



CREDIT CARD FRAUD DEDUCTION

USING MACHINE LEARNING



ABSTRACT

It is the responsibility of the credit card companies to ensure that their customers are given the best security measures when it comes to fraudulent credit card transactions. The main objective of this work is to explore whether a credit card transaction (before being processed) is fraudulent or not. 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



PROBLEM DEFINITION

The problem is to develop a machine-learning based system for real time credit card fraud detection. The goal is to create a solution that can accurately identify fraudulent transaction while minimizing false positive. This project involves data processing, feature engineering, model selection, training and evolution to create a robust fraud detection system



DESIGN THINKING

- 1. Data source**
- 2. Data Preprocessing**
- 3. Model selection**
- 4. Training set**

Data source

Data mining –
Data matching
The sounds like
function

Regression analysis
Clustering analysis and

Gap


Dataset Link: <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>



DATA PREPROCESSING

–

Data preprocessing crucial step in the preparation data for analysis or machine learning .It involves task like cleaning transforming and organizing data to make it suitable for intended use processing step includes






MODEL SELECTION

Model selection is a critical step in the machine learning workflow where you choose the algorithm or model that best fits your data and problem

1. Understand Your Problem : Start by understanding the nature of your problem. Is it a classification, regression, clustering, or some other type of problem? Knowing this will guide your choice of algorithms.
2. Available Data: The amount and quality of your data can influence the choice of models. Some models require large datasets, while others work well with small data.
3. Algorithm Types: Different algorithms have different strengths and weaknesses. For example, decision trees are interpretable, while deep neural networks can capture complex patterns.




4. Model Complexity: Consider the complexity of your problem. Simple models like linear regression may work well for straightforward tasks, while complex problems might require ensemble methods or deep learning.

5. Overfitting and Underfitting: Be aware of overfitting (model is too complex and fits noise) and underfitting (model is too simple and can't capture the underlying patterns). Choose models that balance this trade-off.

6. Cross-Validation: Use techniques like cross-validation to estimate a model's performance on unseen data. This helps prevent over-optimistic evaluations.

7. Domain Knowledge: Your knowledge of the problem domain can be invaluable. It might lead you to choose specific models or feature engineering techniques that are known to work well.

8. Resources: Consider the computational resources available. Some models, like deep neural networks, can be computationally intensive.






TRAINING SET

In machine learning, the training set is a crucial component of the data used to train a model. It consists of a set of labeled examples or data points, where each data point includes both input features and the corresponding target or output label. The model learns from this training data by adjusting its parameters to minimize the error between its predictions and the actual labels in the training set.

1. Input Features : These are the variables or attributes that the model uses to make predictions. For example, if you're building a model to predict housing prices, input features could include square footage, number of bedrooms, location, etc.

2. Output Labels: These are the values that the model is trying to predict or classify. In regression tasks, the output label is typically a continuous value (e.g., predicting house prices), while in classification tasks, it's a category or class label (e.g., classifying emails as spam or not spam).



3. Labeled Data: Each data point in the training set consists of both input features and their corresponding output label. For instance, if you have a dataset of 100 houses, each data point might include the features of a house (e.g., square footage, bedrooms) along with its actual sale price (the output label).

The goal during the training process is to find a model that generalizes well from the training data to make accurate predictions on new, unseen data. This process involves adjusting the model's internal parameters iteratively until the error (the difference between predictions and actual labels) is minimized on the training set. Once trained, the model can be used to make predictions on new, unlabeled data.

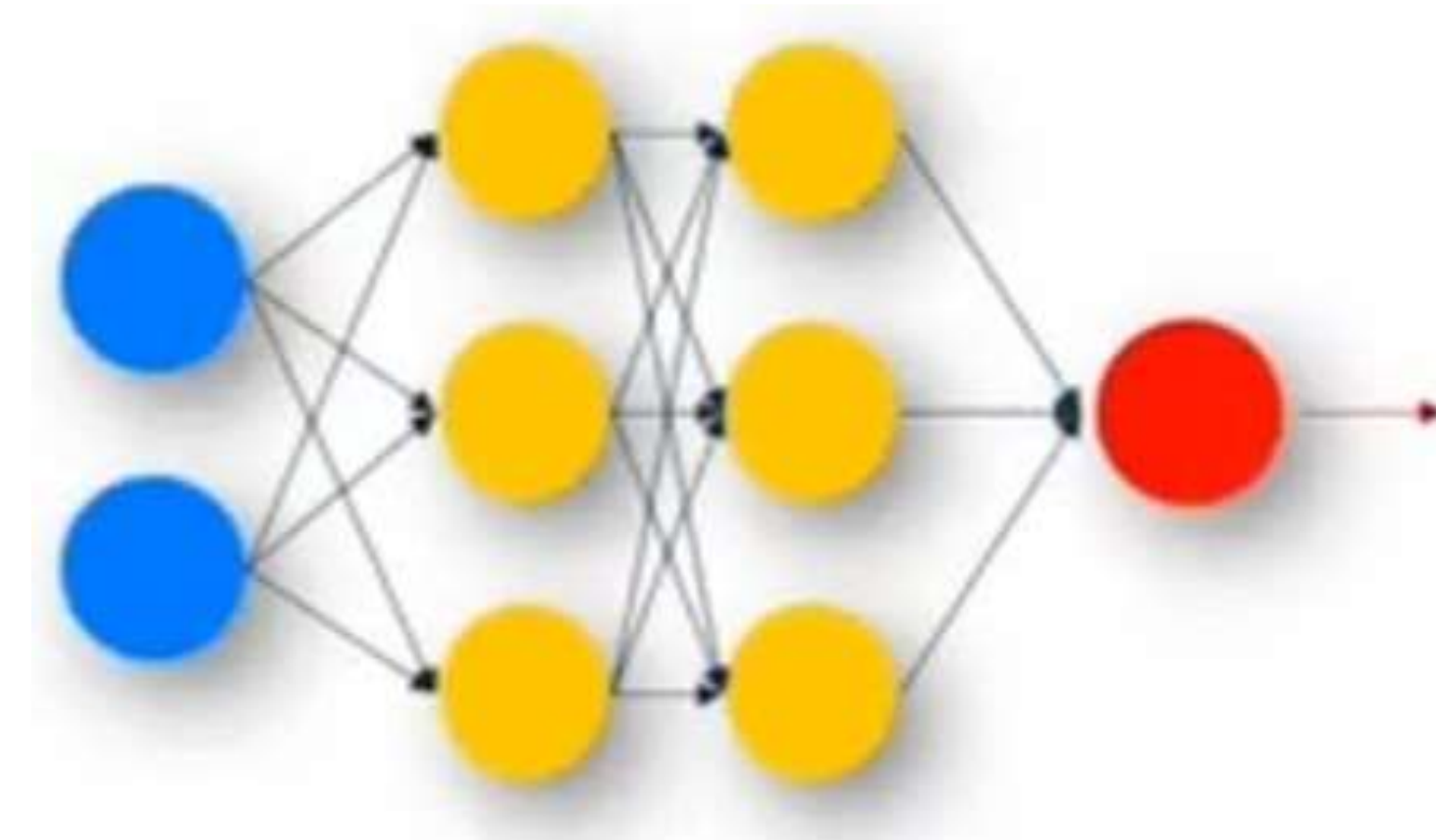
It's important to note that a good training set should be representative of the data the model will encounter in the real world and should be large enough to capture the underlying patterns in the data. Additionally, the training set is typically split into subsets for training, validation (to tune hyperparameters), and testing (to evaluate the model's performance) to ensure reliable model development.

Data Science Approach

Machine learning




Deep Neural Network (DNN)





CLASSIFICATION


Classification is a supervised machine learning method where the model tries to predict the correct label of a given input data. In classification, the model is fully trained using the training data, and then it is evaluated on test data before being used to perform prediction on new unseen data.





Regression

Machine Learning Regression is a technique for investigating the relationship between independent variables or features and a dependent variable or outcome. It's used as a method for predictive modelling in machine learning, in which an algorithm is used to predict continuous outcomes.






CLUSTERIN

G – Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.



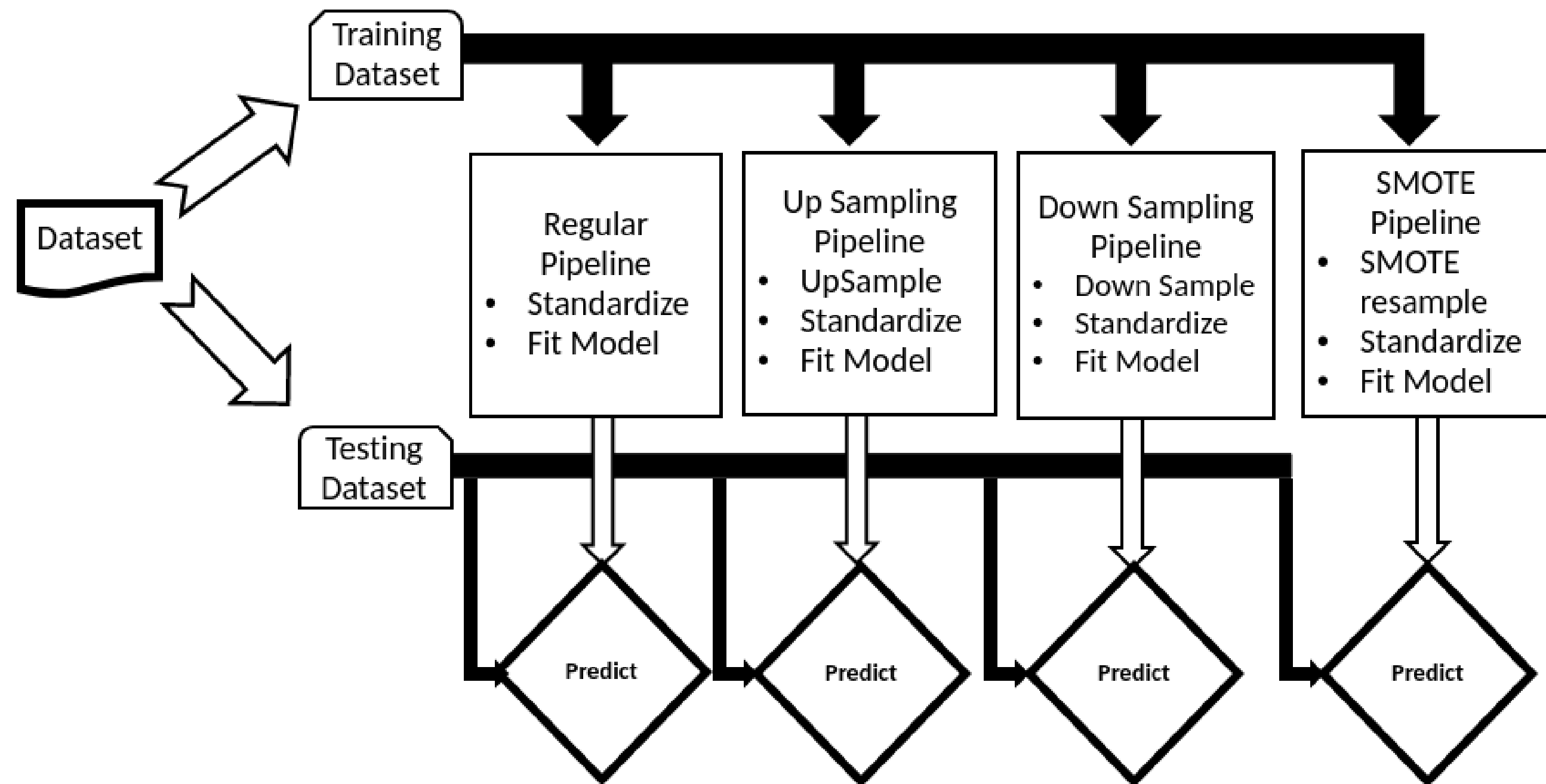
Anomaly Detection

Anomaly Detection is the technique of identifying rare events or observations which can raise suspicions by being statistically different from the rest of the observations. Such “anomalous” behaviour typically translates to some kind of a problem like a credit card fraud, failing machine in a server, a cyber attack, etc.



What other Data Scientists got

Method Used	Frauds	Genuines	MCC
Naïve Bayes	83.130	97.730	0.219
Decision Tree	81.098	99.951	0.775
Random Forest	42.683	99.988	0.604
Gradient Boosted Tree	81.098	99.936	0.746
Decision Stump	66.870	99.963	0.711
Random Tree	32.520	99.982	0.497
Deep Learning	81.504	99.956	0.787
Neural Network	82.317	99.966	0.812
Multi Layer Perceptron	80.894	99.966	0.806
Linear Regression	54.065	99.985	0.683
Logistic Regression	79.065	99.962	0.786
Support Vector Machine	79.878	99.972	0.813





CONCLUSION

–

Credit card fraud is most common problem resulting in loss of lot money for people and loss for some banks and credit card company.

This project want to help the peoples from their wealth loss and also for the banked company and trying to develop the model which more eciently separate the fraud and fraud less transaction by using the time and amount feature in data set given in the Kegel. rst we build the model using some machine learning algorithms such as logistic regression, decision tree, support vector machine, this all are supervised machine learning algorithm in machine learning.