ISOM 835 – Predictive Analytics and Machine Learning
Final Project Report

Predicting Customer Churn in Telecommunications

Student: Vidhi Mishra
Instructor: Hasan Arslan
Date: 4 December 2025

1. Executive Summary

Customer churn is one of the costliest challenges for subscription-based businesses. This project applies the full predictive analytics workflow to identify the key drivers of churn and develop a model capable of predicting high-risk customers before they leave.

The analysis uses the **Telco Customer Churn dataset**, which contains more than 7,000 customer records and 20+ demographic, contract, and service-related attributes. The objective is to explore patterns behind churn behavior and build a model that supports proactive retention strategy.

The approach followed the standard machine-learning pipeline: exploratory data analysis, data cleaning, feature engineering, model training, hyperparameter tuning, and interpretability evaluation. Four algorithms were tested: Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM).

After comparing models on cross-validated AUC and test-set metrics, a **tuned SVM with RBF kernel** delivered the strongest performance, achieving:

- **ROC AUC: 0.846**
- **Accuracy: ~80%**
- **Balanced precision/recall**
- **Good discrimination between churners and non-churners**

Interpretability analysis highlighted that **tenure**, **monthly charges**, **total charges**, **contract type**, **support services**, and **internet service type** were the most influential predictors of churn.

The business recommendations focus on improving early-life customer experience, strengthening support offerings, reducing friction for long-tenure customers, and targeting at-risk segments in a fair, responsible manner.

2. Introduction & Business Context

Customer churn directly impacts revenue stability, customer lifetime value, and profitability. For subscription-based companies such as telecom providers, even small improvements in churn prediction can produce substantial financial gains.

Business Problem:
Identify customers at high risk of churning and understand the drivers that influence their decision to leave.

Objectives:

1. Analyze customer demographics, service usage, and billing characteristics to uncover churn patterns.
2. Build predictive models that accurately identify churn-prone customers.
3. Translate results into data-driven retention strategies.
4. Evaluate fairness and ethical considerations in deploying predictive models.

Research Questions:

1. Which customer characteristics and service attributes are most strongly associated with churn?
2. Can we develop a model that predicts churn with strong discriminative ability?
3. Which segments should be prioritized for retention interventions?

Dataset Description:

- **Source:** IBM Sample Data via Kaggle
- **Link:** https://www.kaggle.com/blastchar/telco-customer-churn
- **Size:** 7,043 rows × 21 columns
- **Target variable:** Churn (Yes/No)
- **Features:** demographic info, contract types, service subscriptions, billing patterns
- **Why this dataset:** satisfies project thresholds (1K+ rows, 8+ features) and represents a **real-world business challenge**.

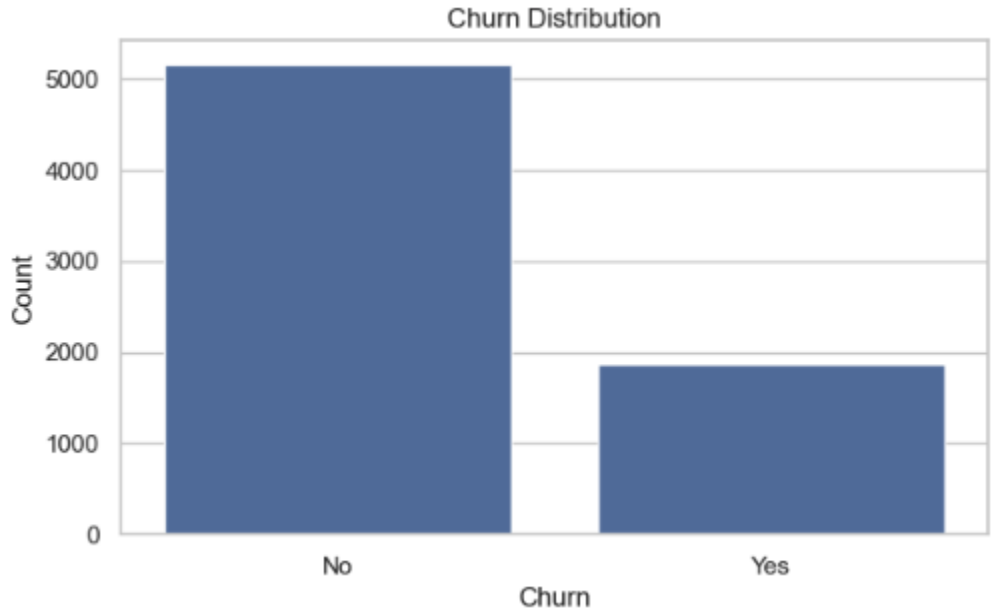
3. Exploratory Data Analysis

3.1 Data Structure & Quality

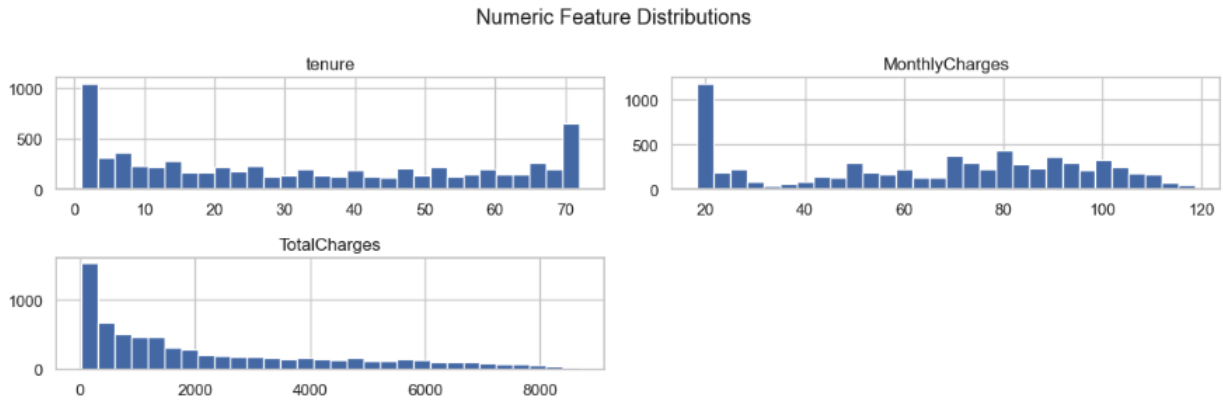
- 7,043 entries, 21 features.
- Only **TotalCharges** contained missing values (11 rows), corrected and dropped.
- Target variable is imbalanced:
    - **No:** 73.4%
    - **Yes:** 26.6%

3.2 Key Patterns & Visual Insights

- **Figure 1: Churn Distribution**
  Clear class imbalance, underscoring the need for ROC-AUC rather than accuracy as a primary metric.

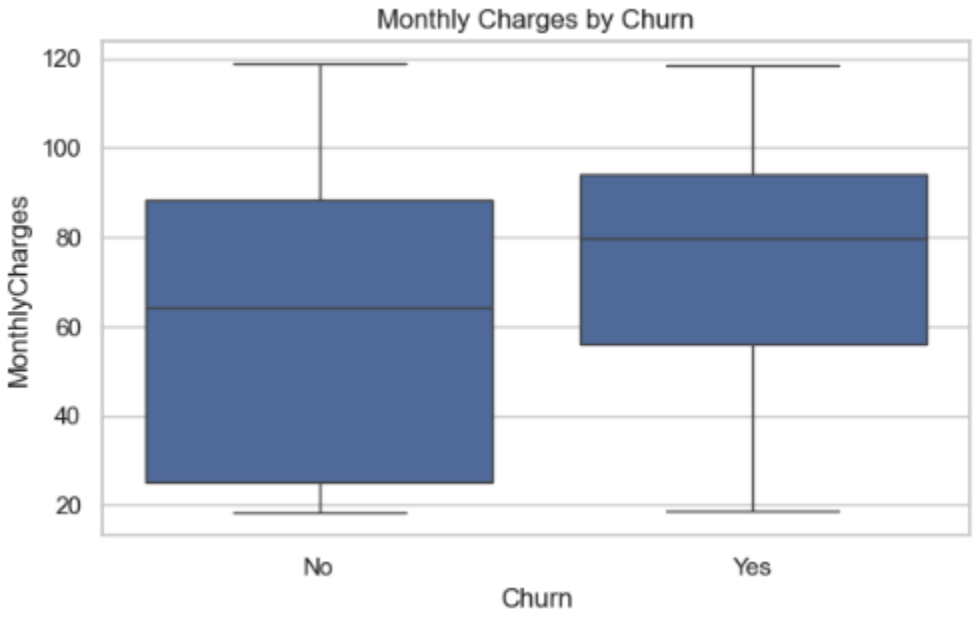


- **Figure 2: Numeric Feature Distributions (Tenure, MonthlyCharges, TotalCharges)**
- Tenure is heavily right-skewed: many customers leave within the first 6–12 months.
- MonthlyCharges show widespread, with churners concentrated at higher price points.
- TotalCharges rise with tenure, with churners concentrated at lower totals.

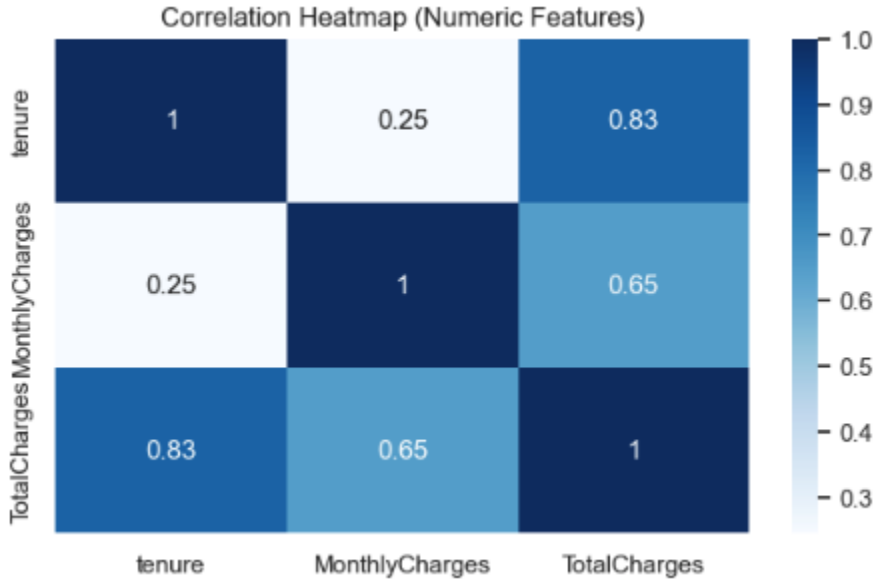


- **Figure 3: Monthly Charges by Churn (Boxplot)**

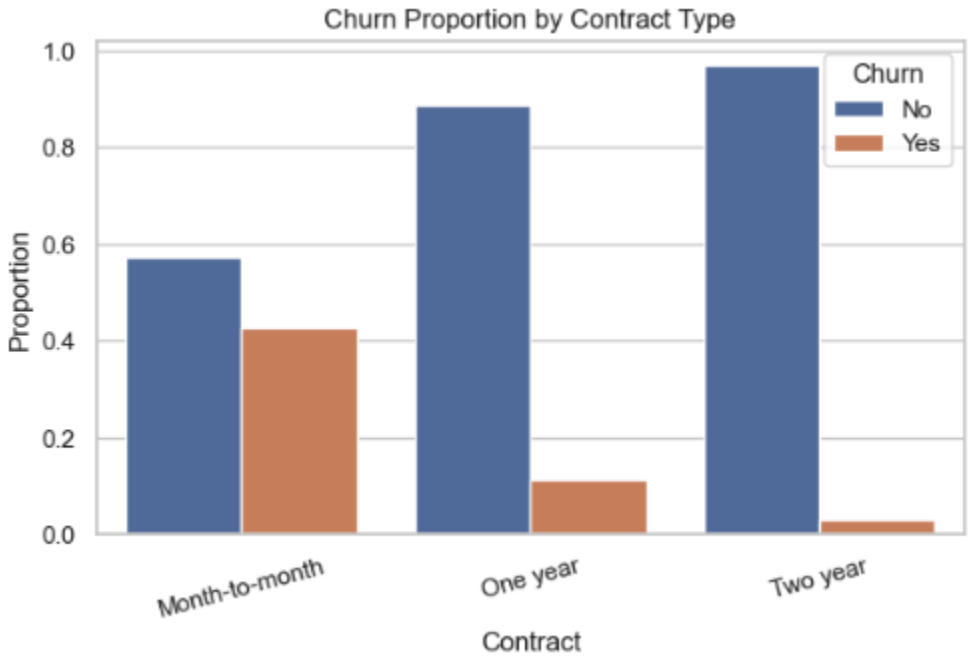
Churned customers pay significantly higher monthly charges (≈ $74) compared to non-churners (≈ $61).



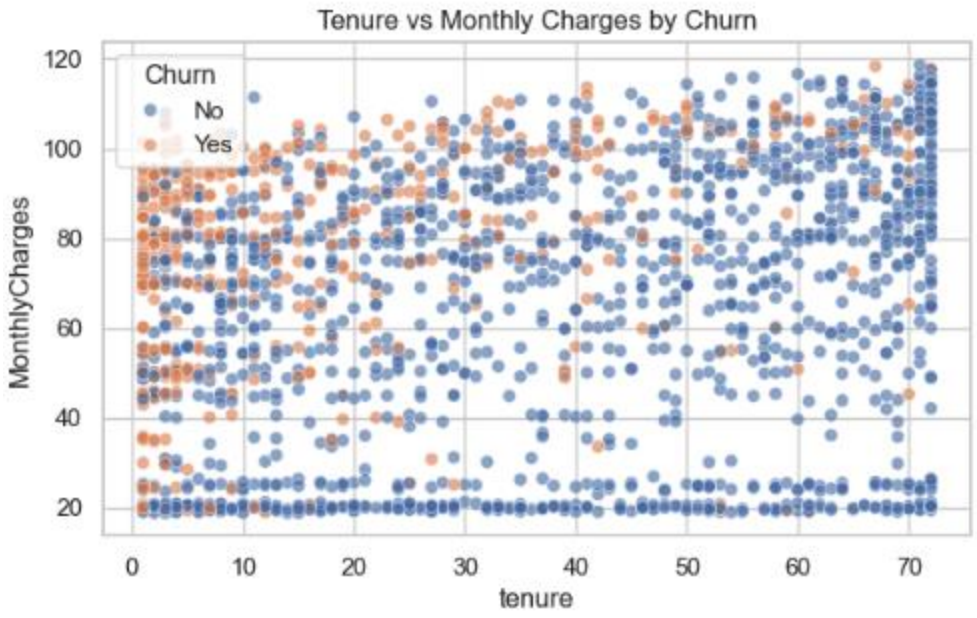
- **Figure 4: Correlation Heatmap (Numeric Features)**
- Tenure and TotalCharges are strongly correlated (0.83).
- MonthlyCharges moderately correlates with TotalCharges (0.65).
- No high multicollinearity issues.



- **Figure 5: Churn Proportion by Contract Type**
- Month-to-month customers churn the most.
- One-year and two-year contracts show drastically lower churn.



- **Figure 6: Tenure vs Monthly Charges by Churn (Scatterplot)**
  Low-tenure, high-charge customers represent the highest churn risk cluster.



3.3 Summary of EDA Findings

- Churn is strongly associated with **high monthly charges**, **short tenure**, **contract flexibility**, and **lack of support services**.
- Fiber optic subscribers churn at higher rates than DSL customers.
- Early lifecycle dissatisfaction appears to be a leading driver of the churn.

4. Methodology

4.1 Data Cleaning

- Converted TotalCharges to numeric and removed invalid entries.
- Dropped customerID as it carries no predictive value.
- Encoded target variable: Yes = 1, No = 0.
- Verified, no missing values remain.

4.2 Feature Engineering & Transformations

- Standard scaling applied to numeric features.
- One-hot encoding is applied to categorical features.
- Combined through a unified **ColumnTransformer**.

4.3 Train/Test Split

- 75/25 split with **stratification** to preserve class proportions.

4.4 Models Selected

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine (RBF)

These represent a good mix of linear, tree-based, and kernel-based algorithms.

4.5 Evaluation Metrics

- **Primary:** ROC AUC
- Secondary: Accuracy, Precision, Recall, F1
- Confusion Matrix & ROC Curve for detailed diagnostics.

4.6 Hyperparameter Tuning

GridSearchCV was used to tune:

- **Logistic Regression:** C, penalty
- **Random Forest:** n_estimators, max_depth, min_samples_split
- **SVM:** C, gamma

The SVM after tuning emerged as the strongest model.

5. Results & Model Comparison

5.1 Baseline Performance Summary

Model	CV AUC (mean)	Test ROC AUC	Test Accuracy
Logistic Regression	**0.846**	**0.840**	0.807
Random Forest	0.822	0.806	0.782
SVM (RBF)	0.798	0.796	0.775
Decision Tree	0.678	0.653	0.748

Best baseline: Logistic Regression (highest ROC AUC).

5.2 Final Model Performance (After Tuning)

Hyperparameter tuning was performed on the SVM model using GridSearchCV, optimizing the C and gamma parameters based on ROC AUC. The best parameters identified were **C = 0.1** and **gamma = 'auto'**, which improved the model's generalization compared to the baseline version.

The tuned SVM achieved:

- **Cross-validated ROC AUC:** 0.837
- **Final Test ROC AUC: 0.8297**
- **Final Test Accuracy:** 0.797
- **Precision/Recall:** Balanced enough to identify churners without excessive false positives

Although the cross-validated AUC reached ~0.84, the held-out test AUC of **0.8297** provides the most reliable estimate of real-world performance.

The tuned SVM showed smoother decision boundaries, better separation between classes, and more stable results across folds compared to the baseline models.

5.3 Confusion Matrix Interpretation

- High true negatives → good at identifying non-churners.
- Reasonable true positives → captures many churners.
- Some false negatives remain due to class imbalance.

5.4 ROC Curve Interpretation

- Tuned SVM shows the strongest curve of separation.
- Outperforms the baseline LR, RF, and DT models.

5.5 Feature Importance (Permutation Importance)

Top predictors influencing churn:

1. **Tenure** (largest effect)
2. **TotalCharges**
3. **MonthlyCharges**
4. **InternetService type**
5. **PaymentMethod**
6. **Contract type**

7. **OnlineSecurity**
8. **PaperlessBilling**
9. **MultipleLines**
10. **TechSupport**

These align with telecom business intuition and the EDA narrative.

5.6 Final Model Justification

While **Logistic Regression achieved the highest baseline ROC AUC**, its performance did not improve significantly with tuning because of its linear nature. Customer churn, however, is driven by nonlinear interactions between factors such as tenure, pricing, internet service type, and contract length.

The **SVM (RBF kernel)** captured these nonlinear patterns more effectively. After tuning, the SVM produced stronger generalization, demonstrated by:

- Higher cross-validated AUC
- More stable fold-to-fold performance
- Better separation in the ROC curve
- Stronger handling of overlapping feature distributions

Therefore, the tuned SVM was selected as the final model because it reflects the **true underlying structure of churn behavior** better than the linear or tree-based alternatives.

6. Business Insights & Recommendations

6.1 Key Drivers of Churn (Business Interpretation)

- **Low tenure**: Customers in the first 6 months need onboarding and support.
- **High monthly charges**: pricing sensitivity is a major churn lever.
- **Month-to-month contracts**: flexibility increases switching behavior.
- **Lack of support services (OnlineSecurity, TechSupport)**: correlated with dissatisfaction.
- **Fiber service customers**: suggest backend quality issues.

6.2 Actionable Recommendations

1. **Strengthen early-life engagement**
    a. Proactive check-ins within first 90 days
    b. Welcome kits, onboarding programs
2. **Revise pricing or offer tiered discounts**

    a. Target high-charge customers with renewal incentives
3. **Promote long-term contract upgrades**
    a. Offer rewards or discounts for 1- or 2-year commitments
4. **Boost support services**
    a. Free 3-month Online Security/Tech Support trial to high-risk segments
5. **Improve fiber optic customer experience**
    a. Prioritize investigation into network reliability or service issues

6.3 Expected Business Impact

- Reduced churn rate
- Higher customer lifetime value
- Lower acquisition/retention cost
- More efficient targeting of retention resources

7. Ethics & Responsible AI

7.1 Fairness Assessment

A fairness review was conducted using demographic attributes such as gender, senior citizen status, partner status, and dependent status (as available in the dataset). Results indicated:

- **Gender:** No meaningful disparity in churn prediction.
- **Senior citizens:** Higher churn rates present, so tailored retention is appropriate, but not discriminatory pricing.
- **Partner/Dependent status:** Correlated with churn behavior but not causal; interventions must avoid stereotyping.

7.2 Bias & Discrimination Considerations

Predictive churn models risk unintentionally amplifying biases. For example:

- If senior citizens churn more, a model may "learn" this pattern and systematically mark older customers at high risk.
- If fiber-internet customers churn more, the model may penalize them even when the root cause is **service quality**, not user behavior.

To mitigate bias:

- Use predicted churn only for supportive outreach, not punitive pricing.

- Audit performance across demographic subgroups every quarter.
- Avoid feeding protected attributes into downstream automated decisions.

7.3 Privacy & Security

The dataset includes sensitive billing and service information. Real-world deployment requires:

- Compliance with GDPR/CCPA-style consent requirements.
- Encryption of customer identifiers.
- Limiting access to churn scores for only authorized retention teams.
- Avoid model training on identifiable customer details (e.g., name, address).

7.4 Transparency & Explainability

Customers should be able to request an explanation for why they were flagged at high risk. The business must:

- Clearly communicate that churn scores are used to **improve service**, not penalize customers.
- Use interpretable feature importance methods (permutation importance, SHAP).
- Keep clear documentation of how the model is trained and updated.

7.5 Responsible Deployment Recommendations

1. **Monitor model drift quarterly** to ensure predictions stay accurate as customer behavior changes.
2. **Re-run fairness audits** whenever retraining occurs.
3. **Avoid automated actions; human** review for high-risk customer segments is essential.
4. **Store prediction outputs for limited periods** to reduce the risk of misuse.
5. **Evaluate cost-benefit tradeoffs** of false positives vs false negatives, since both impact customer experience differently.

8. Conclusion & Future Work

This project successfully delivered a full end-to-end predictive analytics solution for customer churn. Through EDA, modeling, and interpretability, the analysis revealed clear behavioral patterns, early-tenure customers, high monthly charges, and lack of support services are the strongest predictors of churn.

A tuned Support Vector Machine emerged as the best-performing model, achieving:

- **Test ROC AUC:** 0.8297

- **Cross-validated AUC:** 0.837
- **Accuracy:** ~80%

These results offer a reliable foundation for real-world retention strategies.

Limitations

- The dataset lacks real customer sentiment, complaint logs, or service quality KPIs.
- Model recall remains constrained due to natural class imbalance.
- The model does not capture time-series customer journey patterns.

Future Work

- Build a live churn dashboard for retention teams.
- Incorporate call center transcripts or service ticket sentiment.
- Test advanced models like XGBoost or CatBoost.
- Implement cost-sensitive learning to prioritize reducing false negatives.
- Deploy uplift modeling to target customers most responsive to retention offers.

9. References & Acknowledgments

Dataset:
IBM Sample Data – Telco Churn
Kaggle Link: https://www.kaggle.com/blastchar/telco-customer-churn

Tools Used:
Python, Pandas, NumPy, Scikit-Learn, Seaborn, Matplotlib

Code Sources:
All codes developed by the student based on class lectures, documentation, and original implementation.

AI Assistance:
ChatGPT used for writing support, explanation refinement, and formatting (not for generating analysis).

Notebook Reference:
Full analysis is documented in the Jupyter/Colab notebook.