MBAS905: Assessment 2

Business Analytics Research Project

Name: Chinmay Datar

Student No.: 6956361

Exercise 1: Data Mining, Machine Learning and Text Analytics

Question 1. Data Mining and Machine Learning

D-1.

Category A of Status Category 96NK has the highest count of 64,086 as shown in figure 1.

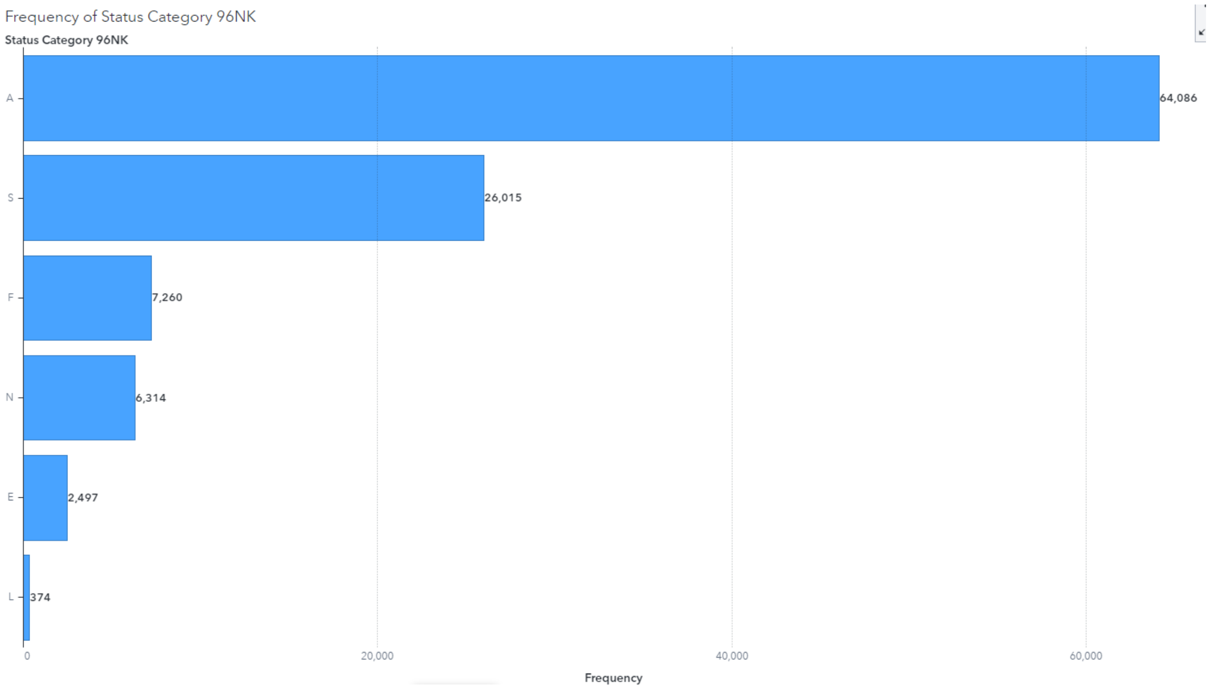
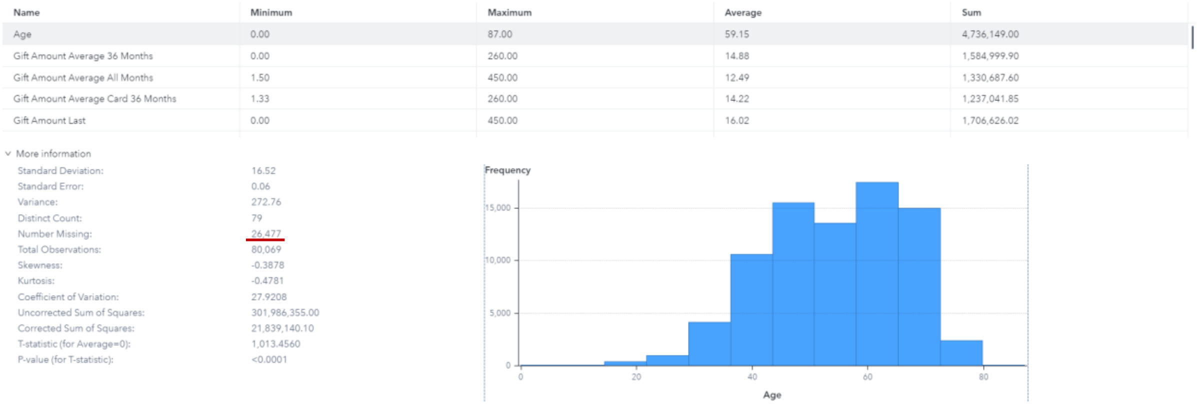


Figure 1

D-2.

Age has 26,477 missing values as can be seen in figure 2.



Name	Minimum	Maximum	Average	Sum
Age	0.00	87.00	59.15	4,736,149.00
Gift Amount Average 36 Months	0.00	260.00	14.88	1,584,999.90
Gift Amount Average All Months	1.50	450.00	12.49	1,330,687.60
Gift Amount Average Card 36 Months	1.33	260.00	14.22	1,237,041.85
Gift Amount Last	0.00	450.00	16.02	1,706,626.02

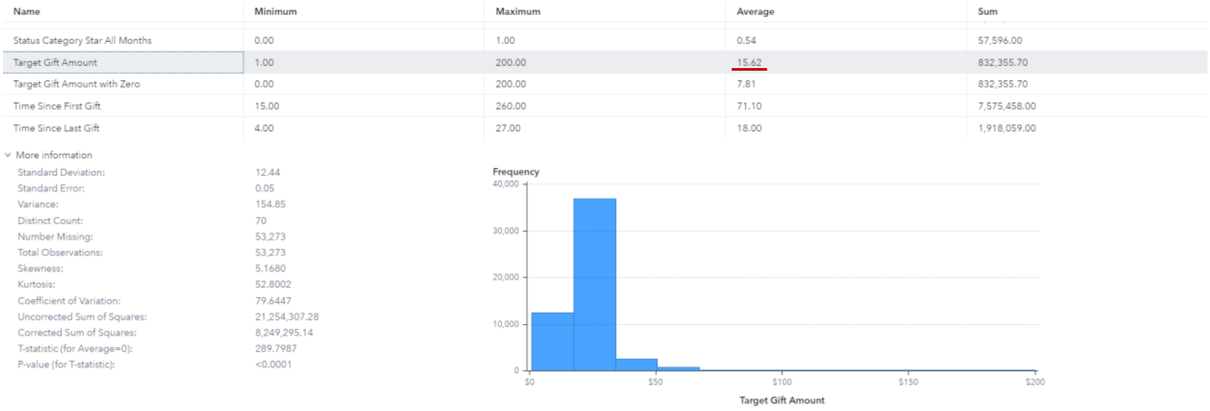
More information

Standard Deviation:	16.52
Standard Error:	0.06
Variance:	272.76
Distinct Count:	79
Number Missing:	26,477
Total Observations:	80,069
Skewness:	-0.3878
Kurtosis:	-0.4781
Coefficient of Variation:	27.9208
Uncorrected Sum of Squares:	301,986,355.00
Corrected Sum of Squares:	21,839,140.10
T-statistic (for Average=0):	1,013.4560
P-value (for T-statistic):	<0.0001

Figure 2

D-3.

The average of Target Gift Amount is 15.62 as seen in figure 3.



Name	Minimum	Maximum	Average	Sum
Status Category Star All Months	0.00	1.00	0.54	57,596.00
Target Gift Amount	1.00	200.00	15.62	832,355.70
Target Gift Amount with Zero	0.00	200.00	7.81	832,355.70
Time Since First Gift	15.00	260.00	71.10	7,575,458.00
Time Since Last Gift	4.00	27.00	18.00	1,918,059.00

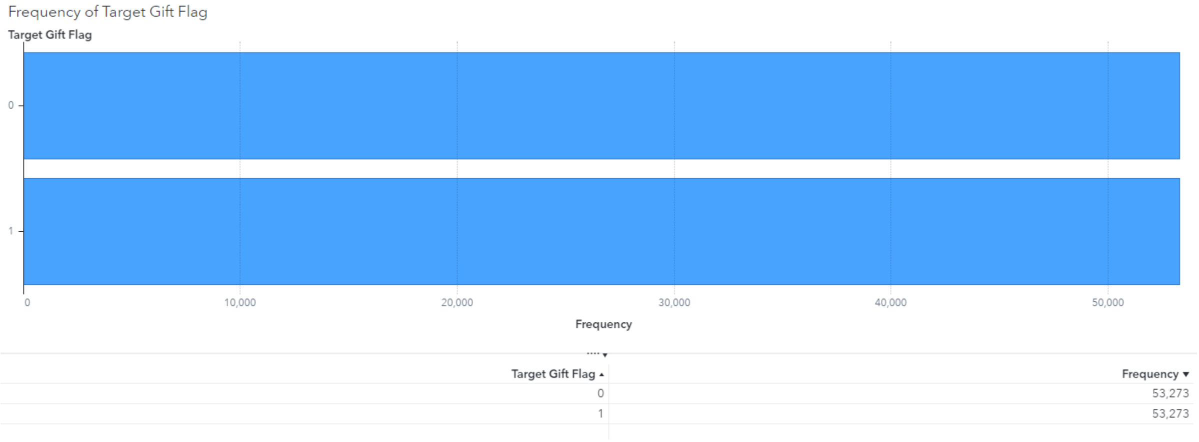
More information

Standard Deviation:	12.44
Standard Error:	0.05
Variance:	154.85
Distinct Count:	70
Number Missing:	53,273
Total Observations:	53,273
Skewness:	5.1680
Kurtosis:	52.8002
Coefficient of Variation:	79.6447
Uncorrected Sum of Squares:	21,254,307.28
Corrected Sum of Squares:	8,249,295.14
T-statistic (for Average=0):	289.7987
P-value (for T-statistic):	<0.0001

Figure 3

E-1.

Respondents and non-respondents are distributed equally. See figure 4.



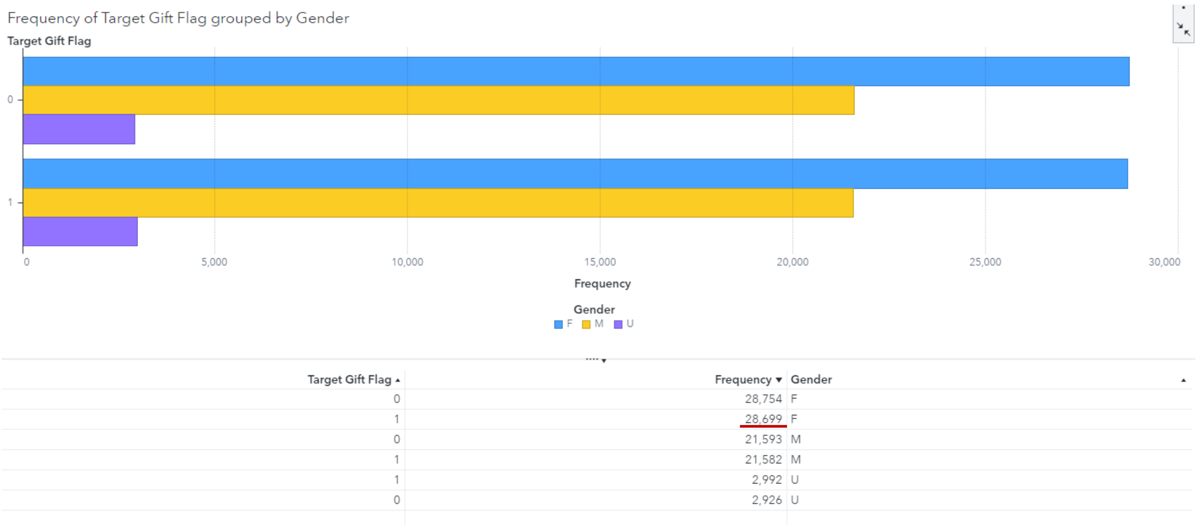
Frequency of Target Gift Flag

Target Gift Flag ▲	Frequency ▼
0	53,273
1	53,273

Figure 4

E-2.

From figure 5 we can see 28,699 females responded to the campaign.



Target Gift Flag ▴	Frequency ▾	Gender
0	28,754	F
1	28,699	F
0	21,593	M
1	21,582	M
1	2,992	U
0	2,926	U

Figure 5

Question 2

H-1.

64,988 observations are used by the neural network algorithm, as shown in figure 6.

H-2

The neural network algorithm does not use all the observations as some of the predictor variables has missing such as ‘Age’ and ‘Gift Amount Average Card 36 months’ has high number of missing values.

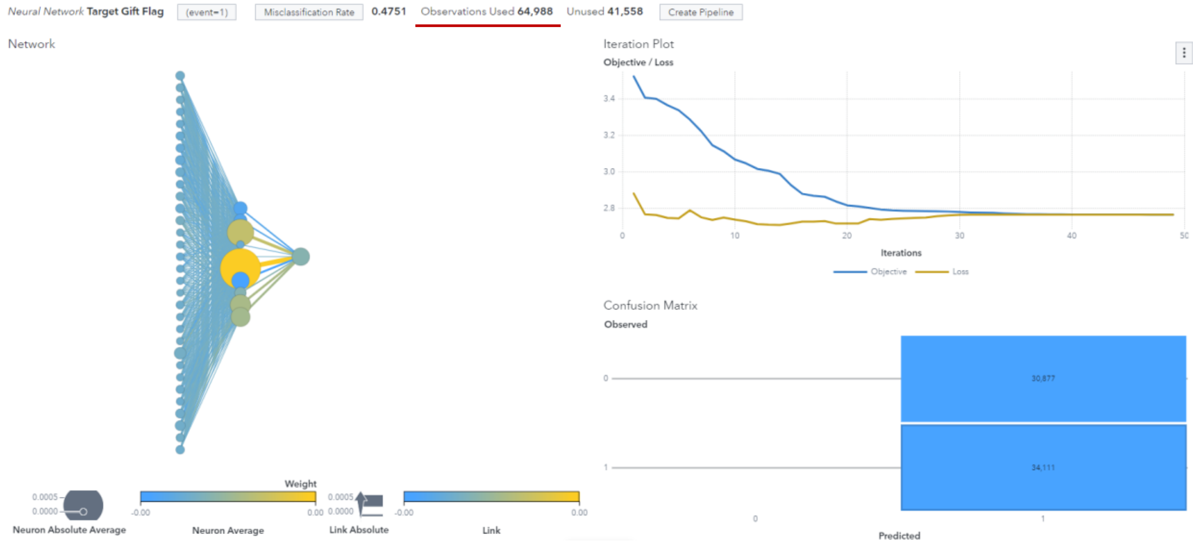

Neural Network Target Gift Flag (event=1) Misclassification Rate 0.4751 Observations Used 64,988 Unused 41,558 Create Pipeline

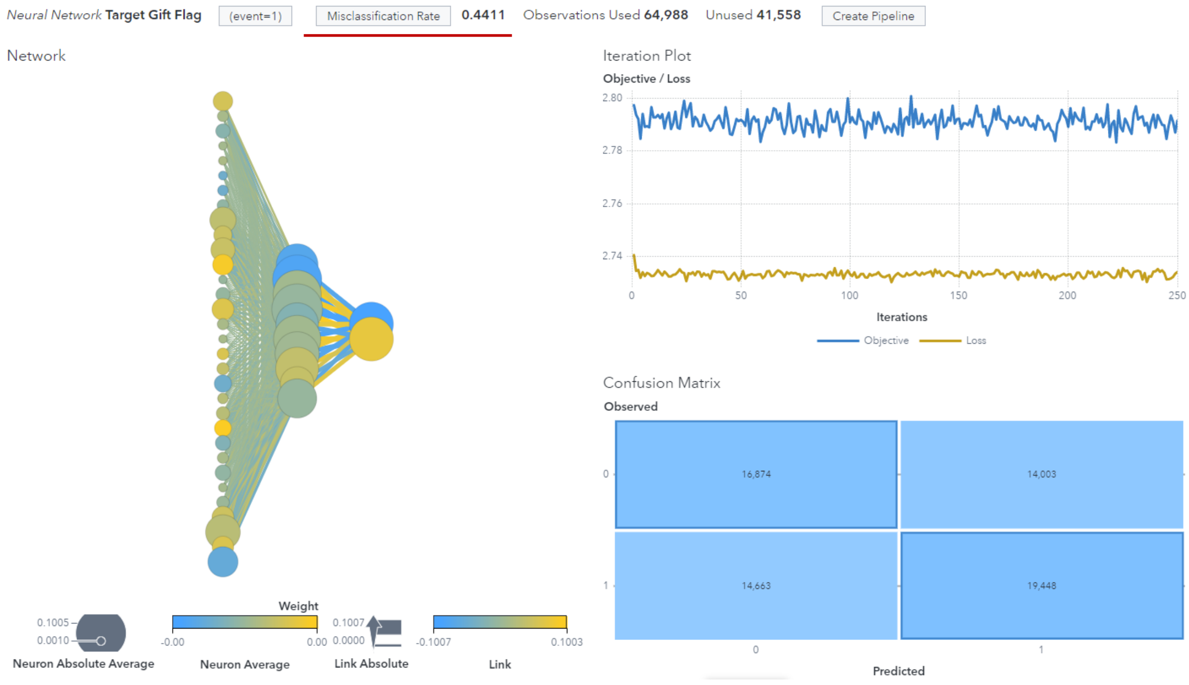
Figure 6

H-3

The misclassification rate of the model at default settings is 0.4751(47.51%) as seen in figure 6.

I.

Changing the optimisation method from LBFGS to SGD reduces the misclassification rate to 0.4411, as shown in figure 7.



Neural Network **Target Gift Flag** (event=1) Misclassification Rate **0.4411** Observations Used **64,988** Unused **41,558** Create Pipeline

Confusion Matrix

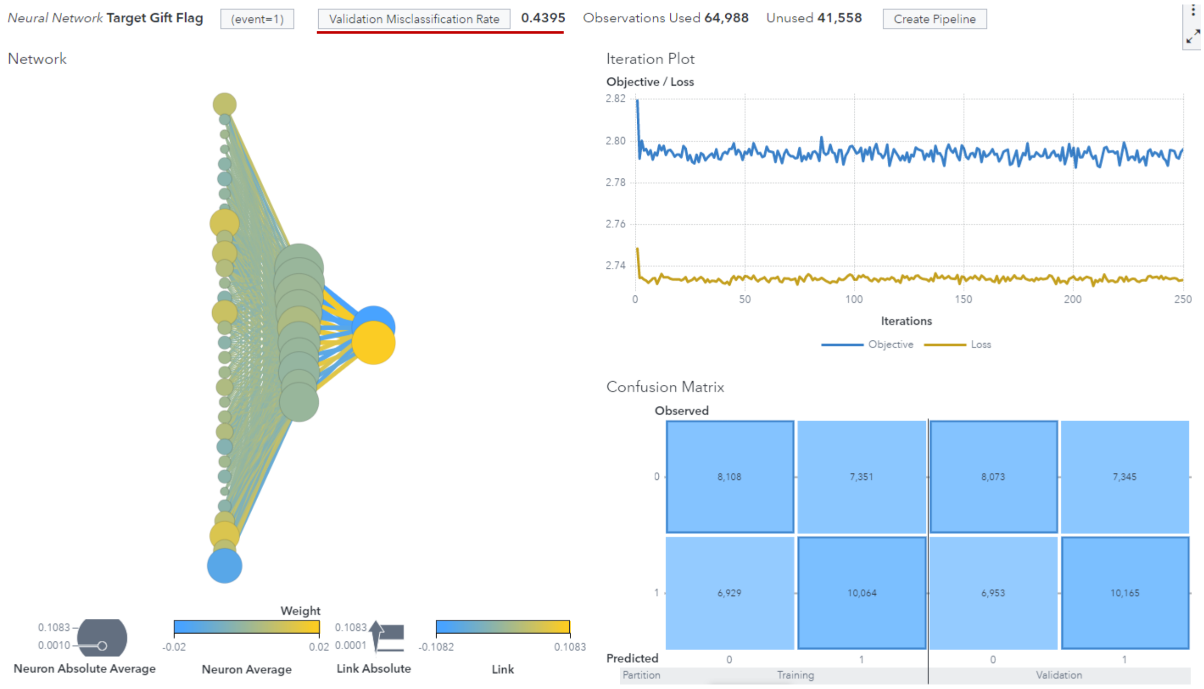
Observed	0	1
0	16,874	14,003
1	14,663	19,448

Predicted

Figure 7

J-2.

After adding the partition, the validation misclassification rate drops down to 0.4395, as shown in figure 8.



Neural Network **Target Gift Flag** (event=1) Validation Misclassification Rate **0.4395** Observations Used **64,988** Unused **41,558** Create Pipeline

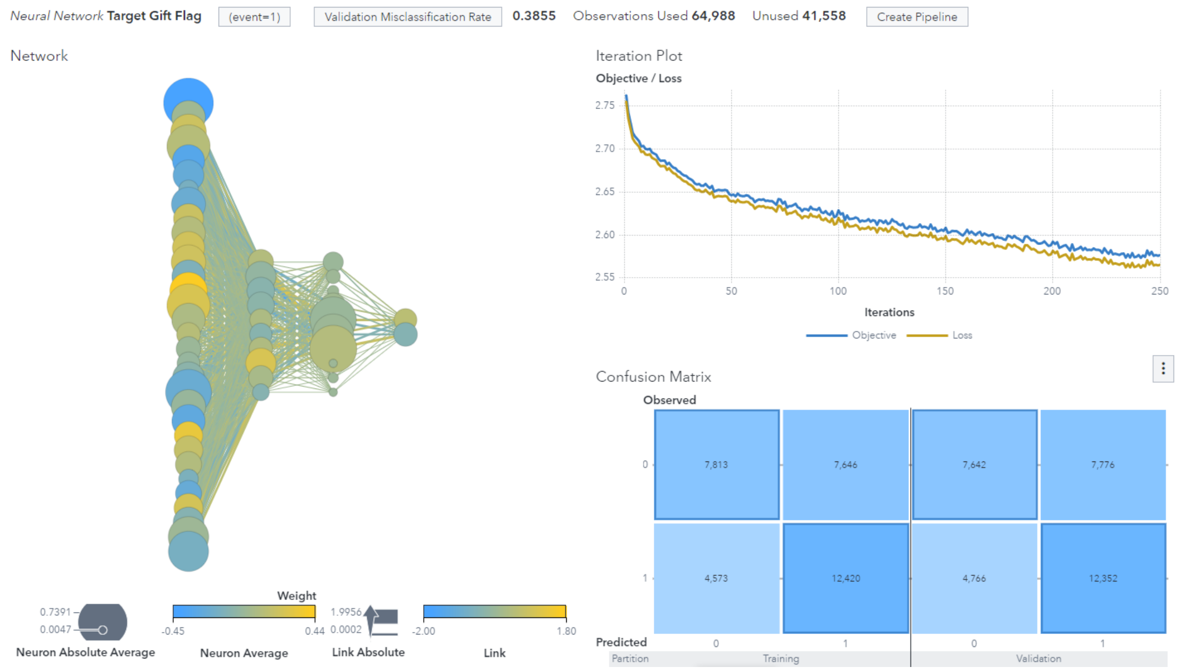
Confusion Matrix

Partition	Training		Validation	
Observed \ Predicted	0	1	0	1
0	8,108	7,351	8,073	7,345
1	6,929	10,064	6,953	10,165

Figure 8

J-3.

After changing L2 regularisation parameter to 0.001 and increasing number of hidden layers to 2 the misclassification rate further drops down to 0.3855, as shown in figure 9.



Neural Network **Target Gift Flag** (event=1) Validation Misclassification Rate **0.3855** Observations Used **64,988** Unused **41,558** Create Pipeline

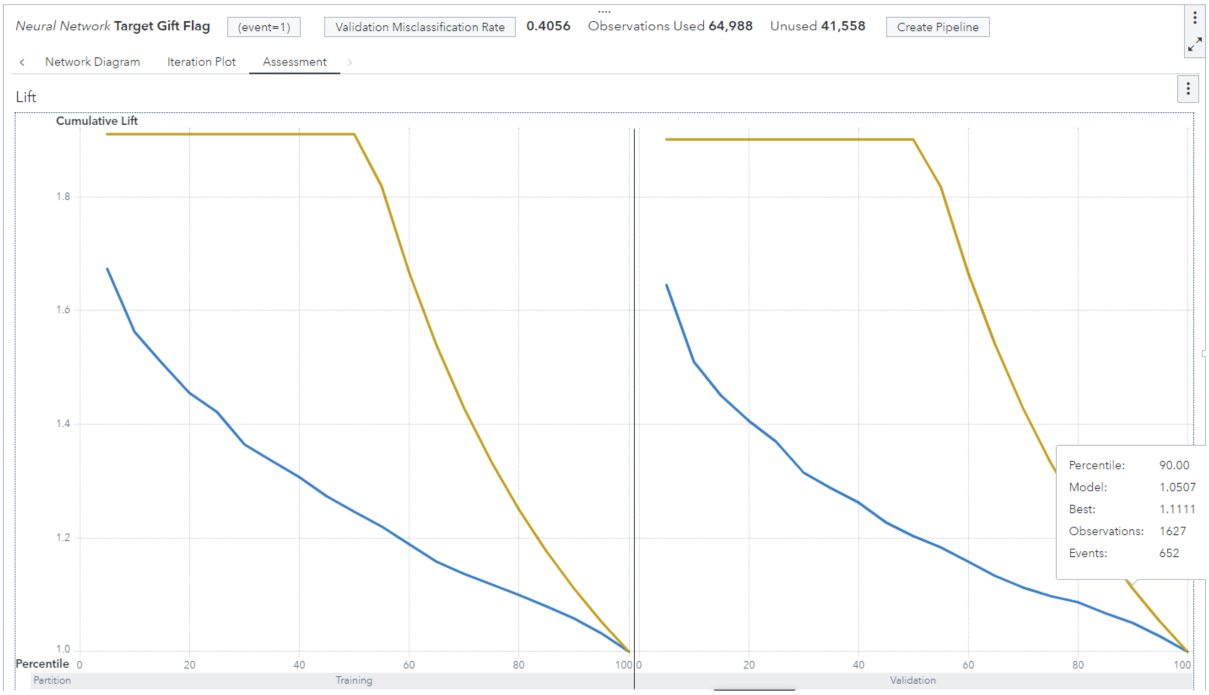
Confusion Matrix

Partition	Training	Training	Validation	Validation
Observed \ Predicted	0	1	0	1
0	7,813	7,646	7,642	7,776
1	4,573	12,420	4,766	12,352

Figure 9

J-4.

The top 10 percentile of the data used in the neural network algorithm contains 1.0507 times more responders in randomly ordered data and 1.1111 times responders in perfect ordered data as shown in figure 10.



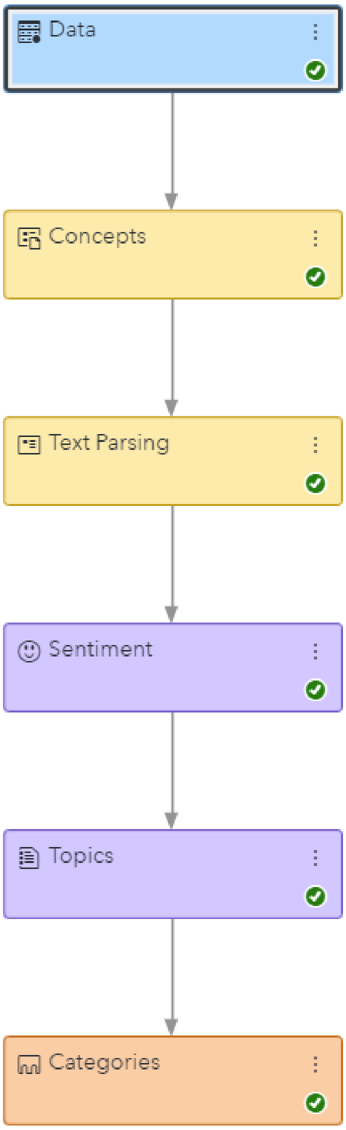
Neural Network **Target Gift Flag** (event=1) Validation Misclassification Rate **0.4056** Observations Used **64,988** Unused **41,558** Create Pipeline

Percentile: 90.00
Model: 1.0507
Best: 1.1111
Observations: 1627
Events: 652

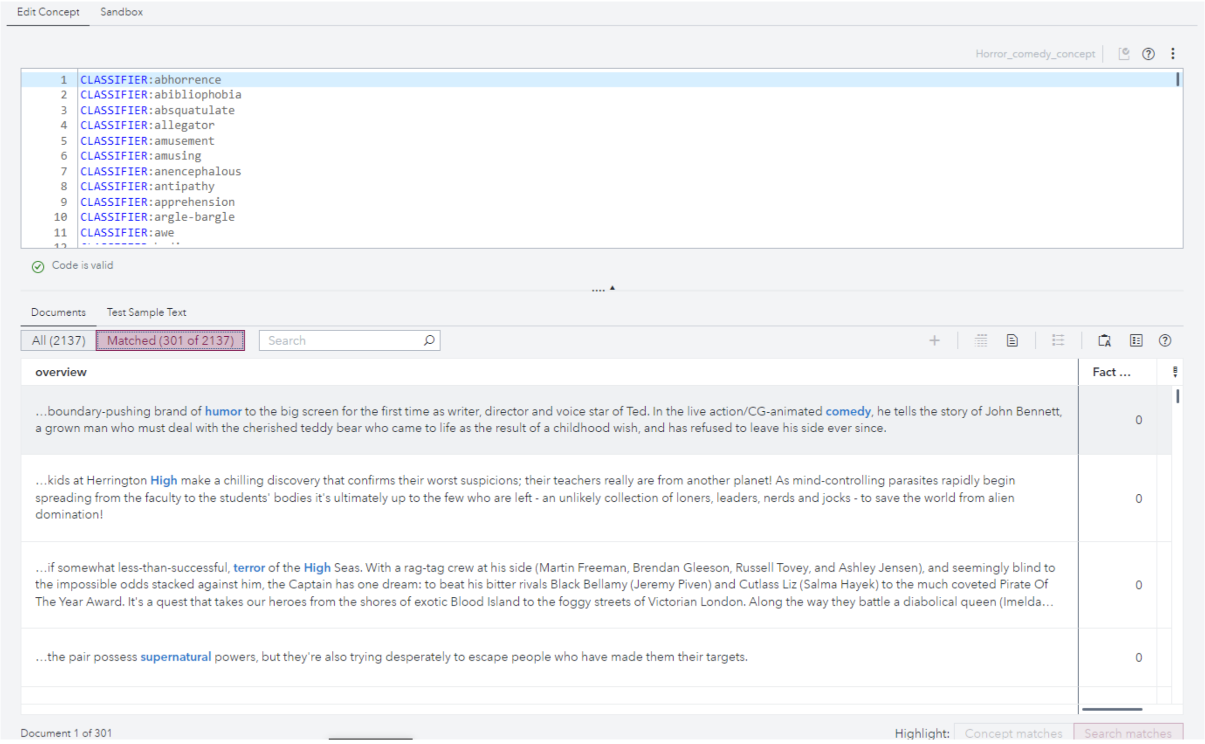
Figure 10

Exercise 2. Text Analytics

The text analytics was performed on ‘MOVIES_PLUS’ dataset. The ‘Overview’ variable was converted to text, ‘Title’ was made the key variable and ‘Made_Money’ was changed to category.



The predefined concepts were of little relevance to our analysis, so we make a new concept named ‘Horror_Comedy_Concept’ and rules are set before running the ‘concept’ node in the pipeline.

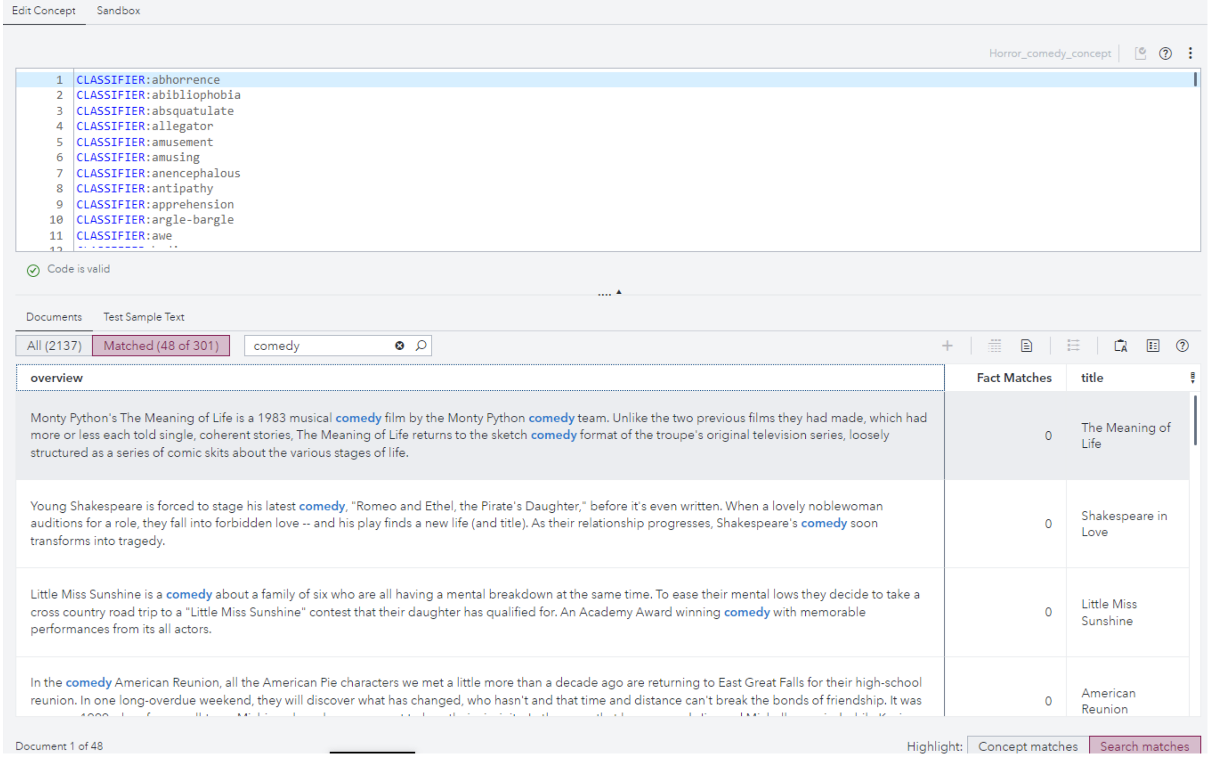


```
CLASSIFIER:abhorrence
CLASSIFIER:abibliophobia
CLASSIFIER:absquatulate
CLASSIFIER:allegator
CLASSIFIER:amusement
CLASSIFIER:amusing
CLASSIFIER:anencephalous
CLASSIFIER:antipathy
CLASSIFIER:apprehension
CLASSIFIER:argle-bargle
CLASSIFIER:awe
```

overview	Fact ...
…boundary-pushing brand of **humor** to the big screen for the first time as writer, director and voice star of Ted. In the live action/CG-animated **comedy**, he tells the story of John Bennett, a grown man who must deal with the cherished teddy bear who came to life as the result of a childhood wish, and has refused to leave his side ever since.	0
…kids at Herrington **High** make a chilling discovery that confirms their worst suspicions; their teachers really are from another planet! As mind-controlling parasites rapidly begin spreading from the faculty to the students' bodies it's ultimately up to the few who are left - an unlikely collection of loners, leaders, nerds and jocks - to save the world from alien domination!	0
…if somewhat less-than-successful, **terror** of the **High** Seas. With a rag-tag crew at his side (Martin Freeman, Brendan Gleeson, Russell Tovey, and Ashley Jensen), and seemingly blind to the impossible odds stacked against him, the Captain has one dream: to beat his bitter rivals Black Bellamy (Jeremy Piven) and Cutlass Liz (Salma Hayek) to the much coveted Pirate Of The Year Award. It's a quest that takes our heroes from the shores of exotic Blood Island to the foggy streets of Victorian London. Along the way they battle a diabolical queen (Imelda…	0
…the pair possess **supernatural** powers, but they're also trying desperately to escape people who have made them their targets.	0

Figure 11

We get 301 matches o 2137 observations for the custom concept is added and the 'concept' node is run in the pipeline. After searching 'comedy' the result narrows down to 48 matches.



```
CLASSIFIER:abhorrence
CLASSIFIER:abibliophobia
CLASSIFIER:absquatulate
CLASSIFIER:allegator
CLASSIFIER:amusement
CLASSIFIER:amusing
CLASSIFIER:anencephalous
CLASSIFIER:antipathy
CLASSIFIER:apprehension
CLASSIFIER:argle-bargle
CLASSIFIER:awe
```

overview	Fact Matches	title
Monty Python's The Meaning of Life is a 1983 musical **comedy** film by the Monty Python **comedy** team. Unlike the two previous films they had made, which had more or less each told single, coherent stories, The Meaning of Life returns to the sketch **comedy** format of the troupe's original television series, loosely structured as a series of comic skits about the various stages of life.	0	The Meaning of Life
Young Shakespeare is forced to stage his latest **comedy**, "Romeo and Ethel, the Pirate's Daughter," before it's even written. When a lovely noblewoman auditions for a role, they fall into forbidden love -- and his play finds a new life (and title). As their relationship progresses, Shakespeare's **comedy** soon transforms into tragedy.	0	Shakespeare in Love
Little Miss Sunshine is a **comedy** about a family of six who are all having a mental breakdown at the same time. To ease their mental lows they decide to take a cross country road trip to a "Little Miss Sunshine" contest that their daughter has qualified for. An Academy Award winning **comedy** with memorable performances from its all actors.	0	Little Miss Sunshine
In the **comedy** American Reunion, all the American Pie characters we met a little more than a decade ago are returning to East Great Falls for their high-school reunion. In one long-overdue weekend, they will discover what has changed, who hasn't and that time and distance can't break the bonds of friendship. It was	0	American Reunion

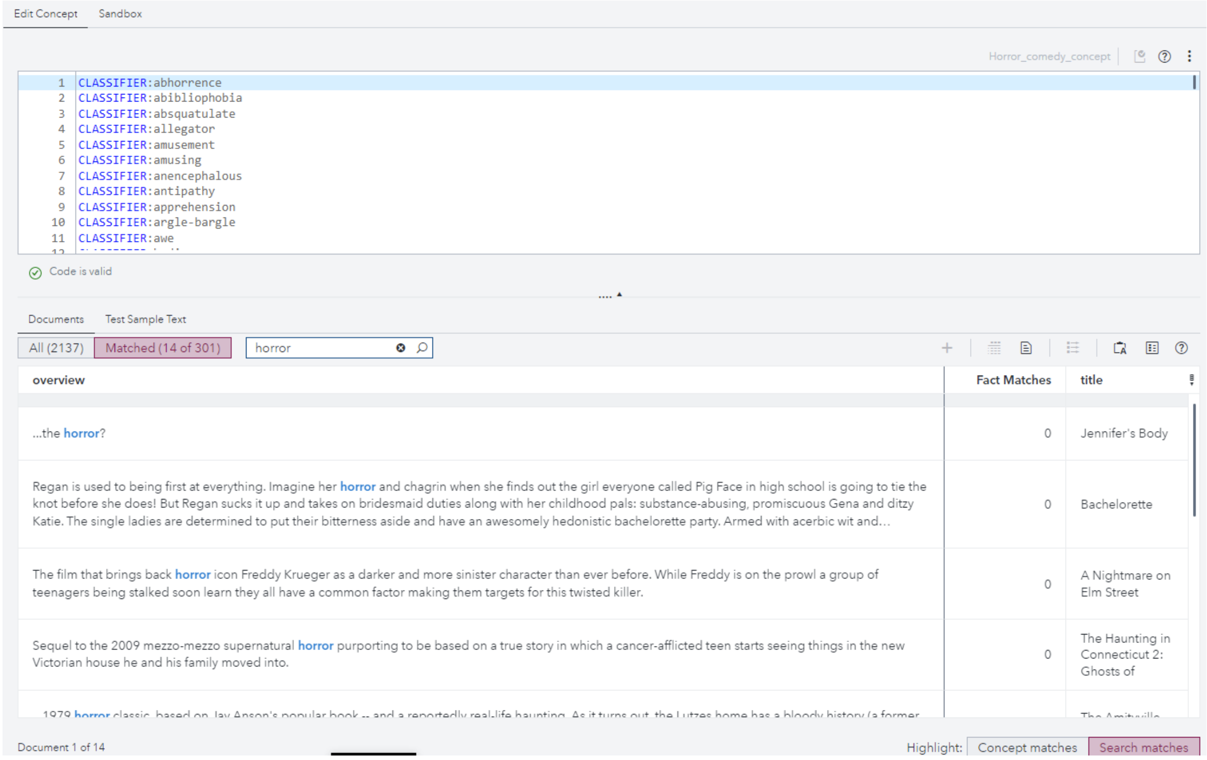
Figure 12

The 3 comedy movies are as follows:

1. The Meaning of Life
2. Little Miss Sunshine
3. American Reunion

Similarly, when we search for horror in the search box, we get 14 matches. The three horror movies are as follows:

1. Texas Chainsaw 3D
2. A Nightmare on Elm Street
3. The Haunting in Connecticut 2: Ghosts of Georgia



```
CLASSIFIER:abhorrence
CLASSIFIER:abibliophobia
CLASSIFIER:absquatulate
CLASSIFIER:allegator
CLASSIFIER:amusement
CLASSIFIER:amusing
CLASSIFIER:anencephalous
CLASSIFIER:antipathy
CLASSIFIER:apprehension
CLASSIFIER:argle-bargle
CLASSIFIER:awe
```

overview	Fact Matches	title
...the horror?	0	Jennifer's Body
Regan is used to being first at everything. Imagine her horror and chagrin when she finds out the girl everyone called Pig Face in high school is going to tie the knot before she does! But Regan sucks it up and takes on bridesmaid duties along with her childhood pals: substance-abusing, promiscuous Gena and ditzy Katie. The single ladies are determined to put their bitterness aside and have an awesomely hedonistic bachelorette party. Armed with acerbic wit and...	0	Bachelorette
The film that brings back horror icon Freddy Krueger as a darker and more sinister character than ever before. While Freddy is on the prowl a group of teenagers being stalked soon learn they all have a common factor making them targets for this twisted killer.	0	A Nightmare on Elm Street
Sequel to the 2009 mezzo-mezzo supernatural horror purporting to be based on a true story in which a cancer-afflicted teen starts seeing things in the new Victorian house he and his family moved into.	0	The Haunting in Connecticut 2: Ghosts of
...1979 horror classic, based on Jay Anson's popular book -- and a reportedly real-life haunting. As it turns out, the Lutzes home has a bloody history (a former		The Amityville

Figure 13

Part 2: Report

Introduction

In the modern age every industry is adopting modern technological advances to stay competitive, reduce administrative loads and automate tasks. These industries have been collecting data for a few decades now. The amount of data generated till date is too large in volume and complex to be analysed by a human. Moreover, there are hidden patterns and insights in the data that can only be found by carrying out long calculations on data. With the advent of industry 4.0 revolution, organisations want to use this raw data and with the modern analytics technologies gain useful insights off it. Every organisation is continuously putting resources in research and development to implement these changes as fast as possible.

To choose the right technology multiple criteria are required. These criteria are: Functionality, compatibility, user-friendliness, reliability, performance, robustness, setup costs and licensing and maintenance cost (Belinda et al. 2021, p. 165). This method of evaluation is called Analytic Hierarchy Process (AHP). This is a generic multi-criteria problem-solving approach that is used to make complex decisions based on characteristic that are not quantifiable. It is a hierarchical structure with objective at the top, criteria in the middle and decision alternatives in the bottom. It is a qualitative methodology to take an unstructured problem and change it to a decision hierarchy (Sureshchandar & Liesten 2006, p. 22).

We need to first define the criteria in order to evaluate a technology.

1. Functionality: It is defined as the ability of the software to perform the tasks that it is intended for.
2. Compatibility: It is the ease with which the software interacts with other existing softwares applications.
3. User-friendliness: It is the capability of the software to enable user to operate and understand the usability for tasks and conditions for use.
4. Reliability: It refers to the probability of software operating in the given time frame without incurring breakdowns.
5. Performance: It refers to the total effectiveness of the software.

6. Robustness: It refers to the ability to deal with any form of error that may occur during the period of operation.
7. Setup Costs and licensing: Price of the software and other infrastructure required, if any, to deploy the software.
8. Maintenance cost: The ability to modify to correct any faults or improve performance.

Artificial Neural Network (ANN)

ANN, also known as Simulated Neural Network (SNN) is a subset of machine learning used primarily for creating deep learning models. Its idea is based on how the human brain works. ANN is composed of node layers for an input layer, one or more hidden layers and an output node layer. Nodes are essentially mimicking neurons. All nodes are connected to nodes in the next layer and each node has its own weight and threshold. If the output of any node is above the threshold the signal is passed to the node in the next layer. A visual representation of neural network is shown in figure 1.

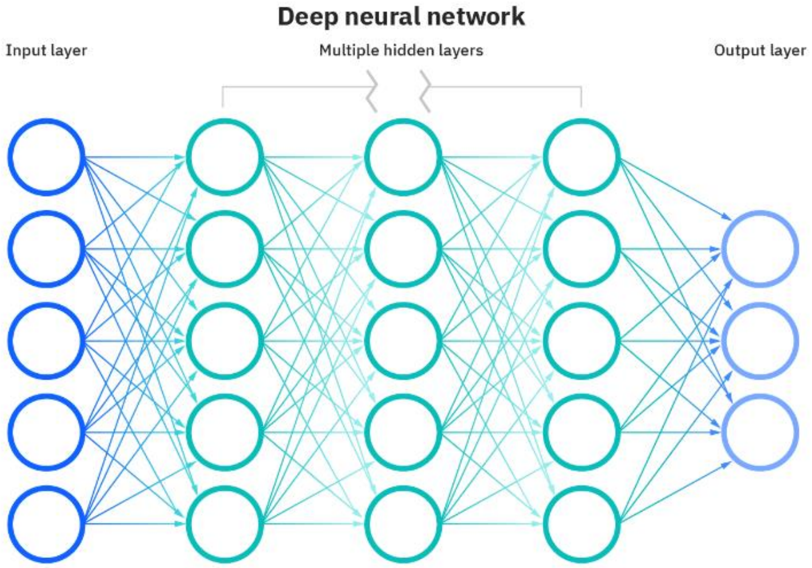


Figure 14 Neural Network Structure

As mentioned earlier, neural networks are modelled based on how human brain learns i.e., way that its accuracy increases over time by training on new data and making itself better over time. Once the weights and thresholds are fine tuned for our use case ANN is a powerful tool for artificial intelligence and computer science. It can be used for clustering and classify data at a high speed.

In ANN each node can be considered a linear regression model which passes a binary value forward when activated. The weight of each node decides the importance of the variable and create bias for model. When the input is received it is multiplied by the weight of the node and if the output is greater than the threshold the node is fired, and the output is passed to another node in the next layer. This mechanism of going from one layer to next is called feedforward neural network. In cases where the output is sent back to calculate the error and reassign weight and threshold values, it is called backpropagation neural network (IBM Cloud Education 2020).

Advantages of Using Artificial Neural network

Over the years ANN has established itself as a default model for deep learning because of its ability to learn any non-linear function and also ability to map any and every interaction between the independent variables.

ANN is preferred over other technology because of the following reason (Mijwil 2018):

- Storing information on the entire network: Unlike traditional programming, all of the information is stored on the network instead of a database, making the network run equally efficiently even if some pieces of information is missing at one place.
- Ability to work with incomplete information: After training the model, even with incomplete data the model can produce output. Although the reduction in performance depends on the importance of the missing value but unlike traditional models the output is generated.
- Having fault tolerance: Malfunction of one or more node cells does not prevent the network from generating output. This reduces performance of the network but makes the network fault resistant.
- Having a distributed memory: Like every machine learning model, ANN needs data to be trained on which includes all desired outputs in order to train the network with the examples of each case. The network's performance is directly affected by the this. The network cannot output correct results if the required event is not given in the training data from all aspects and thus produce false output.

- Gradual Corruption: This states that the network slows down over time and undergoes gradual degradation. The network problem however does not corrode immediately.
- Ability to make Machine Learning: The network is trained and then generates output on similar events.
- Parallel Processing Capability: This is where ANN is superior to other machine learning algorithms. A well train ANN has enough performance to perform more than one task at a time.

Disadvantages of Artificial Neural Network

Even though ANN has a lot of merits to use it over other machine learning algorithms it has its limitations (Mijwil 2018).

- Hardware Dependencies: ANN requires a lot of parallel processing power, depending on the complexity of the network, hence it is highly dependent on the hardware used.
- Unexplained behaviour of the network: Like many other machine learning models when ANN gives an output there is no reasoning or how the model reached to the output. This makes the model less trustworthy.
- Determination of proper network structure: The structure of ANN is achieved by trial-and-error method. This makes is difficult to come up with a structure for different problems.
- Difficulty of showing the problem to the network: ANN requires numerical data to train. How the information is translated to numerical values determines how well the model will perform.

Uses of Artificial Neural Network

Every industry is trying incorporate ANN in everyday activities to reduce workload of professionals and analysts. Some of the industries that have successfully implemented ANN are:

- Manufacturing industry: Manufacturing companies are using ANN to automate defect detection, predictive maintenance, optimize supply chain and forecast energy needs.

- Banking industry: They are using ANN for fraud detection, conducting credit checks, and automating financial advising services.
- Retail industry: Retail companies uses ANN for chatbots, analyse and enhance customer intelligence.
- Public Sector: Public Sector uses ANN to power smart cities, improve facial recognition and security intelligence.

Evaluating Artificial Neural Network on Criteria for Data Security and Privacy

Primary problem of using digital technologies and storing data on databases is the risk of breach in security and threat to people's privacy. In case of a cyber-attack the sensitive information of company and details of employees and customers could be leaked. A single breach in the company network can cause a havoc in the organisation so it is of utmost importance that the organisation takes strict measures to protect their assets.

Barni, Orlandi and Piya (2006, p. 146) discusses about the two major points of information leak in a neural network and how it needs to be protected. First is protection of data while inputting it in the neural network and after getting the output. Second is protecting the network itself as the owner might not want to disclose knowledge embedded in it. The first security concern can be handled by encrypting the data before feeding it to the neural network. They came up framework which helps process data and signals directly in an encrypted domain on which secure and privacy preserving protocols can be built upon. The developed protocol aims to protect both the input data as well as the information embedded in the neural network.

Due to the nature of ANN in an organisation where all processes and data is on a cloud or offsite server the threat of security breach and privacy concerns arise even further as everyone who can access the cloud has the access to the confidential information. Encrypting the data being the first choice does hinder the efficiency of the neural network. Melissourgos et al. (2021) has addressed this issue by working with concept of matrix masking. Using matrix masking the user is able to send the masked data to cloud and train the neural network on this masked data without affecting the performance time and efficiency by a significant level.

Schlitter (2008) suggests that security and privacy protection can be achieved in a different way. He proposes a privacy protocol for neural networks using secure sum and secure matrix addition protocol to iteratively include the weights of the participating nodes global neural network weight matrix. He states that by doing this the security and privacy is not achieved in the strict sense of the words, but the disclosed information is not associated to a single party which is sufficiently secure for practical applications.

In a recent study conducted by Shi and Li (2022) they state that security and privacy protection are lagging behind compared to the advancement in the neural network technology every day. The study focuses on security and privacy concerns related to wireless sensor. They constructed a wireless sensor network intrusion detection system which was based on particle swarm optimization algorithm. It includes modules for data extraction to decision making at different levels of sensor network privacy protection. The results show that the privacy protection and the detection system have practical value and can be implemented in the industry.

From the above studies we can see that research is still being conducted in to make neural networks more secure and protect privacy of customers and employees, but no universal solution has been found so far. Different approaches have been taken by the researchers to come up with a viable solution to the security and privacy and protection issue. The solutions and protocols developed are limited by either their scope or field and are not being applied across the industry. This could be tackled by establishing certain industry standards for the security and privacy protection in order to take major actions for the safety of their assets.

References

- Abidi, MH, Mohammed, MK & Alkhalefah, H 2022, ‘Predictive Maintenance Planning for Industry 4.0 Using Machine Learning for Sustainable Manufacturing’, *Sustainability (Basel, Switzerland),* vol. 14, no. 6, p. 3387–.
- Barni, M., Orlandi, C. and Piva, A., 2006, September. A privacy-preserving protocol for neural-network-based computation. In Proceedings of the 8th workshop on Multimedia and security (pp. 146-151).
- Belinda, BI, Akintoba, A, Solomon, N & Boniface, A 2021, ‘Evaluating Software Quality Attributes using Analytic Hierarchy Process (AHP)’, *International journal of advanced computer science & applications*, vol. 12, no. 3, pp. 165-173.

- IBM Cloud Education 2020, “Neural Networks”, Blog, 17 August 2020, viewed on 6 August 2022, < https://www.ibm.com/cloud/learn/neural-networks>
- Melissourgos, D., Gao, H., Ma, C., Chen, S. and Wu, S.S., 2021, November. On Outsourcing Artificial Neural Network Learning of Privacy-Sensitive Medical Data to the Cloud. In 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI) pp. 381-385. IEEE.
- Mijwil, MM 2018, “Artificial Neural Networks Advantages and Disadvantages”, Blog, 2018 January 2018, viewed on 6 Ausgust 2022, < https://www.linkedin.com/pulse/artificial-neural-networks-advantages-disadvantages-maad-m-mijwel/>
- Schlitter, N., 2008. A protocol for privacy preserving neural network learning on horizontal partitioned data. PSD.
- Shi, L. and Li, K., 2022, “Privacy Protection and Intrusion Detection System of Wireless Sensor Network Based on Artificial Neural Network”, *Computational Intelligence and Neuroscience*, 2022.
- Sureshchandar, G. & Leisten, R 2006, ‘A framework for evaluating the criticality of software metrics: an analytic hierarchy process (AHP) approach’, *Measuring business excellence*, vol. 10, no. 4, pp. 22–33.