IMSE 514 - MULTIVARIATE STATISTICS
PREDICTING REMAINING USEFUL LIFE OF TURBO ENGINES

(*Shanmukh Varma Rudraraju, Yash Kakkad, Ananya Singh*)

INTRODUCTION

An important component of an aircraft is the turbo fan engine. Engine Components are prone to degradation over their life, which in turn affects the reliability of the engine as a whole. Thus, predictive maintenance of these engines is necessary to detect any possible failures early on. So in order to avoid these failures, the calculation of the remaining useful life of the engine is key. This project presents the prediction framework for the engine's life cycle based on a Machine Learning approach. The main aim of the project is to address this issue and explore it using various techniques and make improvements to existing models instead of just taking a bunch of models and comparing them.

DATASET

The dataset we've been given consists of three csv files, which consists of our X_train, that contains data from 50 aircraft turbofan engines from cycle 0 to the failure cycles. Another file is the X_test, which contains data from 50 aircraft turbofan engines from cycle 0 to a random cycle before failure. The final file is RUL_forecast_length.csv, which basically contains the remaining number of cycles for the engines in the testing set to reach the failure cycle. This is the number of cycles forecasted beyond the available data in "test_data.csv". This is thus our y_test. Now, our data has various columns which have their own specific purpose. To go into depth, there are three types of columns. They are:

- **Engine Related Information**: which consists of things like engine id and cycle number
- **Operating Conditions**
- **Sensor Readings**: which consists of values from 21 different sensors, ranging from temperature at fan inlets and outlet, coolant bleed values, etc.

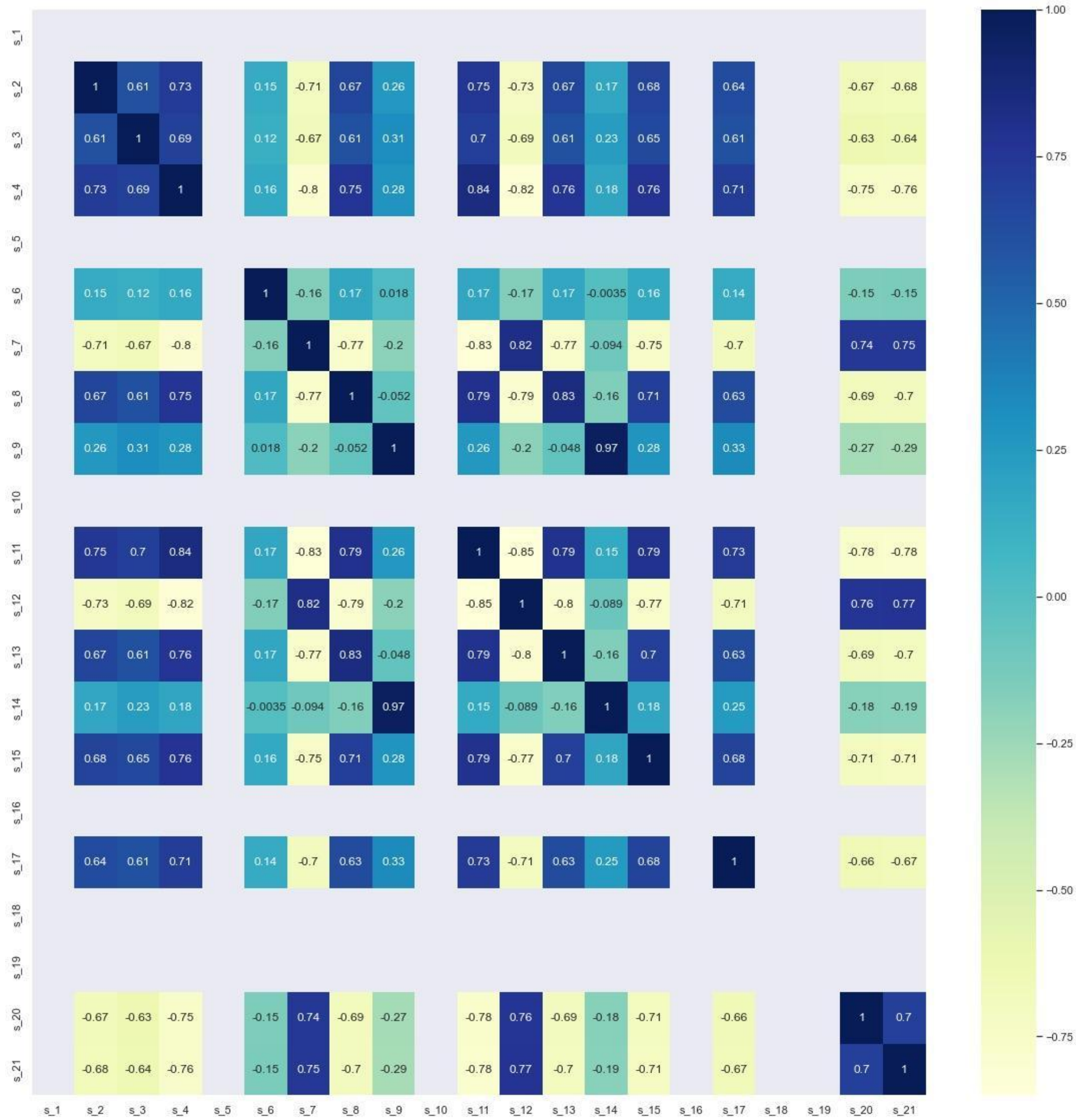
DATA VISUALIZATION

The following is a heat map between the sensor values that we have from the data. This helps us to visualize the data and find the pairs of sensors having highest correlation between them.
We are checking the multicollinearity between the sensors. From the matrix, we can observe that there is multicollinearity between the following pairs of sensors:

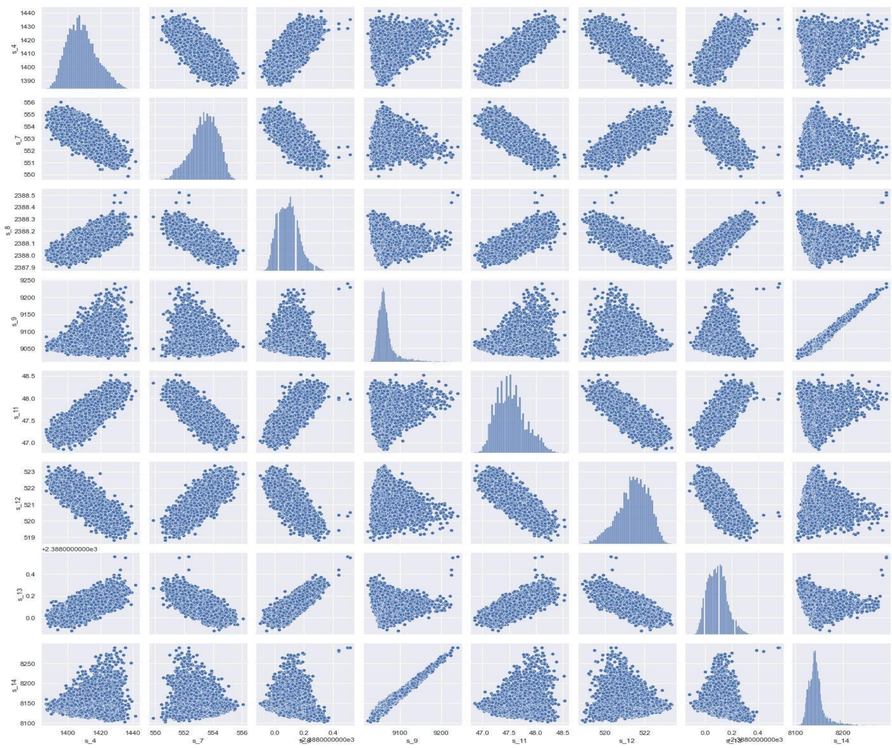
  - (s_11 and s_4)
  - (s_14 and s_9)

- (s_13 and s_8)
- (s_12 and s_7)

Here the correlation threshold for a pair of sensors is 80%

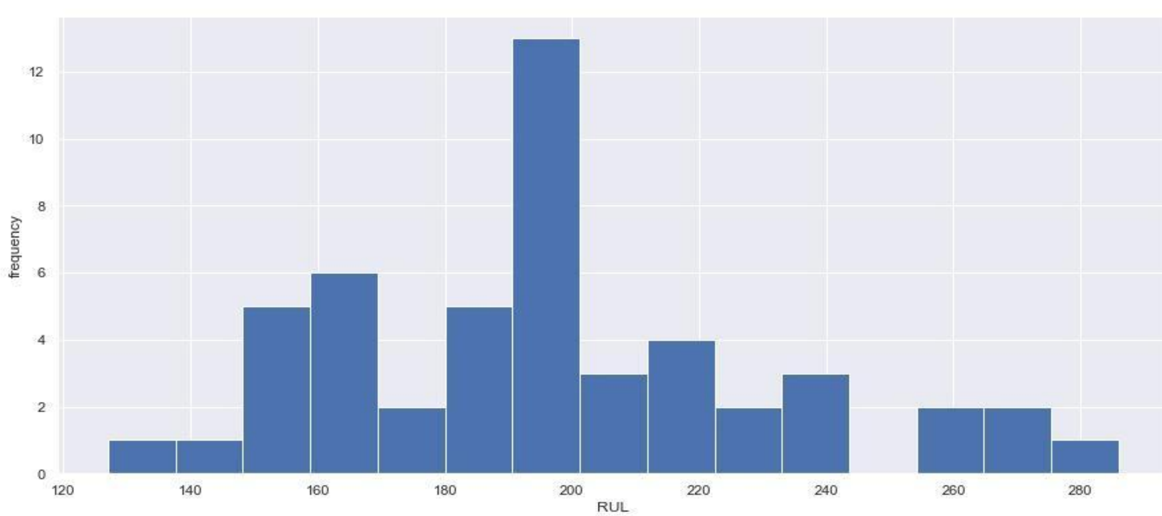


The following graph consists of the distribution of the correlating sensors

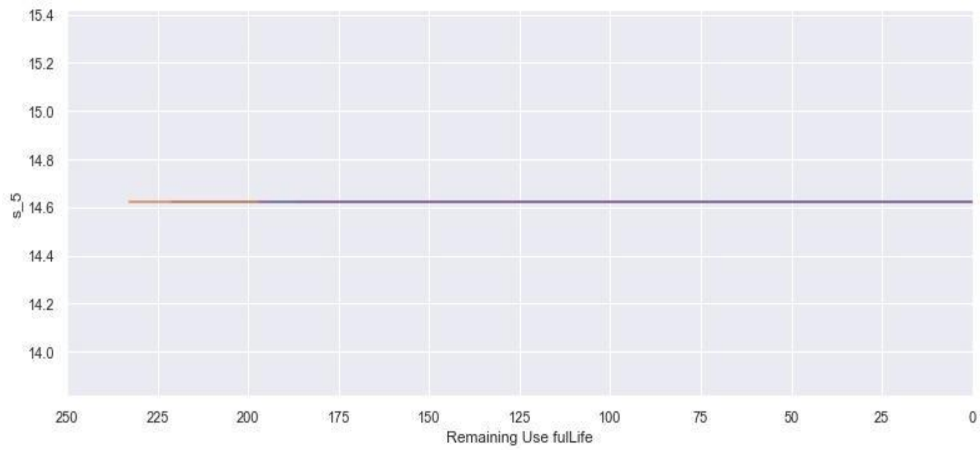


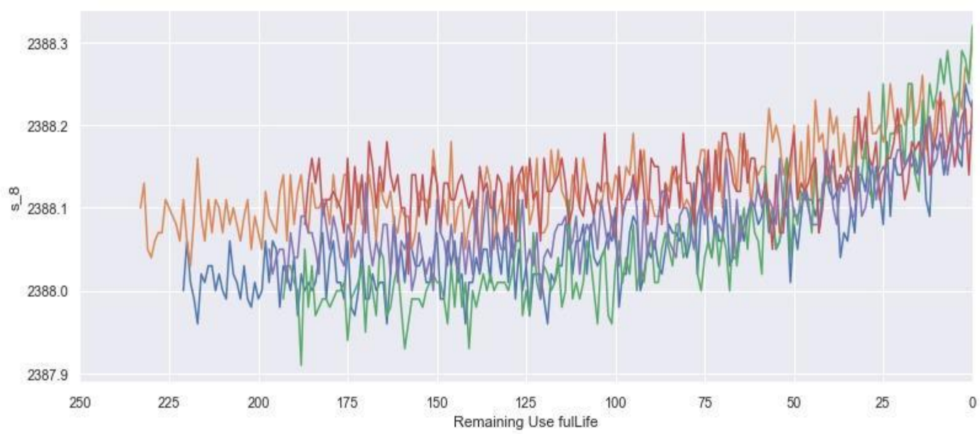
PROBLEM STATEMENT

The problem here is to produce a Machine Learning model which can predict the Remaining Useful Life (RUL) based on the series of data of 20+ sensor measurements which is available from the aircraft turbine engines. We know the formula to calculate the value of remaining useful life as: RUL Calculation: Max cycles – current cycle.



Now we do a comparison between the sensors and the calculated RUL values to check whether particular sensor is useful in the prediction of RUL values. From the above process we identified that there are sensors which do not show any pattern when plotted against the RUL values. These sensors are (1, 5, 10, 16, 18, and 19). Below shows the difference of sensor plots where a pattern is visible or not.





BUILDING MODELS

1. **MULTIPLE LINEAR REGRESSION**

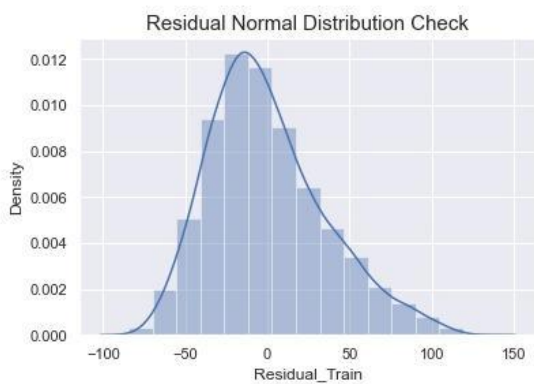
On the basis of the above observations we have removed the sensors (1, 5, 10, 16, 18, and 19). After this we create a linear regression model on the remaining sensors with the assumption that the RUL values will linearly decrease with the increase in the number of cycles completed.

For this we have noted the following evaluation metrics:

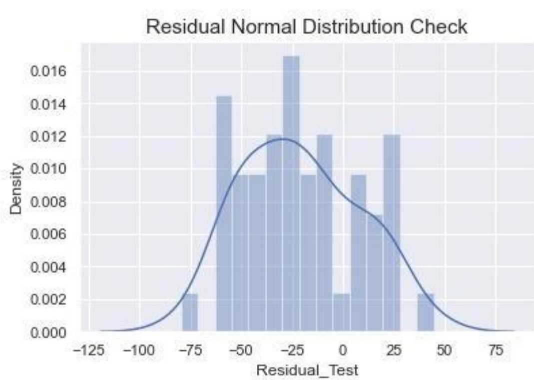
	R Squared	RMSE
Train Data	0.67	35.78
Test Data	0.68	35.53

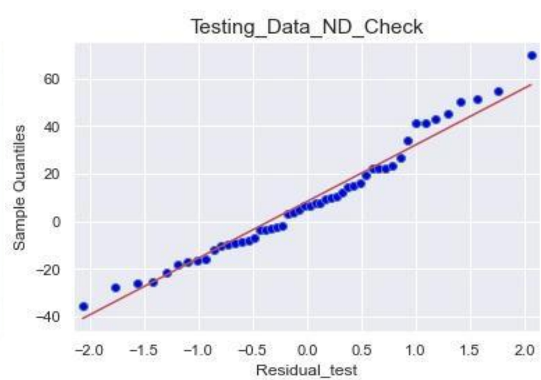
Intuition: RMSE Values of Training and Testing are similar but high so, we need to explore other models.

Train



Test





s

2. **REFRAMED MULTIPLE LINEAR REGRESSION**

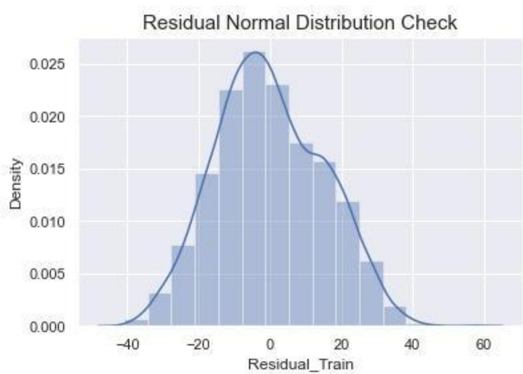
Here as well we create a linear regression model on the relevant sensors with a different assumption that the RUL values will remain constant till a particular number of cycles completed that is the threshold after which it starts decreasing gradually with increase in number of cycles. The assumed threshold is 100.

For this we have noted the following evaluation metrics:

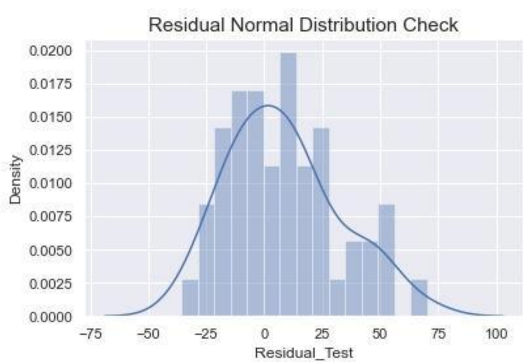
	R Squared	RMSE
Train Data	0.88	15.18
Test Data	0.74	25.21

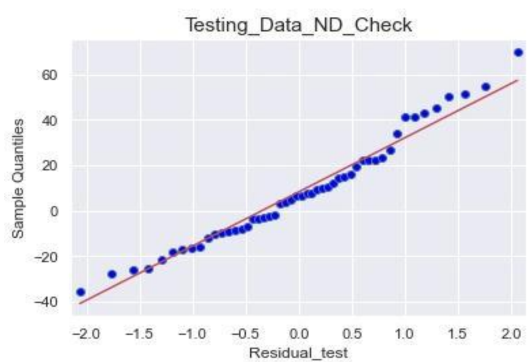
Intuition: RMSE Values of Training and Testing are acceptable and RMSE percentage is reduced by 58% and 29% respectively when compared with the previous MLR model.

Train



Test





3. **SUPPORT VECTOR MACHINES (REGRESSION)**

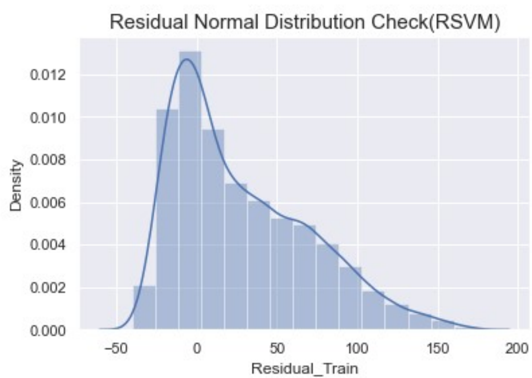
On the basis of the above observations we have removed the sensors (1, 5, 10, 16, 18, 19). After this we have normalized the data as SVM regression is based on Distance and normalizing the data before running SVM is necessary.

For SVM we have noted the following evaluation metrics:

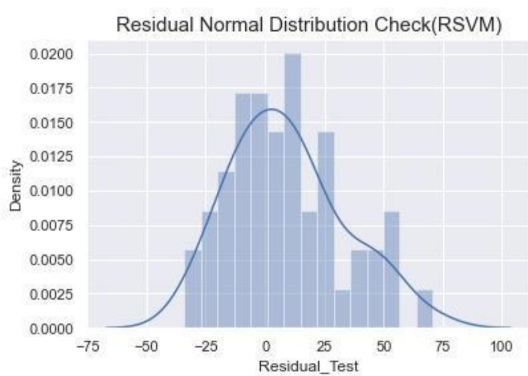
	R Squared	RMSE
Train Data	0.87	15.24
Test Data	0.74	25.49

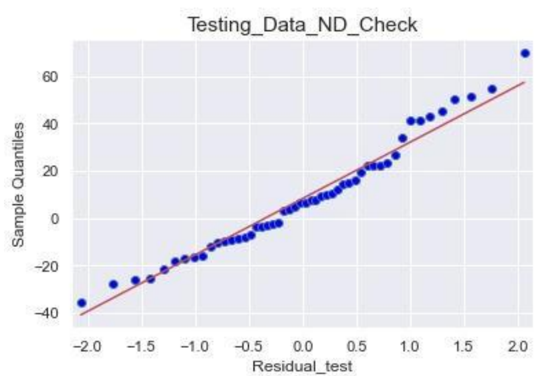
Intuition: RMSE Values of Training and Testing are different but acceptable and RMSE percentage is reduced by 58% and 29% respectively when compared with the MLR model.

Train



Test





4. **TIME SERIES MODEL**

The time series model is a distributed lag type of model in which the effect of the regressor x on y occurs over time and not all at once.

Table 1:

	R Squared	RMSE
Train Data	0.67	37.57
Test Data	0.593	35.34

Table 2:

	R Squared	RMSE
Train Data	0.783	19.81
Test Data	0.770	21.33

Table 1 is a using 1 lag and the Table 2 is of 9 lags, using 1 lag has improved the model a lot when compared to the base model, whereas in order to select the optimal lags we can use the AIC information to tell it which is 8 in this case.

5. LONG SHORT - TERM MEMORY (LSTM)

Long short-term memory (LSTM) in an artificial recurrent neural network(RNN) is an architecture of deep learning. For this we have noted the following evaluation matrix:

	R Squared	RMSE
Train Data	0.87	14.24
Test Data	0.75	22.49

The current LSTM model with a RMSE of 22.5 is already a 20% improvement to the base model, when the insignificant sensors are dropped, when hyper parameter tuning is done the model accuracy can go beyond 90% which we think is quite a gain.

RESULTS

From the various models developed on the given data we can say that the most significant sensors in the process of predicting the remaining useful life values are: T24, T30, T50, P15, P30, Nf, Nc, Ps30, phi, NRf, Nrc, BPR, htBleed, W31, W32

REFERENCES

- https://medium.com/@hamalyas_/jet-engine-remaining-useful-life-rul-prediction-a8989d52f194
- https://curve.carleton.ca/system/files/etd/49de4a69-0eb6-4c6b-a615-72064b5f31a3/etd_pdf/94612acfb1aff511b6662c950f790b41/thakkarremainingusefullifepredictionofaturbofanengine.pdf
- https://escholarship.org/content/qt5ns8r3fs/qt5ns8r3fs_noSplash_7d463f0d432194b63917f89e056f2644.pdf
- https://ieeexplore.ieee.org/document/4711414
- https://www.rrighart.com/blog-gatu/sensor-time-series-of-aircraft-engines
- https://www.kaggle.com/mikenguyen1712/nasa-turbofan-predict-rul-using-lstm-and-ar-model
- https://yajasd.github.io/2018/06/04/Predicting-Engine-Failure/