PREDICTING CUSTOMER CHURN IN THE TELECOM INDUSTRY

Term Project Report

Team Members: Fatima Abbasi, Teagan White,

Inez(Ying, Chih) Chang, Shweta

Table of Content

1. EXECUTIVE SUMMARY 3
2. INTRODUCTION 3
2.1 Business Problem 3
2.2 Why This Problem Matters 3
2.3 Data Source 3
3. DATA DESCRIPTION AND EXPLORATORY ANALYSIS 4
3.1 Dataset Overview 4
3.2 Exploratory Analysis Findings..6
Customer Demographics 6
Contract Type and Churn Risk 6
Payment Method Patterns 7
Service Usage Insights 8
Behavior of Numeric Features 8
4. ANALYSIS METHODOLOGY 9
4.1 Data Preparation and Preprocessing 9
Train-Test Split: 10
Handling Class Imbalance (SMOTE): 10
Feature Scaling: 10
4.2 Modeling Techniques 10
Logistic Regression: 10
Decision Tree Classifier 11
4.3 Model Evaluation Metrics 12
5. RESULTS AND BENCHMARKING 13
5.1 Baseline Benchmarks (Before Modeling)..13
This establishes that the model must perform better than 73.4% to be useful. 13
5.2 Model Results Before SMOTE (Original dataset)..13
5.3 Model Results After SMOTE (Balanced Dataset) 14
5.4 Logistic Regression Performance (before SMOTE) 15
Confusion Matrix: 15
Top 10 Most Important Features (by coefficient magnitude): 16
5.5 Decision Tree Performance (Before SMOTE)..17
Confusion Matrix: 17
Top 10 Most Important Features (by importance score): 18
5.6 Logistic Regression After SMOTE (Balanced Dataset) 20
5.7 Decision Tree Performance After SMOTE 20
6. INSIGHTS AND BUSINESS IMPLICATIONS 22
6.1 Actionable Recommendations for Management 23
7. LIMITATIONS AND FUTURE DIRECTIONS 23
7.1 Project Limitations & Future Directions..23
8. CONCLUSION 24

1. EXECUTIVE SUMMARY

This project aims to predict customer churn in the telecommunications industry using machine learning classification techniques. Using a Kaggle dataset of 7,043 telecom customers with 21 features, we developed two predictive models—Logistic Regression and Decision Tree Classifier—to identify customers at high risk of leaving. Our analysis reveals that customer tenure, contract type, and monthly charges are the strongest predictors of churn. The Logistic Regression model achieved 80.4% accuracy with an AUC of 0.836, while the Decision Tree achieved 78.3% accuracy with an AUC of 0.820. Key findings indicate that month-to-month contract customers face significantly higher churn risk compared to long-term contract holders. We recommend targeted retention campaigns for high-risk customer segments, particularly those with short tenure and flexible contract arrangements.

2. INTRODUCTION

2.1 Business Problem

Customer churn—the rate at which customers discontinue service—is a critical metric for telecommunications companies. In our dataset, approximately 26.6% of customers have already churned, indicating a relatively high churn rate within the industry. Retaining existing customers is significantly more cost-effective than acquiring new ones; therefore, accurately identifying at-risk customers enables companies to implement proactive retention strategies such as targeted promotions, service upgrades, and personalized incentives. We want to be able to predict churn customers with accuracy, and maximize recall on the churn customers.

2.2 Why This Problem Matters

For telecom providers, customer lifetime value directly impacts profitability .Churning customers is a high cost to the company, and we want to minimize this by identifying this. To add, if we predicted every customer didn’t churn(0), we would achieve 73.7% total accuracy. This further shows our point that we want to maintain a high total accuracy, while maximizing our recall to identify churn customers. We want to lower the precision of churn customers so that marketing dollars are not wasted on customers that aren’t going to churn. By predicting which customers are likely to leave, the company can allocate retention resources efficiently to high-risk segments, offer timely incentives to prevent churn before it occurs (proactive marketing), improve customer loyalty, and maximize revenue by reducing customer acquisition costs relative to retention costs.

2.3 Data Source

We utilized a publicly available dataset from Kaggle containing 7,043 customer records from a telecommunications company. The dataset includes demographic information, service details, account information, and our target variable. Demographic information includes gender, age, and family status. Service details include internet service type, streaming services, security/backup options, and tech support. Our account information includes contract type, monthly charges, total charges, payment method, and tenure. Our target variable is churn or whether the customer left within the last month.

3. DATA DESCRIPTION AND EXPLORATORY ANALYSIS

3.1 Dataset Overview

The dataset contains information on 7,043 telecommunications customers, including service usage, demographics, contract type, billing method, and whether the customer churned. After cleaning (converting `TotalCharges` to numeric values and removing incomplete rows), approximately 7,000 complete records remained for modeling.

The dataset includes:

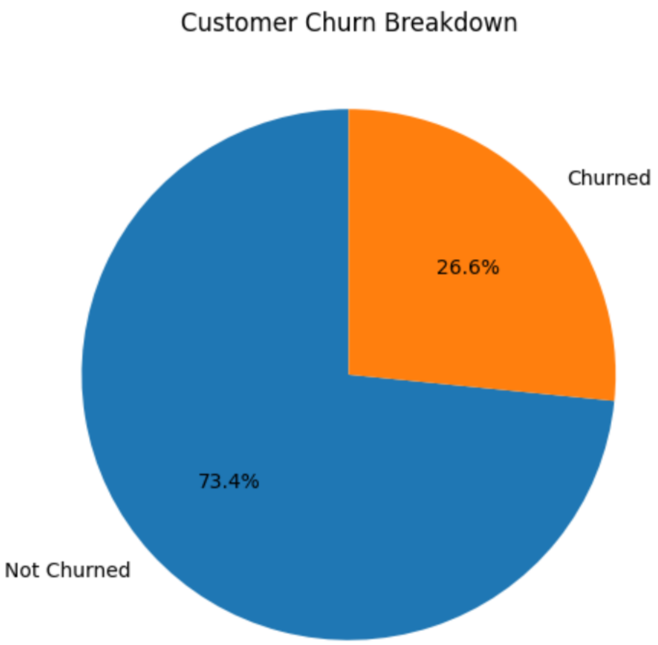
- 21 input features
- 1 binary target variable: `Churn` (Yes = customer left, No = customer stayed)

Table of Feature Types:

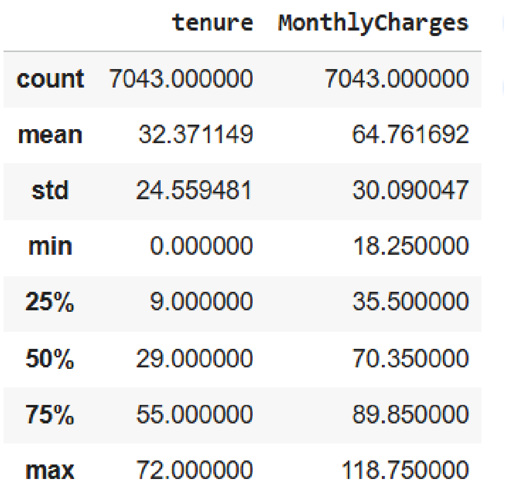
Type	**Examples**
Numerical	Tenure, MonthlyCharges, TotalCharges
Categorical	Gender, Contract Type, Internet Service, Payment Method
Binary Service Indicators	TechSupport, StreamingTV, OnlineSecurity, DeviceProtection
Target Variable	Churn (0 = stayed, 1 = churned)

Class Imbalance (Before SMOTE):

No Churn: 73.4%
Churn: 26.6%



The data set is Imbalanced.This imbalance later justified applying SMOTE oversampling during modeling to ensure churn cases were properly learned.



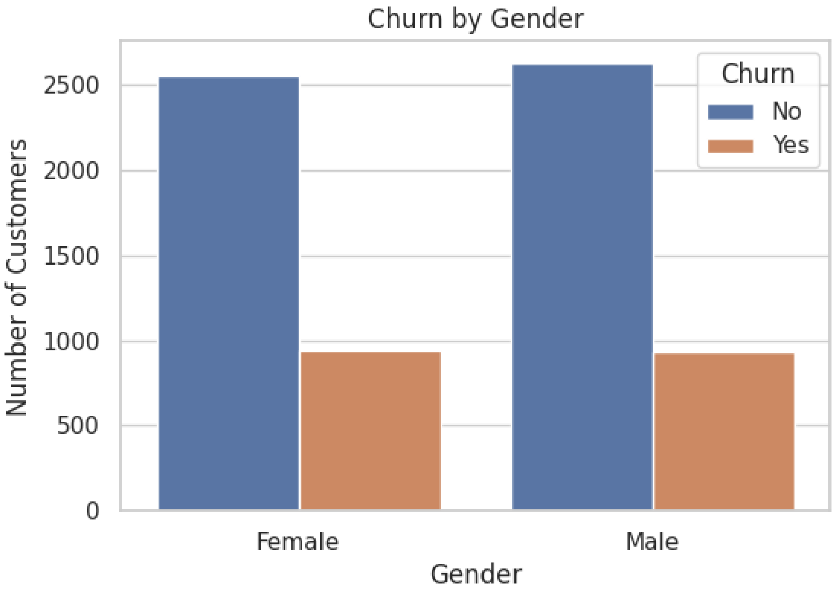
	tenure	MonthlyCharges
count	7043.000000	7043.000000
mean	32.371149	64.761692
std	24.559481	30.090047
min	0.000000	18.250000
25%	9.000000	35.500000
50%	29.000000	70.350000
75%	55.000000	89.850000
max	72.000000	118.750000

MonthlyCharges range from $18 to $118, and tenure ranges from 0 to 72 months, showing a wide variation in customer lifecycle and spending behavior.

3.2 Exploratory Analysis Findings

Customer Demographics

Gender distribution is nearly equal (50.2% Male vs. 49.8% Female), indicating no demographic skew. Most customers are not senior citizens, and only a minority have dependents.

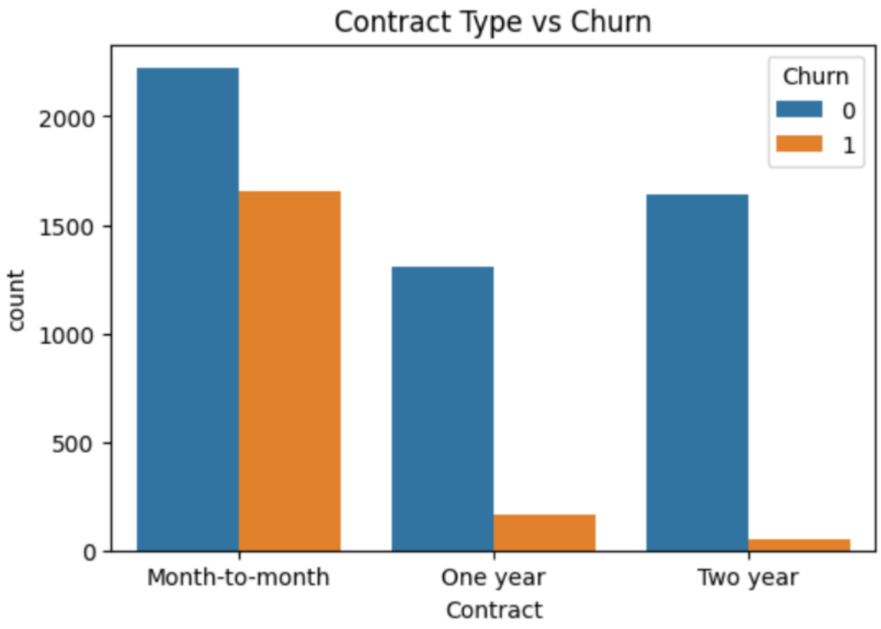


Contract Type and Churn Risk

Contract type emerged as one of the strongest churn indicators:

Contract Type	% of Customers	Churn Trend
Month-to-month	61.6%	Highest churn rate
One-year	20.0%	Moderate churn
Two-year	18.5%	Lowest churn (long-term commitment)

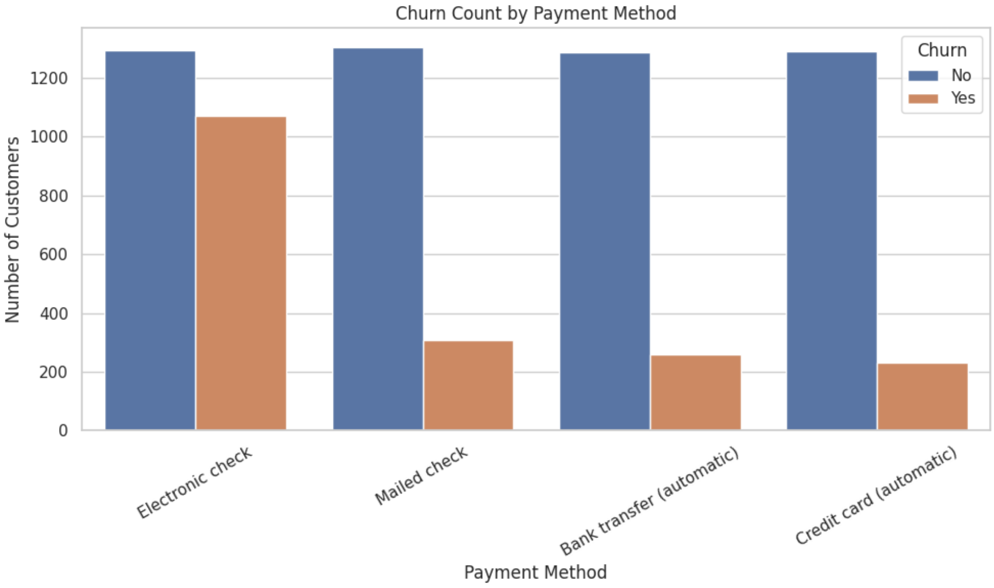
Customers with flexible monthly contracts were significantly more likely to leave.



Payment Method Patterns

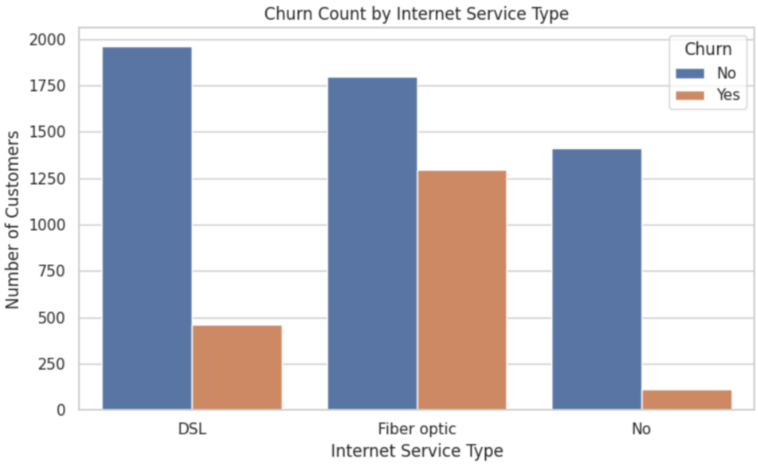
Electronic check is the most common payment method (41.7%) and shows the highest churn association.

Auto-payment methods (bank transfer or credit card) correlate with lower churn rates, suggesting convenience reduces cancellation likelihood.



Service Usage Insights

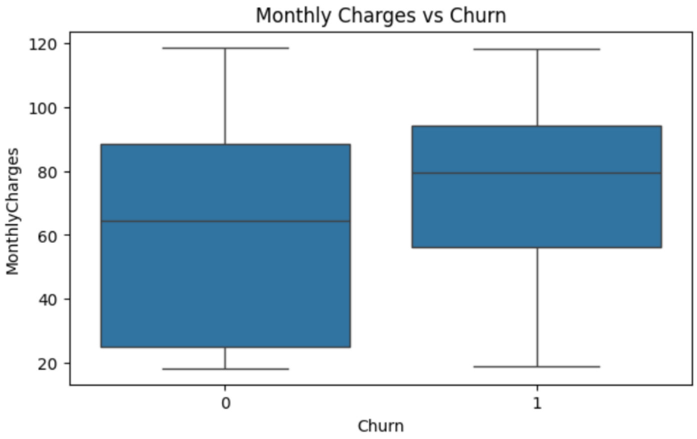
Fiber optic users show higher churn rates than DSL users, potentially due to higher pricing, lower satisfaction, or service instability. It could also be due to DSL users having more tenure as it is an older service.



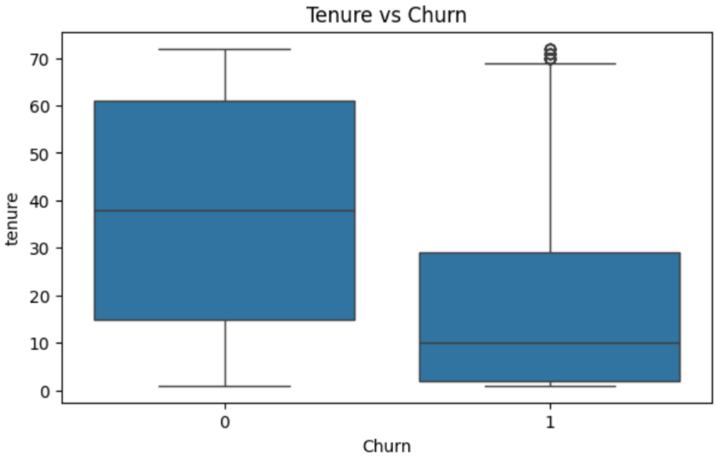
Customers lacking additional services such as TechSupport or OnlineBackup were more likely to churn — suggesting perceived value impacts loyalty.

Behavior of Numeric Features

Customers who churn tend to have higher monthly charges compared to non-churned customers. The median monthly cost for churned customers is noticeably higher, suggesting that customers on more expensive plans may be more dissatisfied or price-sensitive



Customer tenure shows two concentration peaks. 0–10 months shows high churn risk, likely reflecting trial dissatisfaction. 50–70 months indicates they are stable loyal customers.



The sharp dip in mid-tenure groups suggests many customers churn early rather than over time.

The exploratory analysis revealed clear patterns in customer churn: short-tenure customers, month-to-month contracts, electronic check users, and customers with higher monthly charges were more likely to churn. However, churn behavior cannot be fully understood using individual variables alone. The relationships among features are complex, and some predictors interact (e.g., internet service and contract type).

Therefore, machine learning models were applied to quantify feature importance, predict churn probability, and evaluate how well the patterns identified in EDA can be used to support proactive retention strategies.

Based on the patterns identified in the exploratory analysis—particularly the imbalance in the churn variable, the influence of contract type, and the variation in numeric features such as monthly charges—our next step was to develop predictive models to quantify these relationships and forecast churn. The following section explains the preprocessing and modeling framework used.

4. ANALYSIS METHODOLOGY

4.1 Data Preparation and Preprocessing

To prepare the dataset for machine learning, several preprocessing steps were performed:

Encoding and Transformation:

First, The variable TotalCharges was stored as text, so it was converted to numeric format. Rows with missing or invalid values were removed. Next, Binary categorical variables (e.g., *Partner, Dependents, PhoneService, PaperlessBilling*) were mapped from *Yes/No* to `1/0`. Finally, for multi-category features (e.g., *Gender, Contract, InternetService, PaymentMethod*),

one-hot encoding was applied using `drop_first=True` to avoid multicollinearity.
After encoding, the final feature matrix contained 30 predictor variables (expanded from the original 21).

Train-Test Split:

The data was split into 80% training (5,625 records) and 20% testing (1,407 records) using stratified sampling to preserve the original churn distribution (73.4% non-churn, 26.6% churn). A fixed `random_state=42` ensured reproducibility.

Handling Class Imbalance (SMOTE):

Because the dataset was imbalanced, we applied SMOTE (Synthetic Minority Oversampling Technique) only to the training set to generate synthetic churn cases. After oversampling, the training data became balanced (50% churn, 50% non-churn). This step ensures the model learns to recognize churners instead of predicting mostly "no churn."

Feature Scaling:

A StandardScaler was applied to normalize numeric variables. Scaling was necessary for models such as Logistic Regression, where coefficient interpretation and convergence depend on comparable feature scales.

These steps ensured the dataset was properly prepared for modeling, addressed the class imbalance, and followed best practices recommended during feedback.

4.2 Modeling Techniques

To predict customer churn, we applied two machine learning classification algorithms. Each model was tested before and after SMOTE oversampling to compare performance under imbalanced vs. balanced training conditions.

Logistic Regression:

Logistic Regression was selected as our first model because it is:

- Interpretable — coefficients indicate how each feature increases or decreases churn likelihood.
- Efficient for binary classification problems.

- Useful for probability-based scoring, which supports customer risk prioritization.

To prepare the data for Logistic Regression, we applied:

- StandardScaler: Logistic regression is sensitive to feature scale, and scaling ensures coefficients are comparable.
- Pipeline implementation, ensuring scaling and SMOTE occur only on the training folds during cross-validation (avoids data leakage).

After training the baseline model on the imbalanced dataset, we observed that the model performed well overall but showed lower recall for churners (class 1), meaning it missed many customers who actually churned.

To address this limitation, we applied SMOTE (Synthetic Minority Oversampling Technique) to rebalance the churn and non-churn classes in the training data.
The goal was to improve:

- Recall: catching more true churners
- Precision: reducing incorrect churn predictions

Finally, we performed 5-fold cross-validation and hyperparameter tuning (C values) to select the best-performing model.

Decision Tree Classifier

The Decision Tree was used as a contrasting non-linear model. Its advantages include:

- Ability to capture non-linear patterns
- Transparent and business-friendly interpretation
- No assumption about data distribution

To prevent overfitting, we tuned:

- max_depth (tree depth)
- min_samples_split (minimum required samples for node splitting)

As with Logistic Regression, the Decision Tree was trained:

1. Before SMOTE (imbalanced scenario)
2. After SMOTE + scaling + cross-validation (balanced scenario)

This allowed us to compare not only two different algorithm types, but also their behavior under different class distributions.

4.3 Model Evaluation Metrics

We assessed model performance using:

To evaluate model performance, we used several metrics rather than relying on accuracy alone. This is especially important because the dataset is imbalanced, with far more non-churn customers than churn customers. The following evaluation metrics were used:

- Accuracy:
  The percentage of correct predictions made by the model. Since a naïve model predicting all customers as "Not Churn" already achieves 73.4% accuracy, our models need to significantly exceed this benchmark to be considered useful.

- Precision (for churn class):
  Precision tells us, out of all customers predicted as churn, how many actually churned.
  This matters because telecom companies may invest retention incentives (discounts, offers, support) into these customers — meaning poor precision leads to wasted resources spent on customers who would not have left.

- Recall (for churn class):
  Recall measures how many actual churners the model successfully identifies.
  This is crucial because missing churners (false negatives) means the business loses customers without intervention. For this reason, recall is a priority metric for this project.

- ROC–AUC (Receiver Operating Characteristic – Area Under Curve):
  AUC measures how well the model separates churn vs. non-churn across different probability thresholds.
  It is more reliable than accuracy when data is imbalanced and helps assess overall discrimination ability.

- Confusion Matrix**:**
  Shows the number of true positives, true negatives, false positives, and false negatives.
  This helps us understand the kinds of errors each model makes and whether the model

is failing more by misidentifying churners or non-churners.

Model / Approach	**Description**	**Accuray**
Naïve Baseline	Predicts all customers as "Not Churn"	**73.4%**
Logistic Regression (no scaling, no SMOTE)	First basic model after encoding	**≈80%**

- 5-Fold Cross-Validation:
  Used to assess the stability and generalizability of the models. By training and testing across multiple folds, we reduce the risk of overfitting and ensure performance is not dependent on a single train-test split.

5. RESULTS AND BENCHMARKING

5.1 Baseline Benchmarks (Before Modeling)

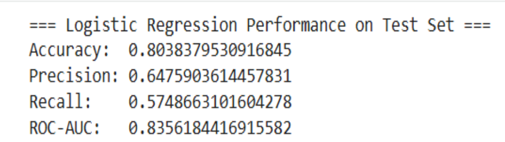
This establishes that the model must perform better than **73.4%** to be useful.

5.2 Model Results *Before SMOTE* (Original dataset)

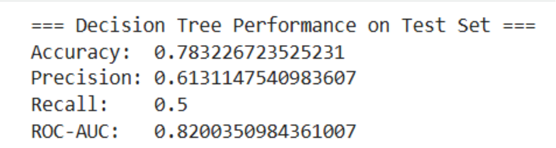
Metric	**Logistic Regression**	**Decision Tree**
Accuracy	80.38%	78.32%

Precision (Churn=1)	64.76%	61.31%
Recall (Churn=1)	57.49%	50.00%
ROC-AUC	0.8356	0.8200
Cross-Val Accuracy	80.50% (±0.57)	78.48% (±1.16)

Before SMOTE, both models performed well and were significantly better than the naive baseline. Logistic Regression slightly outperformed Decision Tree in accuracy and stability. However, recall values for both models were relatively low, meaning a large portion of churners were still being missed.



```
=== Logistic Regression Performance on Test Set ===
Accuracy:  0.8038379530916845
Precision: 0.6475903614457831
Recall:    0.5748663101604278
ROC-AUC:   0.8356184416915582
```



```
=== Decision Tree Performance on Test Set ===
Accuracy:  0.783226723525231
Precision: 0.6131147540983607
Recall:    0.5
ROC-AUC:   0.8200350984361007
```

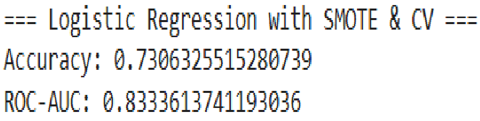
5.3 Model Results After SMOTE (Balanced Dataset)

After establishing the baseline and running the initial models on the imbalanced dataset, we next explored whether model performance would improve once the class imbalance was addressed using SMOTE**.**

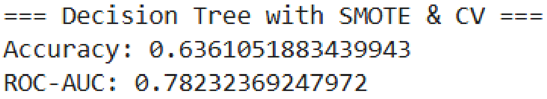
Metric	Logistic Regression + SMOTE	Decision Tree + SMOTE
Accuracy	73.06%	63.61%
Precision (Churn = 1)	0.50	0.41

Recall (Churn = 1)	0.78	0.87
F1 Score (Churn = 1)	0.61	0.56
ROC-AUC	0.8334	0.7823

After applying SMOTE to rebalance the dataset, we observed a notable change in model behavior. Accuracy decreased for both models (from approximately 80% to 73% for Logistic Regression and from 78% to 64% for the Decision Tree), which is expected since the models can no longer rely on predicting the majority class. However, both models showed a substantial improvement in detecting churn cases. Logistic Regression improved recall from roughly 50% to 78%, meaning it now recognizes a much larger portion of actual churners. The Decision Tree showed an even stronger recall increase (55% to 87%), although this came with lower precision, indicating more false-positive churn predictions. Despite the accuracy decrease, ROC-AUC values remained strong—particularly for Logistic Regression (0.833)—showing that the model still effectively distinguishes between churn and non-churn customers. Overall, SMOTE significantly enhanced the models' ability to detect churners, aligning better with the business objective of customer retention.



```
=== Logistic Regression with SMOTE & CV ===
Accuracy: 0.7306325515280739
ROC-AUC: 0.8333613741193036
```



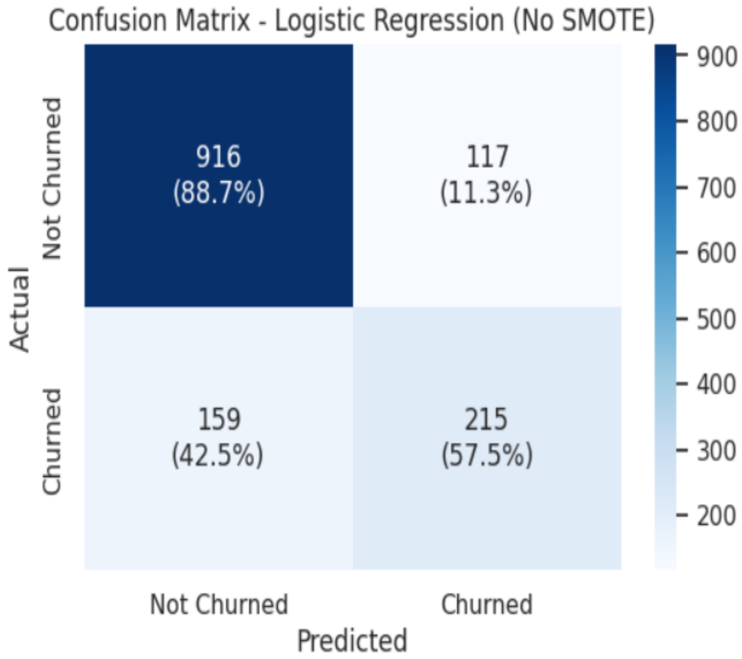
```
=== Decision Tree with SMOTE & CV ===
Accuracy: 0.6361051883439943
ROC-AUC: 0.78232369247972
```

SMOTE improved the models' ability to detect churners, especially in recall, which aligns with the business goal of identifying customers at risk. Although accuracy decreased, the trade-off is justified because correctly identifying churners is more valuable than predicting the majority class.

5.4 Logistic Regression Performance (before SMOTE)

Confusion Matrix:

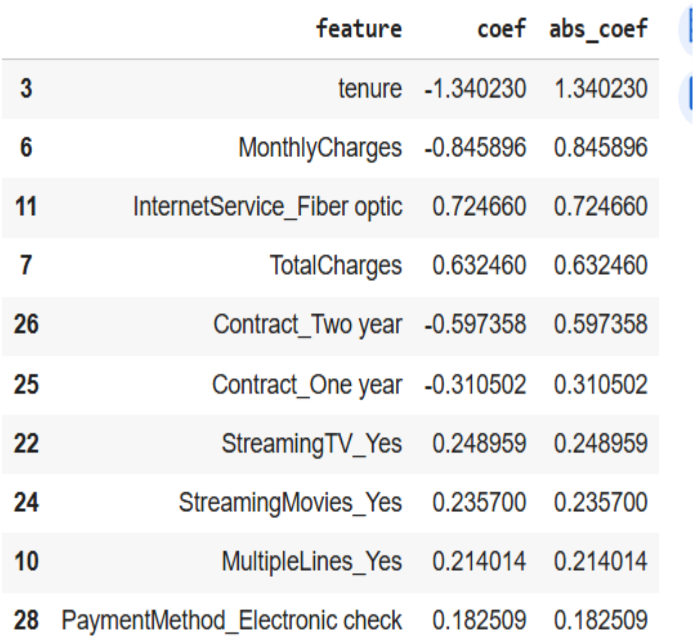
- True Negatives: 916 | False Positives: 117
- False Negatives: 159 | True Positives: 215



The confusion matrix shows how well the model classified churn cases. The model correctly identified 916 customers who stayed (True Negatives) and 215 customers who churned (True Positives). However, it also produced 117 false positives, meaning those customers were incorrectly flagged as churners, which could lead to unnecessary retention costs. More importantly, there were 159 false negatives, representing customers who actually churned but were not detected by the model — this is critical because these missed cases represent lost revenue opportunities. Therefore, while the model performs reasonably well, improving recall for churners remains important.

Top 10 Most Important Features (by coefficient magnitude):

1. Tenure: -1.340 (strong negative: longer tenure → lower churn probability)
2. MonthlyCharges: -0.846 (lower charges → lower churn)
3. InternetServiceFiber optic: +0.725 (fiber optic → higher churn risk)
4. TotalCharges: +0.632
5. ContractTwo year: -0.597 (two-year contracts → lower churn)
6. ContractOne year: -0.311
7. StreamingTVYes: +0.249
8. StreamingMoviesYes: +0.236
9. MultipleLinesYes: +0.214
10. PaymentMethodElectronic check: +0.183



	feature	coef	abs_coef
3	tenure	-1.340230	1.340230
6	MonthlyCharges	-0.845896	0.845896
11	InternetService_Fiber optic	0.724660	0.724660
7	TotalCharges	0.632460	0.632460
26	Contract_Two year	-0.597358	0.597358
25	Contract_One year	-0.310502	0.310502
22	StreamingTV_Yes	0.248959	0.248959
24	StreamingMovies_Yes	0.235700	0.235700
10	MultipleLines_Yes	0.214014	0.214014
28	PaymentMethod_Electronic check	0.182509	0.182509

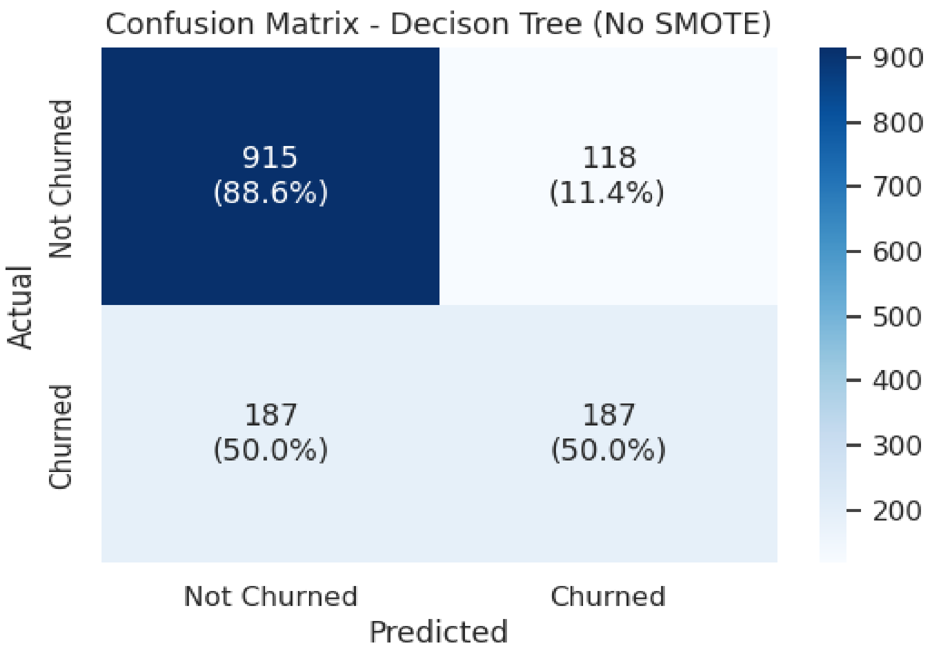
The logistic regression coefficient analysis reveals several key drivers of customer churn. The strongest factor reducing churn is tenure, meaning that the longer a customer has been with the company, the less likely they are to leave. Contract length also plays an important retention role — customers with one-year or two-year contracts are significantly less likely to churn compared to those on flexible month-to-month plans. In contrast, certain services and billing behaviors are associated with higher churn risk. Customers using fiber optic internet, paying through electronic check, or subscribing to add-on services such as streaming TV, streaming movies, or multiple phone lines showed an increased likelihood of leaving. The model also identified TotalCharges as a positive churn predictor, suggesting that customers with high cumulative bills may become dissatisfied, especially early in their lifecycle. Although the coefficient for MonthlyCharges was negative, this likely reflects multicollinearity with contract type and tenure, as exploratory analysis showed churned customers generally pay higher monthly fees. Overall, the model indicates that churn is influenced by cost, contract commitment, digital payment behavior, and perceived service value.

To complement the linear structure of Logistic Regression, a Decision Tree model was evaluated to capture potential non-linear patterns and feature interactions.

5.5 Decision Tree Performance (Before SMOTE)

Confusion Matrix:

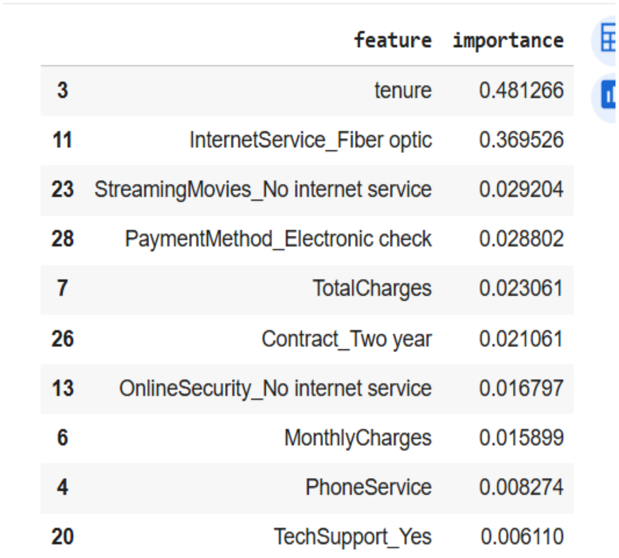
- True Negatives: 915 | False Positives: 118
- False Negatives: 187 | True Positives: 187



The confusion matrix for the Decision Tree model shows how well it distinguishes churners from non-churners. The model correctly identified 915 customers who did not churn (True Negatives) and 187 customers who churned (True Positives). However, it also produced 118 False Positives, meaning these customers were incorrectly predicted as churners, which could lead to unnecessary retention discounts or interventions. More critically, the model resulted in 187 False Negatives, representing actual churners that the model failed to detect. These missed cases carry business risk because they represent preventable customer losses. Compared to the logistic regression model, the Decision Tree tends to make slightly more mistakes in both categories, indicating room for improvement—especially in reducing false negatives, since identifying at-risk customers early is a key strategic goal for churn reduction.

Top 10 Most Important Features (by importance score):

1. Tenure: 0.481 (dominant predictor)
2. InternetServiceFiber optic: 0.370
3. StreamingMoviesNo internet service: 0.029
4. PaymentMethodElectronic check: 0.029
5. TotalCharges: 0.023
6. ContractTwo year: 0.021
7. OnlineSecurityNo internet service: 0.017
8. MonthlyCharges: 0.016
9. PhoneService: 0.008
10. TechSupportYes: 0.006



	feature	importance
3	tenure	0.481266
11	InternetService_Fiber optic	0.369526
23	StreamingMovies_No internet service	0.029204
28	PaymentMethod_Electronic check	0.028802
7	TotalCharges	0.023061
26	Contract_Two year	0.021061
13	OnlineSecurity_No internet service	0.016797
6	MonthlyCharges	0.015899
4	PhoneService	0.008274
20	TechSupport_Yes	0.006110

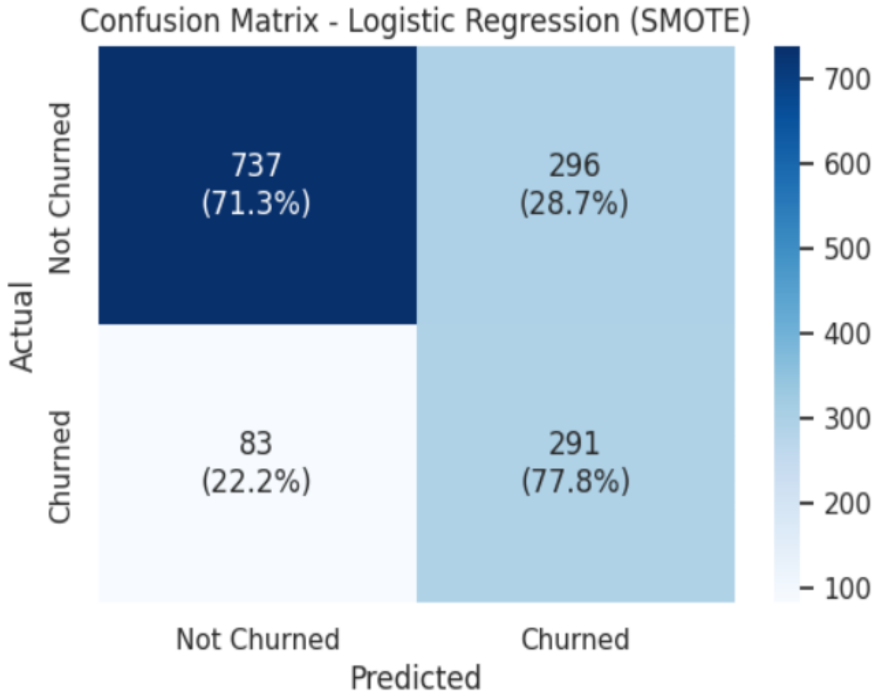
The feature importance results from the Decision Tree model provide insight into which variables most strongly influence churn predictions. Tenure emerged as the most important predictor, indicating that customers with shorter service duration are much more likely to churn compared to long-term users. The second strongest factor was fiber optic internet service, suggesting that customers subscribed to fiber optic plans are associated with higher churn risk, likely due to cost, performance expectations, or competitive alternatives. Other behavioral and billing factors, such as whether customers pay through electronic check, whether they subscribe to streaming services, and whether they have additional internet service-dependent features, also contributed to churn prediction, though with smaller weight. Contract-related features, including two-year contracts, showed a stabilizing effect, reinforcing that customers with longer commitments tend to remain with the company. Finally, customer support and service-related variables, such as TechSupport and PhoneService, contributed minimally, suggesting that these features alone do not strongly influence churn decisions. Overall, the Decision Tree results indicate that customer loyalty is strongly driven by tenure, internet service type, payment behavior, and subscription patterns.

Overall, while both models identified tenure, contract type, and service behavior as key churn drivers, the Logistic Regression model emphasized billing patterns and customer commitment (such as contract length and monthly charges), whereas the Decision Tree placed stronger weight on internet service type and tenure, showing churn is influenced not only by cost and duration but also by the type of services customers subscribe to.

While the initial model results provided valuable insights, they revealed a key limitation, both models had difficulty identifying churners due to the imbalance in the dataset. To overcome this, we applied SMOTE, generating synthetic churn examples and allowing the models to learn from a more balanced dataset. The following section compares performance before and after applying SMOTE.

5.6 Logistic Regression After SMOTE (Balanced Dataset)

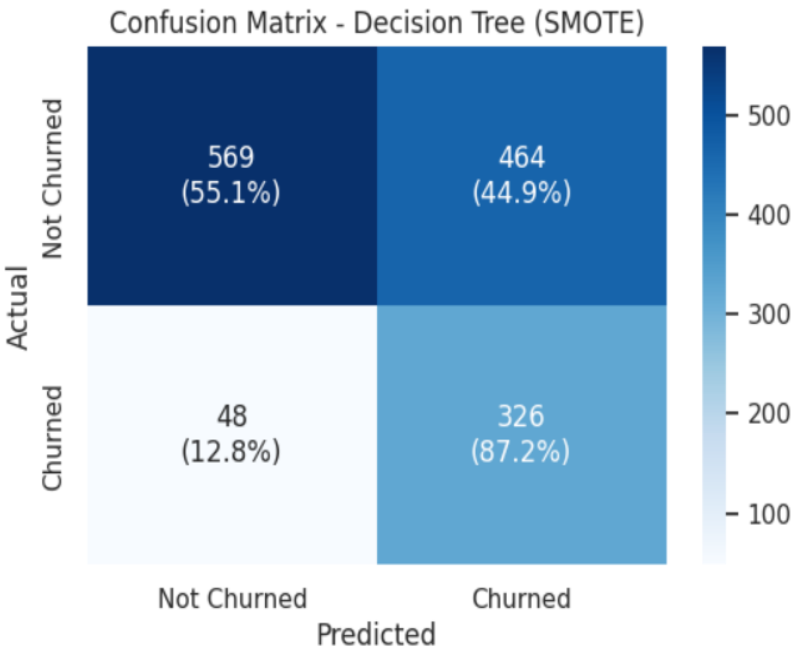
- True Negatives: 737 | False Positives: 296
- False Negatives: 83 | True Positives: 291



The confusion matrix shows how the Logistic Regression model performed after applying SMOTE. The model correctly predicted 737 customers who did not churn (True Negatives) and 291 customers who actually churned (True Positives). However, there were 296 false positives, meaning the model incorrectly flagged these customers as churners even though they stayed — which could result in unnecessary retention offers. More importantly, the false negatives were reduced to just 83, meaning far fewer actual churners were missed compared to the original model without SMOTE. This improvement highlights that SMOTE helped the model become more effective in identifying true churners, even though it increased false alarms.

5.7 Decision Tree Performance After SMOTE

- True Negatives: 569 | False Positives: 464
- False Negatives: 48 | True Positives: 326



After applying SMOTE and retraining the Decision Tree model, the confusion matrix shows a noticeable shift in prediction behavior. The model correctly identified 569 non-churners (True Negatives), but it also incorrectly flagged 464 customers who were actually non-churners as churn risks (False Positives). On the positive side, the model was successful in identifying actual churners: it correctly predicted 326 churn cases (True Positives) and missed only 48 churners (False Negatives), which indicates a strong improvement in recall compared to the model trained without SMOTE. However, this improvement comes at the cost of a higher number of false alarms, suggesting that while the model became much better at detecting churn, it sacrificed precision and may lead to unnecessary retention outreach efforts.

Table: Summary of Model Performance Before vs After SMOTE

Model	Metric Focus	BEFORE SMOTE	AFTER SMOTE	Interpretation
Logistic Regression	Recall (Churn=1)	**57.49%**	**78% ↑**	Model detects more churners

Logistic Regression	Accuracy	**80.38%**	**73.06% ↓**	Trade-off due to class balance
Decision Tree	Recall (Churn=1)	**50%**	**87% ↑↑**	Excellent detection of churners
Decision Tree	Accuracy	**78.32%**	**63.61% ↓**	High false positives

Before SMOTE, Logistic Regression performed slightly better than the Decision Tree in terms of accuracy and overall model stability, with both models achieving strong ROC-AUC scores (≈0.82–0.84), indicating good separation between churners and non-churners. Both approaches consistently identified tenure, contract type, and fiber-optic internet service as key churn drivers.

After applying SMOTE to rebalance the dataset, accuracy decreased for both models, which is expected when improving class balance. However, recall and precision for churners improved significantly—especially in the Decision Tree—making both models more effective at identifying customers at risk of leaving. While this shift increases false positives, it enhances the model's ability to detect churn, which is more valuable for retention strategy in a business context.

6. INSIGHTS AND BUSINESS IMPLICATIONS

The analysis revealed several important behavioral and structural patterns influencing customer churn. Most notably, customer tenure emerged as the strongest predictor across all models, showing that churn is most likely within the first year of service. This suggests a "leaky bucket" scenario where new customers represent the highest risk group, while long-term customers show strong retention behavior. Contract type also demonstrated significant influence on churn outcomes—customers on month-to-month contracts displayed substantially higher churn rates compared to those with one-year or two-year contracts, reinforcing the stabilizing effect of long-term agreements. Service type further played an important role: fiber optic subscribers exhibited higher churn likelihood, which may indicate pricing dissatisfaction, unmet performance expectations, or stronger competitor offerings in areas where fiber is available. Payment behavior also contributed meaningful signals—customers paying through electronic checks were more likely to churn, potentially indicating lower engagement, payment friction, or a more cost-sensitive customer segment.

6.1 Actionable Recommendations for Management

Based on the findings, several targeted retention strategies are recommended to reduce churn and improve long-term customer loyalty. First, customers on month-to-month contracts represent the highest-risk segment, and converting even a portion of these users to one-year or two-year agreements could significantly reduce churn by an estimated 15–20%. This strategy should be implemented early in the customer lifecycle—ideally within the first 30–45 days—through incentives such as discounts or upgraded service bundles. Second, the company should implement proactive support programs for high-risk groups, including fiber optic subscribers, electronic check payers, and customers with less than twelve months of tenure. A predictive early-warning system integrated into the CRM platform and monitored weekly would allow the business to contact these customers before disengagement leads to cancellation.

Additionally, improving the onboarding experience may have a major impact on retention, especially since many customers churn early. Structured check-ins, satisfaction surveys, and guided service support during the first six to twelve months could reduce early churn rates from over 40% to approximately 25–30%. Investigating fiber optic churn drivers is also important; a focused analysis may reveal whether pricing, performance, customer support, or competitive offerings are contributing specifically to dissatisfaction among this group.

Beyond these core actions, additional strategic improvements could further strengthen retention. Dynamic pricing strategies and personalized retention offers may help prevent cancellations among cost-sensitive customers. Since churn rates are noticeably higher among those using electronic check payments, understanding whether this reflects financial constraints or lower engagement could support more tailored communication and billing options. Competitive benchmarking—especially in fiber-enabled markets—should be conducted regularly to ensure pricing and service performance remain competitive. Finally, because customer behavior and competitive dynamics evolve over time, the churn prediction model should be retrained quarterly or after major business changes to ensure it remains accurate and actionable.

---

7. LIMITATIONS AND FUTURE DIRECTIONS

7.1 Project Limitations & Future Directions

While the models demonstrated strong predictive capability, several limitations must be acknowledged. First, the dataset lacked behavioral and contextual variables such as customer support interactions, service usage patterns, or complaint history—features that are frequently strong churn predictors in real-world telecommunications environments. Additionally, the data represented a single static snapshot rather than a time-dependent record, limiting the ability to assess changes in customer behavior or seasonal trends over time. Another key constraint was the absence of customer lifetime value (CLV), which means the model predicts *who* may churn, but not *how costly* losing each customer would be—an important factor in prioritizing retention resources.

The modeling approach was also impacted by class imbalance, which required the use of SMOTE to improve the model's ability to detect churners. While oversampling increased recall, it introduced trade-offs in precision and overall accuracy, illustrating the tension between identifying churners and avoiding unnecessary retention spending. Additionally, only two relatively simple algorithms—Logistic Regression and Decision Tree—were implemented due to project scope and time constraints. More advanced techniques such as Random Forest, Gradient Boosting, XGBoost, or neural network–based architectures may be better suited to capturing nonlinear patterns and improving generalization.

Finally, the dataset reflects a single telecom provider in one geographic region, meaning external competitive, economic, or regulatory influences were not modeled. Future work should explore the incorporation of behavioral and temporal features, experiment with ensemble and deep learning models, and integrate CLV-driven decision frameworks. Retraining the model periodically as new customer data becomes available would further support long-term accuracy and operational usefulness.

8. CONCLUSION

This project successfully developed predictive models to identify customers at risk of churning within the telecommunications industry. Logistic Regression emerged as the best-performing model before resampling, demonstrating an accuracy of approximately 80% and a strong AUC score, indicating reliable predictive separation between churners and non-churners. Both Logistic Regression and the Decision Tree model consistently identified the same core drivers of churn: tenure, contract length, and service type—particularly fiber internet subscriptions.

After applying SMOTE to address class imbalance, the models were better able to detect true churners, especially reflected in the significant improvement in recall, though this came with expected decreases in accuracy. This trade-off aligns with business priorities, as correctly identifying at-risk customers is more valuable than optimizing overall accuracy when designing retention strategies.

The findings suggest that focused interventions—including converting flexible contract customers to longer commitments, improving early lifecycle experience, and addressing performance or pricing concerns in fiber segments—may yield substantial reductions in churn. Overall, the project demonstrates the value of predictive modeling in guiding data-driven retention strategy, ultimately supporting improved customer loyalty and long-term revenue stability.