|  |  |  |
| --- | --- | --- |
|  | **SRI KRISHNA COLLEGE OF TECHNOLOGY**  **(AN AUTONOMOUS INSTITUTION)**  **KOVAIPUDUR, COIMBATORE 641042** |  |

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

# **Detect Credit Card Fraud**

***A DATA ANALYTICS Mini Project Report***

|  |
| --- |
| ***Submitted by*** |
| **18TUCS220-SETHUPATHI.P**  **18TUCS211-SANJAY.S. P**  **18TUCS209-SABARI BALA MURUGAN** |

|  |  |  |
| --- | --- | --- |
| **S. NO** | **CONTENTS** | **PAGE NO.** |
| 1. | ABSTRACT |  |
| 2. | INTRODUCTION |  |
| 3. | CREDIT CARD FRAUD DETECTION USING ML |  |
| 4. | WHAT IS FRAUD DETECTIONSYSTEM? |  |
| 5. | DATASET |  |
| 6. | PROCESS AND WORKFLOW |  |
| 7. | PROGRAM |  |
| 8. | IMPLEMENTATION |  |
| 9. | CONCLUSION |  |

|  |  |
| --- | --- |
|  | |
|  | |
|  |  |
|  |  |
|  | **INDEX** |
|  |  |

**ABSTRACT:**

This project is a **DETECT CREDIT CARD FRAUD;** detection of credit cards. Detecting fraud transactions is of great importance for any credit card company. Credit card fraud detection is presently the most frequently occurring problem in the present world. This is due to the rise in both online transactions and e-commerce platforms. Credit card fraud generally happens when the card was stolen for any of the unauthorized purposes or even when the fraudster uses the credit card information for his use. In the present world, we are facing a lot of credit card problems. To detect the fraudulent activities the credit card fraud detection system was introduced. This project aims to focus mainly on various algorithms like Decision Trees, Logistic Regression, Artificial Neural Networks and finally, Gradient Boosting Classifier.

**INTRODUCTION:**

A credit card is a thin handy plastic card that contains identification information such as a signature or picture, and authorizes the person named on it to charge purchases or services to his account - charges for which he will be billed periodically. Today, the information on the card is read by automated teller machines (ATMs), store readers, bank and is also used in online internet banking system. They have a unique card number which is of utmost importance. Its security relies on the physical security of the plastic card as well as the privacy of the credit card number. There is a rapid growth in the number of credit card transactions which has led to a substantial rise in fraudulent activities. Credit card fraud is a wide-ranging term for theft and fraud committed using a credit card as a fraudulent source of funds in a given transaction. Generally, the statistical methods and many data mining algorithms are used to solve this fraud detection problem. Most of the credit card fraud detection systems are based on artificial intelligence, Meta learning and pattern matching.

In this report we will see the detect credit card fraud using R in R studio.

**CREDIT CARD FRAUD DETECTION PROJECT USING ML:**

The main goal of this machine learning project is to build a detection engine that detect fraud in credit card transactions. This R project is designed to help you understand the functioning of how a Fraud detection system works. We will be developing an Item Based Collaborative Filter and various machine learning algorithms.

**WHAT IS A FRAUD DETECTION SYSTEM?**

Fraud detection is a set of activities undertaken to prevent money or property from being obtained through false pretenses. Fraud detection is applied to many industries such as banking or insurance. In banking, fraud may include forging checks or using stolen credit cards. Other forms of fraud may involve exaggerating losses or causing an accident with the sole intent for the payout.

With an unlimited and rising number of ways someone can commit fraud, detection can be difficult to accomplish. Activities such as reorganization, downsizing, moving to new information systems or encountering a cybersecurity breach could weaken an organization's ability to detect fraud. This means techniques such as real-time monitoring for frauds is recommended. Organizations should look for fraud in financial transactions, location, devices used, initiated sessions and authentication systems. Fraud is typically an act which involves many repeated methods; making searching for patterns a general focus for fraud detection. Fraud detection can be separated by the use of statistical data analysis techniques or artificial intelligence.

Statistical data analysis techniques include the use of:

* Calculating statistical parameters
* Regression analysis
* Probability distributions and models.
* Data matching

AI techniques used to detect fraud include the use of:

* [Data mining](https://searchsqlserver.techtarget.com/definition/data-mining)- Which can classify, group and segment data to search through up to millions of transactions to find patterns and detect fraud.
* [Neural networks](https://searchenterpriseai.techtarget.com/definition/neural-network)- Which can learn suspicious looking patterns, and use those patterns to detect them further.
* [Machine learning](https://searchenterpriseai.techtarget.com/definition/machine-learning-ML)- Which can automatically identify characteristics found in fraud.
* [Pattern recognition](https://whatis.techtarget.com/definition/pattern-recognition)- Which can detect classes, clusters and patterns of suspicious behavior.

**DATASET:**

In order to build our fraud detection system, we have used the credit card Dataset. This dataset contain transactions made by credit cards. Card Transactions dataset that contains a mix of fraud as well as non-fraudulent transactions.

The dataset contains 492 frauds out of 284,807 transactions. Thus, it is highly unbalanced, with the positive (frauds) accounting for only 0.17%.

**PROCESS AND WORKFLOW:**

The main steps in a data science project include:

Importing the Datasets: importing the datasets that contain transactions made by credit cards.

Data Exploration:

In this section of the fraud detection ML project, we will explore the data that is contained in the credit card data frame. We will proceed by displaying the credit card data using the head() function as well as the tail() function. We will then proceed to explore the other components of this data frame.

Data Manipulation:

In this section of the R data science project, we will scale our data using the scale() function. We will apply this to the amount component of our creditcard\_data amount. Scaling is also known as feature standardization. With the help of scaling, the data is structured according to a specified range. Therefore, there are no extreme values in our dataset that might interfere with the functioning of our model.

Data Modeling:

After we have standardized our entire dataset, we will split our dataset into training set as well as test set with a split ratio of 0.80. This means that 80% of our data will be attributed to the train\_data whereas 20% will be attributed to the test data. We will then find the dimensions using the dim() function.

Fitting Logistic Regression Model:

In this section of credit card fraud detection project, we will fit our first model. We will begin with logistic regression. A logistic regression is used for modeling the outcome probability of a class such as pass/fail, positive/negative and in our case – fraud/not fraud.

Fitting a Decision Tree Model:

In this section, we will implement a decision tree algorithm. [Decision Trees](https://data-flair.training/blogs/r-decision-trees/) to plot the outcomes of a decision. These outcomes are basically a consequence through which we can conclude as to what class the object belongs to. We will now implement our decision tree model and will plot it using the rpart.plot() function. We will specifically use the recursive parting to plot the decision tree.

Artificial Neural Network:

[Artificial Neural Networks](https://data-flair.training/blogs/artificial-neural-network/) are a type of machine learning algorithm that are modeled after the human nervous system. The ANN models are able to learn the patterns using the historical data and are able to perform classification on the input data. We import the neuralnet package that would allow us to implement our ANNs. Then we proceeded to plot it using the plot() function. Now, in the case of Artificial Neural Networks, there is a range of values that is between 1 and 0. We set a threshold as 0.5, that is, values above 0.5 will correspond to 1 and the rest will be 0.

### Gradient Boosting (GBM):

[Gradient Boosting](https://data-flair.training/blogs/gradient-boosting-algorithm/) is a popular machine learning algorithm that is used to perform classification and regression tasks. This model comprises of several underlying ensemble models like weak decision trees. These decision trees combine together to form a strong model of gradient boosting.

**PROGRAM:**

library(ranger)

library(caret)

library(data.table)

creditcard\_data <- read.csv("/home/dataflair/data/Credit Card/creditcard.csv")

dim(creditcard\_data)

head(creditcard\_data,6)

tail(creditcard\_data,6)

table(creditcard\_data$Class)

summary(creditcard\_data$Amount)

names(creditcard\_data)

var(creditcard\_data$Amount)

sd(creditcard\_data$Amount)

head(creditcard\_data)

creditcard\_data$Amount=scale(creditcard\_data$Amount)

NewData=creditcard\_data[,-c(1)]

head(NewData)

library(caTools)

set.seed(123)

data\_sample = sample.split(NewData$Class,SplitRatio=0.80)

train\_data = subset(NewData,data\_sample==TRUE)

test\_data = subset(NewData,data\_sample==FALSE)

dim(train\_data)

dim(test\_data)

Logistic\_Model=glm(Class~.,test\_data,family=binomial())

summary(Logistic\_Model)

plot(Logistic\_Model)

library(pROC)

lr.predict <- predict(Logistic\_Model,train\_data, probability = TRUE)

auc.gbm = roc(test\_data$Class, lr.predict, plot = TRUE, col = "blue")

library(rpart)

library(rpart.plot)

decisionTree\_model <- rpart(Class ~ . , creditcard\_data, method = 'class')

predicted\_val <- predict(decisionTree\_model, creditcard\_data, type = 'class')

probability <- predict(decisionTree\_model, creditcard\_data, type = 'prob')

rpart.plot(decisionTree\_model)

library(neuralnet)

ANN\_model =neuralnet (Class~.,train\_data,linear.output=FALSE)

plot(ANN\_model)

predANN=compute(ANN\_model,test\_data)

resultANN=predANN$net.result

resultANN=ifelse(resultANN>0.5,1,0)

library(gbm, quietly=TRUE)

system.time(

model\_gbm <- gbm(Class ~ .

, distribution = "bernoulli"

, data = rbind(train\_data, test\_data)

, n.trees = 500

, interaction.depth = 3

, n.minobsinnode = 100

, shrinkage = 0.01

, bag.fraction = 0.5

, train.fraction = nrow(train\_data) / (nrow(train\_data) + nrow(test\_data))

)

)

# Determine best iteration based on test data

gbm.iter = gbm.perf(model\_gbm, method = "test")

model.influence = relative.influence(model\_gbm, n.trees = gbm.iter, sort. = TRUE)

#Plot the gbm model

plot(model\_gbm)

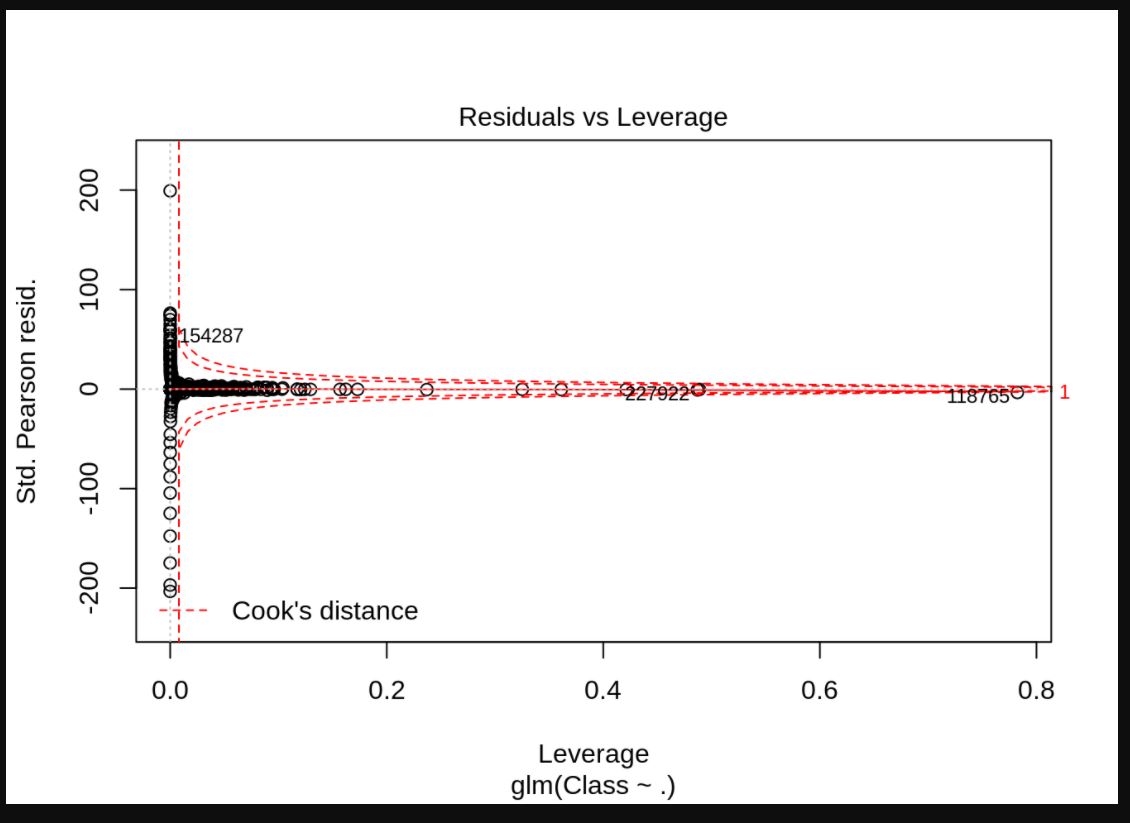
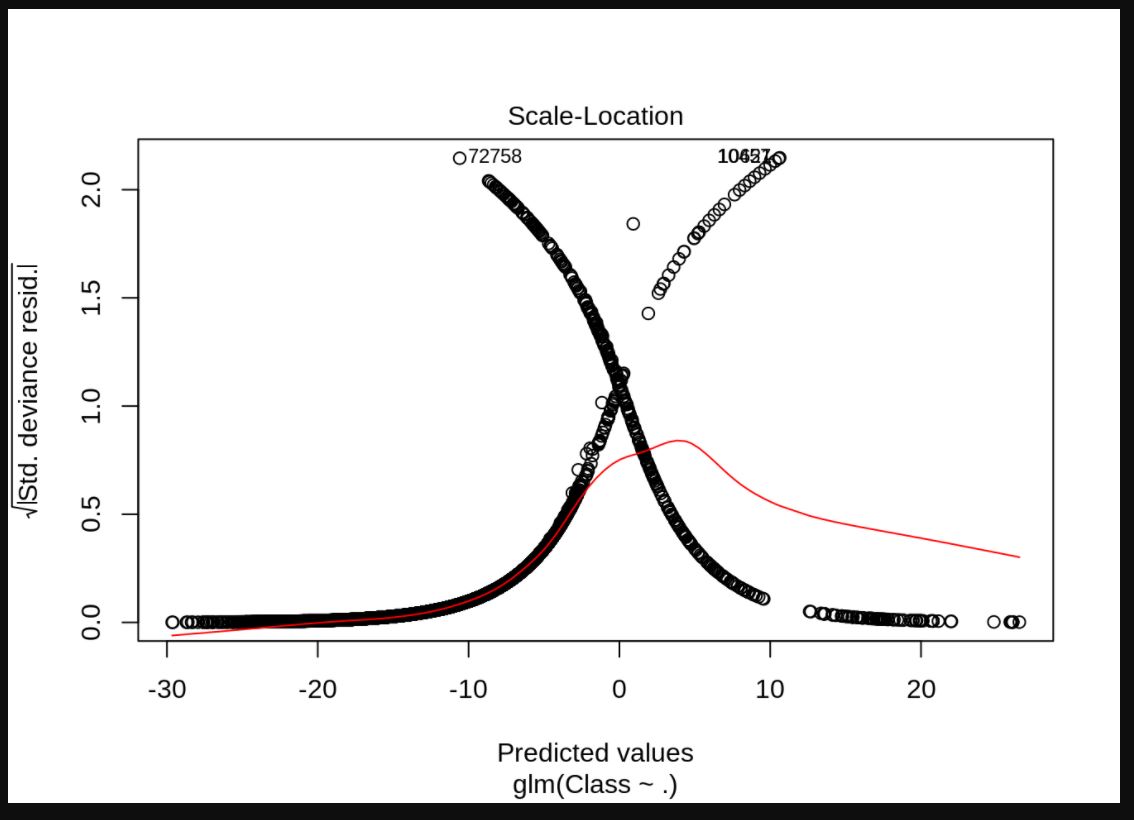
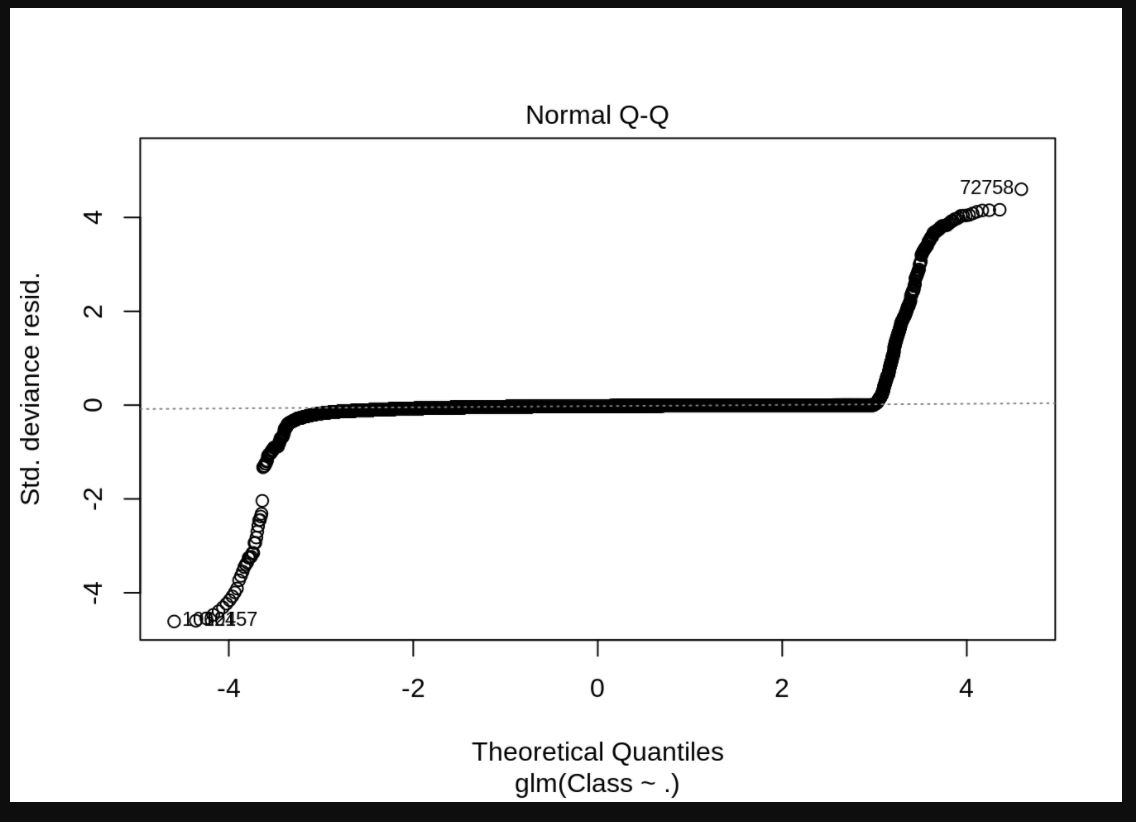
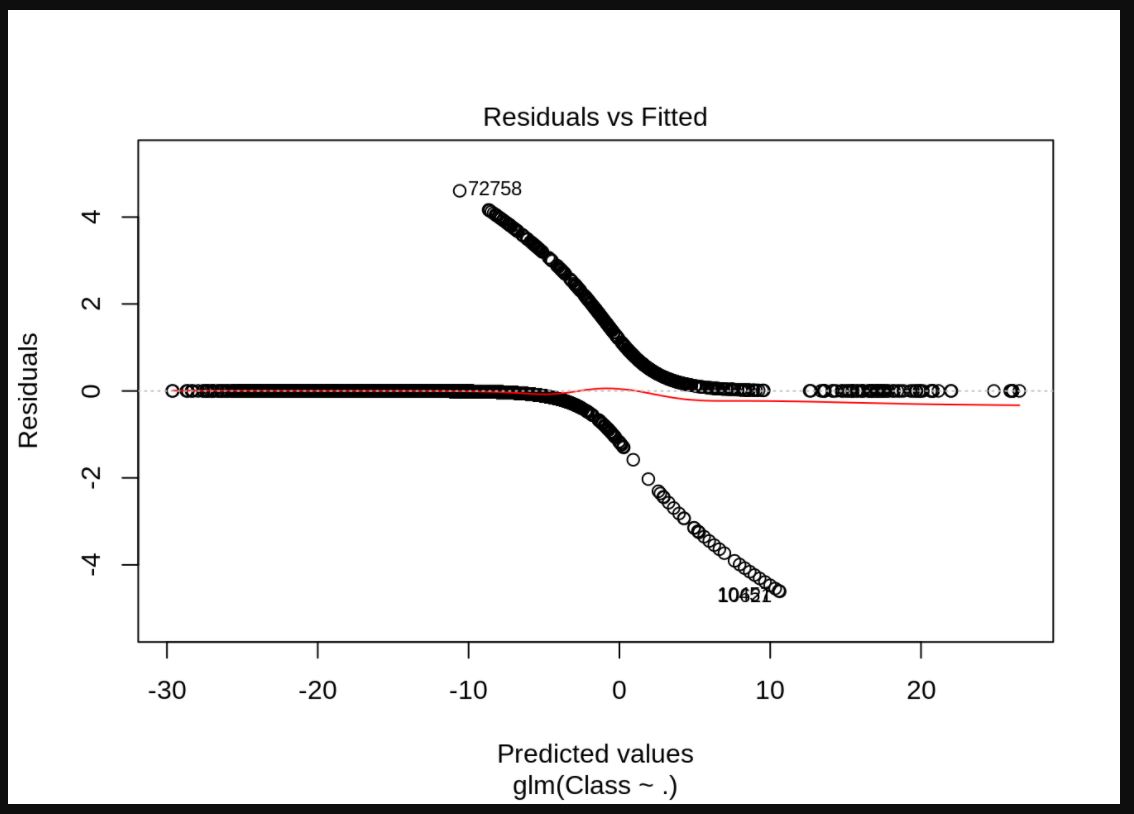
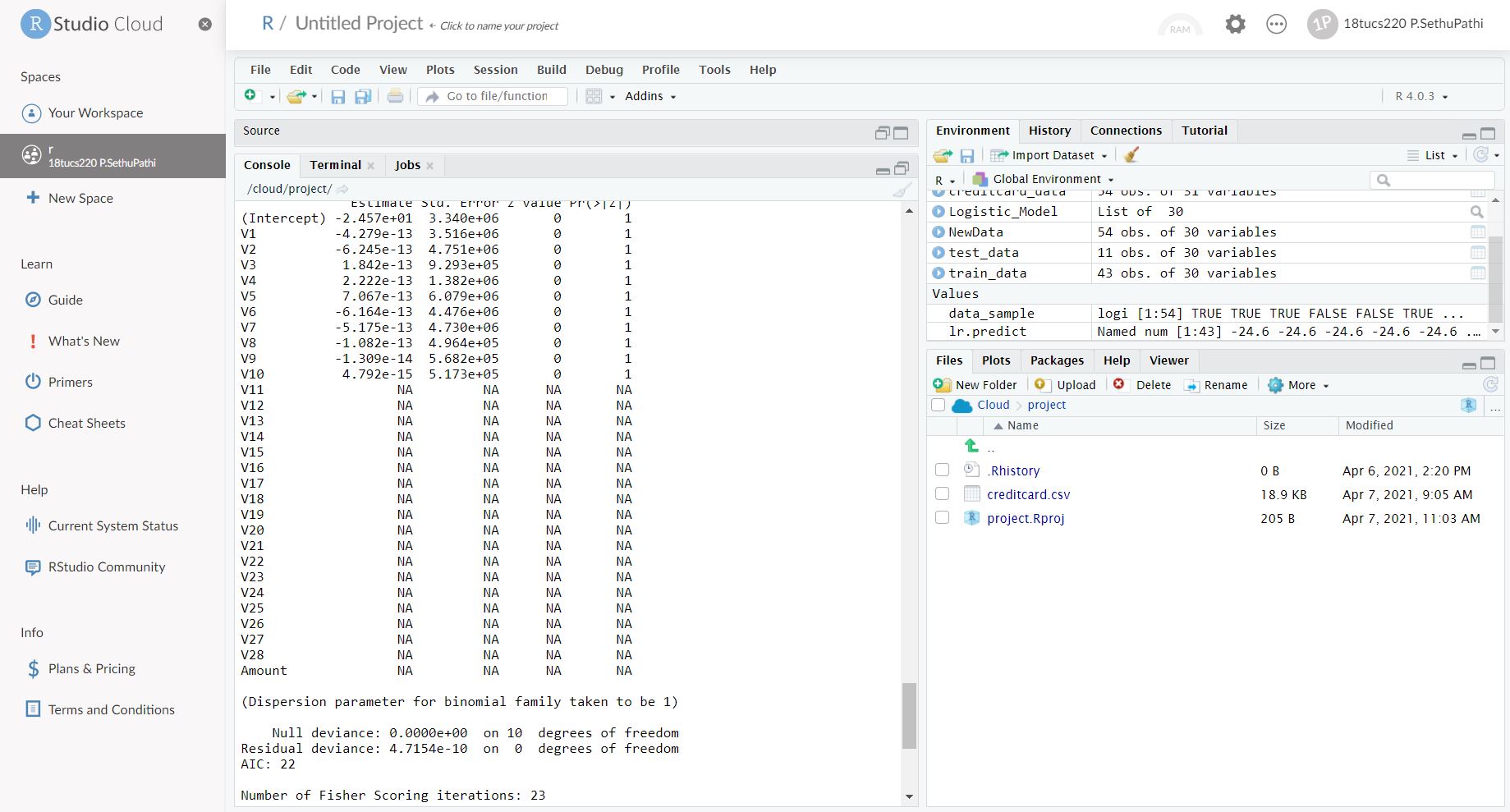
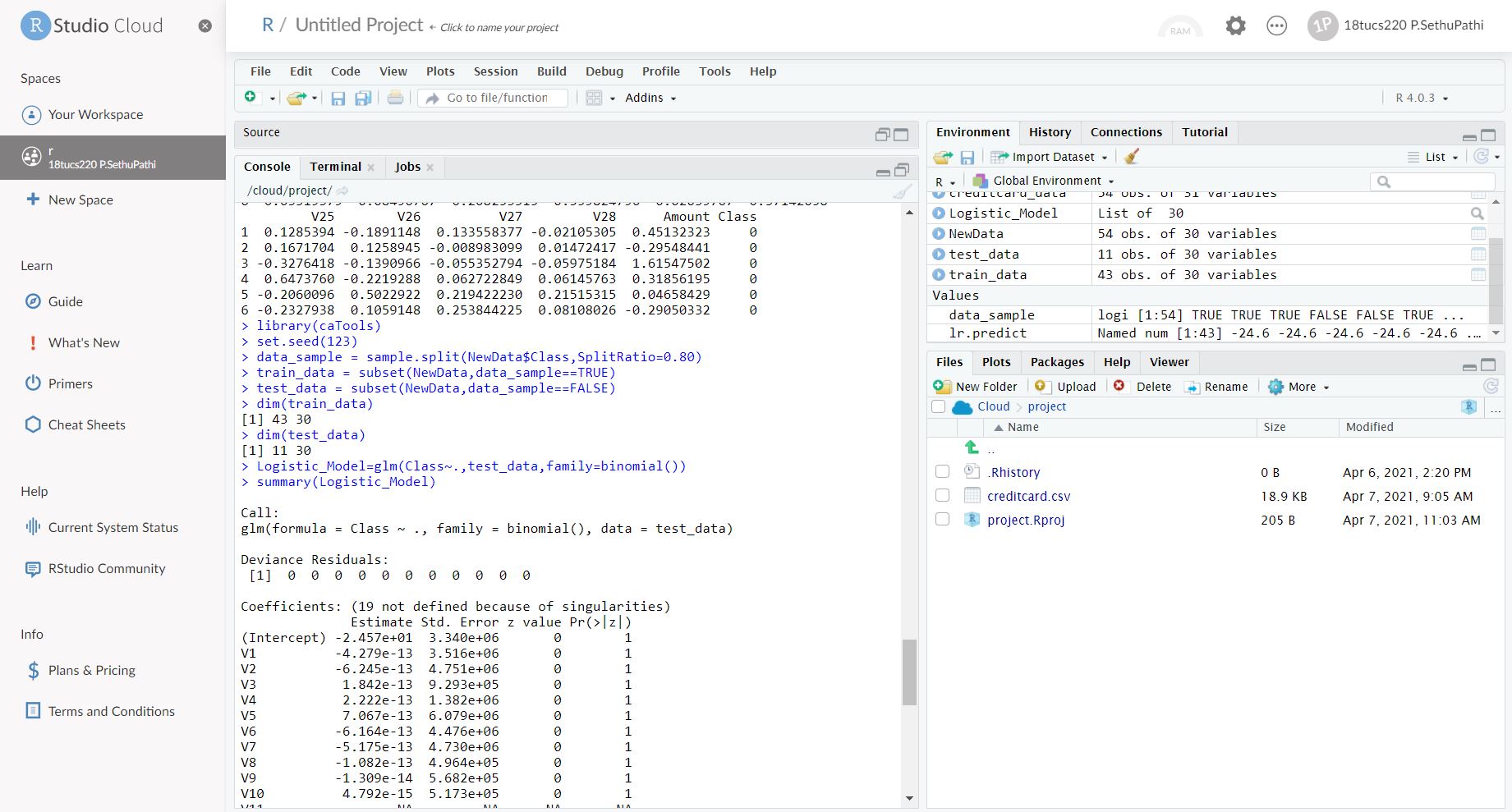
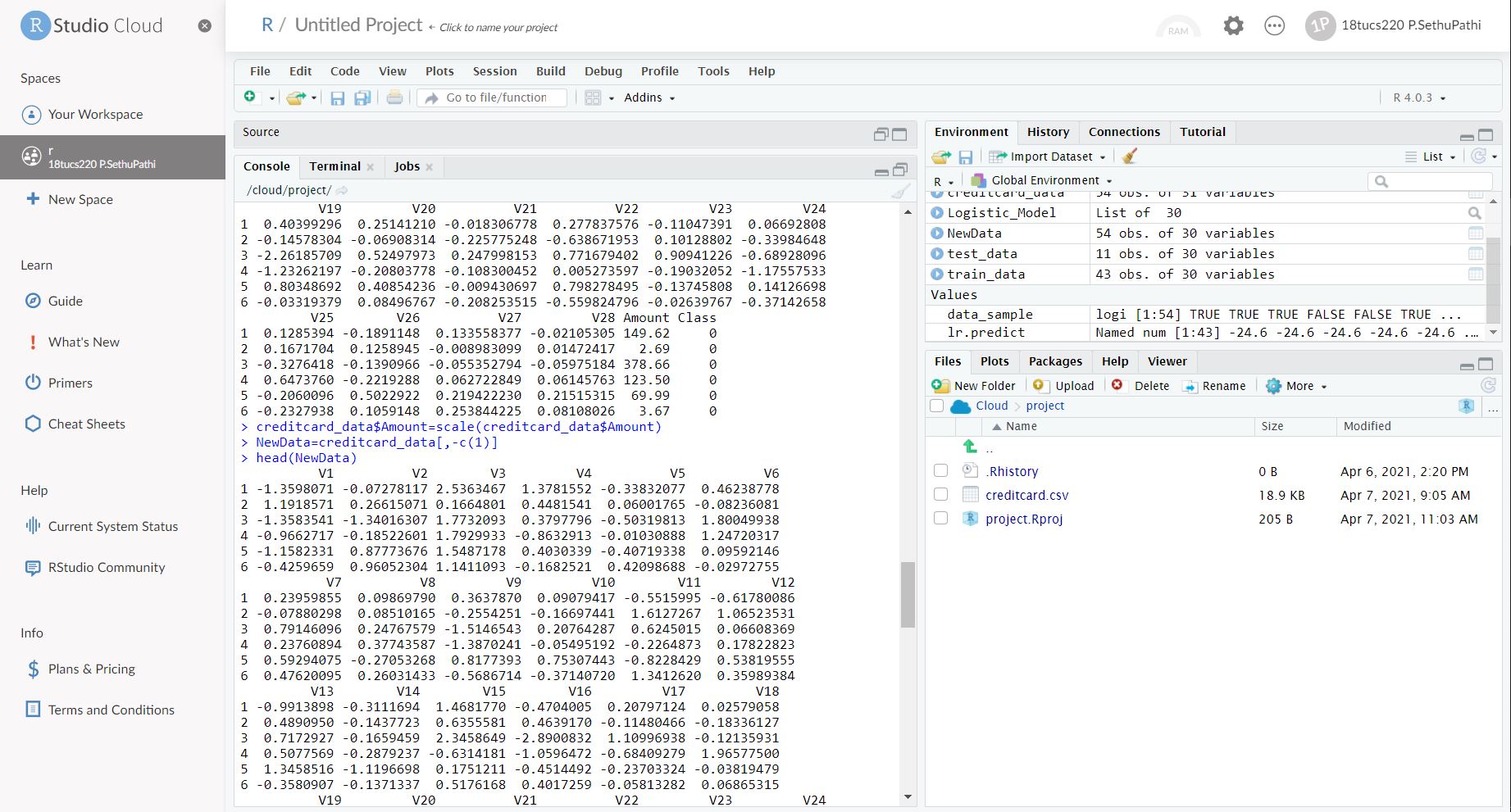
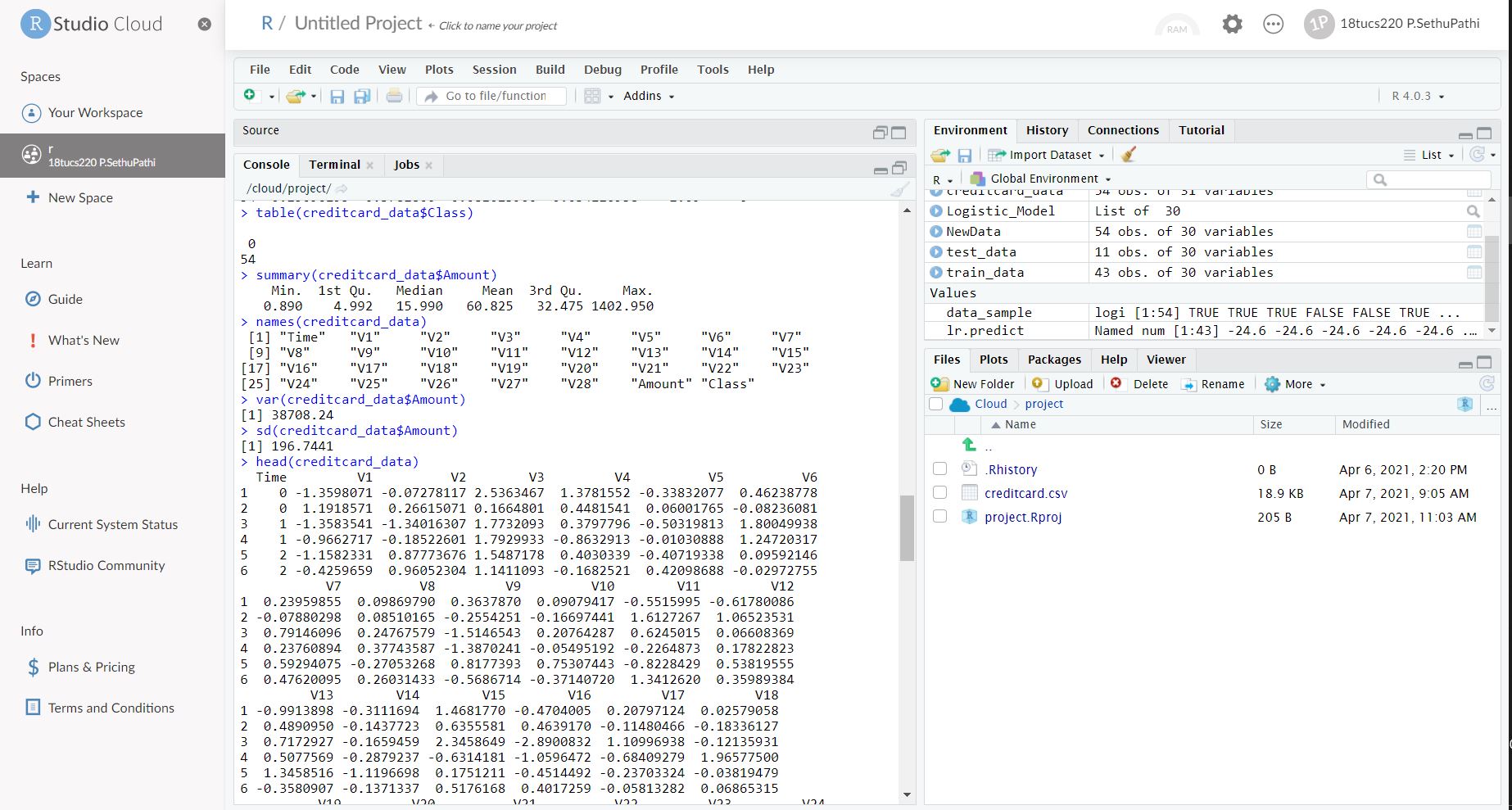
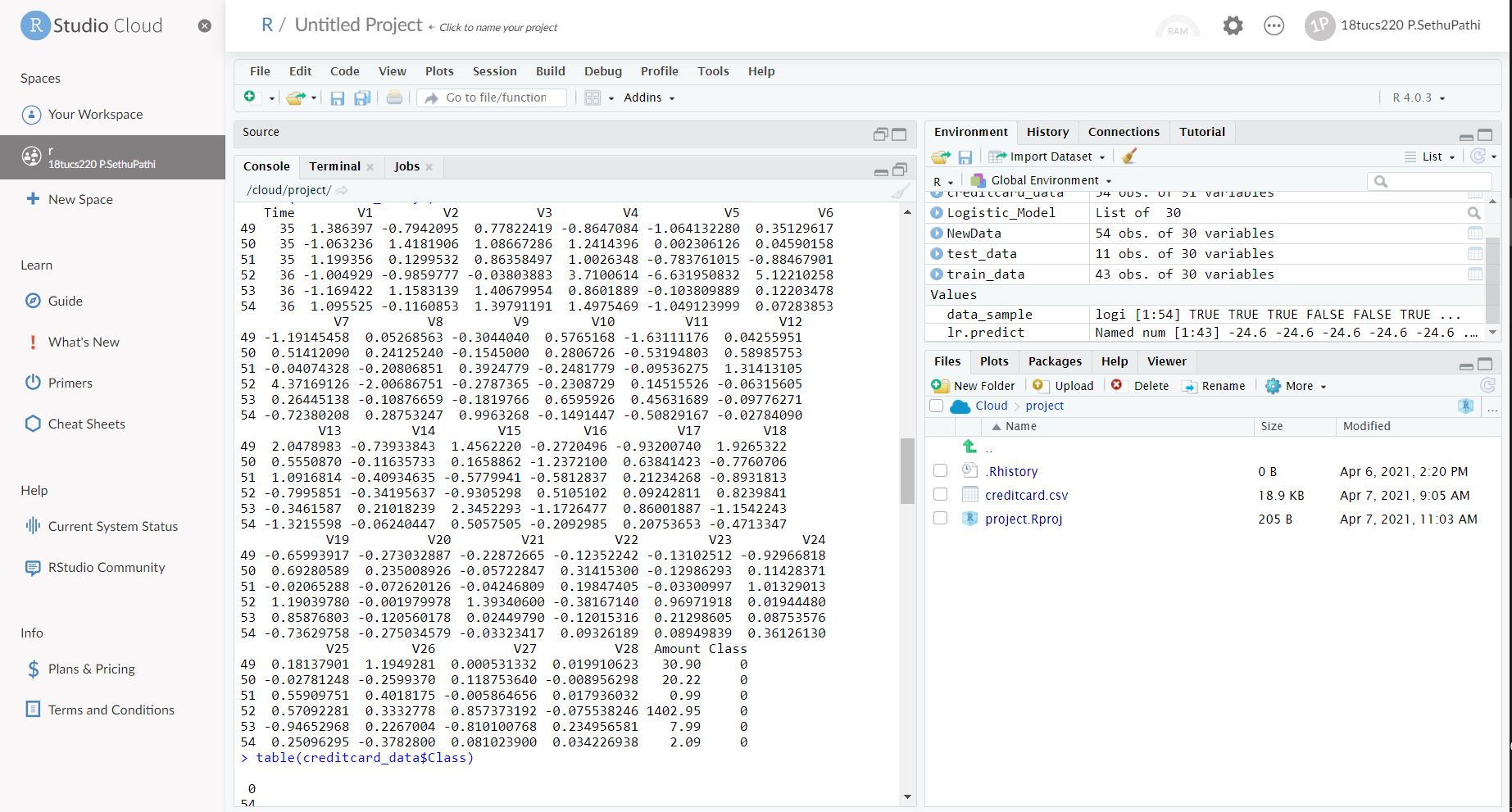
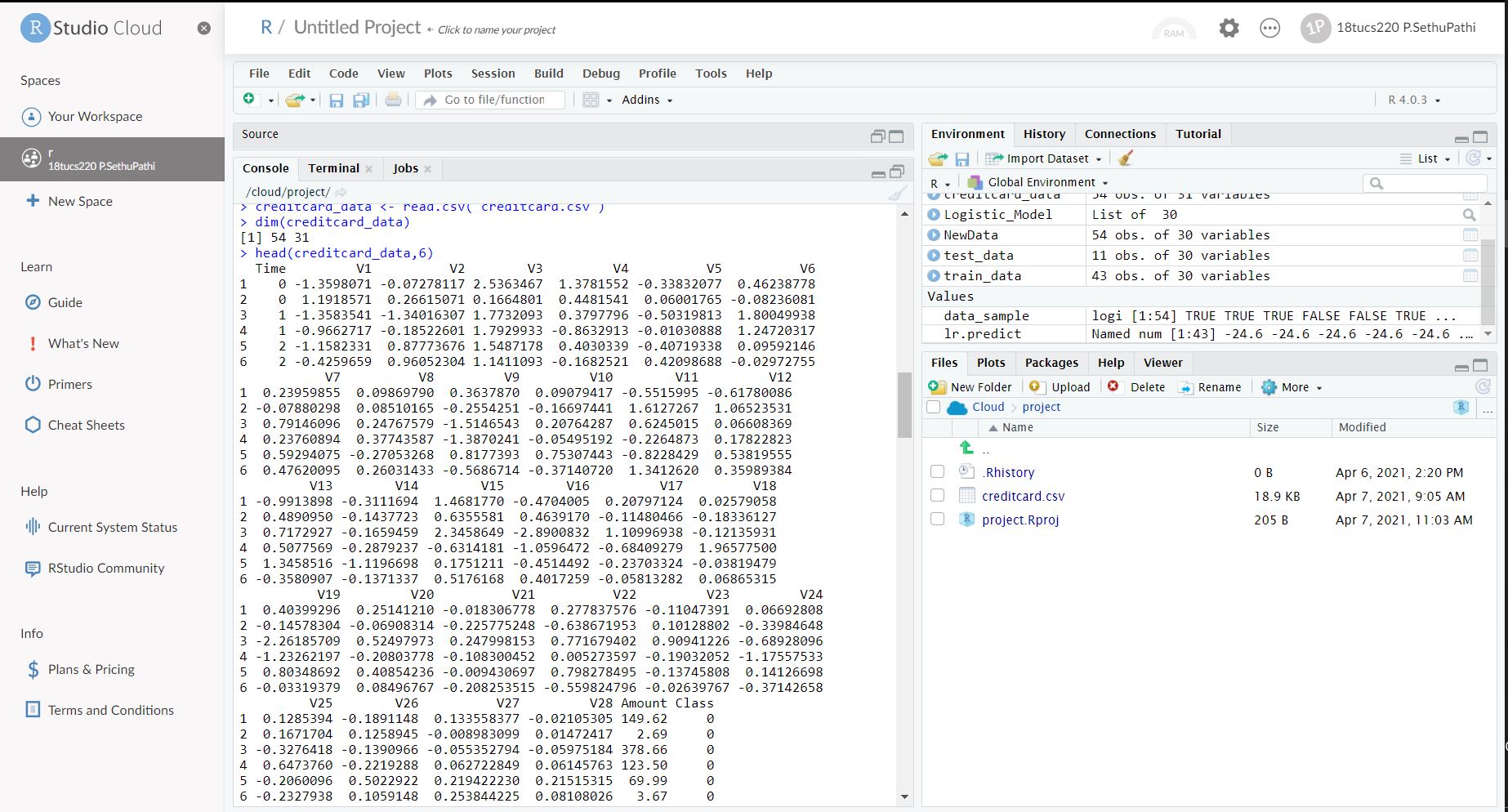
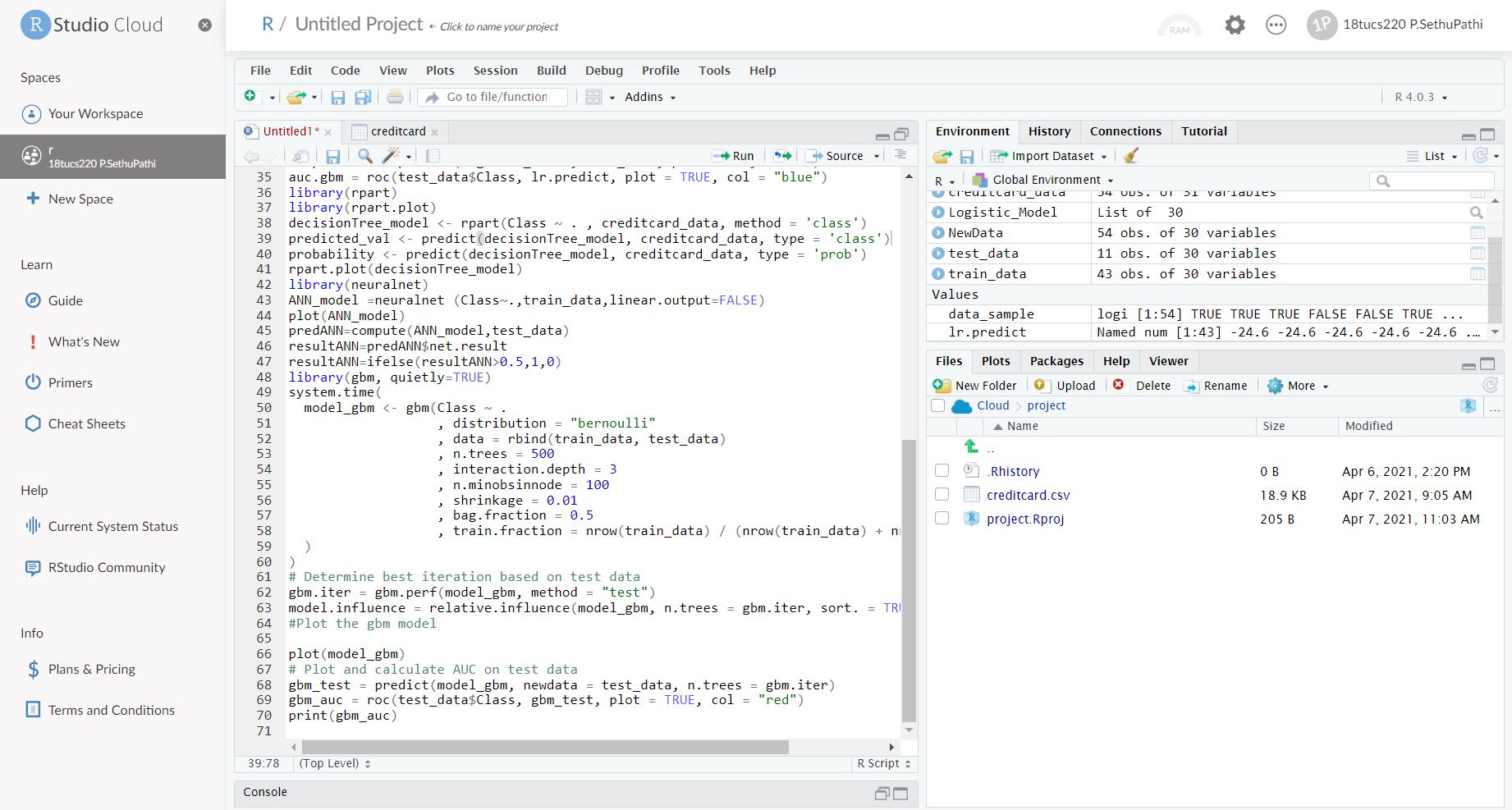
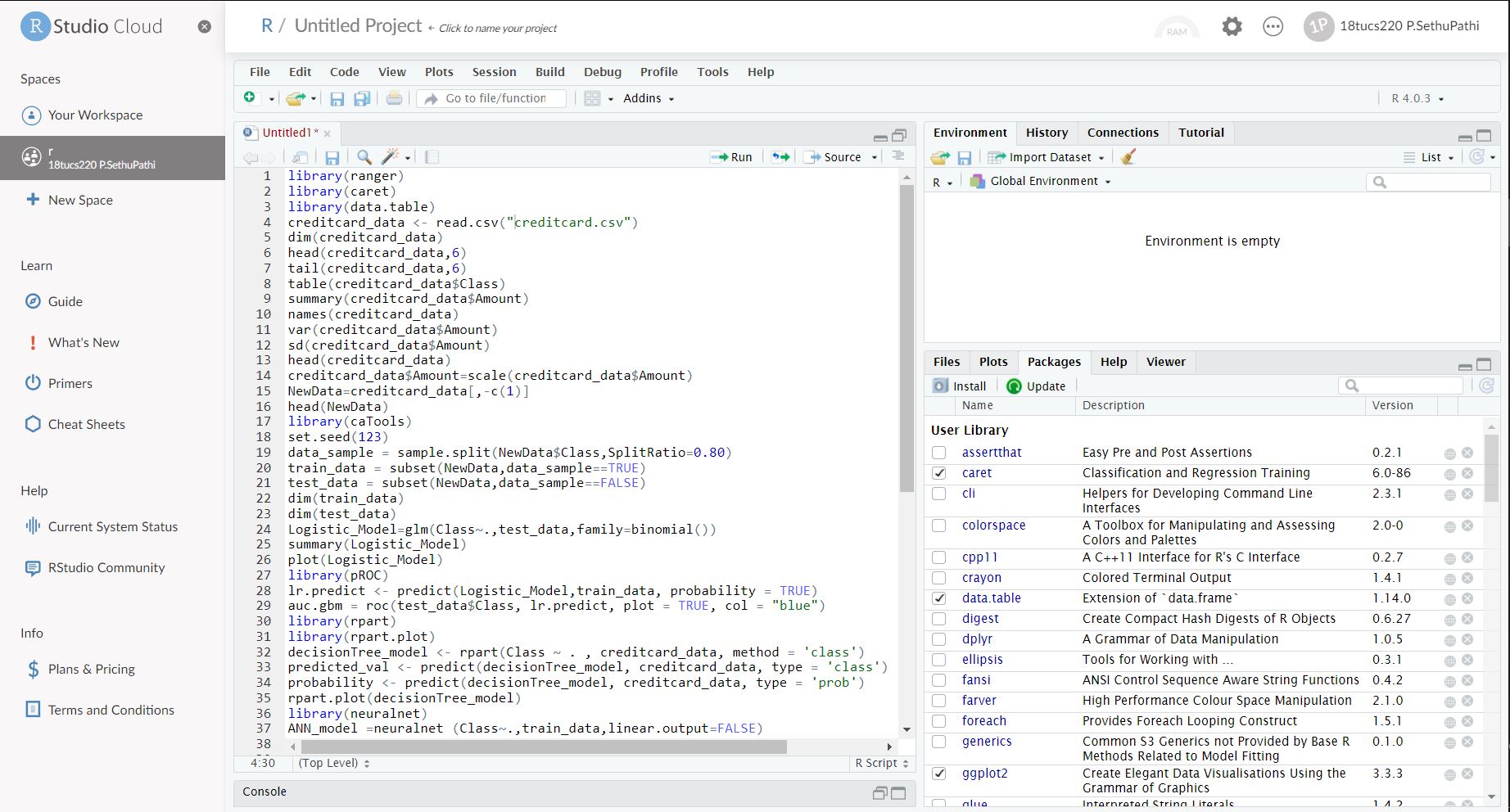
# Plot and calculate AUC on test data

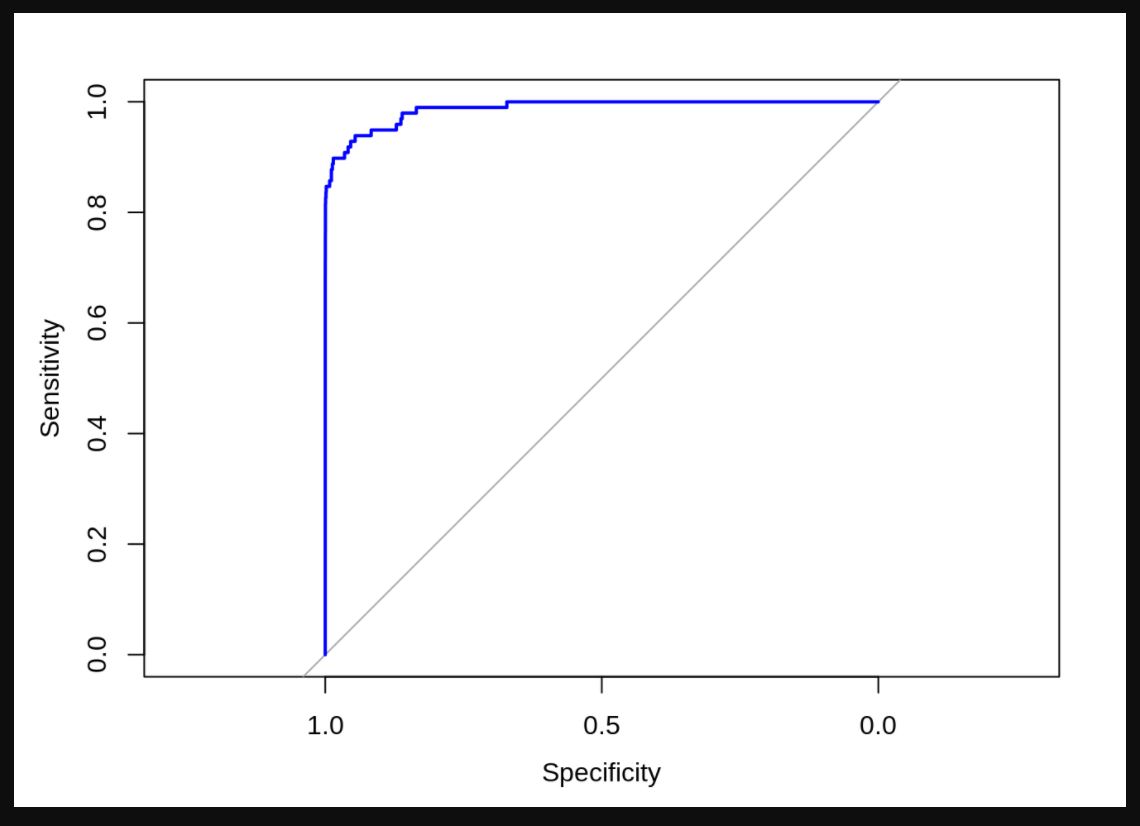
gbm\_test = predict(model\_gbm, newdata = test\_data, n.trees = gbm.iter)

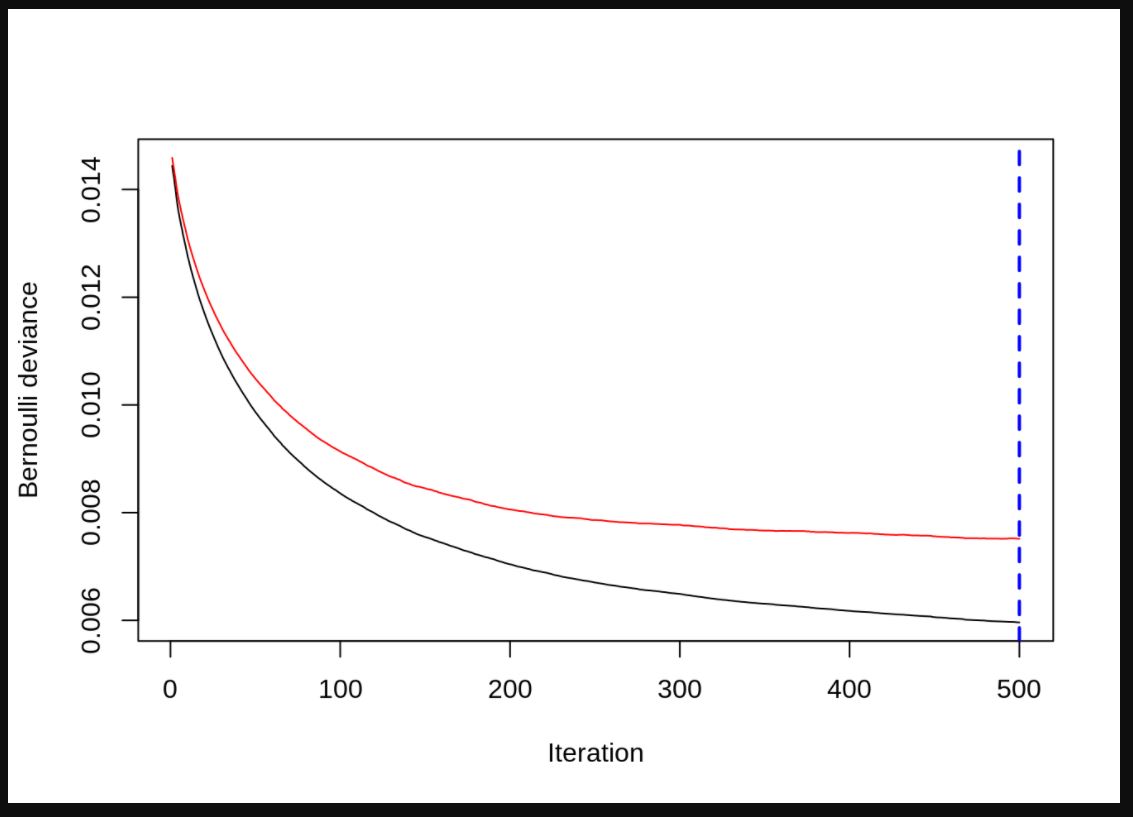
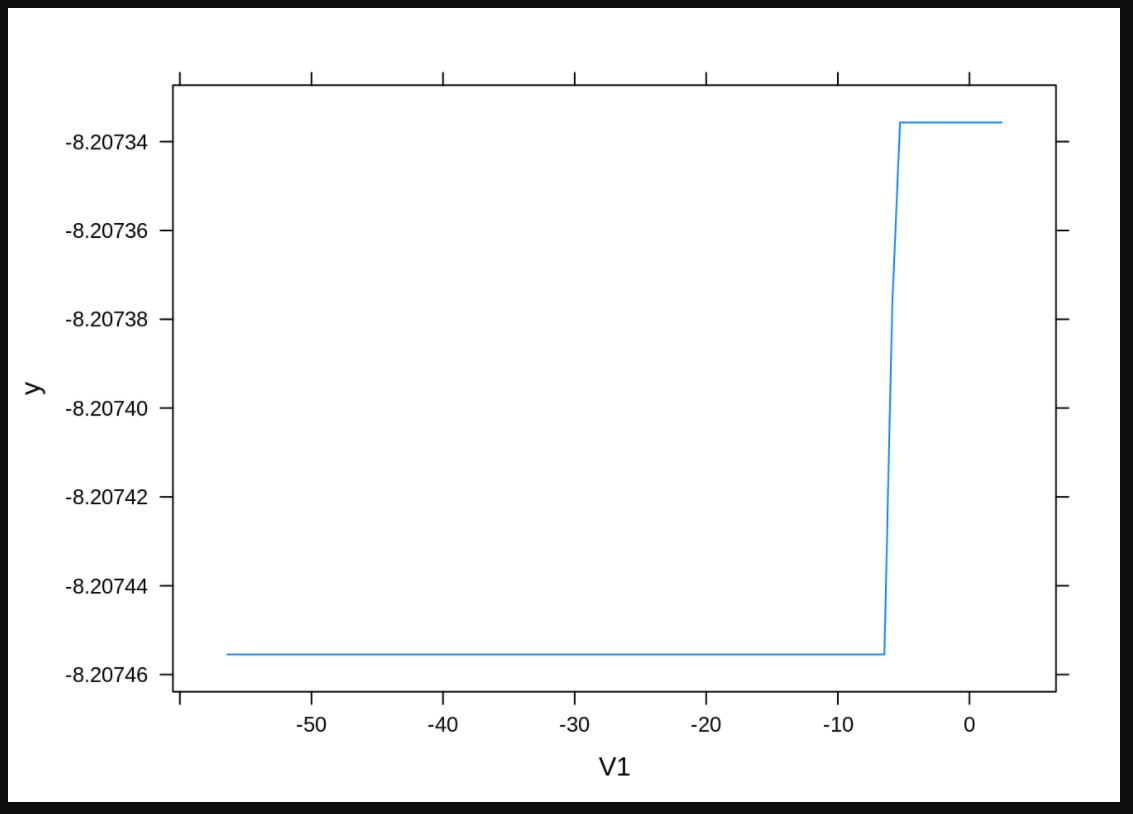
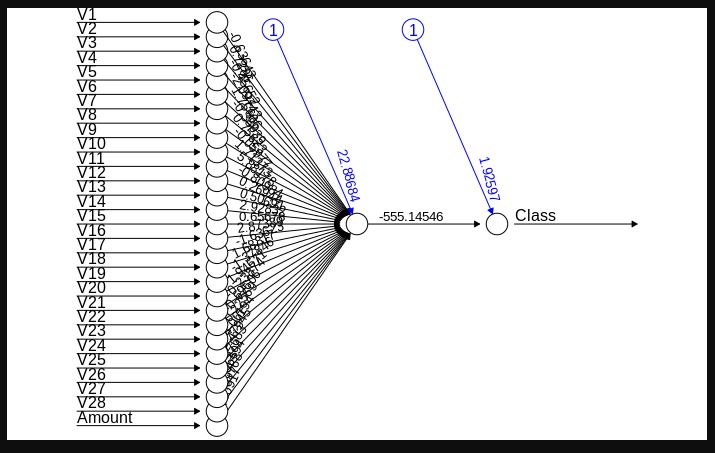
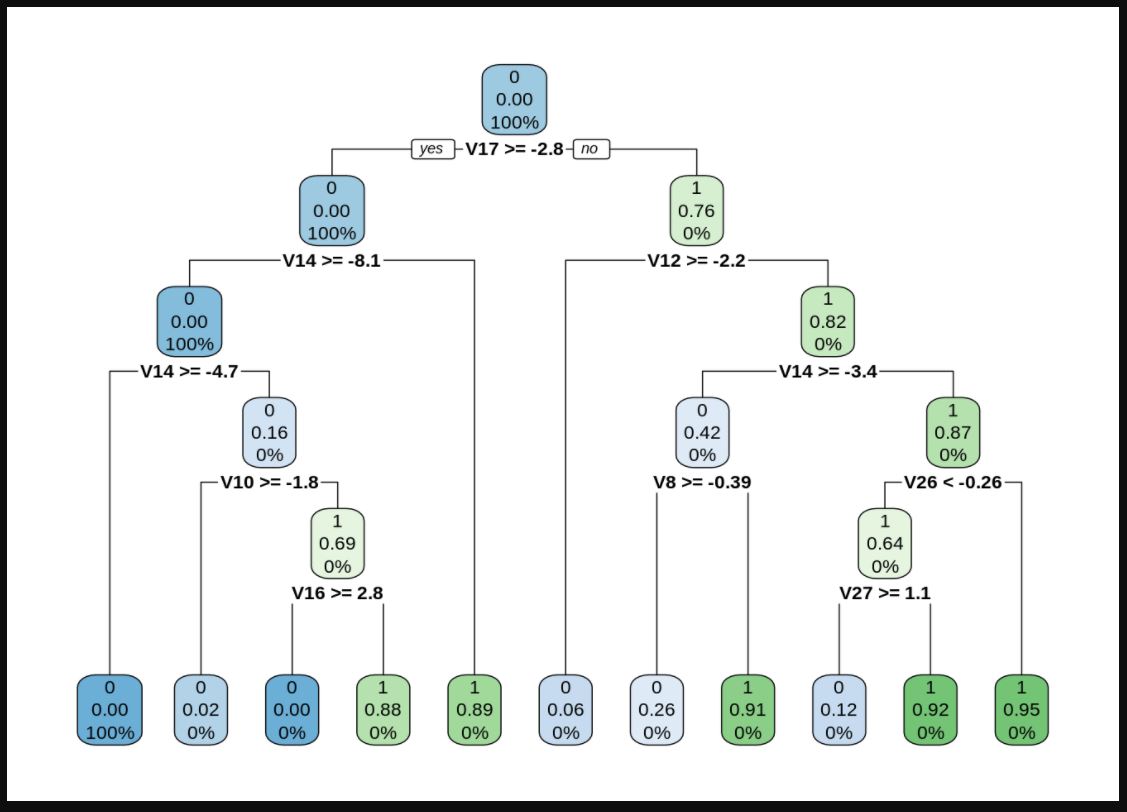
gbm\_auc = roc(test\_data$Class, gbm\_test, plot = TRUE, col = "red")

print(gbm\_auc)

**IMPLEMENTATION**



****



**CONCLUSION:**

Recommendation Systems are the most popular type of machine learning applications that are used in all sectors. They are an improvement over the traditional classification algorithms as they can take many classes of input and provide similarity ranking based algorithms to provide the user with accurate results. These recommendation systems have evolved over time and have incorporated many advanced machine learning techniques to provide the users with the content that they want.