INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 06 ISSUE: 05 | MAY - 2022 IMPACT FACTOR: 7.185 ISSN: 2582-3930

PREDICTION OF CUSTOMER CHURN USING ML

Harsh Vardhan, Ankit Kumar Singh, - Dr. R.Yamini.

[1,2]Student, Dept. of Computer Engineering, SRM Institute of Science And Technology Kattankulathur - Chengalpattu District, Tamil Nadu.

[3]Professor, Dept. of Computer Engineering, SRM Institute of Science And Technology, Kattankulathur - Chengalpattu District, Tamil Nadu.

Abstract-- No business can thrive without it's customers. On the flip side, customers leaving the business is a nightmare that every business owner dreads!In fact, one of the key metrics to measure a business' success is by measuring its customer churn rate - the lower the churn, the more loved the company is.Typically, every user of a product or a service is assigned a prediction value that estimates their state of churn at any given time. This value may be based on multiple factors such as the user's demographic, their browsing behavior and historical purchase data, among other details.This value factors in unique and proprietary predictions of how long a user will remain a customer and is updated every day for all users who have purchased at least one of the products/services. The values assigned are between 1 and 5

I. INTRODUCTION

Abundance of information available has also led to consumers facing a higher supply of subscription-based services. This can be viewed as a challenge for companies since retaining customers can potentially become more difficult. Digitalization within companies can lead to a decrease in labor costs, an increase in efficiency and a better overview of the company's operations within the organization . All of this is essential for staying competitive, and to gain an edge over other companies.

e available amount of data and information has increased significantly during the past years. This rapid growth has enabled storage and processing of great amounts of data while increasing the necessity of automatically finding valuable information and creating knowledge. With meaningful information extracted from the stored data, firms can make appropriate decisions in order to grow the business. With this growth, the use of data mining techniques and machine learning has increased, due to its ability of handling and analyzing great amounts of data . The digitalization has also brought forward an ongoing trend to improve current data processing activities as a part of customer relationship management (CRM) strategies. The idea of knowledge management and customer relationship management has lately obtained more attention in the subscription-based business model and the concepts focus on distribution of resources to activities that are customer-centric, to be able to increase competitive advantages . Customer knowledge is the knowledge that businesses can obtain through interaction with their customers . Customer relationship management systems are systems that support interaction between businesses and customers with the objective of collecting, storing and analyzing data to get an overview of their customers . Such systems have evolved through the past years and by using technology and different data analysis tools, enterprises can find patterns in customer's behavior, which would be almost impossible to discover manually [3]. Such patterns could vary from a customer's purchasing behavior or patterns related to customer churning. In a subscription-based business model, a fundamental part of success is to minimize the rate of customers ending their subscriptions, in other words, to minimize churn [3]

.

II. STATE OF THE ART

We conducted a literature study using two search engines: BTH summon and google scholar. We also used a database with a built-in search engine,. In order to filter the results and obtain a high reliability, we have mainly used books and peer-reviewed articles published in scientific journals and conferences. As our study aims to investigate customer churn prediction, we used search words such as customer churn prediction, customer churning, customer relationship management, churn management in subscription-based services, churn prediction in B2B, and customer churning in B2B.

Even though the researchers propose several churn approaches using variety of techniques and methodology, still very effective CHURN is not achieved till now. With this concern, a novel approach for churn application is proposed in this paper.

III. PROPOSED WORK

As mentioned above, we decided to use the machine learning process proposed by. The proposed machine learning process includes data collection and preparation, feature selection, algorithm choice, parameter and model selection, training, and evaluation. We do not consider the proposed machine learning process as a linear process; there is a lot of back and forth between the steps in order to create the most optimal model, in our case a churn prediction model.

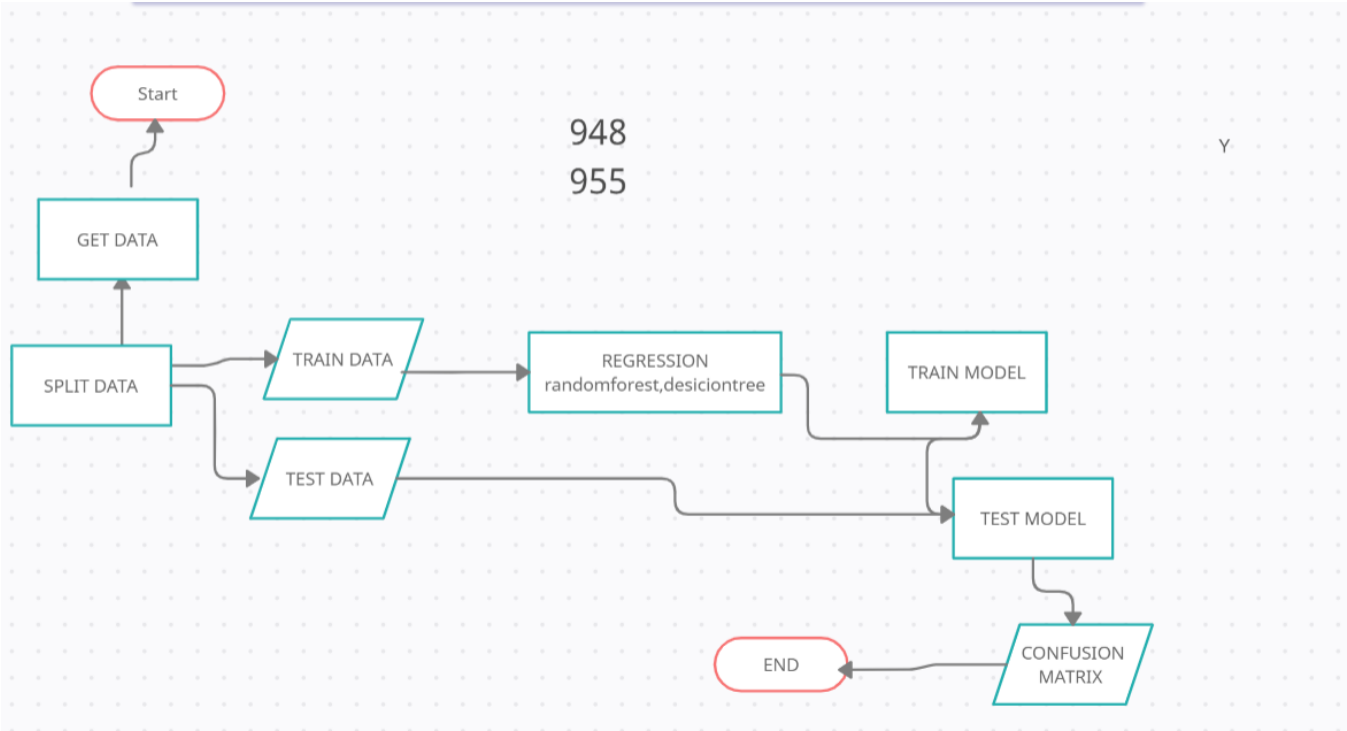

Fig.1. Architecture Diagram of the Proposed Work

A. *Data Pre-Processing Using Stop Words*

This study was performed using a dataset given by Fortnox AB, which is a company operating in the domain of B2B and provides a cloud-based platform for accounting and administration within finance. Fortnox AB provides subscription-based services in the form of licenses to different customers such as small businesses, associations, schools, and accounting firms [57]. All customer data is stored in raw data files and includes cross-sectional data of customers. Meaning that the dataset contains observations of customer specific data at a certain point. In order to use the data in our study, it was necessary to prepare the data. We aggregated all the raw data files into one dataset including information about 96129 customers, including 94487 non-churners and 1642 churners, see Figure 6. The aggregated dataset contains Customer ID, Customer information, and class labels, that is, churn/non-churn for each customer.

B. *Featuer Extraction Uisng Count Vectorizer*

The text data after removing the stop words are to be coverted into a matrix of tokens. For such vectorization, Count Vectorizer is employed to transform the data into format of vector. Count Vectorizer converts the input text data into the vectors depending on the frequency or count of each word that occurs in the whole text. Count vectorizer transforms the text document collection into a token count matrix. This function builds a sparse count matrix. The Count Vectorizer matrix is constructed in the following example. Assume there is a document that contains the following sentences. “This car is beautiful”, “This car is dirty”, and “This car is speedy”. The above sentences form a Count Vectorizer matrix of size 3*6 since there are three documents and six different features (“This”, “car”, “is”, “beautiful”, “dirty”, “speedy”). The vector matrix of the above example is tabulated in Table. 1.

TABLE 1: Example of Count Vectorizer matrix

Features / **Sentences**	**1**	**2**	**3**	**4**	**5**	**6**
1	1	1	1	1	0	0
2	1	1	1	0	1	0
3	1	1	1	0	0	1

In the above table, the presence of a feature is indicated by a '1', while the absence is indicated by a '0'. The first four features/words in "Sentence 1" are marked as '1', but "Feature 4" is marked as '0' in "Sentence 2" and "Sentence 3", and "Feature 5" and 'Feature 6" are marked as '1' in "Sentence 2" and "Sentence 3" respectively. Thus, the words in the text data are converted into vector format using Count Vectorizer. In the same procedure, the count vectorizer converts the data into matrix even for a large document.

C. *Algorithm Choice*

Based on the scope of the problem, the optimal algorithm for the considered problem can vary. There are numerous algorithms to choose from when faced with a problem, for instance prediction tasks. However, not every algorithm is suitable and fits the criteria for the task at hand, which is why several algorithms are rejected. Since the problem we are trying to solve is a classification problem, all algorithms used for regression problems are rejected. Common classifiers, such as Naïve Bayes, Random Forest, XGBoost, Support Vector Machine, Logistic Regression, and Hidden Markov's Model are used in prediction problems, however previous studies show that ensemble learners are preferred for classification problems and they are rapidly becoming the standard choice among algorithms due to their performance enhancing characteristics

IV. IMPLEMENTATION

We chose to implement our machine learning models in Python using different libraries for data analysis, such as NumPy, Pandas, and Scikit-learn [66]. Scikit-learn is a tool used for predictive data analysis in Python with built-in functions for implementation of algorithms, training, and evaluation.

V Results and Discussion

A. *Evaluation Metric*

Accuracy is evaluated to verify the efficacy of the proposed SA application. The rate of accuracy is defined as the ratio of the number of results that are truly predicted to the total number of results predicted by the proposed SA application. The higher percentage of accuracy indicates that the proposed SA application is good in sentiment analysis. The formula for calculating the accuracy is given below.

$$Accuracy = \frac{n_{correct}}{n_{total}} \quad (4)$$

Where, $n_{correct}$ indicates the number of correct results predicted and n_{total} indicates the total number of results predicted by the application.

B. *Predicted Results*

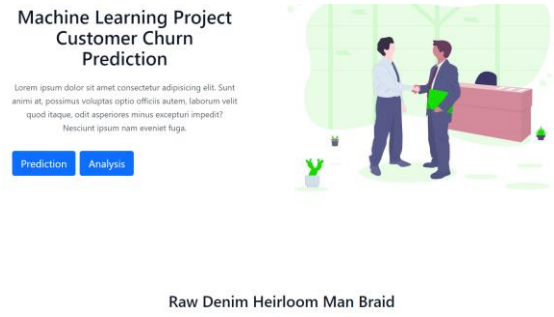

Fig. 3. Home Page

We have home page here which shows the what our project is about ..that is managing customer relationships with clients by using ml model

The Fig. 5 shows the 'Contact' page, which shows the location, contact information like, mail Id and phone number of the company or admin who use it. If the users have any queries related to the application, they are able to contact or clarify by sending the message. For this purpose, in the 'contact' page, the name, Mail ID, Subject and the Message box is given to describe the queries or doubts related to the application.

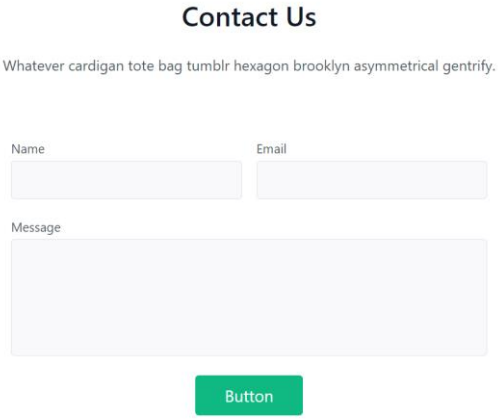

Fig. 5. Contact page

The Fig.6 shows the 'Registration' page, which asks to register the user when using this application for the first time. Once registered, the user has to login for visiting it next time. For registration, the details such as User name, Password and Password confirmation are to be filled and by clicking the Sign up button, the registration has done successfully.

Fig. 6. Registration Page

Finally, the page for predicting or testing the review is shown in Fig. 9. In this page, we fill in all the details of he customers and then the button, 'PREDICT' is clicked. Thus, the final prediction of the comment is predicted as 'Positive'.

V. CONCLUSION

We have studied different aspects related to customer churn prediction using machine learning. By studying these aspects, which are captured in the machine learning process, we contribute to building a complete understanding on how machine learning can be used for churn prediction. Hence, we provide an answer to our research question.

REFERENCES

1. L. Yang, Y. Li, J. Wang, and R. S. Sherratt, "Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning", in IEEE Access, vol. 08, pp. 23522-23530, 2020.
2. Cambria, and Erik, "Affective Computing and Sentiment Analysis", IEEE Intelligent Systems, vol. 31, no. 02, pp. 102–107, 2016.
3. Zhiying Ren, Guangping Zeng, Liu Chen, Qingchuan Zhang, Chunguang Zhang, and Dingqi Pan, "A Lexicon-Enhanced Attention Network for Aspect-Level Sentiment Analysis", IEEE Access, Vol: 08, DOI: 10.1109/ACCESS.2020.2995211, pp: 93464 – 93471,2020.
4. W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey", Ain Shams engineering journal, vol. 05, No: 04, pp. 1093-1113, 2014.
5. B. Zhang, X. Li, X. Xu, K. C. Leung, Z. Chen, and Y. Ye, "Knowledge Guided Capsule Attention Network for Aspect-Based Sentiment Analysis", in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2538-2551, Doi: 10.1109/TASLP.2020.3017093, 2020.
6. S. M. Mohammad, "Challenges in sentiment analysis. In A practical guide to sentiment analysis", Springer Cham, pp. 61-83, 2017.
7. F. A. Pozzi, E. Fersini, E. Messina, and B. Liu, "Challenges of sentiment analysis in social networks: an overview", Sentiment analysis in social networks, pp. 01-11, 2017.
8. D. H Farias, and P. Rosso, "Irony, sarcasm, and sentiment analysis In Sentiment Analysis in Social Networks", Morgan Kaufmann, pp. 113-128, 2017.
9. Y. Gao, M. Gong, Y. Xie, and A. K. Qin, "An Attention-Based Unsupervised Adversarial Model for Movie Review Spam Detection", in IEEE Transactions on Multimedia, vol. 23, pp. 784-796, Doi: 10.1109/TMM.2020.2990085, 2021.
10. F. Alattar, and K. Shaalan, "Using Artificial Intelligence to Understand What Causes Sentiment Changes on Social Media", in IEEE Access, vol. 09, pp. 61756-61767, Doi: 10.1109/ACCESS.2021.3073657, 2021.
11. Y. Fang, H. Tan, and J. Zhang, "Multi-Strategy Sentiment Analysis of Consumer Reviews Based on Semantic Fuzziness", in IEEE Access, vol. 06, pp. 20625-20631, Doi: 10.1109/ACCESS.2018.2820025, 2018.
12. T. Gu, G. Xu, and J. Luo, "Sentiment Analysis via Deep Multichannel Neural Networks with Variational Information Bottleneck", in IEEE Access, vol. 08, pp. 121014-121021, Doi: 10.1109/ACCESS.2020.3006569, 2020.
13. Y.Gao, J. Liu, P. Li, and D. Zhou, "CE-HEAT: An Aspect-Level Sentiment Classification Approach with Collaborative Extraction Hierarchical Attention Network", in IEEE Access, vol. 07, pp. 168548-168556, Doi: 10.1109/ACCESS.2019.2954590, 2019.
14. H. T. Phan, V. C. Tran, N. T. Nguyen, and D. Hwang, "Improving the Performance of Sentiment Analysis of Tweets Containing Fuzzy Sentiment Using the Feature Ensemble Model", in IEEE Access, vol. 08, pp. 14630-14641, Doi: 10.1109/ACCESS.2019.2963702, 2020.
15. Y. Wang, G. Huang, J. Li, H. Li, Y. Zhou, and H. Jiang, "Refined Global Word Embeddings Based on Sentiment Concept for Sentiment Analysis", in IEEE Access, vol. 09, pp. 37075-37085, Doi: 10.1109/ACCESS.2021.3062654, 2021.