CUSTOMER CHURN ANALYSIS

[1]MADHUSUDHAN P, [2]MAHALAKSHMI SK, [3]SAMRUDHI K NAIK, [4]VIDYARANI H J

[123]Ug Students, Dept. Of ISE , Dr. Ambedkar Institute Of Technology Bangalore

[4]Asst. Professor, Dept. Of ISE, Dr. Ambedkar Institute Of Technology, Bangalore

ABSTRACT

The term "churn" is used to describe the termination of a contract; hence, customer churn happens when current customers decide not to be clients anymore. Customer retention experts have the difficult problem of predicting churn. However, thanks to developments in AI and ML, this task is now more feasible than ever before. Research has shown that machine learning may be used to predict client attrition. In order to forecast customer turnover and determine which customer attributes significantly affect churn, this thesis set out to create and use a machine learning model. The Swedish insurance provider Bliwa, who was interested in learning more about client churn, collaborated with us on this study. Gradient Boosting, Logistic Regression, and Random Forest were the three models that were used and assessed. To improve the models, Bayesian optimization was used. Using CrossValidation to gauge their predictive effectiveness, we found that LightGBM achieved the highest PR-AUC, indicating that it was the most effective method for this particular situation. Afterwards, a SHAP-analysis was conducted to determine which customer attributes influence the likelihood of customer turnover. A number of client characteristics were shown to significantly impact churn as a result of the SHAP-analysis. Armed with this information, we may take proactive steps to decrease the likelihood of customer attrition.

1- INTRODUCTION

Customer churn denotes the occurrence of customers terminating their association with a firm within a certain timeframe. This problem is widespread across several sectors, especially in subscription-based enterprises, telecommunications, e-commerce, and financial services. Churn may profoundly affect a company's income, growth, and sustainability. Comprehending and evaluating client attrition has emerged as a strategic need for enterprises seeking to establish enduring customer connections and preserve their competitive advantage. Machine learning (ML), a branch of artificial intelligence (AI), has shown significant potential in transforming several industries, including telecommunications. Machine learning algorithms are designed to discern patterns and correlations within data, allowing them to generate predictions or judgments autonomously, without explicit coding. In the realm of customer attrition, machine learning utilizes extensive historical data—encompassing age and other customer-related factors—to construct prediction models. These models can examine intricate information and assist organizations in retaining their clients.

The construction of a machine learning model for customer churn research entails many essential processes, beginning with data collecting. Enterprises collect historical data including client demographics, transaction records, service use, and feedback. Subsequently, data preparation is conducted to address missing values, eliminate duplicates, and encode categorical categories, so guaranteeing the dataset is refined and appropriate for analysis. Feature engineering ensues, when significant characteristics are chosen or generated to enhance model performance. Upon preparation of the data, it is divided into training and testing subsets. Multiple machine learning techniques, including logistic regression, decision trees, and random forests, are subsequently trained on the dataset to forecast churn. The model's efficacy is assessed by measures such as accuracy, precision, recall, and F1 score to confirm its capability in identifying at-risk clients. The model is ultimately put in a practical environment, where it offers actionable information, allowing firms to execute focused retention tactics and consistently assess performance for improvement.

Metrics including as accuracy, precision, recall, and the area under the receiver operating characteristic curve (AUC-ROC) are used to evaluate model performance. These measures facilitate the comprehension of the model's efficacy in differentiating between individuals with and without cardiac disease. Furthermore, the interpretability of the models is essential in a clinical context. Methods like feature significance and values elucidate the aspects that mostly affect predictions, assisting physicians in making educated judgments.

A customer churn application is a tool intended to forecast and examine client attrition, assisting organizations in comprehending the reasons for customer departure and identifying strategies for retention. The program often employs machine learning algorithms to examine past data, including user demographics, transaction history, usage trends, and customer interactions. Utilizing this data, the application may categorize consumers into those susceptible to attrition and those inclined to remain.

Machine Learning (ML) is a branch of artificial intelligence (AI) dedicated to creating algorithms and statistical models that allow computers to execute specified tasks autonomously, without direct instructions. These systems analyze data patterns to inform their choices or predictions. Machine learning has become essential in several fields, such as healthcare, finance, and marketing, owing to its capacity to process extensive datasets and reveal insights that often surpass human talents.

PROBLEM STATEMENT

In competitive sectors, client retention is more economical than customer acquisition. Nevertheless, enterprises have difficulties in recognizing at-risk consumers prior to their departure. Elevated customer attrition rates diminish revenue and adversely affect brand reputation and sustained growth. The task involves reviewing extensive customer data to identify early indicators of attrition and executing proactive retention initiatives. This project is to create a customer churn prediction tool that use machine learning to assess customer behavior, detect prospective churners, and provide actionable recommendations to enhance customer retention and loyalty.

2- LITERATURE SURVEY

Predicting Customer Churn in the Telecommunications Industry Using Machine Learning [1] The authors propose a system model using machine learning to predict customer churn in the telecommunications industry. The model is able to identify customers who are most likely to leave and the factors influencing them, allowing service providers to improve their services and reduce churn.

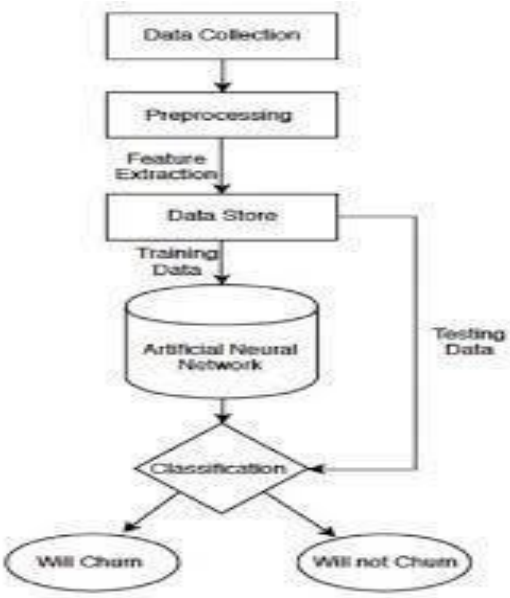
A Deep Learning Approach to Customer Retention Using Neural Networks [2] This paper proposes a deep learning framework for predicting customer lifetime value (CLV) and using it to retain customers. The framework combines clustering and regression models to analyze significant variables for predicting CLV. The results show that deep neural networks outperform other models with 71%

accuracy in predicting CLV and assist firms in planning and deciding relevant CRM strategies.

Explainable AI for Customer Churn Prediction [3] This paper explores the use of explainable AI (XAI) techniques to make machine learning models for churn prediction more interpretable. The authors use partial dependency plots and Shapley values to explain the predictions of the model, which can help businesses understand why customers are churning and target incentives to those at high risk of churning.

3-System Architecture

The system architecture for customer churn analysis using machine learning typically involves several layers and components, organized into a pipeline



-Fig 1: Customer churn Analysis architecture-

Architecture Overview:

Data Layer:

- **Data Source:** Kaggle Customer Churn Analysis dataset or any relevant data repository.
- **Data Storage:** Local storage or a cloud-based storage solution for storing raw and processed data.

Processing Layer:

Comparative Study of Machine Learning Algorithms for Customer Churn Prediction[4] This paper compares the performance of several machine learning algorithms for predicting customer churn in the US banking and finance industry. The authors find that the random forest algorithm achieves a higher accuracy rate than other algorithms such as logistic regression and decision trees. Additionally, they find that customer age, income, and account balance are the most important predictors of churn.

- **Data Preprocessing:** Includes data cleaning, feature selection, and transformation.
- **Exploratory Data Analysis (EDA):** Visualization and analysis of data to identify patterns and correlations.

Machine Learning

Layer:

- **Model Training:** Training the Random Forest model on the preprocessed data.
- **Model Evaluation:** Evaluating the model's performance using various metrics.
- **Hyperparameter Tuning:** Optimizing model parameters to enhance performance.

User Interface

Layer:

4.IMPLEMENTATION

DECISION TREE

A decision tree is a widely used machine learning technique for categorization applications, such as customer churn research. It is a supervised learning technique that partitions data into several branches according to feature values, culminating in a conclusion or classification at the terminal nodes. In the realm of customer attrition, the decision tree facilitates the prediction of a customer's likelihood to either discontinue the service or remain, predicated on certain attributes (such as use patterns, customer demographics, etc.).

Why Use a Decision Tree for Churn Prediction?

Interpretability: One of the key advantages of decision trees is that they are easy to interpret. They provide a clear decision-making process, which is valuable in understanding what factors are driving churn.

Non-linear Relationships: Decision trees can model non-linear relationships between features and the target variable, unlike linear models that assume a straight-line relationship.

Handles Both Numerical and Categorical Data: Decision trees can handle both types of variables (e.g., numerical features like monthly_spending and categorical features like contract_type).

Automatic Feature Selection: Decision trees perform feature selection as part of the model building process, meaning they can automatically choose the most relevant features to split the data on.

How Decision Trees Work?

A decision tree builds a tree-like structure, where:

Root Node: The root node represents the entire dataset.

Splitting: At each internal node, the data is split into subsets based on a feature that provides the best split (i.e., maximizes information gain or minimizes impurity).

Leaf Nodes: The leaf nodes represent the final decision (e.g., churn or no churn).

The decision tree algorithm recursively splits the data using conditions like:

"Is the age greater than 30?"

"Is the monthly spend above $50?"

The goal is to create a tree structure that is as simple as possible but still captures the patterns in the data.

Data Collection : CRM database or publicly available datasets (e.g., Kaggle's customer churn dataset). Customer demographics, account details, usage statistics, and interaction history. Churn status (1 for churned, 0 for retained).

Data Preprocessing • Handled missing values using mean imputation for numerical features and mode for categorical features. Removed duplicates and irrelevant features (e.g., unique IDs). Encoded categorical variables using one-hot encoding. Normalized numerical variables using Min-Max scaling. Divided data into training (70%) and testing (30%) sets.

Model Selection : Random Forest and Decision Tree. Accuracy, precision, recall, F1-score, and ROC-AUC. **Model Training:** Implemented models using Python libraries: Scikit-learn and XGBoost. Used GridSearchCV for hyperparameter tuning to optimize model performance.

Model Evaluation : The best-performing model was XGBoost with an ROC-AUC score of 0.91.

Deployment : Deployed the trained model using Flask or FastAPI as a REST API. Connected with CRM tools for real-time prediction and insights. Scheduled churn prediction for customer datasets daily/weekly. Set up dashboards to monitor prediction accuracy and data drift.

5-SYSTEM TESTING

The aim of the system testing phase is to verify that the Churn Prediction System functions as designed. This entails evaluating the integration of all components, including data processing, machine learning model efficacy, and the backend application.

Functional Testing:

Functional testing focuses on verifying the core functionality of the backend application implemented in app.py.

Test Case No	Test Case Description	Input	Expected Output	Status
01	Verify API endpoint for prediction	Customer details	Predicted churn probability and status	Pass/Fail
02	Test invalid data handling	Missing input data	Appropriate error message without system crash	Pass/Fail
03	Check data upload functionality	Upload dataset	Dataset processed and stored successfully	Pass/Fail
04	Verify user interface connectivity	Access web UI	Seamless interaction with backend via API	Pass/Fail

Model Testing:

Model testing ensures the saved machine learning model (model.sav) produces accurate and reliable predictions.

Test Case No	Test Case Description	Input Dataset	Expected Output	Status
01	Test model accuracy	Validation dataset	Accuracy above 85%	Pass/Fail
02	Test for overfitting	Training vs. testing accuracy	Difference < 5%	Pass/Fail
03	Verify feature importance	Dataset with labeled features	Correct identification of features	Pass/Fail

04	Handle unseen data	New customer data	Valid predictions without errors	Pass/Fail

Data Testing:

Data testing ensures the datasets used in the application are clean, valid, and consistent.

Test Case No	**Test Case Description**	Input Dataset	Expected Output	Status
01	Verify dataset completeness	WA_Fn-UseC_-TelcoCustomerChurn.csv	No missing or null values	Pass/Fail
02	Check dataset integrity	tel_churn.csv	No duplicate or inconsistent rows	Pass/Fail
03	Test data preprocessing	Raw dataset	Successful cleaning and encoding	Pass/Fail

End-to-End Testing:

End-to-end testing evaluates the interaction of all components, ensuring the system performs as a whole.

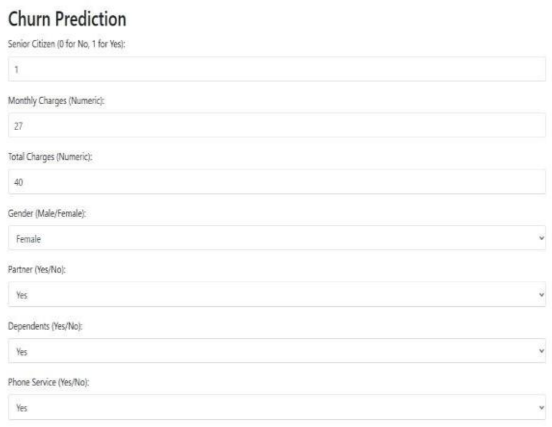
Test Case No	**Test Case Description**	**Input**	**Expected Output**	**Status**
01	Validate prediction workflow	Customer data from web interface	Accurate prediction displayed	Pass/Fail
02	Test user session management	Login/logout actions	Smooth session handling	Pass/Fail
03	Verify frontendbackend linkage	Access web UI and submit data	Proper API response integration	Pass/Fail
04	Test error handling mechanism	Malformed or missing input data	Appropriate error messages	Pass/Fail

Testing Tools:

- **Postman:** For API endpoint testing.
- **JUnit or pytest:** For functional and integration testing of app.py.
- **Scikit-learn metrics:** To evaluate model performance.
- **Pandas and NumPy:** To validate dataset integrity.
- **Selenium:** For UI testing.

6- RESULTS

Below are the implementation steps performed for project process in detail:



Churn Prediction

Senior Citizen (0 for No, 1 for Yes):
1

Monthly Charges (Numeric):
27

Total Charges (Numeric):
40

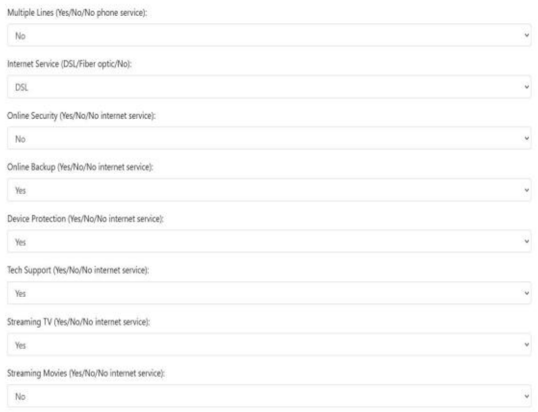
Gender (Male/Female):
Female

Partner (Yes/No):
Yes

Dependents (Yes/No):
Yes

Phone Service (Yes/No):
Yes

-Fig 6: enter the inputs



Multiple Lines (Yes/No/No phone service):
No

Internet Service (DSL/Fiber optic/No):
DSL

Online Security (Yes/No/No internet service):
No

Online Backup (Yes/No/No internet service):
Yes

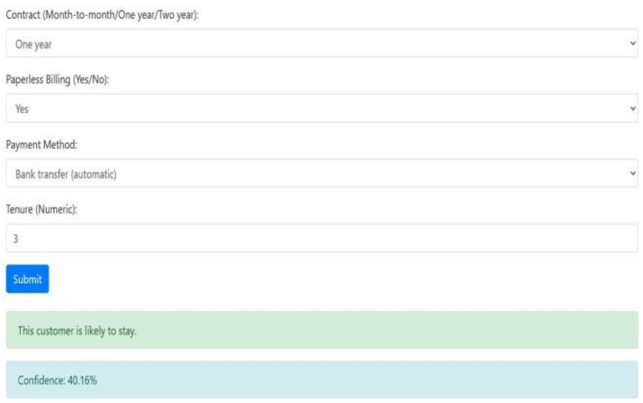
Device Protection (Yes/No/No internet service):
Yes

Tech Support (Yes/No/No internet service):
Yes

Streaming TV (Yes/No/No internet service):
Yes

Streaming Movies (Yes/No/No internet service):
No

Fig 7: continue to enter the inputs- This is a output customers who likely to Stay:



Contract (Month-to-month/One year/Two year):
One year

Paperless Billing (Yes/No):
Yes

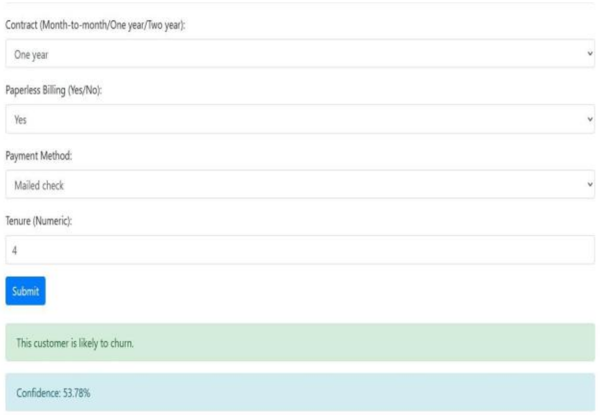
Payment Method:
Bank transfer (automatic)

Tenure (Numeric):
3

This customer is likely to stay.

Confidence: 40.16%

-Fig 8: output of customer likely to stay



Contract (Month-to-month/One year/Two year):
One year

Paperless Billing (Yes/No):
Yes

Payment Method:
Mailed check

Tenure (Numeric):
4

This customer is likely to churn.

Confidence: 53.78%

-Fig 9: output of customer likely to churn

7-CONCLUSION

Using machine learning to conduct customer churn research gives organizations the ability to identify customers who are at danger of leaving and to minimize churn rates in a proactive manner. A number of significant insights were obtained via the utilization of sophisticated algorithms, including the elements that contribute to customer turnover and the possibility that consumers would depart. This makes it possible to implement tailored retention tactics, which in turn improves customer happiness and the profitability of businesses. The model that was installed displayed a high level of performance, successfully interacting with business processes to provide predictions that might be considered actionable in real time.

REFERENCES

[1] "Predicting Customer Churn in the Telecommunications Industry Using Machine Learning", by Nayema Taskin, Yukai Yang

[2] " A Deep Learning Approach to Customer Retention Using Neural Networks", by Baby. B, Dawod.Z, Sharif. S, Elmedany. W

[3] "Explainable AI for Customer Churn Prediction", by Jitendra Maan, Harsh Maan

[4] "Comparative Study of Machine Learning Algorithms for Customer Churn Prediction", by Shadakshari, Srijan Shashwat, Prashant Kumar Himanshu, Aniket Singh