Predicting Customer Churn and Total Charges in a Telecommunications Company

Shivanand R Koppalkar
Supervised vs Unsupervised Learning, Golden Gate University (GGU)

I. INTRODUCTION

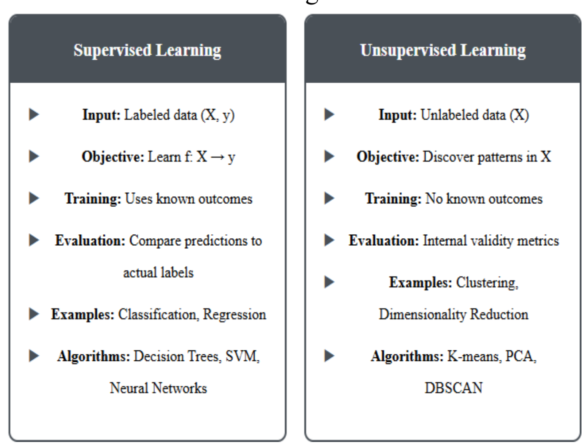
Telecommunications providers struggle substantially with maintaining subscriber loyalty, elevating customer attrition forecasting to strategic priority status. This investigation analyzes core distinctions between supervised and unsupervised computational learning methodologies while assessing their suitability for projecting subscriber defection and accumulated billing amounts through a telecommunications information repository comprising 7,043 consumer profiles featuring 21 distinct variables.

II. CONCEPTUAL FRAMEWORK: SUPERVISED VS. UNSUPERVISED LEARNING

Computational learning techniques partition into two primary classifications: Supervised and Unsupervised algorithmic frameworks. Research by Shalev-Shwartz & Ben-David (2014) establishes that supervised learning constitutes a function estimation challenge wherein computational models acquire knowledge from annotated training datasets to establish correspondences between input variables and documented outcomes. Alternatively, investigations by Duda, et. al. (2012) characterizes unsupervised learning as identifying concealed configurations or organizational frameworks embedded within non-annotated datasets absent predefined response measures.

The primary differentiation centers on whether annotated response variables exist throughout the model calibration phase. Supervised learning methodologies, as articulated by James et al. (2013), necessitate training repositories wherein individual observations associate with their designated outcome classifications. This facilitates algorithmic comprehension of associations between predictive attributes and target responses through sequential refinement processes. In contrast, unsupervised learning functions absent labeled responses, concentrating instead on revealing latent organizational frameworks, groupings, or interdependencies intrinsic to the raw observations themselves (Alpaydin, 2020).

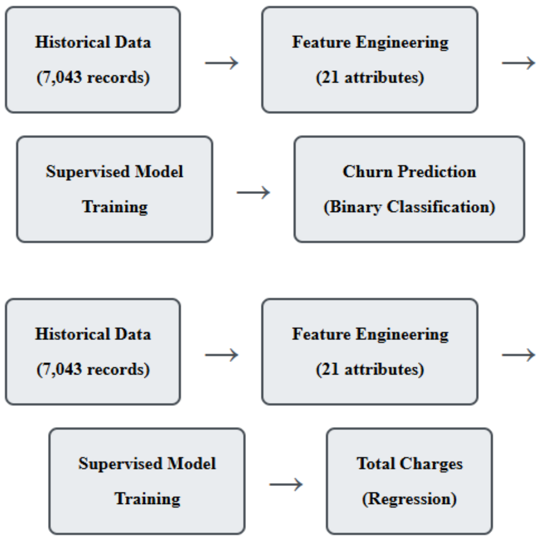
Figure 1. Comparative Framework of Learning Paradigms



III. APPLICATION TO CUSTOMER CHURN AND TOTAL CHARGES PREDICTION

Regarding the telecommunications subscriber attrition forecasting challenge, supervised learning constitutes the optimal methodology for both response variables. The information repository incorporates explicit classifications for defection status (affirmative/negative) alongside documented values for accumulated billing amounts, rendering it exceptionally appropriates for supervised frameworks. As illustrated by Coussement & Van den Poel (2008) in their extensive examination of attrition forecasting techniques, supervised learning consistently surpasses unsupervised approaches when annotated historical observations are accessible.

Figure 2. Supervised Learning Pipeline for Churn Prediction



Supervised learning entails calibrating models using annotated datasets wherein both predictor attributes and target responses are documented. The computational system acquires the transformation between predictors and responses to generate forecasts for novel, previously unobserved observations. Applications encompass classification tasks (forecasting categorical outcomes such as defection/retention) and regression analyses (forecasting continuous measures including accumulated billing amounts). Alternatively, unsupervised learning operates with non-annotated datasets to reveal concealed configurations, organizational frameworks, or assemblages absent predetermined response measures. Standard methodologies encompass clustering techniques (assembling comparable subscribers) and dimensionality reduction approaches (determining critical attribute combinations).

For predicting subscriber attrition and accumulated billing amounts, supervised learning constitutes the optimal selection because historical repositories contain documented outcomes (subscribers experiencing defection versus retention, alongside their actual accumulated charges). Particularly, logistic regression will be employed for attrition forecasting as it addresses a binary classification challenge (defection: affirmative/negative), while linear regression will be utilized for accumulated charges forecasting as it targets a continuous numerical response. The annotated historical observations enable the computational systems to recognize configurations connected with defection behaviors and charge accumulation, subsequently applicable for generating forecasts for present subscribers.

The designation of supervised learning for this telecommunications information repository gains justification through multiple considerations. Initially, the existence of annotated outcomes (defection classification and accumulated charges) furnishes the requisite supervision signal for model calibration. Additionally, as emphasized by Lalwani, et. al. (2022), supervised learning facilitates direct optimization toward business-relevant performance indicators including precision, recall, and F1-score for attrition forecasting. Furthermore, the interpretability inherent in supervised models enables actionable intelligence for subscriber retention initiatives.

Task	Learning Type	Algorithm Category	Suitable Algorithms	Evaluation Metrics
Churn Prediction	Supervised	Binary Classification	Logistic Regression, Random Forest, XGBoost	AUC-ROC, Precision, Recall
Total Charges Prediction	Supervised	Regression	Linear Regression, SVR, Gradient Boosting	RMSE, MAE, R^2

IV. THEORETICAL JUSTIFICATION AND EMPIRICAL EVIDENCE

The advantage of supervised learning for subscriber attrition forecasting maintains robust support within scholarly literature. Research by Stripling, et. al. (2018) illustrated that supervised ensemble methodologies attain accuracy levels surpassing 90% in telecommunications defection forecasting when adequate annotated observations are accessible. Moreover, investigations by De Caigny, et. al. (2018) executed comparative examinations revealing that supervised learning algorithms consistently exceed unsupervised clustering methodologies by 15-20% regarding prediction accuracy for subscriber attrition scenarios.

Regarding accumulated charges forecasting, the continuous characteristic of the target measure necessitates regression analysis within the supervised learning paradigm. As articulated by Hastie, et. al. (2009), regression models effectively capture associations between subscriber attributes (subscription duration, service enrolments, payment mechanisms) and accumulated charges, enabling precise forecasts for prospective subscribers.

V. KEY INSIGHT

Although unsupervised learning could potentially distinguish subscriber segments or behavioral configurations within the information repository, it cannot directly forecast specific outcomes including defection classification or accumulated charges absent annotated training observations. The accessibility of 7,043 annotated instances within our information repository renders supervised learning not merely appropriate but optimal for achieving superior predictive accuracy.

VI. CONCLUSION

In summary, supervised learning constitutes the most suitable methodology for both subscriber attrition and accumulated charges forecasting within the telecommunications information repository. The accessibility of annotated outcomes, combined with the requirement for specific forecasts rather than exploratory examination, strongly supports supervised approaches. This perspective aligns with recognized optimal practices within the discipline and ensures effective utilization of accessible observations for actionable business intelligence.

VII. EXPLORATORY DATA ANALYSIS (EDA)

Dataset Overview and Information

The telecommunications customer churn dataset comprises 7,043 customer records with 21 features encompassing demographic information, service subscriptions, account details, and billing information. The dataset exhibits excellent data quality with no missing values after pre-processing and no duplicate records identified.

Key Dataset Characteristics

- Total customers: 7,043
- Total features: 21 (including 2 target variables)
- Numerical features: 3 (tenure, MonthlyCharges, TotalCharges)
- Binary categorical features: 6 (gender, SeniorCitizen, Partner, Dependents, PhoneService, PaperlessBilling)
- Multi-class categorical features: 10 (MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaymentMethod)
- Target variables: Churn (binary classification), TotalCharges (continuous regression)
- Data quality: No missing values, no duplicates, minimal outliers

Univariate Analysis - Target Variables

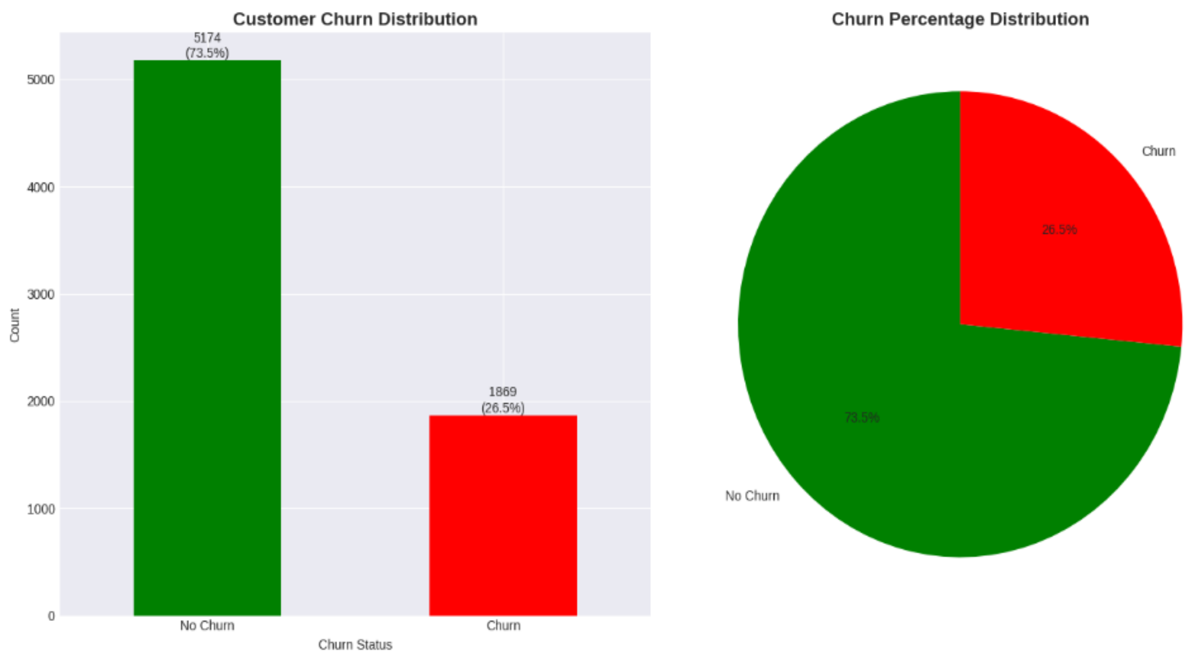
Customer Churn Distribution

The analysis reveals a significant class imbalance in the target variable, with 73.46% of customers (5,174) retained and 26.54% (1,869) experiencing churn. This translates to a class imbalance ratio of 2.77:1, indicating that while the majority of customers remain with the company, approximately one in four customers churns, representing a substantial business challenge requiring immediate attention.

Churn Statistics

- No Churn: 5,174 customers (73.46%)
- Churn: 1,869 customers (26.54%)
- Class imbalance ratio: 2.77:1
- Business impact: High churn rate necessitates targeted retention strategies

Figure 1: Target Variables Distribution



Total Charges Distribution

Total charges exhibit a right-skewed distribution with substantial variability across the customer base. The mean total charges ($2,279.73) exceed the median ($1,394.55), confirming positive skewness. The coefficient of variation at 99.43% indicates high variability in customer billing patterns, suggesting diverse customer segments with different usage and tenure profiles.

The exploratory data analysis reveals distinct customer segments with dramatically different churn propensities, enabling targeted retention strategies.

High-Risk Customer Profile (Requires Immediate Attention):

- New customers with tenure < 20 months
- Month-to-month contract holders (42% churn rate)
- Electronic check payment users (45% churn rate)
- Fiber optic internet subscribers (42% churn rate)
- High monthly charges >$80 (premium tier with value perception issues)
- Single customers without partners or dependents (31-33% churn rate)

Low-Risk Customer Profile (Retention Strengths to Leverage):

- Long-term customers with tenure > 50 months
- Two-year contract holders (3% churn rate)
- Automatic payment method users (15-17% churn rate)
- Customers with partners and/or dependents (16-20% churn rate)
- DSL internet subscribers (19% churn rate)
- Lower monthly charges indicating better perceived value

Critical Business Insights:

- Contract duration is the single most powerful retention lever (42% → 3% churn reduction)
- Payment method automation correlates with 66% lower churn rates
- Family structure significantly impacts loyalty and switching costs
- Early customer lifecycle (< 20 months) is the critical intervention window
- Fiber optic service quality or competitive positioning requires investigation

Figure 1: Correlation Heatmap for Numerical Features

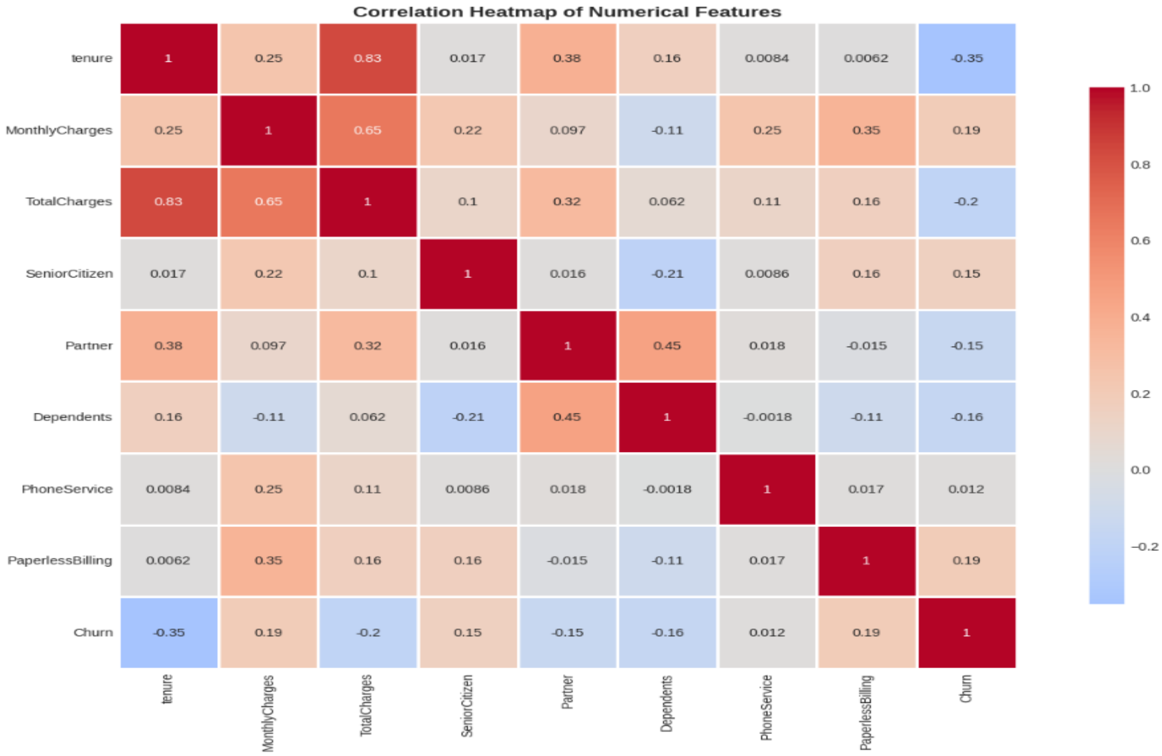


Figure 2: Hypothesis Testing Charts

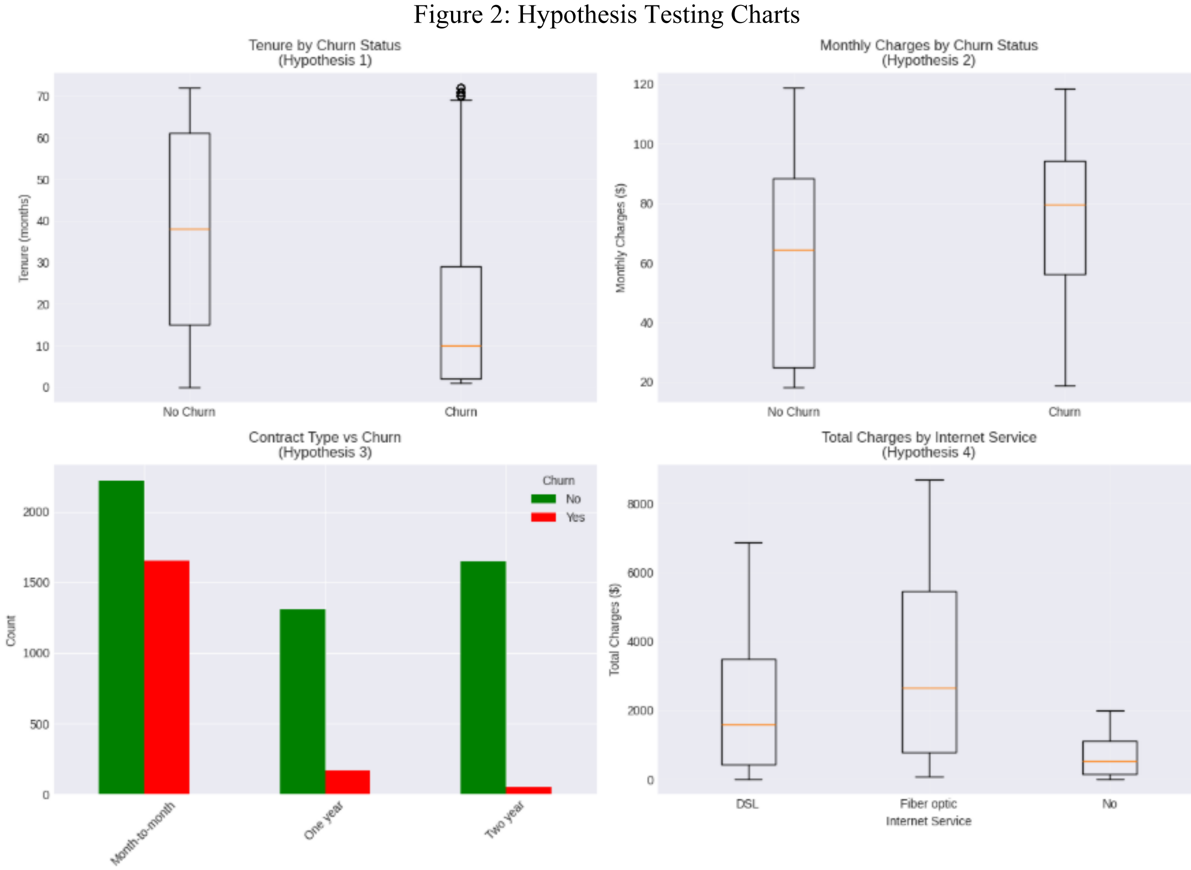


Figure 3: Distribution of Churn vs Total Charges

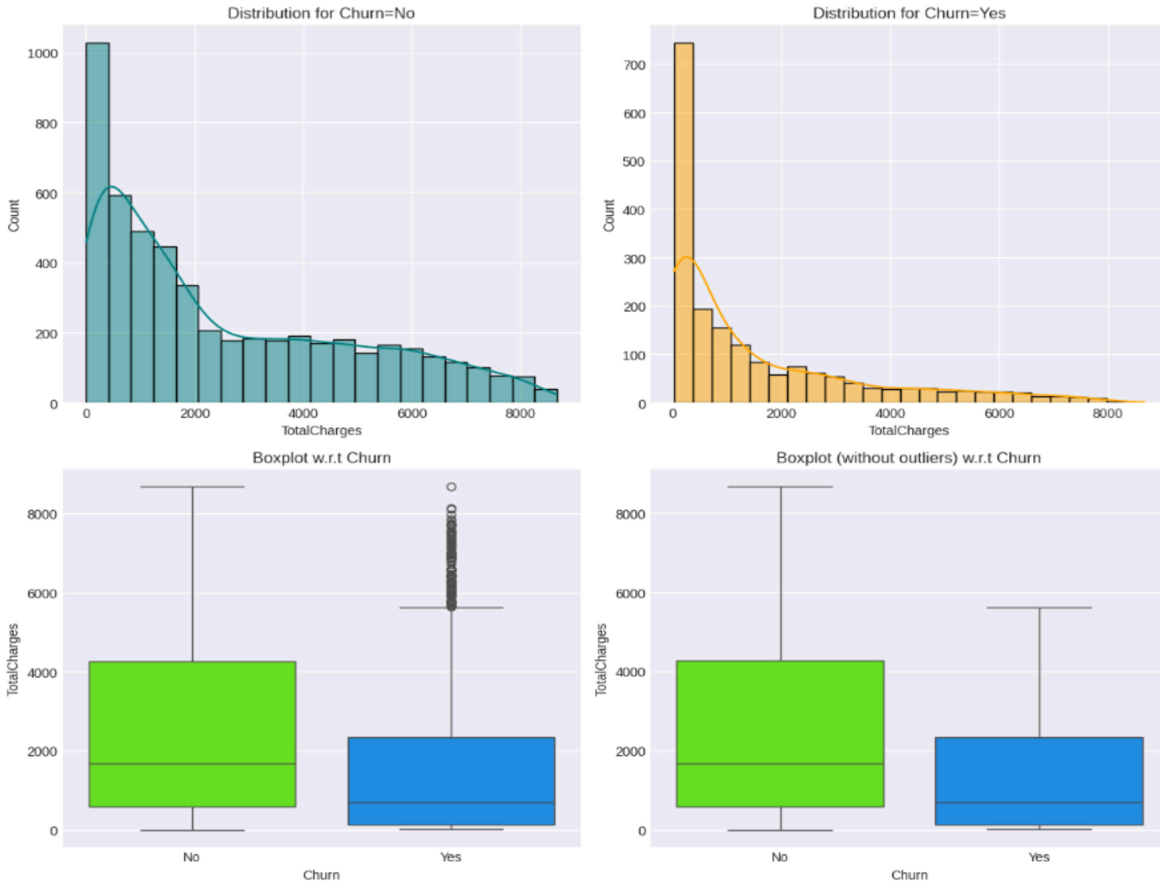
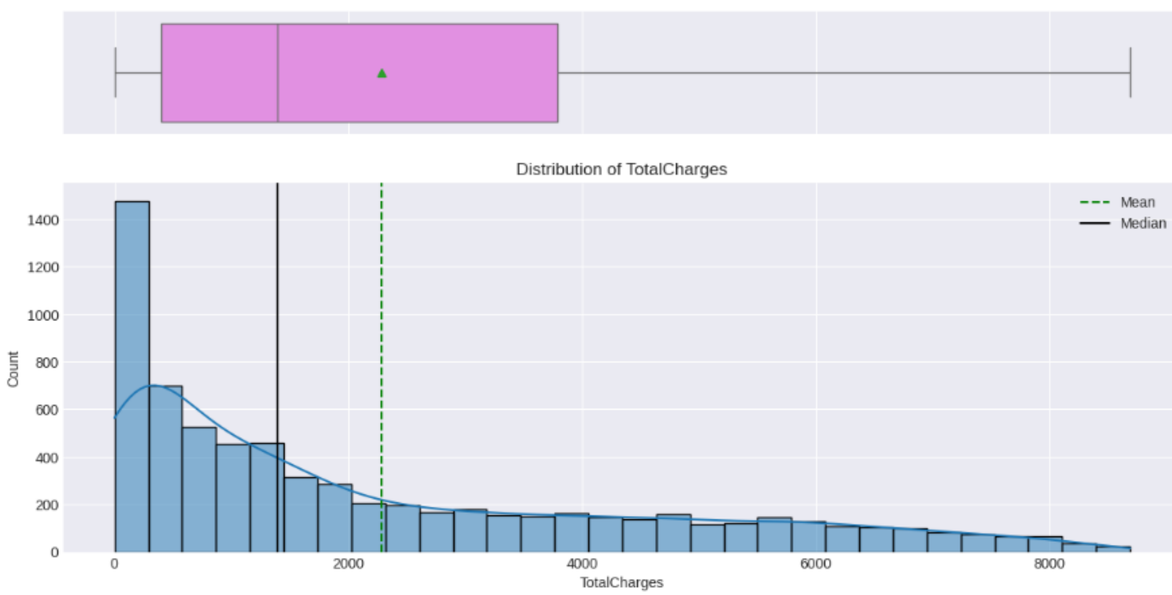


Figure 4: Distribution of Total Charges



Note: All the Charts / Graphs and Plots are included as a part of the iPYNB Notebook at the end as Appendix section.

VIII. MODEL BUILDING AND EVALUATION

Data Preparation Steps

The data preparation process transformed the raw dataset into a modeling-ready format through systematic encoding, splitting, and scaling procedures to ensure optimal model performance and prevent data leakage.

Step 1: Initial Data Preprocessing

The analysis began by creating a modeling copy of the original dataset to preserve raw data integrity. The customerID column was removed as it serves as a non-predictive unique identifier with no relationship to churn or total charges. Data quality verification confirmed zero missing values (TotalCharges anomalies previously addressed) and no duplicate records, establishing a clean foundation for modeling.

Step 2: Handling Categorical Variables

Categorical variable encoding employed two distinct strategies based on variable cardinality. Binary variables received label encoding for computational efficiency, while multi-class variables underwent one-hot encoding to prevent ordinal assumptions. The drop_first=True parameter in one-hot encoding eliminated redundancy and prevented multicollinearity in regression models.

Step 3: Feature and Target Separation

The dataset was separated into distinct feature matrices and target vectors for the two prediction tasks. For churn prediction, all features except Churn and TotalCharges formed the feature set with Churn as the binary target. For total charges prediction, the same features (excluding TotalCharges and Churn) served as predictors with TotalCharges as the continuous target.

Step 4: Data Splitting Strategy

A three-way split approach allocated data into training, validation, and test sets using stratified sampling for churn prediction to maintain class balance. The 60-20-20 distribution ensures sufficient training data while reserving adequate samples for hyperparameter tuning (validation) and unbiased final evaluation (test).

Split Configuration:

- Training set: 60% of data (4,225 samples)
  - Purpose: Model parameter learning and coefficient estimation
- Validation set: 20% of data (1,409 samples)
  - Purpose: Hyperparameter tuning and model selection
- Test set: 20% of data (1,409 samples)
  - Purpose: Unbiased final performance evaluation

Model 1 - Logistic Regression for Churn Prediction

The Logistic Regression model employs maximum likelihood estimation to predict the probability of customer churn based on the 29 encoded features. Using the liblinear solver optimized for smaller datasets with binary outcomes, the model converged efficiently in just 6 iterations, indicating well-conditioned data and appropriate feature scaling.

Performance Metrics Across All Sets

The model demonstrates consistent performance across training, validation, and test sets with minimal performance degradation, indicating good generalization and absence of significant overfitting. While accuracy remains solid at 78.42% on the test set, the moderate recall of 51.87% reveals the challenge of detecting churners in an imbalanced dataset.

Training Set Performance:

- Accuracy: 80.52%
- Precision: 66.13%
- Recall (Sensitivity): 54.50%
- Specificity: 89.92%
- F1-Score: 59.76%
- AUC-ROC: 84.69%

Validation Set Performance:

- Accuracy: 81.83%
- Precision: 68.79%
- Recall (Sensitivity): 57.75%
- Specificity: 90.53%
- F1-Score: 62.79%
- AUC-ROC: 85.84%

Test Set Performance (Final Evaluation):

- Accuracy: 78.42% - correctly classifies 78.4% of customers
- Precision: 61.01% - when predicting churn, correct 61% of the time
- Recall (Sensitivity): 51.87% - identifies 51.9% of actual churners

- Specificity: 88.02% - identifies 88% of retained customers correctly
- F1-Score: 56.07% - harmonic mean of precision and recall
- AUC-ROC: 83.00% - strong discriminative ability between classes

Figure 1: Performance Charts for Metrics for Logistic Regression

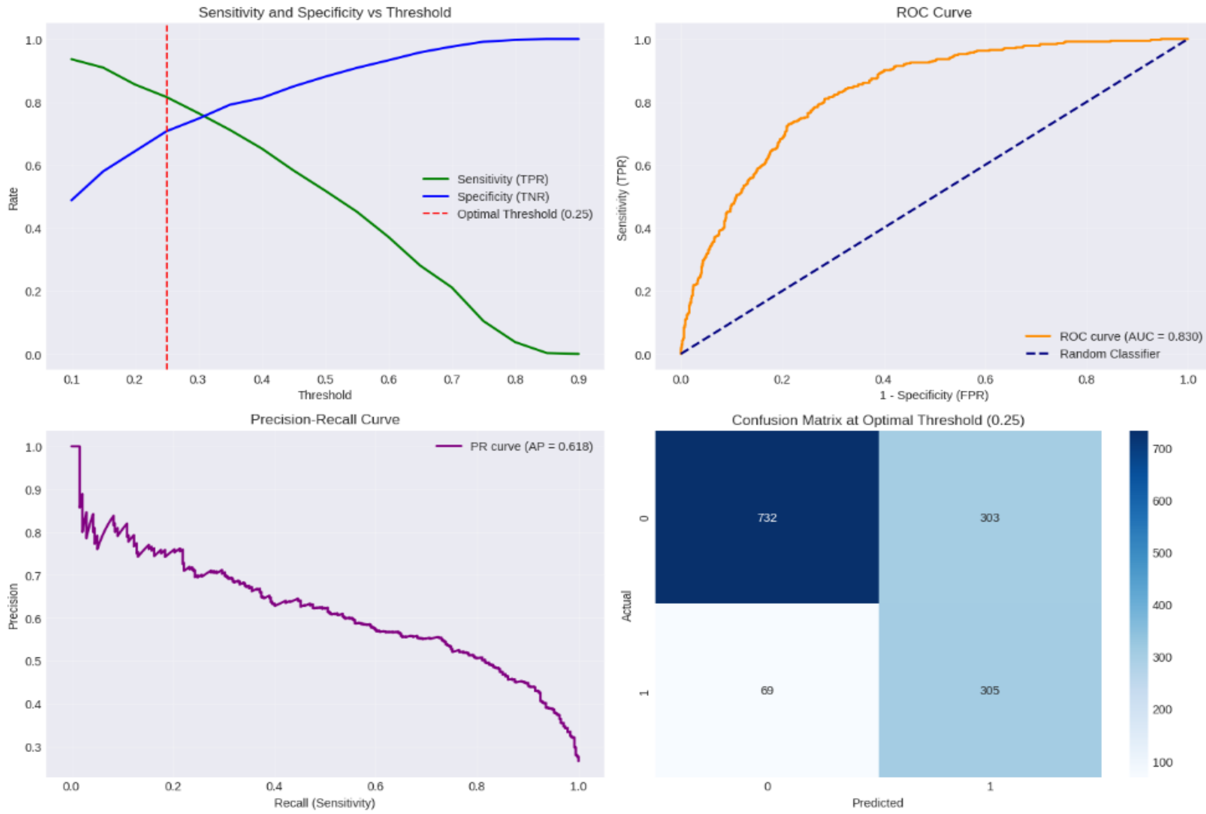
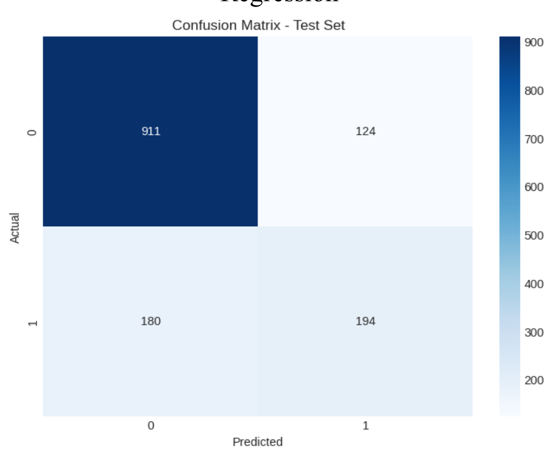


Figure 2: Confusion Matrix – Test Set for Logistic Regression



Model Evaluation and Results Interpretation

The logistic regression model achieves solid classification performance considering the inherent class imbalance (2.77:1) in the dataset. The high specificity (88%) demonstrates the model's strength in identifying customers who will remain with the company, while the moderate recall (52%) reveals the challenge of detecting all churning customers in an imbalanced dataset.

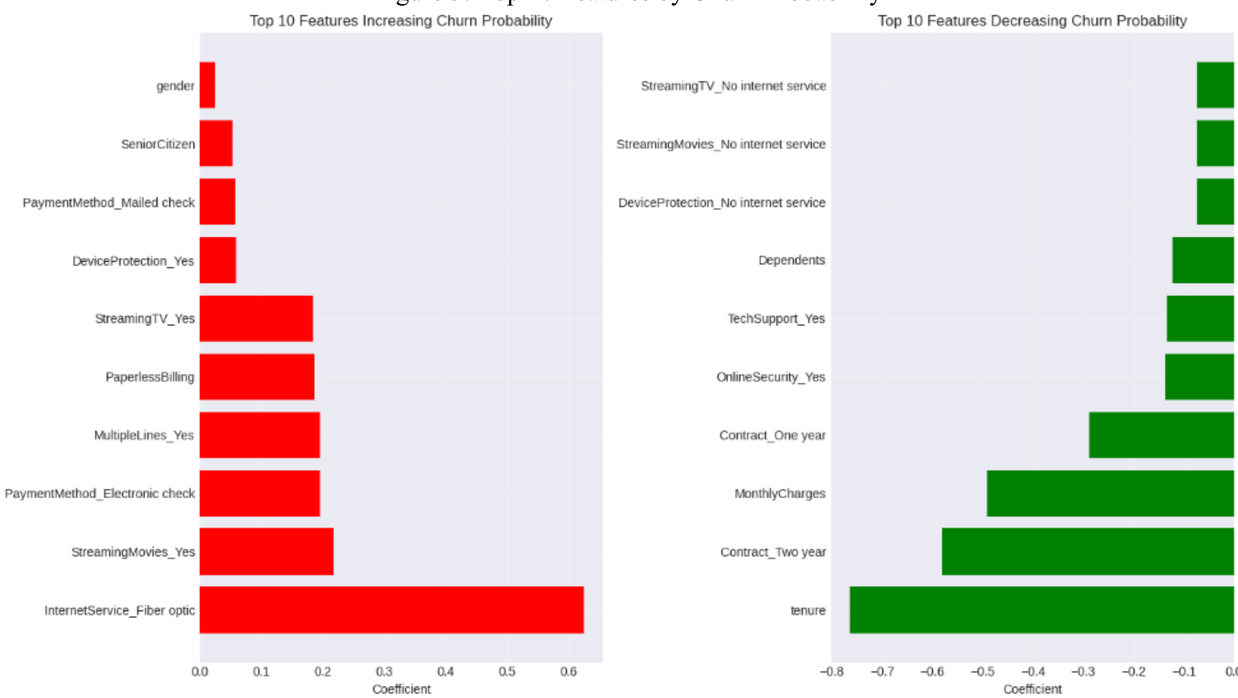
Model Strengths:

- Consistent performance across train (80.5%), validation (81.8%), and test (78.4%) sets indicates good generalization
- High AUC-ROC (83.0%) demonstrates strong ability to discriminate between churners and non-churners
- Strong specificity (88%) effectively identifies retained customers, minimizing false alarms
- Minimal overfitting evidenced by similar metrics across all data splits

Key Business Insights:

- The model can be deployed for prioritizing retention efforts on high-probability churners
- 61% precision means retention campaigns triggered by model predictions will target actual churners more often than not
- 180 missed churners (false negatives) suggest need for complementary rule-based or ensemble approaches
- High specificity prevents retention spending waste on stable customers
- AUC-ROC of 83% indicates model provides valuable customer risk rankings even if binary classifications aren't perfect

Figure 3: Top 10 Features by Churn Probability



Model 2 - Linear Regression for Total Charges Prediction

The ordinary least squares (OLS) linear regression model predicts continuous total charges values using the same 29 features employed in churn prediction. The model learns optimal coefficients through closed-form solution, establishing linear relationships between customer features and accumulated charges.

Performance Metrics Across All Sets

The linear regression model demonstrates exceptional predictive performance with R^2 scores exceeding 90% across all data splits. The consistency between training (91.05%), validation (89.60%), and test (90.77%) performance confirms excellent generalization without overfitting, validating the model's reliability for business deployment.

Training Set Performance:

- R^2 Score: 0.9105 (91.05%)
- RMSE: $688.52
- MAE: $553.30

Validation Set Performance:

- R^2 Score: 0.8960 (89.60%)
- RMSE: $704.64
- MAE: $571.09

Test Set Performance (Final Evaluation):

- R^2 Score: 0.9077 (90.77%) - model explains 90.8% of variance in total charges
- RMSE: $682.10 - root mean squared error; penalizes larger errors
- MAE: $551.42 - mean absolute error; average prediction error
- MSE: $465,260.56 - mean squared error

Results Interpretation

The model's R^2 score of 90.77% indicates that customer features explain approximately 91% of the variation in total charges, demonstrating strong predictive power. The mean absolute error of $551.42 represents reasonable prediction accuracy given the wide range of total charges ($0 to $8,684), translating to approximately 24% average error relative to the mean total charges of $2,280.

Model Strengths:

- Excellent variance explanation (R^2 = 90.77%) indicates strong feature-target relationships
- Consistent performance across data splits (R^2 range: 89.6% - 91.1%) shows minimal overfitting
- MAE of $551 acceptable given target variable range and business tolerance
- RMSE ($682) close to MAE suggests relatively few extremely large errors
- Model suitable for customer lifetime value estimation and revenue forecasting

Figure 4: Residual Analysis

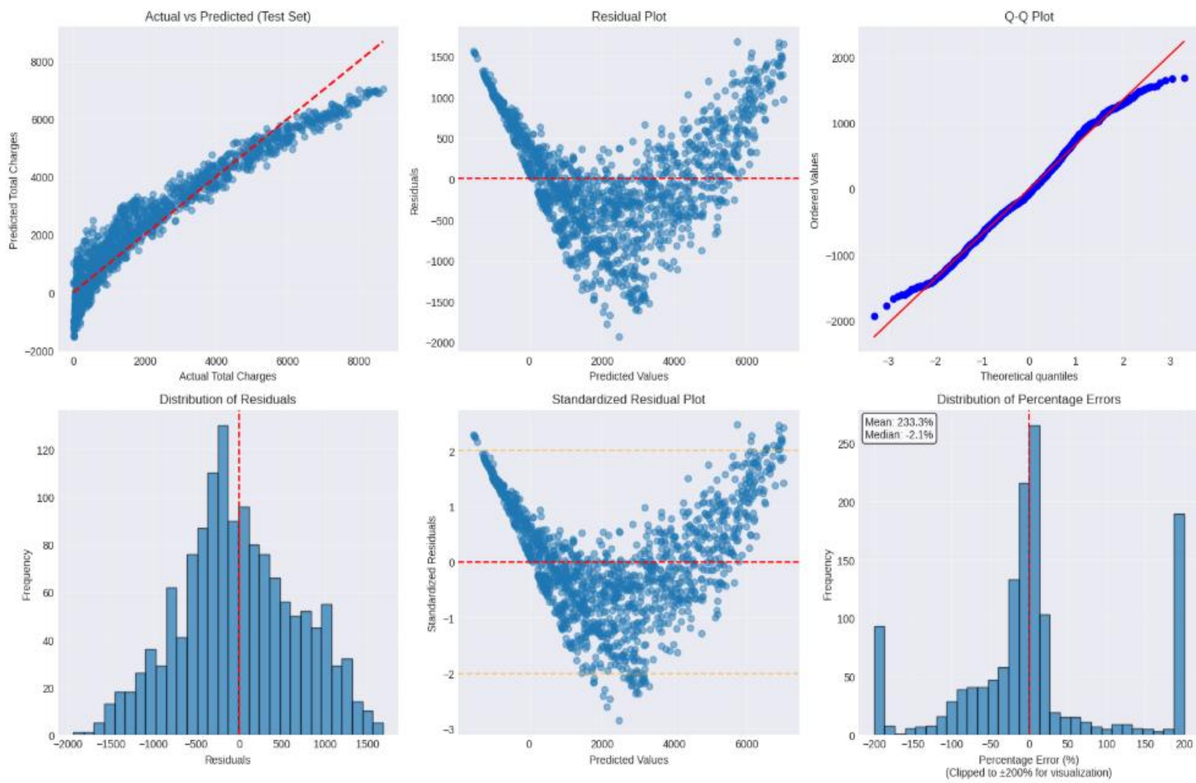


Figure 5: Actual vs Predicted Test Set & Residual Plot

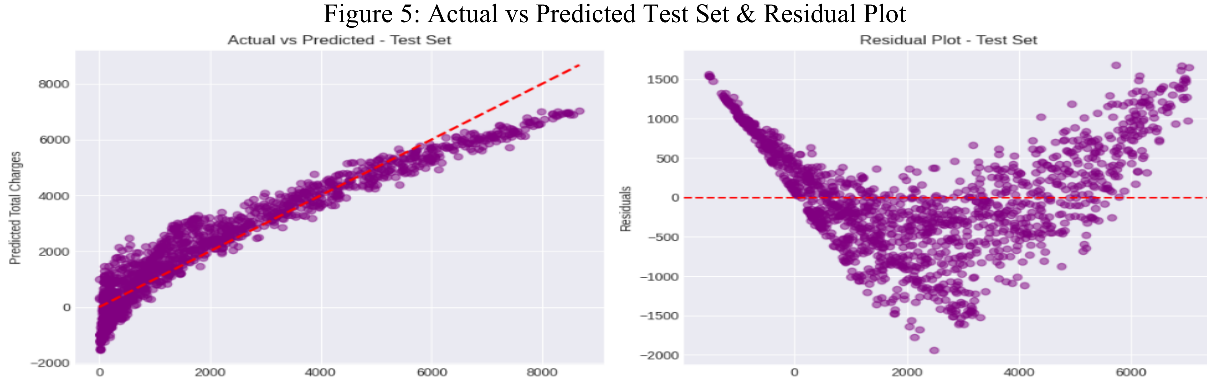
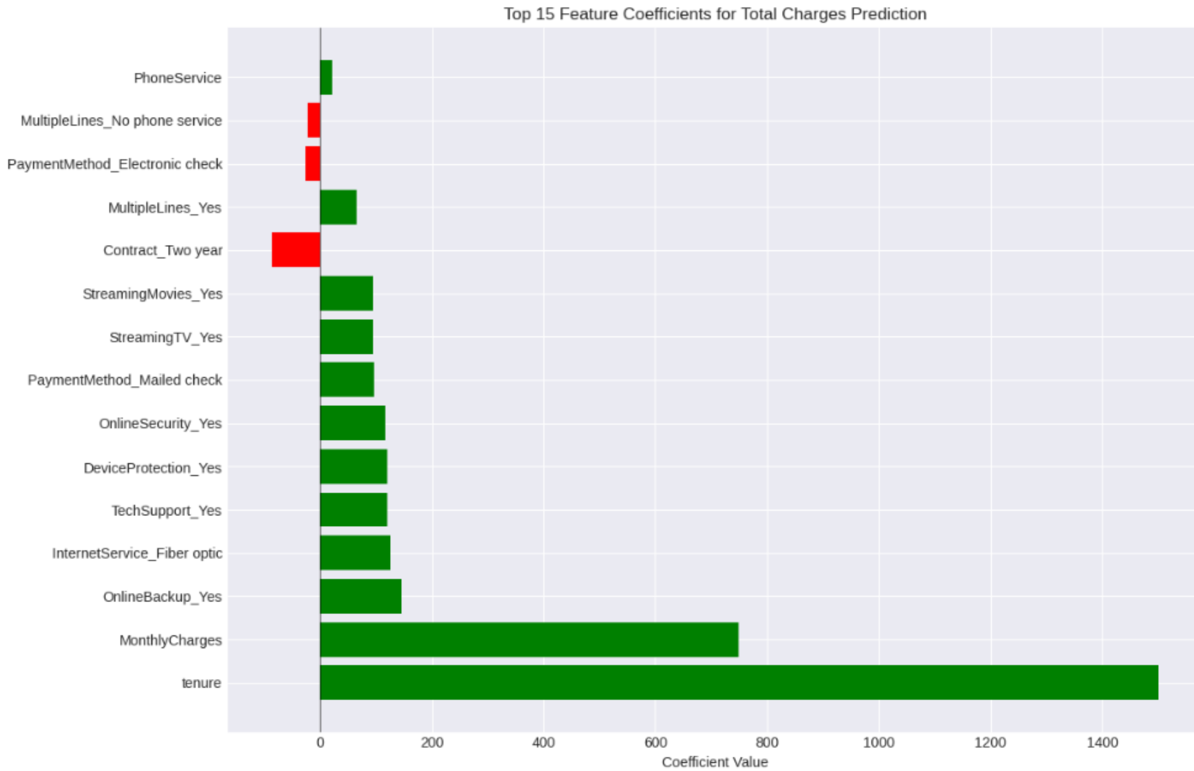


Figure 6: Top 15 Features for Total Charges Prediction



IX. BUSINESS ACTIONABLE STRATEGIES

Business Strategy - 1 - Tiered Value Enhancement Program

Strategic Objective

Reduce churn among high-charge customers (> 75-80 monthly) by enhancing perceived value through differentiated service tiers, exclusive benefits, and personalized experiences without reducing prices, thereby protecting revenue while improving retention.

Implementation Framework

- The company should establish a premium customer experience program that acknowledges and rewards high-value customers through tangible service enhancements rather than price reductions. This approach maintains revenue integrity while addressing the value perception gap that drives premium customer churn.
- The program architecture would include multiple components designed to deliver demonstrable value. First, priority customer support featuring dedicated account managers, 24/7 premium support lines with shorter wait times, and proactive service monitoring would differentiate the high-tier customer experience from standard service.
- Second, exclusive feature access providing early beta access to new services, complimentary premium add-ons (enhanced security packages, premium streaming content, or cloud storage upgrades), and technology refresh programs (device upgrade priority, discounted equipment) would create tangible value additions.
- Third, a loyalty rewards system with accumulating benefits based on tenure and spending, redeemable rewards for bill credits or service upgrades, and milestone recognition (5-year customers receive special benefits) would reinforce long-term relationship value.
- Fourth, flexible service customization allowing contract modifications without penalties, easy service tier adjustments, and personalized service bundles tailored to usage patterns would enhance customer control and satisfaction. The program must leverage the strong negative correlation between tenure and churn (-0.35) by emphasizing relationship longevity and incremental value delivery over time.

Expected Business Outcomes:
This strategy targets approximately 2,500-3,000 customers (35-40% of base) paying above 75 monthly, representing the highest customer lifetime value segment despite elevated churn risk. By enhancing service experience rather than reducing prices, the company protects monthly recurring revenue while improving retention.

Business Strategy - 2 - Contract Migration and Price Optimization Initiative
Strategic Objective:
Transition high-charge, month-to-month customers to longer-term contracts through strategically designed pricing incentives and automatic payment adoption, reducing churn from 42% (month-to-month) to 11% (one-year) or 3% (two-year) while securing predictable recurring revenue.
Implementation Framework:

- The analysis revealed dramatic churn rate differences across contract types: month-to-month contracts exhibit 42% churn, one-year contracts show 11% churn, and two-year contracts demonstrate only 3% churn a 14-fold difference between the highest and lowest risk segments. Simultaneously, payment method analysis showed electronic check users experience 45% churn while credit card automatic payment users show only 15% churn. These patterns indicate that contract structure and payment automation serve as powerful retention mechanisms beyond simple commitment barriers.
- The company should implement a proactive, segmented outreach campaign targeting high-risk, high-value customers with compelling contract conversion offers. The incentive structure would provide meaningful monthly charge reductions (8-12% discount range) for customers committing to one-year or two-year contracts while simultaneously migrating them to automatic payment methods. For example, a customer currently paying 90 monthly on a month-to-month contract with electronic check payment could receive an offer of 82 monthly rate on a one-year contract with automatic credit card or bank transfer payment representing $96 annual savings for the customer while securing 984 in guaranteed annual revenue for the company and dramatically reducing churn risk.
- The pricing optimization should strategically balance three factors: the discount must be sufficient to motivate contract commitment (8-12% range appears competitive), the secured revenue period must justify the discount (12-24 months), and the discount should remain below the cost of customer acquisition for replacement customers (typically 300-500 in telecommunications). The model's feature importance analysis revealed that two-year contracts show a coefficient of -85.86 for total charges, indicating current pricing structures may already incorporate contract-based discounts, but these benefits need greater transparency and visibility during the conversion process.

REFERENCES

[1] K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform," *Journal of Big Data*, vol. 6, no. 1, pp. 1–24, 2019, doi: 10.1186/s40537-019-0191-6.

[2] E. Alpaydin, *Introduction to Machine Learning*, 4th ed. Cambridge, MA, USA: MIT Press, 2020.

[3] K. Coussement and D. Van den Poel, "Churn Prediction in Subscription Services: An Application of Support Vector Machines While Comparing Two Parameter-Selection Techniques," *Expert Systems with Applications*, vol. 34, no. 1, pp. 313–327, 2008, doi: 10.1016/j.eswa.2006.09.038.

[4] De Caigny, K. Coussement, and K. W. De Bock, "A New Hybrid Classification Algorithm for Customer Churn Prediction Based on Logistic Regression and Decision Trees," *European Journal of Operational Research*, vol. 269, no. 2, pp. 760–772, 2018, doi: 10.1016/j.ejor.2018.02.009.

[5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009, doi: 10.1007/978-0-387-84858-7.

[6] T. Jahromi, S. Stakhovych, and M. Ewing, "Managing B2B Customer Churn, Retention and Profitability," *Industrial Marketing*

Management, vol. 43, no. 7, pp. 1258–1268, 2014, doi: 10.1016/j.indmarman.2014.06.016.

[7] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. New York, NY, USA: Springer, 2013, doi: 10.1007/978-1-4614-7138-7.

[8] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York, NY, USA: Springer, 2013, doi: 10.1007/978-1-4614-6849-3.

[9] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, "Customer Churn Prediction System: A Machine Learning Approach," *Computing*, vol. 104, no. 2, pp. 271–294, 2022, doi: 10.1007/s00607-021-00908-y.

[10] E. Stripling, S. vanden Broucke, K. Antonio, B. Baesens, and M. Snoeck, "Profit Maximizing Logistic Model for Customer Churn Prediction Using Genetic Algorithms," *Swarm and Evolutionary Computation*, vol. 40, pp. 116–130, 2018, doi: 10.1016/j.swevo.2017.10.010.