Customer Churn Prediction

A Machine Learning Approach to Reduce Revenue Loss

Presented by: Group 4
HAN, Qingying
PIAO, Zhuying
ZHENG, Shuyu
CHEN, Zhiying
XU, Chenjunxiu

Project Overview & Business Impact

Challenge

- Customer acquisition costs 5-25x more than retention in telecom
- Churn directly impacts Monthly Recurring Revenue and Customer Lifetime Value
- Need proactive identification of at-risk customers

Objective

- Predict customer churn using ML classification (binary: Churn=1, Stay=0)
- Achieve >75% accuracy with balanced precision-recall performance

Business Value

- Early identification of at-risk customers allows targeted offers and improved retention

Business Questions

1.Which customer segments exhibit the highest churn propensity across demographic and behavioral dimensions?

2.Which service features that correlate most strongly with customer retention patterns?

3.Optimization strategies for retention spending allocation based on individual customer churn probability assessments.

Data Understanding

Data Source: From Kaggle (https://www.kaggle.com/datasets/blastchar/telco-customer-churn)

Dataset Characteristics:

- Size: 7,043 customer records with 21 original features, after encoding is 30 features
- Target Variable: Churn (binary: Yes/No, binary: 1 = Yes, 0 = No)
- Feature Categories:
  - Demographic: Gender, SeniorCitizen, Partner, Dependents
  - Account Information: Tenure, Contract, PaperlessBilling, PaymentMethod
  - Services: PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies
  - Financial: MonthlyCharges, TotalCharges

Data Quality Assessment:

- Missing values identified in TotalCharges column (11 records with blank values)
- No duplicate records found
- Balanced target distribution: 26.5% churn rate (1,869 churned customers)
- Numeric features (e.g. 'tenure', 'MonthlyCharges') show appropriate ranges and were standardized
- Categorical features were encoded for modeling

Initial data insights

Segment	Churn Rate
Month-to-month contract	42.7%
Tenure < 12 months	47.4%
Electronic check payment	45.3%
Senior citizens	41.7%

- **Insight: Short-tenure, monthly-payment, and digitally disengaged users churn more often.**

Data Preprocessing

- Cleaning: Handled missing TotalCharges (11 entries)
- Encoding: One-hot for categorical; Label for target
- Scaling: StandardScaler for numerical
- Split: 80/20 train-test with random_state=42

Models and Rationale

Model	Rationale
Logistic Regression	Fast, interpretable, and provides clear feature coefficients as a baseline
Decision Tree	Handles non-linear relationships; easy to visualize and explain
Random Forest	Ensemble method that improves stability and accuracy by combining multiple trees
SVM (RBF Kernel)	Performs best in detecting churners; effective with high-dimensional, complex patterns

Model Performance Comparison

Model	Accuracy	Precision (Churn)	Recall (Churn)	F1-Score (Churn)	ROC-AUC
Logistic Regression	0.79	0.62	0.52	0.56	**0.83**
Decision Tree	0.78	0.58	0.59	0.58	0.81
Random Forest	0.79	**0.64**	0.45	0.53	0.81
SVM	0.75	0.51	**0.79**	**0.62**	0.82

Key Insight:
SVM maximizes recall (79%), crucial for identifying customers most likely to churn and enabling early intervention.

Model Selection Summary

Models	Type	Top Features (Coef / Importance)	ROC-AUC	F1-Score (Class 1)	Comments
Logistic Regression	Linear	tenure, Contract_Two year, Contract_One year	0.8319	0.56	Most interpretable
Decision Tree	Non-linear	tenure, InternetService_Fiber, opticTotalCharges,	0.81	0.58	Visualizable tree
Random Forest	Non-linear	MonthlyCharges, tenure, InternetService_Fiber optic	0.81	0.53	Robust, good precision
SVM	Non-linear	Contract_One year, PaymentMethod_Electronic check, OnlineBackup_Yes	0.82	0.62	Best at recall

Key Insight:

SVM provides the best balance of recall and F1 for churn prediction, while Logistic Regression offers business-friendly interpretability and clear feature coefficients.

Strategic Recommendation

Primary Model: Support Vector Machine

- Deploy for production churn prediction
- Captures 79% of actual churning customers
- Minimizes revenue loss from missed at-risk customers

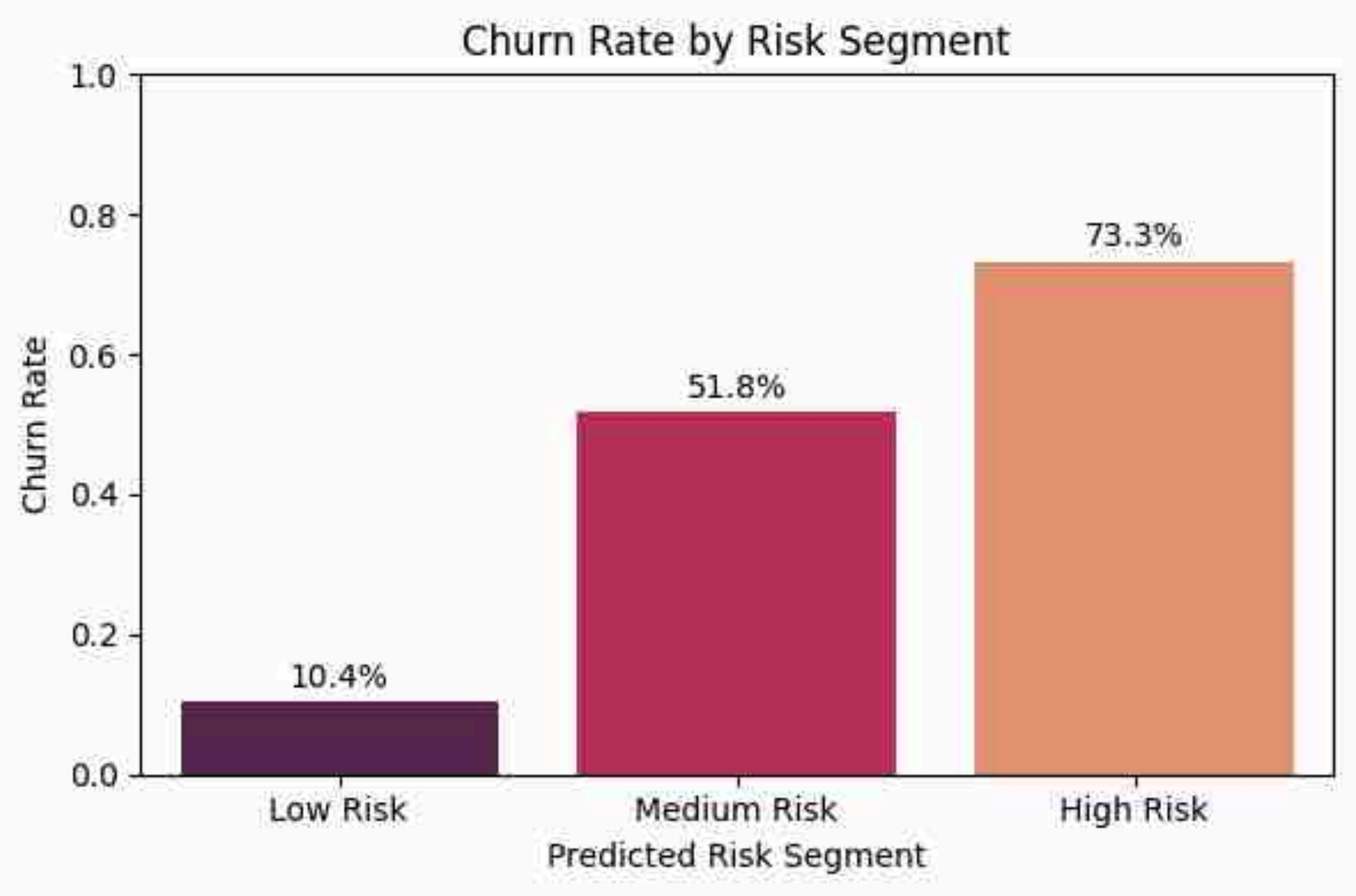
Secondary Model: Logistic Regression

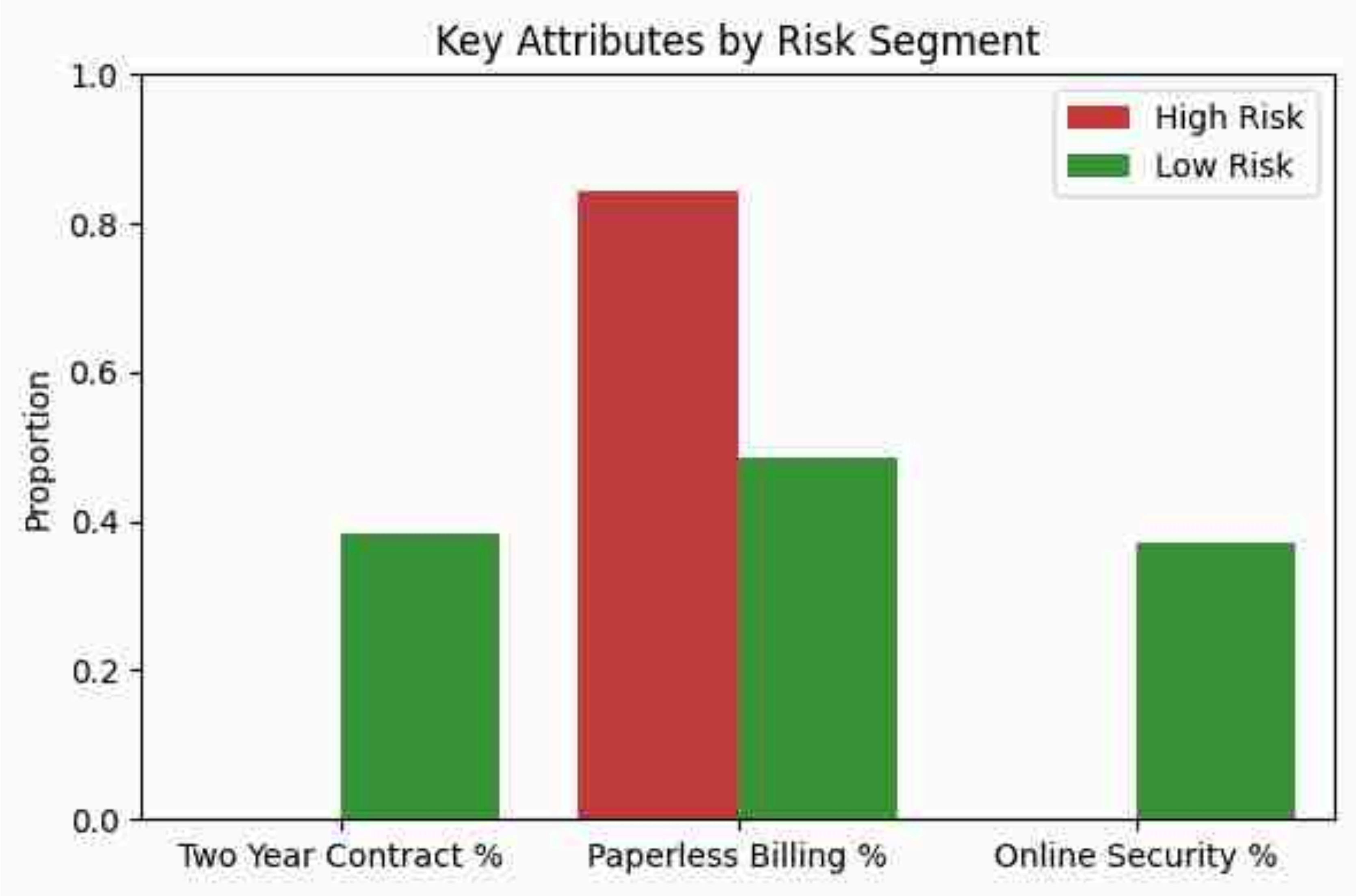
- Use for business analysis and stakeholder insights
- Most interpretable coefficients for strategy development
- Highest ROC-AUC (0.83) for overall discrimination

Implementation Impact

- Proactive retention campaigns targeting high-risk customers
- Reduced acquisition costs through improved retention
- Data-driven customer success interventions

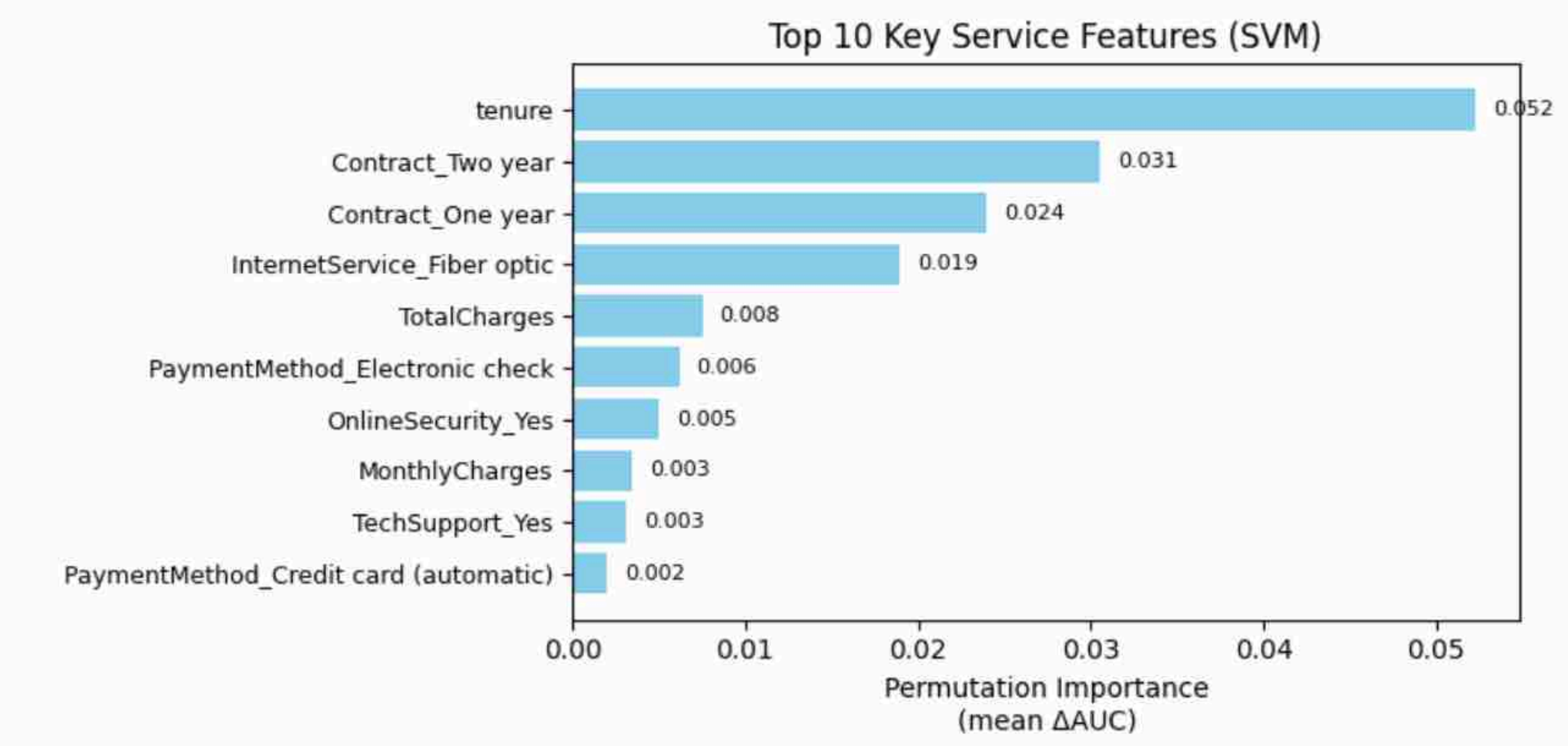
Q1. Who is Most Likely to Churn?





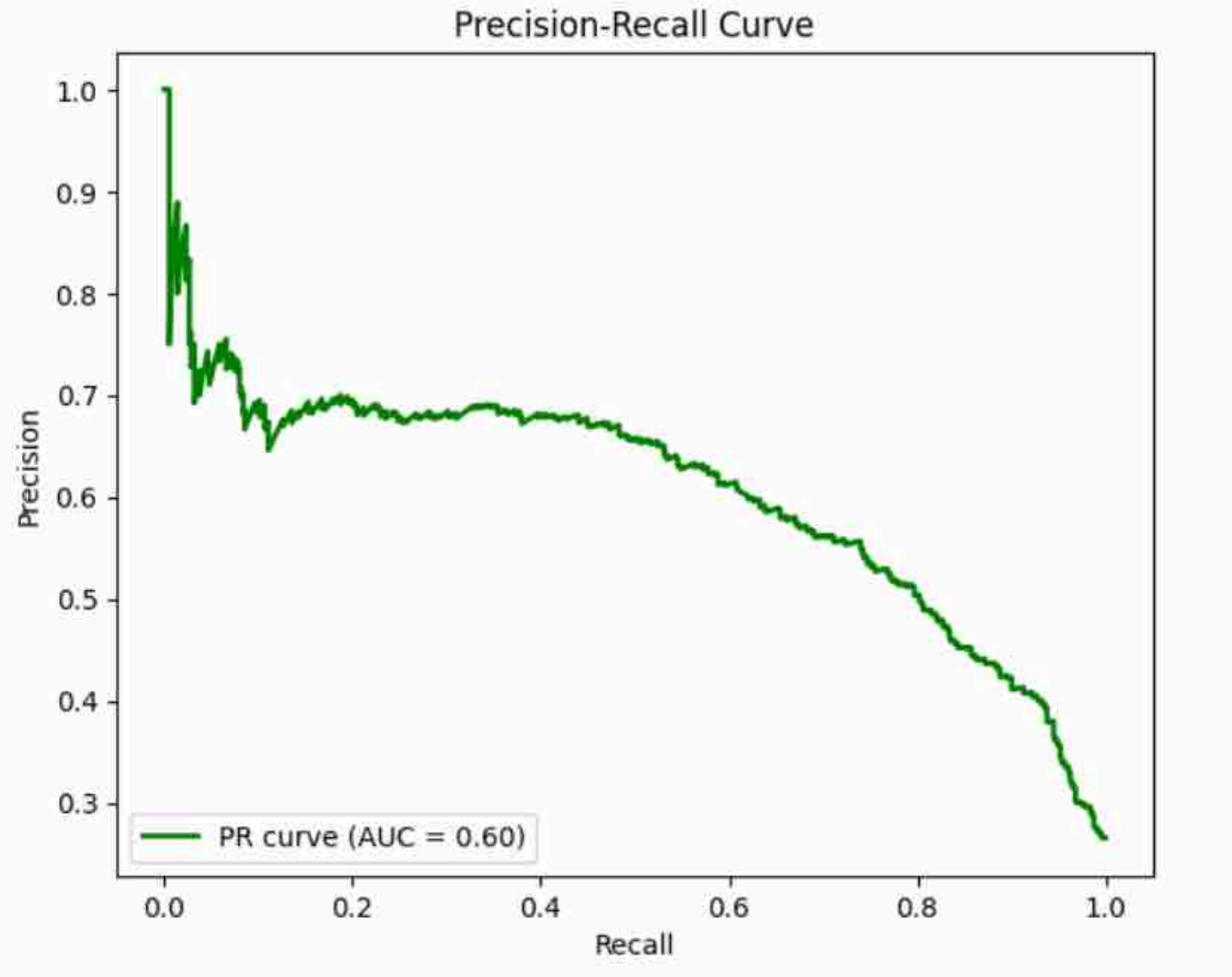
- Customers on shorter contracts and with no security bundle are far more likely to leave.
- Target retention offers should focus on high-risk segment.

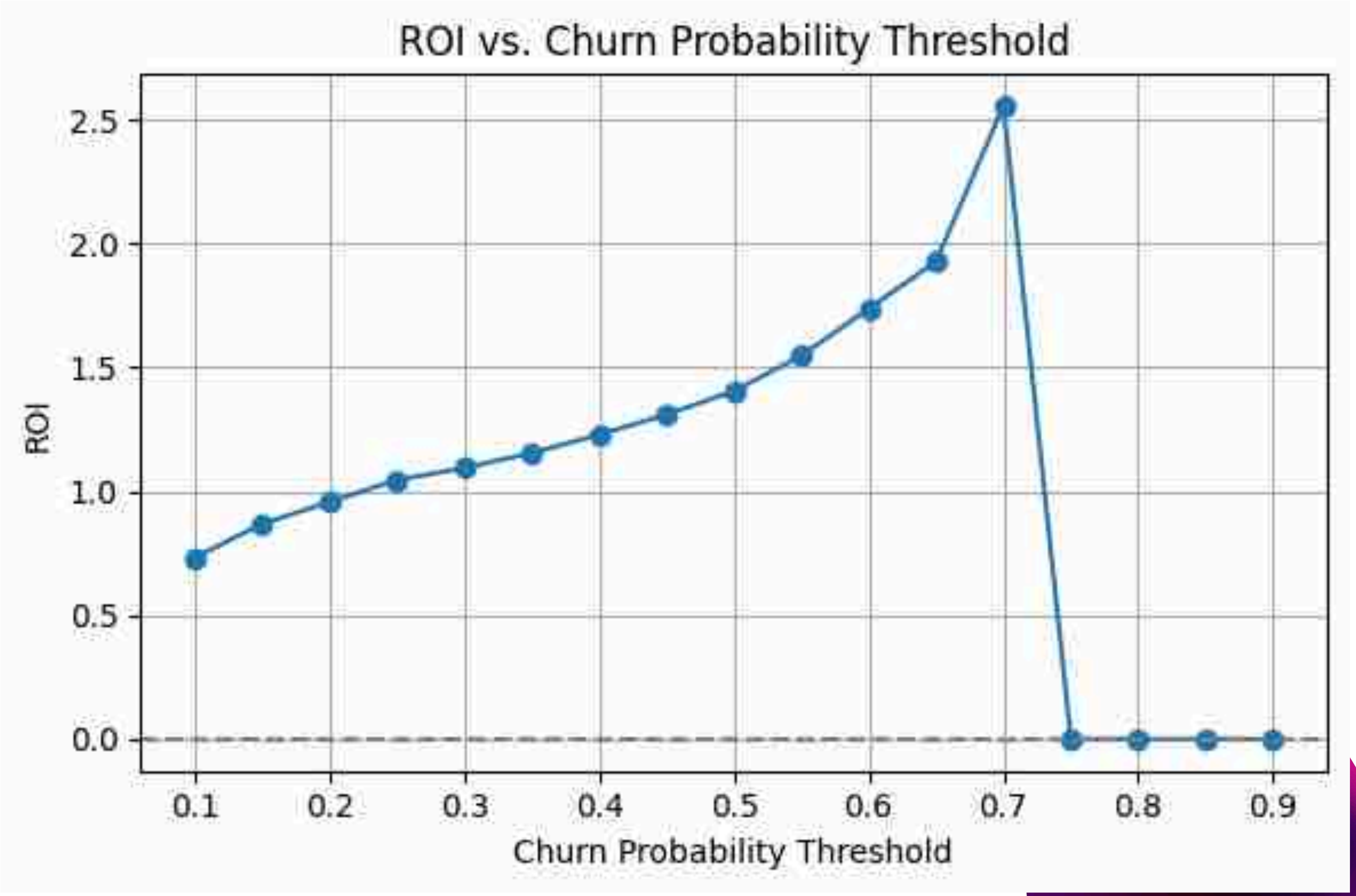
Q2. Which Features Drive Churn Most?



- Incentivize longer-term contracts
- Target fiber-optic customers with loyalty perks or competitively priced bundles
- Encourage automated payments
- Upsell security and support add-ons to at-risk segments to boost stickiness
- Tailor retention offers heavily toward new customers (low tenure), whose churn risk is by far the highest.

Q3. Strategy for optimizing retention spending based on churn probability





Business Recommendations

💡 Promote Long-Term Contracts

Offer price discounts or loyalty points to incentivize annual contracts
→ reduces churn linked to monthly contracts (key churn driver)

💳 Improve Electronic Payment Experience

Redesign payment portal UI/UX and add flexible options (e.g. PayPal, Apple Pay)
→ addresses friction in electronic check users (high churn subgroup)

📦 Offer Bundled Service Packages

Encourage multi-service adoption (e.g. Internet + Streaming)
→ combats churn among single-service customers

🎯 Target High-Risk Customers Early

Use churn probability scores to launch early retention offers
→ focus on customers with <12 months tenure, identified as most vulnerable

Ethical Considerations

Privacy & Data Protection

- Handle sensitive personal & financial data with care
- Comply with GDPR & CCPA
- Apply anonymization & secure storage practices

Algorithmic Fairness

- Monitor bias across demograph ic groups
- Conduct regular fairness audits
- Adjust models based on fairness metrics

Transparency

- Clearly explain how« churn scores are used
- Provide opt-out options for data-based targeting
- Document decisions from automated models

Discrimination Prevention

- Avoid targeting based on protected attributes (e.g. age, gender)
- Ensure equitable access to retention offers

Model Explainability

- Use interpretable models (e.gLogistic Regression)
- Ensure decisions are auditable & justified

NEXT STEP

- Deploy SVM in CRM for real-time churn scoring.
- Use scores to target high-risk users with personalized offers.
- Retrain models regularly with updated data.
- Run A/B tests to optimize retention campaigns.
- Ensure fairness and transparency in all prediction-based actions.

Conclusion

- This project demonstrated how machine learning can turn customer data into actionable retention strategies.

- By comparing four models, we found that SVM excels at detecting churners, while Logistic Regression offers valuable business insights.

- Our analysis revealed clear churn drivers and delivered data-backed recommendations to reduce customer loss and protect revenue.

Thank you for listening!