Telco Customer Analysis

Feras Abdel-Rahim
Jacob Nava

INSY 5339 Project

Executive Summary

Customer churn analysis report provides the in-depth analysis of customer behavior that determines the customer behavior patterns and assists the predictive model development to support strategic decision-making. Using this report, telecom operators will be able to control churn in advance, improve retention, and lower business expenses. Customer data exploratory data analysis using visualization tools is employed on identifying key trends, and a classification model is implemented to forecast the risk of churn. Additionally, the report examines how different prediction thresholds impact business outcomes and offers customer retention recommendations based on evidence.

Project Motivation

The prime objective of this project is to reduce customer churn and reduce business cost for telecommunication companies. Retaining existing customers is less costly than acquiring new customers, and early churn forecasting allows proactive customer retention. Reducing churn can allow companies to reduce costly re-acquisition initiatives and optimize marketing budget utilization. The project further aims at leveraging customer data for strategic decision-making. By transforming raw information into actionable intelligence, the analysis facilitates the development of a data culture in telecommunication operations. Not only does it equip decision-makers with greater understanding of the behavior of their customers, but it also aims to improve customer experience. Having knowledge on which customers are at-risk and why they are at-risk means that they can be intervened with specifically targeted contracts or services offered that ultimately drive customer satisfaction and loyalty.

Data Description

Source**:** Publicly available Telco Customer Churn dataset (Kaggle)
Total Records**:** 7,043 customer entries
Target Variable**:** Churn – whether a customer left the company (Yes/No)
Total Churn Distribution:

- Did not Churn: 73.46%
- Churn: 26.54%

The dataset includes the following variables:

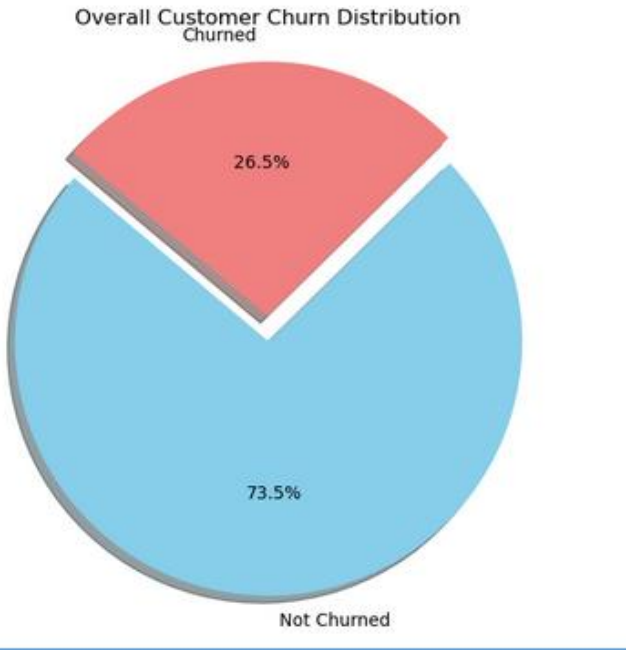
- **customerID**: Unique customer identifier
- **gender**: Gender of the customer (Male/Female)
- **SeniorCitizen**: Whether the customer is a senior citizen (0 = No, 1 = Yes)
- **Partner**: Whether the customer has a partner (Yes/No)
- **Dependents**: Whether the customer has dependents (Yes/No)
- **tenure**: Number of months the customer has stayed with the company
- **PhoneService**: Whether the customer has phone service (Yes/No)
- **MultipleLines**: Whether the customer has multiple phone lines
- **InternetService**: Type of internet service (DSL, Fiber optic, None)
- **OnlineSecurity**: Whether the customer has online security add-on
- **OnlineBackup**: Whether the customer has online backup add-on
- **DeviceProtection**: Whether the customer has device protection
- **TechSupport**: Whether the customer has tech support
- **StreamingTV**: Whether the customer has streaming TV
- **StreamingMovies**: Whether the customer has streaming movies
- **Contract**: Contract term (Month-to-month, One year, Two year)
- **PaperlessBilling**: Whether the customer uses paperless billing (Yes/No)
- **PaymentMethod**: Payment method used (Electronic check, Mailed check, Bank transfer, Credit card)
- **MonthlyCharges**: Monthly bill amount
- **TotalCharges**: Total amount charged to the customer
- **Churn**: Whether the customer left the company (Yes/No) — this is the target variable

Data Type Breakdown:

- Binary: 7 features
- Categorical: 11 features
- Continuous: 3 features

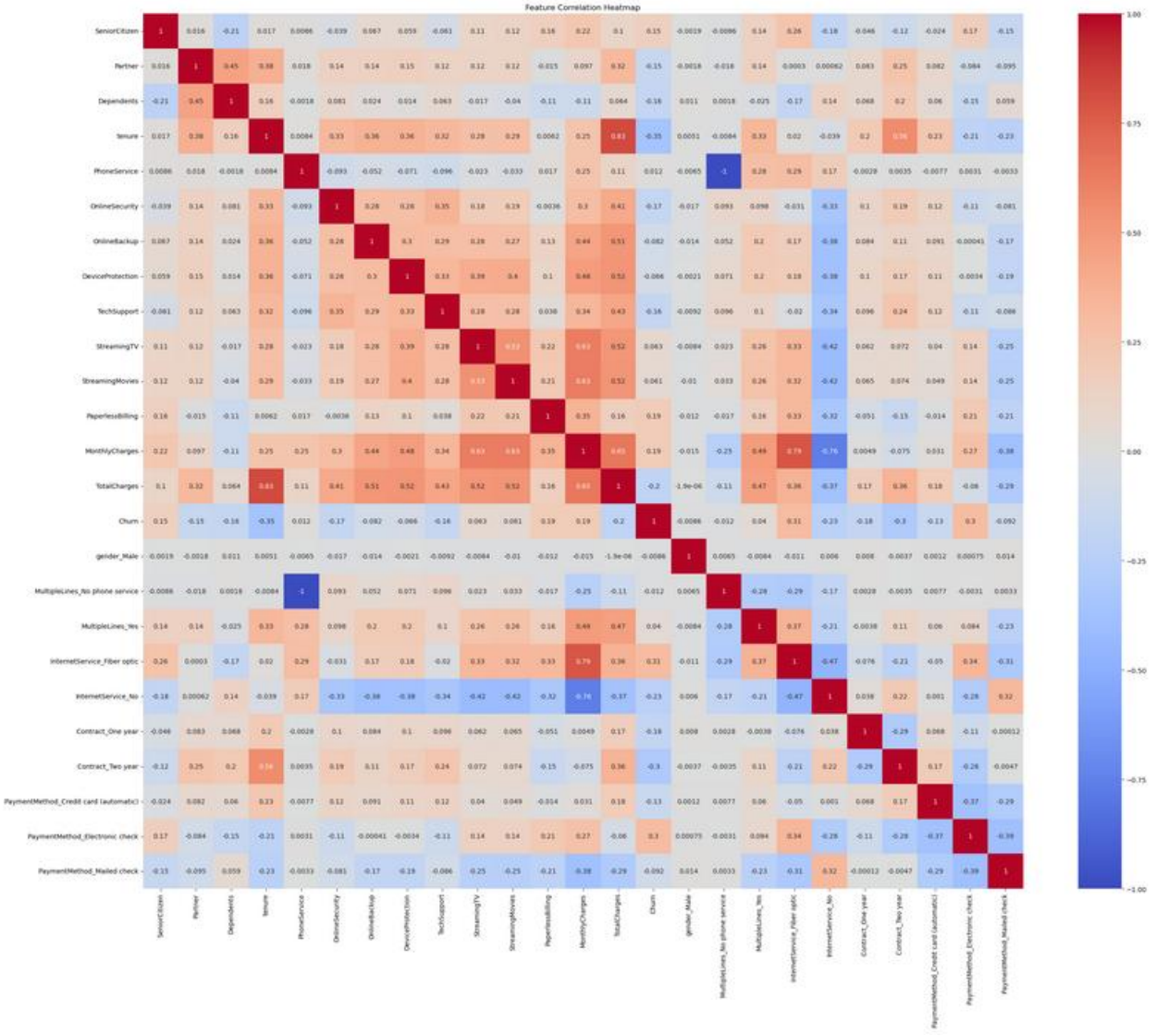
Data Analysis Visualization and Findings

1. Overall Churn Rate Distribution
    - A pie chart of churn vs non-churn shows that approximately 26.5% of customers have churned, while 73.5% have not
    - This imbalance highlights the importance of targeting the minority churn group without over-predicting churn (which leads to high false positives)
    - Understanding the baseline churn rate is essential for evaluating any classification model's usefulness



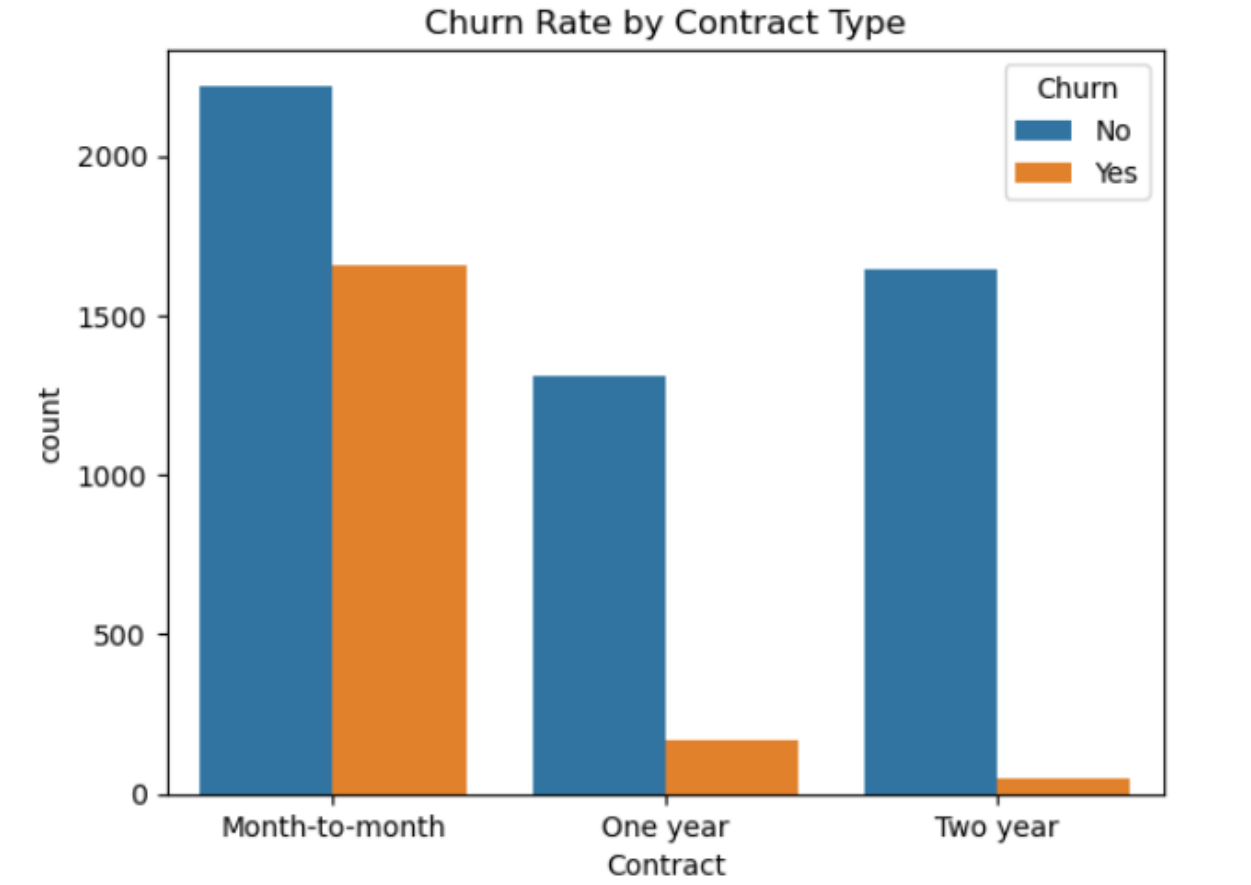
2. Correlation heatmap
    - The correlation heatmap visualizes relationships between numeric features
    - Positive correlations:

    - Internet service fiber optic & payment method fiber optic have the strongest positive correlation
- Negative correlations:
    - Tenure & Contract 2 year have the strongest negative correlation

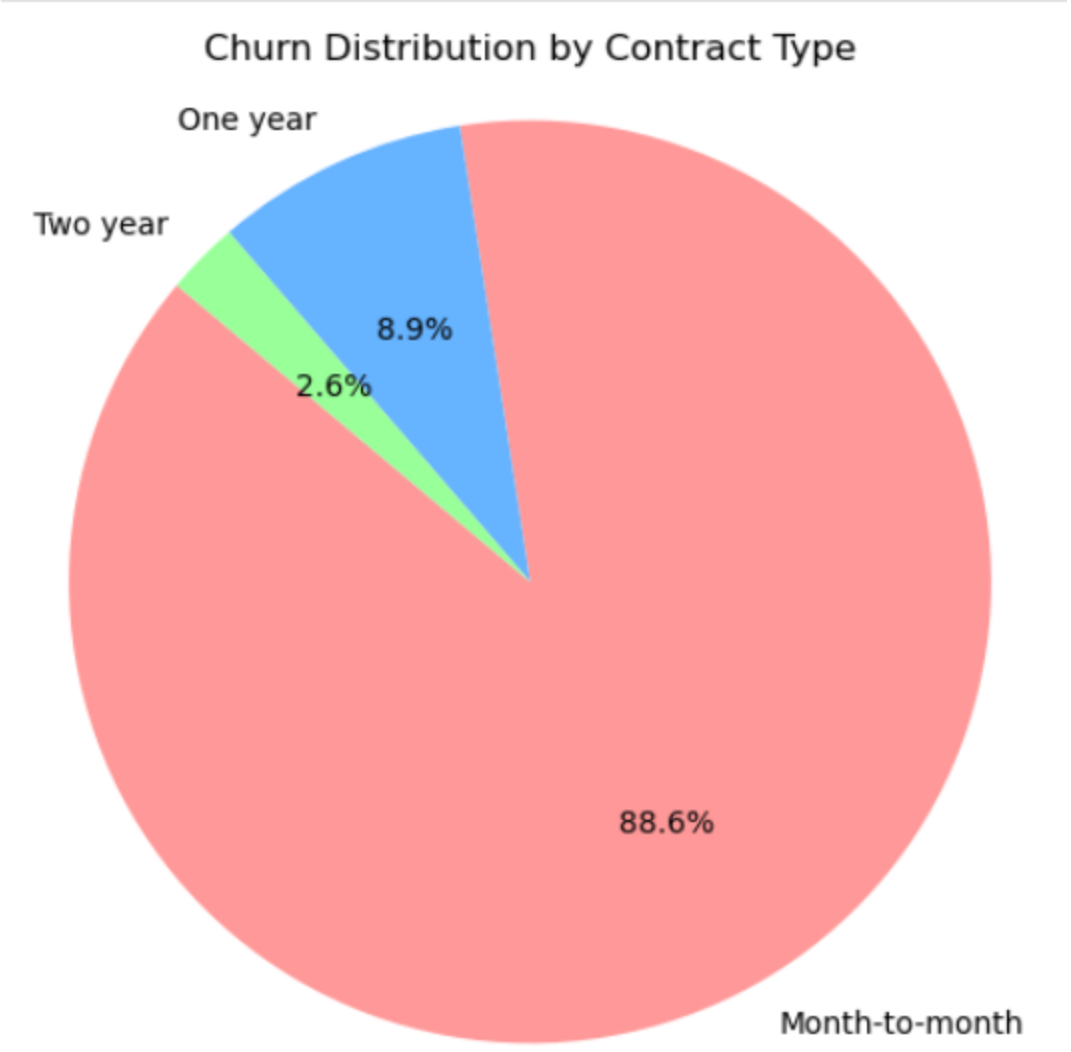


3. <u>Churn Rate by Contract Type (Bar Chart)</u>
    - Customers with Month-to-Month contracts have the highest churn rate, with a significantly larger churn population than One Year or Two Year contract types
    - Longer contract commitments correlate with customer retention

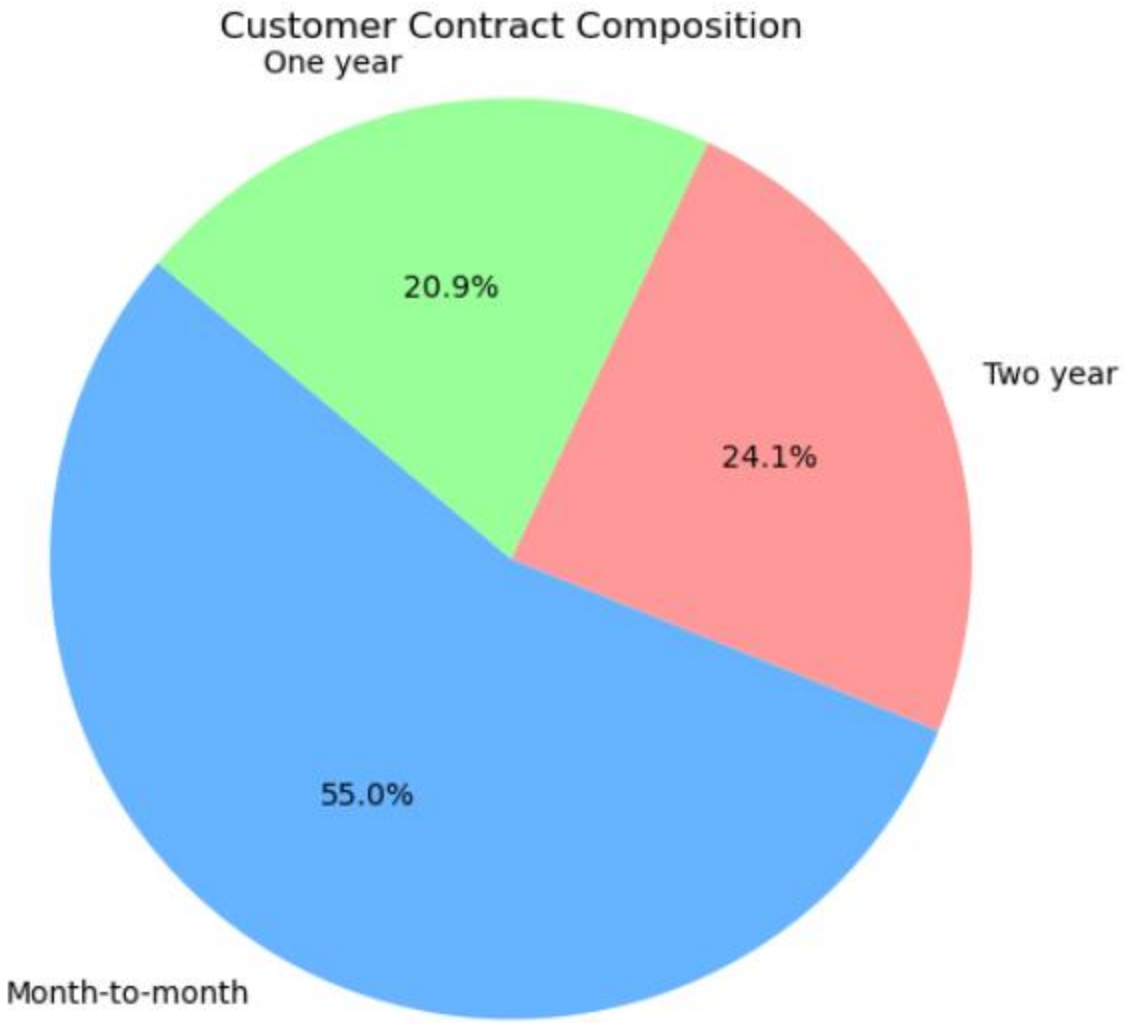
- Business Insight: Month-to-month contracts could benefit from targeted retention offers or loyalty incentives



4. <u>Churn Distribution by Contract Type (Pie Chart)</u>
    - Of all churned customers:
        - 88.6% are on Month-to-Month plans
        - Only 8.9% on One Year, and 2.6% on Two Year
    - Month-to-month users make up the bulk of churners, even though they represent just over half the customer base
    - Business Insight: Focusing retention efforts on this group yields the greatest impact



5. Customer Contract Composition (Pie Chart)
    - Shows the overall distribution of contract types among all customers:
        - 55% are on Month-to-Month contracts
        - 21% One Year
        - 24% Two Year
    - Despite Month-to-Month being the majority, their churn contribution is disproportionately high
    - Business Insight: Opportunity exists to convert Month-to-Month customers into longer contracts through bundled pricing or loyalty programs

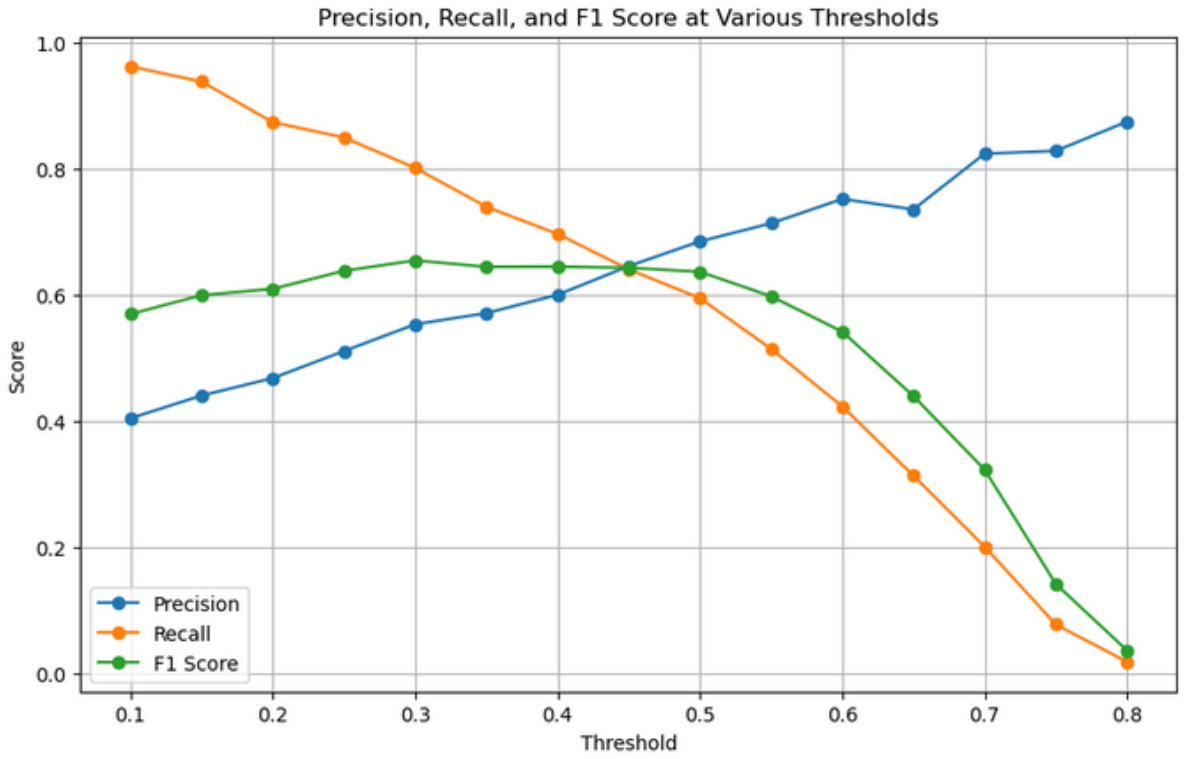


6. Summary

The visualizations confirm that churn is heavily concentrated among customers on Month-to-Month contracts, which comprise the majority of the user base. Contract length and tenure are strong predictors of churn, and the imbalance in the churn distribution makes model evaluation more nuanced. These findings guided the model design and informed the threshold selection process to minimize false negatives and target high-risk customers effectively.

Predictive Modeling and Findings

1. Log Regression Threshold Tuning and Business Cost Trade-Offs
    - This line chart compares how precision, recall, and F1 score vary across different classification thresholds
    - At low thresholds (e.g., 0.1–0.3), the model captures more churners (high recall) but also makes many false positive predictions (low precision)
    - As the threshold increases, precision improves while recall declines, indicating a trade-off between over-flagging and under-catching churners



- To determine the most effective prediction threshold, we compared model performance at different cutoff points. The key focus was minimizing business cost from two types of prediction errors:
    - False Positive (FP): Predicting a customer will churn when they won't → unnecessary retention offer
        - Estimated cost per FP: $50
    - False Negative (FN): Failing to identify a true churner → lost customer and revenue

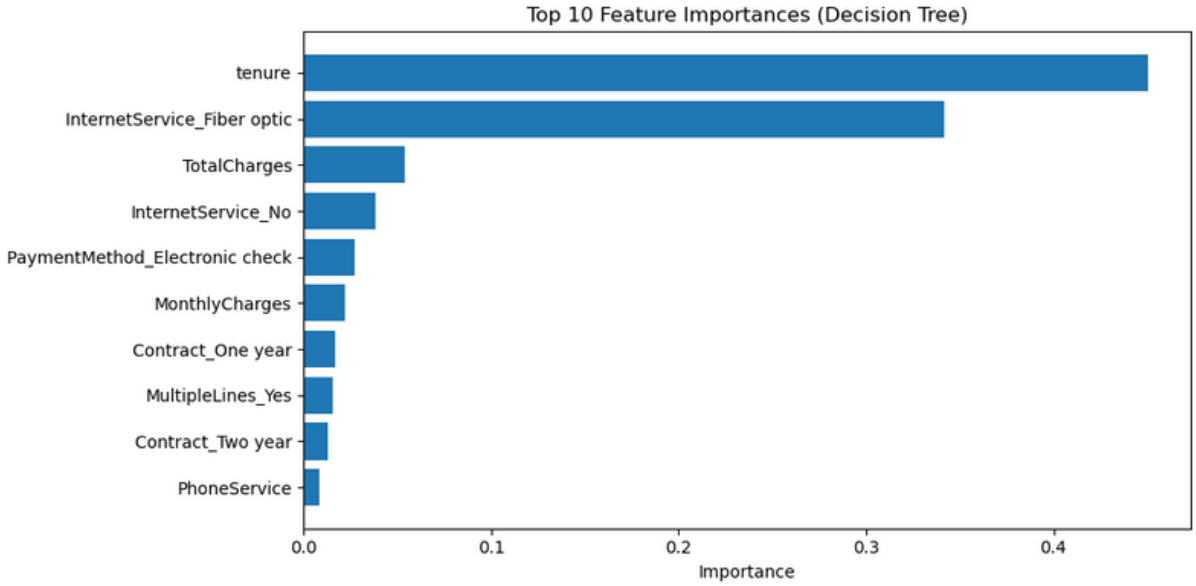
- Estimated cost per FN: $500

Threshold	False Positives (FP)	False Negatives (FN)	Total Cost Calculation	Total Cost
0.3	296	84	(296 × $50) + (84 × $500) = $14,800 + $42,000	**$56,800**
0.4	177	111	(177 × $50) + (111 × $500) = $8,850 + $55,500	**$64,350**

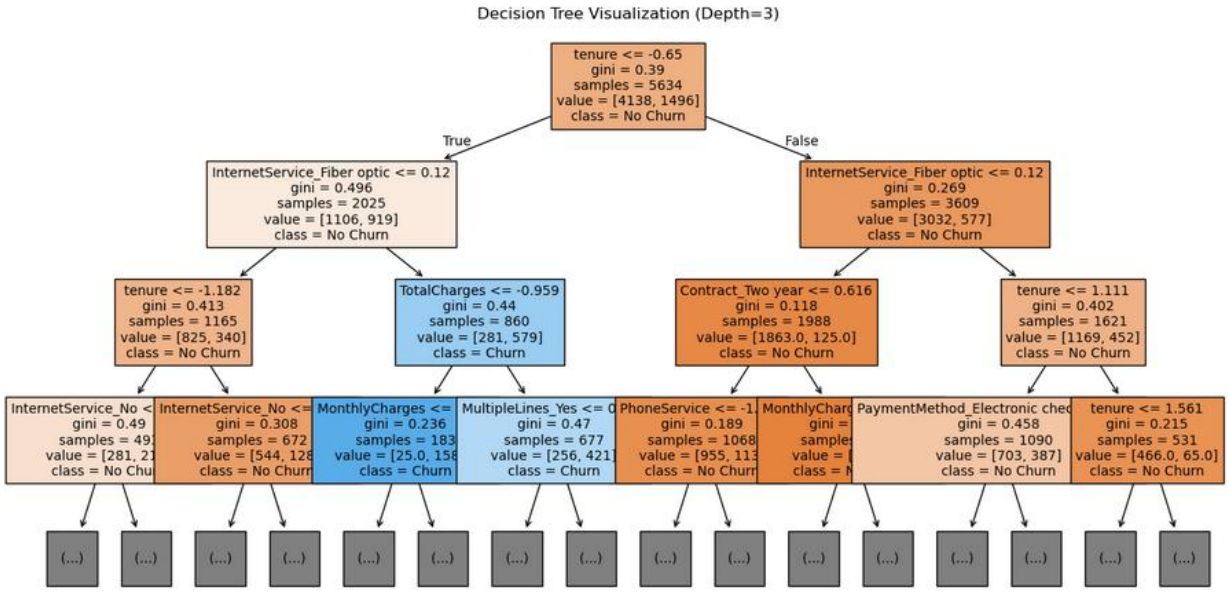
- Threshold 0.3 is more cost-effective overall, even though it increases the number of false positives slightly
- It reduces high-cost false negatives (missed churners), which likely leads to better retention and business savings in the long run

2. Decision Tree Performance at Threshold = 0.4
   - Confusion Matrix Summary:
     - True Negatives (TN): 859
     - False Positives (FP): 177
     - False Negatives (FN): 111
     - True Positives (TP): 262
   - Metrics:
     - Accuracy: 80%
     - Precision (churn): 0.60
     - Recall (churn): 0.70
     - F1 Score (churn): 0.65
   - Top Features:
     1. Tenure
     2. InternetService_Fiber optic
     3. TotalCharges
     4. Electronic Check
     5. Contract One/Two Year

```
📊 Decision Tree (Threshold=0.4)
Confusion Matrix:
 [[859 177]
 [111 262]]

Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.83      0.86      1036
           1       0.60      0.70      0.65       373

    accuracy                           0.80      1409
   macro avg       0.74      0.77      0.75      1409
weighted avg       0.81      0.80      0.80      1409
```



3. <u>Churn Decision Path (Tree Visualization)</u>
    - This tree shows how feature splits affect churn classification:
        - Low tenure and fiber optic service are early indicators of churn
        - Two-year contracts and consistent payments reduce churn likelihood
    - Business Insight: Can be used to design targeted rules for loyalty program triggers



4. Cross-Validation Results (Threshold = 0.4)
    - 10-Fold Average:
        - Accuracy: 0.789 ± 0.017
        - Precision: 0.618 ± 0.035
        - Recall: 0.609 ± 0.037
    - Model performs consistently across folds, suggesting good generalization

```
🔁 10-Fold Cross-Validation with Threshold = 0.4
Fold 1: Accuracy=0.807, Precision=0.636, Recall=0.636
Fold 2: Accuracy=0.804, Precision=0.637, Recall=0.610
Fold 3: Accuracy=0.787, Precision=0.585, Recall=0.679
Fold 4: Accuracy=0.793, Precision=0.608, Recall=0.608
Fold 5: Accuracy=0.793, Precision=0.605, Recall=0.631
Fold 6: Accuracy=0.778, Precision=0.579, Recall=0.610
Fold 7: Accuracy=0.743, Precision=0.515, Recall=0.556
Fold 8: Accuracy=0.798, Precision=0.618, Recall=0.631
Fold 9: Accuracy=0.797, Precision=0.626, Recall=0.583
Fold 10: Accuracy=0.790, Precision=0.617, Recall=0.551

📈 Average 10-Fold Scores (Threshold=0.4):
Accuracy : 0.789 ± 0.017
Precision: 0.603 ± 0.035
Recall   : 0.609 ± 0.037
```

5. Summary

We built a Decision Tree model using scaled features and a maximum depth of 3 for interpretability, initially evaluating it at a 0.4 threshold. However, after cost analysis showed that false negatives ($500 cost) were significantly more expensive than false positives ($50 cost), we adjusted the decision threshold to 0.3 to increase recall and better capture potential churners. Although the tree structure remains unchanged, this threshold shift enhances the model's business impact by prioritizing churn detection. The tree reveals that short tenure, fiber optic internet service, and low total charges are strong indicators of churn, while longer-term customers on non-fiber plans, especially those with contract commitments, are more likely to stay. These patterns highlight clear segmentation opportunities: newer customers on high-speed plans with minimal billing history are at higher churn risk and should be prioritized for retention efforts.

Managerial Policy Implementations

1. Optimal Threshold (0.3): Yields higher sensitivity (recall), ensuring fewer lost customers at the cost of more retention actions.
2. Retention Campaign: Focus retention resources on customers predicted to churn at or above the 0.3 threshold.
3. Loyalty Program: Consider implementing a loyalty program targeting medium-tenure customers to improve customer lifetime value.
4. Contract Incentives: Promote longer-term contracts, especially for fiber-optic internet users, who showed higher churn risk.

Conclusion

This project successfully identified key drivers of customer churn in a telecom dataset and applied predictive modeling to guide retention strategies. Initially, a 0.4 classification threshold was used, but after conducting a cost analysis that weighted false negatives at $500 and false positives at $50, a 0.3 threshold was determined to minimize overall

business impact. Analysis revealed that customers on month-to-month contracts—particularly those using fiber optic internet—were most likely to churn, while those with longer tenure, higher total charges, and contract-based services were more likely to stay. Logistic regression provided probability-based classification, and the decision tree model (visualized at depth 3) revealed interpretable patterns such as low-tenure fiber optic users being at high risk of churn. These insights led to actionable recommendations, including loyalty programs for mid-tenure customers, converting month-to-month users to longer-term contracts, and designing targeted retention efforts for high-risk segments. This work demonstrates how cost-sensitive modeling and interpretable machine learning can drive data-informed strategies to reduce churn and improve customer lifetime value.