Sponsored by:

DATA SCIENCE INSTITUTE

The World Data Science Institute is a Specialized Consulting Agency offering DSaaS (Data Science as a Service)

Meet Data Science Cohort Team 2

Anade Davis - Data Science Manager @ LinkedIn

Ragavendhra Ramanan - Data Science Researcher @ LinkedIn

Hafizah Ab Rahim - Data Scientist @ LinkedIn

Mukovhe Lugisani - Quantitative Analyst @LinkedIn

Raques McGill - Data Scientist @ LinkedIn

Brandon Oppong-Antwi-Data Engineer @LinkedIn

CREDIT CARD FRAUD DETECTION ANALYSIS

Table of Contents

Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection

- Many companies use Machine Learning Techniques to create Fraud Detection Systems
- The goal of a **FDS (Fraud Detection System)** is to identify irregular purchase(s)
- It is recommended that a **one-time password (OTP)** of purchases over a certain amount can increase security even more
- When evaluating/creating a system for Credit Card Fraud as a Data Scientist these are some of the variables to consider:
    - Traveling
    - Seasonal spending
    - Customer behavior (ex. Starts eating healthy, starts exercising, starts a new business)
    - Real time Machine Learning classification capabilities

Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection

- One of the biggest issues in dealing with Credit Card data is the imbalanced ratio of Fraud Transactions compared to Non-Fraud Transactions.

- The **Number 1 is commonly used as detecting Fraud.**

- The **Number 0 is commonly used as detecting Normality.**

    - It could even be considered to create a Number 2 to detect Outliers and a OTP is sent to user to properly identify whether it should be categorized as Fraud or Normality.

Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection

- We are looking for a model with good Accuracy, Precision, Specificity, Sensitivity to Outliers, and Recall.
- **Diagnostic Odd Ratio** - is a calculation used in Supervised Learning Classification to measure the effectiveness of a Test or model.
- **DOR** is used with heavily imbalanced Data Sets (**the higher the number the better**)
    - In this example it would be a measure of having the test say Fraud to those being Non-Fraud.

Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection

- Below we look at a Credit Card DataSet that was highly imbalanced and several machine learning techniques were compared to find the best Model to identify credit card fraud.

TABLE II
PERFORMANCE EVALUATION OF SUPERVISED MACHINE LEARNING ALGORITHMS

Techniques Used	Classifier Metrics Values					
	Positive Predictive Value (Precision)	*Negative Predictive Value*	*Prevalence*	*True Negative Rate (Specificity)*	*True Positive Rate (Sensitivity/ Recall)*	*Diagnostic Odd Ratio (DOR)*
NB	0.06	0.99	0.001	0.97	0.82	6.319
RF	0.99	0.16	0.998	0.91	0.99	18.857
K-NN	NaN	0.99	0.001	1.0	0.0	0
LR	0.99	0.63	0.99	0.87	0.99	168.56
XGBT	0.99	0.92	0.99	0.81	0.99	1138.5
SVM	NaN	NaN	NaN	0.92	0.93	0
ANN	0.99	0.84	0.99	0.77	0.99	462
DL	0.98	0.93	0.86	0.98	0.91	651
QDA	0.97	0.89	0.97	0.42	0.88	261.60
NN	0.99	NaN	0.99	0.0	1.0	0

TABLE IV
PERFORMANCE EVALUATION OF UNSUPERVISED MACHINE LEARNING ALGORITHMS

Techniques Used	Classifier Metrics Values					
	Positive Predictive Value (Precision)	*Negative Predictive Value*	*Prevalence*	*True Negative Rate (Specificity)*	*True Positive Rate (Sensitivity/ Recall)*	*Diagnostic Odd Ratio (DOR)*
SOM Hybrid	0.92	0.84	0.99	0.83	0.92	60.375
Isolation Forest	0.99	0.99	0.99	1.0	1.0	9801
Local Outlier Factor	0.99	0.99	0.998	1.0	1.0	9801
K-Means	0.99	NaN	0.998	1.0	0.0	0

Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection

- The results of the experiment show us Below that **Local Outlier Factor and Isolation Forest gives us our best results when detecting Credit Card Fraud.**

- Based on our research it has been found that Unsupervised Learning techniques handle the identification of Fraud better than traditional Supervised Learning techniques and Hybrid Machine Learning techniques.

TABLE IV
PERFORMANCE EVALUATION OF UNSUPERVISED MACHINE LEARNING ALGORITHMS

Techniques Used	Classifier Metrics Values					
	Positive Predictive Value (Precision)	*Negative Predictive Value*	*Prevalence*	*True Negative Rate (Specificity)*	*True Positive Rate (Sensitivity/ Recall)*	*Diagnostic Odd Ratio (DOR)*
SOM Hybrid	0.92	0.84	0.99	0.83	0.92	60.375
Isolation Forest	0.99	0.99	0.99	1.0	1.0	9801
Local Outlier Factor	0.99	0.99	0.998	1.0	1.0	9801
K-Means	0.99	NaN	0.998	1.0	0.0	0

A Comparative Analysis of Various Credit Card Fraud Detection Techniques

- Credit Card related frauds cause a loss of billions of dollars globally.
- With continued advancement in fraudulent strategies, it is important to develop effective models to combat these frauds.
- The properties of any good Fraud Detection System should
  - be able to identify the frauds accurately that means the number of wrong classifications should be minimum.
  - be able to detect the fraud while it is in transit--before they can take to completion.
  - not term any genuine transaction as fraudulent.

A Comparative Analysis of Various Credit Card Fraud Detection Techniques

- The metrics used for this evaluation are Accuracy, Precision, False Alarm Rate, Sensitivity, Specificity, and Cost where Cost = 100 * False negative rate + 10 * (False positive rate +True Positive rate)

Techniques	Accuracy	Detection Rate (Precision)	False Alarm Rate
Support Vector Machine (SVM)	94.65%	85.45%	5.2%
Artificial Neural Networks (ANN)	99.71%	99.68%	0.12%
Bayesian Network	97.52%	97.04%	2.50%
K- Nearest Neighbour (KNN)	97.15%	96.84%	2.88%
Fuzzy Logic Based System	95.2%	86.84%	1.15%
Decision Trees	97.93%	98.52%	2.19%
Logistic Regression	94.7%	77.8%	2.9%

A Comparative Analysis of Various Credit Card Fraud Detection Techniques

Training Expense Categories	Models
Expensive to train	Artificial Neural Networks and Naive Bayesian Networks
Somewhat expensive to train	KNN, SVM, Fuzzy Logic Based Systems, and Decision Trees
Not at all expensive to train	Logistic Regression

A Comparative Analysis of Various Credit Card Fraud Detection Techniques

- **Artificial Neural Networks perform best** in this comparative evaluation.
- However, they are very expensive to train and can be easily overtrained.
- **To minimize their expense, one should create a hybrid of neural network with an optimization technique.**
    - Examples of optimization techniques include Genetic Algorithm, Artificial Immune System, and Case Based Reasoning.

A Comparative Analysis of Various Credit Card Fraud Detection Techniques

- The major gaps in the current models and methods for fraud detection techniques are:
  - Unavailability of complete data for credit cards as they are private property and neither banks nor customers wish to disclose their information, thus leading to improperly and undertrained systems.
  - Unavailability of a single powerful algorithm that can perform consistently in all environments and can outperform all other algorithms.
  - A lack of good, efficient evaluation parameters that can not only describe the accuracy of the system but also can give a better comparative result among different approaches.
  - Inability of a system to adapt itself effectively to a changing environment, new fraudulent techniques, and genuine changes in the purchase habits of a user.

Predictive Modelling For Credit Card Fraud Detection Using Data Analytics

- Fraud detection is very important to save the financial losses for the banks as they issue credit card to the customer.
- Without the knowledge of credit card holder use of credit card information is a credit card fraud.There are of two types of fraud detection approaches: misuse detection and anomaly detection.
- In **misuse detection**, the system trains on normal and fake transactions, it will identify the known frauds.
- In **anomaly detection**, normal transactions are used for training so it has potential to identify the novel frauds.

Predictive Modelling For Credit Card Fraud Detection Using Data Analytics

Designing a framework for pre-preprocessing:

- These are responsible for processing big data and giving it to analytical server for predictive modelling.
- Hadoop network stores data in **Hadoop Distributed File System (HDFS)** and data from Hadoop is read by **SAS** using data step and proc hadoop step and converted into **raw data file**.
- The fields in raw data are separated by delimiter and the raw file is given to analytical model for building data model.

Dataset:

- German credit card fraud dataset is taken .
- It consists of 20 attributes out of which 7 are numerical attributes and 13 are categorical attributes and almost 1000 transactions.

Predictive Modelling For Credit Card Fraud Detection Using Data Analytics

Table1. Exploratory Analysis of Credit Card Dataset

Attribute: Status Of Checking Account	Fraud	Genuine	Total
'<0'	135	139	274
'0<=X<200	105	164	269
'>=200'	14	49	63
'No checking'	46	348	394
Grand Total	300	700	1000
Attribute: Credit History			
'all paid'	28	21	49
'critical/other existing credit'	50	243	293
'delayed previously'	28	60	88
'existing paid'	169	361	530
'no credit/all paid'	25	15	40
Grand Total	300	700	1000
Attribute: Property			
'Car'	102	230	332
'Life Insurance'	71	161	232
'No known Property'	67	87	154
'Real estate'	60	222	282
Grand Total	300	700	1000

Logistic Regression:

- Logistic analytical model is used for fraud detection.
- The fraud is response variable and the rest all are predictors.
- After training, it is tested with the test data with the threshold cut off value of 0.5.
- The model is tuned by choosing **most significant variables**.
- ~Status.of.existing.checking.account, Duration.in.months, Savings.account.bonds, Present.employment.since, Other.installment.plans.
- The **optimal cut off is 0.18** is used by above model with most significant variable and it gives best performance compared to 0.5.

Decision Tree:

- **ID3** algorithm is used.
- It is tuned by using most significant variables.

  Status.of.existing.checking.account

  Duration.in.months

  Savings.account.bonds

  Purpose.

Predictive Modelling For Credit Card Fraud Detection Using Data Analytics

Decision Tree:

Table 3. Confusion Matrix for Decesion Tree

	Actual Condition		Accuracy	
	Condition +ve	Condition – ve	72%	
Predicted Condition +ve	156	19	Precision	False discovery rate
			89%	11%
Predicted Condition – ve	51	24	False omission rate	Negative predictive value
			68%	32%
Prevalence	Sensitivity, Recall TPR	Fallout FPR	+ve likelihood ratio	- ve likelihood ratio
83%	75%	44%	1.71	0.44
	Miss Rate FNR	Specificity TNR	F1 Score	
	25%	56%	71	

Random Forest:

- The model is built with most significant variables based on feature selection.
- Significant variables:

  Status.of.existing.checking.account

  Credit.History

  Savings.account.bonds

  Present.employment.since

  Other.installment.plans

Predictive Modelling For Credit Card Fraud Detection Using Data Analytics

Random Forest:

Table 4. Confusion Matrix for Random Forest Decesion Tree

	Actual Condition		Accuracy	
	Condition +ve	Condition – ve	76%	
Predicted Condition +ve	163	12	Precision	False discovery rate
			93%	7%
Predicted Condition - ve	48	27	False omission rate	Negative predictive value
			64%	36%
Prevalence	Sensitivity, Recall TPR	Fallout FPR	+ve likelihood ratio	- ve likelihood ratio
84%	77%	31%	2.51	0.33
	Miss Rate FNR	Specificity TNR	F1 Score	
	23%	69%	61	

Predictive Modelling For Credit Card Fraud Detection Using Data Analytics

- To improve the analytical accuracy of fraud detection, we have implemented three models: Logistic Regression, Decision Tree, and Random Forest.
- Among the three**, Random Forest** model performs better in terms of **precision, recall and accuracy.**
- The only problem with decision tree is **overfitting of tree** in memory as data increases.
- The future scope of this work is to remove overfitting problem of decision tree and to detect real time fraud transaction for high streaming real-time data.

Credit Card Fraud Detection using Machine Learning Algorithms

- With different frauds, credit card frauds are in top of mind for most the world's population.
- Multiple supervised and unsupervised techniques are used to solve fraud detection.
- There are **three major problems** involved, i.e.) strong class imbalance, the inclusion of labelled and unlabelled samples, and to increase the ability to process large amount of transactions.
- Although **random forest** method performs good it suffers from the **problem of class imbalance.**

Credit Card Fraud Detection using Machine Learning Algorithms

- It suffers from a main problem called **concept drift.**
- Concept drift can be described as a variable which changes over time and in unforeseen ways.
- Thus it leads to highly **imbalanced** data.
- The dataset contains transactions made by a card holder in a duration of 2 days, i.e.) two days in the month of september 2013.
- 284,807 transactions are made out of which 0.172% i.e.) 492 transactions are fraudulent which accounts for high class imbalance.
- Since providing transactions details of customer is issue related to confidentiality therefore most of the transactions are converted into principal component analysis.
- Thus v1 to v28 are **PCA features** and ‘time’,’amount’,and ‘class’ are **non PCA features**

Credit Card Fraud Detection using Machine Learning Algorithms

Table 1: Raw features of credit card transactions

Attribute name	Description
Transaction id	Identification number of a transaction
Cardholder id	Unique Identification number given to the cardholder
Amount	Amount transferred or credited in a particular transaction by the customer
Time	Details like time and date, to identify when the transaction was made
Label	To specify whether the transaction is genuine or fraudulent

Table 2: Attributes of European dataset

S. No.	Feature	Description
1.	Time	Time in seconds to specify the elapses between the current transaction and first transaction.
2.	Amount	Transaction amount
3.	Class	0 - not fraud 1 – fraud

Credit Card Fraud Detection using Machine Learning Algorithms

- Firstly, we use cluster methods to divide the customers based on the **transaction amount**, i.e.) high, low, medium range.
- Using **sliding windows** we aggregate transactions into respective groups for extracting some feature from window to find card holders behavioural patterns.
- After that each group is trained on **different classifiers**.
- Features like maximum amount, minimum amount, average amount followed by time elapsed are found.
- We perform **SMOTE** (Synthetic Minority Over-Sampling Technique) operation on dataset.
- Oversampling does not provide good results.
- Thus there are two ways of dealing with class imbalance, **Matthew correlation coefficient** on original dataset or we can make use of **one-class classifiers**.

Credit Card Fraud Detection using Machine Learning Algorithms

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

TP- True Positive
TN- True Negative
FP- False Positive
FN- False Negative

Credit Card Fraud Detection using Machine Learning Algorithms

- We have experimented few models on original as well as SMOTE dataset.
- Original dataset

Table 3: Accuracy, Precision and MCC values before applying SMOTE,

Methods	Accuracy	Precision	MCC
Local Outlier factor	0.8990	0.0038	0.0172
Isolation forest	0.9011	0.0147	0.1047
Support vector machine	0.9987	0.7681	0.5257
Logistic regression	0.9990	0.875	0.6766
Decision tree	0.9994	0.8854	0.8356
Random forest	0.9994	0.9310	0.8268

- We observed that **Matthew Correlation** coefficient was better parameter to deal with imbalanced dataset.
- It was not the only solution.
  - By applying **SMOTE** we tried to balance the dataset and find classifiers that were performing better than before.
  - The other way of handling class imbalance is one-class SVM.
- We finally observed that **Logistic Regression, Decision Tree, and Random Forest** were the algorithms that gave the better results.

Credit Card Fraud Detection using Machine Learning Algorithms

- SMOTE dataset.

Table 4: Accuracy, Precision and MCC values after applying SMOTE,

Methods	Accuracy	Precision	MCC
Local Outlier factor	0.4582	0.2941	0.1376
Isolation forest	0.5883	0.9447	0.2961
Logistic regression	0.9718	0.9831	0.9438
Decision tree	0.9708	0.9814	0.9420
Random forest	0.9998	0.9996	0.9996

Comparative Analysis of Machine Learning Algorithms through Credit Card Fraud Detection

- Researchers analyzed correlation of certain factors with fraudulence by examining numerous **classification models** trained on the public dataset.
- Researchers proposed the following:
  - Better metrics to determine false negative rate.
  - Measure the **effectiveness of random sampling** to diminish the imbalance of the dataset.
- Conclusion: **Support Vector Machine** algorithm had the highest performance rate for detecting credit card fraud under realistic conditions.

Comparative Analysis of Machine Learning Algorithms through Credit Card Fraud Detection

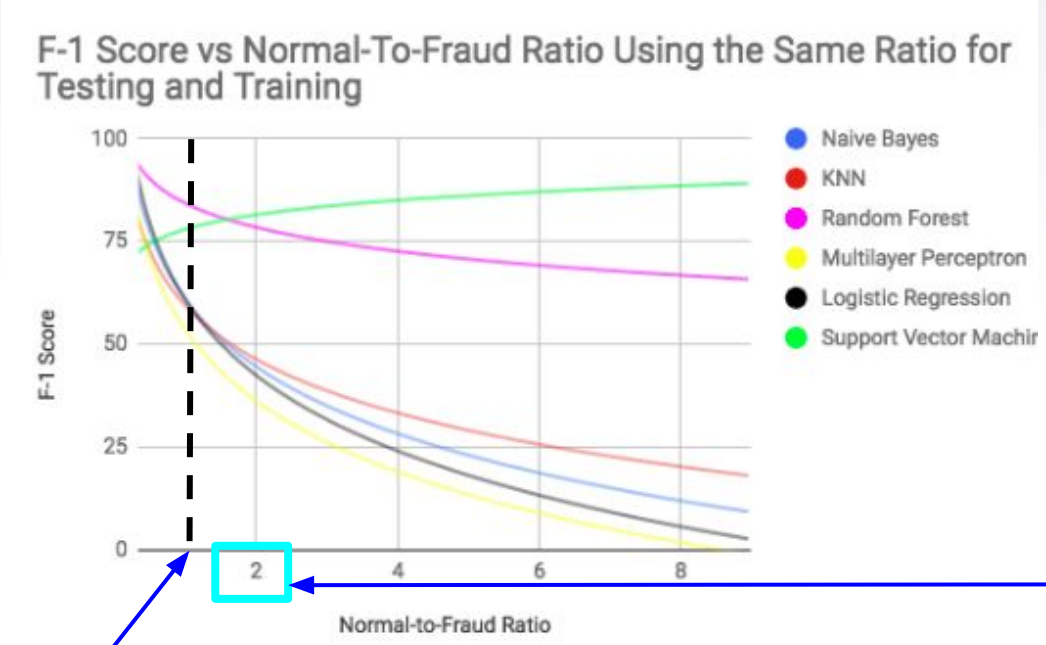
- Credit-card fraud is the term used for an unauthorized use of funds in a transaction by mean of credit or debit card.
- There are 2 types of fraud: **card-present** (stolen card) and **card-not-present** (online shopping)
- Credit card companies need robust detection of fraudulent transactions to minimize monetary losses. This can be done by optimizing companies algorithmic solutions.

Comparative Analysis of Machine Learning Algorithms through Credit Card Fraud Detection

Classification Model Considered

- Random Forest Classifier
- K Nearest Neighbor
- Logistic Regression
- Naive Bayes Classifier
- Multilayer Perceptron
- Support Vector Machine

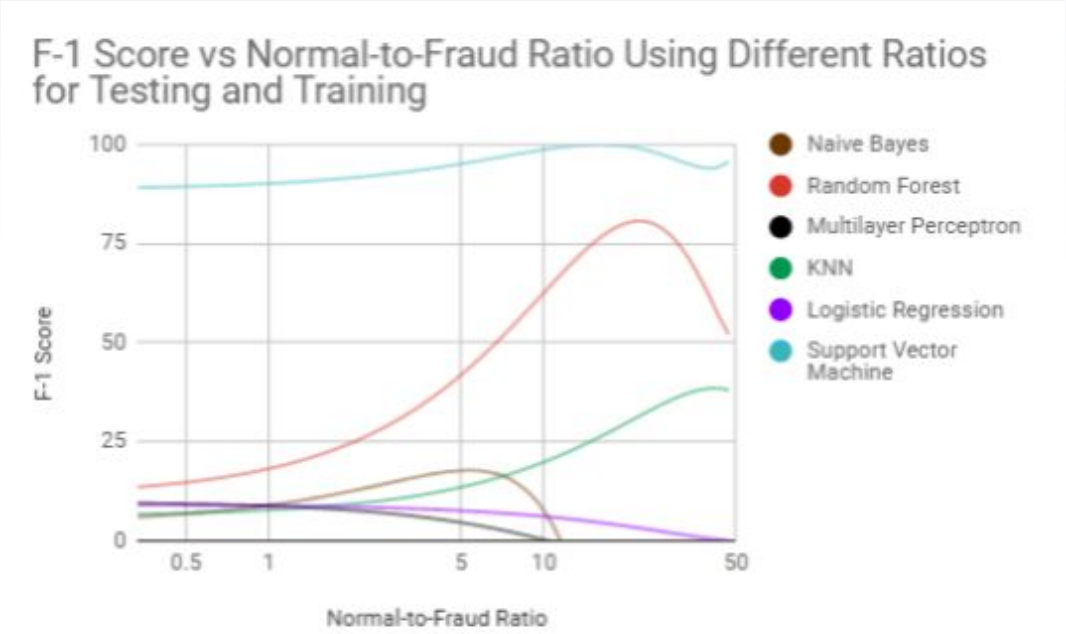
Comparative Analysis of Machine Learning Algorithms through Credit Card Fraud Detection



Balanced data (Ratio 1:1)
Random Forest produced the highest F1 score when balanced dataset was used.

- Figure shows the **effectiveness** (measured by F1-score) of the algorithms as the ratio of non-fraudulent to fraudulent data ratio increases.
- Datasets were created using **undersampling** method. The same ratio of fraudulent to non-fraudulent was used for training and testing processes.
- Number 2 means that the ratio is 2:1
- **SVM** increased in effectiveness as the ratio become higher compared to other algorithms.

Comparative Analysis of Machine Learning Algorithms through Credit Card Fraud Detection



All of the algorithms created were next tested on **highly skewed datasets** with fraudulent to non-fraudulent ratio of 98:2.

SVM produced the **highest F1 score** when highly skewed dataset was used.

Unlike most models tested, the F1 score of the SVM model did not decrease as the normal-to-fraud ratio decreased.

Comparative Analysis of Machine Learning Algorithms through Credit Card Fraud Detection

Random Forest	K Nearest Neighbor	Logistic Regression	Naive Bayes	Multilayer Perceptron	Support Vector Machine
The most efficient algorithm for testing on biased datasets at the **optimal dataset ratio.** This is because the algorithm does not require much processing **power and time** to train efficiently, unlike SVM.	The algorithm is not very efficient because the algorithms calculates **Euclidean distance** between all of the points in the dataset and all of the points in the testing set. Model takes less time to acquire results.	The algorithm works better when it is trained and tested on the **same dataset ratio** and size. When forced to test on very biased dataset, a training dataset split 75% normal transactions and 25% fraudulent transactions is best.	When ratio of fraudulent to non-fraudulent is high, the model is provided with **evidence of trends** pointing towards normal transactions. The model simplicity prevents it from observing patterns of great complexity within the data.	The model shown to be very unstable when imbalanced datasets were used. Model failed at any fraudulent to non-fraudulent ratio when there were **more than one attribute** being tested at a time.	Model works well when imbalanced datasets were used. Model **requires more time** and power to compute fitting the model.

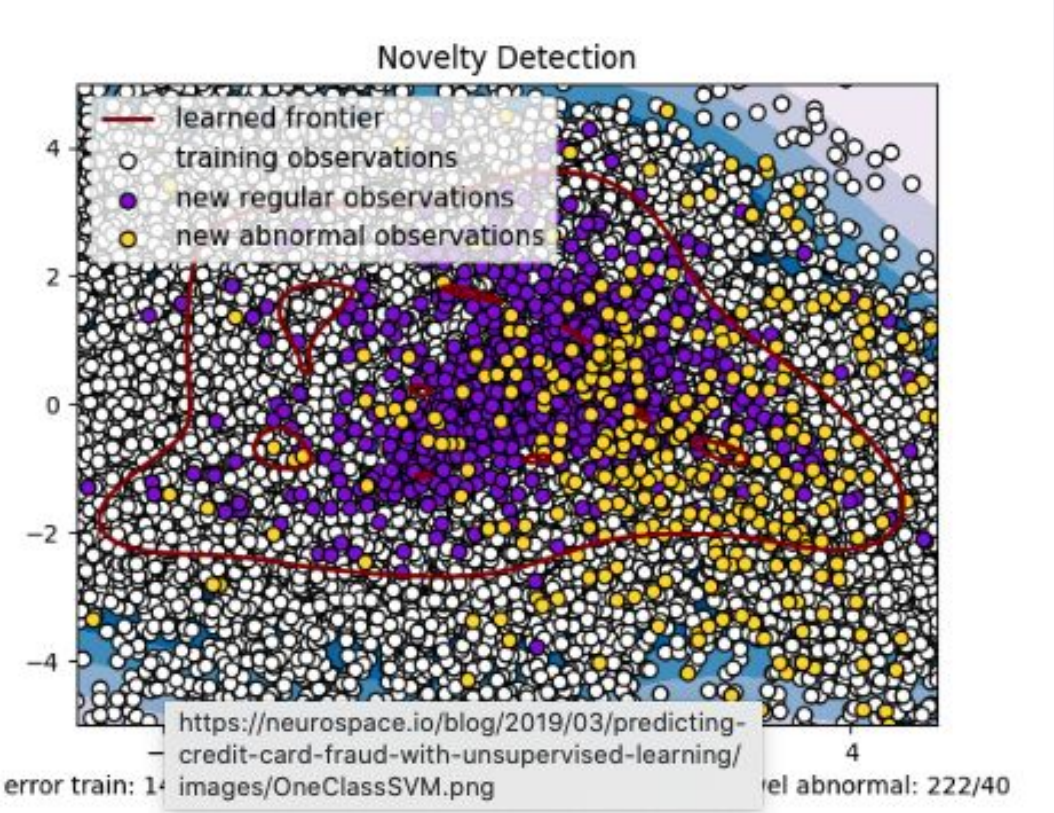
Conclusion

- The best algorithm for analysis of datasets with a **close to 1:1 ratio** of fraudulent to non-fraudulent transactions is the **Random Forest Classifier**, assuming the fraud-to-not fraud distribution of the testing and training set is the same.
- The optimal machine learning algorithm that a credit card company should use is dependent on the F1 scores of algorithms tested with highly skewed datasets.
- **Support Vector Machine** was the most successful in the detection of credit card fraud when tested under more realistic conditions.
- A credit card company should consider implementing a Support Vector Machine algorithm that analyzes the purchase time in order to most accurately detect whether a credit-card transaction is fraudulent or not.

Introduction

- Fraudulent transaction can be detected using **One Class Support Vector Machine**, an unsupervised learning algorithm.
- The algorithm can detect fraudulent transactions when it has been trained only on **normal transaction**.
- Unsupervised learning is when the algorithm train on data without output value or labeled data. It uses **correlations and patterns** to classify data.

Predicting Credit Card Fraud with Unsupervised Learning



One Class SVM:

- One Class SVM is one of a few algorithms designed for **outlier detection**. Others includes IsolationForest and LocalOutlierFactor.
- One class is **semi-supervised learning**. It trains on “healthy” data (normal transactions) and learn its pattern. When data with abnormal pattern is introduced, algorithm will classify this new pattern as an outlier.

Conclusion

- One class SVM is inspired by how SVM separates different classifications by the help of a **hyperplane margin**.

- One Class SVM is great for solving **imbalanced** problems. It can predict correctly with accuracy of 95% opposed to 99% labeled artificial neural network.

Handling Class Imbalance in Credit Card Fraud Detection using Resampling

Introduction

- As online based credit card payments have grown rapidly so has the need for effective fraud detection systems that handle large amounts of user data.
- Current methods of credit card fraud detection have problems in that the **data is heavily imbalanced in nature**. It means that very small percentages of all credit card transactions are fraudulent. This causes the detection of fraud transactions to be very difficult and imprecise.
- As previously mentioned classification techniques do not perform well when it comes to **huge numbers of differences in the minority and majority cases.**
- We use **resampling methods** on the data in order to combat the data imbalance.
- We will show the difference in resampling the data for four different machine learning classification techniques

Handling Class Imbalance in Credit Card Fraud Detection using Resampling

What exactly is Resampling?

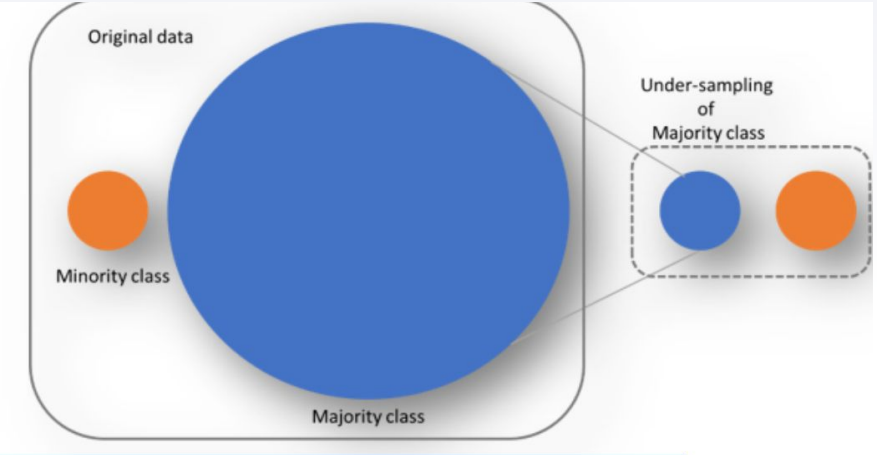
- Resampling Methods or the process of changing our data is when one repeatedly draws samples from a training set and refits a model of interest on each sample in order to obtain additional information about the new fitted model.
- The purpose is that it helps to derive a more accurate picture of the data where a dataset could be heavily imbalanced.
- As previously mentioned classification techniques do not perform well when it comes to **huge numbers of differences in minority and majority cases so resampling becomes essential in order to fix this.**
- In credit card fraud detection, a great system should be able to handle noise,**avoid the overlapping data,** adapt themselves to new kinds of frauds, evaluate the classifier using good metrics, and detect the behaviour of the frauds.
- We use the resampling methods like, oversampling,undersampling, and SMOTE in order to avoid imbalanced data and create more accurate fraud detection systems.

What is imbalanced data?

Fraud	592	0.15%
Non-Fraud	384315	99.85%

- Data imbalance means that the distribution of samples across different classes is unequal. In our Credit Card dataset which is a binary classification(either a transaction is fraud given a 0 or non-fraudulent given a 1) we identify two classes:
- a) Majority class- the non-fraudulent/genuine/real transactions
- b) Minority class- the fraudulent transactions
- As an example, the figure below shows that only 0.15% of the observations being in the Fraudulent category
- We use resampling methods to bypass this bias in the data.

Resampling (Sampling Methods)



RANDOM UNDER-SAMPLING

Method that randomly deletes examples in the majority class. Random draws are taken from the non-fraud observations which is the majority class to match it with the Fraud observations the minority class. This means, we are throwing away (deleting) some information from the dataset which might not be ideal always.

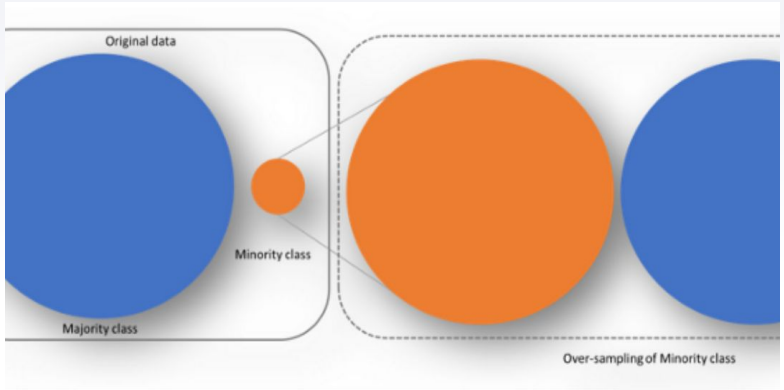
Advantages:

- Used to alleviate imbalance in the dataset
- Improves the runtime
- Reduces the number of training data samples when the training data set is huge

Disadvantages:

- Can increase the variance of the classifier, which could bias the dataset toward certain values
- It could potentially discard useful or important samples.

Resampling (Sampling Methods)



RANDOM OVER-SAMPLING

Method that randomly deletes examples in the minority class. The opposite of under-sampling. This duplicates the minority class or the Fraudulent observations at random to increase the number of the minority class till we get a balanced dataset.

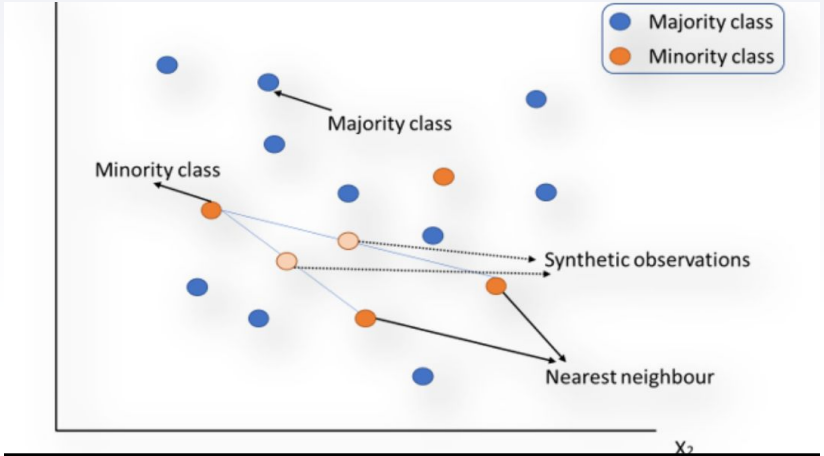
Advantages:

- Unlike under sampling this method leads to minimal information loss.
- Outperforms under sampling

Disadvantages:

- It increases the likelihood of overfitting since it replicates the events that happen in the minority class..

Resampling (Sampling Methods)



SMOTE (SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE)

SMOTE is a technique that generates synthetic samples from the minority class. Method uses synthetic data with KNN instead of using duplicate data. Each minority class example along with their k-nearest neighbours is considered.Then the sythentic examples are created along the line segments that join all the minority class examples and their k-nearest neighbours

Advantages:

- Alleviates overfitting caused by random oversampling as synthetic examples are generated rather than replication of instances.
- No loss of information.
- It's simple to implement and interpret

Disadvantages:

- SMOTE does not take into consideration neighboring examples can be from other classes. This can increase the overlapping of classes and can introduce additional noise.
- SMOTE is not very practical for high dimensional data.

Handling Class Imbalance in Credit Card Fraud Detection using Resampling

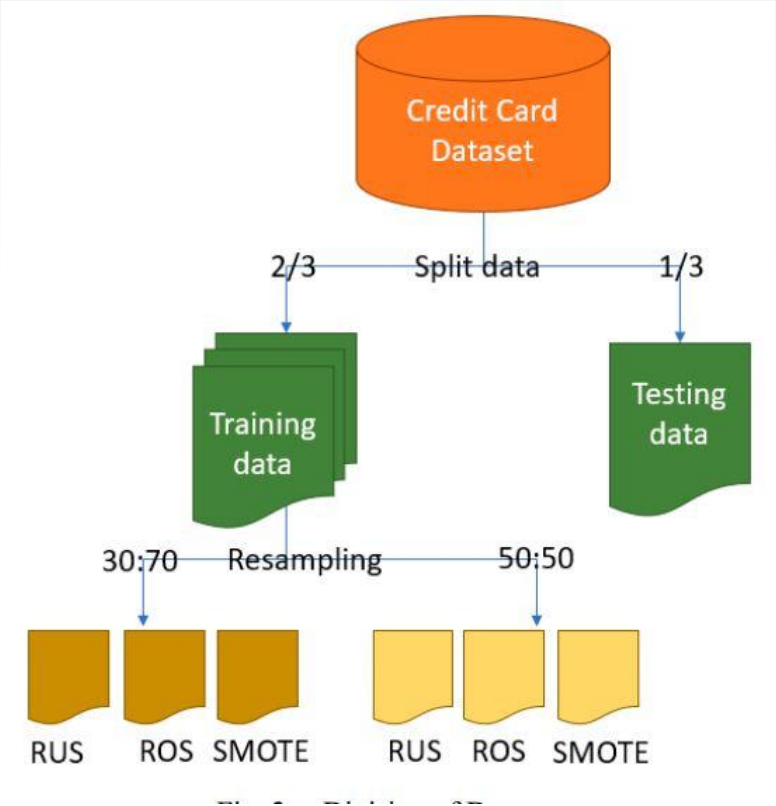
The Dataset

- The dataset that was used to test whether our resampling methods improved the accuracy of fraud detection system is a publicly available dataset containing a total of 284,807 transactions made in September 2013 by European cardholders.

- The dataset contains 492 fraud transactions, making it highly imbalanced.

- We test the accuracy by doing a comparison of the training data on the regular data versus the resampled data.

Handling Class Imbalance in Credit Card Fraud Detection using Resampling



- In the study, the dataset was resampled using the three different techniques
- Then four different classification techniques were then explored in order by performance: Naïve Bayes (NB), Linear Regression (LR), Random Forest (RF) and Multilayer Perceptron (MLP).
- The research handled the imbalance problem in churn prediction by resampling the minority and majority classes based on ratios 10:90, 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, 80:20 and 90:10 where churners proportionate with non-churners.

Handling Class Imbalance in Credit Card Fraud Detection using Resampling

Metrics used to analyze performance:

In order to test the performance of our resampling versus the regular unsampled data, we create a confusion matrix. The confusion matrix tells us the performance of a classification models.

TABLE V. CONFUSION MATRIX OF CREDIT CARD DATASET

	Classified as Fraud	**Classified as Non-Fraud**
Fraud	True Positive (TP)	False Negative (FN)
Non-Fraud	False Positive (FP)	True Negative (TN)

- True Positive(TP)- represents the number of cases correctly classified as fraud
- True Negative(TN)-represents the number of cases correctly classified as non-fraud
- False Negative(FN)- represents the cases that are actually fraud but the model predicted as non-fraud
- False Positive(FP)- those cases that are actually-non fraud but the model predicted as Fraud

Handling Class Imbalance in Credit Card Fraud Detection using Resampling

Additional Metrics used to analyze performance:

The confusion matrix and its values allow us to create mathematical formulas to represent different aspects of our dataset performance under resampling. The formulas listed are used to detail the results:

$$Sensitivity = \frac{TP}{TP+FN} \quad (1)$$

$$Specificity = \frac{TN}{TN+FP} \quad (2)$$

$$Accuracy = \frac{TP+TN}{\sum(TP+FP+TN+FN)} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$F\text{-}Measure = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall} \quad (5)$$

$$AUC = \frac{1}{2} \cdot (Sensitivity + Specificity) \quad (6)$$

$$Error = \frac{FP+FN}{\sum(TP+FP+TN+FN)} \quad (7)$$

Handling Class Imbalance in Credit Card Fraud Detection using Resampling

COMPARISON RESULTS OF CLASSIFICATION TECHNIQUES BY RATIO 50:50

Resampling Methods: Random Under Sampling						
Techniques	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**F-Measure**	**Time (s)**
Naïve Bayes	0.85321101	0.97009103	0.93521898	0.92384106	0.88712242	0
Linear Regression	0.89602446	0.9869961	0.95985401	0.9669967	0.93015873	0.08
Random Forest	0.90825688	1	**0.97262774**	1	0.95192308	0.5
Multilayer Perceptron	0.9204893	0.98829649	0.96806569	0.97096774	0.94505495	4.52
Resampling Methods: Random Over Sampling						
Techniques	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**F-Measure**	**Time (s)**
Naïve Bayes	0.84758845	0.97382652	0.93595468	0.93279053	0.88815077	2.11
Linear Regression	1	0.99763642	0.9983455	0.99451532	0.99725012	53.5
Random Forest	1	1	**0.99995199**	1	1	343.43
Multilayer Perceptron	0.96569171	0.9989079	0.98894293	0.99736822	0.98127439	896.31
Resampling Methods: Synthetic Minority Over Sampling Technique						
Techniques	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**F-Measure**	**Time (s)**
Naïve Bayes	0.833294	0.97426441	0.93194925	0.93283212	0.88025831	2.51
Linear Regression	0.98990147	0.9940594	0.99281131	0.98620166	0.9880481	88.16
Random Forest	0.99924968	0.99957266	**0.99947571**	0.99900392	0.99912679	716.2
Multilayer Perceptron	0.98723232	0.99297785	0.9912532	0.98368713	0.98545653	1034.57

Handling Class Imbalance in Credit Card Fraud Detection using Resampling Conclusion

- Study aimed to discover whether the classification models could identify fraud and if the resampling methods would improve performance of the models
- The comparison results for each classification techniques in three resampling methods by ratio 30:70 and 50:50, correspondingly
- Resampling proved effective in accurately detecting fraudulent activity within the data
- Of all the four classifiers, Random Forest(RF) proved to have robust performance in all three resampling methods

Handling Class Imbalance in Credit Card Fraud Detection using Resampling Conclusion

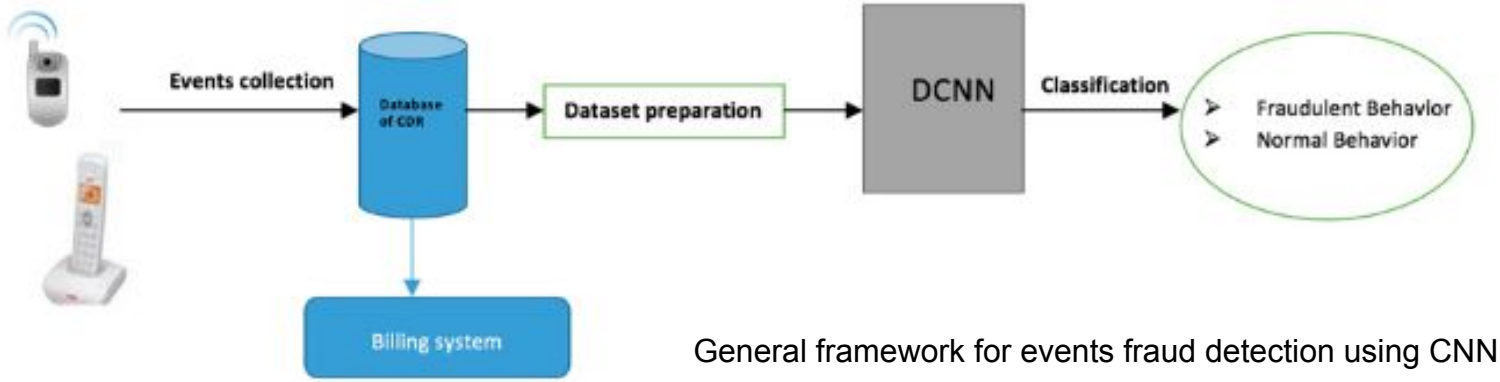
- **(ROS) was found to give convincing results if compared to SMOTE**
- ROS, Logistic Regression(LR) and RF have an accuracy of 99% which they correctly identified the fraud and non fraud of the credit card dataset.
- Random Sampling drastically **improved all metrics on training data.**
- These results may provide further support to the organisation to build better fraud detection system which can handle the skewed distribution and noise

ConvNets for Fraud Detection Analysis

Introduction

- Telecommunication companies are incredibly reliant on developing efficient algorithms that detect early potential frauds and/or prevent them
- As previously mentioned fraudulent behaviour is defined as illegal use of telecom infrastructure like mobile communications with an intention for not paying services or misuse of voice calls. Neural Technologies, a software company, estimated that in 2016 the telecom industry lost $249 billion dollars USD due to fraud activities.
- The use of convolutional neural networks (CNN) can be useful in finding fraudulent behavior
- Comparison of CNN's versus other machine learning algorithms will evaluate the performance of which is best in fraud detection

ConvNets for Fraud Detection analysis



General framework for events fraud detection using CNN architecture

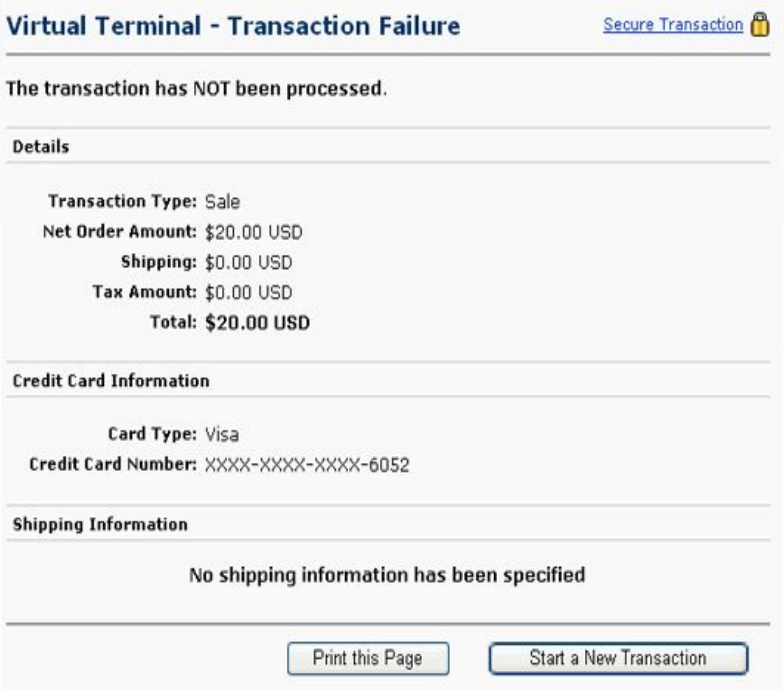
The dataset represents the artificial images of different subscribers during two months of behaviour. This historical data will be used as indicator capturing any change from a normal behaviour to fraud. The dataset contains 18000 generated images representing the profile of 300 users during the period of 60 days.

ConvNets for Fraud Detection analysis

Classifying Fraudulent Behaviour

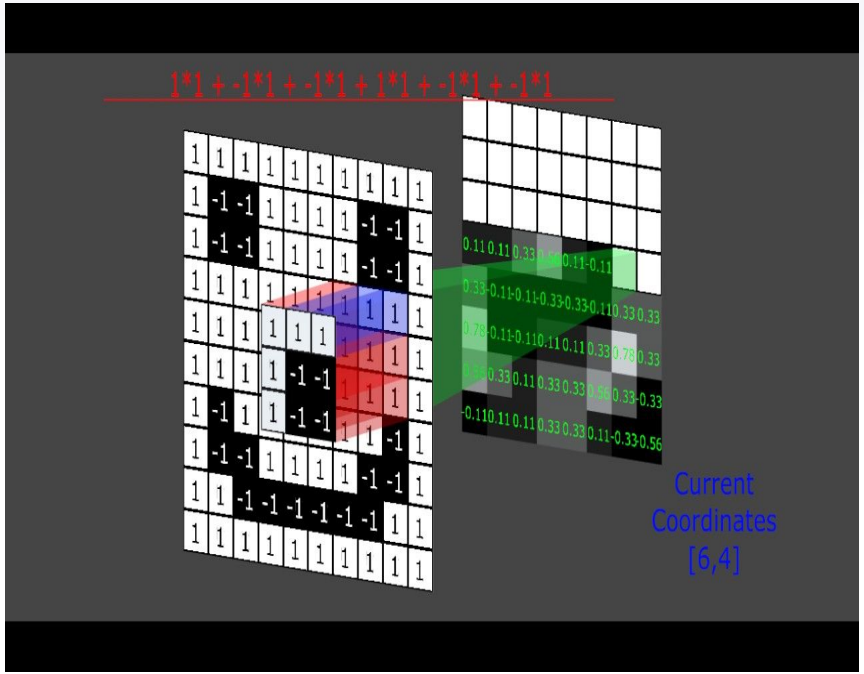
- A convolutional neural network (or ConvNet) is a type of feed-forward artificial neural network
- The architecture of a CNN is designed to take advantage and classify the 2 dimensional structure of an input image
- In our dataset containing 18000 generated images, we would be looking for minute details in the image to detect whether the transaction was fraudulent or real.

Advantage of ConvNets



- Consider we have a fraudulent transaction image that is of size 300x300x3 (300 wide, 300 high, 3 color channels) .
  - It should be noted for our detection system we rescaled all the artificial images to 28x 28.
- If we were to create – a single fully-connected neuron in the first hidden layer of a regular Neural Network we would have 300*300*3 = 270,000 weights.
- Due to the presence of several such neurons and weights, this full connectivity would be too complex and wasteful and overfitting would become inevitable

Advantage of ConvNets

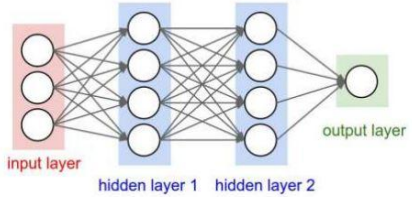


- However, if we use a ConvNet, the neurons in a layer will only be connected to a small region of the layer before it, instead of all of the neurons in a fully-connected manner
- The final output layer would have dimensions 1x1xN.
- This is because at the end of the architecture we will reduce the full image into a single vector of class scores (for N classes) all arranged along the depth dimension

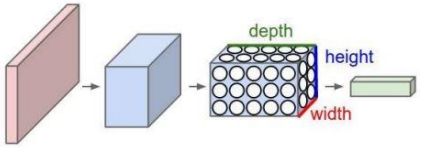
Advantage of ConvNets

- This is why in our fraud detection analysis system we use convolutional neural networks as opposed to a regular neural network which work better on image data
- The diagram below shows the advantage and ease in creating a convNet versus a regular neural network for an n-dimensional system

▪ A regular 3-layer Neural Network.



▪ A ConvNet arranges its neurons in three dimensions (width, height, depth), as visualized in one of the layers.



ConvNets for Fraud Detection Analysis

Convolutional Neural Network (CNN) Architecture

We create a CNN with the following components for the architecture in order to process our data into fradulent or non -fraudulent

Layer	Layer's specification
1	Convolution Layer 1
2	Convolution Layer 2+RELU
3	Max Pooling
4	Convolution Layer 3+RELU
5	Max Pooling
6	Fully connected layer
7	Softmax

ConvNets for Fraud Detection analysis

Hyperparameters

Before we train our deep convolutional neural network we outline very important parameters so, that our model can accurately learn which activity is fraudulent or not. The model was trained for different epochs using cross-entropy loss function. The employed stochastic gradient descent optimization method served as a momentum based learning rate using a mini batch size of 200.

Parameter	Value
Number of Epoch	46
Learning rate	0.1
Momentum	0.7
weight decay	0.045
Mini-batch size	200

number of times all of the training vectors are used to update the weights

optimization algorithm that determines the step size

improves both training speed and accuracy.

Weight update. Used to update training weights

number of training examples utilized in one iteration

ConvNets for Fraud Detection Analysis

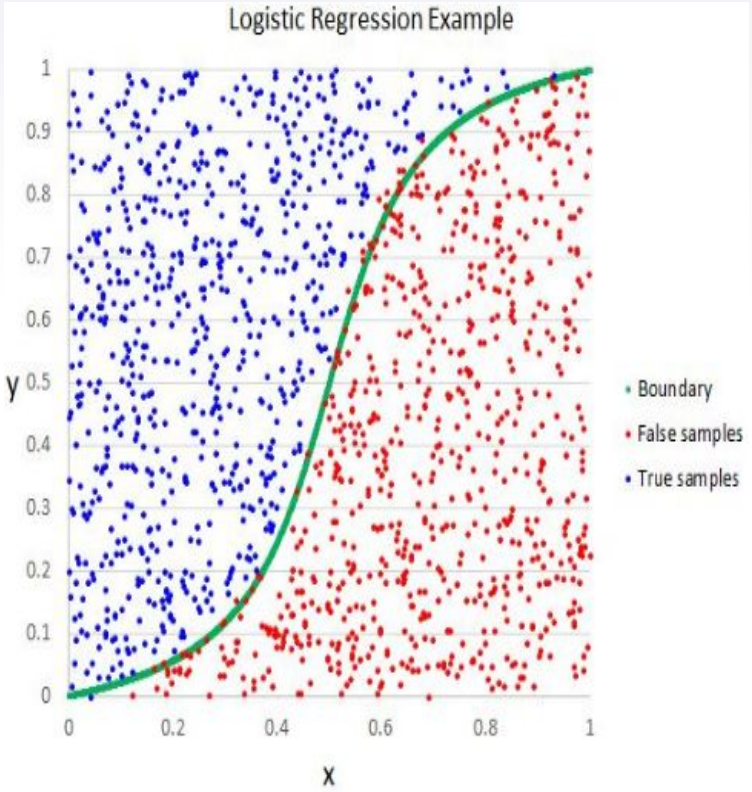
- In order to evaluate our proposed model, we have calculated the accuracy of our CNN model during different epochs.
- The performance of the CNN model, was evaluated using three different traditional machine learning algorithms: Support vector machines (SVM), Random Forest (RF) and Gradient Boosting Classifier (GBC)
- The CNN model proved superior in accuracy to any of the other models with an **accuracy of 82%.**
- This details that CNN prove more effective than other traditional machine learning algorithms in fraud detection

Model	Accuracy (%)
SVM	77
RF	72
GBC	79
DCNN	82

Credit Fraud Detection Based on Hybrid Credit Scoring Model

- Credit risk rating is the valuation of the credit risk of a prospective debtor, predicting their ability to pay back the debt,
- Credit risk rating can be described by several economic activity indicators. Utilizing these financial movement markers to build a tenable credit scoring model will enormously improve the precision of the model.
- In this paper, the logistic regression algorithm is joined with weighted evidence to fabricate another credit score model.
- In practice, due to numerous weaknesses in the records, there is significant error in the logistic regression.
- Hence, building of hybrid scoring model can increase the accurateness of credit score. Thus improved the prediction rate of user credit scores and reducing the occurrence of credit fraud.

Credit Fraud Detection Based on Hybrid Credit Scoring Model



- Credit fraud detection is mainly used to distinguish the trustworthy and the untrustworthy.
- The model is based on **Logistic Regression**, the advantage is that due to the high precision of the weight of evidence, its credit rating results contain more information to explain the relationship between variables and dependent variables.
- The logistic regression model can be built to guarantee fortifies and clarification of the model.
- As a result, this hybrid model is more accurate than the model recognized with a single process.

Credit Fraud Detection Based on Hybrid Credit Scoring Model

- This figure shows an analysis of the distribution of selected attributes to choose **the best amalgamation of features for better results, when the information fit to singular loans consumption for the accompanying essential properties:**
    - the age of the mortgagor;
    - dissolvability and credit (unpaid days);
    - the situation of the property;
    - loans past due property factors, including time, window; and the size of family borrower.

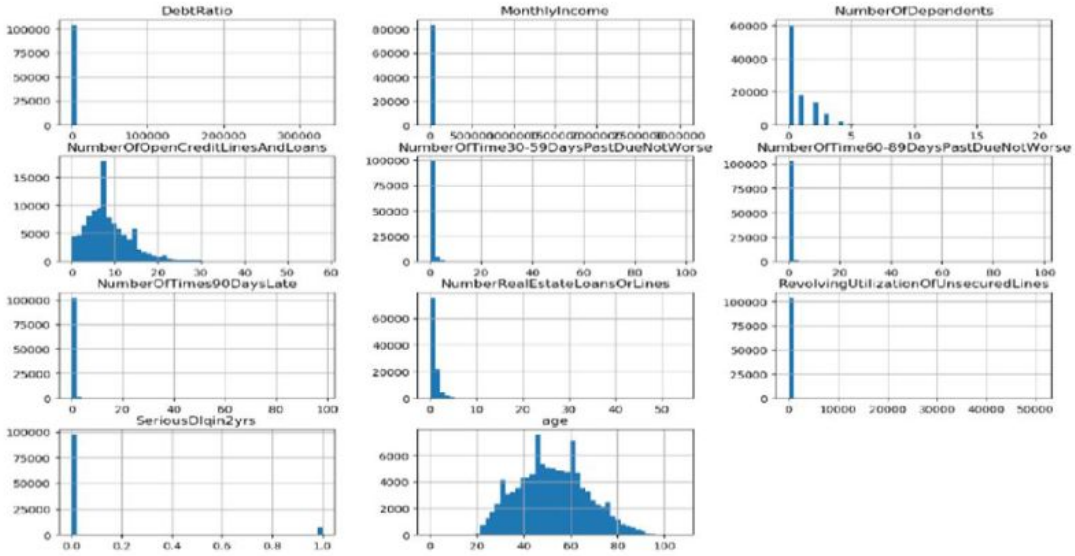


Fig. 1. Borrower Data Distribution Map

Credit Fraud Detection Based on Hybrid Credit Scoring Model

- The selection of variables was tested using model Information Value (IV), also called evidence weight.
- The IV values of X1 and X7 are large; the training effect on the model is significant.
- The IV values of X4, X5, X6, X8, and X10 are small; the training effect on the model is not significant.

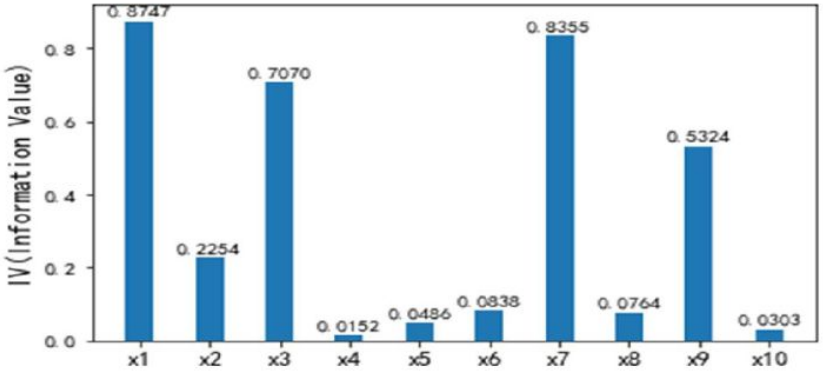


Fig. 2. IV Diagram (X1-X10 corresponds to the borrower's attributes)

Credit Fraud Detection Based on Hybrid Credit Scoring Model

Compared with the hybrid model, the mixed model prediction results are greatly improved.

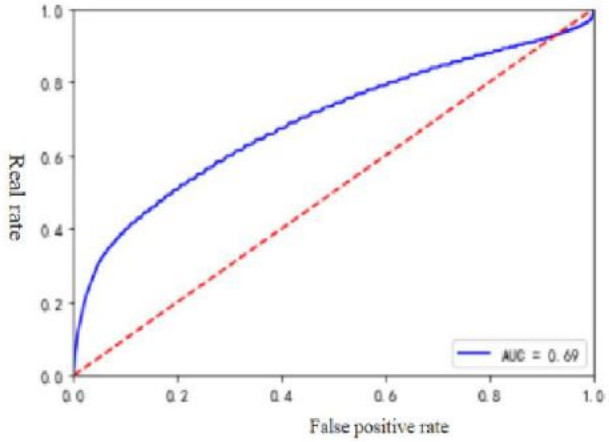


Fig. 4. AUC curve after traditional logistic regression test

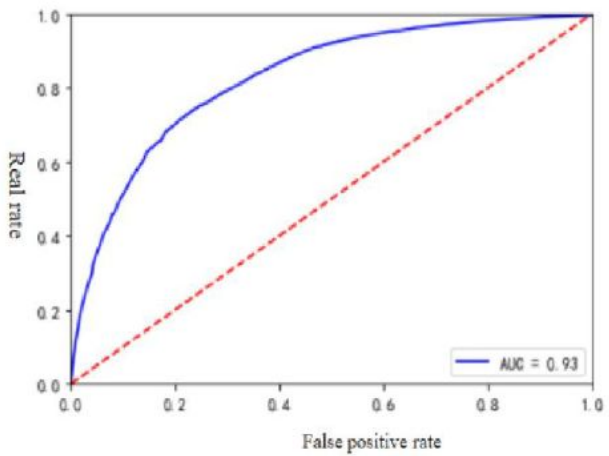


Fig. 5. AUC curve after mixed model test

Credit Fraud Detection Based on Hybrid Credit Scoring Model

- **A progressively powerful forecast model can be accomplished by combination of the evidence weight and Logistic regression model**. This mixed scoring model can get progressively exact credit score assessment yet, in addition, decrease risk of credit.
- For economical associations, it is hard to equitably accomplish all the proof of individual clients, which signs to the lacking proof asymmetry. This is one of the main reasons for credit risk and it is believed that this study will help to minimize such risk.
- For future examination, this exploration will be stretched out to accentuation on disclosure of treasured information from a lesser amount of information or an enormous amount of missing informational collections. At the same time, the model will be optimized to maintain the stability and robustness of the model and improve the computing speed. Then analysing and predicting of these data will be done to improve the level of credit fraud detection.