DATA SCIENCE : Summer Training & Project on Machine Learning at INFOSEEK MY5 TECHNOLOGIES PVT LTD

A Project Report

by

Aishwarya Mishra (170102263)

Under the Guidance of

Dr. Radha Guha
Professor
Dept. Computer Science

In partial fulfilment of the requirements for the Degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING,

DIT UNIVERSITY, DEHRADUN
(State Private University through State Legislature Act No. 10 of 2013 of Uttarakhand
and approved by UGC)
Mussoorie Diversion Road, Dehradun, Uttarakhand - 248009, India.

May 2019

DECLARATION

This is to certify that the Project entitled **"CREDIT CARD FRAUD DETECTION & BOARD GAME REVIEW PREDICTION"** in partial fulfillment of the requirement for the award of the **Degree of B.TECH** in Computer Science, submitted to **DIT University, Dehradun, Uttarakhand, India,** is an authentic record of bonafide work carried out by me Aishwarya Mishra, under the guidance of Dr Radha Guha.

The matter embodied in this Project/Thesis/Dissertation has not been submitted for the award of any other degree or diploma to any University/Institution. The results embodied in this project report have not been submitted to any other University or Institute for the award of any Degree or Diploma.

Students Name & Signature:

Prof. Vishal Bharti
Head of Department

Date:

Place: Dehradun

CERTIFICATE

This is to certify that the Project entitled **“CREDIT CARD FRAUD DETECTION & BOARD GAME REVIEW PREDICTION”** in partial fulfillment of the requirement for the award of the **Degree B.Tech** in Computer Science, submitted to **DIT University, Dehradun, Uttarakhand, India,** is an authentic record of INFOSEEK SYSTEM SOFTWARES carried out by Ms. Aishwarya Mishra, having roll no.170102263 respectively, under my supervision/ guidance.

Dr. Radha Guha
Professor
Dept. Computer Science

Date:

Place: Dehradun

ACKNOWLEDGEMENT

First and foremost, we express our gratitude to the Almighty and the Management of DIT University for providing all the facilities needed for this project. We express our gratitude to the Head of Department, Prof. Vishal Bharti, for his immense support and encouragement. We express my gratitude to our project coordinator, Dr Radha Guha, Department of Computer Science and Engineering for her valuable suggestions and guidance. I would like to extend my gratefulness to my external project director Mr. MIRZA GAZANFAR BAIG (SR. SOFTWARE ENGINEER at INFOSEEK MY5 TECHNOLOGIES. Last but not the least, I wish to thank all the staff of Computer Science and Engineering Department for their help and support.

ABSTRACT

I) CREDIT CARD FRAUD DETECTION

Financial fraud is an ever growing menace with far consequences in the financial industry. Data mining had played an imperative role in the detection of credit card fraud in online transactions. Credit card fraud detection, which is a data mining problem, becomes challenging due to two major reasons - first, the profiles of normal and fraudulent behaviours change constantly and secondly, credit card fraud data sets are highly skewed. The performance of fraud detection in credit card transactions is greatly affected by the sampling approach on dataset, selection of variables and detection technique(s) used. This paper investigates the performance of naïve bayes, k-nearest neighbor and logistic regression on highly skewed credit card fraud data. Dataset of credit card transactions is sourced from European cardholders containing 284,807 transactions. A hybrid technique of under-sampling and oversampling is carried out on the skewed data. The three techniques are applied on the raw and preprocessed data. The work is implemented in Python. The performance of the techniques is evaluated based on accuracy, sensitivity, specificity and precision. The results shows of optimal accuracy for naïve bayes, k-nearest neighbor and logistic regression classifiers are 97.92%, 97.69% and 54.86% respectively. The comparative results show that k-nearest neighbour performs better than naïve bayes and logistic regression techniques.

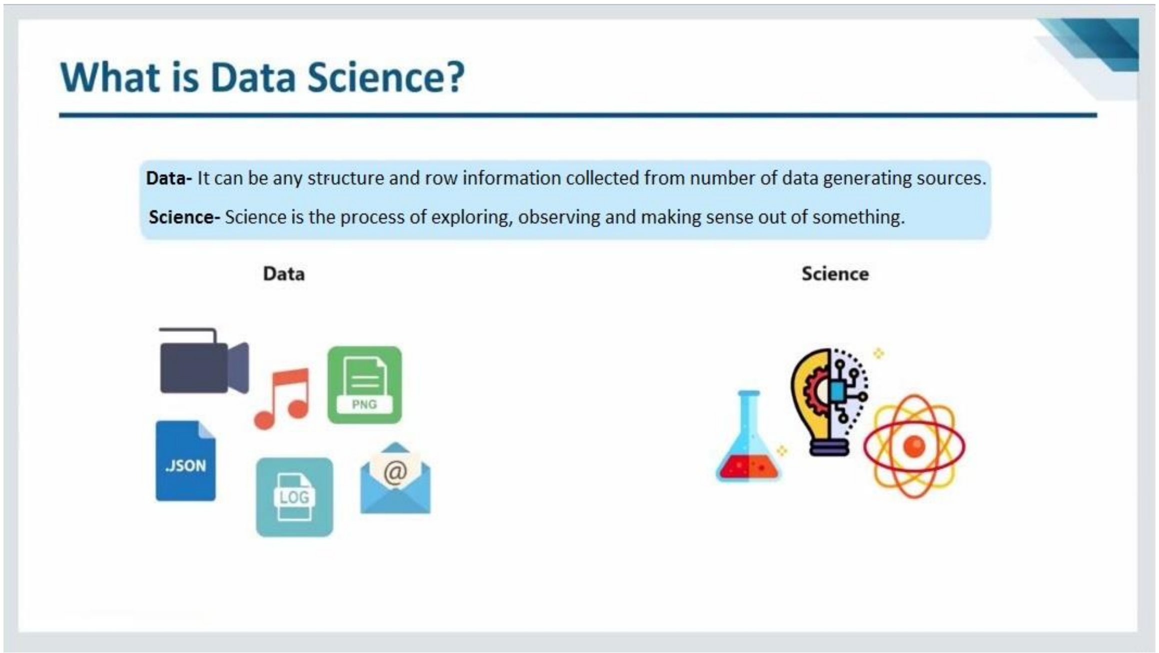
II) GAME BOARD REVIEW PREDICTION

With the growth of the World Wide Web, recommender systems have received an increasing amount of attention. Many recommender systems in use today are based on collaborative filtering.

Content-based recommender systems suggest documents, items, and services to users based on learning a profile of the user from rated examples containing information about the given items. Text categorization methods are very useful for this task but generally rely on unstructured text. We have developed a board game recommending system that utilizes semi-structured information about items gathered from the web using simple information extraction techniques. Initial experimental results demonstrate that this approach can produce fairly accurate recommendations.

TABLE OF CONTENTS

SUMMER TRAINING ON DATA SCIENCE



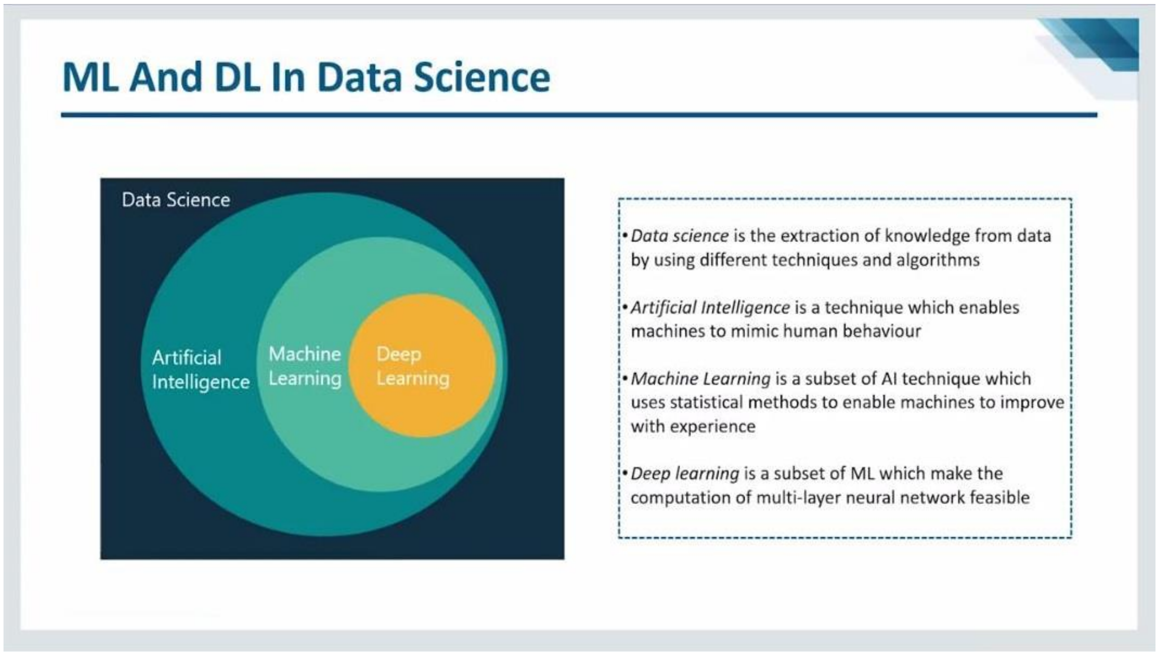
What is Data Science?

Data- It can be any structure and row information collected from number of data generating sources.

Science- Science is the process of exploring, observing and making sense out of something.

Data

Science



ML And DL In Data Science

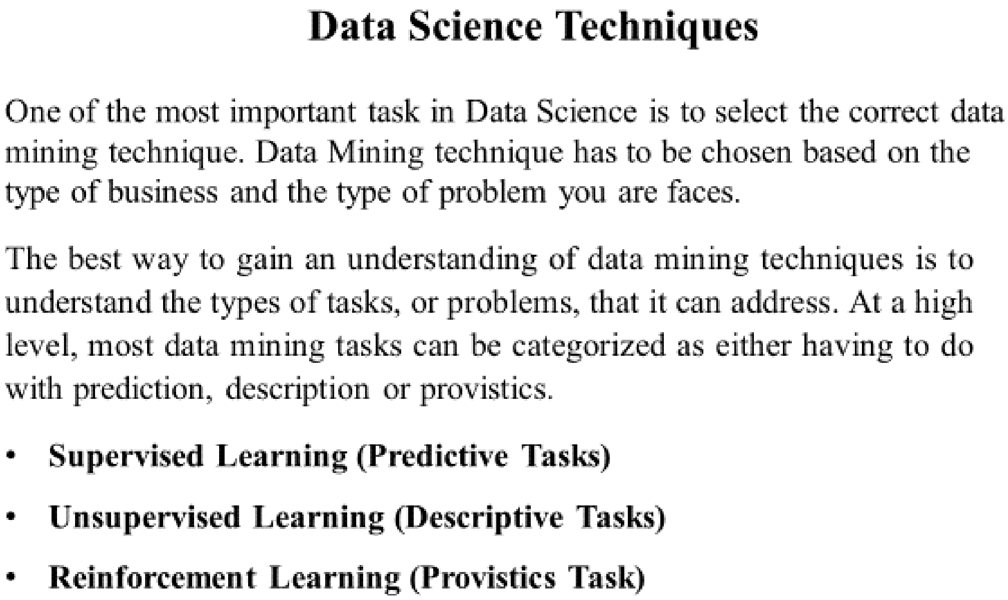
- *Data science* is the extraction of knowledge from data by using different techniques and algorithms
- *Artificial Intelligence* is a technique which enables machines to mimic human behaviour
- *Machine Learning* is a subset of AI technique which uses statistical methods to enable machines to improve with experience
- *Deep learning* is a subset of ML which make the computation of multi-layer neural network feasible

Data Science (Data Mining and Analysis) has been an active research area for a couple of decades, yet the complicated nature of data mining is still not fully understood .

Data Science is a multi-disciplinary field which combines Mathematics, Statistics, Analytics, Artificial Intelligence, Machine Learning, and Big Data Technology.

Data from different sources may seem irrelevant to each other. Once information generated from different sources is integrated, as new and useful knowledge may emerge.

Data mining is defined as a process of discovering hidden valuable knowledge by analyzing large amounts of data, which is stored in databases, data repositories and data warehouse, using various data mining techniques.

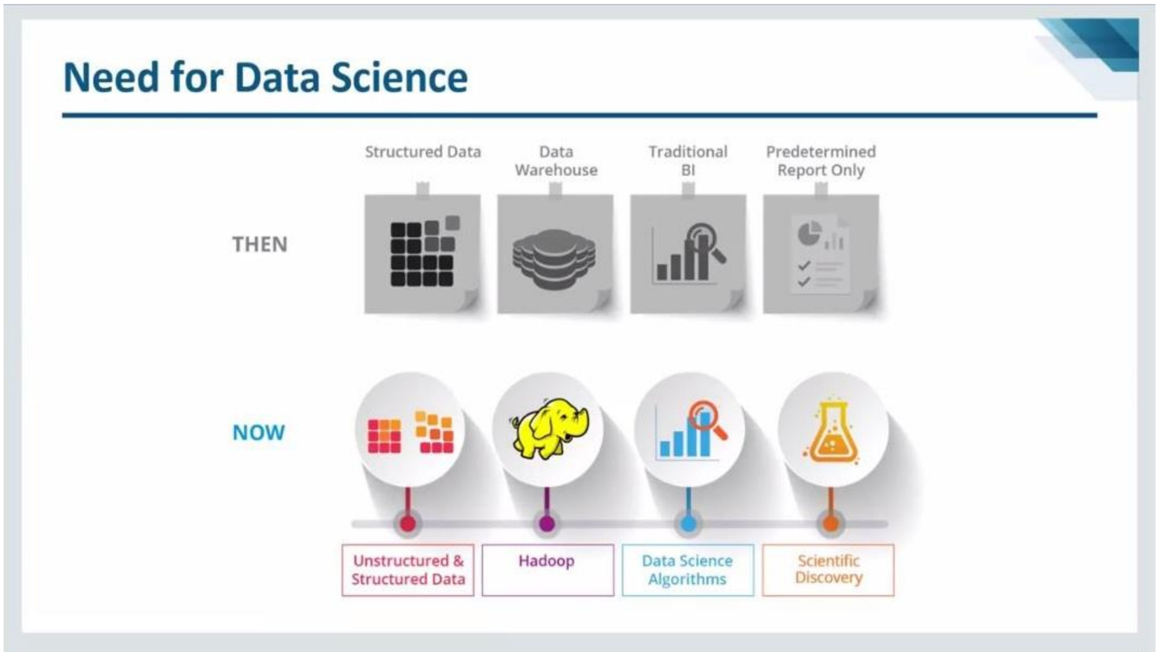


Data Science Techniques

One of the most important task in Data Science is to select the correct data mining technique. Data Mining technique has to be chosen based on the type of business and the type of problem you are faces.

The best way to gain an understanding of data mining techniques is to understand the types of tasks, or problems, that it can address. At a high level, most data mining tasks can be categorized as either having to do with prediction, description or provistics.

- **Supervised Learning (Predictive Tasks)**
- **Unsupervised Learning (Descriptive Tasks)**
- **Reinforcement Learning (Provistics Task)**



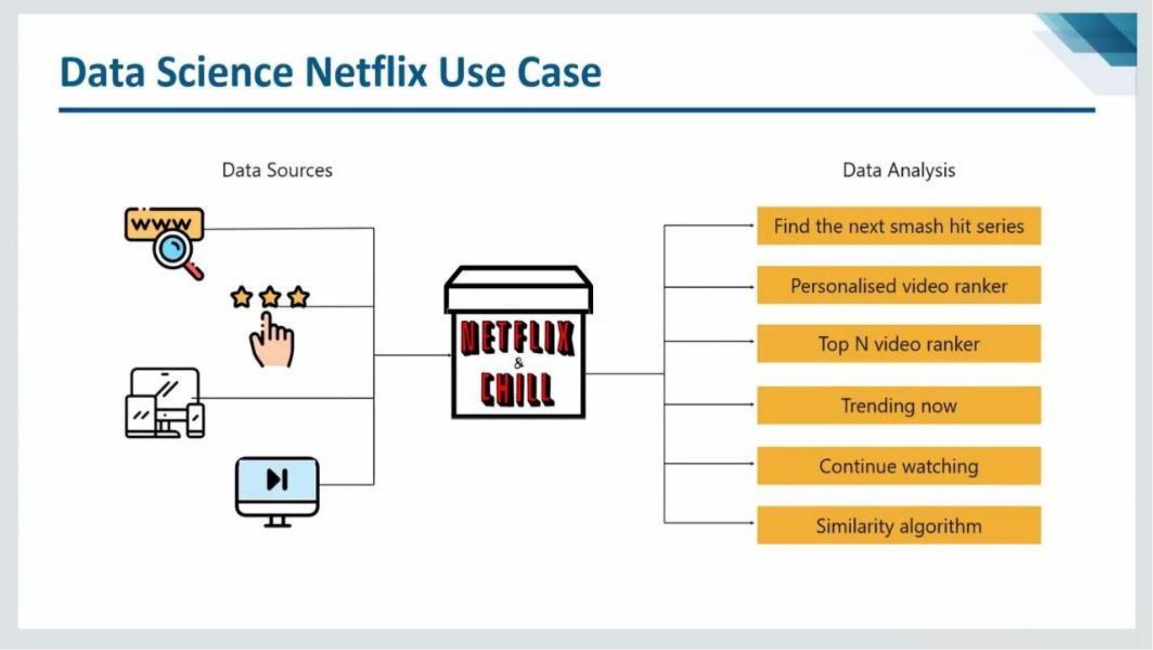
Need for Data Science

THEN

- Structured Data
- Data Warehouse
- Traditional BI
- Predetermined Report Only

NOW

- Unstructured & Structured Data
- Hadoop
- Data Science Algorithms
- Scientific Discovery



Data Science Netflix Use Case

Data Sources

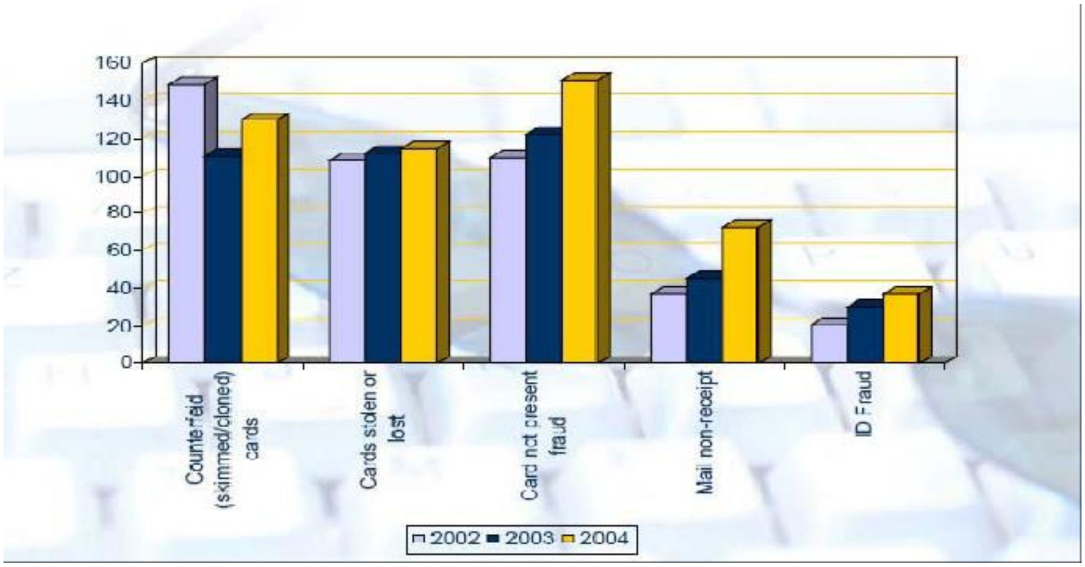
Data Analysis

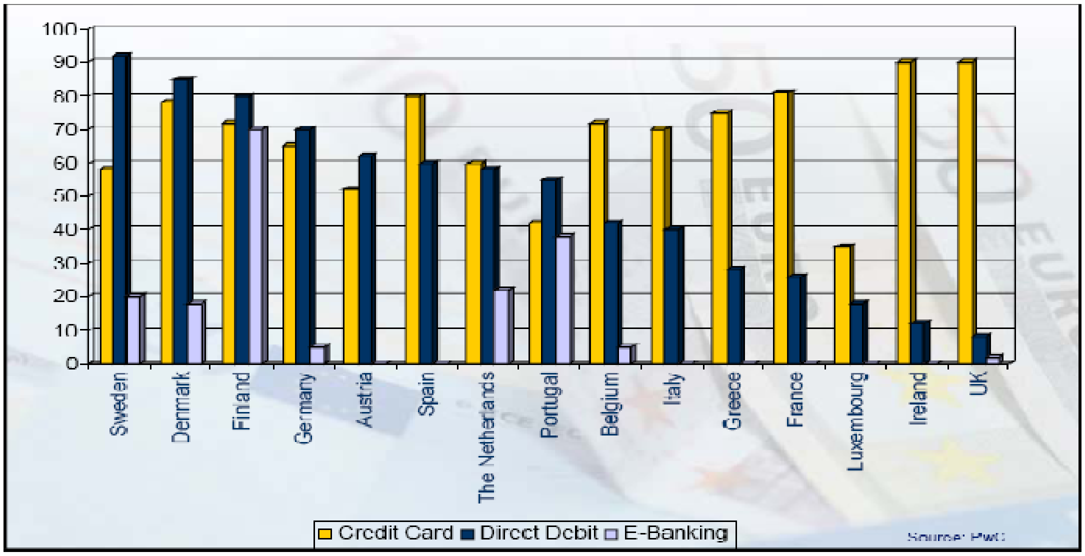
- Find the next smash hit series
- Personalised video ranker
- Top N video ranker
- Trending now
- Continue watching
- Similarity algorithm

CHAPTER 1

INTRODUCTION

Financial fraud is a growing concern with far reaching consequences in the government, corporate organizations, finance industry, In Today’s world high dependency on internet technology has enjoyed increased credit card transactions but credit card fraud had also accelerated as online and offline transaction. As credit card transactions become a widespread mode of payment, focus has been given to recent computational methodologies to handle the credit card fraud problem. There are many fraud detection solutions and software which prevent frauds in businesses such as credit card, retail, e-commerce, insurance, and industries. Data mining technique is one notable and popular methods used in solving credit fraud detection problem. It is impossible to be sheer certain about the true intention and rightfulness behind an application or transaction. In reality, to seek out possible evidences of fraud from the available data using mathematical algorithms is the best effective option. Fraud detection in credit card is the truly the process of identifying those transactions that are fraudulent into two classes of legit class and fraud class transactions, several techniques are designed and implemented to solve to credit card fraud detection such as genetic algorithm, artificial neural network frequent item set mining, machine learning algorithms, migrating birds optimization algorithm, comparative analysis of logistic regression, SVM, decision tree and random forest is carried out. Credit card fraud detection is a very popular but also a difficult problem to solve. Firstly, due to issue of having only a limited amount of data, credit card makes it challenging to match a pattern for dataset. Secondly, there can be many entries in dataset with truncations of fraudsters which also will fit a pattern of legitimate behavior. Also the problem has many constraints. Firstly, data sets are not easily accessible for public and the results of researches are often hidden and censored, making the results inaccessible and due to this it is challenging to benchmarking for the models built. Datasets in previous researches with real data in the literature is nowhere mentioned. Secondly, the improvement of methods is more difficult by the fact that the security concern imposes a limitation to exchange of ideas and methods in fraud detection, and especially in credit card fraud detection. Lastly, the data sets are continuously evolving and changing making the profiles of normal and fraudulent behaviors always different that is the legit transaction in the past may be a fraud in present or vice versa. This paper evaluates four advanced data mining approaches, Decision tree, support vector machines, Logistic regression and random forests and then a collative comparison is made to evaluate that which model performed best. Credit card transaction datasets are rarely available, highly imbalanced and skewed. Optimal feature (variables) selection for the models, suitable metric is most important part of data mining to evaluate performance of techniques on skewed credit card fraud data.





Source: PwC

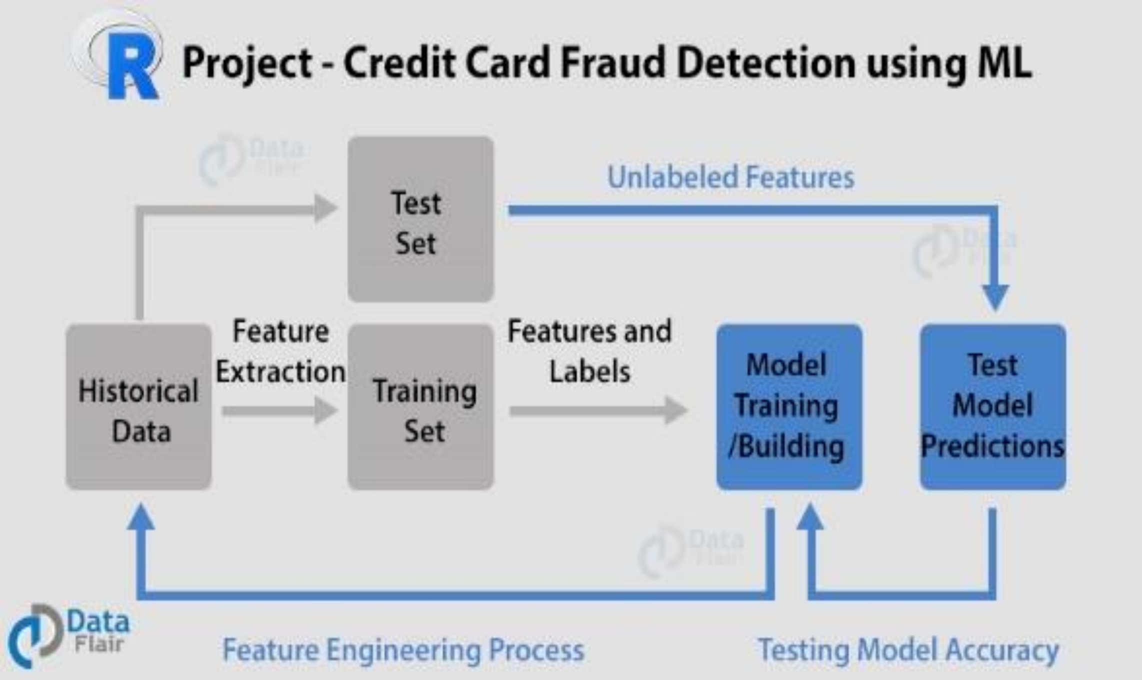
A number of challenges are associated with credit card detection, namely fraudulent behavior profile is dynamic, that is fraudulent transactions tend to look like legitimate ones, Credit card fraud detection International Journal of Pure and Applied Mathematics Special Issue 827 performance is greatly affected by type of sampling approach used, selection of variables and detection technique used. In the end of this paper, conclusions about results of classifier evaluative testing are made and collated. From the experiments the result that has been concluded is that Logistic regression has a accuracy of 97.7% while SVM shows accuracy of 97.5% and Decision tree shows accuracy of 95.5% but the best results are obtained by Random forest with a precise accuracy of 98.6%. The results obtained thus conclude that Random forest shows the most precise and high accuracy of 98.6% in problem of credit card fraud detection with dataset provided by ULB machine learning.

1) RELATED WORK

In [1] This report represents a project work about a case study involving credit card fraud detection, where data normalization is applied before Cluster Analysis and with results obtained from the use of Cluster Analysis and Artificial Neural Networks on fraud detection has shown that by clustering attributes neuronal inputs can be minimized. And promising results can be obtained by using normalized data and data should be MLP trained. This research was based on unsupervised learning. Significance of this paper was to find new methods for fraud detection and to increase the accuracy of results.

In [2] In this paper a new collative comparison measure that reasonably represents the gains and losses due to fraud detection is proposed. A cost sensitive method which is based on Bayes minimum risk is presented using the proposed cost measure. Improvements up to 23% is obtained when this method and other state of art algorithms are compared. The data set for this paper is based on real life transactional data by a large European company and personal details in data is kept confidential., accuracy of an algorithm is around 50%. Significance of this paper was to find an algorithm and to reduce the cost measure. The result obtained was by 23% and the algorithm they find was Bayes minimum risk.

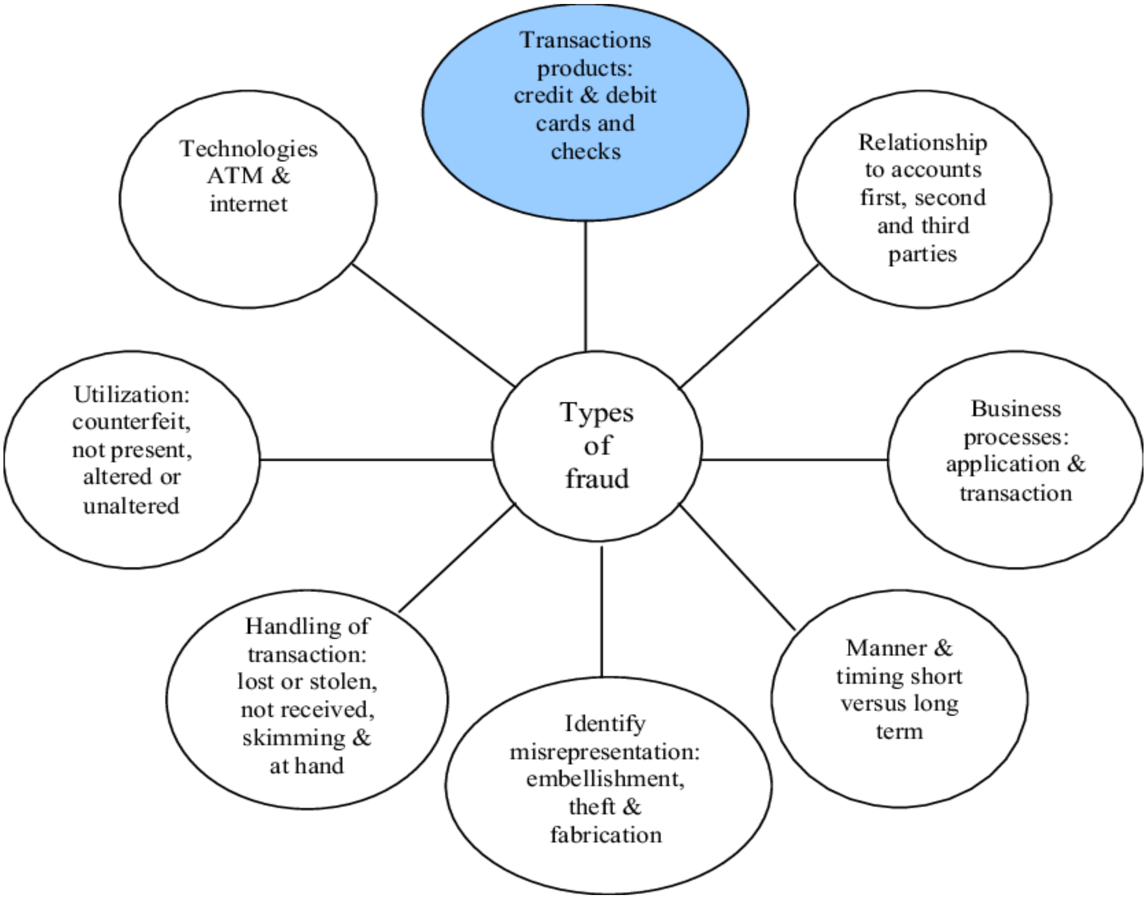
In [3] Various modern techniques based on Sequence Alignment, Machine learning, Artificial Intelligence, Genetic Programming, Data mining etc. has been evolved and is still evolving to detect fraudulent transactions in credit card. A sound and clear understanding on all these approaches is needed that will certainly lead to an efficient credit card fraud detection system. Survey of various techniques used in credit card fraud detection mechanisms has been shown in this paper along with evaluation of each methodology based on certain design criteria. Analysis on Credit Card Fraud Detection Methods has been done. The survey in this paper was purely based to detect the efficiency and transparency of each method. Significance of this paper was conduct a survey to compare different credit card fraud detection algorithm to find the most suitable algorithm to solve the problem.



Credit card is a plastic-card issued by a bank or nonbanking financial company (NBFC) ready to lend money (give credit) to its customer. It is a suitable alternative for cash payment. It is used to execute transactions which are compiled through electronic devices like a card swapping machine, computer with internet facility, etc. Basically, it is a synthetic-card made from a laminated plastic sheet and other materials like paints, magnetic stripe, microchip (IC), gelatin, hologram, etc. It entitles (authorizes) the customers to buy goods and services, based on credit sanctioned to them. It shall be used among a prescribed credit limit. This limit relies on the earning capability. It gives a customer a suitable choice to plan payments for goods and services that may be most necessary to him on a day-to-day basis. By using it, customer promises the repayment of credit transactions executed by him. Such a repayment along with interest shall be paid to bank or NBFC at a later agreed date. Generally, repayments along with an applicable interest are made either after a period of 30-45 days or are done on a monthly billing basis.

2) BACKGROUND

Fraud is one of the major ethical issues in the credit card industry. The main aims are, firstly, to identify the different types of credit card fraud, and, secondly, to review alternative techniques that have been used in fraud detection. The sub-aim is to present, compare and analyze recently published findings in credit card fraud detection. This article defines common terms in credit card fraud and highlights key statistics and figures in this field. Depending on the type of fraud faced by banks or credit card companies, various measures can be adopted and implemented. The proposals made in this paper are likely to have beneficial attributes in terms of cost savings and time efficiency. The significance of the application of the techniques reviewed here is in the minimization of credit card fraud. Yet there are still ethical issues when genuine credit card customers are misclassified as fraudulent.



Types of fraud
Transactions products: credit & debit cards and checks
Relationship to accounts first, second and third parties
Business processes: application & transaction
Manner & timing short versus long term
Identify misrepresentation: embellishment, theft & fabrication
Handling of transaction: lost or stolen, not received, skimming & at hand
Utilization: counterfeit, not present, altered or unaltered
Technologies ATM & internet

3) MOTIVATION

In recent years, topics such as fraud detection and fraud prevention have received a lot of attention on the research front, in particular from payment card issuers. The reason for this increase in research activity can be attributed to the huge annual financial losses incurred by card issuers due to fraudulent use of their card products. A successful strategy for dealing with fraud can quite literally mean millions of dollars in savings per year on operational costs. Artificial neural networks have come to the front as an at least partially successful method for fraud detection. The success of neural networks in this field is, however, limited by their underlying design - a feedforward neural network is simply a static mapping of input vectors to output vectors, and as such is incapable of adapting to changing shopping profiles of legitimate card holders. Thus, fraud detection systems in use today are plagued by misclassifications and their usefulness is hampered by high false positive rates. We address this problem by proposing the use of a dynamic machine learning method in an attempt to model the time series inherent in sequences of same card transactions. We believe that, instead of looking at individual transactions, it makes more sense to look at sequences of transactions as a whole; a technique that can model time in this context will be more robust to minor shifts in legitimate shopping behaviour. In order to form a clear basis for comparison, we did some investigative research on feature selection, preprocessing, and on the selection of performance measures; the latter will facilitate comparison of results obtained by applying machine learning methods to the biased data sets largely associated with fraud detection. We ran experiments on real world credit card transactional data using two innovative machine learning techniques: the support vector machine (SVM) and the long short-term memory recurrent neural network (LSTM).

4) DEFINITION AND OVERVIEW

Card fraud begins either with the theft of the physical card or with the compromise of data associated with the account, including the card account number or other information that would routinely and necessarily be available to a merchant during a legitimate transaction. The compromise can occur by many common routes and can usually be conducted without tipping off the cardholder, the merchant, or the issuer at least until the account is ultimately used for fraud. A simple example is that of a store clerk copying sales receipts for later use. An advanced example would be when a card scammer purchases from a certain store with the stolen numbers and hacks into the system, taking that money back into their possession. The rapid growth of credit card use on the Internet has made database security lapses particularly costly; in some cases, millions of accounts have been compromised.

Stolen cards can be reported quickly by cardholders, but a compromised account can be hoarded by a thief for weeks or months before any fraudulent use, making it difficult to identify the source of the compromise.

CHAPTER 2: OVERALL DESCRIPTION

2.1 PROJECT PRESPECTIVE

Existing System:

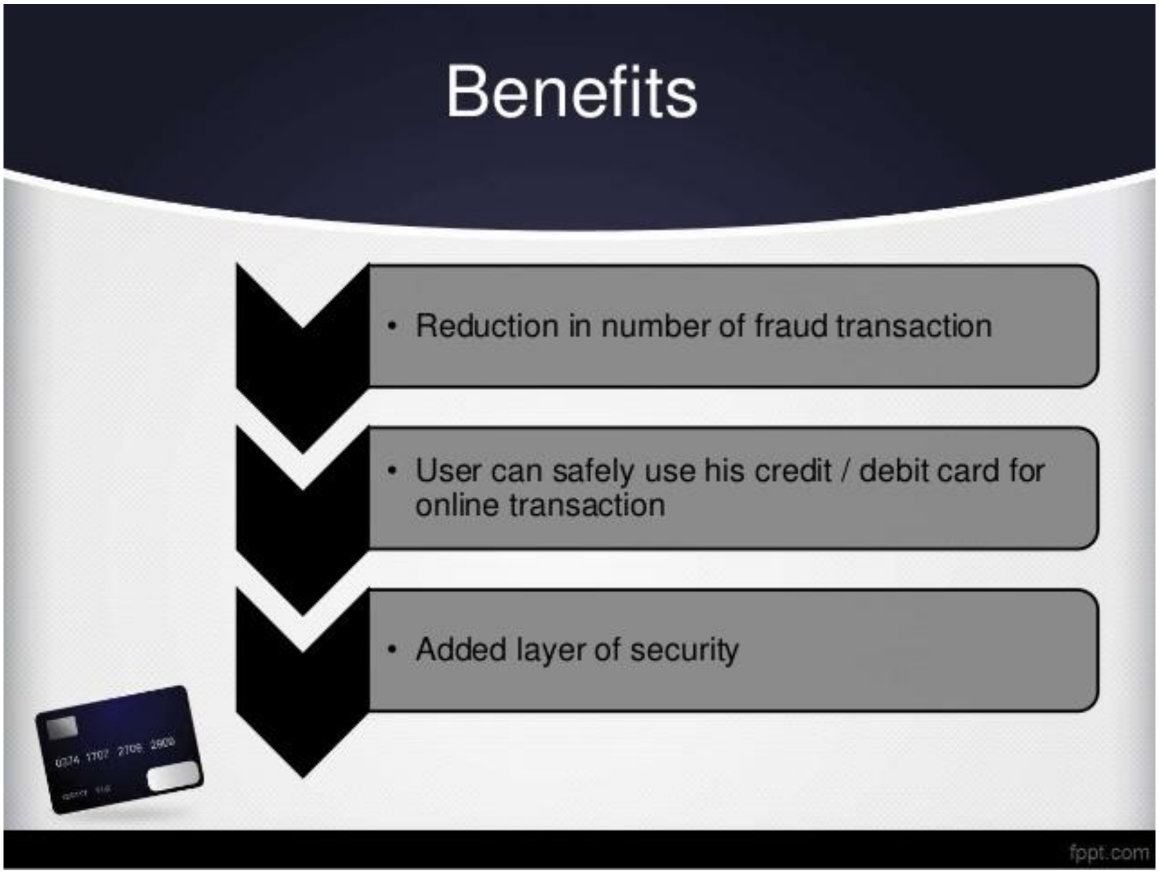
In case of the existing system the fraud is detected after the fraud is done that is, the fraud is detected after the complaint of the card holder. And so the card holder faced a lot of trouble before the investigation finish. And also as all the transaction is maintained in a log, we need to maintain a huge data. And also now a days lot of online purchase are made so we don't know the person how is using the card online, we just capture the IP address for verification purpose. So there need a help from the cyber crime to investigate the fraud. To avoid the entire above disadvantage we propose the system to detect the fraud in a best and easy way.

Proposed System:

In proposed system, we present a Hidden Markov Model (HMM).Which does not require fraud signatures and yet is able to detect frauds by considering a cardholder's spending habit. Card transaction processing sequence by the stochastic process of an HMM. The details of items purchased in Individual transactions are usually not known to any Fraud Detection System (FDS) running at the bank that issues credit cards to the cardholders. Hence, we feel that HMM is an ideal choice for addressing this problem. Another important advantage of the HMM-based approach is a drastic reduction in the number of False Positives transactions identified as malicious by an FDS although they are actually genuine. An FDS runs at a credit card issuing bank. Each incoming transaction is submitted to the FDS for verification. FDS receives the card details and the value of purchase to verify, whether the transaction is genuine or not. The types of goods that are bought in that transaction are not known to the FDS. It tries to find any anomaly in the transaction based on the spending profile of the cardholder, shipping address, and billing address, etc. If the FDS confirms the transaction to be of fraud, it raises an alarm, and the issuing bank declines the transaction.

Advantage

1. The detection of the fraud use of the card is found much faster that the existing system.
2. In case of the existing system even the original card holder is also checked for fraud detection. But in this system no need to check the original user as we maintain a log.
3. The log which is maintained will also be a proof for the bank for the transaction made.
4. We can find the most accurate detection using this technique.
5. This reduce the tedious work of an employee in the bank.



Benefits

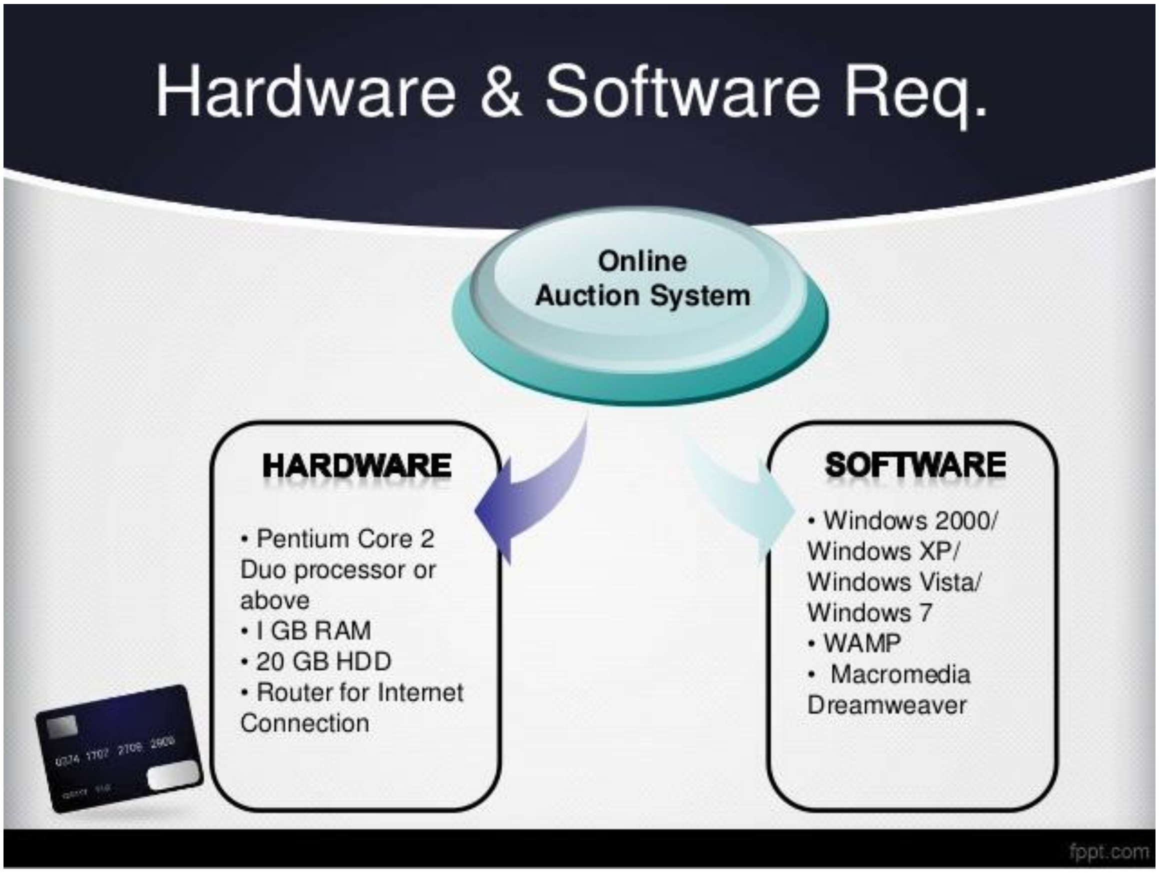
- Reduction in number of fraud transaction
- User can safely use his credit / debit card for online transaction
- Added layer of security

Hardware Requirements:

- System : Pentium Iv 2.4 Ghz
- Hard Disk : 40 Gb
- Floppy Drive : 1.44 Mb
- Monitor : 15 Vga Colour
- Mouse : Logitech.
- Ram : 256 Mb

Software Requirements:

- Operating system :- Windows XP Professional
- Front End : – Asp .Net 2.0.
- Coding Language :- Visual C# .Net
- Back-End : – Sql Server 2000.



Hardware & Software Req.

Online Auction System

HARDWARE

- Pentium Core 2 Duo processor or above
- I GB RAM
- 20 GB HDD
- Router for Internet Connection

SOFTWARE

- Windows 2000/ Windows XP/ Windows Vista/ Windows 7
- WAMP
- Macromedia Dreamweaver

fppt.com

Modules:

1. New card
2. Login
3. Security information
4. Transaction
5. Verification

Module Description

New card

In this module, the customer gives there information to enroll a new card. The information is all about there contact details. They can create there own login and password for there future use of the card.

Login

In Login Form module presents site visitors with a form with username and password fields. If the user enters a valid username/password combination they will be granted access to additional resources on website. Which additional resources they will have access to can be configured separately.

Security information

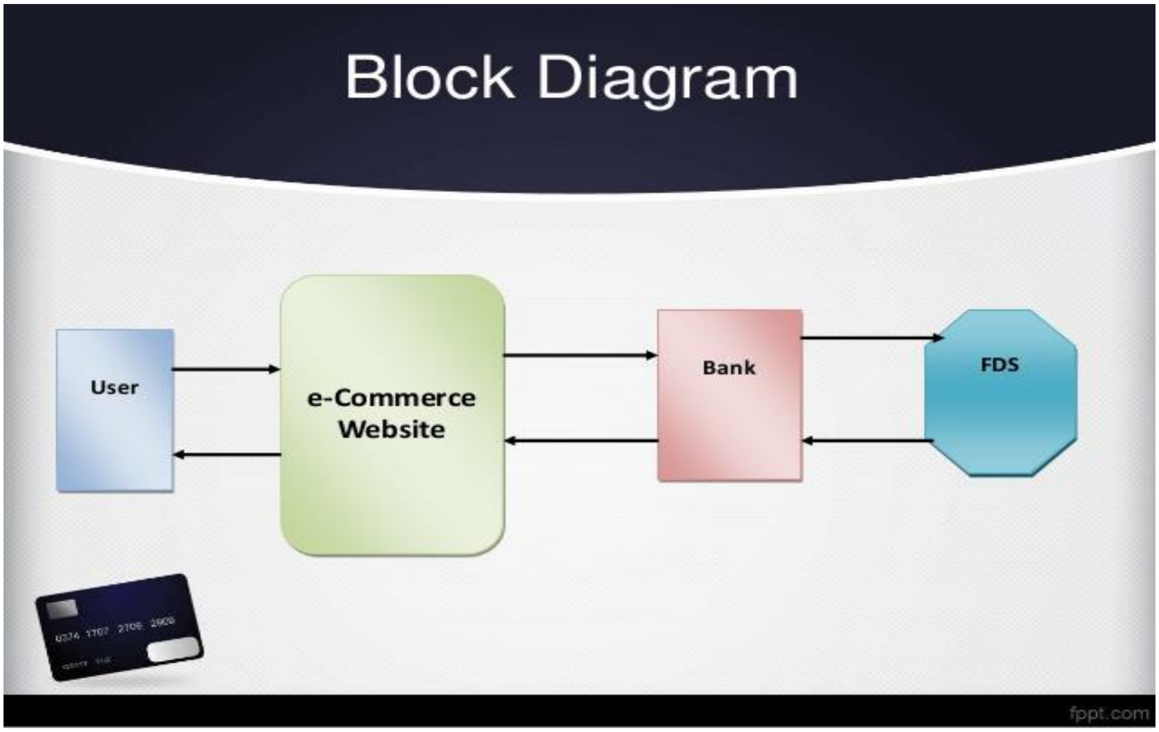
In Security information module it will get the information detail and its store's in database. If the card lost then the Security information module form arise. It has a set of question where the user has to answer the correctly to move to the transaction section. It contain informational privacy and informational self-determination are addressed squarely by the invention affording persons and entities a trusted means to user, secure, search, process, and exchange personal and/or confidential information.

Transaction

The method and apparatus for pre-authorizing transactions includes providing a communications device to a vendor and a credit card owner. The credit card owner initiates a credit card transaction by communicating to a credit card number, and storing therein, a distinguishing piece of information that characterizes a specific transaction to be made by an authorized user of the credit card at a later time. The information is accepted as "network data" in the data base only if a correct personal identification code (PIC) is used with the communication. The "network data" will serve to later authorize that specific transaction. The credit card owner or other authorized user can then only make that specific transaction with the credit card. Because the transaction is pre-authorized, the vendor does not need to see or transmit a PIC.

Verification

Verification information is provided with respect to a transaction between an initiating party and a verification-seeking party, the verification information being given by a third, verifying party, based on confidential information in the possession of the initiating party. In verification the process will seeks card number and if the card number is correct the relevant process will be executed. If the number is wrong, mail will be sent to the user saying the card no has been block and he can't do the further transaction.



Block Diagram

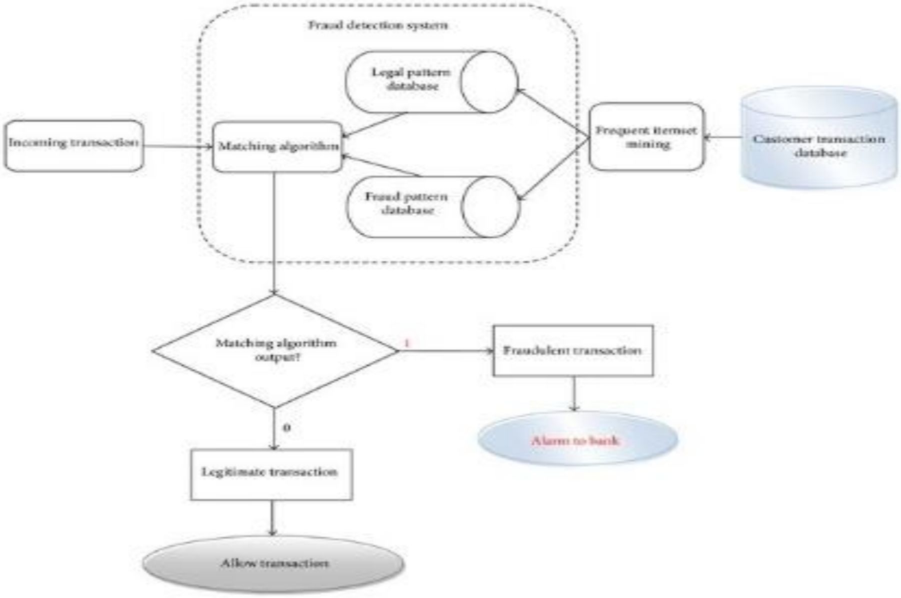
User

e-Commerce Website

Bank

FDS

fppt.com



PROPOSED CREDIT CARD FRAUD DETECTION MODEL

CHAPTER 3

CONCLUSION AND FUTURE WORK

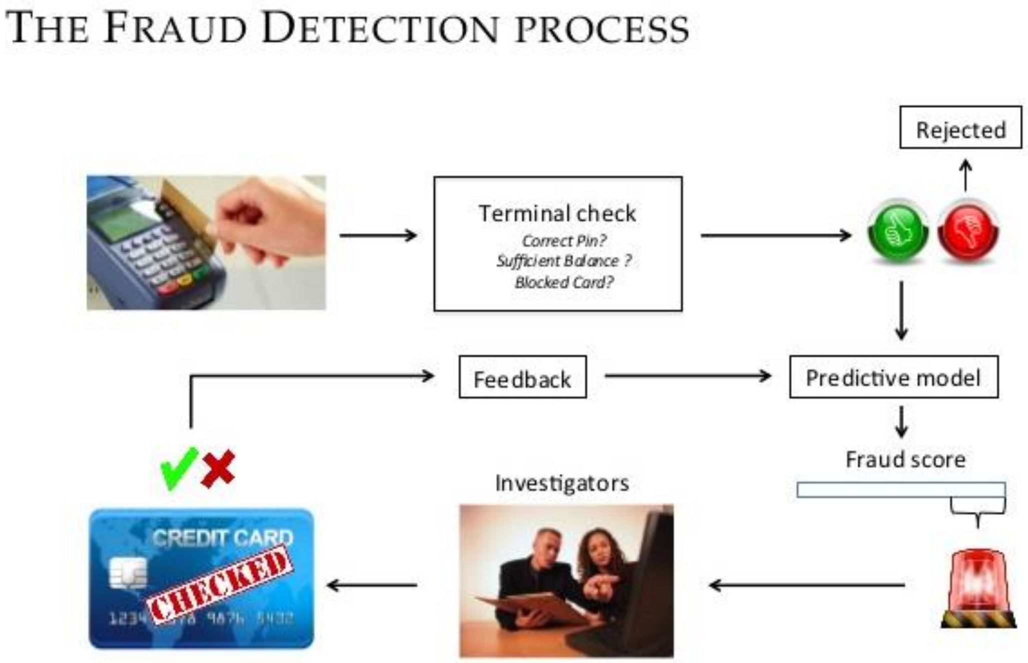
1. CONCLUSION

This report presents classification of credit card the challenges faced by cardholder as well as the card issuer, verity of fraud implemented by the persons who commit that fraud, some latest news regarding credit card fraudster and provide some prevention techniques that should be followed by the cardholder against the fraudulent activity. In recent times credit cards becomes the most popular means of payment and if credit card transactions increase, so too do frauds. The good news is that technology for preventing credit card frauds is also increasing in recent times and reducing cost of computing helps in introducing complex systems, which can analyze a fraudulent activity in a matter of fraction of a second.

From the experiments the result that has been concluded is that Logistic regression has a accuracy of 97.7% while SVM shows accuracy of 97.5% and Decision tree shows accuracy of 95.5% but the best results are obtained by Random forest with a precise accuracy of 98.6%. The results obtained thus conclude that Random forest shows the most precise and high accuracy of 98.6% in problem of credit card fraud detection with dataset provided by ULB machine learning. The Random forest algorithm will perform better with a larger number of training data, but speed during testing and application will suffer. Application of more pre-processing techniques would also help. The SVM algorithm still suffers from the imbalanced dataset problem and requires more preprocessing to give better results at the results shown by SVM is great but it could have been better if more preprocessing have been done on the data.

2. SCOPE FOR FUTURE WORK

Advances in technology give criminals increasingly powerful tools to commit fraud, especially using credit cards or internet bots. To combat the evolving face of fraud, researchers are developing increasingly sophisticated tools, with algorithms and data structures capable of handling large-scale complex data analysis and storage.



REFERENCES

APPENDIX

References

[1] Raj S.B.E., Portia A.A., Analysis on credit card fraud detection methods, Computer, Communication and Electrical Technology International Conference on (ICCCET) (2011), 152-156.

[2] Jain R., Gour B., Dubey S., A hybrid approach for credit card fraud detection using rough set and decision tree technique, International Journal of Computer Applications 139(10) (2016).

[3] Dermala N., Agrawal A.N., Credit card fraud detection using SVM and Reduction of false alarms, International Journal of Innovations in Engineering and Technology (IJIET) 7(2) (2016).

[4] Phua C., Lee V., Smith, Gayler K.R., A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119 (2010).

[5] Bahnsen A.C., Stojanovic A., Aouada D., Ottersten B., Cost sensitive credit card fraud detection using Bayes minimum risk. 12th International Conference on Machine Learning and Applications (ICMLA) (2013), 333-338

PROJECT 2

GAME BOARD REVIEW PREDICTION

Wouldn’t it be nice to know if a board game is good *before* you buy it? Let alone before you spend several hours with your friends plodding through the rule book, stumbling through the metaphorical dark. For games that already exist, BoardGameGeek (BGG) serves as the light. BGG is an only forum, marketplace, and wiki for all things board game related. With a large active community and a database of over 80,000 board games, it is the best place to find a new game to play. The goal of this project was to create a model that can accurately predict game ratings.

The Toolkit:

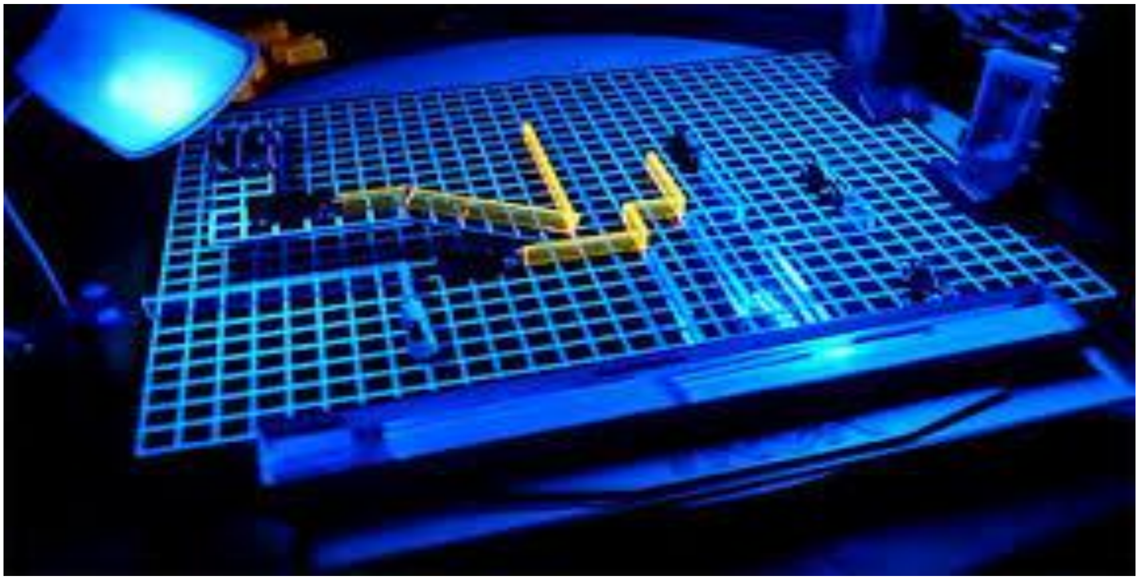
- scikit-learn
- AWS
- seaborn
- pandas
- Web scrapers written by Sean Beck

INTRODUCTION

This report will focus on predicting the reviews of over 80,000 different board games. This project based course is a fun way for you to master two important Machine Learning algorithms that can be applied on a much grander scale to other data sets - Linear Regression Model and a Random Forest Regressor.

However, for this project we will focus on board games. The information that we will work on was scrubbed from a database of 80,000 board games and includes information such as minimum players, maximum players, minimum playtime, maximum playtime, etc. We will use the models to ensure that using all of this information, it gives an accurate prediction regarding the reviews.

In the course of this project, we will walk you through all of the steps needed to generate this output including how to train the models, how to load and preprocess the dataset appropriately, and so much more. At the end of this, you will not only have an accurate prediction of the board game reviews, but also hands-on experience to learn how you can actually train two significant Machine Learning algorithms to learn and sort data, as well as make accurate predictions using a data set.



OVERVIEW

Machine learning is a field that uses algorithms to learn from data and make predictions. Practically, this means that we can feed data into an algorithm, and use it to make predictions about what might happen in the future. This has a vast range of applications, from self-driving cars to stock price prediction. Not only is machine learning interesting, it's also starting to be widely used, making it an extremely practical skill to learn. In this tutorial, we'll guide you through the basic principles of machine learning, and how to get started with machine learning with Python. Luckily for us, Python has an amazing ecosystem of libraries that make machine learning easy to get started with. We'll be using the excellent Scikit-learn, Pandas, and Matplotlib libraries in this tutorial.

BOARD GAME REVIEW PREDICTION

A)DATASET

Before we dive into machine learning, we’re going to explore a dataset and figure out what might be interesting to predict. The dataset is from BoardGameGeek , and contains data on 80000 board games. Here’s a single board game on the site. This information was kindly scraped into csv format by Sean Beck.The dataset contains several data points about each board game. Here’s a list of the interesting ones:
(i)name
– name of the board game.
(ii)playingtime
– the playing time (given by the manufacturer).
(iii)minplaytime
– the minimum playing time (given by the manufacturer).
(iv)maxplaytime
– the maximum playing time (given by the manufacturer).
(v)minage
– the minimum recommended age to play.
(vi)user_rated
– the number of users who rated the game.
(vii)average_rating
– the average rating given to the game by users. (0-10)
(viii)total_weight
– Number of weights given by users.
(x)average_weight
– the average of all the subjective weights (0-5).

- ## INTODUCTION TO PANDAS

The first step in our exploration is to read in the data and print some quick summary statistics. In order to do this, we'll us the Pandas library. Pandas provides data structures and data analysis tools that make manipulating data in Python much quicker and more effective. The most com-mon data structure is called a data-frame.

```
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split


import pandas as pd
games= pd.read_csv("games.csv")

print(games.columns)
print(games.shape)
```

The dataset can be given as:

id,type,name,yearpublished,minplayer,maxplayer,playingtime
12333,boardgame,Twilight struggle,2005,2,2,180
120677,boardgame,Terra Mystica,2012,2,5,150

This is in a format called csv, or comma-separated values, which you can read more about here. Each row of the data is a different board game, and different data points about each board game are separated by commas within the row. The first row is the header row, and describes what each data point is. The entire set of one data point, going down, is a column.

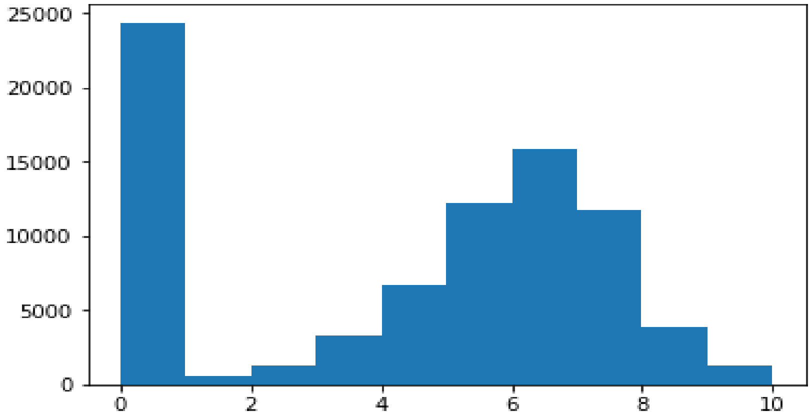
- READING OUR DATASET

```
Index(['id', 'type', 'name', 'yearpublished', 'minplayers', 'maxplay
ers',
       'playingtime', 'minplaytime', 'maxplaytime', 'minage', 'users
_rated',
       'average_rating', 'bayes_average_rating', 'total_owners',
       'total_traders', 'total_wanters', 'total_wishers', 'total_com
ments',
       'total_weights', 'average_weight'],
      dtype='object')
(81312, 20)
```

- PLOTTING THE TARGET VARIABLES

```
plt.hist(games["average_rating"])
plt.show()
```

It could be interesting to predict the average score that a human wouldgive to a new, unreleased, board game. This is stored in the average_rating column, which is the average of all the user ratings for a board game. Predicting this column could be useful to board game manufacturers who are thinking of what kind of game to make next, for instance



What we see here is that there are quite a few games with a 0 rating.There's a fairly normal distribution of ratings, with some right skew, and a mean rating around 6 (if you remove the zeros).

- PRINTING RESULTS

```
# Print the first row of the games with zero ratings
print(games[games["average_rating"]==0].iloc[0])

# Print the first row of games with scores greater than zero
print(games[games["average_rating"]>0].iloc[0])

print(games.columns)
```

```
Index(['id', 'type', 'name', 'yearpublished', 'minplayers', 'maxplay
ers',
       'playingtime', 'minplaytime', 'maxplaytime', 'minage', 'users
_rated',
       'average_rating', 'bayes_average_rating', 'total_owners',
       'total_traders', 'total_wanters', 'total_wishers', 'total_com
ments',
       'total_weights', 'average_weight'],
      dtype='object')
```

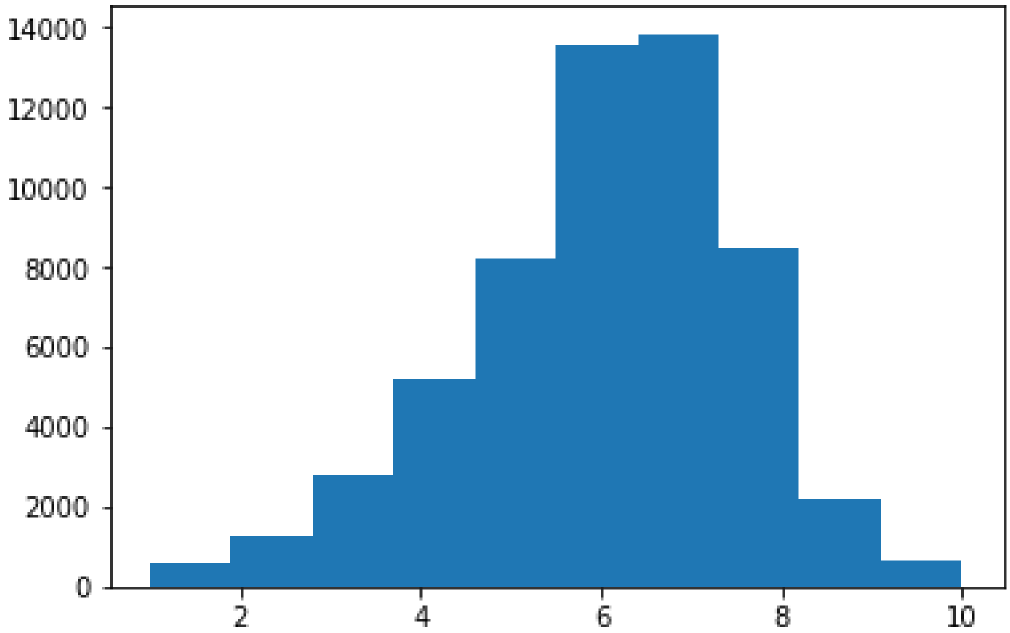
- EXPLORING THE RATINGS

```
id                             318
type                     boardgame
name                    Looney Leo
yearpublished                    0
minplayers                       0
maxplayers                       0
playingtime                      0
minplaytime                      0
maxplaytime                      0
minage                           0
users_rated                      0
average_rating                   0
bayes_average_rating             0
total_owners                     0
total_traders                    0
total_wanters                    0
total_wishers                    1
total_comments                   0
total_weights                    0
average_weight                   0
Name: 13048, dtype: object
id                                  12333
type                            boardgame
name                    Twilight Struggle
yearpublished                        2005
minplayers                              2
maxplayers                              2
playingtime                           180
minplaytime                           180
maxplaytime                           180
minage                                 13
users_rated                         20113
average_rating                    8.33774
bayes_average_rating              8.22186
total_owners                        26647
total_traders                         372
total_wanters                        1219
total_wishers                        5865
total_comments                       5347
total_weights                        2562
average_weight                     3.4785
Name: 0, dtype: object
```

- REMOVING GAMES WITH O REVIEWS

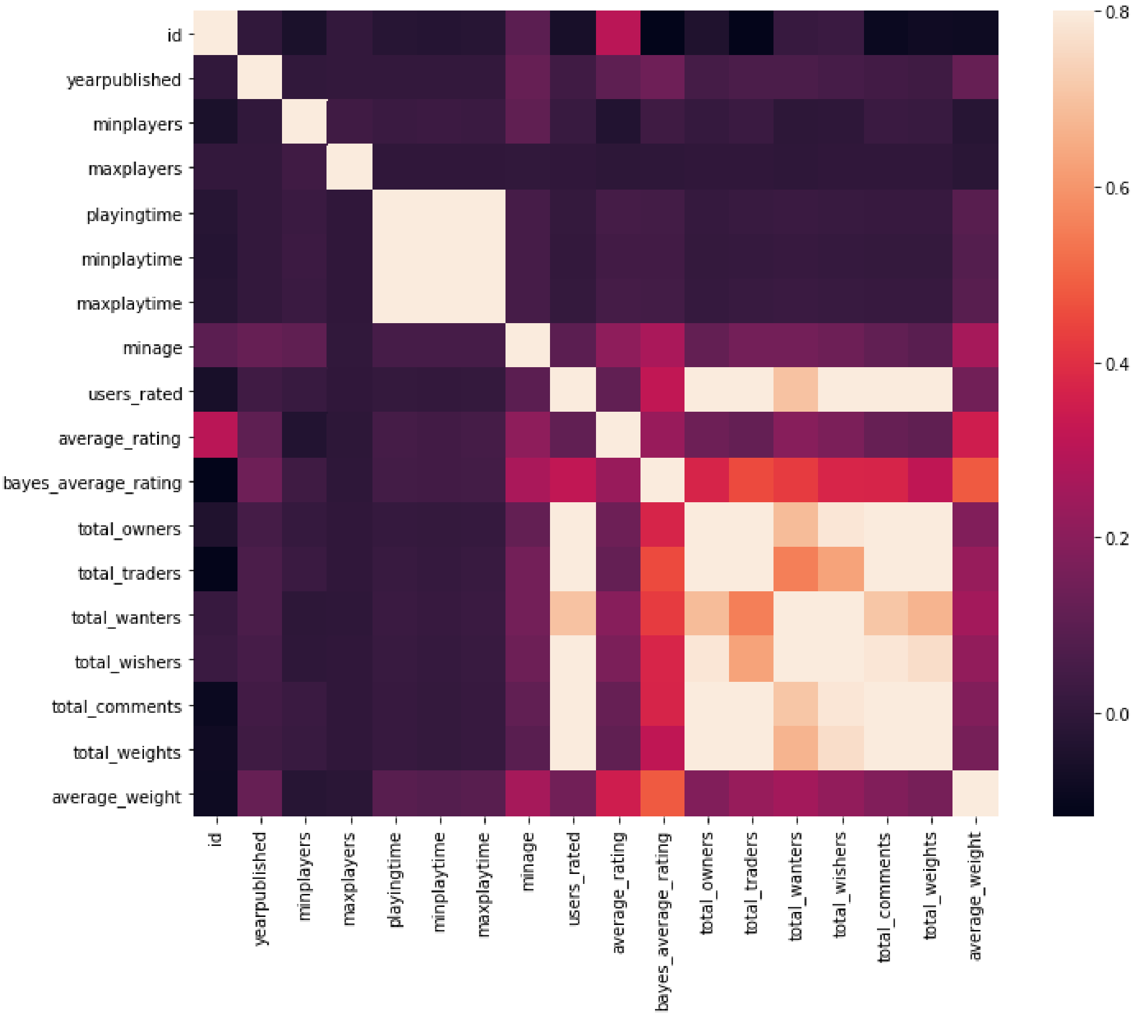
We just filtered out all of the rows without user reviews. While we were at it, we also took out any rows with missing values. Many machine learning algorithms can’t work with missing values, so we need some way to deal with them. Filtering them out is one common technique, but it means that we may potentially lose valuable data.

```
# Remove any rows without zero reviews
games= games[games['users_rated']>0]

# Remove any rows without user reviews
games=games.dropna(axis=0)

#Histogram of average rating of all games
plt.hist(games["average_rating"])
plt.show()
```



- CORRELATION MATRIX AND ACCURACY

```
# Corelation Matrix
corrmat = games.corr()
fig = plt.figure(figsize= (12,9))

sns.heatmap(corrmat, vmax= .8, square= True)
plt.show()
```



- MAKING PREDICTIONS

```
#Make predictions with both models 1st row
rating_lr= lr.predict(test[columns].iloc[1].values.reshape(1,-1))
rating_rfr= rfr.predict(test[columns].iloc[1].values.reshape(1,-1))

#Print the predictons
print(rating_lr)
print(rating_rfr)
```

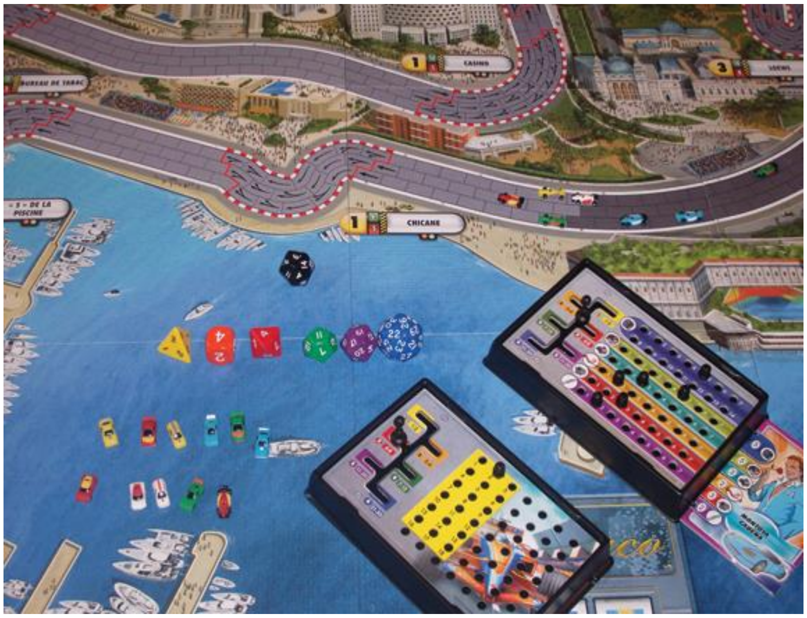
```
[9.20860328]
[7.85532168]
```

Machine Learning is slowly spreading its tentacles into all aspects of technology and even further. Better algorithms are helping devices become smarter and users to become more informed.

Board games have been a great way to pass the time, from simple ones like The Game of Life to more complicated ones like Dungeons & Dragons

CONCLUSION AND SCOPE OF FUTURE WORK

1. CONCLUSION: When the connection between games and analytics is used for educational purposes, the games teachers and authors select tend to be classics such as poker or chess. These are beloved games, and there is value in the fact that many students are already familiar with how they are played. Instead of spending time explaining the rules, the teacher can spend time explaining the actual analytic concepts he or she wants to illustrate. However, these classic games are relatively static; some of them, like chess, have been studied to the point that the full depth of the game is only evident to an expert. Others, like checkers, have been "solved," meaning that the optimal move is known from any position..



2. SCOPE FOR FUTURE: In my opinion, these are not attractive qualities for games used for teaching analytics and decision-making. I like the examples that I use in teaching to not necessarily have a right answer, for the moves to be debatable from multiple points of view, and for the intricacies of the decision-making to be readily apparent.

NOTES & REFERENCES

1. By “board game,” I really mean any game that is played around a table either with a group of friends or by a single player. The presence of a physical board isn’t a necessary component.
2. **http://viewer.zmags.com/publication/e0d5d66f#/e0d5d66f/84**
3. See the excellent introductory talk by Quintin Smith entitled “Board Gaming’s Golden Age” for more discussion (**http://vimeo.com/52293009**).
4. This is just one of many frameworks with which board games can be analyzed.
5. **www.boardgamegeek.com**
6. **www.susd.pretend-money.com**
7. **www.tabletop.geekandsundry.co**