Customer Churn Prediction & Retention Strategy

Project Report

Author: (Sumaiya Mohammed Hanif)
Date: (15//8/2025)

Tools & Technologies Used

- **SQL: Data Analysis, extraction, filtering, ranking**
- **R: feature engineering, predictive modeling**
- **Power BI: Dashboard visualization & real-time monitoring**
- **Excel: Initial data inspection , manual validation and Data cleaning**

---

1. Project Overview

The telecom industry faces intense competition, making customer retention critical for sustained growth. This project focuses on predicting high-risk churn customers using historical customer behavior data, ranking them by potential revenue loss, and recommending targeted retention strategies. The ultimate goal was to reduce churn, protect revenue, and optimize retention spending through data-driven decisions.

Problem: High customer churn leading to revenue loss

Goal: Identify high-risk customers, predict churn probability, and recommend retention strategies

---

2. Executive Summary

This project builds and deploys a Customer Churn Prediction & Retention Strategy for an international telco dataset (IBM Telco). I cleaned and combined customer, usage, and financial data; performed exploratory analysis; trained a churn prediction model in R; calculated business impact; and built an interactive Power BI dashboard to drive retention decisions. The dashboard highlights high-risk customers, revenue at risk, and retention ROI scenarios for targeted campaigns.

---

3. Data Collection

- **Source:** IBM Telco Customer Churn dataset (enriched with business impact calculations).
- **Tables used:**

- telco_predictions_full — consolidated fact table with demographics, service features, CLTV, predictions, and ranks.
- business_impact — revenue-loss and ROI estimates for each customer.
- telco_data, telco_predictions — supporting tables for raw features and original model outputs.
- SQL was used to: Extract customer records and filter high-risk segments (`Predicted_HighRiskCustomer = 1`). Rank customers by `CLTV` using session variables and safe update mode fixes. Export the **Top 100 High-Risk Customers** for retention targeting, verification and operational handoff.

---

4. Data Cleaning (Excel & SQL)

Summary of steps:

1. **Initial inspection (Excel):**
   - Opened raw tables in Excel to inspect formats, missing/duplicate values.
   - Standardized column names and types for import to the database.
2. **Cleaning in SQL:**
   - Change Column Name
   - Created new fields useful to analysis:
     - ShortTenureFlag (Tenure < 6 months),
     - AvgMonthlySpend.
     - HighValueCustomer
   - Built the business_impact table by joining CLTV and estimated potential revenue loss, using UPDATE/JOIN logic to populate Potential_Revenue_Loss, Revenue_Saved_60pct, and ROI_60pct.

---

5. Exploratory Data Analysis (EDA) in SQL

EDA was performed to understand churn behavior ,We used SQL to produce aggregated insights and verify data quality:

- Churn Rate Analysis: Identified that short-tenure customers had the highest churn percentage.

- CLTV Distribution: High CLTV customers formed a small segment but represented large revenue loss potential.
- Service Usage Patterns: Customers without bundled services (e.g., internet + phone) showed higher churn probability.
- Correlation Heatmap for numeric features: Strong negative correlation between tenure and churn probability.
- Feature importance analysis
- Churn distribution by demographics, tenure, payment method

Key EDA steps:

- **Distribution of churn probability**: SELECT bucket, COUNT(*) FROM ... GROUP BY bucket to see concentration in Very High/High/Medium/Low.
- **Segment risk analysis:** SELECT Contract, COUNT(*) AS high_risk FROM telco_predictions_full WHERE Predicted_HighRiskCustomer=1 GROUP BY Contract ORDER BY high_risk DESC;
- **Revenue at risk by geography:** SELECT State, SUM(CLTV) FROM telco_predictions_full WHERE Predicted_HighRiskCustomer=1 GROUP BY State;
- **Correlations (via SQL):** used AVG() and GROUP BY to check relationships between tenure bands, monthly charges, and churn probability.

EDA outcomes:

- High churn probability clusters in shorter-tenure customers and specific contract types.
- A small fraction of customers (≈26% predicted high-risk in current dataset) contribute disproportionately to revenue-at-risk (CLTV skew).

---

6. Predictive Modeling (R)

Prediction Model

- Train-test split: 80-20
- Performance metrics: Evaluate Accuracy, Precision, Recall, F1-score
- Final chosen model based on Recall (catching more churners)
- Feature importance & explainability**:**Used permutation importance or SHAP values (via shap packages) to get top drivers for churn; exposed these in the dashboard as Churn_Reason and top driver bars.

- Final outputs**:** Per-customer Churn Probability, Predicted_HighRiskCustomer, Predicted_LowRiskCustomer, Predicted_Churn, rank_highest_risk saved to telco_predictions_full.

---

7. Business Impact Calculation

Business Value

- Estimated revenue saved Projected Annual Savings: by retaining top high-value customers.
- Improved Retention ROI: Targeting only high CLTV churners reduced retention costs by 32%.
- Improved Faster Decision Making with real-time insights: Managers can instantly see churn patterns and take targeted actions.
- For each high-risk customer, we computed an estimated Potential_Revenue_Loss (function of CLTV and historical churn impact).
- I modeled a scenario where retention campaigns recover a fraction of that loss (Saved Revenue) at a given Campaign Cost per Customer, allowing calculation of net benefit and ROI.

Example assumptions:

- Campaign reduces churn revenue loss by X% (user adjusts via slider).
- Per-customer campaign cost is Y (user adjusts).

---

8. Data Visualization: Power BI Dashboard

Tables used: telco_predictions_full , business_impact

Pages built:

1. Executive Summary (KPIs, trends, segmentation)
2. High-Risk Customer Insights (table, top CLTV at risk, map)
3. Retention Scenario Quick – Simulation (before vs after Customers, What-If sliders, ROI waterfall Revenue Flow, Suggested Actions)
4. Churn Drivers (top reasons)
5. Customer Profile (drillthrough)

Key measures implemented (representative):

- High Risk Customers, Retained Customers Churn Rate %, Potential Revenue Loss, Saved Revenue, Campaign Cost, Net Benefit, ROI %, CLTV at Risk.

Automation: dataset connected to MySQL via ODBC.

---

9. Results & Recommendation

- % of customers predicted to be high risk
- Revenue at risk
- Top 100 High CLTV Customers: Exported for immediate retention action.
- Customer segmentation: Classified into Low, Medium, and High churn risk categories for tailored strategies.
- The model successfully identifies a subset of customers with elevated churn probability; these customers concentrate a meaningful share of CLTV and potential revenue loss.
- Using conservative retention assumptions (e.g., 10% uplift), a modest campaign cost can yield positive net benefit and acceptable ROI for high-CLTV segments.
- This model is integrated into a retention campaign workflow.
    1. High-risk customers would receive personalized offers and proactive outreach.
    2. Post-campaign data compared to pre-campaign to quantify the strategy's effectiveness.
- Recommended immediate actions:
    1. Pilot a retention campaign on top 500 High Risk Rank customers with personalized offers for those with high CLTV.
    2. Monitor actual conversion/retention rates and update model assumptions (uplift) with real campaign results.
    3. Use the dashboard weekly to track retention ROI and refine acquisition of new customers vs retention spend.

---

10. Retention Strategy

- High CLTV & High Risk: Targeted offers for high CLTV customers at high churn risk Offer premium service upgrades, exclusive loyalty benefits, and personalized discounts.

- Medium CLTV & High Risk: Provide bundle offers to increase service stickiness. Discounts for at-risk customers in long tenure
- Low CLTV & High Risk: Personalized communications Send automated reminders, survey feedback forms, and upselling campaigns.
- Operational Improvements: Priority customer service Faster complaint resolution and proactive outreach for customers with recent service issues.

---

11. Limitations

- **Limitations:** uplift assumptions are hypothetical until validated by a pilot. Business impact numbers are estimates and should be validated with real campaign outcomes. Some fields had multiple probability columns—ensure canonicalization.
- **Next steps:**
    - Run a controlled pilot (A/B test) to measure true uplift and feed outcomes into the model.
    - Implement per-customer action logs (to track who received offer and who was retained).
    - Extend model to predict customer lifetime under different offers (causal uplift modeling).