# Module 9

## Affairs

**Package**

**Package Usage in R**

AER - To load the Affairs dataset

Car - To check the VIF (multi collinearity) of the variables

InformationValue - To check the misClassError and plot ROC

**Package Usage in Py**

Pandas - Data Manipulation

numpy - Scientific Calculation

matplotlib.pyplot - Plot and Visualization

statsmodels.formula.api - Used to run Regression Models

sklearn.model\_selection import train\_test\_split- Used in data splitting to training and testing.

sklearn import metrics - Used to calculate true and false prediction of the model

sklearn.metrics import classification\_report - Used to check precision , recall by classification report

**Loading the data**

Installed AER package and loaded the dataset

install.packages('AER')

data(Affairs,package="AER")

**Plot**

**Box plot** : Used to find the Outliner in the variables. Outliner found in Age and affairs

**Plot**

Age and Affairs Plot : Every age group are in affairs . We cant to group into a age group

Affairs and Years Marriage : Newly wedded couples to old age are in affairs.

**Histogram:**

Affairs: Dataset has less than 5% of the dataset has affairs

Age: Dataset has majority of age between 21 to 35

Years Marriage: Dataset has more 15 years marriage people then newly marriaged

**EDA**

1. No Missing Value found
2. Outliers

Found in Age and Affairs Column - No treatment was conducted

Used 3Rd IQR to found the Outlier

1. Dummy Variables

Created dummy variable for gender and childrens

Deleted the extra dummy variable to avoid Dummy trap

1. Typecasting

Converted Affairs entry to 0 or 1. Entries equal or greater than 1 is converted to 1.

**Modeling**

**Build-ed varies model to find the Less AIC value model**

**model** <- glm(affairs ~ ., data = Affairs, family = "binomial")

**#AIC 627.51**

**model2** <- glm(affairs ~ age+ yearsmarried +education+ occupation +rating +gendermale +childrenyes , data = Affairs, family = "binomial")

**#AIC 638.92**

**model3** <- glm(affairs ~ age+ yearsmarried +gendermale +childrenyes, data = Affairs, family = "binomial")

**#AIC 663.87**

**model4** <- glm(affairs ~ age+ yearsmarried, data = Affairs, family = "binomial")

**#AIC 665.4**

**model5** <- glm(affairs ~ age+ yearsmarried +religiousness +occupation+ rating + gendermale +childrenyes, data = Affairs, family = "binomial")

**#AIC 625.68**

**model6** <- glm(affairs ~ age+ yearsmarried + occupation+ rating+ gendermale +childrenyes, data = Affairs, family = "binomial")

**#AIC 637.27**

**model7** <- glm(affairs ~ yearsmarried +occupation +rating +childrenyes, data = Affairs, family = "binomial")

**#AIC 640.37**

**model8** <- glm(affairs ~ religiousness+ education +occupation +rating, data = Affairs, family = "binomial")

**#AIC 636.93**

**Model 5 has less AIC value compared to other models with the accuracy of 0.7670549**

**Data Partitioning**

Splitting the data into 80:20 Ratio. Training : Testing

**Finalmodel** <- glm(affairs ~ age+ yearsmarried +religiousness +occupation+ rating + gendermale +childrenyes, data = train, family = "binomial")

**AIC: 488.94**

**Prediction the modeling in Test data.**

**Model Accuracy and Other Test**

**Calculating the Confusion matrix to calculate the accuracy of the model**

**Final Test Model Accuracy: 0.7438017**

**Calculating the VIF : All the variables are less than 10 in the VIF test**

**Miscalculation Error: 0.1736**

**ROC Curve: 0.7597**

1. **Bank**

**Package Usage in R**

Car - To check the VIF (multi collinearity) of the variables

InformationValue - To check the misClassError and plot ROC

**Package Usage in Py**

Pandas - Data Manipulation

numpy - Scientific Calculation

matplotlib.pyplot - Plot and Visualization

statsmodels.formula.api - Used to run Regression Models

sklearn.model\_selection import train\_test\_split- Used in data splitting to training and testing.

sklearn import metrics - Used to calculate true and false prediction of the model

sklearn.metrics import classification\_report - Used to check precision , recall by classification report

**Loading the data**

Loading the Bank dataset

**Plot**

**Box plot** : Used to find the Outliner in the variables. Outliner found in Balance and Duration

**Plot**

Age and Balance Plot : Data is too scattered at the lower side of the plot. More Negative Balance found in the data balance.

Age and Loan Plot : loan is gotten by he age group of 20 to 65 and above 65 age has not gotten or applied for the loan

Age and Duration: Data is too scattered at the lower side of the plot.

**Histogram:**

Affairs: Dataset has less than 5% of the dataset has affairs

Age: Dataset has majority of age between 21 to 35

Duration: Majority of the dataset has lies between 01 to 500

**EDA**

1. No Missing Value found
2. Outliers

Found in Balance and Duration Column

Used 3Rd IQR to found the Outlier

4712 outliers Entries found in Balance- More Negative balance found and influencing the data set ,No Action taken as bank Balance and duration always varies and no limit bar can be put on it

3235 outliers Entries found in duration - No treatment was conducted. Duration varies to customers to customers and places a major role and Replacing the outliers will result in change the original result

**Modeling**

**Model1** <- glm(y ~ age + balance + housing + loan + duration + campaign + divorced + married, data = bank, family = "binomial")

**AIC 25684**

**Model2** <- glm(y ~ balance+housing+loan+duration+campaign+poutfailure+poutsuccess+married+jostudent, data = bank, family = "binomial")

**AIC 23461**

**Model3** <- glm(y ~ balance+housing+loan+duration+campaign+jostudent, data = bank, family = "binomial")

**AIC 25692**

**Model4** <- glm(y ~ age+default+balance+housing+ loan+duration+campaign+pdays+previous+poutfailure+ poutother+ poutsuccess+ poutunknown+ con\_cellular+ con\_telephone+ con\_unknown+ divorced+ married+ single+ joadmin.+ joblue.collar+ joentrepreneur+ johousemaid+ jomanagement+ joretired+ joself.employed+ joservices+ jostudent+ jotechnician+ jounemployed+ jounknown, data = bank, family = "binomial")

**AIC 22696**

**Model5** <- glm(y ~ age + balance + duration + campaign + previous + default + housing + loan + poutfailure + poutother + poutsuccess + con\_cellular + con\_telephone + divorced + married + joadmin. + joblue.collar + johousemaid + jomanagement + joretired + jostudent + jotechnician + jounemployed, data = bank, family ="binomial" )

summary(model5)

**AIC 22689**

**Model6** <- glm(y ~ age + balance + duration + campaign + previous + default + housing + loan + poutfailure + poutother + poutsuccess + con\_cellular + con\_telephone + divorced + married + joadmin. + joblue.collar + johousemaid + jomanagement + joretired + jostudent + jotechnician + jounemployed, data = bank, family ="binomial" )

summary(model6)

**AIC 22689**

**Model7** <- glm(y~ balance + duration + campaign + previous + housing + loan + poutfailure + poutother + poutsuccess + con\_cellular + con\_telephone + divorced + married + joadmin. + joblue.collar + johousemaid + jomanagement + joretired + jostudent + jounemployed, data = bank, family = "binomial")

summary(model7)

**AIC 22686**

**Model 7 has less AIC value compared to other models with the accuracy of 0.900**

**Data Partitioning**

Splitting the data into 80:20 Ratio. Training : Testing

finalmodel <- glm(y ~ balance + duration + campaign + previous + housing + loan + poutfailure + poutother + poutsuccess + con\_cellular + con\_telephone + divorced + married + joadmin. + joblue.collar + johousemaid + jomanagement + joretired + jostudent + jounemployed, data = train, family = "binomial")

**AIC: 18185**

**Prediction the modeling in Test data.**

**Model Accuracy and Other Test**

**Calculating the Confusion matrix to calculate the accuracy of the model**

**Final Test Model Accuracy: .8988**

**Calculating the VIF : All the variables are less than 10 in the VIF test**

**Miscalculation Error: 0.099**

**ROC Curve: 0.8883**

1. **Election**

**Package Usage in R**

Car - To check the VIF (multi collinearity) of the variables

InformationValue - To check the misClassError and plot ROC

**Package Usage in Py**

Pandas - Data Manipulation

numpy - Scientific Calculation

matplotlib.pyplot - Plot and Visualization

statsmodels.formula.api - Used to run Regression Models

sklearn.model\_selection import train\_test\_split- Used in data splitting to training and testing.

sklearn import metrics - Used to calculate true and false prediction of the model

sklearn.metrics import classification\_report - Used to check precision , recall by classification report

**Loading the data**

Loading the Election dataset

**Plot**

**Box plot** : Used to find the Outliner in the variables. Outliner found in Amount Spent.

**Plot**

Result and Year : Data shows 4 wins and 6 lose

Result and Money Spent : Need more data to draw a insight

Year and Popularity.Rank: Negative Linearity. Slow the popularity has raised

**Histogram:**

Amount.Spent: Majority money has been spent in minimal amount

Popularity.Rank:

**EDA**

1. NaN Value found in the dataset and row has been removed
2. Outliers

Found in Amount Spent

Used 3Rd IQR to found the Outlier

1 outliers Entries found in Amount Spent - Outlier Treatment not done as the Money plays a major role in election.

**Modeling**

**Model** <- glm(Result ~ Year + Amount.Spent+ Popularity.Rank, data = election, family = "binomial",control=glm.control(maxit=50))

AIC 8

**Model2** <- glm(Result ~ Popularity.Rank, data = election, family = "binomial")

summary(model2)

AIC 7.8191

**Model3** <- glm(Result ~ Year, data = election, family = "binomial")

**AIC 14.715**

**Model4** <- glm(Result ~ Amount.Spent, data = election, family = "binomial")

**AIC 16.501**

**Model5** <- glm(Result ~ Amount.Spent+ Popularity.Rank, data = election, family = "binomial")

**AIC 9.8177**

**Model6** <- glm(Result ~ Year + Popularity.Rank, data = election, family = "binomial")

**AIC 6**

**Model7** <- glm(Result ~ Year + Amount.Spent, data = election, family = "binomial")

**AIC 15.809**

**Model 6 has less AIC value compared to other models with the accuracy of 1**

**Data Partitioning**

Splitting the data into 70:30 Ratio. Training : Testing

**Finalmodel** <- glm(Result ~ Year + Popularity.Rank, data = train, family = "binomial")

**AIC: 6**

**Prediction the modeling in Test data.**

**Model Accuracy and Other Test**

**Calculating the Confusion matrix to calculate the accuracy of the model**

**Final Test Model Accuracy: .66**

**Calculating the VIF : All the variables are less than 10 in the VIF test**

**Miscalculation Error: 0.5**

**ROC Curve: 0.75**