Predicting Life Expectancy Report

Data Exploration and Analysis

Jack Sydenham - s3841816

This report details a comprehensive study undertaken to predict human life expectancy based on a variety of regional attributes. The endeavour begins by scrutinizing a dataset comprising approximately 2000 instances, each described by 20 distinct features. The objective is to employ machine learning techniques to establish a model that can accurately predict life expectancy, utilizing regression analysis as the principal methodological framework.

Exploratory Data Analysis

Initial data exploration is a crucial step in our machine learning workflow, enabling us to understand the dataset's underlying characteristics. Through careful analysis using histograms for each feature, insight was gained into the data's distribution, identifying key patterns and potential outliers crucial for informed modelling.

Our examination revealed varied feature distributions: some exhibited close to Gaussian-like symmetry around the mean [Figure 1.1], suggesting compatibility with linear model assumptions and possibly requiring minimal preprocessing. In contrast, other features showed significant skewness [Figure 1.2], indicating the presence of outliers and non-uniform distributions. Such skewness implies that without proper adjustments, these features could bias the model, affecting its predictive performance.

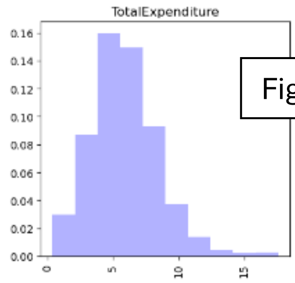

Figure 1.1

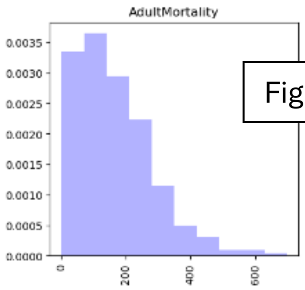

Figure 1.2

This diversity in distributions guided our preprocessing strategy, highlighting the need for tailored approaches to ensure data compatibility with linear modelling techniques and improve prediction accuracy.

Box plots were then used to analyse the spread of the data on each attribute and spot outliers. It was found that on many attributes, many outliers were present. To remedy this, the IQR method was applied to drop the exterior 30% of values from all numerical attributes. This is a good start to generalise the given data. Other methods will be used for further processing. The before [Figure 1.3] and after [Figure 1.4] of this application on attribute 'Diphtheria' are as follows:

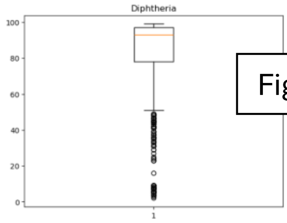

Figure 1.3

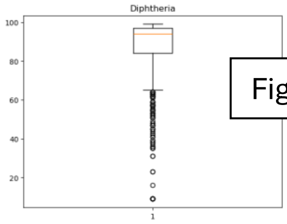

Figure 1.4

Feature Scaling

Feature scaling, a vital preprocessing step, standardizes the range of dataset features to improve algorithm performance. Our methodology applied Min-Max scaling to features with symmetrical distributions, adjusting their values to a [0, 1] scale to ensure uniformity in scale across all features. A Power Transformation (Yeo-Johnson method) followed by Min-Max scaling to the same scale of [0, 1] was applied to those with skewness, aiming to mitigate outliers and normalize distributions.

We can measure how much variance in the dependant variable of life expectancy is predictable by our model based on the independent variables using the coefficient of determination, R^2. A higher R^2 indicates a larger portion of variance in life expectancy, hence, a more effective model.

Utilizing this scaled data, a baseline linear regression model was trained, valued for its straightforwardness and interpretability. This initial model, assessing the scaled features, achieved an R^2 score of 0.739, indicating a substantial proportion of life expectancy variance explained by the model. Conversely, training with unscaled data yielded a lower R^2 of 0.725, underlining the advantages of feature scaling in model efficacy.

Initial Modelling and Potential Improvements

For the baseline model, we selected linear regression, a choice grounded in its simplicity, interpretability, and efficiency. This model serves as a benchmark due to its widespread application and the linear relationship it assumes between features and the target variable, which, as suggested by the exploratory data analysis (EDA), applies to several features in our dataset.

In enhancing our model, we'll prioritize exploring regularization techniques to mitigate overfitting and emphasize informative. Concurrently, establishing a robust cross-validation strategy is paramount for validating model performance and ensuring its generalizability to unseen data. These steps collectively aim to refine our model's accuracy and reliability, addressing initial limitations and leveraging insights gained from our exploratory analysis.

Regularisation

The most appropriate first step improving our model is regularisation. Through this, we aim to avoid overfitting through generalisation of our training data, not allowing our model to become too complex. We want to generalise our training data so that predictions aren't made too specific through a basis of outliers and other noise.

Ridge regularisation (L2) will be implemented to help maintain model simplicity by applying a penalty that uniformly shrinks the coefficients, in turn retaining all features but reducing their potential to individually dominate the prediction. The use of L2 regularisation is motivated primarily by the fact that our data contains many attributes which harbour small but meaningful effects, and we aim for our model to capture essential patterns in our data without becoming too complex. After regularising with a strength value of 1.0, our R^2 value sits at 0.707. a significant drop, but an appropriate loss considering the importance of generalising our data.

Hyperparameter Tuning

For our model, the only hyperparameter that will be tuned is the 'alpha' variable of our regularisation function, which in ridge regularisation, directly influences the strength of the regularisation. If we were using a combination of both L1 and L2 regularisation, the 'l1_ratio' variable could be tuned, controlling the mix between L1 and L2 regularisation techniques, however of course, this is not the case, as we are using L2 alone.

Focussing only on the alpha value for our regularisation is both impactful and efficient for our model, as we are able to test a vast range of strengths to find an optimal value quickly, despite us working with a large data set which requires much computation.

To efficiently find an optimal value for alpha, we implement a broad range of potential values through the use of NumPy's logspace function to generate a set of alpha values evenly on a logarithmic scale from 10^{-6} to 10^{6} . This way, we are able to assess a wide range of regularisation strengths from very weak to very strong, ensuring no potential optimal value is left unnoticed. Through this, we improve the model's ability to make accurate predictions on new data it hasn't seen before, ensuring it's not weighed too heavily by any unusual or extreme values found in the training data.

Through implementation of this optimisation, our tuning techniques revealed an optimal alpha value of 1e-06. This indicated that regularisation in fact has little effect on our dataset at this stage. While this came as a surprise initially, a logical explanation can be aligned with this outcome. This finding suggests that the dataset, along with the preprocessing steps made earlier through outlier management and feature scaling, appear to have effectively mitigated the risk of overfitting, thereby practically nullifying the effects of regularisation. This in turn suggests that our data and the relationships among the variables have been sufficiently captured by the model already and there is no need for additional complexity. This alpha value being so close to 0 reveals that our model's current complexity is already well-tuned to the patterns in the data, and stricter regularisation would not lead to any greater predictions.

Validation

To validate our regression model, a combination of ridge regularisation and K-Fold cross-validation was employed. This approach involves splitting the training set into five distinct subsets, using each to evaluate the model trained on the remaining four.

Upon evaluation, our model demonstrated very consistent predictions across the five data folds as follows; 0.760, 0.736, 0.722, 0.722, 0.759, leading to an average R^2 score of 0.740, suggesting our model explains around 74% of the variance in life expectancy across the given data folds. This tells us that our model is well fitted, capturing a large portion of the underlying data without being too complex or overly fitted to the training data.

On top of this, a Root Mean Squared Error (RMSE) was calculated to provide a secondary perspective to our model. With a result of 4.495, we can say that on average our predictions of life expectancy deviate by about 4 and a half years. This shows that we can be confident that our model will consistently predict the life expectancy within a range of about 4 and a half years.

Evaluation

In evaluating the performance of a model, we are really testing how well it is able to make predictions for unseen data. The test set is used as a basis for this evaluation, providing a solid measure of how well the model generalises.

The efficacy of our predictions is measured through comparison between the generated life expectancy values and the actual life expectancy values. This direct comparison is crucial in validating the accuracy and consistency of our model. A fundamental measure of our model's success is how closely the predictions align with real-world values; the closer they align, the more effective our model is. We utilise metrics such as R^2 score, quantifying the proportion of variance in life expectancy produced by the model, and RMSE, measuring the average magnitude of the model's prediction errors, to quantitatively measure the model's predictive efficacy. This way, we ensure that our model's predictions are not only theoretically effective, but also practically.

The Ultimate Judgement

In closing this study with an ultimate judgement on the Ridge Regression model developed, we must first reflect on the model's alignment with our analysis's foundational assumptions and its demonstrated performance through the lens of the predictions made. Our Ridge Regression model was selected for its ability to manage relationships within data featuring a multitude of predictors; a capability that was especially important for us, considering the model's need to handle a wide range of attributes impacting life expectancy.

Looking at the predictions made by our final model, the Ridge Regression model's ability to predict life expectancy with a commendable degree of accuracy is evident. The model's predictive efficacy, evidenced by a clear balance between predicted and actual life expectancy values, attests to its adeptness at capturing the essence of the underlying data and its relationships. This alignment of predictions made both makes clear the model's practical utility, and validates our analytical approach centred around linear assumptions.

The model's performance metrics are key to our ultimate judgement. The combination of the R^2 score and RMSE collectively prove our model's reliability and power. These metrics highlight our model's ability to both grasp and accurately represent the linear relationships within the given data, which is a core assumption driving our analytical strategy. We can also be assured of our model's resilience against overfitting, as our application of Ridge Regression, along with regularisation work to enhance its generalisability for unseen data.

Because of these findings and justifications, our decision to favour the Ridge Regression model as the optimal choice for predicting life expectancy is clearly rooted in a comprehensive evaluation of its predictive efficacy. This model, through its theoretical foundations and practical validation, has proven to be well-suited for our objective, encapsulating the linear dynamics of the data while ensuring a reliable projection of life expectancy based on all given attributes. This ultimate judgement then is not only a reflection of our model's success for this dataset, but an indication of its potential application to real-world scenarios, where predictive analysis can drive decision-making processes.