FRAUD DETECTION IN CREDIT CARD TRANSACTIONS USING DEEP LEARNING



Under the guidance of

Professor CHAIYAPORN MUTSALKLISANA

TEAM DATA EXCAVATORS

Arvind Venkatasubramanian

Manakkattumadom Subramani Vaisakh Srinivas

Table of Contents

[Introduction 2](#_Toc449629120)

[Types of Credit Card Frauds 3](#_Toc449629121)

[SOLUTION PROPOSED 4](#_Toc449629122)

[Deep learning – the concept 5](#_Toc449629123)

[Deep learning in Credit Card Fraud Analysis 5](#_Toc449629124)

[DataSet 6](#_Toc449629125)

[Algorithms 7](#_Toc449629126)

[Random Forest Algorithm 7](#_Toc449629127)

[Logistic Regression 7](#_Toc449629128)

[Decision Trees 8](#_Toc449629129)

[Neural Networks 8](#_Toc449629130)

[Code Snippets and Analysis 9](#_Toc449629131)

[Random Forest Analysis 9](#_Toc449629132)

[Regression Tree Analysis: 12](#_Toc449629133)

[Neural Network Analysis: 19](#_Toc449629134)

[Python Implementation: 22](#_Toc449629135)

[CONCLUSION 23](#_Toc449629136)

## Introduction

* Banks and financial services have millions of data and transactions to deal with every day. One of the most important transaction that banks have to deal with is credit card.
* As defined in businessdictionary.com, credit cards are defined as a convenient substitute for cash with a magnetic strip that contains a machine readable code. This code is what authorizes for a payment of any commodity that is purchased. It is a very important and essential component for the ecommerce and internet usage.
* Credit card usages do not require authorization immediately. This allows for a scope of fraud transactions to happen. This may result in losses to be incurred by banks and financial services on nonpayment of the credit card amount due. The amount may seem small in terms of one person. But when considered as a whole lot, it can result in huge losses for the financial terms.
* Fraud means obtaining services/goods and/or money by unethical means

# Types of Credit Card Frauds

* Bankruptcy Fraud:

This is one of the most difficult fraud to predict. It means that the purchaser is using the credit card with the knowledge that they do not have sufficient funds to return the amount that needs to be paid. In the end, the bank would need to incur losses and settle the amount with the merchant. One of the ways to check this method is by pre checking their credit record, before sanctioning any credit usage on their credit card.

* Theft Fraud:

This type of fraud is linked to stealing of the credit card details. The thief would use the credit card till the period when the owner can block the card and avoid any other further transactions. This is also occurred when credit cards are used remotely. The credit card can be used from any remote location for online transactions. Online transactions suffer a lot through this fraud as they require only the credit card information and no authorization as such.

* Application Fraud:

This type of fraud is occurred when the person applying for a credit card with false information. This usually occurs in two stages – when a person is applying for a duplicate card from the same location and when other users are applying for new cards using same information.

# SOLUTION PROPOSED

To avoid these frauds we can use machine learning algorithms to predict the behavior pattern of the customers. The accuracy of the prediction depends on the dataset that is fed based on the location and the banking rules defined.

There are various steps involved in the final prediction based on the information.

We propose to use a sequence of algorithms to begin with the three main steps of any data science project. The 3 main steps involved are

* Clustering
* Classification
* Approximation

# Deep learning – the concept

Deep learning is an advanced version of Advanced Neural Networks (ANN). It enables us to solve complex problems by training the machine like the human brain. It refers to a neural network with *more than one hidden layer.*

*As per Wikipedia, Deep learning is defined as ‘high-level abstractions in data by using multiple processing layers, with complex structures or otherwise, composed of multiple non-*[*linear transformations*](https://en.wikipedia.org/wiki/Linear_transformation)*’.*

This concept has been used to solve various complex problems. Some of the real time applications based on this application are

* Face/Image recognition
* Voice Search (Google voice/ SIRI)
* Speech to text
* Spam Filtering

Deep learning is based on interpretation, similar to the sense organs of our body. It will help classify, cluster and predict by reading signals, or by understanding any input. Similar to way humans would too.

# Deep learning in Credit Card Fraud Analysis

We can predict and prevent credit card frauds by using deep learning algorithms. Distributed deep learning systems can help create accurate models for detecting fraud and other anomalies. Based on the training model, the machine can learn from the new data automatically and can be made to do so without human intervention.

Based on the different kinds of frauds that can happen, the machine can be made to interpret and perceive to estimate what was possibility to avoid this fraud.

# DataSet

To ensure that the data is of top quality there has to be a sequence of steps that need to be followed to ensure that the dataset gives us the best possible result. The following conditions were kept in consideration before data cleansing part.

* Data is accurate
* Data is consistent
* Data does not contain any garbage values.
* The data is complete.

After cleansing and processing of the data, a dataset of 90 observations was chosen as the data to be fed to design the algorithm

Sample\_data.csv



# Algorithms

# Random Forest Algorithm

A Random Forest developed by Breiman, consists of an arbitrary number of simple trees, which are used to determine the final outcome. Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

**Why this Algorithm?**

Randomforest will help predict the right attributes that are required to determine the status of fraud or not. Based on the result of this attribute, the data will be fed into a classification system to ensure that what are the right attributes to determine the correct result.

# Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable. It is a regression model where the dependent variable (DV) is categorical.

**Why this Algorithm?**

Logistic regression is one of the best classification regression analysis. In the current problem statement it is helping us to identify based on the attributes that have been classified as significant from random forest, what is the classification based on the ‘credit\_class’ of the user.

# Decision Trees

Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making.

**Why this Algorithm?**

This algorithm has been chosen to denote the values and classifications based on the input of the significant variables. These clustering methods help understand the outliers and the dataset in a much defined way.

# Neural Networks

In machine learning and cognitive science, artificial neural networks (ANNs) are a family of models inspired by biological neural networks (the central nervous systems of animals, in particular the brain) which are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural networks are generally presented as systems of interconnected "neurons" which exchange messages between each other. The connections have numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning. Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A trained neural network can be thought of as an "expert" in the category of information it has been given to analyse.

**Why this Algorithm?**

This algorithm is the benchmark for detecting fraud and anomalies. After a logistic regression has established the connection between the attributes on how significant they are, neural network will help predict the model on how accurate it is between the predicted and the actual model. In this project, neural network establishes a link to help understand based on the training set and the test set, if the designed model fits for all the data set completely.

# Code Snippets and Analysis

# Random Forest Analysis

#import the library

library(randomForest)

#Change the Working Directory to modify the working Directory

## Modify the home directory to your home directory

getwd();

setwd("C:/Users/Arvind/Desktop/NEU/Semester 2/Advanced Data Science/Datasets")

mydata <- read.csv("Sample\_Data.csv",header=TRUE,strip.white = TRUE)

#Setting the Data into a dataframe, so that the main data can stay intact

x <- data.frame(mydata)

#Bind the data

xBind <- cbind(x)

#Fit the model

## Fitting the model through randomForest based on Credit Score.

fit <- randomForest(class ~ ., data=xBind,ntree=500,proximite=TRUE)

#Check the summary of the fit model

summary(fit)

#Predicted model between the credit score and fitted model using Random Forest

predicted <- predict(fit,x)

#Summary of the predicted value

summary(predicted)

#The correct recommended values for the classification based on credit score and

#fitted value

fit$predicted

fit$proximity

fit$importance

#Plot the datato find the right attributes for the information

varImpPlot(fit)

**OUTPUT ANALYSIS**

#Random Forest Algorithm is such that it can identify the most significant elements.

#In our case study we can clearly see that factors such as Age, Annual\_Sal,

#Avg\_Salary and balance are significant in predicting the Fraud Analysis.

#Based on this result we can clearly say that

#Age

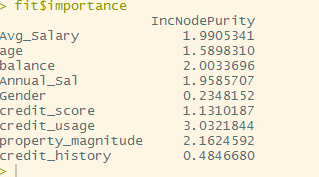
#Annual\_Sal

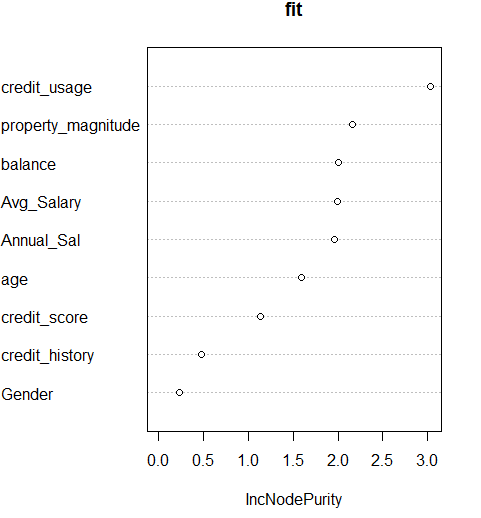
#Avg\_Salary

#balance

#credit\_usage

#are the important parameters in predicting credit card fraud based on the given dataset



Based on the above interpretation of the fitness model, we can clearly state that the attributes credit\_usage,balance,avg\_salary,annual\_sal and property magnitude will play a significant role in determining the class of the credit card user.   


# Regression Tree Analysis:

### Creating a Regression Tree

library(rpart)

setwd("C:/Users/Arvind/Desktop/NEU/Semester 2/Advanced Data Science/Datasets")

mydata <- read.csv("Sample\_Data.csv",header=TRUE,strip.white = TRUE)

#Trying to fit the regression Tree information

fit <- rpart(class ~ credit\_usage + property\_magnitude + Annual\_Sal + Avg\_Salary + balance + age + credit\_score,method = "anova",data = mydata)

printcp(fit)

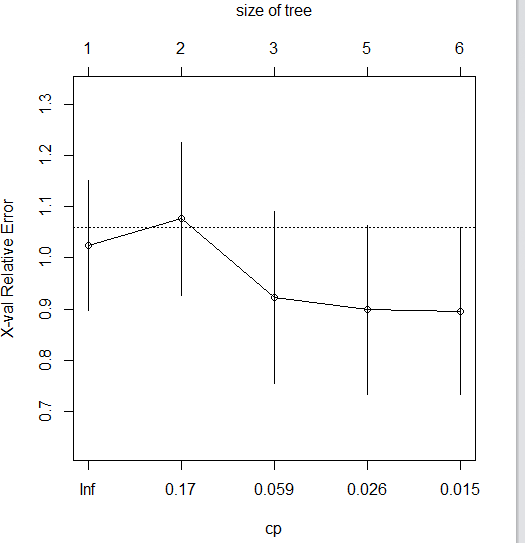
plotcp(fit)

summary(fit)

plot(fit, uniform=TRUE, main="Regression Tree for Credit Class ")

text(fit, use.n=TRUE, all=TRUE, cex=.8)

OUTPUT



Logistic Regression Analysis:

setwd("C:/Users/Arvind/Desktop/NEU/Semester 2/Advanced Data Science/Datasets")

mydata <- read.csv("Sample\_Data.csv",header=TRUE,strip.white = TRUE)

# Creating a dataset for the data

train <- mydata[1:65,]

#Creating a test dataset

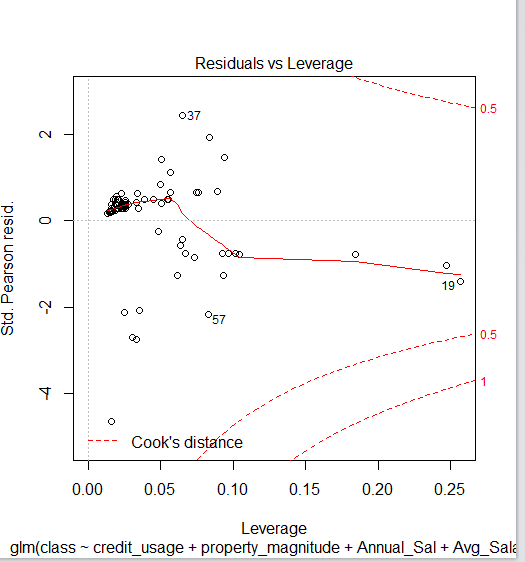
test <- mydata[66:89,]

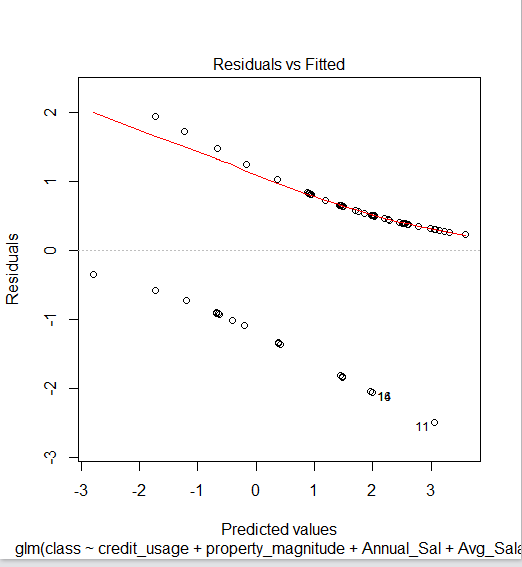
#Fitting a Model for the logistic regression

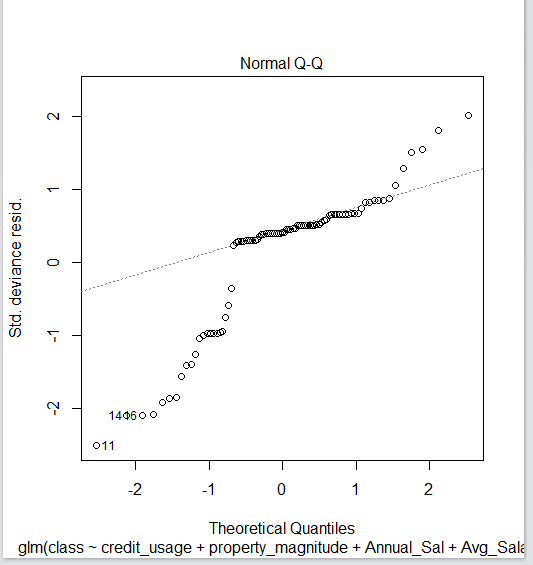
model <- glm(class ~ credit\_usage + property\_magnitude + Annual\_Sal + Avg\_Salary + balance + age + credit\_score,family=binomial(link='logit'),data=mydata)

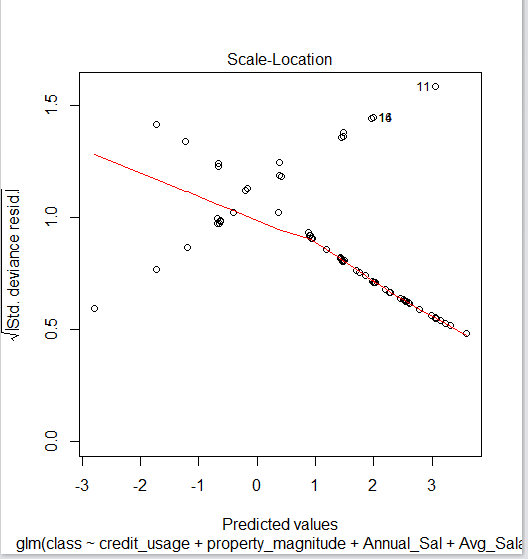
#Plotting the model

plot(model)









# Neural Network Analysis:

library("neuralnet")

setwd("C:/Users/Arvind/Desktop/NEU/Semester 2/Advanced Data Science/Datasets")

mydata <- read.csv("Sample\_Data.csv",header=TRUE,strip.white = TRUE)

# Creating a dataset for the data

train <- mydata[1:65,]

#Creating a test dataset

test <- mydata[66:89,]

#Train the Neural Network based on the most probably input parameters which have been

# detected from the randomForest algorithm

creditnet <- neuralnet(class ~ credit\_usage + property\_magnitude,train,hidden=4,lifesign = "minimal",linear.output = FALSE, threshold = 0.1)

# Plotting of the neural networkbased on the predicted values

plot(creditnet, rep = "best")

#temp\_test <- subset(test, select = c("credit\_usage", "property\_magnitude"))

creditnet.results <- compute(creditnet, temp\_test)

head(temp\_test)

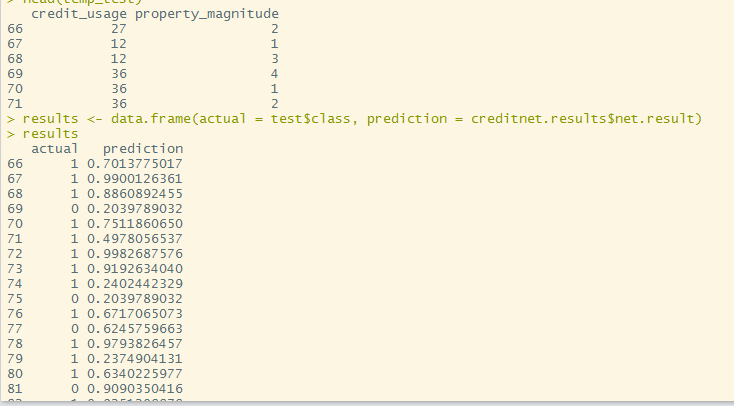
results <- data.frame(actual = test$class, prediction = creditnet.results$net.result)

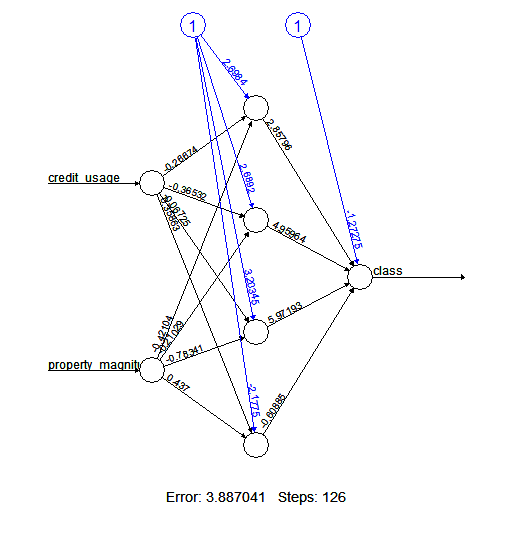
results

results$prediction <- round(results$prediction)

results

results





# Python Implementation:

For deep learning, we tried to implement the Artificial neural network using Library Pandas and Theano.

import pandas as pd

import theano

fields=['Avg\_Salary','age','balance','Annual\_Sal','Gender','credit\_score','credit\_usage','property\_magnitude ','credit\_history','class']

df = pd.read\_csv('Sample\_Data.csv',usecols=fields)

X = df.iloc[:, 1:].values

y = df.iloc[:, 0].values

df = 0

print(X.shape)

print(y.shape)

import matplotlib.pyplot as plt

%matplotlib inline

def plot\_digit(X, y, idx):

img = X[idx].shape

plt.imshow(img, cmap='Greys', interpolation='nearest')

plt.title('true label: %d' % y[idx])

plt.show()

plot\_digit(X, y, 10)

# CONCLUSION

Based on the analysis done on the dataset, the classification of significant parameters, it can be safe to say that the fraud transactions are based out of credit\_trasactions and property\_magnitude. Depending on the information of the customer, they can be more or less prone to committing fraud.

# References

http://www.businessdictionary.com/definition/credit-card.html

<https://www.tigurl.org/images/resources/tool/docs/3269.pdf>

<http://deeplearning4j.org/>

<https://www.quora.com/What-are-the-practical-applications-of-deep-learning-What-are-all-the-major-areas-fields>

<https://gigaom.com/2015/03/06/how-paypal-uses-deep-learning-and-detective-work-to-fight-fraud/>

<http://www.analyticsvidhya.com/blog/2015/09/random-forest-algorithm-multiple-challenges/>

<http://blog.christianperone.com/2015/08/convolutional-neural-networks-and-feature-extraction-with-python/>

<http://searchdl.org/public/conference/2014/ITC/65.pdf>

<http://deeplearning.net/tutorial/lenet.html#lenet>

http://www.analyticbridge.com/group/ForensicAndFraudAnalytics/forum/topics/fraud-detection-using-logistic