IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CUSTOMER CHURN PREDICTION THROUGH EDA

Harshita Yajjala , Thota Thanu Sri , Nalla Deekshitha , Aravind kumar Mallavarapu

Department of Computer Science Engineering, GITAM Deemed to be University, Visakhapatnam, India-530045

I. ABSTRACT:

The "churn model" is a prediction method that determines which consumers are most likely to terminate their accounts. By applying a machine learning model to previous client data, a churn model may forecast possible new clients by identifying associations between qualities and goals. Organizations may detect and stop customer churn before it happens by taking proactive steps to keep clients. This includes focused reactivation tactics, customized client education, and more. Getting the right information, including product usage statistics and direct customer input, is the first stage in creating a churn model. Analyzing data trends will be the next stage in determining the main reasons for client attrition. Exploratory data analysis is required to find relevant data points and inform decisions going forward. By identifying the clients who are most likely to vanish, we can more efficiently focus our rescue efforts. For example, we may send these clients a marketing campaign informing them that they haven't bought from us in a while or maybe offering them a reward.

KEYWORD : churn analysis , machine learning. Exploratory data analysis , random forest regression.

II. INTRODUCTION:

Churn may be a term that alludes to the number of clients that an organization/company loses over a particular time period. CC is a calculable notice in benefit area with tall furious administrations. Anticipating clients who will take off the company early can be a huge income source. CC dataset is utilized to look at the promoting inclination of clients from the expansive databases. One way to think approximately client whittling down is as a churn rate, it is the rate of shoppers that cease employing a benefit inside a indicated time outline.

Subscriber-based benefit models, which have a legally binding client base, frequently utilize this metric to evaluate their monetary practicality. Broadcast communications could be a major industry in created countries. Specialized enhancements and more administrators boost competition. Employing a unused dataset, the machine learning (ML) model affect on churn will be tried. The commitments show CC expectation demonstrate is 1. Joining of different preprocessing methods together with SMOTE-ENN normalize the information 2. Applying distinctive classification methods to anticipate the reasonable demonstrate ".

Forecasting the loss of clients is a critical area of concern for businesses trying to develop a loyal client base and see long-term success. Understanding how customers behave and pinpointing the essential elements that give consumers their competitive edge need the application of data analysis, or EDA. Customer prediction is helped by the crucial information that EDA uncovers through data set analysis, including user usage, usage trends, and interaction history. Analysts may make well-informed judgments on selection and design by using EDA to display data distributions, identify deviations, and assess correlations between variables.

The primary goal of EDA is to make the agreement more widely known so that companies may employ cost-effective and customer-focused storage methods. We will use EDA to a dataset containing data on customer behavior, demographics, and interactions with the product or service in this project. The goal of EDA is to analyze data, identify critical characteristics, and ascertain how they affect customer attrition. This will enable us to choose important characteristics, preprocess the data, and develop a predictive model that will accurately predict client churn. EDA helps companies identify at-risk customers, comprehend the main causes of customer churn, and put proactive retention plans in place. Moreover, EDA makes it possible to find patterns and trends in client data, which may be utilized to enhance current offerings or develop brand-new ones that better meet consumer needs. All things considered, EDA is a strong tool that may assist businesses in lowering loss of talent, raising client satisfaction, and ultimately achieving long-term profitability.

III. LITERATURE SURVEY :

1. Customer churn prediction in telecommunication industry: A review of recent research"* - Authors: Al-Fahad, F. N., & Mohammed, A. S. - Published in: Journal of Telecommunications and Information Technology, 2019 - Summary: This review paper offers an extensive overview of recent studies on customer churn prediction within the telecommunication sector using machine learning methods. It delves into diverse methodologies, datasets, and evaluation metrics employed in these studies, highlighting both the challenges and opportunities in this field of research.

2. "Machine learning techniques for customer churn prediction: A systematic literature review" - Authors: Verma, A., Pandey, S., & Shukla, A. K. - Published in: Expert Systems with Applications, 2020 - Summary: This systematic literature review thoroughly explores the utilization of machine learning techniques for predicting customer churn across various industries. It synthesizes findings from numerous studies, categorizing the machine learning algorithms used for churn prediction and identifying critical factors affecting prediction accuracy, such as feature selection and data preprocessing techniques.

3. "A survey on customer churn prediction using machine learning techniques" - Authors: Haddadnia, J., & Sheikhtaheri, A. - Published in: 2019 4th International Conference on Web Research (ICWR) - Summary: This survey paper presents an in-depth overview of machine learning techniques applied to customer churn prediction in diverse industries, including telecommunications, banking, e-commerce, and insurance. It addresses challenges associated with churn prediction, such as data imbalance and feature selection, while also proposing future research directions in this area.

4. "Customer churn prediction using machine learning techniques: A systematic review and future research directions" - Client churn expectation utilizing machine learning strategies: A efficient survey and future inquire about bearings"* - Creators: * Gunasekaran, M., & Selvi, S. T. - *Distributed in: * 2020 5th Worldwide Conference on Communication and Hardware Frameworks (ICCES) - *Summary: This systematic review paper critically examines the application of machine learning techniques for customer churn prediction across various domains. It evaluates the strengths and limitations of different machine learning algorithms, discusses the impact of feature selection and data preprocessing on prediction accuracy, and suggests future research directions to enhance churn prediction models.

5. "Recent trends in customer churn prediction using machine learning techniques: A systematic literature review" - Authors: Jain, V., Kumar, A., & Kumar, S. - Published in: 2021 6th International Conference on Computing for Sustainable Global Development (INDIACom) - Summary: This literature review paper discusses recent trends in customer churn prediction utilizing machine learning techniques. It covers advancements in feature engineering, model selection, and evaluation metrics for churn prediction, as well as emerging technologies such as deep learning and ensemble methods within this domain. These literature surveys

provide valuable insights into the current state-of-the-art in customer churn prediction using machine learning, highlighting methodologies, challenges, and future research directions in this evolving field.

IV. SYSTEM METHODOLOGY

In arrange to avoid making less precise or inadmissible discoveries, it is basic that the information be made usable by dispensing with undesirable or invalid values. The information collection contains a expansive number of lost and wrong values. After looking at the entire dataset, we as it were included the foremost useful qualities. Way better exactness may emerge from a highlight list that as it were incorporates imperative components required to distinguish specific data, such as the proprietor, enrollment area, and address. Choosing the vital components from the information set based on information requires a basic step called highlight determination. There are a part of characteristics within the dataset that we utilized for this, but we as it were chosen the highlights that are fundamental for moving forward execution appraisal and making a difference us make choices. The other highlights will be less critical. When there are fair profitable and highly unsurprising factors within the dataset, classification execution progresses.

As a result, the execution of categorization is made strides by having fair vital characteristics and less insignificant qualities. Thus, constraining the sum of insignificant qualities and keeping fair imperative highlights makes strides classification execution. Within the telecom segment, a few strategies have been put out for foreseeing client whittling down. The likelihood of a churn, or the chance that a client would cancel their membership, may be anticipated here utilizing calculated relapse, random forest and KNN. Execution measures like exactness, exactness, and review score can be utilized to evaluate the models.

4.1. Data Collection :

In arrange to conduct this think about, we were able to get a telecom dataset that contained different customer-related information, such as socioeconomics, exchange histories, utilize designs, client back logs, and account data. There were around 7,043 sections with 20 characteristics in the test. To guarantee buyer protection, the information was recovered from a telecom company's framework and scrambled. Utilizing SQL inquiries, we were able to extricate the information from the database and spare it in a CSV record organize for afterward processing.

4.2. Information Cleaning and Pre-processing:-

Some time recently beginning the investigation, we performed information cleaning to expel any lost values, copy records, and exceptions. The cleaned dataset was utilized for exploratory information examination and building the client churn expectation models. Encode categorical factors: Change over categorical factors into numerical representations utilizing strategies like one-hot encoding or name encoding. By carefully following these pre -treatment processes, we can ensure that the data is clean, consistent, and appropriately structured for EDA, resulting in more accurate insights into customer turnover behaviour and more robust prediction models.

4.3. Exploratory Data Analysis :-

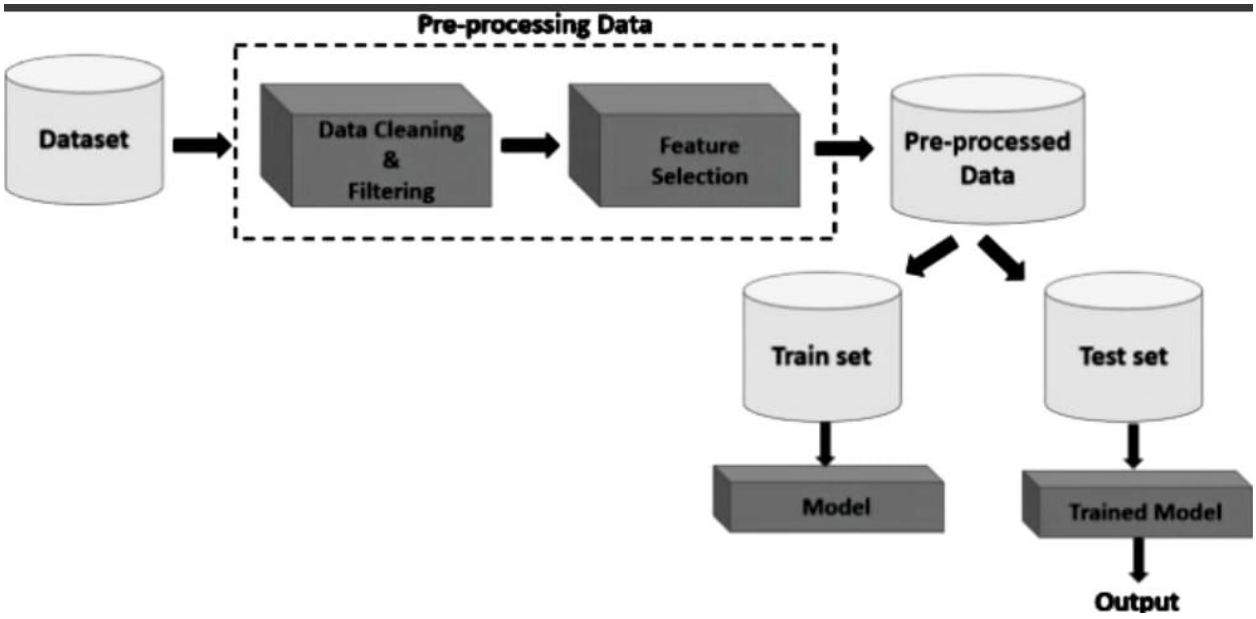
In the EDA stage, we investigated the information through different factual and visual strategies. We analyzed the dissemination of factors, checked for lost values, and recognized exceptions. We too performed highlight designing to extricate significant highlights and expelled insignificant highlights. Additionally, we analyzed the relationship between factors to maintain a strategic distance from multicollinearity issues. We made a few box plots to visualize the relationship between factors and their affect on the target variable. We recognized the most critical factors that influence client churn utilizing univariate and bivariate investigation. By and large, the EDA stage made a difference us to pick up experiences into the information and get it the relationship between factors, which is basic for building an compelling churn forecast demonstrate. Another critical component of EDA is data transformation, which entails changing data into a more acceptable format for analysis. Visualization is an effective technique in EDA, allowing analysts to visually analyse data and detect patterns or trends. Scatter plots, histograms, and box plots are common visualizations used in EDA to help understand variable distributions and interactions.

4.4. Feature selection :-

Including choice is an critical errand to discover the client is chur or not. Highlight determination can be done on two perspectives some time recently applying the classification calculations. Data's are plotted utilizing visualization methods and following is utilized to discover the esteem utilizing "lasso coefficient". Three highlights are (residency, month to month charges, add up to charges) chosen after applying the two diverse sorts of include determination strategies. Select the most important highlights that are likely to contribute to foreseeing churn. Strategies like relationship investigation, include significance from models, or space information can be accommodating. Furthermore, feature engineering may entail altering existing features to make them more suited for modeling, such as converting continuous variables to categorical ones or scaling features to a standard range.

4.5. Model selection :-

Choosing suitable machine learning calculations for churn forecast. Commonly utilized calculations incorporate calculated relapse, choice trees, irregular timberlands, back vector machines, or slope boosting strategies. We utilized a 5-fold cross-validation strategy to assess the execution of each show based on exactness, exactness, review, and F1 score. Based on the assessment comes about, we chosen the top-performing models for assist examination. By using insights from EDA to understand the dataset's characteristics and the underlying relationships between features, analysts can choose the best model or combination of models to accurately predict customer churn, assisting businesses in implementing targeted retention strategies and reducing customer attrition.



4.6. Model training :-

Utilize the preparing information to train the chosen demonstrate. while in this stage, the demonstrate learns the designs in the information that are related with churn. Use the preparing information to fit the chosen machine learning calculation to the designs in the information. Amid preparing, the show learns the connections between the input highlights (e.g., client graph , exchange history) and the target variable (churn). Training for models such as Random Forest and Decision Tree, as well as approaches such as SMOTEEN, entails fitting the selected model to the training data while guided by EDA insights. To test the trained model's performance, a validation set is used. Accuracy, precision, recall, F1 score, are some common churn prediction evaluation criteria.

4.7. Model assessment :-

In the model assessment step, we surveyed the execution of the prepared models utilizing different measurements. To begin with, we calculated the exactness score to decide the extent of accurately classified occasions out of all occasions. Be that as it may, exactness alone is not continuously the best metric for assessing classifier models, particularly when the dataset is imbalanced. Hence, we too calculated the F1-score, which considers both exactness and review, to assess the in general execution of the models.

4.8. Deployment :-

The model with the highest performance metrics on the test dataset is chosen as the final model for customer churn prediction. The chosen model should strike a compromise between accuracy and generalization, ensuring that it can accurately forecast customer attrition without overfitting to the training data. Deploy the trained model into production. This could involve integrating it into existing systems or applications where churn predictions are needed. Set up monitoring to track model performance over time and retrain periodically if necessary.

Advance bits of knowledge like electronic check medium, tech assistance , and online security can be looked an decided based on churning as well.For client churn forecast, a combination of exploratory information investigation (EDA), the SMOTE-ENN strategy for rectifying course imbalance, and the utilize of choice tree and irregular forest models created extraordinary results. Following broad information examination, pre-processing, and model preparing, the random forest and decision tree models come to a 94% prediction accuracy. The EDA stage given profitable experiences into the dataset, counting the dispersion of churned and non-churned customers, major qualities affecting churn, and likely relationships between factors. This data impacted the pre handling methods, which included dealing with lost information, encoding category factors, different categories for it to be and scaling numerical features.

The SMOTE-ENN approach was basic for adjusting the dataset, especially given the considerable class imbalance in client churn data. By oversampling the minority course (churned clients) and under examining the majority class (non-churned clients), SMOTE-ENN expanded the precision and capacity so as to construct models' ability to foresee churn. The pre-processed dataset was used to prepare choice tree and arbitrary woodland models, which were then assessed utilizing measures such as accuracy, precision, recall, and the F1-score. Both models performed well, with the random forest model imperceptibly outperforming the decision tree model due to its capacity to diminish overfitting and deal with noise in information.

V. REFERENCES

A. Journals/Articles

1.Abhishek and Ratnesh ,"*Predicting Customer Churn Prediction in Telecom Sector Using Various Machine Learning Techniques",* In the proceedings of 2017 International Conference on Advanced Computation and Telecommunication, **Bhopal, India, 2017.**

2.Abinash and Srinivasulu U ,"*Machine Learning techniques applied to prepaid subscribers: case study on the telecom industry of Morocco",*

In the proceedings of 2017 International Conference on Inventive Computing and Informatics , Coimbatore, India, **pp. 721-725, 2017.**

3.Trupti S. Gaikwad; Snehal A. Jadhav; Ruta R. Vaidya; Snehal H. Kulkarni. "*Machine learning amalgamation of Mathematics, Statistics and Electronics"*. International Research Journal on Advanced Science Hub, **2, 7, 2020, 100-108. doi: 10.47392/irjash.2020.72**

4. Ali, A. 2001.Macroeconomic variables as common pervasive risk factors and the empiricaIntent of the Arbitrage Pricing Theory. Journal of Empirical finance, 5(3): 221–240.

5**.** Basu, S. 1997. The Investment Performance of Common Stocks in Relation to their Price to Earnings Ratio: A Test of the Efficient Markets Hypothesis. Journal of Finance, 33(3): 663-682.

B. Books

Fighting Churn with Data: The science and strategy of customer retention **1st Edition**

by Carl S. Gold (Author),2022