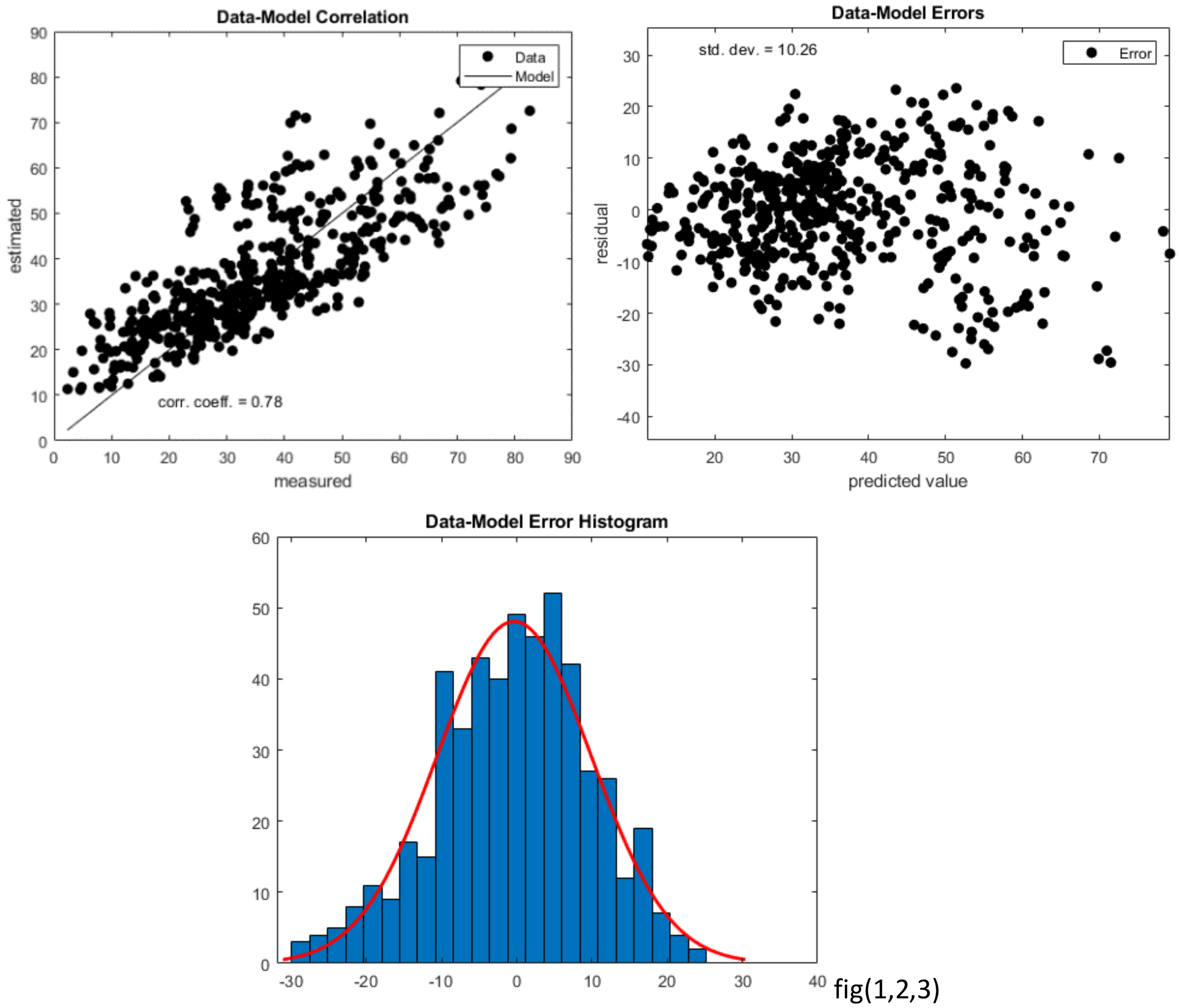
Reg No. 160140802

The Task set was to develop a mode capable of predicting the compressive strength of concrete after being trained on data providing the relationship between its constituent material components, its age and the resultant strength. In Pursuing this task, 3 data models are presented. A General Linear Model (GLM) which serves as a performance benchmark, a Multi-Layer Perceptron (MLP), and a model based on the Gaussian Radial Basis Function structure (GRBF). All 3 models where trained and on the data and validated using an out of sample set extracted from the data.

The Data

The Data presented consists of 9 data columns, 8 of which were model inputs consisting of real value data, with the final column being the model output, The input data set was all normalised to constrain the magnitude of the numerical operations performed and to show variance about the mean operational values. Afterwards the data set was split into 2 parts of similar size, one for training the model ,and an out of sample data set for validating the model, each point in the new data sets being chosen at random at the start of the program. This means different data points are chosen upon every single run, leading to slightly different results being produced on each run.

After preparing the data, the GLM was initialised and run using the normalised sum of squared errors and an indicator of its performance. Three data plots were produced showing the error and correlation between the model output prediction and validation data set output.



fig(1,2,3)

The GLM benchmark model was able to produce a model capable of a 0.78 correlation to the out of sample output data and error with 10.36 standard deviation.

The MLP

Given the high dimensionality of the input data set, It would have been unwise to attempt to apply a polynomial model as it does not scale very well leading to the number of regressor terms being very large and increasing the complexity in terms of both the number of required samples as well as computational expense. Another issue raised by the data set is its black box nature, as the data structure as well as the nature linearity of the relationship between input variables and the output are unknown. This generated the need for a model with the ability to generalise itself to any structure. Knowing this as well as understanding the universal approximation capabilities of MLPs, they were the natural choice.

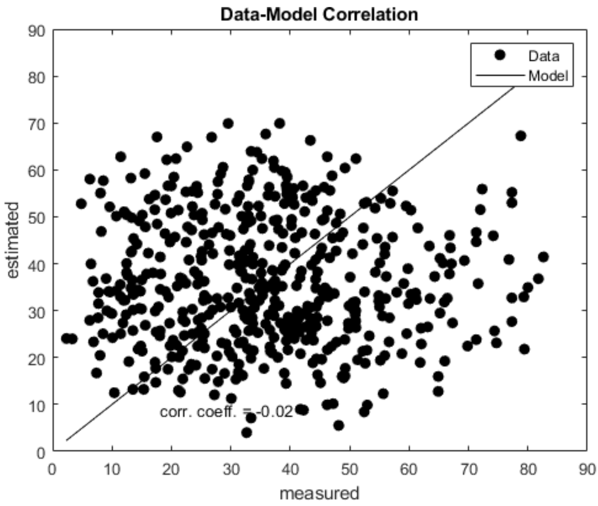
However, due to this lack of pre-knowledge about the characteristics presented by the data, the MLP was set up such that it had very high flexibility with 20 hidden layer elements so as to maximise the model's ability to conform to the data. This however creates a very complex model. In order to control the complexity of the model, it was regularised using L2 regularisation where very large weights are penalised minimising their occurrence hence reducing the model complexity.

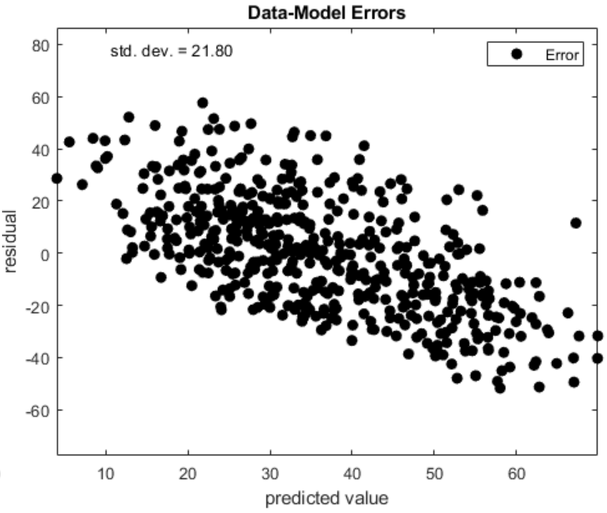
In order to find the appropriate value of the regularisation term, rho, 2 grid searches were run, first a coarse search, using only 5 values logarithmically increasing values between -3 and 3 to allow for the range of possible value to narrow before applying a finer search with 50 linearly increasing values falling between -1 and 1. The model was then trained using each of the value until the one which produce the minimum normalised least squared error between the model output and the out of sample output data was chosen. This value was then used to in training the final version of the model.

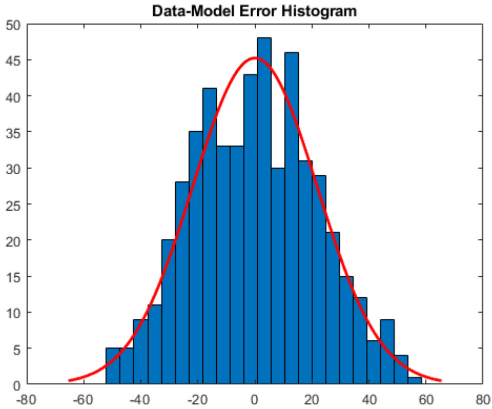
Model Validation

The model was validated using the K-fold Cross Validation technique which sub divides the training data into K sub sections then trains the model on most of the subset while reserving one for validating , before replacing it back into the set and repeating for each set. I utilised a k = 3 fold model where the cross validation was run for each value of the rho regularisation parameter as the data set is rather large. Also, in trying to validate the choice of an MLP model, a GRBF was also trained on the data set so its performance could be compared to that of the MPL.

Firstly, the GRBF model produced results inferior to both that of the MLP model and GLP benchmarks with a correlation value of near 0 and error standard deviation of over 20.

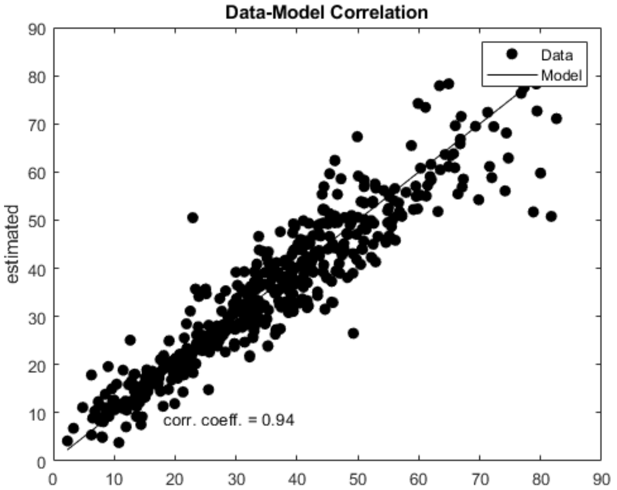


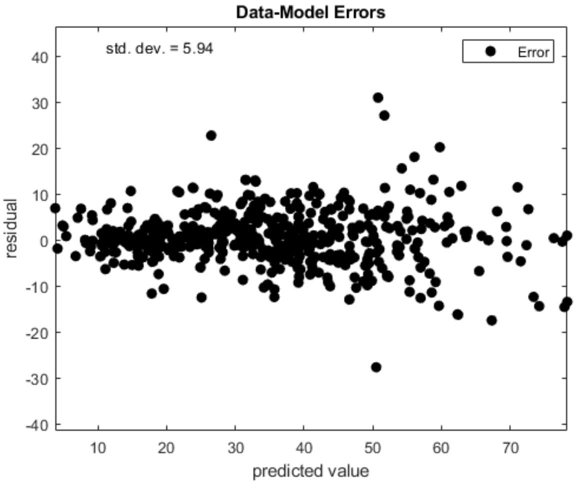


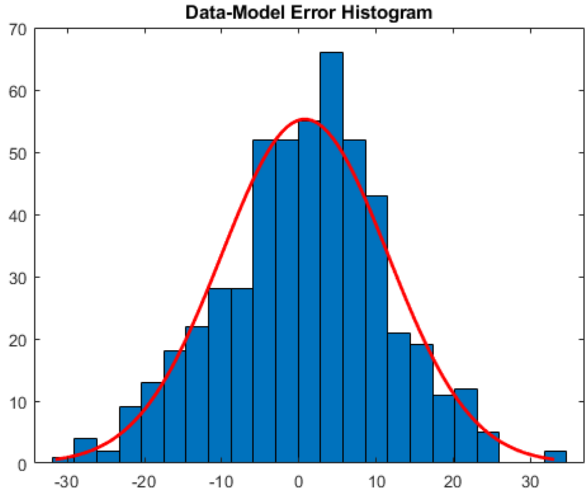


fig(4,5,6)

The MLP model was able to produce results of a higher correlation of upwards of 0.9 in many cases and with an error standard deviation of less than 6 throughout. Error as used throughout the investigation was measured using the normalised sum of squared errors.







This high level of correlation between the trained model when exposed to unseen data shows the model has successfully been able to capture the system dynamics presented by the data. In a formal setting the model might be considered one of high fidelity and would be presentable for consideration in an industrial endeavour.