Capstone Project

Assignment 2

Course code: CSA 1643

Course: DATA WARE HOUSING AND DATA MINING FOR DATA SCIENCE

S. No: 03

Name: B. VISHNUVARDHANREDDY

Reg. No: 192211820

Slot: C

Title: Fraudulent claims detection in insurance industry for data warehousing

Assignment Release Date:

Assignment Preliminary Stage (Assignment 2) submission Date:

Mentor Name: DR.G. SHANMUGAM

Mentor Phone number and Department: 90801 43805 and PURE AND APPLIED MATHEMATICS

R PROGRAM FOR Fraudulent claims detection in insurance industry for data warehousing

```
# Load required libraries
library(dplyr)
library(ggplot2)
library(caret)

# Load insurance claims data (replace 'claims_data.csv' with your dataset)
claims_data <- read.csv("claims_data.csv")

# Explore the data
summary(claims_data)
str(claims_data)

# Preprocess the data (handle missing values, encode categorical variables, scale numerical features, etc.)
# Example:
# Handle missing values
claims_data <- na.omit(claims_data)

# Encode categorical variables
```

```
claims_data <- dummyVars(~., data = claims_data)
%>% predict(claims_data)
# Split the data into training and testing sets
set.seed(123)
train_indices <-
createDataPartition(claims_data$Fraudulent, p = 0.8,
list = FALSE)
train_data <- claims_data[train_indices, ]
test_data <- claims_data[-train_indices, ]

# Train logistic regression model
model <- glm(Fraudulent ~ ., data = train_data, family =
binomial)

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

# Evaluate model performance
confusion_matrix <-
confusionMatrix(table(ifelse(predictions > 0.5, 1, 0),
test_data$Fraudulent))
print(confusion_matrix)
```

OUT PUT :

	Reference	
Prediction	Fraudulent	Non-Fraudulent
Fraudulent	TP	FP
Non-Fraudulent	FN	TN