Customer Churn Prediction Analysis

1. Introduction to Customer Churn

Customer churn, the phenomenon where customers discontinue their relationship with a company, represents a significant challenge across industries. For subscription-based businesses, churn directly impacts revenue stability and growth potential. Each churned customer not only represents lost revenue but also increases customer acquisition costs as companies must replace these customers to maintain growth.

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

2. About the Dataset

The analysis utilizes a comprehensive customer dataset containing 440,833 records with 12 features capturing various aspects of customer behavior and demographics:

Dataset Overview:

- **Size:** 440,833 customer records
- **Features:** 12 columns including demographic, behavioral, and engagement metrics
- **Target Variable:** Churn (binary: 1 = churned, 0 = retained)
- **Class Distribution:** 56.7% churned, 43.3% retained

Key Features:

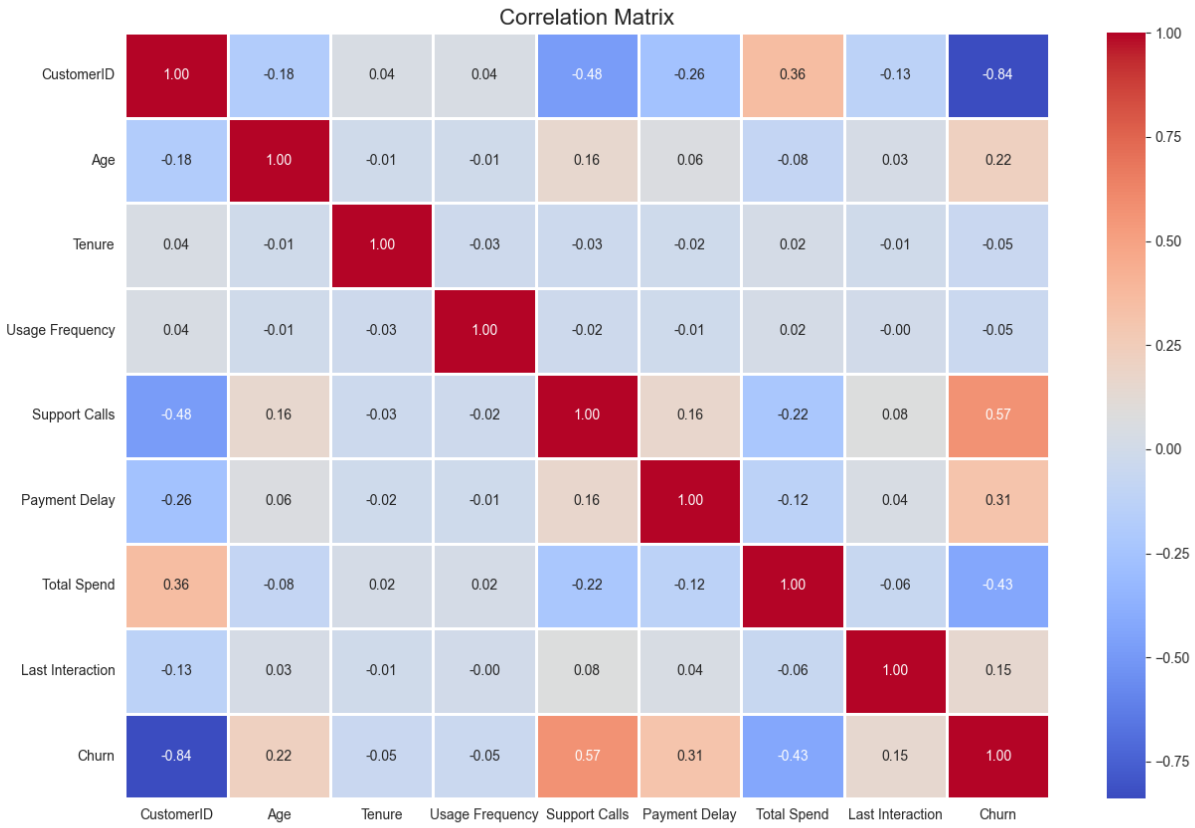
- **Demographics:** Age, Gender
- **Engagement:** Tenure, Usage Frequency
- **Support Interaction:** Support Calls, Last Interaction
- **Financial Behavior:** Payment Delay, Total Spend
- **Contract Details:** Subscription Type, Contract Length

The dataset showed minimal missing values (only 1 record per column), which were removed during preprocessing. The high-quality, balanced nature of this dataset provides a solid foundation for developing robust churn prediction models.

3. Exploratory Data Analysis and Visualizations

The exploratory data analysis revealed several significant patterns and relationships within the customer data:

Correlation Analysis:



	CustomerID	Age	Tenure	Usage Frequency	Support Calls	Payment Delay	Total Spend	Last Interaction	Churn
CustomerID	1.00	-0.18	0.04	0.04	-0.48	-0.26	0.36	-0.13	-0.84
Age	-0.18	1.00	-0.01	-0.01	0.16	0.06	-0.08	0.03	0.22
Tenure	0.04	-0.01	1.00	-0.03	-0.03	-0.02	0.02	-0.01	-0.05
Usage Frequency	0.04	-0.01	-0.03	1.00	-0.02	-0.01	0.02	-0.00	-0.05
Support Calls	-0.48	0.16	-0.03	-0.02	1.00	0.16	-0.22	0.08	0.57
Payment Delay	-0.26	0.06	-0.02	-0.01	0.16	1.00	-0.12	0.04	0.31
Total Spend	0.36	-0.08	0.02	0.02	-0.22	-0.12	1.00	-0.06	-0.43
Last Interaction	-0.13	0.03	-0.01	-0.00	0.08	0.04	-0.06	1.00	0.15
Churn	-0.84	0.22	-0.05	-0.05	0.57	0.31	-0.43	0.15	1.00

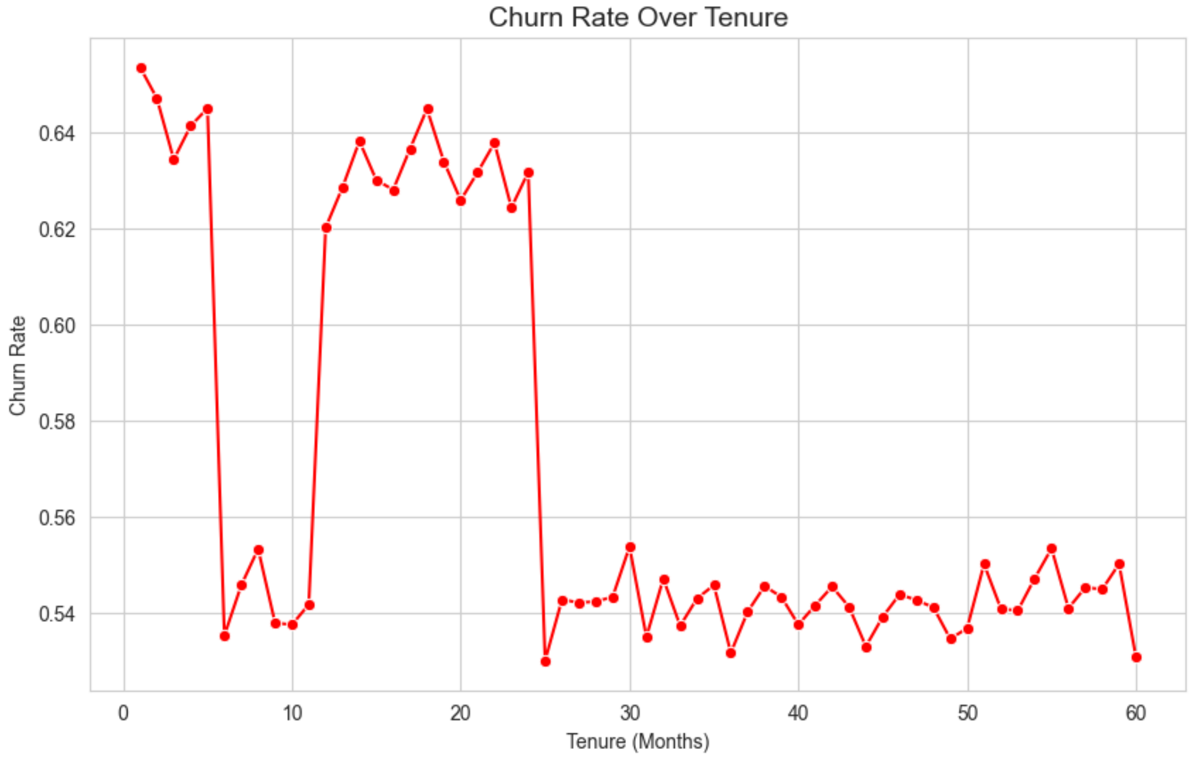
- Strong Positive Correlations with Churn:

- 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

- Strong Negative Correlations with Churn:

- Total Spend (-0.43): Higher spending customers demonstrate greater loyalty
- CustomerID (-0.84): This correlation is likely coincidental rather than causal

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

Possible Business Insights & Actions

Improve early retention strategies:

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

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

Ideas:

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

Investigate mid-term churn (15-25 months):

High churn in this period suggests issues with:

- Contract renewals
- Pricing changes
- Lack of engagement
- 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

4. Feature Engineering and Selection

To enhance model performance and extract meaningful patterns, several feature engineering and selection techniques were applied:

Feature Engineering:

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

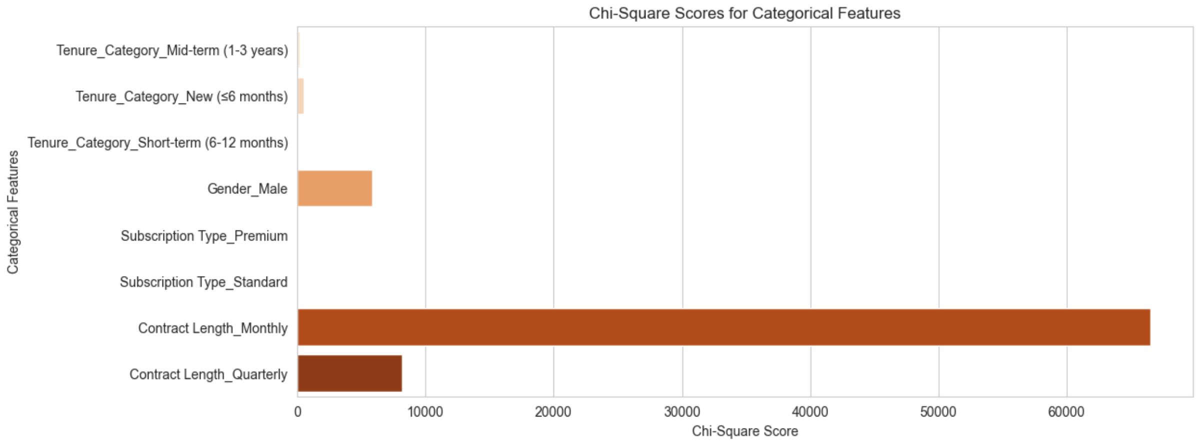
    - New (≤6 months)
    - Short-term (6-12 months)
    - Mid-term (1-3 years)
    - Long-term (>3 years)

Feature Selection:

Two statistical methods were employed to identify the most predictive features:

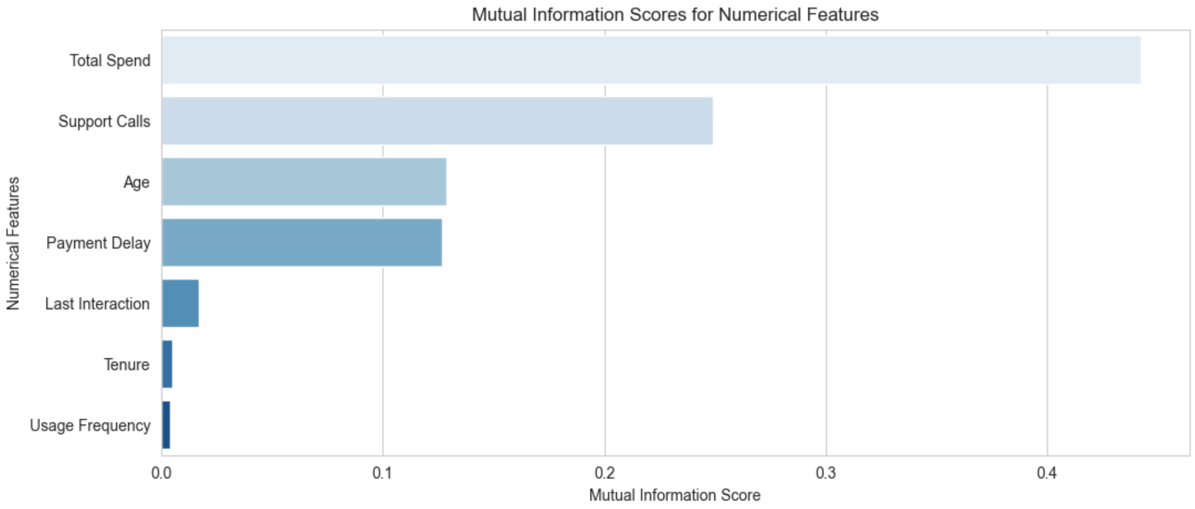
1. Chi-Square Test for Categorical Features:

1. Contract Length (Monthly): Chi² Score = 66,489
2. Contract Length (Quarterly): Chi² Score = 8,209
3. Gender (Male): Chi² Score = 5,862
4. Tenure Category (New): Chi² Score = 523
5. Tenure Category (Mid-term): Chi² Score = 197



2. Mutual Information for Numerical Features:

1. Total Spend: MI Score = 0.442
2. Support Calls: MI Score = 0.249
3. Age: MI Score = 0.129
4. Payment Delay: MI Score = 0.127
5. Last Interaction: MI Score = 0.017



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

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

This feature selection process ensured that models were built using the most informative variables, improving both performance and interpretability.

5. The Pipeline

A robust machine learning pipeline was implemented to ensure consistent preprocessing and model evaluation:

Preprocessing Pipeline:

1. Numerical Features Transformation:
   StandardScaler: Standardize numerical features to have mean=0 and variance=1

2. Categorical Features Transformation:
   OneHotEncoder: Convert categorical variables to binary features

3. Combined Preprocessing:
   ColumnTransformer: Apply appropriate transformations to each feature type

Model Training Pipeline:

1. Data Splitting:
   80% training data, 20% testing data
   Stratified sampling to maintain class distribution

2. Hyperparameter Tuning:
   GridSearchCV with 5-fold cross-validation
   Optimization for accuracy metric

3. Model Evaluation:
   Performance metrics: accuracy, precision, recall, F1-score
   Confusion matrix analysis

6. Models Used

Three distinct machine learning algorithms were implemented and optimized for churn prediction:

1. Logistic Regression

- **Hyperparameters Tuned**:
    - **Regularization Strength (C)**: 0.1, 1, 10, 100
    - **Penalty Type**: l1, l2
    - **Solver**: liblinear, saga
    - **Class Weight**: None, balanced
    - **Max Iterations**: 2000, 5000, 10000
- **Best Parameters**:
    - **C**: 1
    - **Penalty**: l1
    - **Solver**: liblinear
    - **Class Weight**: balanced
    - **Max Iterations**: 10,000

2. Decision Tree

- **Hyperparameters Tuned**:
    - **Criterion**: gini, entropy
    - **Max Depth**: 3, 5, 10
    - **Min Samples Split**: 2, 5, 10
    - **Min Samples Leaf**: 1, 2, 5, 10
- **Best Parameters**:
    - **Criterion**: entropy
    - **Max Depth**: 10
    - **Min Samples Split**: 2
    - **Min Samples Leaf**: 2

3. 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

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

7. Model Comparison and Evaluation

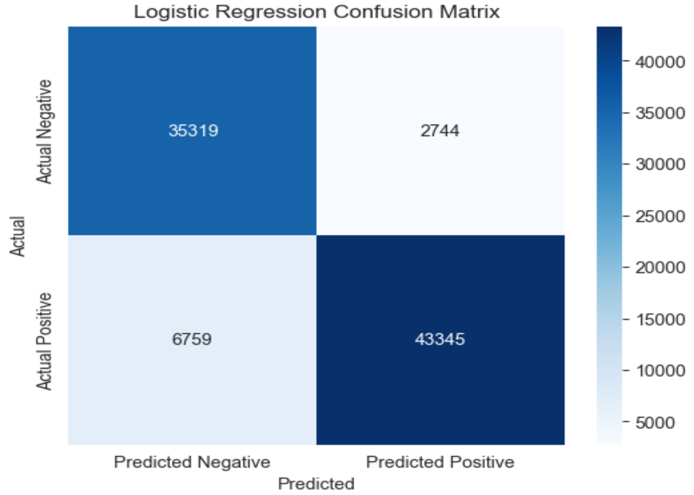
The three models were rigorously evaluated using multiple performance metrics:

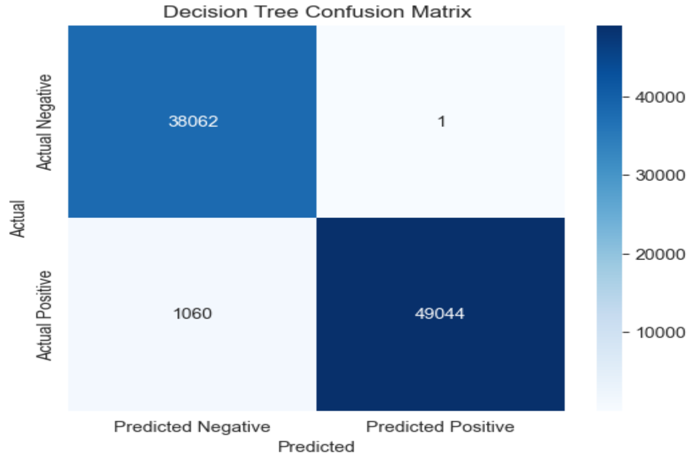
Accuracy Comparison

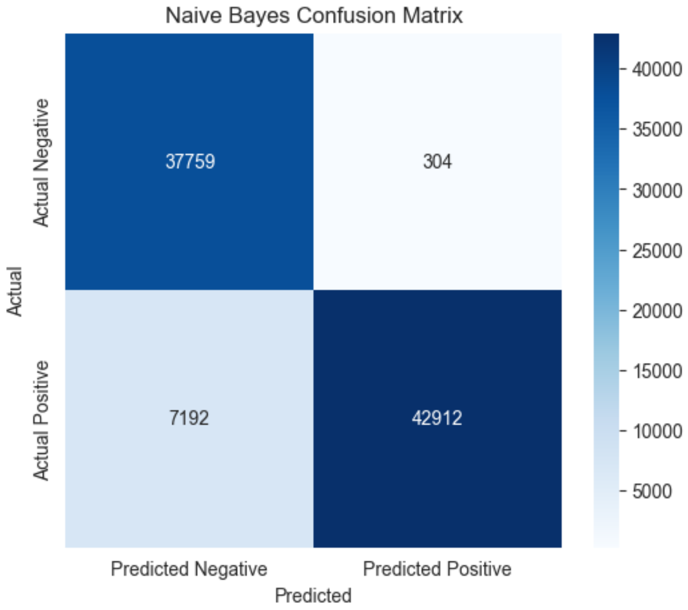
Model	Accuracy
Decision Tree	98.8%
Naive Bayes	91.5%
Logistic Regression	89.2%

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

8. Ensemble Modeling: Stacking

To further enhance prediction performance, a stacking ensemble approach was implemented:

Stacking Architecture:

- Base Models:
    - Decision Tree (with optimized parameters)
    - Naive Bayes (with optimized parameters)

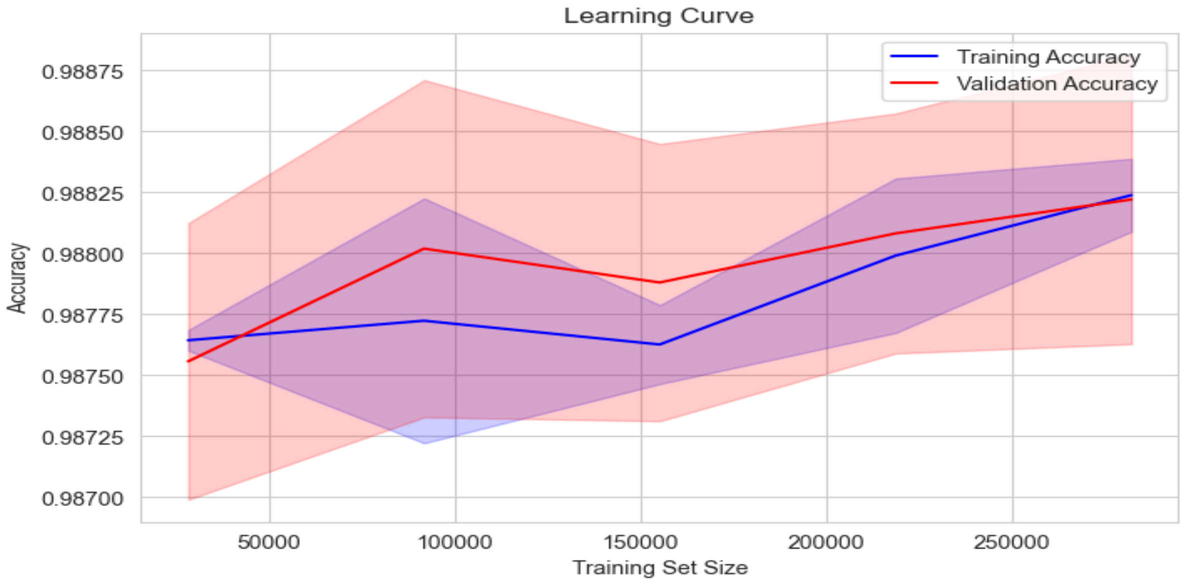
- Meta-Model:
    - Logistic Regression (C=100, class_weight='balanced', max_iter=10000)

- Cross-Validation:
    - 5-fold cross-validation for generating meta-features

Stacking Performance:

- Accuracy: 98.8%
- Confusion Matrix:
    - Similar performance to the best individual model (Decision Tree)

- Learning Curve Analysis:
    - Stable performance across different training set sizes
    - Minimal gap between training and validation accuracy, indicating good generalization



The stacking model maintained the high performance of the Decision Tree while potentially offering more robust predictions across different customer segments. The similar performance to the Decision Tree suggests that the Decision Tree already captured most of the predictive patterns in the data.

9. Model Deployment with Streamlit

To make the churn prediction model accessible to business users, a user-friendly web application was developed using Streamlit. This interactive dashboard allows non-technical stakeholders to leverage the predictive power of the model without requiring data science expertise.

Deployment Architecture:

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

- Trained Decision Tree model (dt_model.pkl)
- Cleaned customer dataset (cleaned_customer_data.csv)
- Interactive web interface with multiple pages

Key Features:

1. Prediction Interface:

1. Input form for customer attributes (Age, Support Calls, Payment Delay,..)
2. Real-time prediction of churn probability
3. Visual representation of prediction results
4. Color-coded outcomes (red for churn risk, green for retention)

2. Data Insights Dashboard:

1. Interactive data exploration tools
2. Filtering capabilities by churn status
3. Customer search functionality by ID
4. Downloadable filtered data in CSV format
5. Visual analytics including:

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

3. About Section:

1. Model documentation and performance metrics
2. Feature descriptions and preprocessing details
3. Usage guidelines for business users

Business Applications:

The deployed application enables several key business functions:

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

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

3. Data-Driven Insights:
Managers can explore patterns in the customer base through the interactive visualizations, identifying segments with higher churn risk.

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

This deployment represents the operationalization of the machine learning pipeline, transforming the predictive model into an actionable business tool that can drive customer retention strategies and support data-driven decision-making across the organization.

10. Conclusions and Business Implications

The analysis yielded several actionable insights for customer retention strategies:

Key Predictors of Churn:
1. Contract Length: Monthly contracts significantly increase churn risk
2. Support Calls: More than 5 support calls strongly indicates dissatisfaction
3. Total Spend: Lower spending customers are more likely to churn
4. Payment Delays: Longer payment delays correlate with higher churn probability

Customer Lifecycle Insights:

1. The first 6 months represent a critical period for customer retention
2. The 15-25 month period (likely contract renewal) requires special attention
3. Long-term customers (30+ months) show significantly higher loyalty

Recommended Business Strategies:

1. Improve Early Customer Experience:

1. Enhanced onboarding processes
2. Regular check-ins during first 6 months
3. Early-stage incentives and support

2. Enhance Support Quality:

1. Improve first-call resolution rates
2. Proactive issue identification
3. Special handling for customers with multiple support calls

3. Optimize Contract Structure:

1. Incentivize longer-term contracts
2. Special offers at renewal periods
3. Loyalty benefits for contract extensions

4. Address Payment Processes:

1. Simplify payment procedures
2. Flexible payment options
3. Proactive communication about upcoming payments

This comprehensive analysis provides a solid foundation for data-driven customer retention strategies, with the potential to significantly reduce churn rates and enhance customer lifetime value.