**Churn Prediction Analysis**

Barath Anandaraman

MS in Data Science, Bellevue University

DSC 630 T301: Predictive Analytics

Milestone5 Final Project

Professor Frank Neugebauer

May 30, 2025

**Churn Prediction Analysis**

**Problem Statement:**

Telecom and Cable companies face multiple challenges in the current market with customer churn. Churn refers to the rate at which customers stop using the services and move to a competitor. Predicting customer churn allows companies to proactively identify customers at risk of leaving, enabling them to implement targeted retention strategies.

**Importance and Need:**

* Customer acquisition is expensive, while retaining existing customers is significantly cheaper. Minimizing churn directly translates to increased revenue and profitability.
* By understanding the factors that lead to churn, companies can improve their services and offerings to better meet customer needs.
* Churn prediction allows for focused marketing campaigns towards at-risk customers, offering personalized incentives to keep them engaged.

**Overview of Data Used:**

The dataset selected originates from a HackerEarth (Syed, 2021) churn risk competition. It contains 36,992 rows and 24 variables, covering customer demographics, service usage, complaints, feedback, and engagement indicators, like:

* The target variable churn\_risk\_score is already binary (0 = low risk, 1 = high risk).
* No null values were present, though some preprocessing (e.g., object-type conversion) was necessary.
* Key features include points\_in\_wallet, membership\_category, avg\_transaction\_value, and feedback.

**Data Preparation & Feature Engineering**

* Dropped non-informative columns (security\_no, referral\_id, joining\_date, etc.).
* Encoded categorical variables using LabelEncoder.
  + Data before EncodingA screenshot of a computer code

    AI-generated content may be incorrect.
  + Data after EncodingA white background with black numbers and letters

    AI-generated content may be incorrect.
* Impute missing numeric fields with mean using Simple Imputer.
* Engineered meaningful features:
  + age\_group binning by age groups.
  + high\_spender with avg\_transaction\_value above its mean.
  + active\_user based avg\_frequency\_login\_days that is above its mean.
  + Tenure based on number of years from 2025.
  + wallet\_x\_transaction with points\_in\_wallet times avg\_transaction\_value.

A screenshot of a computer code

AI-generated content may be incorrect.

**Exploratory Data Analysis**

* Histograms of Numerical columns. **A group of graphs showing the value of a graph

  AI-generated content may be incorrect.**
  + avg\_time\_spent right skewed but has negative values.
  + avg\_transaction\_value, right skewed with possible outliers.
  + avg\_frequency\_login\_days also has negative values.
  + points\_in\_wallet bell shaped with some outliers.
* Box plots**A close-up of a graph

  AI-generated content may be incorrect.** A comparison of a diagram

  AI-generated content may be incorrect.
  + avg\_transaction\_value and points\_in\_wallet tend to decrease as churn risk increases.
  + avg\_frequency\_login\_days tend to increase as churn risk increases.
* Correlation matrices helped identify feature relationships. **A screenshot of a graph

  AI-generated content may be incorrect.**
  + avg\_frequency\_login\_days shows mild positive correlation with churn risk.
  + points\_in\_wallet and avg\_transaction\_value show negative correlation.
  + age, days\_since\_last\_login, avg\_time\_spent have weak correlation.
* Bar plots A close-up of a graph

  AI-generated content may be incorrect. A close-up of a graph

  AI-generated content may be incorrect. A comparison of a number of blue bars

  AI-generated content may be incorrect. A graph of a certain amount of data

  AI-generated content may be incorrect.
  + membership\_category: Customers with No Membership show higher churn risk on average.
  + joined\_through\_referral: those who did not join through referral tend to have higher churn.
  + Customers who did not use discounts have slightly higher churn.
  + offer\_application\_preference, those who dont prefer offers churn more.
  + Higher complaint history tends to churn more.
  + Unsolved complaints are strongly associated with churn.
  + Negative feedback has higher churn as expected.
* Scatter plots A group of blue and orange dots

  AI-generated content may be incorrect.
  + Cluster of low churn scores exists in high spending and high engagement regions.
  + Customers with lower spending and less time spent are more likely to have higher churn risk.
  + Higher churn risk tends to have fewer login days.
  + Active users with more wallet points have lower churn risk.

**Top Contributors to churn**

1. Points In Wallet - Strong predictor of churn
2. Membership Category - Plays a significant role
3. Average Transaction Value - Tied to engagement impacts churn
4. Feedback - Customer sentiment
5. Average Time Spent - Indicates how active a user is.

**Model Building:**

Trained three classification models:

1. Logistic Regression Model
2. Decision Tree
3. Random Forest

### **Modeling & Evaluation Summary**

#### **1. Models Trained**

##### Trained and evaluated three classification models

##### **Logistic Regression** (Linear, interpretable)

**Findings:** A basic LogisticRegression model is used. It's a good baseline model due to its interpretability.

##### **Decision Tree Classifier** (non-linear, rule based)

**Findings:** A DecisionTreeClassifier is initialized with random\_state=42 for reproducibility. Decision trees are prone to overfitting, and typically parameters like max\_depth or min\_samples\_leaf would be tuned to control this.

##### **Random Forest Classifier** (ensemble of decision trees for improved accuracy)

**Findings:** A RandomForestClassifier is used with n\_estimators=100 and random\_state=42. Random Forest is an ensemble method that generally performs well and reduces overfitting compared to a single Decision Tree.

**2. Feature Importance based on each Model trained**A screenshot of a graph

AI-generated content may be incorrect.A graph with a bar chart

AI-generated content may be incorrect.

#### Top 3 important features based on Decision Tree and Random Forest are points\_in\_wallet, membership\_category, avg\_transaction\_value, while the top 3 in Logistic Regression are avg\_frequency\_login\_days, membership\_category, feedback.

#### **3. Cross-Validation Results**

##### Used 5-fold cross-validation to assess model robustness and generalization using F1 Score (weighted)

##### **Findings:**

##### - Random Forest consistently outperformed the others across all folds.

##### - Logistic Regression underperformed, possibly due to inability to model complex patterns.

| **Model** | **F1Score (Mean)** | **F1Score (Std)** |
| --- | --- | --- |
| Logistic Regression | ~0.69 | Higher Variance |
| Decision Tree | ~0.91 | Low Variance |
| Random Forest | ~0.93 | Very Low Variance |

#### **4. Confusion Matrices**

##### A blue squares with white text AI-generated content may be incorrect.

##### - Random Forest and Decision Tree had strong diagonals - they correctly predicted most churn vs non-churn cases.

##### - Logistic Regression misclassified a larger number of churn cases (low recall).

##### **Findings:**

##### Random Forest minimized false negatives – critical in churn prediction, where missing a churner is costlier than flagging a loyal customer.

#### **5. Classification Metrics**

##### Evaluated models using Precision, Recall, F1Score and ROC AUC

##### **Findings:**

##### - Random Forest: Excellent across all metrics, particularly Recall (0.94) and ROC AUC (0.97), it is very good at correctly identifying churners.

##### - Logistic Regression: Lower recall suggests many true churners are being missed

| **Model** | **Precision** | **Recall** | **F1 Score** | **ROC AUC** |
| --- | --- | --- | --- | --- |
| Logistic Regression | 0.69 | 0.79 | 0.73 | 0.74 |
| Decision Tree | 0.92 | 0.90 | 0.91 | 0.90 |
| Random Forest | 0.92 | 0.94 | 0.93 | 0.97 |

#### **6. ROC Curve Analysis** (GeeksforGeeks, 2025)

##### - Random Forest's curve hugged the top-left indicating strong separation between churn and non-churn classes.

##### - Decision Tree performed well, but slightly under Random Forest.

##### - Logistic Regression had a curve closer to diagonal, indicating weaker performance.

##### **Findings:**

##### Random Forest had the highest AUC and the most desirable ROC curve shape.

##### A graph of different colored lines AI-generated content may be incorrect.

**Ethical Implications:**

While the datasets are publicly available and suitable for educational purposes, it's essential to acknowledge the broader ethical considerations surrounding churn prediction. The models created will be built from customer data, and if deployed in a real-world application, it is important to consider the following.

* **Data Privacy:** Even with publicly available data, the deployed model will be handling customer data. It is important to make sure that any data used is anonymized, and that all data privacy laws are followed.
* **Bias:** The model may inherit biases present in the training data, potentially leading to discriminatory outcomes. For example, the model might unfairly predict churn for specific demographic groups. We will implement fairness-aware modeling techniques and conduct thorough bias analysis to mitigate these risks.
* **Transparency:** Customers should be informed about how their data is being used for churn prediction. We will advocate for transparent communication practices and ensure that customers have control over their data.
* **Data Security:** Data used in the model will be stored and handled with care, so that no data breaches occur.
* It is vital that any deployment of this model is done with the best interest of the customer in mind.

**Expected Outcome:**

The primary expected outcome of this project is the development of a robust binary classification model capable of accurately predicting customer churn within a telecommunications company. This model will generate a clear binary output: 'Churn' (indicating a high probability of customer departure) or 'No Churn' (indicating a low probability). This output will be derived from a comprehensive analysis of various customer-related variables, including service usage patterns, billing details, and customer support interactions. The goal is to provide company leadership with actionable insights, enabling them to proactively implement targeted retention strategies. By accurately identifying at-risk customers, the company can optimize resource allocation, personalize customer engagement, and ultimately reduce churn rates, leading to increased customer lifetime value and overall business profitability.

**Potential Applications:**

The idea is to have a system in place on a telecommunications company that can be used to tag the customers in potential risk of being churned out. Once deployed, the model will continuously analyze incoming customer data, providing real-time churn risk assessments. Customers identified as high-risk will be flagged within the system, triggering automated or manual interventions. These interventions may include:

* Personalized offers and discounts tailored to individual customer needs.
* Proactive customer support outreach to address potential issues.
* Targeted communication campaigns highlighting the value of continued service.
* Implementation of loyalty programs to incentivize customer retention.
* This system will allow for a more efficient and data-driven approach to customer retention, enabling the company to focus its resources on the customers most likely to churn, thereby maximizing the impact of its retention efforts.

**Potential Risks:**

A significant potential risk is the limited regional classification information within the chosen datasets. This deficiency could result in the model's inability to capture regional variations in customer behavior, which may be critical predictors of churn. For example, service quality, competitive offerings, and economic conditions can vary significantly across regions, influencing customer churn rates. Without sufficient regional data, the model's accuracy and generalizability may be compromised. To mitigate this risk, following steps will be taken:

* Explore external data sources that may provide supplemental regional information.
* Conduct sensitivity analysis to assess the model's performance across different regions.
* Clearly document the limitations of the model due to the lack of regional data.

**Recommendations**

1. **Launch a Wallet bonus campaign**
   1. Customer with low wallet balances is more likely to churn.
   2. Offer loyalty credit, point-matching, or wallet recharge bonuses to high-risk customers.
2. **Segment and Engage Based on Membership**
   1. Customers with no or low tier memberships churn more.
   2. Provide proactive upgrade incentives or early engagement for non-members.
3. **Targeted Outreach Using Feedback**
   1. Prioritize customers who gave negative feedback for retention campaigns or follow-up support.
4. **Real-Time Scoring dashboard**
   1. Integrate the trained Random Forest model into a CRM system to flag high-risk customers daily.
   2. Setup automated alerts for customer support teams to intervene.

#### **Conclusion**

##### **Random Forest** is the best performing model

* 1. Among the models tested, Random Forest delivered the highest performance across all key metrics like F1 Score, ROC AUC, and confusion matrix balance.
  2. It also exhibited low variance across folds, indicating strong generalization and robustness.
* **Key Churn Drivers identified**

The most important features contributing to churn are:

* + - Points\_in\_wallet – Low wallet balances correlate with higher churn.
    - Membership\_category – Certain membership types (lack thereof) drive churn risk.
    - Avg\_transaction\_value and feedback – provide insight into engagement and satisfaction levels.
* **Visualization Supports Interpretability**
  1. Bar Plots, Scatterplots clearly show how behavioral and transactional patterns differ across churn outcomes,
  2. Feature importance visualizations provide business insights for customer retention strategies.
* **Dataset was Model Ready** with minor adjustments
  1. The churn label was already binary.
  2. Minimal preprocessing (encoding, missing value imputation) and feature engineering were required to enhance model performance.

References

Syed, H. (2021). *Predict the Churn Risk Rate*. HackerEarth. https://www.hackerearth.com/problem/machine-learning/predict-the-churn-risk-rate-11-fb7a760d/

GeeksforGeeks. (2025, April 28). *How to plot ROC curve in Python.* GeeksForGeeks. <https://www.geeksforgeeks.org/how-to-plot-roc-curve-in-python/>