Department of Mathematics

Student's Name	Mahek Jain Kruthi MN			**USN**	1RV18CS082 1RV18CS073
Semester	**VI**	**Branch**	**Common to all**	**Section**	B
Course Title	**ADVANCED STATISTICAL METHODS (Global Elective-Theory)**			**Course Code**	**18G6E15**

Experiential Learning _ASSIGNMENT

Course Outcomes	CO1	CO2	CO3	CO4
CO Attainment in marks				
COs	Explore the fundamental concepts of mathematics involved in machine learning techniques	Orient the basic concepts of mathematics towards machine learning approach	Apply the linear algebra and probability concepts to understand the development of different machine learning techniques	Analyze the mathematics concepts to develop different machine learning models to solve practical problems

Certificate

This is to certify that **Mahek Jain and Kruthi MN** (**USN: 1RV18CS082 and 1RV18CS073**) of VI Semester **CSE** Branch have satisfactorily completed the **Assignment** prescribed by the Institution in the course **ADVANCED STATISTICAL METHODS** (18G6E15) for the academic year 2020 – 2021 (EVEN SEM 2021).

Max. Marks	Marks Obtained
20	

Identifying Fraudulent Activities in e-Commerce Websites using Random Forest Algorithm

Kruthi MN[1], Mahek Jain[2],

Department of Computer Science and Engineering

RV College of Engineering

Bengaluru, India

kruthimn.cs18@rvce.edu.in[1],mahekjain.cs18@rvce.edu.in[2]

***Abstract*—E-Commerce is undeniably one of the biggest sectors in online business. As a sector that continues to blow up in size, volume and influence, it is no surprise that fraud still exists through the loopholes in the system. With more sophisticated technology available to fraudsters, it has become even more difficult for e-commerce businesses to keep a track of the tactics used to defraud online businesses, hence there arises a need to use sophisticated machine learning algorithms to combat it. Predictive machine learning algorithms like random forest turns out to be a potential solution for online fraud detection in ecommerce websites. Precision, recall and F1 score can be used to evaluate the performance of the random forest model.**

Keywords—Random forest, Decision tree, F1 score, Entropy, Information gain, Receiver Operating characteristic curve.

I. Introduction

Machine Learning has always been useful for solving real-world problems. Nowadays, it is widely used in every field such as medical, e-commerce, banking, insurance companies, etc. Earlier, all the reviewing tasks were accomplished manually. But with the increase in the processing power of systems and the advancement in statistical modeling, the acceptance of Machine Learning in every sector has increased.

For years, fraud has been a major issue in sectors like banking, medical, insurance, and many others. E-commerce websites often transact huge amounts of money. And whenever a huge amount of money is moved, there is a high risk of users performing fraudulent activities, e.g. using stolen credit cards, doing money laundry, etc. Due to the increase in online transactions through availability of different payment options, such as credit/debit cards, PhonePe, Gpay, Paytm, etc., fraudulent activities are becoming intense and are increasing rapidly in recent days. Moreover, fraudsters or criminals have become very skilled in identifying and exploiting all possible vulnerabilities in online transactions. Since no system is perfect and there is always a loophole, it has become a challenging task to make a secure system for authentication and preventing customers from fraud. So, Fraud detection algorithms based on machine learning turns out to be very useful in making online transactions as secure as possible for the customers of ecommerce websites.

Machine Learning really excels at identifying fraudulent activities. Any website where you put your credit card information has a risk team in charge of avoiding frauds via machine learning. The goal is to build a machine-learning model that predicts the probability that the first transaction of a new user is fraudulent.

II. Use of machine learning for ecommerce fraud detection

A. Speed

Machine Learning is widely used because of its fast computation. It analyzes and processes data and extracts new patterns from it within no time. For human beings to evaluate the data, it will take a lot of time and evaluation time will increase with the amount of data. Rule-based fraud prevention systems are based on written rules for permitting which type of actions are deemed safe and which one's must raise a flag of suspicion. Now, this Rule-based system is inefficient because it takes much time to write these rules for different scenarios. And that's exactly where Machine Learning based Fraud Detection algorithms succeed in not only learning from these patterns it is capable of detecting new patterns automatically. And it does all of this in a fraction of the time that these rule-based systems could achieve.

B. Scalability

As more and more data is fed into the Machine Learning-based model, the model becomes more accurate and effective in prediction. Rule-based systems don't evolve by themselves as professionals who developed these systems must write these rules meeting various circumstances.

C. Efficiency

As more and more data is fed into the Machine Learning-based model, the model becomes more accurate and effective in prediction. Rule-based systems don't evolve by themselves as professionals who developed these systems must write these rules meeting various circumstances.

III. Methodology

A. Data

This study uses an e-commerce fraud dataset sourced from Kaggle. The dataset consists of 151,112 records in total. The number of records classified as fraud are 14,151, hence, the ratio of fraud data is 0.093.

The two tables considered in the dataset are as follows :

1. Fraud_Data -contains information about each user first transaction with attributes :

   • user_id : Id of the user. Unique by user

   • signup_time : the time when the user created her account (GMT time)

• purchase_time : the time when the user bought the item (GMT time)

• purchase_value : the cost of the item purchased (USD)

• device_id : the device id. You can assume that it is unique by device. I.e., 2 transactions with the same device ID means that the same physical device was used to buy

• source : user marketing channel: ads, SEO, Direct (i.e. came to the site by directly typing the site address on the browser).

• browser : the browser used by the user.

• sex : user sex: Male/Female

• age : user age

• ip_address : user numeric ip address

• class : this is what we are trying to predict: whether the activity was fraudulent (1) or not (0).

2. IpAddress_to_Country - contains mapping of each numeric ip address to its country. For each country, it gives a range. If the numeric ip address falls within the range, then the ip address belongs to the corresponding country. The attributes are as follows :

   • lower_bound_ip_address : the lower bound of the numeric ip address for that country

   • upper_bound_ip_address : the upper bound of the numeric ip address for that country

   • country : the corresponding country. If a user has an ip address whose value is within the upper and lower bound, then she is based in this country.

B. Decision Tree for classification

Decision tree algorithm is a supervised machine learning algorithm. Decision tree algorithms belong to the family of predictive modelling approaches that are used in statistics, data mining and machine learning. It uses a decision tree as a predictive model to traverse from observations about an item represented by branches to conclusions about the item's target value represented by leaf nodes. Decision tree algorithms can be used for solving both regression and classification problems. The goal of using decision tree algorithms is to create a training model that can predict target class or value for input variables by simple decision rules inferred from prior training data. In the prediction process of the decision tree, we start from the root of the tree, values of the root and record's attribute are compared, on the basis of results of comparison we follow the branch that corresponds to that value and jump to the next node. Decision tree which has a categorical target variable is called a categorical variable decision tree and the one with a continuous target variable is known as a continuous variable decision tree.

C. Random Forest Classifier

Random forest is a supervised machine learning algorithm for classification and regression. Random forest creates multiple decision trees at training time and outputs the mode of the classes for classification problems and mean prediction for regression problems of the individual trees. Random forest builds an ensemble of decision trees that are trained by bagging methods and the trees are randomised by precision techniques, which substantially increases the performance of the model.

D. Information Gain

Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute. It calculates how much information a feature provides us about a class. According to the value of information gain, we split the node and build the decision tree. A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first. It can be calculated using the below formula:

Information Gain = Entropy(S) - [Weighted Average * Entropy of each feature]

Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

$$\text{Entropy} = \sum_{j=1}^{k} P_j \log_2 P_j$$

where, k is the number of classes of the target attribute. P_j is the number of occurrences of class j divided by the total number of instances i.e. the probability of occurrence of j..

E. Gini Index

Gini index is a measure of impurity or purity used while creating a decision tree in the CART(Classification and Regression Tree) algorithm. An attribute with the low Gini index should be preferred as compared to the high Gini index. It only creates binary splits, and the CART algorithm uses the Gini index to create binary splits. Gini index can be calculated using the below formula:

$$\text{Gini Index} = 1 - \sum_j P_j^2$$

F. Precision

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. High precision relates to the low false positive rate. Value of Precision can be evaluated as

$$\text{Precision} = TP/TP+FP$$

where, TP is the number of true positives, FP is the number of false positives.

G. Recall

Recall is the ratio of correctly predicted positive observations to the all observations in actual class. Value of Recall can be evaluated as

$$\text{Recall} = TP/TP+FN$$

where, TP is the number of true positives, FN is the number of false negatives.

H. F1 Score

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false

negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if there is an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall. In this case, F1 score is .

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

I. Algorithm

Random forest Classifier which is an ensemble of decision trees classifiers is the algorithm used in the project. Python is the programming language used for implementation. Algorithm is as follows :

1. Import Sklearn, numpy, matplotlib, pandas and other required Python libraries.

2. Loading the datasets.

3. Pre-processing the loaded data by converting the attributes to proper format.

4. Converting categorical data into numerical data by label encoding as shown in Fig-1.

5. Splitting data into training and testing subsets.

6. Implementing Random forest algorithm to the formatted data as shown in Fig-2.

7. Evaluating the model by testing the dataset and plotting ROC curve as shown in Fig-3.

8. Analysing the testing results and retraining the model with updated parameters.

```
[ ]  lb = LabelEncoder()
     X['device_id'] = lb.fit_transform(X['device_id'])
     X['source'] = lb.fit_transform(X['source'])
     X['browser'] = lb.fit_transform(X['browser'])
     X['sex'] = lb.fit_transform(X['sex'])
     X['country_revised'] = lb.fit_transform(X['country_revised'])
```

Fig-1. Python code snippet for label encoding

```
pipeline = Pipeline(steps = [('clf', RandomForestClassifier(criterion = 'entropy'))])
```

```
clf_forest = RandomForestClassifier(n_estimators= 20, criterion = 'entropy',
    max_depth= 50, min_samples_leaf= 3,min_samples_split= 3, oob_score= True)
```

```
clf_forest.fit(X_train, y_train)
```

Fig-2. Implementation of Random Forest Algorithm in Python.

```
[ ]  preds = clf_forest.predict(X_test)
     preds #predicting y using X_test
     print(classification_report(y_test, preds))
     #ROC Curve
     prob_score = clf_forest.predict_proba(X_test)
     prob_score = DataFrame(prob_score).iloc[:,0]
     fpr,tpr,thresholds = roc_curve(y_test,1-prob_score)
```

Fig-3. Python code snippet for calculating precision, recall,F1 score and plotting ROC curve.

IV. Results

Precision gives a measure of fraudulent activities that the proposed system identifies correctly out of the total number of fraudulent activities that actually occur, for the system. The value of precision obtained is 0.98 for the proposed system, hence, whenever the system identifies a fraudulent activity it will be correct 98% of time. For the system, Recall gives the percentage of actual fraudulent activities that are correctly identified. The value of recall i.e., true positive rate obtained for the system is 0.75, hence the system can correctly identify 75% of all fraudulent activities. To fully evaluate the effectiveness of a model, both precision and recall must be evaluated, but there is a trade-off between both, hence, F1 score, which is a combination of both, is calculated. For the system, the F1 score obtained is 0.68.

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. An ROC curve plots True positive rate vs. False positive rate at different classification

thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. ROC curves obtained are shown in Fig 4 and Fig 5.

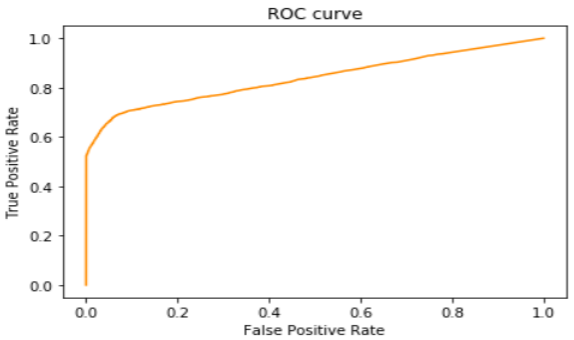

Fig-4. ROC curve.

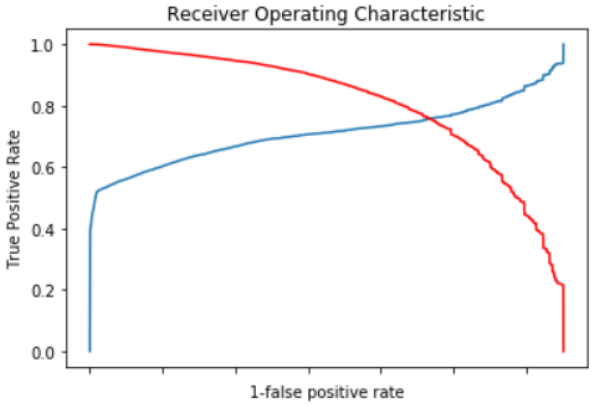

Fig 5. ROC curve after modification

V. Conclusion

Ecommerce is undoubtedly the biggest contributor of revenue. Unfortunately, increased money comes at the cost of increased fraud. According to the reports, eCommerce fraud numbers have grown rapidly over the last few years making online frauds twice the rate of eCommerce sales. Ecommerce fraud is sophisticated and ever-evolving, as fraudsters leverage more advanced tactics with every passing year. Hence, e-commerce fraudulent activities prediction models like the proposed system powered by machine learning helps ecommerce companies better combat fraud attempts. The proposed fraud detection system requires minimal input data as it relies on only the information of the first transaction. There is a scope for further improvement by incorporating additional input data.

References

[1] Renjith, S. “Detection of Fraudulent Sellers in Online Marketplaces using Support Vector Machine Approach”, International Journal of Engineering Trends and Technology, 2018.

[2] Roy, Abhimanyu, "Deep learning detecting fraud in credit card transactions." 2018 Systems and Information Engineering Design Symposium (SIEDS). IEEE, 2018.

[3] Zhao T, Jie H, . "Extracting and reasoning about implicit behavioral evidences for detecting fraudulent online transactions in e-Commerce.", Decision support systems 86, 20161.

[4] Pumsirirat, Apapan, Liu Yan. "Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine.", International Journal of advanced computer science and applications, 2018.

[5] Srivastava, Abhinav, "Credit card fraud detection using hidden Markov model.", International Conference Of Transactions on dependable and secure computing , 2008 .

[6] Lakshmi, S. V. S. S, S. D. Kavilla. "Machine Learning For Credit Card Fraud Detection System.", International Journal of Applied Engineering Research 2018.

[7] Aljarah, Ibrahim, Hossam Faris, Seyedali Mirjalili. "Optimizing connection weights in neural networks using the whale optimization algorithm.", Soft Computing 2018.

[8] Xuan, Shiyang, Guanjun Liu, and Zhenchuan Li. "Refined weighted random forest and its application to credit card fraud detection.", International Conference on Computational Social Networks. Springer, Cham, 2018.

[9] Hong, Haoyuan, et al. "Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China).", 2018.

[10] Sharma, Shiven, et al. "Synthetic oversampling with the majority class: A new perspective on handling extreme imbalance." 2018 IEEE International Conference on Data Mining (ICDM). IEEE, 2018.

[11] Kim, Jaekwon, Youngshin Han, and Jongsik Lee. "Data imbalance problem solving for smote based oversampling: Study on fault detection prediction model in semiconductor manufacturing process." International Conference On Advanced Science and Technology Letters, 2016.