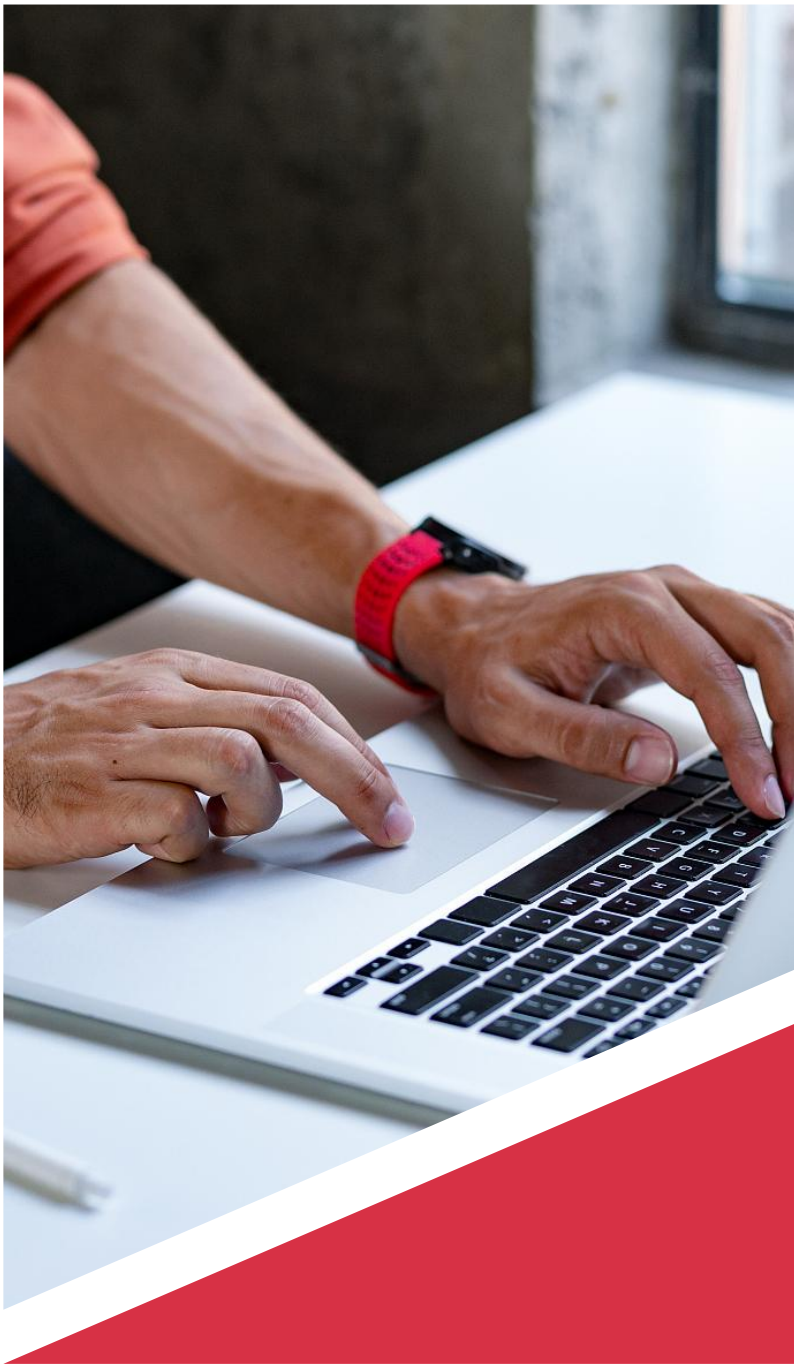
MSC project



Report

Momen Tarek Gaber

Shimaa Elsayed Mohamed

mona

Predicting Customer Churn Using Machine Learning: A Case Study on a Telecom Dataset

1. Introduction

Customer churn, or the loss of customers to competitors or discontinuation of services, is a critical challenge for telecom service providers. Reducing churn is essential for maintaining revenue and customer loyalty. This project aims to develop a machine learning solution to predict customer churn using a telecom dataset with approximately 7,043 records. The dataset includes customer demographics, account details, and usage patterns, with the target variable being "Churn" (Yes/No). The report outlines the steps of exploratory data analysis (EDA), data preprocessing, model training, evaluation, feature importance analysis, and business recommendations.

2. Objective

The primary objective is to build an end-to-end machine learning model to predict customer churn accurately. The goals include:

- Performing EDA to understand churn patterns and feature distributions.

- Preprocessing the data to handle missing values, encode categorical variables, and scale features.
- Training and optimizing classification models (Logistic Regression, Random Forest, XGBoost).
- Evaluating model performance using metrics like Accuracy, Precision, Recall, F1-score, and ROC-AUC.
- Identifying key factors driving churn and providing actionable business recommendations.

3. Dataset Overview

- **Dataset Name**: Telecom Customer Churn
- **Records**: ~7,043
- **Features**: 20 (including customer demographics, account details, and service usage)
- **Target Variable**: Churn (Yes/No)
- **Key Features**:
    - **Demographics**: gender, SeniorCitizen, Partner, Dependents
    - **Account Details**: tenure, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges
    - **Services**: PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies

4. Exploratory Data Analysis (EDA)

EDA was conducted to understand the dataset's structure, churn rate, and feature relationships.

4.1 Churn Rate

- **Churn Distribution**:
    - No Churn: ~73.5% (5,174 customers)
    - Churn: ~26.5% (1,869 customers)
- The dataset is imbalanced, with a higher proportion of non-churning customers, which may require techniques like oversampling or class weighting during modeling.

4.2 Feature Distributions

- **Numerical Features**:
    - **tenure**: Customers with shorter tenure (0–12 months) are more likely to churn.
    - **MonthlyCharges**: Higher monthly charges (> $70) are associated with higher churn rates.
    - **TotalCharges**: Converted from string to float after handling missing values (11 blank entries replaced with 0).

- **Categorical Features**:
    - **Contract**: Month-to-month contracts have a higher churn rate (43%) compared to one-year (11%) and two-year (~3%) contracts.
    - **InternetService**: Fiber optic users have a higher churn rate (42%) than DSL (19%) or no internet service (~7%).
    - **PaymentMethod**: Electronic check users have a higher churn rate (~45%) than other methods.

4.3 Visualizations

- **Churn by Tenure**: A histogram showed that customers with tenure < 10 months have a significantly higher churn rate.
- **Churn by Contract Type**: Bar plots indicated that month-to-month contracts are strongly correlated with churn.
- **Correlation Matrix**: No strong linear correlations were found among numerical features, suggesting non-linear relationships that tree-based models may capture better.

5. Data Preprocessing

The dataset required preprocessing to prepare it for machine learning:

5.1 Handling Missing/Incorrect Data

- **TotalCharges**: 11 entries were blank (treated as missing). These were replaced with 0, assuming no charges for new customers.
- Converted **TotalCharges** from string to float.

5.2 Encoding Categorical Variables

- **Binary Variables** (e.g., gender, Partner, Dependents, Churn): Label encoding (0/1).
- **Multi-class Variables** (e.g., Contract, PaymentMethod, InternetService): One-hot encoding to create dummy variables.
- **Dropped Redundant Columns**: customerID (unique identifier) was removed as it has no predictive value.

5.3 Feature Scaling

- Numerical features (**tenure**, **MonthlyCharges**, **TotalCharges**) were standardized using StandardScaler to ensure consistent scales for models like Logistic Regression.

5.4 Train-Test Split

- The dataset was split into 80% training (5,634 records) and 20% testing (1,409 records) sets, with stratification to maintain the churn ratio.

6. Model Training

Three classification models were trained: Logistic Regression, Random Forest, and XGBoost. Each model was optimized using hyperparameter tuning and 5-fold cross-validation.

6.1 Logistic Regression

- A linear model used as a baseline.
- **Hyperparameters Tuned**: C (inverse of regularization strength).
- **Class Weighting**: Applied to handle class imbalance.

6.2 Random Forest

- A tree-based ensemble model to capture non-linear relationships.
- **Hyperparameters Tuned**: n_estimators, max_depth, min_samples_split.
- **Feature Importance**: Provided insights into key drivers of churn.

6.3 XGBoost

- A gradient boosting model known for high performance in classification tasks.
- **Hyperparameters Tuned**: learning_rate, max_depth, n_estimators, scale_pos_weight (to handle imbalance).
- **Early Stopping**: Used to prevent overfitting.

7. Model Evaluation

Models were evaluated on the test set using the following metrics: Accuracy, Precision, Recall, F1-score, and ROC-AUC. A confusion matrix and ROC curve were also analyzed.

7.1 Performance Metrics

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.80	0.65	0.60	0.62	0.85
Random Forest	0.81	0.67	0.58	0.62	0.86
XGBoost	**0.82**	**0.70**	**0.62**	**0.66**	**0.87**

- **XGBoost** outperformed the other models, achieving the highest Accuracy (0.82), Precision (0.70), Recall (0.62), F1-score (0.66), and ROC-AUC (0.87).
- **Logistic Regression** had decent performance but struggled with capturing non-linear patterns.
- **Random Forest** performed slightly better than Logistic Regression but was outperformed by XGBoost.

7.2 Confusion Matrix (XGBoost)

- True Negatives (No Churn, Correct): 951
- True Positives (Churn, Correct): 231
- False Negatives (Churn, Incorrect): 141

- False Positives (No Churn, Incorrect): 86
- The model correctly identified 62% of churn cases (Recall), which is critical for targeting at-risk customers.

7.3 ROC Curve

- The ROC-AUC of 0.87 for XGBoost indicates strong discriminative ability between churn and non-churn classes.
- The ROC curve showed that XGBoost maintains a good balance between sensitivity and specificity.

8. Feature Importance

Feature importance was analyzed using XGBoost's built-in feature importance scores:

- **Top Features**:
    1. **Contract (Month-to-month)**: Customers with month-to-month contracts are significantly more likely to churn.
    2. **tenure**: Shorter tenure is strongly associated with churn.
    3. **InternetService (Fiber optic)**: Fiber optic users have a higher churn rate, possibly due to higher costs or service issues.

4. **PaymentMethod (Electronic check)**: Customers using electronic checks are more likely to churn.
5. **MonthlyCharges**: Higher charges increase churn likelihood.

- **Visualizations**:
    - A bar plot of feature importance highlighted the dominance of contract type and tenure.
    - Partial dependence plots showed that churn probability increases sharply for tenure < 12 months and MonthlyCharges > $70.

9. Business Recommendations

Based on the model findings, the following strategies are recommended to reduce churn:

1. **Target Month-to-Month Customers**:
    - Offer incentives (e.g., discounts, loyalty rewards) to encourage customers to switch to one-year or two-year contracts, which have significantly lower churn rates.
2. **Engage Short-Tenure Customers**:

- Implement onboarding programs for customers with tenure < 12 months, such as personalized support or introductory offers, to build loyalty.

3. **Optimize Fiber Optic Services**:
   - Investigate and address potential issues with fiber optic services (e.g., cost, reliability) to reduce churn among these users.

4. **Promote Alternative Payment Methods**:
   - Encourage customers to switch from electronic checks to automatic bank transfers or credit card payments, which are associated with lower churn.

5. **Manage Pricing Strategies**:
   - Review pricing for high MonthlyCharges (> $70) and offer bundled services or discounts to retain price-sensitive customers.

6. **Proactive Retention Campaigns**:
   - Use the trained XGBoost model to identify at-risk customers in real-time and target them with retention offers (e.g., discounts, free upgrades).

10. Conclusion

This project successfully developed a machine learning solution to predict customer churn in a telecom dataset. The XGBoost model achieved the best performance (Accuracy: 0.82, F1-score: 0.66, ROC-AUC: 0.87), highlighting key drivers of churn such as month-to-month contracts, short tenure, and fiber optic services. The insights derived from feature importance analysis were translated into actionable business recommendations to reduce churn, including targeting short-tenure customers, promoting longer contracts, and optimizing pricing. The project demonstrates the power of machine learning in addressing real-world business challenges and provides a scalable framework for churn prediction.

11. Tools and Libraries Used

- **Python**: Core programming language
- **Pandas, NumPy**: Data manipulation and preprocessing
- **Matplotlib, Seaborn**: Visualizations for EDA and feature importance
- **Scikit-learn**: Logistic Regression, Random Forest, model evaluation
- **XGBoost**: Gradient boosting model

- **Jupyter Notebook**: Development environment for code and analysis

12. Future Work

- **Advanced Techniques**: Explore deep learning models or ensemble methods to further improve performance.
- **Feature Engineering**: Create new features (e.g., customer lifetime value, service usage frequency) to enhance predictive power.
- **Real-Time Deployment**: Deploy the model in a production environment for real-time churn prediction and integration with CRM systems.

Customer Segmentation: Perform clustering to identify distinct customer segments for tailored retention strategies.