SOMAIYA VIDYAVIHAR UNIVERSITY
K J Somaiya College of Engineering

Machine Learning Algoriththm used for detection of frauds in E-Commerce transactions

What is a fraud detection Algorithm in machine learning?

Fraud Detection Using Machine Learning deploys a machine learning (ML) model and an example dataset of credit card transactions to train the model to recognize fraud patterns. The model is self-learning which enables it to adapt to new, unknown fraud patterns.

What is the main purpose of fraud detection algorithm in E-commerce Transactions?

- Fraud is a lie that is used to illegally limit the rights of other people, entities, or money.
- In 2017 to 2018 total of 911 credit card frauds amounting to 65.26 crores according to RBI report which has been illegally transferred from different banks.
- Due to cashless transaction every people use ATM card and credit card for transaction, so fraud can also be increased.
- Billions of dollars are lost every year by fraudulent activities.

So the purpose of Ecommerce fraud detection and prevention covers all the tools and processes an online store can put in place to reduce the costs and resources lost to fraud.

Thus, you provide a better customer experience.

Types of fraud detection Algorithm used in machine learning

There are various different algorithms or classification techniques used in fraud detection such as Decision tree algorithm,logistics regression,K-nearest Neighbour,Support vector Machine and Random forest algorithms.

It was found that Logistic regression is often preferred over other algorithms because of fllowing reasons:

- It is easier to understand and explain. We can easily see which factors are most important in detecting fraud.
- It requires less data than other tools.
- It works well when the data is not too complicated.
- It is fast and efficient.

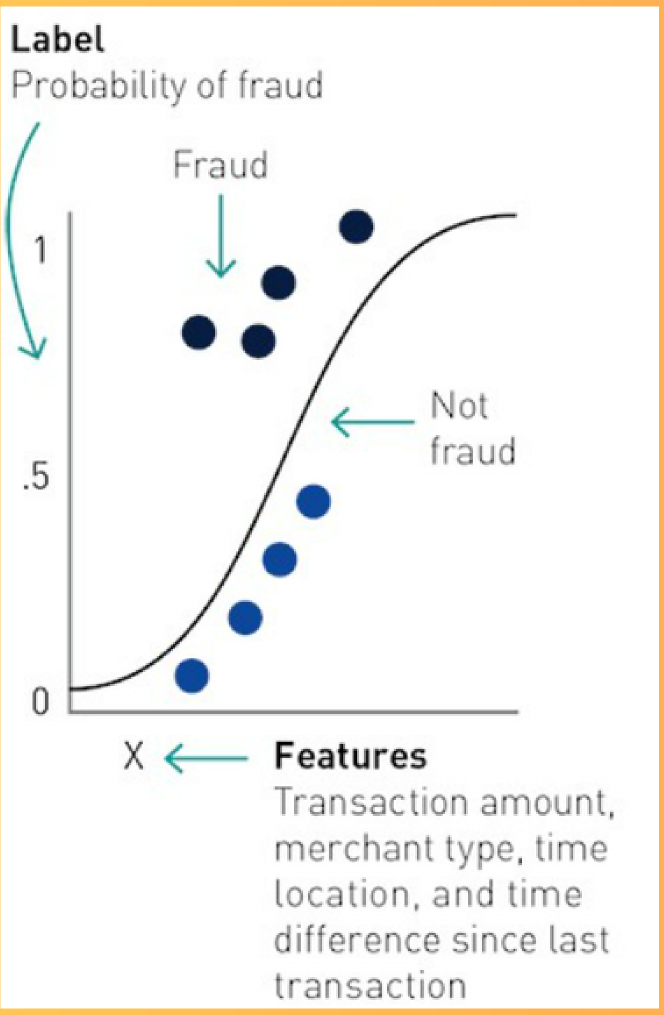
Credit Card Fraud Detection

References

- An Approach for Detecting Frauds inE-Commerce Transactions using ML Techniques
- https://www.spiceworks.com/tech/artificial intelligence/articles/what-is-logistic-regression/
- https://favtutor.com/blog-details/credit-card-fraud-detection-using-machine-learning

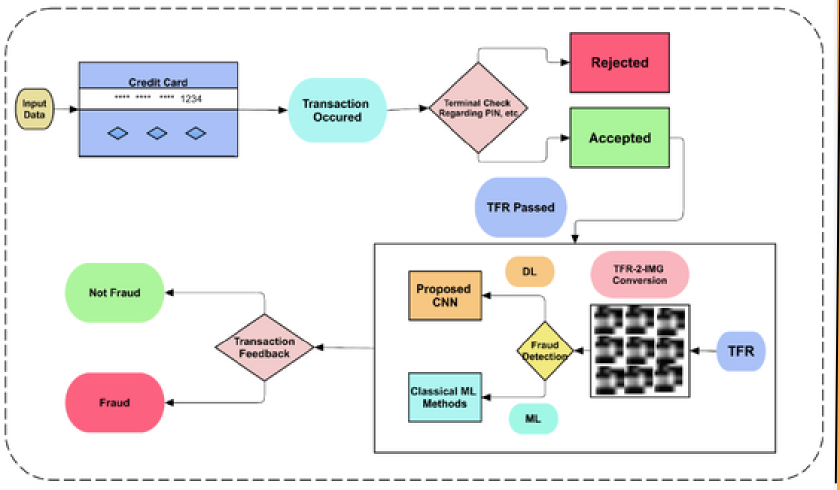
Logistic Regression

Logistic regression is a statistical algorithm that works by estimating the probability of a binary outcome - in this case, whether a transaction is fraudulent or not. The algorithm uses a set of input variables, or features, to calculate the probability of the outcome.



Working of Logistics Regression Algorithm

1. Suppose we have a dataset of 10,000 transactions, 1000 of which are fraudulent. We split the data into a training set (80%) and a testing set (20%).
2. We preprocess the data by cleaning it, removing any missing values.
3. Engineer use new features, such as the number of previous purchases made by the user, the average purchase amount, or the distance between the user's billing and shipping addresses.
4. We train the logistic regression model on the training set using the input variables and the known outcomes (fraudulent or not fraudulent) of each transaction. The model learns to calculate the probability of a transaction being fraudulent based on the input variables.
5. We evaluate the model's performance on the testing set by comparing its predictions to the actual outcomes of the transactions. We can use metrics such as accuracy, precision, recall, and F1-score to measure how well the model is able to detect fraudulent transactions.
6. We can then use the logistic regression model to predict the likelihood of fraud for new transactions based on their input variables.

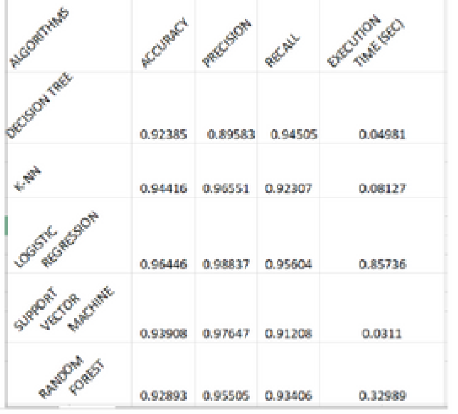


Logistics Regression Algorithm steps

```
Input: Training data
Begin
For i = 1 to k
For each training data instance $d_i$.
Set the target value for the regression to $z_i = \frac{y_i - P(1|d_j)}{[P(1|d_j)(1-P(1|d_j))]}$
Initialize the weight of instance $d_j$ to $[P(1|d_j)(1 - P(1|d_j))]$
Finalize a $f(j)$ to the data with class value $(Z_j)$ and weight $(w_j)$
Classical label decision
Assign (class label: 1) if $P_{id} > 0.5$, otherwise (class label: 2)
End
```

Time and Space complexity analysis of logistics regression algorithm

- Time complexity : The time complexity of the logistic regression algorithm is O(kn^2), where k is the number of input variables and n is the number of observations in the dataset. This means that as the number of input variables or the size of the dataset increases, the time required to train the model will increase quadratically. However, logistic regression is generally considered to be computationally efficient and can handle large datasets.
- Space complexity : The space complexity of the logistic regression algorithm is O(k), which means that the amount of memory required to store the model coefficients increases linearly with the number of input variables.



ALGORITHMS	ACCURACY	PRECISION	RECALL	EXECUTION TIME (SEC)
DECISION TREE	0.92385	0.89583	0.94505	0.04981
K-NN	0.94416	0.96551	0.92307	0.08127
LOGISTIC REGRESSION	0.96446	0.98837	0.95604	0.85736
SUPPORT VECTOR MACHINE	0.93908	0.97647	0.91208	0.0311
RANDOM FOREST	0.92893	0.95505	0.93406	0.32989

Confusion Matrix For Comparison of various Fraud detection Algorithms

Logistics Regression Algorithm Real life Applications

- ***Determine the probability of heart attacks***: With the help of a logistic model, medical practitioners can determine the relationship between variables such as the weight, exercise, etc., of an individual and use it to predict whether the person will suffer from a heart attack or any other medical complication.
- ***Possibility of enrolling into a university*** : Application aggregators can determine the probability of a student getting accepted to a particular university or a degree course in a college by studying the relationship between the estimator variables, such as GRE, GMAT, or TOEFL scores.
- ***Identifying spam emails*** : Email inboxes are filtered to determine if the email communication is promotional/spam by understanding the predictor variables and applying a logistic regression algorithm to check its authenticity.

Logistics regression formula

$$y = \frac{e^{(b_0 + b_1 X)}}{1 + e^{(b_0 + b_1 X)}}$$

Name : Keyur Patel
Roll No : 16010421073
Batch : A-2