Project Protocol

Bank Customer Churn

Almog Cohen

Introduction

Customer churn prediction is a common machine learning problem, but in real-world settings, prediction alone is insufficient. Organizations do not act on probabilities - they act on **priorities, costs, and expected outcomes**.

This project builds an end-to-end **Customer Churn Risk Scoring and Action Optimization system** using anonymized customer data from a multinational bank. Rather than stopping at binary churn classification, the system is designed to support **decision-making** by:

- Assigning individual churn risk scores
- Ranking customers by urgency and likelihood to churn
- Translating model outputs into concrete retention actions
- Evaluating the financial impact and efficiency of those actions

The result is a practical, decision-oriented pipeline that bridges the gap between machine learning outputs and business execution. The project emphasizes **risk-based prioritization**, **cost-aware evaluation**, and **operational realism**, making it suitable as a blueprint for real-world deployment rather than a purely academic exercise.

Objectives

What am I trying to find out?

- How can churn predictions be transformed into actionable, ranked risk scores?
- Which customers should be targeted, with which retention strategy, and at what cost?
- How much churn can be prevented when interventions are applied selectively rather than uniformly?

What do I already know?

- Retaining customers is significantly cheaper than acquiring new ones, but intervening on every customer is inefficient.
- High churn probability alone does not imply that intervention is always optimal - cost and expected lift matter.
- In business settings, **ranking quality and calibration** are often more valuable than a single classification threshold.

What am I aiming to achieve?

- Build a churn modeling pipeline that supports **prioritization and decision-making**, not just prediction.
- Compare multiple classification models and select one based on probability quality and ranking power, not accuracy alone.
- Design a policy layer that maps risk levels to concrete retention actions.
- Quantify the financial impact of different intervention strategies using cost and outcome simulations.

What factors influence the results?

- The dataset is anonymized and represents a snapshot rather than longitudinal behavior.

- Success rates and action costs are estimated assumptions rather than observed outcomes.
- Model performance depends on probability calibration and ranking stability, not only raw metrics.
- Business conclusions are sensitive to chosen thresholds and cost definitions.

Data Preparation

Data Preparation (Concise)

The dataset contains 10,000 customer records combining demographic, behavioral, and account-level information. Prior to analysis, the data was cleaned and standardized to ensure consistency and prevent leakage.

Non-informative identifiers (e.g., customer IDs, surnames, row indices) were removed, and the churn indicator (Exited) was isolated as the target variable. Feature types were normalized: binary flags were cast to boolean values, categorical attributes standardized, and numeric variables preserved in their native formats.

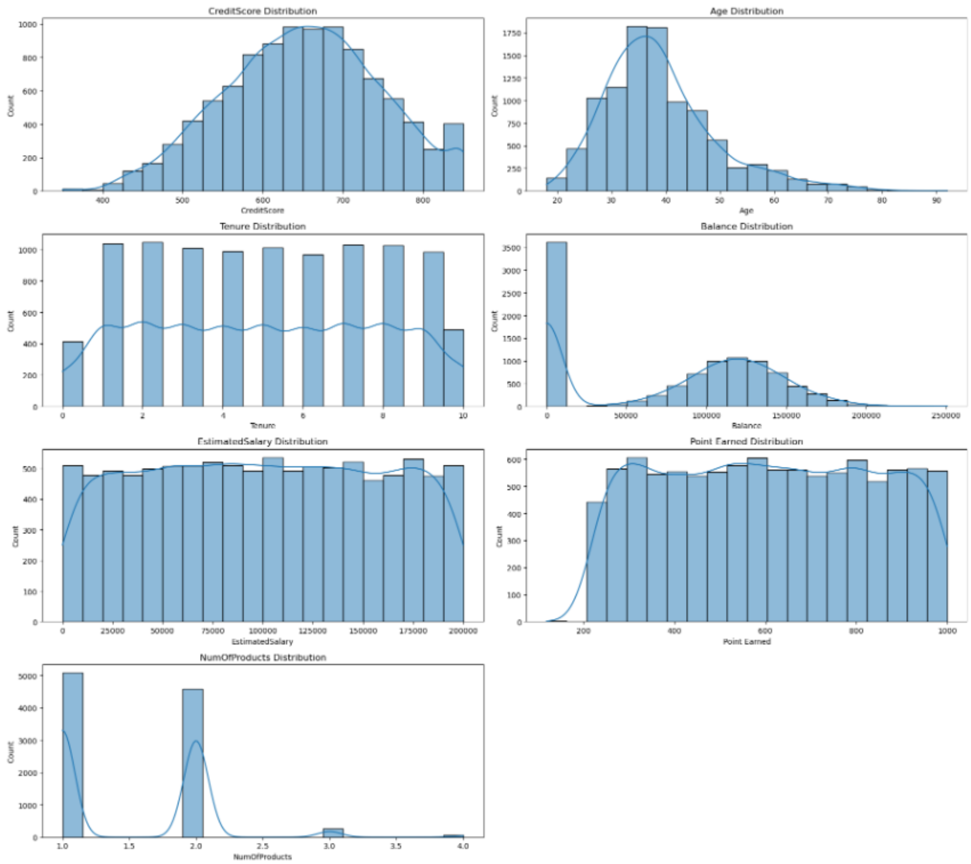
After preprocessing, the dataset consists of a clean, fully populated feature set covering customer demographics, engagement behavior, product usage, and financial context. The target variable is moderately imbalanced (~80% non-churn vs. ~20% churn), reflecting realistic churn dynamics and informing downstream evaluation choices.

Exploratory Data Analysis (EDA)

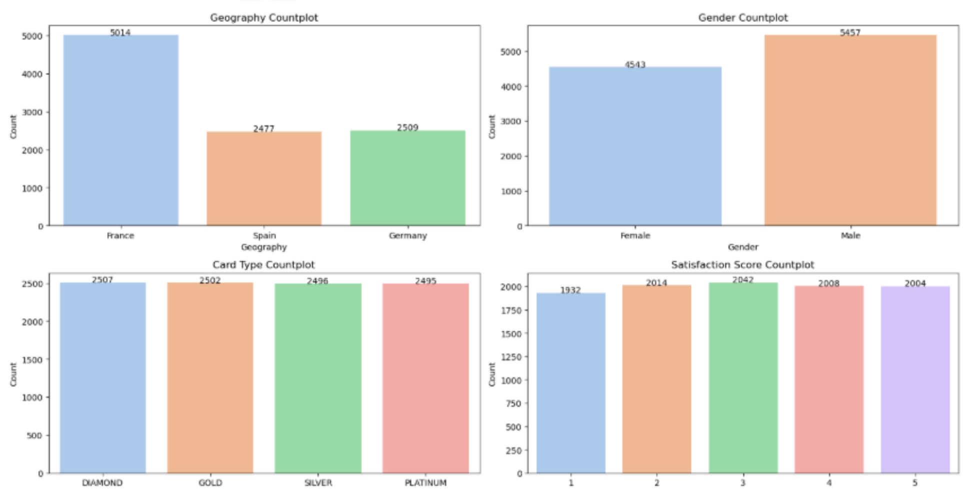
EDA was conducted to validate data integrity, assess feature behavior, and identify patterns relevant to churn. Rather than exhaustive analysis, this section highlights the key structural insights supported by visual evidence.

Feature Distributions

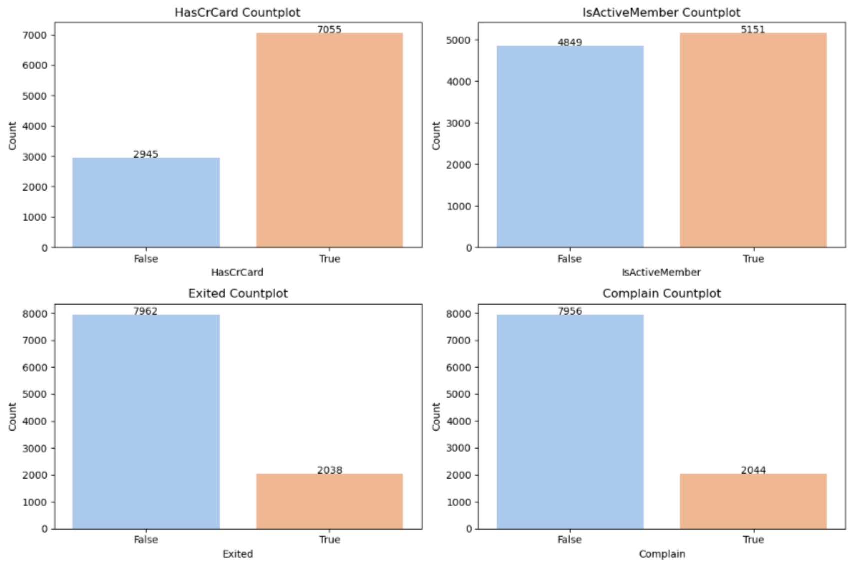
Numerical and categorical feature distributions show generally well-behaved shapes, with limited skewness across most financial and engagement variables. Age exhibits right-skew, suggesting non-linear effects across lifecycle segments. No anomalies or data quality issues were observed.



CreditScore Distribution
Count
CreditScore
Age Distribution
Count
Age
Tenure Distribution
Count
Tenure
Balance Distribution
Count
Balance
EstimatedSalary Distribution
Count
EstimatedSalary
Point Earned Distribution
Count
Point Earned
NumOfProducts Distribution
Count
NumOfProducts

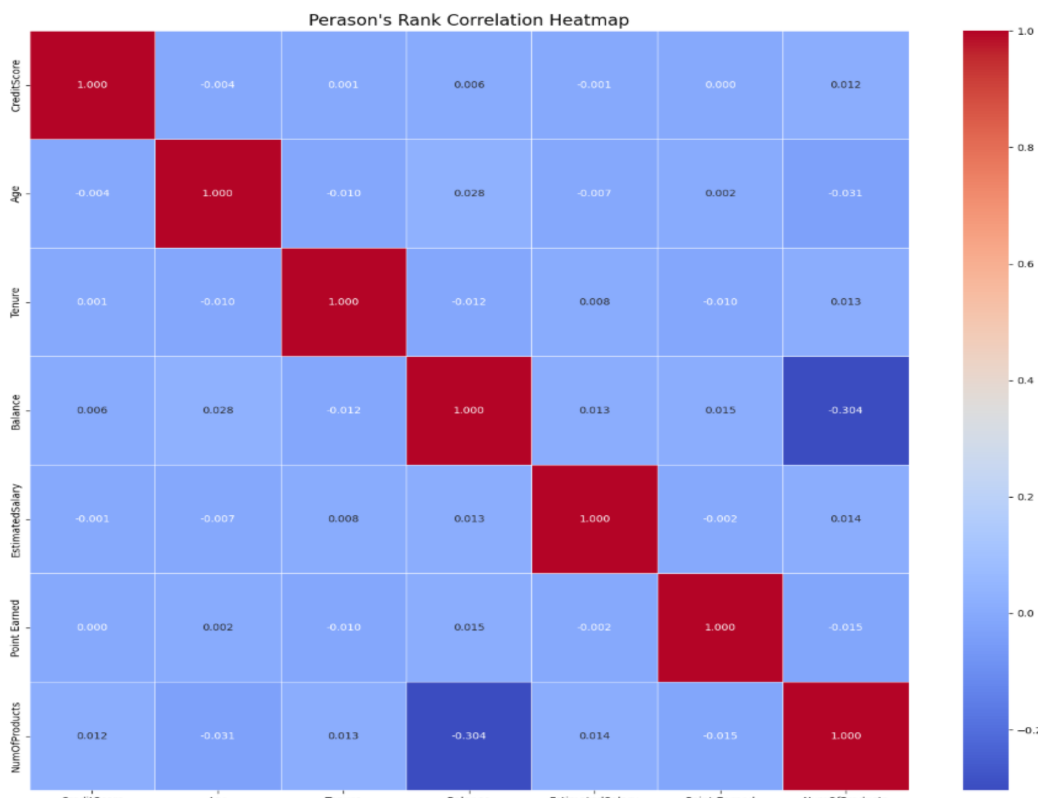


Geography Countplot
5014
2477
2509
Count
France
Spain
Germany
Geography
Gender Countplot
4543
5457
Count
Female
Male
Gender
Card Type Countplot
2507
2502
2496
2495
Count
DIAMOND
GOLD
SILVER
PLATINUM
Card Type
Satisfaction Score Countplot
1932
2014
2042
2008
2004
Count
1
2
3
4
5
Satisfaction Score



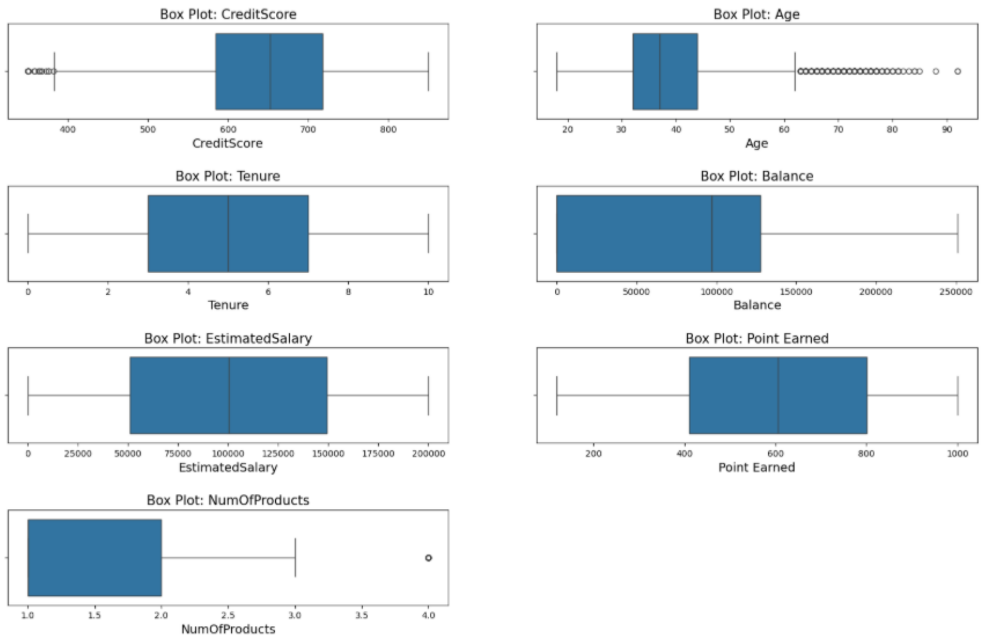
Correlation Structure

Pearson correlation analysis confirms that most numerical features exhibit weak linear relationships, both with each other and with churn. One moderate exception appears between **Balance** and **Number of Products**, indicating distinct engagement profiles rather than monotonic behavior. This reinforces the need for interaction-aware and non-linear models.



Outlier Assessment

Outliers were evaluated using IQR-based inspection. While several features contain statistical outliers, visual inspection confirms these represent legitimate customer behavior. Removing them did not improve distributional clarity or separability, so all observations were retained to preserve real-world variability.



EDA Takeaway

Across distributions, correlations, and outlier checks, churn signal appears to arise primarily from **behavioral patterns and feature interactions**, not from simple linear effects or extreme values. These findings directly informed feature engineering decisions and motivated the use of flexible, non-linear models.

Feature Engineering

Feature engineering was performed in two stages:

1. encoding raw categorical variables into model-ready representations
2. constructing higher-level behavioral and risk signals informed by EDA insights and business intuition.

The goal was not feature proliferation, but **expressive signal creation** that improves ranking power and interpretability.

---

1. Encoding Categorical Features

Categorical variables were encoded using mappings aligned with their semantic meaning and downstream modeling needs.

- **Nominal categories** such as geography and gender were mapped to numeric labels without implying ordinal distance.
- **Ordinal categories**, such as card type, were encoded to reflect increasing status or value.
- **Binary indicators** were normalized into consistent 0/1 representations.

This approach avoids one-hot explosion while remaining compatible with tree-based models and correlation-based analysis.

Design considerations:

- Preserve interpretability and ordering where it exists
- Maintain a compact feature space
- Avoid introducing artificial hierarchy into purely nominal variables

---

2. Derived Behavioral & Risk Features

Beyond raw attributes, new features were constructed to capture **behavioral patterns and interaction effects** that are not visible in individual columns.

These features were explicitly motivated by findings from EDA and business logic.

Behavioral Engagement Flags

Several features were designed to capture mismatches between engagement and satisfaction:

- **Active but dissatisfied customers**, potentially signaling early churn risk despite engagement
- **Inactive but highly satisfied customers**, indicating latent retention opportunities
- **Explicit complainer flags**, isolating customers who have expressed dissatisfaction

These features help distinguish *surface engagement* from *true loyalty*.

Credit & Risk Signals

Credit-related features were transformed into interpretable risk buckets:

- Low and high credit score indicators
- Combined risk flags capturing customers with both poor credit and complaint behavior

This allows the model to reason about **risk compounding**, not just absolute credit levels.

Tenure & Loyalty Patterns
Tenure was discretized into meaningful segments:

- Short-tenure customers, who may still be in onboarding
- Long-tenure customers, where inactivity may signal disengagement
- Inactive long-tenure customers, a particularly important churn-risk segment

These transformations reflect the idea that **time alone is not loyalty - behavior matters**.

Product Engagement Intensity
Product usage was reframed into engagement tiers:

- Single-product customers
- Multi-product customers
- Active customers with minimal product depth

This directly builds on EDA insights showing interaction between balance and product count, and helps the model capture different engagement strategies.

Financial Behavior Features
Balance-related features were engineered to reflect both presence and scale:

- Binary indicators for having any balance
- Relative balance positioning (above/below median)
- A salary-to-balance ratio to contextualize liquidity relative to income

These features help distinguish *capacity* from *actual financial engagement*.

Age-Based Risk Segmentation
Age was transformed into coarse risk segments rather than treated purely as a continuous variable:

- Younger customers
- Senior customers
- Senior inactive customers, combining demographic and behavioral risk

This supports non-linear effects and improves interpretability around lifecycle-related churn patterns.

Feature Selection

Feature selection was performed using a consensus-based approach across multiple model families, including Lasso, Ridge, Random Forest, Gradient Boosting, and XGBoost. Each feature was evaluated for consistent contribution under different inductive assumptions, and aggregate signals were used to assess robustness rather than optimize for any single model.

Although some engineered features showed lower consensus scores, **no features were removed at this stage**. Given the dataset size and the interaction-driven nature of several engineered variables, premature pruning risked removing conditional or recall-oriented signals.

Final feature reduction was intentionally deferred to later stages, after baseline modeling, probability calibration, and cost-aware evaluation provided empirical evidence of non-contributing features.

Model Training & Selection

Multiple classification models were evaluated, including Logistic Regression, Random Forest, Gradient Boosting, XGBoost, SVM, and KNN, using a consistent training and evaluation pipeline.

Initial results showed near-perfect performance across several model families. Given the dataset size, such uniformity raised concerns about potential information leakage. To address this, the feature set was refined to exclude post-event and sentiment-based variables, and the entire pipeline was re-run under a stricter, more realistic prediction scenario.

After refinement, linear and tree-based models continued to perform strongly and consistently, confirming that the churn signal is genuine rather than leakage-driven. In contrast, SVM and KNN performed poorly, as expected in a high-dimensional, engineered feature space.

Following hyperparameter tuning, Logistic Regression and Random Forest achieved nearly identical performance. Since further gains from model complexity were unlikely, the focus shifted from classification accuracy to probability quality and ranking behavior.

Random Forest was selected for the next phase due to its stable probability estimates, strong ranking performance, and suitability for risk scoring and segmentation.

Decision Modeling

Probability - Based Risk Scoring

The final model was evaluated using predicted churn probabilities rather than binary outputs. Probability inspection confirmed a clear separation between low-risk and high-risk customers, indicating that the model functions effectively as a risk scorer.

Customers were grouped into risk tiers (Low, Medium, High, Critical) based on predicted probabilities. Most customers fall into the Low Risk segment, while a small, well-defined

high-risk group enables targeted and cost-effective intervention strategies.
This probability-based framing shifts the problem from pure classification to actionable decision-making.

Scoring Validation - Ranking Quality

The churn model was validated as a **ranking system**, not just a classifier. Customers were sorted by predicted churn probability and evaluated by risk segments to assess whether higher scores correspond to higher observed churn.
Results show extreme concentration of churn at the top of the score distribution. Nearly all churn events occur within the highest-risk segment, while churn probability drops sharply beyond it. Targeting the top 20% highest-risk customers captures almost all churn cases, demonstrating strong lift and prioritization power.
This confirms the model is suitable for decision-making scenarios where limited retention resources must be focused on the highest-risk customers.

From Scores to Decisions

The churn model was converted from a prediction tool into a decision-support system by linking risk scores to cost-aware business actions.
Key outcomes:

- Targeting ~20% of customers captures ~99% of churn events.
- Precision remains extremely high, minimizing unnecessary outreach.
- Cost-sensitive analysis identifies an optimal decision threshold (~0.7), balancing missed churn against intervention cost.

Rather than relying on a fixed classification cutoff, customers are segmented into risk tiers, each mapped to an appropriate retention strategy. High-risk customers warrant immediate intervention, while low-risk customers are correctly excluded.
This step completes the transition from modeling to decision-making, demonstrating how probabilistic scores can drive practical, capacity-aware business actions.

Business Impact & Value Simulation

To assess real-world value, the churn scores were translated into a cost-sensitive retention simulation under realistic business assumptions:

- Retention actions succeed only for a subset of churners

- Each action incurs a fixed cost
- Only actual churners can be saved

Simulation Results

- **Total customers saved:** 170
- **Total intervention cost:** ~20,350
- **Cost per saved customer:** ~$119.7

Key Insights

- All saved customers come from the highest-risk segment.
- Retention incentives drive the entire impact, while softer actions add cost without measurable benefit.
- Low-risk customers are correctly excluded, preventing unnecessary spend.

Strategic Takeaway

The model operates as an **economic decision filter**, not just a churn predictor.
It enables targeted retention with measurable financial outcomes, ensuring that intervention budgets are deployed where they generate real value.

Project Conclusion

This project began as a churn prediction problem and intentionally evolved into a **risk-based decision system**.
Rather than optimizing for classification accuracy alone, the focus shifted toward building a model that supports real operational decisions. The churn model was transformed into a probabilistic scoring engine that enables:

- Customer-level risk assessment
- Actionable risk segmentation
- Threshold-driven targeting under capacity constraints
- Cost-aware intervention planning
- Measurable business impact evaluation

By integrating exploratory analysis, feature engineering, model validation, scoring diagnostics, and intervention simulation, the project demonstrates how data science moves beyond prediction into **prioritization and action**.

The final outcome is not a static model, but an end-to-end pipeline that reflects how churn decisions are made in practice — balancing risk, cost, and limited resources.
Ultimately, this project illustrates a core principle of applied data science:
value is created not by maximizing metrics in isolation, but by aligning models with business reality, economic trade-offs, and decision-making constraints.