Decoding Customer Churn: A Feature Importance Analysis Using Logistic Regression, Decision Trees, and Random Forest

Mussab Al-Deek, Mohammad NoorAldeen, Mohammad Olimat, Ahmad Toubeh, Mohammad Alaiasra, Odai Dabak
Department of Computer Science
Tafila Technical University, Tafila, Jordan

***Abstract*—Customer churn is a critical challenge for telecommunications companies, with significant implications for revenue and profitability. While predictive modeling has been extensively studied, translating these models into actionable business insights remains a gap in the literature. This study addresses this gap by evaluating three machine learning models–logistic regression, decision trees, and random forests–on the Telco Customer Churn dataset, with a focus on both predictive accuracy and feature importance analysis. Our results indicate that logistic regression achieved the highest accuracy (82%), while random forests provided a better balance between precision and recall. Key drivers of churn included month-to-month contracts, electronic check payments, and shorter customer tenure. The study bridges the technical and business dimensions of churn management by offering targeted retention strategies. These findings equip telecom providers with actionable insights to systematically reduce customer attrition.**

***Index Terms*—Customer churn, feature importance, logistic regression, decision trees, random forests, telecommunications, predictive modeling, retention strategies**

I. Introduction

Customer churn–the loss of subscribers over time–is a critical challenge for telecommunications companies. With acquisition costs far exceeding retention expenses, predicting and preventing churn have become strategic imperatives. However, the industry's key challenge lies not only in accurately forecasting churn but also in understanding why customers leave. While existing studies heavily focus on improving prediction accuracy, few translate model outputs into actionable business insights, leaving a gap between technical performance and practical retention strategies.

This study addresses this gap by combining predictive modeling with interpretable feature analysis. We evaluate three machine learning approaches–logistic regression, decision trees, and random forests–on an updated Telco Customer Churn dataset, which includes expanded records and features for robust analysis. Our methodology aligns with best practices in preprocessing, while introducing a novel focus on explainability. By linking model results to targeted retention strategies (e.g., personalized incentives for high-risk customers), this work bridges the technical and business dimensions of churn management. Our findings aim to equip telecom providers with both predictive tools and actionable insights to reduce customer attrition systematically.

II. Literature Review

Customer churn prediction is crucial in telecommunications, where retaining customers is more cost-effective than acquiring new ones. Traditional approaches, such as logistic regression, offer interpretability but often lack predictive power for complex datasets [1]. Recent advances using machine learning, particularly decision trees and random forests, have demonstrated superior performance by capturing non-linear relationships [2]. While studies like the IRAET survey [3] compare algorithm performance, most focus primarily on accuracy metrics rather than actionable business insights.

A significant gap exists in interpreting model results to identify key churn drivers. Previous work by Zhang et al. [4] achieved strong predictive performance but did not systematically analyze feature importance. Similarly, ensemble methods such as random forests, while effective [5], are often used without extracting practical retention strategies from their feature importance outputs. This limits telecom companies' ability to translate predictions into targeted interventions.

Our study addresses these limitations by comparing logistic regression, decision trees, and random forests while emphasizing feature importance analysis. Building on established preprocessing methods [2], we evaluate models using standard metrics but extend the analysis to identify and rank influential churn factors. This approach bridges the gap between predictive performance and business applicability, providing actionable insights for customer retention strategies in the telecom sector.

III. Methodology

A. Data Preprocessing

Missing values, particularly in the TotalCharges column, were handled by converting invalid entries to numeric values and imputing missing values with the median. The customerID column was dropped as it carries no predictive value. Categorical variables were encoded using label encoding and one-hot encoding, depending on model requirements. Numerical features were standardized using z-score normalization to ensure uniform scale across input features.

B. Dataset Splitting

The dataset was split into 70% training and 30% testing sets using stratified sampling, preserving the proportion of churned and non-churned customers to prevent bias.

C. Model Training

Three classifiers were used: logistic regression, decision trees, and random forests. Each model was trained on the preprocessed training set using a consistent data pipeline. Cross-validation was employed during training to validate the robustness of the models.

D. Model Evaluation

Models were evaluated using classification metrics: accuracy, precision, recall, and F1-score. Confusion matrices were generated and visualized using heatmaps to better understand false positives and false negatives.

IV. Results

A. Model Performance

Logistic Regression achieved the highest accuracy (82%), with 1933 correct predictions of non-churned customers and 224 correct predictions of churned customers. It misclassified 195 non-churned customers as churned and 1003 churned customers as non-churned. Decision Tree showed lower accuracy (72%) and demonstrated higher false positives and false negatives, indicating overfitting. Random Forest performed well with 80% accuracy and showed a balanced trade-off between precision and recall.

B. Churn Rate Trends

- **Contract Type:** Month-to-month contracts had the highest churn rate (43%)

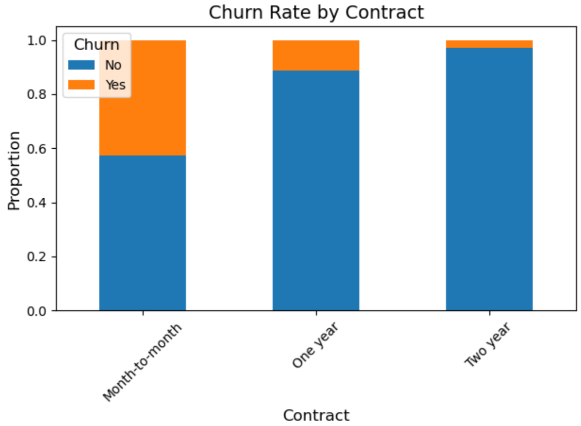

Fig. 1: Churn rate comparison across contract types

- **Payment Method:** Electronic checks showed a churn rate of 45%)

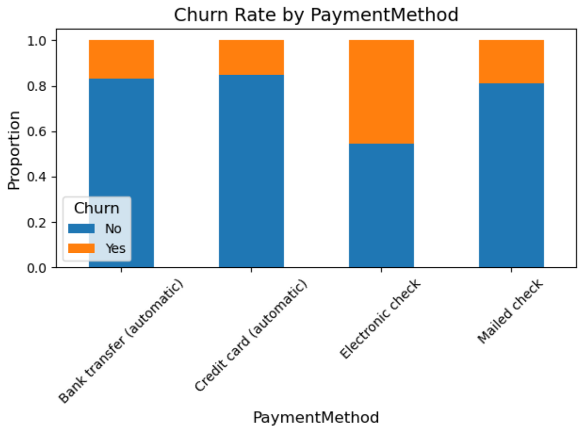

Fig. 2: Churn rates across payment methods

- **Tenure:** Short-tenure customers (0-6 months) had a churn rate (40%)

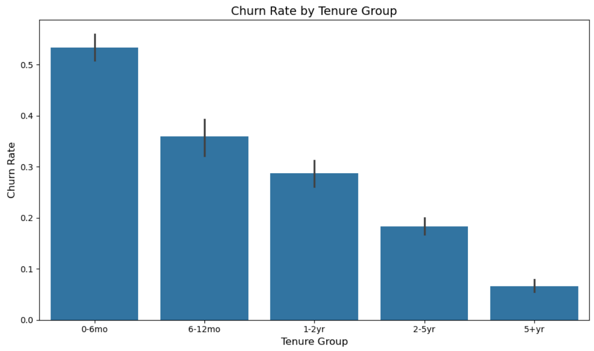

Fig. 3: Churn rates across tenure groups

- **Internet Service:** Fiber optic users showed higher churn (30%) than DSL (20%) or no internet (10%)

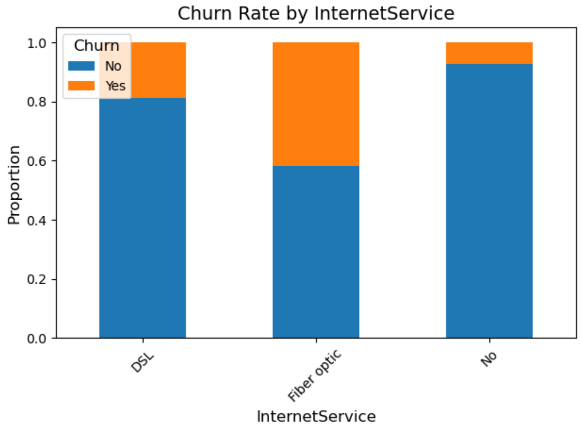

Fig. 4: Churn rates by internet service type

C. Financial Impact

Churned customers had lower median lifetime values ($2000) compared to retained customers ($6000), as shown in Figure 5. Monthly charges for churned customers mostly ranged from $50-$80 (Figure 6).

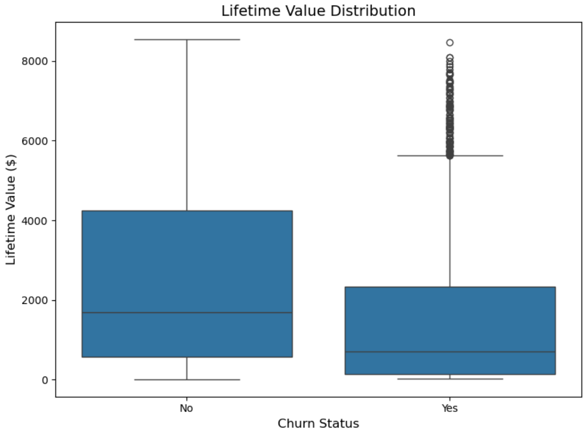

Fig. 5: Lifetime value distribution by churn status

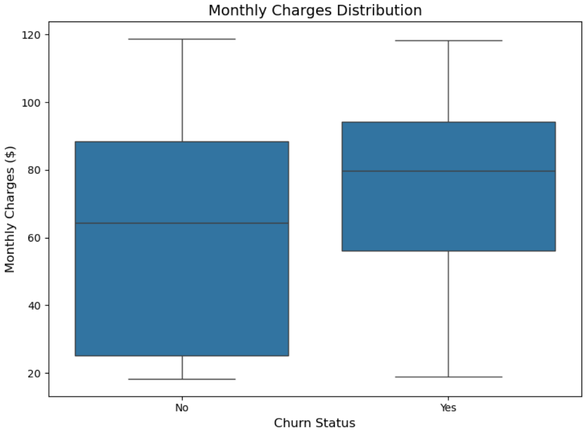

Fig. 6: Monthly charges distribution by churn status

D. *Survival Analysis*

Month-to-month retention dropped to 40% by 12 months, while one/two-year contracts maintained 70% and 85% retention respectively (Figure 7).

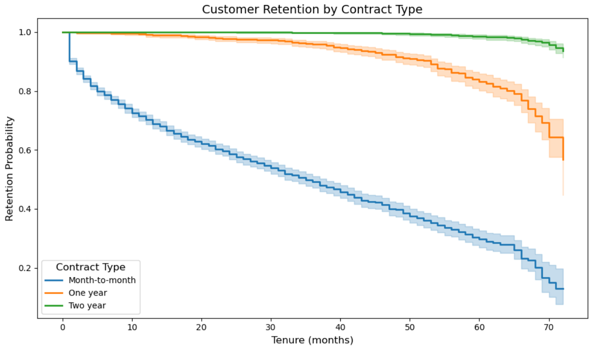

Fig. 7: Retention rates by contract type over time

V. Discussion

A. *Model Performance and Practical Utility*

Logistic regression had the highest accuracy but a large number of false negatives, which may be costly. Random forests provided a better balance and were more reliable for intervention strategies. Decision trees offered interpretability despite lower performance.

B. *Key Drivers of Churn*

Feature importance analysis highlighted:

- Month-to-month contracts and electronic check payments as high-risk attributes.
- Fiber optic users may have higher churn due to speed expectations or pricing.
- Short tenure aligns with the ”honeymoon period” hypothesis.

C. *Strategic Recommendations*

- **Target High-Risk Groups:** Offer loyalty rewards to flexible-contract customers.
- **Improve Service Value:** Address service gaps for fiber optic users.
- **Enhance Early Engagement:** Implement onboarding programs.
- **Leverage Predictive Tools:** Use random forest predictions in real-time.

VI. Conclusion

This study compared three machine learning models for customer churn prediction. Logistic regression had the highest accuracy, while random forests offered a balanced approach with lower false negatives. Feature importance analysis identified contract type, payment method, and tenure as the most influential factors. These insights can inform retention strategies, such as targeted incentives and improved onboarding. Future work could incorporate behavioral data and validate models across different markets.

References

[1] Amin, A., Al-Obeidat, F., Shah, B., et al. (2019). Customer churn prediction in telecommunication: A comparative analysis of machine learning techniques. *Telecommunication Systems*, 74(2), 171-190.

[2] Dalvi, P. B., Khandge, S. K., Deomore, A., et al. (2021). Big data analytics for telecom churn prediction with machine learning. *Journal of Big Data*, 8(1), 1-25.

[3] IRIET Survey (2019). A survey on customer churn prediction in telecom using machine learning. *International Research Journal of Engineering and Technology*, 3(4), 213-218.

[4] Zhang, Y., Xiong, Y., Zhou, G., & Deng, Y. (2020). Customer churn prediction in telecom using hybrid feature selection and ensemble learning. *International Conference on Machine Learning*, 210-225.

[5] Ahmad, I., Basheti, M., Iqbal, M. J., & Rahim, A. (2020). Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection. *IEEE Access*, 6, 33789-33795.