Project Final Report: Customer Churn Prediction in the Telecom Industry

Team Members: Seokhoon Shin, Joshua Nahm, Jiwon Choi, Shinyeong Park, Jaejoong Kim

I. Introduction

In this project, we decided to build a predictive model for customer churn rates using telecommunication data. We considered several potential datasets, including retail data, e-commerce data, craft beer data, and historical sales and active inventory data. However, we ultimately chose telecommunication data because most of our teammates have marketing internship experience, making it an interesting and engaging topic to analyze. Additionally, through these internships, we understand the critical importance of churn rates in the telecom industry.

In fact, studies by Bain & Company have shown that a 5% reduction in customer churn can increase operating profits by as much as 25 to 29%. This highlights that acquiring new customers is typically more costly than retaining existing ones. Also, by proactively identifying and addressing potential issues before customers leave, companies can not only retain customers but also deliver a better overall customer experience.

The goal of this project is to develop a predictive model that can be used by companies to identify customers who are more likely to churn. This enables the company to optimize the effective usage of resources in targeted retention strategies or personalized offers for customers considered at risk. In the process, the profitability becomes long-term and growing sustainably for the businesses.

II. Exploratory Data Analysis

2.1 Teleco Customer Churn Dataset

The Teleco Customer Churn dataset had 21 columns (features) and 7043 rows (customers/observations). The variables are described in Appendix A.

2.2 Data Preprocessing

We first checked for missing data. There were some missing values in the ‘TotalCharges’ column, which were either imputed or removed to maintain data integrity. The next step was to drop unnecessary columns in the dataset such as customerID and gender columns. CusomerID was a unique identifier of customers that did not carry any predicting power. Gender column was dropped because this feature did not have a significant impact on churns. When we created a bar graph of total churned customers by gender, the distribution was nearly identical.

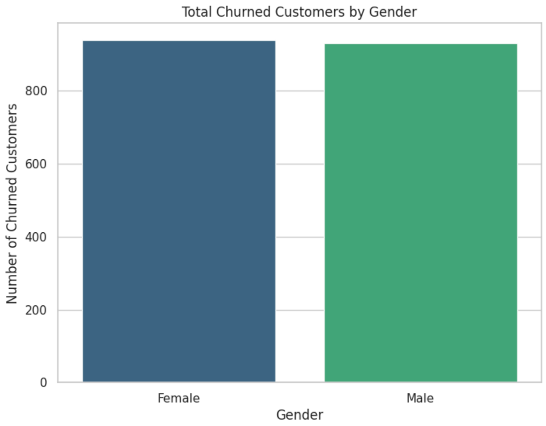


Figure 1. Total Churned Customers by Gender Bar Graph

Next, we encoded categorical variables using **Label Encoding** to make them compatible with machine learning algorithms, and converted binary responses (e.g., ‘Yes/No’) into **0 and 1** values for numerical processing.

2.3 Correlation Matrix

After cleaning our data, we checked for correlation to avoid multicollinearity:

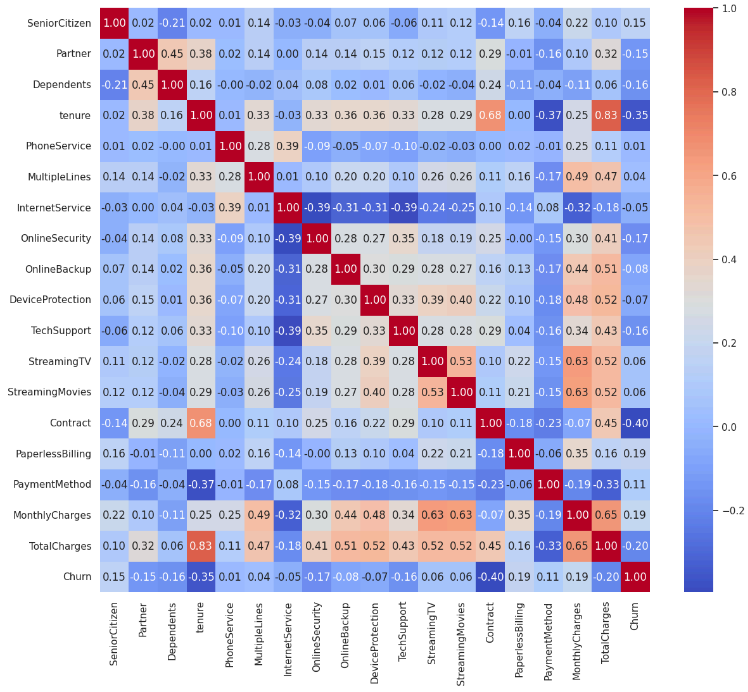
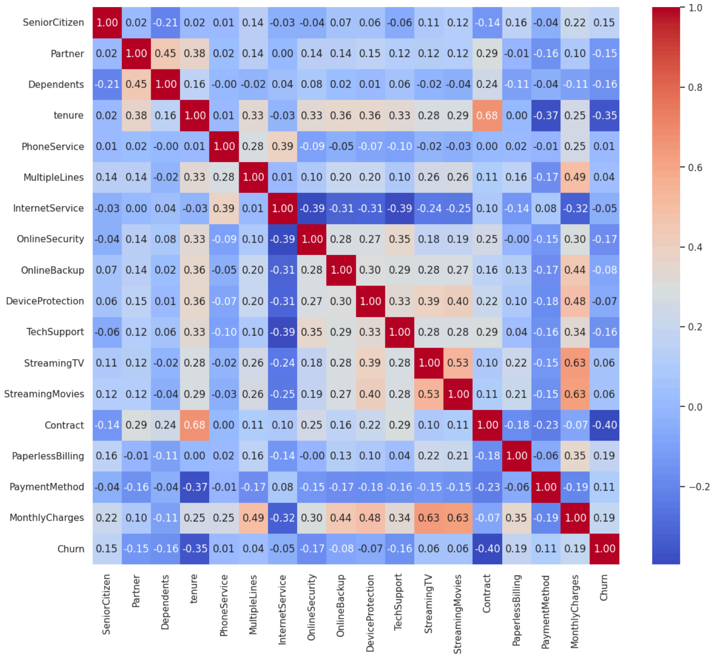


Figure 2. Correlation matrix of all features

There was a high correlation between tenure and total charges with a correlation coefficient of 0.83. We decided to drop total charges because there was a higher correlation between tenure and churn (-0.35) than total charges and churn (-0.20). Without the TotalCharges column, there weren’t significant correlations between variables.

After dropping total charges, the new correlation matrix was as follows:



	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	Churn
SeniorCitizen	1.00	0.02	-0.21	0.02	0.01	0.14	-0.03	-0.04	0.07	0.06	-0.06	0.11	0.12	-0.14	0.16	-0.04	0.22	0.15
Partner	0.02	1.00	0.45	0.38	0.02	0.14	0.00	0.14	0.14	0.15	0.12	0.12	0.12	0.29	-0.01	-0.16	0.10	-0.15
Dependents	-0.21	0.45	1.00	0.16	-0.00	-0.02	0.04	0.08	0.02	0.01	0.06	-0.02	-0.04	0.24	-0.11	-0.04	-0.11	-0.16
tenure	0.02	0.38	0.16	1.00	0.01	0.33	-0.03	0.33	0.36	0.36	0.33	0.28	0.29	0.68	0.00	-0.37	0.25	-0.35
PhoneService	0.01	0.02	-0.00	0.01	1.00	0.28	0.39	-0.09	-0.05	-0.07	-0.10	-0.02	-0.03	0.00	0.02	-0.01	0.25	0.01
MultipleLines	0.14	0.14	-0.02	0.33	0.28	1.00	0.01	0.10	0.20	0.20	0.10	0.26	0.26	0.11	0.16	-0.17	0.49	0.04
InternetService	-0.03	0.00	0.04	-0.03	0.39	0.01	1.00	-0.39	-0.31	-0.31	-0.39	-0.24	-0.25	0.10	-0.14	0.08	-0.32	-0.05
OnlineSecurity	-0.04	0.14	0.08	0.33	-0.09	0.10	-0.39	1.00	0.28	0.27	0.35	0.18	0.19	0.25	-0.00	-0.15	0.30	-0.17
OnlineBackup	0.07	0.14	0.02	0.36	-0.05	0.20	-0.31	0.28	1.00	0.30	0.29	0.28	0.27	0.16	0.13	-0.17	0.44	-0.08
DeviceProtection	0.06	0.15	0.01	0.36	-0.07	0.20	-0.31	0.27	0.30	1.00	0.33	0.39	0.40	0.22	0.10	-0.18	0.48	-0.07
TechSupport	-0.06	0.12	0.06	0.33	-0.10	0.10	-0.39	0.35	0.29	0.33	1.00	0.28	0.28	0.29	0.04	-0.16	0.34	-0.16
StreamingTV	0.11	0.12	-0.02	0.28	-0.02	0.26	-0.24	0.18	0.28	0.39	0.28	1.00	0.53	0.10	0.22	-0.15	0.63	0.06
StreamingMovies	0.12	0.12	-0.04	0.29	-0.03	0.26	-0.25	0.19	0.27	0.40	0.28	0.53	1.00	0.11	0.21	-0.15	0.63	0.06
Contract	-0.14	0.29	0.24	0.68	0.00	0.11	0.10	0.25	0.16	0.22	0.29	0.10	0.11	1.00	-0.18	-0.23	-0.07	-0.40
PaperlessBilling	0.16	-0.01	-0.11	0.00	0.02	0.16	-0.14	-0.00	0.13	0.10	0.04	0.22	0.21	-0.18	1.00	-0.06	0.35	0.19
PaymentMethod	-0.04	-0.16	-0.04	-0.37	-0.01	-0.17	0.08	-0.15	-0.17	-0.18	-0.16	-0.15	-0.15	-0.23	-0.06	1.00	-0.19	0.11
MonthlyCharges	0.22	0.10	-0.11	0.25	0.25	0.49	-0.32	0.30	0.44	0.48	0.34	0.63	0.63	-0.07	0.35	-0.19	1.00	0.19
Churn	0.15	-0.15	-0.16	-0.35	0.01	0.04	-0.05	-0.17	-0.08	-0.07	-0.16	0.06	0.06	-0.40	0.19	0.11	0.19	1.00

Figure 3. Correlation matrix after dropping TotalCharges column

2.4 Preliminary Insights

Before moving on to our models, we explored our data for preliminary insights. Specifically, we looked into contract type, payment method, tenure, and monthly charges.

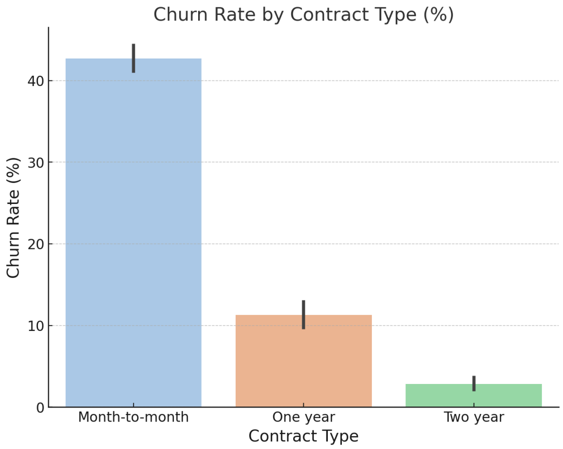


Figure 4. Churn rate by contract type in percent

Customers on month-to month contracts are more likely to churn, which accounts for **~89%** of churned customers.

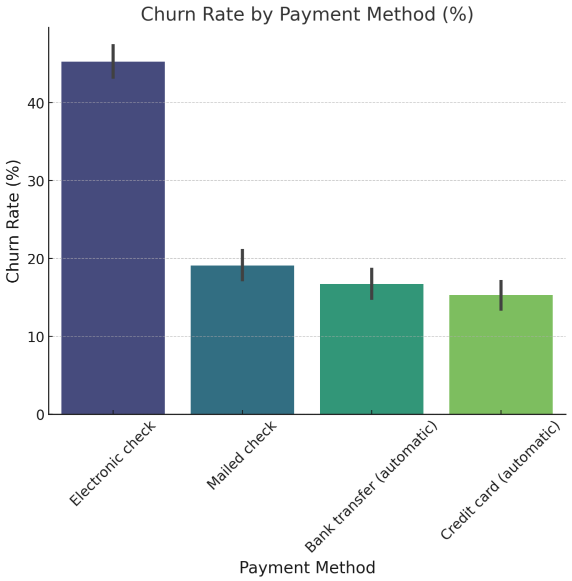


Figure 5. Churn Rate by Payment Method

Customers who pay with **electronic checks** are more likely to churn, which accounts for **~57%** of churned customers.

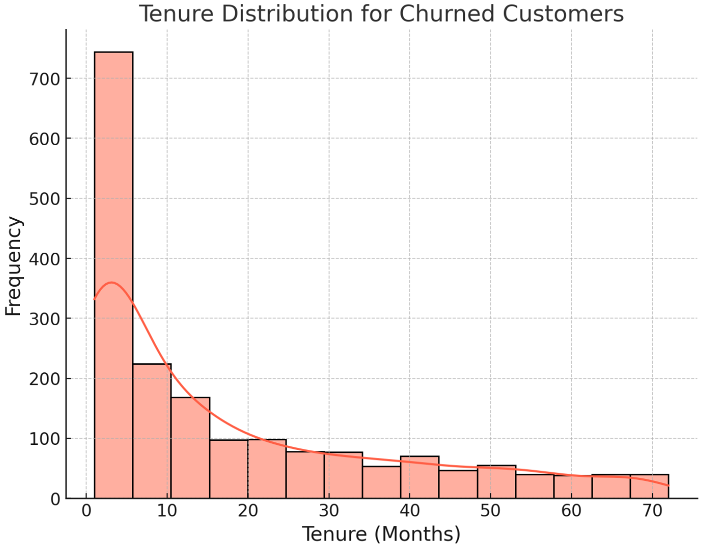


Figure 6. Tenure Distribution for Churned Customers

Customers with short tenures (0-6 months) are at the highest risk, and ~43% of churned customers fall within this range.

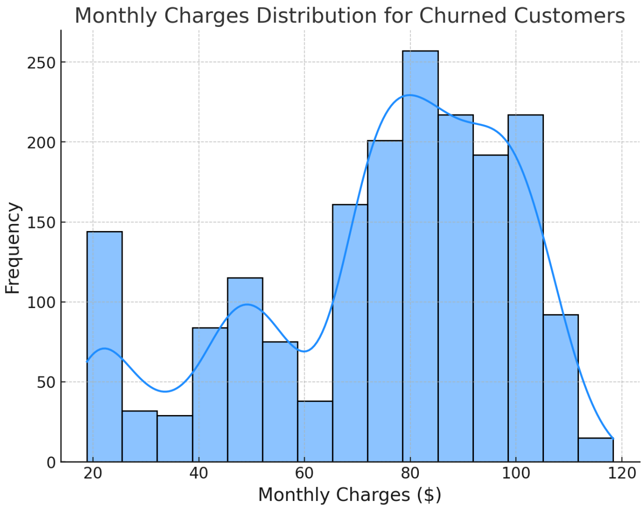
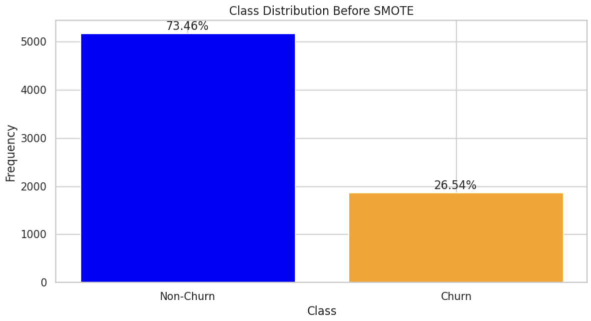


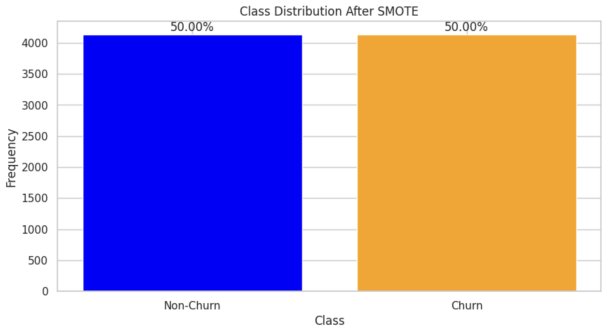
Figure 7. Monthly Charges Distribution

Higher monthly charges correlate with increased churn likelihood. Customers with lower charges (<$40) show lower churn rates.

2.5 Modeling Concerns and Hyperparameter Tuning

We normalized features such as 'MonthlyCharges' to ensure fair model performance. We also applied both grid search and random search to optimize the parameters for **Random Forest, XGBoost, CatBoost, and LightGBM** models. This tuning helped improve each model's accuracy, particularly for Random Forest and LightGBM, which demonstrated the best performance. Finally, we used SMOTE to account for class imbalance.





III. Modeling

1. What models did you choose and why? (3-5 models)

1) **Logistic Regression:** Selected as a baseline due to simplicity, efficiency, and interpretability, making it ideal for the initial exploration of binary classification tasks.
2) **Random Forest:** Chosen for its robustness, ability to handle numerical and categorical data, and feature importance insights. It performs well with non-linear relationships and minimizes overfitting through ensemble learning.
3) **XGBoost:** Preferred for its strong performance on structured data, built-in regularization, and ability to handle missing values. It is suitable for capturing complex patterns in tabular data.
4) **CatBoost:** Optimized for categorical features, this model required minimal preprocessing and demonstrated robust performance, especially for datasets with significant categorical variables.
5) **LightGBM:** Selected for its computational efficiency, fast training speeds, and scalability, making it ideal for large datasets and real-time processing.

2. What accuracy measure are you optimizing for? Why?

Accuracy Metric: Optimizing for F1-Score

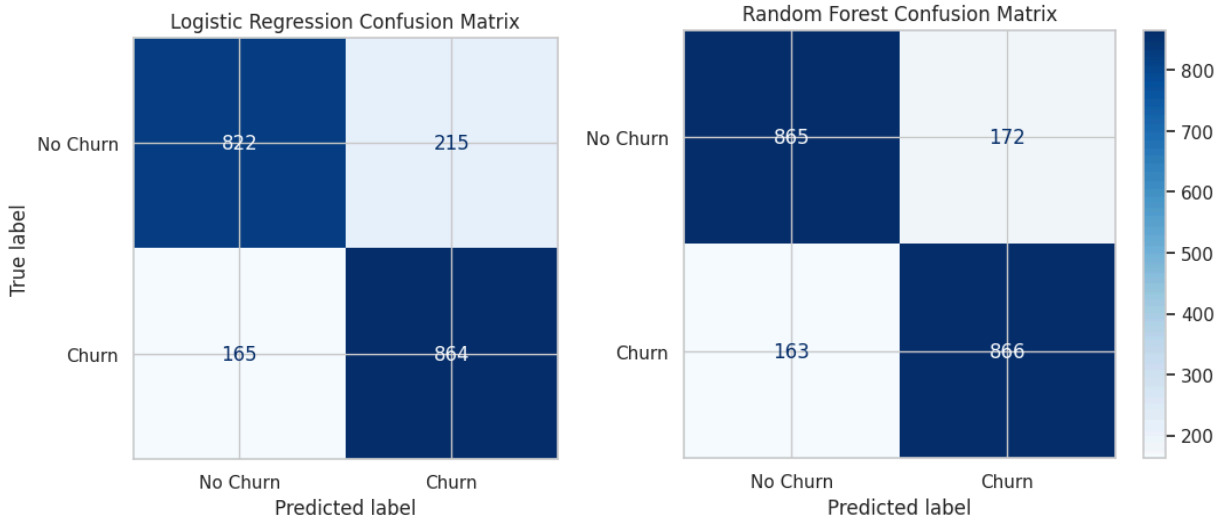
1) The **F1 score** was chosen to balance **precision** (minimizing unnecessary retention efforts) and **recall** (capturing as many actual churners as possible), which is crucial for an imbalanced dataset.
2) **Cost of False Negatives:** Losing potential churners results in revenue loss and missed retention opportunities.
3) **Cost of False Positives:** Inefficient resource allocation for non-churning customers reduces profitability but is less severe than false negatives.
4) Baseline accuracy using the naive rule (predicting the majority class) was 73%.

3. For each model:

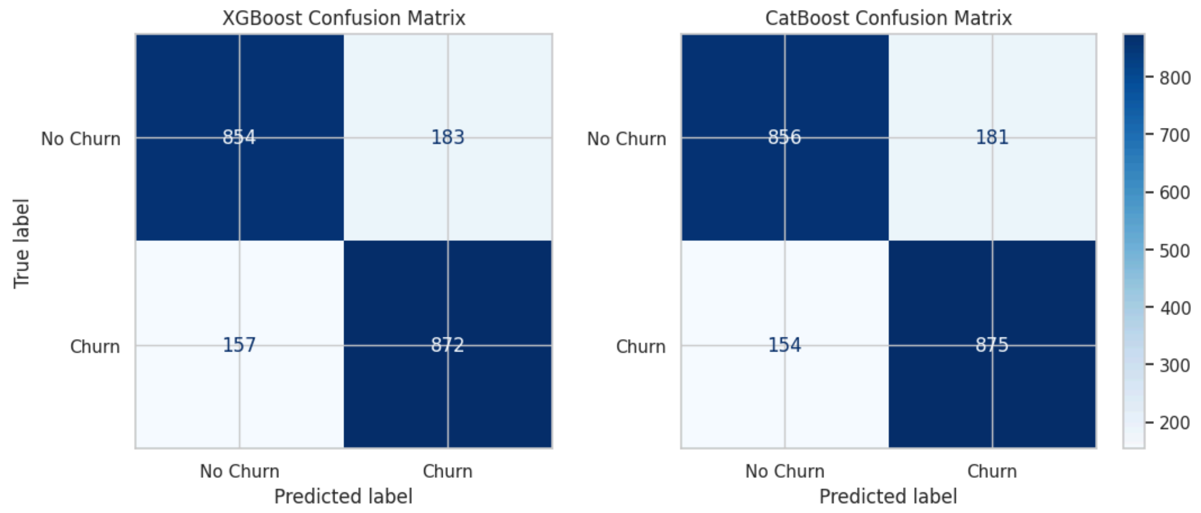
a. Explain its drawbacks and advantages
b. Your approach to implementation (hyperparameter tuning)
c. How do you know the model is not overfit?
d. Visualize the output (confusion matrix, graph)

Model Evaluation For each model, we assessed advantages, drawbacks, hyperparameter tuning, overfitting, and performance visualization:

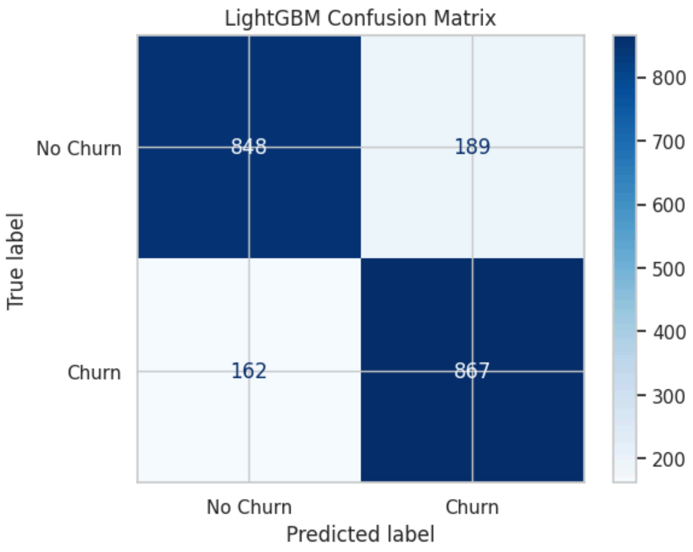
1) **Logistic Regression:** Simple and interpretable but struggled with complex patterns. We achieved 81.5% training accuracy and 80.8% testing accuracy with minimal overfitting (0.7% difference).
2) **Random Forest:** Robust and interpretable but computationally intensive. Grid search optimized parameters (300 trees, depth 25), achieving 86.2% training and 84.5% testing accuracy (1.7% difference).



3) **XGBoost:** Excellent for complex patterns but computationally heavy. Tuned parameters (learning rate 0.1, max depth 6), achieving 85.8% training and 83.8% testing accuracy (2.0% difference).
4) **CatBoost:** Strong with categorical data but slower training. Optimized with 800 iterations, depth 6, and a learning rate of 0.05, achieving 86.0% training and 84.2% testing accuracy (1.8% difference).



5) **LightGBM:** Highly efficient but slightly less interpretable. Tuned parameters (63 leaves, learning rate 0.05), achieving 85.5% training and 84.4% testing accuracy (1.1% difference).



4. Overfitting Assessment and Conclusion: All models showed minimal overfitting, with training and testing accuracies differing by no more than 2-3%. To sum up, Random Forest and LightGBM emerged as top performers, balancing predictive accuracy, interpretability, and computational efficiency. These models are recommended for deployment to mitigate customer churn in the telecom industry.

IV. Findings and Implications

1. The Role of SMOTE

SMOTE (Synthetic Minority Oversampling Technique) is a method for addressing class imbalance by creating synthetic examples of the minority class (in this case, churned customers). This ensures that machine learning models do not become biased towards the majority class and improves their ability to predict the minority class accurately.

Impact of SMOTE:

- Without SMOTE, our models favored predicting non-churn customers due to the imbalance in the dataset (73% non-churn vs. 27% churn).
- After applying SMOTE, the performance metrics, particularly recall for churned customers, improved significantly, ensuring that the models were better at identifying customers likely to churn. For instance:
    - Random Forest achieved an accuracy of **84%**, with a balanced F1-score for both churned and non-churned customers.
    - CatBoost and XGBoost also displayed improvements in classifying churned customers effectively.

2.

V. Conclusion

1) General limitations of your data

Our model currently achieves high accuracy in predicting customer churn, with **Random Forest** and **LightGBM** providing the best results. However, we recognize that the dataset has limitations, particularly regarding class imbalance. Although techniques like SMOTE have been used to address this, they might have introduced some noise into the data, impacting model performance. Additionally, the dataset primarily reflects customer behavior within a specific telecom context, which may not fully represent broader, **real-world scenarios across various industries**.

2) Areas for Improvement

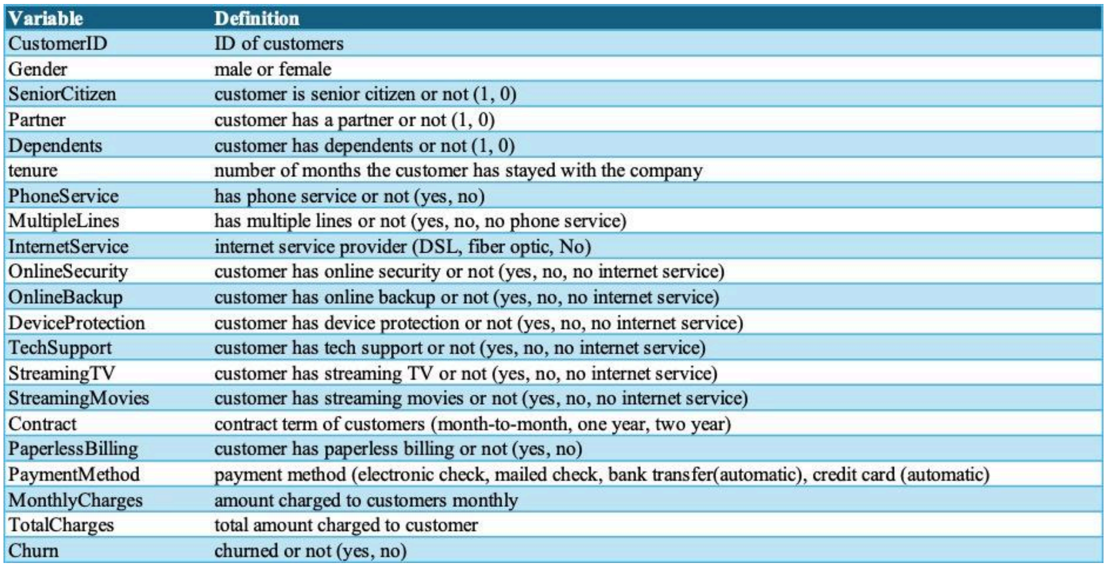
If given the chance to redo the project, we would focus on further **refining our handling of class imbalance.** Exploring alternative methods like **NearMiss or Tomek Links** could help **reduce** the **noise introduced by SMOTE**. Furthermore, we would experiment with more sophisticated feature selection techniques to improve model interpretability and reduce overfitting. It would also be beneficial to integrate external datasets, such as customer service interactions or regional market data, to provide a more holistic view of churn behaviors.

3) Answer to Your Problem Statement

Our analysis successfully identifies the key factors influencing customer churn, including **payment methods, contract types, and customer tenure**. Through our models, we are able to predict churn with reasonable accuracy, providing actionable insights to telecom companies for targeted retention strategies.

VI. Appendix

6.1 Appendix A: Data Dictionary



Variable	Definition
CustomerID	ID of customers
Gender	male or female
SeniorCitizen	customer is senior citizen or not (1, 0)
Partner	customer has a partner or not (1, 0)
Dependents	customer has dependents or not (1, 0)
tenure	number of months the customer has stayed with the company
PhoneService	has phone service or not (yes, no)
MultipleLines	has multiple lines or not (yes, no, no phone service)
InternetService	internet service provider (DSL, fiber optic, No)
OnlineSecurity	customer has online security or not (yes, no, no internet service)
OnlineBackup	customer has online backup or not (yes, no, no internet service)
DeviceProtection	customer has device protection or not (yes, no, no internet service)
TechSupport	customer has tech support or not (yes, no, no internet service)
StreamingTV	customer has streaming TV or not (yes, no, no internet service)
StreamingMovies	customer has streaming movies or not (yes, no, no internet service)
Contract	contract term of customers (month-to-month, one year, two year)
PaperlessBilling	customer has paperless billing or not (yes, no)
PaymentMethod	payment method (electronic check, mailed check, bank transfer(automatic), credit card (automatic)
MonthlyCharges	amount charged to customers monthly
TotalCharges	total amount charged to customer
Churn	churned or not (yes, no)

Table 1. Data Dictionary

VII. Biggest Pitfalls

예상 질문

Why use SMOTE rather than Random Over-sampling?

Both techniques aim to address class imbalance, but **SMOTE** is generally preferred when you're looking to improve model robustness and avoid overfitting caused by duplicate data points.

However, SMOTE can introduce noise if the minority class has outliers or poorly defined boundaries.