# Customer Churn Prediction: A Machine Learning Approach to Reduce Revenue Loss

**Prepared by Group 4:**

HAN, Qingying

PIAO, Zhuying

ZHENG, Shuyu

CHEN, Zhiying

XU, Chenjunxiu

# 1. Introduction and Objective

Customer churn represents one of the most critical challenges facing telecommunications companies, directly impacting revenue streams and long-term business sustainability. The goal of this project is to use machine learning to predict whether a customer will churn based on their usage, contract type, payment method, tenure, and other attributes.

This is a classic binary classification problem. The target variable, 'Churn', takes values of 'Yes' or 'No'. We reframed this as a binary target: 1 = Churned, 0 = Not Churned.Accurate predicting and early identification of at-risk customers allows for targeted retention campaigns, personalized offers, and improved customer service interventions, ultimately reducing churn rates and increasing profitability.

# 2. CRISP-DM (Business & Data Understanding)

## 2.1 Business Understanding

**Problem Context**: Customer acquisition costs in telecommunications are significantly higher than retention costs, with studies showing that acquiring a new customer costs 5-25 times more than retaining an existing one. Churn directly impacts Monthly Recurring Revenue (MRR) and Customer Lifetime Value (CLV).

**Stakeholders**: Our stakeholders include Customer Success teams, Marketing departments, Executive leadership, and Data Science teams that require actionable insights to implement retention strategies effectively.

**Success Criteria**: This research establishes four primary success benchmarks. First, models must achieve accuracy rates exceeding 75% while maintaining balanced precision and recall performance. Second, the analysis must identify key churn indicators that inform strategic business planning. Third, results must provide interpretable insights that facilitate executive decision-making. Fourth, the framework must enable proactive customer retention interventions through predictive modeling.

**Business Questions**: Three core research questions guide this investigation. The first examines which customer segments exhibit the highest churn propensity across demographic and behavioral dimensions. The second analyzes service features that correlate most strongly with customer retention patterns. The third explores optimization strategies for retention spending allocation based on individual customer churn probability assessments.

## 2.2 Data Understanding

**Data Source**: Telco Customer Churn Dataset 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

# 3. Dataset Description and Summary Statistics

**Summary Statistics**:

- **Tenure**: Mean = 32.4 months, Range = 0-72 months
- **Monthly Charges**: Mean = $64.76, Range = $18.25-$118.75
- **Total Charges**: Mean = $2,283.30, Range = $18.80-$8,684.80
- **Senior Citizens**: 16.2% of customer base
- **Contract Distribution**: Month-to-month (55.0%), One year (21.0%), Two year (24.0%)

**Key Observations**:

- Customers with month-to-month contracts show higher churn rates (42.7%)
- Senior citizens demonstrate elevated churn propensity (41.7%)
- Fiber optic internet users exhibit higher churn rates than DSL users
- Electronic check payment method correlates with increased churn risk

# 4. Data Preprocessing Steps

## 4.1 Data Cleaning

1. **Load and Inspect:** Loaded 7,043 rows × 21 columns; confirmed most features are categorical.
2. **Missing Value Treatment**: Found 11 blank/non-numeric entries in TotalCharges, coerced them to NaN.
3. **Data Type Conversion**: Ensured TotalCharges is now a float column.
4. **Redundant Column Removal**: Removed customerID, as it does not carry predictive information.

## 4.2 Feature Engineering

1. **Exploratory Plots:** Visualized churn distribution, monthly charges vs. churn, and correlation matrix
2. **Categorical Encoding**: Applied one-hot encoding to all categorical variables (including Yes/No fields).
3. **Feature Scaling**: Implemented StandardScaler for numerical features (tenure, MonthlyCharges, TotalCharges)
4. **Target Encoding**: Transformed Churn labels (Yes → 1, No → 0)

## 4.3 Data Splitting Strategy

- **Training Set**: 80% (5,634 samples)
- **Testing Set**: 20% (1,409 samples)
- **Random State**: Fixed seed (42) for reproducibility

# 5. Model Development and Evaluation

## 5.1 Model Selection Rationale

**Logistic Regression**: Chosen as baseline model for its interpretability and efficiency. Provides coefficient insights for business understanding and handles binary classification naturally.

**Decision Tree**: Selected for its ability to capture non-linear relationships and provide interpretable decision rules. Offers clear visualization of decision boundaries important for business stakeholders.

**Random Forest:**Included for its robustness and ensemble power. By aggregating the predictions of multiple decision trees, it reduces overfitting and improves generalization. It handles feature importance ranking well and performs reliably on moderately imbalanced datasets.

**Support Vector Machine (SVM)**: Implemented with RBF kernel to handle high-dimensional feature space effectively. Excellent performance on classification tasks with clear margin separation.

### 5.2 Hyperparameter Optimization

- **Logistic Regression**: Tuned the regularization parameter **C** to balance model complexity and avoid overfitting.
- **Decision Tree**: Adjusted **max depth**, **min samples split**, and **min samples leaf** to control tree size and prevent overfitting.
- **SVM: Optimized C and gamma using grid search to achieve better classification performance in high-dimensional space.**

### 5.3 Model Registration

All models were registered using jrjModelRegistry (PyPI) with version control and metadata tracking for reproducibility and deployment readiness.

### 5.4 Evaluation Metrics Justification

- **Accuracy**: Overall model performance measure
- **Precision**: Critical for retention campaign efficiency (minimizing false positives)
- **Recall**: Essential for capturing actual churn cases (minimizing false negatives)
- **F1-Score**: Balanced metric addressing class imbalance in churn prediction
- **ROC-AUC**: Comprehensive evaluation of classification performance across thresholds

## 6. Model Comparison and Results

| Model | Accuracy | Precision (Class 1) | Recall (Class 1) | F1-Score (Class 1) | ROC-AUC |
|---|---|---|---|---|---|
| Logistic Regression | 0.79 | 0.62 | 0.52 | 0.56 | 0.83 |
| Random Forest | 0.79 | 0.64 | 0.45 | 0.53 | 0.81 |
| Decision Tree | 0.78 | 0.58 | 0.59 | 0.58 | 0.81 |
| SVM | 0.75 | 0.51 | 0.79 | 0.62 | 0.82 |

**Model Performance Analysis**:

**Best Recall and F1-Score:** The Support Vector Machine (SVM) achieved the highest recall (0.79) and F1-score (0.62) for Class 1 (Churn), making it the most effective model at correctly identifying customers who are likely to leave — the key goal in churn prediction.

**Best ROC-AUC:** Logistic Regression showed the highest ROC-AUC (0.83), indicating strong overall discrimination between churners and non-churners, despite its lower recall.

**Business Interpretability:** Logistic Regression remains the most interpretable model, with clear coefficients that help identify which features contribute most to customer churn. This makes it particularly valuable for generating actionable business insights.

**Efficiency and Balance:** Decision Tree offers a solid balance between performance and interpretability, with decent recall (0.59) and F1 (0.58), and is typically fast to train, making it suitable for real-time or iterative applications.

### Recommendation

This analysis recommends SVM for production deployment given its superior recall and F1-score performance in capturing churning customers. Logistic Regression should be leveraged for business analysis and stakeholder communication due to its interpretability of churn factors. Decision Trees warrant consideration in resource-constrained or explainability-critical environments given their intuitive decision pathways.

**Model Feature 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 |

# 7. Insights and Ethical Reflections

## 7.1 Business Insights

- **Contract Type Impact**: Month-to-month contracts show 42.7% churn rate vs. 2.85% for two-year contracts
- **Payment Method**: Electronic check users demonstrate 45.3% churn rate, indicating payment friction
- **Service Bundling**: Customers with multiple services show lower churn propensity
- **Tenure Effect**: Customers with tenure < 12 months exhibit 47.4% churn rate

## 7.2 Actionable Recommendations

- **Contract Type Impact**: Launch personalized discounts or loyalty rewards to encourage customers to switch from month-to-month plans to longer-term contracts, improving retention and forecasting revenue more accurately.
- **Payment Experience Optimization**: Redesign the electronic payment flow to reduce friction, introduce seamless auto-pay options, and provide incentives for digital payment adoption—especially targeting electronic check users.
- **Service Bundling Strategies**:Recommend relevant service bundles based on customer usage patterns (e.g., streaming + internet) to increase engagement and reduce the risk of churn due to single-service dissatisfaction.
- **Onboarding & Early Retention Programs**: Deploy targeted welcome campaigns, tutorials, and check-ins within the first 3 months of a customer's journey—especially for users with tenure <12 months—to build stickiness early.

## 7.3 Ethical Considerations

- **Privacy and Data Protection**: Customer data contains sensitive personal and financial information requiring strict adherence to data protection regulations (GDPR, CCPA). Implement data anonymization and secure storage practices.
- **Algorithmic Bias**: Monitor for demographic bias in churn predictions to ensure fair treatment across customer segments. Regular bias audits and fairness metrics evaluation required.
- **Transparency**: Provide customers with clear communication about data usage for retention purposes and offer opt-out mechanisms.
- **Discrimination Prevention**: Ensure retention offers based on churn predictions don't discriminate against protected classes or create unfair pricing disparities.
- **Model Explainability**: Maintain interpretable models to ensure retention decisions can be justified and audited for fairness.

## 7.4 Implementation Considerations

- **Real-time Scoring**:Integrate the churn prediction model into the CRM system to score customers in real time, enabling timely interventions before disengagement occurs.
- **Performance Monitoring**:Set up dashboards and alerts to monitor model drift, data distribution shifts, and prediction accuracy over time, ensuring continued relevance.
- **Continuous Learning Loop**:Establish mechanisms to regularly retrain the model using newly acquired customer behavior and churn data to improve adaptability and accuracy.
- **Payment Experience Optimization**: churn data to improve adaptability and accuracy.,m**A/B Testing for Strategy Validation**: Use controlled experiments (A/B tests) to test the effectiveness of retention campaigns based on churn probability scores and continuously optimize them.

# 8. Team Contribution Breakdown

**PIAO, Zhuying – Business Understanding + Model Evaluation Lead (20%)** defined the business problem and authored the comprehensive Introduction and Business Understanding section of the report. This member selected and justified the evaluation metrics including Accuracy, F1-score, and ROC-AUC based on business relevance, compiled the model comparison table, and recommended SVM as the best-performing model.

**CHEN, Zhiying – Data Understanding + Logistic Regression (20%)** explored the dataset by describing its Kaggle source, variables, and data types while conducting exploratory data analysis with summary statistics and visualizations. This member trained and evaluated the Logistic Regression model, and wrote both the Data Understanding and Logistic Regression sections in the report.

**HAN, Qingying – Data Preprocessing + Decision Tree (20%)** cleaned the dataset by handling missing values, encoding categorical variables, and normalizing numerical features. They split the dataset into training and testing sets, applied cross-validation methodology, and trained and evaluated the Decision Tree model. This member shared preprocessing notebooks and code via GitHub.

**XU Chenjunxiu – SVM Modeling + Ethical Reflections (20%)** trained and evaluated the Support Vector Machine model while comparing SVM performance with other models using the selected metrics. This member wrote the SVM Results and Ethical Reflections sections in the report, addressing privacy, fairness, and potential algorithmic bias concerns.

**ZHENG, Shuyu – Demo App + Final Report Assembly (20%)** built the prediction API using FastAPI and tested live endpoints for functionality. This member created a clear and reproducible README.md for GitHub including code setup and usage instructions, assembled the final report in PDF format using content from all team members, ensured final submission met all formatting, GitHub, and demo requirements.

# Conclusion

This project successfully demonstrates the application of machine learning techniques to solve a real-world business problem. The SVM model achieved 75% accuracy in predicting customer churn, providing actionable insights for retention strategies. The CRISP-DM framework ensured systematic approach to problem-solving, while ethical considerations guide responsible implementation. The developed models offer significant business value through improved customer retention and revenue optimization.