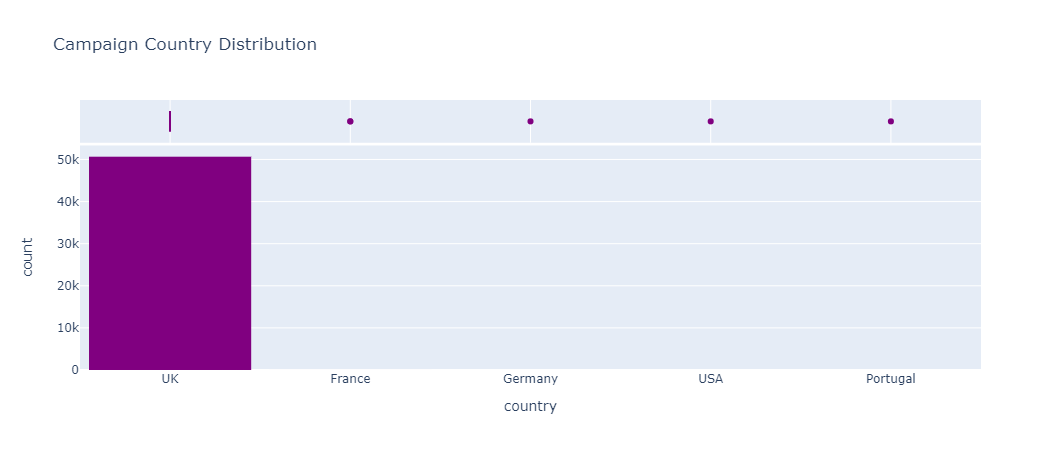
**Customer response classification model**

**Business Understanding**

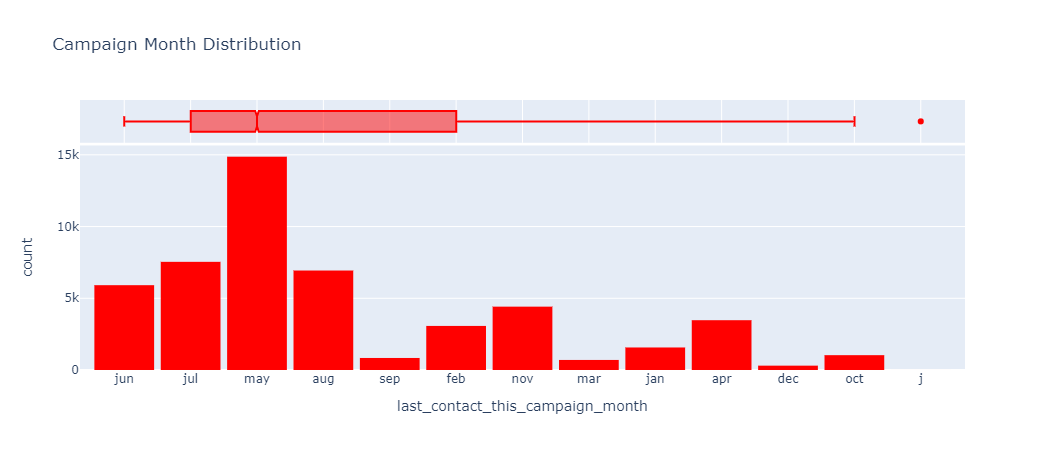
The company Stircom intends to identify which are customers are more likely to respond to a current study based on the data from previous studies. A solution to this problem will help the company narrow the scope for the customers to conduct for the current study. The provided data, for more than 50,000 respondents, covers customers from a wide geographical range and some from different countries. Traditional methods of analysis are likely to take more time and resources while trying to identify the most suitable customers for the current study. Hence, a machine learning approach to predict the likelihood that a customer will respond to a call is a more suitable approach. The provided business problem is a classification problem in which customers will be classified in two categories; whether they will be given a new contract for the current campaign or not. A bivariate or multivariate regression approach can also be applied to the problem upon selecting key variable that have a suitable correlation. The machine learning model will be able to classify new customer data provided to it as to whether customer is fit for a new contract or not. This model will frequently use the terms model, variable, hyperparameter tuning among other machine learning jargons. A model is a mathematical and or scientific approach to solving a conventional problem. A model is trained on variable (features), which are the dependent and independent variable for fitting the model. Hyperparameters are model tuning sub-variables that are used to tweak the model’s performance and accuracy. This report is prepared for academic purposes and it is in no way violating the use of personal data. Data protection and copyrights are fully reserved to stircom. It is to the understanding of the researcher that the data herein was collected in an ethical manner without violation of civil rights of any individual.

**Data Summary**

ID: unique identifier for this record; this is a record of each of the customers that contacted in the campaign. The ID might not have key significance in data modelling as it is used to identify records in relational databases. Town: home town for customer; identifies where the customers hails from; might be a key feature since customers from certain could have a higher likelihood of responding than customers from other towns. Country: country for customer's home address; most of the customers are from the United Kingdom, hence this may not serve as a significant variable in the modelling process. Attached here is the visualization for this variable.



Age: customer's age – a continuous variable and job -a discrete variable: customer's job, these variables have a business and statistical significance will be useful in the training of the model. Married: customer's marital status, education: customer's highest educational qualification level obtained. Like the former two features, these two features are of a significant importance to determining whether customers of a certain educational background or marital status should be given a new contract in the current campaign. Arrears: has the customer failed to pay a recent bill, current balance: current amount in customer's landline account in pounds. These two features – discrete and continuous, respectively, are important for the modelling since a customer who has pending arrears on their current subscription is more likely to extend this trend even into the new contract, and hence, it could be statistically justifiable to deny them a contract on the new campaign. Similarly, a customer who a good balance on their current subscription is more likely to pay for a new subscription since they afford the previous package comfortably. Housing: is the customer a homeowner; houseowners are more likely to take long term subscription even for new packages as compared to non-house owner who are more likely to move with uncertainty. Has\_tv\_package: has the customer got an additional TV and data package on their landline. This is a discrete variable that could be useful for modelling the problem as an additional subscription. Last\_contact: type of communication used for previous call to customer; a discrete and ordinal variable - the type of communication is important since some customers may no longer be reachable using certain methods of communication. Conn\_tr: connection type grouping ID (related to the connection for their landline); a discrete variable for the connection type; key variable since it may be easier to establish connection to a customer using a certain type of connection than other types; hence this will be a useful feature in predicting whether a customer should be a contract call for the current campaign. Last\_contact\_this\_campaign\_month: last contact month of year, last\_contact\_this\_campaign\_day: last contact day of the month, these two continuous variables describe the specific day and month the customer was contacted for a campaign. These two features did not take any known distribution in the visualization hence they may not as significant to the modelling process.

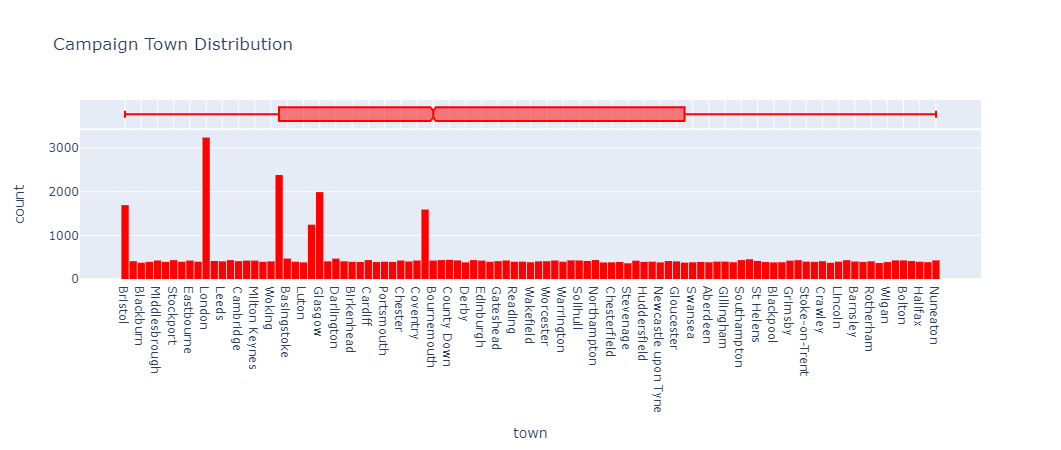


this\_campaign: number of contacts performed during this campaign and for this client, days\_since\_last\_contact\_previous\_campaign: number of days that passed by after the client was last contacted from a previous campaign, (-1 means the client has never been contacted before), a continuous variable that provides insight on whether the company has contacted the customer since they participated in the previous campaign. Actively contacted customers are statistically more likely to respond to a call from the company. Contacted\_during\_previous\_campaign: number of contacts performed before this campaign and for this client. An ordinal feature that gives the number of times a customer has participated in a previous campaign. Outcome\_previous\_campaign: outcome of the previous marketing campaign, a discrete variable for whether the customer was preferred in the previous campaign or not.

new\_contract\_this\_campaign: has the client taken out a new contract. This is the target variable, it is a discrete variables that states whether the customer contacted, whose data is in the variable is best suited for a new contract or not. The model will be trained and tested to predict the accuracy with which the customer can be assessed and given the contract for the new campaign.

**Data Preparation**

Most machine learning models require that the data using for training and testing the model, be numerical arrays of either one or two dimensions depending on the specific model. Hence, the first data preparation method was determining the most suitable way of converting non-numerical data columns into n-dimensions numpy arrays. However, it is important to visualize categorical data before embarking on the encoding and imputing techniques. The visualized data did not contain any missing values per se and hence the modelling process saw no importance of using imputation techniques to fill up missing values. However, features that did not present any statistical distribution on visualization were deemed insignificant to the modelling process and the respective columns were deleted. This included columns for the ID, the country, month for last contact and day for last contact. In addition, the towns columns were as well deleted as most town had a generally similar response across the different location in the United Kingdom.

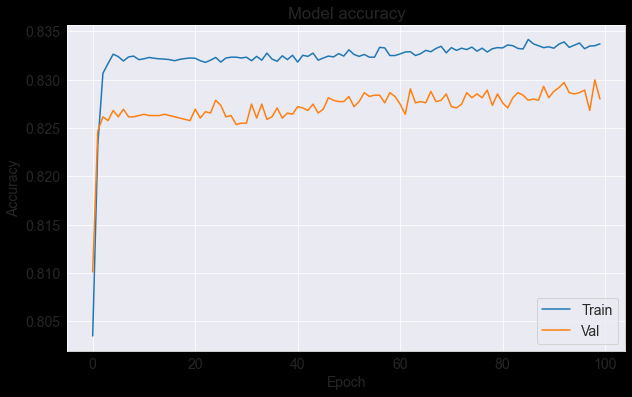


Most of the variables that were visualized did not have any outliers thence eliminating the need for outlier removal. However, during the model training, the StandardScaler was used to scale the data to a usable range for machine learning models. This is an important technique as it eliminates statistical constants such as the mean and median from the data – this ensures that training the model is not biased on the mean or median. This technique is also important in reducing the total time required to train the model - standardized data is faster to process at it in smaller ranges and lower dimensions.

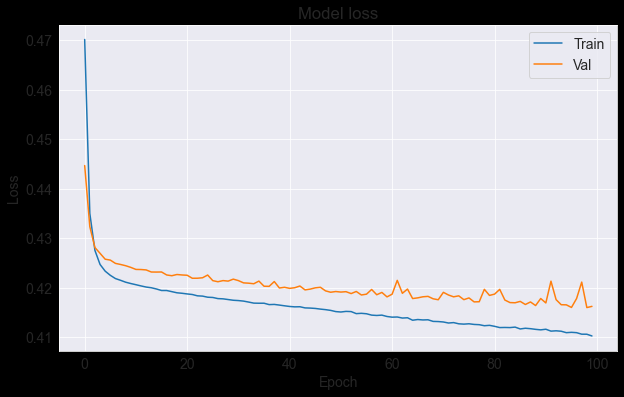
**Modelling**

***Neural Networks models***

Neural Networks prediction models perform relatively well in classification modelling. Using a specified loss, they traverse through the dataset to classify the inputs into feature categories based on weights, depth of training, regularization parameters and number of epochs. The model used 70% of the data for training the model and 30% testing the model. Of the 70% training data, 50% percent was used to model evaluation while the remaining 50% was used specifically for model training. A validation set is important to ‘test’ and tune the model before the final testing on an unknown ‘testing set. An initial neural network was fitted to the training data and produced a root mean squared error of 0.5051. A regularization parameter was used to fine tune the model for a better score. Regularization prevents the model from overfitting – it maintains a fair value for the variance and bias tradeoff. Fitting the model with this hyperparameter significantly improved the score to 0.3761.



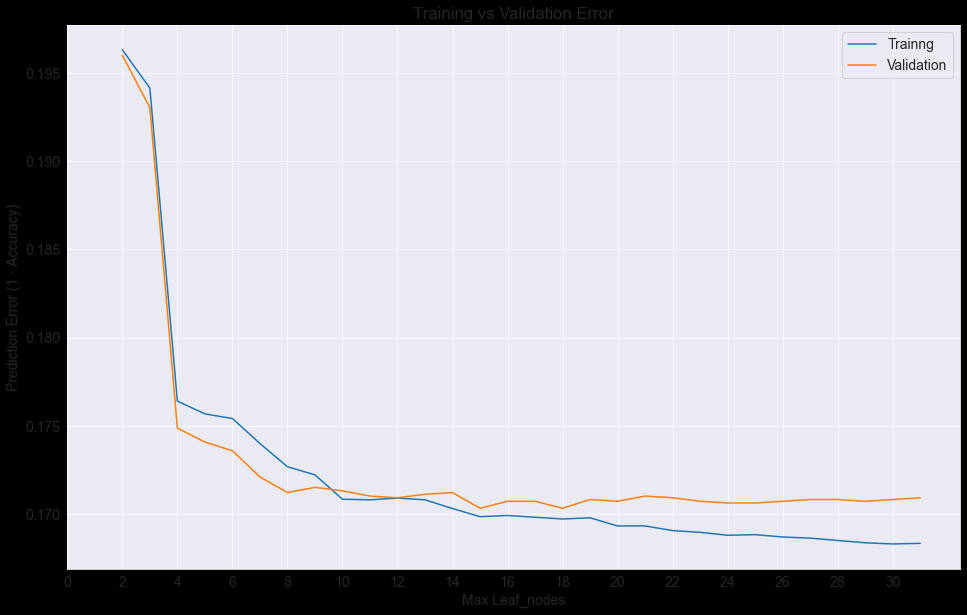
Model accuracy on the training and validation data.



A visualization of the loss minimization for the ANN model.

***Random forests models***

As a tree-based machine learning model, the RF model is a collection or trees that split the data and vote on the best possible classification of the data as per the provided features. An RF provides a feature importance function that was used in this modelling to determine which features were mostly used in taking the classification vote. For this model, 80% of the data was used for training the model of which, 50% was used as validation set for training time testing of the model. The base model scored 99% on the training set and 86% on the validation setting. An insignificantly high is an implication for model overfitting which calls for hyperparameter tuning to improve model performance. Key hyperparameters for the RF model are; the number of leaf nodes – the number of leaf nodes that will be on each tree on the RF, a significant value should minimize the loss function – passing an a range of values to the model and selecting the one with the least loss was the approach used in this mode, number of estimators – this is the number of trees in the RF, a low values could result in underfitting and low performance and a high value could lead to overfitting and more training time – passing a range of values and choosing the best was the approach used herein. The maximum traversing depth is how far each tree is traversed before a vote is made; also, a range of values was passed to the model and a suitable one chosen for the final model. The model scored an accuracy of 82% using the confusion matrix evaluation metric. Using the sklearn classification\_report method, the model scored 83% accuracy with a precision of 83%, recall of 99% and F1-score of 90%.



Leaf nodes and their prediction error. A value of 4 was chosen to train the final model.

**Linear Regression Models**

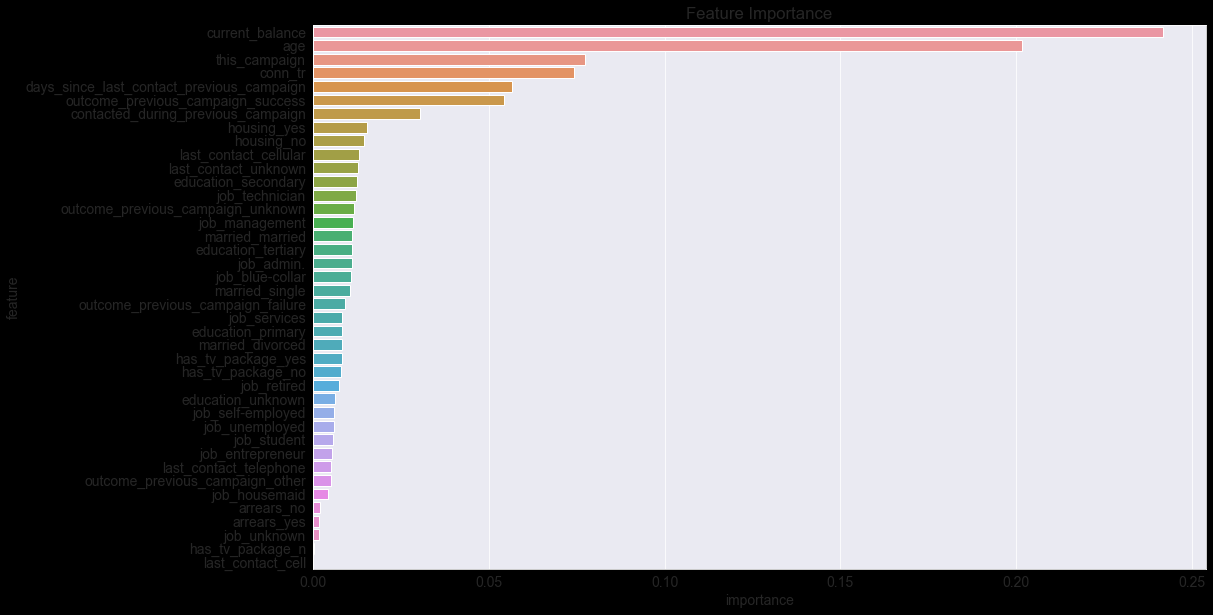
Three regression models were trained for this problem and the performance of each was assessed. A traditional linear multivariate linear regressor, Random Forest regressor and an Xtreme Gradient Boost regressor (XGB). The metrics for the models were mean squared error, root mean squared error and mean absolute percentage error. The linear regressor had a RMSE of 2.059, while the XGB regressor scored 0.13 and finally the random forest regressor score a RMSE of 0.1.

Clearly, of the three models, the linear regression models are seen to have outperformed the DNN (Deep Neural Network) and RF model. However, it important to note that the further tuning and feature engineering could significantly improve the performance of the RF model and the DNN model. Also, the method of encoding used for this model (ordinal encoding) might have affected the performance of the DNN model. Use of a different approach to encode the data could as well boost the performance of this model. Using the most important features to retrain the RF model could as well significantly improve its performance. The RF model stands out as the best approach to solving despite having been outperformed by the regressor models. This is so because the model outrightly classify a customer as fit the contract or not despite the numerous unneeded features that were used to fit the model.

**Evaluation**

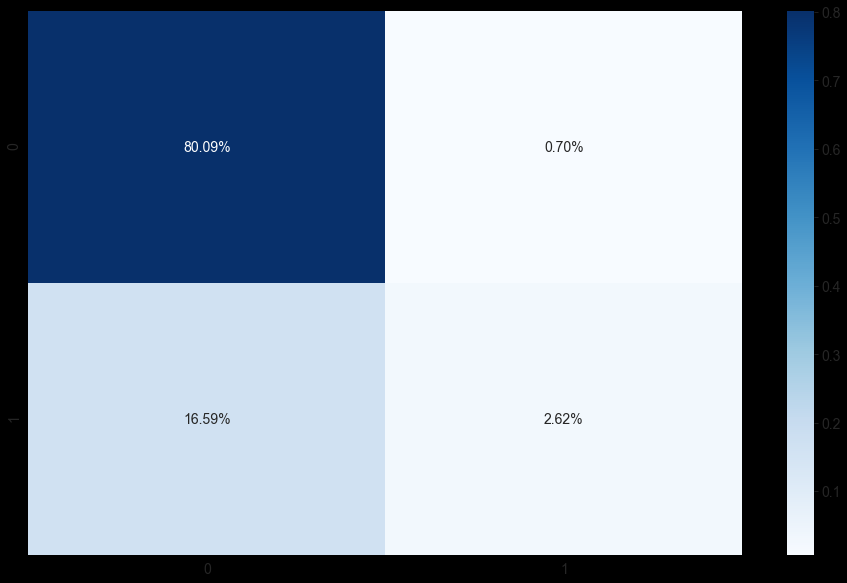
As for the RF model, a plot of the most useful features implied that most of the features used in the training process were insignificant. This is a likely huge cause of loss for the model that performed significantly well with a score of 82%.

The model in contrast to the other modelling approaches show a very minimal loss between the training and the validation data of the model. Using the proba\_ function, the model can give probabilities and classifications for each individual user input variable data the model is tested with.



Feature importance visualization. The current balance on subscription and the age had the highest weight in making the classification vote for the RF model.

Visualized here are the features used to training the RF model and significant they were in determining the final classification vote. A decrease in the accuracy could be affected by a significantly high number of features that have little or no essence to predicting power of the model. This is so because the model traverses each leaf node and creates a huge number of trees with the maximum depth passed to model at the expense of training time and accuracy.



A confusion matrix for the classification of the RF model.

From the confusion matrix, the model correctly determined 80.09% values as true positives, hence correctly classified. A 2.62% of the data previously classified as negative, was classified by the model to be acceptably contractable customers. Only 0.70% of the provided was wrongly classified by the model as acceptable truth. [of 50,000 customers, this represents 350 customers], although significantly small, the report would recommend more screen to reduce the number of customers who might be wrongfully denied the contract.