Service Portfolio Optimization Bundling Profitability and Retention

1.Introduction

Customer retention is a central priority for subscription-based industries, particularly telecommunications, where acquisition costs are high and competitive pressure is intense. The Telco Customer Churn dataset provides a rich foundation for studying churn behavior because it captures demographic attributes, service configurations, pricing, and billing information. In this project, we treat churn prediction as a methodological exercise: our goal is to construct, compare, and interpret supervised-learning models that estimate the probability a customer will discontinue service. The target variable, Churn, is a binary indicator reflecting whether a customer left during the most recent billing cycle, making the problem naturally suited to binary classification techniques.

This dataset is widely studied because churn is not random—it is shaped by customer tenure, contract structure, service usage, and pricing. Accurate prediction enables firms to identify at-risk customers early and deploy targeted interventions. However, the dataset poses several challenges. First, the churn label is imbalanced (≈26% churn), which can bias models toward predicting "No churn." Second, the dataset contains many categorical variables with multiple levels, requiring careful encoding to prevent information loss. Third, the TotalCharges variable contains inconsistent formatting and missing values that must be cleaned to avoid distorted baseline estimates. These difficulties require a modeling pipeline that integrates data cleaning, feature engineering, and comparison of linear versus nonlinear learners.

Finally, because customer profitability is a key business consideration, we complement the churn-prediction models with two additional analyses: (1) a linear LTV estimation based on billing patterns, and (2) K-Means clustering to identify meaningful customer personas. Together, these analyses provide a comprehensive view of customer value and attrition risk.

2.Related Work

The Telco churn problem has been explored extensively across machine-learning and marketing analytics research. Coussement & Van den Poel (2008) applied Support Vector Machines for churn prediction and showed that nonlinear decision boundaries can outperform classical statistical models. Their work emphasizes the importance of feature preprocessing and hyperparameter tuning—elements we incorporate into our pipeline. Verbeke et al. (2014) used social-network features to improve churn classification, demonstrating that relationship-driven attributes can outperform traditional tabular variables. While their approach yields strong performance, it requires network data not present in the Telco dataset, making our supervised-learning focus more appropriate.

Tianqi Chen & Carlos Guestrin (2016) introduced XGBoost, a scalable gradient-boosting algorithm that has since become a state-of-the-art method for tabular classification. Their findings motivate our use of XGBoost as a nonlinear counterpart to Logistic Regression. Fader, Hardie & Lee (2005) proposed probabilistic models for estimating customer lifetime value, showing that deterministic formulas can be insufficient when customer purchasing patterns vary widely. While our LTV model is simpler, the comparison highlights why high R^2 scores arise in deterministic scenarios.

Other studies, such as Ahn et al. (2006) and Idris et al. (2012), have demonstrated that ensemble models consistently outperform single learners in churn prediction. These findings align with our results, where XGBoost slightly outperforms Logistic Regression but both capture substantial predictive signal. Relative to the literature, our approach contributes a streamlined methodological comparison paired with a value-based segmentation framework that enables actionable business recommendations.

3.Related Implementations

Kaggle implementations of the Telco dataset provide useful methodological baselines. Pingli (2025) follows a CRISP-DM structure, emphasizing systematic preprocessing and classical models such as Logistic Regression and Random Forests. While clear and well-structured, the notebook relies primarily on baseline models and minimal feature engineering. Afram (2025) conducts extensive EDA and tests SVMs and Decision Trees, highlighting important predictors such as contract type and monthly charges. However, the models are largely untuned and do not incorporate boosted-tree ensembles.

Compared with these implementations, our work advances in two ways. First, we engineer behavioral features (e.g., service-count and charge-intensity metrics) that summarize usage patterns rather than relying solely on one-hot expansions. Second, we integrate XGBoost as a nonlinear learner capable of capturing higher-order feature interactions. Our pipeline therefore builds on community practices while offering a more systematic comparison between linear and nonlinear predictive frameworks.

4. Data Analysis

The dataset includes 7,043 customers and 21 variables. We cleaned TotalCharges by converting the field to numeric and imputing missing values using the formula:

$$TotalCharges = MonthlyCharges \times tenure \quad (\text{for customers with tenure } = 0)$$

We engineered two behavioral features:

$$service\_count = \sum_i service_i, \qquad avg\_charge\_per\_service = \frac{MonthlyCharges}{\max(1, service\_count)}.$$

These capture breadth and intensity of service usage. Figure 1 (kept) illustrates clear differences in tenure and accumulated charges between churned and non-churned customers, suggesting that customer longevity and financial engagement are strong churn indicators. Figure 2 (kept) shows churn rates across contract types, with month-to-month customers exhibiting substantially higher churn, reinforcing the importance of commitment structure.

Categorical-variable analysis revealed that fiber-optic internet, electronic-check payment, and lack of security add-ons correlate with elevated churn. Numerical correlations showed that MonthlyCharges and TotalCharges are related yet represent distinct behavioral dimensions—current burden vs cumulative value—justifying retention of both. This analysis informed the feature set, revealing both linear trends and nonlinear interactions that motivate our choice of models.

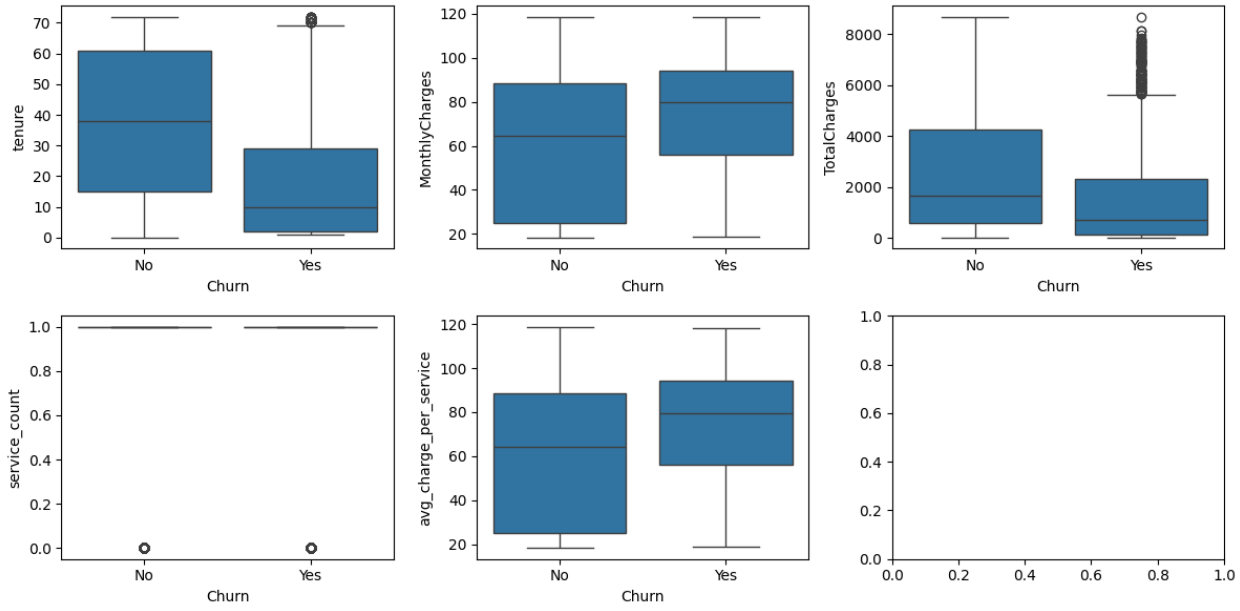

Figure 1: Boxplots Comparing Churn vs. Non-Churn Customers for Numerical Features

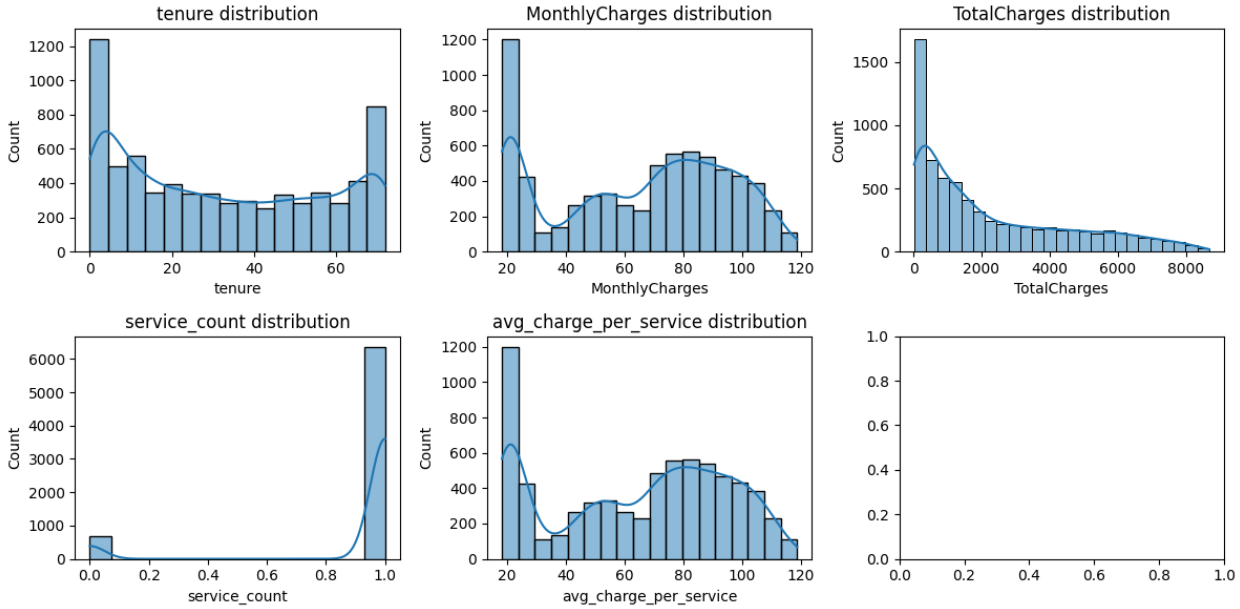

Figure 2: Distribution of Key Numerical Features

5.Proposed Method

5.1 Logistic Regression

Logistic Regression models the log-odds of churn as a linear combination of predictors:

$$P(churn = 1 \mid x) = 1 / (1 + e^{-(\beta_0 + \beta^T x)})$$

We apply L2 regularization to mitigate multicollinearity introduced by one-hot encoding and set class_weight="balanced" to address label imbalance. Logistic Regression is well-suited for this dataset because many predictors—contract type, monthly charges, service count—exhibit monotonic relationships with churn risk. The model produces interpretable coefficients, enabling direct identification of factors that increase or decrease churn likelihood.

5.2 XGBoost

XGBoost builds an ensemble of decision trees using gradient boosting. The model

$$L = \Sigma\, \ell(y_i, \hat{y}_i) + \Sigma\, (\gamma\, T_k + \tfrac{1}{2}\, \lambda\, ||w_k||^2)$$

where each tree corrects errors from previous ones. We tune learning rate (0.1), max depth (4), and number of boosting rounds (200). XGBoost is appropriate for churn because it captures nonlinear interactions—e.g., how contract type interacts with service count—without requiring manual feature-interaction engineering.

5.3 Linear Model for Lifetime Value (LTV)

We use a deterministic formula:

LTV = MonthlyCharges X tenure X 0.35

reflecting churn-adjusted gross margin. Because LTV is nearly a direct transformation of two variables, linear regression achieves $R^2 \approx 0.999$. Although simple, this model provides a stable profitability target used later in segmentation.

5.4 K-Means Clustering

To uncover customer personas, we apply K-Means (k=4) using standardized numerical features. The algorithm minimizes:

$$\sum_{i=1}^{n} \|x_i - \mu_{c(i)}\|^2,$$

and partitions customers into interpretable clusters based on spending level, engagement, and churn patterns.

6.Experimental Setup

We standardize continuous features and one-hot encode categorical variables. All models use the same 80/20 train–test split to ensure comparability. Logistic Regression uses L2 regularization, class balancing, and max_iter=2000 to guarantee convergence. XGBoost uses the hyperparameters described earlier. The linear LTV model uses MonthlyCharges and tenure directly. K-Means is applied to normalized numerical features to prevent disproportionate scaling effects.

Evaluation metrics include accuracy, F1 score, and ROC-AUC for classification, and R^2/RMSE for LTV. All experiments are fully documented in the public GitHub repository.

7.Results

7.1 Churn Prediction Performance

Model	**Accuracy**	**F1**	**ROC-AUC**
Logistic Regression	**0.7409**	**0.6162**	**0.8410**
XGBoost	**0.7544**	**0.6256**	**0.8362**

Both models perform similarly. Logistic Regression slightly outperforms XGBoost in ROC-AUC, indicating that linear effects explain much of the churn structure. XGBoost achieves marginally higher accuracy and F1, capturing subtle nonlinearities. Together, these models provide complementary strengths: interpretability vs flexible pattern learning.

7.2 Lifetime Value Modeling

Model Performance：

Metric	**Train**	**Test**
R^2	0.9985	0.9988
RMSE	30.60	27.31

High performance reflects the deterministic structure of LTV, confirming the suitability of a simple linear approach for estimating customer profitability.

7.3 K-Means Segmentation

We applied K-Means clustering (k = 4) to identify customer personas based on service usage, billing patterns, and churn risk.

The clustering produced four clearly differentiated segments (total n = 7,043):

Cluster	**MonthlyCharges**	**Tenure**	**LTV**	**Churn Rate**	**Interpretation**
0	92.43	56.65	1831.40	0.17	**High-value, long-tenure premium users**
1	74.00	13.20	345.54	0.46	**High-charge but newly onboarded high-risk users**

2	42.03	31.74	523.51	0.25	**Mid-tier, moderately stable users**
3	22.61	30.65	250.36	0.08	**Low-ARPU but highly loyal users**

The segmentation reveals clear personas that will be directly used in the optimization stage to analyze profit–risk trade-offs and design targeted retention or upsell strategies.

8.Discussion

Logistic Regression performs well because the dataset contains many additive, monotonic relationships with churn—shorter contracts, higher monthly charges, and certain service bundles consistently raise churn risk. These effects are naturally captured in a linear model, leading to strong ROC-AUC performance.

XGBoost succeeds because it captures nonlinear interactions, such as how internet type interacts with security add-ons or how contract form interacts with tenure. Although performance gains are modest, they confirm that small but meaningful nonlinearities exist.

The LTV and segmentation analyses extend churn prediction toward business decision-making. Because LTV is stable and deterministic, it provides a reliable profitability anchor. K-Means reveals heterogeneity in customer value and churn risk, highlighting subgroups that require differentiated strategies.

Overall, our findings demonstrate that while complex models provide marginal predictive improvements, simpler interpretable models already capture most of the actionable churn structure. Combining churn prediction with customer segmentation produces insights that generalize to other subscription industries where contract type, service bundling, and spending intensity shape retention.

9.Conclusion

This project demonstrates that both linear and nonlinear models effectively predict churn in the Telco dataset. Logistic Regression provides interpretable coefficients that clarify the core drivers of attrition, while XGBoost delivers slightly stronger predictive accuracy by modeling interactions. If an organization required transparent and operationally simple deployment, Logistic Regression would be recommended; however, for maximum predictive power, XGBoost would be preferable.

Lifetime value estimation and clustering extend churn prediction into actionable strategy. The LTV model identifies high-value customers, and K-Means highlights segments with distinct risk–value profiles. Retention strategies should prioritize high-value and moderate-tenure customers while addressing onboarding issues for high-risk new users.

With more data (e.g., usage sequences, network metrics), additional techniques such as survival analysis or sequence-based deep models could be employed. For large-scale production deployment, automated retraining pipelines and real-time scoring would make the proposed solution suitable for high-volume telecom environments.

6.Reference

[1] Kaggle. *Telco Customer Churn Dataset.* Available at: https://www.kaggle.com/blastchar/telco-customer-churn

[2] Tianqi Chen, Carlos Guestrin. XGBoost:w A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2016.

[3] Lloyd, S. Least squares quantization in PCM. IEEE Transactions on Information Theory, 28(2):129–137, 1982. (K-Means original algorithm)

[4] Fader, P. S., Hardie, B. G. S., & Lee, K. L. "Counting Your Customers" the Easy Way: An Alternative to the Pareto/NBD Model. Marketing Science, 24(2), 275–284, 2005. (Classical SOTA for LTV estimation)

[5] Coussement, K., & Van den Poel, D. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. Expert Systems with Applications, 34(1), 313–327, 2008. (Representative churn modeling literature)

[6] Verbeke, W., Martens, D., & Baesens, B. Social network analysis for customer churn prediction. Applied Soft Computing, 14, 431–446, 2014. (Example of advanced SOTA churn modeling approaches)

[7] Pingli, "CRISP-DM Methodology for a Customer Churn," Kaggle Notebook, 2025.

[8] Afram, "Telco Customer Churn Analysis and Prediction," Kaggle Notebook, 2025.

7.Appendix

The public Github repository link for this project: https://github.com/NiziJennyHe/Service-Portfolio-Optimization-Bundling-Profitability-and-Retention

8.Statement of Contribution

The development of this project reflects a coordinated division of labor. Jiani He was primarily responsible for data engineering and exploratory analysis, including data cleaning, categorical-encoding design, construction of derived behavioral features, and the creation of descriptive visualizations. She also implemented and evaluated the Logistic Regression and XGBoost models, conducted performance assessment, and contributed to the interpretation of model outputs.

Zongyang Li implemented the project's additional analytical components, including the linear Lifetime Value (LTV) model and the K-Means clustering analysis, and generated the associated visualizations and interpretations. He also contributed to organizing the modeling workflow and refining the structure and clarity of the final report.

Both members collaborated on synthesizing findings and preparing the final submission.