Churn Forecasting in E-commerce: A Predictive Approach

**Mr.Ezra Vethamani *S***

*Dept of CSE*

*SRM Institute of Science and Technology* [***ezravets@srmist.edu.in***](mailto:ezravets@srmist.edu.in)***\****

**Mr. Madhavan R**

*Dept of CSE*

*SRM Institute of Science and Technology* [***mr5906@srmist.edu.in***](mailto:mr5906@srmist.edu.in)***\****

**Mr. Hariharan A S**

*Dept of CSE*

*SRM Institute of Science and Technology* [***ah2513@srmist.edu.in***](mailto:ah2513@srmist.edu.in)

**Mr. Gowtham S**

*Dept of CSE*

*SRM Institute of Science and Technology* [***gs7902@srmist.edu.in***](mailto:gs7902@srmist.edu.in)

***Abstract*— The relationship between customers and the businesses they engage with is ever-evolving. In certain circumstances, customers may discontinue their relationship with a business for a multitude of reasons. This major challenge is faced by most businesses, irrespective of the measure of the success of the brand. Customer churn is the phenomenon that causes the outflow of customers from brands they regularly associate with. Businesses face numerous issues such as increasing intensity of competition, and exponentially growing customer acquisition rates, to state a few. Hence, businesses resort to using data-driven analytical approaches to resolve this challenge. This abstract aims to present a detailed overview of the primary aspects of Customer Churn Prediction, with a focus its importance, methodologies and prospective assets. The predictive modelling technique used in this project identifies potential customers who could be lost by the business and ensures that all measures are taken to retain their participation. Proactive strategies are implemented to mitigate customer churn in a step-by-step manner.**

***Keywords***— **churn , training ,testing**

#### Introduction

In the world of business, the word “prediction” refers to anticipation of future events and occurrences. Anyone who purchases a product-based or service-based commodity is a customer. If the product you created needs to sustain in this market you must collect the feedback either its good or bad and the that collective feedback must be analyzed and need for change must be done. It’s fierce competition in today's business landscape, retaining a customer is essential for an establishment's growth. The word churn signifies the figure at which a customer terminates doing business with the establishment over some time. Retaining such customers in a competitive world is vital for the substantial growth and revenue of the establishment. Every company has a separate unit of employees working with different churn models predicting mostly using manual outdated model applications which takes significant amount of time and error may occur too. In this project, we subjected the use of Deep Learning concepts which is a subset of Artificial Intelligence that supports the long-term run in this Artificial Intelligence Era. Over time the customer tends to shift over to other commodities. Our project stands outside the box as we use Deep Learning which lowers gaps between the theoretical and practical insights. This project reduces the processing time and helps the unit of employees to acquire precise predictions. Although the key feature of this model is the precious feedback of the customer. The proposed models range from

algorithms, like logistic regression and Support vector machine. Since the heart of the project depends on customer feedback it may vary the availability of data for various situations.

#### Literature Survey

In [1] this Ahmad, had chosen four tree based algorithms namely Decision Tree, Random Forest, GBM tree algorithm, and XGBOOST algorithm in SyriaTel. In which XGBOOST has shown 93.301% accuracy and he also believe Social Network Analysis plays major role in customer churn.

1. Senthilnayaki B,have proposed to built an automated application for long run .Stratified k-fold cross validation technique is used to select the best fit model by calculating the performance metrics.
2. Hong Xue and Wen-chao Lu works on implementing churn prediction model in the Super Markets for retaining customers by using C4.5 Decision Tree. This model speaks about the improved C4.5 model which is consider to be a prediction model with better validity and stability. This mode; provides more effective support for decision making in Super Market.
3. Mishra, took Ensemble based Classifiers such as Bagging, Boosting and Random Forest .Comparing Ensemble based Classifiers with the well-known classifiers The Random forest has greater accuracy of 91.66%.
4. David, proposes on using a deep feedforward neural network for classification accompanied by a sequential pattern mining method on high-dimensional sparse data.A causal Bayesian network to predict cause probabilities that lead to customer churn.
5. In this paper, cluster based model is compared against conventional model of churn prediction. The author signifies the use of K-means clustering by targeting substantial data about a customer to make decisions.
6. Yihui Deng et al, states that using ensemble model in bank user churn can score upto 80% accurate values. The authors used ensemble models like Catboost, Lightgbm to formulate feasible user retention scheme.
7. Shinjin Kang, conducted survey how churn works in various industries like business administration, marketing, IT, telecommunications, newspapers, insurance and psychology.he used log base dataset as an input .
8. The Authors Qi Tang et al., has used XGBoost and MLP to create a hybrid prediction model. The MLP is used to deal with one-hot vector transformed from the leaf number.
9. this paper, Sajad Fathi Hafshejani et al. says about the new kernal function which improves the efficiency of SVM. The Gaussian Kernal is mixed to derive an efficient result in term of classification.
10. In this paper, many regression models are discussed and evaluated. This research used various metrics such as R- Squared, Root Mean Square Error (RMSE), and Cross- Validation to measure the accuracy of these models. This research outcome provides valuable guidance for informed decision making using SVM.
11. The authors I. Sakthidevi et al. in this paper speaks about Resource Manager and Model Repository. The simulation model provides valuable insights into the behaviour and performance of the system under different workload scenarios and thereby improving the efficiency and automation of training and deploying Distributed Machine Learning.

In this paper [13], the author speaks about the retention of employee model using machine learning. In this paper, algorithms used f are Decision tree (DT), Ensemble with boosted tree, K-nearest neighbor (KNN) and Support Vector machine (SVM) are used. This provides insights about the retention model applications.

1. The authors, Efendi Efendi et al. speak about customer behavior and retention necessity of customer for an industry.
2. Weiyun Yin, speaks about customer churn prevention using the Random Forest model. This proves to give a good accuracy rate.
3. **EXISTING AND PROPOSED METHODS**

The existing system for customer churn prediction involves collecting historical customer data, preprocessing it to handle missing values and outliers, and engineering relevant features. The data is then divided into training and testing sets for model development and evaluation. ML algorithms like Logistic Regression or Random Forest are employed for classification. After training and fine-tuning, the model is deployed in a production environment. Continuous monitoring and periodic retraining ensure its accuracy over time. Feedback from stakeholders aids in refinement, and integration with business processes allows for timely action, such as targeted marketing campaigns. Adherence to data privacy regulations is crucial throughout the process.

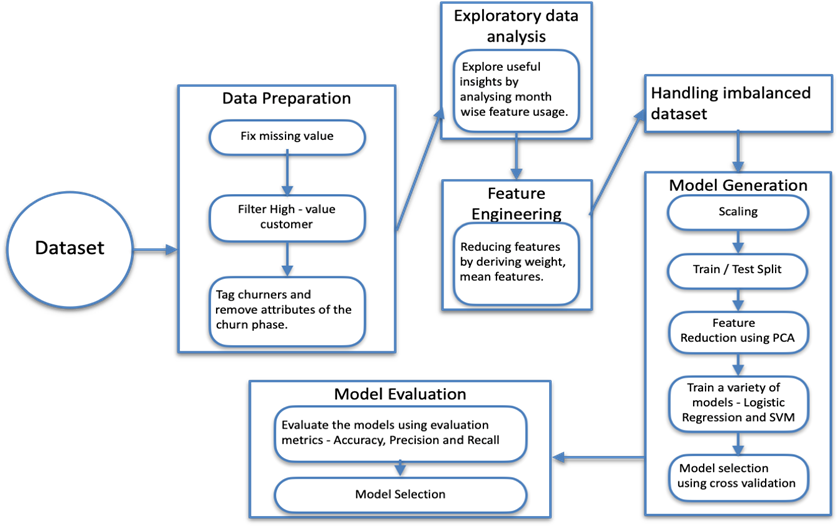
The proposed system employs advanced machine learning techniques to predict customer churn in real-time. It consists of key stages: data collection and preprocessing, feature selection and engineering, machine learning model training, real-time prediction, model monitoring, visualization and reporting, intervention strategies, and performance evaluation. Customer data is gathered with important features identified. A predictive model is trained on historical data, enabling real- time churn prediction. Regular monitoring and updates maintain model accuracy. Insights are visualised for stakeholders, and intervention strategies like tailored offers are implemented. Ultimately, the system empowers very proactive customer engagement and management, boosting retention and business performance.

#### Objective

The primary objective of this project is to develop an accurate and robust customer churn prediction model using advanced machine-learning techniques. This model will leverage historical customer data to identify potential churners, allowing businesses to proactively implement targeted

retention strategies. This project employs a systematic approach to predict customer churn. It refines data through cleaning, preprocessing, and feature engineering. Various machine learning algorithms are used for accurate churn prediction. The project emphasizes identifying crucial factors for targeted interventions. Clear documentation and ethical considerations are maintained. Overall, it aims to develop a precise churn prediction model that significantly strengthens customer retention efforts.

#### Architecture Diagram



**Fig.1**

**Figure.1** represents the architecture where the customer data is processed. We are splitting the data features for training, testing and validation. By training the model using the SVM

algorithm, the model is now trained for producing results with higher accuracy (88%). After the deployment of the model, we can predict if the customer will churn or not using the features used for training, testing and validation of the model.

#### Methodology

The methodology of this work contains various steps as given in the architecture diagram. The E- Commerce dataset had a total of around 5500 records which is used for training the model. The dataset is analysed next for missing values and then preprocessing is done. Then, the data is separated into training and testing dataset. Then these datasets are used in two different machine learning algorithms to find the customer churn. Finally, after training these models the probability of the churn is predicted and the model’s performance metrics like accuracy, precision and recall score are calculated for the evaluation.

#### Data Preparation:

The E COMMERCE dataset of customers is taken from Kaggle and they are trained using machine learning algorithms for prediction. The database consists of 5630 entries and three important features are taken for the prediction of the customer churn.

The top 3 features used are:

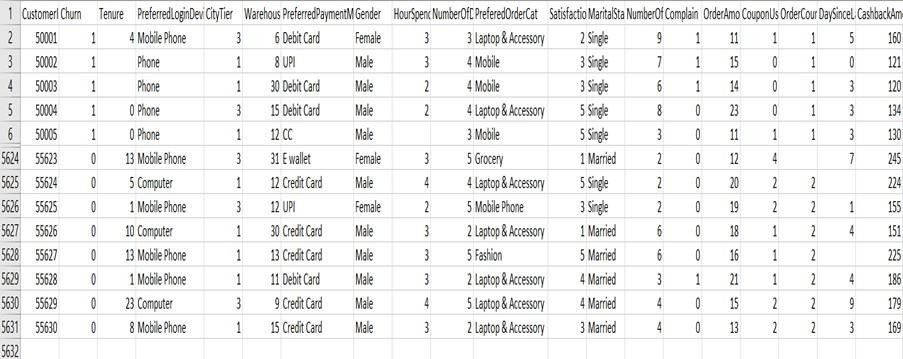
* + - Tenure
    - Distance from warehouse to home
    - Cash back Amount

The number of customers who churned is as follows:

|  |  |
| --- | --- |
| Churn | Non – churn |
| 948 | 4682 |

#### Table 1

The different parameters employed in this dataset are:



#### Fig.2 Datasets Sample

* 1. **Data Preprocessing:**

This is a very important building a model. The raw data is changed into the correct form for computation. Irrelevant and incomplete data is removed from the dataset to remove the overall complexity of the model. This step is where exploratory data analysis, feature engineering and handling imbalanced dataset is done as shown in the architecture diagram. The following steps are done to preprocess the data to build the machine learning algorithm:

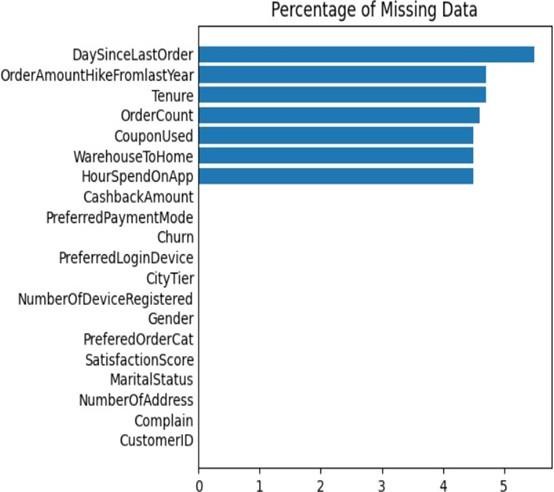
The number of table values missing in the table is mentioned below column- wise:

|  |  |  |
| --- | --- | --- |
| **S. No** | **Column** | **N u m b e r o f missing values** |
| 1. | CustomerID | 0 |
| 2. | Churn | 0 |
| 3. | Tenure | 265 |
| 4. | PreferredLoginDevice | 0 |
| 5. | CityTier | 0 |
| 6. | WarehouseToHome | 252 |
| 7. | PreferredPaymentMode | 0 |
| 8. | Gender | 0 |
| 9. | HourSpendOnApp | 256 |
| 10. | NumberOfDeviceRegistered | 0 |
| 11. | PreferedOrderCat | 0 |
| 12. | SatisfactionScore | 0 |
| 13. | MaritalStatus | 0 |
| 14. | NumberOfAddress | 0 |
| 15. | Complain | 0 |
| 16. | OrderAmount Last Year | 266 |
| 17 | OrderUsed | 257 |
| 18 | OrderCount | 259 |
| 19 | DaySinceLastOrder | 308 |
| 20 | CashbackAmount | 0 |

#### Table 2

A Graph for the percentage of missing data is shown in **Fig.3**

|  |  |  |
| --- | --- | --- |
| 19. | DaySinceLastOrder | {0, 46} |
| 20. | CashbackAmount | {0, 325} |

**Fig.3** Percentage of missing

**The lower and upper limits of all the parameters in this dataset are listed below:**

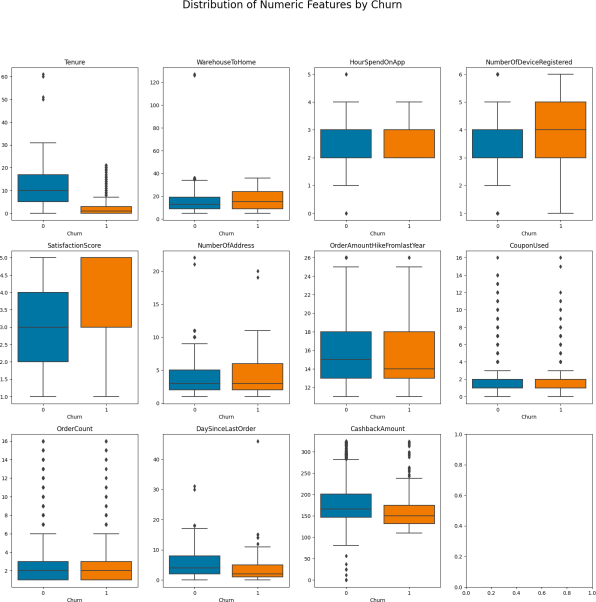
**Table 3**

### Distribution of numeric features by churn values

This representation is based on the distribution of numeric features by churn values. The categorical features are standardized in the next step.

**Fig.4** shows this distribution of numeric features in a plot format.

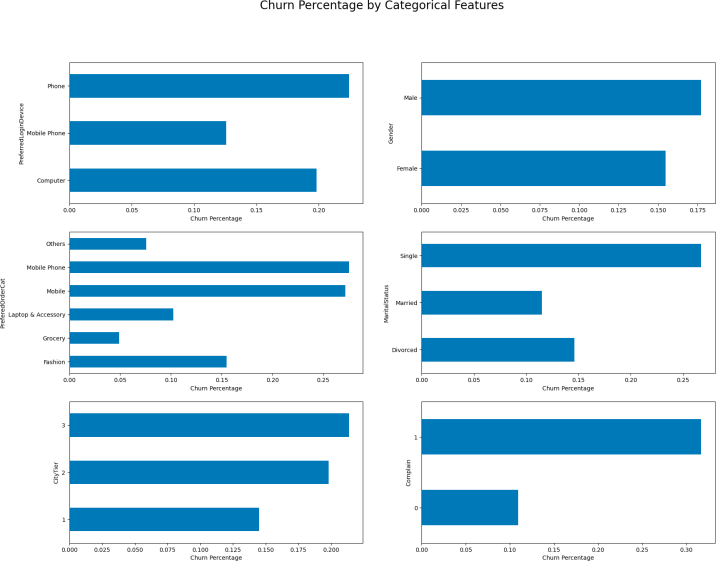
|  |  |  |
| --- | --- | --- |
| **S.**  **No** | **Column** | **Lower and Upper Limits** |
| 1. | CustomerID | {50001,  55630} |
| 2. | Churn | {0, 1} |
| 3. | Tenure | {0, 30} |
| 5. | CityTier | {1, 3} |
| 6. | WarehouseToHome | {6, 35} |
| 9. | HourSpendOnApp | {0, 3} |
| 10. | NumberOfDeviceRegistered | {1, 5} |
| 12. | SatisfactionScore | {2, 5} |
| 14. | NumberOfAddress | {1, 21} |
| 15. | Complain | {0, 1} |
| 16. | OrderAmountHikeFromlastYear | {11, 26} |
| 17. | CouponUsed | {0, 26} |
| 18. | OrderCount | {1, 16} |



**Fig.4** Numeric features by churn values

## CHURN PERCENTAGE BY CATEGORICAL FEATURES

This representation gives the churn percentage of categorical features. These features are standardized for better precision and accuracy levels. Categorical features are the ones which are composed of different categories. Fig.5 shows the different categorical features and their churn percentage in each category.



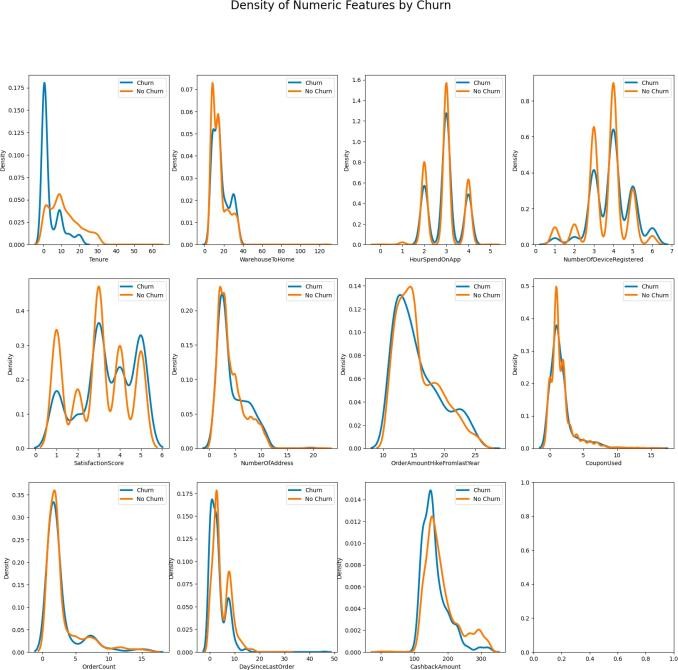
**Fig.5** Categorical features by churn

#### Impute missing values:

The missing value data must be treated for better accuracy. So, the median value is filled in the missing values in the table for better training of the model.

## DENSITY OF NUMERIC FEATURES BY CHURN

There are 11 numerical features taken in the dataset and these 11 are plotted in a graph with the values of churn and non- churn. Fig.6 shows the graphical representation of the density of numerical features given by churn.



**Fig.6** Density of numeric features

#### Remove Outliers:

The outliers are removed next, using IQR. An IQR is used to detect the outliers in the dataset. An outlier is a feature that is not related and is different from the dataset. The blank spaces and logical operators are replaced with different symbols, and text is converted into a regular expression, which helps the model be trained effectively.

text is converted into a regular expression, which helps the model be trained effectively.

## MODEL GENERATION:

The first part is splitting the dataset for training and testing. It is important to separate training and testing data to build a properly working machine learning model. The issue of overfitting can be avoided if this process is used. The dataset is separated into two different parts, namely training and testing data.

#### Training data:

* All the instances in the training set contain input features and their respective outputs.
* 80% of the data has been taken for training.
* The training of the model has been performed for 5630 epochs.
* The errors observed while training the model are rectified by repeatedly re-adjusting the weights and biases.

#### Testing data:

* Testing is performed on the data that was excluded from the training data.
* The remaining 20% of the data has been taken for testing.

#### MinMaxScaler:

The Min-Max scaler is like a handy tool in the world of data analysis and machine learning. It's a bit like a chef who adjusts the seasoning in a recipe to make sure it tastes just right. What it does is take a bunch of numbers and make sure they're all in a specific range, usually between 0 and 1. It does this by first figuring out the smallest number in the group and then subtracting it from each number. After that, it divides everything by the difference between the biggest and smallest numbers. This way, it keeps the relationships between the numbers intact but makes sure they're all playing nicely in the same range.

Imagine you have a group of friends, each with their own height. The Min-Max scaler would be like asking everyone to stand on their tiptoes so that the shortest person's head a touches the ceiling and the tallest person's head is just a little

below it. This way, everyone's height is still accurate relative to each other, but they're all within a certain range.

This technique is especially helpful when you're dealing with data that comes in all sorts of sizes. It's like making sure all the ingredients in a recipe are measured in the same units. This makes it easier for the computer to learn from the data

Where,

* x = Input value

***e***( ***β***0+***β***1***x***)

1 + ***e***(***β***0+***β***1***x***) (1)

and figure things out. However, it's worth mentioning that the Min-Max scaler can get a bit particular if there are some really extreme values in the mix. In those cases, there are other techniques, like standardisation, that might be a better fit.

#### StandardScaler:

The Standard Scaler is like a language translator for data, making numbers from different groups understand each other better. It adjusts them to center around a common value, making comparisons easier. It's like using the same measuring tape on a construction project to ensure everyone's on the same page. However, it works best with data that follows a regular pattern. If the data is scattered, other techniques like Min-Max scaling might be more suitable. It's all about choosing the right tool.

MinMaxScaler is used here for training this model. The top 3 features used to predict churn in this model are: tenure, cash back amount, and distance from warehouse to home. The train and test models are transformed and fit into the model.

The next step is choosing the algorithm and using the model with the highest accuracy to train the model.

The two algorithms chosen for training the model are **Logistic Regression** and **SVM.**

### Logistic Regression

Logistic Regression is a machine learning model which is used as classification algorithm. In this, it mainly used to check if the given dataset belongs to class or not. It is mainly used to solve classification problem. In Logistic Regression, we do not use regression line instead we use “S” shaped function to predict the value. They are of three types, namely Binomial, multinomial and ordinal. In our dataset we used Binomial which can have only two dependent variables e.g., a customer will continue or not.

#### Sigmoid function:

A sigmoid or squashing function, which takes input from previous hidden layer and squash it between 0 or 1. Similarly in Logistic Regression, it takes output of linear regression as input and use sigmoid function to calculate the probability of the class.

The formula for Logistic Regression using sigmoid function is as follows:

* e = Natural logarithm
* β0 = Bias or intercept term •
* β1 = Coefficient of input (X)

# Training the logistic regression model:

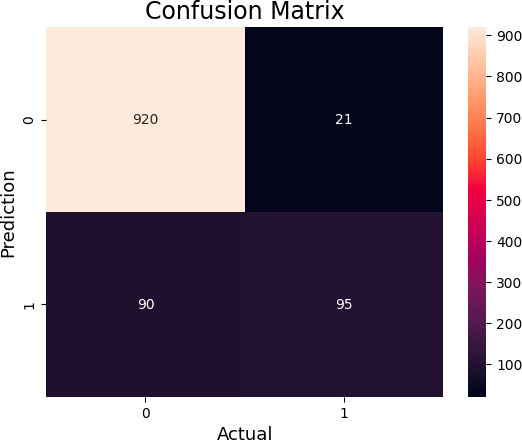
This logistic regression model gives a CV score of about: 0.8782461189647567 which equals to 87%.

**The below table gives the classification report of the same.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1 score | Support |
| 0 | 0.91 | 0.97 | 0.94 | 941 |
| 1 | 0.80 | 0.54 | 0.64 | 185 |
| Accuracy |  |  | 0.90 | 1126 |
| M a c r o avg | 0.86 | 0.75 | 0.79 | 1126 |
| Weighted avg | 0.90 | 0.90 | 0.89 | 1126 |

### Confusion Matrix

Confusion Matrix is a N X N matrix which is used for evaluating the performance of the model. Fig.8 shows the confusion matrix of the SVM model the performance ofthe model. Fig.7 shows the confusion matrix of the logistic regression model.



**Fig.7** Confusion Matrix of Logistic Regression

## SVM

SVM or Support Vector Machine is a primordial learning algorithm used in machine learning models. It is one of the most robust prediction methods for handling complex datasets and a well known algorithm for practical applications. SVM categories even when the data is not linearly separable, by mapping data to the high dimensional feature space.

#### LINEAR SVM:

The hyperplane is determined by maximising the margin, which represents the distance between the hyperplane and the nearest data points from each class. These critical points are called Support Vectors.

The formula for the classification is as follows:

**f(x)=sin(w****x+b)** …… (2)

Where,

* + w is the weight vector that is orthogonal to the hyperplane.
  + x is the input feature vector.
  + b is the bias term (also known as the offset).

# Training the SVM model:

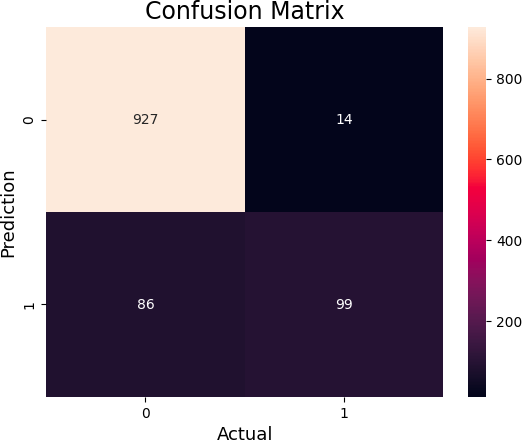
Training the SVM model gives a CV score of about: 0.8804764460647728 which equals to 88%.

**The below table gives the classification report of the same.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1 score | Support |
| 0 | 0.92 | 0.99 | 0.95 | 941 |
| 1 | 0.88 | 0.54 | 0.66 | 185 |
| Accuracy |  |  | 0.91 | 1126 |
| M a c r o avg | 0.90 | 0.76 | 0.81 | 1126 |
| Weighted avg | 0.91 | 0.91 | 0.90 | 1126 |

### Confusion Matrix

Confusion Matrix is a N X N matrix which is used for evaluating the performance of the model. Fig.8 shows the confusion matrix of the SVM model.

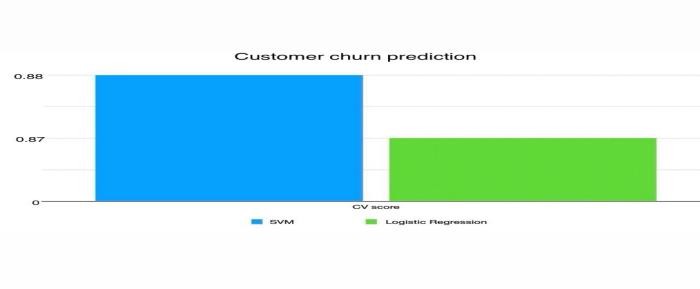
We use SVM model to predict the customer likeliness to continue certain products and it clearly gives an accurate prediction for the customer churn. The commodity analysis is essential for the provider to sustain and maintain the growth.

Finally, the ‘classification\_report ’is imported and the ‘cv\_score ’is calculated. The Logistic Regression model gives an accuracy of about 87% and SVM model gives an accuracy

of about 88%. **Fig.8** Confusion Matrix of SVC

## MODEL EVALUATION:

Using the above CV scores, classification reports and confusion matrix, the model with more accuracy is found to be SVM with an accuracy rate of 88%. This is shown in Fig.10. This SVM model is trained to know if the customer will churn or continue to use the specified E-Commerce website.



**Fig.9** Accuracy of the models

#### RESULTS AND CONCLUSIONS

In the era of AI and the brink of the emergence of super intelligence, industries have to keep up their phases to sustain in the corporate world. If a company is to grow and sustain itself among its competitors investing in acquiring new customers and retaining the old customers is vital. Both time and effort need to be put in to fill the leaving customer. Being able to predict when a customer is about to leave can give a huge saving to the business. Customer churn prediction is an existing application that provides the concern of customer attrition rate.

In this paper, we have included two such prediction models for the churn prediction application. SVM, known for its effectiveness in complex classification tasks, demonstrated promising results in accurately predicting customer churn giving an accuracy of 88%. In addition, the Regression model provides valuable insights into the quantitative aspects of churn prediction, enabling a more profound comprehension of the fundamental patterns and tendencies providing an accuracy of 87%.

In conclusion, our study underscores the effectiveness of combining SVM and Regression techniques for customer churn prediction. We have developed a robust and accurate model that empowers businesses to proactively manage customer retention. With an impressive accuracy rate of 88%, the proposed models can effectively help industrial meters identify potential churners thereby making these models more accurate than other existing model

1. **REFERENCES**
2. Abdelrahim Kasem Ahmad, Assef Jafar and Kadan Aljoumaa ,” Customer churn prediction in telecom using machine learning in big data platform “,2019.
3. Senthilnayaki B, Swetha M, Nivedha D, “CUSTOMER CHURN PREDICTION”,2021

International Advanced Research Journal in Science, Engineering and Technology (IARJSET), Chennai, India,2021.

1. Hong Xue and Wen-chao Lu, "Research of customer churn prediction model in a supermarket", 2011 International Conference on E-Business and E- Government (ICEE), Shanghai, China,2011.
2. Abinash Mishra, U. Srinivasulu Reddy,”A comparative study of customer churn prediction in telecom industry using ensemble based classifiers”, 2017 International Conference on Inventive Computing and Informatics (ICICI),Coimbatore ,India ,2017.
3. David Hason Rudd, Huan Huo, Guandong Xu,”Causal Analysis of Customer Churn Using Deep Learning”,2023 International Conference on Advancement inComputation & Computer Technologies (InCACCT),Australia,2023.
4. Harish A S and Malathy C, "Evaluative study of cluster based customer churn prediction against conventional RFM based churn model",2023 Second International Conference on Electrical, Electronics, I n f o r m a t i o n a n d C o m m u n i c a ti o n T e c h n o l o g i e s (ICEEICT),Trichirappalli, India,2023.
5. Yihui Deng, Dingzhao Li, Lvqing Yang, Jintao Tang, Jiangsheng Zhao, "Analysis and prediction of bank user churn based on ensemble learning algorithm",2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA), Shenyang, China,2021.
6. Jaehyun Ann, Junsik Hwang, Doyoung Kim, Hyukgeun Choi and Shinjin Kang, “A Survey on Churn Analysis in Various Business Domains”,2020 IEEE Access,2020.
7. Qi Tang, Guoen Xia, Xianquan Zhang and Feng Long, "A Customer Churn Prediction Model Based on XGBoost and MLP", 2020 International Conference on Computer Engineering and Application (ICCEA), Guangzhou, China,2020.
8. Sajad Fathi Hafshejani and Zahra Moaberfard, "A new trigonometric kernel function for support vector machine",2022 Iran Journal of Computer Science, Iran,2022.
9. Mindi Richia Putri, I Gede Pasek Suta Wijaya, Fritzie Primananda Adi Praja, Abdul Hadi and Faqih Hamami, "The Comparison Study of Regression Models (Multiple Linear Regression, Ridge, Lasso, Random Forest, and Polynomial Regression) for House Price Prediction in West Nusa Tenggara", 2023 International Conference on Advancement in Data Science, E-learning and Information System (ICADEIS), Bali, Indonesia,2023.
10. Sakthidevi, G.Vinoth Rajkumar, R.Sunitha, A. Sangeetha, R.Santhana Krishnan and S. Sundararajan, "Machine Learning Orchestration in Cloud Environments: Automating the Training and Deployment of Distributed Machine Learning AI Model", 2023 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Kirtipur, Nepal,2023.
11. Bobbinpreet Kaur and Ayush Dogra, "A Machine Learning Model for Predicting Employees

Retention: An Initiative towards HR through Machine", 2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan, Himachal Pradesh, India, 2022.

1. Efendi Efendi, Gabriel Michael Ivan Santosa, Christopher Jourdan and Anderes Gui, “Customer Relationship Management, Customer Retention, and the Mediating Role of Customer Satisfaction on a Healthcare Mobile Applications”, 2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE), Jakarta, Indonesia, 2023.
2. Weiyun Yin, “Prevention of Customer Churn using Random Forest model”, 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT), Australia, 2023.