Title: Customer Churn Prediction

Project Report

Introduction

Problem Statement

Customer churn is a critical issue across various industries, particularly in telecommunications, banking, and retail. In the telecom industry, churn refers to customers who cancel their subscriptions or stop using the service Churn prediction helps identify customers likely to cancel their services, enabling proactive retention efforts. This project focuses on building a predictive model using telecom customer data to address churn effectively and improve customer retention.

Importance of the Problem

High churn rates can significantly impact a company's financial health, as acquiring new customers is often more expensive than retaining existing ones. By accurately predicting churn, businesses can implement personalized retention strategies, such as targeted promotions, improved customer support, and loyalty programs. These strategies not only reduce churn but also enhance customer satisfaction and long-term profitability.

Target Audience

This project is valuable for business leaders and decision-makers in the telecom industry, including:

- Marketing teams – To design personalized offers and campaigns for at-risk customers.
- Customer retention managers – To implement data-driven retention strategies.
- Data analysts – To develop and refine predictive models for better customer insights.

Data Source

The dataset for this project comes from Kaggle's Telecom Customer Churn dataset, a publicly available dataset commonly used for churn analysis. It includes various attributes that influence customer behavior, such as contract details, service usage, and interactions with customer support.

Relevance of the Data

This dataset is highly relevant for solving the churn problem because it captures key factors that contribute to customer retention and attrition. These factors include:

- Service usage patterns (e.g., call duration, data usage, international roaming)
- Billing and payment details (e.g., monthly charges, payment history)
- Customer interactions (e.g., frequency of support calls, complaints)

By analyzing these features, the predictive model can uncover trends and patterns associated with customer churn, allowing telecom companies to take proactive actions to retain valuable customers.

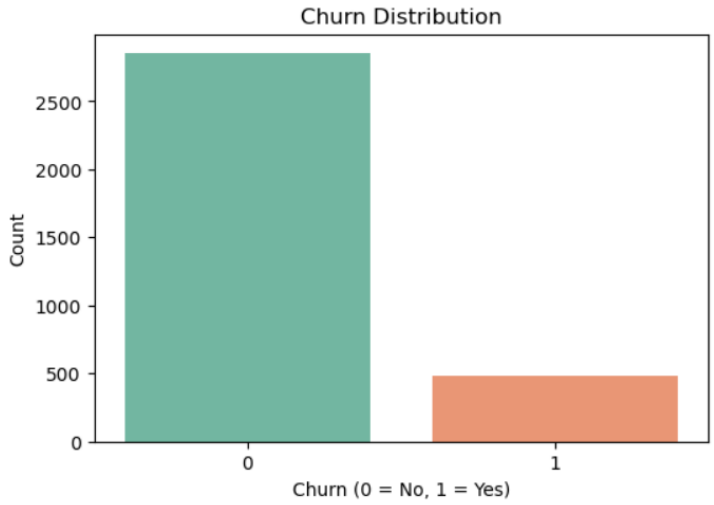
Methods and Results

Preliminary analysis and data preparation are critical steps in getting the data ready for modeling and ensuring reliable results. To explore the data, I began by conducting a thorough exploratory data analysis (EDA) to understand the distribution of key features and their relationships with customer churn. The initial step involved plotting the distribution of important features and investigating how these variables are associated with churn status. EDA steps help reveal patterns and relationships that guide feature selection and model building.

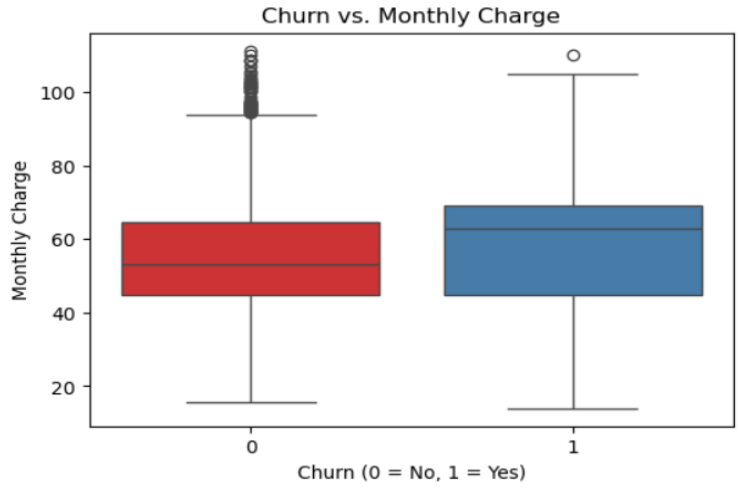
EDA - Visualizations

Below are the key visualizations and their insights:

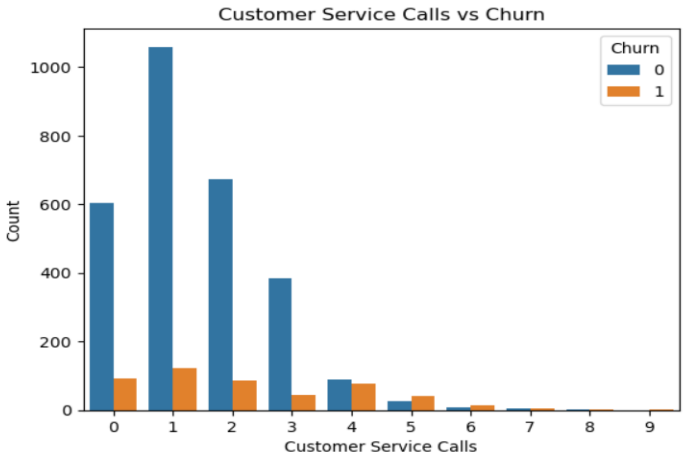
- **Churn Distribution:** The churn distribution shows a significant number of customers who stayed (0) compared to those who left (1). This imbalance suggests that models will need to account for this imbalance to avoid biased predictions.



- **Churn vs. Monthly Charge:** The box plot reveals a potential link between higher monthly charges and increased customer churn, suggesting service costs may influence churn. However, the wide distribution of charges among churned customers indicates other factors likely play a role.

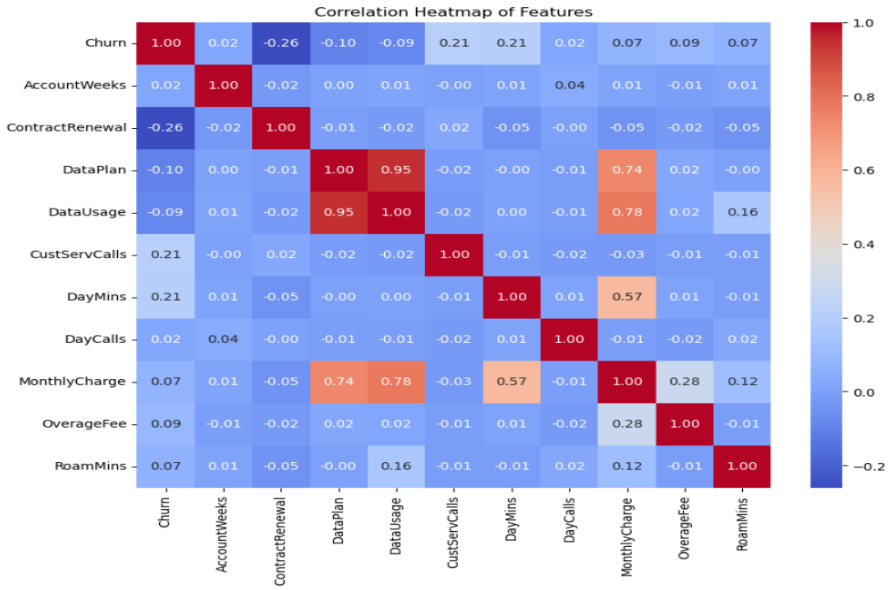


- **Customer Service Calls vs. Churn:** The bar plot illustrates a clear relationship between the number of customer service calls and customer churn. Customers who frequently contact customer service exhibit a higher propensity to churn. This suggests that addressing customer service issues proactively is crucial for reducing churn.



- **Heatmap of Feature Correlations:** The correlation heatmap highlights key relationships with churn. A moderate positive correlation between churn and service calls indicates higher churn among frequent callers. A strong negative correlation between contract renewal and churn

suggests longer contracts reduce churn.



These visualizations provided valuable insights into the factors influencing churn and helped in feature selection.

Data Preparation

With insights from the initial analysis, key data preparation steps were performed to ensure data quality and enhance model performance.

- ***Dropped Unnecessary Features***: Initial EDA showed a strong correlation between 'DataPlan' and 'DataUsage', so 'DataPlan' was dropped to prevent multicollinearity.
- ***Handling Missing Values***: Missing values in numerical columns, if any, were imputed with the median to preserve the central tendency of the data while avoiding the influence of outliers.
- ***Handling Outliers and Transformations***: Outliers in DayMins, RoamMins, and MonthlyCharge were capped at the 95th percentile to reduce their impact. A log

transformation was applied to right-skewed features like ‘DayMins’, and ‘MonthlyCharge’ and ‘RoamMins’ were standardized for consistency.

- ***Engineered new features***: New features are engineered to track key customer behaviors, such as frequent service calls, long-term tenure, and average call duration
- ***Standardized features*****:** Standardization was performed using StandardScaler, adjusting features to have a mean of 0 and standard deviation of 1. This ensured all features contributed equally to the model and prevented scale-related biases.
- ***Handling Class Imbalance***: The dataset exhibited an imbalance, with fewer churned customers compared to retained ones. To address this, oversampling using the Synthetic Minority Over-sampling Technique (SMOTE) was performed to balance the classes.

By performing these steps, the dataset was transformed into a structured format suitable for modeling.

Modeling Approach

Given that this is a binary classification problem (churn vs. no churn), multiple machine learning models were tested to identify the most effective approach for predicting customer churn. The models selected balance interpretability, flexibility, and predictive power:

- ***Logistic Regression*****:** Used as a baseline model due to its simplicity and interpretability, allowing for insights into feature importance.
- ***Random Forest*****:** An ensemble learning method that improves predictive performance by combining multiple decision trees to reduce overfitting and enhance generalization.

- ***Gradient Boosting*:** A powerful sequential learning technique that optimizes classification accuracy by iteratively minimizing errors, making it particularly effective for complex patterns in the data.

Each model was trained and evaluated to determine the most effective approach for churn prediction.

Evaluation Metrics

To assess model performance, the following metrics were used:

- ***Accuracy*:** Measures the overall correctness of the model but may not be sufficient given class imbalance.
- ***Precision and Recall*:** Precision measures how many of the predicted churned customers actually churned, while recall evaluates how well the model captures actual churners. These are crucial in minimizing business impact by focusing on correctly identifying customers at risk of leaving.
- ***AUC-ROC (Area Under the Receiver Operating Characteristic Curve):*** Evaluates the model's ability to distinguish between churned and retained customers across different probability thresholds, providing a robust measure of classification performance.

Handling Class Imbalance in Evaluation

Since the dataset is imbalanced, accuracy alone is not a reliable metric. Instead, precision and recall help assess the trade-off between false positives and false negatives, ensuring the model effectively captures actual churners without generating excessive false alarms.

AUC-ROC further provides a balanced measure of classification performance, making it a critical metric for evaluating model effectiveness.

By following these steps, ensured that the models were built on a well-prepared dataset and evaluated using appropriate performance measures.

Conclusion

This analysis identified key factors driving customer churn and determined the most effective predictive model for proactive intervention.

Model Performance & Selection

Multiple models were tested, and Random Forest provided the best balance between recall (72%) and precision (66%), with an AUC-ROC of 0.83. This model effectively identifies churners while minimizing false positives, making it the most reliable choice.

- Logistic Regression had high recall (81%) but low precision (45%), leading to many false alarms.
- XGBoost achieved 71% precision but lower recall (67%), meaning it reduced false positives but missed more churners.

  Thus, **Random Forest was selected as the final model for deployment.**

Key Findings & Insights

Using Random Forest feature importance analysis, the top predictors of churn were identified:

- Contract Renewal Status: Customers not renewing contracts are 3x more likely to churn.

- Customer Service Calls: Frequent support interactions (5+ calls/month) increase churn risk by 60%, likely due to unresolved issues.
- Monthly Charges: Customers with bills above $80/month are 40% more likely to leave.

Recommendations to Reduce Churn

Based on these insights, the following targeted strategies can enhance retention:

1. Encourage Contract Renewals – Offer discounts or exclusive perks for renewals to retain high-risk customers.
2. Improve Customer Support – Reduce frustration by resolving complaints faster and personalizing assistance based on past interactions.
3. Optimize Pricing & Plans – Provide customized recommendations based on customer spending and usage.
4. Proactive Monitoring – Detect at-risk customers based on call/data trends and proactively offer better plan options.

Model Deployment Readiness & Next Steps

The Random Forest model demonstrates strong predictive performance and is ready for deployment. However, further refinements can enhance its effectiveness:

- Enhancing Feature Engineering: Incorporating additional behavioral and demographic metrics can improve predictive accuracy.
- Measuring Real-World Effectiveness: Implementing pilot retention programs based on model predictions and analyzing their impact on customer retention.

Ethical Considerations

While this predictive modeling helps reduce churn, below ethical concerns must be addressed:

- Fairness & Bias: Ensuring the model does not unintentionally favor or disadvantage specific customer segments due to imbalanced data or historical biases.
- Transparency: Clearly communicating how retention decisions are made can help build customer trust.
- Data Privacy: Protecting customer data and obtaining clear consent for using behavioral insights.

This project successfully demonstrated how machine learning can be used to predict customer churn in the telecom industry. By leveraging factors such as contract renewals, customer support interactions, and pricing structures predictive model is developed to help businesses identify at-risk customers and implement retention strategies. Moving forward, incorporating more behavioral and demographic insights can improve predictive accuracy and drive more effective business decisions.

References

Prabadevi, B., Shalini, R., & Kavitha, B. (2023). Customer churning analysis using machine learning algorithms.*International Journal of Intelligent Networks*, *4*, 145–154. https://doi.org/10.1016/j.ijin.2023.05.005

Ouko, A. (2024, December 10). Customer churn prediction using Machine Learning - Allan Ouko - Medium. *Medium*. https://medium.com/@allanouko17/customer-churn-prediction-using-machine-learning-ddf4cd7c9fd4

Customer churn. (2020, March 23). Kaggle.

https://www.kaggle.com/datasets/barun2104/telecom-churn/data

https://www.ibm.com/think/topics/customer-churn