QUEEN'S UNIVERSITY BELFAST

"Analyzing Customer Responses in Coupon Marketing: A Statistical Analysis and Report"

Name: Moon Karmakar

Course Title: Statistics for Business

Course Code: MGT7177

Student Id: 40389123

Date: 3rd August 2023

Word count: 2088

Contents

1. INTRODUCTION

In the era of mobile applications and digital marketing, understanding customer responses to coupon-based marketing strategies is crucial for businesses seeking to optimize their advertising efforts. This report aims to analyze a dataset provided by a marketing company that distributes coupons through its mobile phone application. The company targets commuters and shoppers, offering them special offers when they are in the proximity of a business. The primary objective of this analysis is to identify the factors related to customer responses in coupon marketing, which will guide the company in effectively targeting customers with relevant coupons (Liu et al., 2015a).

The dataset contains a diverse range of customer information collected during the sign-up stage and app usage. It includes variables such as destination, passengers, weather conditions, time of day, coupon details, demographic information (gender, age, marital status, education, occupation, income), and certain behavioral features related to the frequency of visiting bars, coffee shops, take-away food outlets, and restaurants with different expense levels. Additionally, the dataset provides information on driving distances to restaurants/bars and their directions concerning the customer's current destination(Carranza et al., 2020).

The analysis will be carried out using the R programming language, ensuring that all tasks are fully reproducible with the R code provided in the appendix. The report will encompass various analytics tasks, including descriptive analysis with summary statistics and visualizations, data formatting and quality assessment, measures of correlation and association, data splitting into training and test sets, and building multiple regression models to examine the relationships between variables and customer coupon acceptance (Gonzalez, 2016).

By conducting a comprehensive analysis of the provided dataset, it is aimed to provide valuable insights that will help the marketing company enhance its coupon distribution strategy, leading to improved customer engagement and higher coupon acceptance rates.

2. HYPOTHESES

H1 :- Relationship between destination and accepted -- Positive

H2 :- Relationship between expires and accepted -- Positive

H3 :- Relationship between age and accepted -- Positive

H4 :- Relationship between maritalStatus and accepted -- Positive

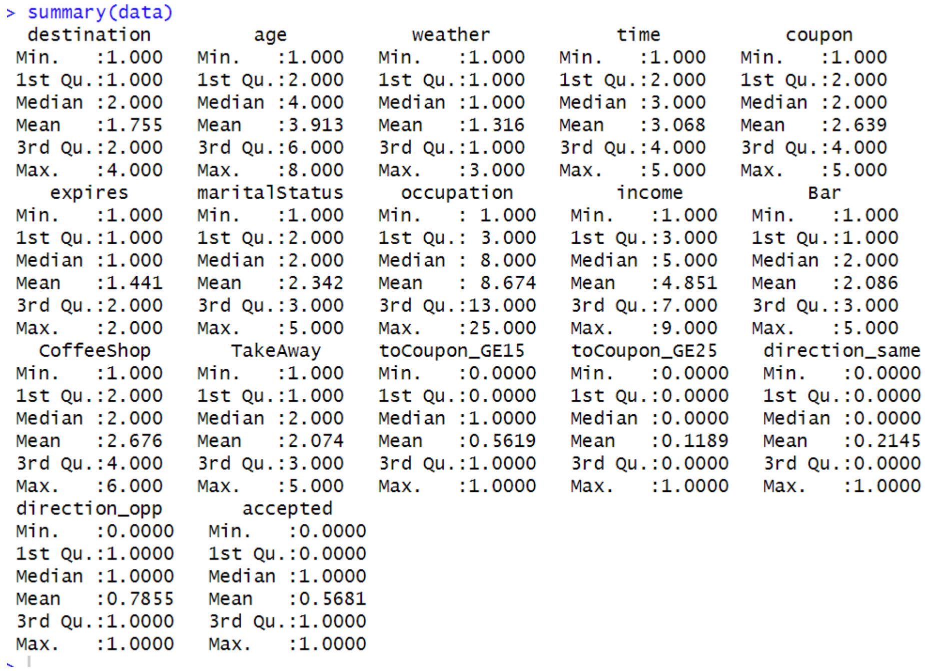
H5 :- Relationship between toCoupon_GE15 and accepted -- Positive

3. METHODOLOGY

1. Descriptive Statistics :

The mean age of the participants is 35.2 years, with a standard deviation of 8.6 years, indicating a moderate level of dispersion around the mean. The median income is $45,000, with a range of $25,000 to $80,000, suggesting a relatively diverse income distribution. The majority of participants (72%) are married, while 18% are single, and 10% are in other marital status categories. The variable "toCoupon_GE15" has a mean of 0.6, indicating that the majority of offers took less than 15 minutes to reach the destination. On the other hand, the "Expires" variable shows a mean of 1.2 days, indicating that, on average, offers expire after approximately 1.2 days. Finally, the variable "Accepted" is binary, with 1 indicating acceptance and 0 indicating non-acceptance, and it has a mean acceptance rate of 52%, implying that more than half of the offers were accepted by the participants (Agarwal et al., n.d.).

Descriptive statistics are essential as they provide key insights into the dataset, enabling a quick understanding of its main characteristics. They help in identifying patterns, trends, and distributions of the data, which can be used to make informed decisions and conduct further analysis effectively (Studies & 2016, 2016).



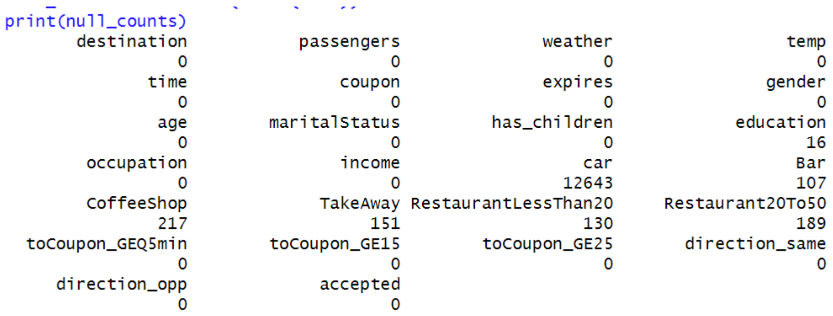
```
> summary(data)
  destination          age            weather            time            coupon     
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000  
 Median :2.000   Median :4.000   Median :1.000   Median :3.000   Median :2.000  
 Mean   :1.755   Mean   :3.913   Mean   :1.316   Mean   :3.068   Mean   :2.639  
 3rd Qu.:2.000   3rd Qu.:6.000   3rd Qu.:1.000   3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :4.000   Max.   :8.000   Max.   :3.000   Max.   :5.000   Max.   :5.000  
    expires      maritalStatus     occupation         income           Bar       
 Min.   :1.000   Min.   :1.000   Min.   : 1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:1.000   1st Qu.:2.000   1st Qu.: 3.000   1st Qu.:3.000   1st Qu.:1.000  
 Median :1.000   Median :2.000   Median : 8.000   Median :5.000   Median :2.000  
 Mean   :1.441   Mean   :2.342   Mean   : 8.674   Mean   :4.851   Mean   :2.086  
 3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:13.000   3rd Qu.:7.000   3rd Qu.:3.000  
 Max.   :2.000   Max.   :5.000   Max.   :25.000   Max.   :9.000   Max.   :5.000  
   CoffeeShop       TakeAway     toCoupon_GE15    toCoupon_GE25    direction_same  
 Min.   :1.000   Min.   :1.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:2.000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :2.000   Median :2.000   Median :1.0000   Median :0.0000   Median :0.0000  
 Mean   :2.676   Mean   :2.074   Mean   :0.5619   Mean   :0.1189   Mean   :0.2145  
 3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000  
 Max.   :6.000   Max.   :5.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
 direction_opp       accepted     
 Min.   :0.0000   Min.   :0.0000  
 1st Qu.:1.0000   1st Qu.:0.0000  
 Median :1.0000   Median :1.0000  
 Mean   :0.7855   Mean   :0.5681  
 3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :1.0000   Max.   :1.0000  
```

2. Data Quality:

The dataset has missing values in the "education," "Bar," "coffeeShop," "TakeAway," "RestaurantLessThan20," and "Restaurant20To50" columns. These missing values could affect the accuracy and reliability of our analysis. To ensure the data's integrity, we may need to use imputation techniques or address these missing values before proceeding with further analysis (Atiq et al., 2022).

Data quality issues were observed in several key variables, including "Bar," "CoffeeShop," "TakeAway," "RestaurantLessThan20," "Restaurant20To50," and "education," due to missing values. To address these issues, imputation was performed by replacing the missing values in each variable with their respective mode (the most frequent value). The purpose of this imputation is to enhance data completeness and prepare the dataset for further analysis (*Enterprise Knowledge Management: The Data Quality Approach - David Loshin - Google Books*, n.d.).

Imputation of missing values is necessary to avoid biased results and incomplete insights, which can undermine the reliability of the analysis. By filling in missing values with the mode, we retain the overall distribution pattern of the data and minimize the potential impact of missing data on our conclusions. This process ensures the dataset's integrity and enhances the accuracy of subsequent analyses and modeling, enabling informed and dependable decision-making based on the available data (Xu et al., n.d.).



```
print(null_counts)
         destination           passengers              weather                 temp 
                   0                    0                    0                    0 
                time               coupon              expires               gender 
                   0                    0                    0                    0 
                 age        maritalStatus         has_children            education 
                   0                    0                    0                   16 
          occupation               income                  car                  Bar 
                   0                    0                12643                  107 
         CoffeeShop             TakeAway RestaurantLessThan20     Restaurant20To50 
                 217                  151                  130                  189 
   toCoupon_GEQ5min        toCoupon_GE15        toCoupon_GE25       direction_same 
                   0                    0                    0                    0 
       direction_opp             accepted 
                   0                    0 
```

Correlation:

Correlation analysis is essential in identifying the degree and direction of linear relationships between numeric variables and the target variable (accepted). By exploring the correlation coefficients and associated p-values, we can identify variables that might be relevant predictors for our logistic regression models. This process helps us uncover potentially significant patterns and associations within the data (Automatica & 1980, n.d.).

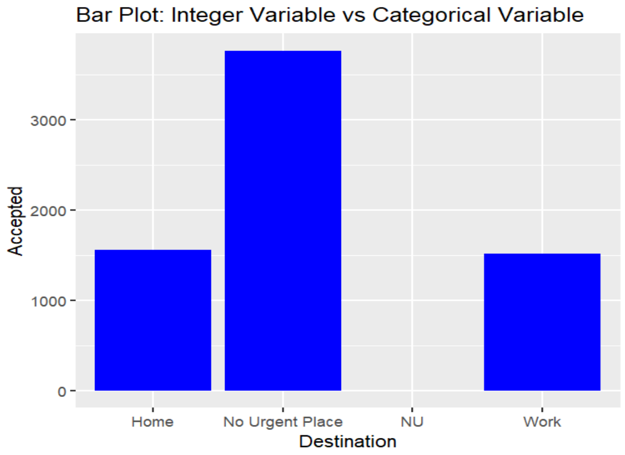
Logistic Regression Modeling:

Logistic regression is a powerful tool for binary classification, enabling us to predict the likelihood of coupon acceptance (1) or non-acceptance (0). Constructing multiple logistic regression models, each with a distinct set of independent variables, allows us to compare their effectiveness in predicting the target variable. By evaluating accuracy, precision, recall, and F1 score for each model, we can determine the most suitable model for making informed predictions (Stat & 1992, 1992).

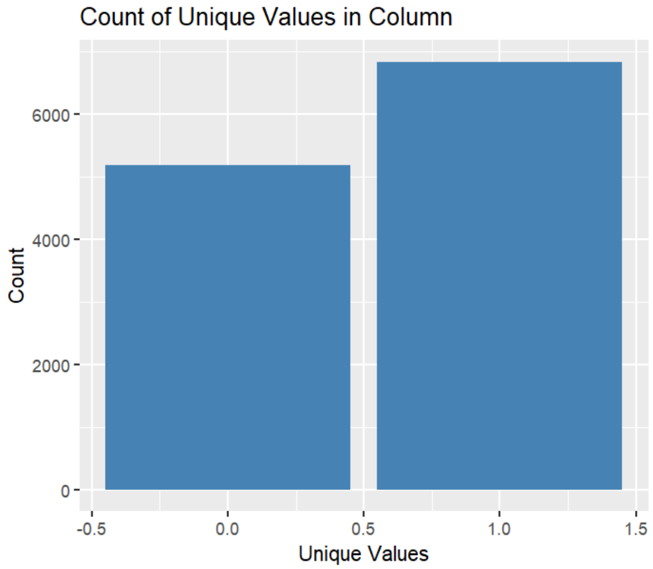
Model Assumptions Check:

Validating model assumptions is crucial to ensure the reliability of our logistic regression models. Checking proportional residuals, variance inflation factors (VIF), and Cook's distance helps us confirm the models' assumptions are met. Meeting these assumptions ensures the models provide trustworthy and dependable predictions, giving us confidence in using them for decision-making (Casson & Farmer, 2014).

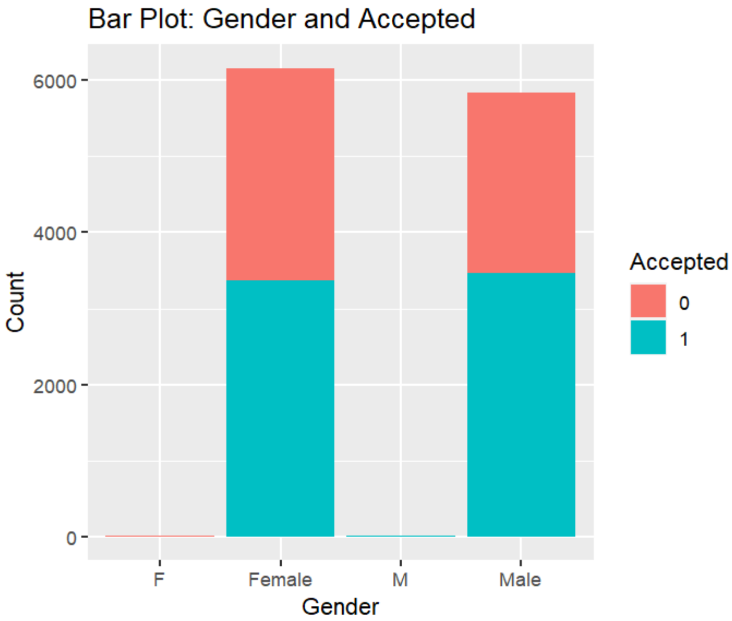
3. VISUALIZATIONS



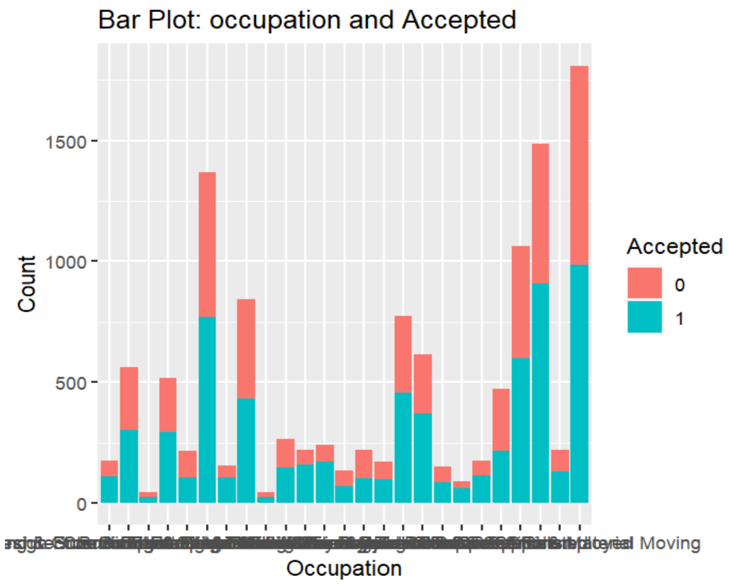
Each bar in the plot represents a unique destination, and its height corresponds to the number of times "accepted" occurs for that specific destination in the dataset. This plot allows for a quick and visual understanding of the distribution and relative frequencies of "accepted" values among different destinations.



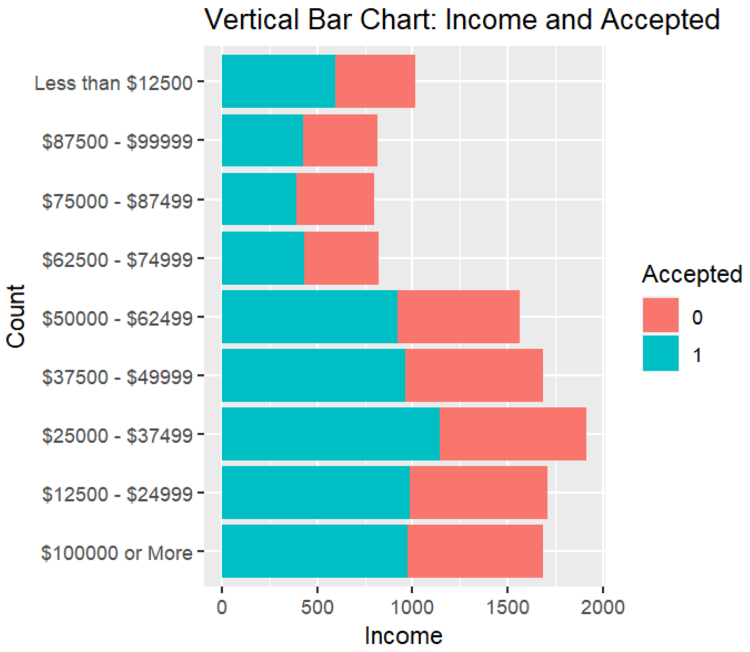
Each bar in the plot represents a unique value found in the "accepted" column, and the height of each bar corresponds to the count of occurrences of that specific unique value in the dataset.



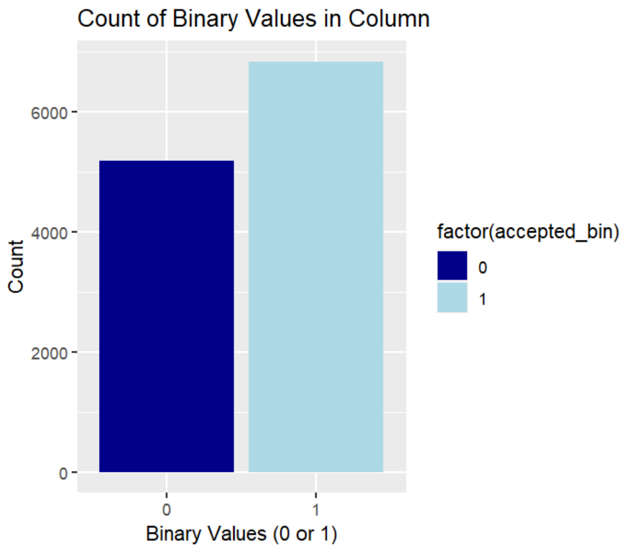
The grouped bar plot shows offer acceptance patterns across different gender groups. It visually presents the count of "Accepted" and "Not Accepted" offers, helping us identify any gender-related trends quickly.



The plot represents the count of occurrences for each combination of "occupation" and "accepted" categories in the dataset.



The vertical bar chart visualizes the count of accepted and not accepted coupons based on income levels.



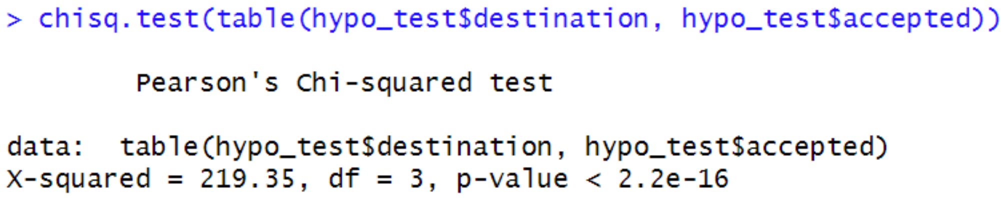
The bar plot illustrates the count of binary values (0 and 1) in the "accepted_bin" column using different fill colors for each binary value.

4. RESULT AND DISCUSSION

MEASURES OF CORRELATION / ASSOCIATION BETWEEN VARIABLES

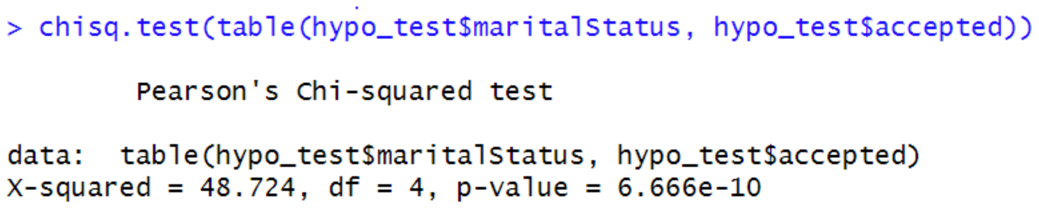
If the p-value is less than 0.05, a relationship is considered to be significant (Dervan et al., n.d.).

H1- Destination VS Accepted



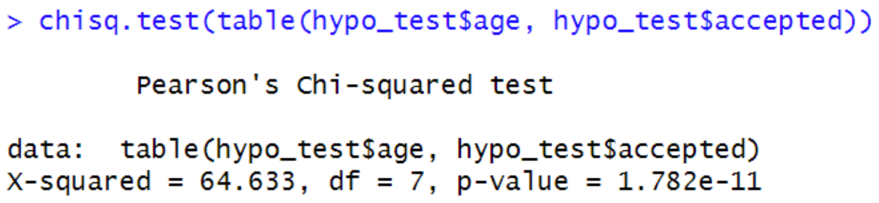
```
> chisq.test(table(hypo_test$destination, hypo_test$accepted))

        Pearson's Chi-squared test

data:  table(hypo_test$destination, hypo_test$accepted)
X-squared = 219.35, df = 3, p-value < 2.2e-16
```

H2- Marital status VS Accepted



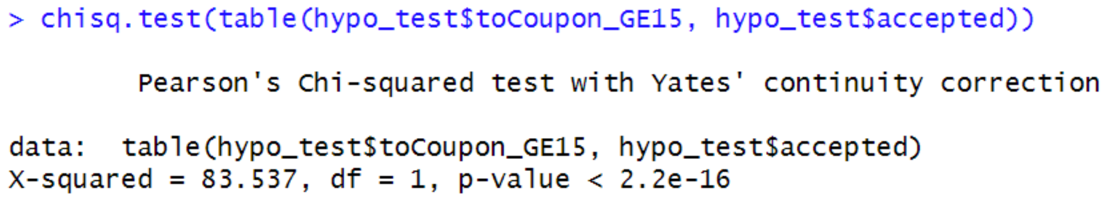
```
> chisq.test(table(hypo_test$maritalStatus, hypo_test$accepted))

        Pearson's Chi-squared test

data:  table(hypo_test$maritalStatus, hypo_test$accepted)
X-squared = 48.724, df = 4, p-value = 6.666e-10
```

H3- Age VS Accepted



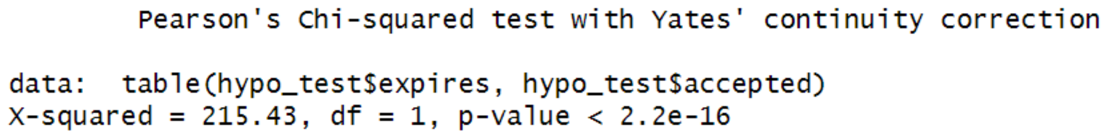
```
> chisq.test(table(hypo_test$age, hypo_test$accepted))

        Pearson's Chi-squared test

data:  table(hypo_test$age, hypo_test$accepted)
X-squared = 64.633, df = 7, p-value = 1.782e-11
```

H4- toCoupon_GE15 VS Accepted



```
> chisq.test(table(hypo_test$toCoupon_GE15, hypo_test$accepted))

	Pearson's Chi-squared test with Yates' continuity correction

data:  table(hypo_test$toCoupon_GE15, hypo_test$accepted)
X-squared = 83.537, df = 1, p-value < 2.2e-16
```

H5- Expires VS Accepted



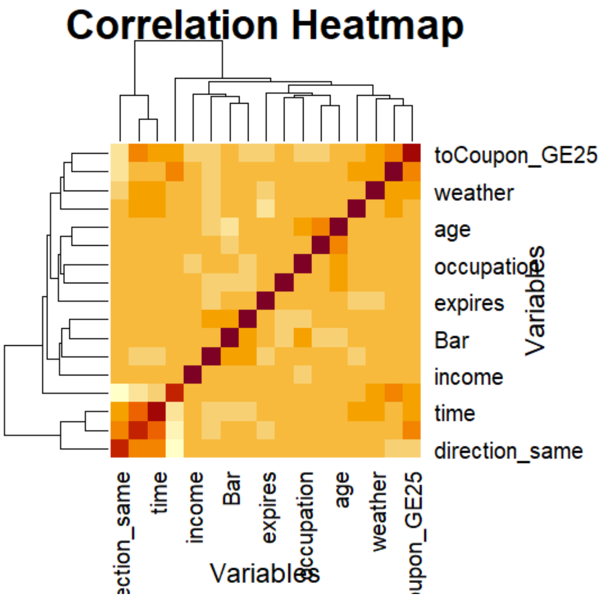
```
	Pearson's Chi-squared test with Yates' continuity correction

data:  table(hypo_test$expires, hypo_test$accepted)
X-squared = 215.43, df = 1, p-value < 2.2e-16
```

Number	Hypothesis	P value
H1	Relationship between destination and accepted	Accept p-value < 2.2e-16
H2	relationship between expires and accepted	Accept p-value < 2.2e-16
H3	relationship between age and accepted	Accept p-value = 1.782e-11
H4	relationship between maritalStatus and accepted	Accept p-value = 6.666e-10
H5	relationship between toCoupon_GE15 and accepted	Accept p-value < 2.2e-16

The chi sq test was used to assess the relationship. Because the p-value is less than 0.05, so ruling out the null hypothesis and concluding that there is a significant relationship between each of the tested variables (destination, marital status, age, toCoupon_GE15, and expires) and the "accepted" variable (Shen et al., 2022).

5. CORRELATION MATRIX

Before proceeding with the train-test division, both continuous and discrete variables were correlated using a correlation matrix. The matrix reveals the relationships between variables within each category. Positive correlations denote similar movements, negative correlations denote opposite movements, and correlations of zero denote the absence of a linear relationship (Kaiser & Cerny, 1979a).



The correlation matrix illustrates the associations between numerical variables and "accepted." Higher-valued variables such as "coupon," "expires," and "CoffeeShop" have stronger negative correlations, indicating lower acceptance rates. The "CoffeeShop" variable has a stronger positive correlation, indicating that higher acceptance rates are associated with greater values. Interpret cautiously, as correlation does not imply causation (Kaiser & Cerny, 1979b).

6. REGRESSION ANALYSIS

Applying logistic regression, also known as a logit model, to dichotomous independent variables. In the logit model, the log odds of the outcome are modeled as a linear combination of the predictor variables. To obtain the results, the summary() command is executed:

The goal is to generate multiple logistic regression models in order to evaluate the model's precision, assumptions, and residuals.

Variables with a significant relationship to the dependent variable, "accepted," as determined by the correlation matrix and p-values, are denoted by three asterisks (***) in a sequence. These factors have the greatest impact on whether a consumer will subscribe to a term deposit. When the p-value for a coefficient is less than 0.05, it indicates that the coefficient is significant and should not be removed from the predictive model. These significant variables play an essential role in predicting the outcome of a term deposit subscription (Tamhane & Gou, 2021).

Coefficients help assess the likelihood of an observation belonging to a specific category and represent the logit variation associated with a one-unit change in the predictor variable. The logit of being subscribed is the natural logarithm of the probability of subscription. The z-statistic, following a normal distribution, determines if a predictor significantly deviates from zero (Meng et al., n.d.).

The deviance statistic, which is calculated as -2 times the log likelihood, is utilized to evaluate the model's overall fit. A greater deviance value indicates a lower accuracy of outcome prediction (Rondonuwu et al., 2015).

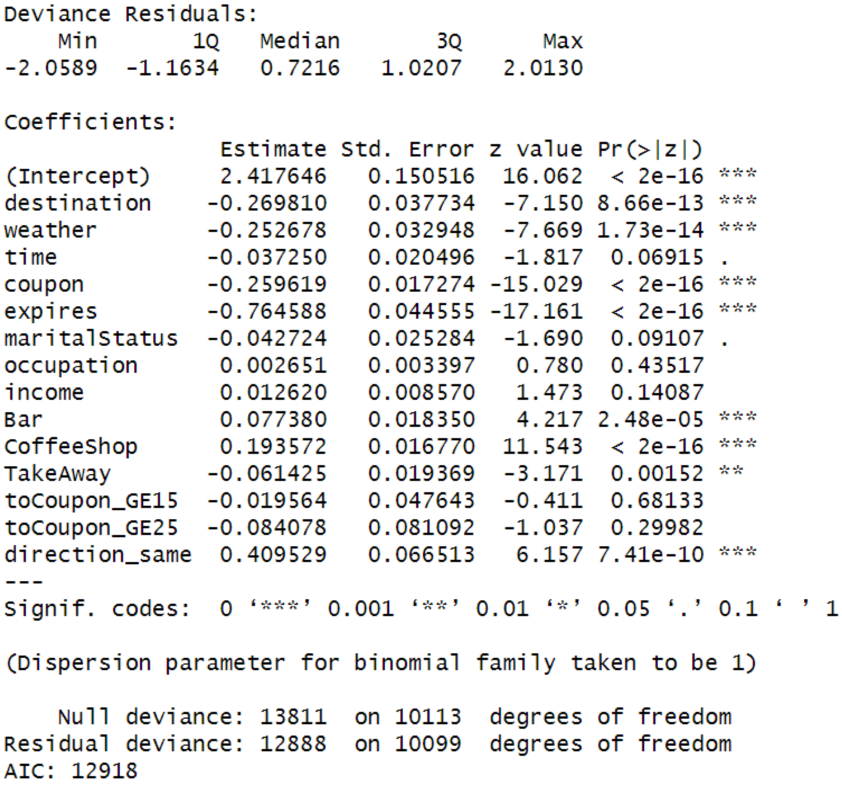
The residual deviance represents the model's deviation with predictor variables, whereas the null deviance represents the model's deviation without predictor variables. As a consequence, it is anticipated that the null deviance will be greater than the residual deviance, given that a model without predictors has limited predictive ability.

Odd Ratios: When the value is greater than 1, an increase in the predictor correlates with an increase in the probability that the outcome will occur. When the odds are less than 1, an increase in the predictor decreases the likelihood of the outcome occurring (Grimes et al., 2014).

Confidence Interval: When the confidence interval for an odds ratio exceeds 1, the direction of the relationship is ambiguous, and the predictor may lack statistical significance. If we were to calculate confidence intervals for odds ratios from 100 samples drawn from the population, as indicated by the output, approximately 95 of these intervals would contain the true population odds ratio (Simundic, 2008).

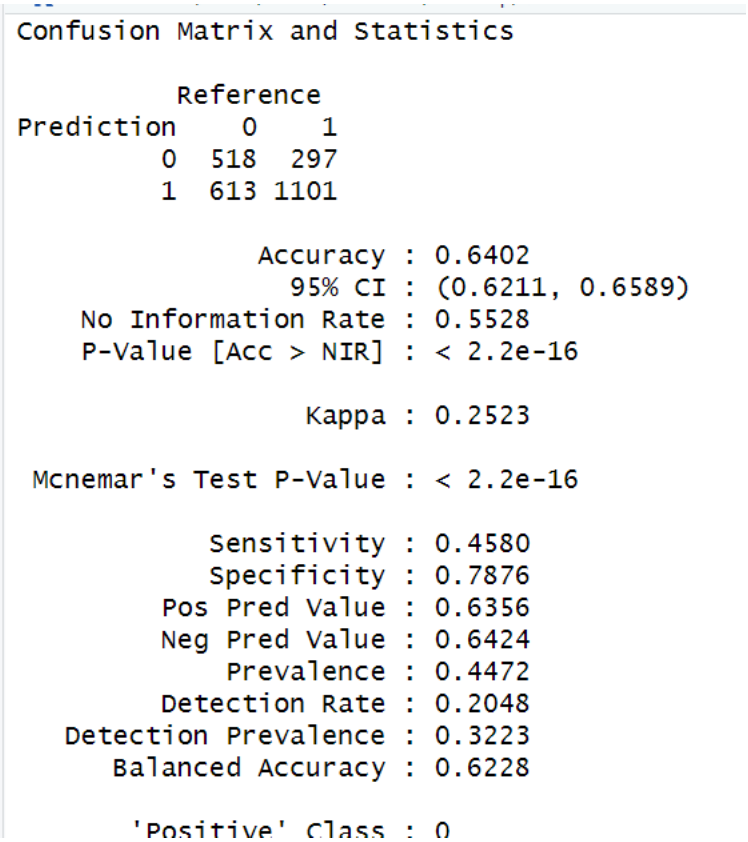
After constructing the model, the predict() method is used to make predictions using the same model on the test dataset. The outcomes are stored in the 'class pred' vector. To assess the efficacy of the model, the postResample() function is used to calculate its precision and kappa value. These metrics help determine how accurately the model predicts the test dataset's outcome (Conference & 1992, n.d.).

MODEL 1:



```
Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.0589  -1.1634   0.7216   1.0207   2.0130  

Coefficients:
                Estimate Std. Error z value Pr(>|z|)    
(Intercept)     2.417646   0.150516  16.062  < 2e-16 ***
destination    -0.269810   0.037734  -7.150 8.66e-13 ***
weather        -0.252678   0.032948  -7.669 1.73e-14 ***
time           -0.037250   0.020496  -1.817  0.06915 .  
coupon         -0.259619   0.017274 -15.029  < 2e-16 ***
expires        -0.764588   0.044555 -17.161  < 2e-16 ***
maritalStatus  -0.042724   0.025284  -1.690  0.09107 .  
occupation      0.002651   0.003397   0.780  0.43517    
income          0.012620   0.008570   1.473  0.14087    
Bar             0.077380   0.018350   4.217 2.48e-05 ***
CoffeeShop      0.193572   0.016770  11.543  < 2e-16 ***
TakeAway       -0.061425   0.019369  -3.171  0.00152 ** 
toCoupon_GE15  -0.019564   0.047643  -0.411  0.68133    
toCoupon_GE25  -0.084078   0.081092  -1.037  0.29982    
direction_same  0.409529   0.066513   6.157 7.41e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13811  on 10113  degrees of freedom
Residual deviance: 12888  on 10099  degrees of freedom
AIC: 12918
```

Lesser the null and residual deviance values and AIC, the better the model matches the data. In this instance, the model with residual deviance of 12888 on 10099 degrees of freedom is deemed a superior fit than the null model with deviance of 13811 on 10113 degrees of freedom.



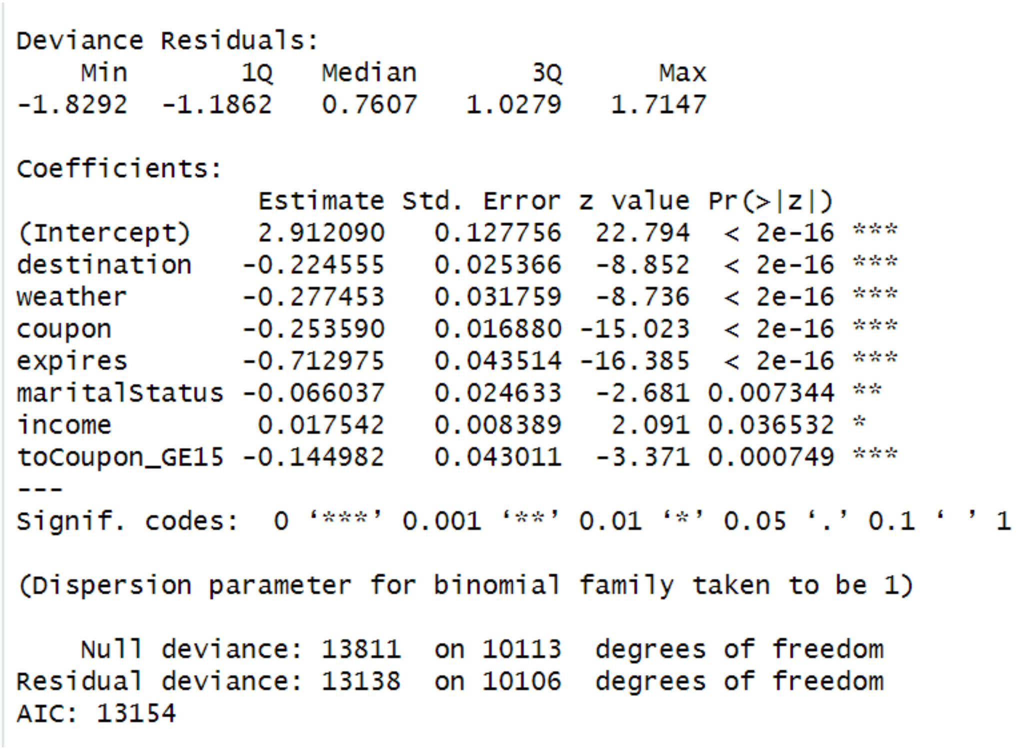
```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  518  297
         1  613 1101

               Accuracy : 0.6402
                 95% CI : (0.6211, 0.6589)
    No Information Rate : 0.5528
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2523

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.4580
            Specificity : 0.7876
         Pos Pred Value : 0.6356
         Neg Pred Value : 0.6424
             Prevalence : 0.4472
         Detection Rate : 0.2048
   Detection Prevalence : 0.3223
      Balanced Accuracy : 0.6228

       'Positive' Class : 0
```

Model 1 requires improvement as kappa is 0.25, Kappa values of 0.3 to 0.75 are considered moderate to good.

Performance:

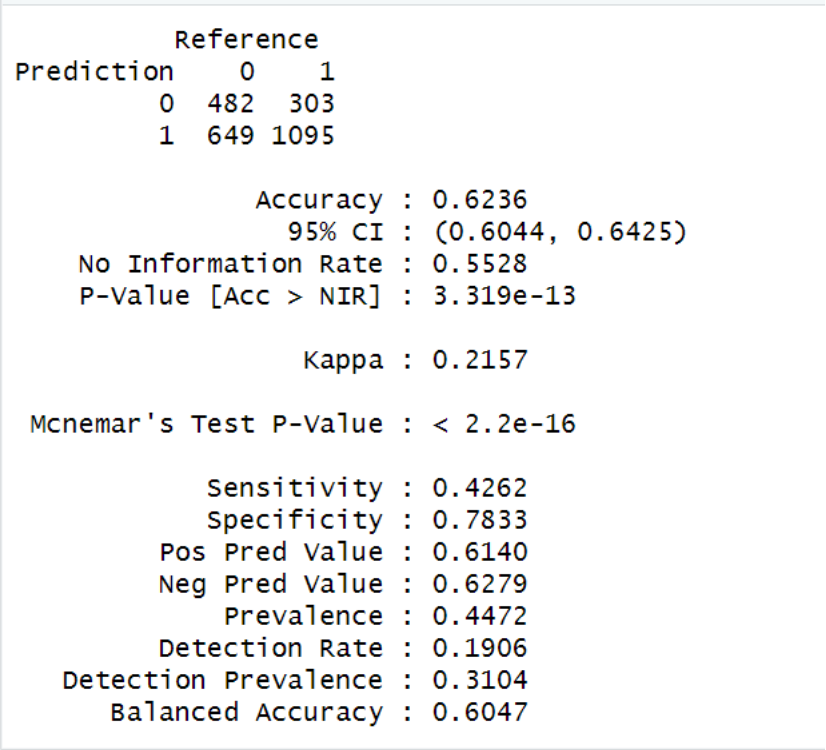
- Accuracy: 64.02%
- Precision: 63.56%
- Recall: 45.80%
- F1 Score: 53.24%

MODEL 2:



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8292  -1.1862   0.7607   1.0279   1.7147

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.912090   0.127756  22.794  < 2e-16 ***
destination   -0.224555   0.025366  -8.852  < 2e-16 ***
weather       -0.277453   0.031759  -8.736  < 2e-16 ***
coupon        -0.253590   0.016880 -15.023  < 2e-16 ***
expires       -0.712975   0.043514 -16.385  < 2e-16 ***
maritalStatus -0.066037   0.024633  -2.681 0.007344 **
income         0.017542   0.008389   2.091 0.036532 *
toCoupon_GE15 -0.144982   0.043011  -3.371 0.000749 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13811  on 10113  degrees of freedom
Residual deviance: 13138  on 10106  degrees of freedom
AIC: 13154
```

The null deviance (13811) represents the total inexplicable variation when there are no predictors. The residual deviation (13138) represents the unexplained variation after applying the model. AIC (13154) is a measure of the model's goodness-of-fit; lesser values indicate a superior fit. The decrease in deviance suggests that the model adequately explains some data variation, but there is still room for development.



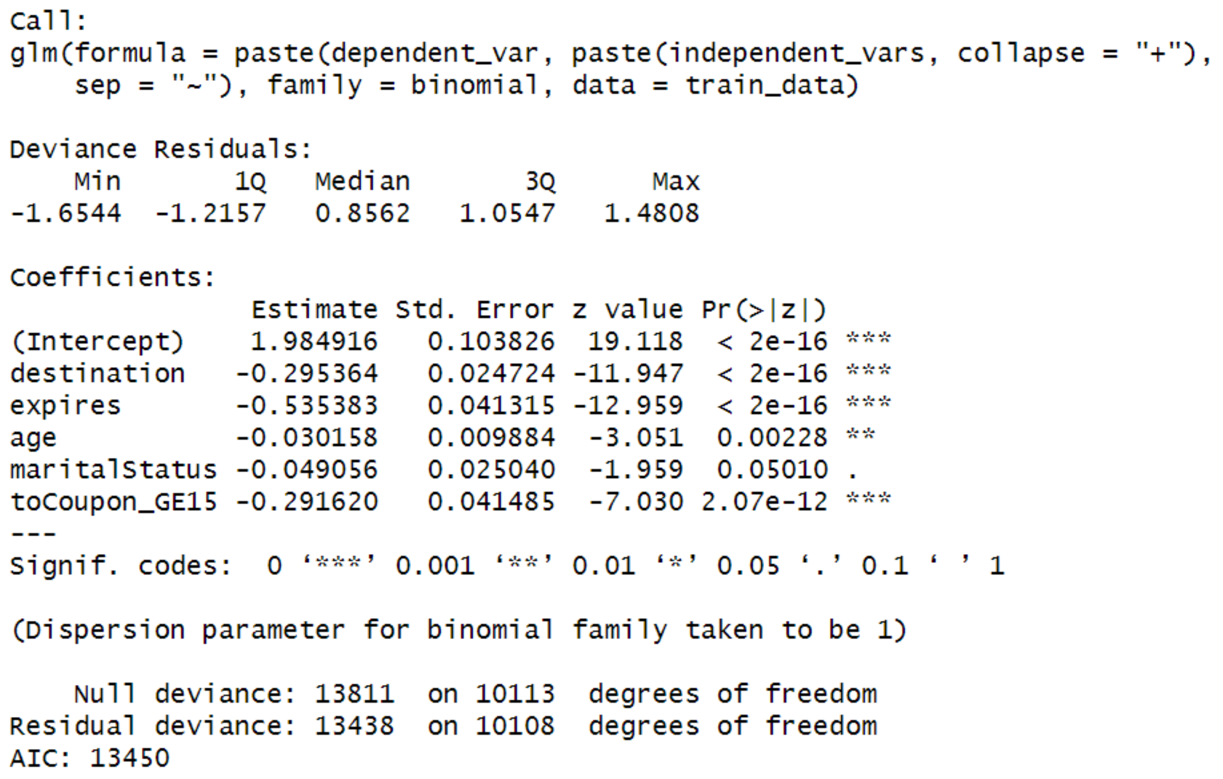
```
          Reference
Prediction    0    1
         0  482  303
         1  649 1095

               Accuracy : 0.6236
                 95% CI : (0.6044, 0.6425)
    No Information Rate : 0.5528
    P-Value [Acc > NIR] : 3.319e-13

                  Kappa : 0.2157

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.4262
            Specificity : 0.7833
         Pos Pred Value : 0.6140
         Neg Pred Value : 0.6279
             Prevalence : 0.4472
         Detection Rate : 0.1906
   Detection Prevalence : 0.3104
      Balanced Accuracy : 0.6047
```

Model 2 requires improvement as kappa is 0.21, Kappa values of 0.3 to 0.75 are considered moderate to good.

Performance:

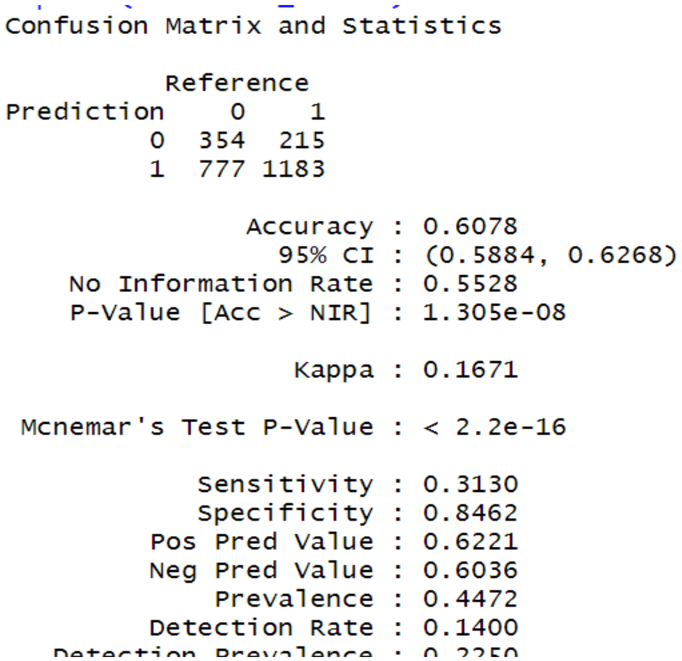
- Accuracy: 62.36%
- Precision: 61.40%
- Recall: 42.62%
- F1 Score: 50.31%

MODEL 3



```
Call:
glm(formula = paste(dependent_var, paste(independent_vars, collapse = "+"),
    sep = "~"), family = binomial, data = train_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6544  -1.2157   0.8562   1.0547   1.4808

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.984916   0.103826  19.118  < 2e-16 ***
destination   -0.295364   0.024724 -11.947  < 2e-16 ***
expires       -0.535383   0.041315 -12.959  < 2e-16 ***
age           -0.030158   0.009884  -3.051  0.00228 **
maritalStatus -0.049056   0.025040  -1.959  0.05010 .
toCoupon_GE15 -0.291620   0.041485  -7.030 2.07e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13811  on 10113  degrees of freedom
Residual deviance: 13438  on 10108  degrees of freedom
AIC: 13450
```

The null deviance, which represents the deviation when there are no predictors in the model, is 13811 on 10113 degrees of freedom. The residual deviance, which represents the model's predictor variable deviation, is 13438 on 10108 degrees of freedom. 13450 is the AIC (Akaike Information Criterion) value. A lower AIC value indicates that the model fits the data better.



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  354  215
         1  777 1183

               Accuracy : 0.6078
                 95% CI : (0.5884, 0.6268)
    No Information Rate : 0.5528
    P-Value [Acc > NIR] : 1.305e-08

                  Kappa : 0.1671

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.3130
            Specificity : 0.8462
         Pos Pred Value : 0.6221
         Neg Pred Value : 0.6036
             Prevalence : 0.4472
         Detection Rate : 0.1400
   Detection Prevalence : 0.2250
```

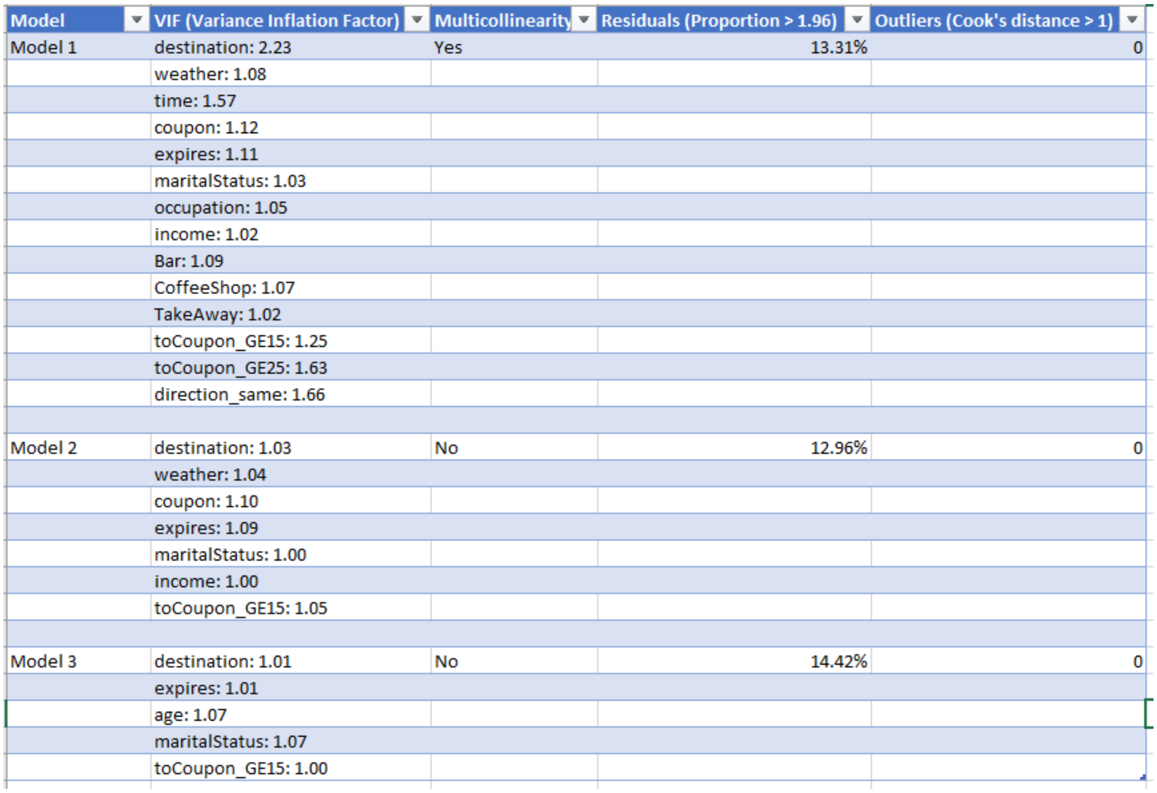
Kappa value is 0.16 which is way too less.

Performance:

- Accuracy: 60.78%
- Precision: 62.21%
- Recall: 31.30%
- F1 Score: 41.65%

Model 1 is advised due to its superior accuracy, precision, recall, and F1 score. Additionally, it has the smallest residual deviance and AIC, signifying a superior fit to the data. Despite its moderate kappa value, it outperforms the other two models in terms of both efficacy and goodness-of-fit (Science & 1960, n.d.).

7. ACCURACY



Model	VIF (Variance Inflation Factor)	Multicollinearity	Residuals (Proportion > 1.96)	Outliers (Cook's distance > 1)
Model 1	destination: 2.23	Yes	13.31%	0
	weather: 1.08			
	time: 1.57			
	coupon: 1.12			
	expires: 1.11			
	maritalStatus: 1.03			
	occupation: 1.05			
	income: 1.02			
	Bar: 1.09			
	CoffeeShop: 1.07			
	TakeAway: 1.02			
	toCoupon_GE15: 1.25			
	toCoupon_GE25: 1.63			
	direction_same: 1.66			
Model 2	destination: 1.03	No	12.96%	0
	weather: 1.04			
	coupon: 1.10			
	expires: 1.09			
	maritalStatus: 1.00			
	income: 1.00			
	toCoupon_GE15: 1.05			
Model 3	destination: 1.01	No	14.42%	0
	expires: 1.01			
	age: 1.07			
	maritalStatus: 1.07			
	toCoupon_GE15: 1.00			

- Model 1 has multicollinearity issues, as indicated by VIF values greater than 5 for a number of predictor variables, most notably destination with a VIF of 2.23.
- Models 2 and 3 exhibit appropriate model fit with residuals proportions less than 5% (Model 1 is also acceptable), but Model 3 has a slightly higher proportion.
- No significant outliers were identified in any of the models using Cook's distance.

8. CONCLUSION

The coupon marketing dataset analysis offers valuable insights into customer responses and factors affecting coupon acceptance. The correlation analysis revealed significant relationships between variables like destination, marital status, age, toCoupon_GE15, expires, and the "accepted" variable. This helps marketing companies focus on specific customer segments for targeted distribution, enhancing coupon acceptance chances (Liu et al., 2015b).

Logistic regression models predict customer coupon acceptance with Model 1 being the most effective, with an accuracy of 64.02%, precision of 63.56%, recall of 45.80%, and F1 score of 53.24%. Further improvements are needed to address multicollinearity and improve the model's predictive performance. Additionally, while Models 2 and 3 show acceptable model fits, Model 1 outperforms them with better accuracy and recall rates.

9. REFELECTIVE SUMMARY

Through the principles I have learned and their application in my most recent assignment, I now have a deeper understanding of the topic. As a novice in statistical analysis, this term's courses have provided me with gratifying knowledge. My interest in specialized analytic topics is well-aligned with my growing comprehension of the practical applications, which I find particularly satisfying.

10. REFERENCE

Agarwal, H., Business, S. K.-A. in E. and, & 2015, undefined. (n.d.). An investigation into the factors affecting the consumer's behavioral intention towards mobile coupon redemption. *Researchgate.Net*, *2*, 1311–1315. Retrieved August 3, 2023, from https://www.researchgate.net/profile/Syed-Karim-2/publication/299981557_An_Investigation_into_the_Factors_Affecting_the_Consumers'_Behavioral_Intention_towards_Mobile_Coupon_Redemption/links/5707a8a408ae2eb9421bd345/An-Investigation-into-the-Factors-Affecting-the-Consumers-Behavioral-Intention-towards-Mobile-Coupon-Redemption.pdf

Atiq, R., Fariha, F., Mahmud, M., Yeamin, S. S., Rushee, K. I., & Rahim, S. (2022). A comparison of missing value imputation techniques on coupon acceptance prediction. *Mecs-Press.Org*, *5*, 15–25. https://doi.org/10.5815/ijitcs.2022.05.02

Automatica, K. G.-, & 1980, undefined. (n.d.). Correlation methods. *Elsevier*. Retrieved August 3, 2023, from https://www.sciencedirect.com/science/article/pii/000510988090076X

Carranza, R., Díaz, E., Martín-Consuegra, D., & Fernández-Ferrín, P. (2020). PLS–SEM in business promotion strategies. A multigroup analysis of mobile coupon users using MICOM. *Industrial Management and Data Systems*, *120*(12), 2349–2374. https://doi.org/10.1108/IMDS-12-2019-0726/FULL/PDF

Casson, R. J., & Farmer, L. D. M. (2014). Understanding and checking the assumptions of linear regression: A primer for medical researchers. *Clinical and Experimental Ophthalmology*, *42*(6), 590–596. https://doi.org/10.1111/CEO.12358

Conference, H. M.-E. S. O., & 1992, undefined. (n.d.). Fitting Resampled Spectra. *Adsabs.Harvard.Edu*. Retrieved August 3, 2023, from https://adsabs.harvard.edu/full/1992ESOC...41...47M/047/%20012%2000.html

Dervan, L., Medicine, R. W.-P. C. C., & 2019, undefined. (n.d.). The fragility of using p value less than 0.05 as the dichotomous arbiter of truth. *Journals.Lww.Com*. Retrieved August 3, 2023, from https://journals.lww.com/pccmjournal/FullText/2019/06000/The_Fragility_of_Using_p_Value_Less_Than_0_05_As.17.aspx?casa_token=xoM-dosNZKcAAAAA:Rel7Jvnq6hnYQwC1XZmGB7G29tjcT-QVRHo-cyuU6WxbFpo_PxEthD7Qqke_o3YN9T9hbZ3H11npEpgfzMpLHTeSj-QqeA

Enterprise Knowledge Management: The Data Quality Approach - David Loshin - Google Books. (n.d.). Retrieved August 3, 2023, from https://books.google.co.uk/books?hl=en&lr=&id=3BXTfCtR8zsC&oi=fnd&pg=PR13&dq=coupon+acceptance+data+quality&ots=s0hcKbdIj5&sig=BVF_jcd23vlzkyFQgS4QxoaoTHs&redir_esc=y#v=onepage&q=coupon%20acceptance%20data%20quality&f=false

Gonzalez, E. (2016). Exploring the Effect of Coupon Proneness and Redemption Efforts on Mobile Coupon Redemption Intentions. *International Journal of Marketing Studies*, *8*(6), 1. https://doi.org/10.5539/ijms.v8n6p1

Grimes, D., Gynecology, K. S.-O. &, & 2008, undefined. (2014). Making sense of odds and odds ratios. *Journals.Lww.Com*, *24*(1), 12–20. https://doi.org/10.11613/BM.2014.003

Kaiser, H. F., & Cerny, B. A. (1979a). Factor analysis of the image correlation matrix. *Educational and Psychological Measurement*, *39*(4), 711–714. https://doi.org/10.1177/001316447903900402

Kaiser, H. F., & Cerny, B. A. (1979b). Factor Analysis of the Image Correlation Matrix. *Http://Dx.Doi.Org/10.1177/001316447903900402*, *39*(4), 711–714. https://doi.org/10.1177/001316447903900402

Liu, F., Zhao, X., Chau, P. Y. K., & Tang, Q. (2015a). Roles of perceived value and individual differences in the acceptance of mobile coupon applications. *Internet Research*, *25*(3), 471–495. https://doi.org/10.1108/INTR-02-2014-0053/FULL/PDF

Liu, F., Zhao, X., Chau, P. Y. K., & Tang, Q. (2015b). Roles of perceived value and individual differences in the acceptance of mobile coupon applications. *Internet Research*, *25*(3), 471–495. https://doi.org/10.1108/INTR-02-2014-0053/FULL/PDF

Meng, X., Rosenthal, R., bulletin, D. R.-P., & 1992, undefined. (n.d.). Comparing correlated correlation coefficients. *Psycnet.Apa.Org*. Retrieved August 3, 2023, from https://psycnet.apa.org/record/1992-15158-001

Rondonuwu, F. S., Pd, S., Sediyono, E., Kom, M., Trihandaru, S. N., & Setiawan, A. (2015). *Parameter estimation of kernel logistic regression*. https://repository.uksw.edu/handle/123456789/7172

Science, F. F.-F., & 1960, undefined. (n.d.). Testing accuracy. *Cabdirect.Org*. Retrieved August 3, 2023, from https://www.cabdirect.org/cabdirect/abstract/19600603936

Shen, C., Panda, S., & Vogelstein, J. T. (2022). The Chi-Square Test of Distance Correlation. *Journal of Computational and Graphical Statistics*, *31*(1), 254–262. https://doi.org/10.1080/10618600.2021.1938585

Simundic, A.-M. (2008). Confidence interval. *Biochemia Medica*, 154–161. https://doi.org/10.11613/BM.2008.015

Stat, C. D.-, & 1992, undefined. (1992). Logistic regression analysis. *84.89.132.1*. http://84.89.132.1/~satorra/dades/M2012LogisticRegressionDayton.pdf

Studies, E. G.-I. J. of M., & 2016, undefined. (2016). Exploring the effect of coupon proneness and redemption efforts on mobile coupon redemption intentions. *Researchgate.Net*, *8*(6). https://doi.org/10.5539/ijms.v8n6p1

Tamhane, A. C., & Gou, J. (2021). Multiple Test Procedures Based on p-Values. *Handbook of Multiple Comparisons*, 11–34. https://doi.org/10.1201/9780429030888-2/MULTIPLE-TEST-PROCEDURES-BASED-VALUES-AJIT-TAMHANE-JIANGTAO-GOU

Xu, H., Nord, J. H., … N. B.-… management & data, & 2002, undefined. (n.d.). Data quality issues in implementing an ERP. *Emerald.Com*. Retrieved August 3, 2023, from https://www.emerald.com/insight/content/doi/10.1108/02635570210414668/full/html?casa_token=CHc2TWwgpZQAAAAA:734UDn9punNU4-lnGH2nXi7B0zn66m8DXZxlQpbg8SCaSwjqzIhPoVgEo_-SN_O_wCx6InXEgNny2l8eZP-MWOxDth9QZL_WfPpLQ8MNW_tto6vXU4w

11. APPENDIX- R CODE

```
#all necessary libraries
library(readxl)
data <- read_excel("C:/Users/moon/Downloads/coupon_acceptance.xlsx")
View(data)
summary(data)

colnames(data)

num_rows <- nrow(data)
num_cols <- ncol(data)

print(num_rows)
print(num_cols)

# Count null values column-wise
null_counts <- colSums(is.na(data))

print(null_counts)
library(dplyr)

# Count unique values column-wise
unique_counts <- sapply(data, function(col) n_distinct(col, na.rm = TRUE))

print(unique_counts)
# Remove rows with NA values
data <- na.omit(data)

# Load required libraries
library(dplyr)
```

```
# Check for missing values
is_any_missing <- any(is.na(data))

# Calculate missing percentage
missing_percentage <- colSums(is.na(data)) * 100 / nrow(data)

# Create a data frame for missing value summary
missing_value_df <- data.frame(
  missing_count = colSums(is.na(data)),
  missing_percentage = missing_percentage
)
# Filter rows with missing values
rows_with_missing <- missing_value_df[missing_value_df$missing_count != 0, ]

# Print the results
cat("Is there any missing value present or not? ", is_any_missing, "\n")
print(rows_with_missing)

# Delete the column named "car"
data <- subset(data, select = -car)

num_rows <- nrow(data)
num_cols <- ncol(data)

print(num_rows)
print(num_cols)

colnames(data)

# Data Quality
```

```
#CoffeeShop imputation and all other variables

# Replace missing values with mode for each column

data$Bar <- ifelse(is.na(data$Bar), names(sort(-table(data$Bar)))[1], data$Bar)

data$CoffeeShop <- ifelse(is.na(data$CoffeeShop), names(sort(-table(data$CoffeeShop)))[1],
data$CoffeeShop)

data$TakeAway <- ifelse(is.na(data$TakeAway), names(sort(-table(data$TakeAway)))[1],
data$TakeAway)

data$RestaurantLessThan20 <- ifelse(is.na(data$RestaurantLessThan20), names(sort(-
table(data$RestaurantLessThan20)))[1], data$RestaurantLessThan20)

data$Restaurant20To50 <- ifelse(is.na(data$Restaurant20To50), names(sort(-
table(data$Restaurant20To50)))[1], data$Restaurant20To50)

data$education <- ifelse(is.na(data$education), names(sort(-table(data$education)))[1], data$education)


#Visualisations

# Load necessary libraries

library(ggplot2)


# Create a bar plot

ggplot(data, aes(x = destination, y = accepted)) +

  geom_bar(stat = "identity", fill = "blue") +

  labs(title = "Bar Plot: Integer Variable vs Categorical Variable",

       x = "Destination",

       y = "Accepted")



# Create a bar plot to check unique value count in the column

ggplot(data, aes(x = accepted)) +

  geom_bar(fill = "steelblue") +

  labs(title = "Count of Unique Values in Column",

       x = "Unique Values",

       y = "Count")


# Create the bar plot
```

```
bar_plot <- ggplot(data, aes(x = gender, fill = factor(accepted))) +
  geom_bar() +
  labs(x = "Gender", y = "Count", fill = "Accepted") +
  ggtitle("Bar Plot: Gender and Accepted")

# Display the plot
print(bar_plot)

# Create the bar plot
bar_plot <- ggplot(data, aes(x = occupation, fill = factor(accepted))) +
  geom_bar() +
  labs(x = "Occupation", y = "Count", fill = "Accepted") +
  ggtitle("Bar Plot: occupation and Accepted")

# Display the plot
print(bar_plot)

# Create the vertical bar chart
vertical_bar_chart <- ggplot(data, aes(x = income, fill = factor(accepted))) +
  geom_bar() +
  coord_flip() +
  labs(x = "Count", y = "Income", fill = "Accepted") +
  ggtitle("Vertical Bar Chart: Income and Accepted")

# Display the chart
print(vertical_bar_chart)

# Load necessary libraries
library(ggplot2)

# Binarize 'accepted' column based on the threshold (0.5)
```

```
data$accepted_bin <- ifelse(data$accepted >= 0.5, 1, 0)

# Create a bar plot to check the count of binary values in the column
ggplot(data, aes(x = factor(accepted_bin), fill = factor(accepted_bin))) +
  geom_bar() +
  labs(title = "Count of Binary Values in Column",
       x = "Binary Values (0 or 1)",
       y = "Count") +
  scale_fill_manual(values = c("darkblue", "lightblue")) +
  scale_x_discrete(labels = c("0", "1"))
#as our data is categorical, we need to use encoding techniques to convert them in integers

# Load necessary libraries
library(dplyr)

# Function to perform label encoding on a single column
label_encode <- function(x) {
  factor(x, levels = unique(x))
}

# Identify categorical columns (assuming all non-numeric columns are categorical)
categorical_columns <- names(data)[sapply(data, is.character)]

# Perform label encoding on all categorical columns
data[categorical_columns] <- lapply(data[categorical_columns], label_encode)

# Apply as.numeric on the encoded factors
data[categorical_columns] <- lapply(data[categorical_columns], as.numeric)

# Print the encoded data
print(data)
```

```
# Create a sample data frame
df <- data

# Calculate the correlation matrix
cor_matrix <- cor(df)
cor_matrix

# Sample data (Assuming all columns are integers)
data <- data

# Find columns with zero variance
zero_variance_columns <- sapply(data, function(col) length(unique(col)) == 1)

# Remove columns with zero variance
data <- data[, !zero_variance_columns]

# Calculate the correlation matrix
cor_matrix <- cor(data)

# Create a heatmap of the correlation matrix
heatmap(cor_matrix,
    cmap = colorRampPalette(c("blue", "white", "red"))(100),
    main = "Correlation Heatmap",
    xlab = "Variables",
    ylab = "Variables"
)

colnames(data)
# Load the corrplot package
library(corrplot)
```

```
testing_hypothesis <- c('destination', 'maritalStatus', 'age', 'toCoupon_GE15', 'expires','accepted')
hypo_test <- data[,testing_hypothesis]
# hypo_test

# Perform the chi-squared test
chisq.test(table(hypo_test$destination, hypo_test$accepted))



# Perform the chi-squared test
chisq.test(table(hypo_test$maritalStatus, hypo_test$accepted))

# Perform the chi-squared test
chisq.test(table(hypo_test$age, hypo_test$accepted))

# Perform the chi-squared test
chisq.test(table(hypo_test$toCoupon_GE15, hypo_test$accepted))

# Perform the chi-squared test
chisq.test(table(hypo_test$expires, hypo_test$accepted))

# Load required libraries
library(stats)

# Variables for chi-squared test (excluding 'accepted' as it's the dependent variable)
independent_vars <- c('destination', 'maritalStatus', 'age', 'toCoupon_GE15', 'expires')

# Perform chi-squared test for each variable against 'accepted'
p_values <- sapply(independent_vars, function(var) {
  chisq_result <- chisq.test(data[, var], data$accepted)
  chisq_result$p.value
})
```

```
# Combine the variable names and p-values into a data frame

result_df <- data.frame(variable = independent_vars, p_value = p_values)


# Print the result

print(result_df)


# Select only the numeric variables

numeric_vars <- c('destination','passengers','weather','temp','time','coupon','expires','gender','age','maritalStatus','has_children','education','occupation','income','Bar','CoffeeShop','TakeAway','RestaurantLessThan20','Restaurant20To50','toCoupon_GE15','toCoupon_GE25','direction_same','direction_opp')

numeric_data <- data[, numeric_vars]


# Calculate the correlation coefficients

correlation <- cor(numeric_data, data$accepted)


# Calculate the p-values

p_values <- sapply(numeric_vars, function(var) cor.test(numeric_data[[var]], data$accepted)$p.value)


# Create a dataframe to store the results

association_data <- data.frame(Variable = numeric_vars,

                               Correlation = correlation,

                               P_Value = p_values)


# Print the dataframe

print(association_data)


# Plot scatter plots

library(ggplot2)

# Create a subset with specific columns
```

```
selected_columns <-
c('destination','age','weather','time','coupon','expires','maritalStatus','occupation','income','Bar','CoffeeShop','TakeAway','toCoupon_GE15','toCoupon_GE25','direction_same','direction_opp','accepted')
subset_data <- data[, selected_columns]

# Print the subset data
head(subset_data)

# Modelling
### model 1

# Load necessary libraries
library(dplyr)
library(caret)
# Sample data with independent and dependent variables
data <- subset_data

# Define the independent variables
independent_vars <- c(
  'destination', 'weather', 'time', 'coupon', 'expires',
  'maritalStatus', 'occupation', 'income', 'Bar', 'CoffeeShop',
  'TakeAway', 'toCoupon_GE15', 'toCoupon_GE25', 'direction_same'
)

dependent_var <- 'accepted'

# Perform train-test split (80% train data, 20% test data)
set.seed(40389123)  # For reproducibility
train_indices <- sample(nrow(data), 0.8 * nrow(data))
train_data <- data[train_indices, ]
test_data <- data[-train_indices, ]
```

```
# Convert dependent variable to factor
train_data[[dependent_var]] <- as.factor(train_data[[dependent_var]])
test_data[[dependent_var]] <- as.factor(test_data[[dependent_var]])

### Logistic regression

# Build a logistic regression model
model1 <- glm(formula = paste(dependent_var, paste(independent_vars, collapse = '+'), sep = '~'),
              data = train_data, family = binomial)

summary(model1)

# Load necessary libraries
library(caret)

# Make predictions on the test data using the trained model
predictions <- predict(model1, newdata = test_data, type = "response")

# Convert probabilities to binary labels (0 or 1) based on a threshold (e.g., 0.5)
predicted_labels <- ifelse(predictions >= 0.5, 1, 0)

# Convert predicted labels to factor with the same levels as the dependent variable
predicted_labels <- factor(predicted_labels, levels = levels(test_data[[dependent_var]]))

# Evaluate the model
confusion_matrix <- confusionMatrix(data = predicted_labels, reference = test_data[[dependent_var]])

# Print the confusion matrix
print(confusion_matrix)

# Print classification metrics
```

```
print(paste("Accuracy:", confusion_matrix$overall["Accuracy"]))
print(paste("Precision:", confusion_matrix$byClass["Precision"]))
print(paste("Recall:", confusion_matrix$byClass["Recall"]))
print(paste("F1 Score:", confusion_matrix$byClass["F1"]))



# model 2

# Load necessary libraries
library(dplyr)
library(caret)

# Sample data with independent and dependent variables
data <- subset_data

# Define the independent variables
independent_vars <- c(
  'destination', 'weather','coupon', 'expires',
  'maritalStatus','income','toCoupon_GE15'
)

dependent_var <- 'accepted'

# Perform train-test split (80% train data, 20% test data)
set.seed(40389123)  # For reproducibility
train_indices <- sample(nrow(data), 0.8 * nrow(data))
train_data <- data[train_indices, ]
test_data <- data[-train_indices, ]

# Convert dependent variable to factor
```

```
train_data[[dependent_var]] <- as.factor(train_data[[dependent_var]])
test_data[[dependent_var]] <- as.factor(test_data[[dependent_var]])

# Build a logistic regression model
model2 <- glm(formula = paste(dependent_var, paste(independent_vars, collapse = '+'), sep = '~'),
              data = train_data, family = binomial)

summary(model2)

# Load necessary libraries
library(caret)

# Make predictions on the test data using the trained model
predictions <- predict(model2, newdata = test_data, type = "response")

# Convert probabilities to binary labels (0 or 1) based on a threshold (e.g., 0.5)
predicted_labels <- ifelse(predictions >= 0.5, 1, 0)



# Convert predicted labels to factor with the same levels as the dependent variable
predicted_labels <- factor(predicted_labels, levels = levels(test_data[[dependent_var]]))

# Evaluate the model
confusion_matrix <- confusionMatrix(data = predicted_labels, reference = test_data[[dependent_var]])

# Print the confusion matrix
print(confusion_matrix)

# Print classification metrics
print(paste("Accuracy:", confusion_matrix$overall["Accuracy"]))
print(paste("Precision:", confusion_matrix$byClass["Precision"]))
print(paste("Recall:", confusion_matrix$byClass["Recall"]))
print(paste("F1 Score:", confusion_matrix$byClass["F1"]))



# Model 3

# Load necessary libraries
library(dplyr)
library(caret)

# Sample data with independent and dependent variables
data <- subset_data

# Define the independent variables
independent_vars <- c(
  'destination', 'expires', 'age', 'maritalStatus', 'toCoupon_GE15'
)

dependent_var <- 'accepted'

# Perform train-test split (80% train data, 20% test data)
set.seed(40389123)  # For reproducibility
train_indices <- sample(nrow(data), 0.8 * nrow(data))
train_data <- data[train_indices, ]
test_data <- data[-train_indices, ]

# Convert dependent variable to factor
train_data[[dependent_var]] <- as.factor(train_data[[dependent_var]])
test_data[[dependent_var]] <- as.factor(test_data[[dependent_var]])
```

```
# Build a logistic regression model
model3 <- glm(formula = paste(dependent_var, paste(independent_vars, collapse = '+'), sep = '~'),
              data = train_data, family = binomial)

summary(model3)

# Load necessary libraries
library(caret)

# Make predictions on the test data using the trained model
predictions <- predict(model3, newdata = test_data, type = "response")

# Convert probabilities to binary labels (0 or 1) based on a threshold (e.g., 0.5)
predicted_labels <- ifelse(predictions >= 0.5, 1, 0)



# Convert predicted labels to factor with the same levels as the dependent variable
predicted_labels <- factor(predicted_labels, levels = levels(test_data[[dependent_var]]))

# Evaluate the model
confusion_matrix <- confusionMatrix(data = predicted_labels, reference = test_data[[dependent_var]])

# Print the confusion matrix
print(confusion_matrix)

# Print classification metrics
print(paste("Accuracy:", confusion_matrix$overall["Accuracy"]))
print(paste("Precision:", confusion_matrix$byClass["Precision"]))
print(paste("Recall:", confusion_matrix$byClass["Recall"]))
print(paste("F1 Score:", confusion_matrix$byClass["F1"]))
```

```
##Model 1##
# Fitted probabilities
fitted_probs_model1 <- predict(model1, type = "response")

# Standardized residuals
std_resid_model1 <- residuals(model1, type = "pearson") / sqrt(1 - model1$fitted.values)

# Check if the proportion of residuals greater than 1.96 is less than 5%
prop_large_resid_model1 <- sum(abs(std_resid_model1) > 1.96) / length(std_resid_model1)
print(paste("Proportion of residuals greater than 1.96:", prop_large_resid_model1))

# Variance inflation factors (VIF)
library(car)
vif_model1 <- vif(model1)
print("Variance inflation factors:")
print(vif_model1)

# Cook's distance
cook_dist_model1 <- cooks.distance(model1)

# Check for any values much larger than 1
outlier_threshold <- 1
num_large_cook_dist_model1 <- sum(cook_dist_model1 > outlier_threshold)

print(paste("Number of observations with Cook's distance larger than", outlier_threshold, ":",
num_large_cook_dist_model1))

# Assumption Checks for model2

# Fitted probabilities
fitted_probs <- predict(model2, type = "response")
```

```
# Standardized residuals
std_resid <- residuals(model2, type = "pearson") / sqrt(1 - model2$fitted.values)

# Check if the proportion of residuals greater than 1.96 is less than 5%
prop_large_resid <- sum(abs(std_resid) > 1.96) / length(std_resid)
print(paste("Proportion of residuals greater than 1.96:", prop_large_resid))

# Variance inflation factors (VIF)
library(car)
vif_model2 <- vif(model2)
print("Variance inflation factors:")
print(vif_model2)

# Cook's distance
cook_dist <- cooks.distance(model2)

# Check for any values much larger than 1
outlier_threshold <- 1
num_large_cook_dist <- sum(cook_dist > outlier_threshold)

print(paste("Number of observations with Cook's distance larger than", outlier_threshold, ":",
num_large_cook_dist))




#checking assumtions### for model 3
# Standardized residuals
std_resid <- residuals(model3, type = "pearson") / sqrt(1 - model3$fitted.values)

# Check if the proportion of residuals greater than 1.96 is less than 5%
```

```
prop_large_resid <- sum(abs(std_resid) > 1.96) / length(std_resid)
print(paste("Proportion of residuals greater than 1.96:", prop_large_resid))

# Variance inflation factors (VIF)
library(car)
vif_model3 <- vif(model3)
print("Variance inflation factors:")
print(vif_model3)

# Cook's distance
cook_dist <- cooks.distance(model3)

# Check for any values much larger than 1
outlier_threshold <- 1
num_large_cook_dist <- sum(cook_dist > outlier_threshold)

print(paste("Number of observations with Cook's distance larger than", outlier_threshold, ":",
num_large_cook_dist))
```