Customer Churn Prediction for a Telecom Company Using Machine Learning

By
Nitin Dukare (Data Analyst)

Business Problem –

A telecom company is experiencing customer attrition (churn), which leads to revenue loss and increased customer acquisition costs.
The company wants to identify customers who are likely to churn in advance so that proactive retention strategies can be applied.

Business Questions –

Churn Prediction (WHO will churn?)

- Which customers are most likely to churn from the telecom service?
- Can high-risk customers be identified in advance using historical customer data?

Churn Drivers (WHY do they churn?)

- What are the key factors influencing customer churn?
- Does contract type significantly impact customer churn?
- Does customer tenure affect churn behavior?
- Is there a relationship between monthly charges and customer churn?
- Does the type of internet service influence churn rates?
- Does payment method have an impact on customer churn?

Business Action (SO WHAT?)

- How effectively can machine learning models predict customer churn (using recall and ROC-AUC)?
- How can churn prediction help the business take proactive retention actions?

Steps Performed in the Customer Churn Prediction Project

Step 1: Data Understanding

- Loaded the customer churn dataset and reviewed its structure.
- Identified the total number of records (6,000) and columns.
- Identified Churn as the target variable.

- Reviewed each column to understand its business meaning and relevance.
- Assessed the dataset for overall suitability for churn prediction.

Step 2: Data Exploration

- Identified categorical and numerical variables.
- Analyzed value counts for categorical variables such as Gender, Internet Service, and Payment Method.
- Examined the distribution of the target variable (Churn).
- Checked class imbalance between churned and non-churned customers.
- Observed initial data patterns and distributions.

Step 3: Data Cleaning

- Checked for missing values and confirmed that no missing data was present.
- Checked for duplicate records and confirmed there were none.
- Removed the non-informative customerID column.
- Standardized column names by renaming gender to Gender and tenure to Tenure.
- Encoded the target variable Churn into binary format (0 = No, 1 = Yes).
- Verified data types to ensure consistency after cleaning.

Step 4: Exploratory Data Analysis (EDA)

- Analyzed churn distribution across different contract types.
- Analyzed churn behavior across internet service types.
- Analyzed churn behavior across payment methods.
- Compared average tenure between churned and non-churned customers.
- Compared average monthly charges between churned and non-churned customers.
- Identified key patterns and relationships influencing churn.

Step 5: Feature Engineering

- Identified categorical features requiring transformation.
- Applied one-hot encoding to convert categorical variables into numerical format.
- Used drop_first=True to avoid multicollinearity.
- Separated the feature set (X) and target variable (y).
- Prepared a final model-ready dataset consisting only of numerical features.

Step 6: Model Building and Evaluation

- Split the dataset into training and testing sets using an 80–20 ratio.
- Built a Logistic Regression model as a baseline.
- Built a Random Forest model to capture non-linear relationships.

- Generated predictions and churn probabilities.
- Evaluated models using Accuracy, Recall, and ROC-AUC.
- Focused on Recall as the primary business metric due to the cost of missing churned customers.
- Identified limitations related to class imbalance.

Step 7: Business Question Answering

- Identified high-risk churn customers using model prediction probabilities.
- Evaluated the ability of models to identify churn customers in advance.
- Determined key churn drivers using feature importance.
- Assessed the impact of contract type, tenure, monthly charges, internet service, and payment method on churn.
- Compared model performance to determine effectiveness.
- Explained how churn prediction can support proactive customer retention strategies.

Insights

- The dataset shows a moderate churn rate (~29%), which reflects a realistic business scenario where most customers are retained.
- Monthly Charges and Tenure emerged as the most important features from the Random Forest model, indicating that pricing and customer duration influence churn when combined with other factors.
- Tenure alone does not strongly differentiate churn behavior, as the average tenure of churned and non-churned customers is nearly the same.
- Monthly charges alone also do not strongly drive churn, since both churned and retained customers have similar average monthly charges.
- Contract type impacts churn behavior, with churn rates varying across month-to-month, one-year, and two-year contracts, indicating that contract structure influences retention.
- Internet service type shows only minor variation in churn rates, suggesting that service type alone is not a strong churn driver.

- Payment methods show small differences in churn rates, indicating limited standalone impact on churn behavior.
- Feature importance analysis confirms that churn is driven by a combination of multiple factors rather than a single variable.
- Logistic Regression failed to identify churn customers due to class imbalance, despite showing higher accuracy.
- Random Forest performed better than Logistic Regression in identifying churn customers, achieving a higher recall, though overall recall remains low.
- Accuracy alone is misleading for churn prediction, and recall is the most critical metric for this business problem.
- The model successfully identifies high-risk customers with high churn probabilities, enabling targeted retention actions.
- Current model performance indicates that early churn identification is possible but limited, highlighting the need for further model tuning or class balancing.
- Churn prediction can help businesses prioritize retention efforts, reducing unnecessary costs by focusing only on high-risk customers.

Business Recommendations

- **Focus on High-Risk Customers Identified by the Model**
  Customers with high churn probability (≥ 0.7) should be prioritized for retention efforts, as targeting these customers is more cost-effective than applying strategies to the entire customer base.
- **Implement Proactive Retention Strategies**
  The business should proactively reach out to high-risk customers with personalized offers, discounts, or service improvements before they decide to leave.
- **Review Pricing and Value Proposition Together**
  Since monthly charges alone do not strongly drive churn, pricing should be reviewed in combination with service quality and customer experience rather than offering blanket price reductions.
- **Improve Contract-Based Retention Policies**
  Variations in churn across contract types indicate that contract structure influences retention. Encouraging customers to move toward longer-term contracts may help reduce churn.

- **Enhance Customer Engagement Across Tenure Groups**
  Both new and long-term customers appear in the high-risk group. Retention programs should therefore target customers across different tenure levels, not only new customers.
- **Improve Model Performance for Better Churn Detection**
  The current models show low recall due to class imbalance. Applying techniques such as class weighting, threshold tuning, or resampling can improve churn identification and business impact.
- **Use Recall as the Primary Performance Metric**
  Business decisions should focus on recall rather than accuracy, as missing churn-prone customers has a higher cost than incorrectly flagging non-churn customers.
- **Integrate Churn Predictions into Business Workflow**
  Churn prediction outputs should be integrated into CRM systems so that customer support and marketing teams can take timely, data-driven actions.