**Individual Project 1: Fraud Detection**

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

With an unlimited and rising number of ways someone can commit fraud, detection can be difficult to accomplish. Activities such as reorganization, downsizing, moving to new information systems or encountering a cybersecurity breach could weaken an organization's ability to detect fraud. This means techniques such as real-time monitoring for frauds is recommended. Organizations should look for fraud in financial transactions, location, devices used, initiated sessions and authentication system

**Contents**

[**1.** **Use Regression, Decision Tree and Neural Net on the “Fraud” data.** 2](#_Toc51883382)

[**2.** **Analyse the results including which model perform the best.** 5](#_Toc51883383)

[**2.1.** **Detailed Classification Report:** 6](#_Toc51883384)

[**2.2.** **Analysis:** 7](#_Toc51883385)

[**3.** **Analysis of Models** 8](#_Toc51883386)

[**3.1.** **Logistic Regression** 8](#_Toc51883387)

[**3.2.** **Decision Tree** 9](#_Toc51883388)

[**3.3.** **Neural Network** 11](#_Toc51883389)

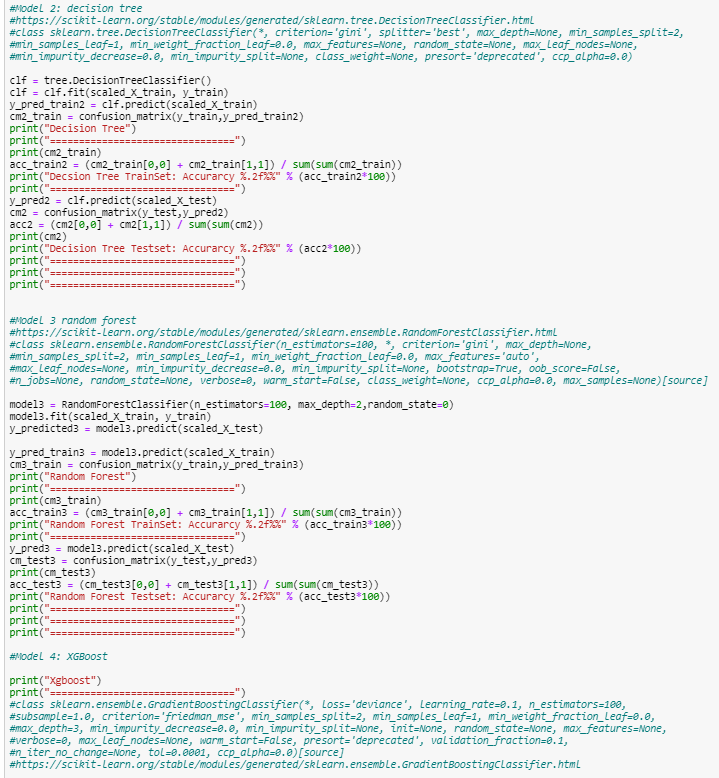
[**4.** **Suggest ways to improve the results** 11](#_Toc51883390)

[**5.** **Please analyse the business aspect of fraud detection** 11](#_Toc51883391)

[**6.** **References:** 11](#_Toc51883392)

# **Use Regression, Decision Tree and Neural Net on the “Fraud” data.**







Refer to .txt file for result data.

# **Analyse the results including which model perform the best.**

All 5 models (Logistic Regression, decision tree, random forest, XGBoost and neural network) all performed well generally with test set accuracy of => 80%.

|  |  |
| --- | --- |
| **Model** | **Test set Accuracy** |
| **Logistic Regression** | 85.50% |
| **Decision Tree** | 95.50% |
| **Random Forest** | 87.50% |
| **XGBoost** | 98.50% |
| **Neural Network** | 80.00% |

## **Detailed Classification Report:**

|  |  |
| --- | --- |
| **Model** | **Test set Accuracy** |
| **Logistic Regression** |  |
| **Decision Tree** |  |
| **Random Forest** |  |
| **XGBoost** |  |
| **Neural Network** |  |

# **Analysis:**

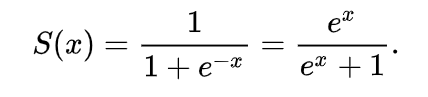
XGBoost and decision tree model performed the best, while neural network model performed the worst among the 5. With reference to the table below, the logistic regression, decision tree, XGBoost and Neural Network models are more accurate in predicting positive output, while Random Forest model is more accurate in predicting negative output.

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy** | |
| **Positive** | **Negative** |
| **Logistic Regression** | 86.3% | 82% |
| **Decision Tree** | 100% | 85.7% |
| **Random Forest** | 85.4% | 100% |
| **XGBoost** | 100% | 94.7% |
| **Neural Network** | 81.9% | 70.6% |

# **Analysis of Models**

# **Logistic Regression**

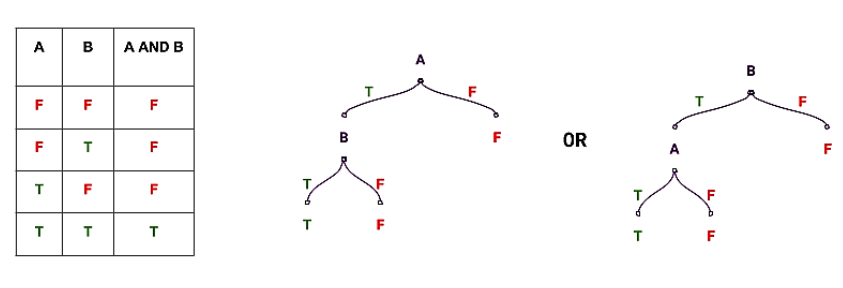
Logistic regression is a statistical model used to predict the probability of a certain class or event happening, in this case whether a fraudulent transaction is determined based on several input variables. Logistic regression models have a certain fixed number of parameters that depend on the number of input features and output categorical prediction. These models model the data using the sigmoid function, as shown in the equation below.



For this fraud detection problem, binary logistic regression is being used to predict the chances of a transaction being fraudulent, and the high accuracy numbers can be attributed to the suitability of the model for this scenario. Binary logistic regression requires the dependent variable to be binary, which is 1/0 for this case where 1 represents a fraud detection. Furthermore, the independent variables are required to be independent of each other and that the model has little or no multicollinearity.

# **Decision Tree**

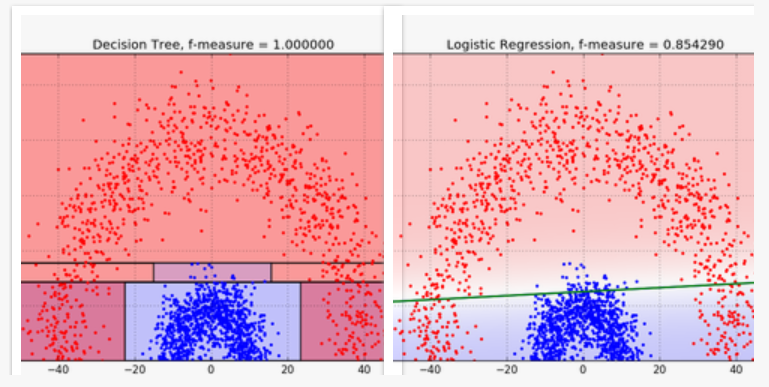
Decision Trees are a type of supervised learning method used for both classification and regression tasks, with the aim of creating a model that predicts the outcome of a target variable, in this case whether or not a transaction is fraudulent, by learning simple decision rules from features obtained from data.



The diagram above illustrates how decision trees can input any Boolean function of input attributes, and shows how decision trees classify problems by sorting them down the tree from root to the respective leaf nodes.

The advantages of decision trees models are that they can handle both numerical and categorical data, their excellent performance on large datasets and their general speed.

With reference to results in section 2, we realise that decision trees models generally outperform logistic regression models with respect to this classification problem. To account for the relatively distinct gap in performance, we have to first the fundamental differences between the two models.

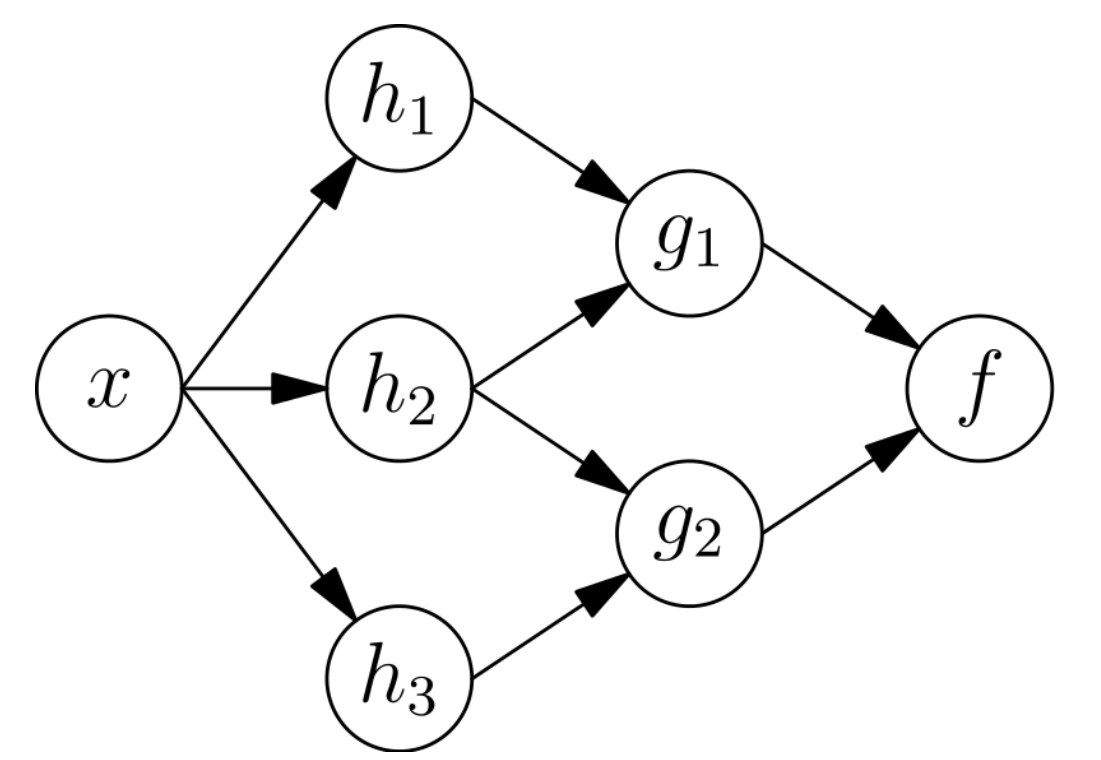


From the above diagram, we can see how differently the two models create decision boundaries. Logistic Regression (right) assumes that the data is linearly or curvy linearly separable in space and uses a single line to segregate the data into two regions. However, Decision Trees (left) are non-linear classifiers and do not require data to be linearly separable. Thus, we can assume that the two classes in this case (fraud and non-fraud) are separated by a decidedly non-linear boundary, and that decision trees can better capture the division and segregation, leading to superior classification performance and higher accuracy figures.

XGBoost stands for extreme gradient boosting; it is an implementation of gradient boosted decision trees designed for speed and performance. It is an extremely popular and successful model and is the go-to algorithm for competition winners for predictive modelling problems. The reason for XGBoost’s success and popularity is the fact that the model is an ensemble technique where new models are added to correct errors made by existing models. Models are added sequentially until no further improvements can be made. It uses a gradient descent algorithm to minimise the loss when adding new models. Thus, the model is more accurate compared to single decision tree models.

# **Neural Network**

A neural network is a circuit of artificial neurons and nodes. The connections of neurons are modelled as weights, with all inputs being modified by a weight and summed. An activation function then controls the amplitude the output. The figure below is a simple illustration of a neural network.



The strengths of neural networks include their ability to deal with extremely complex and high dimensionality problems. However, we see in this case that the performance of neural network is actually the worst among the 5. This may be due to the tuning of the model not being the most optimised. We also know that neural network models are good at handle binary data, while being unable to handle categorical values (which are present in our data set in this case). This explains why models like decision trees and logistic regression have performed better.

# **Suggest ways to improve the results**

For decision tree models, the data can be easily overfitted. Thus, we can prevent this issue by pruning the model by growing superfluous branches or removing them after the tree is grown. Furthermore, for the neural network implementation, we consider increasing the number of epochs, and increase more layers.

# **Please analyse the business aspect of fraud detection**

Fraud detection is an extremely fundamental and critical department in financial institutions and a single case being undetected can lead to large sums of losses for organisations. It is worth nothing that it is more costly to detect a false negative than to detect a false positive.

# **References:**

* <https://blog.bigml.com/2016/09/28/logistic-regression-versus-decision-trees/#:~:text=Decision%20Trees%20bisect%20the%20space,the%20space%20exactly%20into%20two.&text=A%20single%20linear%20boundary%20can%20sometimes%20be%20limiting%20for%20Logistic%20Regression>.
* <https://www.geeksforgeeks.org/understanding-logistic-regression/>
* <https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/ml-decision-tree/tutorial/#:~:text=Decision%20Trees%20are%20a%20non,inferred%20from%20the%20data%20features>.
* <https://towardsdatascience.com/classification-using-neural-networks-b8e98f3a904f#:~:text=Neural%20networks%20are%20complex%20models,passing%20outputs%20to%20further%20layers>.