**Business question:** A company (e.g., telecom or SaaS) wants to identify customers who are likely to churn so they can target retention efforts and reduce revenue loss.

**Success metric (business):** Reduce churn by focusing offers on high-risk customers; measure success by expected retention revenue saved vs. cost of retention.

**Churn Risk:**

* **Churn** = when a customer leaves (e.g., cancels a subscription, closes a bank account).
* **Churn risk** = the **probability** that a given customer will churn.
  + Example: if our model predicts that **Customer A has a 0.80 churn probability**, then there’s an **80% risk** they will leave.
  + If **Customer B has 0.20**, then they are at low churn risk.

Steps:

1. Load the CSV ( [Churn\_Modelling.csv](https://indianinstituteoftechnol299-my.sharepoint.com/:x:/g/personal/24mb0051_iitism_ac_in/ERWq8CBUXmhHnKhnDPZCShkBkpDJ-IgjV5IvNRZxxC9yiA?e=JTsrW7) )
2. Exploratory Data Analysis (EDA)

EDA is a preliminary approach to analyzing datasets to identify general patterns, anomalies, and unexpected features within the data.

* Descriptive Statistics
* No. of customers who have exited

1. Identify categorical and numeric features

Different statistical and visualization methods are appropriate for different data types. Using the wrong one can lead to faulty analysis.

* **Categorical data**: Can be visualized with bar charts and pie charts to show frequency and proportion.
* **Numerical data**: Can be explored with histograms, box plots, and scatter plots to understand distribution and relationships.

Preprocessing for machine learning algorithms

Many machine learning algorithms, particularly linear models, can only process numerical inputs. Categorical data, represented by text labels like "red" or "Paris," must be converted into a numerical format through a process called encoding. Different encoding methods are used depending on the nature of the categorical data.

* **One-Hot Encoding**: Used for *nominal* data (categories with no inherent order, like "city"). This creates a new binary column for each category, preventing the model from assuming a false ordinal relationship.
* **Label Encoding**: Can be used for *ordinal* data (categories with a natural rank, like "low," "medium," "high"). This method assigns a unique integer to each category, preserving the order.
* Any missing (numerical) values in the features processed by this SimpleImputer will be replaced by the most frequent (mode) value of that particular feature.
* Any missing (categorical) values in the features processed by this SimpleImputer will be replaced by the median value of that particular feature.

1. Splitting into train and test data( 80-20)
2. Logistics Regression
3. Random Forest for improvement
4. ROC Curve and PR Curve

ROC Curve PR Curve

A graph of a positive label

AI-generated content may be incorrect.A graph of a positive rate

AI-generated content may be incorrect.

1. Business Oriented Cost-benefit analysis

**Why Logistic Regression was used:**

**Logistic regression is used to predict the categorical dependent variable using a given set of independent variables.** It is used for solving classification problems.

1. Logistic Regression gives **clear, explainable coefficients** that show the **direction and magnitude of impact** of each feature on churn probability.

For eg, If the coefficient for IsActiveMember is negative, it means being active reduces churn likelihood.

1. It’s a **simple, fast model** → great for a **baseline comparison** against more complex models like Random Forest or XGBoost.
2. Logistic Regression outputs **probabilities (0–1)** instead of just “yes/no churn.”

This allows you to:

Rank customers by risk

Set thresholds based on business profit (e.g., target top 10% high-risk customers)

Do a cost-benefit analysis of retention campaigns

We used Random Forest, which improved ROC AUC from 0.75 (LogReg) to 0.85

(A higher AUC indicates better overall performance)

Interpret the Odds Ratio: (Odds are the ratio of the probability of an event happening to the probability of it not happening.)

* **Odds Ratio = 1**: There is no association; the odds of the outcome are the same for both groups.
* **Odds Ratio > 1**: The odds of the outcome are higher for the group of interest.
* **Odds Ratio < 1**: The odds of the outcome are lower for the group of interest.

**Eg, Gender\_Male (odds ratio ≈ 0.71):** Males are slightly less likely to churn compared to females.

In your example, the odds ratio for Gender\_Male is approximately **0.71**. This means the odds of a male churning are 0.71 times the odds of a female churning. Since 0.71 is less than 1, males have lower odds of churning compared to females, confirming your initial interpretation.

Eg: **Age (odds ratio ≈ 2.35):** For each unit increase in scaled age, churn likelihood more than doubles.

**Geography\_Germany (odds ratio ≈ 1.77):** Customers in Germany are ~77% more likely to churn compared to the baseline (other geographies).

**IsActiveMember (odds ratio ≈ 0.64):** Active members are ~36% less likely to churn.

*Business Perspective:*

*Inactive older German customers are the highest churn risk group.*

*Retention offers should prioritize these segments first.*

Why random forest not XG boost:

Random Forest because it’s a strong, robust model that works well on structured data like churn, and requires less fine-tuning than XGBoost. This let us focus on business interpretation and cost-benefit analysis rather than hyperparameter optimization. XGBoost might squeeze out a bit more accuracy, but for this dataset, Random Forest gave us reliable results faster and was easier to explain.

I tested the data with XG boost as well and the results were similar to Random forest.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1** | **ROC AUC** |
| Logistic Regression | 0.710145 | 0.384409 | 0.70098 | 0.496528 | 0.771197 |
| Random Forest | 0.842579 | 0.622691 | 0.578431 | 0.599746 | 0.849999 |
| XGBoost | 0.854073 | 0.648718 | 0.620098 | 0.634085 | 0.864473 |

**COST-BENEFIT ANALYSIS:**

**Best threshold by profit: 0.13**

* This means if we classify customers with a **churn probability ≥ 0.13** (13%) as “at-risk,” the **expected profit** is maximized.
* In other words, the business makes the most money if it starts targeting customers once their churn probability is **just 13% or higher**.
* Why so low? Because in this dataset, even moderately risky customers are valuable enough to try saving.

**Max profit: 260,200**

* Based on the assumptions we set:
  + Each retained customer brings **$1000 in lifetime value**
  + Retention offer costs **$100** per customer
* The model estimates that **if we use 0.13 as the cutoff**, the bank could expect to save about **$260,200 net profit** compared to doing nothing.

Profit=(TP×(Value saved−Retention cost))−(FP×Retention cost)

* Result:
  + Best threshold = **0.13**
  + Max profit = **260,200**

**Why this matters for business**

1. **Churn probability ≠ churn decision**
   * Just because someone has 20% risk doesn’t mean we ignore them.
   * If the cost-benefit math works out, it can be profitable to act even for low-to-medium risk customers.
2. **Profit-based thresholding**
   * Instead of just using the default 0.5 threshold (prob > 50%), we choose the cutoff that **maximizes business value**.
   * That’s how analytics bridges into real business strategy.
3. **Actionable outcome**
   * “At 0.13 threshold, we maximize ROI from retention offers.”
   * So, the recommendation is: *Target all customers with predicted churn probability ≥ 0.13 with our retention campaign.*

Evaluating the logistic regression model helps assess its performance and ensure it generalizes well to new, unseen data. The following metrics are commonly used:

**1. Accuracy:**[Accuracy](https://www.geeksforgeeks.org/data-analysis/techniques-to-evaluate-accuracy-of-classifier-in-data-mining/) provides the proportion of correctly classified instances.

*Accuracy=(TruePositives+TrueNegatives)/ Total*

*Accuracy = (TP+TN)/(TP+FP+TN+FN)*

**2. Precision:** [Precision](https://www.geeksforgeeks.org/r-language/precision-recall-and-f1-score-using-r/) focuses on the accuracy of positive predictions.

*Precision=TruePositives/TruePositives+FalsePositives*

*Precision=TP/(TP+FP)*

**3. Recall (Sensitivity or True Positive Rate):** [Recall](https://www.geeksforgeeks.org/machine-learning/precision-and-recall-in-machine-learning/) measures the proportion of correctly predicted positive instances among all actual positive instances.

*Recall=TruePositives/TruePositives+FalseNegatives*

*Recall=TP/(TP+FN)*

**4. F1 Score:**[F1 score](https://www.geeksforgeeks.org/machine-learning/f1-score-in-machine-learning/) is the harmonic mean of precision and recall.

*F1Score=2∗[(Precision∗Recall)/(Precision+Recall)]*

**5. Area Under the Receiver Operating Characteristic Curve (AUC-ROC):** The ROC curve plots the true positive rate against the false positive rate at various thresholds. [AUC-ROC](https://www.geeksforgeeks.org/machine-learning/auc-roc-curve/) measures the area under this curve, which provides an aggregate measure of a model's performance across different classification thresholds.

**6. Area Under the Precision-Recall Curve (AUC-PR):** Similar to AUC-ROC, [AUC-PR](https://www.geeksforgeeks.org/machine-learning/precision-recall-curve-ml/) measures the area under the precision-recall curve helps in providing a summary of a model's performance across different precision-recall trade-offs.

Code Link: [https://colab.research.google.com/drive/1anMhkx\_\_HejWa4bg5WFGus9wS8SmKtrw#scrollTo=02425b3e](https://colab.research.google.com/drive/1anMhkx__HejWa4bg5WFGus9wS8SmKtrw%23scrollTo=02425b3e)