Predicting Customer Churn

Telecommunications Company

(Project 01 – Milestone 03)

Bellevue University
DSC680 – Applied Data Science

Submitted By:
Debabrata Mishra

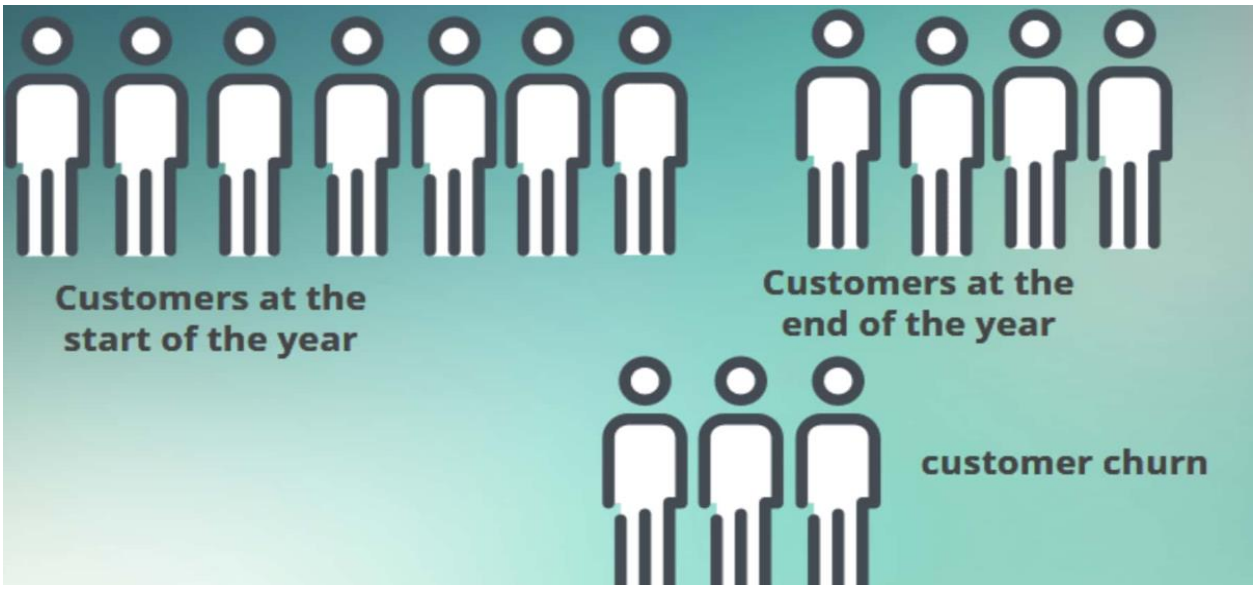
Instructor:
Amirfarrokh Iranitalab

Table of Contents

1.0 Introduction

1.1 Business Process Overview

Customer churn refers to the scenario where a client discontinues their association with a particular entity. There are various reasons prompting users to cease using a company's product or service, including affordability concerns, dissatisfaction with offerings, and subpar customer service. Typically, customers who churn from one company often transition to a competitor. For instance, if dissatisfied with the sluggish Internet speed provided by their current mobile service provider, individuals are inclined to switch to an alternative service. The process of churning seldom occurs abruptly; instead, it unfolds gradually. When encountering issues like low network bandwidth, customers might endure it for a month or two. During this period, they may reach out to customer support, check their network speed, or voice their discontent on social media platforms.



(Figure 1: Customer Churn)

Customer churn prediction stands as one of the pivotal applications of data science in marketing. Companies face substantial costs when users churn, given the expense associated with replacing an existing customer. Consequently, a majority of mid to large-sized organizations employ some form of churn prediction mechanism.

In the telecom industry, customers have a plethora of service providers to choose from and frequently switch between them. With an annual churn rate ranging from 15 to 25 percent in this fiercely competitive market, retaining individualized customer relationships poses a challenge. The sheer volume of customers renders personalized attention impractical for most companies, considering the costs would outweigh the additional revenue.

However, if a company could anticipate which customers are more likely to depart in advance, it could strategically channel its customer retention efforts towards these "high-risk" clients. The goal remains to expand coverage and regain customer loyalty. The crux of success in this market lies within understanding and catering to customer needs. Customer churn serves as a crucial metric, emphasizing that retaining existing customers is notably more cost-effective than acquiring new ones.

Telecom companies aim to curtail customer churn by accurately predicting which customers are prone to leaving their services.

Detecting early indications of potential churn involves obtaining a comprehensive understanding of customers and their engagements across various touchpoints. This encompasses store or branch visits, purchase histories, interactions with customer service, online transactions, and social media engagement, among other channels.

By proactively addressing churn, these companies not only safeguard their market standing but also foster growth and prosperity. A larger customer base translates to reduced initiation costs and amplified profits. Hence, the primary focus for the company's success lies in minimizing client attrition and implementing robust retention strategies.

1.2 Business Problem

The main goal of this project is to predict customer churn in the telecommunication industry using machine learning algorithms. By accurately identifying customers who are at risk of leaving, telecom companies can proactively implement strategies to retain them, leading to higher customer retention rates and better overall business performance.
Key questions addressed 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 examining these questions alongside the primary objective of predicting churn, this project aims to provide a comprehensive understanding of the factors driving customer attrition in the telecommunications sector. This holistic approach enables telecom companies to develop informed strategies to reduce churn and build long-term customer loyalty.

1.3 Importance/usefulness of solving the problem

Telecom companies grapple with an ongoing challenge: customer churn. It's not solely about customer loss; it equates to revenue depletion, reduced market share, and missed opportunities. High churn rates directly impact profitability, diminish customer lifetime value, and tarnish brand perception.

In the competitive market landscape, telecom churn prediction emerges as a critical pursuit. It transcends mere customer retention; it's about safeguarding revenue streams, nurturing brand allegiance, and fostering sustained growth.

The essence lies in predictive modeling, particularly churn prediction, as the linchpin for proactive customer retention. Leveraging historical data and sophisticated analytics enables the anticipation of potential churners before their departure. This strategic foresight allows tailored interventions, customized offerings, and the optimization of customer retention strategies.

2.0 Dataset

2.1 Dataset 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 fictitious telecommunications company.

Within this dataset, the "Churn" column serves as the response variable, signifying whether a customer terminated their subscription within the preceding month. The dataset is divided into two classes: "No" encompasses clients who retained their services last month, while "Yes" includes those who opted to discontinue their affiliation with the company. The primary aim of this analysis is to discern the connection between customer characteristics and churn rates.

Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges Demographic info about customers – gender, age range, and if they have partners and dependents.

In summary the dataset includes below category of attributes/features:

- Customers who left – the column is called Churn.
- Serviced customers opted for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies.
- Account Information - how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges.
- Customer Demographics – gender, age range, and if they have partners and dependents.

2.2 Data Dictionary

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

- Customer ID: A unique identifier for each customer.
- Gender: The gender of the customer (Male, Female).
- Senior Citizen: Indicates if the customer is a senior citizen (1) or not (0).
- Has Partner: Indicates if the customer has a partner (Yes, No).
- Has Dependents: Indicates if the customer has dependents (Yes, No).
- Tenure Months: The number of months the customer has been with the company.
- Has Phone Service: Indicates if the customer has a phone service (Yes, No)
- Multiple Lines: Indicates if the customer has multiple lines (No, Yes, No phone service).
- Internet Service: Type of internet service the customer has (DSL, Fiber optic, No).
- Online Security: Indicates if the customer has online security (Yes, No, No internet service).
- Online Backup: Indicates if the customer has online backup (Yes, No, No internet service).
- Device Protection: Indicates if the customer has device protection (Yes, No, No internet service).
- Tech Support: Indicates if the customer has tech support (Yes, No, No internet service).
- Streaming TV: Indicates if the customer has streaming TV (Yes, No, No internet service).
- Streaming Movies: 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).
- Paperless Billing: Indicates if the customer has paperless billing (Yes, No).
- Payment Method: The customer's payment method (Electronic check, Mailed check,
- Bank transfer (automatic), Credit card (automatic).
- Monthly Charge: The amount charged to the customer monthly.
- Total Charge: The total amount charged to the customer.
- Churn: Indicates if the customer has churned (Yes) or not (No).

2.2 Data Preprocessing for Analysis and Modeling

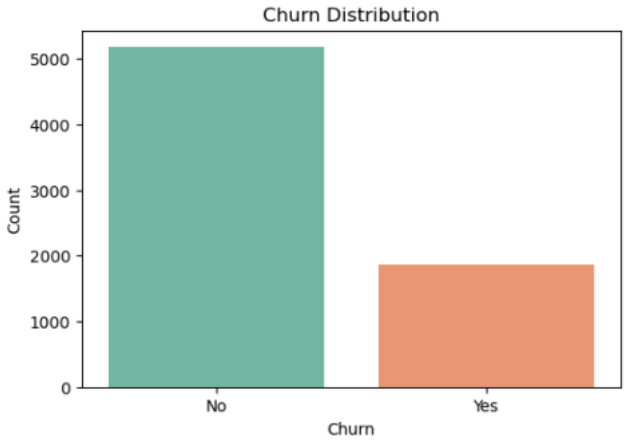
To prepare the dataset for analysis and modeling, several key steps were undertaken. Missing values, especially in the Total Charges column, were handled by either imputing them with relevant statistics or removing rows with missing data. Data types were adjusted, converting the Total Charges column from a string to a numeric format for accurate analysis. Categorical variables were transformed into numerical values using Label Encoding and One-Hot Encoding. To address class imbalance between churned and non-churned customers, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. Finally, the dataset was divided into an 80-20 ratio for training and testing to assess the model's performance.

3.0 Comprehensive Analysis Summary

3.1 Data Exploration and Initial Insights

Churn Distribution:

This bar plot visualizes the distribution of churned vs. non-churned customers in the dataset. The "Churn" variable is binary, with "Yes" indicating churned customers and "No" indicating non-churned customers.

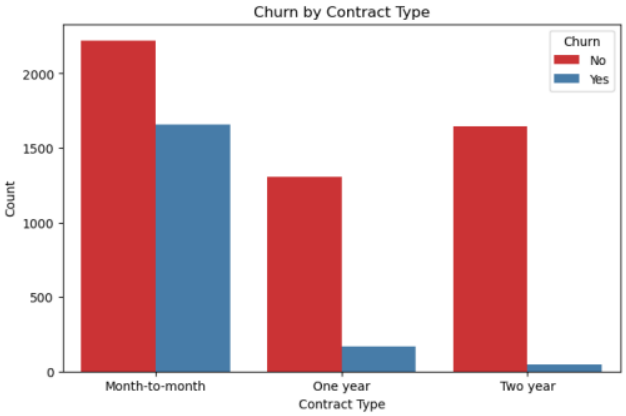


(Figure 2: Churn distribution)

The graph shows that the dataset has an imbalanced distribution of churned and non-churned customers, with a higher count of non-churned customers (represented by "No"). Understanding this imbalance is important because it may affect the performance of machine learning models. In cases of imbalanced data, model accuracy alone can be misleading, and other metrics like precision and recall become more critical for evaluation.

Churn by Contract Type:

This count plot illustrates how churn varies based on different contract types ("Month-to-month"," One year," and "Two year").

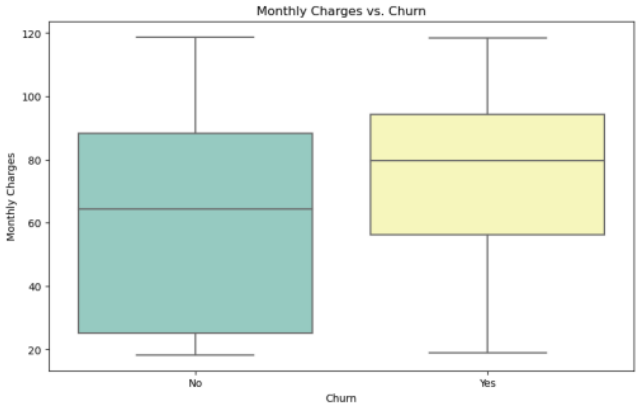


(Figure 3: Churn by contract type)

Customers with "Month-to-month" contracts have a higher likelihood of churning compared to those with longer-term contracts. "Two years" contract customers have the lowest churn rate, indicating that longer contract durations may lead to higher customer retention. This graph highlights the potential impact of contract type on customer churn, which can be valuable information for decision makers in the telecommunications company

Monthly Charges vs. Churn:

This box plot compares the distribution of monthly charges for churned and non-churned customers.

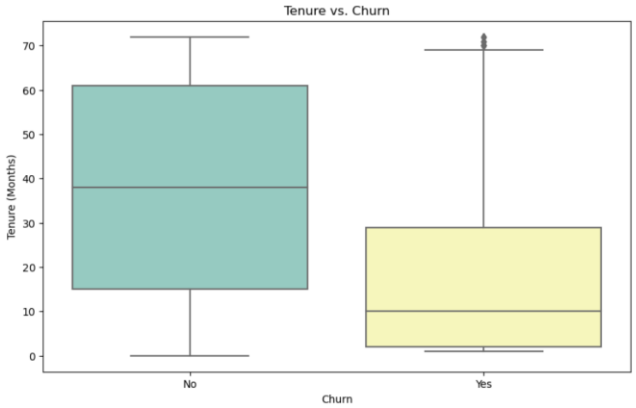


Churned customers tend to have higher median monthly charges compared to non-churned customers. The interquartile range (IQR) for churn customers is also wider, indicating a broader range of monthly charges among those who churn. This suggests that customers with higher monthly charges are more likely to churn, which is a critical finding for the company's pricing and retention strategies.

(Figure 4: Monthly Charges vs Churn)

Tenure vs. Churn:

This box plot displays the distribution of customer tenure (in months) for churned and non-churned customers.

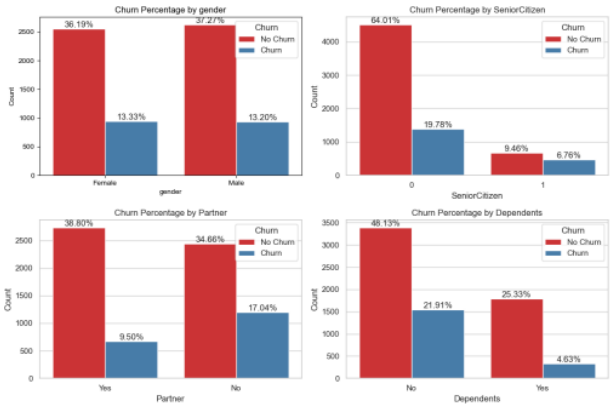


Churned customers generally have shorter tenures (lower median) compared to non-churned customers. Non-churned customers tend to have longer-lasting relationships with the company. Shorter tenure appears to be associated with a higher likelihood of churn, which emphasizes the importance of retaining customers during their early stages with the company.

(Figure 4: Tenure vs Churn)

Churn percentile w.r.t Demographics:

This visualizes how customer demographics ('gender', 'Senior Citizen', 'Partner', 'Dependents']) are associated with churn.



The distribution of churn based on gender is very similar or nearly identical. The distribution of churn based on Senior Citizen is very similar or nearly identical..

(Figure 5: Churn percentile w.r.t demographics)

3.2 Data Preparation

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

- Transformed “Total Charges” feature from “object” to numerical datatype.
- “Total Charges” has 11 missing values. The same 11 rows also have 0 value for “tenure” column even though “Monthly Charges” is not null for these entries, this information seems contradictory. Therefore, I have chosen to exclude these observations from the dataset.
- The “Customer ID” column does not provide any valuable information for predicting whether a customer will churn. Consequently, we have opted to remove this column from the dataset.
- Dummy variables were generated for the categorical variable 'Contract' to prevent bias within the model. This process safeguards against the algorithm assigning inappropriate weightage to numeric values when dealing with categorical data. The 'Contract' column originally encompassed three categories: 'Month-to-Month,' 'One Year,' and 'Two Year’.
- Employing Scikit-Learn's label encoder, transformed all categorical variables into a numerical format.
- To optimize model performance and mitigate potential issues associated with feature scaling, standardization was applied to all numerical columns.
- Oversampling technique implemented on the training dataset to address the class imbalance.

3.3 Model Building and Evaluation

Dataset attributes/features:

Numerical Features:

To optimize model performance and mitigate potential issues associated with feature scaling, standardization was applied to all numerical columns. These features are likely to include columns such Total Charges, Monthly Charges, Senior Citizen, and tenure.

Categorical Features:

The categorical features are encoded using the Scikit-Learn's label encoder. These features include 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 columns.

Target Variable:

The target variable is 'Churn", which represents whether a customer left or not.

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 0.742**. Here is the classification report:

	Precision	Recall	F1-Score	Support
0	0.89	0.74	0.81	1033
1	0.51	0.76	0.61	374
Accuracy	0.74	-	-	1407
Macro Avg	0.70	0.75	0.71	1407
Weighted Avg	0.79	0.74	0.75	1407

The accuracy score for the **Random Forest Classifier is 0.763**. Here is the classification report:

	Precision	Recall	F1-Score	Support
0	0.85	0.82	0.84	1033
1	0.55	0.61	0.58	374

	Precision	Recall	F1-Score	Support
Accuracy	0.76	-	-	1407
Macro Avg	0.70	0.72	0.71	1407
Weighted Avg	0.77	0.76	0.77	1407

The accuracy score for the **Gradient Boosting Classifier is 0.743**. Here is the classification report:

	Precision	Recall	F1-Score	Support
0	0.89	0.75	0.81	1033
1	0.51	0.73	0.60	374
Accuracy	0.74	-	-	1407
Macro Avg	0.70	0.74	0.71	1407
Weighted Avg	0.79	0.74	0.75	1407

This report provides insights into the precision, recall, and F1-Score for both classes (0 and 1) along with the overall accuracy metrics for the model.

4.0 Conclusion

4.1 Outcome of analysis and model building

Model Outcome Analysis:
The model analysis across Logistic Regression, Random Forest, and Gradient Boosting classifiers reveals varying performance levels. The Random Forest Classifier exhibits the highest accuracy at 0.763, showcasing a stronger predictive ability than the Gradient Boosting and Logistic Regression models. However, while the Random Forest model achieves higher accuracy, it's notable that its precision, recall, and F1-Score for class 1 (churned customers) are lower compared to the Logistic Regression model.
Despite the Random Forest's higher overall accuracy, the Gradient Boosting Classifier demonstrates competitive results, especially in identifying churned customers (class 1), with higher recall and F1-Score. Conversely, the Logistic Regression model offers a balanced precision-recall trade-off for both classes but at a slightly lower overall accuracy.

In summary, while the Random Forest model boasts the highest accuracy, the Gradient Boosting model showcases stronger recall for detecting churned customers. The Logistic Regression model, although with a marginally lower accuracy, presents a balanced performance across precision and recall for both classes. The choice of the most suitable model depends on the specific emphasis placed on accurately identifying churned customers or achieving a balanced predictive performance overall.

Model Performance:

This performance analysis suggests that while the Random Forest model achieves the highest accuracy, the Gradient Boosting model excels in recall for identifying churned customers. The Logistic Regression model showcases a balanced performance across both precision and recall. Selecting the most suitable model should consider specific priorities, such as accurately identifying churned customers or achieving a balanced predictive performance overall.

4.2 Model Deployment and Implementation Decision

Deployment Decision:

Model Selection: Assess the trade-offs between models based on specific business needs:

- Random Forest: Higher overall accuracy but lower performance in identifying churned customers (class 1).
- Gradient Boosting: Competitive recall for churned customers but slightly lower accuracy than Random Forest.
- Logistic Regression: Balanced precision-recall trade-off but marginally lower overall accuracy.

Model Suitability: Select the model aligning with the priority of accurately identifying churned customers or achieving a balanced predictive performance.

Recommendations:

- Potential Deployment: Consider deploying the Gradient Boosting model due to its competitive performance in detecting churned customers.
- Ensemble Approach: Explore an ensemble method combining models to leverage the strengths of each classifier.
- Continuous Monitoring: Implement continuous model monitoring to assess performance post-deployment and consider retraining models periodically with new data.
- Explain ability and Interpretability: Prioritize models offering interpretability (like Logistic Regression) for easier understanding by stakeholders.

4.3 Potential challenges and additional opportunities.

Challenges:

- Data Quality: Addressing inconsistencies or missing data, ensuring data quality for robust model performance.
- Class Imbalance: Handling imbalanced classes, especially when dealing with churn prediction where positive cases (churned customers) might be relatively fewer.
- Model Interpretability: Balancing model complexity with interpretability, especially crucial for stakeholders' comprehension and acceptance.
- Scalability: Ensuring scalability of the model deployment process, particularly for larger datasets or increased prediction demands.
- Adaptation to Changes: Models might need regular retraining to adapt to evolving customer behaviors or market dynamics.

Opportunities:

- Feature Engineering: Exploring further feature engineering or extraction techniques to enhance model performance.
- Ensemble Techniques: Leveraging ensemble methods to combine multiple models for improved predictive power.
- Customer Segmentation: Utilizing clustering techniques to create customer segments for targeted retention strategies.
- Personalization: Developing personalized marketing strategies tailored to individual customer needs, based on model insights.

4.4 Final Conclusion

In conclusion, the telecom churn prediction project delved into crucial insights aiming to mitigate customer attrition. Leveraging predictive modeling techniques such as Logistic Regression, Random Forest, and Gradient Boosting, we sought to identify potential churners in the customer base. The evaluation showcased diverse model performances, highlighting trade-offs between accuracy, precision, and recall for different classifiers. While the Random Forest model exhibited the highest accuracy, the Gradient Boosting model displayed competitive recall for identifying churned customers, and the Logistic Regression model offered a balanced performance.

Deploying the Gradient Boosting model or exploring ensemble methods presents opportunities to enhance retention strategies. Addressing challenges such as data quality, class imbalance, and ensuring model interpretability remains pivotal. The project emphasizes continuous improvement, advocating for regular model monitoring, adaptation to evolving customer behaviors, and potential feature engineering.

Ultimately, the telecom churn prediction initiative underlines the significance of data-driven strategies in retaining customers, providing actionable insights crucial for sustaining market position, and fostering long-term growth in the dynamic telecom industry.

5.0 Ethical Considerations

In terms of ethical considerations, this project prioritizes the privacy and confidentiality of customer data. Rigorous measures are in place to ensure that sensitive information is protected and anonymized throughout the data analysis process. By following strict data protection protocols and regulatory standards, the project upholds the integrity and trustworthiness of its data handling practices.

Transparency is also a crucial ethical consideration in this project. Detailed documentation of the model development process including data preprocessing steps, feature engineering techniques, and model evaluation methodologies promotes transparency and accountability. This comprehensive approach provides stakeholders with a clear understanding of how the predictive models operate, fostering trust and confidence in the decision-making process.

6.0 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/