Predicting Customers Churn in a Telecom Company

I. Introduction

In the competitive landscape of modern business, telecommunications industries are in constant pursuit of innovative strategies to sustain growth by retaining customers.

Telecom firms confront distinct challenges in managing customer churn. Amidst a saturated market, customers possess the freedom to switch providers, impacting revenue. Precise churn prediction is essential for devising proactive retention strategies and enhancing overall satisfaction.

This project leverages machine learning to predict churn in telecom firms. By assessing historical customer data, usage patterns, and service interactions, a predictive model identifies at-risk customers. This empowers proactive resolution of concerns and personalized retention tactics, ultimately curtailing churn. The objectives of this project are as follows:

i. Identifying key features behind churn in the telecommunications industry.

ii. Employing machine learning to predict churn based on historical data.

iii. Assisting telecoms in crafting data-driven retention and service strategies.

As the telecommunications sector evolves, harnessing machine learning's potential for churn prediction emerges as a strategic advantage and necessity, enabling informed decision-making and enduring customer relationships. Through its focus on telecom churn, this project contributes to data-driven decision-making and underscores the importance of proactive customer management, offering telecommunication enterprises insights to comprehend and address the challenges of customer attrition.

II. Dataset Overview

Dataset was obtained by Kaggle originally obtained from IBM Sample Datasets as a fictional telecom company that provided home phone and Internet services in California in Q3. The dataset contains 33 attributes and 7043 observations. Attributes include:

- Customer with Geographic details
- Service Usage
- Contact and Bill
- Churn Indicators

Exploratory data analysis (EDA) was conducted to understand more about the problem.

It can be said From the Churn Distribution plot that provides valuable preliminary information about the customer base's churn behavior:

- Non-Churned Customers: Approximately 73.5% of customers fall into the category of "Non-Churned," indicating that the majority of customers have not terminated their services within the observed period.
- Churned Customers: About 26.5% of customers are categorized as "Churned," suggesting that a significant portion of customers have discontinued their subscription or services during the specified timeframe.

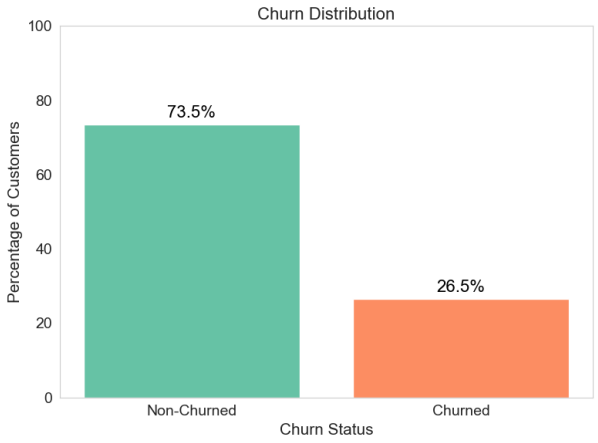

Figure 1. Churn Distribution

The boxplot of Tenure vs. Churn illustrates the relationship between customer tenure and churn behavior within the telecom company dataset. The visualization reveals key insights:

- 50% of the customers who left the service did so in the first 10 months.
- Longer-tenured customers are generally more likely to remain loyal to the company, while shorter-tenured customers are more prone to churning.

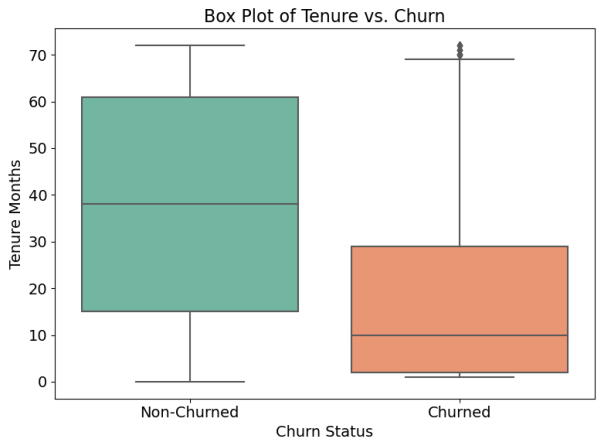

Figure 2. Boxplot of Tenure vs Churn

The distribution of churn reasons provides valuable insights into the factors influencing customer churn within the telecom company dataset.

From the churn reasons distribution, the following insights emerge:

- Diverse Churn Reasons: The diverse range of churn reasons reflects the multifaceted nature of customer attrition. Factors span from service quality and pricing concerns to interactions with customer support and offers by competitors.
- Top Churn Reasons: Some prominent churn reasons include "Attitude of support person," "Competitor offered higher download speeds," and "Competitor offered more data." These reasons collectively account for a significant portion of churn, highlighting the importance of addressing service quality, competitive offerings, and customer support experiences.

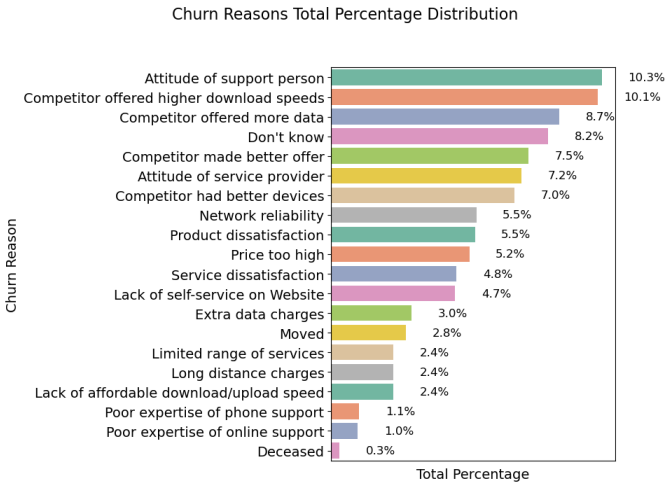

Figure 3. Churn Reasons Distribution by Percentage

The churn rate by contract type plot offers compelling insights into the dynamics of customer attrition based on the duration of contractual agreements within the telecom company dataset. Key observations emerge from the presented table:

- Notably, "Month-to-month" contracts exhibit the highest churn rate at approximately 42.71%
- Longer contract terms, such as "One year" and "Two year," are associated with notably lower churn rates compared to the more flexible "Month-to-month" contracts.

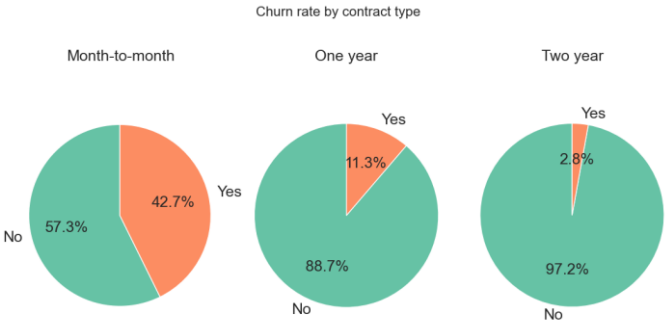

Figure 4. Churn Rate by Contract Type

III. Problem Definition

In the context of our project, the central goal revolves around binary classification—a fundamental aspect of supervised learning, where labeled data guides our model to accurately predict outcomes. The primary aim is to categorize customer behavior, offering insights into their actions and motivations. Customer attrition is a multifaceted phenomenon influenced by a multitude of factors, such as service quality, customer interactions, contract terms, and pricing concerns.

Given this intricate nature of customer churn, a multivariate analysis approach proves to be the most suitable strategy for addressing this problem effectively. The decision to adopt a multivariate analysis approach stems from the need to comprehensively address the complexities of customer churn. By considering the collective impact of various features and their interactions, we aim to achieve a more accurate and insightful predictive model that captures the intricate realm of churn behavior.

Data Preparation for Prediction

In the process of preparing our dataset for accurate customer churn prediction within the telecom industry, several vital steps were undertaken. These actions ensure that our data is ready for modeling, addresses class imbalance, and leverages a powerful technique to enhance our predictions:

- Converting Categorical Variables: A key initial step involved transforming categorical variables, such as the 'Churn Label' feature, into numeric values.

- Addressing Data Imbalance: Observing the class distribution of the 'Churn Label,' it became evident that the dataset was unbalanced. A higher count of non-churning customers (0) existed compared to those who churned (1). This imbalance could lead to biased model performance and inaccurate predictions, particularly in scenarios where the minority class (Churn) is of interest, we proceeded to rectify this issue.
- Utilizing Synthetic Minority Oversampling Technique (SMOTE): To ensure our dataset is balanced and representative, we opted for SMOTE – a technique adept at oversampling the minority class.
- Adhering to recommended practices, we split the dataset into distinct training and test sets, employing SMOTE exclusively on the training data. This approach augments the model's proficiency in adapting to new, production-like data, all while preserving the integrity of the test dataset. Through this deliberate strategy, our models are better poised to effectively discern both churning and non-churning behaviors.

Machine Learning Algorithm selection

Our primary goal centers around binary classification. An intriguing aspect of our approach is the utilization of multivariate analysis, a method that keenly examines the interplay among different features. This approach enriches our understanding of the intricate dynamics behind customer attrition.

Selected Models:

- Logistic Regression: Logistic Regression stands out as a fundamental and interpretable model, serving as an ideal baseline for binary classification tasks such as churn prediction. Its strength lies in scenarios where the relationship between features and the target demonstrates a degree of linearity, providing valuable insights into feature significance.
- RF**:** Random Forest is an ensemble method that combines multiple decision trees to improve predictive accuracy and handle non-linear relationships. It's robust and can capture interactions between features, making it suitable for more complex datasets.
- Gradient Boosting: Gradient Boosting algorithms are powerful ensemble methods that build decision trees sequentially, focusing on correcting the errors made by previous trees. They can capture complex patterns in the data and often result in high predictive performance.
- SVM: It is effective for binary classification tasks, especially when the decision boundary is not necessarily linear. It can map the data into higher-dimensional space to find a separating hyperplane.

In conclusion, the models we've chosen – Logistic Regression, RF, Gradient Boosting, and SVM– are well-suited to handle the intricate nature of predicting customer churn. Each of these models has unique strengths that help us deal with complexities and gain valuable insights from our dataset.

Not Selected Models:

- KNN: This model's effectiveness diminishes in high-dimensional datasets with complex relationships, such as customer churn prediction. The "curse of dimensionality" hampers its accuracy, while imbalanced class distributions and the challenge of selecting appropriate distance metrics further limit its performance.
- Decision Trees: Though intuitive, Decision Trees struggle with overfitting noisy data, hindering generalization to new instances like churn behavior. Their inability to capture intricate relationships and handle imbalanced data restricts their suitability for the complexities of customer churn prediction.

In summary, KNN and Decision Trees, though having potential in certain contexts, are not well-suited for our predictive model in the context of customer churn prediction. Their limitations in capturing nuanced relationships and addressing class imbalances render them less suitable for the intricate nature of customer churn prediction.

Evaluation Methodology

For the evaluation methodology, we employed essential performance metrics to assess the effectiveness of our models. These metrics encompass:

- Accuracy: This measures the overall correctness of predictions.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Predictions}$$

- Precision: This measure emphasizes on minimizing false positive predictions.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

- Recall (Sensitivity): This measure focuses on capturing true positive predictions.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

- F1-Score: The F1-score harmonizes precision and recall, providing a balanced metric that accounts for both false positives and false negatives.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

IV. Analysis and Evaluation

A comprehensive analysis and evaluation of the four chosen models were conducted to gauge their predictive prowess in the intricate landscape of customer churn prediction. The models—Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machine—underwent rigorous scrutiny using essential performance metrics, namely Precision, Recall, and F1-Score.

```
Logistic Regression Accuracy: 0.7572746628814763

Logistic Regression Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.74      0.81      1009
           1       0.55      0.80      0.65       400

    accuracy                           0.76      1409
   macro avg       0.73      0.77      0.73      1409
weighted avg       0.80      0.76      0.77      1409
```

Figure 5. Performance Report of Logistic Regression

```
Random Forest Accuracy: 0.7963094393186657

Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.89      0.86      1009
           1       0.66      0.57      0.61       400

    accuracy                           0.80      1409
   macro avg       0.75      0.73      0.74      1409
weighted avg       0.79      0.80      0.79      1409
```

Figure 6. Performance Report of Random Forest

```
SVM Accuracy: 0.7849538679914834

SVM Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.82      0.85      1009
           1       0.61      0.70      0.65       400

    accuracy                           0.78      1409
   macro avg       0.74      0.76      0.75      1409
weighted avg       0.80      0.78      0.79      1409
```

Figure 7. Performance Report of SVM

```
Gradient Boosting Accuracy: 0.8076650106458482

Gradient Boosting Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.86      0.87      1009
           1       0.66      0.67      0.67       400

    accuracy                           0.81      1409
   macro avg       0.76      0.77      0.77      1409
weighted avg       0.81      0.81      0.81      1409
```

Figure 8. Performance Report of Gradient Boosting

Cross Validation

Having meticulously assessed the predictive performance of our selected machine learning models – LR, RF, GB, and SVM – through a comprehensive evaluation of various metrics including accuracy, precision, recall, and F1-score, we recognize the significance of further validating their robustness. To accomplish this, we have chosen to employ a well-regarded technique known as cross-validation.

To ensure a thorough assessment of each model's performance, we adopted a 5-fold cross-validation approach.

Table 1. Summary of Cross Validation Performance Metrics

Cross Validation Performance Metrics				
Model Selected	**Mean Accuracy**	**Mean Precision**	**Mean Recall**	**Mean F1**
Logistic Regression (LR)	0.8106	0.6600	0.5630	0.6073
Random Forest (RF)	0.7916	0.6310	0.4833	0.5472
Support Vector Machine (SVM)	0.8043	0.6690	0.4929	0.5674
Gradient Boosting (GB)	0.8078	0.6627	0.5344	0.5916

Model's Selection Discussion

The evaluation of Cross Validation Performance Metrics for four predictive models, reveals distinct characteristics in their performance:

- Logistic Regression (LR): Demonstrating a robust mean accuracy of 0.81 and a balanced F1-score of 0.61,

LR showcases remarkable precision, indicating its ability to minimize false positives. However, its recall score of 0.56 suggests a scope for enhancing its capability to identify all genuine churn cases.

- Random Forest (RF): RF maintains consistent performance with a mean accuracy of 0.79. Its mean precision of 0.63 underscores its strength in mitigating false positives. However, the model's recall score of 0.48 highlights limitations in capturing all true churn cases. The F1-score of 0.54 emphasizes the trade-off between precision and recall.
- Support Vector Machine (SVM): SVM attains a mean accuracy of 0.80, aligning it with the other models. Notably, its mean precision of 0.67 highlights its capability in minimizing false positives. Yet, the recall score at 0.49 suggests the need for improvement in capturing true churn cases. The model achieves an F1-score of 0.57, reflecting a moderate balance.
- Gradient Boosting (GB): GB stands out with a mean accuracy of 0.81, matching the highest accuracy achieved. Its mean precision of 0.66 is in line with LR and SVM. The recall value of 0.54 demonstrates its proficiency in identifying true churn cases. The balanced F1-score of 0.59 signifies its harmonized precision-recall trade-off.

Considering these findings, both LR and GB emerge as the most promising contenders. While LR excels in precision, GB's balanced approach aligns well with the telecom company's objective of capturing genuine churn instances while minimizing false positives.

ROC curve

The subsequent step of constructing ROC curves and calculating the AUC values contributes another layer of understanding to our model evaluation process.

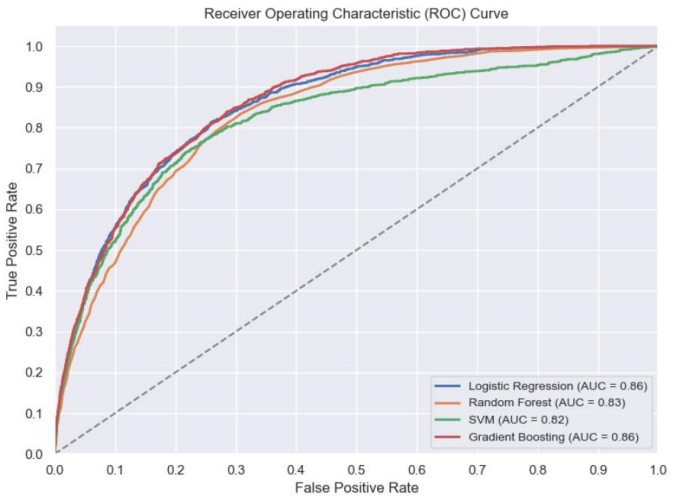

Figure 9. ROC curve

Notably, both LR and GB exhibit high AUC values of 0.86, underlining their robust discrimination power. These values complement their strong performance as identified through cross-validation.

Considering the telecom company's focus on precision, recall, and the objective of reducing false positives while detecting actual churn instances, the decision between LR and GB should reflect the company's priorities. LR's notable precision makes it a strong contender for retaining valuable customers and minimizing incorrect alerts. On the other hand, GB's balanced approach aligns well with the aim of accurately pinpointing churn cases while maintaining a reasonable level of false positives.

V. Conclusions

In this project, we have undertaken a comprehensive exploration of predicting customer churn within the telecommunications industry through the application of machine learning techniques.

By analyzing historical customer data, service interactions, and usage patterns, our predictive model enables proactive measures for customer retention. Our meticulous data preparation, SMOTE application solely to the training set, and careful algorithm selection, including Logistic Regression, Random Forest, Gradient Boosting, and SVM, reflect a holistic approach.

Models were subjected to rigorous evaluation metrics encompassing accuracy, precision, recall, and F1-score. Our commitment to robustness is bolstered by 5-fold cross-validation, enhancing the credibility of our results. Ultimately, our evaluation showcased the prowess of both Gradient Boosting and Logistic Regression. While Gradient Boosting displayed balanced performance and an ensemble-based approach, Logistic Regression stood out for its remarkable precision.

References

[1] Geiler, L., Affeldt, S. and Nadif, M. (2022) An effective strategy for churn prediction and customer profiling. Data & Knowledge Engineering [online]. 142, p. 102100. Available from: https://www.sciencedirect.com/science/article/pii/S0169023X2200091X [Accessed 16 August 2023].

[2] Kavitha, V., Kumar, G., Kumar, S. and Harish, M. (2020) Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms. International Journal of Engineering Research and [online]. V9.

[3] Telco customer churn: IBM dataset. (no date) [online]. Available from: https://www.kaggle.com/datasets/yeanzc/telco-customer-churn-ibm-dataset [Accessed 16 August 2023].

[4] Hadden, J., Tiwari, A., Roy, R. and Ruta, D. (2006) Churn Prediction: Does Technology Matter. International Journal of Intelligent Technology. 1, pp. 104–110.

[5] Ahmad, A.K., Jafar, A. and Aljoumaa, K. (2019) Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data [online]. 6 (1), p. 28. Available from: https://doi.org/10.1186/s40537-019-0191-6 [Accessed 16 August 2023].

[6] Huang, B., Kechadi, M.T. and Buckley, B. (2012) Customer churn prediction in telecommunications. Expert Systems with Applications [online]. 39 (1), pp. 1414–1425. Available from: https://www.sciencedirect.com/science/article/pii/S0957417411011353 [Accessed 16 August 2023].

[7] Mustafa, N., Lew, L. and Abdul Razak, S.F. (2021) Customer churn prediction for telecommunication industry: A Malaysian Case Study. F1000Research [online]. 10, p. 1274.

[8] Shumaly, S., Neysaryan, P. and Guo, Y. (2020) Handling Class Imbalance in Customer Churn Prediction in Telecom Sector Using Sampling Techniques, Bagging and Boosting Trees. In: 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE) [online]2020 10th International Conference on Computer and Knowledge Engineering (ICCKE). pp. 082–087.

[9] Preetha, S. and Rayapeddi, R. (2018) Predicting Customer Churn in the Telecom Industry Using Data Analytics. In: 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT) [online]2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT). pp. 38–43.

[10] Reducing churn in telecom through advanced analytics | McKinsey. (no date) [online]. Available from: https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/reducing-churn-in-telecom-through-advanced-analytics [Accessed 16 August 2023].

[11] Eria, K. and Poolan Marikannan, B. (2018) Systematic Review of Customer Churn Prediction in the Telecom Sector. 2, pp. 7–14.