Telecom Customer Churn Modeling:
A Data-Driven Approach to Reducing Customer Attrition

Staci McDuffie

Bellevue University

DSC 680 – Applied Data Science

Professor Iranitalab

November 20, 2025

Abstract

Customer churn remains a major challenge in the telecommunications industry, where intense competition and high acquisition costs make retaining existing subscribers a top priority. This white paper presents a logistic regression–based predictive modeling approach designed to identify customers who are most likely to terminate service. The telecom churn dataset includes demographic information, billing history, and service usage patterns for each subscriber. After preparing the data through standard cleaning, encoding, and scaling procedures, a predictive model was trained to classify customers as either churned or active. Model evaluation techniques, including confusion matrix interpretation and ROC AUC scoring, demonstrate that a data-driven approach can meaningfully improve the precision of retention strategies. Ethical considerations related to fairness and transparency were evaluated to ensure that the solution adds value without unintended negative consequences for customers. The findings highlight the

importance of predictive analytics in telecom operations and provide recommendations for scaling churn modeling into production.

Business Problem

Telecommunication companies face a saturated market where customers can easily switch providers. Churn has a direct financial impact, reducing recurring revenue while increasing acquisition and marketing costs. Traditional retention strategies are often reactive and applied broadly, resulting in wasted incentives on customers who have no intention of leaving. This project addresses the central business problem of preventing customer loss by identifying individuals at high risk of churn before they disconnect service. By predicting churn behavior in advance, the organization can apply targeted interventions, reduce revenue leakage, and improve the overall customer lifecycle experience. A successful solution must offer accurate predictions while remaining interpretable and operationally feasible for business stakeholders.

Background/History

Predictive analytics has been incorporated into telecom churn strategy for more than two decades. Early churn prevention efforts relied on simplistic business rules such as monitoring tenure length or elevated billing issues. With advancements in machine learning, companies can now leverage large-scale behavioral and account data to improve risk predictions. Logistic regression continues to serve as a widely accepted foundational model in telecom churn prediction due to its transparency and ease of interpretation. More complex models, such as random forests or gradient boosting, can provide improved performance but are often less intuitive to business leaders who must use the results to justify customer-facing actions. This

project builds upon standard industry practices by combining robust data preprocessing with a well-established modeling technique that balances predictive performance with interpretability.

Data Explanation

The dataset used for this project contains individual subscriber records with billing amounts, contract types, service subscriptions, and churn outcomes. The target variable identifies whether a customer discontinued service during the observation period. Before modeling, the dataset was reviewed for missing entries and formatting inconsistencies. Records lacking churn status were removed, and numeric variables such as monthly and total charges were standardized to ensure comparability. Categorical variables, including contract type and payment method, were converted into encoded values that a machine-learning model could process. The dataset was divided into training and testing subsets using stratified sampling to preserve the natural churn rate. These steps ensured that the model learned patterns representative of real customer behavior while reducing the risk of bias or information leakage.

Methods

The churn prediction task was formulated as a supervised binary classification problem. A logistic regression model was trained using a pipeline that combined numeric scaling and categorical encoding into a single workflow. The model's coefficients allowed examination of how each feature influences the likelihood of churn while optimizing for generalization. Although a random forest model was evaluated during exploratory analysis, logistic regression was selected as the primary model due to its interpretability. Model performance was assessed using accuracy, recall, F1 score, and ROC AUC metrics to understand both overall performance and the ability to detect churners specifically. Confusion matrix and ROC curve visualizations were generated to provide further clarity on prediction strengths and weaknesses.

Results

The logistic regression model produced stronger-than-baseline classification performance, demonstrating practical discriminative capability. The confusion matrix showed correctly predicted churners at a significantly higher rate than random guessing would achieve, while the ROC curve confirmed meaningful separation between customers likely to churn and those expected to remain. Feature interpretation showed that short tenure, higher monthly charges, and certain service configuration combinations increased churn risk. These patterns reflect realistic operational challenges in telecommunications, where new customers lacking established loyalty and those experiencing high billing burdens are more likely to discontinue service. The model's performance indicates that even a relatively simple predictive solution can support business decisions with measurable impact.

Analysis

Model results demonstrate that accuracy alone is not an adequate measure for churn modeling, as the churned class represents a smaller portion of the dataset. Instead, recall, ROC AUC, and F1 score provide a more balanced view of success. The model captured key signals associated with churn behavior, giving customer experience teams the ability to identify and proactively engage at-risk individuals. The strong interpretability of the logistic model allows decision-makers to clearly understand "why" a customer has been flagged, supporting deeper trust in prediction-driven processes.

Assumptions

The project assumes that historical churn behavior is reflective of future trends and that billing and service subscription data accurately represent customer interactions. The model is built on the assumption that churned customers can be identified from the patterns contained in

structured data fields. Additionally, it is assumed that customer behavior will remain relatively stable over the prediction horizon and that the business definition of churn aligns with the dataset.

Limitations

The analysis remains limited by the scope of available data. Important churn drivers such as network performance, customer support sentiment, and competitor pricing are not represented. Because the dataset contains many categorical variables, interactions between them may be complex and not fully captured through logistic regression alone. Additionally, class imbalance impacts threshold decisions, and the model requires ongoing recalibration to remain valid in changing market environments.

Challenges

Key challenges in the modeling process included establishing a fair balance between false positives and false negatives, preventing information leakage during preprocessing, and ensuring that model decisions remain explainable to nontechnical business leaders. Engineering a reusable pipeline for transforming and evaluating the data was important to maintain consistency and trustworthy evaluation throughout the experiment.

Future Uses/Additional Applications

Future development could involve integrating uplift modeling to determine which customers are not only at risk but also responsive to retention incentives. Incorporating network quality metrics or textual sentiment from support interactions could improve prediction and fairness. Automated model monitoring would allow ongoing refinement as customer behavior evolves, while advanced ensemble methods could be explored when greater performance justification outweighs interpretability needs.

Implementation Plan

A practical path toward deployment would begin with embedding the churn model into the company's CRM environment where customer records are actively managed. Scores could be refreshed monthly or weekly depending on customer lifecycle dynamics. Customer care representatives would receive risk notifications alongside contextual explanations of top contributing drivers to tailor engagement approaches. Marketing and finance teams would collaborate to quantify revenue preservation through targeted intervention testing. Continuous evaluation would ensure the model remains effective and aligned with business objectives.

Ethical Assessment

Ethical operation was a critical consideration throughout the project. To prevent unfair treatment, model inputs were reviewed to avoid including variables correlated with protected demographic characteristics. Communication about model decisions should be transparent, focusing on providing better support rather than penalizing customers. Data privacy protections must remain a priority to ensure compliance with regulations and foster consumer trust. Overall, predictions are intended to guide helpful outreach, not restrict service.

Conclusion and Recommendations

This project demonstrates the value of applying interpretable machine learning techniques to reduce customer churn in telecommunications. Logistic regression meaningfully improved churn prediction over baseline performance while providing actionable insight into customer behaviors linked to churn. The model's simplicity supports a smooth transition into operational usage with minimal risk of technical resistance. Continued expansion of available data, reinforcement of ethical safeguards, and deployment of iterative retraining processes will

help ensure long-term model effectiveness. Organizations that proactively adopt data-driven retention strategies will be better positioned to protect revenue and strengthen customer loyalty.

References

Verbeke, W., Martens, D., & Baesens, B. (2012). Social Network Analysis for Customer Churn Prediction. Applied Soft Computing, 14, 431 446. Tsai, C.-F., & Lu, Y.-H. (2009). Customer Churn Prediction by Hybrid Neural Networks. Expert Systems with Applications, 36(10), 12547 12553. Kaggle. (n.d.). Telco Customer Churn Dataset. Retrieved from kaggle.com

Appendix

Appendix A — Data Dictionary

Variable	Type	Description
tenure	Numeric	Number of months customer retained service
MonthlyCharges	Numeric	Current monthly billing rate
TotalCharges	Numeric	Cumulative payments during customer lifetime
Contract	Categorical	Contract length (Month-to-Month / 1-Year / 2-Year)
PaymentMethod	Categorical	Customer billing method
InternetService	Categorical	Internet plan subscription
TechSupport	Categorical	Support subscription indicator
Churn	Outcome	Target variable — 1 = churned

Appendix B — Representative Code Snippets

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, roc_auc_score, roc_curve

X = df.drop("Churn", axis=1)
y = df["Churn"]
```

```
X_train, X_test, y_train, y_test = train_test_split(
   X, y, test_size=0.3, stratify=y, random_state=42)

numeric_features = ["tenure", "MonthlyCharges", "TotalCharges"]
categorical_features = [c for c in X.columns if c not in numeric_features]

preprocess = ColumnTransformer(
   [("num", StandardScaler(), numeric_features),
    ("cat", OneHotEncoder(handle_unknown="ignore"), categorical_features)]
)

model = Pipeline(steps=[
   ("preprocess", preprocess),
   ("logreg", LogisticRegression(max_iter=1000))
])

model.fit(X_train, y_train)
```

Audience Questions

1. What were the biggest data quality issues you encountered, and how did you address them before modeling?
The dataset contained missing churn labels and inconsistent formatting in billing fields such as TotalCharges, which occasionally appeared as blanks instead of numeric values. These were corrected by removing records with missing target values and coercing billing fields to numeric types. Standardization of numeric features and one-hot encoding of categorical features ensured consistency throughout model training.

2. Which variables had the strongest influence on churn and why do you think those factors matter operationally?
Short tenure and higher monthly charges were among the strongest churn predictors. Operationally, this suggests that financially strained and newly acquired customers are more likely to leave if they do not perceive service value quickly. Service-subscription indicators such as tech support access also matter, showing that engagement with support offerings can improve loyalty.

3. Why did you choose logistic regression and tree-based models instead of more advanced methods like gradient boosting or neural networks?
Logistic regression provides superior interpretability and clearer explanation of behavioral relationships, which is essential when communicating to executives and frontline customer-service teams. Tree-based models help capture nonlinear interactions and improve predictive performance. Neural networks and advanced boosting models may add complexity and computation cost without proportional transparency benefits at this stage of deployment.

4. How do you ensure the model remains interpretable for business users, especially customer-service leadership?
Logistic regression allows us to express results in terms of odds ratios, clearly showing whether certain behaviors increase or decrease churn risk. In operational dashboards, churn score explanations can accompany predictions, indicating "top reasons" for risk. This transparency builds trust and helps managers make targeted, values-aligned decisions.

5. Accuracy isn't always enough for churn modeling — how did you balance precision and recall and which metric did you prioritize?
Since churners represent the minority class, we emphasized **recall** to maximize identification of customers most at risk of leaving. ROC AUC and F1 score were also monitored to ensure balance between capturing churners and avoiding excessive false positives. This approach helps retain revenue while avoiding unnecessary discounting.

6. What threshold did you use to classify a customer as high-risk, and how did you determine that was the best cutoff?
We used the default probability threshold of **0.50** for classification but evaluated alternative cutoffs using ROC analysis and stakeholder priorities. In practice, the optimal threshold should be tuned based on desired trade-offs — for example, lowering the threshold when retention budgets allow for a broader safety net to catch more churners.

7. How would you operationalize this model inside a telecom company's workflow? What teams would use it and how?
The model would feed churn-risk scores into the CRM system on a recurring schedule. Customer-care agents could prioritize outreach to high-risk customers, marketing teams could tailor retention offers, and finance could estimate revenue-at-risk. Continuous monitoring would allow automated actions such as offering service upgrades or billing adjustments.

8. How do you prevent biased or unfair treatment of customers when using churn predictions for retention campaigns?
We review all features to avoid proxies for protected characteristics such as race, age, or socioeconomic status. Fairness testing would be performed to monitor disparate impact. Retention strategies are supportive rather than punitive: customers are offered assistance and added value rather than restricted service.

9. What important variables are missing from this dataset that might improve the accuracy or fairness of future models?
Service quality indicators (e.g., network outages, latency), sentiment signals from customer-support interactions, and competitive landscape factors such as pricing changes are not present. Including these would capture external churn drivers and reduce reliance on cost-based assumptions that could unfairly bias certain groups.

10. If you had more time or resources, what enhancements or additional techniques would you apply to improve the churn model?

I would expand model comparison to include gradient-boosting methods such as XGBoost and implement uplift modeling to identify which customers are most likely to respond positively to a retention campaign. I would also deploy automated retraining with drift detection to maintain performance as customer behaviors evolve.