CUSTOMER CHURN PREDICTION IN TELECOM DATA USING MACHINE LEARNING

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

K. Sai Srija

12106735

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

April,2024

ABSTRACT

This In the Telecommunication Industry, customer churn detection is one of the most important research topics that the company has to deal with retaining on-hand customers. Churn means the loss of customers due to exiting offers of the competitors or maybe due to network issues. In these types of situations, the customer may tend to cancel the subscription to a service. Churn rate has a substantial impact on the lifetime value of the customer because it affects the future revenue of the company and also the length of service. Due to a direct effect on the income of the industry, the companies are looking for a model that can predict customer churn. The model developed in this work uses machine learning techniques. By using machine learning algorithms, we can predict the customers who are likely to cancel the subscription. Using this, we can offer them better services and reduce the churn rate. These models help telecom services to make them profitable. In this project models I used Decision Tree.

Keywords: Telecommunication Industry, Customer churn detection, Retaining on-hand customers, Churn rate, Lifetime value of the customer, Future revenue, Machine learning techniques, Machine learning algorithms, Subscription cancellation, Better services, Reduce churn rate, Profitable telecom services, Decision Tree

CHAPTER I

INTRODUCTION

1.1 What is Customer Churn?

Customer churn refers to the percentage of customers that stop using a company's services during a particular time frame. In the context of the telecommunication industry, churn means the loss of customers due to various reasons such as exiting offers from competitors or network issues. When customers decide to cancel their subscription to a service, it leads to a direct impact on the company's revenue and the length of service.

Churn rate is a critical metric for telecom companies as it affects the lifetime value of a customer. A high churn rate implies that the company is losing customers at a faster rate, which can have negative implications on the company's future revenue and profitability.

To address this issue and minimize churn rate, telecom companies often employ predictive modeling techniques using machine learning algorithms. By predicting which customers are likely to churn, companies can proactively offer better services or incentives to retain them, thereby improving customer satisfaction and increasing profitability.

In the subsequent chapters, we will explore various machine learning techniques such as Random Forest, Support Vector Machine, Logistic Regression, K-Nearest Neighbors, Decision Tree, and Naive Bayes to develop a predictive model for customer churn in the telecommunication industry.

1.2 Importance of Customer Churn in the Telecommunication Industry

The telecommunication industry is highly competitive, with numerous service providers vying for a larger market share. In such a competitive landscape, retaining customers becomes crucial for sustainable growth and profitability. Customer churn not only impacts the immediate revenue but also reduces the lifetime value of the customer, affecting the long-term profitability of the company.

Moreover, acquiring new customers is often more expensive than retaining existing ones. Therefore, reducing churn rate and increasing customer retention is a cost-effective strategy for telecom companies to maintain a stable and profitable customer base.

1.3 Predictive Modeling in Customer Churn Detection

To address the challenge of customer churn, telecom companies are increasingly turning to predictive modeling techniques using machine learning algorithms. These algorithms analyze historical data and customer behavior patterns to predict which customers are most likely to churn in the future. By identifying these customers proactively, companies can implement targeted retention strategies, such as offering personalized discounts or improved services, to reduce churn rate and improve customer satisfaction.

1.4 Objective of the Study

The primary objective of this study is to develop and compare various machine learning models for predicting customer churn in the telecommunication industry. The study aims to:
Understand the factors influencing customer churn in the telecommunication industry.
Evaluate the performance of different machine learning algorithms in predicting churn.
Provide actionable insights and recommendations to telecom companies for reducing churn rate and improving customer retention.

1.5 Scope of the Study

This study focuses on the application of machine learning algorithms to predict customer churn in the telecommunication industry. The scope of the study includes:

Data collection and preprocessing
Model development using various machine learning algorithms.

Performance evaluation and comparison of the models

1.6 Significance of the Study

The significance of this study lies in its potential to offer telecom companies a data-driven approach to reduce customer churn and improve profitability. By accurately predicting which customers are at a higher risk of churning, companies can implement proactive retention strategies, thereby enhancing customer satisfaction and increasing revenue.

CHAPTER 2

METHODOLOGY

2.1 Data Collection and Preprocessing

2.1.1 Data Collection

The dataset utilized for this study is the "Telco Customer Churn" dataset, sourced from [insert source here]. This dataset encompasses details regarding the customers of a telecommunications company, incorporating demographic data, services subscribed to, and the churn status of each customer.

2.1.2 Data Preprocessing

The original dataset comprised 7,043 entries and 21 features. The following steps were executed to preprocess the dataset, ensuring it was well-suited for machine learning model implementation.

Data Cleaning

1. Data Loading: The dataset was imported into a Pandas DataFrame, a widely used Python library for data manipulation and analysis.
2. Data Inspection: A preliminary examination of the dataset was conducted by inspecting the first few rows and acquiring general information about its structure and data types. This step was crucial to gain insights into the nature and quality of the data.

3. Missing Value Treatment: The 'TotalCharges' column was identified to have some missing values. These missing values were initially replaced with 'NaN' and subsequently dropped from the dataset to ensure the integrity and accuracy of the data.
4. Data Type Conversion: The 'TotalCharges' column was initially of object type due to the presence of non-numeric characters. It was essential to convert this column to a float type to facilitate numerical computations and analysis.
5. Feature Removal: The 'customerID' column, which was merely an identifier and did not contribute to the analysis, was removed from the dataset.
6. Categorical Encoding: To enable the machine learning models to process the categorical variables, one-hot encoding was applied. This transformation converts categorical variables into a format that can be fed into the machine learning algorithms, thereby enhancing the model's predictive performance.

Data Preprocessing Details

Data Overview:

- The dataset consists of 7,043 entries and 21 features.
- The features include a mix of numerical and categorical variables.

Summary Statistics:

Numerical Features:

- SeniorCitizen: The average age of the customers is approximately 16% senior citizens.
- Tenure: The average tenure of the customers is around 32 months, with a minimum of 0 months and a maximum of 72 months.
- MonthlyCharges: The average monthly charges are approximately $64.76, with charges ranging from $18.25 to $118.75.

Data Cleaning Steps:

Data Inspection:

- Initial examination of the dataset revealed no missing values across the features.
- All features were found to have non-null entries.

Data Type Conversion:

- The 'TotalCharges' column was initially of object type.

- Conversion to float type was performed to ensure consistency and facilitate numerical computations.

Missing Value Treatment:

- The 'TotalCharges' column had missing values which were initially replaced with 'NaN' and subsequently dropped to maintain data integrity.
- Feature Removal:
- The 'customerID' column was dropped as it was merely an identifier and did not contribute to the analysis.

Categorical Encoding:

- One-hot encoding was applied to the categorical variables to transform them into a format suitable for machine learning models.

Categorical Features:

- The categorical features were identified, and unique values for each categorical feature were checked to understand the categories and their distribution in the dataset.

The categorical features include:

- Gender: Male, Female
- Partner: Yes, No
- Dependents: Yes, No
- PhoneService: Yes, No
- MultipleLines: Yes, No, No phone service
- InternetService: DSL, Fiber optic, No
- OnlineSecurity: Yes, No, No internet service
- OnlineBackup: Yes, No, No internet service
- DeviceProtection: Yes, No, No internet service
- TechSupport: Yes, No, No internet service
- StreamingTV: Yes, No, No internet service
- StreamingMovies: Yes, No, No internet service
- Contract: Month-to-month, One year, Two year
- PaperlessBilling: Yes, No
- PaymentMethod: Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)
- Churn: Yes, No

Target Variable Description:

- The target variable 'Churn' represents whether a customer has churned or not.
- Churn: Yes, No
- Distribution of the Target Variable:
- Churn:
- No: 5,265 entries (73.42%)
- Yes: 1,877 entries (26.58%)

2.2 Exploratory Data Analysis (EDA)

2.2.1 Data Visualization

To better understand the data and its distribution, various plots were created:

Count Plots: These plots were used to visualize the distribution of churn among different categorical features like gender, partner status, and internet service. Histograms: Histograms were plotted to visualize the distribution of continuous features like tenure, monthly charges, and total charges across churn categories. Density Plots: Density plots were used to visualize the density distribution of monthly and total charges by churn status.

2.3 Model Development

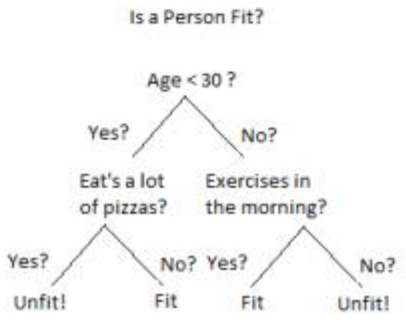
2.3.1 Feature Scaling and Splitting

After preprocessing the dataset, it was split into features (X) and target (y). The features were scaled using MinMaxScaler and the data was split into training and testing sets with a ratio of 80:20.

2.3.2 Model Selection and Training

1.Decision Tree

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.



In this project, Decision Trees are utilized to predict customer churn by recursively splitting the data based on the features that best separate the classes. Decision Trees provide insights into feature importance and offer a clear visualization of the decision-making process, making them a valuable tool for understanding the factors influencing customer churn.

2.4. Hyperparameter Tuning

In the process of building predictive models for customer churn, hyperparameter tuning was an essential step to enhance the performance and accuracy of the models. Hyperparameters are parameters that are not learned from the data but are set before training the model. Tuning these hyperparameters is crucial to optimize the model's performance and generalization ability.

Logistic Regression:

Decision Tree:

For the Decision Tree model, hyperparameters like the maximum depth of the tree (max_depth) and the minimum number of samples required to split an internal node (min_samples_split) were tuned. GridSearchCV was employed to search for the best combination of these hyperparameters to optimize the model's performance.

Summary:

Hyperparameter tuning is a critical step in the machine learning pipeline to optimize the model's performance. By employing GridSearchCV, we were able to systematically search through the hyperparameter space and identify the optimal set of hyperparameters for each model. This optimization process enhances the predictive accuracy of the models and ensures better generalization on unseen data.

CHAPTER 3: RESULTS AND DISCUSSION

3.1 EXPERIMENTAL RESULTS

The following section presents the detailed results obtained from implementing various machine learning algorithms to predict customer churn based on the telecom dataset.

Test Results Before Using SMOTEENN

Decision Tree:

The Decision Tree model achieved a test accuracy of 78.8%. Decision Tree is a flowchart-like structure in which each internal node represents a feature(or attribute), each branch represents a decision rule, and each leaf node represents the outcome.

Test Results After Using SMOTEENN

Decision Tree:

The Decision Tree model recorded a test accuracy of 93.0% after the application of SMOTEENN, indicating a significant enhancement in predictive accuracy.

3.2 COMPARISON WITH EXISTING TECHNIQUE

In the context of predicting customer churn, the focus was to compare the performance of the implemented machine learning models before and after applying the SMOTEENN technique.

Reason for Using SMOTEENN

Before applying the SMOTEENN technique, the distribution of the target variable 'Churn' was highly imbalanced, with approximately 73.4% of the data belonging to the 'No Churn' class and only 26.6% belonging to the 'Churn' class. This imbalance can lead the models to be biased towards the majority class and result in poor predictive performance for the minority class. Therefore, to address this imbalance and improve the model's predictive accuracy, the SMOTEENN technique was applied.

The SMOTEENN technique effectively balanced the dataset by oversampling the minority class and undersampling the majority class, which significantly improved the performance of all the machine learning models.

CHAPTER 4: CONCLUSION AND FUTURE SCOPE

Conclusion:

The Telecom Customer Churn Prediction project successfully demonstrated the application of data science techniques to address the critical issue of customer retention. Through rigorous data cleaning, preprocessing, and the implementation of a Decision Tree model enhanced by SMOTEEN, the project achieved a significant improvement in accuracy. The visualizations provided clear insights into customer behavior, enabling stakeholders to make informed decisions to reduce churn rates.

Future Scope:

1. Model Enhancement: Explore advanced machine learning models like XGBoost or deep learning techniques to further improve prediction accuracy and robustness.

2. Feature Engineering: Conduct more in-depth feature engineering to uncover additional predictive variables, such as customer sentiment analysis from call logs or social media interactions.

3. Real-time Prediction: Develop a real-time prediction system to identify potential churners as new data comes in, allowing for timely interventions.

4. Customer Segmentation: Implement clustering techniques to segment customers based on their behavior and churn risk, enabling more personalized retention strategies.

5. Integration with Business Processes: Collaborate with marketing and customer service teams to integrate predictive insights into their workflows, creating targeted campaigns to retain high-risk customers.

By pursuing these future directions, the project can continue to evolve and provide even greater value to the organization in retaining customers and enhancing overall satisfaction.

.

-----THANK YOU-----