**Synopsis**

**on**

**Telecom Churn Prediction Using Machine Learning**

in partial fulfilment for the award of the degree of

**BACHELOR OF ENGINEERING**

IN

**Computer Science Engineering**

**(Data Science)**

**Submitted by:**

Tushar (24BDA70353)

Rajarshri (24BDA70351)

Ayush (24BDA70346)

Chirag (24BDA70332)

Aniket (24BDA70336)

**Under the Guidance of**

Mrs. Neeru

**INSTITUTE – UIE**

**ACADEMIC UNIT-I**

****

**Chandigarh University**

**January 2025**

**Index**

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Content** | **Page No.** |
| 1 | Introduction and Literature review | 1 |
| 2 | Problem definition and objectives |  |
| 3 | Scope |  |
| 4 | Planning and Task definition |  |

1. **Introduction and Literature Review :**

The telecommunication sector is competitive, with business entities always working to gain new customers while also keeping the old ones. Customer churn happens when customers stop availing themselves of a provider's service, either by moving over to a rival or by ceasing their subscription altogether. Churn has direct effects on an entity's revenue, customer acquisition expenses, and general business viability.

As the availability of big data and machine learning has increased, telecommunications organizations can now predict at-risk customers prior to their churn through the use of predictive analytics. This project targets building a high-accuracy model by using Random Forest and XGBoost for predicting whether the customers are "Churn" or "Not Churn", thereby enabling enterprises to take preventative measures like tailored offers, loyalty initiatives, and enhanced customer service.

This dataset for this project, (WA\_Fn-UseC\_-Telco-Customer-Churn.xlsx), holds 7,043 customer entries comprising 21 variables covering demographics, service, type of contract, method of payment, and churning. All these will be processed with Google Colab model deployment for ready availability without necessarily needing installation software on a user's system.

A number of research studies have investigated the application of machine learning algorithms for predicting customer churn in the telecommunication sector. Some of the key findings from previous studies are:

1. Machine Learning for Churn Prediction:

Decision Trees, Random Forest, and Gradient Boosting are popular classification models for predicting churn.

XGBoost has been shown to perform better than conventional models because it can deal with missing values and feature selection efficiently.

1. Feature Engineering for Churn Prediction:

Research indicates that contract type, monthly charges, tenure, and payment method are significant predictors of churn.

More sophisticated feature selection methods such as SHAP (SHapley Additive exPlanations) and Permutation Importance are frequently applied to identify the most critical factors.

1. Imbalanced Datasets Handling:

As churn datasets tend to be imbalanced (less churn cases compared to non-churn cases), methods such as SMOTE (Synthetic Minority Over-sampling Technique) are employed to enhance model performance.

1. Applications in the Real World:

Organizations such as AT&T, Verizon, and Vodafone leverage churn prediction models to minimize customer attrition and enhance engagement.

This project improves upon prior research by utilizing Random Forest and XGBoost for predicting customer churn from real-world telecom service data.

1. **Problem definition and objectives :**

* Customer churn is a serious problem in the telecom sector, where businesses lose a large number of customers due to reasons such as price, service quality, competition, and customer dissatisfaction. Churn has a direct impact on revenue, marketing expenses, and brand image, so it is crucial for telecom businesses to implement proactive measures.
* Why is Churn a Problem?

1. Revenue Loss

When clients exit, telecommunications firms lose continuous income from subscription fees. Acquiring new customers costs more than holding on to current ones.

1. Rising Customer Acquisition Expenses

Telecommunications firms spend lots of money on marketing, advertisement, and selling efforts to recruit new users.When customers continuously exit, firms need to spend more on new purchases, hence raising operational expenditures.

1. Brand Image and Client Trust

High churn rates indicate low-quality service, bad prices, or dissatisfaction. Leavers may post negative feedback, deterring potential users from signing up. Word-of-mouth is key to retaining and acquiring new customers.

1. Existing Industry Strategy: A Reactive Approach

Most telecom operators employ a reactive approach, i.e., they try to fix churn once the customer has already departed. This involves:

* Providing discounts or promotions to regain lost customers.
* Carrying out exit surveys to determine why users departed.
* Spending more on advertising campaigns to win back lost customers.

Create a machine learning model with Random Forest and XGBoost to forecast customer churn.

* Assess the primary factors driving churn, including contract type, monthly fees, and tenure.
* Conduct data preprocessing, such as missing value handling, categorical encoding, and feature scaling.
* Assess model performance based on accuracy, precision, recall, and F1-score.
* Solve class imbalance if necessary by SMOTE.
* Run the project in Google Colab for smooth execution.

1. **Scope :**

Application Scope:

* Telecom Sector: Assists in determining vulnerable customers and enhancing customer retention programs.
* Business Decision-Makers: Supports data-driven decisions for enhanced customer interaction and tailored offers.
* Data Science Students: Is a real-life case study on using machine learning for business analytics.

Technical Scope:

* Platform: Google Colab (No local setup required).
* Dataset: WA\_Fn-UseC\_-Telco-Customer-Churn.xlsx (Customer information, service information, and churn labels).
* Algorithms Used:
* Random Forest: A method of ensemble learning that enhances prediction accuracy.
* XGBoost: An enhanced gradient boosting model that is good at structured data.
* Data Processing Steps:
* Impute missing values.
* Transform categorical variables into numbers with One-Hot Encoding.
* Scale numerical features for improved model performance.
* Model Evaluation Metrics:
* Accuracy, Precision, Recall, F1-score, and Confusion Matrix.
* Final Output: Classify whether a customer is likely to churn or not based on his/her features.

1. **Planning and Task definition :**

* Phase 1: Data Collection & Preprocessing (Week 1-2):
* Load the dataset into Google Colab and check missing values.
* Convert TotalCharges to numeric and drop unnecessary columns (customerID).
* One-Hot Encoding for categorical features.
* Scale numerical features (tenure, MonthlyCharges, TotalCharges).
* Phase 2: Model Selection & Training (Week 3-4):
* Split data into 80% training & 20% testing.
* Train Random Forest and XGBoost models.
* Test models on accuracy, precision, recall, and F1-score.
* Phase 3: Model Optimization (Week 5-6):
* Perform hyperparameter tuning using GridSearchCV.
* Solve class imbalance with SMOTE (if necessary).
* Phase 4: Visualization & Reporting (Week 7):
* Create feature importance plots, confusion matrices, and ROC curves.
* Finalize a comprehensive project report (PDF) and presentation (PPT).
* Phase 5: Final Submission (Week 8):
* Finalize documentation and code and submit the project.

**Only Following Points need to cover in the synopsis:**

1. **Introduction and Literature review:**
2. **Problem definition and objectives:**
3. **Scope:**
4. **Planning and Task definition:**

**Important Information:**

* The headings must be in times new roman-14, bold
* Paragraphs under headings must be in Times new roman-12, Unbold and justified.
* Synopsis must be typed no hand written

**\*** The Student will propose a project based on their unique ideas depend upon the concepts studies during the lab sessions.

\* Each group has to present a synopsis of the project that they will submit towards the end of the semester.