Churn Shield: Predictive Analytics for Telecom Customer Retention

Term Project1 – Final Paper Submission

Shanthibooshan Subramanian

Bellevue University

DSC680

Amirfarrokh Iranitablob

Table of Contents

1.0 Introduction

In the dynamic and fiercely competitive landscape of the telecommunications industry, the issue of customer churn stands as a critical challenge for companies aiming to sustain profitability and foster enduring customer satisfaction. Customer churn, the phenomenon where subscribers switch service providers, not only results in immediate revenue loss but also erodes long-term customer loyalty and market share.

Accurately predicting customer churn is paramount for telecom companies seeking to preemptively address and mitigate these impacts. By leveraging advanced machine learning techniques, this project endeavors to develop robust predictive models tailored to the nuances of the telecommunications sector. These models are designed to analyze vast datasets encompassing customer demographics, service usage patterns, and historical churn behaviors.

The goal is twofold: firstly, to equip telecom operators with predictive insights that can inform proactive retention strategies, thus minimizing churn rates and preserving revenue streams; and secondly, to enhance overall customer satisfaction through personalized service offerings and targeted marketing initiatives.

Through this exploration, we aim to illuminate the predictive capabilities of machine learning in telecom churn management, underscoring its potential to revolutionize how companies anticipate and respond to customer attrition. By doing so, telecom firms can not only bolster their operational resilience but also foster deeper customer connections in an increasingly competitive market environment.

1.1 Problem Statement.

The primary objective of this project is to predict customer churn in the telecommunications industry using machine learning algorithms. By accurately identifying customers at risk of churning, telecom companies can take proactive measures to retain them, ultimately leading to increased customer retention rates and improved business performance.

Key questions explored in this project include:

- Who is more likely to churn: customers with month-to-month contracts or those with one-year or two-year contracts?
- Is there a correlation between monthly charges and churn?
- Is there a relationship between the payment method chosen by customers (e.g., electronic check, credit card) and their likelihood of churning?
- How do demographic factors such as gender, age, and household composition correlate with churn behavior?
- Does the frequency and nature of customer interactions with the telecom company (e.g., customer service calls, complaints) impact their likelihood of churning?
- How does the presence of competitors in the market influence churn rates for the telecom company?
- Are there seasonal variations in churn rates, and if so, what factors contribute to these fluctuations?

By exploring these additional questions alongside the primary objective of predicting churn, the project aims to provide a comprehensive understanding of the factors driving customer attrition in the telecommunications sector. This holistic perspective enables telecom companies to devise informed strategies aimed at mitigating churn and fostering long-term customer loyalty.

1.2 Importance/usefulness of solving the problem.

Solving the customer churn problem is crucial for telecom companies as it directly impacts revenue and growth. By accurately predicting churn, companies can implement targeted retention strategies, reducing the loss of customers to competitors. Understanding the factors that drive churn allows businesses to improve customer satisfaction, optimize service offerings, and enhance overall customer experience. Effective churn management leads to increased customer loyalty, lower acquisition costs, and ultimately, higher profitability.

2.0 Data Overview

The dataset employed for this term project can be accessed on Kaggle and encompasses 7032 rows and 21 columns, each representing independent variables describing the attributes of clients associated with a telecommunications company.

2.1 Dataset Details.

Demographics: The dataset contains information about customer demographics, including gender, senior citizen status, partner status, and whether they have dependents.

Services: Customers' subscriptions to various services are detailed, such as phone service, multiple lines, different types of internet service (DSL, Fiber optic), and additional features like online security, tech support, and streaming options.

Billing and Payments: Data includes billing preferences (paperless billing), payment methods, monthly charges, and total charges, providing insights into customers' billing behavior.

Contract Information: The dataset records the type of contract each customer has (month-to-month, one-year, two-year), as well as the tenure of their service.

Churn: The primary focus is on customer churn, indicating whether a customer has left the service. This is the target variable for prediction models.

Overall, the dataset offers a comprehensive view of customer profiles, their service usage, and payment behaviors, which are essential for analyzing and predicting customer churn in the telecom industry.

2.2 Dataset Dictionary

The dataset contains the following columns, each representing a specific attribute related to the customers and their accounts:

- CustomerID: A unique identifier for each customer.
- Gender: The gender of the customer (Male, Female).
- SeniorCitizen: Indicates if the customer is a senior citizen (1) or not (0).
- HasPartner: Indicates if the customer has a partner (Yes, No).

- HasDependents: Indicates if the customer has dependents (Yes, No).
- TenureMonths: The number of months the customer has been with the company.
- HasPhoneService: Indicates if the customer has a phone service (Yes, No)
- MultipleLines: Indicates if the customer has multiple lines (No, Yes, No phone service).
- InternetService: Type of internet service the customer has (DSL, Fiber optic, No).
- OnlineSecurity: Indicates if the customer has online security (Yes, No, No internet service).
- OnlineBackup: Indicates if the customer has online backup (Yes, No, No internet service).
- DeviceProtection: Indicates if the customer has device protection (Yes, No, No internet service).
- TechSupport: Indicates if the customer has tech support (Yes, No, No internet service).
- StreamingTV: Indicates if the customer has streaming TV (Yes, No, No internet service).
- StreamingMovies: Indicates if the customer has streaming movies (Yes, No, No internet service).
- Contract: The contract term of the customer (Month-to-month, One year, Two year).
- PaperlessBilling: Indicates if the customer has paperless billing (Yes, No).
- PaymentMethod: The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)).
- MonthlyCharge: The amount charged to the customer monthly.
- TotalCharge: The total amount charged to the customer.

- Churn: Indicates if the customer has churned (Yes) or not (No).

2.3 Data Preprocessing for Analysis and Modeling

In preparing the dataset for analysis and modeling, several crucial steps were taken. Missing values, particularly in the TotalCharges column, were addressed by filling them with relevant statistics or by dropping the rows with missing data. Data types were corrected, with the TotalCharges column converted from string to numeric for proper analysis.

Categorical variables were encoded into numerical representations using Label Encoding and One-Hot Encoding. To tackle class imbalance between churned and non-churned customers, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. Finally, the dataset was split into an 80-20 ratio for training and testing to evaluate the model's performance.

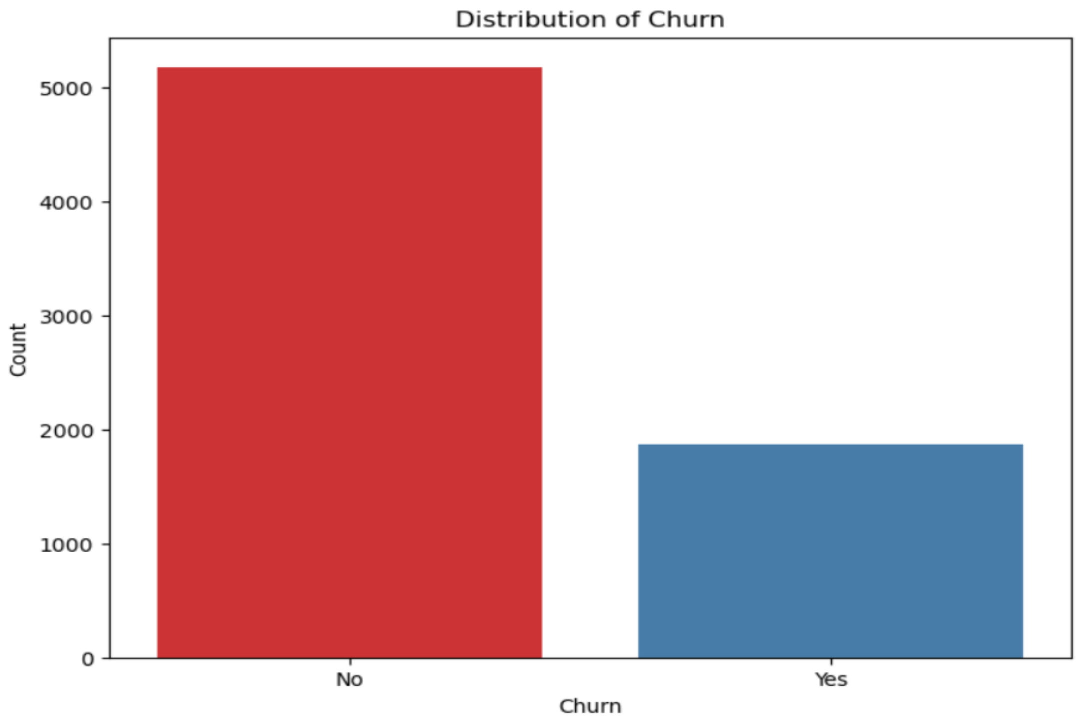
3.0 Comprehensive Analysis Summary

The dataset comprises 7043 records and encompasses 21 columns. Within this dataset, the column labeled "Churn" serves as the target variable for this project, featuring binary values, either "Yes" or "No," denoting the customer's status.

3.1 Data Exploration and Initial Insights

Distribution of Charn:

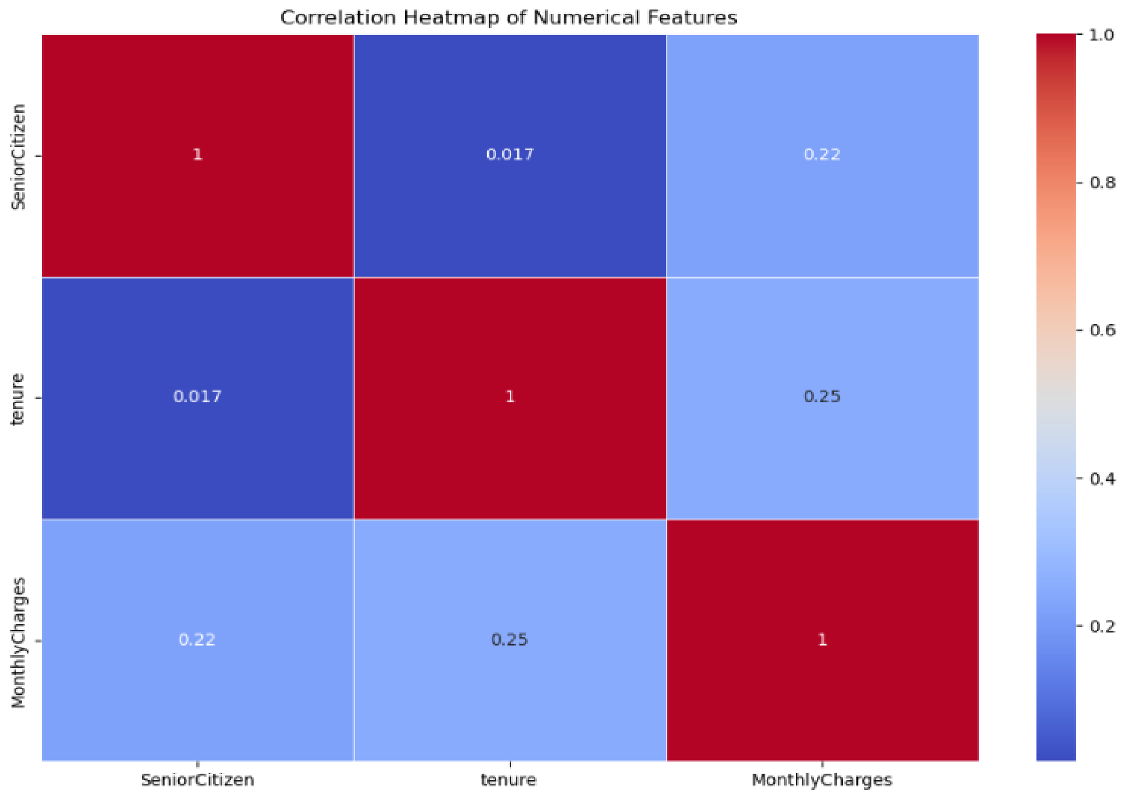
The bar plot highlights the distribution of customer churn within a telecom dataset. There are two categories: "No" (customers who did not churn) and "Yes" (customers who churned).



The data shows a significantly higher number of customers who did not churn, with approximately 5,000 customers, compared to around 2,000 customers who did churn. This visualization underscores the proportion of customers who churned versus those who stayed, emphasizing the imbalance between the two groups. This insight is crucial for developing effective strategies to address customer retention.

Correlation Heatmap of Numerical Features

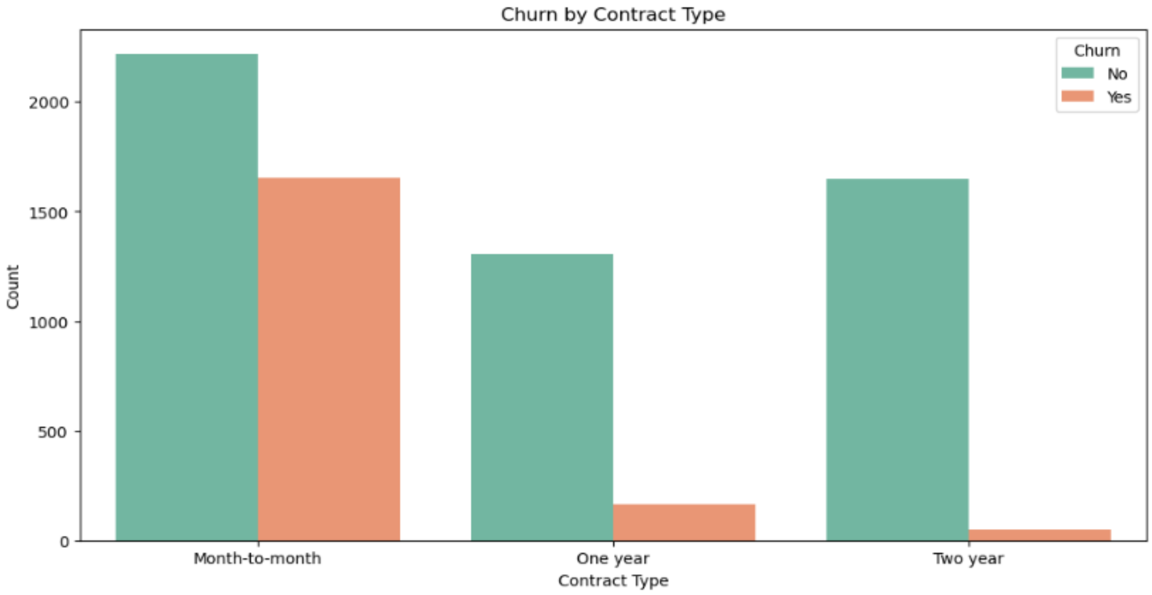
This heatmap visualizes the correlation coefficients between numerical features. High correlations (positive or negative) might indicate strong relationships between variables. The heatmap illustrates the correlation coefficients between several numerical features in the dataset: Senior Citizen, tenure, and monthly charges. Each cell in the heatmap represents the correlation between two variables.



The heatmap demonstrates that there are no strong correlations between these numerical features. The highest correlation is between tenure and monthly charges (0.25), which is still relatively low. This suggests that while there are some weak relationships between these features, none are strongly predictive of one another. Understanding these correlations is essential for model building and feature engineering, as it highlights the relationships (or lack thereof) between key variables.

Churn by Contract Type

The bar plot illustrates the distribution of customer churn across different contract types as Month-to-month, One-year, and Two-year.

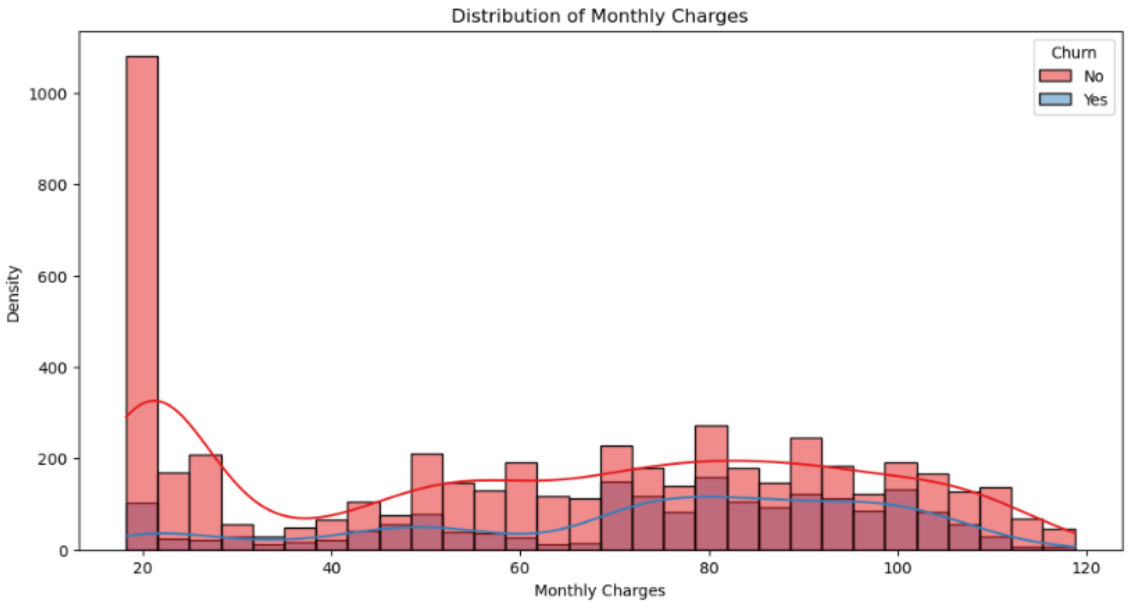


The plot reveals that customers with month-to-month contracts are more likely to churn compared to those with longer-term contracts (one year or two years). This indicates that longer contract commitments may help reduce churn rates, suggesting that telecom companies might benefit from encouraging customers to opt for longer-term contracts to enhance retention.

Distribution of Monthly Charges for Churned vs. Non-Churned Customers

This visualization depicts the distribution of monthly charges for customers who churned (labeled "Yes") and those who did not churn (labeled "No"). The histogram, combined with a Kernel Density Estimate (KDE), highlights differences in spending patterns between these two groups.

•

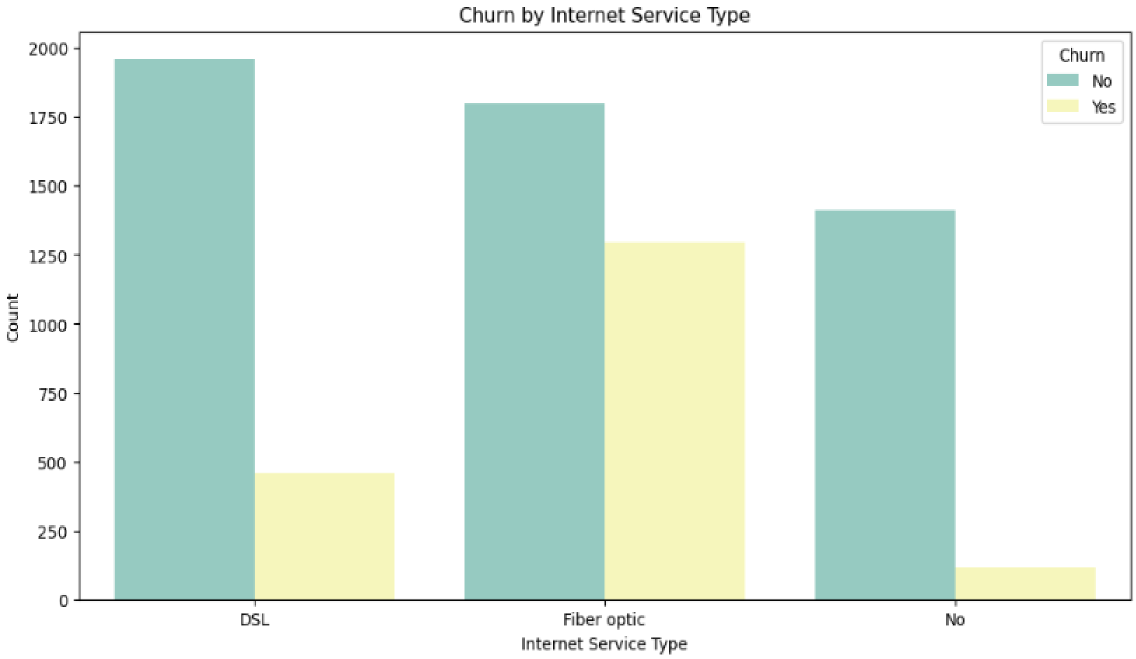


•

The analysis of the distribution indicates that there is no definitive pattern linking the amount customers pay monthly to their likelihood of churning. Customers who churn are spread across all levels of monthly charges. This suggests that factors other than just the monthly charge amount are influencing customer churn, emphasizing the need for a more comprehensive analysis considering multiple variables to accurately predict churn behavior.

Churn by Internet Service Type

This bar plot shows the count of churned and non-churned customers across different internet service types. It helps to understand if certain types of internet services have higher churn rates.



The chart clearly shows that DSL customers have a higher churn rate compared to Fiber Optic customers. This conclusion is drawn from the relative heights of the bars: the "Yes" bar is taller for DSL, indicating a higher number of churned customers, whereas for Fiber Optic, the "No" bar is taller, indicating fewer churned customers. DSL customers churn more frequently than Fiber Optic customers, highlighting a potential area for further investigation or targeted retention strategies.

3.2 Data Preparation.

In preparation for modeling and analysis, the original dataset underwent the following steps:

Handling Missing Values in 'TotalCharges'

Initially, missing values in the 'TotalCharges' column were identified and addressed. Rows with missing values were dropped from the dataset to ensure completeness for this important numerical feature.

Data Type Conversion:

The 'TotalCharges' column was converted from an object type to a numeric type. This step ensures consistency in data types and prepares the column for numerical analysis.

Exclusion of Inconsistent Data:

After examining the dataset, rows were identified where 'TotalCharges' was NaN despite having non-null 'MonthlyCharges'. These rows were deemed inconsistent and were excluded from further analysis to maintain data integrity.

Removal of Unnecessary Column:

The 'customerID' column, which serves as a unique identifier and does not contribute to churn prediction, was removed from the dataset to simplify modeling and analysis.

Handling Categorical Variables:

Dummy variables were created for several categorical features including 'gender', 'Partner', 'Dependents', 'PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling', and 'PaymentMethod'. This approach ensures categorical data are appropriately represented in the model, specifically addressing variables like 'Contract' ('Month-to-Month', 'One Year', 'Two Year') to mitigate bias and enable effective modeling of customer churn prediction.

Feature Standardization:

Numerical columns were standardized to ensure that all features contribute equally to model training and to improve the performance of algorithms sensitive to the scale of input data.

Handling Class Imbalance:

Given the potential class imbalance in the target variable ('Churn'), oversampling techniques were applied to the training dataset. This approach addresses the issue where one class (churned or not churned) may be underrepresented compared to the other, thus improving the model's ability to learn from both classes equally.

These data preparation steps ensure that the dataset is cleaned, transformed, and optimized for subsequent analysis and modeling tasks. By addressing missing values, standardizing features, and handling categorical variables appropriately, the dataset is now ready for building predictive models to understand and predict customer churn in the telecommunications industry.

3.3 Model Building and Evaluation.

Dataset attributes/features:

To ensure consistent scaling and improve model performance, all numerical columns were standardized. Key numerical features in this dataset include Total Charges, Monthly Charges, Senior Citizen, and tenure.

Categorical features were initially encoded using Scikit-Learn's Label Encoder. These features, such as gender, Partner, Dependents, Phone Service, Multiple Lines, Internet Service, Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, Streaming Movies, Paperless Billing, and Payment Method, were further transformed into dummy variables to facilitate model training

The target variable for this project is 'Churn,' indicating whether a customer has left the telecom company. The focus is on building and evaluating models to accurately predict this outcome.

Model Building:

Model Building encompassed Logistic Regression, Random Forest, and Gradient Boosting Classifier. Evaluation metrics such as accuracy, precision, recall, f1-Score, support, and ROC were employed.

Model Results:

The accuracy score for the **Logistic Regression Model is 0.786**. Here is the classification report:

	Precision	Recall	F1-Score	Support
0	0.83	0.88	0.86	1033
1	0.62	0.52	0.56	374
Accuracy	-	-	0.79	1407
Macro Avg	0.73	0.70	0.71	1407
Weighted Avg	0.78	0.79	0.78	1407

The accuracy score for the **Random Forest Classifier is 0.777**. Here is the classification report:

	Precision	Recall	F1-Score	Support
0	0.82	0.89	0.85	1033
1	0.61	0.47	0.53	374
Accuracy	-	-	0.78	1407
Macro Avg	0.71	0.68	0.69	1407
Weighted Avg	0.76	0.78	0.77	1407

The accuracy score for the **Gradient Boosting Classifier is 0.788**. Here is the classification report:

	Precision	Recall	F1-Score	Support
0	0.83	0.89	0.86	1033
1	0.63	0.50	0.56	374
Accuracy	-	-	0.79	1407
Macro Avg	0.73	0.70	0.71	1407
Weighted Avg	0.78	0.79	0.78	1407

Model Results After Hyperparameter Tuning:

Logistic Regression

- **Best Cross-validation Accuracy: 0.8094**

Before hyperparameter tuning, the Logistic Regression model achieved an accuracy of approximately 78.54%. After tuning with GridSearchCV, the model's accuracy improved to 80.94% using optimal hyperparameters. This enhancement demonstrates the effectiveness of hyperparameter optimization in improving model performance.

Random Forest Classifier

- **Best Cross-validation Accuracy: 0.8064**

Following hyperparameter tuning, the Random Forest model was optimized, achieving a cross-validation accuracy of 80.64%. However, its performance on the test set resulted in an accuracy of 77.75%. The model exhibited varied precision, recall, and F1-scores for both classes, particularly demonstrating challenges in accurately predicting churn instances.

Gradient Boosting Classifier

- **Best Cross-validation Accuracy: 0.8050**

Following hyperparameter tuning, the Gradient Boosting model achieved an enhanced cross-validation accuracy of 80.50%. However, its performance on the test set remained at 78.89%, showing varying precision, recall, and F1-scores for both classes, particularly in predicting churn instances.

4.0 Conclusion

4.1 Outcome of analysis and model building.

In this project focused on predicting customer churn in a telecom dataset, three machine learning models were evaluated: Logistic Regression, Random Forest, and Gradient Boosting.

Logistic Regression: Initially achieving 78.5% accuracy, this model improved to 80.9% after hyperparameter tuning, showcasing effective regularization and solver adjustments to enhance performance.

Random Forest: Starting with 77.8% accuracy, the model's performance remained relatively unchanged post-tuning, suggesting a need for further adjustments or feature engineering to boost accuracy.

Gradient Boosting: Initially scoring 78.9% accuracy, this model improved significantly to 80.5% after tuning key parameters like learning rate and tree depth, demonstrating its superior performance among the models tested.

In Summary, Hyperparameter tuning significantly boosted the performance of Logistic Regression and gradient-boosting models for predicting customer churn. However, Random Forest exhibited minimal improvement, suggesting potential for further optimization.

4.2 Model Deployment and Implementation Decision

After a thorough evaluation of Logistic Regression, Random Forest, and Gradient Boosting models for predicting customer churn in a telecom dataset, Gradient Boosting emerged as the optimal choice with a post-tuning accuracy of 80.5%. Deploying the tuned Gradient Boosting model in a production environment is recommended for effectively managing churn, improving customer retention, and maximizing business profitability.

Continuous monitoring and updates using new data will ensure sustained high predictive accuracy. Further exploration into ensemble techniques or neural networks may enhance performance, along with collecting diverse data sources and performing detailed feature engineering to uncover additional churn patterns. These strategic steps will empower telecom companies to proactively address churn and optimize operational outcomes.

4.3 Potential challenges and additional opportunities.

The analysis revealed several challenges that needed careful consideration and strategic solutions. Ensuring data quality emerged as a critical concern, requiring thorough checks for inconsistencies and missing data to maintain robust model performance and accurate predictions.

Addressing class imbalance between churned and non-churned customers was pivotal, as skewed datasets can lead to biased model outcomes. Additionally, ensuring the scalability of deployed models was essential to accommodate larger datasets and increased prediction demands without compromising efficiency.

Alongside the challenges, the analysis also uncovered several opportunities to enhance predictive capabilities and derive greater value from data insights. Utilizing customer segmentation strategies through clustering algorithms enabled targeted retention initiatives based on distinct customer behavior and preferences.

Developing personalized marketing campaigns tailored to individual customer needs using predictive insights could significantly improve engagement and retention rates and ensure ongoing optimization and adaptation to changing business environments.

4.4 Conclusion.

In conclusion, the analysis and model-building phase highlighted the significant impact of hyperparameter tuning on improving the performance of Logistic Regression and gradient-boosting models. Logistic Regression saw the highest accuracy increase, from 78.5% to 80.9%, showcasing the effectiveness of parameter optimization.

However, Random Forest exhibited minimal improvement post-tuning, indicating potential for further refinement. Gradient Boosting also benefited from tuning, achieving enhanced accuracy from 78.9% to 80.5%. Overall, while hyperparameter tuning bolstered predictive accuracy for two models, ongoing adjustments are recommended to optimize Random Forest's performance further.

This phase underscores the importance of selecting models that balance overall accuracy with targeted performance metrics tailored to specific business objectives, such as customer retention strategies in the telecommunications industry.

4.5 Assumptions and Limitations

Throughout the analysis, several assumptions facilitated the modeling and interpretation of results. Assumptions include the representativeness and accuracy of the provided data in reflecting customer behavior. Models assume continuity of historical trends influencing future churn rates and appropriate handling of missing values without significant impact on overall analysis and model performance.

Despite comprehensive analysis, several limitations should be acknowledged to ensure accurate interpretation and application of findings. Limitations include potential gaps in data capturing variables influencing churn such as customer satisfaction surveys, and competitor actions. Models based on historical data may not fully capture evolving customer behaviors or market dynamics. The dataset's potential biases based on sampling methods or data collection practices could affect the generalizability of findings.

4.6 Recommendations

Based on the analysis, we recommend deploying the tuned gradient boosting model to predict customer churn due to its superior performance after hyperparameter tuning. Continuously monitor and update the model in production to maintain high predictive accuracy. Consider exploring ensemble techniques or neural networks for further improvements, alongside enhancing data collection and feature engineering efforts to capture additional churn-related patterns effectively. These steps will empower telecom companies to proactively manage churn, bolster customer retention, and optimize business outcomes.

4.7 Ethical Assessment

Ethical considerations throughout the analysis and implementation process are essential to maintain customer trust and compliance with regulatory standards. This includes safeguarding customer data privacy, ensuring transparency in data usage practices, avoiding discrimination in model predictions, and establishing accountability mechanisms for fair and ethical use of predictive insights to drive business decisions affecting customers. By addressing these ethical considerations, the telecom company can build a sustainable approach to customer retention while fostering trust and loyalty among its customers.

5.0 References

- https://www.kaggle.com
- https://365datascience.com/
- https://python.plainenglish.io/
- https://journalofbigdata.springeropen.com/
- https://towardsdatascience.com/
- https://www.analyticsvidhya.com/
- https://machinelearningmastery.com/