ISSN : 2582-7464

Computational Intelligence and Machine Learning

Volume No. 4

Issue No. 2

May - August 2023

ENRICHEDPUBLICATIONSPVT.LTD

S-9,IIndFLOOR,MLUPOCKET,
MANISHABHINAVPLAZA-II,ABOVEFEDERALBANK,
PLOTNO-5,SECTOR-5,DWARKA,NEWDELHI,INDIA-110075,
PHONE:-+(91)-(11)-47026006

ISSN : 2582-7464

Computational Intelligence and Machine Learning

Aims and Scope

The primary objective of the Computational Intelligence and Machine Learning is to serve as a comprehensive, open-access platform that is dedicated solely to facilitating the progress and advancement of the field of Artificial Intelligence & Machine Learning by -

offering gifted and talented researchers engaged within the domain of Artificial Intelligence & Machine Learning a unique setting for them to get their work published and elevate their reputations/standing within the global community, as well as,

providing professionals, students, academics, and scholars free access to the latest and most advanced research outcomes, findings, and studies, being carried out in the field of Artificial Intelligence & Machine Learning, all across the world.

Related Topics

Soft computing

Fuzzy Logic

Artificial Neural Networks

Evolutionary Computing

Artificial Intelligence and Machine Learning

Artificial Immune Systems

Probabilistic Methods

Cognitive Robotics

Data mining

Computational Intelligence Methods for Bioinformatics and Biostatistics

ISSN : 2582-7464

Other emerging topics in Computational Intelligence

Nanobioscience

Information Forensics and technology

Nanotechnology

Cybersecurity

Big Data

Bioengineering and Biotechnology

Computational Neuroscience

ISSN : 2582-7464

Advisor

Editor-in-chief

Editorial Board Members

ISSN : 2582-7464

ISSN : 2582-7464

ISSN : 2582-7464

Computational Intelligence and Machine Learning

(Volume No. 4, Issue No 2., May - August 2023)

Contents

Prediction of Customer Churn in Telecom Industry:A Machine Learning Perspective

Lopamudra Hota1, Prasant Kumar Dash jer2*
1 M Department of Computer Science and Engineering, National Institute of Technology, Rourkela, India.
2 Department of Computer Science and Engineering, C. V. Raman Global University, Bhubaneswar, India.

ABSTRACT

The business world is becoming increasingly saturated in today's competitive environment. There is a great deal of competition in the telecommunications industry, especially due to various vibrant service providers. As a result, they have had difficulty retaining their existing customers. As attracting new customers is much more costly than retaining current ones, now is the time to ensure the telecom industry maintains value by retaining customers over acquiring new ones. Numerous machine learning and data mining methods have been proposed in the literature to predict customer churners using heterogeneous customerrecords overthe past decade. This research gives a brief idea on the Customer Churn problem, and explores how various machine learning techniques can be used to predict customer churn via models such as XGBoost, GradientBoost, AdaBoost, ANN, Logistic Regression and Random Forest, and also compare the effectiveness of the models in term of accuracy.

KeywordsAdaBoost, Customer Churn, GBoost, Machine Learning, Prediction Model, Random Forest, XGBoost.

INTRODUCTION

It is much more expensive to acquire new customers than to retain existing ones. The cost of acquiring a new customer is six to seven times greater than retaining an existing customer [1]. Customers are regarded as the most significant assets in any industry or sector because they provide the majority of the profit. Companies today are putting a greater emphasis on convincing and retaining their existing customers. Consumers' churn can be reduced if the firm correctly predicts customer behavior, expands the link between consumer attrition, and has factors under its control. By determining the difference between churners and non-churners, you can predict churn [2]. Customer churn is perceived as a significant issue in service-based firms due to its direct effect on revenues. As a result, many businesses focus on reducing churn and identifying appropriate processes for doing so. Firms intend to keep their customers through spending and minimizing profits. The best way to retain consumers is to reduce the rate of churn, which refers to the phenomenon of a consumer switching from one service provider to another or ceasing to use a particular service over a certain period. A variety of reasons could be identified in the past if a firm analyzed its history of data and adopted machine learning technology, which can identify theconsumers who are likely to churn [3]. Almost every firm now has data about its

clients and about the behavior of their customers thanks to the development of data management. A major advantage of big data is the high quality and diversity of consumer data and the ability to provide a strategic benefit to the company. Data mining assists in identifying,identifying, and understanding the behavior of the consumer,thus optimizing business operations and enhancing customermanagement effectiveness [4].

Many factors can lead to a firm losing its customers. Cost,quality, and service quality all play a significant role in that.A huge outflow of consumers affects the valuation of anyfirm. Market reputations and stakeholder trust are destroyed mainly by most firms. However, it is also essential to determine the level of customer satisfaction to retain attract customers.

Identifying how satisfied consumers are can be a challenging task. As the base of consumers grows, it becomes more challenging. The value-added service is another cause of consumer churn. The telecom industry hasintroduced a new offer called Triple Play, which spans television, phone, and internet services. In addition to adding value, this offer helps retain consumers.

Furthermore, it maximizes the revenue allocated to eachuser directly by the company [5]. Telecom companies face a unique challenge in predicting churn. Telecom analytics is a type of business intelligence explicitly used to satisfy the demands of the telecom sector. Analytics in telecom is primarily focused on maximizing profits, minimizing costs, and decreasing fraud. The purpose of telecom analytics is to forecast, multidimensionally, and optimize. Most companies suffer from customer churn, affecting their revenues when a customer moves from one service provider to another in the telecom sector. To grow their revenue-generating base, Telco companies must both attract new customers and avoid terminations (churn). According to churn analysis, customers terminate their contracts for various reasons, including better price offers, more exciting packages, poor service experiences, and changes in their personal circumstances.

The paper is organized as, Section 2 presents the review of literature related to customer churn based on machine learning, Section 3 states the churn problem in details.Section 4, describes the proposed work comprising of the model design and result analysis; finally, section 5 depicts the conclusion of the work.

LITERATURE STUDY

Kassem et al. [6] identified the main factors influencing customer churn and identified customers likely to churn byanalyzing social media. The results are analyzed using various ma chine learning algorithms such as Deep Learning, Logistic Regression, and Naïve Bayes. In [7], the study's main goal is to predict customer churn in telecom by using machine learning and big data platforms. Consumer churn can be predicted using machine learning methods. Consumer churn prediction using KNN and big data depicts the study results shows an accuracy rate of 0.80 percent for predicting consumer churn, and 1.01 percent for the area under the curve.

With specific reference to SyriaTel Telecom Company,Ahmad et al. [3] developed a mechanism for predicting the churn of consumers. Random Tree, Decision Tree, extreme gradient boosting algorithm, and GSM tree algorithm were chosen in this research. Selecting the features as well as adding the features of the mobile social network have had a significant influence on the success of the developed model as SyriaTel's area under the curve (AUC) value has reached 93.301 percent. In all measurements, the extreme gradient boosting algorithm achieved the best results. Almuqren et al. [8], offers a new approach to predicting churn and comparesmthe telecom industry using social media mining. In this study, Arab Twitter mining was used to predict churn in Saudi Telecom companies for the first time. Based on various standard metrics to the ground-truth actual outcomes offered by a telecom company, the newly proposed method is proven to be effective.

Different techniques have been used to predict customerchurn, including data mining, machine learning, and hybrid technologies. Churn predicting, and retention techniques help companies identify, predict, and prevent churn. Most of them used decision trees because it is a recognized method for determining customer churn, but it's a challenge to solving complex problems this way. The study shows that reducing the data improves the accuracy of the decision tree [9]. Customer prediction algorithms and historical analysis are sometimes used in data mining. In addition to discussion of regression trees, decision trees, neural networks, and some other data mining methods were examined in [10]. Our system is designed based on the data analysis and visualization of data collected from telecom department. The churn prediction and analysis of the machine learning models is done based on performance metrics such as precision, recall, f1-score and accuracy.

CHURN PROBLEM

When it comes to a business environment, customer attrition simply refers to customers switching services. Subscriber churn or customer churn is similar to attrition,when a customer switches from one service provider to another anonymously. In machine learning terms, churn prediction is a supervised (i.e. labeled) problem: Given a predetermined forecast horizon, one goal is to predict the number of subscribers that will churn over that time frame. Churn Prediction identifies churners in advance, before they leave the network. Therefore, the CRM department is able to prevent subscribers from churning in the future by implementing retention policies that attract and retain likely churners. Thus, the company would not suffer a potential loss. A mobile subscriber's past calls, along with his or her personal details and business information, are inputs into this problem. A list of churners is also provided for the training phase. When a model has been trained to the highest level of accuracy, it must be able to predict the churners from the real dataset which does not include any churn labels. The knowledge discovery process categorizes this problem as predictive modeling or data-mining. Figure 1 portray a model of churn prediction with four steps; 1) Preprocessing ofcustomer data 2) Feature extraction for model

design 3) Model design by classifiers and validation4) Computation of performance metrics for model comparison.

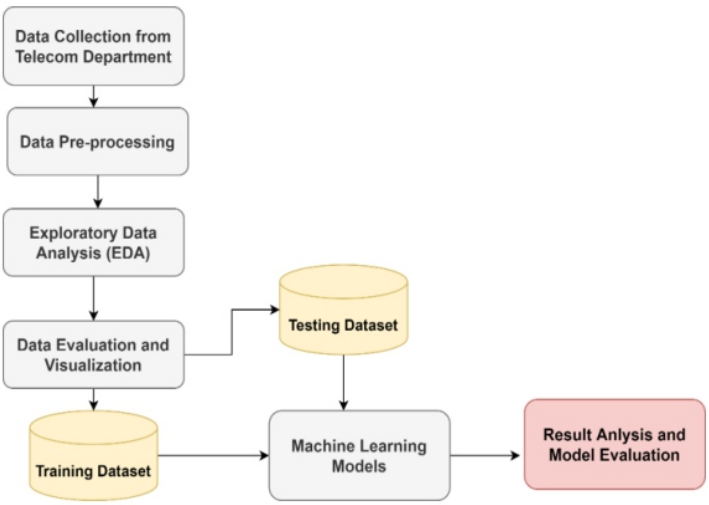


Figure 1. Flow Diagram of Proposed Work

In this research, some of the addressed questions will be;analysis of the most important feature for customer churn,which type of customers are leaving more, and which machine learning model is the best one for result analysis and prediction. We explored classification techniques, compared their accuracy, as well as other metrics, precision, recall, f1-score, True/False Positive Rates. Data Preprocessing checks for missing values, correlated variables, and outliers;EDA for hypothesis generation; data scaling to improvise data accuracy; train and test dataset generation; training models for cross-validation and plotting data accuracy results from test data.

Dataset

IBM Telecom's Kaggle Dataset was used in this research paper. Several extremely important parameters for predictive churn analysis were included in the dataset, and the data is extremely large. 7043 instances of 21 attributes are contained in the dataset. Features include details about demographicinformation like gender, age, and dependents, services they have signed up for, contract information, payment methods, paperless billing, monthly charges, and a variable in which we anticipate which customers have left within the past month. Input data is in CSV format and visualized using various visual elements such as graphs, helping to identify trends, outliers, and patterns in the data. The analysis starts with data cleaning followed by graphical analysis, machine learning model, estimation and result analysis.

Methodology

Data Pre-processing

A data set consists of features and N rows. There are many formats used for values. Duplicate values and null values can lead to loss of accuracy in a dataset and dependent values. Various data sources have been used to collect data, so uses a different format to represent a single value, such as whether someone represents Male/Female or M/F. In order to avoid noisy data, null values, and incorrect sizes, an image in 3-dimension should be reduced to a 2-dimension format by reducing it to 0 and 1. Images can be cleaned with OpenCV or Panda's tabular data. Making the data useful is paramount since generating unsatisfactory results or achieving less accurate results can be affected by unwanted or null values. Missing and incorrect values are prevalent in the data set. The entire dataset was analyzed and only the most useful features were listed. By listing features, the listing will be more accurate and contain only useful features. For a knowledge-based approach to data selection, feature selection is a crucial step. Out of the dataset here, we chose the features necessary for improving performance and helpful for decision-making, while the rest of the features have less importance.

Data Exploration

Explorative Data Analysis (EDA) provides a clear and better understanding of data patterns and potential hypothesis. The distribution of feature is an essential for trend analysis of dataset.

a. The gender distribution graph depicts that male and female distribution is nearly same as per figure 2.

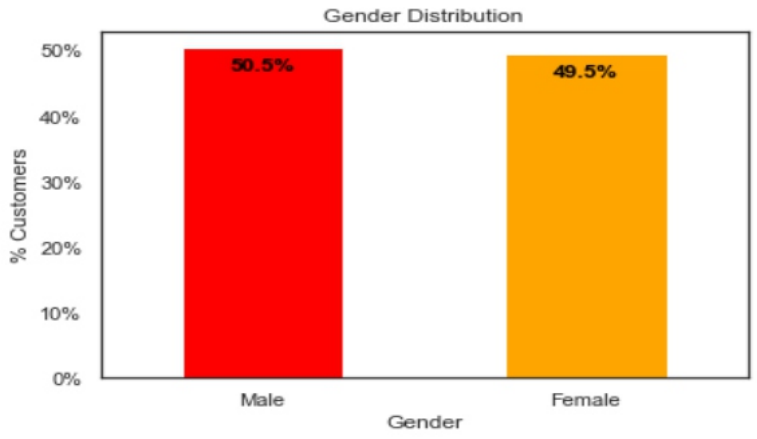


Figure 2. Gender Distribution

b. Most of the customers are youngsters rather than senior citizens as depicted in graph (figure 3).

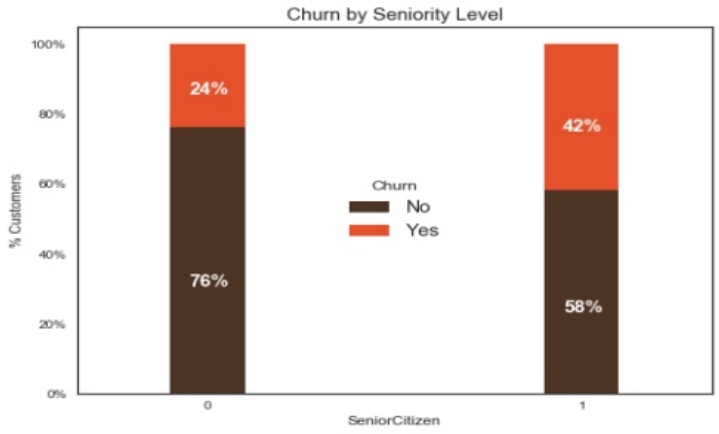


Figure 3. Customer Churn based on Seniority Level

c. Figure 4 depicts 48 percentage of the customers partner dependent, while 30 percentage have dependents. Fascinatingly, only about half of the customers who have a partner have a dependent, while the other half do not. Furthermore, a majority (80%) of the customers without a partner do not have dependents.

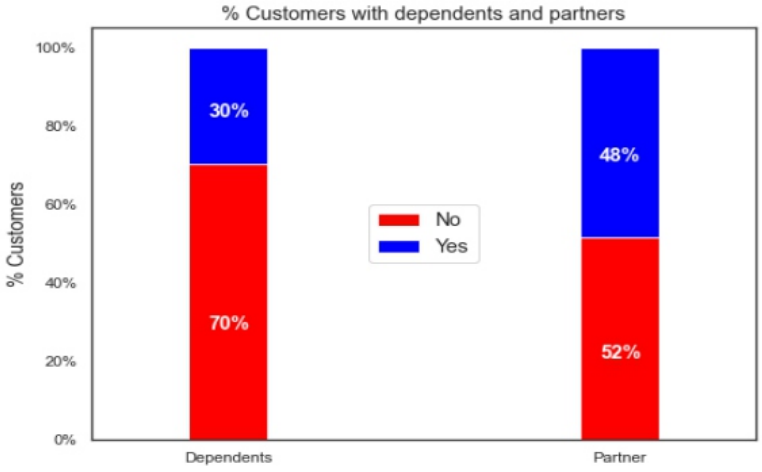


Figure 4. Customer Churn based on Dependents and Partners

d. Customer Tenure and Account Information: From the histogram as in figure 5, we can see that many customers have been with the telecom company just for a month, while many others have been with the company for about 72 months. Different contracts may apply to different customers. Due to this, depending on the contract they are in, it might be easier or harder for customers to stay or leave the telecom company.

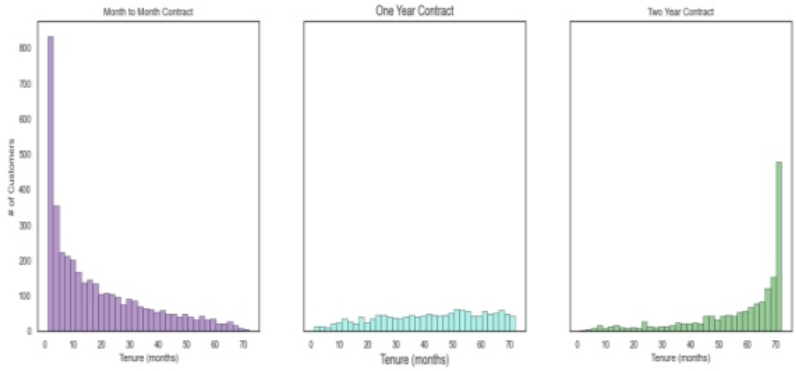


Figure 5. Customer Churn based on tenure and year of contract

Contrary to popular belief, the typical monthly contract usually lasts between one and two months, while the typical two-year contract lasts around seventy months. People who sign longer contracts and stick with the company longer show customer loyalty to the company.

e. Customer Distribution: The graph depicts the distribution of services used by customers. From the graph of the relation between monthly and total charges, the total charges are directly proportional to customers' monthly bills. Finally, the rate of churn is depicted in figure6.

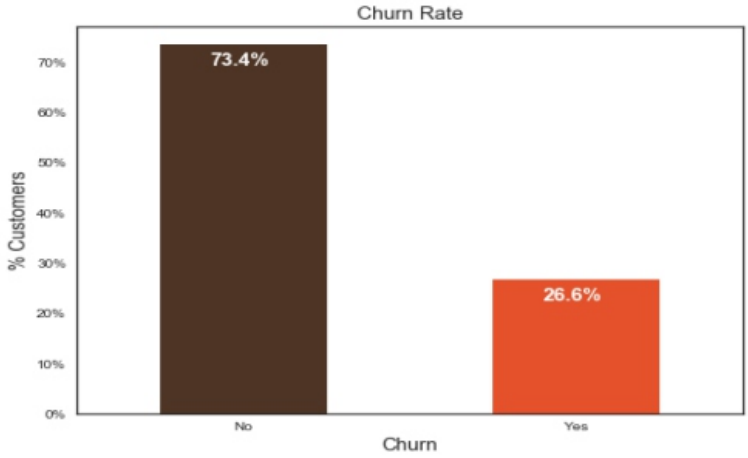


Figure 6. Customer Churn Rate Visualization

f. Churn by Monthly Charges vs Total Charges: customer's churn directly proportional to monthly charges; and churn rate is inversely proportional to total charges depicted in figure

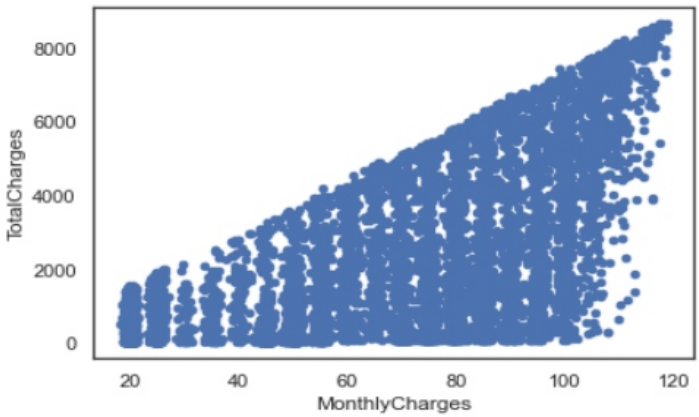


Figure 7. Monthly vs Total Charges

Model Design and Analysis

Machine Learning Models

ML (machine learning) is a form of artificial intelligence, in which software applications make better predictions without being explicitly programmed. Here, historical data are trained to predict the test result or output [11]. There are typically three types of ML algorithms Supervised, Un-supervised and Reinforcement Learning algorithms.

Logistic regression:

Logistic regression is a supervised learning approach for predicting a target variable's probability. Since the variable of interest is dichotomous, there are only two possibilities. In other words, the dependent variable comprises binary data that can either be coded as 0 (for failure) or as 1 (for success). As a function of X, logistic regression predicts the value of P(Y=1). Thisis one of the simplest ML algorithmsthat can applied to a variety of classification problems such as diabetes prediction, cancer detection, fraud detection, spam detection and many more.

Random Forest:

The Random Forest algorithm is both a classification and mregression learning algorithm that is used for supervised learning. This method is mainly used to solve classification problems. Forests can be thought of as a forest of trees, and a forest that is more robust has more trees. In the same way, random forests create decision trees using data samples and then obtain their predictions. Ultimately, they vote on which solution is the best. By averaging the result, it allows us to reduce over-fitting to a minimum.

AdaBoost:

It is short for Adaptive Boosting - is one of many Ensemble Methods used to improve neural networks.

The Adaptive Boosting method assigns higher weights to incorrectly classified instances since the weights are re-assigned. A boost is used in supervised learning to reduce bias and variance. The system is based on the principle sequential growth of arners. In all cases, except the first, the subsequent learners are grown from the previous ones. To putit simply, weak learners are transformed into strong ones.

GradientBoost:

Gradient Boosting Machines combine predictions from multiple decision trees into a final prediction. It is important to remember that gradient boosting machines only use weak learners. To select the best split in every decision tree, different nodes take into account different factors. In other words, each tree is different, and thus it can capture different signals from the data. Additionally, each new tree corrects previous errors. As a result, every subsequent decision tree builds upon the mistakes of the previous trees. An algorithm for gradient boosting machines builds trees sequentially in this manner.

XGBoost:

It is a gradient boosting-based ensemble Machine Learning algorithm based on decision trees. In unstructured data prediction problems (images, text, etc.); it is suitable for solving regression and classification problems, ranking problems, and user-defined prediction problems.

ANN:

Artificial Neural Network (ANN) can be considered the core element of Deep Learning. In addition to their versatility, adaptability, and scalability, ANNs are also suitable for handling large datasets and highly complex Machine Learning problems, like image classification, speech recognition, or video recommendation. In ANN algorithms, the aim is to create the most minimal error function possible by selecting the optimal weights and bias terms [12]. thought of as the most sophisticated version of Machine Learning.

Result Analysis

Performance Metrics Co-relation Matrix: A correlation is a description of how variables are related to one another. Feature variables such as these can be used as input for forecasting our target variable.A correlation is a statistical technique that determines how one variable moves/changes in relation to another. A correlation matrix is a table presenting multiple variables and their 'correlations'. Rows and columns in this matrix represent variables, and each value in the matrix represents a correlation coefficient between variables depicted in figure 7.

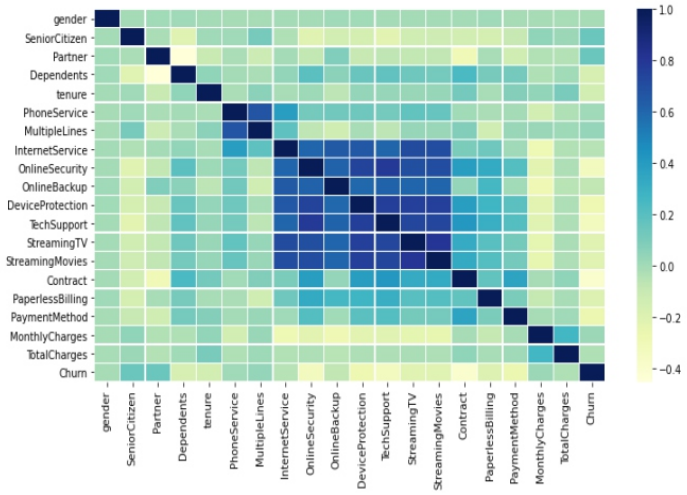


Figure 8. Confusion Matrix

Confusion Matrix shows how many True Positives/True Negatives and False Positives/False Negatives there are in a prediction.

TP: Number of customers who will actually default is alsopredicted as defaulting

TN: Number of customers not expected to default is also reported as non-default

FP: Number of customers who are predicted to default but won't actually default

FN: Number of customers predicted to default but actually defaulting

A telecom company needs to understand which customers will default. Therefore, we should keep the number of False Positives (FP's) as low as possible, as this will predict that the riskier s will not be too risky. All of these factors should be considered as we assess every classification model. ‘

Accuracy' is the easiest performance metric to grasp and issimply the ratio of rightly predicted observations to the total observations. It is easy to assume that the best model is the one with high accuracy. It is true that accuracy is an important measure, but only if you have symmetric datasets with relatively equal values of false positives and false negatives. As a result, you have to evaluate the model'sperformance by looking at other parameters.

Accuracy = TP+TN/TP+FP+FN+TN

Precision is the percent of positively predicted observations among all predicted positive observations. This metric answers the question, how many of all passengers th labeled as survivors actually survived? Precision is related to a low false positive rate.

Precision = TP/TP+FP

Sensitivity (Recall) - Recall is measured by how many of the correctly predicted positive observations have actually occurred in the class - yes. How many passengers did welabel from all of those that truly

survived? is the question that can be answered here.

Recall = TP/TP+FN

An F1 score is calculated by summing Precision and Recall. This score considers both false positives and false negatives. Although F1 is not intuitively as easy to understand as accuracy, it is usually more useful than accuracy if your classes are unevenly distributed. True positives and false negatives have similar costs when it comes to accuracy. Precision and recall should be considered along with the cost of false positives and false negatives.

F1-Score = 2 TP / (2TP + FP + FN)

Model Comparison

Machine learning algorithms were used on the dataset to perform several experiments on the proposed churn model. The results were observed pertaining to precision, recall, f1-score, and accuracy values. Table 1 presents the details of the performance results of all the models based on the metrics. As per the result, XGBoost model has the highest accuracy value of 81.14%, followed by RF, AdaBoost, GBM, LR and ANN. But, for the rest of the metrics like precision, f1-score and recall ANN has a higher value followed by three of the boosting models (XGBoost, AdaBoost, GBM) with approximately the same result, then LR and RF.

Table 1. Model Analysis

ML Models	Accuracy	Precision	Recall	F1-Score
GradientBoost	80.41%	0.66	0.58	0.59
AdaBoost	80.59%	0.67	0.57	0.58
XGBoost	82.20%	0.66	0.45	0.56
ANN	79.98%	0.89	0.84	0.85
LR	80.29%	0.68	0.56	0.60
RF	81.10%	0.66	0.49	0.56

CONCLUSION

As the telecommunications industry continues to grow,the problem of customer churn has significant growth as well. Retaining customers is a critical challenge in the telecommunications industry since it reduces customer churn through increased customer satisfaction. The use of predictive analytics can help tackle this threat by identifying vulnerable customers and implementing customer-centricretention measures. The proposed prediction analysis can be solved with the help of machine learning models. The paper describes the problem of churn and the importance of preventing the same. The paper investigated

through increased customer satisfaction. The use of predictive analytics can help tackle this threat by identifying vulnerable customers and implementing customer-centric retention measures. The proposed prediction analysis can be solved with the help of machine learning models. The paper describes the problem of hurn and the importance of preventing the same. paper investigated the realm of machine learning hods and applied them to the dataset. By modeling and testing, ANN and XGBoost models were found to out-perform other models in terms of precision,recall, f1-score and accuracy. We can extend the same for various other datasets from the telecom department and implement methodologies to achieve better results in future work. Big-data analytics with machine learning approach can also be implemented for the datasets.

REFERENCES

[1] Suguna, R., M. Shyamala Devi, and Rincy Merlin Mathew."Customer churn predictive analysis by componentminimization using machine learning." International journalof innovative technology and exploring engineering 8.8(2019): 3229-3233.

[2] Vafeiadis, Thanasis, et al. "A comparison of machine learning techniques for customer churn prediction." Simulation Modelling Practice and Theory 55 (2015): 1-9.

[3] Ahmad, Abdelrahim Kasem, Assef Jafar, and Kadan Aljoumaa. "Customer churn prediction in telecom using machine learning in big data platform." Journal of Big Data 6.1 (2019): 1-24.

[4] Ahn, Hyunchul, et al. "Facilitating cross-selling in a mobile telecom market to develop customer classification model based on hybrid data mining techniques." Expert Systems with Applications 38.5 (2011): 5005-5012.

[5] Kumar.N and Naik.C. "Comparative analysis of machine learning algorithms for their effectiveness in churn prediction in the Telecom industry". International research journal of engineering and technology, (2017), vol 4, iss 8, pp: 485-489.

[6] Abou el Kassem, Essam, et al. "Customer Churn Prediction Model and Identifying Features to Increase Customer Retention based on User Generated Content", (IJACSA) International Journal of Advanced Computer Science andApplications, (2020), 11. 5.

[7] M. A. Hassonah, A. Rodan, A. Al-Tamimi and J. Alsakran, "Churn Prediction: A Comparative Study Using KNN and Decision Trees," 2019 Sixth HCT Information Technology Trends (ITT), (2019), 182-186, doi:10.1109/ITT48889.2019.9075077.

[8] Almuqren, L.A.; Moh'd Qasem, M.; Cristea, A.I. Using Deep Learning Networks to Predict Telecom Company Customer Satisfaction Based on Arabic Tweets; ISD: Tolerance, France,(2019).

[9] Mahajan, Vishal, and Renuka Mahajan. "Variable Selection of Customers for Churn Analysis in Telecommunication Industry." International Journal of Virtual Communities and Social Networking (IJVCSN) 10.1 (2018): 17-32..

[10] Hudaib, Amjad, et al. "Hybrid data mining models for predicting customer churn." International Journal of Communications, Network and System Sciences 8.05 (2015): 91-96.

[11] Labhsetwar, Shreyas Rajesh. "Predictive Analysis of Customer Churn in Telecom Industry using Supervised Learning." ICTACT Journal on Soft Computing 10.2 (2020): 2054-2060..

[12] Hota, L., Dash, P. "Comparative Analysis of Stock Price Prediction by ANN and RF Model", Computational Intelligence and Machine Learning, 2.1, (2020), 1-9

Faster-RCNN Based Deep Learning Model for Pomegranate Diseases Detection and Classification

Aziz Makandar1 , Syeda Bibi Javeriya2*

1 Professor, Department of Computer Science, KSAW University Vijayapura, Karnataka, India.

2 Research Scholar, Department of Computer Science, KSAW University Vijayapura, Karnataka, India.

A B S T R A C T

India is the largest producer of pomegranates in the world which earns a high profit. However, due to atmospheric conditions such as temperature variations, climate, and heavy rains, pomegranate fruits become infected with various diseases, resulting in agricultural losses. The two most common diseases seen in the Karnataka region are bacterial blight and anthracnose, both of which cause a significant production loss. This paper has detected and classified these two diseases by extracting knowledge from custom trained models using Deep Learning. To overcome the traditional methods, Faster-RCNN helps us to do better object detection.

Keywords Anthracnose, Deep Learning, Faster-RCNN, Object detection, Tensorflow Bacterial blight.

INTRODUCTION

Asian countries have been manufacturing pomegranates to a larger extent. The exports of pomegranates are growing year by year. Over the past few years, agriculture has swung and is turning into a supply of financial benefit generation. In India, 11.0 lakh tones of pomegranate are produced on 1.5 lakh hectares of land. Maharashtra is India's leading pomegranate producer, India grant 2/3 rd. of the total.

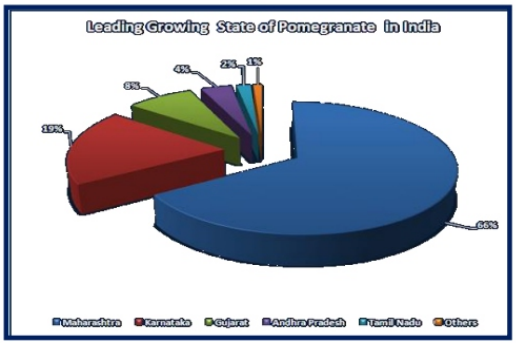


Fig -1: Productivity of Leading Pomegranate Growing States in India.

Importance of Disease Detection in Fruits:

India is an agricultural dependent country as it stands second largest producer of fruits and there is a high demand for quality of fruits in market. The cultivation of fruits faces threat of several diseases caused by pest, micro-organs,weather conditions, soil profile and deficiency of nutritionetc. Which leads to significant reduction in crops when it comes to fruits preservation from diseases diagnosis is

very essential to enhance crop production and thus, improve the economic growth [12].

Two Most Common Diseases in Pomegranate Are: 1)

Bacterial blight: Dark color irregular spots appear on fruits, and the leaves start dropping, and fruit crack appears in V and L shape and spreads rapidly throughout the farm and cause severe destruction. 2)Anthracnose: it's a kind of fungithat causes irregular brown spots and this disease also leads to severe fruit loss. In the present situation, Farmers in India lack knowledge about how to use pesticides properly; as a result, a proper agriculture system would assist farmers in crop management and decision-making using advanced technology. The intelligent system will detect and diagnose diseases in the fruits for their purpose, and it will restrict the growth of the diseases. Researchers have developed machine learning technology to solve e problems of the farmers [1]. Deep learning is one of the most commonly used subfields ofmachine learning. It helps in the prediction of various problems and provides solutions [2][3].

LITERATURE SURVEY

One of the important research areas is the automated method for detecting disease-affected fruits, as it offers numerous benefits in terms of fruit preservation. Although a lot of research is done in this area, Artificial Intelligence is rarely used for this purpose. To detect multi-fruit classification, the authors proposed a Deep learning approach that uses a faster R-CNN. Fruitssuch as mango and pitaya are used as ingredients. The dataset was actual data obtained from a farmer during harvest time, and it was divided into two classes for object detection training: mango and pitaya. On the TensorFlow platform, authors used the MobileNet model. In this study, they achieved 99 % accuracy rate [4]. In this paper, using plant leaf photos, the authors propose a deep-learning-based approach for detecting leaf diseases in a variety of plants. They identified and developed deep learning methodologies for good results, and they considered three major detector families: The Faster Region-based Convolutional Neural Network (Faster R-CNN), the Region-based Fully Convolutional Network (R-FCN), and the Single Shot Multibox Detector (SSD). The proposed system capable of identifying various types of diseases and dealing with complex scenarios from within a plant's area [5]. In a deeper analysis using deep learning techniques, Rismayati and Rahari SN [6] investigated CNN's sorting ofsalak fruits. authors used neural networks to analyze the salak image and classification scheme in a region of interest (RoI).With 3x5x5, they make six filter layers in the first layer. The second layer generates 18 filters size of 6x3x3. The accuracy rate was 81.45%. To solve image classification problemsfaster, the R-CNN and Quick R CNN methods are used. This method was chosen because it has the highest level of precision in a variety of tests at 1 frame per second (Frame Per Second).

Table -1: Comparison table of various versions of RCNN.

	R-CNN	Fast R-CNN	Faster R-CNN
Time taken for per image	50 seconds	2 seconds	0.2 seconds
Speed up	1x	2x	250x

PROPOSED METHOD

In this article, we propose a system for detecting pomegranate diseases like anthracnose andbacterialblight via TensorFlow for object detection on a Faster R-CNN. Based on the literature survey, we create our own dataset. For each classifier, i.e., each object label, we collected almost 200-300 images. We used online tool for Image Annotation process re we have uploaded all our dataset, and set the object names (Classifiers) as anthracnose and bacterialblight and used rectangle for creating xml files as annotation directories. After labeling images or Annotations we converted theminto CSV (train.csv, test.csv) format because of tensorflow specifications. CSV files are converted into ord formatto enhace the training. Once the training has been completed successfully, the protocol buffer(.pb) file is generated with the python inference graph. This graph file can create a user interface on Android or a web application in which a camera is used to detect an object using the trained TensorFlow model.

Convolutional neural network

In [15] CNN's architecture as consisting of an input layer followed by a Conv layer. The dimensions of the conv layer vary depending on the data and problem, so they must be adjusted accordingly. There is an activation layer after the Conv Layer, which is normally ReLU because it produces better performance. A pooling layer is used to minimise the scale after certain Conv and Relu combinations. The flattening layer is used to flatten the input for the completely connected layer after some variation of previously established architecture. The third layer, after the first two, is the output layer.

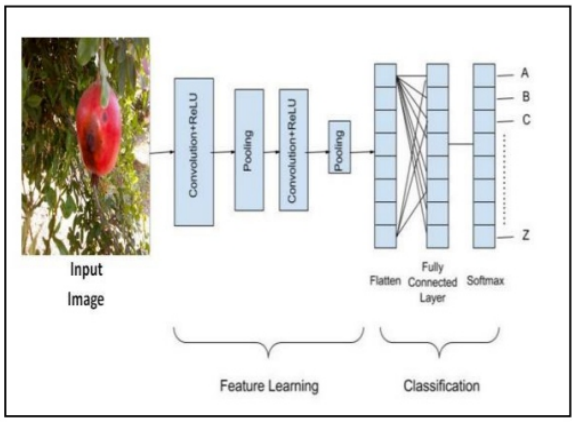


Fig -2: CNN's architecture

Faster Region-Based Convolutional Neural Network (Faster R-CNN)

Faster R-CNN is a Convolutional Neural Network-based object recognition architecture that uses a Region Proposal Network (RPN). It is commonly used in Deep Learning and Computer Vision and is considered one of the most effective object detection architectures.

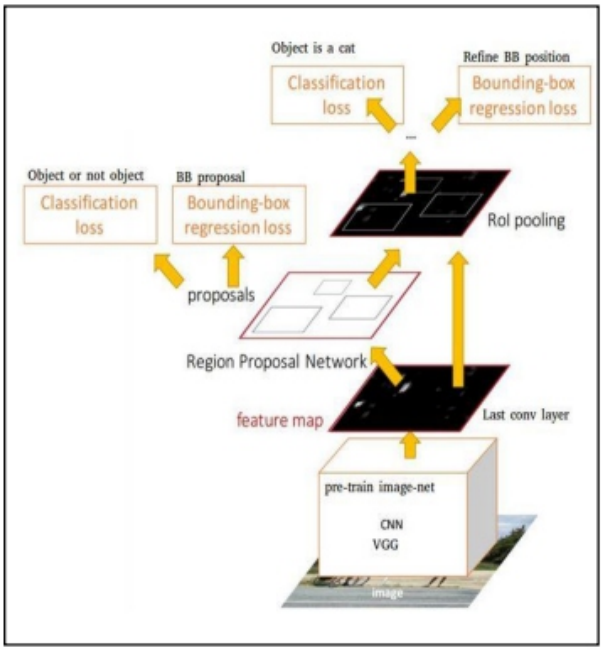
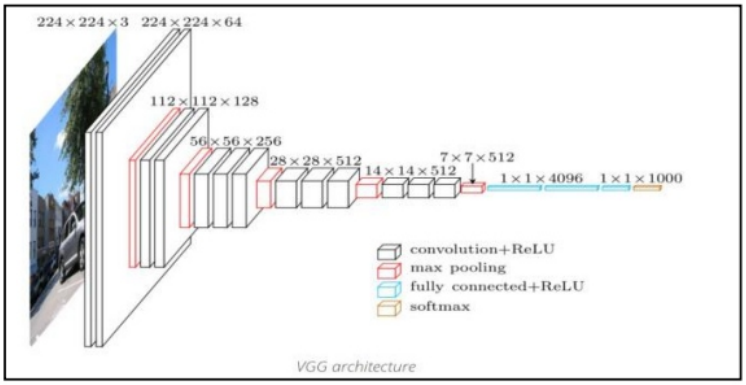


Fig -3: Faster RCNN

It takes an image and sends it to the ConvNet, which creates feature maps for it. Use the Region Proposal Network (RPN) to generate object proposals from these feature maps,and use the ROI pooling layer to make all of the osals the same size. Finally, submit these suggestions to a fully linked layer in order to define and predict the bounding boxes of the image.

(Visual Geomerty Group) VGG 16

In [14] It's a 16-layer deep network that's used for feature extraction. We can load a pre-trained version of the network that can be trained on millions of images from the ImageNet database. The network has been pre-trained to classify images into 1000 different object categories.



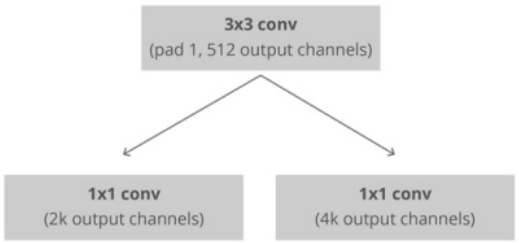


Fig -4: VGG 16 Architecture

VGG16 will eliminate the pre-trained network's bottleneck (classifier) layer. Then, with the exception of the last few convolutional layers, all weights are frozen, and we attach our own classifier with a very low learning rate

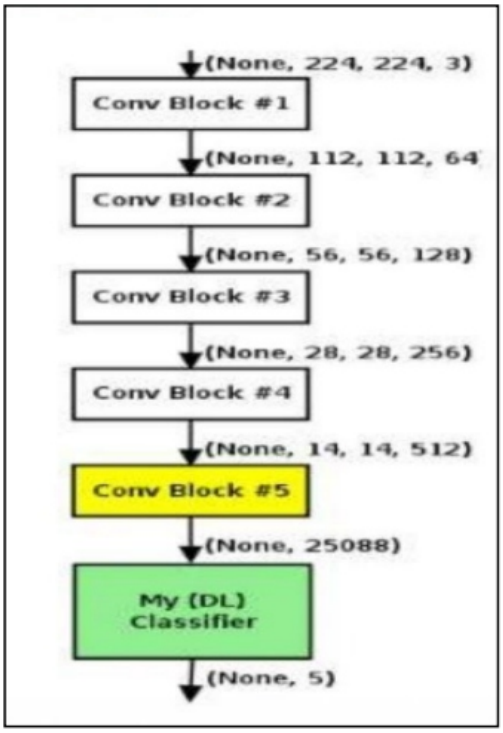


Fig-5: VGG16 Model

Region Proposal Network (RPN)

The area proposal network will take all the anchors (reference boxes) and produce two different outputs for each of the anchors, resulting in a list of good object proposals. The first is a "objectness" score, which indicates how likely the anchor is to be an entity; RPN is unconcerned about the type of object. We'll use this objectness score to weed out the bad predictions in the second step. The bounding box regression is the second production, which is used to modify anchors to match the items that are being predicted. The function map, which is convoluted returned by the network as an imput, is used by RPN to implement in a completely convolutional way. With 512 channels and a 3x3 kernel dimension, the convolutional layer is used. Then, using a 1x1 kernel, we'll have two parallel layers of convolution, with the number of channels determined by the number of anchors per point. We get two performance predictions per anchor for classification. Its score isn't an object (background), but it is an object (foreground).Adjustment layer for regression or bounding box. We generate four predictions: Δxcenter, Δycenter, Δwidth, and Δheight, which we combine with theanchors to formfinal proposals We have a strong set of object proposals using the final proposal co-ordinates and their"objectness rating."

Anchors

The network generates the maximum number of k- anchor boxes for each sliding window. For each of the different sliding positions in the image, the default value of k=9 (3 scales of (128*128, 256*256, and 512*512) and 3 aspect ratios of (1:1, 1:2, and 2:1) is used. As a result, we get N = W* H* k anchor boxes for a convolution feature map of W * H. These region suggestions were then passed through an intermediate layer with 3*3 convolution and 1 padding, as well as 256 (for ZF) or 512 (for VGG-16) output channels. This layer's output is passed through two 1*1 convolution layers, the classification layer, and the regression layer. The classification layer has 2*N (W * H * (2*k) output parameters, while the regression layer has 4*N (W * H * (4*k) output parameters (denoting the coordinates of bounding boxes) (denoting the probability of object or not object).

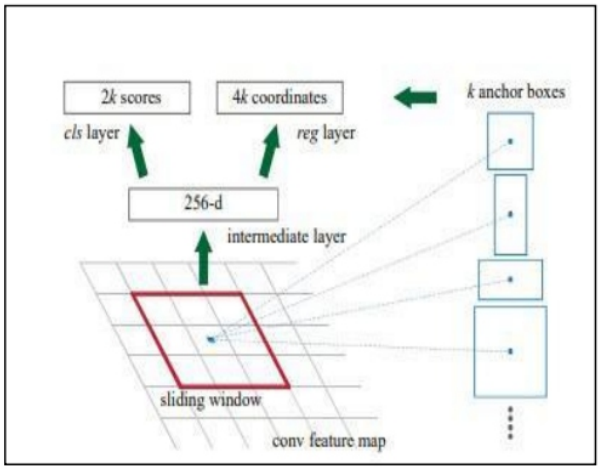


Fig -6 Anchors.

ROI Pooling

Region of interest pooling (also known as RoI pooling) is a popular operation in convolutional neural network object detection tasks. The problem of a fixed image size requirement for an object detection network is solved by ROI pooling. By doing max-pooling on the inputs, ROI pooling creates fixed-size function maps from non-uniform inputs. The number of output channels is equal to the number of input channels for this layer.

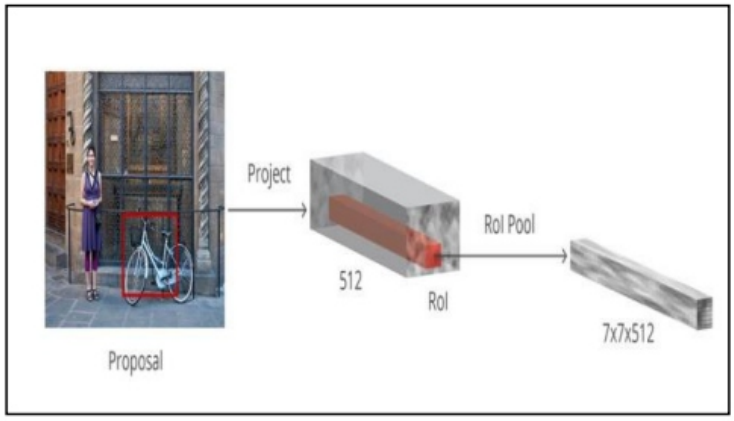


Fig -7 Region of interest pooling

APPROACH

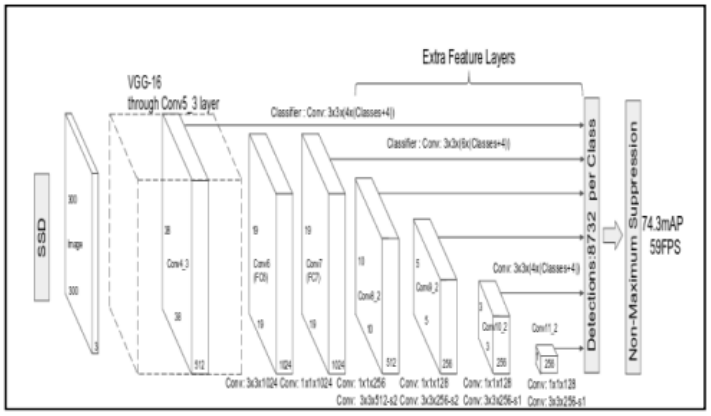


Fig-8: SSD Architectur

This project's network is focused on single-shot detection (SSD). Normally, the SSD begins with a VGG [8] model that has been transformed to a completely convolutional network. Then we add some additional convolutional layers to better manage larger subjects. A 38x38 feature map (conv4 3) is generated by the VGG network. The additional layers result in function maps that are 19x19, 10x10, 5x5, 3x3, and 1x1. As seen in the following diagram, both of these feature maps are used to dict bounding boxes at different scales (later layers are responsible for larger objects).

IMAGE ANNOTATION

PASCAL VOC [9] offers structured image datasets for object type recognition as well as a common collection of resources for accessing the datasets and annotations. Our PASCAL VOC dataset has two classes and a task that is based on it. The PASCAL VOC dataset is well-marked and of good quality, allowing for evaluation and comparison of various approaches. The PASCAL VOC dataset has a smaller amount of data than the ImageNet dataset, making it ideal for researchers evaluating network programmes. As shown in the following figure, our dataset is also based on the PASCALVOC dataset norm.

```
<?xml version="1.0"?>
- <annotation>
    <folder>images</folder>
    <filename>3p.jpg</filename>
    <path>images/3p.jpg</path>
  - <source>
      <database>Unknown</database>
    </source>
  - <size>
      <width>300</width>
      <height>168</height>
      <depth>3</depth>
    </size>
    <segmented>0</segmented>
  - <object>
      <name>Bacterialblight</name>
      <pose>Unspecified</pose>
      <truncated>0</truncated>
      <difficult>0</difficult>
    - <bndbox>
        <xmin>55.000003814697266</xmin>
        <ymin>0</ymin>
        <xmax>257.0000305175781</xmax>
        <ymax>167</ymax>
      </bndbox>
    </object>
</annotation>
```

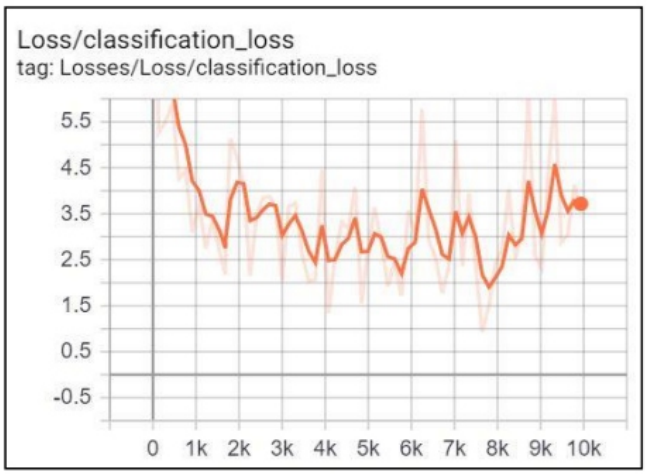
Fig -9 Image Annotation

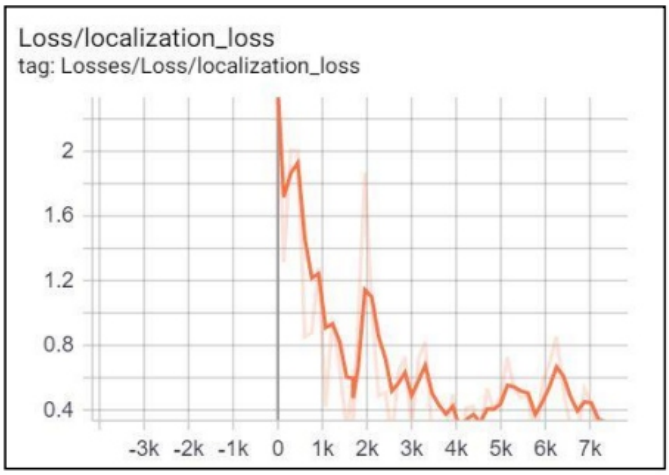
Fig -10 Labeling Tool

Fig -4: Table example of the labeled dataset.

filename	width	height	class	xmin	ymin	xmax	ymax
1 (1).jpg	1024	974	anthracno	65	55	961	973
1 (10).jpg	300	400	anthracno	72	163	244	357
1 (11).jpg	896	504	anthracno	102	55	725	462
1 (12).jpg	1024	576	anthracno	129	44	855	576
1 (13).jpg	450	300	anthracno	20	52	214	265
1 (13).jpg	450	300	anthracno	214	51	424	240
1 (14).jpg	450	300	anthracno	115	12	420	300
1 (15).jpg	480	360	anthracno	130	43	387	279
1 (16).jpg	800	600	anthracno	194	156	618	547
1 (17).jpg	1300	956	anthracno	4	132	217	424
1 (17).jpg	1300	956	anthracno	221	379	539	757
1 (17).jpg	1300	956	anthracno	509	197	992	749
1 (18).jpg	1280	720	anthracno	45	27	644	683
1 (18).jpg	1280	720	anthracno	553	176	1246	720
1 (18).jpg	1280	720	anthracno	779	9	1280	428
1 (19).jpg	3184	3184	anthracno	347	297	2776	2922

RESULT AND DISCUSSION





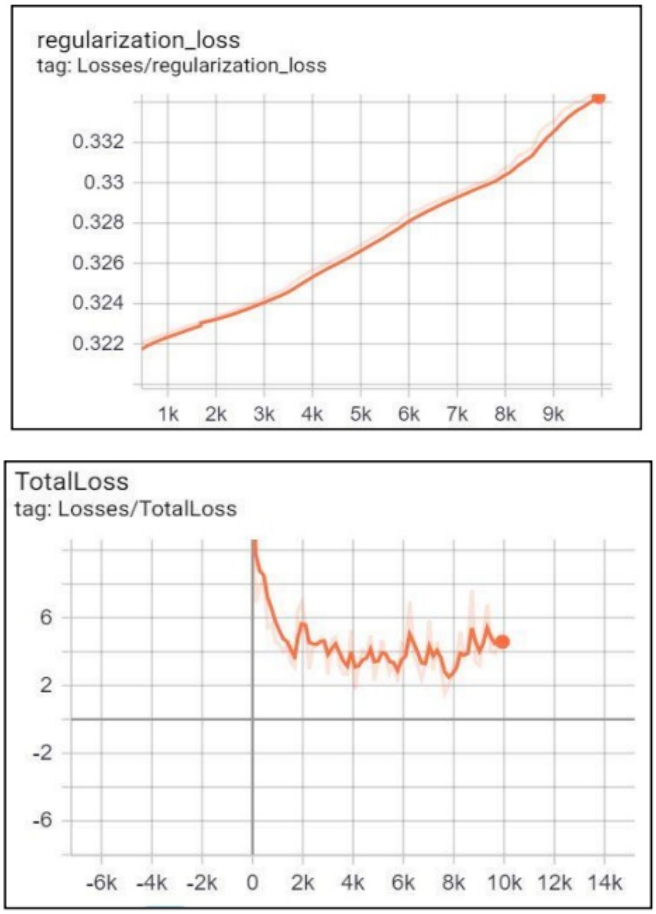
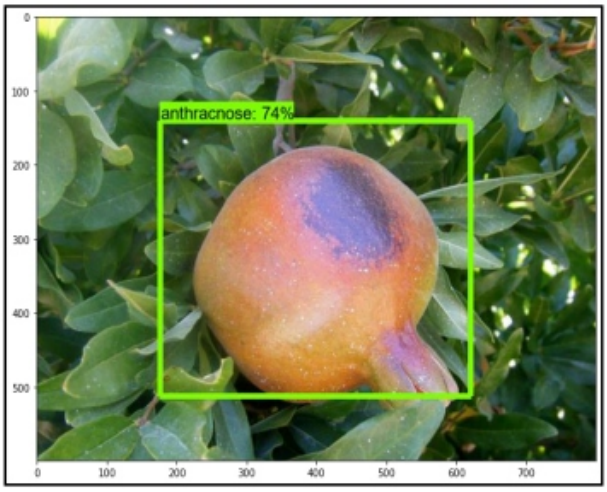


Fig-11: Total Losses of Faster-R-CNN

The number and consistency of the dataset will influence the neural network performance accuracy after the images are trained [10]. Deep learning approaches [11] are growing every day in popularity it enables rapid and efficient solutions, especially in the analysis of large amounts of data. This study used a custom dataset to identify pomegrana diseases such as anthracnose and bacterialblight for deep learning applications. Tensorflow played a major role in this.



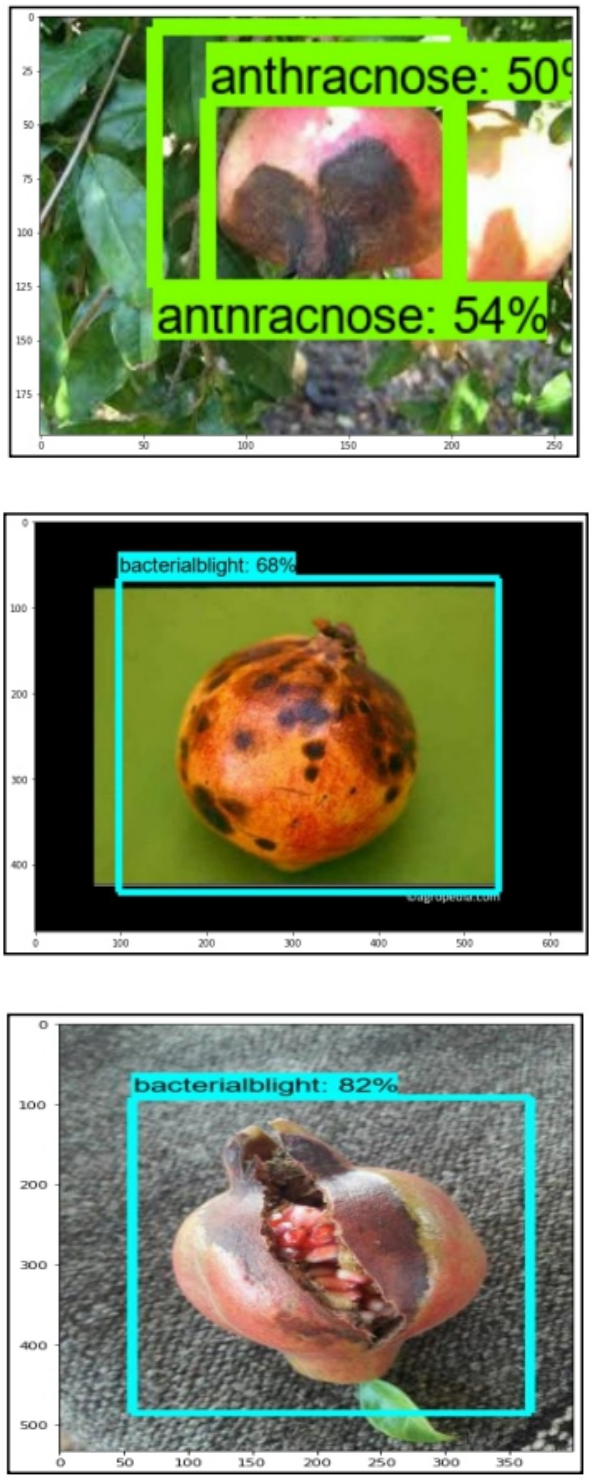


Fig -12: Experimental results.

CONCLUSION

The proposed system is able to detect the diseases in pomegranate and can able to classify them into different categories here we have identified two kinds of diseases anthracnose and bacterialblight . In this study we considered deep learning methodology based on Faster RCNN modelwhich gave an

accurate and efficient object detection system.

The goal for the future is to figure out how to overcome the issue of low image resolution causing detection failures. Another choice is to apply this approach to crops other thanpomegranates.

REFERENCES

[1] L. Ma, S. fadillah Umayah, S. Riyadi, C. Damarjati, and N. A.Utama, "Deep Learning Implementation using Convolutional Neural Network in Mangosteen Surface Defect Detection," no. November, pp. 24–26, 2017.

[2] K. N. Ranjit, H. K. Chethan, and C. Naveena, "Identification and Classification of Fruit Diseases," vol. 6, no. 7, pp. 11–14, 2016.

[3] H. Jang, H. Yang, and D. Jeong, "Object Classification using CNN for Video Traffic Detection System," Korea-

[4] Japan Jt. Work. Front. Comput. Vis., no. 1, pp. 1–4, 2015.

[5] Hasan Basri, Iwan Syarif and Sritrustra Sukaridhoto,

[6] "Faster R-CNN Implementation Method for Multi-Fruit Detection Using Tensorflow Platform," 2018 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC), pp.978-1-5386-8079-7,2018.

[7] M. Akila, P. Deepan, "Detection and Classification of Plant Leaf Diseases by using Deep Learning Algorithm," Volume 6, Issue 07, International Journal of Engineering Research &Technology (IJERT), ISSN: 2278-0181.

[8] Rismiyati and S. N. Azhari, "Convolutional Neural Network implementation for image-based Salak sortation," in Proceedings - 2016 2nd International Conference on Science and TechnologyComputer, ICST 2016, 2017, pp. 77–82

[9] Tensorflow Framework www.tensorflow.org

[10] R. Girshick. Fast R-CNN. arXiv:1504.08083, 2015

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. IJCV, 2010.

[12] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks, vol. 61, pp. 85–117, Jan. 2015.

[13] M. Abadi et al., "TensorFlow: A System for Large-Scale Machine Learning TensorFlow: A system for large-scale machine learning," in 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI' 16), 2016, pp. 265–284.

[14] Aziz Makandar,Syeda BibiJaveriya,"Survey On Fruit Disease Detection Using Image Processing Techniques", International Conference On Artificial Intelligence and Soft Computing (ICAISC-2021),ISBN:978-93-88929-53-0.

[15] Hands-On Machine Learning with Scikit-Learn,Keras,and Tensorflow: Concepts, Tools, and

Techniques to Build Intelligent Systems 2nd Edition by Audrelien Geron.

[16] https://tryolabs.com/blog/2018/01/18/faster-r-cnn-down-the-r abbit-hole-of-modern-object-detection/

[17] https://towardsdatascience.com/a-comprehensive-guide-to-co nvolutional-neural-networks-the-eli5-way-3bd2b1164a53.

[18] Ranjit K N,Chethan H K,Naveena C ,"Identification andClassification of Fruit Diseases", Int. Journal of Engineering Research and Application,pp. 11-14,2016.

[19] Dakshayini Patil,"Fruit Disease Detection using Image Processing Techniques", International Journal for Research in Engineering Application & Management (IJREAM),pp.2454-9150,2018.

[20] Liu, W., Anguelov, D, Erhan, D., Szegedy, C., Reed, S.; Fu,C.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision ECCV, Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.

[21] Ren, S., He, K., Girshick, R., Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans. Pattern Anal. Mach. Intell. 2016, 39, 1137–1149

Deception Recognition Method Based on Machine Learning

Siddth Kumar Chhajer 1 *, Rudra Bhanu Satpathy 2
1MBA,Marketing, St. Peter's University, Chennai, Tamil Nadu, India.
2 M.Tech, Electrical, Electronics and Communications Engineering, St. Peter's University, Chennai, Tamil Nadu,

A B S T R A C T

Money extortion is a developing issue with far results in the budgetary business and keeping in mind that numerous procedures have been found. Information removalis effectively functional to back records to computerize the investigation of colossal volumes of multifaceted information. Information removal has additionally assumed a notable job in the location of Visa deception in online exchanges. Deception recognition in credit card is an information mining issue, it gets testing because of two significant reasons–first, the profiles of typical and deceitful practices change much of the time and besides because of the reason that Mastercard extortion informational collections are exceptionally slanted. This paper examines and analyze the presence of the Decision tree, Random Forest, SVM, and strategic regression on exceptionally slanted credit card extortion information. Dataset of Visa exchanges is sourced from European cardholders containing 274,335 exchanges. These function are usedto crude and preprocessed information. The presentation of the strategies is assessed dependent on exactness, affectability, explicitness, accuracy. The outcomes demonstrate the ideal accuracy for logistic regression, decision tree, Random Forest and SVM classifiers are 96.8%, 94.4%,99.5%, and 96.6%.

Keywords Credit Card, Deception Recognition, Decision Tree, and Support Vector Machine.

INTRODUCTION

Money related extortion is a emergingconcern with broad results in the administration, corporate associations, fund industry, In the reality of extremely dynamic dependency on online innovation, increased visa transactions have been praised, but credit card deception has also escalated as on the network and disconnected trade. When credit card transactions become a much-reaching installment system, the center has been tasked with ongoing technological ideologies to resolve the problem of Visa extortion. There are various extortion discovery and programming agreements that forestall fakes in organizations such as credit card, retail, web-based business, security, and businesses.

The information mining procedure is one prominent and well-known technique utilized in tackling credit extortion discovery issues. It is hard to be completely sure of the true intent and validity behind a request or transaction. In actuality, the best convincing solution is to check for possible verifications of deception from the available knowledge using numerical calculations. Extortion discovery in Visa is the real way to identify certain transactions that are deceptive in two types of legitimate class and deception class transactions, a few methods are designed and implemented to understand credit card deception, such as hereditary recognition calculation, counterfeit neural system visit thing set mining, AI calculations, relocating feathered creatures advancement calculation, near examination of strategic

decision tree and irregular woodland is done. [1]Mastercard deception recognition is a well-known yet additionally a troublesome issue to unravel. Firstly, due to the question of having only a small measure of knowledge, a credit card attempts to arrange an example for the dataset. In addition, there could be several parts in the database with deceptionster truncations that also match an indication of credible behavior. The problem has a range of specifications, in fact firstly, informational indexes are not effectively open for open and the aftereffects of looks into are frequently protected and managed, rendering the results out of reach and attempting to compare the integrated models for that purpose. Datasets of legitimate knowledge in the documentation in previous inquiries are not referenced. Furthermore, the development of techniques is gradually disturbing in that security forces confinement to the exchange of thoughts techniques in discovery of deceptions, and particularly in credit and debit cards extortion identification. [2]

In addition, the knowledge databases are constantly developing, and changing process issues of constantly exceptional traditional and deceitful activities that are the legitimate interaction in the past may be manipulation throughout the present or another way around. This paper assesses four propelled information mining draws near, Decision tree, support vector machines, Logistic regression, and arbitrary woodlands, and afterward a collative correlation is made to assess what model performed best. [3]

Mastercard exchange datasets are seldom accessible, profoundly imbalanced, and slanted. Ideal element (factors) decision for the models, reasonable measurement is the most significant piece of information mining to assess the execution of methods on slanted credit card extortion information. Various difficulties are related to Visa discovery, to be specific deceitful conduct profile is dynamic, that is fake exchanges will, in general, appear as though real ones, Credit card extortion location execution is incredibly influenced by the kind of examining approach utilized, the decision of factors and recognition strategy utilized. Toward the finish of this paper, decisions about aftereffects of classifier evaluative testing are made and ordered. From the prosecutions, the result that has already been concluded is that regression analysis has an accuracy of 96.8 percent, while Classifier demonstrates an accuracy of 96.6 percent and the Decision tree demonstrates an accuracy of 94.4 percent, while Random Forest achieves the best results with an accuracy of 99.5 percent. The results obtained in this way suggest Random Forest reveals the most consistent and efficient accuracy of 99.5 percent in the issue of Artificial Intelligence bank card deception identification with data set gave by Artificial Intelligence (AI).

There was also a desire to push forward from a whole new perspective. Attempts were made in the case of deception transactions to enhance alert-feedback interaction. In the case of a deceitful transaction, the authorized program would be informed, and a request would be submitted to repudiate the constantoperation. The Artificial Hereditary Algorithm countered deception from a different direction, unique approaches which shed novelbright in this field.[4]

Methods for detecting deception are constantly being built to protect criminals by responding to their deceitful tactics. These deceptions are categorized as:

Bank Card Deception: Connected and Disconnectedfrom internet

Identity Burglary

Computer Infringement

Application Deception

Deception Identity

Telecommunications Deception

Following are some of the different methodology which is used for detecting the deception:

Artificial Neural Network,

Logistic Regression,

Bayesian Network,

K-Nearest Neighbor,

Genetic Algorithm, and

Fuzzy Logic.

LITERATURE REVIEW

Scam acts as either an illegitimate or unethical deception designed to gain economic or social benefit. It is a premeditated tactic that is clearly illegal, regulation or policy of achieving illegal financial benefit. authors argue related to the detection of anomalies or deception has been already published in this field and seems to be available for public use. The researcher showed that the techniques employed this field are information accumulation technologies, advanced criminal identification and antagonistic detection .While in some places these methods and algorithms have created an unforeseen success, they have botched to deliver a enduring and reliable explanation to detect deception. The author presented a related research domain where they used Outlier mining, Outlier detection mining and Distance amount algorithms to reliably predict deception transactions in an emulation experiment of a certain commercial bank's credit card transaction data collection.

Outlier removal is a field of data mining that is used primarily in the monetary and internet sectors. This deals with the identification of items disconnected from the main network, i.e. transactions not genuine. They took consumer behavior attributes and based on the value of those attributes they measured the difference between that attribute's observed value and its predicted value.[5]

Unconventional techniques such as hybrid data mining / complex network classification algorithm may perceive illegal instances in an actual data set of card transactions, based on network reconstruction algorithm.This paper proposes a new collative comparison measure which reflects fairly the gains and losses resulting from deception detection. Using the proposed cost estimate, a cost-sensitive approach

based on the Bayes minimal risk is introduced. Improvements of up to 22 percent are made as compared with this approach and other state-of-the-art algorithms. The data collection for this paper is focused on a large European company's real-life transactional data and personal data in data is kept secret, an algorithm's accuracy is around 49.9%.

The purpose of this paper[6] was to find an algorithm and lower the estimate of costs. The result obtained was 22 percent and Bayes' minimal risk algorithm was the one they noticed.Efforts have also been made to advance from a whole new perspective. In the case of deception lent transactions, attempts were made to enhance alert-feedback interaction.The approved system would be notified in case of deception lent transactions, and feedback would be sent to reject the ongoing transaction. One of the approaches which shed new light in this area, the Artificial Genetic Algorithm countered deception from a different direction.

This paper[7], such as the Naive Bayesian Classifier and the model based on Bayesian Networks, the clustering model, offers a contrast between models based on artificial intelligence and a general overview of the evolved deception detection system. And in the end conclusions are based on the findings of the evaluative testing of the models. The number of legal truncations was estimated to be greater than or equal to 0.64, which is their accuracy using Bayesian Network was 64 percent. This paper aims to compare models based on artificial intelligence with a general description of the system developed and to state the accuracy of each model.

METHODOLOGY

The author does experiments by using four types of classifiers. These are:

Logical Regression,

SVM (Support Vector Machine)

Decision Tree Classifier, and Random Forest.

Logistic Regression:

Logistic Regression is a supervised identification process that calculates the likelihood of binary predictor variables predicted from the independent dataset variable that is logistic regression predicts the likelihood of an effect that has two values either zero or one, yes or no and false or actual.Logistic regression has parallels to linear regression but, as a straight line is obtained in linear regression, a gradient is seen in logistic regression. Use one or more determinants or an individual variable is based on what prediction, logistic regression generates logistic curves that show values around 0 & 1.

In another type of problems known as classification tasks, logistic regression is used. The goal here is to determine the category to which the actual object under investigation belongs. Classification is about dividing the data into classes with us, depending on some characteristics.

Let's look at the most widely used example: a tumor must be categorized as malignant or benign based on different characteristics such as size, position, etc. So, the Logistic Regression is a regression framework that has attribute variables such as accurate / inaccurate or 0/1 in the response variable (dependent variable). This basically calculates the likelihood of a binary answer as the answer variable value based on computational equation which relates this to the predictors.

The expression which is used for logistic regression is:

$$\pi(x) = \exp(\alpha + \beta x) / (1 + \exp(\alpha + \beta x))$$

Where,

alpha and beta are the parameters that are constant

numerical values,

Y – response variable, and

x- predictor.

The logistic regression is of two types such as:

1. Binary,
2. Multinomial, and
3. Ordinal

Binary:

Under such a category a variable of interest can only have two possible forms either 1 or 0. Such variables, for example, may indicate final outcome, yes or no, win and lose etc

.Multinomial:

In such a category, variable of interest can have 3 or more potential unsorted types, or forms without quantitative sense. Such factors, for example, may reflect "Type A" or "Type B" or "Type C."

Ordinal:

Under these categories, variable of interest can have 3 or more potential ordered types, or forms with a numerical value. Such parameters, for example, may reflect "bad" or "nice," "very nice," "outstanding," and each category may have values such as 0,1,2,3.

Regression Models of Binary, Multinomial:

The basic example of logistic regression is binary or binomial logistical regression for which the objective or explanatory variables should only have two viable categories, either 1 or 0. A further important feature of logistic regression is multivariate regression logistic regression wherein the objective or explanatory variables may have 3 or maybe more feasible unsorted types, i.e. the types that

that have no quantitative significance.

Support Vector Machine (SVM):

Support Vector Machine "(SVM) is a administered algorithm that may be utilized to address bothclassificationand deterioration. It's mainly utilized in identification issues. In this system, the author plots respectively element of data as a point in i-dimensional space (in which i is the number of options) with the value of each function becoming the value of a unique coordinate. Then they perform segregation by finding the hyperplane which very well differentiates the two groups (see below figure 1). The parts of support vector machine are:

Support Vector:

The points which are nearest to the plane is called support vectors. The line are separated by the data point for finding the best result.

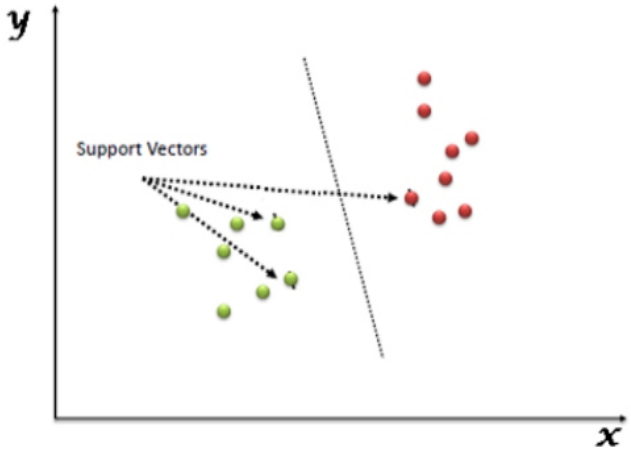


Figure 1: Support Vector Machine classifier

Hyperplane:

It is a plane which is knows as decision plane because it takes decision for giving the outcomes. It is separated between a set of multiple class objects.

Margin:

It can be described as the interface between two lines at the various classes' closet datasets. This could be computed as the distance of the object between both the line and the testing set. Wide margins are called good margins, and low margins are known as poor margins.Support Vectors Machine are merely the coordinates of distinctcomment. Support Vector Machine is used for segregating of two classes (hyper-plane/ line).SVM can produce the hyperplane iteratively, so the error can be reduced. SVM 's objective is to split the data into groups in order to find a maximum marginal hyperplane (MMH).SVM classifiers

have moderately successful and function well enough with high - dimensional feature space. Basically,SVM classifiers are using a set of training samples and thus use far less memory in the end.

Decision Tree:

The decision tree builds a tree structure in the context of classification or regression models. It breaks down a collection of data into smaller and smaller subsets while at the same time incrementally creating a related decision tree. The result is a tree with nodes for decision and nodes for leaves. [8]–[11]A decision node has two or more branches, and a classification or decision is represented by a leaf node. The top decision node in a tree that coincides with the strongest predictor called the root node. Decision trees are capable of handling both categorical and numerical data.

Decision Tree Classifier repetitively divides the area(plot) into subpart by identifying lines. (repetitively because there may be two distant regions of the same class divided by other as shown in figure 2 below).

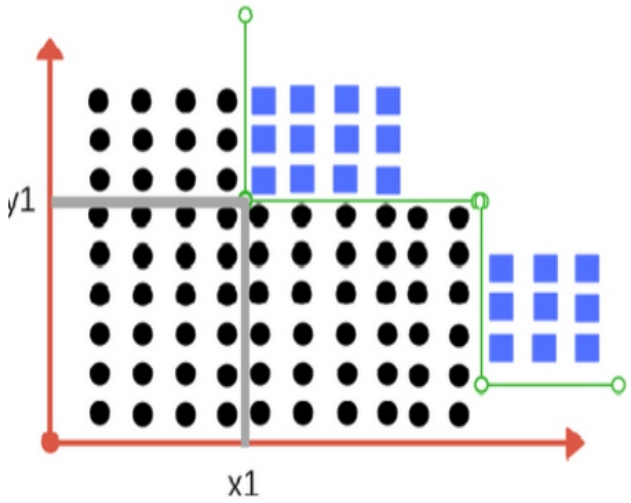


Figure 2: Decision Tree Classifier

Construction of Decision Tree:

A tree can be "learned" by splitting the source set into subsets based on a check of the value of the attribute. This cycle is replicated in a recursive manner, called recursive partitioning on each derived subset. The recursion is completed when the subset at a node all has the same target variable value, or when splitting does not add value to the predictions anymore.

The construction of a decision tree classifier involves no domain knowledge or set of parameters and is therefore ideal for the exploration of explorative knowledge. Decision trees can manage data of large dimensions. Tree classifier has excellent accuracy in general decision making. Decision tree inference is a characteristic inductive approach to learn knowledge on classification.

Terminal Nodes:

While developing tree based terminal nodes, one crucial fact is to determine when and how to stop growing node or generate more terminal nodes. It can even be achieved utilizing two parameters, namely maximum tree depth and minimum node records as follows:

Maximum Tree Depth: This is the maximum number of nodes in a tree after root node, as the name implies. So, people avoid adding terminal nodes until a tree reaches a certain depth, i.e. if a tree has a maximum number of terminal nodes.

Minimum Node Records: The total number of training patterns for which a given node is responsible can be specified. This is necessary to avoid adding terminal nodes once tree is reached at or below these minimum node records.

Now that it is cleared that when to construct terminal nodes, and start building the list. Recursive splitting is a tree-construction method. In this process, once a node is created, recurring is created the child nodes (nodes added to an existing node) on each data group, generated by splitting the dataset, by repeatedly calling the same function.

It is necessary to make a guess about that after creating a decision tree. Prediction essentially includes navigating the decision tree with the data row explicitly given. With the assistance of recursive method, make a prediction, same as above. The same prediction routine is reappointed with left or right child node.

Decision Tree Representation:

Decision trees identify instances by sorting them from the root to some leaf vertex down the tree which gives the instance classification. An instance is defined by beginning at the tree's core point, checking the attribute stated by that node, then heading down the tree branch corresponding to the attribution's value as shown in figure 2 above. For the subtree rooted in the new node this cycle is then repeated.

The instance of a tree structure is given below to predict how often an individual is fit or unfit to provide numerous characteristics such as age, food patterns and exercise levels, offering great accuracy and continuing to work well with high - dimensional feature space.

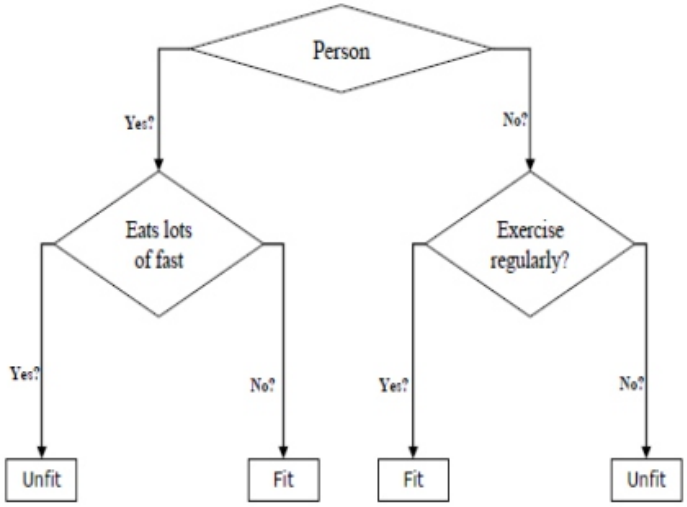


Figure 3: Example of Decesion Tree

According to an above figure 3, the author defined the persons wellness by analyzing many factors. The person has two choices: if he eats lots of fast food or excise regularly. Now, if the individual eats lots of fast food, then again this has two option such as: If yes, then it is sure that an individual is unfit. If no, then an individual is fit.

Now the author came on second option in order to know that an individual is doing exercise regularly or not. If yes, then definitely he is fir or if no, then surely, he is unfit. So, in this way the machine learning decision tree algorithm works.

Random Forest:

Random Forest is a Classification and Regression algorithm. In short, it's a set of classifiers for the decision tree. The random forest has an advantage over the decision tree because it corrects their training collection with the habit of overfitting.

Likelihood Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

Posterior Probability Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

A subset of the training set is sampled randomly so that each tree is trained and then a decision tree is constructed, then each node splits on a feature selected from a random subset of the full feature set. In a random forest, training is extremely fast even for large data sets with many features and data instances, and since each tree is trained independently the others.It has been found that the Random algorithm provides a good estimate of the generalization error and is prone to overfitting.

Figure 4 shows the random forest algorithm. There are two sets:Training Set, and Test Set. The training set has n samples. It has following steps:

Firstly, choose the sample from datasets,

Then develop a decision tree for each sample. It provides the prediction result from each decision tree.

Then Voting is done for each result, and

The final prediction result is the most voted prediction result.

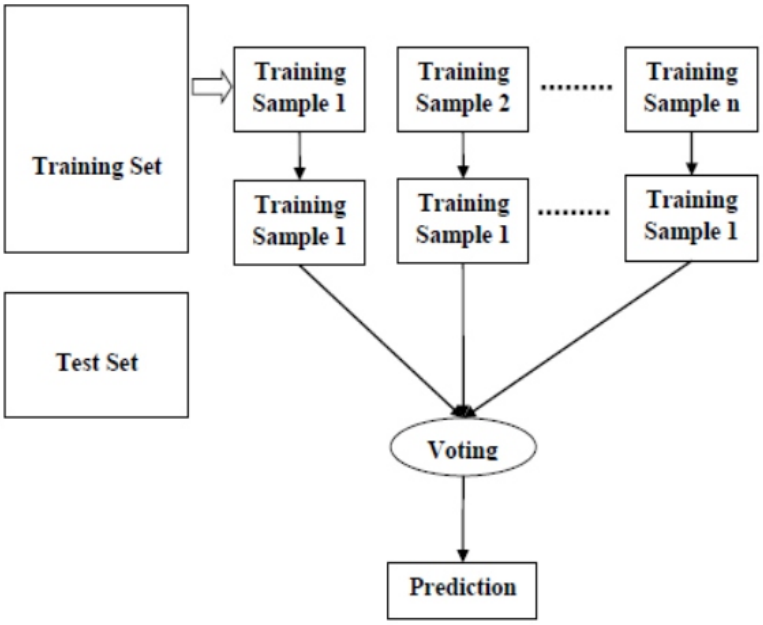


Figure 4: Random Forest Algorithm

The below diagram (Figure 5) helps in understanding the basic rough architecture of this system. Firstly, the incoming transaction is done by the user. This transaction is associated with the bank card deception recognition system and decision maker. These two provides result by comparing all the result and provide best output with help of transaction database as shown in below figure.

The dataset is currently being structured and analyzed. The time and distance and the volume column is uniform, and the column is removed to ensure quality equity. The statistics are processed by a selection of modular algorithms. Above four algorithms is used for this purpose. These data fit into a model.

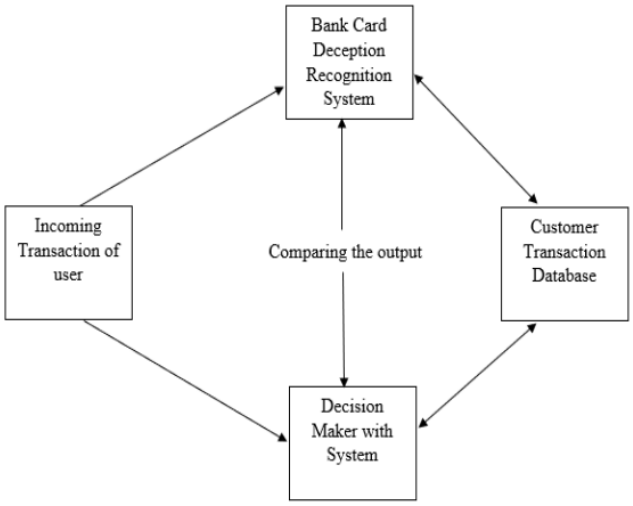


Figure 5: System Architecture

All the people listed in this list have their cards closed in order to evite some risk because of their high-risk profile. Prerequisite is more complicated to the other half. The list is saved only in the restricted data to be properly audited on a case-by - case basis.Credit and collection officers judged half the cases. This list may be regarded as suspect fraudulent comportment.

RESULT& DISCUSSION

This idea is hard to execute in real life, since it needs bank cooperation that is reluctant to share knowledge because of their competitiveness in the market, and for legal reasons, and for the protection of their users ' data.So the authorhas looked up some research papers that is similar methods followed, and findings gathered to make this system useable in a practical manner.

The code shows the number of false positives that it has found and then equate it to real values. Using this it is necessary to calculate the algorithm accuracy, sensitivity and precision.The fraction of the data which the used used for quicker testing is 10% of the full dataset. The complete dataset is also used at the end and the reports are also written out.

These results, along with the report for each classification in the output, the algorithm is as follows, where class 0 is given that means the transaction was legitimate, and 1 means the transaction was assessed as fraud.

Table 1 shows the performance analysis of different classifiers. The accuracy, sensitivity, and precision of all classifiers.In this paper four algorithms are developed for machine learning for identifying credit card fraud. To appraise the algorithms, 70% of the dataset is used for training and 30% is used for testing. It is used for the validation and checking.

Table 1: Performance Analysis

Metrics	Logistic Regression	SVM (Support Vector Machine)	Decision Tree	Random Forest
Accuracy	0.968	0.966	0.944	0.995
Sensitivity	0.965	0.964	0.944	0.993
Precision	0.995	0.995	0.994	0.996

The author performs statistical process in order to find the accuracy, sensitivity and precision for deception recognition method. When the author perform analysis while using linear regression the accuracy, sensitivity and precision is 0.968, 0,965, and 0,995 respectively.

The support vector machine is another algorithm in machine learning that is used for this purpose. Here, the accuracy, sensitivity and precision is 0.966, 0,944, and 0,995 respectively. When decision tree is used for this purpose then the accuracy, sensitivity and precision is 0.944, 0,944, and 0,994 respectively. After that random forest is used for finding the best result then the accuracy, sensitivity and precision is 0.965, 0,963, and 0,996 respectively.

So, after comparing all the analysis and result the author found that the accuracy of logistic regression is the highest i.e. 96.8%. The sensitivity of the logistic regression is againhighest sensitivity, i.e. 96.5% but the precision of random forest is the highest precision, i.e. 99.6%.

For understanding the accuracy, sensitivity and precision, firstly people have to understand what is these factor that affect the system in a various manner.

Accuracy:

It shows the difference between the indicated value and the real value. If the indicated value is Ai and the real value is Ar then the accuracy is:

Accuracy= Ai-Ar

It shows the closeness of the indicated value towards the real value.

Sensitivity: The ratio of change in output response is called sensitivity, S. Out of a system for a specified change in input, anywhere this is to be evaluated. This can be expressed mathematically as

$$S = \frac{\Delta A_{out}}{\Delta A_{in}}$$

ΔA_{in} ..., ΔA and are the change in input and output respectively.

The word sensitivity implies the slightest amount in quantifiable input forced to answer to an instrument. If the standard curve is linear, then the instrument's responsiveness is a perpetual and the standard curve

is linear, then the instrument's responsiveness is a perpetual and the standard curve is equal to the slope. When the standard curve is variable, then the instrument's responsiveness will not be a fixed and will differ with the data.

Precision:

When a device consistently shows a certain value and is often used to calculate a certain quantity for any couple of iterations in the same conditions, so it is assumed that the system has high precision.

CONCLUSION

Misusing the Credit Card is a criminal offense in society. This research has drilled down the most widely recognized strategies for deception alongside their recognition techniques and assessed late discoveries in this field. This paper has additionally clarified in detail, how AI can be applied to show signs of improvement brings about extortion discovery alongside the calculation, pseudocode, clarification its execution, and experimentation results. From all the results the author concluded that the accuracy of the random forest is the highest i.e. 99.5 % and precision is also highest i.e. 99.6%. The sensitivity of the logistic regression is the highest among all the classifiers, i.e. 96.5%.

With a greater number of training data, the Random Forest Algorithm will work better, but speed will suffer during testing and implementation. This will also help to incorporate more pre-treatment procedures. The SVM algorithm still suffers from the problem of the imbalanced dataset and needs more preprocessing to provide better results in the results shown by SVM is good but it could have been better if more preprocessing was done on the data.

REFERENCES

[1] N. Khare and S. Yunus Sait, "Credit Card Deception Detection Using Machine Learning Models and Collating Machine Learning Models," Int. J. Pure Appl. Math., vol. 118, no. 20, pp. 825–838, 2018.

[2] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, "Credit card deception detection using machine learning techniques: A comparative analysis," in Proceedings of the IEEE International Conference on Computing, Networking and Informatics, ICCNI 2017, 2017, vol. 2017-January, pp. 1–9, doi: 10.1109/ICCNI.2017.8123782.

[3] L. S. V S S and S. Deepthi Kavila, "Machine Learning For Credit Card Deception Detection System," 2018. Accessed: 06-May-2020. [Online]. Available: http://www.ripublication.com

[4] A. Oza, "Deception Detection using Machine Learning."

[5] "(PDF) A Review On Credit Card Deception Detection Using Machine Learning." https://www.researchgate.net/publication/336552027_A_Review_On_Credit_Card_Deception_Detection_Using_Machine_Learning (accessed May 06, 2020).

[6] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, and A. Anderla, "Credit Card Deception Detection - Machine Learning methods," in 2019 18th International Symposium INFOTEH-JAHORINA, INFOTEH 2019 - Proceedings, 2019, doi: 10.1109/INFOTEH.2019.8717766.

[7] S P Maniraj, Aditya Saini, Shadab Ahmed, and Swarna Deep Sarkar, "Credit Card Deception Detection using Machine Learning and Data Science," Int. J. Eng. Res., vol. 08, no. 09, Sep. 2019, doi: 10.17577/ijertv8is090031.

[8] S. Yaram, "Machine learning algorithms for document clustering and deception detection," in Proceedings of the 2016 International Conference on Data Science and Engineering, ICDSE 2016, 2017, doi: 10.1109/ICDSE.2016.7823950.

[9] R. A. Bauder and T. M. Khoshgoftaar, "Medicare deception detection using machine learning methods," in Proceedings -16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017, 2017, vol. 2017-December, pp. 858–865, doi: 10.1109/ICMLA.2017.00-48.

[10] P. Raghavan and N. El Gayar, "Deception Detection using Machine Learning and Deep Learning," in Proceedings of 2019 International Conference on Computational Intelligence and Knowledge Economy, ICCIKE 2019, 2019, pp. 334–339, doi: 10.1109/ICCIKE47802.2019.9004231.

[11] O. S. Yee, S. Sagadevan, and N. H. A. H. Malim, "Credit card deception detection using machine learning as data mining technique," J. Telecommun. Electron. Comput. Eng., vol. 10, no. 1–4, pp. 23–27, 2018.

Electronic Mail Classification System Based on Machine Learning Approach

Subhrajyoti Ranjan Sahu1*, J.Sunil Gavaskar2
1 B. Tech.,ECE, Nalanda Institute of Technology, Bhubaneswar, India.
2Assistant Professor, Lord Jegannath College of Engineering & Technology, Ramanathichenputhur, Tamil Nadu, India.

ABSTRACT

In current times, users depend comprehensively on electronic communication ways such as electronic mails as it is considered a foremost source of communication. A vast amount of time is investedin electronic mail for communication in the information technology field, due to which electronic mail management has become a prominent feature among the mailing applications. Electronic mail classification comes under this type of management which helps the expert to eliminate the time invested during un-necessary mail reading. Also, the content of electronic mail is further used in the analysis for future prediction and reading behaviors in which a good mail classification system would reduce a lot of time and resources. Conventionally many other systems or methods are present and widely popular in the market but there is no such system that achieves high accuracy. This paper proposes a novel electronic mail classification system that is based ensemble technique which combines the result of many classifiers to achieve good accuracy.

KeywordsClassifiers, Content Analysis, Electronic communication, Electronic mail, Feature extraction

INTRODUCTION

Communications is part and parcel of everyday operations run smoothly, run smoothly. Good communication is not only within the company but rhe customer also extends. Telecommunications Consumer Care, E-mail and talk communications are primarily based today. Every e-mail a service ticket is considered the client. It refers to small or medium businesses might be enough for the whole support team to have a common email inbox works together on service tickets for consumers.

The method is not scalable though, the support team is also rising as the business expands. Take a look at a situation teams, each running a broad support team errands. To optimize performance, and minimize time the support ticket is spent in the system and incoming tickets must be sorted and assign right support team to them.

This task is time consuming and Intensive labor but automation is not a trivial task because of the complexity of the Natural languages which the software needs to understand. Any system that does the processing of languages, that is to say a language used by humans to Communicate, does Natural Language Processing (NLP).

Automating email labelling and sorting requires a model that can differentiate between different types of

errands and support requests. Such models must be able to do this even if the email contains spelling mistakes, previous conversations, irrelevant information, different formatting or simply rubbish. The LSTM model is an extended version of the Recurrent Neural Network (RNN) network which is a sequential model often used in text classification. Word embedding models aim to model the words of a language in a vector space and placing words with similar semantic meaning close to each other.

Electronic Mail (E-mail) is commonly a procedure for transferring and receiving electronic messages by using electronic devices such as smartphones or laptops.Emails were accessible in the 1960's, at that period only administrators could only transfer applications on the same device, so early email networks needed to submit both the sender and the receiver electronically, equivalent to immediate texting. In 1971, Ray Tomlinson created a first program that could transfer mail between theuses on various servers around the ARPANET using the @ sign for connecting the username to a destination site and that was recognized as email in the mid-1970s.

Mail operates on computer networks, primarily uses the Internet. The email networks today are focused on a store-and-forward approach i.e. Allow, forward, send, and store messages from email servers. Neither of the consumers or their machines must be concurrently online, nor will they sign in normally with a mail server or a webmail system for writing, receiving or uploading information.Original ASCII text only, Internet email has been expanded to express text through certain character sets and interactive material attachments via Multipurpose Internet Mail Extensions (MIME). The history of current Internet email systems dates back to earlier ARPANET and specifications for email encoding released in 1973 (RFC 561). This is the consequence of the fact that foreign email addresses using UTF-8 are universal, although not generally accepted. In the early 1970's, a text address is identical to a foundational email received today.

Email was a popular and efficient communication mechanism increases the number of users on the Internet. Email control is also necessary and relevant the rising issue for citizens and businesses since it's vulnerable to violence. The Blind Layout Spam is a spam-known unwanted email a mishandling case.

A vast volume of data from different channels such as existing / potential clients, suppliers and internal contact inside the organization, product/service requests, other private and government organizations, etc. is delivered to businesses including IT firms, company institutes (such as investment banks), production sectors, and process sectors.Such unstructured communications are handily categorized by most organizations, with the assistance of trained customer support personnel, depending on the skills needed to answer and respond to the information contents of the document. The vast complexity of incoming e-mails nevertheless renders this strategy difficult to handle, time-consuming and misunderstood.

Email content Analysis is the term given to the process of exploration of the content present in the electronic mails; it is a useful process and one of the foremost applications of this type of content analysis

analysis lies under digital forensic investigation for distinguishing the abnormal activities like crimes, frauds. Also one of the application of this type of analysis is organizations get to know about the behavior of the email users. For instance, if there is company named “Nutrition SPL” and it sends mails for advertising there products, now with the help of the email content analysis, the company would get to know about how customers react to the emails such as, they read it or the customers leave the mails unread.

For conducting a good content analysis, the electronic mails should be categorized properly. Though some electronic mail applications provide features like filters. In which the user applies a filter to get some specific types of mail. But if the vast amount of content is considered then it is not easy to filter out the mails and content according to some particular requirements.

The most common application of the email classification is spam classification, it has been observed in many surveys and feedbacks, that spam in the mail is considered as one of the furthermost complex problems in the email services.Spam e-mails are any unintended e-mails not meant for a single recipient and submitted for marketing reasons, fraud, hoaxes etc. In 2009-2010 about 97% are reported to have been spammed[1].Therefore, several academic publications or reviewing emails centered on this topic (e.g. spam classification). But there is an ongoing conflict between spammers and spam detection devices, in which each party attempts to develop different methods of beating the other's techniques.

Some local papers [2], [3]conducting a spam evaluation have shownthat the problem was realistic. Researchers also carried out studies to determine existing spam delivery status in KSA. Writers have sought to sum up crucial explanations for spam communications and e-mails including pornographic content, advertisement, phishing, faith, etc. Of course, overuse and bandwidth and resources for no good purposes are a major disadvantage of spam spread.

An e-mail spam classifier is not only required to identify spam as junk mail correctly but also to recognize non-spam or regular e-mails in this regard. It is when the criteria for determining its definition or its estimation are all known.Then the accuracy of the email forecast is measured by four forecast metrics. True Positive (TP) claims the spam detector method assumes spam is spam, it always was spam,and it is spam. True Negative (TN) means the device or email program predicts the email is usual, not spam and it was right. The method inappropriately predicts that spam

(alleging false alarms) is a positive e-mail[4]–[6].

“Eventually False Negative (FN) often leads to a further mistake in which spam email is supposed to be ordinary. The identification method would also include the following values: TP 100%, TN 100%, FP 0% and FN 0%. It is difficult and unrealistic to attain this optimal condition. TP and FP balance each other by 100% (i.e. 100% of them total). Some email classification systems face the difficulty of limiting TP across multiple spam detection functions, but also many false alarms. On the other side, very lean rules that earn very high TN yet FN.Another challenge in emails’ spam detection is speed. Insecurity,

speed or performance is always in a trade-off withsecurity where too many roles may slow down the system." In addition to the classification that is based on spam messages, some papers have conducted researches in the content of the email, discussed other aspects for instance: automatic folder classification, contacts and email ID classification and alike.

The emails could be structured into chains and contain several emails sent back and forth between the client and the support staff. The models could use this in theory to determine the context and how the subject changes conversations. These chains are not included in the emails our model use, they are isolated and labeled individually. The model will be restricted to, and no classification or training will be done on any sort of document other than emails.

The field of machine learning is a substratum for the broad artificial intelligence field, aimed at ensuring that machines learn as human. Training implies that certain mathematical processes are known, identified and described.

Furthermore, this research paper is divided into various segments to address the issue and solution along with related work related to the email classification. Second segment of this paper describes and cites some related work like publications, patents and articles related to the field of email classification. Third segment of this paper discloses about the proposed work and fourth segment talks about the testing and relative results of the proposed work, lastly, the fifth segment of the researcher paper concludes all the work and addresses some advantages of the proposed work.

RELATED WORK

This segment of the research paper deals with some research work related to email classifications.

Table 1 : Some research work related to E-mail Classification

1. 2015- [4]	Discloses about frequency and term frequency combined feature selection method (DTFS) to improve the performance of email classification.
2. 2015- [5]	Discloses about classification task is called feature selection, which is used to reduce the dimensionality of word frequency without affecting the performance of the classification task.
3. 2016- [6]	Discloses about fuzzy logic techniques for email clustering. Extract concept and feature, same feature keyword goes into one cluster if a new keyword is found and not matched with any existing cluster than a new cluster is defined for that. Based on these clustering techniques authors wants to update that calendar for real time information and hassle free for reading unnecessary emails.
4. 2017-[7]	Discloses about two different approaches for classifying emails based on their categories. Naive Bayes and Hidden Markov Model (HMM), two different machine learning algorithms, both have been used for detecting whether an email is important or spam.
5. 2018-[8]	Discloses about different representation schemas for emails and a large set of features are also adopted for the purpose of experiment. Proposed Cascaded SOM based classification model performs well in email-classification compared to standard classification approaches and classical SOM based model.
6. 2019-[9]	Discloses about the use of semi-supervised learning can help leverage both labeled and unlabeled data. In the evaluation, we investigate the performance of our proposed approach with two datasets and in a real network environment.

As it is discussed that email classification is divided according to the situation requirements, so this segment cites paper according to the field such as spam-non-spam email classification, email data analysis research goals, ontology classification of email contents. Table 1 presents the work done related to email classification [7][8][9][10][11][12]. Figure 1, below, illustrates the distribution of research paper according to the email classification. The areas are categorized into five domains: spam, phishing, spam and phishing, multi-folder categorization, and others. Figure 2 depicts the frequency of email classification techniques according to the domains.

Categorization of single-label text (classification) is defined as the task of assigning a category to a document given a predefined set of categories. The goal is to approximate the representation of the text, so that it coincides with the text 's actual category. If a document will consist of multiple categories, then we need to adapt our algorithm to multiple category performance, which is called multilabel classification. The task then is to assign a appropriate number of labels that correspond to the document 's actual labels.

A fundamental categorization objective is to categorize documents in the same set that have the same context and documents that do not have the same context in separate sets. This can be related with various approaches involving algorithms for the machine learning. Machine learning algorithms learn to generalize categories from documents previously seen that are later used to predict the category of documents previously unseen.

The non-sequential models are used as a baseline for comparison with the LSTM network. The parameters of these models are not optimised and do not use preprocessing techniques such as lemming, stemming or stop word removal. The BoW or Average Word Vector (AvgWV) text representation models used for the non-sequential models are also not optimised. The BoW hyperparameters filter the numbers of words within a relative range, i.e sub- and supersampling. This is done to make the experiments possible with all non-sequential models

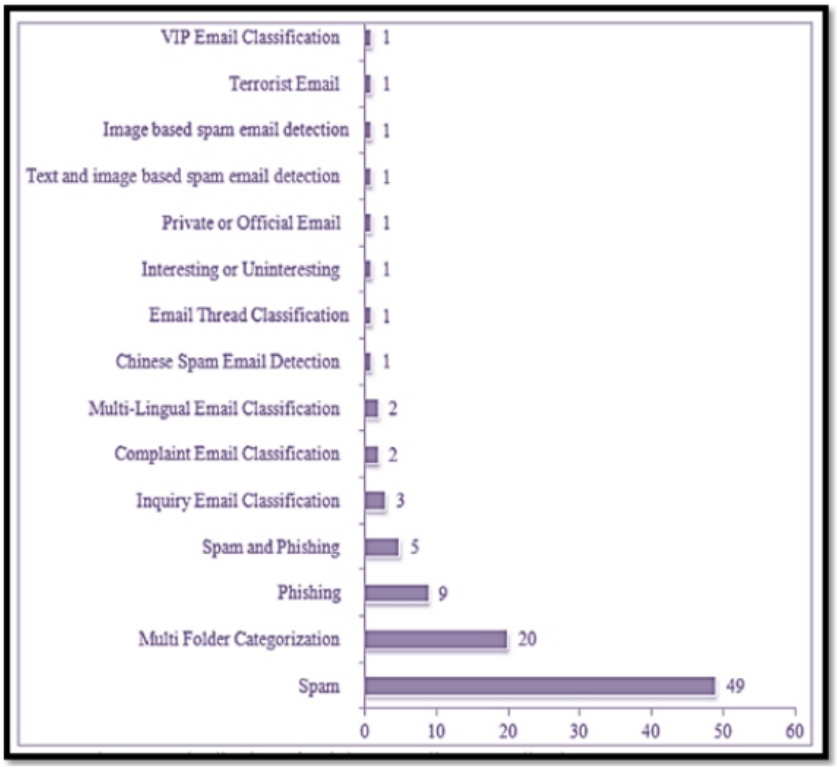


Figure 1: Distribution of Research paper According To Email Classification

Several types of email classification systems and methods are developed by using various techniques of machine learning. But still some challenges are still there, i.e. the challenges are dynamic, which is caused to the heterogeneous behavior of the content present in the email and some challenges are caused due to CPU limitations during implementation, due to which the required accuracy i.e. goal is not accomplished. Due to the aforementioned drawbacks, there is a need to develop an effective email classification system.

Several of the topics needed to effectively classify emails, i.e. models that interpret natural language and classifiers that use word relationships in a time series, are investigated. The bulk of the recent literature on the English language has been conducted while conducting similar studies and no study has been performed on the Swedish language. In terms of email classification and how to better use the NLP and machine learning models within that context, little work has also been done.

PROPOSED WORK

This segment of the research paper talks about the proposed system for email classification based on machine learning technique. Figure 3, shows the proposed system, various modules used in the system are explained below.The Python Application and Gmail API Bridge between the email client and the classification system connect to an email client and transmits an email to the rating system (in HTML form).

The same GUI is used after grouping to route messages. Extractor software removes and blends email

material, e.g. Email body and name. The purpose of the preprocessing module is to clean and prepare data that can be transmitted through the machine learning algorithm for the development of apps. Feature vectors are built from textual content in the function creation and representation module such that a learning algorithm can be understood (textual material must be expressed in number).The classifier is just a machine-learning algorithm to be equipped to do well in an unreleased dataset (unseen email). The professional classifier is used to determine in real time which e-mail ID a given e-mail will be routed to.

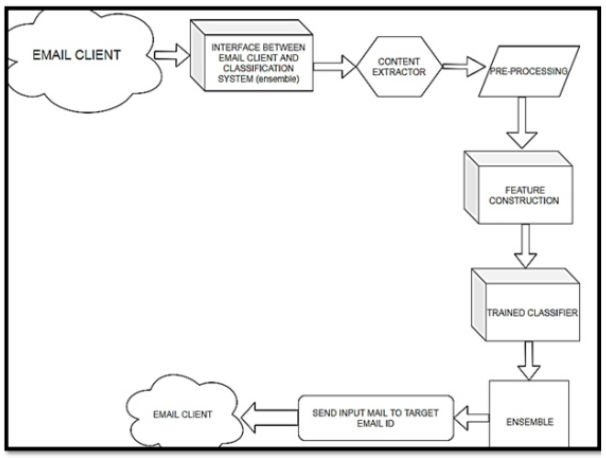


Figure 3 : Structure of machine learning-based email classification system

Additionally, the proposed system has an ensemble module for increasing accuracy. A series of learning algorithms are used by an actor to blend their results to simulate. There was a misunderstanding. This method works very well when the simple algorithms are unbalanced: they are typically unique, depending on the sub-set of the results.Ensemble learning as a whole is the way several models are systematically created and merged, such as classifiers or specialists, to solve a particular computational issue. The ensemble approach is used primarily for classification enhancement. Figure 4 shows the working of the ensemble module.

Ensemble approach/modeling is considered as a powerful technique to ameliorate the performance of the system. According to many subject experts relation machine learning, the ensemble is an art of integrating a diverse set of classifiers to ameliorate the stability and predictive power of the system. In simple terms, the prediction results of each classifier are combined and an average result is considered as the final result.

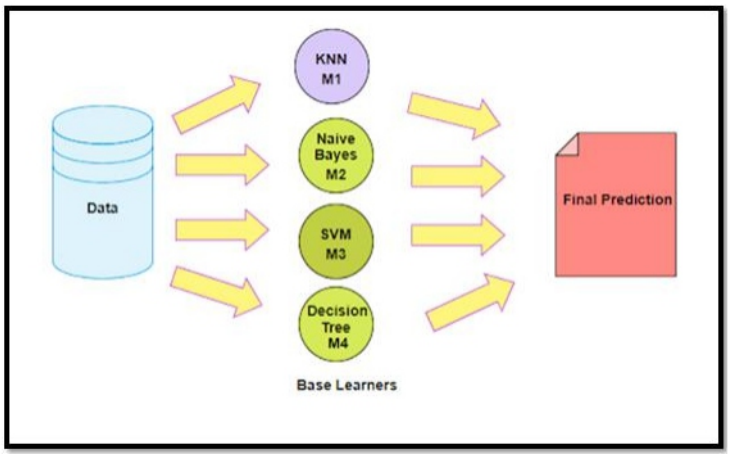


Figure 4 : Working of Ensemble Module

The simulations are performed to determine the correct e-mail classifier, function design and structural techniques. The impact of numerous preprocessing methods on the efficiency of the classifier will be studied (i.e. stopping words elimination, stamping, translating text to a single case of letters, different tokenization schemes).

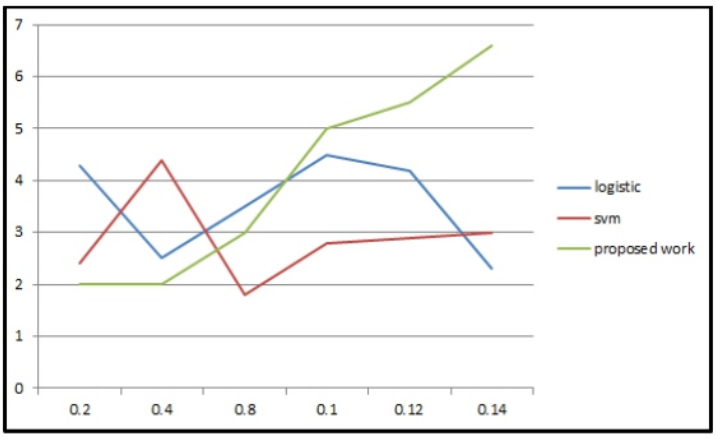


Figure 5 : Various proportion of training data in accordance to the techniques

The effect of connection terminology and the decrease in the classifier's output are also evaluated. Also, best attributes (i.e. hyper-parameter tuning) are used for each classifier along with grid search over a range of attributes. All the experiment work is conducted on two kinds of datasets i.e. one of them is 20 newsgroups and the other is of demo email dataset that has more than 350 emails, which demonstrates the effectiveness of the proposed system in real-time. Figure 5 shows the graph representing the test error rate for various proportions of training data. In this case, the measure of test error rate is taken as log loss. Dataset used is 20 newsgroup datasets. Figure 6, illustrates a Bar chart showing a comparison between different classifiers based on score/accuracy.

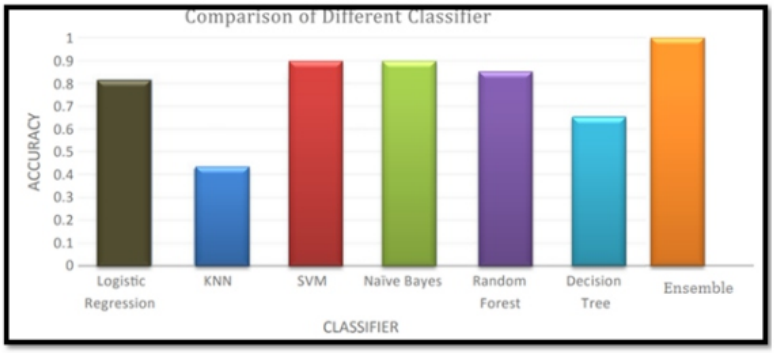


Figure 6 : Bar Chart showing a comparison between different classifiers based on score/accuracy

CONCLUSION

For big corporations, corporations, sectors, and businesses to evaluate vast e-mail details, an integrated real-time e-mail classification system would be quite useful. Once trained on a named dataset, the proposed email classification system can be used to accomplish real-time classification and even be compatible with any email user. A stand-alone framework that can be tailored to the needs of any enterprise is established as the approach suggested. It can be observed through the results that the addition done in the form of the ensemble in the proposed system defiantly boost up the accuracy of the system.

The research can be further extended to analyze the efficacy of deep learning techniques to solve the classification problem because of the massive nature.Extending the email classification to identify emotions can help the support team address angry or dissatisfied clients. That will improve customer service as support staff can cope with customer emotions. This would improve customer loyalty and raising the number of customers changing provider.

REFERENCES

[1] M. A. Al-Kadhi, "Assessment of the status of spam in the Kingdom of Saudi Arabia," J. King Saud Univ. - Comput. Inf. Sci., 2011, doi: 10.1016/j.jksuci.2011.05.001.

[2] I. Alsmadi and I. Alhami, "Clustering and classification of email contents," J. King Saud Univ. - Comput. Inf. Sci., 2015, doi: 10.1016/j.jksuci.2014.03.014.

[3] G. Mujtaba, L. Shuib, R. G. Raj, N. Majeed, and M. A. Al-Garadi, "Email Classification Research Trends: Review and Open Issues," IEEE Access, 2017, doi: 10.1109/ACCESS.2017.2702187.

[4] K. Karthik and R. Ponnusamy, "Adaptive machine learning approach for emotional email classification," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2011, vol. 6763 LNCS, no. PART 3, pp. 552–558, doi: 10.1007/978-3-642-21616-9_62.

[5] N. Sutta, Z. Liu, and X. Zhang, "A Study of Machine Learning Algorithms on Email Spam Classification," 2020, vol. 69, pp. 170–159, doi: 10.29007/qshd.

[6] S. Kranthi Reddy, P. B. Tarun, S. Rushika, P. D. Redd, and E.Anjala, "Spam email classification using machine learning algorithms," Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 7, pp. 1748–1752, 2019.

[7] Y. Wang, Y. Liu, L. Feng, and X. Zhu, "Novel feature selection method based on harmony search for email classification," Knowledge-Based Syst., 2015, doi: 10.1016/j.knosys.2014.10.013.

[8] M. Mohamad and A. Selamat, "An evaluation on the efficiency of hybrid feature selection in spam email classification," in I4CT 2015 - 2015 2nd International Conference on Computer, Communications, and Control Technology, Art Proceeding, 2015, doi: 10.1109/I4CT.2015.7219571.

[9] T. Suma and S. Y. S. Kumara, "Email classification using adaptive ontologies Learning," in 2016 IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, RTEICT 2016 - Proceedings, 2017, doi: 10.1109/RTEICT.2016.7808210.

[10] S. R. Gomes et al., "A comparative approach to email classification using Naive Bayes classifier and hidden Markov model," in 4th International Conference on Advances in Electrical Engineering, ICAEE 2017, 2017, doi: 10.1109/ICAEE.2017.8255404.

[11] N. Saini, S. Saha, and P. Bhattacharyya, "Cascaded SOM: An Ameliorated Technique for Automatic Email Classification," in Proceedings of the International Joint Conference on Neural Networks, 2018, doi: 10.1109/IJCNN.2018.8489584.

[12] W. Li, W. Meng, Z. Tan, and Y. Xiang, "Design of multi-view based email classification for IoT systems via semi-supervised learning," J. Netw. Comput. Appl., 2019, doi: 10.1016/j.jnca.2018.12.002

Use of Machine Learning for Continuous Improvement and Handling Multi-Dimensional Data in Service Sector

Sriram Lohit 1*, Mohammed Mutahar Mujahid 2, Galipally Kushal Sai 3
1, 2, 3 Student, Department of Electronics and Communication Engineering,
Chaitanya Bharathi Institute of Technology, India

ABSTRACT

Machine learning is known as a significant pattern of AI that gives an effective allowance to the software applications to become precise at forecasting outcomes without explicitly programmed in doing that. In addition, machine learning is important as this gives service sectors a suitable view of trends in "business operational patterns" and consumer behaviors. Service sectors are mainly known as the healthcare sectors, tourism sectors, and transportation sectors. In several developed countries, AI is maximizing labor productivity by more than 30% in the coming 15 years. The requirement of showing the usage of machine learning and the way it handles the multi-dimensional data have also been shown in this entire work.

Machine learning shows some ways through that it helps in providing improvement to all the service sectors such as enhancing consumer analytics, giving rapid and effective assistance, providing effective personalization, identifying the fraud cases and also enhancing customer experiences. Though, in this research work it has been highlighted that, in terms of implementing ML in service sectors, service sectors are facing several challenges. Moreover, in terms of showing the effectiveness of ML two algorithms with flowcharts have been shown in this work. On the other hand, in this research work, a secondary data collection method has been utilized and a qualitative data analysis method has also been used in this research work. In addition, secondary data resources have been assembled from books, scholarly articles, journals, and newspapers.

Keywords AI, Machine learning, secondary data resources, Service sectors.

INTRODUCTIONMachine learning can be recognized as an effective pattern of artificial intelligence that provides a better allowance to software applications to become more precise at predicting results instead of being explicitly programmed to do that. In addition, "machine learning algorithms" utilize historical information as input towards predicting latest output values. Machine learning supports service sectors to enhance their logistics through increasing efficiency in each step of storage, shipping and sales procedures. On the other hand, this technology is providing the allowance to all the "forward thinking businesses" to effectively integrate autonomous driving into their fleets. Machine learning helps in managing multi-dimensional data in all the service sectors with the help of some steps and these steps have been shown in this work. In addition, in this research work, the usage of machine learning for constant improvement and process of handling data have been shown through flowcharts, graphs, diagrams, and tables.

LITERATURE REVIEW

Machine learning brings continuous improvement in service sector Machine learning is necessary as it provides service sectors a view of effective trends in consumer behavior along with "business operational patterns" and helps the establishment of new products. In addition, it can be stated that, in present times, machine learning is increasingly used in logistics, healthcare, transportation and tourism companies. On the other hand, several of the leading companies mainly known as Google, Facebook, Uber all are making machine learning a pivotal part of their operations [1]. The market share of AI of the IT services sectors in India has reached approximately 51.8% in the year of 2021. It is not unknown that machine learning is the subset of AI therefore, machine learning's market share across the service sectors is also increasing. Moreover, in numerous developed countries, this AI might effectively maximize labor productivity by more than 30% in the upcoming 15 years [2].

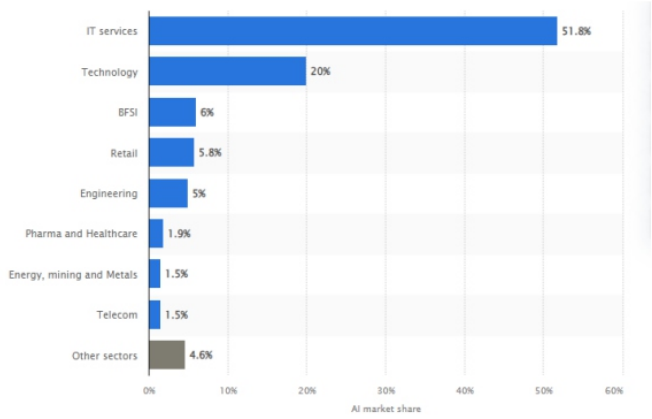


Figure 1. Market share of AI in India [2]

In the service sectors, proper usage of machine learning can bring continuous improvement through providing efficient consumer service. It can be highlighted that there are several ways that machine learning is bringing improvement in service sectors and also enhancing consumer service.

These ways have been shown in below:

Offering superior personalization Machine learning might be utilized to assess several previous interactions with a proper prospect and utilize this effective information to give highly personalized experiences to consumers, and empower effective consumer engagements.

Providing rapid and more effective assistance

Having theaid of "machine learning capabilities" consumers might be able to utilize their natural words and language to illustrate what they require assistance with [3].

Improving consumer analytics

Machine learning pulls information from consumers and utilizes it to effectively predict all the behavioral patterns along with several trends. This can support in identifying and contacting prospects and also help in enhancing sales and improving the consumer experience.

Reaching the right consumers at the appropriate place

As the service sector begins to gain more consumers and gathers more data, this provides a better allowance of"machine learning tools" to scrutinize and offer some effective ways to sell products, services and markets that would inevitably attract all the attention of consumers.

Identifying frauds

Fraud cases are growing the concerns for all the service sectors, mainly after the COVID 19 pandemic situation.In addition, machine learning supports guard against all fraud cases and provides an additional layer of protection.

Continuously improving consumer experiences

Machine learning provides a better allowance to programs to always remember and gain knowledge from previous experiences with consumers [4].

Machine learning manage to handle multidimensional data in service sector

Machine learning can be recognized as an effective type of AI software that is aiming to simplify procedure, with more simple programs. In addition, it is more ubiquitous in this modern world, being utilized in nearly several service sectors. Machine learning can be recognized as the large market that is encompassing the majority of this AI software along with projects. This market has been forecasted to expand from approximately 22.6 billion US dollars to nearly 126 billion US dollars by 2025.

Figure 2. Machine Learning [5]

It has been analyzed thatpredictive maintenance is havingenormous market focused opportunities along with that ML is known as the innovative solution to this “predictive maintenance implementation”. There are few challenges that are increasing barriers in implementing “machine learning algorithms” and these challenges have been effectively identified in table 1. In the result section, some effective utilization of machine learning has been highlighted through the support of two flow charts.

Table 1. Challenges of implementing machine learning

Challenges	Remarks
Getting needed dataset	• Unclear business planning and goal. • Not clear evidence of data that is giving proper value. • Launch of inter-linked machines.
Proper identification of needed data to gather	• Require resources and time to establish all ML solutions. • Selecting incorrect ML algorithms causes loss in price and time.
Improved data	• Selecting an effective method of illustrating the “data driven insights”. • Determine a suitable method of illustrating the data [6].
Security	• Safeguarding the admission to crucial equipment. • Proactive approach towards cybersecurity during protecting inter-linked assets.

MATERIALS AND METHODS

In this section, it is necessary to highlight that, in the requirement of highlighting “use of machine learning for continuous improvement and handling multi-dimensional data in the service sector”, a secondary data collection procedure has been selected. In addition, the main reason for choosing the secondary data collection method is that it provides authentic information about the research topic and cost effective [7]. On the other hand, the primary data collection method has not been utilized as it is costly and requires more time than usual. On the other hand, secondary qualitative analysis has been chosen and these secondary data sources have been gathered from books, scholarly articles, journals, newspapers and some reports.

It is not unknown that service sectors are known as the healthcare sectors, tourism sectors and transportation sectors. Some effective methods have been shown in this section in an effective manner and in this section the data resource along with the “preliminary data processing” have been utilized to effectively predict the task. Furthermore, the approaches of ML feature significant measures utilized in this work and its weaknesses and strengths. “Machine learning methods” have been highlighted in this

section in an effective manner [8]. In order to highlight the status of service sectors, an effective establishment of classification model overarch (y) (X) trained on a particular labeled set of proper training examples s, {yi,Xi}Ni=1.

Each of these N examples are representing anyone, where X ∈Rd can be recognized as the "d-dimensional vector of predictors" and ∈{0,1} is another person's outcome, that is encoded as 1 in case that person is diagnosed including that 0 otherwise. In this section, three machine learning methods for predicting the outcomes of the service sectors have been illustrated and these are "penalized logistics regression (LR)", "random forest (RF)", along with "extreme gradient boosting (XGBoost)". In addition, in this work it has been selected to utilize LR along with RF due to their pervasiveness along with accessibility to all the researchers. XGBoost has been selected to also due to its powerful performance in several recent competitions along with subsequent adaptation mainly as the "out of the box classifier of effective choice" [9]. On the other hand, for these methods, a summarization of methodology has been discussed below.

"Penalized Logistics Regression"

Logistic regression helps in training a linear model on the "log-odds ratio" of the result being positive.

$$\log\left(\frac{\Pr(y_i = 1|X_i)}{\Pr(y_i = 1|X_i)}\right) = \beta^T{}_{X_i}$$

Figure 3. "Penalized Logistics Regression" [9]

There "β = [β1, βd]"can be recognized as coefficients related to each predictor. In this analysis, it can be assumed that predictors are effectively standardized towards the unit variance along with mean-considered therefore, the intercept is specifically zero.

In this "standardized logistic regression", the coefficients are selected to maximize this log-likelihood of the main observations. In addition, penalized regression is effectively applied an extra penalty term that is mainly proportional towards the magnitudes of this coefficients, as:

$$\max_{\beta}\left\{\sum_{i=1}^{N}\left[y_i\left(\beta^T{}_{X_i}\right) - \log\left(1+\exp\left(\beta^T{}_{X_i}\right)\right)\right] - \lambda\sum_{j=1}^{d}\|\beta_j\|\right\}$$

Figure 4. "standardized logistic regression" [10]

On the other hand, random forest can be analyzed as an ensemble "machine learning" model that mainly trains numerous decision trees utilizing an amalgamation of bootstrap aggregating along with "random feature selection". Apart from that, in this section, XGBoost can be recognized as an ensemble "machine learning method" depending on gradient boosting of each decision tree [10].

In this section, in the modern world, the size of this AI market globally from 2021 and 2030 has been

shown in this section in an effective manner. As per the NMSC, the "global explainable AI market" has been valued at 4.4 billion US billion in the year of 2021. In addition, by the year 2030, this market has been estimated to have a value of approximately 21 billion US dollars.

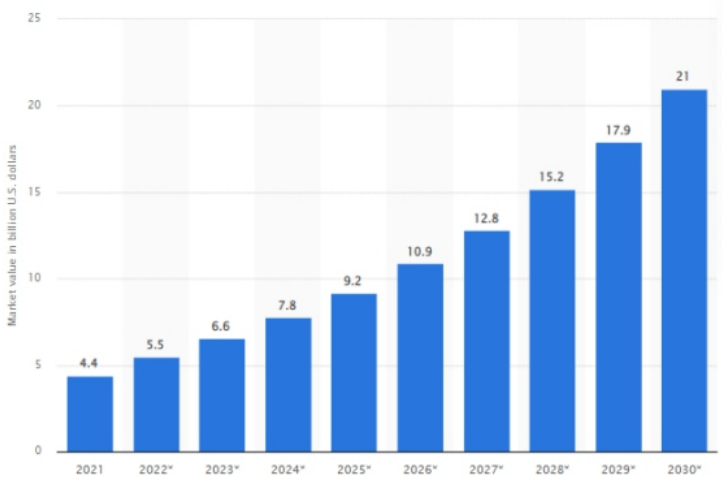


Figure 5. Size of this AI market [11]

Apart from thematerials and methods, the benefits of "machine learning" have been shown in an effective manner. In this section, it can be stated that machine learning helps in wide applications, reducing human intervention, identifying patterns and trends, managing multi-variety information and bringing continuous improvement. Proper utilization of machine learning can support service sectors in bringing continuous improvement. Moreover, there are some disadvantages of this ML such as higher chances of having errors, data acquiring, outcomes interpretation, resource along with timing [11]. On the other hand, proper use of materials and methods are required, otherwise ML cannot bring continuous improvement to the service sectors. Moreover, machine learning supports managing multinational data in the service sectors and this has been highlighted in this section in an effective manner.

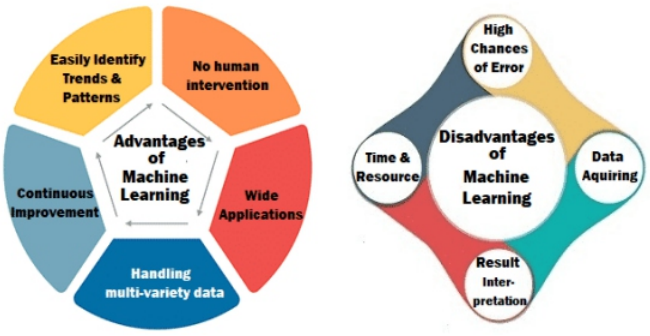


Figure 6. Advantages of ML [11]

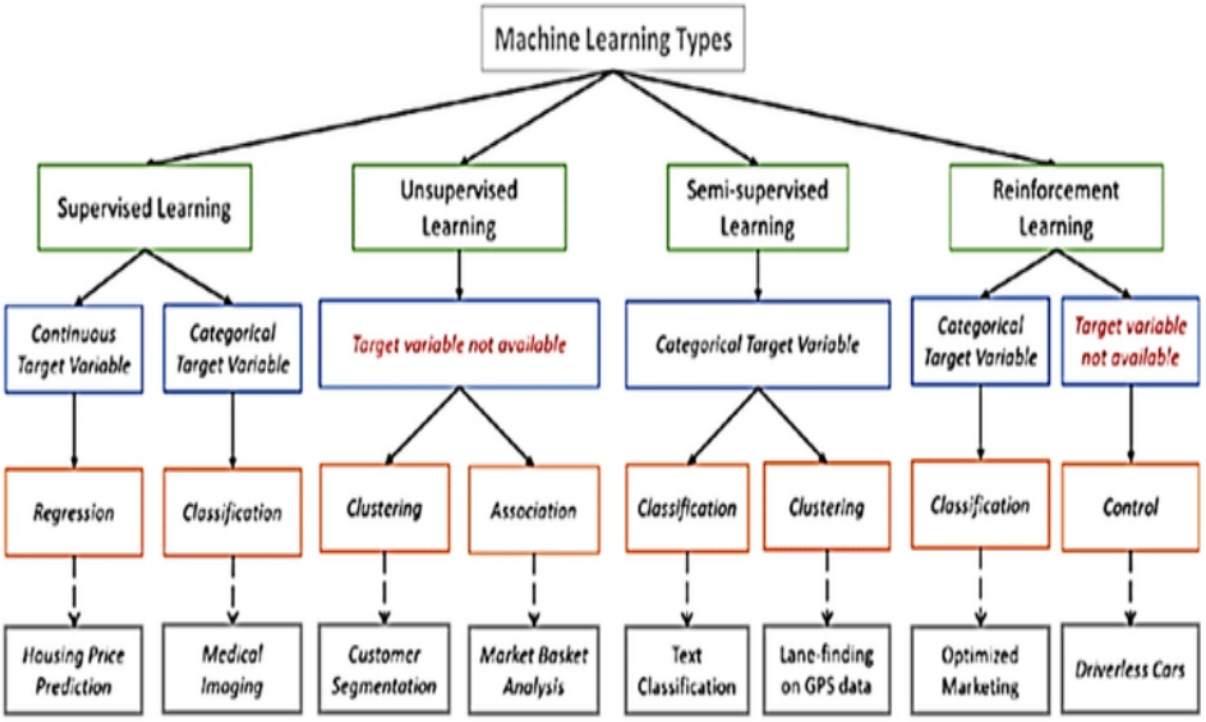


Figure 7. Types of machine learning [12]

In the result section, some effective utilization of machine learning has been highlighted through the support of two flow charts. From this above flowchart, it can be stated that, machine learning has various types such as reinforcement learning, unsupervised learning, supervised learning and semi-supervised learning. In addition, supervised learning has two subsections such as "continuous target variable" and "categorical target variable". On the other hand, unsupervised learning has one subsection that is "target variable is not available", "semi-supervised learning" has one subsection that is "categorical target variable". Apart from that, "reinforcement learning" has two subsections such as "categorical target variable' and "target variable not available" [12]. On the other hand, some subsections have also been shown in this flowchart in an effective manner.

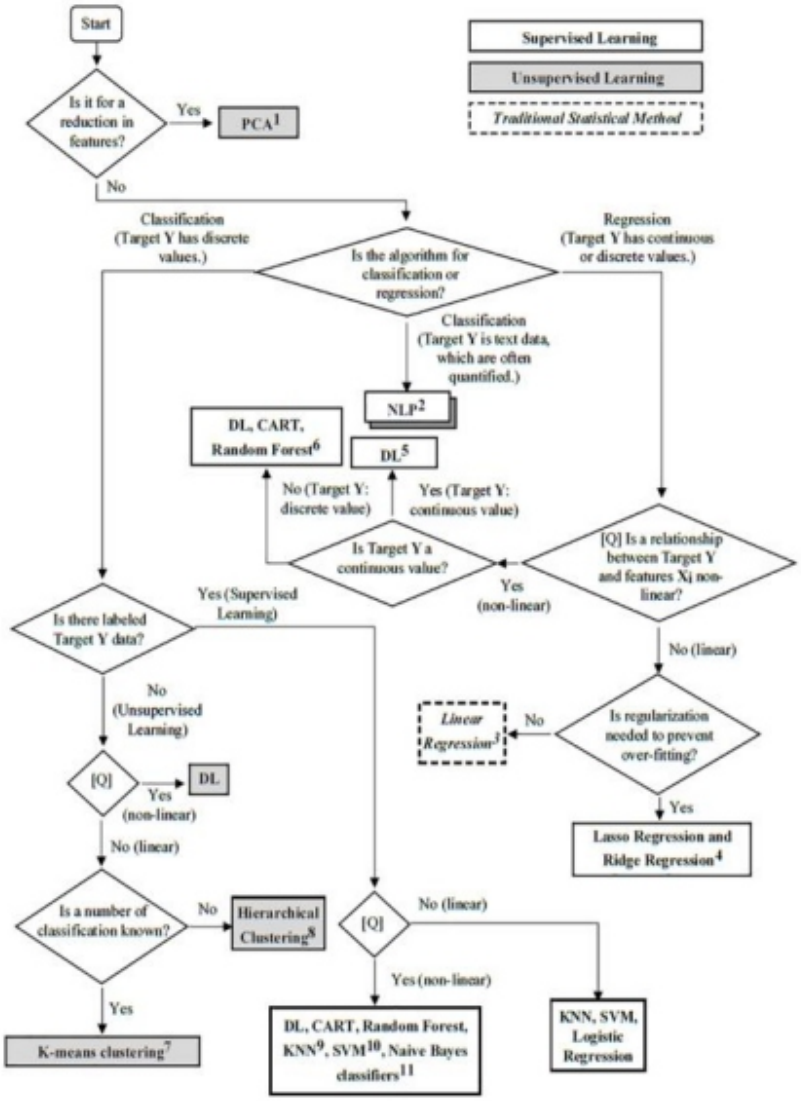


Figure 8. Flow chart of machine learning [12]

From this above figure, it can be stated that there are few types of ML such as supervised learning, and unsupervised learning. In this flowchart some methods have been highlighted such as CART, DI and Random Forest. Apart from that, "K-means clustering" and Hierarchical Clustering have also been shown in an effective manner [13].

Machine learning is necessary in the service sectors as several pieces of multi-dimensional data can be handled and it provides service sectors a proper view of some trends in the behavior of consumers and patterns of business operational and helps in developing new products. On the other hand, training can be recognized as the most significant portion of machine learning and data cleaning can be recognized as the most significant portion of this machine learning [14]. On the contrary, in this result section an effective comparison of the "machine learning technologies" have also been highlighted in terms of gaining a better understanding about machine learning. The "comparison of machine learning technologies" has been shown through the support of the below table.

Table 2. Comparison between "machine learning technologies"

Learning types	Tasks related to data processing	Distinction norm	Learning algorithms	References
Supervised learning	Regression or estimation or classification	Statistical classifiers	"Support vector machine"	[15]
		Computational classifiers	"Hidden Markov model"	[16]
			Naive Bayes	
			Bayesian networks	
		Connectionist classifiers	Neutral networks	[17]
Unsupervised learning	Prediction or clustering	Parametric	K-means	[18]

DISCUSSION

In this section, it can be discussed that, machine learning handles the "multi-dimensional data" in all the service sectors through acquiring the dataset, encoding the entire categorical data, importing all the critical libraries, importing the dataset and handling along with analyzing the missing values. In this section various comparisons between the "machine learning technologies" have been highlighted through a table form. In this table two learning types have been illustrated such as unsupervised learning and supervised learning [19]. Machine learning might be considered as a significant pattern of AI software and it aims to simplify the entire procedure with simple programs. The market of machine learning has been forecasted to increase nearly 22.6 billion US dollars up to 126 billion US dollars by the year of 2025 [20]. In this research, several advantages of machine learning have been shown such as wide applications, identify patterns and trends, and bring constant improvement.

CONCLUSIONIn the end it can be concluded that the entire research is showing the usage of machine learning in service sectors and the way it manages all the "multi-dimensional data". There are several ways that machine learning is bringing improvement such as providing superior personalization, giving quick and more beneficial assistance, enhancing customer analytics and proper identification of frauds. In this modern era, usage of machine learning is effectively increasing in tourism, healthcare and transportation companies. This market share of IT and AI in service sectors have reached nearly 51.8% in the year of 2021. The usage of machine learning is increasing in the telecom, engineering, pharma and healthcare, technology, retail, metals and mining organizations and it is handling these sector's "multi-dimensional data".

REFERENCE

[1] Min, Q., Lu, Y., Liu, Z., Su, C. and Wang, B., 2019. Machine learning based digital twin framework for production optimization in petrochemical industry. International Journal of Information

Management, 49, pp.502-519.

[2] Sun, S, 2022. AI market share in India in 2021, by industry. [Online]. Available at:https://www.statista.com/statistics/ 1180858/india-ai-market-share-by-industry/ [Accessed on: 11 September, 2022].

[3] Thormundsson, B, 2022. Machine learning - Statistics & Facts. [Online]. Available at:https://www.statista.com/topics/9583/ machine-learning/#dossierKeyfigures [Accessed on: 11 September, 2022].

[4] Gomez, O., Holter, S., Yuan, J. and Bertini, E., 2020, March. ViCE: visual counterfactual explanations for machine learning models. In Proceedings of the 25th International Conference on Intelligent User Interfaces (pp. 531-535).

[5] Liu, Y., Yu, F.R., Li, X., Ji, H. and Leung, V.C., 2020. Blockchain and machine learning for communications and networking systems. ieee communications surveys & tutorials, 22(2), pp.1392-1431.

[6] Çınar, Z.M., Abdussalam Nuhu, A., Zeeshan, Q., Korhan, O., Asmael, M. and Safaei, B., 2020. Machine learning in predictive maintenance towards sustainable smart manufacturing in industry 4.0. Sustainability, 12(19), p.8211

[7] Zekić-Sušac, M., Mitrović, S. and Has, A., 2021. Machine learning based system for managing energy efficiency of public sector as an approach towards smart cities. International journal of information management, 58, p.102074.

[8] UsugaCadavid, J.P., Lamouri, S., Grabot, B., Pellerin, R. and Fortin, A., 2020. Machine learning applied in production planning and control: a state-of-the-art in the era of industry 4.0. Journal of Intelligent Manufacturing, 31(6), pp.1531-1558.

[9] Tanizaki, T., Hoshino, T., Shimmura, T. and Takenaka, T., 2019. Demand forecasting in restaurants using machine learning and statistical analysis. Procedia CIRP, 79, pp.679-683.

[10] Benbelkacem, S., Kadri, F., Atmani, B. and Chaabane, S., 2019. Machine learning for emergency department management. International Journal of Information Systems in the Service Sector (IJISSS), 11(3), pp.19-36.

[11] Thormundsson, B, 2022. Revenues from the artificial intelligence (AI) software market worldwide from 2018 to 2025(in billion U.S. dollars). [Online]. Available at:https://www.statista.com/statistics/607716/worldwide-artificial-intelligence-market-revenues/ [Accessed on: 11 September, 2022].

[12] Lee, I. and Shin, Y.J., 2020. Machine learning for enterprises: Applications, algorithm selection, and challenges. Business Horizons, 63(2), pp.157-170.

[13] Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M. and Zhao, S., 2019. Applications of machine learning in drug discovery and

development. Nature reviews Drug discovery, 18(6), pp.463-477.

[14] Graur, D., Aymon, D., Kluser, D., Albrici, T., Thekkath, C.A. and Klimovic, A., 2022. Cachew: Machine learning input data processing as a service. In 2022 USENIX Annual Technical Conference (USENIX ATC 22) (pp. 689-706).

[15] Shehadeh, A., Alshboul, O., Al Mamlook, R.E. and Hamedat, O., 2021. Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression. Automation in Construction, 129, p.103827.

[16] Wu, H., Lubbers, N., Viswanathan, H.S. and Pollyea, R.M., 2021. A multi-dimensional parametric study of variability in multi-phase flow dynamics during geologic Co2 sequestration accelerated with machine learning. Applied Energy, 287, p.116580.

[17] Wan, S., Zhao, Y., Wang, T., Gu, Z., Abbasi, Q.H. and Choo, K.K.R., 2019. Multi-dimensional data indexing and range query processing via Voronoi diagram for internet of things. Future Generation Computer Systems, 91, pp.382-391.

[18] Sengan, S., Kamalam, G.K., Vellingiri, J., Gopal, J., Velayutham, P. and Subramaniyaswamy, V., 2020. Medical information retrieval systems for e-Health care records using fuzzy based machine learning model. Microprocessors and Microsystems, p.103344.

[19] He, R., Li, X., Chen, G., Chen, G. and Liu, Y., 2020. Generative adversarial network-based semi-supervised learning for real-time risk warning of process industries. Expert Systems with Applications, 150, p.113244.

[20] Perera, A.T.D. and Kamalaruban, P., 2021. Applications of reinforcement learning in energy systems. Renewable and Sustainable Energy Reviews, 137, p.110618

Instructions for Authors

Essentials for Publishing in this Journal

1 Submitted articles should not have been previously published or be currently under consideration for publication elsewhere.

2 Conference papers may only be submitted if the paper has been completely re-written (taken to mean more than 50%) and the author has cleared any necessary permission with the copyright owner if it has been previously copyrighted.

3 All our articles are refereed through a double-blind process.

4 All authors must declare they have read and agreed to the content of the submitted article and must sign a declaration correspond to the originality of the article.

Submission Process

All articles for this journal must be submitted using our online submissions system. http://enrichedpub.com/ . Please use the Submit Your Article link in the Author Service area.

Manuscript Guidelines

The instructions to authors about the article preparation for publication in the Manuscripts are submitted online, through the e-Ur (Electronic editing) system, developed by **Enriched Publications Pvt. Ltd**. The article should contain the abstract with keywords, introduction, body, conclusion, references and the summary in English language (without heading and subheading enumeration). The article length should not exceed 16 pages of A4 paper format.

Title

The title should be informative. It is in both Journal's and author's best interest to use terms suitable. For indexing and word search. If there are no such terms in the title, the author is strongly advised to add a subtitle. The title should be given in English as well. The titles precede the abstract and the summary in an appropriate language.

Letterhead Title

The letterhead title is given at a top of each page for easier identification of article copies in an Electronic form in particular. It contains the author's surname and first name initial .article title, journal title and collation (year, volume, and issue, first and last page). The journal and article titles can be given in a shortened form.

Author's Name

Full name(s) of author(s) should be used. It is advisable to give the middle initial. Names are given in their original form.

Contact Details

The postal address or the e-mail address of the author (usually of the first one if there are more Authors) is given in the footnote at the bottom of the first page.

Type of Articles

Classification of articles is a duty of the editorial staff and is of special importance. Referees and the members of the editorial staff, or section editors, can propose a category, but the editor-in-chief has the sole responsibility for their classification. Journal articles are classified as follows:

Scientific articles:

1. Original scientific paper (giving the previously unpublished results of the author's own research based on management methods).

2. Survey paper (giving an original, detailed and critical view of a research problem or an area to which the author has made a contribution visible through his self-citation);

3. Short or preliminary communication (original management paper of full format but of a smaller extent or of a preliminary character);

4. Scientific critique or forum (discussion on a particular scientific topic, based exclusively on management argumentation) and commentaries. Exceptionally, in particular areas, a scientific paper in the Journal can be in a form of a monograph or a critical edition of scientific data (historical, archival, lexicographic, bibliographic, data survey, etc.) which were unknown or hardly accessible for scientific research.

Professional articles:

1. Professional paper (contribution offering experience useful for improvement of professional practice but not necessarily based on scientific methods);
2. Informative contribution (editorial, commentary, etc.);
3. Review (of a book, software, case study, scientific event, etc.)

Language

The article should be in English. The grammar and style of the article should be of good quality. The systematized text should be without abbreviations (except standard ones). All measurements must be in SI units. The sequence of formulae is denoted in Arabic numerals in parentheses on the right-hand side.

Abstract and Summary

An abstract is a concise informative presentation of the article content for fast and accurate Evaluation of its relevance. It is both in the Editorial Office's and the author's best interest for an abstract to contain terms often used for indexing and article search. The abstract describes the purpose of the study and the methods, outlines the findings and state the conclusions. A 100- to 250-Word abstract should be placed between the title and the keywords with the body text to follow. Besides an abstract are advised to have a summary in English, at the end of the article, after the Reference list. The summary should be structured and long up to 1/10 of the article length (it is more extensive than the abstract).

Keywords

Keywords are terms or phrases showing adequately the article content for indexing and search purposes. They should be allocated heaving in mind widely accepted international sources (index, dictionary or thesaurus), such as the Web of Science keyword list for science in general. The higher their usage frequency is the better. Up to 10 keywords immediately follow the abstract and the summary, in respective languages.

Acknowledgements

The name and the number of the project or programmed within which the article was realized is given in a separate note at the bottom of the first page together with the name of the institution which financially supported the project or programmed.

Tables and Illustrations

All the captions should be in the original language as well as in English, together with the texts in illustrations if possible. Tables are typed in the same style as the text and are denoted by numerals at the top. Photographs and drawings, placed appropriately in the text, should be clear, precise and suitable for reproduction. Drawings should be created in Word or Corel.

Citation in the Text

Citation in the text must be uniform. When citing references in the text, use the reference number set in square brackets from the Reference list at the end of the article.

Footnotes

Footnotes are given at the bottom of the page with the text they refer to. They can contain less relevant details, additional explanations or used sources (e.g. scientific material, manuals). They cannot replace the cited literature.
The article should be accompanied with a cover letter with the information about the author(s): surname, middle initial, first name, and citizen personal number, rank, title, e-mail address, and affiliation address, home address including municipality, phone number in the office and at home (or a mobile phone number). The cover letter should state the type of the article and tell which illustrations are original and which are not.

Address of the Editorial Office:

Enriched Publications Pvt. Ltd.
S-9,IInd FLOOR, MLU POCKET,
MANISH ABHINAV PLAZA-II, ABOVE FEDERAL BANK,
PLOT NO-5, SECTOR -5, DWARKA, NEW DELHI, INDIA-110075,
PHONE: - + (91)-(11)-45525005

Note