TERM PROJECT- SCMA 851 PREDICTIVE ANALYTICS

(Due on October 8th 2021)

**Aims: Data analysis, transformation, model development and assessment. Please submit a word document that includes all of your answers for each question and submit your R scripting files with the report.**

MidWest.csv Prediction Scoring.xlsx

MidWest Airlines has entered into a partnership with the wireless phone company Telcom to sell the latter's service via direct mail. The MidWest-Airlines data file contains a subset of a data sample who has already received a test offer. About 13% accepted. You are asked to develop a model to classify MidWest customers as to whether they purchase a wireless phone service contract (target variable Phone\_sale), that can be used to predict classifications for additional customers.

1. Explain the role of Data Dictionary below using the CRISP-DM process for a successful predictive analytics project in the MidWest Co.

The Data Dictionary is part of the data preparation process of CRISP-DM. The data needs to be easily understandable especially to the average business user. The using the data dictionary helps to ensure data fields are not duplicated and helps to identify which data can be tied to each other.

1. First, transform variables based on the data dictionary given below and form a dataframe in RStudio. Present summary statistics for *Balance, Bonus\_miles, and Topflight* variables.

> summary(midwest.df$Balance)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0 15832 36227 63931 79211 1704838

> summary(midwest.df$Bonus\_miles)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0 1729 8000 17980 24356 263685

> summary(midwest.df$Topflight)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.0000 0.0000 0.0000 0.1721 0.0000 1.0000

1. Develop a logistic regression model to predict customer purchase behavior (Target variable= Phone\_sale). Report the model output. Which factors are statistically significant predictors of the purchase decision. Present Odds Ratios of the significant variables with 95% Confidence Intervals. Plot the ROC curve of the model.

Call:

glm(formula = Phone\_sale ~ ., family = binomial(link = "logit"),

data = train.df.LR2)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.2015 -1.0712 -0.7816 1.1904 2.6603

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.0255561021 0.0610843231 -16.789 < 0.0000000000000002 \*\*\*

Balance -0.0000026921 0.0000004118 -6.537 0.0000000000628 \*\*\*

Bonus\_miles 0.0000040524 0.0000014805 2.737 0.006199 \*\*

Bonus\_trans 0.0246716526 0.0040215801 6.135 0.0000000008526 \*\*\*

Flight\_miles\_12mo 0.0000423464 0.0000228972 1.849 0.064399 .

Online\_12 0.1727066288 0.0551152023 3.134 0.001727 \*\*

Email 0.1973611081 0.0584974790 3.374 0.000741 \*\*\*

Any\_cc\_miles\_12mo 0.6212114912 0.0695256748 8.935 < 0.0000000000000002 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 8316.8 on 6083 degrees of freedom

Residual deviance: 7901.3 on 6076 degrees of freedom

AIC: 7917.3

Number of Fisher Scoring iterations: 4

statistically significant variables are: Balance, Bonus\_miles,Bonus\_trans, Flight\_miles\_12mo, Online\_12,Email, Any\_cc\_miles\_12mo

Chart, line chart

Description automatically generated

|  |  |  |  |
| --- | --- | --- | --- |
|  | OR | 2.50% | 97.50% |
| (Intercept) | 0.358597 | 0.3179293 | 0.4039577 |
| Balance | 0.9999973 | 0.9999965 | 0.9999981 |
| Bonus\_miles | 1.0000041 | 1.0000012 | 1.000007 |
| Bonus\_trans | 1.0249785 | 1.0169572 | 1.0331207 |
| Flight\_miles\_12mo | 1.0000423 | 0.9999979 | 1.0000879 |
| Online\_12 | 1.1885174 | 1.0729515 | 1.3332712 |
| Email | 1.2181839 | 1.0863713 | 1.3663889 |
| Any\_cc\_miles\_12mo | 1.8611815 | 1.6242842 | 2.1332387 |

1. Use a step() function in R for feature selection. Which factors are significant in this final model? Present the ORs for these variables. Plot the ROC curve of the model.

statistically significant variables are: Balance, Bonus\_miles, Bonus\_trans, Online\_12, Email, Any\_cc\_miles\_12mo

|  |  |  |  |
| --- | --- | --- | --- |
|  | OR | 2.50% | 97.50% |
| (Intercept) | 0.3586907 | 0.3180305 | 0.4040418 |
| Balance | 0.9999974 | 0.9999966 | 0.9999982 |
| Bonus\_miles | 1.000004 | 1.0000012 | 1.000007 |
| Bonus\_trans | 1.0276192 | 1.0200717 | 1.0352927 |
| Online\_12 | 1.1867492 | 1.0719779 | 1.3307876 |
| Email | 1.2284429 | 1.0959032 | 1.3774173 |
| Any\_cc\_miles\_12mo | 1.792184 | 1.5734042 | 2.0418631 |

Chart, line chart

Description automatically generated

1. Chart, line chart

   Description automatically generatedChart, line chart

   Description automatically generatedConstruct a Neural Network model. Change the number of nodes in the hidden layer from 20 to 50 and rerun the model. Plot the ROC curves of both models.

Hidden Layer 50

Hidden Layer 20

1. Fill the following table for the models developed.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **MODEL** | **ACCURACY** | **PRECISION** | **SENSITIVITY** | **F1** | **AUC** |
| Logistic regression 1 | 0.6009 | 0.7436 | 0.4122 | 0.5779 | 64.9% |
| Logistic regression 2 | 0.5976 | 0.7425 | 0.4061 | 0.5743 | 64.7% |
| Neural Network (20 Nodes) | 0.6029 | 0.7321 | 0.4321 | 0.5821 | 63.4% |
| Neural Network (50 Nodes) | 0.6075 | 0.7656 | 0.3985 | 0.5820 | 64.2% |

1. Which model would you suggest marketing department should use for targeting customers? Summarize your results in a discussion.

The Neural Network (50 Nodes) is the best model according to the above table. The model has better statistics than both logistic regressions and better precision, accuracy and AUC than the Neural Network (20 nodes).

1. Use one of the models you developed above and the provided records of new customers in the *PredictionScoring.xlsx* data file for predicting if they would by the Telcom service as a result of the direct mailing campaign or not.

Using Logistic Regression 2 it shows that customers 1, 5, and 10 would buy but the other customers would not.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Customer | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| Result | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |

**Supplement: Data Dictionary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **FIELD NAME** | **DATA TYPE** | **MAX DATA LENGTH** | **RAW DATA OR TELCOM CREATED FIELD?** | **DESCRIPTION** |
| ID# | NUMBER |  | Telcom | Unique ID |
| Topflight | CHAR | 1 | Raw | Indicates whether flyer has attained elite "Topflight" status, 1  = yes, 0 = no |
| Balance | NUMBER | 8 | Raw | Number of miles eligible for award travel |
| Qual\_miles | NUMBER | 8 | Raw | Number of miles counted as qualifying for Topflight status |
| cc1\_miles? | CHAR | 1 | Raw | Has member earned miles with airline freq. flyer credit card in  the past 12 months (1=Yes/0=No)? |
| cc2\_miles? | CHAR | 1 | Raw | Has member earned miles with Rewards credit card in the past 12  months (1=Yes/0=No)? |
| cc3\_miles? | CHAR | 1 | Raw | Has member earned miles with Small Business credit card in the past 12 months (1=Yes/0=No)? |
| Bonus\_miles | NUMBER |  | Raw | Number of miles earned from non-flight bonus transactions in  the past 12 months |
| Bonus\_trans | NUMBER |  | Raw | Number of non-flight bonus transactions in the past 12 months |
| Flight\_miles\_12mo | NUMBER |  | Raw | Number of flight miles in the past 12 months |
| Flight\_trans\_12 | NUMBER |  | Raw | Number of flight transactions in the past 12 months |
| Online\_12 | NUMBER |  | Raw | Number of online purchases within the past 12 months |
| Email | CHAR | 1 | Raw | E-mail address on file. 1= yes, 0  =no? |
| Club\_member | NUMBER |  | Telcom | Member of the airline's club (paid membership), 1=yes, 0=no |
| Any\_cc\_miles\_12mo | NUMBER |  | Telcom | Dummy variable indicating whether member added miles on  any credit card type within the past 12 months (1='Y', 0='N') |
| Phone\_sale | NUMBER |  | Telcom | Dummy variable indicating whether member purchased Telcom service as a result of the direct mail campaign (1=sale,  0=no sale) |