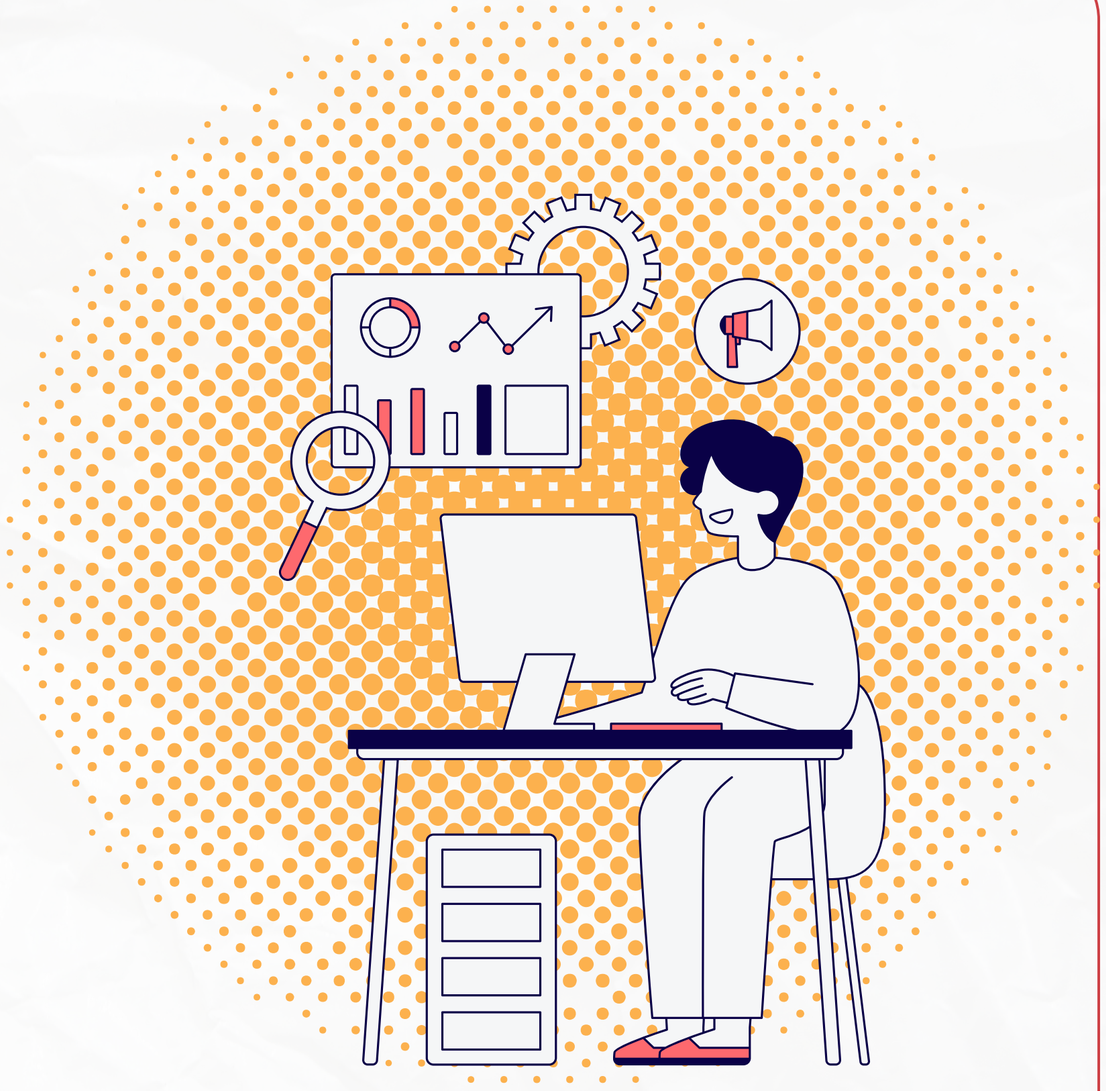
CUSTOMER CHURN PREDICTION

Naive Bayes Application



Rishi Ashok Kumar
Aleksandra Tatko
Nicolas Gonzalez Gort

BUSINESS PROBLEM

Business Problem

The telecom company (IBM) aims to predict which customers are likely to churn, enabling timely retention actions.

The challenge is to balance marketing efficiency and customer acquisition by minimizing both false alarms and missed churners.

Research Question

How can we predict customer churn effectively without over-targeting loyal customers or missing true churners?

WHAT IS NAIVE BAYES?

→ **Naive Bayes** is a probabilistic classifier based on **Bayes' Theorem.**

→ **Naive:** assumes that **features are independent**

→ Calculates the probability that a customer belongs to each class

→ Naive Bayes Formula:

$$P(Class \mid features) \propto P(Class) \times \Pi\, P(feature\ i \mid Class)$$

→ Predict the class with the highest resulting value.

DATA DESCRIPTION

1 Overview of Data

The Telco Customer Churn dataset contains **7,043** customer records and **21 variables** describing each customer's demographics, contract type, and service usage.

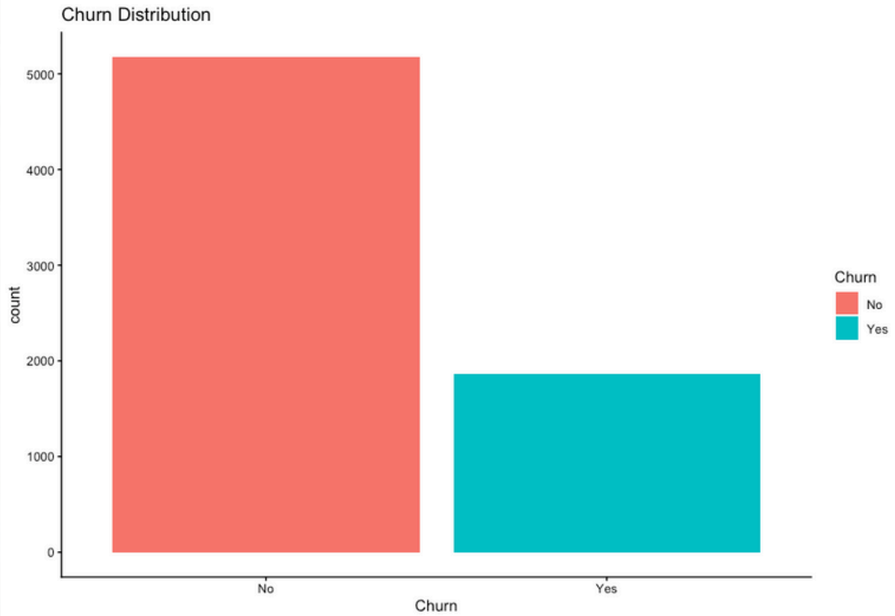
2 Target Variable

Churn (Yes/No) indicates whether the customer left or not the company.

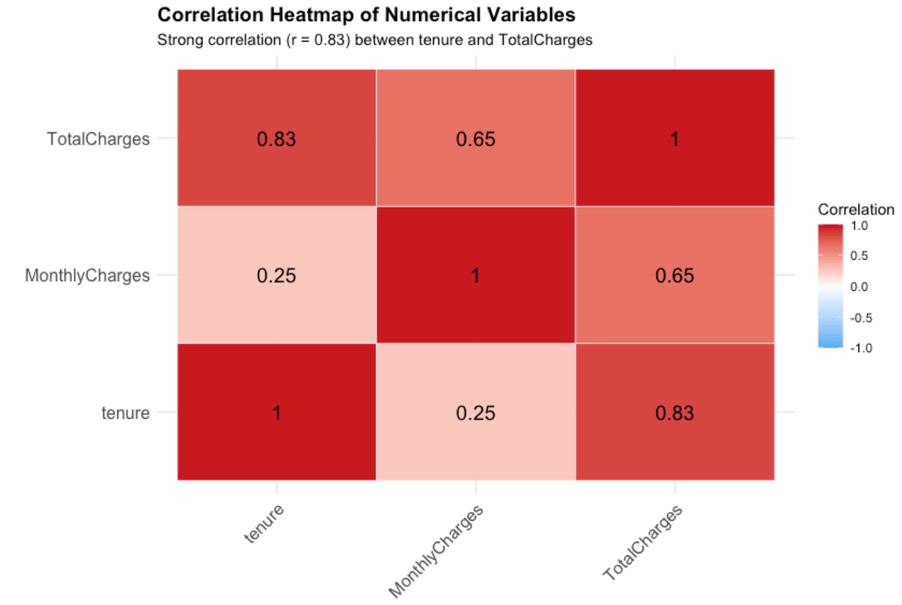
Group	Variables
Customer Demographics	customerID, gender, SeniorCitizen, Partner, Dependents
Service Subscriptions	PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies
Account Information	Contract, PaperlessBilling, PaymentMethod
Financial Information	MonthlyCharges, TotalCharges
Customer Lifecycle	tenure
Target Variable	Churn

DATA DESCRIPTION

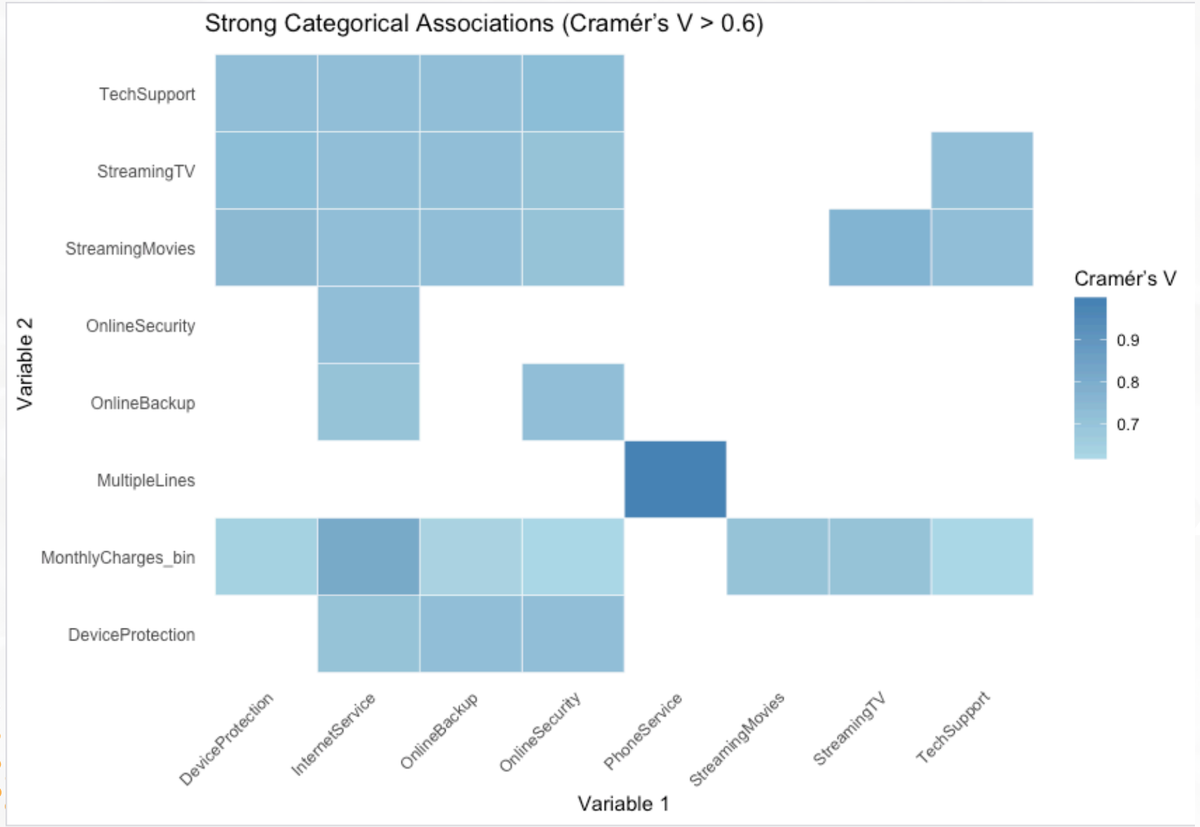
1 Churn Distribution



2 Correlation Heat Map



DATA DESCRIPTION

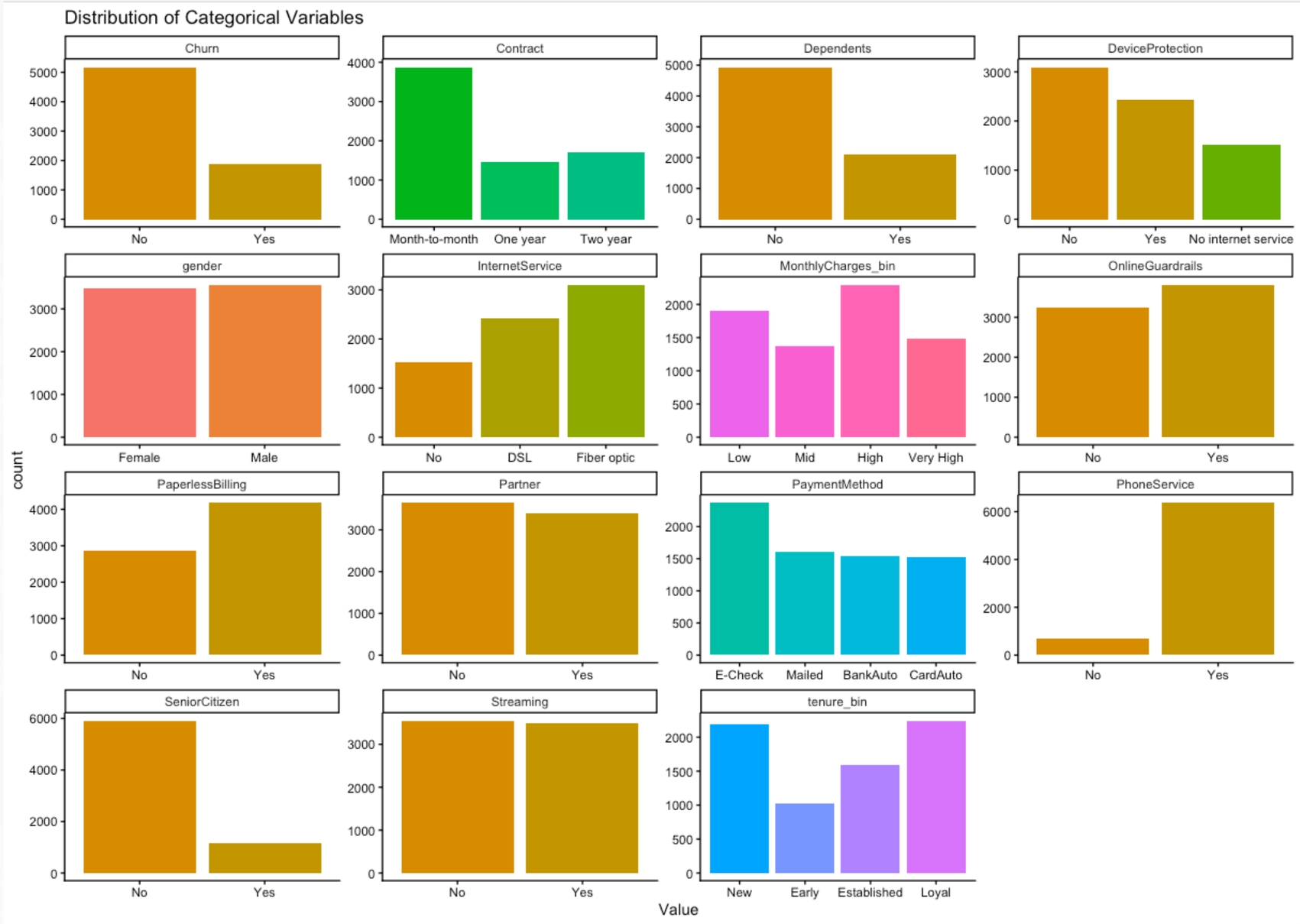


A **Chi-Square Test of Independence** was used to check whether pairs of categorical variables are statistically dependent (i.e., not independent).

Cramér's V was then calculated to measure the **strength** of those relationships (from 0 = weak to 1 = very strong).

Only variable pairs with significant **Chi-Square results (p < 0.05)** and **Cramér's V > 0.6** are shown in the heatmap.

FINAL DATASET

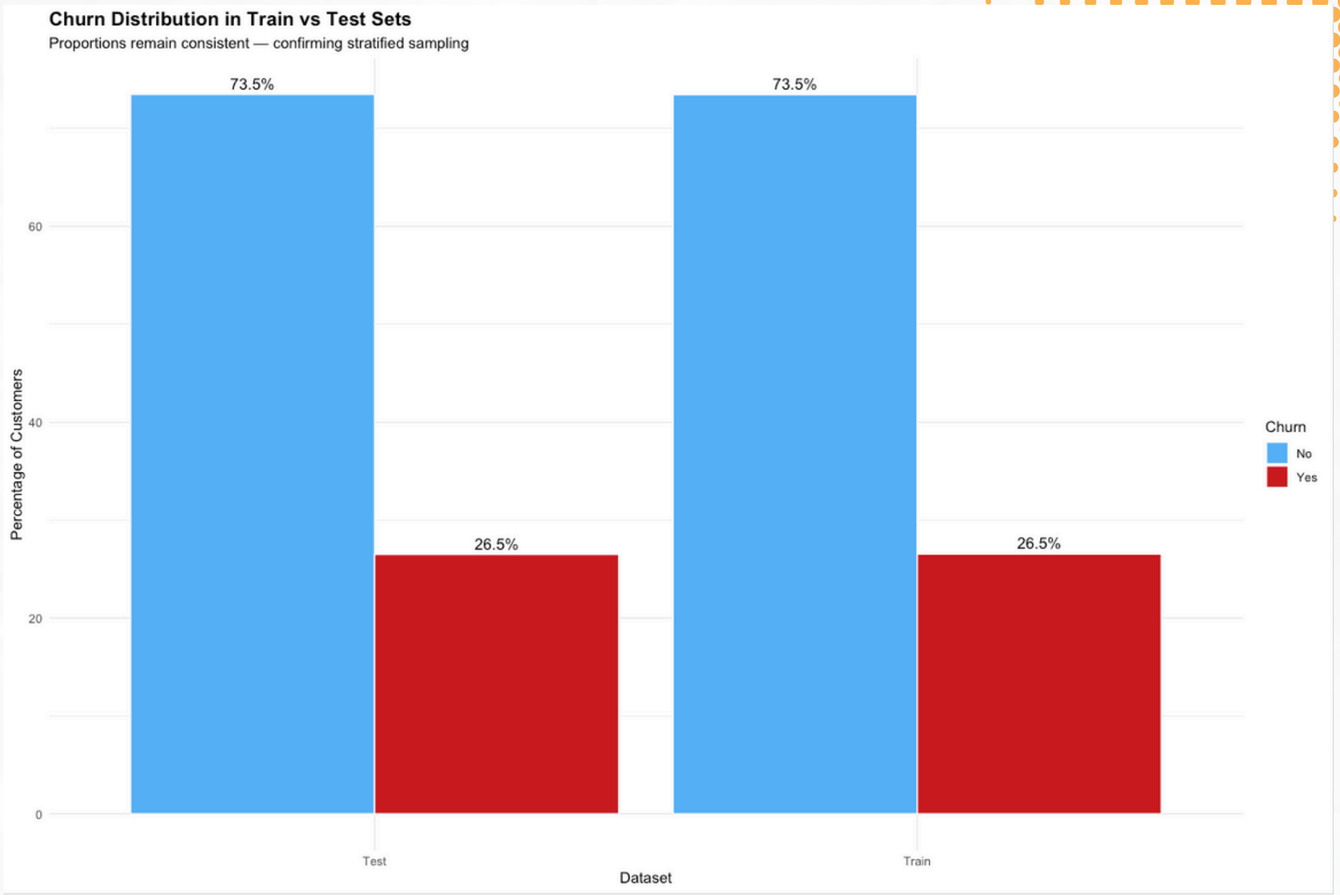


NAIVE BAYES MODEL

All variables categorical:
Categorical Naive Bayes

7043 observations:
80 train/ 20 test

Laplace Smoothing →+1



RESAMPLING

Method	What It Does
Baseline	Uses the **original** data without changing class proportions.
Downsampling	**Removes** some non-churn cases to match the number of churners
Upsampling	**Duplicates** existing churn cases until both classes are balanced.
ROSE (Random Over-Sampling Examples)	Creates **synthetic examples** of the minority class (churners) using smoothed bootstrapping.

NAIVE BAYES RESULTS

- **TP** (True Positive): Customers correctly predicted to churn — model caught real churners.
- **TN** (True Negative): Customers correctly predicted to stay — loyal customers identified correctly.
- **FP** (False Positive): Customers predicted to churn but actually stayed — wasted retention effort.
- **FN** (False Negative): Customers predicted to stay but actually churned — missed churners the model failed to catch.

Model	TP	TN	FP	FN
Baseline	226	**863**	147	**171**
ROSE	284	733	89	301
Upsample	273	760	100	274
Downsample	277	750	96	284

NAIVE BAYES RESULTS

- **Accuracy:** Overall % of correct predictions.
- **Precision**: % of predicted churners that actually churned (controls marketing waste).
- **Recall**: % of real churners correctly identified (controls missed churners).
- **Specificity**: % of loyal customers correctly identified as non-churners.
- **F1:** Balance between precision and recall.
- **Kappa**: Overall model reliability vs random guessing.

Model	Best Threshold	Accuracy	Precision	Recall	Specificity	F1	Kappa
Baseline	0.64	**0.774**	0.606	**0.569**	0.854	0.587	**0.432**
ROSE	0.56	0.723	**0.761**	0.485	0.892	0.593	0.398
Upsample	0.62	0.734	0.732	0.499	0.884	0.593	0.406
Downsample	0.6	0.73	0.743	0.494	**0.887**	0.593	0.403

INTERPRETATION

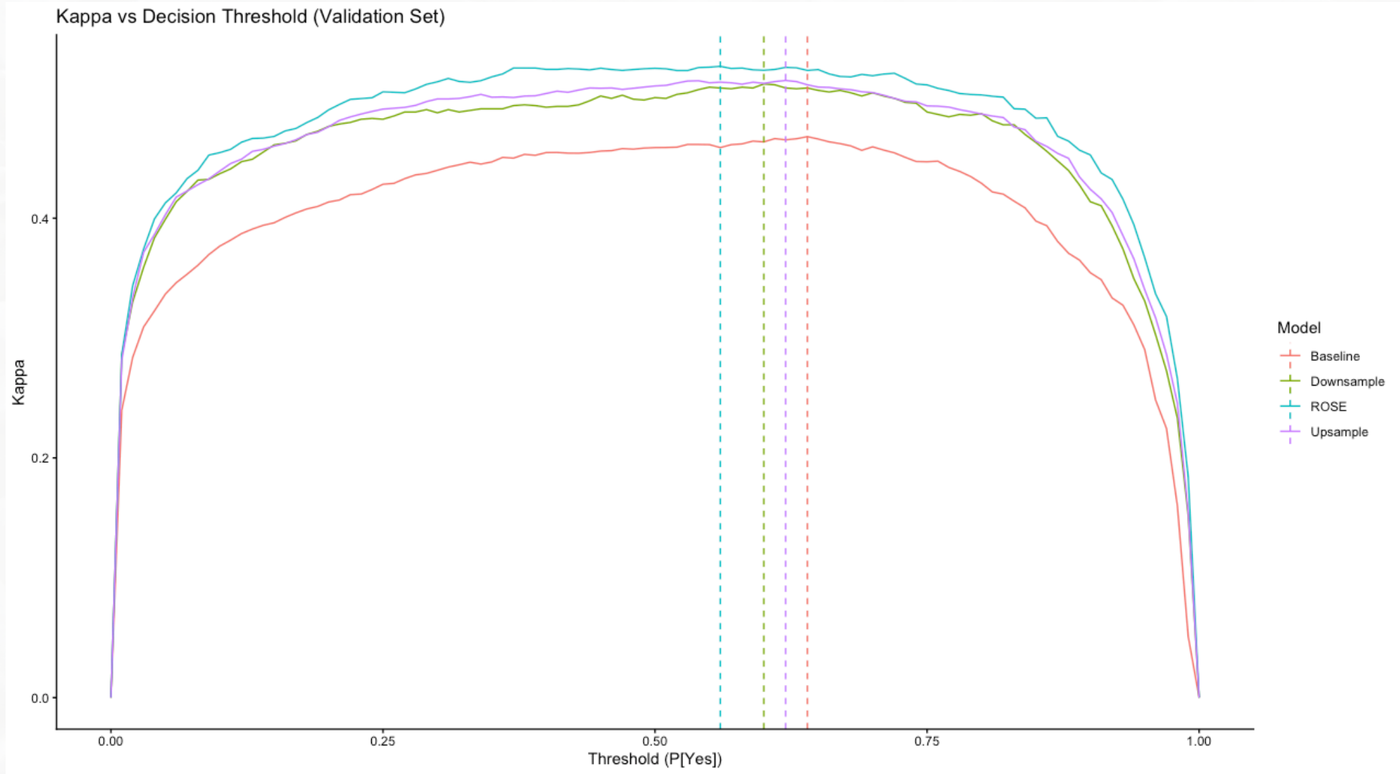
The Baseline model provides the most balanced results

- solid accuracy (0.774),
- strong specificity (0.854),
- F1 score (0.587)

→ **It correctly identifies most loyal customers and a fair share of churners, without overalerting the marketing team.**

Minimizes wasted retention spending while keeping churn under control — ideal for steady operations.

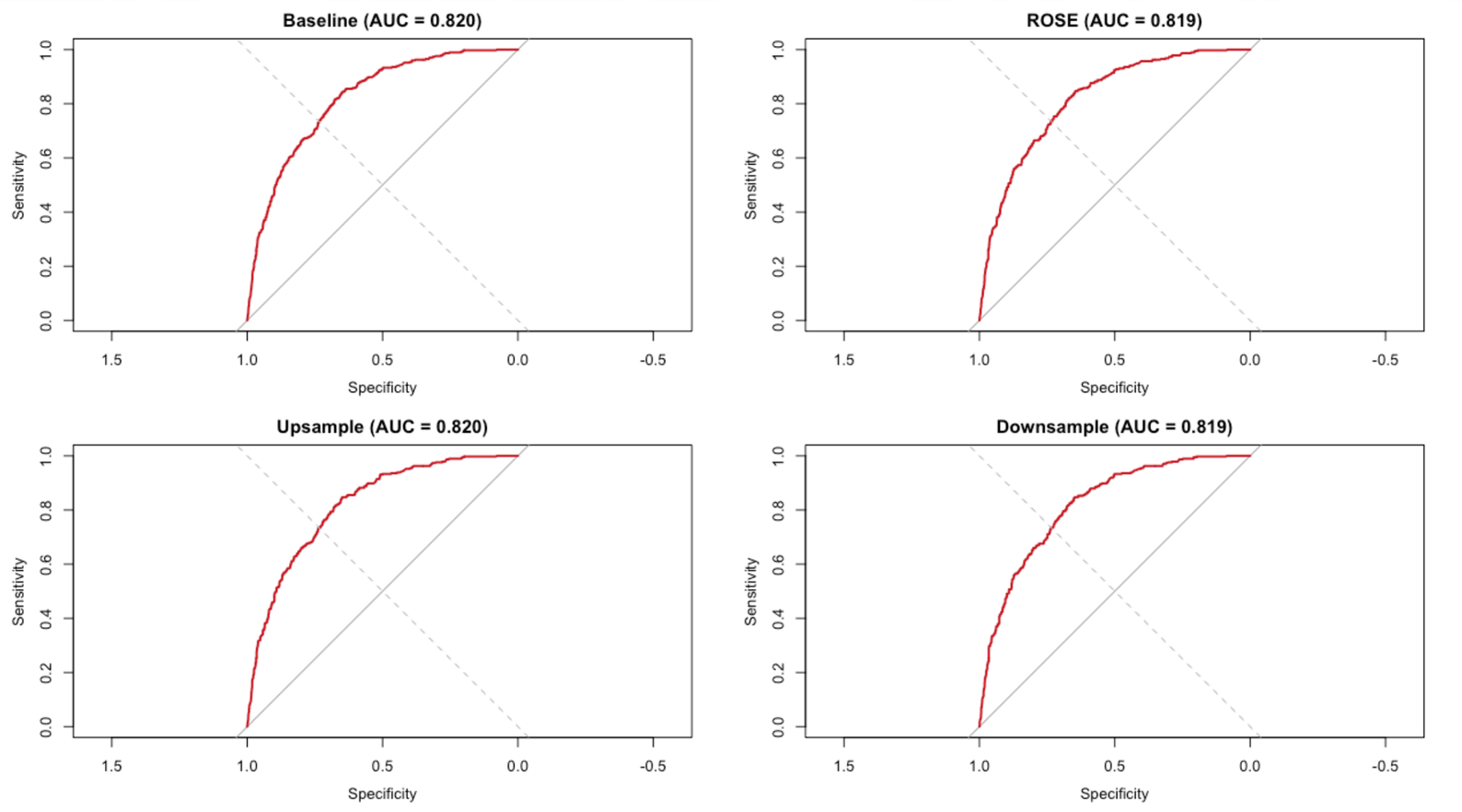
THRESHOLD ROBUSTNESS



Performance remains stable across thresholds and resampling — **robust calibration**.

Naive Bayes provides **stable predictions** even if we slightly change how we define a churner

DISCRIMINATIVE POWER



Almost **identical AUC** values, around 0.82

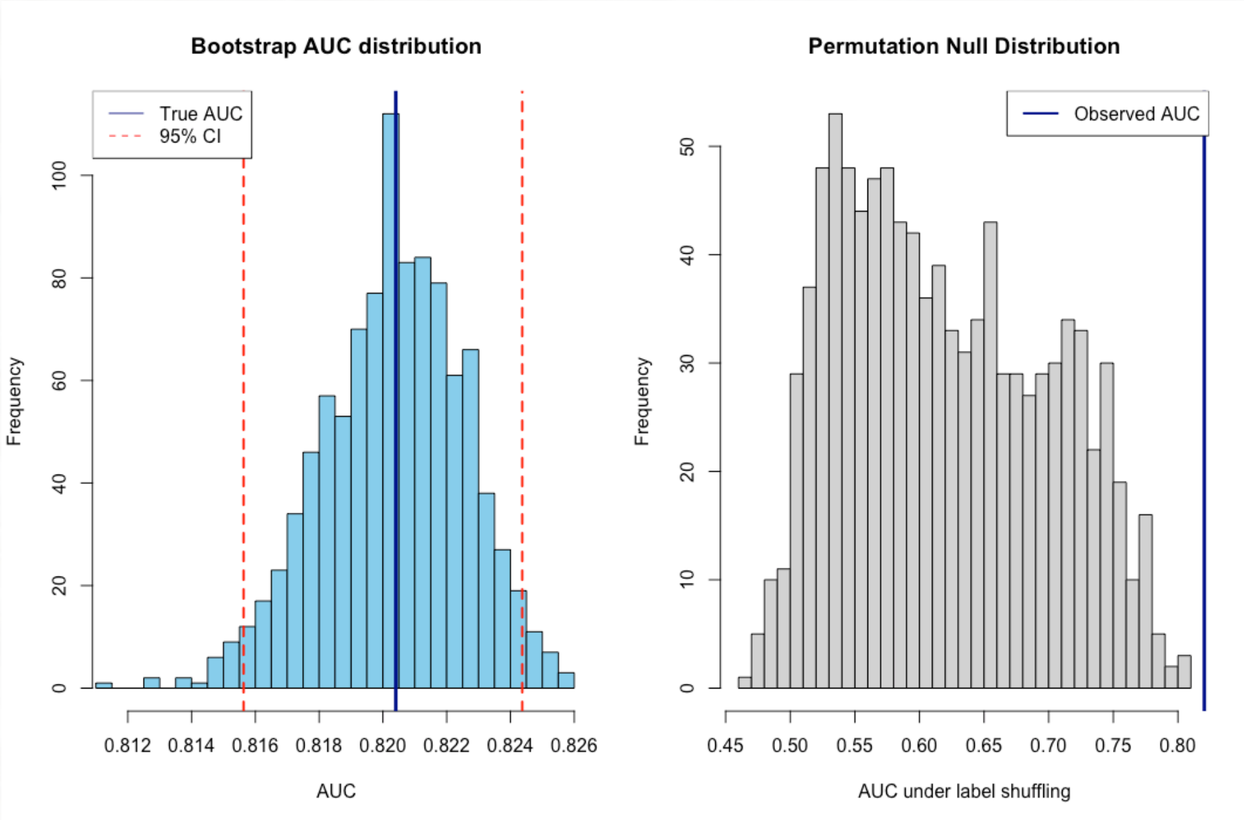
Good discriminative power, **clearly separating** churners from non churners

Performance is **consistent across data resampling** methods.

STATISTICAL AND RANDOMNESS ROBUSTNESS

Stability of performance under resampling. The model is **stable and reproducible (Internal Stability).**

Observed AUC **beyond** null distribution (external significance)



LIMITATIONS

1. Independence Assumption (unrealistic)
2. Poor Handling of Correlated or Reduntant Features
3. Sensitivity of feature encoding
4. Limited Expressiveness (linear relationships)
5. Imbalances in Y affects the model outcomes

CONCLUSION & BUSINESS IMPLICATIONS

- Naive Bayes performed as a solid baseline model. Kappa (0.432)
- Resampling increased precision but reduced recall and increased false negatives.
- Allocate budget efficiently
- The model avoids unnecessary contact to loyal customers
- Suggestions (create different models to assess feature importance)
- Lower customers predicted to stay but actually churned (FN)

References

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: With Applications in R(2nd ed.). Springer

Murphy, K. P. (2006). Naive bayes classifiers. University of British Columbia, 18(60), 1-8.

Thank You

Questions?