FRAUD DETECTION

Submitted To:
Ms. Sonia
Department of IT, DTU

Submitted By:
Aaryaman Bajaj (2K17/IT/02)
Abhiyudhya Pratap (2K17/IT/10)
Himanshi Nimesh (2K17/IT/52)

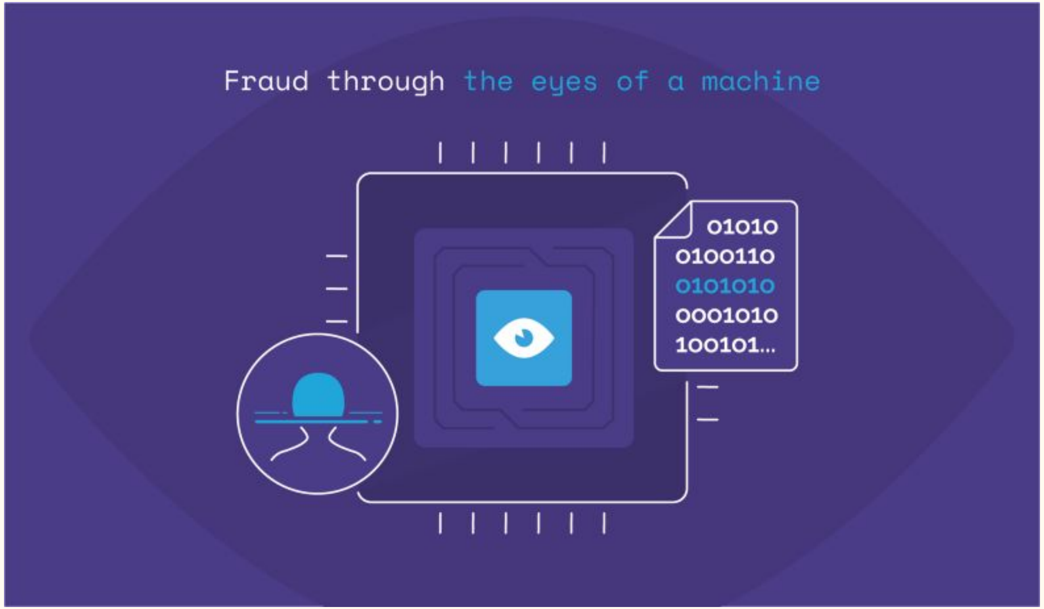
INTRODUCTION

Fraud detection is a set of activities undertaken to prevent money or property from being obtained through false pretenses. 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 payout.

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 systems.

Fraud Detection using Machine Learning

The idea is that there are certain characteristics of fraudulent transactions that differentiate them from legitimate ones. Machine Learning algorithms recognize patterns in the data that allow them to discern fraudsters from legitimate clients, based on thousands of pieces of information, that sometimes may seem completely unrelated to a human being. The algorithm is searching for patterns in fraudsters' behaviour, their hardware characteristics etc.



About the Dataset

- This dataset is obtained from Kaggle: [in reference 1]
- A major limitation for the validation of existing fraud detection methods is the lack of public datasets.
- The datasets contains transactions made by credit cards in September 2013 by European cardholders.
- This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions.
- The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.
- It contains only numeric input variables which are the result of a PCA transformation.
- Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset.
- The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning.
- Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Methods Used

We have applied 3 machine learning techniques:

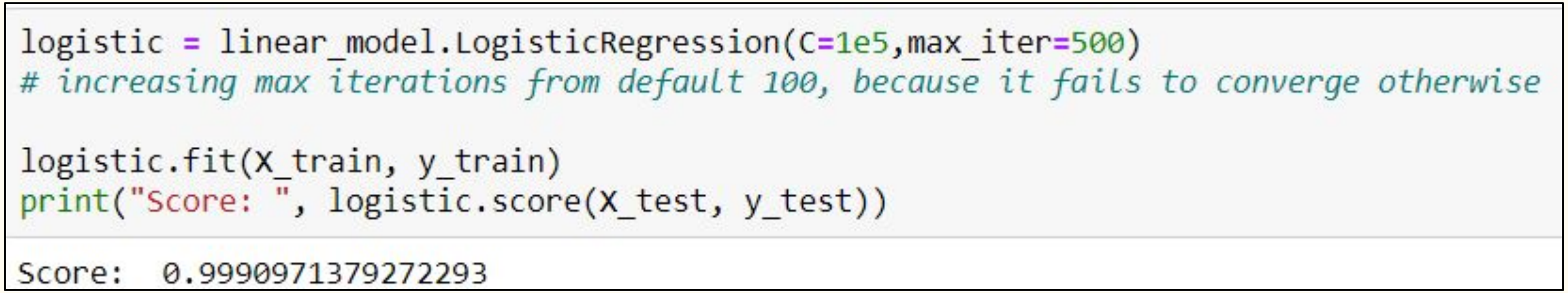
- Logistic regression
- K-means Clustering
- Neural Networks

Other algorithms such as decision trees, random forests, and support vector machines (SVMs) could also similarly be applied.

Logistic Regression

Logistic regression is executed when the dependent variable is binary. It is a predictive analysis technique. It describes data and explains the relationship between one dependent binary variable and one or more independent variables. The model is usually easy to interpret, and we can know which feature is important for us. We need to scale our data before applying this model to it.

Below is the code that we have implemented in our project:



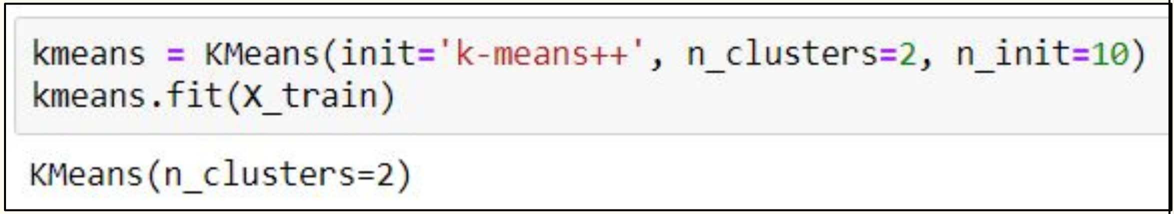
```
logistic = linear_model.LogisticRegression(C=1e5,max_iter=500)
# increasing max iterations from default 100, because it fails to converge otherwise

logistic.fit(X_train, y_train)
print("Score: ", logistic.score(X_test, y_test))
```

```
Score:  0.9990971379272293
```

K-Means Clustering

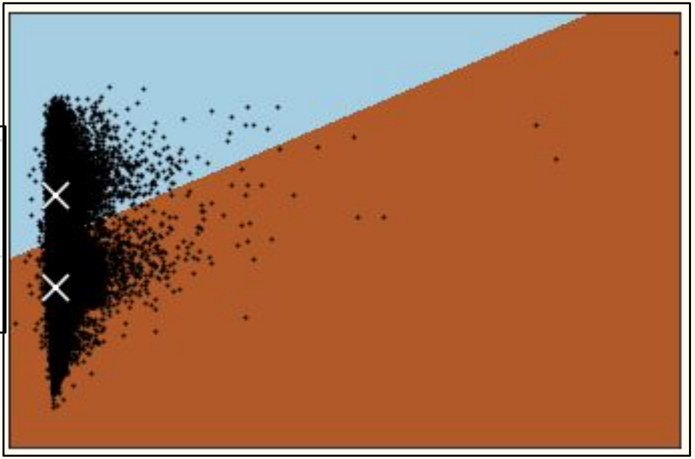
K-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.

Below is the code used in our project, with 2 clusters:
(Along with the output on the right)



```
kmeans = KMeans(init='k-means++', n_clusters=2, n_init=10)
kmeans.fit(X_train)
```

```
KMeans(n_clusters=2)
```



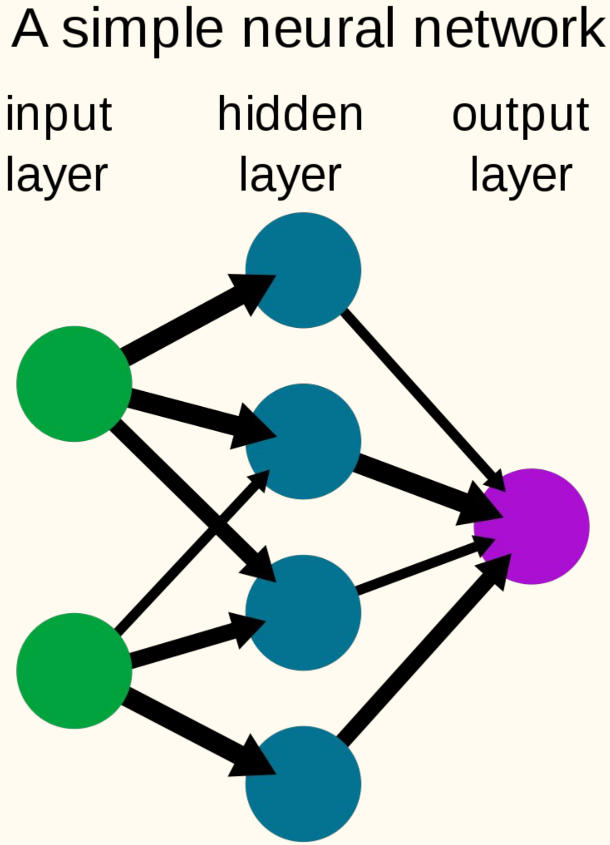
Neural Networks

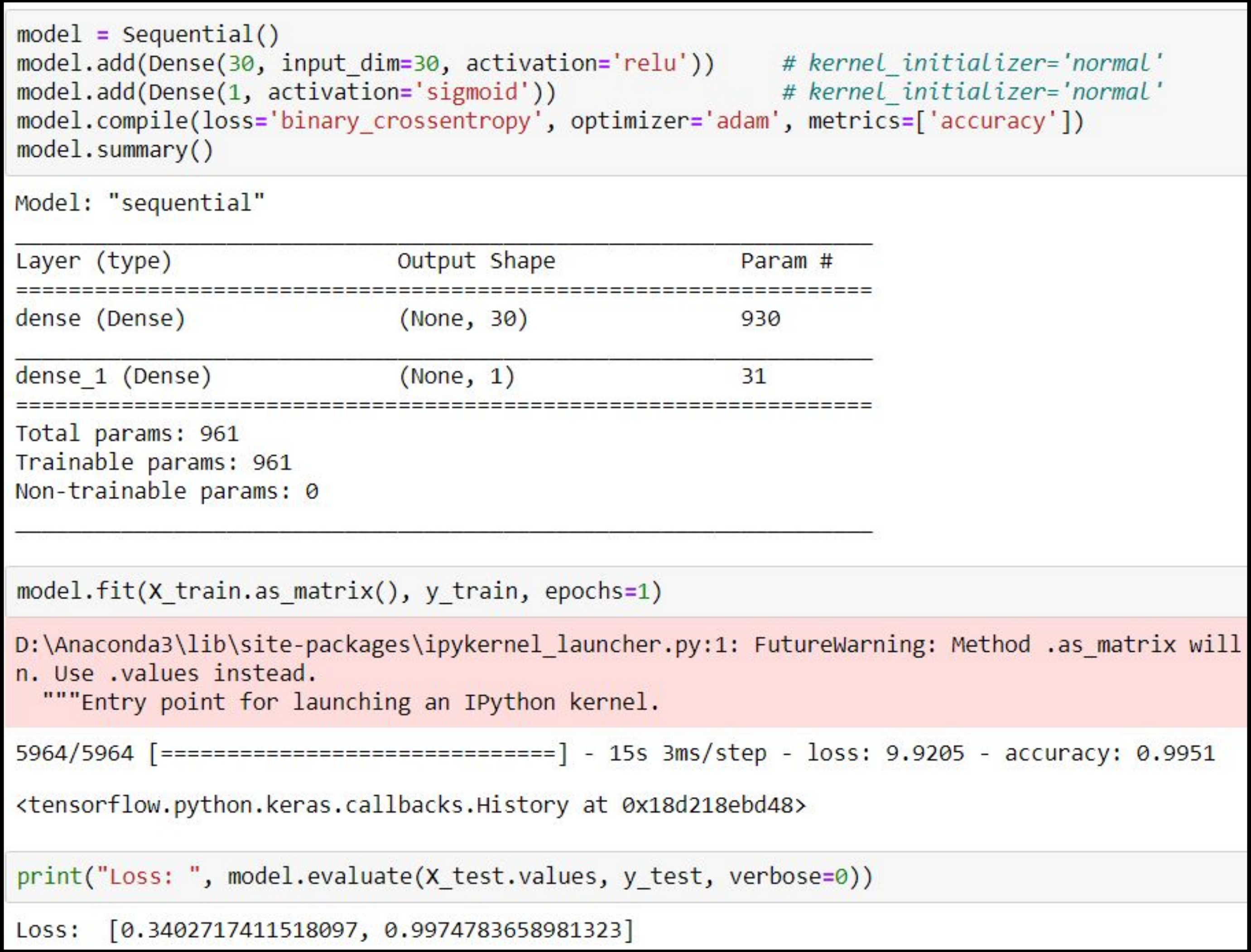
A neural network is a network or circuit of neurons composed of artificial neurons or nodes. It is used for solving artificial intelligence problems.

The connections of the biological neuron are modeled as weights. A positive weight reflects an excitatory connection, while negative values mean inhibitory connections.

All inputs are modified by a weight and summed.

The Code used in our project is given in the next slide.





```
model = Sequential()
model.add(Dense(30, input_dim=30, activation='relu'))     # kernel_initializer='normal'
model.add(Dense(1, activation='sigmoid'))                  # kernel_initializer='normal'
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
model.summary()
```

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
dense (Dense)                (None, 30)                930       
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 31        
=================================================================
Total params: 961
Trainable params: 961
Non-trainable params: 0
_________________________________________________________________
```

```
model.fit(X_train.as_matrix(), y_train, epochs=1)
```

```
D:\Anaconda3\lib\site-packages\ipykernel_launcher.py:1: FutureWarning: Method .as_matrix will
n. Use .values instead.
  """Entry point for launching an IPython kernel.
```

```
5964/5964 [==============================] - 15s 3ms/step - loss: 9.9205 - accuracy: 0.9951

<tensorflow.python.keras.callbacks.History at 0x18d218ebd48>
```

```
print("Loss: ", model.evaluate(X_test.values, y_test, verbose=0))
```

```
Loss:  [0.3402717411518097, 0.9974783658981323]
```

Comparison of Models

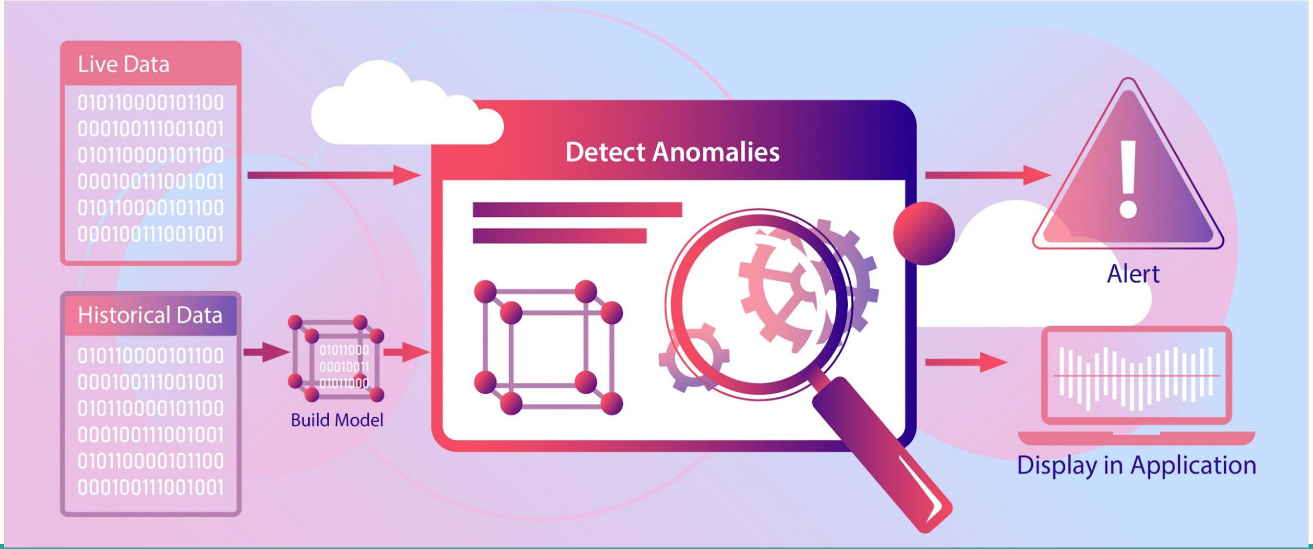
Logistic regression outperformed both the K-means and neural network. We believe that it is because of how the decision boundary changed with the class weights features.

The neural network was next, and K-means performed the poorest. We believe this is due to the fact that clustering relies entirely on the similarities and differences of features of the dataset. Since fraud transactions can look very similar to regular transactions, it is difficult to put them into a separate group based on features alone.

APPLICATIONS

❏ Credit card fraud detection is one of the most explored domains of fraud detection and relies on the automatic analysis of recorded transactions to detect fraudulent behaviour.

❏ Fraud detection has become a vital activity in order to decrease the impact of fraudulent transactions on service delivery, costs, and reputation of the company.



CONCLUSION

In recent years, the development of new technologies like the internet has provided further ways in which fraudsters can commit fraud. Fraud is a very skilled crime; therefore a special method of intelligent data analysis to detect and prevent it is necessary.

To prevent the increasingly numerous frauds spawned by the information age, management must know its vulnerabilities and be able to mitigate risk in a cost - effective manner.

Fraud detection is a complex issue that requires a substantial amount of planning before throwing machine learning algorithms at it. Nonetheless, it is also an application of data science and machine learning for the good, which makes sure that the customer’s money is safe and not easily tampered with.

REFERENCES

- https://web.archive.org/web/20171105235151/https://www.kaggle.com/dalpozz/creditcardfraud
- https://searchsecurity.techtarget.com/definition/fraud-detection
- https://www.kaggle.com/mlg-ulb/creditcardfraud
- https://github.com/georgymh/ml-fraud-detection
- https://towardsdatascience.com/tagged/fraud-detection
- https://towardsdatascience.com/detecting-credit-card-fraud-using-machine-learning-a3d83423d3b8
- https://medium.com/@Nethone_/a-beginners-guide-to-machine-learning-in-payment-fraud-detection-prevention-360c95a9ca54
- https://www.hindawi.com/journals/tswj/2014/252797/
- https://en.wikipedia.org/wiki/Data_analysis_techniques_for_fraud_detection

THANK YOU!