Customer Churn Prediction Using Machine Learning

Abstract

Customer churn is a major concern for businesses, impacting revenue and long-term sustainability. This study presents a machine learning-based approach to predict customer churn using classification models such as Logistic Regression, Random Forest, and XGBoost. The dataset utilized for this study originates from a telecom service provider. The implemented models were evaluated using standard performance metrics, and the best-performing model was deployed using a Flask-based API. This research highlights the advantages and challenges of using machine learning for churn prediction and provides insights into improving customer retention strategies.

1. Introduction

Customer churn occurs when customers discontinue using a service, posing a critical challenge for businesses across industries such as telecom, banking, and e-commerce. Predicting churn helps organizations take proactive measures to retain at-risk customers. Traditional statistical models often fail to capture complex customer behavior patterns, making machine learning (ML) a promising alternative.

This project focuses on building a **churn prediction model** using ML algorithms, evaluating their performance, and deploying the best model as a REST API. The main objectives include:

- Data preprocessing and feature engineering.
- Training multiple ML models and selecting the best one.
- Deploying the best model for real-time predictions.

---

2. Dataset Overview

The dataset used in this study is the **Telco Customer Churn Dataset**, which contains customer demographics, account details, and service usage data. The dataset comprises:

- **Total records:** 7043 customers
- **Features:** 19 independent variables, including:
    - **Demographics:** Gender, Senior Citizen, Partner, Dependents

    - **Account Information:** Contract type, Paperless billing, Payment method
    - **Usage Details:** Monthly charges, Total charges, Tenure
    - **Target Variable:** Churn (Yes/No, later converted to 1/0)

Data Preprocessing

- Converted categorical variables to numerical values.
- Handled missing values in the `TotalCharges` column.
- Standardized numerical features for models that require normalization.
- Split dataset into **80% training** and **20% testing** sets.

3. Machine Learning Approaches

Three different machine learning models were trained and evaluated:

3.1 Logistic Regression

- A statistical model commonly used for binary classification.
- Suitable for understanding how each feature impacts churn probability.

3.2 Random Forest Classifier

- An ensemble learning method that constructs multiple decision trees.
- Reduces overfitting compared to individual decision trees.

3.3 XGBoost Classifier

- A powerful gradient boosting algorithm known for high performance.
- Handles complex patterns and interactions between features.

Why These Models?

- Logistic Regression provides interpretability.
- Random Forest balances performance and robustness.
- XGBoost delivers high accuracy and efficiency.

---

4. Evaluation Metrics

To assess model performance, the following metrics were used:

- **Accuracy**: Measures overall correctness.
- **Precision**: Ratio of correctly predicted churn cases.
- **Recall (Sensitivity)**: Ability to detect churn cases.
- **F1-Score**: Harmonic mean of precision and recall.
- **Confusion Matrix**: Visual representation of predictions vs actual values.

Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	79.2%	76.5%	72.3%	74.3%
Random Forest	85.1%	83.4%	80.7%	82.0%
XGBoost	**87.4%**	**85.9%**	**83.1%**	**84.4%**

5. Deployment Approach

After training, the best-performing model (XGBoost) was deployed using a **Flask API**. The API allows external applications to send customer data and receive churn predictions in real-time.

Deployment Steps:

1. **Save trained model:**
    - `joblib.dump(model, "models/xgboost_best_model.pkl")`
2. **Create Flask API (`churn_api.py`)**
3. **Expose a POST endpoint (`/predict`)** to accept customer data.
4. **Deploy on Render** for public access.

Example **API Request:**

```
{
  "features": [1, 0, 1, 0, 12, 1, 0, 2, 0, 1, 1, 0, 1, 1, 0, 1, 2, 45.3, 540.5]
}
```

Example **API Response:**

```
{
  "churn_prediction": 1,
  "churn_probability": 0.87
}
```

---

6. Advantages and Challenges

Advantages

✅ Automates customer churn detection with high accuracy. ✅ Helps businesses make data-driven retention decisions. ✅ Can be improved with real-time customer feedback.

Challenges

⚠️ Imbalanced dataset may affect recall. ⚠️ Feature selection impacts model interpretability. ⚠️ Requires periodic retraining as customer behavior evolves.

7. Conclusion

This study demonstrated how machine learning can effectively predict customer churn. Among the models tested, **XGBoost performed best** with **87.4% accuracy**. The model was deployed using a **Flask API**, enabling real-time predictions for customer retention strategies. Future work includes optimizing hyperparameters further and integrating deep learning models.

---

8. References

[1] J. Brownlee, "Machine Learning Mastery," 2021.
[2] Kaggle Dataset: "Telco Customer Churn".
[3] XGBoost Documentation: https://xgboost.readthedocs.io