

Telecommunications Customer Churn Prediction Using Data Mining and Predictive Models

CSCE 5380 – Data Mining - Final Project | Group 3

Abhiram Reddy Pudi - ID: 11817072

Ramanaji Boggarapu - ID: 11718646

Shiva Sayi Shesshank Mamidi - ID: 11752768

Sai Kumar Ramisetti - ID: 11822974

Presentation Outline

- ⌚ Problem Definition & Hypothesis
- 📘 Methodology Overview
- 🔍 Part 1: Discovering Patterns in the Data
- ⚙️ Part 2: Building and Validating a Predictive Model
- ☑️ Conclusions & Actionable Recommendations

The Business Problem & Dataset

- **Business Goal:** Proactively identify and retain customers at risk of churning to reduce revenue loss.
- **Dataset:** Telco Customer Churn (7,043 Customers, 20 Features).
- **Target Variable:** Churn (1 = Yes, 0 = No).

```
First 5 rows of the dataset:
   gender SeniorCitizen Partner Dependents tenure PhoneService \
0  Female                 0      Yes        No     1        No
1   Male                  0       No        No    34       Yes
2   Male                  0       No        No     2       Yes
3   Male                  0       No        No    45       No
4  Female                 0       No        No     2       Yes

   MultipleLines InternetService OnlineSecurity OnlineBackup \
0  No phone service           DSL        No      Yes
1            No                DSL       Yes      No
2            No                DSL       Yes      Yes
3  No phone service           DSL       Yes      No
4            No      Fiber optic     No      No

   DeviceProtection TechSupport StreamingTV StreamingMovies      Contract \
0             No          No         No        No Month-to-month
1            Yes          No         No        No    One year
2             No          No         No        No Month-to-month
3            Yes          Yes        No        No    One year
4             No          No         No        No Month-to-month

   PaperlessBilling          PaymentMethod MonthlyCharges  TotalCharges \
0            Yes Electronic check      29.85        29.85
1            No  Mailed check        56.95  1889.50
2            Yes  Mailed check        53.85       108.15
3            No Bank transfer (automatic)  42.30  1840.75
4            Yes Electronic check      70.70       151.65

   Churn
0      0
1      0
2      1
3      0
4      1
```

Our Research Questions & Hypothesis

Our Research Questions

- 1 What are the key patterns and customer segments that
· drive churn?
- 2 Can we build a reliable predictive model to identify these
· at-risk customers?

Our Hypothesis

Churn is driven by identifiable factors like contract terms and service choices, not random chance.

Methodology Overview



1. Data Mining & EDA

- Visual Analysis of Churn Drivers
- K-Means Clustering to find segments
- Association Rule Mining to find service bundles



2. Predictive Modeling

- Models: Logistic Regression, Decision Tree, SVM
- Method: GridSearchCV with 5-fold Cross-Validation



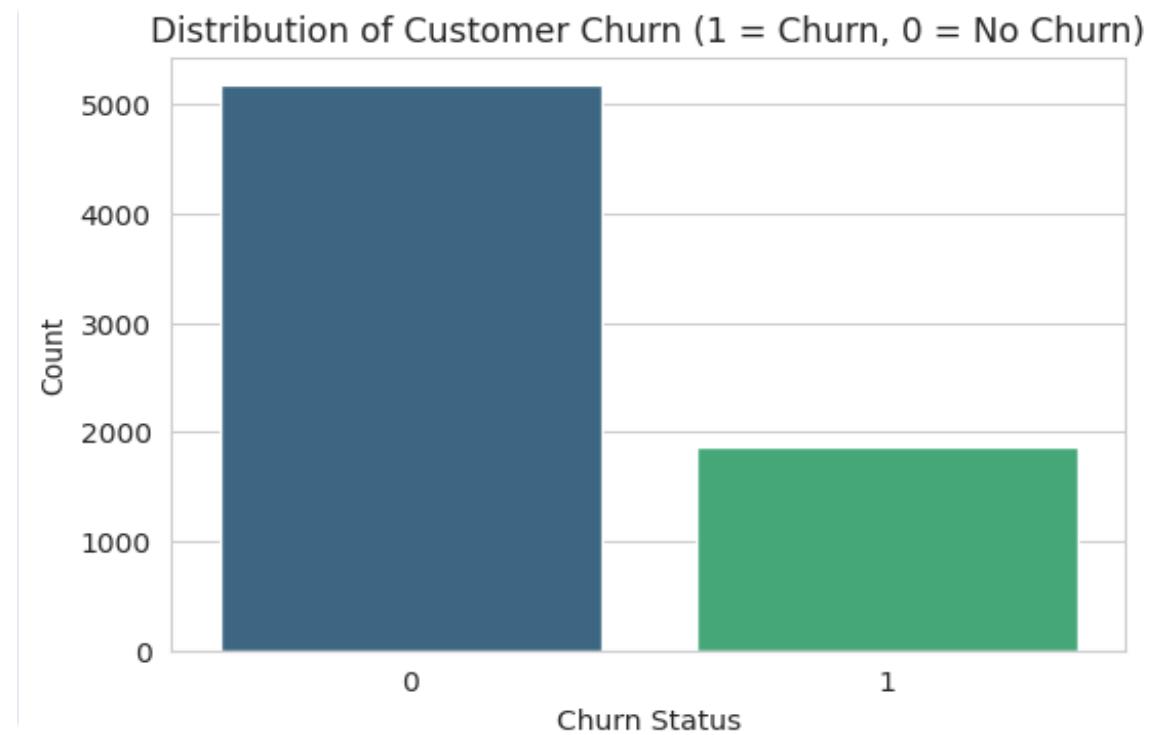
3. Evaluation & Interpretation

- Test Set Performance (AUC, Classification Report)
- Statistical Significance Testing (Paired t-test)

Part 1: Discovering Patterns

Dataset Overview: An Imbalanced Problem

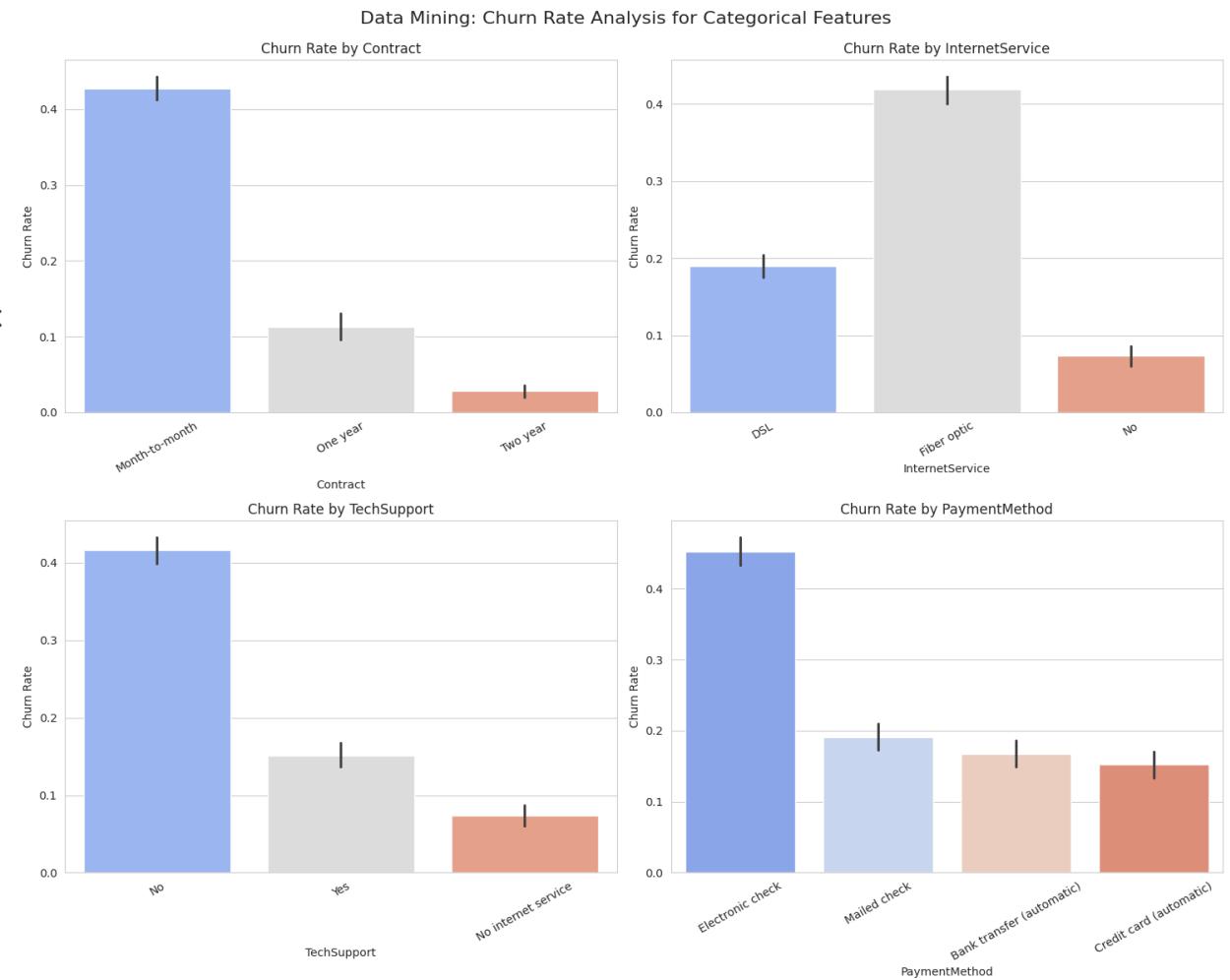
- The dataset is **imbalanced**, with a **26.5%** overall churn rate.
- This requires careful handling during model training (e.g.,
`'class_weight='balanced'`) to avoid bias.



Finding #1: Churn Varies Significantly by Category

Key Insight:

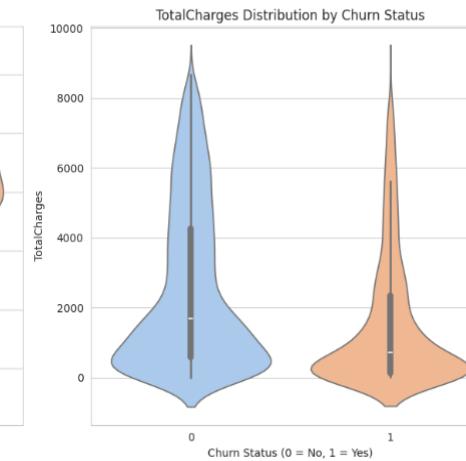
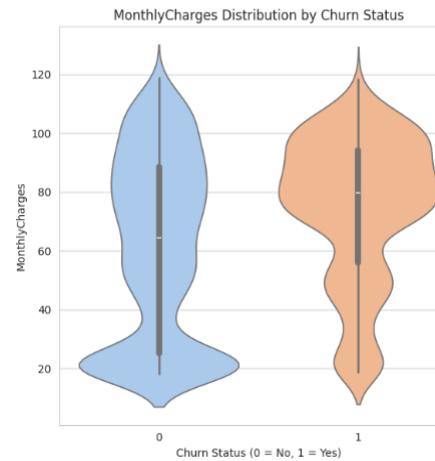
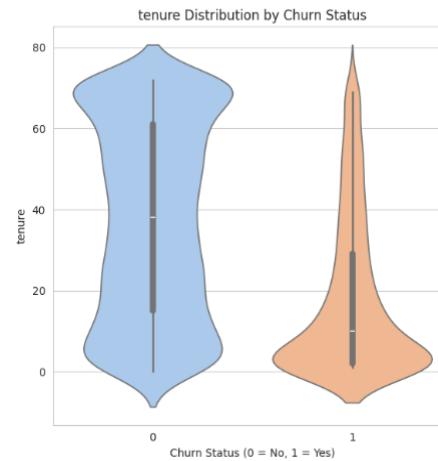
Customers with Month-to-Month Contracts and Fiber Optic Service have the highest churn rates.



Finding #2: The Profile of a Churning Customer

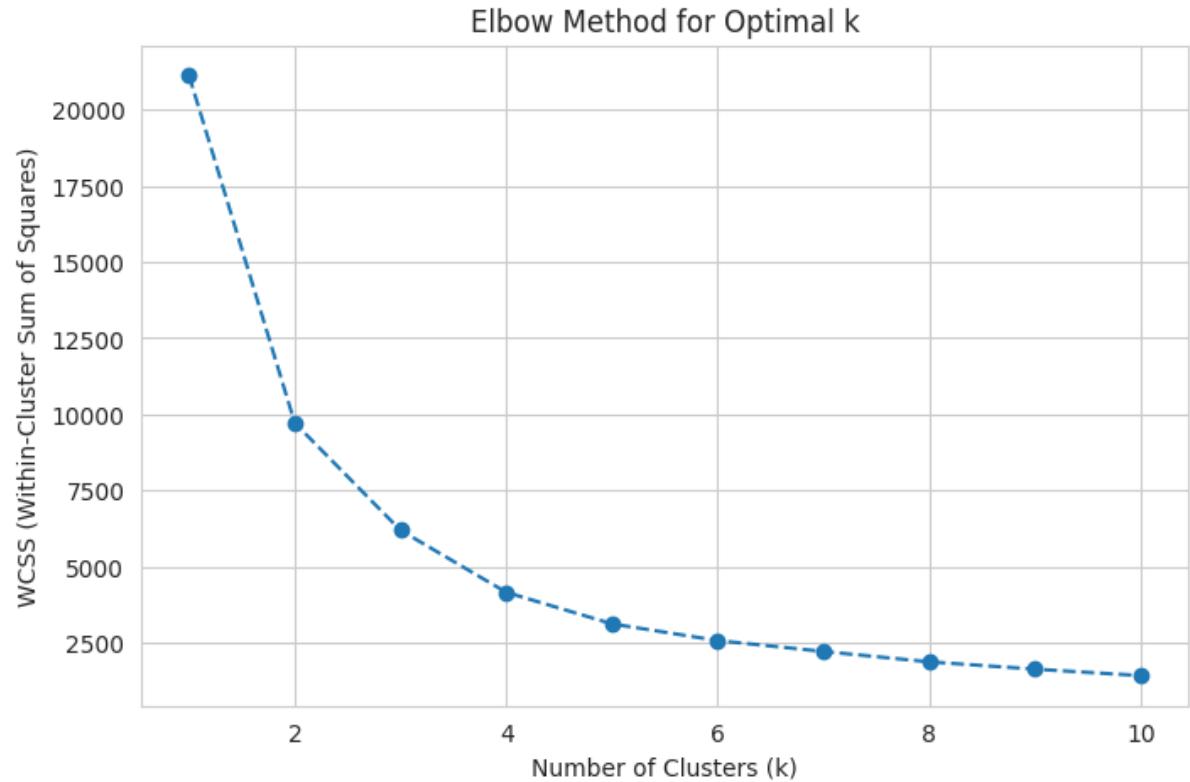
Customers who churn (Status = 1) typically have:

- Lower Tenure
- Higher Monthly Charges



Finding #3: Identifying Natural Customer Segments

- We used **K-Means clustering** to find distinct customer groups based on their tenure and charges.
- The **Elbow Method** plot shows a clear "elbow" at **k=3**, justifying our choice of three segments.



Uncovering the "High-Risk Newcomer" Segment

Characteristics and Churn Rate of Each Segment:

Customer Segment	SegmentSize	tenure	MonthlyCharges	TotalCharges	Churn Rate
0	2154	29.50	26.57	809.42	12.30%
1	2200	58.56	89.70	5246.13	15.36%
2	2689	13.25	74.95	1030.57	47.08%

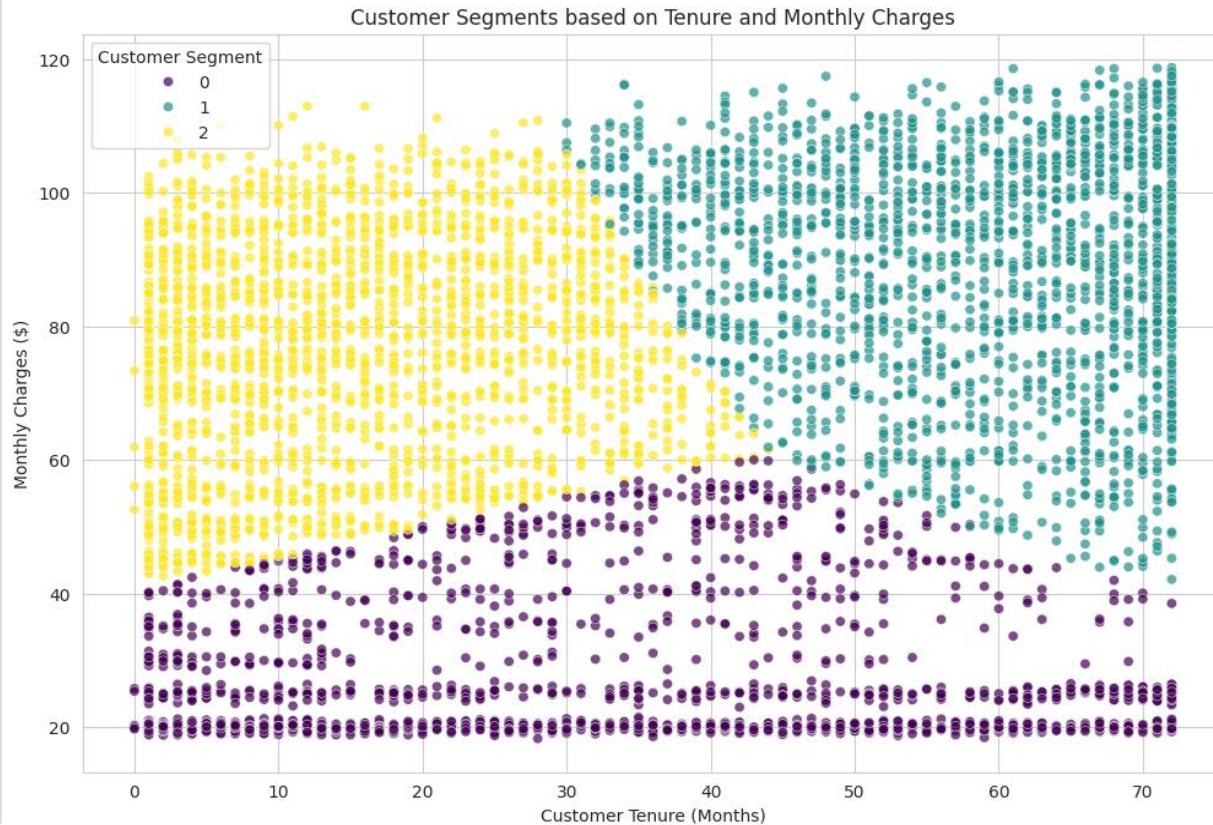
- Segment 2 is the critical group:
- Low Average Tenure: 13 months
- High Average Monthly Bill: \$75
- Alarming Churn Rate: 47%

Visual Confirmation of the High-Risk Segment

This scatter plot visually confirms the location of our high-risk segment.

The yellow cluster (Segment 2) is clearly concentrated in the low-tenure, high-monthly-charge area.

High-Risk Segment
(47% Churn)



Finding #4: Risky Service Combinations

- Association Rule Mining finds "if-then" patterns with high confidence.

Top 5 Rules Leading to Customer Churn (Confidence > 0.68):

Antecedents (Service Combinations)	Consequents	Support	Confidence	Lift
{No OnlineBackup, PaperlessBilling, Fiber Optic}	{Churn=Yes}	0.0574	0.7038	2.652
{PaperlessBilling, Fiber Optic, No TechSupport}	{Churn=Yes}	0.0585	0.6913	2.605
{No OnlineBackup, No TechSupport, Fiber Optic}	{Churn=Yes}	0.0586	0.6883	2.594
{PaperlessBilling, Fiber Optic, No DeviceProtection}	{Churn=Yes}	0.0625	0.6864	2.587
{No OnlineBackup, PaperlessBilling, Fiber Optic, ...}	{Churn=Yes}	0.0628	0.6842	2.578

Top Rule Found: Customers with {No Online Backup, No Tech Support, Fiber Optic, etc.} have a ~70% probability of churning.

Part 2: Predictive Modeling & Validation

Model Performance Comparison

- We trained and tuned three models to automatically identify at-risk customers.

Summary of Final Model Performance on Test Set:

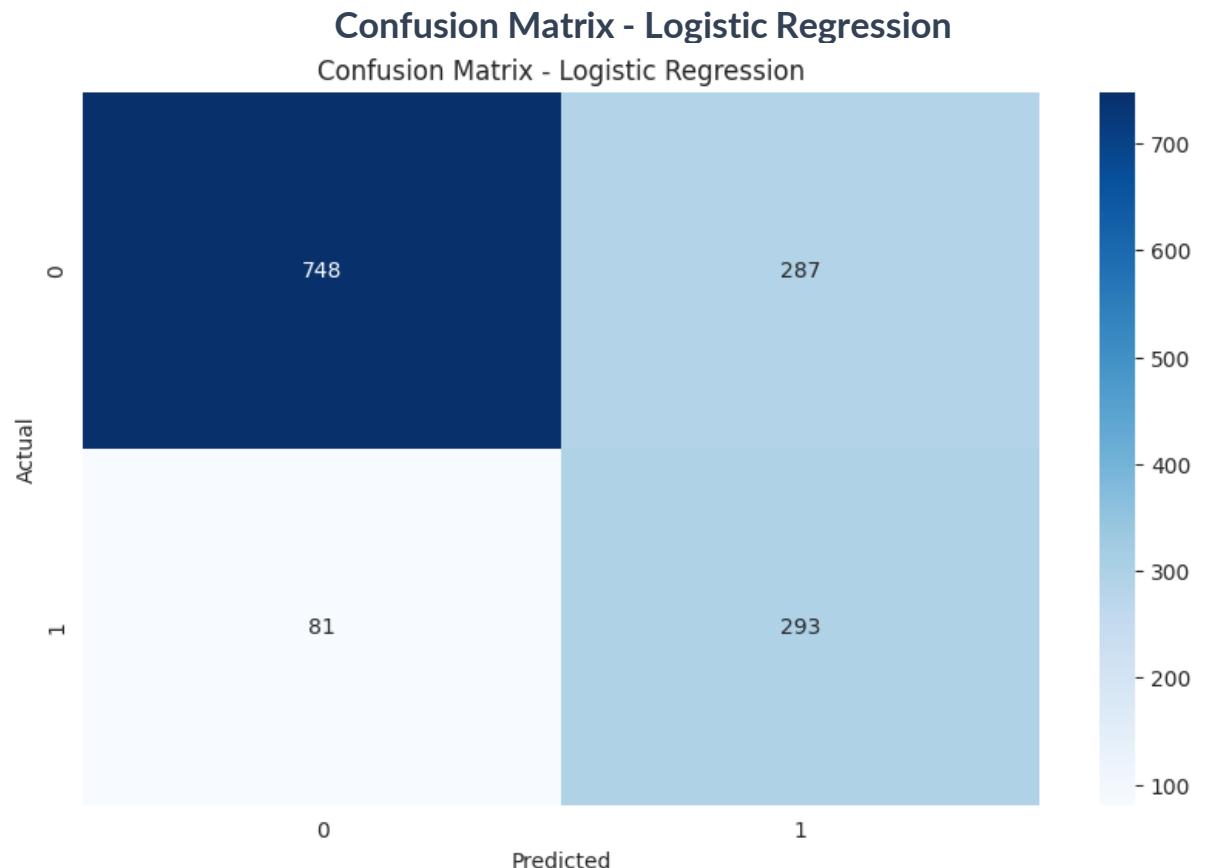
Model	Best CV AUC Score	Test Set Accuracy	Test Set Recall (Churn=1)	Test Set AUC
Logistic Regression	0.8456	0.74	0.78	0.841
Decision Tree	0.8339	0.74	0.81	0.832
Support Vector Machine (SVM)	0.8385	0.73	0.79	0.835

Logistic Regression achieved the best overall performance with an **AUC of 0.841**.

Deep Dive into the Best Model's Performance

- Recall for Churn (1): 0.78 -> Correctly identifies 78% of all customers who actually churned.
- **293 True Positives:** Successfully flagged 293 future churners.

Classification Report for Logistic Regression:



Statistical Validation: Is the Result Reliable?

- Method: Paired t-test on 10-fold cross-validation scores.
- T-statistic: 4.1433
- P-value: 0.0025 (which is < 0.05)
- Conclusion: The superior performance of **Logistic Regression** is statistically significant.

--- 5b. Statistical Significance Testing ---

Paired t-test between Logistic Regression and Decision Tree (based on 10-fold CV AUC scores):

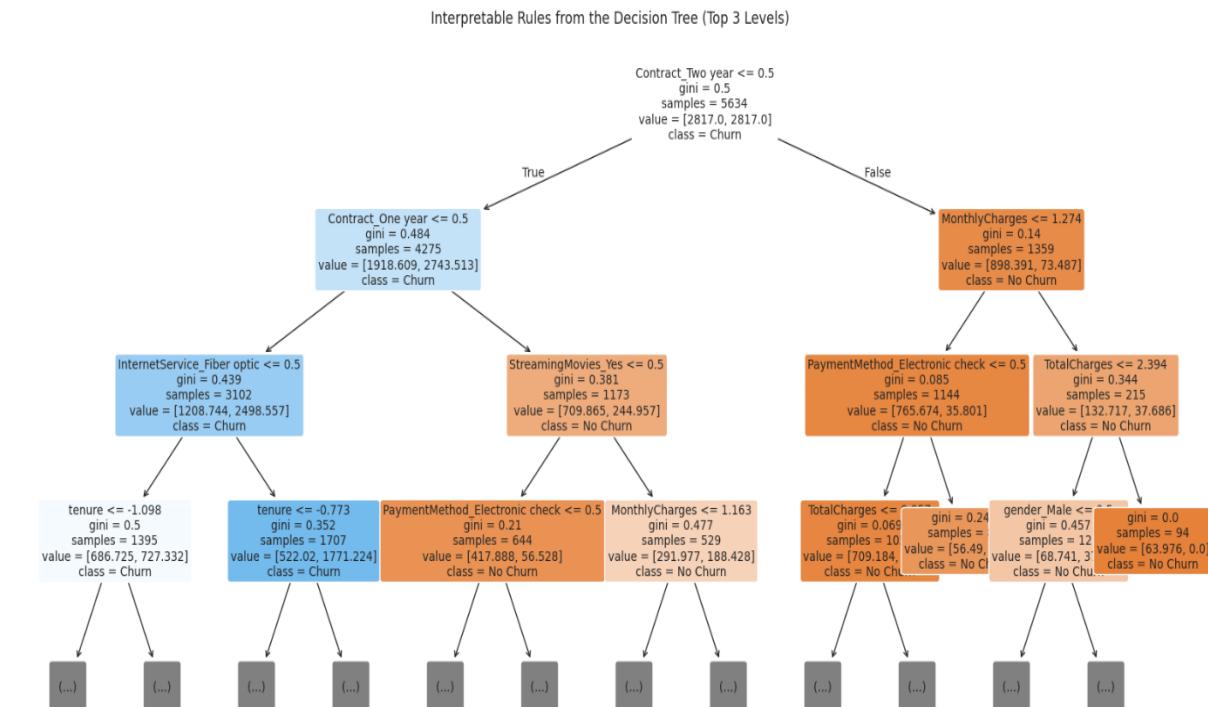
T-statistic: 4.1433

P-value: 0.0025

Conclusion: The performance difference is statistically significant. We can be confident that Logistic Regression is superior.

Interpreting Model Logic with a Decision Tree

- The Decision Tree provides a simple, human-readable flowchart.
- The most important feature at the top is **Contract Type**, confirming our initial EDA findings.



Interpreting the Winning Model's Logic

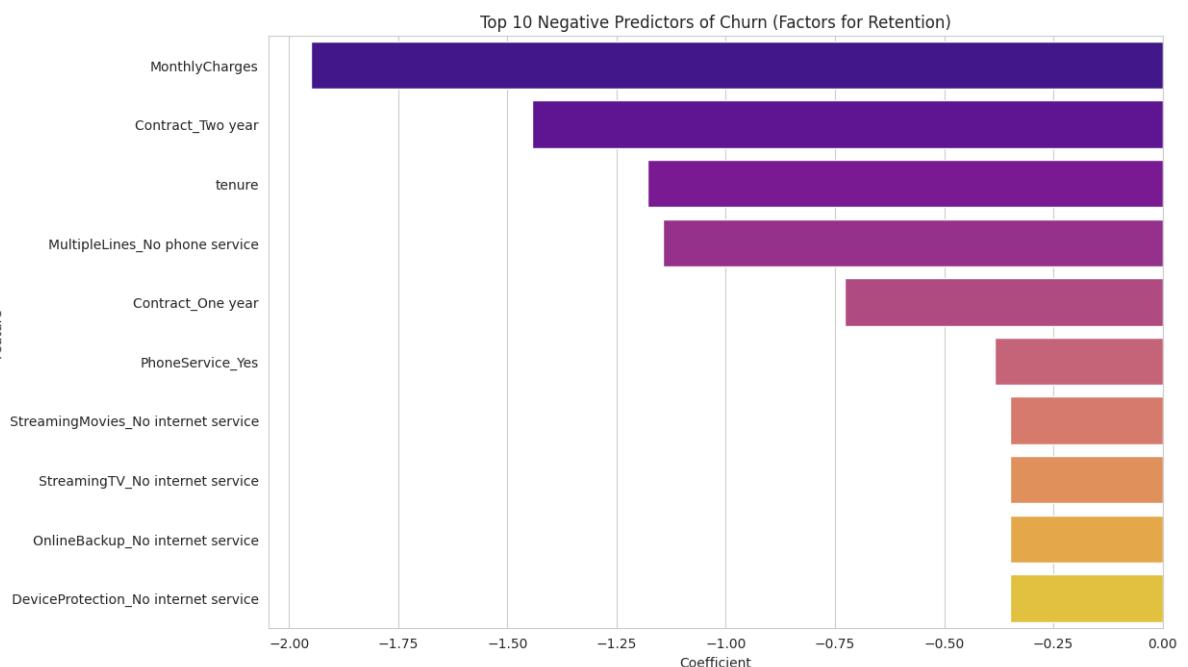
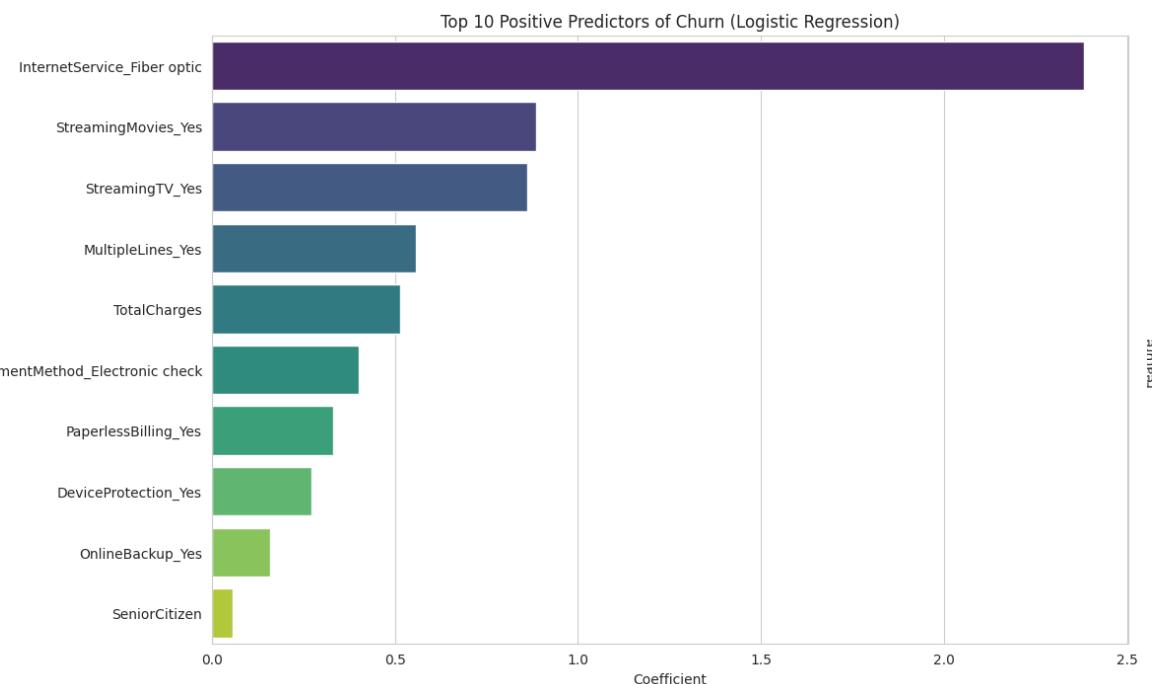
The model's logic confirms our data mining insights:

Strongest Churn Driver:

Fiber Optic Service

Strongest Retention Driver:

Two-Year Contract



Part 4: Conclusion & Future Work

Summary of Key Findings



1. Discovered Key Drivers

Churn is primarily driven by **Month-to-Month contracts** and **Fiber Optic service**.



2. Identified High-Risk Segment

Uncovered a "High-Risk Newcomer" segment (low tenure, high bill) with a **47% churn rate**.



3. Built a Statistically Superior Model

A tuned **Logistic Regression** model is the best predictor with a statistically significant **AUC of 0.841**.

Actionable Recommendations

- ⌚ **Targeted Campaign:** Proactively offer customers in the "**High-Risk Newcomer**" segment a discount to switch to an annual contract or provide complimentary protective services (e.g., Tech Support).
- 📶 **Improve Fiber Onboarding:** Review the initial experience for new **Fiber Optic** customers to address potential service or cost concerns early on.
- 🤝 **Promote Loyalty:** Incentivize **long-term contracts** at sign-up, as this is the strongest factor for customer retention.

Reflection & Lessons Learned

-  **Hybrid Approach is Powerful:** Combining unsupervised discovery (K-Means) with supervised prediction leads to richer, more defensible insights.
-  **Interpretability is a Key Feature:** An interpretable model (Logistic Regression) allows us to understand the "why" and build trust in the results.
-  **Statistical Validation is Crucial:** The t-test was essential for moving from an observation ("this model seems better") to a confident conclusion ("this model **is** better").

Future Directions



Survival Analysis

Instead of predicting **if** a customer will churn, predict **when**.



Incorporate Unstructured Data

Analyze customer support transcripts with NLP to capture sentiment.



Explore Advanced Models

Use models like XGBoost to automatically capture more complex feature interactions.

Limitations

Static Dataset: The analysis is based on a single snapshot in time and does not capture the customer's journey or trends leading up to the churn event..

Limited Feature Scope: The dataset lacks external data like customer service logs or competitor offers, which could provide deeper insights into customer satisfaction.

Model Choice Trade-off: We prioritized an interpretable model (Logistic Regression). More complex models might offer slightly higher accuracy but would be less transparent and harder to explain.

Individual Contributions

Abhiram Reddy Pudi

Data Mining & Pattern Discovery: EDA, K-Means Clustering, and Association Rule Mining.

Shiva Sayi Shesshank Mamidi

Predictive Modeling & Optimization: Model Configuration and GridSearchCV Tuning.

Ramanaji Boggarapu

Data Preparation & Preprocessing: Data Cleaning and Column Transformer Pipeline.

Sai Kumar Ramisetti

Model Evaluation & Interpretation: Test Set Evaluation, Paired t-test, and Model Logic Interpretation.

References

- 1 "Telco Customer Churn," Kaggle, 2020. [Online]. Available: <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>
- 2 F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, Oct. 2011.
- 3 J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2012.
- 4 W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction...," *European Journal of Operational Research*, vol. 218, no. 1, pp. 211-229, Apr. 2012.
- 5 K. Coussement and D. Van den Poel, "Churn prediction in subscription services...," *Expert Systems with Applications*, vol. 34, no. 1, pp. 489-498, Jan. 2008.

Thank You

Questions?