Predicting Customer Churn in the Telecom Industry: A Data-Driven Approach to Retention Strategies

Business Application of Machine Learning

Table of Contents

1. Background, Motivation, and Business Question

Background

Customer churn is a major challenge for businesses, especially in industries with strong competition. Telecom companies, in particular, struggle with retaining customers due to factors like pricing, service quality, and better offers from competitors. Losing customers is costly, as acquiring new ones requires significant marketing and operational expenses.

In my experience with Telus home WiFi, I encountered frequent service disruptions—about 5 to 6 times a month—despite reaching out to customer care multiple times. This recurring issue raised a broader question: how do telecom companies manage customer satisfaction and retention? It also led me to wonder what strategies could be used to predict and prevent customer churn more effectively.

Churn prediction has become a key focus for telecom businesses. By identifying the factors that cause customers to leave, companies can take early action to improve their services and keep valuable customers. In recent years, data analysis and machine learning have become powerful tools for understanding customer behavior. These methods allow businesses to predict churn with greater accuracy and develop targeted retention strategies.

Motivation

Retaining customers is more profitable than acquiring new ones. Studies show that even a small increase in customer retention can lead to a significant rise in profits. However, predicting which customers are likely to leave is not always straightforward. Many different factors—such as contract length, billing issues, internet service type, and customer satisfaction—play a role in churn.

Understanding these factors can help businesses:

- Reduce revenue loss by retaining more customers.
- Improve customer experience by addressing service pain points.
- Optimize marketing efforts by focusing on at-risk customers.
- Gain a competitive advantage by acting before customers leave.

By using historical customer data, we aim to build a predictive model that helps businesses proactively reduce churn and improve customer satisfaction.

Business Question

This report seeks to answer the following key question:
“How can we predict customer churn in the telecom industry and enable proactive customer retention?”

To explore this, we analyze a dataset of telecom customers, examine key factors influencing churn, and test different machine learning models to identify the best predictive classification approach. Our findings will provide insights that can help telecom companies take data-driven actions to reduce customer loss and improve long-term business success.

2. Data Selection and Statistical Question

2.1 Dataset overview: For this study, we used the Telco Customer Churn dataset last month from IBM, which provides comprehensive insights into customer behavior, demographics, billing details, and service usage. The churn column indicates whether the customer departed within the last month. This dataset enables us to analyze the factors that contribute to customer churn and develop predictive models to identify high-risk customers.

Key Tables and Features

The dataset consists of five interconnected tables, each containing attributes that provide a detailed view of customer profiles and their interaction with telecom services:

Table Name	Key Features
Demographics	Gender, Age, Senior Citizen, Married, Dependents, Number of Dependents
Location	Country, State, City, Zip Code, Latitude, Longitude
Population	Zip Code, Estimated Population of the Area
Services	Tenure in Months, Offer Accepted, Internet Type, Streaming Services, Contract Type, Payment Method, Monthly Charges, Total Charges
Status	Satisfaction Score, Churn Value, Churn Reason, CLTV, Churn Score

2.2 Statistical Question

To guide our analysis, we formulated the following statistical question:

"Can we predict if a customer is likely to churn based on factors such as their usage patterns, tenure, and service characteristics?"

This question serves as the foundation of our study, aiming to uncover key predictors of churn and leverage them to enhance customer retention strategies.

2.3 Why is This Important?

Predicting customer churn is crucial for businesses, as it allows them to take proactive measures to improve customer experience and retain high-value clients. The benefits of churn prediction include:

- **Identifying Key Factors** – Helps recognize major contributors to churn, such as tenure, monthly charges, and contract type.
- **Customer Segmentation** – Enables businesses to target high-risk customers with personalized retention strategies.
- **Improving Decision-Making** – Supports better pricing strategies, service enhancements, and loyalty program development.
- **Reducing Revenue Loss** – Retaining customers is more cost-effective than acquiring new ones.
- **Enhancing Customer Experience** – Identifies pain points (such as pricing, service issues, and contract terms) to improve overall satisfaction.
- **Optimizing Marketing & Retention** – Helps target at-risk customers with special offers and promotions.
- **Gaining a Competitive Advantage** – Enables proactive interventions before customers switch to competitors.

3 Exploratory Data Analysis

3.1 Data Splitting

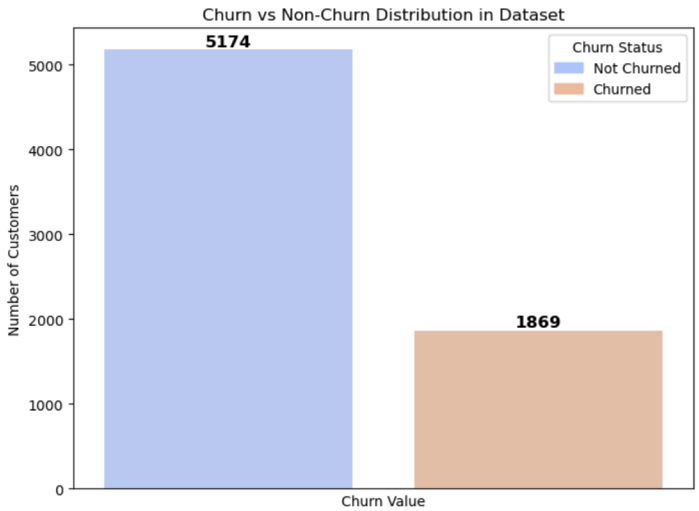
The dataset was split into training (80%) and test (20%) sets to ensure an unbiased evaluation of our predictive models. Stratified sampling was applied based on the churn label to maintain class balance, preventing any skewed distribution in the training and test sets.

3.2 Key Findings from Data Exploration

1. Class Imbalance- Churned VS non churned count
   The dataset exhibits class imbalance, with a smaller proportion of customers (26.5%) labeled as churned. This imbalance may impact model performance. If a model is biased toward the majority class, it might have high accuracy but fail to correctly identify customers at risk of leaving. To fix this issue, we did try oversampling on churned customers with SMOTE (Synthetic Minority Over-sampling Technique)
2. Feature Importance Indicators
   - Satisfaction Score, Contract Type, Monthly Charges, and Tenure in Months show strong correlations with churn. Customers with low satisfaction scores, month-to-month contracts, and higher monthly charges are more likely to churn.
   - Features like Internet Type and Payment Method also exhibit trends that may impact churn behavior.
3. Missing Data
   Certain key features, such as Offer and Internet Type, have missing values. We later addressed these issues by using imputation and adding missing flags to retain information.
4. Skewed Features
   Several numerical features, including Total Charges, Monthly Charges, and Tenure, exhibited high skewness. Log transformation was applied to stabilize variance and improve model learning.

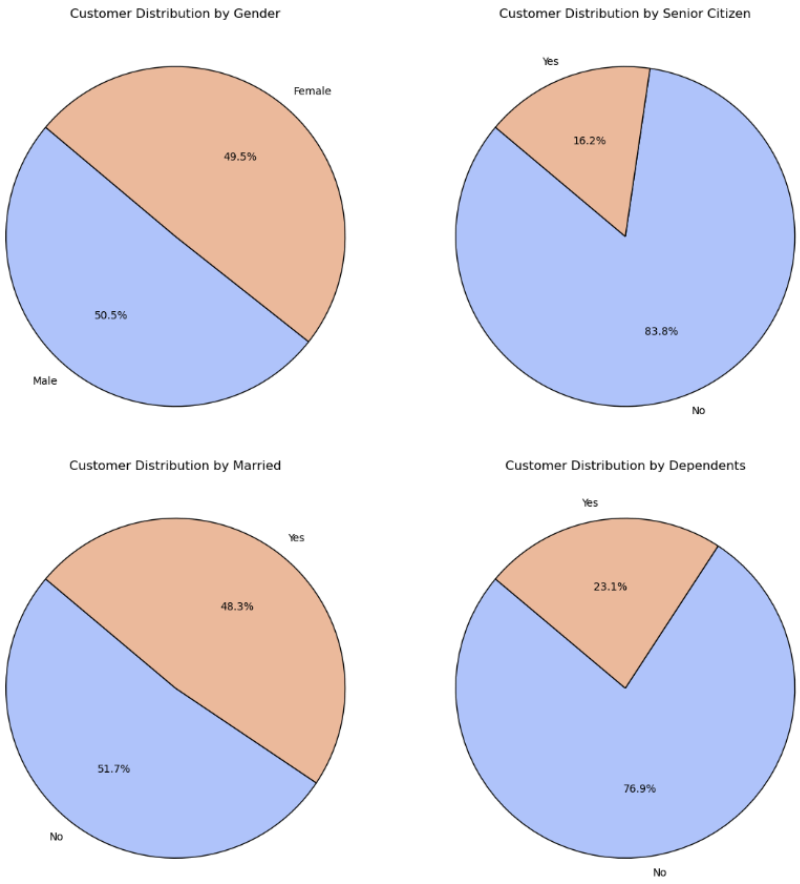
3.3 Visualization Highlights

1. Churn Distribution: The analysis showed us that the dataset is heavily imbalanced, with 73.5% of customers not churning (5,174) and only 26.5% churning (1,869). This class imbalance may cause machine learning models to favor predicting "Not Churned," leading to poor identification of actual churners. To address this, techniques such as resampling or adjusting class weights can improve model performance.



2. Customer Demographics: The customer base consists mostly of younger, independent individuals with no dependents, and gender and marital status are fairly balanced. The results are summarized in the pie plots below:
   - **Gender Distribution:** The customer base is almost evenly split between males (50.5%) and females (49.5%), suggesting no significant gender imbalance.

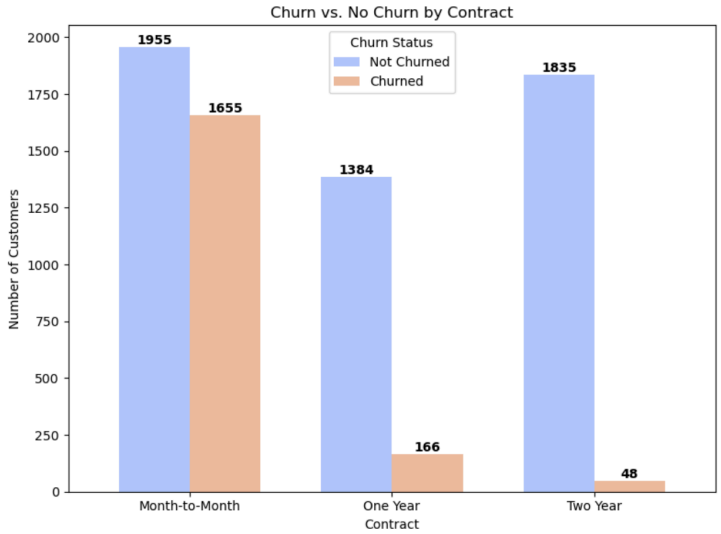
- **Senior Citizen Status:** The majority (83.8%) of customers are not senior citizens, while only 16.2% are seniors, indicating that younger customers form the bulk of the customer base.
- **Marital Status:** Customers are fairly balanced between married (48.3%) and not married (51.7%), showing that marital status is roughly evenly distributed.
- **Dependent Status:** Most customers (76.9%) do not have dependents, while only 23.1% do, suggesting that single or independent individuals make up the majority.



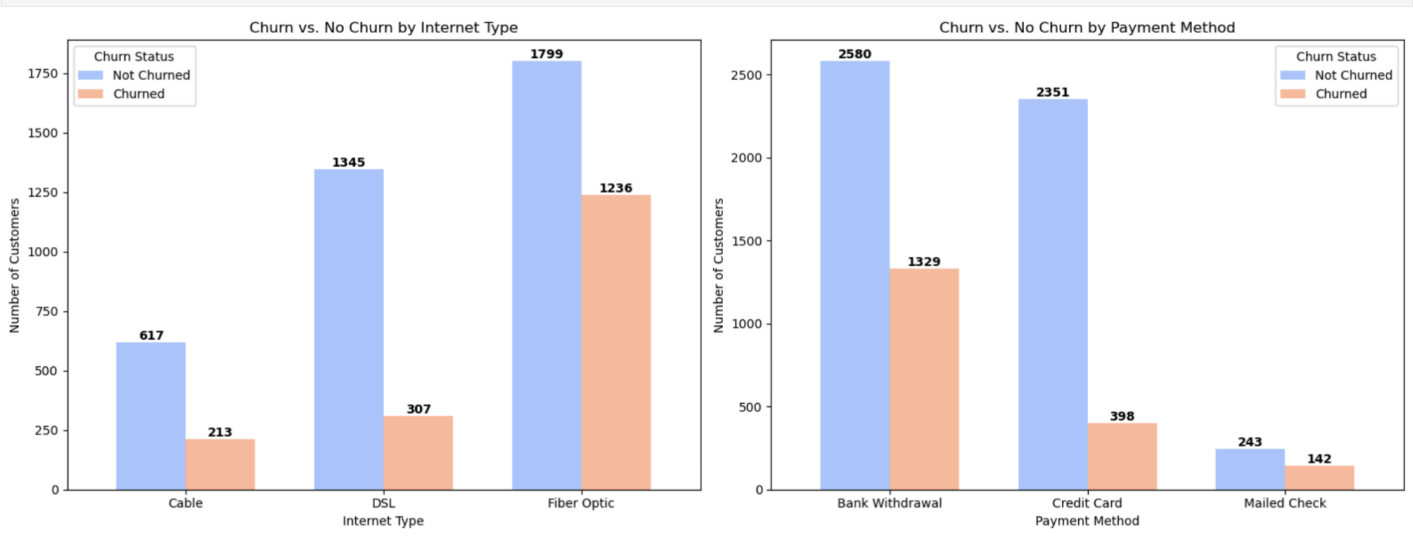
3. Churn by Contract Type: **Month-to-Month contracts** have the highest churn rate (45.3%), with 1,655 churned customers—almost as many as those who stayed (1,955).

   One-Year and Two-Year contracts show significantly lower churn rates. Customers on **One-Year contracts** have a **10.7% churn rate**, while those on **Two-Year contracts** have only a **2.5% churn rate**.

   This suggests that **longer contracts increase customer retention**, likely due to incentives, pricing benefits, or higher commitment levels. Businesses can leverage this insight by offering discounts on extended contracts to reduce churn.

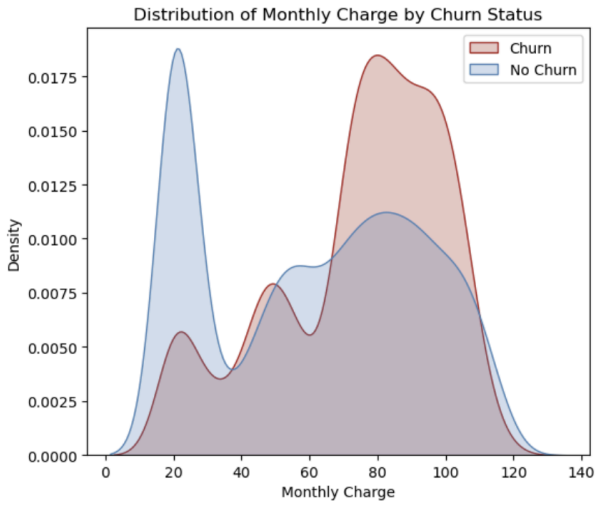


4. Churn by Internet Type & Payment Method
    - Internet Type: Fiber Optic users exhibit the highest churn rate at **40.72%**, indicating potential dissatisfaction with service reliability or pricing. Cable follows with a **25.66%** churn rate, suggesting moderate retention challenges, while DSL has the lowest churn rate at **18.58%**, implying greater stability in customer retention. These insights suggest that improving service quality, offering better pricing, or enhancing customer support for Fiber Optic users could help reduce churn rates.
    - Payment Method: Mailed Check users have the highest churn rate (36.9%), followed by Bank Withdrawal (34.0%), while Credit Card users churn the least (14.5%). This indicates that manual payment users may find the process inconvenient, while auto-payment users may feel less committed, suggesting the need for auto-pay incentives or engagement strategies.



5. Churn by monthly charges : Customers with higher monthly charges are more likely to churn, while those with lower charges tend to stay.

    The chart below shows the distribution of monthly charges for customers who churned versus those who stayed. The blue curve (No Churn) has a peak at lower monthly charges, meaning many customers who stayed had lower bills. In contrast, the red curve (Churned) is more dominant at higher monthly charges, showing that customers with higher bills are more likely to leave. This suggests that pricing strategies can play a big role in retention, offering discounts or better value at higher price points might help reduce the churn rate.



4 Method & Results

4.1 Feature Pre-processing

4.1.1 Feature Selection & Reduction

The dataset was reduced from 62 to 37 columns by removing irrelevant features, redundant features, and non-predictive features.

- Identifiers & Non-Predictive Fields: Unique IDs (Customer ID, Service ID, Location ID) and reporting fields (Count variants) do not help predict churn.
- Duplicate or Highly Correlated Features: Redundant features like Tenure Months, detailed location fields (Country, State, etc.), and alternative churn indicators (Phone Service_churn, etc.) should be removed.
- Churn-Related Fields (Data Leakage Risk): Direct churn indicators (Churn Label, Churn Reason, etc.) and post-churn information can leak future data that are unknown during the prediction stage, making them unsuitable for modeling.
- Non-Informative Features: Time-based fields (Quarter), redundant demographic attributes (Under 30, already covered by Age), and similar variables (Partner vs. Married) add little value.

Removing these features prevented redundancy, data leakage, and irrelevant inputs, leading to a more reliable churn prediction model.

4.1.2. Handling Missing Data & Outliers

We then handled the missing values and reduce the effect of outliers on training:

- For missing numerical features, we filled the missing values using mean imputation.
- For categorical features, we created missing value flags for Offer and Internet Type to capture missingness as a potential signal.
- For skewed features (e.g., Number of Referrals, Total Extra Data Charges), we applied the log transformation to normalize distribution, reduce skewness and stabilize variance.

4.1.3 Data Transformation

The next step was to apply transformations to some features (i.e. feature engineering)

- Numerical Features (e.g., CLTV, Monthly Charge) – Standardized using mean imputation and scaling.
- Ordinal Features (Satisfaction Score) – Processed with median imputation and ordinal encoding, to preserve its natural ranking.

- Nominal Features (categorical variables) – One-hot encoded with unknown categories handled safely.

4.2 Feature Selection (optional)

To enhance model interpretability and reduce redundancy, we considered using Lasso Regression (LassoCV) for feature selection. However, Lasso (L1 regularization) selects features by shrinking less important coefficients to zero, which may lead to the unintended removal of valuable features. In our customer churn prediction project, feature selection using Lasso was not necessary, as our best-performing models are tree-based (e.g., Random Forest, LightGBM, Gradient Boosting). These models inherently perform feature selection by prioritizing the most predictive features, making Lasso redundant for our approach.

4.3 Reflection on Selected Performance Metrics

To evaluate model performance, we considered **Accuracy, F1-score, and AUC-ROC**. While **accuracy** provides a general measure of correctness, it is not a reliable metric for our churn prediction task due to class imbalance—where the majority of customers do not churn. A high accuracy score might be misleading if the model predominantly predicts the majority class (non-churned customers).

Instead, we focused on **F1-score** and **AUC-ROC**:

- **F1-score** balances precision and recall, making it a better metric for identifying churned customers, ensuring both false positives and false negatives are minimized.
- **AUC-ROC** measures the model's ability to distinguish between churned and non-churned customers across different threshold values, providing a more comprehensive assessment of classification performance.

4.4 Model Selection

To achieve the best prediction results, we applied five different types of machine learning models and compared their results. The candidates and their strengths are summarized as follows:

- Logistic Regression
  - A simple baseline model to predict if a customer churns (l) or stays (O).
  - Helps explain how features like monthly charges or contract type influence churn probability.
- Random Forest
  - Combines decision trees to improve accuracy and handle mixed data types
  - Identifies key drivers of churn, like satisfaction scores and internet service.
- Support Vector Machine (SVM)
  - Effective for separating churners and non-churners in complex data patterns.
  - Handles imbalanced datasets well, such as the higher number of non-churners.
- Gradient Boosting
  - Focuses on hard-to-predict cases by correcting errors in sequential models.
  - Captures subtle patterns, such as churn trends in medium-tenure customers.
- LightGBM
  - Optimized for large datasets with fast training and high accuracy.
  - Handles imbalanced data effectively, perfect for our churn prediction task.

4.5 Hyperparameter Tuning

To determine the best parameters for each model, we did parameter tuning. We performed hyperparameter tuning using GridSearchCV with 10-fold cross-validation, optimizing for the F1-score. The best parameters and best trained models are stored for testing later.

Below are the tuning parameters we searched over for each model:

- Random Forest
  - Hyperparameters tested:
  - n_estimators: [100, 200] (number of trees)
  - max_depth: [5, 10, None] (tree depth)
  - min_samples_split: [2, 5] (min samples needed to split)
- Gradient Boosting

    - Hyperparameters tested:
    - n_estimators: [100, 200] (number of boosting iterations)
    - learning_rate: [0.01, 0.1] (step size shrinkage)
    - max_depth: [3, 5] (max tree depth)
- LightGBM
    - Hyperparameters tested:
    - n_estimators: [100, 200]
    - learning_rate: [0.01, 0.1]
    - num_leaves: [20, 31] (leaf nodes per tree)
- SVM (Support Vector Machine)
    - Hyperparameters tested:
    - C: [0.1, 1] (regularization strength)
    - kernel: [‘linear’, ‘rbf’] (kernel type)
- Logistic Regression
    - Hyperparameters tested:
    - C: [0.1, 1] (inverse of regularization strength)

4.6 Cross-Validation to Assess Model Generalization

After hyperparameter tuning, we performed 10-fold cross-validation to evaluate how well the models generalize. While Random Forest showed strong performance on the training data, it struggled on the test set, indicating potential overfitting. In contrast, boosting models, such as Gradient Boosting and LightGBM, demonstrated more consistent performance, making them better candidates for our final selection.

4.7 Model Fitting and Test Data Predictions

Once cross-validation confirmed the best-performing models, we trained them on the full training set and evaluated their performance on the test data. LightGBM emerged as the top-performing model across key metrics, closely followed by Gradient Boosting. Given that telecom datasets can be large, we selected LightGBM as the final model due to its faster training speed and efficiency in handling large-scale data.

Model	Test Accuracy	Test F1-Score	Test AUC-ROC
Random Forest	0.954578	0.908832	0.978460
Gradient Boosting	0.955287	0.911888	0.990790
LightGBM	0.956707	0.914446	0.990944
SVM	0.948190	0.894964	0.985874
Logistic Regression	0.953868	0.908322	0.989865

Note- SMOTE (Synthetic Minority Over-sampling Technique) was applied to address class imbalance. However, it did not significantly improve model performance, so it was excluded from the final process.

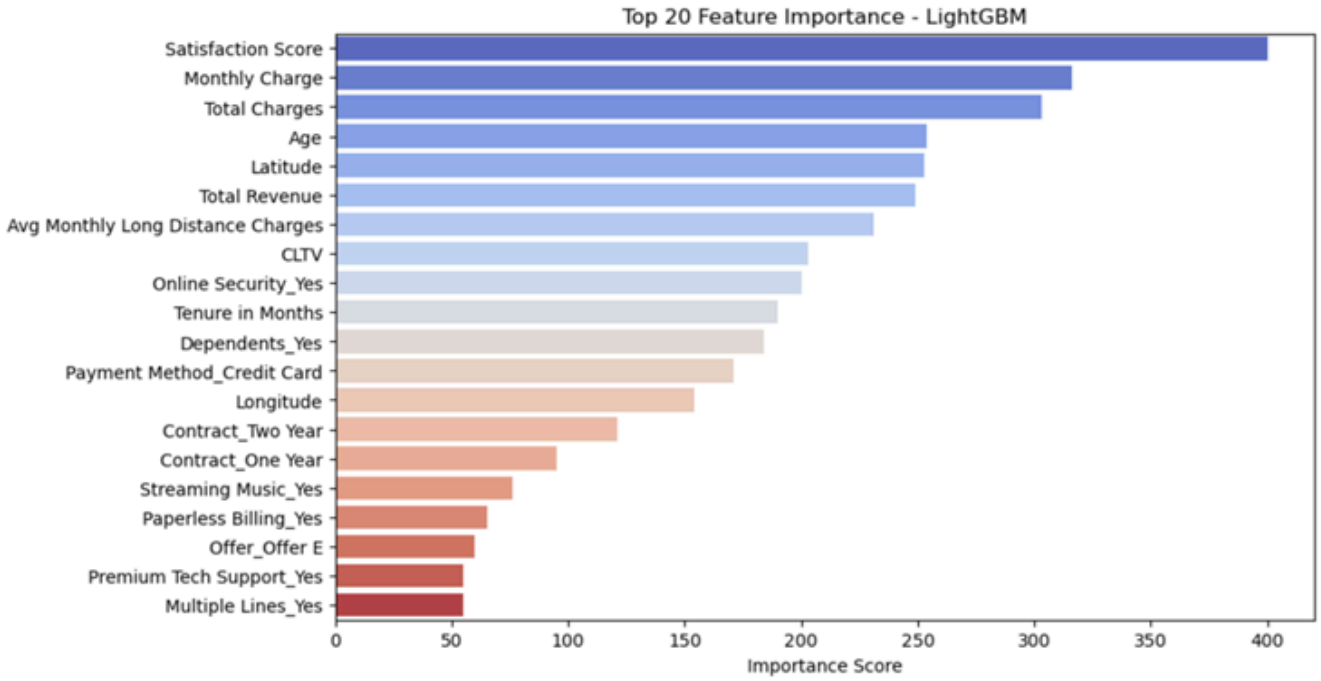
5 Communication of Results, and Advice to a Non-expert

5.1 Top Drivers of Churn and Their Business Impact

Our analysis identifies key factors driving customer churn in the telecommunications industry and provides actionable insights to reduce customer loss. After evaluating a broad set of variables, we found that Satisfaction Score is the most influential predictor of churn. Unlike individual factors such as pricing or contract type, Satisfaction Score captures the overall customer experience, including perceptions of service quality, pricing fairness, customer support effectiveness, and network reliability. Since this score is collected proactively through surveys, it acts as an early warning signal, enabling businesses to address dissatisfaction before customers decide to leave.

Beyond Satisfaction Score, Monthly Charges and Total Charges emerged as critical factors. Customers with higher monthly bills were more likely to churn, particularly if they perceived limited value in the service. However, long-term customers with higher total payments showed greater loyalty, suggesting that tenure strengthens customer retention. Demographic factors such as age and geographic location also played a role, with older customers and those with long-term contracts demonstrating lower churn rates due to a preference for stability. Additionally, regional variations in churn suggest that local competition and service availability influence customer retention, highlighting the need for location-based strategies.

While these were the primary drivers, other factors also contributed to churn. Customer tenure, contract type, and additional service features like online security played a role, with customers on longer contracts or those using security services showing lower churn rates. Payment preferences, such as credit card payments and paperless billing, may also indicate engagement levels, with tech-savvy customers potentially being more loyal. While these factors had a smaller impact compared to satisfaction and pricing, they still shaped a customer's overall experience and decision to stay or leave.



5.2 Recommendations

To effectively reduce churn, the telco company must adopt a **proactive and customer-centric strategy** focused on improving customer experience, service quality, and retention initiatives. Below are our key recommendations:

1. **Enhancing Customer Support & Engagement**
   - Implement **proactive customer support** to address concerns before dissatisfaction escalates.
   - Introduce **loyalty programs and personalized incentives** to retain at-risk customers.
   - Improve **technical support response times** and invest in **online security enhancements** to build trust.
2. **Optimizing Pricing Strategies**

- Develop **cost-effective plans** tailored to **price-sensitive customers** while staying competitive.
- Offer **flexible contract options** to encourage long-term commitments.
- Introduce **discounted bundles** for multi-service subscriptions to increase retention.

3. **Targeted Marketing & Personalized Service Offerings**
   - Tailor telecom service packages to different customer segments:

     Students: Affordable high-speed internet plans.
     Professionals: Business packages with enhanced reliability.
     Seniors: Simplified, cost-effective plans with easy customer support.

   - Utilize **demographic data and behavioral insights** to deliver **personalized offers**.
4. **Regional Service Optimization**
   - Conduct **market research** to understand **local demand and competition** in telecom services.
   - Strategically place customer support centers near campuses, business hubs, and residential communities
   - **Optimize staffing** in high-churn areas to provide better service experiences.

5.3 Limitations & Future Enhancements

While our model demonstrates strong predictive performance, there are areas for further improvement. Addressing missing data through better imputation techniques, tackling class imbalance with advanced resampling methods or improved data collection of churned customers, and mitigating overfitting by reducing model reliance on just one or two dominant features can enhance overall reliability. Additionally, external factors such as competitor pricing and customer sentiment were not included in the dataset but could significantly impact churn behavior

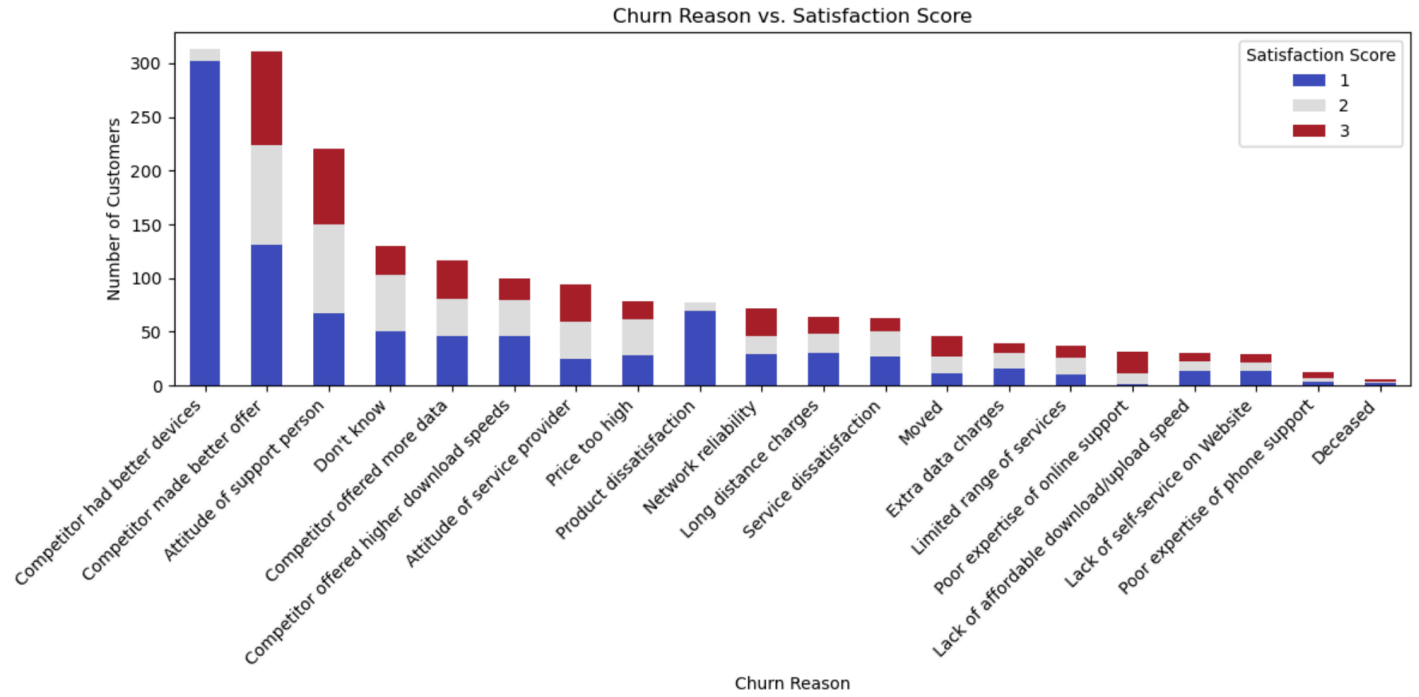
To further refine churn prediction, incorporating **time-series analysis** can help track behavioral trends over time, while **sentiment-based models** leveraging customer feedback can provide deeper insights into dissatisfaction drivers. Moreover, utilizing **multi-class classification on churn reasons** can help businesses understand the specific factors leading to customer churn and implement more targeted retention strategies.

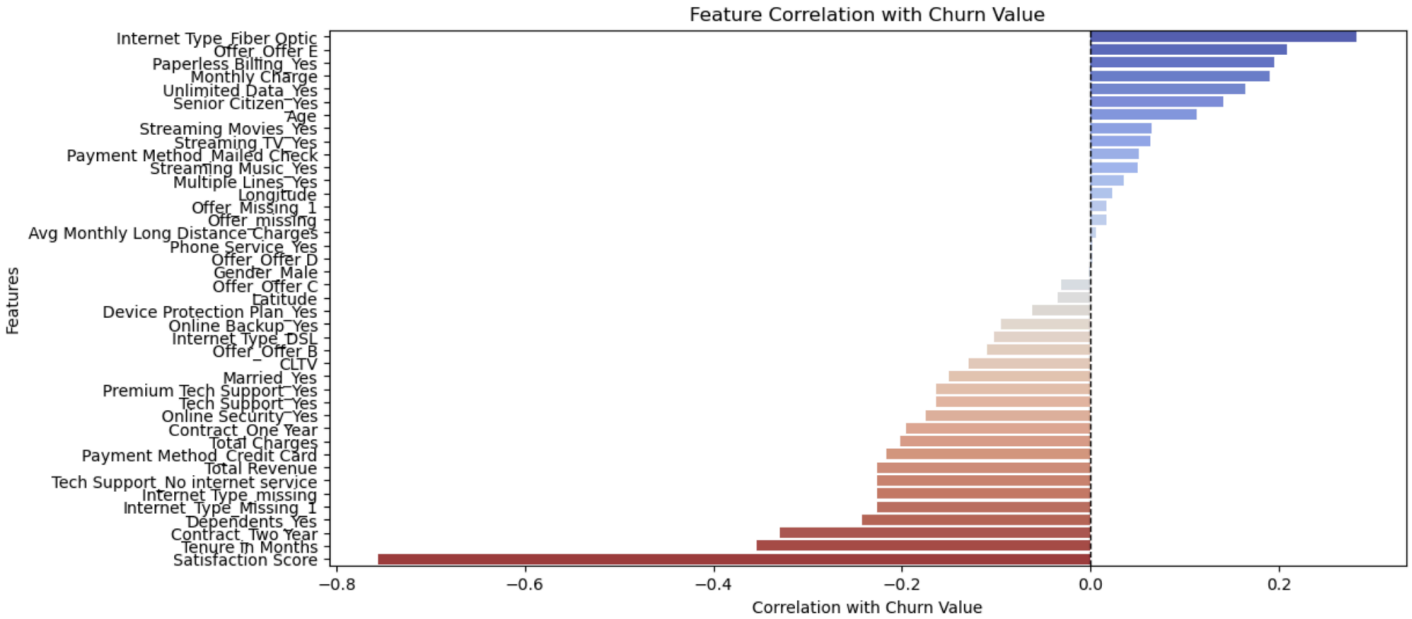
6 Reference

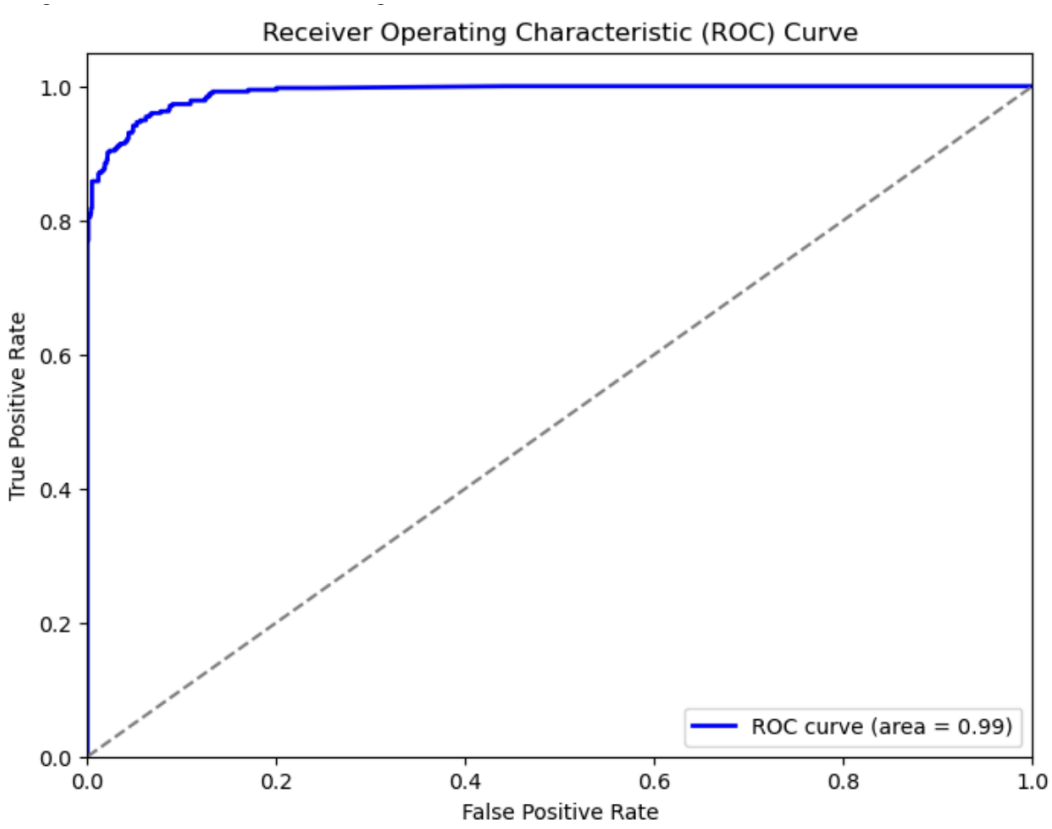
Learning about customer churn (IBM)
https://www.ibm.com/think/topics/customer-churn
https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113

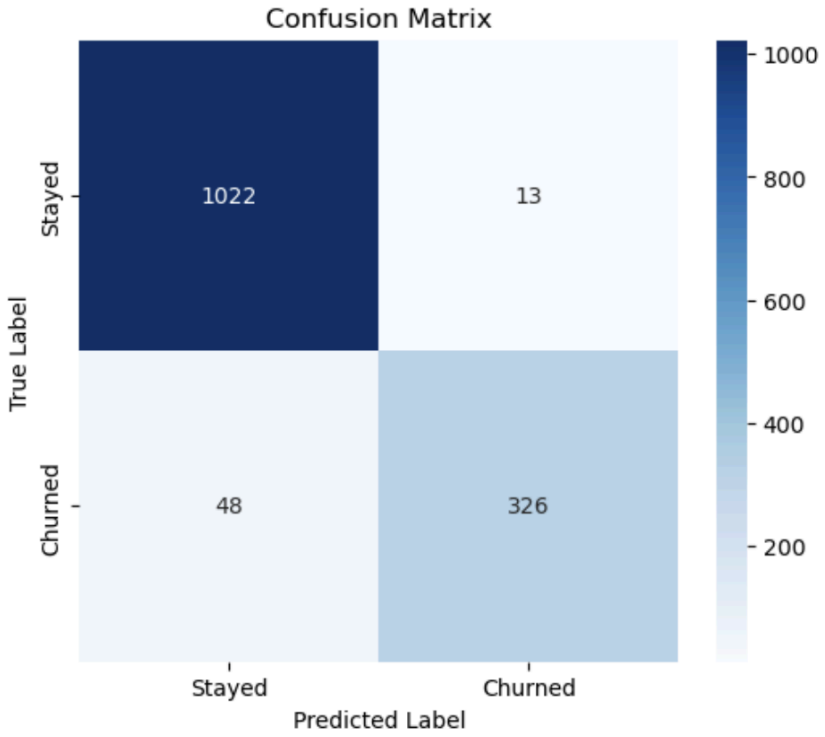
ChatGPT was used as a supplementary resource for concept explanations and for exploring topics beyond the scope of class learning- https://chatgpt.com/share/67afe635-b0b0-8006-b085-31040a2cfd22

Appendix:


Churn Reason vs. Satisfaction Score
Satisfaction Score
1
2
3
Number of Customers
0
50
100
150
200
250
300
Competitor had better devices
Competitor made better offer
Attitude of support person
Don't know
Competitor offered more data
Competitor offered higher download speeds
Attitude of service provider
Price too high
Product dissatisfaction
Network reliability
Long distance charges
Service dissatisfaction
Moved
Extra data charges
Limited range of services
Poor expertise of online support
Lack of affordable download/upload speed
Lack of self-service on Website
Poor expertise of phone support
Deceased
Churn Reason


Feature Correlation with Churn Value
Internet Type_Fiber Optic
Offer_Offer E
Paperless Billing_Yes
Monthly Charge
Unlimited Data_Yes
Senior Citizen_Yes
Age
Streaming Movies_Yes
Streaming TV_Yes
Payment Method_Mailed Check
Streaming Music_Yes
Multiple Lines_Yes
Longitude
Offer_Missing_1
Offer_missing
Avg Monthly Long Distance Charges
Phone Service_Yes
Offer_Offer D
Gender_Male
Offer_Offer C
Latitude
Device Protection Plan_Yes
Online Backup_Yes
Internet Type_DSL
Offer_Offer B
CLTV
Married_Yes
Premium Tech Support_Yes
Tech Support_Yes
Online Security_Yes
Contract_One Year
Total Charges
Payment Method_Credit Card
Total Revenue
Tech Support_No internet service
Internet Type_missing
Internet Type_Missing_1
Dependents_Yes
Contract_Two Year
Tenure in Months
Satisfaction Score
Features
−0.8
−0.6
−0.4
−0.2
0.0
0.2
Correlation with Churn Value


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.99)


Confusion Matrix
True Label
Predicted Label
Stayed
Churned
1022
13
48
326
1000
800
600
400
200