Customer Churn Prediction and Survival Profitability Analysis

using Random Survival Forests (RSF) and Streamlit Dashboard

Sohan Ghosh
M.Sc. in Data Science & Artificial Intelligence, University of Calcutta
Email: sohan562001@gmail.com

Objective

The objective of this project is to build an end-to-end analytical system that predicts **when** a customer is likely to churn and quantifies the **financial benefit of retention actions**. The work integrates **statistical survival analysis**, **machine learning (Random Survival Forest)**, and **business profitability modeling (NPV)** into an interactive **Streamlit** web app.

Phase 1 — Data Understanding & Preprocessing

Dataset: Contained over 10,000 customer records with attributes such as `Age`, `CreditScore`, `Balance`, `Salary`, `NumOfProducts`, `IsActiveMember`, and `Geography`.

Steps:

- Cleaned missing values and outliers.
- Converted data into survival format: `duration` = time until churn/censoring, `event` = churn flag (1 = churned, 0 = active).
- Applied quantile-based binning for numeric features (e.g., Salary, Credit Score).
- Encoded categorical variables using one-hot encoding.
- Train-test split: 75% train, 25% test.

Output: `processed_churn_step1_with_bins.csv`, `processed_churn_step1_ohe.csv`

Interpretation: A customer with `duration=24`, `event=0` stayed 24 months and remains active (right-censored).

Phase 2 — Exploratory Data Analysis (EDA)

Performed detailed visualization using histograms, KDE plots, correlation heatmaps, and countplots.

Key Insights:

- Customers with **CreditScore ¡ 500** and **Salary ¡ 50k** showed higher churn.
- **Inactive members** had nearly twice the churn rate.
- The **Germany** region exhibited higher churn tendency.

Inference: Both financial stability and engagement behavior are significant churn drivers.

Phase 3 — Survival Analysis (Kaplan–Meier & Cox PH)

Kaplan–Meier Estimator

Estimated the probability of customers staying over time. At 12 months, survival probability $S(t) = 0.78 \Rightarrow 78\%$ likely to stay, 22% likely to churn. Steep drop during the first 6–10 months indicates **early-life churn**.

Cox Proportional Hazards Model

Measured influence of factors on churn timing:

Variable	Hazard Ratio	Interpretation
IsActiveMember = 0	2.3×	Inactive customers twice as likely to churn
Salary_Q1	1.8×	Lowest salary group has higher churn risk
CreditScore ¡ 400	1.5×	Poor credit linked to faster attrition

Conclusion: Cox PH confirms that activity and financial health strongly affect churn probability.

Phase 4 — Random Survival Forest (RSF) Modeling

Why RSF: Handles nonlinear interactions and non-proportional hazards effectively.

Modeling Steps:

- Trained RSF using `scikit-survival`.
- Tuned parameters (`n_estimators`, `max_depth`, `min_samples_split`).
- Evaluated using **Harrell's C-index**.

Results:

- Training C-index = 0.83
- Testing C-index = 0.81

Interpretation: C-index = 0.81 means 81% of random pairs were correctly ranked by churn timing, reflecting excellent model discrimination.

Prediction Example: If $S(12) = 0.68$, then $p_{churn,12} = 1 - 0.68 = 0.32$ (32% chance of churn in 12 months).

Phase 5 — Profitability & Targeting (NPV Model)

Objective: Convert churn probabilities into actionable profit estimates.

User Inputs (via Streamlit sliders):

- Monthly revenue = 500
- Offer cost = 200
- Retention uplift = 25%

Formula:

$$INV = (p_{churn,H} - p_{churn,H}(1 - uplift)) \times (Revenue \times 12) - OfferCost$$

Example:

Customer	p_{churn}	INV ()	Action
A	0.60	+80	Target
B	0.10	-150	Skip

Insight: Focus retention on customers where INV ¿ 0 for maximum ROI.

Phase 6 — Model Explainability

Feature Importance (RSF):

Feature	Importance	Interpretation
IsActiveMember	0.23	Inactivity major churn driver
CreditScore	0.18	Lower score ⇒ higher risk
Salary_Q1	0.14	Low salary groups unstable
Geography_Germany	0.11	Regional churn variation
Age	0.08	Mid-age customers churn more

Conclusion: Financial and engagement indicators dominate churn behavior.

Phase 7 — Streamlit Dashboard

Purpose: Create an interactive web-based analytics tool.

Functionalities:

- Upload trained RSF model (`rsf_best.pkl`) and processed dataset.
- Adjustable sliders for revenue, offer cost, and uplift.
- Real-time churn probability and profitability visualization.
- Kaplan–Meier curve (overall & grouped) and feature importance chart.
- Downloadable CSV of predictions and PDF summary report.

Impact: Converts analytical results into accessible business intelligence for decision makers.

Example Business Insight

By offering a 200 incentive to customers with more than 40% churn probability, the company can retain approximately 25% of them, yielding an incremental net value of 57,000 over a 12-month horizon.

Overall Interpretation Summary

Perspective	Insight
Statistical	Early churn occurs mostly within 10 months
Behavioral	Low-salary, inactive customers drive attrition
Predictive	RSF achieves C-index of 0.81 (high accuracy)
Economic	Target positive INV customers to ensure ROI
Visual	KM and feature charts clarify churn patterns

Results Overview

Metric	Value
Training C-index	0.83
Testing C-index	0.81
Average p_churn (12 mo)	0.41
Mean INV / Customer	57.8
High-Value Segment	3.2× higher expected INV

Practical Applications

- Telecom, Banking, Subscription retention systems
- CRM-based customer prioritization
- Early churn alert generation
- Real-time profitability dashboards

Tech Stack

Category	Tools Used
Programming	Python 3.13 (VS Code / Jupyter / Colab)
Libraries	pandas, numpy, matplotlib, seaborn, lifelines, scikit-survival, joblib
Modeling	Kaplan–Meier, Cox PH, Random Survival Forest
Deployment	Streamlit 1.32, ReportLab for PDF export
Version Control	Git & GitHub

Conclusion

This project demonstrates the complete data-science pipeline: from **EDA → Survival Analysis → Machine Learning → Business Modeling → Deployment**. With a C-index of **0.81** and an interactive **Streamlit dashboard**, it effectively merges analytical depth with real-world business value.

Predict not only who will churn, but when, why, and how much it costs to lose them.