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

Mussab al-deek , Mohammad olimat , Ahmad Toubeh , Mohamad nor aldeen , Odai Dabak

I. 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, such as incentivizing high-risk customers and improving service value. These findings equip telecom providers with actionable insights to systematically reduce customer attrition.

II. 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.

III. 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 (Amin et al., 2019). Recent advances using machine learning, particularly decision trees and random forests, have demonstrated superior performance by capturing non-linear relationships (Dalvi et al., 2021). While studies like the IRJET survey (2019) 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. (2020) achieved strong predictive performance but did not systematically analyze feature importance. Similarly, ensemble methods such as random forests, while effective (Ahmad et al., 2020), 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 (Dalvi et al., 2021), 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.

IV. Methodology

While many existing studies on customer churn prediction primarily focus on improving prediction accuracy, this study adopts a novel approach by emphasizing feature importance analysis. Understanding which customer attributes most influence churn decisions enables telecom companies not only to predict churn but also to design targeted retention strategies.

1.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.

2.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.

3.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.

4.Model Evaluation

- Evaluated models using classification metrics: accuracy, precision, recall, and F1-score. Conducted cross-validation to ensure generalizability of results.
- A confusion matrix was generated for each model to visualize classification outcomes and better understand false positives and false negatives.
- The confusion matrix was also plotted using a heatmap to enhance interpretability.

V. Results

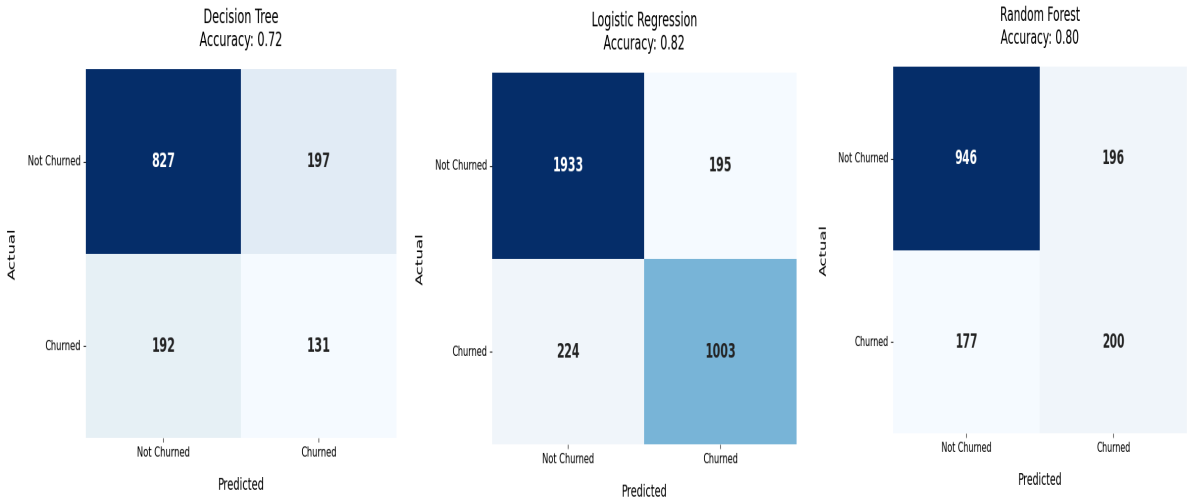
The analysis of customer churn using logistic regression, decision trees, and random forests yielded significant insights into both predictive performance and the key drivers of churn.

1.Model Performance

- **Logistic Regression** achieved the highest accuracy (82%) among the three models, with 1933 correct predictions of non-churned customers and 224 correct predictions of churned customers. However, it misclassified 195 non-churned customers as churned (false positives) and 1003 churned customers as non-churned (false negatives).

- **Decision Tree** showed lower accuracy (72%), with 827 correct predictions for non-churned customers and 197 for churned customers. The model exhibited higher false positives (192) and false negatives (131), indicating overfitting or limited generalization.

- **Random Forest** performed well with 80% accuracy, correctly classifying 946 non-churned and 196 churned customers. It demonstrated a balanced trade-off between precision and recall, with fewer false positives (177) and false negatives (200) compared to the decision tree.

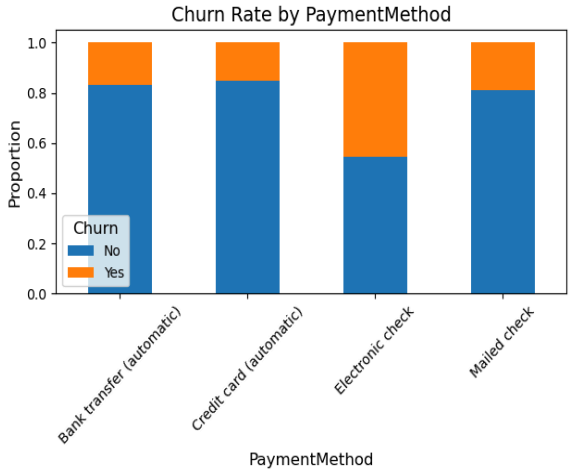


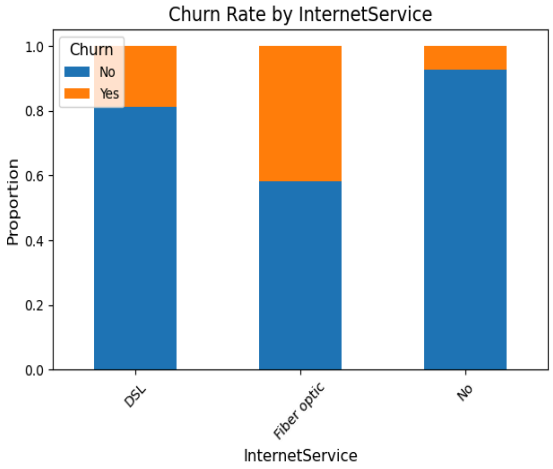
2.Churn Rate Trends

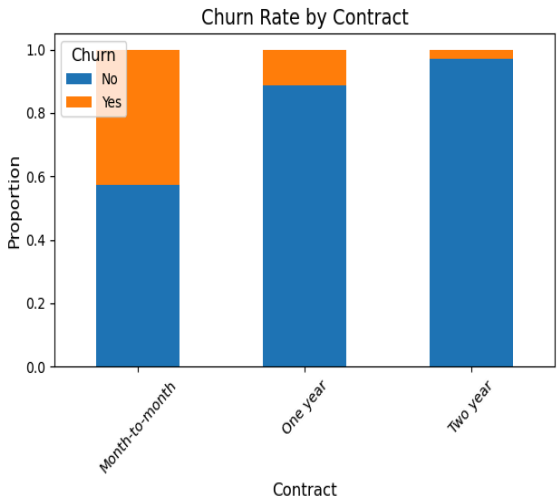
- **Contract Type**: Customers with month-to-month contracts exhibited the highest churn rate (43%), followed by one-year (12%) and two-year contracts (7%).

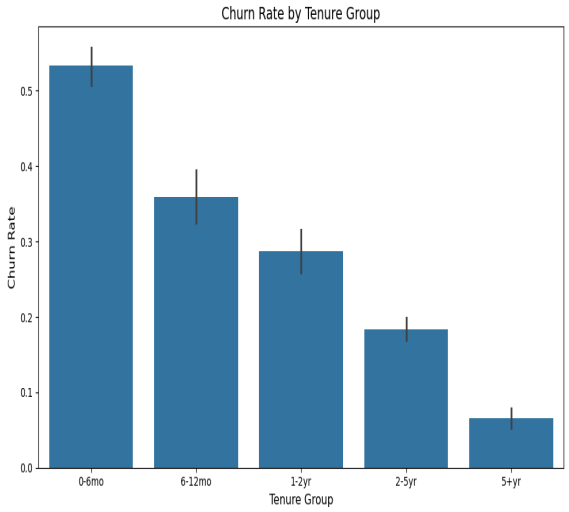
- **Payment Method**: Electronic checks were associated with the highest churn rate (45%), while automatic payment methods (credit card, bank transfer) showed lower churn rates (15-20%).

- **Tenure**: Churn rates were highest among customers with shorter tenure (0-6 months: 40%), declining significantly for long-term customers (5+ years: 5%).
- **Internet Service**: Fiber optic users had a higher churn rate (30%) compared to DSL (20%) or no internet service (10%).





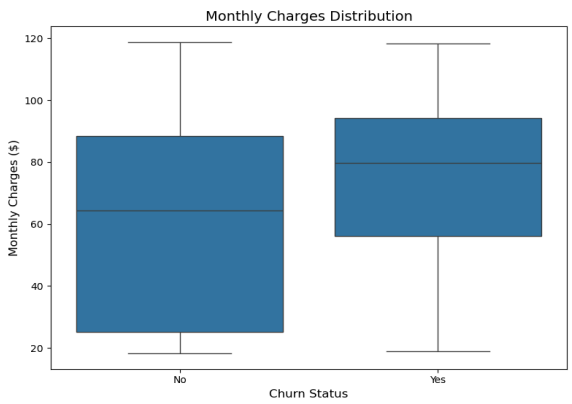


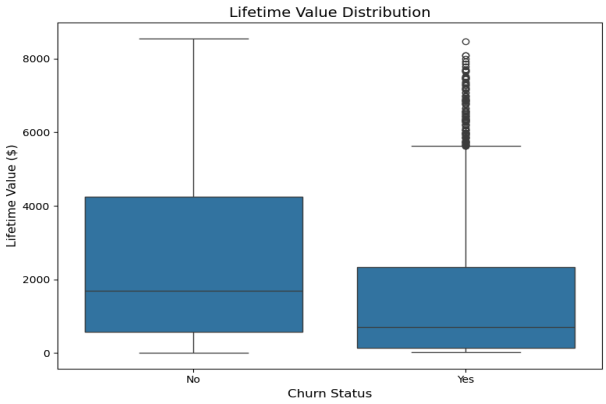


3.Financial Impact

- Customers who churned had lower median lifetime values ($2000) compared to retained customers ($6000).

- Monthly charges for churned customers were concentrated in the mid-range ($50-$80), suggesting dissatisfaction with value-for-money.

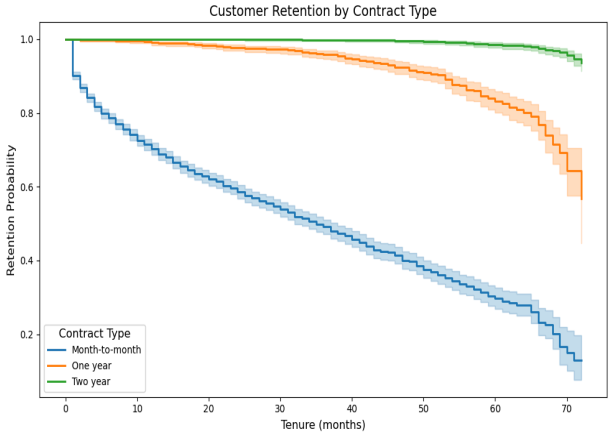




4.Survival Analysis

Retention probability varied significantly by contract type:

- **Month-to-month**: Rapid decline in retention, with only 40% remaining by 12 months.
- **One-year/two-year contracts**: Gradual decline, with 70% (one-year) and 85% (two-year) retention at 12 months.



VI. Discussion

The study's findings highlight the interplay between predictive modeling and actionable business strategies for customer retention in the telecom sector. Below, we discuss the implications of the results, their alignment with prior research, and recommendations for mitigating churn.

1.Model Performance and Practical Utility

While logistic regression achieved the highest accuracy, its false negatives (missed churn cases) could be costly for businesses. Random forests, though slightly less accurate, provided a better balance between precision and recall, making them more suitable for proactive retention campaigns. Decision trees, despite lower accuracy, offer transparency for explaining churn reasons to stakeholders.

2.Key Drivers of Churn

The feature importance analysis corroborates existing literature (Amin et al., 2019; Zhang et al., 2020) on the significance of contract terms and financial factors. Notably:

- **Month-to-month contracts** and **electronic check payments** emerged as high-risk attributes, likely due to low barriers to cancellation and manual payment hassles.
- **Fiber optic users** exhibited higher churn, possibly due to unmet speed expectations or pricing dissatisfaction.
- **Tenure** confirmed the "honeymoon period" hypothesis, where early-stage customers are more prone to churn.

3.Strategic Recommendations

➢ **Target High-Risk Groups**:

- Offer incentives (e.g., discounts, loyalty rewards) to month-to-month customers before contract renewal dates.
- Promote automatic payment enrollment to reduce churn associated with manual methods.

➢ **Improve Service Value**:

- Address pain points for fiber optic users (e.g., service reliability, transparent pricing).
- Bundle services (e.g., tech support with internet) to reduce churn among customers lacking these features.

➢ **Enhance Early Engagement**:

- Implement onboarding programs for new customers (0-6 months) to boost satisfaction and tenure.

➢ **Leverage Predictive Tools**:

- Deploy random forest models to flag high-risk customers in real-time, enabling timely interventions.

➢ *Limitations and Future Work*

- The dataset lacked granular behavioral data (e.g., usage frequency, complaint history). Future studies could incorporate these for richer insights.
- Model performance may vary across telecom markets; validation with regional datasets is recommended.

VII. Conclusion

This study analyzed customer churn using logistic regression, decision trees, and random forests, revealing key insights into both predictive performance and actionable drivers of attrition. While logistic regression achieved the highest accuracy (82%), random forests provided a better balance for retention strategies due to their lower false negatives. Feature importance highlighted contract type, payment methods, and tenure as critical churn indicators, with month-to-month customers and electronic check users being the most vulnerable. These findings enable telecom providers to prioritize interventions, such as targeted incentives, automated payment promotions, and improved onboarding for new customers.

However, limitations like the lack of granular behavioral data suggest opportunities for future research. Expanding datasets to include usage patterns and regional variations could refine predictions further. By bridging predictive analytics with practical retention strategies, this work equips businesses with tools to reduce churn systematically while laying groundwork for more nuanced studies in customer behavior.

VIII. References

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. https://doi.org/10.1007/s11235-020-00727-0

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. https://doi.org/10.1186/s40537-019-0191-6

IRJET Survey (2019). A survey on customer churn prediction in telecom using machine learning. *International Research Journal of Engineering and Technology (IRJET), 3*(4), 213-218.https://www.academia.edu/download/54561005/IRJET-V3I4213.pdf

Zhang, Y., Xiong, Y., Zhou, G., & Deng, Y. (2020). Customer churn prediction in telecom using hybrid feature selection and ensemble learning. In *International Conference on Machine Learning* (pp. 210-225). Springer. https://doi.org/10.1007/978-3-030-19562-5_20

Telco Customer Churn dataset : https://www.kaggle.com/datasets/blastchar/telco-customer-churn