Customer Churn Prediction Using Machine Learning

Dr. S. Brinthakumari
Department of Computer Engineering
New Horizon Institute of Technology And Management
Thane, Maharashtra, India
brinthakumaris@nhitm.ac.in

Ketan Patil, Nikhil Pawar, Sanket Jogale
Department of Computer Engineering
New Horizon Institute of Technology And Management
Thane, Maharashtra , India
{ketansir100, nikhilpawar954, sanketjogale2002}@gmail.com

***Abstract*— Customer churn, the phenomenon where customers cease their relationship with a business, is a critical concern for e-commerce platforms striving for sustained growth and profitability. Predicting churn in advance can empower businesses to implement proactive retention strategies, thereby mitigating revenue loss and enhancing customer satisfaction. In this study, we propose a machine learning-based approach to predict customer churn in e-commerce settings. We begin by collecting extensive data encompassing various customer attributes, transactional history, browsing behavior, and engagement metrics. Leveraging this rich dataset, we employ state-of-the-art machine learning algorithms such as logistic regression, random forests, gradient boosting machines, and neural networks for predictive modeling. Feature engineering techniques are applied to extract meaningful patterns and insights from the raw data, enhancing the predictive performance of the models. By deploying the developed predictive model into production environments, businesses can proactively identify at-risk customers and tailor targeted retention strategies to mitigate churn. The results demonstrate the effectiveness of machine learning in accurately predicting customer churn in e-commerce, enabling businesses to proactively implement retention strategies and enhance customer engagement.**

***Index Terms*— Machine learning, E-commerce, Logistic regression, Decision tree, Random forest algorithm.**

I. INTRODUCTION

Client churn or waste is one of the most pivotal problems for any business that directly sells or serves guests,E-Commerce or SaaS businesses it’s important to track and assay how numerous guests are leaving the platform and how numerous are sticking and the reasons behind them[1]. Knowing client geste can greatly enhance decision- making processes and can further help reduce churn to ameliorate profitability. In this composition, we’re going to assay ane-commerce dataset and find the stylish model to prognosticate client churn. But before probing into analysis let’s have a brief look at what’s churn client churn can be defined as the rate at which guests leave a platform or service. And client churn analysis is the system of analysing the rate. There are generally two kinds of churn. E-Commerce situation can be in one of four countries new, active, inactive, or churned[2]. The assiduity’s success is entirely grounded on its capacity to keep guests engaged for an extended period. Because of the assiduity’s fierce contest, carrying a new consumer is fairly expensive. In the event of a new client, a establishment’s fiscal break indeed(return on investment) is generally reached only after the client completes a many deals over time[3]. Active consumers are the backbone of every E- Commerce business, and businesses should pay special attention to those who may come inactive and churn latterly. Statistical ways and machine literacy strategies are exemplifications of methodologies that can be used to anticipate prospective customer development. client data collected by B2CE-Commerce companies are a great source of information because it allows for assaying different consumers copping patterns[4]. The liability of a customer churn is anticipated via churn vaticination. It reduces the cost of acquiring new guests while also aiding in client retention. It takes more marketing time and plutocrat to develop a new customer than to keep an living consumer[5]. guests who are reluctant to make a purchase or are prepared to switch shopping spots due to fiscal enterprises can be converted and gripped. They can anticipate norms and variety in product immolations[6]. Guests leaving for essential and necessary reasons are free to do so. In a contract script, a client relationship indicates that the establishment and the customer work together[7]. Churn Analysis is one of the world wide used analysis on Subscription acquainted diligence to client actions to prognosticate the guests which are about to leave the service agreement from a company. Both parties rights and scores are easily stated in the contract. The customer must complete the necessary responsibility following the contract to enjoy the applicable freedom after subscribing an agreement with the establishment[8].

II. METHODOLOGY

A. *Background*

Customer churn prediction is a critical task for businesses across various industries. It involves identifying customers who are likely to stop using a company's products or services. Machine learning techniques have emerged as powerful tools for predicting churn, enabling businesses to take proactive measures to retain customers and maintain for customer churn prediction using machine learning, covering data collection, prepro- cessing, feature engineering, model selection, evaluation, and deployment considerations.

Customer churn, the phenomenon where customers discontinue their relationship with a business, poses significant challenges for companies seeking to maintain sustainable growth and profitability. Predicting churn allows businesses to intervene with targeted retention efforts, thereby reducing revenue loss and fostering customer loyalty. Machine learning offers powerful tools to analyze historical data and uncover patterns indicative of potential churn[10]. This paper provides a comprehensive methodology for leveraging machine learning in customer churn prediction.

B. *Methodology*

1. Data Collection :- The first step in customer churn prediction is to gather relevant data. This includes customer demographics, transaction history, product usage, customer service interactions, and any other information that may be indicative of churn. Data sources may vary depending on the industry and the nature of the business[11]. For example, an e-commerce company may collect website browsing behavior, purchase history, and customer feedback, while a telecommunications company may gather call records, subscription plans, and customer support interactions.

2. Data Preprocessing :- Clean the data by handling missing values, removing duplicates, and dealing with outliers. Additionally, preprocess the data by encoding categorical variables and scaling numerical features.

3. Data Cleaning :- Handle missing values: Missing data can adversely affect model performance. Techniques such as mean imputation, median imputation, or deletion of missing records can be employed.Remove duplicates: Duplicate records may skew the analysis and lead to biased results. Removing duplicates ensures the integrity of the dataset.

4. Feature Engineering :- Feature extraction: Extract relevant features from the raw data that can capture the underlying patterns related to churn. This may include aggregating transaction data, calculating customer tenure, or deriving behavioral metrics. In Feature transformation transform features to make them more suitable for modeling[12]. This may involve scaling numerical features, encoding categorical variables, or creating binary indicators. In Feature selection Select a subset of features that are most predictive of churn. Use techniques such as correlation analysis, feature importance ranking, or model-based selection methods.

5. Model Selection :- After preprocessing the data and engineering relevant features, the next step is to select an appropriate machine learning model for churn prediction. Classification Algorithms:-

- Logistic Regression: It is a simple algorithm for binary classification tasks. It models the probability of churn as a function of the input features.
- Decision Trees: Non-linear models that partition the feature space into regions, making them interpretable and easy to visualize.
- Random Forests: Ensemble learning technique that combines multiple decision trees to improve predictive performance and robustness.
- Gradient Boosting Machines (GBM): It is a powerful ensemble learning technique used in both regression and classification tasks.They belong to boosting algorithms, which combine multiple weak learners sequentially to create a strong predictive model.
- Support Vector Machines (SVM): Effective for high-dimensional data and non-linear decision boundaries

6. Model Evaluation :- Once the model is trained, it needs to be evaluated using appropriate metrics to assess its performance and generalization capabilities.

- Performance Metrics: Assess model performance using a combination of evaluation metrics tailored to the specific business context and objectives. Interpret the implications of evaluation metrics in the context of business costs, benefits, and operational constraints to prioritize model performance objectives.
- Cross-Validation: Employ cross-validation techniques such as k-fold cross-validation or stratified cross-validation to estimate model generalization performance and mitigate overfitting.

7. Deployment Considerations:- After selecting and evaluating the model, it needs to be deployed in a production environment for real-time churn prediction.

- Scalability: Design scalable architectures and workflows capable of handling large volumes of data and supporting real-time or batch predictions.
- Real-time Prediction: Implement streaming data processing pipelines or micro services architectures to enable real-time churn prediction and intervention. Integrate with operational systems, marketing platforms, or customer engagement channels to trigger timely retention actions based on churn predictions.
- Model Monitoring and Maintenance: Establish monitoring mechanisms to track model performance degradation, data drift, and concept drift over time.

By following the methodology outlined in this paper, businesses can leverage machine learning techniques to accurately predict churn, implement targeted retention strategies, and ultimately improve customer satisfaction and loyalty.

III. PROPOSED SYSTEM

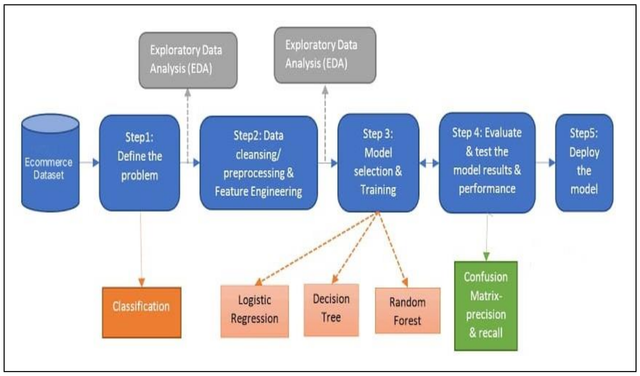

Exploratory Data Analysis (EDA)
Exploratory Data Analysis (EDA)
Ecommerce Dataset
Step1: Define the problem
Step2: Data cleansing/ preprocessing & Feature Engineering
Step 3: Model selection & Training
Step 4: Evaluate & test the model results & performance
Step5: Deploy the model
Classification
Logistic Regression
Decision Tree
Random Forest
Confusion Matrix- precision & recall

Fig. 1. Proposed System

We used machine learning models to create a single-page web application for customer churn prediction. We taken here exploratory data predict to identify missing values and other some variables, and columns that have a high impact on customer churn in some few years. Our dataset includes 5630 unique client IDs, and all columns with n = 5630 have no missing values. We then split the data into a 90 percent training dataset and a 10 percent test dataset. We trained two base learners - Logistic Regression, Decision Trees, Random Forests[13].

The proposed system for customer churn prediction using machine learning encompasses a comprehensive methodology aimed at effectively anticipating and mitigating customer attrition. At the core of this system lies the integration of machine learning techniques with robust data preprocessing, feature engineering, model selection, evaluation metrics, and deployment considerations.

Initially, data collection from diverse sources including customer demographics, transaction history, and interaction logs forms the foundation. Following this, meticulous data preprocessing techniques ensure data integrity by handling missing values, removing duplicates, and treating outliers[14]. Feature engineering plays a pivotal role in enriching the dataset with derived features such as customer tenure, transaction frequency, and engagement metrics. Subsequently, a range of classification algorithms including logistic regression, decision trees, random forests, and support vector machines are evaluated for their efficacy in predicting churn. Hyperparameter tuning further optimizes model performance to achieve the desired predictive accuracy.

Model evaluation employs a suite of performance metrics such as accuracy, precision, recall, and F1-score, providing insights into the model's effectiveness. Cross-validation techniques ensure robustness and generalization of the predictive models. Deployment considerations encompass scalability, real-time prediction capabilities, and interpretability of the model, ensuring its practical applicability in real-world scenarios[15]. By implementing this proposed system, businesses can proactively identify customers at risk of churn, formulate targeted retention strategies, and foster long-term customer relationships, thereby enhancing profitability and sustaining competitive advantage in dynamic market environments.

IV. ALGORITHM

A. *Decision Tree*

Decision Tree is a Regulated learning procedure that can be utilized for both characterization and Relapse problems, however for the most part it is liked for taking care of Order issues. It is a tree-organized classifier, where interior hubs address the highlights of a dataset, branches rep- despise the choice guidelines and each leaf hub addresses the outcome[16].

In a Decision tree, there are two hubs, which are the Choice Hub and Leaf Hub. Choice hubs are utilized to go with any choice and have different branches, while Leaf hubs are the result of those choices and contain no further branches. The choices or the test are performed based on elements of the given dataset. It is a graphical portrayal for getting every one of the potential answers for an issue/choice in view of given conditions. It is known as a choice tree in light of the fact that, like a tree, it begins with the root hub, which develops further branches and builds a tree-like construction.

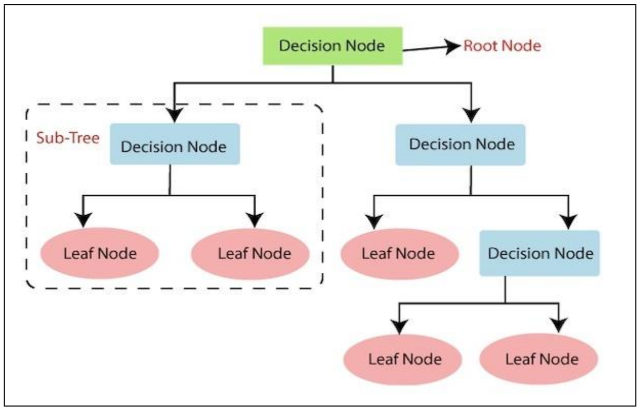


Fig. 2. Decision Tree

Step-1: Start the tree with the node of root

Step-2: Track down the best quality in the dataset utilizing Trait Determination Measure (ASM).

Step-3: Partition the S into subsets that contains potential qualities for the best ascribes.

Step-4: Produce the choice tree hub, which contains the best quality.

Step-5: Recursively settle on new choice trees utilizing the subsets of the dataset made in step - 3. Proceed with this interaction until a phase is reached where you can't further characterize the hubs and called the last hub as a leaf hub.

B. *Random Forest*

Random is a famous AI calculation that has a place with the managed learning procedure. It can be utilized for both Order and Relapse issues in AI. It depends on the idea of en-semble realizing, which is a course of joining numerous classifiers to tackle a perplexing issue and to work on the presentation of the model[17].

"Random Forest is a classifier that contains various choice trees on subsets of the given dataset " Rather than depending on one choice tree, the arbitrary timberland takes the forecast from each tree and in light of the larger part votes of expectations, and it predicts the last result.

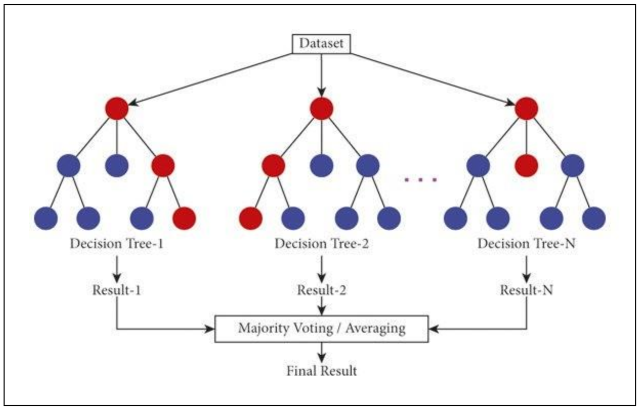


Fig. 3. Random Forest

Step-1: Select arbitrary K data of interest from the preparation set.

Step-2: Fabricate the choice trees related with the chose pieces of information (Subsets).

Step-3: Pick the number N for choice trees that you need to construct.

Step-4: Re-hash Stage 1 and 2.

Step-5: For new pieces of information, find the expectations of every choice tree, and appoint the new information focuses to the classification that wins the larger part casts a ballot.

C. *Logistic Regression*

Logistic regression is utilized for paired order where we utilize sigmoid capability, that accepts input as in-subordinate factors and delivers a likelihood esteem somewhere in the range of 0 and 1.

For instance, we have two classes Class 0 and Class 1 in the event that the worth of the strategic capability for an info is more noteworthy than 0.5 (edge esteem) then it has a place with Class 1 it has a place with Class 0. It's alluded to as relapse since it is the expansion of straight relapse however is mostly utilized for arrangement problems[18]. Logistic regression is widely used because of its simplicity, interpretability, and efficiency. It's also a fundamental building block in more complex machine learning algorithms and techniques.

Step 1: Information Pre-handling

Step 2: Fitting Calculated Relapse to the Preparation

Step 3: Foreseeing the Experimental outcome

Step 4: Test precision of the outcome

Step 5: Picturing the preparation set outcome

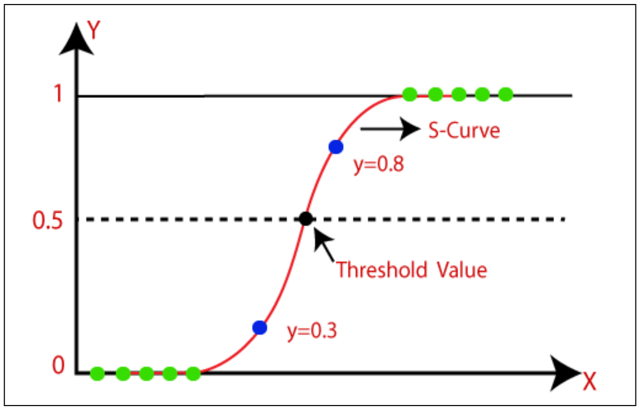
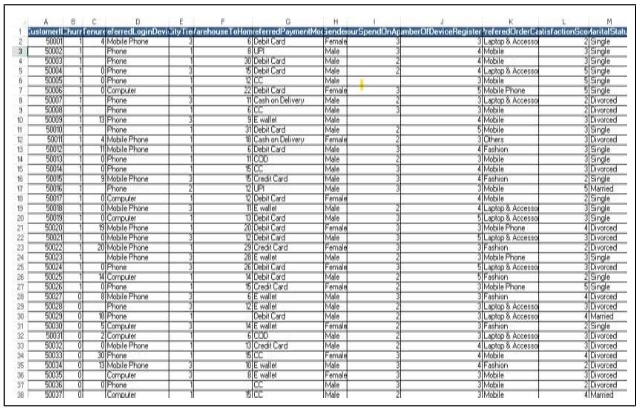


Fig. 4. Logistic Regression

V. RESULTS

A. *Dataset*

For this project, we have used the following dataset from Kaggle. Dataset contains 5630 rows and 20 columns.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	CustomerI	Churn	Tenure	eferredLoginDevi	CityTie	arehouseToHom	referredPaymentMo	Gende	ourSpendOnAp	umberOfDeviceRegister	referedOrderCa	atisfactionSco	aritalStatu
2	50001	1	4	Mobile Phone	3	6	Debit Card	Female	3	3	Laptop & Accesso	2	Single
3	50002	1		Phone	1	8	UPI	Male	3	4	Mobile	3	Single
4	50003	1		Phone	1	30	Debit Card	Male	2	4	Mobile	3	Single
5	50004	1	0	Phone	3	15	Debit Card	Male	2	4	Laptop & Accesso	5	Single
6	50005	1	0	Phone	1	12	CC	Male		3	Mobile	5	Single
7	50006	1	0	Computer	1	22	Debit Card	Female	3	5	Mobile Phone	5	Single
8	50007	1		Phone	3	11	Cash on Delivery	Male	2	3	Laptop & Accesso	2	Divorced
9	50008	1		Phone	1	6	CC	Male	3	3	Mobile	2	Divorced
10	50009	1	13	Phone	3	9	E wallet	Male		4	Mobile	3	Divorced
11	50010	1		Phone	1	31	Debit Card	Male	2	5	Mobile	3	Single
12	50011	1	4	Mobile Phone	1	18	Cash on Delivery	Female	2	3	Others	3	Divorced
13	50012	1	11	Mobile Phone	1	6	Debit Card	Male	3	4	Fashion	3	Single
14	50013	1	0	Phone	1	11	COD	Male	2	3	Mobile	3	Single
15	50014	1	0	Phone	1	15	CC	Male	3	4	Mobile	3	Divorced
16	50015	1	9	Mobile Phone	3	15	Credit Card	Male	3	4	Fashion	2	Single
17	50016	1		Phone	2	12	UPI	Male	3	3	Mobile	5	Married
18	50017	1	0	Computer	1	12	Debit Card	Female		4	Mobile	2	Single
19	50018	1	0	Mobile Phone	3	11	E wallet	Male	2	4	Laptop & Accesso	3	Single
20	50019	1	0	Computer	1	13	Debit Card	Male	3	5	Laptop & Accesso	3	Single
21	50020	1	19	Mobile Phone	1	20	Debit Card	Female	3	3	Mobile Phone	4	Divorced
22	50021	1	0	Mobile Phone	3	12	Debit Card	Male	3	5	Laptop & Accesso	3	Divorced
23	50022	1	20	Mobile Phone	1	29	Credit Card	Female	3	3	Fashion	2	Divorced
24	50023	1		Mobile Phone	3	28	E wallet	Male	2	3	Mobile Phone	3	Single
25	50024	1	0	Phone	3	26	Debit Card	Female	3	5	Laptop & Accesso	3	Divorced
26	50025	1	14	Computer	1	14	Debit Card	Male	2	5	Fashion	2	Single
27	50026	1	0	Phone	1	15	Credit Card	Female	2	3	Mobile Phone	5	Single
28	50027	0	8	Mobile Phone	3	6	E wallet	Male	3	3	Fashion	4	Divorced
29	50028	0		Phone	3	12	E wallet	Male	2	3	Laptop & Accesso	3	Divorced
30	50029	0	10	Phone	1		Debit Card	Male	2	3	Laptop & Accesso	4	Married
31	50030	0	5	Computer	3	14	E wallet	Female	2	3	Fashion	2	Single
32	50031	0	2	Computer	1	6	COD	Male	2	3	Laptop & Accesso	3	Divorced
33	50032	0	0	Mobile Phone	1	13	Credit Card	Male	2	4	Laptop & Accesso	3	Divorced
34	50033	0	30	Phone	1	15	CC	Female	3	4	Mobile	4	Divorced
35	50034	0	13	Mobile Phone	3	10	E wallet	Male	3	4	Fashion	2	Divorced
36	50035	0		Computer	3	8	E wallet	Female	3	3	Mobile	3	Divorced
37	50036	0	0	Phone	1		CC	Male	3	3	Mobile	2	Divorced
38	50037	0		Computer	1	15	CC	Male	2	3	Mobile	4	Married

Fig. 5. Dataset

B. *The distribution of Tenure column*

Less tenure has more probability to churn maximum tenure is more than 20, after that there are low chances that customer will churn. The below graph shows relationship between tenure and churning. The lesser tenure, then most chances to churn. So, one of the best way to reduce customer churning would be to keep the customer for longer tenure, so it lower the chances of churning..

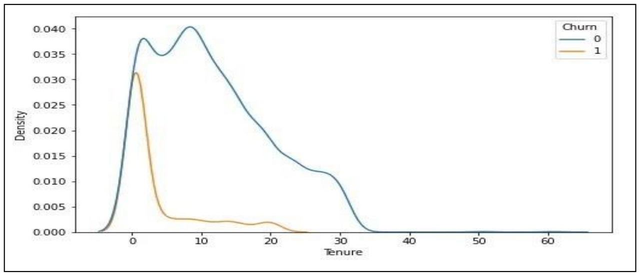
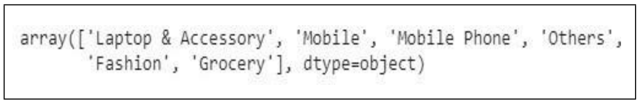


Fig. 6. The distribution of Tenure column

C. *Unqiue values in PreferedOrderCat column.*

As we can see from the figure below, there are 2 redundant categories, Mobile and Mobile Phone.



```
array(['Laptop & Accessory', 'Mobile', 'Mobile Phone', 'Others',
       'Fashion', 'Grocery'], dtype=object)
```

Fig. 7. Unqiue values in PreferedOrderCat column

D. *Preferred Order Category and Churn.*

We can plot the Preferred order category and see variation with Churn. The customers who churn more prefer mobile phones. This means there is some issue with mobile phones. Either the quality is not proper or the services are bad. Maybe customers prefer to buy other things like grocery or laptop or other, in-person.

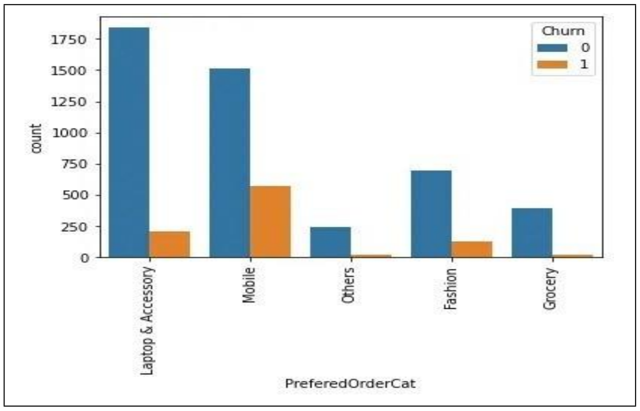


Fig. 8. Preferred Order Category and Churn

E. *Variation of Gender and Churn.*

Here we got 17.73 percent result. Male customers are most likely to churn. On an average almost 18 percent Male customers are churn.

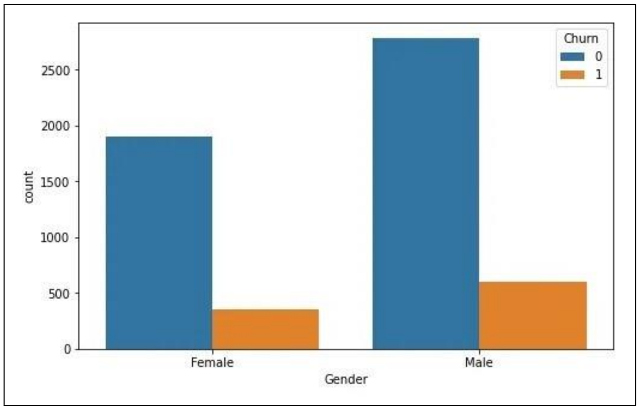


Fig. 9. Variation of Gender and Churn

F. *Variation of Number of Devices Registered and Churn.*

Here customers with most number of registered devices are likely to churn. May be they are facing some issues with multiple devices or the interface doesn't work properly on another devices. For an E-Commerce website, the interface is more important for user engagement.

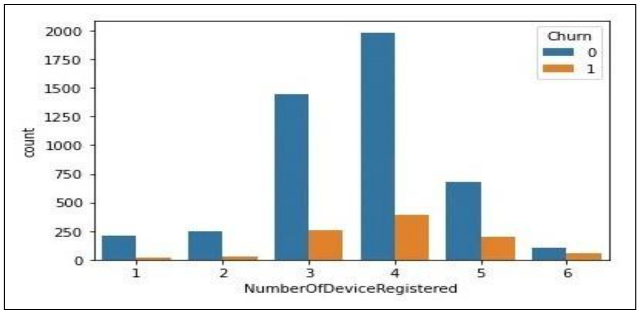


Fig. 10. Number of Devices Registered and Churn

G *Variation of WareHouseToHome columns and Complaint column with Churn.*

From the below figure, we can see that most of the customers that churn are having most of complaints. It shows that the organisation or the company should focus on solving the customer problems. We can say that complaints is one of the valid reasons for customer churn.

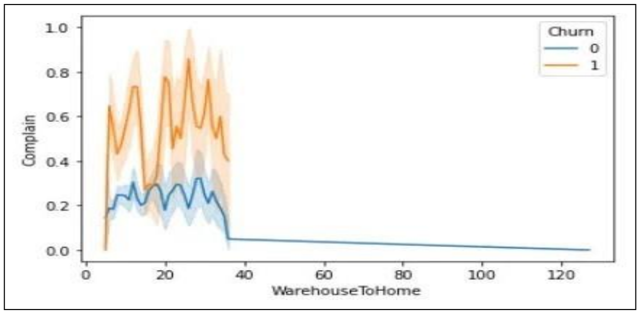


Fig. 11. Variation of WareHouseToHome

H *Which Category is preferred mostly?*

From the graph, we can see here that the preferred category is Laptop and Accessory.

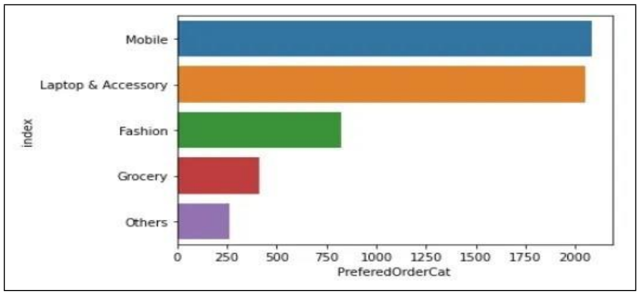


Fig. 12. Category is preferred mostly

I *Correlation of Satisfaction Score and Cashback Amount.*

In below graph, we all can understand that cashback amount is not directly related to satisfaction score.

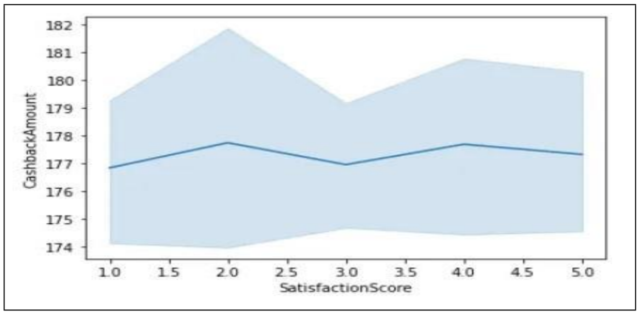
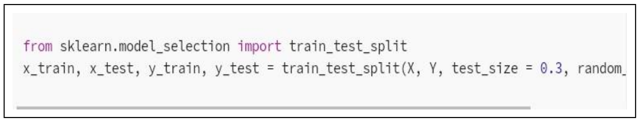


Fig. 13. Correlation of Satisfaction Score and Cashback Amount

J *Splitting into training and testing dataset.*

For splitting the dataset, we have used train test split method from sklearn model selection library. we can import here train test split method and split the dataset. So, for, we have splitted the dataset in a way that 70 percent of the dataset is for training and rest 30 percent is for testing. We can use any ratio, like 60:40, 70:30, 80:20 as per the dataset given.



```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size = 0.3, random_
```

Fig. 14. Split into training and testing dataset

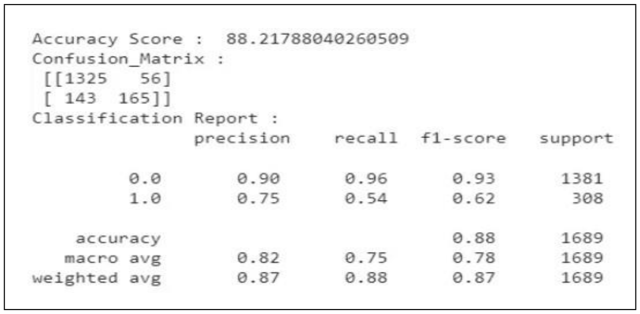
K *Model Training.*

Model training is one of most important steps while doing a prediction. Here we need to try a lot of machine learning algorithms to see which one is performing best on the dataset. The performance of algorithms is largely depend on the quality of data. To evaluate the performance of the different algorithms we use function like accuracy score which calculates the accuracy of model, other function can be confusion matrix[19]. You might find it very intense if you are listening the names of these functions for the first time. If accuracy of the model is 90 percent, it means that it will the predicted 90 percent of the samples or records correctly. Before understanding precision, recall, here we need to know certain terms like true positive, true negative, false positive, false negative.

- True Positive: This means the records that are labeled as positive and are positive in reality.
- True Negative: This means the records that are labeled as negative and are negative in reality.
- False Positive: This means the records that are labeled as positive but are negative in reality.
- False Negative: This means the records that are negative but are actually positive.

L. *Logistic Regression.*

As we have seen, the accuracy of Logistic Regression came out to be 88 percent. However, in an imbalanced dataset, we don't just depend on accuracy to evaluate the performance of the model. So, let's watch at the value of precision and recall. Precision for class 0.0(not churn) came out to be 0.90 and for class 1.0 is (churn) came out to be 0.75. Recall for class 0.0is (not churn) came out to be 0.96 and for class 1.0(churn), it came out be 0.54. The main idea behind project is to predict churning. So, after using logistic regression, we can again see the values of precision and recall were very low for Churn class[20].

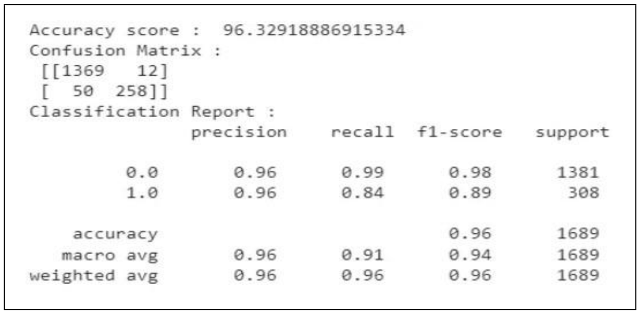


```
Accuracy Score :  88.21788040260509
Confusion_Matrix :
 [[1325   56]
 [ 143  165]]
Classification Report :
              precision    recall  f1-score   support

         0.0       0.90      0.96      0.93      1381
         1.0       0.75      0.54      0.62       308

    accuracy                           0.88      1689
   macro avg       0.82      0.75      0.78      1689
weighted avg       0.87      0.88      0.87      1689
```

Fig. 15. Logistic Regression

M. *Random Forest*

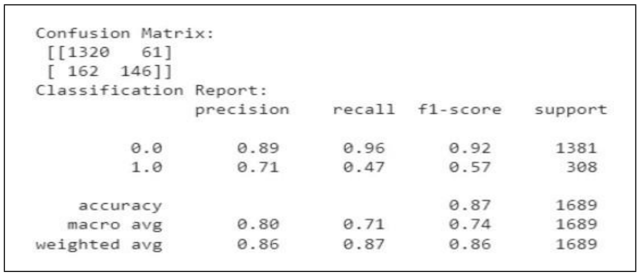
As we can see in the screenshot below, the accuracy for Random Forest came out to be 0.96. Then the values of precision and recall for Churn class were 0.96 and 0.84. So, by looking at these values, we can say that Random forest performed fairly well on the dataset.



```
Accuracy score :  96.32918886915334
Confusion Matrix :
 [[1369   12]
 [  50  258]]
Classification Report :
              precision    recall  f1-score   support

         0.0       0.96      0.99      0.98      1381
         1.0       0.96      0.84      0.89       308

    accuracy                           0.96      1689
   macro avg       0.96      0.91      0.94      1689
weighted avg       0.96      0.96      0.96      1689
```

Fig. 16. Random forest

N. *Decision Tree*



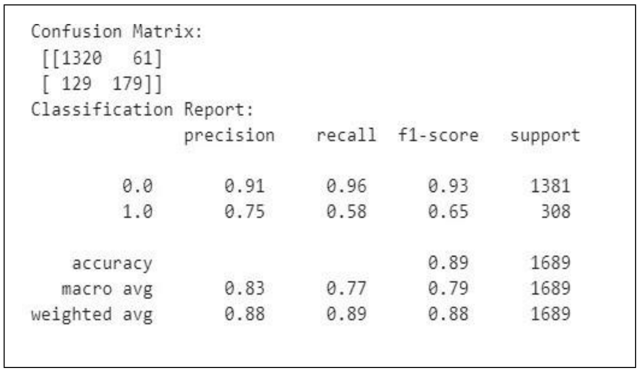
```
Confusion Matrix:
 [[1320   61]
 [ 162  146]]
Classification Report:
              precision    recall  f1-score   support

         0.0       0.89      0.96      0.92      1381
         1.0       0.71      0.47      0.57       308

    accuracy                           0.87      1689
   macro avg       0.80      0.71      0.74      1689
weighted avg       0.86      0.87      0.86      1689
```

Fig. 17. Decision Tree

Decision tree model given an accuracy of the 0.87. The values for the precision and recall are 0.71 and 0.47 for each one. The recall for churn classes is very low, here which shows the bad performance of the model.

O. *AdaBoost*

Here Adaboost model gave an accuracy of 0.89. The precision values and recall values are 0.75 and 0.58 respectively. The recall value is very low, which shows the poor performance of the model here. Till now, we have used here many different supervised machine learning algorithms like Logistic Regression, Decision Trees, Random Forest and AdaBoost. Out of all this many algorithms, the performance of Random Forest has the best with an accuracy of 96 percent and F1-score of 0.87.



```
Confusion Matrix:
 [[1320   61]
 [ 129  179]]
Classification Report:
              precision    recall  f1-score   support

         0.0       0.91      0.96      0.93      1381
         1.0       0.75      0.58      0.65       308

    accuracy                           0.89      1689
   macro avg       0.83      0.77      0.79      1689
weighted avg       0.88      0.89      0.88      1689
```

Fig. 18. AdaBoost

P. *Handling imbalance dataset*

After applying SMOTE algorithm, we got a new set of training and testing dataset, x train sm and y train sm. ADASYN:- After applying ADASYN algorithm, we got a new set of training and testing dataset, x train sm and y train sm. One important thing that remain stable, before using any of the technique's and after using technique's like SMOTE, was that Random Forest outperformed all the algorithms. Random Forest without any im-balancing technique gives an accuracy in form of 0.96. So, here we decided to go with this algorithm and try to improve its performance.

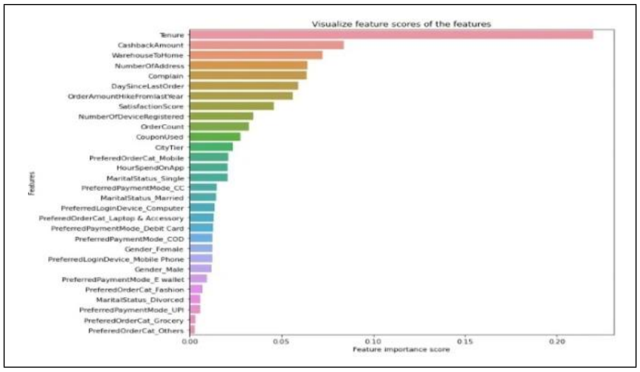
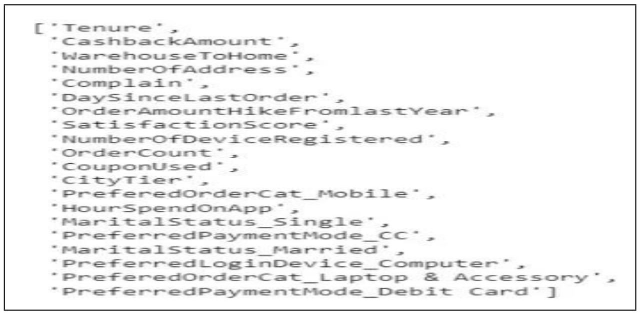


Fig. 19. Handling imbalance dataset

Q. *Extracting top 20 features*

To extract the top 20 features here , first we have to sort all the features as per their contribution in a table.

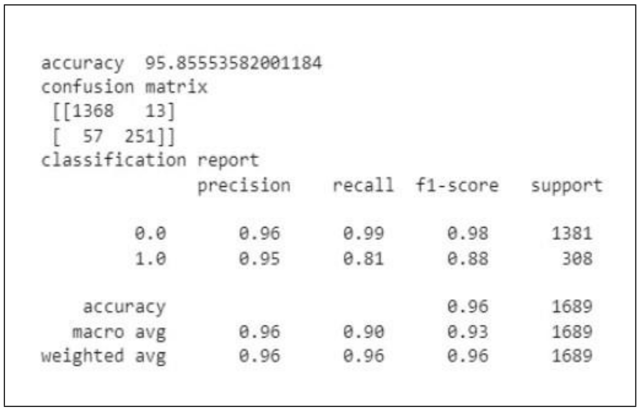


```
['Tenure',
 'CashbackAmount',
 'WarehouseToHome',
 'NumberOfAddress',
 'Complain',
 'DaySinceLastOrder',
 'OrderAmountHikeFromlastYear',
 'SatisfactionScore',
 'NumberOfDeviceRegistered',
 'OrderCount',
 'CouponUsed',
 'CityTier',
 'PreferedOrderCat_Mobile',
 'HourSpendOnApp',
 'MaritalStatus_Single',
 'PreferredPaymentMode_CC',
 'MaritalStatus_Married',
 'PreferredLoginDevice_Computer',
 'PreferedOrderCat_Laptop & Accessory',
 'PreferredPaymentMode_Debit Card']
```

Fig. 20. Extracting top 20 features

R. *Training Random Forest Model using top 20 features*

Random Forest performed properly well on the dataset using these top 20 features as well.



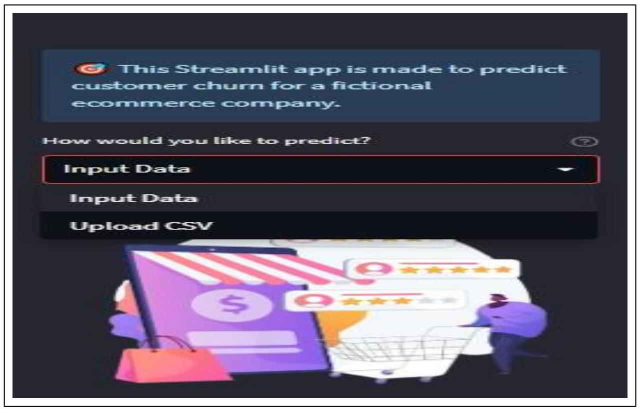
```
accuracy  95.85553582001184
confusion matrix
 [[1368   13]
 [  57  251]]
classification report
              precision    recall  f1-score   support

         0.0       0.96      0.99      0.98      1381
         1.0       0.95      0.81      0.88       308

    accuracy                           0.96      1689
   macro avg       0.96      0.90      0.93      1689
weighted avg       0.96      0.96      0.96      1689
```

Fig. 21. Training Random Forest Model using top 20 features

Now, we are done with part implementation.

VI. OUTPUT OF APPLICATION

1) When we open an application there is an option, whether we have to give input data manually or upload csv file.

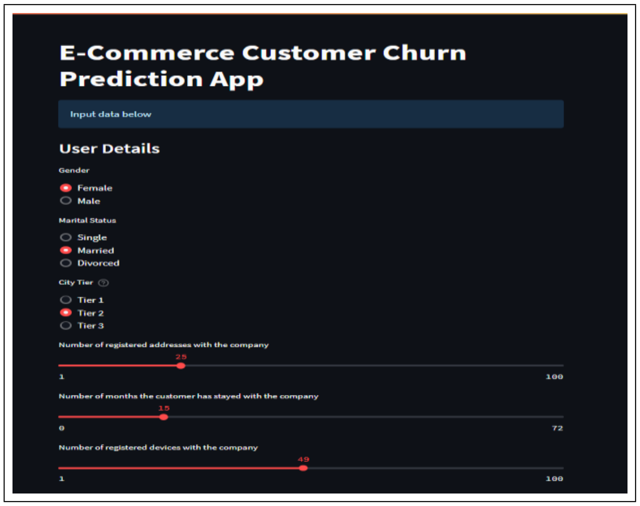


🎯 This Streamlit app is made to predict customer churn for a fictional ecommerce company.

How would you like to predict?

Fig. 22. Select option whether enter detail manually or to upload a csv file.

2) After selecting input data option there is first stage is user details in that we have to select gender , select martial status , select city tier, select number of registered addresses with the company, select number of months the customer has stayed with the company, select number of registered devices with the company, select number of hours spent on app/website in last month, select preferred Login Device of customer, select distance between warehouse to home of customer (in KM), select preferred Payment Mode.


E-Commerce Customer Churn Prediction App

Input data below

User Details

Gender
● Female
○ Male

Marital Status
○ Single
● Married
○ Divorced

City Tier
○ Tier 1
● Tier 2
○ Tier 3

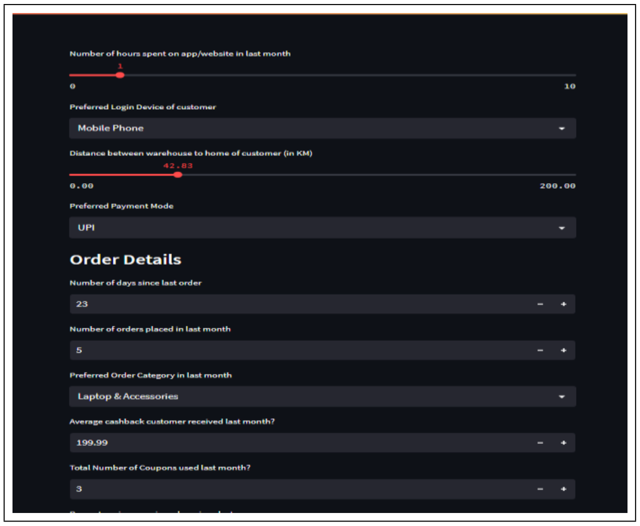
Number of registered addresses with the company
25
1 100

Number of months the customer has stayed with the company
15
0 72

Number of registered devices with the company
49
1 100

Fig. 23. Fill user deatails.

3) Then after this we have select the order details in this we have to select number of days since last order, select number of orders placed in last month, select preferred order category in last month, select average cashback customer received last month?, select total number of coupons used last month?, select percentage increase in orders since last year, select how satisfied customer is with your customer service?, select as customer filed any complaints in last month.


Number of hours spent on app/website in last month
1
0 10

Preferred Login Device of customer
Mobile Phone

Distance between warehouse to home of customer (in KM)
42.83
0.00 200.00

Preferred Payment Mode
UPI

Order Details

Number of days since last order
23

Number of orders placed in last month
5

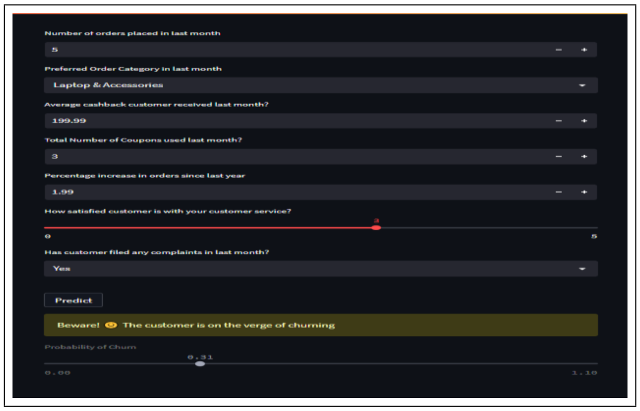
Preferred Order Category in last month
Laptop & Accessories

Average cashback customer received last month?
199.99

Total Number of Coupons used last month?
3

Fig. 24. Fill order details.

4) Then finally it will predict whether the customer will churn or not or either he is on verge of churning.


Number of orders placed in last month
5

Preferred Order Category in last month
Laptop & Accessories

Average cashback customer received last month?
199.99

Total Number of Coupons used last month?
3

Percentage increase in orders since last year
1.99

How satisfied customer is with your customer service?
3
0 5

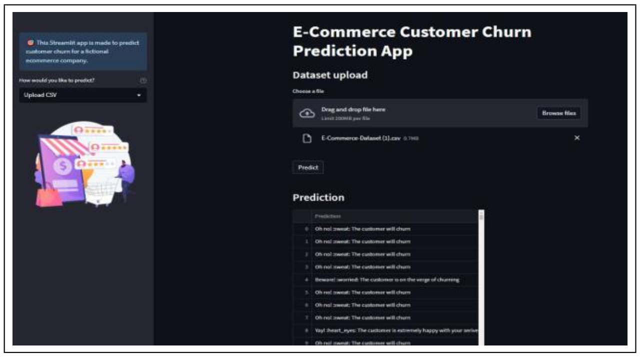
Has customer filed any complaints in last month?
Yes

Beware! 😟 The customer is on the verge of churning

Probability of Churn
0.31
0.00 1.10

Fig. 25. Press predict button it will show the result whether the customer will churn or not.

5) In Upload Csv File option , upload a dataset file then press predict button , it will predict for all the customer whether the customer will churn or not.


This Streamlit app is made to predict customer churn for a fictional ecommerce company.

How would you like to predict?
Upload CSV

E-Commerce Customer Churn Prediction App

Dataset upload

Choose a file

Drag and drop file here
Limit 200MB per file

E-Commerce-Dataset (1).csv 0.7MB

Prediction

	Prediction
0	Oh no! :sweat: The customer will churn
1	Oh no! :sweat: The customer will churn
2	Oh no! :sweat: The customer will churn
3	Oh no! :sweat: The customer will churn
4	Beware! :worried: The customer is on the verge of churning
5	Oh no! :sweat: The customer will churn
6	Oh no! :sweat: The customer will churn
7	Oh no! :sweat: The customer will churn
8	Yay! :heart_eyes: The customer is extremely happy with your servi
9	Oh no! :sweat: The customer will churn

Fig. 26. Upload CSV file then press predict button.

VII. CONCLUSION

Customer churn prediction is a critical task for businesses across various industries, as retaining existing customers is often more cost-effective than acquiring new ones. customer churn prediction is a vital component of customer relationship management and business strategy.By accurately identifying customers at risk of churn, businesses can implement targeted retention efforts, improve customer satisfaction, and drive long-term profitability. However, achieving effective churn prediction requires overcoming various challenges, leveraging advanced methodologies, and adapting to evolving customer behavior. As businesses continue to prioritize customer retention, the field of churn prediction will likely evolve with advancements in technology, data analytics, and customer-centric strategies. Ultimately, successful churn prediction empowers businesses to build stronger relationships with their customers and thrive in an increasingly competitive marketplace. By leveraging predictive analytics and data-driven insights, businesses can proactively mitigate churn, cultivate customer loyalty, and thrive in an increasingly competitive digital landscape.

ACKNOWLEDGMENT

In this survey paper, we extend our heartfelt gratitude to Dr. S. Brinthakumari for his invaluable contributions. His insightful technical discussions regarding blockchain technology and his expertise in real-world identity management issues have significantly enriched this research endeavor. His guidance has been instrumental in shaping the innovative solutions proposed in this paper.

REFERENCES

[1] Chih-Fong Tsai, Yu-Hsin Lu "Customer churn prediction by hybrid neural networks", Expert Systems with Applications 36 (2009) 12547–12553.

[2] V. Kumar, D. Singh, S. V. Singh and T. Anand, "The Role of Converged Network in Disruptive Technology," 2020 International Conference on Intelligent Engineering and Management (ICIEM), 2020, pp. 483-486, doi:10.1109/ICIEM48762.2020.9160337.

[3] P. Surya, P. Pachauri, A. Pachauri, P. Chaturvedi, S. A. Yadav and D. Singh, "Virtualization Risks and associated Issues in Cloud Environment," 2021 International Conference on Technological Advancements and Innovations ICTAI, 2021, pp. 521-525, doi:10.1109/ICTAI53825.2021.9673424

[4] D. Sikka, Shivansh, R. D and P. M, "Prediction of Delamination Size in Composite Material Using Machine Learning," 2022 International Conference on Electronics and Renewable Systems (ICEARS), 2022, pp. 1228-1232, doi:10.1109/ICEARS53579.2022.9752123.

[5] S. Markkandan, R. Logeshwaran, N. Venkateswaran, Analysis of precoder decomposition algorithms for MIMO system design, IETE J. Res. (2021), https:// doi.org/10.1080/03772063.2021.1920848.

[6] Manglani, R., Bokhare, A. 2021. Logistic regression model for loan prediction: A machine learning approach. 2021 Emerging Trends in Industry 4.0 ETI 4.0. https://doi.org/10.1109/eti4.051663.2021.9619201

[7] Ssu-Han Chen, "The gamma CUSUM chart method for online customer churn prediction", Elect

[8] Srinivasan, R., Subalalitha, C.N. Sentimental analysis from imbalanced code-mixed data using machine learning approaches. Distrib Parallel Databases (2021). https://doi.org/10.1007/s10619-021-07331-4.

[9] V. Urbancokova, M. Kompan, Z. Trebulova, M. Bielikova, BEHAVIOR-BASEDcustomer demography prediction in e-commerce, J. Electron. Commer. Res. 21 (2)(2020) 96–112.

[10] G. Klimantaviciute, Customer churn prediction in E-commerce industry, J. Mach.Learn. Res. 1 (2021) 1–14.

[11] Dilip Singh Sisodia, Somdutta Vishwakarma, Abinash Pujahari, Evaluation of machine learning models for employee churn prediction, in: International Conference on Inventive Computing and Informatics, 2017.

[12] Ponnan Suresh, J Robert Theivadas, V.S. HemaKumar, Daniel Einarson, Driver monitoring and passenger interaction system using wearable device in intelligent vehicle, Comput. Electr. Eng. 103 2022, 108323, https://doi.org/10.1016/j. compeleceng.2022.108323. ISSN 00457906.
13P. Routh, A. Roy, and J. Meyer, "Estimating customer churn under competing risks," Journal of the Operational Research Society.

[13] A. R. Nair, "Classification of Cardiac Arrhythmia of 12 Lead ECG Using Combination of SMOTEENN, XGBoost and Machine Learning Algorithms," in International Symposium on Embedded Computing and System Design (ISED), 2019

[14] S. B. Borah, S. Prakhya, and A. Sharma, "Leveraging service recovery strategies to reduce customer churn in an emerging market," J. of the Acad. Mark. Sci., vol. 48, no. 5, pp. 848–868, Sep. 2020, doi: 10.1007/s11747- 019-00634-0.

[15] Anurag Shrivastava; D. Haripriya; Yogini Dilip Borole; Archana Nanoty; Charanjeet Singh; Divyansh Chauhan, High performance FPGA based secured hardware model for IoT devices, International Journal of System Assurance Engineering and Management,2022-03,DOI: 10.1007/s13198-021-01605-x

[16] I. Figalist, C. Elsner, J. Bosch and H. H. Olsson, "Customer Churn Prediction in B2B Contexts," Conference: International Conference on Software Business, 2020

[17] A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo and S. Anwar, "Customer churn prediction in telecommunication industry using data certainty," Journal of Business Research, vol. 94, pp. 290-301, 2019

[18] T. Vafeiadis, K. Diamantaras and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," Simulation Modelling Practice andTheory, vol. 55, pp. 1-9, June 2015

[19] Z. Chen, F. Jiang, Y. Cheng, X. Gu, W. Liu and J. Peng, "XGBoost classifier for DDoS attack detection and analysis in SDN-based cloud," in 2018 IEEE international conference on big data and smart computing, 2018.

[20] P. Lalwani, M. M. Kumar, J. Singh Chadha and P. Sethi, "Customer churn prediction system: a machine learning approach," Computing, pp. 1-24, 2021

[21] . Mehta, M. Bukov, C. H. Wang, A. G. Day, C. Richardson, C. K. Fisher and D. J.Schwab, "A high-bias,l ow-variance introduction to Machine Learning for physicists," Physics Reports, vol. 810, pp. 1-124, 2019

[22] Simeone, "A Very Brief Introduction to Machine Learning With Applications to Communication Systems," IEEE Transactions on Cognitive Communications and Networking, vol. 4, no. 4, pp. 648-664, 2018.

[23] A. Kumar and M. Jain, Ensemble Learning for AI Developers: Learn Bagging, Stacking, and Boosting Methods with Use Cases, Apress, 2020