Predicting and Preventing Churn: An Analytical Approach to Customer Retention for Telco

Jiho Kil, Connor Redding, Xinhui(Mary)XU

December 13, 2024

Abstract

Customer churn is a phenomenon where customers discontinue their relationship with a company or service, and is a critical metric for understanding customer retention and business sustainability. Telco, a fictional company providing home phone and internet services in California, has provided a comprehensive dataset which includes customer demographics, account details, service information, and historical churn data. By leveraging this data, we plan to enhance Telco's competitive edge through strategic, data-driven customer retention initiatives.

We identified 4 unique customer clusters through clustering analysis, and further flagged 2 of those 4 as significantly more likely to churn. For predictive modeling, we employed Vanilla Logistic Regression, LASSO Logistic Regression, CART, and XGBoost algorithms to optimize predictive accuracy and AUC above 95% in both training and testing. For the financial implications of the model we evaluated the financial implications of false positives and negatives. Based on the predictive models and financial analysis, we propose 5 actionable strategies to improve customer retention for Telco.

Contents

1 Introduction

In today's telecommunications industry, customer churn—the phenomenon where customers discontinue their services—represents a critical challenge that directly impacts business sustainability and growth. The cost of acquiring new customers typically far exceeds that of retaining existing ones, making effective churn prediction and prevention essential for service providers[Gre93]. Modern data analytics and machine learning techniques offer promising opportunities to address this challenge through proactive, data-driven approaches.

This paper presents a comprehensive analysis of customer churn for Telco, a California-based provider of home phone and internet services. Through analysis of their customer dataset, which includes demographic information, account details, service usage patterns, and historical churn data, we develop and evaluate sophisticated predictive models and concrete retention strategies.

Our analysis follows a systematic approach, beginning with thorough data preprocessing to address missing values and ensure data quality. Through exploratory analysis, we identify key patterns in customer behavior and characteristics, leading to the discovery of four distinct customer clusters through clustering analysis. Notably, two of these clusters demonstrated significantly higher churn propensity, providing crucial insights for targeted interventions.

The core of our research employs multiple predictive modeling techniques, comparing the effectiveness of Vanilla Logistic Regression, LASSO Logistic Regression, CART, and XGBoost algorithms. These models achieve impressive predictive accuracy and AUC scores exceeding 95% in both training and testing phases. Our modeling approach extends beyond mere prediction accuracy to consider the financial implications of false positives and negatives, ensuring practical utility of our findings.

Building on these analytical insights, we conduct a granular churn analysis that informs five key strategic recommendations: improving customer satisfaction metrics, developing competitive service offerings, creating incentives for longer-term contracts, enhancing online security services, and developing specialized services for senior citizens. Each strategy is directly grounded in our data analysis and designed to address specific patterns identified in our customer clusters and predictive models.

This paper provides a detailed examination of our methodology, findings, and recommendations, offering a blueprint for data-driven customer retention initiatives. We conclude with a thorough discussion of our models' effectiveness and limitations, providing insights for future research and practical implementation of our proposed strategies.

2 Data Inspection

We first inspected the four separate data sheets in Table1 to identify the structure , data types, and any inconsistencies that could impact merging and analysis. The number of rows is consistent among four datasets. So we proceed to

merging.

Name	Row	Column
Demographics.csv	7043	15
Location.csv	7043	10
Services.csv	7043	31
Status.csv	7043	15

Table 1: Structure of the 4 separate datasets

2.1 Data Merging

The four datasets were merged to form a dataset that includes all relevant information. Merging involves aligning four data sheets based on the key identifier - Customer ID to ensure that the relationships between data points are maintained. The 'Count' and 'Quarter' columns were inspected to verify the accuracy of merging, then the redundant columns that appear more than once ere dropped. We also confirmed that every row refers to one unique customer, preventing any potential mistakes resulting from merging the datasets.

2.2 Address Missing Values

In the merged dataset, 4 fields contain missing values: offer, internet_type, churn_category and churn reason, as shown in Table2. According to the data dictionary in Appendix7.3, the missing values in 'offer' column indicate that the customer did not accept any offer in the five categories. Therefore, we replaced the missing values with the string 'No offer'. Similarly, we replaced the missing values in 'internet_type' with the string 'No Internet'. For churn_category and churn_reason, after careful inspection, we discovered that the missing values resulted from those who did not churn, therefore we replaced the missing values with the string 'No churn'.

	Missing Values	Percentage missing
offer	3877	55.047565
internet-type	1526	21.666903
churn-category	5174	73.463013
churn reason	5174	73.463013

Table 2: Missing Value Analysis

3 Exploratory Analysis

After data preprocessing, we proceeded to conduct exploratory data analysis to understand the characteristics and patterns in the dataset. These are key steps to explore and prepare the data for predictive modeling. The cluster analysis also provides insights for strategy making.

3.1 Descriptive Analysis

In the descriptive analysis, means of each column, gender composition by churn, age distribution by churn, satisfaction rate by churn, count by reason with satisfaction score label, count by churn category with satisfaction score label were calculated. Following are insights generated from the analysis:

- No clear patterns discovered in turns of churn status by gender, Fig2.
- The churn rate is noticeably higher in senior population than in younger population, Fig1.
- 3 is the threshold for churn in terms of satisfaction rate, Fig3.
- The main reason for customer churn is competition. Customers have noted that competitors provided better devices, made more attractive offers, and offered more data, Fig4.
- The attitude of the support person is also a key factor contributing to customer churn, Fig4.

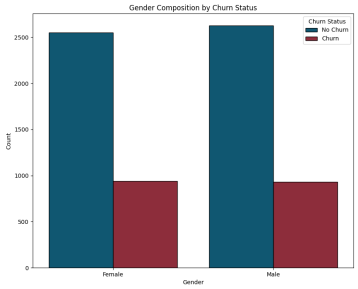


Figure 1: Age Composition by Churn Status

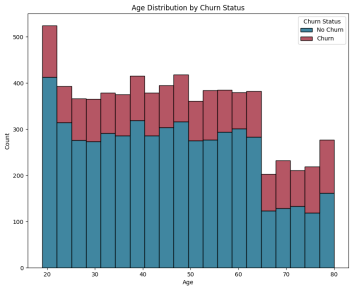


Figure 2: Gender Composition by Churn Status

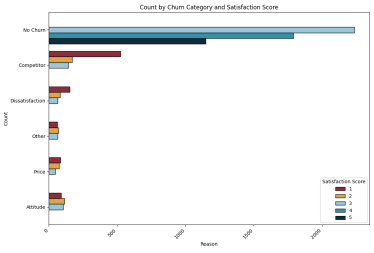


Figure 3: Count by reason with satisfaction score label

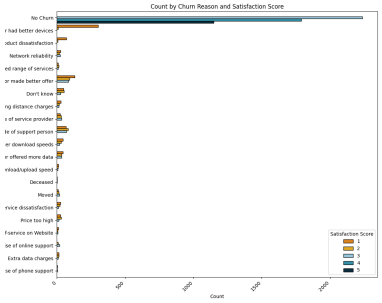


Figure 4: Count by reason and Satisfaction Score label

3.2 Correlation and feature selection

After addressing all binary fields, we computed correlation coefficients to explore relations between variables. According to the correlation heat-map(Fig5), there exist 6 pairs of highly correlated variables.

1. 'Referred_a_friend' and 'married': It's quite surprising to see that one demographic factor and one behavioral factor can be highly related. We chose to focus on 'Referred_a_friend' because we believed it could offer higher strategic value.

2. 'total_charges' and 'total_revenue': With a coefficient of 0.97, they are highly correlated in describing the cash flow from a customer. Since the analysis focuses on customer behavior, we will concentrate exclusively on 'total_charges' in the modeling process.

3. 'monthly_charge' and 'tenure_in_months' and 'total_charges': The product of monthly_charge and tenure is not exactly equal to total-charge, but very similar, with a correlation of 0.99. This is likely because monthly_charge reflects the 'current' monthly charge, while total_charge accounts for variations over time. Therefore, we will remove 'total-charges' from our analysis, as it is derived from the other two variables. 'Monthly-charges' and 'tenure-in-months' offer more actionable insights into customer behavior.

4. 'dependents' and 'number_of_dependents': These two variables are also highly correlated. We chose to keep 'number_of_dependents' as it provides more granular information, including the presence of dependents indicated by the binary variable 'dependents'.

5. 'satisfaction_score' and 'satisfaction_score_label': These are almost identical fields, so we removed 'satisfaction_score_label' and use 'satisfaction_score'.

6. 'churn-label' and 'churn-value': These are almost identical fields, so we removed 'churn_label'.

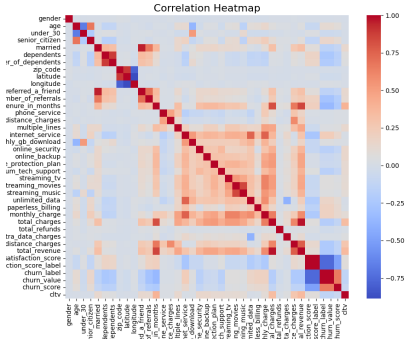


Figure 5: Heatmap

This analysis also helped identify potential predictors and dependencies, preventing the potential issues of multicollinearity and redundancy in terms of interpretation.

3.3 Cluster Analysis

To enhance clustering and prediction outcomes, we omitted fields deemed less relevant, such as 'customer_id', 'location_id', and 'count'. The 'country' and 'state' variables were also excluded since they only contained a single value (US and California, respectively). Additionally, we removed the 'city' variable, which had over 1,000 unique entries, potentially complicating the K-means distance-based clustering method. 'Customer_status' was omitted due to redundancy. We also excluded 'churn_score' and 'churn_score_category' because they relate to an independent prediction score by IBM's predictive tool. Lastly, 'cltv_category' was removed due to its similarity with 'cltv'.

We implemented K-means clustering for customer segmentation. Based on the Elbow method and Silhouette analysis in Fig6, we identified k=4 as the optimal number of clusters. Following are the interpretation of the four clusters, more information about clusters is in the Appendix7.3.

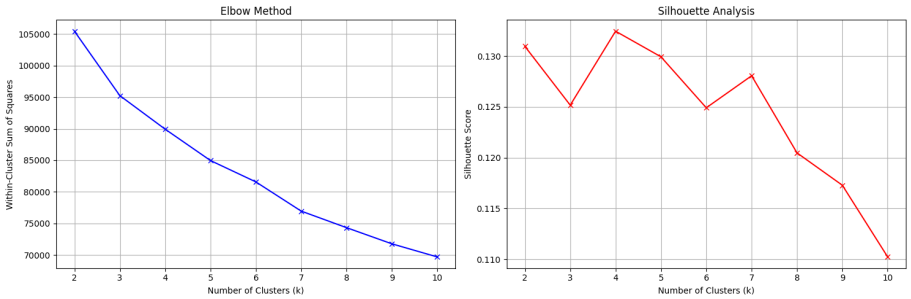


Figure 6: Number of Cluster Selection

Cluster 0: This cluster consists of 1,511 customers with a churn rate of 7.35%. Customers in this group are relatively young, with an average age of 43, and have been with the company for about 30 months, indicating they are established customers. Nearly all (99%) subscribe to phone services, with almost no internet services. Their most common plan is a two-year contract with credit card payments, and they have a high satisfaction score of 3.87. Overall, Cluster 0 represents satisfied, low-risk customers with long-term contracts.

Cluster 1: This cluster includes 2,093 customers with a churn rate of 11.56%. These customers are slightly older than those in Cluster 0, with an average age of 46, and possess the longest average tenure of around 56 months. They show high usage of additional features like streaming. Almost everyone in this cluster subscribes to internet services, distinguishing them from Cluster 0. They have moderate satisfaction rat of 3.45 and the highest value to

the company (5022.52). The common plan is a two-year contract with bank withdrawals. Cluster 1 consists of moderately stable, long-tenured customers subscribing to extensive services, including internet.

Cluster 2: Comprising 3,099 customers, this cluster has a high churn rate of 46.76%, primarily due to competitors offering better devices. Customers here are slightly older (48.18) and are newer, with an average tenure of 17 months. Similar to Cluster 1, they have high internet usage but low engagement in other services. Their satisfaction rate is 2.79, with comparatively lower value to the company. They prefer month-to-month contracts. In all, cluster 2 is a high-risk group, showing a preference for short-term contracts and less commitment, being prone to shifting to competitor offerings.

Cluster 3: The smallest cluster, with 340 customers, has the highest average age (48.68) and includes many senior citizens. The churn rate is 19.71%, with the primary reason being a "better offer" from competitors. Their most accepted offer is 'Offer B', common across all customers. They have high total refunds, possibly due to service issues, and moderate internet and phone usage. Their satisfaction rate is 3.40, with month-to-month contracts being most common. Cluster 3 has moderate risk, potentially needing more engagement or resolution of service issues.

We identified Cluster 2 as the high-risk group with the highest churn rate, and Cluster 3 as the moderate-risk group, which may require more attention be paid to improve their customer experience. Our retention strategy will primarily focus on these two groups.

4 Predictive Modeling

In this section, we implemented several predictive modeling methods to further understand the features of customers that are associated with churn.

4.1 Modeling Preparation

We began by splitting the data into training and testing sets, using an 80/20 split for both the unstandardized and standardized datasets. We also created several helper functions to streamline our modeling process, including functions for hyper-parameter tuning, model evaluation, and plotting confusion matrices and feature importance.

4.2 Vanilla Logistic Regression

We started with the basic Vanilla logistic regression, which provided insights into relationships between features through highly interpretable coefficient.

Based on the model shown in Fig7, the top feature influencing customer churn is the Satisfaction Score, which aligns with expectation and former analysis. Online_security and Number_of_Referrals are surprisingly also among the

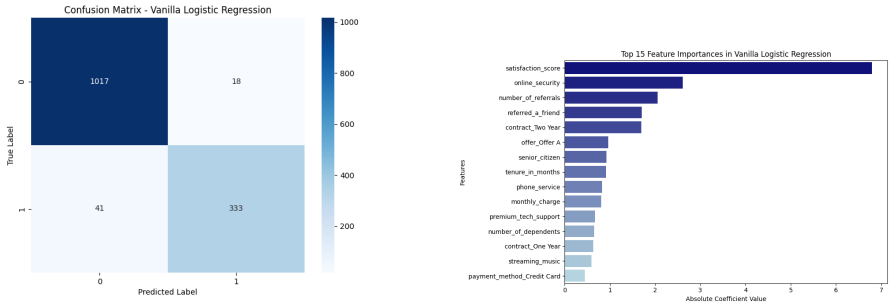


Figure 7: Confusion Matrix and Top Features in Vanilla Logistic Regression

top features, indicating their important influence on churn rate. The standardized coefficient of Online Security is -2.6, which indicates that those who are subscribed to the Online Security Service are less likely to churn. Through this, it can be assumed that they are quite satisfied with the Online Security Service. For model performance, the model exhibits strong performance in predicting customer churn with relatively high accuracy, precision and AUC values, along with quick training times.

	Accuracy	Precision	AUC
Train	0.9657	0.9540	0.9936
Test	0.9581	0.9487	0.9913

Threshold: 0.5 Train Time: 0.0450

Table 3: Vanilla Logistic Regression Performance Metrics

4.3 Penalized Logistic Regression

We then proceeded to implement Penalized Logistic Regression in an attempt to have the model generalize better to unseen data in terms of churn predictions.

The penalized logistics regression model shares the same top three important features as those in Vanilla Regression: Satisfaction Score, Online Security, and Number of Referrals. The standardized coefficient of Number_of_Referrals is -2.44, indicating that the more referrals a customer has, the less likely they are to churn. Also, longer contracts (2 years) decrease the likelihood of churn, while a customer under Offer A are more likely to churn. Compared with the vanilla model, the LASSO model has increased accuracy, potentially reducing overfitting and increasing model interpretability, as is shown in Table4.

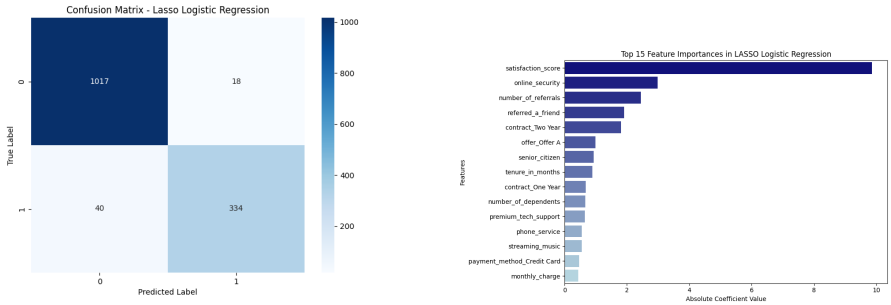


Figure 8: Confusion Matrix and Top Features in LASSO Logistic Regression

	Accuracy	Precision	AUC
Train	0.9656	0.9514	0.9937
Test	0.9588	0.9489	0.9915

Threshold: 0.5 Train Time: 1.581

Table 4: LASSO Logistic Regression Performance Metrics

One critical caveat with these two logistic regressions is that the coefficients for 'number_of_referrals' and 'referred_a_friend' point in opposite directions, making it challenging to determine the exact impact of the referral strategy on customer retention. This suggests that the potential complex relationship between referrals and churn risk is not perfectly captured by simple logistic regression approach, further steps will be discussed in the final conclusion. One possible explanation is that a customer with a high number of referrals is less likely to churn, likely reflecting genuine endorsement of the service. While in contrast, customers with few referrals might be more interested in one-time promo credits rather than true loyalty or endorsement of the service.

4.4 CART

The CART model offers a straightforward decision tree approach that excelled in capturing non-linear patterns, more detail in Appendix7.3. As in the logistic regression models, Satisfaction Score is the dominant factor in predicting churn. Online Security, Monthly Charge, Number_of_Referrals, and Tenure in Months also influence the model's decision. Compared with the regression models, CART is quick and quite precise, though it does show a slight drop in performance on test data compared to training data, which could suggest mild overfitting.

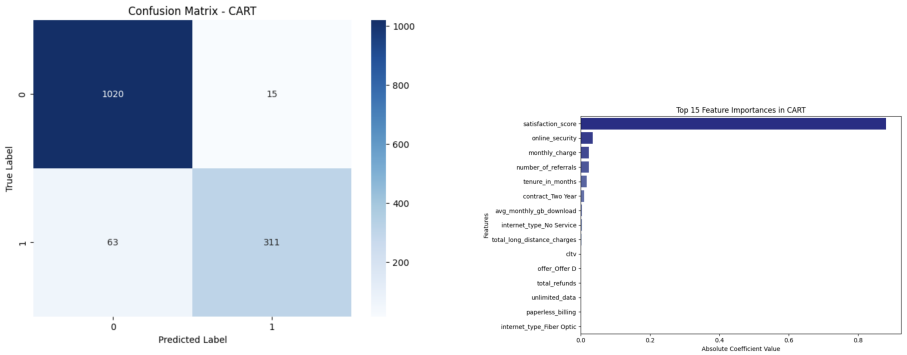


Figure 9: Confusion Matrix and Top Features in CART

	Accuracy	Precision	AUC
Train	0.9533	0.9632	0.9925
Test	0.9446	0.9540	0.9857

Threshold: 0.5 Train Time: 0.0534

Table 5: CART Performance Metrics

4.5 XGBoost

To maximize predictive performance, we implemented XGBoost, which is known to boast great predictive accuracy and robustness among machine learning models. We used GridSearchCV[ld23] which combines 5-fold cross validation and an exhaustive search over the combinations of the specified parameter values to find the best model.

In XGBoost, still, Satisfaction Score is the most influential feature. Other important features include: Internet Type (Fiber Optic), Online Security, Senior Citizen, Contract Type (Two Year). The feature Senior Citizen ranked the fourth in importance, indicating that senior customers are more likely to churn, which aligns with our findings in descriptive analysis. In terms of model performance, the XGBoost model exhibits exceptional predictive performance with high accuracy and precision. High AUC values shows its ability to effectively differentiate churn and non-churn customers. Overall, the model is robust and efficient.

	Accuracy	Precision	AUC
Train	0.9718	0.9771	0.9955
Test	0.9603	0.9676	0.9924

Threshold: 0.5 Train Time: 0.1624

Table 6: XGBoost Performance Metrics

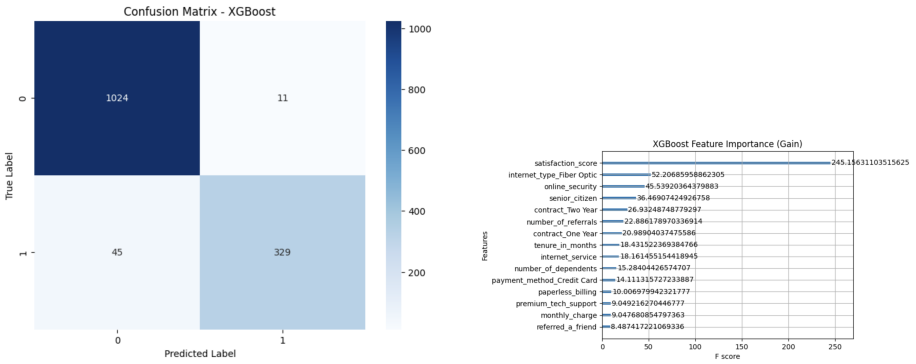


Figure 10: Confusion Matrix and Top Features in XGBoost

4.6 Summary

In general, all of the four models demonstrate strong performance. XGBoost outperforms other models in terms of accuracy, precision, and AUC, while maintaining efficient training time and manageable model complexity; Logistic Regression offers strong performance with simplicity; LASSO adds feature selection to reduce overfitting; the CART model offers great interpretability, accompanied by a slight compromise in model performance.

In terms of key predictors, satisfaction score is undoubtedly the primary driver of customer churn. This is followed by online security, number of referrals, two-year contracts, and senior citizen status, all of which provide valuable insights for strategy making.

5 Granular Churn Analysis

Then we proceeded to a granular analysis of customer churn, aiming to optimize customer retention strategy in the hypothetical case of promoting online security service.

5.1 Setting Description

Suppose that following predictive analysis, Telco conducted market research and is considering using promoted online security services as a strategic incentive to reduce churn and increase profit.

According to results in descriptive analysis(See in Appendix7.3), the mean of monthly charge is $65. Based on this figure, we assume that the monthly profit Telco gains per customer is $50, offering promotion on online security services reduces the profit to $40. It's also assumed that with the churn probability being p without promotion, the churn rate will reduce to 2/p with promotion on online security services. The Figure11 below demonstrates all possible cases and expected profit.

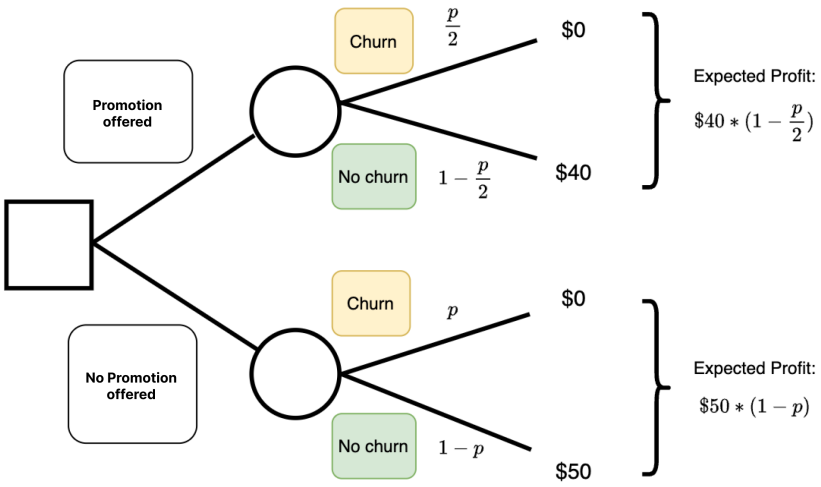


Figure 11: Expected Profit in all possible cases

We calculated the formula $\$40 \times (1 - \frac{2}{p}) > \$50 \times (1 - p)$ to ensure that the promotion is profitable per customer. Consequently, p must be greater than 0.333. This serves as the decision threshold to determine whether Telco should offer the promotion to a customer. We then applied this threshold to the previous model and observed the expected total profit for each case.

5.2 Loss Matrix

Our scenario can also be described by a specific loss matrix in Table7, where the loss for a False Positive (FP) is \$10 and the loss for a False Negative (FN) is \$20. For a False Positive, the loss is \$10 because Telco offers a promotion to a customer who is not going to churn, thereby reducing the profit from \$50 to \$40. For a False Negative, the loss is \$20 because Telco forgoes a 50% chance of earning \$40, resulting in no earnings. Consequently, the decision threshold shifts from 0.5 to 0.33, based on the formula $p > \frac{1}{1+\frac{C_{FN}}{C_{FP}}} = \frac{1}{3}$. This change ensures the model is more inclined to make positive predictions to account for the higher cost associated with False Negatives.

Reality \ Prediction	Churn	No churn
Churn	\$0	\$10
No churn	\$20	\$0

Table 7: Loss Matrix

5.3 Model Performance

According to the Model Performance table in Appendix7.3 , several key findings were identified:

- XGBoost is still the best performing model, with the highest test accuracy, precision, recall, and F1-score among the models.
- The LASSO Logistic Regression model has the longest training time, possibly due to the additional regularization step.
- Changing the threshold from 0.5 to 0.333 generally improves the recall of the models, but decreases the precision and accuracy.
- The AUC values are all above 0.98, indicating that all the models have excellent discriminative power.

5.4 Expected Profit

For each model, we created a new confusion matrix based on our updated threshold of p = 0.333, and calculated the expected profits with the formula where TN = True Negative, FP = False Positive, TP = True Positive:

$$E[total\_profit] = TN * normal\_profit + FP * profit\_after\_promotion + TP * 0.5 * profit\_after\_promotion$$

In our case, normal_profit equals to $50 while profit_after_promotion equals to $40. And in Table8 is the estimated profits in different models. With p=0.333, XGBoost will yield the highest profit of $58290, winning an increase in profit of 13% compared with no offer made. In this hypothetical case, offering promotional online security service is proved to be an effective retention strategy with XGBoost being the best predicting model.

Cases	Expected Profit
No offer for anyone	51750.0
Vanilla Logistic Regression (0.333)	58190.0
LASSO Logistic Regression (0.333)	58230.0
CART (0.333)	58180.0
XGBoost (0.333)	58290.0
Promotion offer for everyone	48880.0

Table 8: Expected Profit per model

6 Customer Retention Strategy

Based on the above analysis, we have developed 5 unique strategies for Telco's customer retention.

6.1 Improve Satisfaction

There is no doubt that satisfaction_score is the most important predictor of all our models. It's also quite obvious from our descriptive analysis that satisfaction score of 3 is the threshold of churn. To improve satisfaction rate, apart from increasing service quality that is mentioned below, our recommendation will be to introduce better customer service to approach those "high risk" customers before they churn, also trying to reduce the chances of refund issues among cluster 3. Training staff to be more supportive is crucial, as it's the third most common reason for customer churn. The support team should also receive enhanced training to assist seniors with technical issues, improving accessibility and inclusiveness in customer service.

6.2 Develop and market about competitive advantages in service offering

By observing the data about reason for churning, we can get a better understanding of what led to the low satisfaction scores, and eventually the churn. For cluster2 and 3, the main reasons for their churn were 'better devices' and 'better offers' from competitors. Especially for cluster3, their most commonly accepted offer(offer B) happened to be the most popular offer provided by Telco, which may be implying that the 'best' offer of Telco is not as attractive as that of its competitors. A clear understanding of the defining traits of competitor services, and connecting them to the characteristics of high-risk groups will provide a pathway for Telco to address the needs of their customers and improve the satisfaction score accordingly.

6.3 Incentivize Longer Term Contracts

In all of our predictive models, the 'Contract_Two Year' feature consistently ranks among the top 10 important features, indicating that two-year contracts tend to lower the risk of customer churn. This is further supported by our K-means clustering analysis, where high-risk groups (Cluster 2 and Cluster 3) predominantly have 'month-to-month contracts' as their most common plan while Cluster 0 and Cluster 1 most common are on 'Two-year Contract'. We recommend that Telco's next action should focus on promotions that incentivize long-term contracts, offering benefits such as better pricing or additional free services (e.g., device protection plans, premium tech support). It's essential to balance the trade-off between reducing churn and budgeting, necessitating further granular churn analysis in this specific case.

6.4 Better Access to Online Security Services

Privacy and data protection are among the most significant needs for customers in today's digital era, and very attractive aspect for marketing[MM17]. As previously mentioned, online security is a significant predictor of churn risk in all

models, with logistic regression results indicating that subscribing to online security services is associated with lower churn risk. However, our high-risk groups have a low proportion of customers subscribed to these services, suggesting that despite the quality of the offering, the pricing is prohibitive. The granular churn analysis conducted earlier provides a valuable starting point. Thorough market research or qualitative analysis using online surveys to evaluate the impact of online security services on churn risk will offer clearer guidance on balancing benefits with costs.

6.5 Services For Senior Citizens

Telco's customers have a high average age across all clusters, with high-risk groups particularly including many senior citizens (customers 65 or older) compared to low-risk groups. Most predictive models have identified 'senior_citizen' as a significant predictor of churn risk, and logistic regression results indicate that senior citizens are more likely to churn. Therefore, it would be beneficial for Telco to devise strategies catering to seniors for customer retention. In addition to enhancing customer service, Telco could create special offer bundles for seniors, including discounted charges for long-distance calls. They should also consider redesigning their app interface and information hierarchy to reduce confusion and improve accessibility for seniors.

7 Conclusion

7.1 Model conclusion

There is no perfect model; we recommend that Telco leverage different models to obtain actionable insights based on their objectives.

Holistic Strategy: LASSO Logistic Regression The major advantage of logistic regression, including LASSO, is interpretability. It helps understand predictor and its directional association, making it ideal for general strategies. For example, the model shows that online security services, referrals, premium tech support, and long-term contracts reduce the risk of churn. Its high AUC score (> 0.99) confirms the reliability and effectiveness of the model in capturing data patterns.

Predictive Performance: XGBoost (0.5) When Telco wants to correctly estimate the probability of a new customer churning to make specific targeting strategies (such as direct marketing, personalized pricing/promotions), XGBoost performs the best. It outperforms all other models on nearly every classification metric, indicating its excellent generalization ability on unseen data. With optimized parameters, XGBoost's training time is reduced, enabling daily data updates. Therefore, XGBoost is a highly suitable choice.

Profit Maximization: XGBoost (0.333) In terms of predicting the effect of a specific strategy for customer retention and profit maximization, the XGBoost (0.333) model is the most appropriate choice. This methodology is

especially suitable when loss matrix that calculates the cost of false and true positive is determined. It can be a perfect way to make predictions for the specific strategy being made with the aim to maximize profit.

7.2 Limitations

The main limitation of our analytics process is that predictive models alone cannot determine the causal relationships between predictors and churn. Our results mainly talk about association and insights, while causation would be a entirely different story. However, it is also true that the interpretations from our analysis provide a clear path for such causal methods by providing plausible candidates and narrowing the options for Telco's future decisions, which is already cost-effective compared to its counterfactual scenarios.

Another limitation is the unclear correlation between the referral strategy and the churn rate. To address this limitation along with the first, it is essential for Telco to undergo an attribution modeling phase to determine whether referrals have a direct positive effect on reducing churn risk. This would include employing more experimental analytical techniques such as A/B testing, or observational methods such as Propensity Score Matching. These would be an excellent improvement to the issue, making the causal relationship between referrals and the churn rate clearer.

Furthermore, due to the complexity of this attribute, we have removed location details from the prediction dataset. However, descriptive analysis has also revealed interesting patterns in user behavior within certain geographical areas, suggesting that location information could bring significant value to the process of predicting and preventing customer churn. In the future, with the inclusion of location data, the process of creating predictive models and designing retention strategies will become more comprehensive and robust.

7.3 Final conclusion

In summary, our models give Telco an analytics advantage by offering a clear understanding of its customers, including characteristics, preferences, distinctions between high-risk and low-risk groups through clustering, and factors linked to churn risk. We successfully trained high-performing predictive models to accurately predict the churn risk of customers and selected the two best to help Telco in their business. These models inform robust strategies tailored to Telco's goal of customer retention. Building on the foundation of this analytics edge, Telco will be able to unlock new frontiers in both research and practical applications.

References

[Gre93] George D. Greenwade. The Comprehensive Tex Archive Network (CTAN). *TUGBoat*, 14(3):342–351, 1993.

[ld23] Scikit learn developers. sklearn.model_selection.gridsearchcv. https://scikit-learn.org/dev/modules/generated/sklearn.model_selection.GridSearchCV.html, 2023. Accessed: 2023-10-11.

[MM17] Kelly D Martin and Patrick E Murphy. The role of data privacy in marketing. *Journal of the Academy of Marketing Science*, 45:135–155, 2017.

Appendix A: Data Dictionary

Demographics Data

CustomerID	*A unique ID that identifies each customer.*
Count	*A value used in reporting/dashboarding to sum up the number of customers in a filtered set.*
Gender	*The customer's gender: Male, Female*
Age	*The customer's current age, in years, at the time the fiscal quarter ended.*
Senior Citizen	*Indicates if the customer is 65 or older: Yes, No*
Married	*Indicates if the customer is married: Yes, No*
Dependents	*Indicates if the customer lives with any dependents: Yes, No. Dependents could be children, parents, grandparents, etc.*
Number of Dependents	*Indicates the number of dependents that live with the customer.*

Location Data

Country	*The country of the customer's primary residence.*
State	*The state of the customer's primary residence.*
City	*The city of the customer's primary residence.*
Zip Code	*The zip code of the customer's primary residence.*
Lat Long	*The combined latitude and longitude of the customer's primary residence.*
Latitude	*The latitude of the customer's primary residence.*
Longitude	*The longitude of the customer's primary residence.*

Population Data

Zip Code	*The zip code of the customer's primary residence.*
Population	*A current population estimate for the entire Zip Code area.*

Service Data

Quarter	*The fiscal quarter that the data has been derived from (e.g. Q3).*
Referred a Friend	*Indicates if the customer has ever referred a friend or family member to this company: Yes, No.*
Number of Referrals	*Indicates the number of referrals to date that the customer has made.*
Tenure in Months	*Indicates the total amount of months that the customer has been with the company by the end of the 3rd quarter.*
Offer	*Identifies the last marketing offer that the customer accepted, if applicable. Values include No Offer, Offer A, Offer B, Offer C, Offer D, and Offer E.*
Phone Service	*Indicates if the customer subscribes to home phone service with the company: Yes, No.*
Avg Monthly Long Distance Charges	*Indicates the customer's average long distance charges, calculated to the end of the quarter specified above.*
Multiple Lines	*Indicates if the customer subscribes to multiple telephone lines with the company: Yes, No.*
Internet Type	*Indicates if the customer subscribes to Internet service with the company: No Subscription, DSL, Fiber Optic, Cable.*
Avg Monthly GB Download	*Indicates the customer's average download volume in gigabytes, calculated to the end of the quarter specified above.*
Online Security	*Indicates if the customer subscribes to an additional online security service provided by the company: Yes, No.*
Online Backup	*Indicates if the customer subscribes to an additional online backup service provided by the company: Yes, No.*
Device Protection Plan	*Indicates if the customer subscribes to an additional device protection plan for their Internet equipment provided by the company: Yes, No.*
Premium Tech Support	*Indicates if the customer subscribes to an additional technical support plan from the company with reduced wait times: Yes, No.*
Streaming TV	*Indicates if the customer uses their Internet service to stream television programming from a third party provider: Yes, No. The company does not charge an additional fee for this service.*

Streaming Movies	*Indicates if the customer uses their Internet service to stream movies from a third party provider: Yes, No. The company does not charge an additional fee for this service.*
Streaming Music	*Indicates if the customer uses their Internet service to stream music from a third party provider: Yes, No. The company does not charge an additional fee for this service.*
Unlimited Data	*Indicates if the customer has paid an additional monthly fee to have unlimited data downloads/uploads: Yes, No.*
Contract	*Indicates the customer's current contract type: Month-to-Month, One Year, Two Year.*
Paperless Billing	*Indicates if the customer has chosen paperless billing: Yes, No.*
Payment Method	*Indicates how the customer pays their bill: Bank Withdrawal, Credit Card, Mailed Check.*
Monthly Charge	*Indicates the customer's current total monthly charge for all their services from the company.*
Total Charges	*Indicates the customer's total charges, calculated to the end of the quarter specified above.*
Total Revenue	*Indicates total revenue of the company. No further details in IBM's data dictionary.*
Total Refunds	*Indicates the customer's total refunds, calculated to the end of the quarter specified above.*
Total Extra Data Charges	*Indicates the customer's total charges for extra data downloads above those specified in their plan, by the end of the quarter specified above.*
Total Long Distance Charges	*Indicates the customer's total charges for long distance above those specified in their plan, by the end of the 'quarter' field.*

Status Data

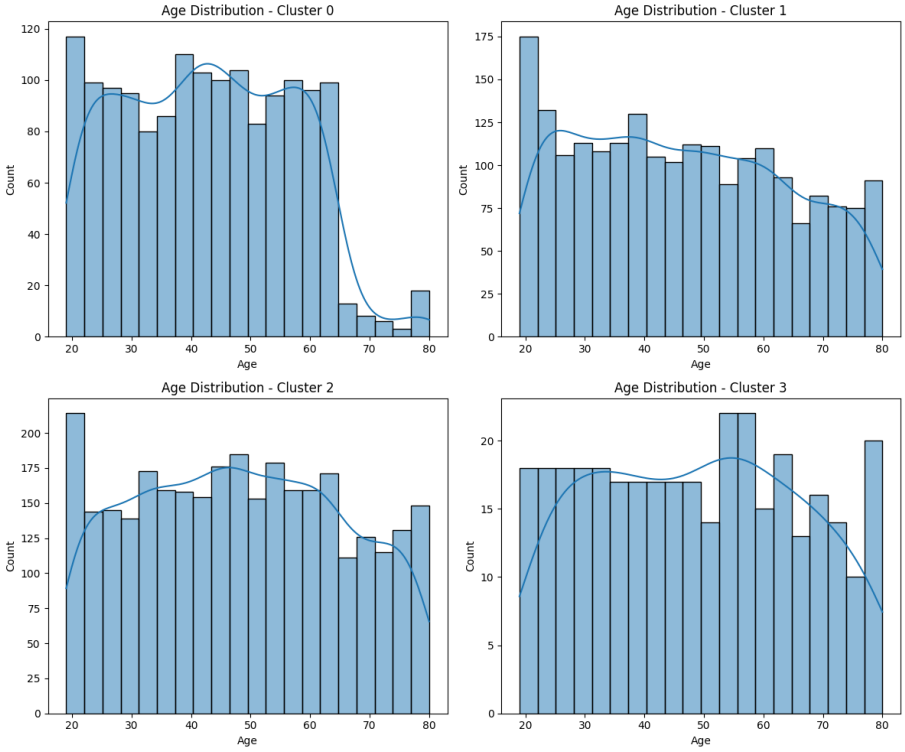
Quarter	*The fiscal quarter that the data has been derived from (e.g. Q3).*
Satisfaction Score	*A customer's overall satisfaction rating of the company from 1 (Very Unsatisfied) to 5 (Very Satisfied).*
Satisfaction Score Label	*Indicates the text version of the score (1-5) as a text string.*
Customer Status	*Indicates the status of the customer at the end of the quarter: Churned, Stayed, or Joined*

Churn Label	*Yes = the customer left the company this quarter. No = the customer remained with the company. Directly related to Churn Value.*
Churn Value	*1 = the customer left the company this quarter. 0 = the customer remained with the company. Directly related to Churn Label.*
Churn Score	*A value from 0-100 that is calculated using the predictive tool IBM SPSS Modeler. The model incorporates multiple factors known to cause churn. The higher the score, the more likely the customer will churn.*
Churn Score Category	*A calculation that assigns a Churn Score to one of the following categories: 0-10, 11-20, 21-30, 31-40, 41-50, 51-60, 61-70, 71-80, 81-90, and 91-100.*
CLTV	*Customer Lifetime Value. A predicted CLTV is calculated using corporate formulas and existing data. The higher the value, the more valuable the customer. High value customers should be monitored for churn.*
CLTV Category	*A calculation that assigns a CLTV value to one of the following categories: 2000-2500, 2501-3000, 3001-3500, 3501-4000, 4001-4500, 4501-5000, 5001-5500, 5501-6000, 6001-6500, and 6501-7000.*
Churn Category	*A high-level category for the customer's reason for churning: Attitude, Competitor, Dissatisfaction, Other, Price. When they leave the company, all customers are asked about their reasons for leaving. Directly related to Churn Reason.*
Churn Reason	*A customer's specific reason for leaving the company. Directly related to Churn Category.*

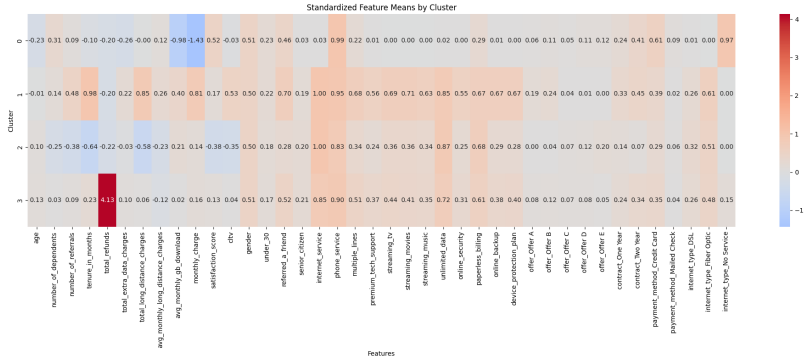
Appendix B: Means of numerical variables

count	1.000000
age	46.509726
number_of_dependents	0.468692
zip_code	93486.071134
latitude	36.197455
longitude	-119.756684
number_of_referrals	1.951867
tenure_in_months	32.386767
avg_monthly_long_distance_charges	22.958954
avg_monthly_gb_download	20.515405
monthly_charge	64.761692
total_charges	2280.381264
total_refunds	1.962182
total_extra_data_charges	6.860713
total_long_distance_charges	749.099262
total_revenue	3034.379056
satisfaction_score	3.244924
satisfaction_score_label	3.244924
churn_value	0.265370
churn_score	58.505040
cltv	4400.295755

Appendix C: Figures of User Clustering



(a) Age Distribution - Cluster



(b) Standardized Feature Means by Cluster

Figure 12: User Clustering Analysis

Appendix D: CART

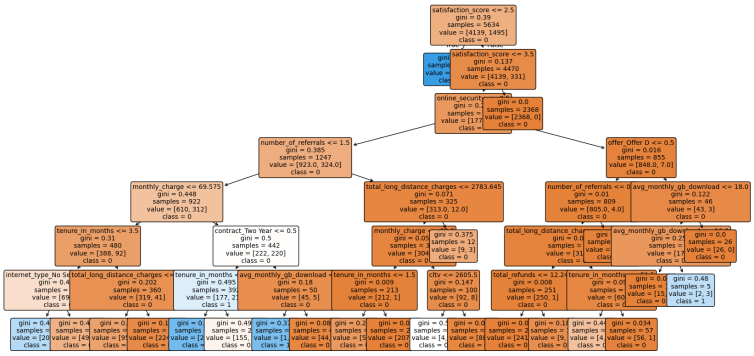


Figure 13: CART Model

Appendix E: Confusion Matrix with p=0.333

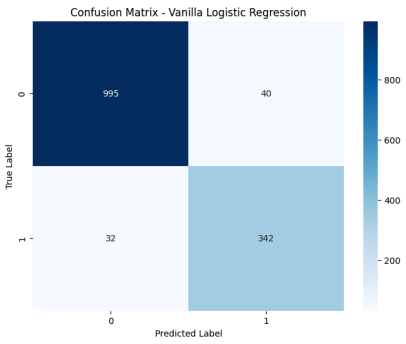


Figure 14: Vanilla

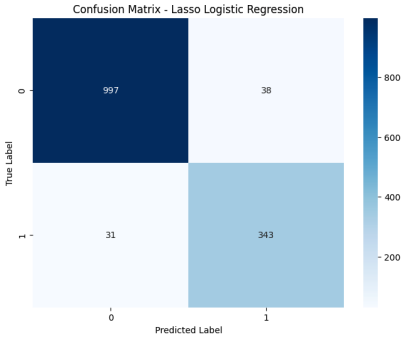


Figure 15: LASSO

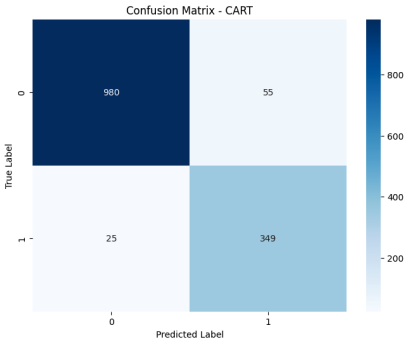


Figure 16: CART

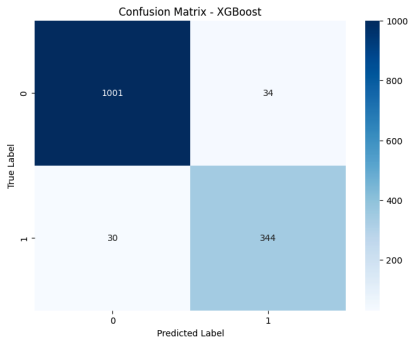


Figure 17: XGBoost

Appendix F: Model Performance Comparison

Model	Threshold	Train_ Accuracy	Test_ Accuracy	Train_ Precision	Test_ Precision	Train_ Recall	Test_ Recall	Train_F1	Test_F1	Train_AUC	Test_AUC	Training_Time
Vanilla Logistic Regression (0.5)	0.5	0.965744	0.958126	0.953975	0.948718	0.91505	0.890374	0.934107	0.918621	0.993572	0.991271	0.054603
LASSO Logistic Regression (0.5)	0.5	0.965566	0.958836	0.951423	0.948864	0.917057	0.893048	0.933924	0.92011	0.993707	0.991483	1.687758
CART (0.5)	0.5	0.953319	0.944642	0.965961	0.953988	0.854181	0.831551	0.906638	0.888571	0.992452	0.98568	0.023865
XGBoost (0.5)	0.5	0.971778	0.960256	0.977143	0.967647	0.91505	0.879679	0.945078	0.921569	0.995499	0.992358	0.136084
Vanilla Logistic Regression (0.333)	0.333	0.958821	0.9489	0.896422	0.895288	0.955184	0.914439	0.92487	0.904762	0.993572	0.991271	0.048309
LASSO Logistic Regression (0.333)	0.333	0.958289	0.951029	0.896725	0.900262	0.952508	0.917112	0.923776	0.908609	0.993707	0.991483	1.919276
CART (0.333)	0.333	0.950302	0.943222	0.872471	0.863861	0.951839	0.933155	0.910429	0.897172	0.992452	0.98568	0.034612
XGBoost (0.333)	0.333	0.965034	0.954578	0.91871	0.910053	0.952508	0.919786	0.935304	0.914894	0.995499	0.992358	0.156008

Figure 18: Model Performance Across Classification Metrics