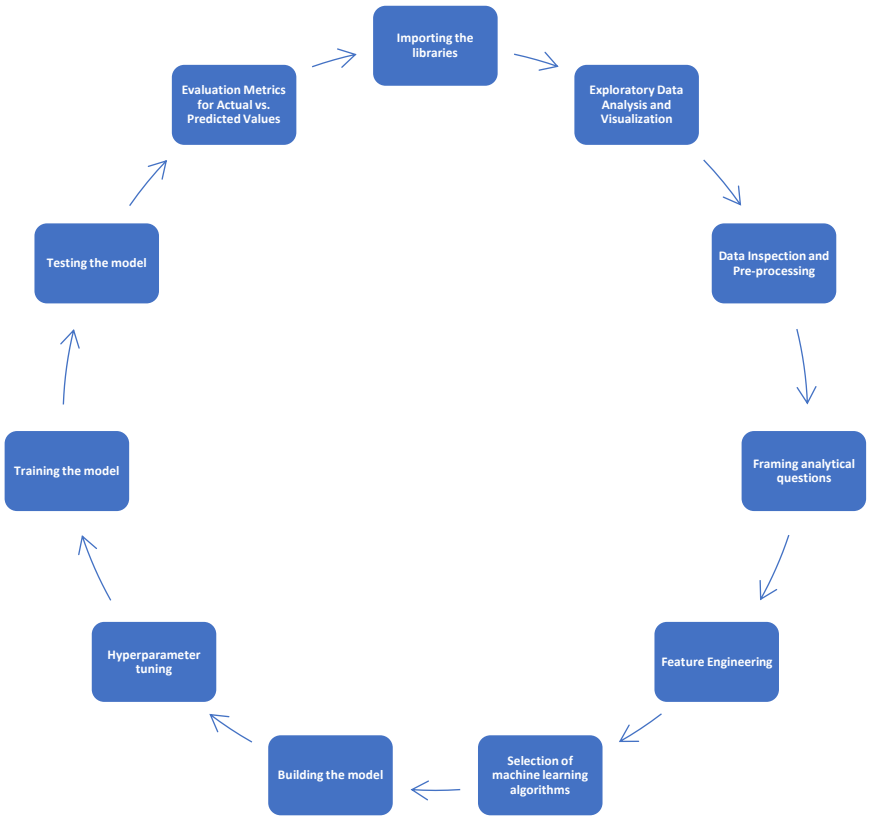
Introduction:

During their lifetime, aircraft components are susceptible to degradation, which affects directly their reliability and performance. This machine learning project will be directed to provide a framework for **predicting the aircraft's remaining useful life (RUL)** based on the entire life cycle data to provide the necessary maintenance behavior. Diverse regression models (Linear Regression, Random Forest, and Support Vector Regression (SVR) are deployed and tested on the NASA's C-MAPSS data-set to assess the engine's lifetime. Please check the report for more theoretical details.

Objective:

To inspect, pre-process, perform Exploratory Data Analysis (EDA), visualize the data, perform feature engineering, build the models, train, and test the models (regression models) and comparing actual vs predicted values.

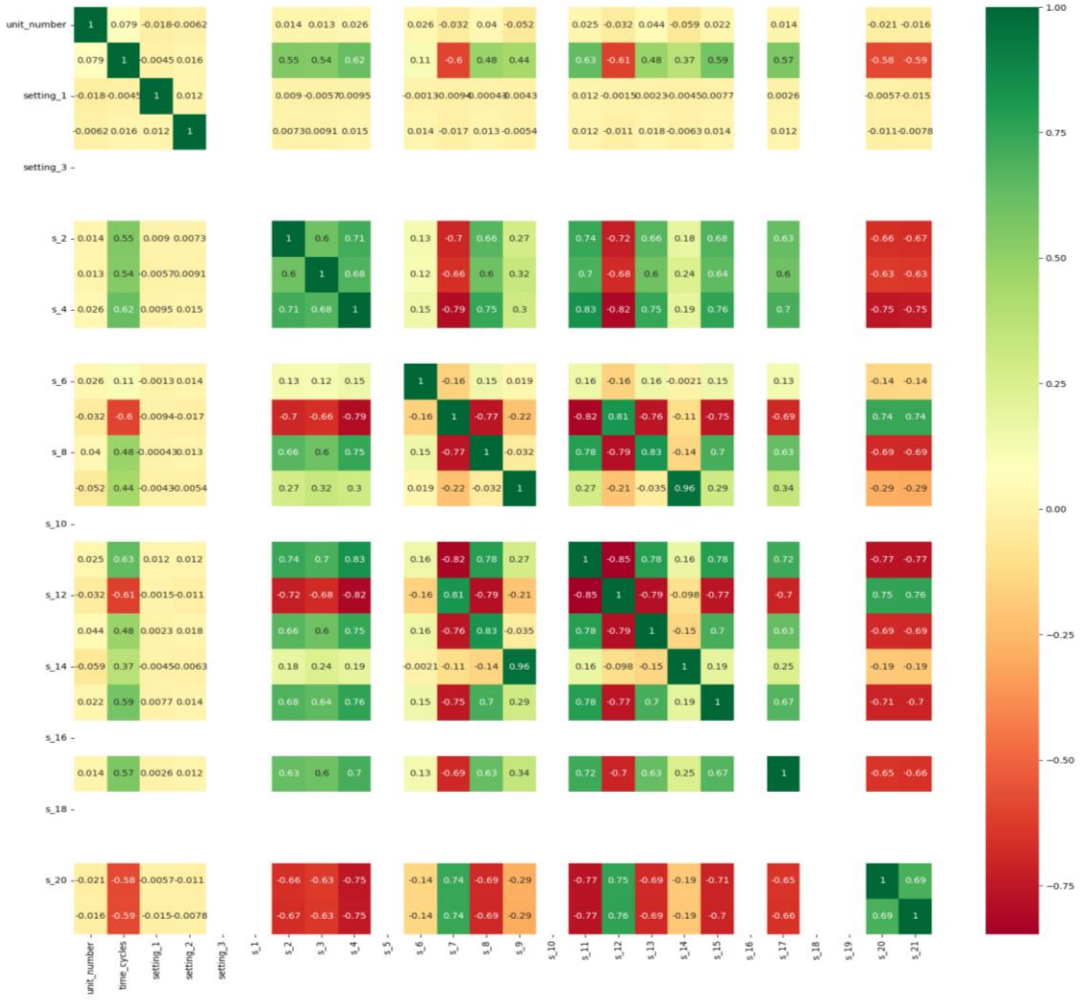
Steps followed in the data science project cycle:



1) **Importing the libraries:**
   For EDA, pre-processing, and visualization, Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn were imported and used. Scikit-learn was imported for data processing, model training, testing, and evaluation.

2) **Loading the dataset:**
The NASA turbofan jet engine train and test Data Set (CSV) file containing around 20631 rows in train and 13096 rows in test dataset with 26 attributes is loaded for analysis.

3) **Exploratory Data Analysis:**
   - The attributes s_11 and s_12 is highly negatively correlated with a value of -0.85.
   - The sensors s9 and s14 have high positive correlation of value 0.96.
   - There are outliers on both ends of the distribution, as the minimum and maximum values are far from the quartiles.



4) **Data Inspection and Pre-processing:**
   - No missing value is detected.
   - The unit numbers range from 1 to 100, but the mean and quantiles are different from a uniform distribution.
   - The units have different max time_cycles and number of rows, which affects the descriptive statistics.
   - The average unit has gone through 108.81 cycles, with a large standard deviation of 68.88.

- The statistical properties of sensors data show that the sensors do not have the same scale and they don't follow a normal distribution, which are solved by performing Minmax scaling on the data.

5) **Framing analytical questions:**
   - Can we efficiently predict the remaining useful time for the engine (error significance)?
   - Which features are the most important for predicting the failure of the turbofan engine?
   - Does adding historical data improve our model?
   - Is the collected data sufficient to give an accurate prediction?
   - How can we turn our problem to a classification one?
6) **Feature Engineering:**
   - The attributes: setting_3, s_1, s_5, s_10, s_16, s_18, and s_19 have constant values and so these columns are dropped/removed from the dataset.
   - Unit number and time cycles were removed.
   - Feature importance is checked for all the attributes with respect to each machine learning algorithm selected.
   - The sensors 3 and 17 have the least importance in feature importance for the models selected and hence are removed from the dataset.
7) **Selection of machine learning algorithms:**
   Random forest regressor, Linear regression and Support Vector Regression (SVR) algorithms were selected and used.
8) **Hyperparameter tuning:**
   Hyperparameter tuning is performed for random forest regressor using grid search cv and the optimal parameters were found to be: 'max_depth': 10, 'n_estimators': 120
9) **Train and test the models:**
   All the 3 ml models i.e., Random Forest regressor, SVR and linear regression were trained and tested with train and test data.
10) **Evaluation Metrics for Actual vs. Predicted Values:**
    - For evaluating all the 3 models with respect to actual and predicted values, R2 score and Root Mean Squared Error (RMSE) metrics are taken.
    - RMSE is a measure of how much the predicted values deviate from the actual values.
    - R2 score, also known as coefficient of determination, is a measure of how well the predicted values fit the actual values in a regression model.
    - For linear regression:
      train set RMSE:44.803728035551984, R2:0.5830135052284944
      test set RMSE:46.11402576435104, R2:0.5357359432947801
      valid set RMSE:43.05877886885164, R2:-0.07365283266128775
    - For random forest regressor with hyper parameter tuning:
      train set RMSE:15.532135723357555, R2:0.9498863496611952
      test set RMSE:44.317894428858324, R2:0.5711975842576446
      valid set RMSE:32.28263096465342, R2:0.39649857638353225
    - For support vector regression,
      train set RMSE:42.54724224579233, R2:0.6239578590036827
      test set RMSE:48.75516379159361, R2:0.48103242301470384

valid set RMSE:25.947912225366206, R2:0.6101071274546097

- The actual values and predicted values of each model are plotted as graphs for better visualization.

Conclusion:

Judging by the difference in RMSE and viewing R2 score, we notice that SVR & RF performs better than the other models when executed on the whole dataset. SVR has the best accuracy on test set of 61% and RF i.e., Random Forest regressor has the best training accuracy of 94.98% which gets dropped dataset due to overfitting i.e., lack of data (data has to be collected more). The valid test RMSE obtained is 26, which will be our score to beat while running the model on the best features found so far.