Customer Churn Prediction Analysis

CONTENTS

01

Introduction to Customer Churn

Introduction to Customer Churn

Definition

Customer churn, the phenomenon where customers discontinue their relationship with a company, represents a significant challenge across industries.

Impact

For subscription-based businesses, churn directly impacts revenue stability and growth potential.

Cost

Each churned customer not only represents lost revenue but also increases customer acquisition costs as companies must replace these customers to maintain growth.

Benefits

Understanding and predicting churn allows organizations to:

- Identify at-risk customers before they leave
- Implement targeted retention strategies
- Optimize resource allocation for customer retention
- Improve overall customer lifetime value

02

About the Dataset

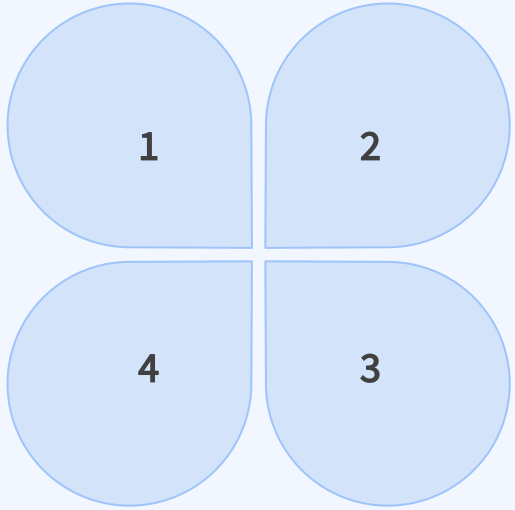
Dataset Overview

Size

440,833 customer records

Class Distribution

56.7% churned, 43.3% retained



Features

12 columns including demographic, behavioral, and engagement metrics

Target Variable

Churn (binary: 1 = churned, 0 = retained)

Key Features

Demographics

Age, Gender

Engagement

Tenure, Usage Frequency

Support Interaction

Support Calls, Last Interaction

Financial Behavior

Payment Delay, Total Spend

Contract Details

Subscription Type, Contract Length

Dataset Quality

The dataset showed minimal missing values (only 1 record per column), which were removed during preprocessing.

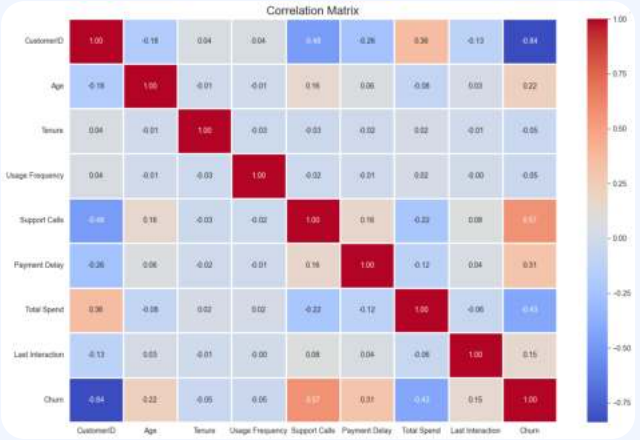
Conclusion

The high-quality, balanced nature of this dataset provides a solid foundation for developing robust churn prediction models.

03

Exploratory Data Analysis and Visualizations

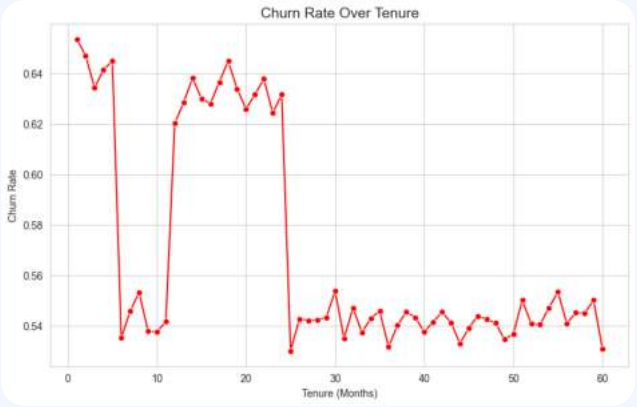
Correlation Analysis



Strong Positive Correlations with Churn:

- Support Calls (0.57): Higher number of support calls strongly indicates increased churn probability
- Payment Delay (0.31): Longer payment delays correlate with higher churn rates
- Age (0.22): Slight positive correlation with churn

Churn Rate Over Tenure



The analysis of churn patterns across customer tenure revealed critical insights:

- High Early Churn: New customers (tenure close to 0 months) exhibit the highest churn rates
- Loyalty Development: Sharp decrease in churn after the first few months (5-10 month mark)
- Renewal Risk Period: Secondary churn spike around 15-25 months, likely corresponding to contract renewal decisions
- Stable Long-term Customers: After 30+ months, churn rates stabilize at relatively low levels

04

Possible Business Insights & Actions

Improve early retention strategies

1

Churn reduction

Since churn is highest in the first few months, focus on:

- Onboarding enhancements
- Better customer support
- Incentives for new customers

2

Retention ideas

Ideas:

- Offer first-month discounts
- Provide tutorials or guides
- Implement loyalty bonuses

Investigate mid-term churn (15-25 months)

Churn issues

High churn in this period suggests issues with:

- Contract renewals
- Pricing changes
- Lack of engagement

Solutions

Possible solutions:

- Loyalty rewards
- Better customer service
- Personalized retention offers

Strengthen loyalty for long-term users

Since churn is lower after 30+ months, it makes sense to:

- Reward loyal customers
- Offer exclusive discounts or premium features

05

Feature Engineering and Selection

Feature Engineering

Tenure Categorization: Created a new feature "Tenure_Category" with four segments:

- New (⩽6 months)
- Short-term (6-12 months)
- Mid-term (1-3 years)
- Long-term (>3 years)

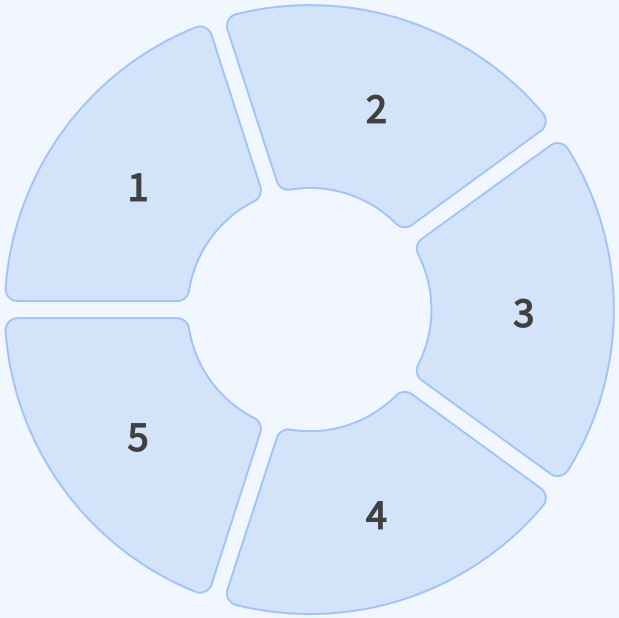
Feature Selection/Chi-Square Test for Categorical Features

Contract Length

Monthly: Chi^2 Score = 66,489

Tenure Category

Mid-term: Chi^2 Score = 197



Contract Length

Quarterly: Chi^2 Score = 8,209

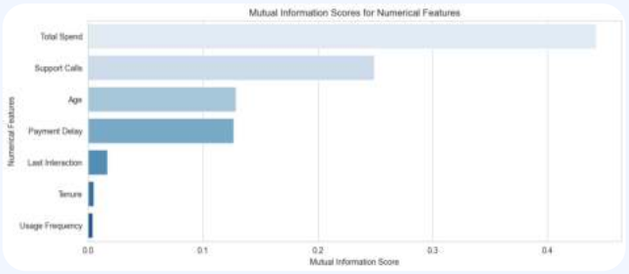
Gender

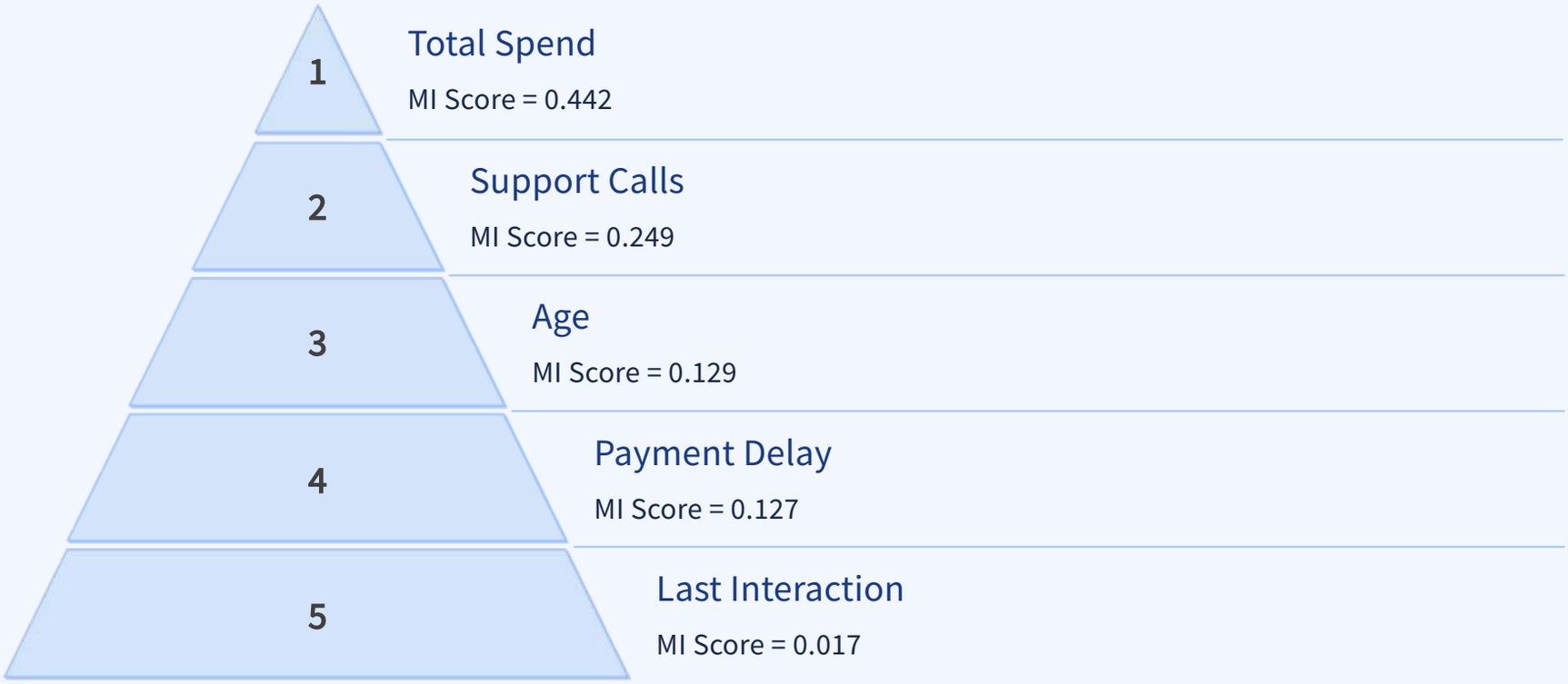
Male: Chi^2 Score = 5,862

Tenure Category

New: Chi^2 Score = 523

Feature Selection/Mutual Information for Numerical Features





Based on these analyses, the final set of selected features included:

- Total Spend
- Support Calls
- Age
- Payment Delay
- Contract Length
- Gender
- Tenure Category

06

The Pipeline

Preprocessing Pipeline

1

Features

Numerical Features Transformation:

2

Standardize

StandardScaler: Standardize numerical features to have mean=0 and variance=1

3

Categorical

Categorical Features Transformation:

4

Convert

OneHotEncoder: Convert categorical variables to binary features

5

Combined

Combined Preprocessing:

6

Apply

ColumnTransformer: Apply appropriate transformations to each feature type

Model Training Pipeline(1)

1

Data Splitting:

2

80% training data, 20% testing data

3

Stratified sampling to maintain class distribution

4

Hyperparameter Tuning:

5

GridSearchCV with 5-fold cross-validation

6

Optimization for accuracy metric

Model Training Pipeline(2)

Model Evaluation:

Performance metrics

accuracy, precision, recall, F1-score

Confusion matrix

analysis

07

Models Used

Logistic Regression

1

Tuned Parameters

- Regularization Strength (C): 0.1, 1, 10, 100
- Penalty Type: l1, l2
- Solver: liblinear, saga
- Class Weight: None, balanced
- Max Iterations: 2000, 5000, 10000

2

Optimal Parameters

- C: 1
- Penalty: l1
- Solver: liblinear
- Class Weight: balanced
- Max Iterations: 10,000

Decision Tree

1 Tuned Parameters

- Criterion: gini, entropy
- Max Depth: 3, 5, 10
- Min Samples Split: 2, 5, 10
- Min Samples Leaf: 1, 2, 5, 10

2 Best Parameters

- Criterion: entropy
- Max Depth: 10
- Min Samples Split: 2
- Min Samples Leaf: 2

Naive Bayes

Hyperparameters

Tuned:
- Var Smoothing: 1e-12, 1e-10, 1e-9, 1e-8, 1e-6, 1e-4, 1e-2

Best Parameters

- Var Smoothing: 0.01

Model Selection

Each model was selected to provide different perspectives on the data:
- Logistic Regression for its interpretability and coefficient insights
- Decision Tree for capturing non-linear relationships and decision rules
- Naive Bayes for its efficiency and probabilistic approach

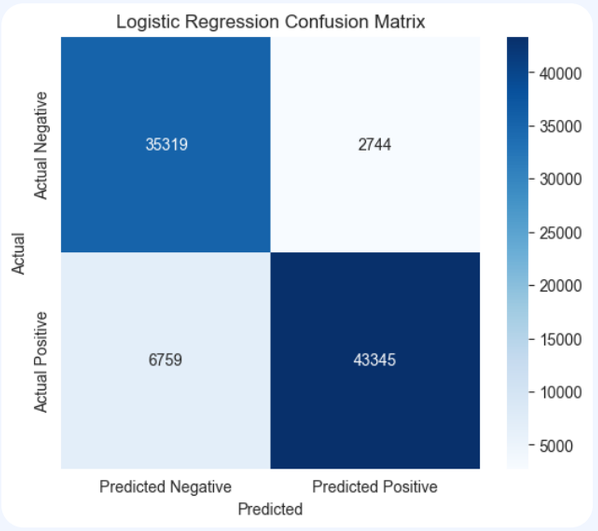
08

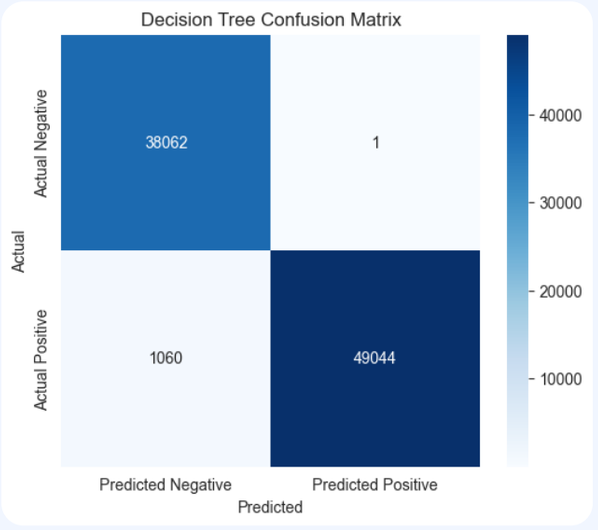
Model Comparison and Evaluation

Accuracy Comparison

Model	Accuracy
Decision Tree	98.8%
Naive Bayes	91.5%
Logistic Regression	89.2%

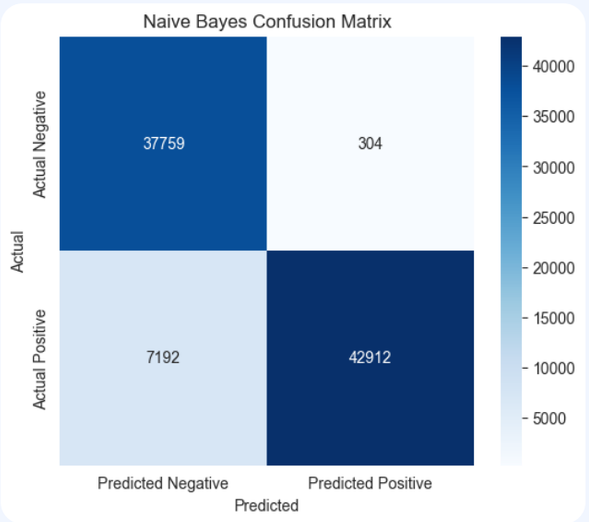
Confusion Matrices





Model	True Negatives (TN)	False Positives (FP)	False Negatives (FN)	True Positives (TP)
Decision Tree	38,062	1	1,060	49,044
Naive Bayes	37,759	304	7,192	42,912
Logistic Regression	35,319	2,744	6,759	43,345

Detailed Classification Metrics



Model	Class	Precision	Recall	F1-score
Decision Tree	0	0.97	1.00	0.99
Decision Tree	1	1.00	0.98	0.99
Naive Bayes	0	0.84	0.99	0.91
Naive Bayes	1	0.99	0.86	0.92
Logistic Regression	0	0.84	0.93	0.88
Logistic Regression	1	0.94	0.87	0.90

Key Observations:

- The Decision Tree model demonstrated exceptional performance across all metrics, with near-perfect precision for churn prediction
- Naive Bayes showed strong precision for churn prediction but lower recall compared to other models
- Logistic Regression provided balanced performance with good interpretability
- All models significantly outperformed random guessing

09

Ensemble Modeling: Stacking

Stacking Architecture

1

Base Models

Decision Tree (with optimized parameters)
・Naive Bayes (with optimized parameters)

2

Meta-Model

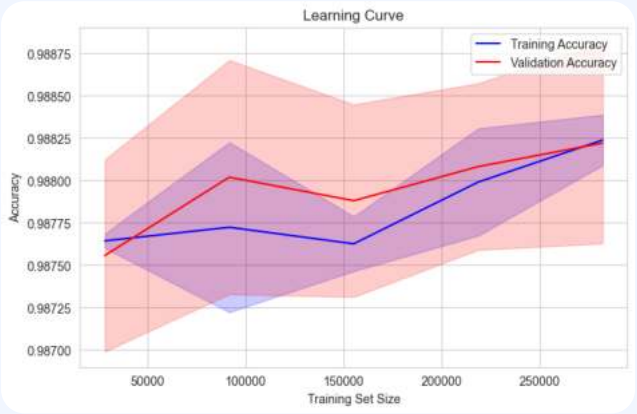
Logistic Regression (C=100, class_weight='balanced', max_iter=10000)

3

Cross-Validation

5-fold cross-validation for generating meta-features

Stacking Performance



1 Accuracy

98.8%

2 Confusion Matrix

・Similar performance to the best individual model (Decision Tree)

3 Learning Curve

・Stable performance across different training set sizes

・Minimal gap between training and validation accuracy, indicating good generalization

10

Model Deployment with Streamlit

Deployment Architecture

The application is built with Streamlit, a Python framework for creating data applications, and includes the following components:

Decision Tree

Trained model (dt_model.pkl)

Customer dataset

Cleaned data (cleaned_customer_data.csv)

Web interface

Interactive interface with multiple pages

Key Features(1)

1

Interface

Prediction Interface:

2

Input form

Input form for customer attributes (Age, Support Calls, Payment Delay,..)

3

Real-time

Real-time prediction of churn probability

4

Visual

Visual representation of prediction results

5

Color-coded

Color-coded outcomes (red for churn risk, green for retention)

6

Dashboard

Data Insights Dashboard:

Key Features(2)

Data Tools

Interactive data exploration tools

Customer Search

Customer search functionality by ID

Visual Analytics

- Distribution histograms for numerical features
- Gender distribution pie chart
- Churn distribution visualization
- Correlation heatmap for numerical features

Filter Options

Filtering capabilities by churn status

Data Export

Downloadable filtered data in CSV format

About

About Section:

Key Features(3)

Model

documentation and performance metrics

Feature

descriptions and preprocessing details

Usage

guidelines for business users

Business Applications(1)

1 Individual Customer Assessment:

2 Customer service representatives can input customer details during interactions to assess churn risk in real-time, allowing for immediate intervention strategies.

3 Scenario Analysis:

4 Business analysts can modify customer attributes to understand how changes in service quality, contract terms, or spending patterns might affect churn probability.

5 Data-Driven Insights:

6 Managers can explore patterns in the customer base through the interactive visualizations, identifying segments with higher churn risk.

Business Applications(2)

1. Exportable Results:
2. Marketing teams can download filtered customer lists for targeted retention campaigns.

11

Conclusions and Business Implications

Key Predictors of Churn

Contract

Monthly contracts significantly increase churn risk

Support

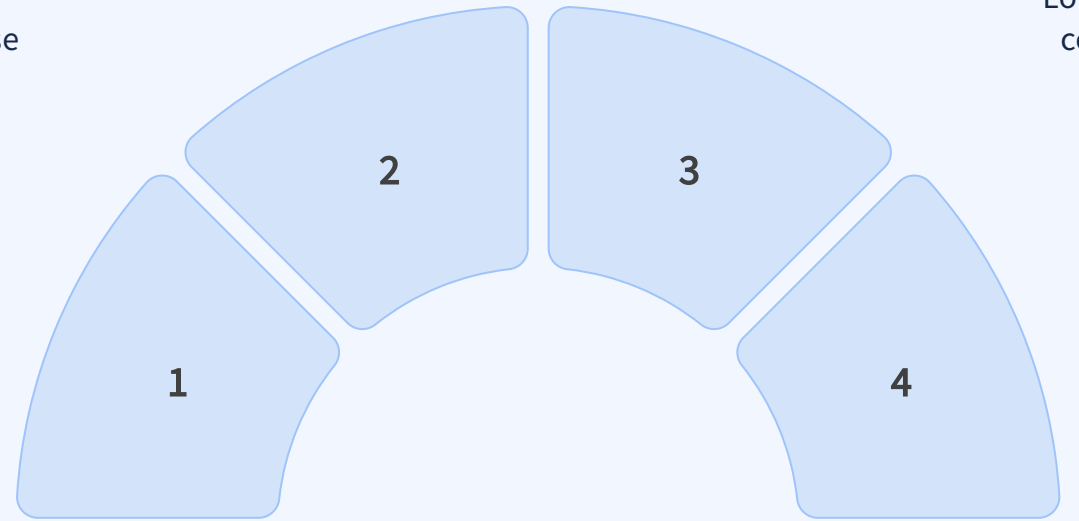
More than 5 support calls strongly indicates dissatisfaction

Spend

Lower spending customers are more likely to churn

Payment

Longer payment delays correlate with higher churn probability



Customer Lifecycle Insights

1

First 6 months

represent a critical period for customer retention

2

15-25 month

period (likely contract renewal) requires special attention

3

30+ months

customers show significantly higher loyalty

Recommended Business Strategies(1)

1 **Customer Experience**

Improve Early Customer Experience:

2 **Onboarding**

Enhanced onboarding processes

3 **Check-ins**

Regular check-ins during first 6 months

4 **Incentives**

Early-stage incentives and support

5 **Support Quality**

Enhance Support Quality:

6 **Resolution**

Improve first-call resolution rates

Recommended Business Strategies(2)

1. **Issue ID**
   Proactive issue identification
2. **Special handling**
   for customers with multiple support calls
3. **Contract**
   Optimize Contract Structure:
4. **Long-term**
   Incentivize longer-term contracts
5. **Special offers**
   at renewal periods
6. **Loyalty**
   benefits for contract extensions

Recommended Business Strategies(3)

Payment Processes

Address Payment Processes:

Simplify

Simplify payment procedures

Flexibility

Flexible payment options

Communication

Proactive communication about upcoming payments

Thank You