# Logistic Regression

# Predicting the Odds of a Homeless Individual being approved for shelter

[1]Makinde Jamiu Olalekan

[1]Federal University of Technology Akure, Nigeria

**Corresponding E-mail:** makindejamiu1011@gmail.com

## Abstract

In this research, we use a logistic regression model to identify the probability of a homeless person being admitted to shelter accommodation depending on the demographic and behavioral factors. The data of 242 homeless applicants acquired through Kaggle was used to analyze nine predictor variables, such as age, gender, veteran status, monthly income, number of nights homeless, substance abuse, completion of in-house training, probation status, and type of assistantship. Following data cleaning and model selection process, five noteworthy predictors were gathered, namely, veteran status, monthly income, number of nights homeless, substance abuse, and type of assistantship. The performance of the model measured by the Hosmer-Lemeshow test, McFadden R 2, and ROC-AUC, showed that the model was well-fitted and predictive (AUC 0.90). The results indicate that homeless veterans, those who receive temporary assistance, those who experience more years of homelessness and their substance abuse problems are more likely to be approved of shelter and higher-income applicants are less likely to be approved. Such lessons can inform the policies of the public health and management of shelters in allocating resources to vulnerable groups in the homeless population.

**Keywords:** Logistic regression; housing; shelter approval; predictive modeling; popular health; socioeconomic variables; veteran status; substance abuse; McFadden R 2; ROC curve.

# I. Introduction

The post-Covid effect on the global economy and the general increase in the price of commodities in the US (Inflation), including housing, are expected to result in a rise in the number of homeless persons living on the streets and in shelters. This suggests that a deeper comprehension of the variables influencing the duration of stay in the homeless shelter could yield important information for public health programmes and policy. The logistic model was used to model the in-house client's approval of the shelter as a function of particular parameters and homeless demographics in order to support the claim.

The data used for this project was obtained from Kaggle. The collection is made up of data collected by a nearby homeless shelter on individuals who are homeless and are applying for approval to stay in the shelter. With nine predictor factors (age, gender, veteran, monthly income, number of nights at homeless, substance abuse, completed in house training, probation, assistantship) and one response variable specified on a binary scale, where 1 denotes a shelter request approved and 0 denotes a shelter request denied, the data set includes information on 242 homeless people.

In order to help prioritise shelter services and, ultimately, result in better public health policies and interventions, the study presents empirical findings based on the data. In order to accomplish this goal, I look at a logistic regression model to estimate the likelihood that a person will be granted shelter approval and to pinpoint important criteria that must be met in order for a homeless person to be granted shelter.

Therefore, I constructed a logistic regression model using all the predictor variables with a stepwise procedure. Also, the Hosmer-Lemeshow test was used to assess the goodness of fit of the model to evaluate how well the predicted probabilities from the model match the observed outcomes. Also, I considered McFadden's R-squared to estimate the predictive power of the model. Then, the cook’s distance plot was used for outlier’s detection, and the variance inflation factor (VIF) was used to assess the severity of multicollinearity in the predictor variables. Lastly, the accuracy, sensitivity, specificity and AUC-ROC measures were used to measure the accuracy of the model.

## The Logistic Model

The logistic model is given as follows

$$P(Shelter = 1|X_1, X_2, \dots, X_9) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_9 X_9)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_9 X_9)}$$

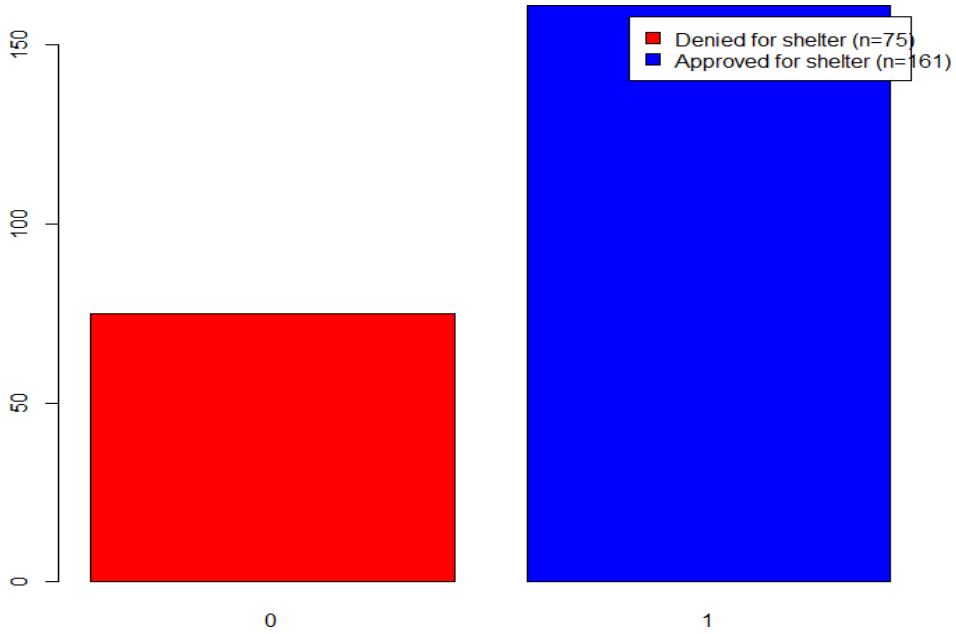
where □□, □ = 0, 1, 2, … 9 are the model coefficients and □□, □ = 1,2, 3, … ,9 are the predictors variables.

For the sake of simplicity, the estimated coefficients and its odd ratios were used in this project to interpret the fitted model parameters. Thus, we are only interested in the direction, rather than the magnitude of the effect, as denoted by the sign attached to each estimate. Furthermore, we pinpoint the important factors that influence the chance of a homeless person being approved for a shelter.

## II. Data analysis and Results

Inconsistencies in the data, such as missing rows, were eliminated through pre-processing and data cleaning, yielding valid cases for 236 participants. Figure 1 shows that 161 people were granted shelter approval while 75 people were denied it.

***Figure 1:*** *Bar plot for the Response variable*




Subsequently, the data was split into a test set and a training set using a split ratio of 7:10. The training set consists of 165 responses, accounting for 84 (50.91%) approved and 81 (49.09%) denied shelter requests; the test set consists of 33 (46.5%) approved and 38 (53.5%) rejected shelter requests (***See Appendix 1***). With the train set, a model building process was applied in order to produce accurate predictions or classifications based on the data.

## Model selection

1. As indicated by the logistic model equation above, the fitted model first incorporates each and every predictor. After that, a stepwise regression model with backward elimination and the AIC criteria was run. Six predictors were found by the logistic regression analysis: the type of assistantship, monthly income, number of nights spent homeless, substance abuse, and veteran status (***see table below***).

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.0819513  0.7224728  -4.266 1.99e-05 ***
VETERAN         1.7451183  0.6565380   2.658  0.00786 **
INCOME         -0.0009769  0.0003668  -2.663  0.00773 **
NIGHTS          0.0159160  0.0057191   2.783  0.00539 **
substanceabuse  1.1398826  0.5805051   1.964  0.04958 *
probation       1.1576948  0.6015941   1.924  0.05431 .
assistancetype  3.3477566  0.5116948   6.542 6.05e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. Following the stepwise procedure and then removing the insignificant terms, we obtain the following significant model (***See table below***).

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.9172655  0.7092798  -4.113 3.91e-05 ***
VETERAN         1.6094324  0.6463826   2.490  0.01278 *
INCOME         -0.0009421  0.0003571  -2.638  0.00834 **
NIGHTS          0.0179621  0.0059105   3.039  0.00237 **
substanceabuse  1.1777510  0.5721741   2.058  0.03955 *
assistancetype  3.1720972  0.4788036   6.625 3.47e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 228.68  on 164  degrees of freedom
Residual deviance: 134.06  on 159  degrees of freedom
AIC: 146.06
```

## III. Interpretation of the Model

The significant model's outcome indicates that, holding other variables constant

a. compared to non-veteran homeless people, being a veteran (□ =1.609, se (□) = 0.646) increases the likelihood that shelter requests will be granted.
b. This also holds true for requests for temporary assistantship shelter (□ =3.172, se (□)=0.479).
c. Additionally, it was discovered that the likelihood of the request being granted increases with the number of nights the applicant stays homeless (□ = 0.018, se (□) = 0.006).
d. According to the model, monthly income decreased the likelihood of shelter approval requests (□ = -0.0009, se (□)=0.0004).
e. Finally, substance abuse raises the likelihood that a request for a shelter will be granted (□ =1.178, se (□) = 0.572).

For an easier interpretation, we can transform these values into odd's ratios:

| (Intercept) | VETERAN | INCOME | NIGHTS | substanceabuse | assistancetype |
|---|---|---|---|---|---|
| 0.05408137 | 4.99997228 | 0.99905834 | 1.01812434 | 3.24706331 | 23.85746475 |

Considering these estimates, we can say (holding the other variables constant);

a. Getting shelter approved for a veteran homeless versus non-veteran, the odds of approval increase by 4.999.
b. Getting shelter approved for a temporary assistant vs permanent assistant, the odds of approval increase by 28.857.
c. Getting shelter approved based on the number of nights the applicant stays homeless, the odds of approval increase by 1.018.
d. Getting shelter approved based on monthly income of the applicant, the odds of approval decrease by 0.999
e. Finally, getting shelter approved based on substance abuse, the odds of approval increase by 3.247.

## ANOVA Test

After that, an ANOVA test based on the log-likelihood was applied to the obtained model.

```
Analysis of Deviance Table

Model 1: required ~ VETERAN + INCOME + NIGHTS + substanceabuse + assistancetype
Model 2: required ~ VETERAN + INCOME + NIGHTS + substanceabuse + probation +
    assistancetype
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       159     134.06
2       158     130.29  1   3.7698  0.05219 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the result is not significant, the ANOVA result showed that model 1 is superior to model 2. This suggests that the logistic model was not the right place for probation to be included in the model. As a result, the best model includes the following variables: veteran, monthly income, number of nights spent homeless, substance abuse, and type of assistance.

## Hosmer-Lemeshow Test

The Hosmer-Lemeshow Goodness of Fit Test to determine model adequacy.

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  logit_model1$y, fitted(logit_model1)
X-squared = 7.337, df = 4, p-value = 0.1191
```

Since the Hosmer-Lemeshow has a p-value = 0.1191 (i.e. p-value > 0.05), we can suggest that there is no significant difference between observed and predicted outcomes, indicating a good model fit.

## Mcfadden R-square

we use the McFadden $R^2$ to assess the predictive power of the model.

```
fitting null model for pseudo-r2
        llh      llhNull           G2     McFadden         r2ML         r2CU
-67.0281063 -114.3420106   94.6278085    0.4137928    0.4364515    0.5819995
```

A McFadden $R^2$ value between 0.2 and 0.4 is considered good. Therefore, since our McFadden $R^2$ is 0.414 we can say that the model selected is an excellent fit for predicting is shelter approval.

## Assumption check

### Collinearity

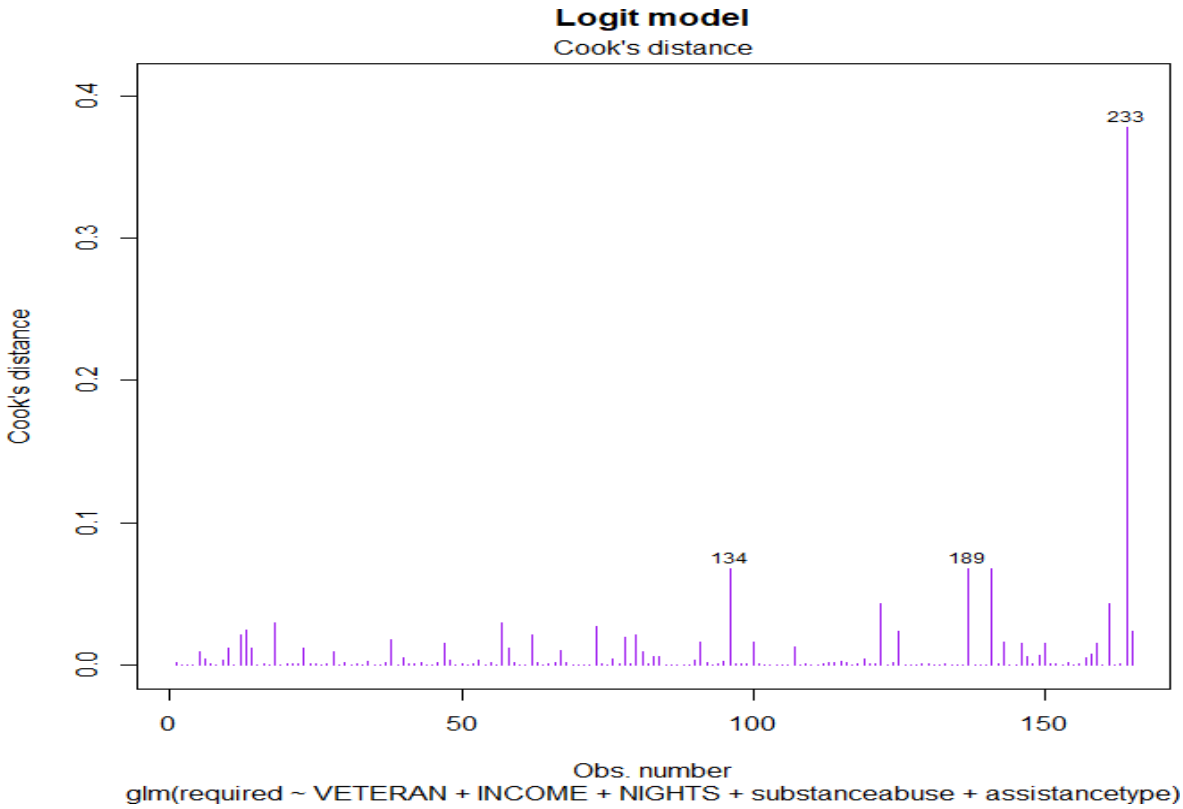
After assessing the goodness of fit of the logistic model, we will check to see if there is any collinearity between the predictor variables. We will check this using variance inflation factors (VIF). If any VIF are greater than 10, we will remove that variable from the model.

| VETERAN | INCOME | NIGHTS | substanceabuse | assistancetype |
|---|---|---|---|---|
| 1.733164 | 1.153007 | 1.753950 | 1.338364 | 1.189282 |

Since none of the VIF values are larger than 10, we can say that there is no collinearity between the predictor variables. This implies that the fitted model is free from the problem of multicollinearity.

### Outliers detection

The cook's distance plot was used to assess the outlier. A cook's distance greater than 1 signifies an influential point.




The model was sensitive to the identification of outliers at points 134, 189, and 233, as demonstrated by the plot above, which showed the existence of outliers in the model. Furthermore, there are no plotted cook's distances larger than 1. Therefore, the model lacks a significant influential point, indicating that it is sensitive to the identification of outliers.

## Predictive Measure

For comparison purposes, the performance measure was computed for both the train set and the test set.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 70 12
         1 11 72

               Accuracy : 0.8606
                 95% CI : (0.7982, 0.9095)
    No Information Rate : 0.5091
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7212

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8571
            Specificity : 0.8642
         Pos Pred Value : 0.8675
         Neg Pred Value : 0.8537
             Prevalence : 0.5091
         Detection Rate : 0.4364
   Detection Prevalence : 0.5030
      Balanced Accuracy : 0.8607

       'Positive' Class : 1
```

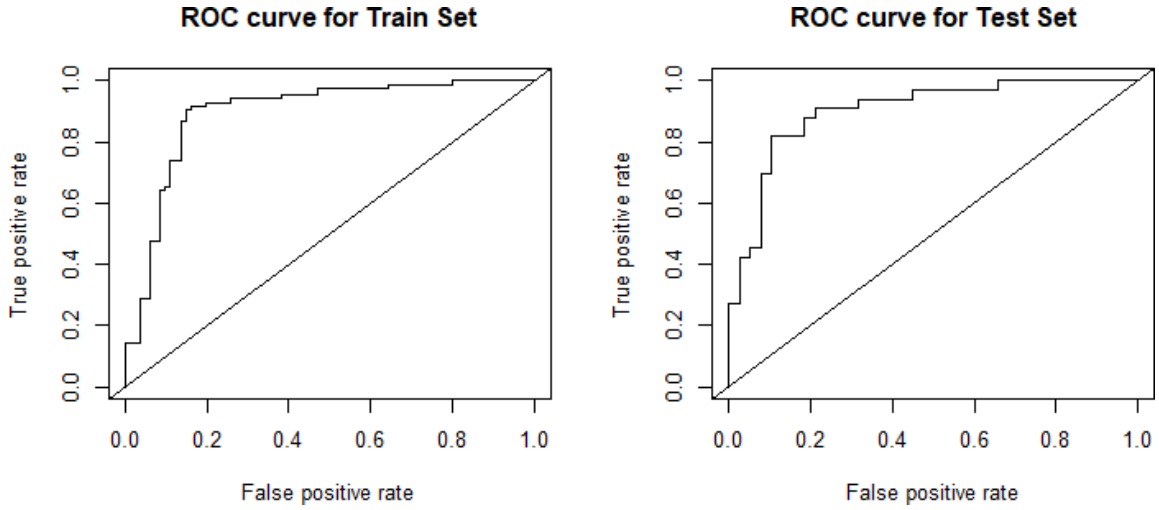
**Train Set Results**

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 31  6
         1  7 27

               Accuracy : 0.8169
                 95% CI : (0.7073, 0.8987)
    No Information Rate : 0.5352
    P-Value [Acc > NIR] : 6.544e-07

                  Kappa : 0.6327

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8182
            Specificity : 0.8158
         Pos Pred Value : 0.7941
         Neg Pred Value : 0.8378
             Prevalence : 0.4648
         Detection Rate : 0.3803
   Detection Prevalence : 0.4789
      Balanced Accuracy : 0.8170

       'Positive' Class : 1
```

**Test Set Results**

According to the results, the logistics model's accuracy is close to 1 (more than 80%), meaning that more than 80% of the response variable was correctly predicted in both the train set (0.861) and the test set (0.817). The logit model also shows a value greater than 80% in both cases according to the sensitivity measure (train: 0.857, test: 0.818), which measures how well a model was able to identify the positive class (the approved shelter request class). Similarly, for the logit model, the specificity measure, which quantifies how well the model identified the negative class (the class of denied shelter requests), likewise yields values higher than 80% for both the train set (0.864) and test set (0.816).

## ROC Curve




The area underneath this ROC curve is .895 for the train set logit model and the it is 0.902 for the test set. The curve is close to the left-hand border yet the top of the curve does not reach the y-value of 1 quickly. This indicates that the test is somewhat accurate. Since the AUC is greater above 89%, the model does a good job of classifying to some appreciable extent, approval requests correctly, and denied request correctly and making predictions using the chosen model.

## IV. Conclusion

In order to determine the factors that determine shelter approval in homeless facilities, this project fit a binary model (the logistic model) on data related to the homeless. Five factors are identified by the logit model: substance abuse, number of nights spent homeless, veteran status, monthly income, and assistantship. According to the model's result, homeless veterans have a higher chance of having their requests approved for shelters, and those who seek out temporary housing have a better chance of getting accepted than those who seek out long-term housing assistance. The outcome also reveals that an individual's likelihood of being approved increases with the number of nights they are homeless. Additionally, the likelihood that a homeless person will be given shelter is decreased for higher earners.

Based on the fitted model, the project also investigates the predictive performance of logistic regression. The outcome reveals that the model is capable of detecting outliers and that the collinearity issue is managed. Based on the analysis of predictive measures, the model demonstrated accuracy in predicting the two classes, achieving over 80% performance across all metrics examined.

# References

1. Data Source: Kaggle Data web: Homeless Dataset (kaggle.com)
2. Jing Zhang (Unpublished): Linear Statistical Analysis Note. Fall semester, 2022. Department of Mathematics and Statistics, Georgia State University.
3. Hao, H., Garfield, M., & Purao, S. (2022). The determinants of length of homeless shelter stays: evidence-based regression analyses. *International Journal of Public Health*, *66*, 1604273.
4. Van Straaten, B., Van der Laan, J., Rodenburg, G., Boersma, S. N., Wolf, J. R., & Van de Mheen, D. (2017). Dutch homeless people 2.5 years after shelter admission: what are predictors of housing stability and housing satisfaction?. *Health & Social Care in the Community*, *25*(2), 710-722.
5. Pourat, N., Yue, D., Chen, X., Zhou, W., & O'Masta, B. (2023). Easy to use and validated predictive models to identify beneficiaries experiencing homelessness in Medicaid administrative data. *Health Services Research*, *58*(4), 882-893.
6. Moxley, V. B., Hoj, T. H., & Novilla, M. L. B. (2020). Predicting homelessness among individuals diagnosed with substance use disorders using local treatment records. *Addictive Behaviors*, *102*, 106160.
7. Mohapatra, A., & Sehgal, N. (2018). Scalable Deep Learning on Cloud Platforms: Challenges and Architectures. *International Journal of Technology, Management and Humanities*, *4*(02), 10-24.
8. Sharma, A., & Odunaike, A. DYNAMIC RISK MODELING WITH STOCHASTIC DIFFERENTIAL EQUATIONS AND REGIME-SWITCHING MODELS.
9. Ojuri, M. A. (2021). Evaluating Cybersecurity Patch Management through QA Performance Indicators. *International Journal of Technology, Management and Humanities*, *7*(04), 30-40.
10. Nkansah, Christopher. (2021). Geomechanical Modeling and Wellbore Stability Analysis for Challenging Formations in the Tano Basin, Ghana.
11. YEVHENIIA, K. (2021). Bio-based preservatives: A natural alternative to synthetic additives. INTERNATIONAL JOURNAL, 1(2), 056-070.
12. Sehgal, N., & Mohapatra, A. (2021). Federated Learning on Cloud Platforms: Privacy-Preserving AI for Distributed Data. *International Journal of Technology, Management and Humanities*, *7*(03), 53-67.
13. Kumar, K. (2022). The Role of Confirmation Bias in Sell-Side Analyst Ratings. *International Journal of Technology, Management and Humanities*, *8*(03), 7-24.
14. Asamoah, A. N. (2022). Global Real-Time Surveillance of Emerging Antimicrobial Resistance Using Multi-Source Data Analytics. INTERNATIONAL JOURNAL OF APPLIED PHARMACEUTICAL SCIENCES AND RESEARCH, 7(02), 30-37.
15. OKAFOR, C., VETHACHALAM, S., & AKINYEMI, A. A DevSecOps MODEL FOR SECURING MULTI-CLOUD ENVIRONMENTS WITH AUTOMATED DATA PROTECTION.
16. Ojuri, M. A. (2022). Cybersecurity Maturity Models as a QA Tool for African Telecommunication Networks. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*, *14*(04), 155-161.
17. Adebayo, I. A., Olagunju, O. J., Nkansah, C., Akomolafe, O., Godson, O., Blessing, O., & Clifford, O. (2019). Water-Energy-Food Nexus in Sub-Saharan Africa: Engineering Solutions for Sustainable Resource Management in Densely Populated Regions of West Africa.

18. Odunaike, A. DESIGNING ADAPTIVE COMPLIANCE FRAMEWORKS USING TIME SERIES FRAUD DETECTION MODELS FOR DYNAMIC REGULATORY AND RISK MANAGEMENT ENVIRONMENTS.
19. Ojuri, M. A. (2022). The Role of QA in Strengthening Cybersecurity for Nigeria's Digital Banking Transformation. *Well Testing Journal*, *31*(1), 214-223.
20. Akomolafe, O. (2022). Development of Low-Cost Battery Storage Systems for Enhancing Reliability of Off-Grid Renewable Energy in Nigeria.
21. Sunkara, G. (2022). AI-Driven Cybersecurity: Advancing Intelligent Threat Detection and Adaptive Network Security in the Era of Sophisticated Cyber Attacks. *Well Testing Journal*, *31*(1), 185-198.
22. Kumar, K. (2023). Capital Deployment Timing: Lessons from Post-Recession Recoveries. *International Journal of Technology, Management and Humanities*, *9*(03), 26-46.
23. Ojuri, M. A. (2023). AI-Driven Quality Assurance for Secure Software Development Lifecycles. *International Journal of Technology, Management and Humanities*, *9*(01), 25-35.
24. Odunaike, A. DESIGNING ADAPTIVE COMPLIANCE FRAMEWORKS USING TIME SERIES FRAUD DETECTION MODELS FOR DYNAMIC REGULATORY AND RISK MANAGEMENT ENVIRONMENTS.

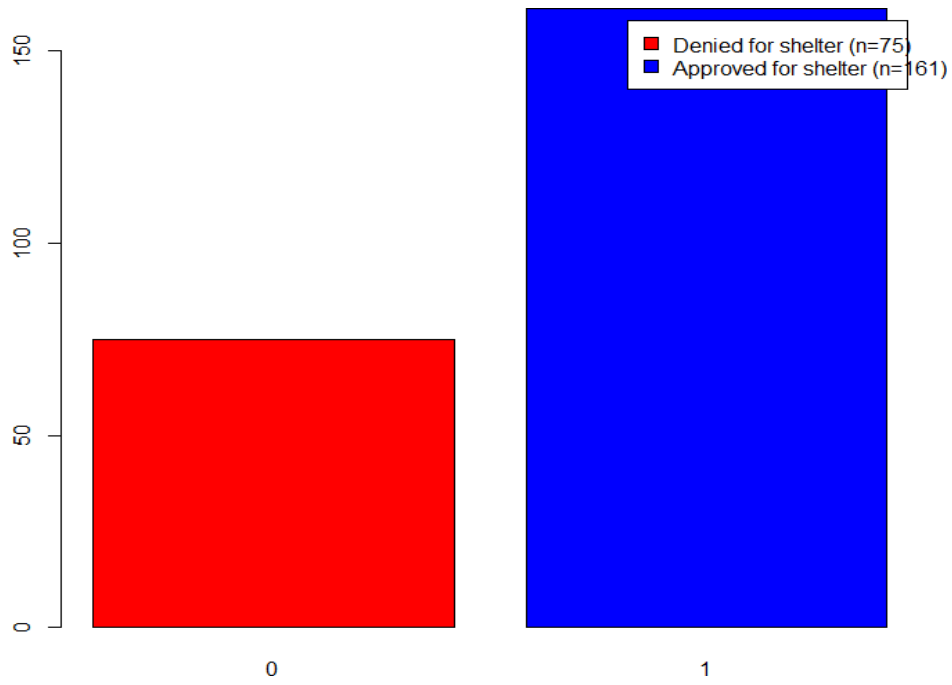
## APPENDIX

### Appendix 1

```
> #estimate the frequency of the train and test set on response variable
> table(train_set$required)

 0  1
81 84
> table(test_set$required)

 0  1
38 33
>
> #estimate the percentage of the train and test set on response variable
> 100*table(train_set$required)/nrow(train_set)

       0        1
49.09091 50.90909
> 100*table(test_set$required)/nrow(test_set)

       0        1
53.52113 46.47887
```

# R-CODE

```
> #Import the homeless data into R studio for analysis
> Data_homeless <- read.csv("C:/Users/makin/Desktop/LINEAR STAT PROJECT/homelessData.c
>
> # Data Cleaning, checking and removing of missing values
> 100*mean(is.na(Data_homeless))
[1] 0.2253944
> Data_homeless <- Data_homeless[complete.cases(Data_homeless), ]
> 100*mean(!is.na(Data_homeless))
[1] 100
>
> # Examining the Imbalance Class data
> table(Data_homeless$required)

  0   1
 75 161
> table(Data_homeless$required)/(nrow(Data_homeless))

        0         1
0.3177966 0.6822034
> barplot(table(Data_homeless$required), col= c("red", "blue"), legend=c("Denied for s
```




```
>
> #Spliting the data into 1=approved and 0 denied
> Appr_1 = Data_homeless[Data_homeless$required==1,]
> Den_0 = Data_homeless[Data_homeless$required==0,]
>
> #To ensure that models are not biased toward the majority class
> #Balance by sampling 50% from the minority class (undersampling) and 50% from the ma
> #install.packages("rpart")
> #install.packages("ROSE")
> library(rpart)
> library(ROSE)
>
> Data_homelesss.both <- ovun.sample(required ~ ., data = Data_homeless, method = "bot
```

```
> head(Data_homelesss.both)
  CLIENT_KEY AGE Male VETERAN INCOME NIGHTS substanceabuse completed probation assist
1     248072  28    0       0 381.94    237              1         0         0
2     104493  25    0       0 500.00    154              1         1         0
3     233477  22    0       0   0.00    477              1         1         1

4     289625  25    0       0 2108.32   275              0         1         1
5     280713  25    1       0 674.00    179              0         1         1
6     291410  30    0       0 690.00    137              1         1         1
> #Check the result of the undersampling and oversampling classes
> table(Data_homelesss.both$required)

  0   1
119 117
> table(Data_homelesss.both$required)/(nrow(Data_homelesss.both))

        0         1
0.5042373 0.4957627
>
> Data_homelesss.both$required = factor(Data_homelesss.both$required)
>
> #spliting the data into train (70%) and test (30%) set
> set.seed(11)
> library(caTools)
> sample <- sample.split(Data_homelesss.both$CLIENT_KEY, SplitRatio = 0.7,set.seed(2))
> train_set <- subset(Data_homelesss.both, sample == TRUE)
> test_set  <- subset(Data_homelesss.both, sample == FALSE)
>
> #check dimensions of training set and test set
> dim(train_set)
[1] 165 11
> dim(test_set)
[1] 71 11
> str(train_set)
'data.frame': 165 obs. of 11 variables:
 $ CLIENT_KEY       : int 248072 104493 289625 280713 291410 318202 248616 263197 96815
 $ AGE              : int 28 25 25 25 30 40 69 26 29 31 ...
 $ Male             : int 0 0 0 1 0 0 0 0 0 0 ...
 $ VETERAN          : int 0 0 0 0 0 0 1 0 0 0 ...
 $ INCOME           : num 382 500 2108 674 690 ...
 $ NIGHTS           : int 237 154 275 179 137 63 1 192 190 103 ...
 $ substanceabuse: int 1 1 0 0 1 0 0 1 1 1 ...
 $ completed        : int 0 1 1 1 1 1 1 0 1 0 ...
 $ probation        : int 0 0 1 1 1 0 0 1 1 0 ...
 $ assistancetype: int 0 1 1 1 0 1 1 1 0 0 ...
 $ required         : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
>
>
> # checking the consistency of data splitting on Response variable
> #estimate the frequency of the train and test set on response variable
> table(train_set$required)

 0  1
81 84
> table(test_set$required)

 0  1
38 33
>
> #estimate the percentage of the train and test set on response variable
> 100*table(train_set$required)/nrow(train_set)

       0        1
49.09091 50.90909
> 100*table(test_set$required)/nrow(test_set)
```

```
        0        1
53.52113 46.47887
>
> # Fitting the logit base model for the data
> logit_model= glm(required~., data=train_set[,-1], family=binomial(link = logit))
> summary(logit_model)

Call:
glm(formula = required ~ ., family = binomial(link = logit),
```

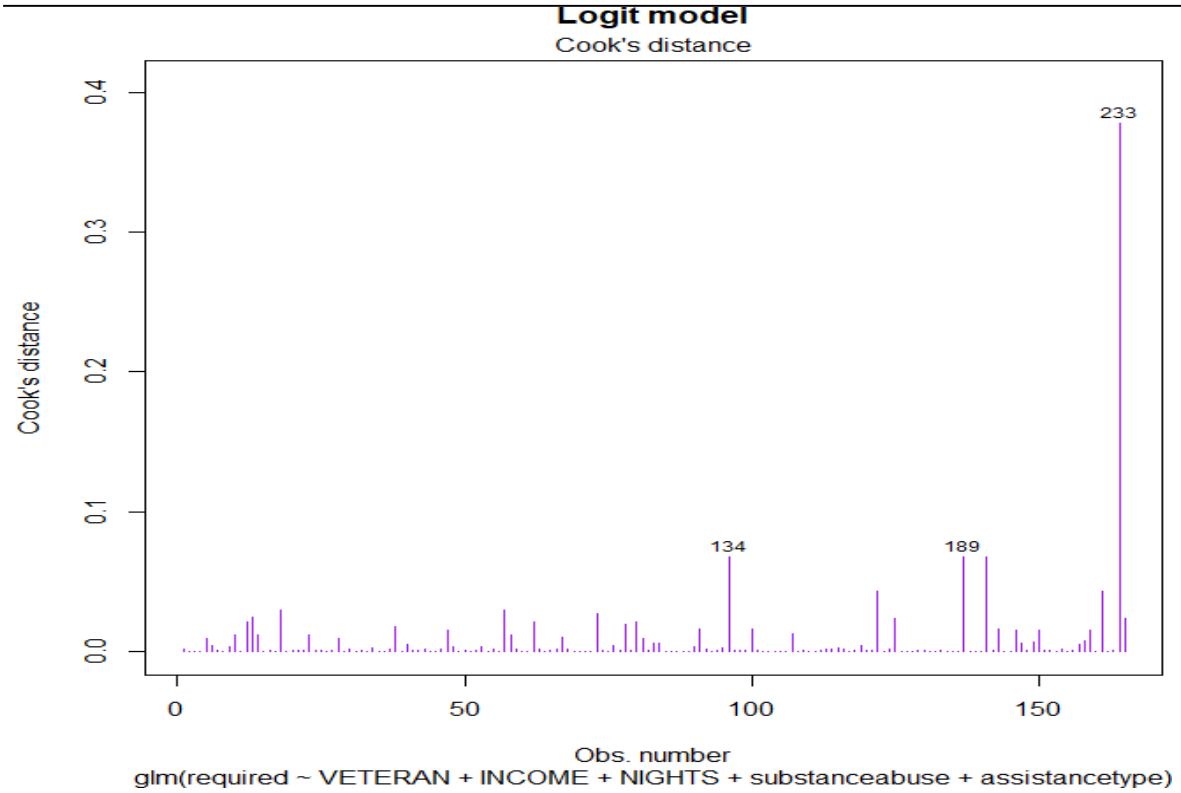
```
      data = train_set[, -1])

Coefficients:
                 Estimate   Std. Error  z value  Pr(>|z|)
(Intercept)   -3.5759450   1.8545509   -1.928   0.0538  .
AGE            0.0194642   0.0302709    0.643   0.5202
Male          -0.1508583   0.6344544   -0.238   0.8121
VETERAN        1.5488819   0.7765394    1.995   0.0461  *
INCOME        -0.0009894   0.0004063   -2.435   0.0149  *
NIGHTS         0.0175977   0.0077125    2.282   0.0225  *
substanceabuse 1.0892081   0.5943185    1.833   0.0668  .
completed     -0.4539838   0.7431643   -0.611   0.5413
probation      1.1868082   0.6189514    1.917   0.0552  .
assistancetype 3.3479199   0.5168837    6.477   9.35e-11 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 (Dispersion parameter for binomial family

taken to be 1)

      Null deviance: 228.68  on 164  degrees  of freedom
Residual deviance: 129.43  on 155  degrees  of freedom
AIC: 149.43

Number of Fisher Scoring iterations: 6

>
> #Stepwise procedure for Logit Model
> step(logit_model)
Start: AIC=149.43
required ~ AGE + Male + VETERAN + INCOME + NIGHTS + substanceabuse + completed + probation +
      assistancetype

                Df  Deviance    AIC
- Male           1    129.49  147.49
- completed      1    129.81  147.81
- AGE            1    129.85  147.85
<none>           1    129.43  149.43
- substanceabuse 1    132.86  150.86
- probation      1    133.15  151.15
- VETERAN        1    133.63  151.63
- INCOME         1    135.85  153.85
- NIGHTS         1    138.87  156.87
- assistancetype 1    193.03  211.03

Step: AIC=147.49
required ~ AGE + VETERAN + INCOME + NIGHTS + substanceabuse + completed + probation +
      assistancetype

                Df Deviance    AIC
- AGE            1    129.88  145.88
- completed      1    129.88  145.88
<none>           1    129.49  147.49
- substanceabuse 1    132.89  148.89
- probation      1    133.17  149.17
- VETERAN        1    133.72  149.72
- INCOME         1    137.16  153.16
- NIGHTS         1    138.99  154.99
- assistancetype 1    193.06  209.06

Step: AIC=145.88
required ~ VETERAN + INCOME + NIGHTS + substanceabuse + completed + probation + assistancetype

                Df Deviance    AIC
```

```
- completed          1   130.29  144.29
<none>                   129.88  145.88
- substanceabuse     1   133.21  147.21
- probation          1   133.46  147.46
- INCOME             1   137.19  151.19

- VETERAN            1   137.63  151.63
- NIGHTS             1   141.18  155.18
- assistancetype     1   194.55  208.55

Step: AIC=144.29
required ~ VETERAN + INCOME + NIGHTS + substanceabuse + probation + assistancetype

                    Df Deviance    AIC
<none>                  130.29  144.29
- probation          1   134.06  146.06
- substanceabuse     1   134.24  146.24
- VETERAN            1   138.00  150.00
- INCOME             1   138.09  150.09
- NIGHTS             1   142.23  154.23
- assistancetype     1   194.70  206.70

Call: glm(formula = required ~ VETERAN + INCOME + NIGHTS + substanceabuse + probation + assistancetype, family =
      binomial(link = logit),

     (Intercept)          VETERAN           INCOME           NIGHTS    substanceabuse              pro
      -3.0819513        1.7451183       -0.0009769        0.0159160        1.1398826                1
      data = train_set[, -1]) Coefficients:


Degrees of Freedom: 164 Total (i.e. Null); 158 Residual Null Deviance:
                                    228.7
Residual Deviance: 130.3                    AIC: 144.3
>
> #fitting the stepwise model to identify the significant predictors
> logit_model2 = glm(required ~ VETERAN + INCOME + NIGHTS + substanceabuse + probation
+                                 data=train_set[,-1], family = binomial(link="logit"))
> summary(logit_model2)

Call:
glm(formula = required ~ VETERAN + INCOME + NIGHTS + substanceabuse + probation +
      assistancetype, family = binomial(link = "logit"), data = train_set[, -1])

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -3.0819513 0.7224728 -4.266 1.99e-05 ***
VETERAN              1.7451183 0.6565380              2.658 0.00786 **
INCOME              -0.0009769 0.0003668 -2.663 0.00773 **
NIGHTS               0.0159160 0.0057191              2.783 0.00539 **
substanceabuse 1.1398826 0.5805051                    1.964 0.04958 *
probation            1.1576948 0.6015941              1.924 0.05431 .
assistancetype 3.3477566 0.5116948                    6.542 6.05e-11 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 (Dispersion parameter for binomial family

taken to be 1)

      Null deviance: 228.68          on 164          degrees of freedom
Residual deviance: 130.29          on 158          degrees of freedom
AIC: 144.29
```

```
Number of Fisher Scoring iterations: 6

>
> ## fitting the logit regression for the significant predictors
> logit_model1 = glm(required ~ VETERAN + INCOME + NIGHTS + substanceabuse + assistanc
+                                    data=train_set[,-1], family = binomial(link="logit"))
> summary(logit_model1)

Call:
glm(formula = required ~ VETERAN + INCOME + NIGHTS + substanceabuse + assistancetype, family =
        binomial(link = "logit"), data = train_set[,
        -1])
```

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.9172655 0.7092798 -4.113 3.91e-05 ***
VETERAN            1.6094324 0.6463826   2.490  0.01278 *
INCOME            -0.0009421 0.0003571 -2.638  0.00834 **
NIGHTS             0.0179621 0.0059105   3.039  0.00237 **
substanceabuse 1.1777510 0.5721741   2.058  0.03955 *
assistancetype 3.1720972 0.4788036   6.625 3.47e-11 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 (Dispersion parameter for binomial family

taken to be 1)

    Null deviance: 228.68  on 164  degrees of freedom
Residual deviance: 134.06  on 159  degrees of freedom
AIC: 146.06

Number of Fisher Scoring iterations: 6

>
>
> # Identify the Best logit Model with ANOVA
> # All Significant predictors Against Initial logit Model
> anova(logit_model1, logit_model, test="Chisq") Analysis of
Deviance Table

Model 1: required ~ VETERAN + INCOME + NIGHTS + substanceabuse + assistancetype Model 2: required ~ AGE +
Male + VETERAN + INCOME + NIGHTS + substanceabuse +
    completed + probation + assistancetype Resid. Df Resid. Dev Df
  Deviance Pr(>Chi)
1       159     134.06
2       155     129.43 4    4.624   0.3281
> # Step Model Against Initial Logit Model
> anova(logit_model2, logit_model, test="Chisq") Analysis of
Deviance Table

Model 1: required ~ VETERAN + INCOME + NIGHTS + substanceabuse + probation + assistancetype
Model 2: required ~ AGE + Male + VETERAN + INCOME + NIGHTS + substanceabuse + completed + probation +
    assistancetype
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       158     130.29
2       155     129.43  3  0.85424   0.8365
> anova(logit_model1, logit_model2, test="Chisq") #Removal of Probation is better Analysis of Deviance Table

Model 1: required ~ VETERAN + INCOME + NIGHTS + substanceabuse + assistancetype Model 2: required ~ VETERAN
+ INCOME + NIGHTS + substanceabuse + probation +
    assistancetype
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       159     134.06
2       158     130.29 1   3.7698  0.05219 .
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> #Therefore, logit_model1 is preferred for the logit model
> #find the summary of logit_model1
> summary(logit_model1)

Call:
glm(formula = required ~ VETERAN + INCOME + NIGHTS + substanceabuse + assistancetype, family =
    binomial(link = "logit"), data = train_set[,
    -1])
```

```
Coefficients:
                    Estimate Std. Error  z value Pr(>|z|)
(Intercept)       -2.9172655 0.7092798   -4.113 3.91e-05 ***
VETERAN            1.6094324 0.6463826    2.490 0.01278 *
INCOME            -0.0009421 0.0003571   -2.638 0.00834 **

NIGHTS             0.0179621 0.0059105    3.039 0.00237 **
substanceabuse     1.1777510 0.5721741    2.058 0.03955 *
assistancetype     3.1720972 0.4788036    6.625 3.47e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 228.68  on 164  degrees of  freedom
Residual deviance: 134.06  on 159  degrees of  freedom
AIC: 146.06

Number of Fisher Scoring iterations: 6

>
> #converting the coefficients to odd ratio
> exp(coef(logit_model1))
     (Intercept)          VETERAN           INCOME         NIGHTS substanceabuse assistancet 0.05408137
      4.99997228       0.99905834       1.01812434       3.24706331       23.85746
>
>
> #Test of assumption
> #install.packages("ResourceSelection")
> library(ResourceSelection)
> library(car)
> library(pscl)
> library(effects)
>
>
> # Test of Model adequacy
> # The choice of g>p+1, in our model p=5, so reasonable g is to choose g>5+1=6
> hoslem.test(logit_model1$y,fitted(logit_model1),g=6)
        Hosmer and Lemeshow goodness of fit (GOF) test data:
logit_model1$y, fitted(logit_model1)
X-squared = 7.337, df = 4, p-value = 0.1191

>
> #Collinearity Check
> #I will use the Variance inflation factor (VIF) to examine this
> vif(logit_model1)
          VETERAN           INCOME         NIGHTS substanceabuse assistancetype 1.733164
         1.153007         1.753950         1.338364         1.189282
>
> #Pseudo R-Squared
> #Estimate the Percent Explained by the model
> pR2(logit_model1)
fitting null model for pseudo-r2
         llh      llhNull           G2     McFadden         r2ML         r2CU
-67.0281063 -114.3420106   94.6278085    0.4137928    0.4364515    0.5819995
>
> #Checking for outliers in the model
> # I will use the Cooks Distance plot
> plot(logit_model1, which = 4, col = "purple", main="Logit model")
```




```
> #Checking the independence assumption
> #I will use the Plot of Deviance Residual
> plot(resid(logit_model1), type="l")

> #linearity: logit against predictor
> pp = logit_model1$fitted.values
> pp.logit = log(pp/(1-pp))
> par(mfrow=c(2,3))
> plot(pp.logit)
> plot(pp.logit~train_set$VETERAN)
> plot(pp.logit~train_set$INCOME)
> plot(pp.logit~train_set$NIGHTS)
> plot(pp.logit~train_set$substanceabuse)
> plot(pp.logit~train_set$assistancetype)
>
>
> # Checking the Accuracy of the model
> # install.packages("caret")
> # Precision and Accuracy for Logit model in Train Set
> p <- predict(logit_model1, train_set, type = "response")
> # If p exceeds threshold of 0.5, 1 else 0
> rule <- ifelse(p > 0.5, 1, 0)
> # Convert to factor: p_class
> p_class <- factor(rule, levels = levels(train_set[["required"]]))
> # Create confusion matrix for Train set
> caret::confusionMatrix(p_class, train_set[["required"]], positive = "1") Confusion Matrix and Statistics

          Reference Prediction 0 1
         0 70 12
         1 11 72

               Accuracy : 0.8606
                 95% CI : (0.7982, 0.9095)
    No Information Rate : 0.5091 P-Value
    [Acc > NIR] : <2e-16

                  Kappa : 0.7212
```

```
          Mcnemar's Test P-Value :  1

                     Sensitivity :  0.8571
                     Specificity :  0.8642
                  Pos Pred Value :  0.8675
                  Neg Pred Value :  0.8537
                      Prevalence :  0.5091
                  Detection Rate :  0.4364
            Detection Prevalence :  0.5030
               Balanced Accuracy :  0.8607

                'Positive' Class :  1

>
>
> # Precision and Accuracy for Logit model in Test Set
> p <- predict(logit_model1, test_set, type = "response")
> # If p exceeds threshold of 0.5, 1 else 0
> rule <- ifelse(p > 0.5, 1, 0)
> # Convert to factor: p_class
> p_class <- factor(rule, levels = levels(test_set[["required"]]))
> # Create confusion matrix for Test set
> caret::confusionMatrix(p_class, test_set[["required"]], positive = "1") Confusion Matrix and Statistics

          Reference Prediction 0 1
        0 31 6
        1 7 27

                     Accuracy :  0.8169
                       95% CI :  (0.7073, 0.8987)
          No Information Rate :  0.5352
          P-Value [Acc > NIR] :  6.544e-07

                        Kappa :  0.6327

          Mcnemar's Test P-Value :  1

                     Sensitivity :  0.8182
                     Specificity :  0.8158
                  Pos Pred Value :  0.7941
                  Neg Pred Value :  0.8378
                      Prevalence :  0.4648
                  Detection Rate :  0.3803
            Detection Prevalence :  0.4789
               Balanced Accuracy :  0.8170

                'Positive' Class :  1

>
>
> # ROC AUC for Train set
> par(mfrow=c(2,2))
> library(pROC)
> library(ROCR)
> prediction = predict(logit_model1, train_set, type="response")
> pred_ROCR = prediction(prediction, train_set$required)
> roc_ROCR        = performance(pred_ROCR, measure = "tpr", x.measure = "fpr")
> plot(roc_ROCR, main = "ROC curve: Train Logit", colorize = F)
> abline(a = 0, b = 1)
> auc_ROCR = performance(pred_ROCR, measure = "auc")
> auc_ROCR = auc_ROCR@y.values[[1]]
> auc_ROCR
[1] 0.8947678
>
> # ROC AUC for Test set
> prediction = predict(logit_model1, test_set, type="response")
```

```
> pred_ROCR = prediction(prediction, test_set$required)
> roc_ROCR        = performance(pred_ROCR, measure = "tpr", x.measure = "fpr")
> plot(roc_ROCR, main = "ROC curve: Logit Test", colorize = T)
> abline(a = 0, b = 1)
> auc_ROCR = performance(pred_ROCR, measure = "auc")
> auc_ROCR = auc_ROCR@y.values[[1]]
> auc_ROCR
[1] 0.9019139


>
```