Predicting Telco Customer Churn

A Business-Centric Approach

Group: 14

Author	**ID**
Farhan Zarif	23301692
H.M. Hasnain Jahangir Aqib	22241077

September 7, 2025

Given Data Set: telco_customer_churn.csv

Contents

1 Introduction

This project aims to build a robust machine learning model to predict customer churn for a telecommunications company. The primary problem being solved is the costly issue of customer attrition. Acquiring a new customer is significantly more expensive than retaining an existing one, making churn prediction a critical business function.

The motivation behind this project is to move beyond simple accuracy metrics and develop a model that provides tangible business value. The final model is selected based on a custom business-centric metric that quantifies the financial impact of retention efforts, balancing the cost of interventions against the revenue saved from preventing churn. This analysis identifies key drivers of churn, builds multiple predictive models, and provides actionable insights for targeted customer retention strategies.

2 Dataset Description

2.1 Dataset Overview

- **How many features?** The dataset initially contains 21 features (columns).
- **How many data points?** There are 7043 data points (rows), each representing a unique customer.
- **Classification or Regression?** This is a **classification problem**. The goal is to predict a categorical outcome: whether a customer will churn ('Yes') or not ('No'). The target variable is binary, not a continuous value, making it unsuitable for regression.
- **Feature Types:** The dataset contains a mix of quantitative and categorical features.
  - **Quantitative (Numerical):** `SeniorCitizen`, `tenure`, `MonthlyCharges`, `TotalCharges`.
  - **Categorical (Object):** `gender`, `Partner`, `Dependents`, `PhoneService`, `MultipleLines`, `InternetService`, `OnlineSecurity`, `OnlineBackup`, `DeviceProtection`, `TechSupport`, `StreamingTV`, `StreamingMovies`, `Contract`, `PaperlessBilling`, `PaymentMethod`, and the target variable `Churn`.
- **Encoding Categorical Variables:** Yes, encoding is necessary. Machine learning algorithms operate on numerical data. Categorical features like 'Contract' (Month-to-month, One year, Two year) must be converted into a numerical format. This is done using One-Hot Encoding to create binary columns for each category, preventing the model from incorrectly assuming an ordinal relationship between non-ordinal categories.

2.2 Correlation of Features

To understand the relationships between features and the target variable (`Churn`), two types of correlation analyses were performed.

2.2.1 Numeric Features (Spearman Correlation)

Spearman correlation was used for numeric features.

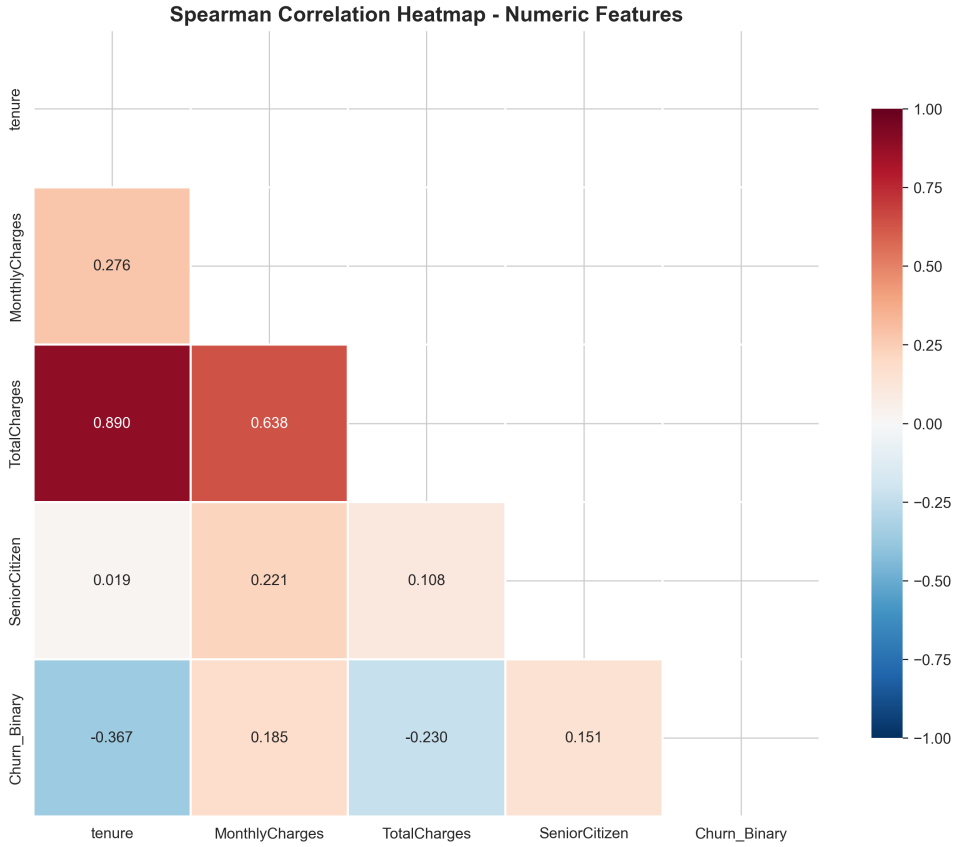

Figure 1: Spearman Correlation Heatmap for Numeric Features.

Understanding: From the heatmap, we observe that `tenure` has the strongest negative correlation with churn (-0.367), indicating that customers with longer tenures are significantly less likely to churn. Conversely, `MonthlyCharges` has a positive correlation (0.185). There is also a very high correlation (0.890) between `tenure` and `TotalCharges`, indicating strong multicollinearity.

2.2.2 Categorical Features (Cramér's V)

Cramér's V was used to measure the association between categorical features.

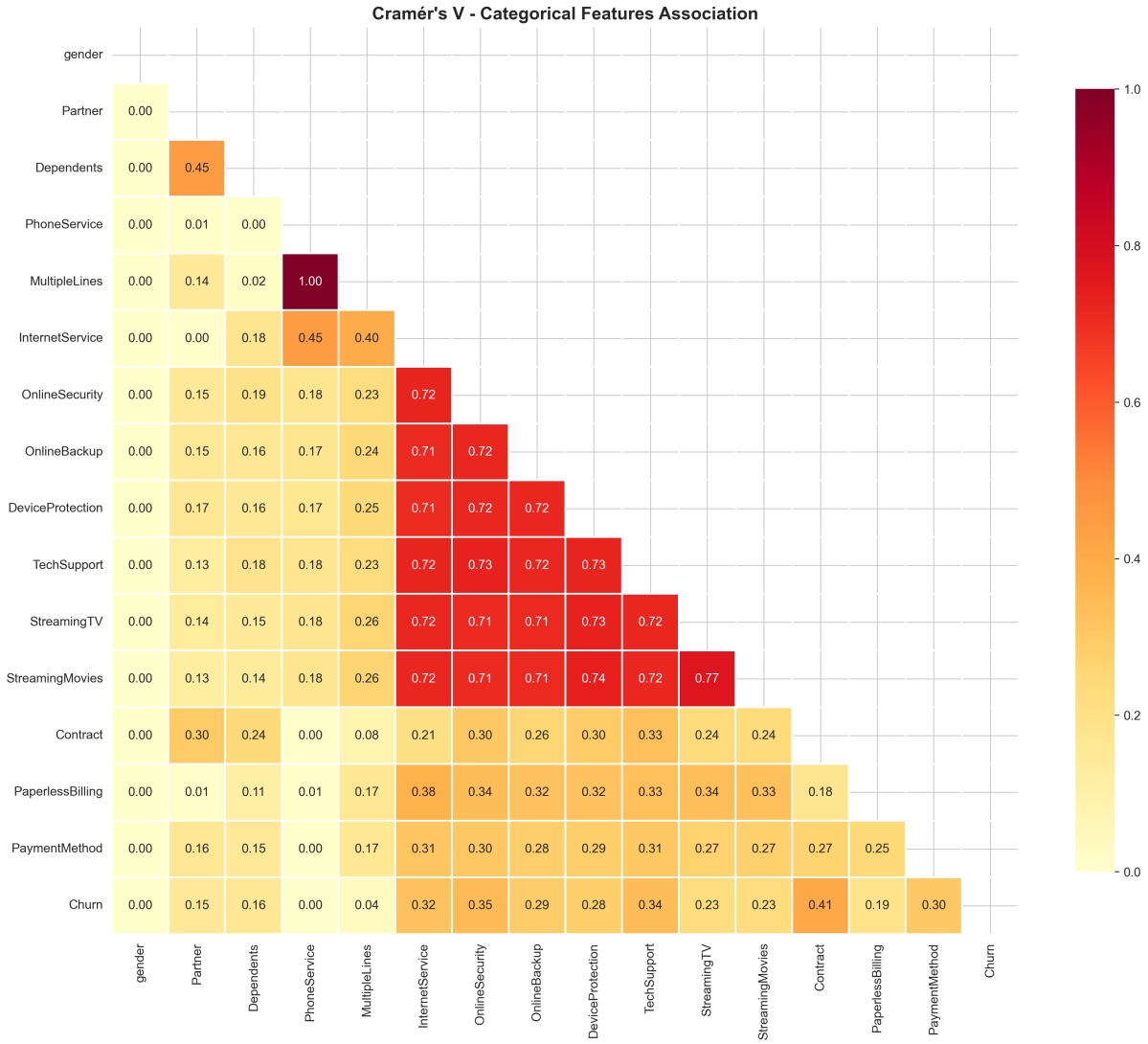

Figure 2: Cramér's V Association Heatmap for Categorical Features.

Understanding: The `Contract` type has the strongest association with churn (0.410). Features like `OnlineSecurity` (0.347), `TechSupport` (0.343), and `InternetService` (0.322) also show strong associations, suggesting that service-related features are major drivers of customer loyalty or churn.

2.3 Imbalanced Dataset

The dataset is imbalanced, which means the classes in the target variable are not represented equally.

- **Class Distribution:** For the output feature `Churn`, the classes are not equal. 5174 instances are 'No' (73.5%) and 1869 instances are 'Yes' (26.5%).
- **Bar Chart Representation:** The class imbalance is visualized below.

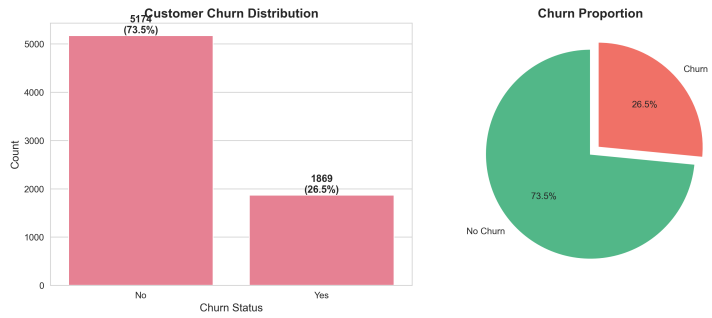

Figure 3: Customer Churn Distribution.

2.4 Exploratory Data Analysis (EDA)

EDA revealed several key relationships:

- **Contract vs. Churn:** Customers on a 'Month-to-month' contract have a very high churn rate (42.7%), while those on a 'Two year' contract have a very low rate (2.8%).
- **Internet Service vs. Churn:** Customers with 'Fiber optic' internet service have a significantly higher churn rate (41.9%) compared to those with 'DSL' (19.0%).
- **Tenure vs. Churn:** The distribution of tenure shows that customers who churn typically have a much shorter tenure (mean of 18 months) compared to those who do not (mean of 37.5 months).

3 Dataset Pre-processing

Several data quality issues were identified and addressed.

Problem 1: Missing Values The `TotalCharges` column contained 11 empty string values, which appeared as missing after attempting to convert the column to a numeric type.

Solution 1: Imputation Upon investigation, all 11 instances corresponded to customers with a `tenure` of 0. These are new customers who have not yet incurred any total charges. Therefore, the missing values were justifiably imputed with 0.

Problem 2: Categorical Values The dataset contains 18 object-type (categorical) columns that cannot be directly used by machine learning models.

Solution 2: One-Hot Encoding To convert these features into a numerical format, `OneHotEncoder` was used. This technique creates new binary columns for each category within a feature, ensuring that the model does not infer any ordinal relationship.

Problem 3: Feature Scaling Numeric features like `tenure`, `MonthlyCharges`, and `TotalCharges` exist on vastly different scales. This can cause models that are sensitive to feature magnitude (like Logistic Regression, KNN, and Neural Networks) to perform poorly.

Solution 3: Standardization `StandardScaler` was applied to all numeric features. This process transforms the data to have a mean of 0 and a standard deviation of 1, ensuring all features contribute equally to the model's learning process.

Problem 4: Multicollinearity A Variance Inflation Factor (VIF) analysis revealed moderate multicollinearity between `TotalCharges` and `tenure` (VIF scores of 8.08 and 6.33, respectively). This can be problematic for linear models as it can inflate the variance of the coefficient estimates.

Solution 4: Feature Selection To address this, two separate feature sets were created. For linear models (e.g., Logistic Regression), which are sensitive to multicollinearity, the `TotalCharges` feature was dropped. For non-linear, tree-based models, which are more robust to this issue, the feature was retained.

4 Dataset Splitting

The dataset was split into a training set and a test set to evaluate the models' performance on unseen data.

- **Ratio:** A 70%-30% split was used, with 70% of the data for training (4930 samples) and 30% for testing (2113 samples).
- **Method: Stratified Splitting** A stratified split was performed based on the target variable `Churn`. This ensures that the proportion of churned to non-churned customers is identical in both the training and test sets. This is a critical step for imbalanced datasets, as it prevents a scenario where one of the sets could end up with a disproportionately low number of minority class samples.

5 Model Training & Testing

5.1 Unsupervised Learning: K-Means Clustering

Before supervised modeling, K-Means clustering was applied as an unsupervised learning technique to identify natural customer segments in the data. The optimal number of clusters was found to be 4.

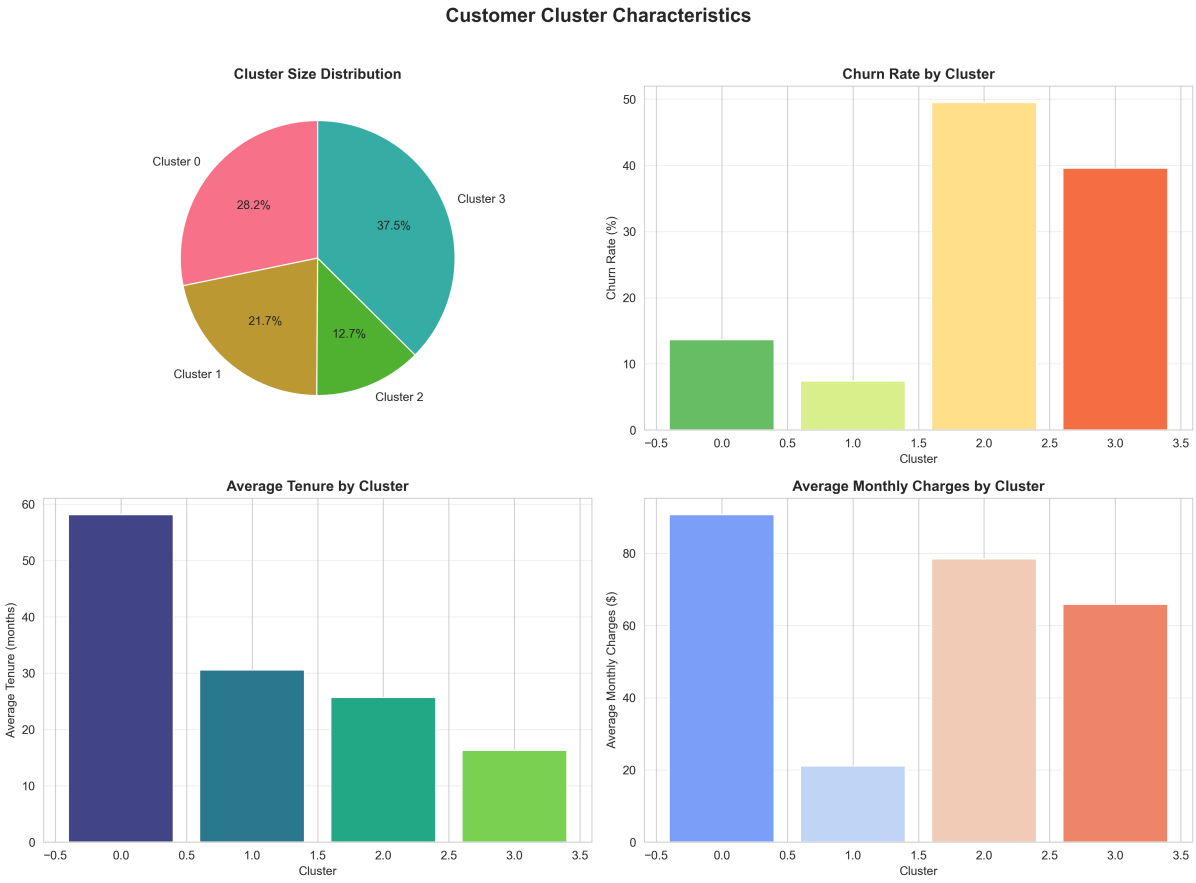

Figure 4: Analysis of the 4 Customer Segments.

The clustering revealed distinct personas:

- **Cluster 0 (Loyal High-Value):** Low churn rate (13.6%), long tenure, high monthly charges, and two-year contracts.
- **Cluster 1 (Loyal Low-Value):** Lowest churn rate (7.4%), moderate tenure, very low monthly charges, often without internet service.
- **Cluster 2 (High-Risk High-Value):** Highest churn rate (49.5%), moderate tenure, high monthly charges, and month-to-month contracts.
- **Cluster 3 (Medium-Risk Newcomers):** High churn rate (39.6%), short tenure, moderate monthly charges, and month-to-month contracts.

5.2 Supervised Learning Models

Five different classification models were trained and tested. Due to the class imbalance, two balancing strategies were applied to the training data: Cluster-Based Undersampling and SMOTE (Synthetic Minority Over-sampling Technique). Each model was trained on both balanced datasets.

The models trained are:

1. **Logistic Regression**
2. **Decision Tree**

3. **Gaussian Naive Bayes**
4. **K-Nearest Neighbors (KNN)**
5. **Neural Network (MLP Classifier)**

6 Model Selection/Comparison Analysis

Models were compared not just on traditional metrics but primarily on a custom "Business Score" designed to maximize financial return from retention campaigns. The score was calculated as: `Score = (TP * $20) + (FP * -$10) + (FN * -$100) + (TN * $0)`

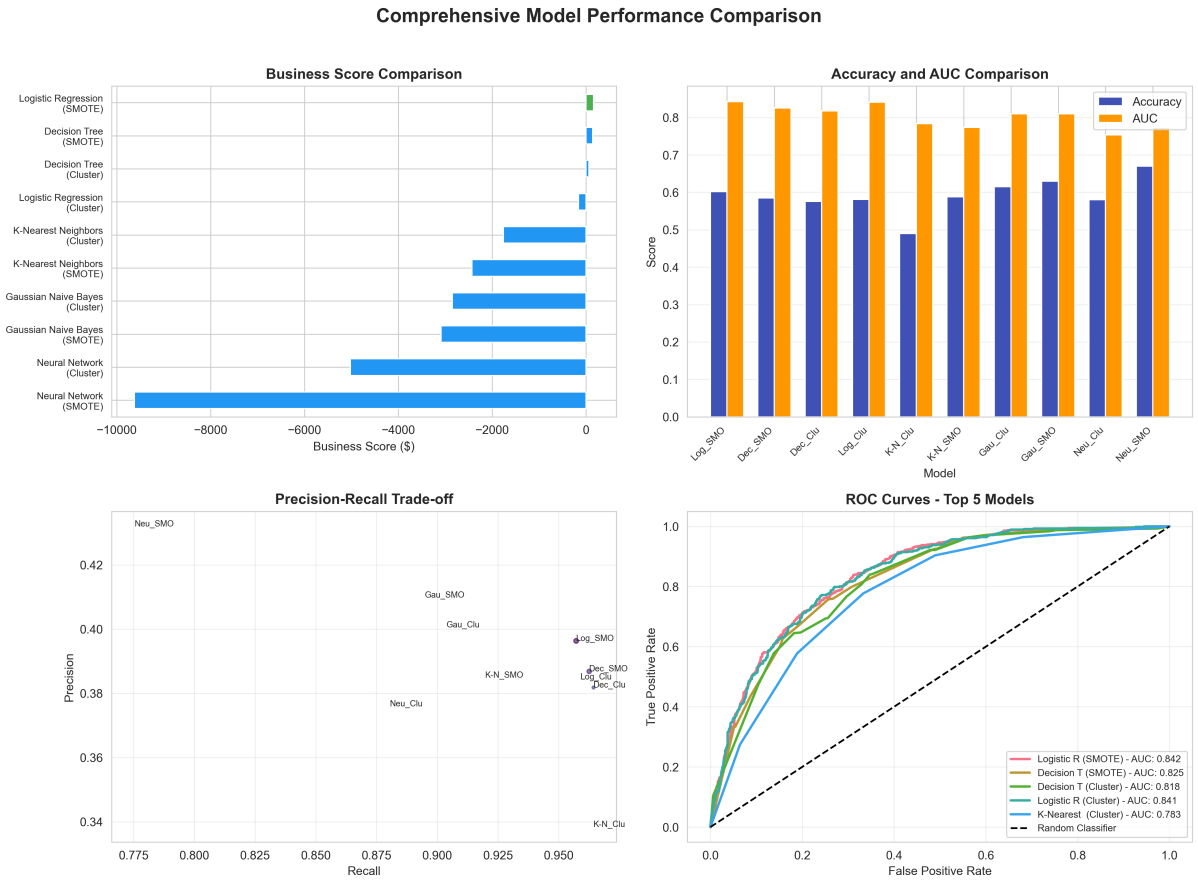

Figure 5: Comprehensive Model Performance Comparison.

Table 1: Final Model Comparison (Sorted by Business Score)

Model	Balancing	Optimal_Threshold	Business_Score	Accuracy	Precision	Recall	AUC
Logistic Regression	**SMOTE**	**0.20**	**$160**	**0.602**	**0.396**	**0.957**	**0.842**
Decision Tree	SMOTE	0.20	$140	0.659	0.428	0.963	0.825
Decision Tree	Cluster	0.25	$60	0.687	0.450	0.964	0.818
Logistic Regression	Cluster	0.20	-$160	0.612	0.403	0.923	0.841
K-Nearest Neighbors	Cluster	0.05	-$1760	0.485	0.340	0.957	0.783
K-Nearest Neighbors	SMOTE	0.05	-$2430	0.435	0.324	0.966	0.774
Gaussian Naive Bayes	Cluster	0.10	-$2850	0.432	0.322	0.954	0.810
Gaussian Naive Bayes	SMOTE	0.05	-$3090	0.413	0.316	0.963	0.810
Neural Network	Cluster	0.05	-$5020	0.334	0.288	0.952	0.753
Neural Network	SMOTE	0.05	-$9620	0.278	0.270	0.982	0.773

Comparison Summary:

- **Best Model:** The **Logistic Regression model trained with SMOTE** achieved the highest Business Score ($160). This model was selected as the final, champion model.
- **Precision vs. Recall:** The best model prioritizes high **Recall (0.957)** over Precision (0.396). This is a strategic choice dictated by the business problem. It is much more costly to miss a potential churner (False Negative) than it is to mistakenly offer a retention incentive to a loyal customer (False Positive). The model is tuned to catch as many potential churners as possible.
- **Confusion Matrix:** The confusion matrix for the best model shows it correctly identified 537 out of 561 churners in the test set, demonstrating its high recall.
- **AUC Score:** The Logistic Regression model also achieved a high AUC score of 0.842, indicating excellent discriminative ability between the positive and negative classes across all thresholds.

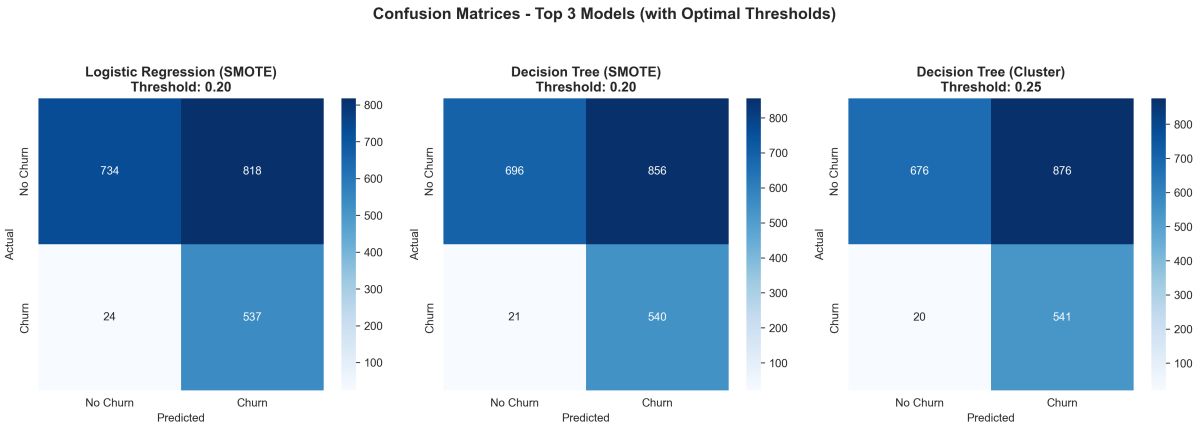

Figure 6: Confusion Matrices for Top 3 Models at Optimal Thresholds.

7 Conclusion

- **Understanding the Results:** The results clearly demonstrate that optimizing for a business-centric metric leads to a more valuable and actionable model than optimizing for accuracy alone. The final Logistic Regression model, despite a modest accuracy of 60.2%, is the most effective because it successfully identifies 95.7% of customers at risk of churning, directly addressing the most expensive aspect of the churn problem.
- **Model Performance Comments:** The model's performance is strong in the context of its business application. The low precision is an accepted trade-off for achieving extremely high recall. This means the marketing team will contact some customers who were not going to churn, but the cost of this is far outweighed by the revenue saved from retaining the high number of correctly identified at-risk customers.

- **Reasons for Results:** The model's success can be attributed to several factors:
    1. **Handling Class Imbalance:** Using SMOTE was crucial. It provided the model with more examples of the minority (churn) class, allowing it to learn the patterns of churning customers more effectively.
    2. **Business-Driven Optimization:** Tuning the decision threshold based on the custom cost matrix allowed us to shift the model's focus from being generally "correct" to being financially "optimal".
    3. **Feature Importance:** The model effectively learned from the strong predictors identified in the EDA, such as contract type and tenure.
- **Challenges Faced:**
    1. **Class Imbalance:** This was the primary challenge. A naive model would have high accuracy by simply predicting "No Churn" for everyone. Overcoming this required specific balancing techniques.
    2. **Defining the Evaluation Metric:** Moving away from standard metrics like accuracy to a custom business score required a clear understanding of the financial implications of the model's predictions.
    3. **Multicollinearity:** The correlation between tenure and total charges required careful feature engineering to ensure the stability of the linear model.