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. During the dataset selection process, we considered several options, 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. Through these internships, we also gained insight into the critical importance of churn rates in the telecom industry. In fact, studies by Bain & Company have shown that reducing customer churn by just 5% can boost profits by more than 25% because retaining an existing customer is less costly than acquiring a new one[1]. This significant impact of reducing churn motivated us to create a predictive model to analyze this important issue.

The goal of this project is to analyze the behavior of the customers to timely and accurately predict the possible churners. By leveraging this model, companies can not only pinpoint at-risk customers but also understand the key factors driving their likelihood to churn. This understanding allows businesses to proactively address potential issues, improving the overall customer experience. Moreover, companies can effectively optimize resource allocation by focusing on customers who are most likely to churn and can design retention policies and efficient customer management strategies. In the long term, these proactive measures lead to enhanced customer satisfaction, increased profitability, and consistent growth.

II. Exploratory Data Analysis

2.1 Telco Customer Churn Dataset

The Telco Customer Churn dataset had 21 columns (features) and 7043 rows (customers/observations). The dataset contains customer demographics, services, account information, and whether the customer has churned or not. Some key features include customer tenure, monthly charges, contract type, and payment methods. All of the variables are described in Appendix A.

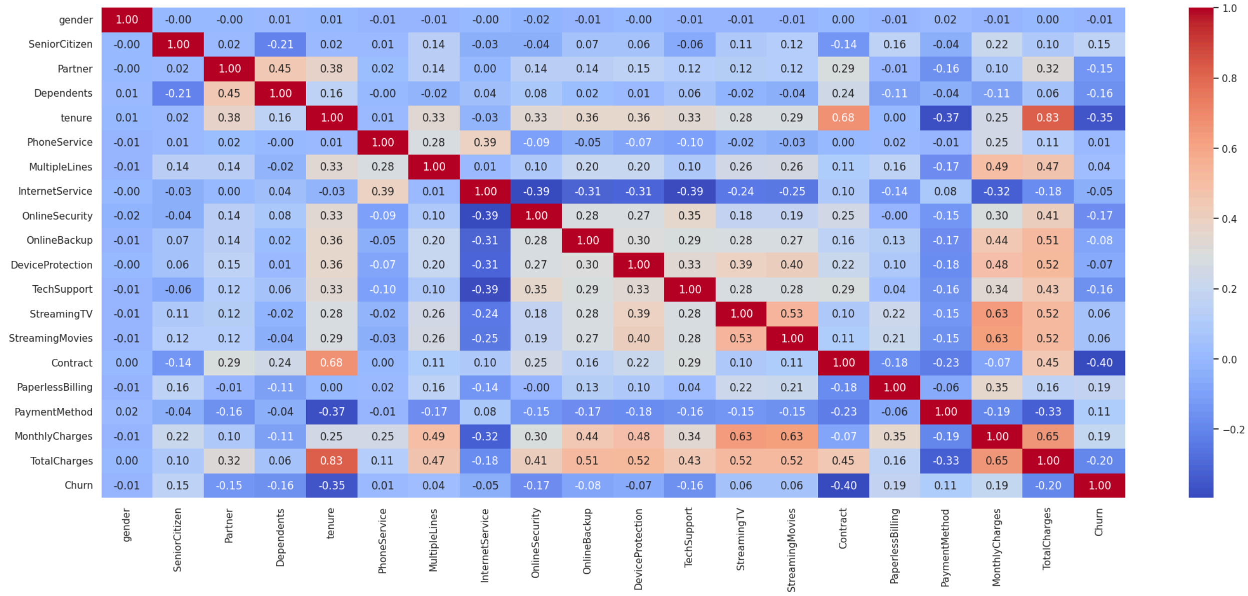
2.2 Data Preprocessing

[1] Bain & Company. 2006. “Retaining customers is the real challenge.” https://www.bain.com/insights/retaining-customers-is-the-real-challenge/.

We first confirmed that there were no missing or NAN values in our dataset. Then, we dropped the customerID column because this was a unique identifier of customers that did not carry any predicting power. Finally, we encoded categorical variables using **Label Encoding** to make them compatible with machine learning algorithms, and converted binary responses (e.g., ‘Yes/No’) into **1 and 0** values for numerical processing.

2.3 Correlation Matrix

After cleaning our data, we checked the correlation between variables to avoid multicollinearity:

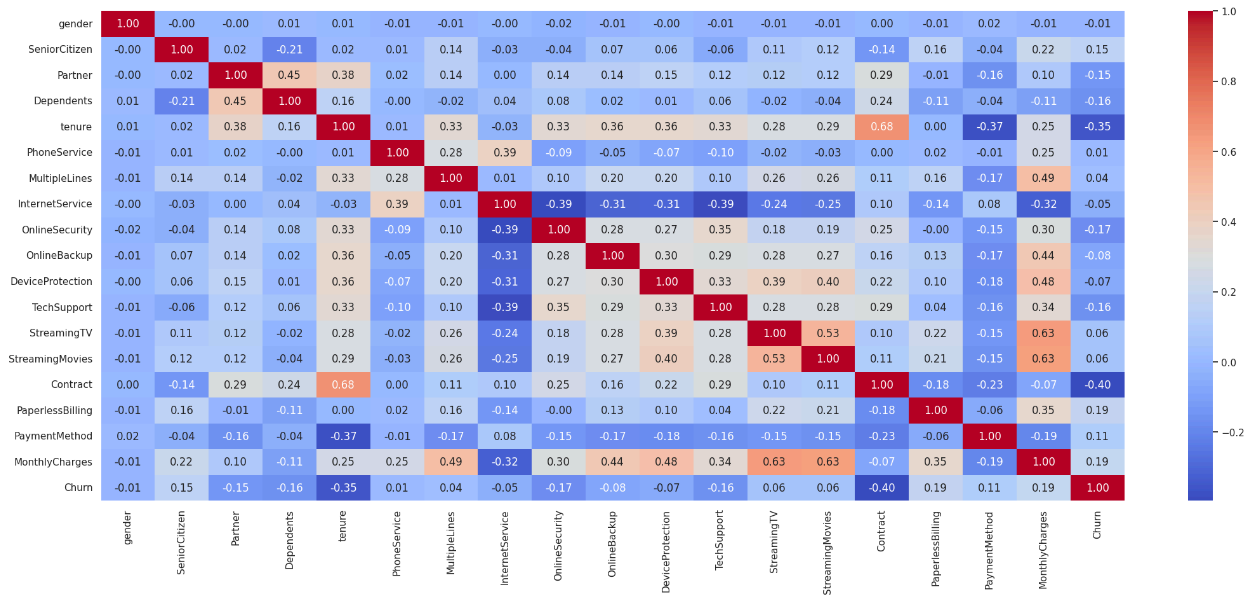


	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
gender	1.00	-0.00	-0.00	0.01	0.01	-0.01	-0.01	-0.00	-0.02	-0.01	-0.00	-0.01	-0.01	-0.01	0.00	-0.01	0.02	-0.01	0.00	-0.01
SeniorCitizen	-0.00	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	0.10	0.15
Partner	-0.00	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	0.32	-0.15
Dependents	0.01	-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	0.06	-0.16
tenure	0.01	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	0.83	-0.35
PhoneService	-0.01	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	0.11	0.01
MultipleLines	-0.01	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	0.47	0.04
InternetService	-0.00	-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	-0.18	-0.05
OnlineSecurity	-0.02	-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	0.41	-0.17
OnlineBackup	-0.01	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	0.51	-0.08
DeviceProtection	-0.00	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	0.52	-0.07
TechSupport	-0.01	-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	0.43	-0.16
StreamingTV	-0.01	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	0.52	0.06
StreamingMovies	-0.01	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	0.52	0.06
Contract	0.00	-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	0.45	-0.40
PaperlessBilling	-0.01	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	0.16	0.19
PaymentMethod	0.02	-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	-0.33	0.11
MonthlyCharges	-0.01	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	0.65	0.19
TotalCharges	0.00	0.10	0.32	0.06	0.83	0.11	0.47	-0.18	0.41	0.51	0.52	0.43	0.52	0.52	0.45	0.16	-0.33	0.65	1.00	-0.20
Churn	-0.01	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	-0.20	1.00

Figure 1. 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.

The new correlation matrix was as follows:



	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	Churn
gender	1.00	-0.00	-0.00	0.01	0.01	-0.01	-0.01	-0.00	-0.02	-0.01	-0.00	-0.01	-0.01	-0.01	0.00	-0.01	0.02	-0.01	-0.01
SeniorCitizen	-0.00	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	-0.00	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	0.01	-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	0.01	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	0.01	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	-0.01	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	-0.00	-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	-0.02	-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	-0.01	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	-0.00	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	-0.01	-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	-0.01	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	-0.01	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	0.00	-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	-0.01	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	0.02	-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	-0.01	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	-0.01	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 2. 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.

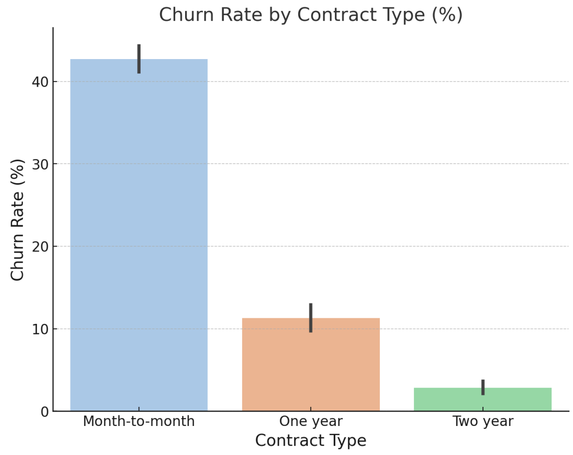


Figure 3. Churn rate by Contract type

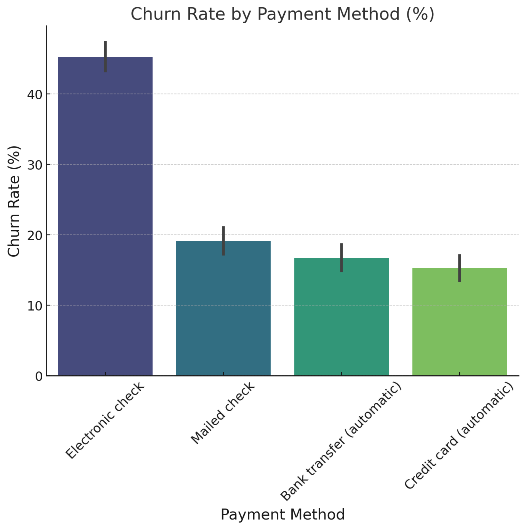


Figure 4. Churn Rate by Payment Method

Customers on **month-to month** contracts are more likely to churn, which accounts for **~89%** of churned customers. Customers who pay with **electronic checks** are more likely to churn, which accounts for **~57%** of churned customers.

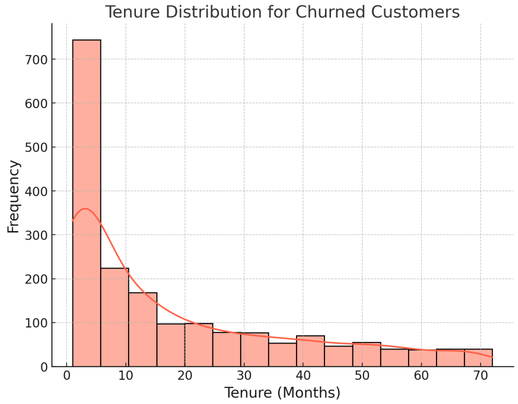


Figure 5. Tenure Distribution for Churned Customers

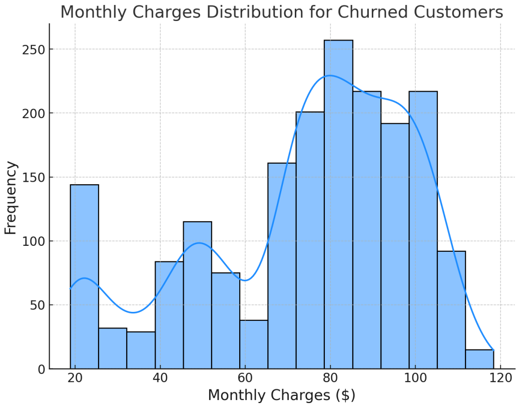


Figure 6. Monthly Charges Distribution for Churned Customer

Customers with **short tenures (0-6 months)** are at the highest risk, and **~43%** of churned customers fall within this range. **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 에'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, our dataset had a significantly higher proportion of non-churners, resulting in a class imbalance. To address this issue, we applied Synthetic Minority Over-sampling Technique (SMOTE). Principal Component Analysis (PCA) was unnecessary because we wanted to prioritize maintaining interpretability.

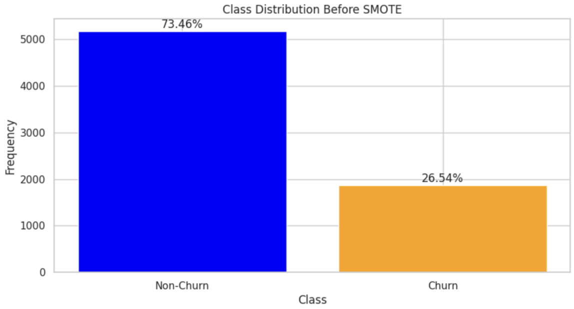


Figure 7. Class Distribution Before SMOTE

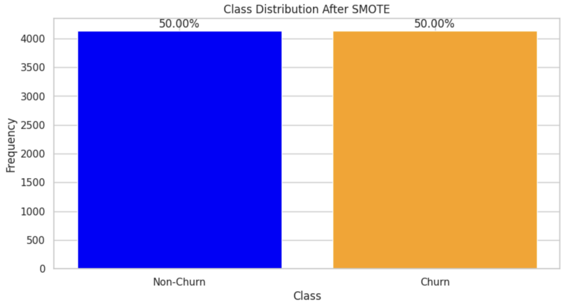


Figure 8. Class Distribution After SMOTE

III. Modeling

3.1 Model Selection

Five models were evaluated based on their strengths and applicability. Logistic Regression served as a baseline model due to its simplicity, efficiency, and ease of interpretation. Random Forest was chosen for its robustness and ability to handle both numerical and categorical data while providing insights into feature importance. XGBoost, known for its superior performance on structured data and built-in regularization, was included to capture complex patterns. CatBoost, optimized for categorical features with minimal preprocessing requirements, demonstrated strong performance with datasets rich in categorical variables. Lastly, LightGBM was selected for its computational efficiency, fast training speeds, and scalability, making it particularly suitable for large datasets and real-time processing.

3.2 Accuracy Measure

The F1 score was prioritized as the primary evaluation metric to balance precision and recall, given the significant costs associated with false negatives (lost revenue and retention opportunities) and false positives (inefficient resource allocation). Baseline accuracy using a naive majority-class prediction was established at 73%.

3.3 Model Implementation and Evaluation

Each model was implemented with hyperparameter tuning to optimize performance. Logistic Regression achieved an accuracy of 81% and an F1-score of 0.81. While interpretable and efficient, it struggled to capture complex patterns, evidenced by a higher rate of false negatives compared to advanced models. Random Forest, with hyperparameters set to 300 trees and a depth of 25, achieved 85% accuracy and an F1-score of 0.85. It demonstrated robustness and reduced false negatives significantly while maintaining interpretability.

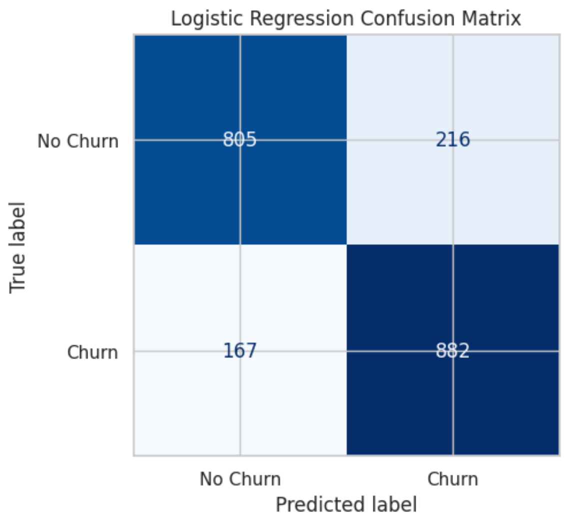


Figure 9. Logistic Regression Confusion Matrix

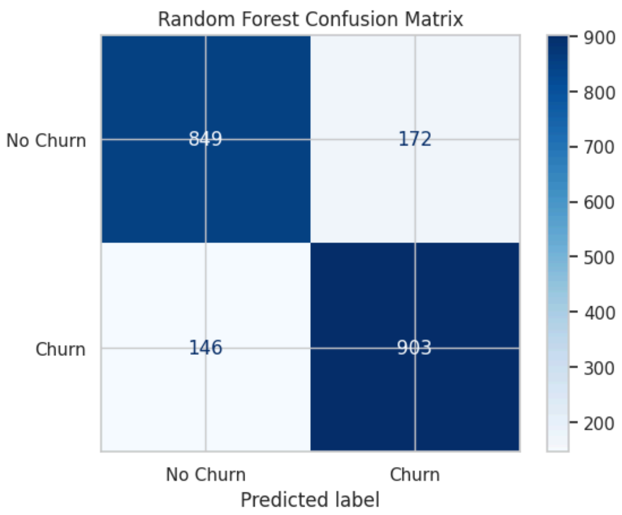


Figure 10. Random Forest Confusion Matrix

XGBoost, configured with a learning rate of 0.1 and max depth of 6, achieved 84% accuracy and an F1-score of 0.84. It captured complex relationships effectively but required higher computational resources. CatBoost, using 800 iterations, a depth of 6, and a learning rate of 0.05, delivered an accuracy of 84% and an F1-score of 0.85. It performed slightly better in reducing false negatives compared to XGBoost, particularly in datasets with significant categorical variables.

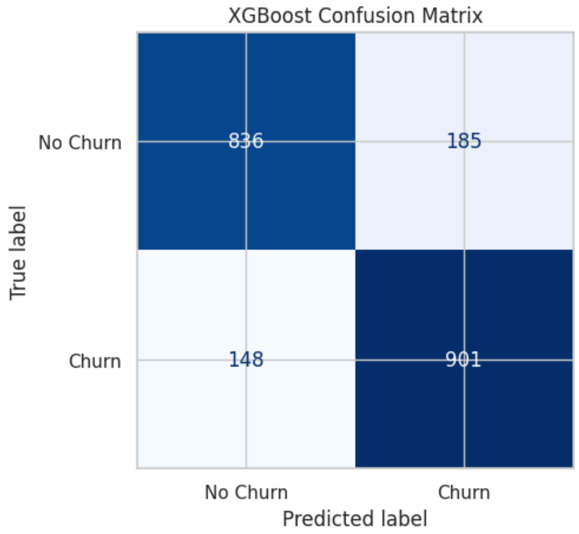


Figure 11. XGBoost Confusion Matrix

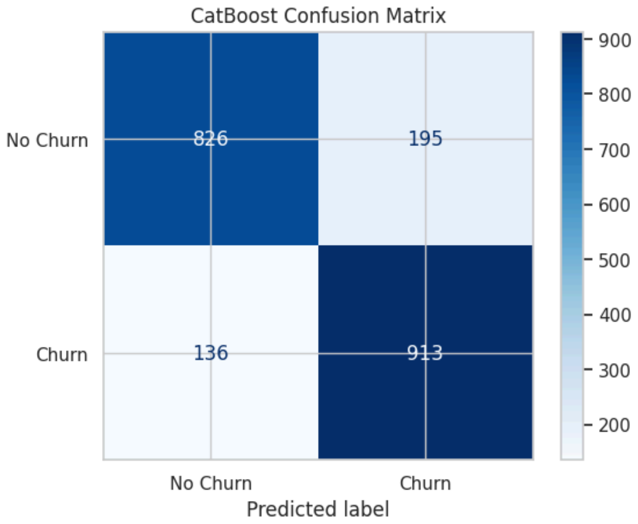


Figure 12. CatBoost Confusion Matrix

LightGBM, with 63 leaves and a learning rate of 0.05, matched CatBoost with an accuracy of 84% and an F1-score of 0.84. It demonstrated excellent computational efficiency and scalability while maintaining competitive performance metrics.

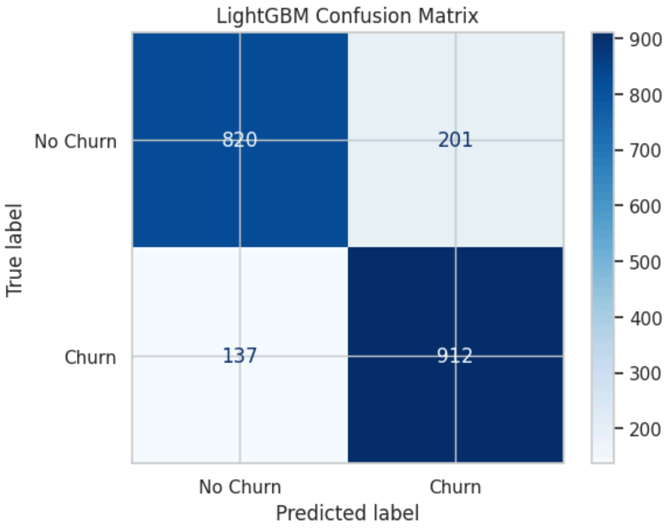


Figure 13. LightGBM Confusion Matrix

3.4 Overfitting Assessment and Visualizations

Training and testing accuracies for all models differed by no more than 2-3%, confirming minimal overfitting. For instance, Random Forest exhibited a training accuracy of 86% and a testing accuracy of 85%, while LightGBM showed a training accuracy of 85% and a testing accuracy of 84%. The consistent performance across training and testing datasets, coupled with balanced confusion matrices, indicates that the models are well-generalized and robust.

3.5 Recommendation

Based on the evaluation, **Random Forest** and **LightGBM** are recommended for deployment in customer churn prediction. Random Forest offers robust performance with strong interpretability, making it valuable for actionable insights into churn drivers. LightGBM excels in computational efficiency and scalability, ideal for large datasets and real-time applications. Both models demonstrated minimal overfitting and strong performance metrics, ensuring reliability and practical applicability in the telecom industry. These models are well-equipped to balance precision and recall, effectively mitigating churn while optimizing resource allocation.

IV. Findings and Implications

4.1 SMOTE

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.

2. 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 **85%**, with a balanced F1-score for both churned and non-churned customers.
- CatBoost and XGBoost also displayed improvements in classifying churned customers effectively.

4.2 Finding Important Features

1. Decision Classification

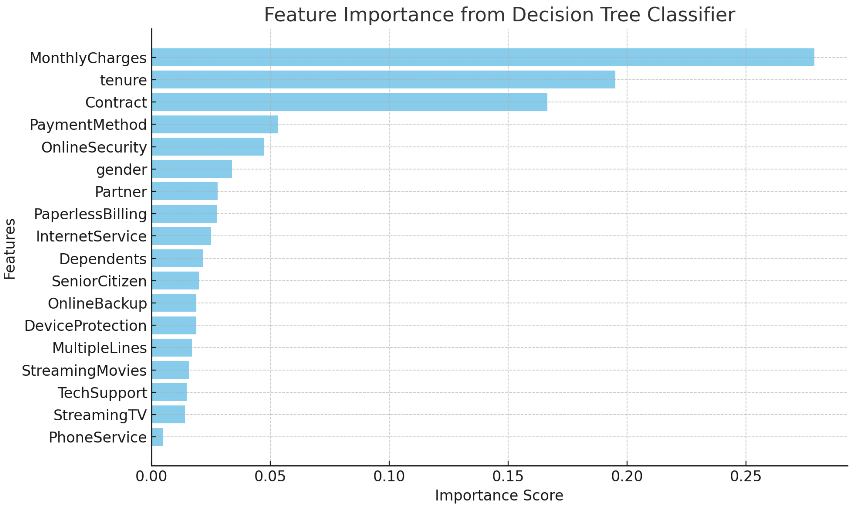


Figure 14. Decision Tree Classifier

After conducting a feature importance analysis using a Decision Tree Classifier, several key factors influencing customer churn were identified. The most significant feature was **MonthlyCharges**, which highlights that customers with higher monthly charges are more likely to churn. This aligns with general observations in the telecom industry, where customers with higher financial commitments may seek better deals elsewhere if they perceive the value of the service does not justify the cost.

The second most influential feature was **tenure**, indicating that customers with shorter durations of service (0-6 months) are at a higher risk of churning. This underscores the critical importance of focusing retention efforts on newly acquired customers by providing attractive offers or enhancing customer experience early in their journey.

Contract type also emerged as a crucial determinant. Customers with month-to-month contracts showed a significantly higher likelihood of churning compared to those with longer-term commitments. This suggests that encouraging customers to switch to yearly or bi-yearly contracts could mitigate churn risks and promote customer loyalty.

The **PaymentMethod** feature was another noteworthy factor, with electronic check payments being associated with higher churn rates. This could indicate that customers using electronic checks might find the payment process less convenient or have a lower engagement with the company's services. Addressing these potential pain points by promoting more seamless payment options, such as autopay or digital wallets, may improve retention.

Lastly, **OnlineSecurity** services appeared as a moderately important feature. Customers without additional online security services were more likely to churn, pointing to the potential for upselling or bundling these services to enhance customer satisfaction and retention.

2. Logistic Regression

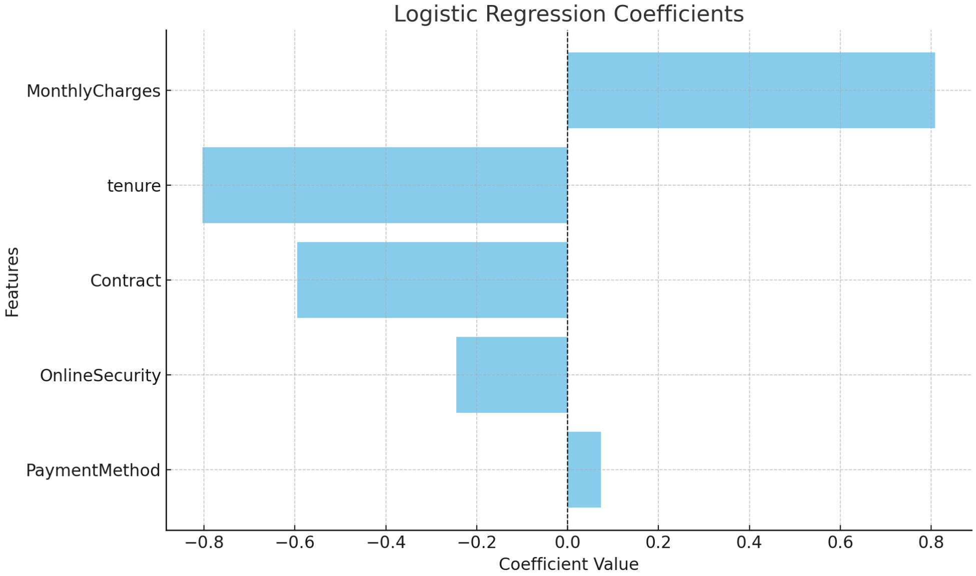


Figure 15. Logistic Regression Coefficients

The logistic regression analysis on the top features identified by the Decision Tree provides a clearer picture of their influence on customer churn. Scaling ensures that the coefficients represent each feature's standardized impact, offering more accurate interpretations.

With the highest positive coefficient of **0.808**, **MonthlyCharges** emerges as a key predictor of customer churn. This indicates that customers with higher monthly charges are significantly more likely to churn. The positive relationship highlights the importance of addressing pricing concerns, as higher charges may lead to dissatisfaction or a perception of reduced value.

Tenure has a strong negative coefficient of **-0.803**, reinforcing that longer-tenured customers are less likely to churn. This aligns with common retention patterns, where established customers exhibit higher loyalty. However, new customers with short tenures remain at the highest risk of churn, emphasizing the need for proactive engagement strategies during the critical early months of a customer's lifecycle.

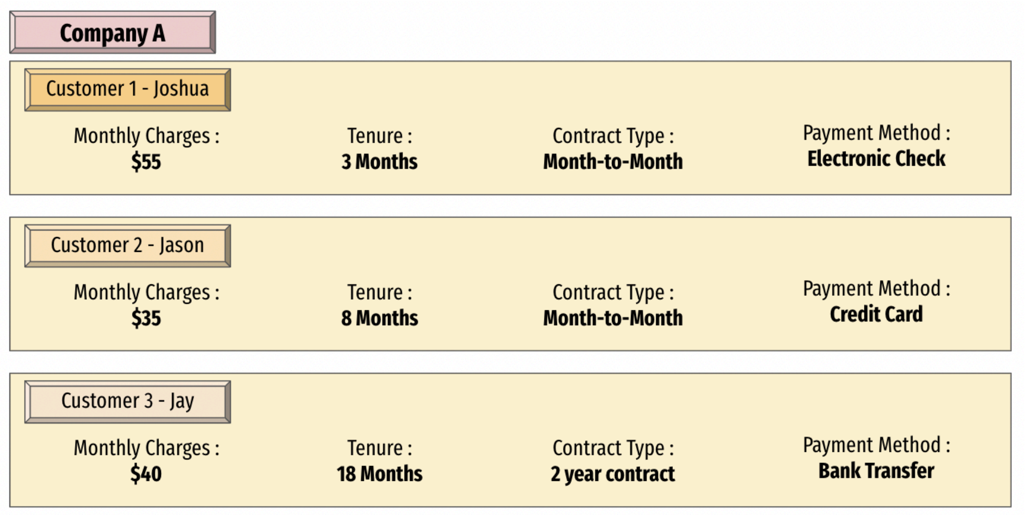
The feature **Contract** also has a notable negative coefficient of **-0.595**, signifying that customers on longer-term contracts (e.g., yearly or bi-yearly plans) are less likely to churn compared to those on month-to-month plans. This highlights the stabilizing effect of long-term agreements.

With a negative coefficient of **-0.244**, **OnlineSecurity** demonstrates that customers who subscribe to additional security services are less likely to churn. This finding underscores the importance of optional, value-added services in increasing perceived value and customer loyalty.

The feature **PaymentMethod** has a smaller positive coefficient of **0.073**, suggesting that certain payment methods, such as electronic checks, are slightly associated with higher churn. While the effect is less pronounced than other features, it indicates potential dissatisfaction or disengagement with this payment method.

4.3 Example Case

We tested our model by applying example cases.



Company A

Customer 1 - Joshua

Monthly Charges : **$55**
Tenure : **3 Months**
Contract Type : **Month-to-Month**
Payment Method : **Electronic Check**

Customer 2 - Jason

Monthly Charges : **$35**
Tenure : **8 Months**
Contract Type : **Month-to-Month**
Payment Method : **Credit Card**

Customer 3 - Jay

Monthly Charges : **$40**
Tenure : **18 Months**
Contract Type : **2 year contract**
Payment Method : **Bank Transfer**

Figure 16. Example Case

In this example case, we used a Random Forest classifier to predict the churn risk of three customers based on their monthly charges, tenure, payment method, and contract type. The case study was done by the Random Forest model providing probability estimates of customers churn. For example, Joshua, with a month-to-month contract, electronic check payment, short tenure (3 months), and higher monthly charges ($55), has a 75% probability to churn. Jason, on the other hand, has a 58% chance of being classified as churn, while Jay, who has a two-year contract, stable payment method, longer tenure (18 months), and moderate charges ($40), shows a 15% probability to churn. This aligns with our findings that customers with month-to-month contracts who pay with electronic checks are more likely to churn. These results, derived using Random Forest, offer valuable insights into customer churn risk based on individual characteristics.

V. Conclusion

5.1 General limitations of your data

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

5.2 Areas for Improvement

1. **Increase focus on Customer Engagement for Month-to-Month Contracts**

Customers on month-to-month contracts are more likely to churn (~89%). To reduce this, increase engagement through personalized offers, loyalty rewards, and proactive outreach. Offering incentives to switch to longer-term contracts, such as discounts or additional benefits, can further decrease churn among this segment.

2. **Increase targeted Retention Strategies for Electronic Check Users**

Customers who pay with electronic checks are more likely to churn (~57%). To address this, provide alternative, convenient payment methods and offer flexible billing cycles. Additionally, implement targeted retention strategies such as dedicated customer support to improve satisfaction and loyalty, reducing churn in this segment.

3. **Increase focus on Monthly Charges and Tenure to Reduce Churn**

Customers with higher monthly charges are more likely to churn, so reducing charges or offering more value through bundled services can help retain them. Additionally, customers with shorter tenures are at a higher risk of churn. To address this, focus on increasing engagement with new customers through personalized onboarding, loyalty programs, and proactive customer support. Long-term customers are more likely to stay, so incentivizing them with rewards for their continued business or offering long-term contract options can further reduce churn in this segment.

4. **Up-to-date Telecom Dataset**

Another way is to regularly update the dataset by adding new data from recent interactions and changes in how customers use services. This ensures the model remains relevant and accurately reflects current customer behavior, improving prediction accuracy over time.

5. **Integrating External Datasets**

Incorporating external datasets, such as regional data or industry trends, will help improve the model's robustness. This broader context will enhance the prediction of customer behavior and make the retention strategies more effective.

6. **Refine handling of class imbalance**

Moreover, If we are 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.

7. **Improvement in real-world**

To apply this model in the real world, we would focus on keeping the model up to **date** by **regularly adding new data**. Customer behavior changes over time, so by using real-time information like recent interactions or changes in how customers use services, the model would stay relevant. We would also set up a system to automatically retrain the model whenever needed, so it can adjust to any shifts in customer behavior.

5.3 Answer to Your Problem Statement

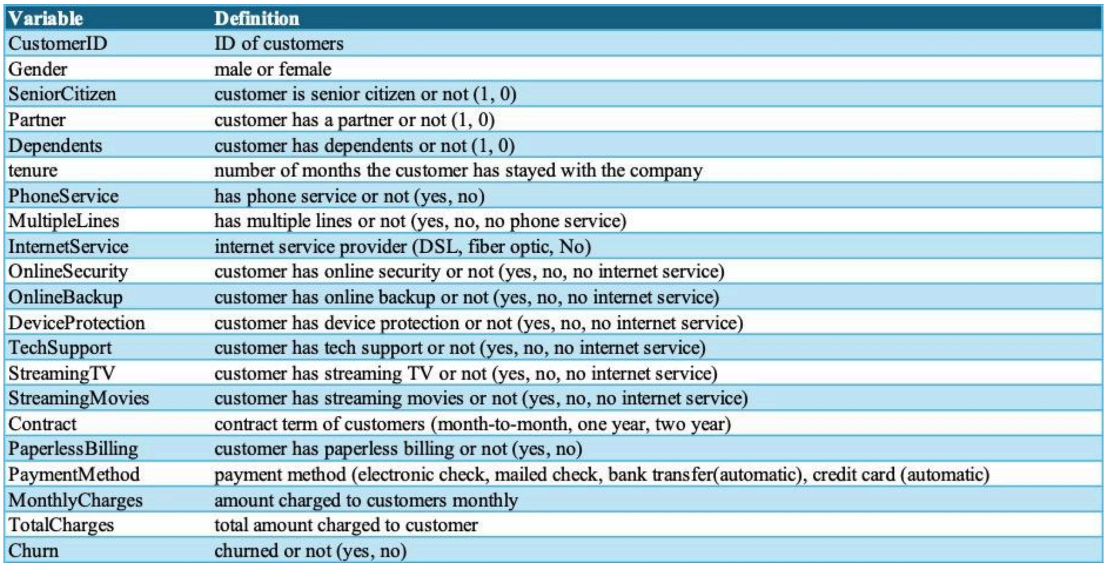
Our analysis successfully identified the key factors influencing customer churn, including payment methods, contract types, and customer tenure. Through our models, we were able to predict churn with reasonable accuracy, providing actionable insights to telecom companies for targeted retention strategies.

5.4 Potential next steps

If we had the opportunity to continue this project, the next steps would involve **integrating external datasets to improve the model's robustness**. We would also look into further fine-tuning the models with more advanced techniques, test them in real-world scenarios, and **explore the possibility of using deep learning approaches for more complex patterns that may arise in larger datasets.**

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