PHASE – 2 NM PROJECT

PROJECT TITLE: CREDIT CARD FRAUD DETECTION.

DATA SCIENCE:

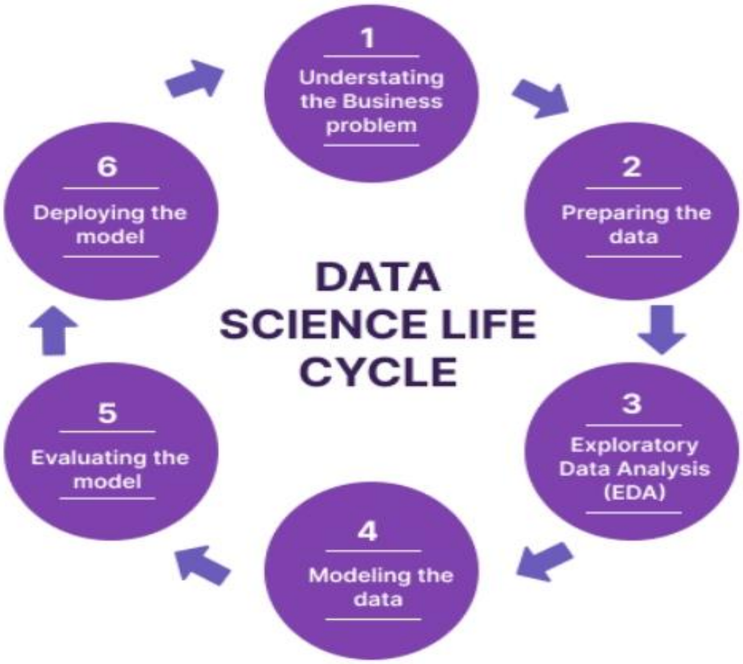
Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. Data Science is not only a synthetic concept to unify statistics, data analysis and their related methods but also comprises its results. It includes three phases, design for data, collection of data, and analysis on data.

The data science life cycle is a complex and iterative process that involves six phases: problem identification, data collection, data preparation; data modeling and analysis, model evaluation, and deployment.

Data Science is about data gathering, analysis and decision-making.Data Science is about finding patterns in data, through analysis, and make future predictions.

By using Data Science, companies are able to make:

1. Better decisions (should we choose A or B)
2. Predictive analysis (what will happen next?)
3. Pattern discoveries (find pattern, or maybe hidden information in the data)



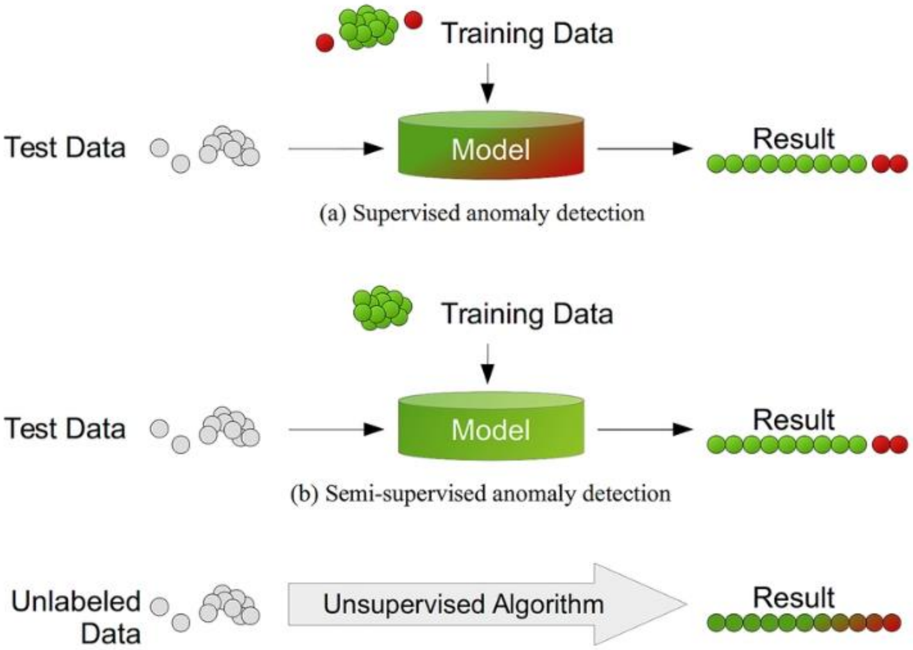
ANAMOLY DETECTION ALGORITHM:

Anomalies are data points that stand out amongst other data points in the dataset and do not confirm the normal behavior in the data. These data points or observations deviate from the dataset's normal behavioral patterns.

Anomaly detection is an unsupervised data processing technique to detect anomalies from the dataset. An anomaly can be broadly classified into different categories:

1. Outliers: Short/small anomalous patterns that appear in a non-systematic way in data collection.
2. Change in Events: Systematic or sudden change from the previous normal behavior.
3. Drifts: Slow, undirectional, long-term change in the data.

Anomalies detection are very useful to detect fraudulent transactions, disease detection, or handle any case studies with high-class imbalance. Anomalies detection techniques can be used to build more robust data science models.



(a) Supervised anomaly detection

(b) Semi-supervised anomaly detection

(c) Unsupervised anomaly detection

1. Supervised Anamoly detection:

Supervised anomaly detection techniques require a data set that has been labeled as "normal" and "abnormal" and involves training a classifier. However, this approach is rarely used in anomaly detection due to the general unavailability of labelled data and the inherent unbalanced nature of the classes.

2. Semi-Supervised Anomaly detection:

Semi-supervised learning is a type of machine learning that falls in between supervised and unsupervised learning. It is a method that uses a small amount of labeled data and a large amount of unlabeled data to train a model.

3. Un-supervised Anomaly detection:

Unsupervised learning is a type of machine learning that does not rely on labeled data to find patterns or clusters in the data. One of the applications of unsupervised learning is anomaly detection, which is the task of identifying outliers or abnormal instances in the data.

LOGISTIC REGRESSION ALGORITHMS:

Logistic regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class or not. It is a kind of statistical algorithm, which analyze the relationship between a set of independent variables and the dependent binary variables. It is a powerful tool for decision-making. For example email spam or not.

It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable.

Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of certain classes based on some dependent variables. In short, the logistic regression model computes a sum of the input features (in most cases, there is a bias term), and calculates the logistic of the result.

The output of logistic regression is always between (0, and 1), which is suitable for a binary classification task. The higher the value, the higher the probability that the current sample is classified as class=1, and vice versa.

He(x)=1/(1+e^(-ex))

As the formula above shows, n is the parameter we want to learn or train or optimize and Equation is the input data. The output is the prediction value when the value is closer to 1, which means the instance is more likely to be a positive sample(y=1). If the value is closer to 0, this means the instance is more likely to be a negative sample(y=0).

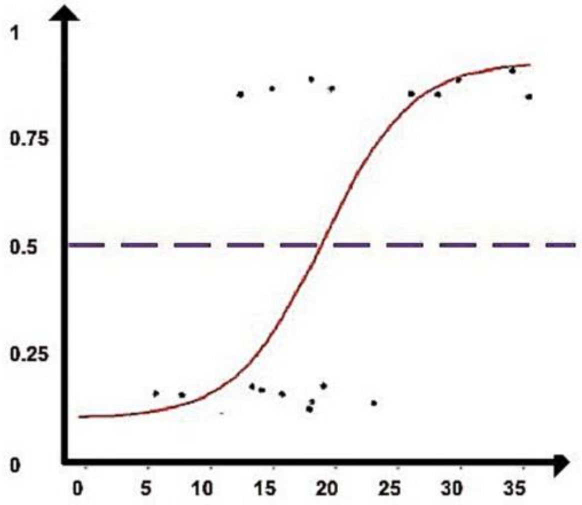
To optimize our task, we need to define a loss function(cost or objective function) for this task. In logistic regression, we use the log-likelihood loss function.

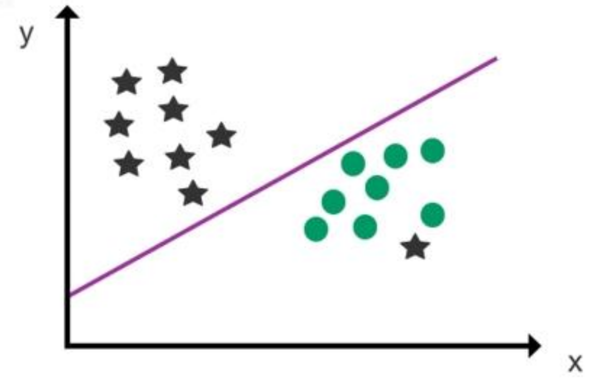
How logistic regression algorithm works:

Machine learning generally involves predicting a quantitative outcome or a qualitative class. The former is commonly referred to as a regression problem. In the scenario of linear regression, the input is a continuous variable, and the prediction is a numerical value. When predicting a qualitative outcome (class), the task is considered a classification problem. Examples of classification problems include predicting what products a user will buy or if a target user will click on an online advertisement.

Not all algorithms fit cleanly into this simple dichotomy, though, and logistic regression is a notable example. Logistic regression is part of the regression family as it involves predicting outcomes based on quantitative relationships between variables. However, unlike linear regression, it accepts both continuous and discrete variables as input and its output is qualitative. In addition, it predicts a discrete class such as “Yes/No” or “Customer/Non-customer”.

In practice, the logistic regression algorithm analyzes relationships between variables. It assigns probabilities to discrete outcomes using the Sigmoid function, which converts numerical results into an expression of probability between 0 and 1.0. Probability is either 0 or 1, depending on whether the event happens or not. For binary predictions, you can divide the population into two groups with a cut-off of 0.5. Everything above 0.5 is considered to belong to group A, and everything below is considered to belong to group B.





DECISION TREE ALGORITHM:

A decision tree is a non-parametric supervised learning algorithm for classification and regression tasks. It has a hierarchical tree structure consisting of a root node, branches, internal nodes, and leaf nodes. Decision trees are used for classification and regression tasks, providing easy-to-understand models.

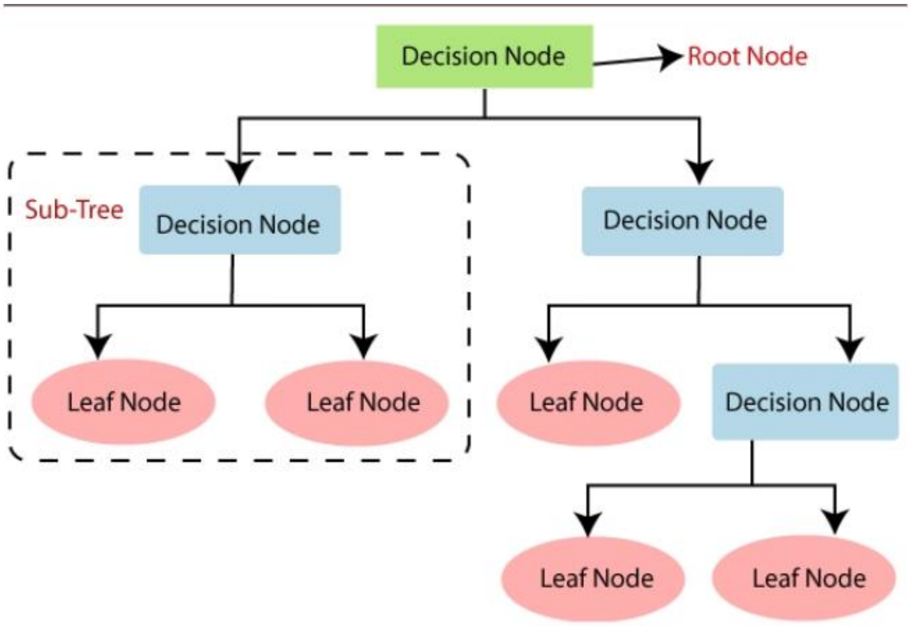
A decision tree is a hierarchical model used in decision support that depicts decisions and their potential outcomes, incorporating chance events, resource expenses, and utility. This algorithmic model utilizes conditional control statements and is non-parametric, supervised learning, useful for both classification and regression tasks. The tree structure is comprised of a root node, branches, internal nodes, and leaf nodes, forming a hierarchical, tree-like structure.

It is a tool that has applications spanning several different areas. Decision trees can be used for classification as well as regression problems. The name itself suggests that it uses a flowchart like a tree structure to show the predictions that result from a series of feature-based splits. It starts with a root node and ends with a decision made by leaves.

Some of the important terminologies in decision tree algorithm are:

1. Root Node: The initial node at the beginning of a decision tree, where the entire population or dataset starts dividing based on various features or conditions.
2. Decision Nodes: Nodes resulting from the splitting of root nodes are known as decision nodes. These nodes represent intermediate decisions or conditions within the tree.
3. Leaf Nodes: Nodes where further splitting is not possible, often indicating the final classification or outcome. Leaf nodes are also referred to as terminal nodes.
4. Sub-Tree: Similar to a subsection of a graph being called a sub-graph, a sub-section of a decision tree is referred to as a sub-tree. It represents a specific portion of the decision tree.
5. Pruning: The process of removing or cutting down specific nodes in a decision tree to prevent overfitting and simplify the model.

6. Branch / Sub-Tree: A subsection of the entire decision tree is referred to as a branch or sub-tree. It represents a specific path of decisions and outcomes within the tree.
7. Parent and Child Node: In a decision tree, a node that is divided into sub-nodes is known as a parent node, and the sub-nodes emerging from it are referred to as child nodes. The parent node represents a decision or condition, while the child nodes represent the potential outcomes or further decisions based on that condition.



ISOLATION FOREST:

Isolation Forest is an unsupervised machine learning algorithm for anomaly detection. As the name implies, Isolation Forest is an ensemble method (similar to random forest). In other words, it use the average of the predictions by several decision trees when assigning the final anomaly score to a given data point.

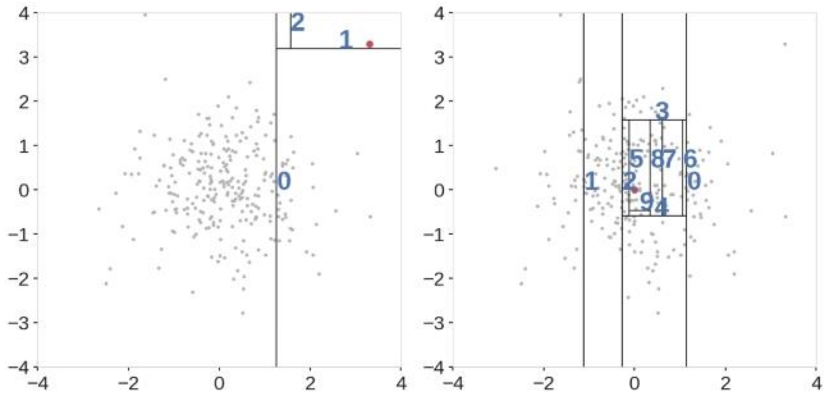
Isolation Forests(IF), similar to Random Forests, are build based on decision trees. And since there are no pre-defined labels here, it is an unsupervised model.

In an Isolation Forest, randomly sub-sampled data is processed in a tree structure based on randomly selected features. The samples that travel deeper into the tree are less likely to be anomalies as they required more cuts to isolate them. Similarly, the samples which end up in shorter branches indicate anomalies as it was easier for the tree to separate them from other observations.

How Isolation forest algorithm works:

1. When given a dataset, a random sub-sample of the data is selected and assigned to a binary tree.
2. Branching of the tree starts by selecting a random feature (from the set of all N features) first. And then branching is done on a random threshold (any value in the range of minimum and maximum values of the selected feature).

3. If the value of a data point is less than the selected threshold, it goes to the left branch else to the right. And thus a node is split into left and right branches.
4. This process from step 2 is continued recursively till each data point is completely isolated or till max depth(if defined) is reached.
5. The above steps are repeated to construct random binary trees.



(a) Anomaly point

(b) Nominal point

CREDIT CARD FRAUD DETECTION:

Credit card fraud detection is a set of methods and techniques designed to block fraudulent purchases, both online and in-store. This is done by ensuring that you are dealing with the right cardholder and that the purchase is legitimate.

In this project we are using

1. PYTHON as programming language,
2. PYTHON PANDAS for working with data sets.
3. Virtual Studio Code for compiling.
4. Kaggle website for data sets about credit card fraud detection.

Now we are going to take a deep look into the processes involved in this project.

Data Understanding:

Data understanding involves accessing the data and exploring it using tables and graphics.

Data preparation is the process of preparing raw data so that it is suitable for further processing and analysis. Key steps for understanding the datasets of credit card fraud is collecting, cleaning, and labeling raw data into a form suitable for machine learning algorithms and then exploring and visualizing the data,

Importing dependencies:

It is necessary to import the software and requirements for further process, the requirements are

1. Importing numpy

2. Importing Python Pandas
3. Importing matplotlib
4. Importing Seaborn
5. Importing xgboost
6. Importing scipy

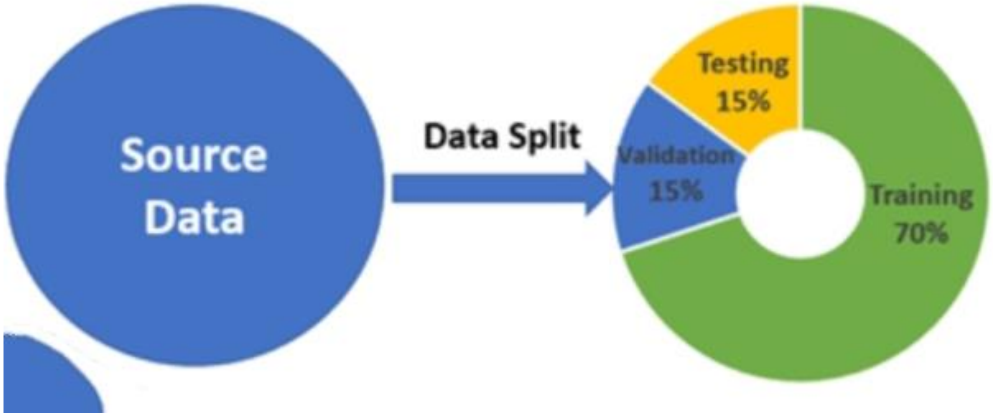
Data Understanding:

Data analysis is the process of cleaning, changing, and processing raw data and extracting actionable, relevant information that helps businesses make informed decisions. The procedure helps reduce the risks inherent in decision-making by providing useful insights and statistics, often presented in charts, images, tables, and graphs.

Data Splitting:

Data splitting is when data is divided into two or more subsets. Typically, with a two-part split, one part is used to evaluate or test the data and the other to train the model.

Data splitting is an important aspect of data science, particularly for creating models based on data. This technique helps ensure the creation of data models and processes that use data models – such as machine learning – are accurate.



Model Building:

Model building is an essential part of data analytics and is used to extract insights and knowledge from the data to make business decisions and strategies. In this phase of the project data science team needs to develop data sets for training, testing, and production purposes. These data sets enable data scientists to develop an analytical method and train it while holding aside some of the data for testing the model. Model building in data analytics is aimed at achieving not only high accuracy on the training data but also the ability to generalize and perform well on new, unseen data.

Cross Validation:

Cross-validation is a statistical technique used by data scientists for training and evaluating a machine learning model, with a focus on ensuring reliable model performance. To understand how cross-validation supports model development, we first need to understand how data is used to train and evaluate models.

Over Sampling:

Data Sampling is done by the following sampling techniques they are,

1. Random sampling: This data sampling method protects the data modeling process from bias toward different possible data characteristics. However, random splitting may have issues regarding the uneven distribution of data.
2. Stratified random sampling: This method selects data samples at random within specific parameters. It ensures the data is correctly distributed in training and test sets.
3. Nonrandom sampling:This approach is typically used when data modelers want the most recent data as the test set.

Hyper parameter tuning:

Hyperparameter tuning allows data scientists to tweak model performance for optimal results. This process is an essential part of machine learning, and choosing appropriate hyperparameter values is crucial for success. For example, assume you're using the learning rate of the model as a hyperparameter.

CONCLUSION:

By following the above process the data sets for credit card fraud occured can be detected successfully.

THANK YOU