DATA ANALYTICS USING R - PREDICTIVE ANALYSIS

Introduction

This section requires a predictive analysis using multiple models for comparison. Multiple linear regression and a binomial logistic regression model was used to carry out this analyses. For the multiple linear regression, an outcome variable was chosen on the several predictors. While, the binomial logistic regression model which is common for classification problems a binary variable was selected. The findings are then used to facilitate data-driven recommendations for real-life decision-making processes.

Procedures of Multiple Linear Regression Model

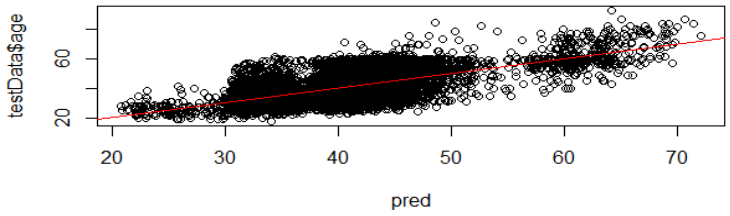
Multiple linear regression is a wing of simple linear regression that predicts an outcome of a continuous variable on the basis of multiple distinct predictor variables (James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. *An Introduction to Statistical Learning: With Applications in R*). For this reason a multiple linear regression model was conducted on the dataset. The outcome variable used was age, therefore, the analyses was the effect of age of customers on all other variables.

The initial step involved building a full model that had all the variables to determine the F-statistic and the relating to the P-value which is at the tail end of the summary. The overall result of the full model showed an R^2 or the coefficient of determination of 42%, the $F(42,45168) = 777$, $P<0.001$ which is highly significant. This means that, at least, one of the predictor variables is significantly related to the age. In order to reduce the parametric, the dataset needed to be spilt into two. The TrainData takes 80% and TestData takes 20%. Following this, a reduction in parameter was necessary to attain the significant variables for the analyses. The Stepwise involves the selection of independent variables to be used in a final model. It involves adding or removing potential explanatory variables and testing for statistical significance after each iteration (Hayes, A. (2022). The final model which from the step wise regression showed a coefficient of determination of 42% and $F (36, 45174) = 906$, $P<0.001$ which is significant. The final model showed 10 significant variables

Consequently, a model comparison was done to attain the better fitting model. The Bayesian Information Criterion also BIC, is an index that is used to compare the different regression models to identify which is the most appropriate (Zach (2021) How to calculate BIC in R, Statology). The BIC for the full model shows 317821 while the final model shows 317770.2. The Bayes factor using the Wagenmaker shows 105797348274

this means that the stepwise model is 105797348274 times more likely to fit the dataset than the full model.

The co-efficient of determination or R-squared for the final model on the TestDate which is the unseen data on the outcome variable (age) is 0.44 which explains that the 44% of the variance in Age was explained by all the 10 variables. Finally, the relationship between the predicted age and the actual age was attained. Based on the figure 11.1, the estimated regression line is the diagonal line in the center of the plot. Because each data point is quite close to the projected regression line, we may conclude that the regression model fits the data reasonably well.



Diagnostic plots.

Based on the test result, there were four different diagnostic plots which were derived to visually represent the data

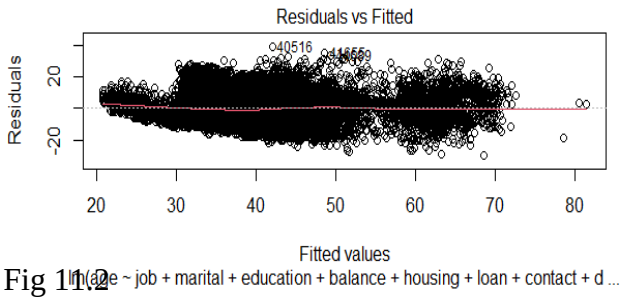


Fig 11.2

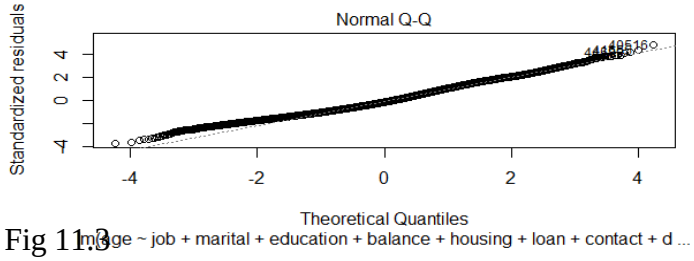


Fig 11.3

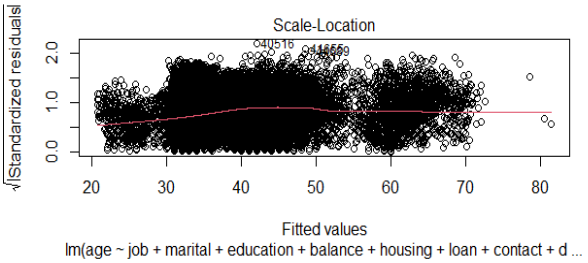


Fig 11.4

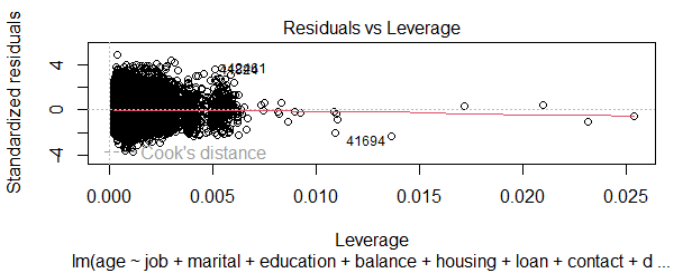


Fig 11.5

The diagnostic plots above showed the following;

1. Residuals vs Fitted (Fig 11.2) – The plot shows a horizontal line without any unique pattern which indicates a good linear relationship between the predictors and age.
2. Normal Q-Q (Fig 11.3) - It is a fairly normal distribution as the residual points follows a straight line.
3. Scale-Location (Fig 11.4) - This plot shows that the residuals are spread equally along the ranges of predictors which is a good indication of homoscedasticity.
4. Residuals vs Leverage (Fig 11.5) - The plot above highlights the top 2 most extreme points (#41694, #42241, with a standardized residuals below 0.

Procedures of Logistic Regression Model

Multiple linear regression predicts the outcome of continuous variables however, logistic regression predicts the outcome of categorical variables. Logistic regression or binomial logistic regression is used to predict the class (or category) of individuals based on one or multiple predictor variables (x). It is used to model a binary outcome that is a variable, which can have only two possible values: yes or no, diseased or non-diseased. (Kassambara et al. (2018).

In this analyses uses the 'Y' categorical variable which shows if the client has subscribed to a term deposit. In order to have a more realistic predictive model, duration needed to be discarded because it severs no purpose when using Y which means that only 16 variables were considered for this analyses. Following this, the dataset was spilt into a TrainData used to train the dataset observation and TestData used for testing. Splitting improves the accuracy of the dataset and helps to avoid overfitting. Also, a full model was built which had all the variables. The overall result of the full model showed a

PseudoR 0.16 using deviance which means that 16% of the variables observed is explained in the model. Pseudo-R2 measures the goodness of fit of categorical variables. Also the stepwise model which shows the most significant variables showed a PseudoR 0.16 using deviance which means that 16% of the variables observed is explained in the model and reveal 8 significant variables after reduction. A model comparison was done to attain the better fitting model, the BIC showed that the stepwise model is 1.57 x 10^19 times more likely to fit the dataset than the full model.

Interpreting the odd ratio of the final model estimates with the categorical variable. An odd ratio represents the ratio of the odds that an event will occur given the presence of the predictor as supposed to it occurring in its absence (Kassambara et al. (2018).

Odd Ratio Interpretation – Final Model

Variables	**Estimates**	**Coefficients**	**Interpretation**
maritalmarried	-8.814e-01	0.8019165	when there is a married client there is a 20% decrease in the likelihood of term deposit being yes
balance	2.015e-05	1.00002	when there is an account balance there is a 100% increase in the likelihood of y being yes
housingyes	-5.894e-01	0.5546499	when there is a housing loan there is a 45% decrease in the likelihood of term deposit being yes
loanyes	-4.063e-01	0.6661394	when there is a personal loan there is a 33% decrease in the likelihood of term deposit being yes
contacttelephone	-2.331e-01	0.792041	when contact by telephone there is a 21% decrease in the likelihood of term deposit being yes
contactunknown	-1.382e+00	0.2509958	when the contact is unknown there is a 75% decrease in the likelihood of term deposit being yes
monthaug	-8.064e-01	0.4464598	when there is contact In August there is a 56% decrease in the likelihood of term deposit being yes
monthdec	6.146e-01	1.84893	when there is contact In December there is a 84% increase in the likelihood of term deposit being yes
monthfeb	-4.526e-01	0.6359646	when there is contact In February there is a 37% decrease in the likelihood of term deposit being yes
monthjan	-1.089e+00	0.3365347	when there is contact In January there is a 67% decrease in the likelihood of term deposit being yes
monthjul	-7.399e-01	0.4771744	when there is contact In July there is a 52% decrease in the likelihood of term deposit being yes
monthmar	1.118e+00	3.058438	when there is contact In March there is a 200% increase in the likelihood of term deposit being yes

monthmay	-4.947e-01	0.609753	when there is contact In May there is a 39% decrease in the likelihood of term deposit being yes
monthnov	-8.732e-01	0.4176035	when there is contact In November there is a 58% decrease in the likelihood of term deposit being yes
monthoct	7.105e-01	2.035097	when there is contact In October there is a 100% increase in the likelihood of term deposit being yes
monthsep	6.526e-01	1.920595	when there is contact In September there is a 90% increase in the likelihood of term deposit being yes
campaign	-8.541e-02	0.9181329	when there is contact In Campaign there is a 10% decrease in the likelihood of term deposit being yes
poutcomesuccess	2.144e+00	8.531895	when there is contact In outcomesuccess there is a 750% increase in the likelihood of term deposit being yes

Confusion Matrix for Model Performance

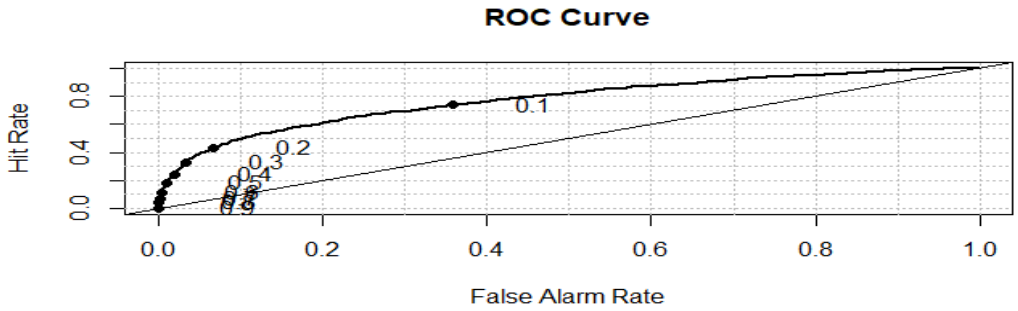
A confusion matrix is used to define the performance of the used model. Based on the matrix outcome on R where 0 – not default and 1 – default, the tables shows the confusion matrix of the full model.

	Not Default	**Default**
Negative	7887	81
Positive	878	196

The full model showed a poor sensitivity model performance of 18% which led to plotting the ROC and applied a cut-off of 0.1 in order to increase sensitivity.

Receiver Operator Characteristic (ROC) Curve

A ROC plot The ROC curve shows the trade-off between sensitivity and specificity. Based on the curve below, the best balance between sensitivity and specificity is 0.1. It is also the best threshold to maximize sensitivity which did not do well.



Following this a final confusion matrix was conducted and it showed that the matrix successfully predicted 5104(True negative) cases that the not default when they really did not. While, it recorded 2864(False positive) cases where they did not default when they actually defaulted. They were 282(False negative) cases were they did default but the algorithm showed not defaulting, while, 792(True positive) cases where it successfully predicted.

Final Confusion Matrix table

	Not Default	**Default**
Negative	5104	282
Positive	2864	792

Model Performance of Final Confusion Matrix

- Accuracy- This indicated that 65% of observations were correctly classified and 35% were misclassified. However, this matrix cannot be trusted.
- Sensitivity (true positive rate) – It indicated that only 73% of defaulters were correctly classified. The model performed well compared to 18%.
- Specificity (true negative rate) – this showed that 64% were correctly classified as non-defaulters. This perform well, however the previous matrix was higher. We traded it for more sensitivity, an increase in sensitivity means a decrease in specificity.
- Precision – It expressed 22% of positive classification were correct
- Recall – 74% chance that the defaulters were correctly classified
- F1-Score – It is a balance score between precision and recall, it showed 33%

Finally, the dataset was test for multicollinearity using variance Inflation Factor (VIF). Normally, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity because the higher the VIF the more standard error is inflated (Kassambara and U, M

(2018). Based on the results none of the predictors is correlated with other variables because they are well below 5

Diagnostic Plot for Logistic Regression Mode