Customer Churn Prediction Using Machine Learning

Author: [Amanuel Getachew]
id: [1500016]
Course: [Fundamentals of Machine Learning]

Abstract

This project aims to predict customer churn in the telecommunications industry using machine learning. By analyzing a dataset containing customer demographics, service usage patterns, and account details, we developed a Random Forest classifier to identify customers at risk of leaving. The dataset was preprocessed to handle missing values, scale numerical features, and balance class distribution using SMOTE. Hyperparameter tuning was performed using Grid Search with cross-validation to optimize model performance. The final model achieved an F1-score of [X%] on the test set and was deployed as a RESTful API using FastAPI. Additionally, a user interface was created using HTML/JavaScript for easy interaction with the API. This project demonstrates the practical application of machine learning in solving real-world business problems.

Problem Definition

Customer churn is a critical issue for businesses, especially in competitive industries like telecommunications. Predicting which customers are likely to leave allows companies to take proactive measures, such as offering discounts or improving customer service. The goal of this project is to develop a machine learning model that accurately identifies customers at risk of churning based on their demographic information, service usage patterns, and account details. This is a binary classification problem where the target variable indicates whether a customer will churn (`1`) or not (`0`).

Data Understanding and Exploration

The dataset used in this project contains historical records of customer interactions and churn status. It includes features such as:
- Demographic features: `gender`, `SeniorCitizen`, `Partner`, `Dependents`
- Service features: `PhoneService`, `MultipleLines`, `InternetService`, etc.
- Account features: `tenure`, `MonthlyCharges`, `TotalCharges`

Key observations from exploratory data analysis include:
- The dataset is highly imbalanced, with only [X%] of customers labeled as churned.
- Features like `tenure`, `MonthlyCharges`, and `TotalCharges` show strong correlations with the target variable.
- Outliers were identified in the `MonthlyCharges` and `TotalCharges` features but retained due to their potential importance in churn detection.

Data Preprocessing

The following preprocessing steps were performed:
1. **Handling Missing Values**:
   - Missing values in the `TotalCharges` column were handled by converting them to numeric and dropping rows with missing data.
2. **Encoding Categorical Variables**:
   - Categorical variables were encoded using `LabelEncoder`.
3. **Scaling Numerical Features**:
   - Numerical features (`MonthlyCharges`, `TotalCharges`, `tenure`) were scaled using `StandardScaler`.
4. **Balancing Class Distribution**:
   - Class imbalance was addressed using SMOTE to oversample the minority class.

Model Implementation and Training

A Random Forest classifier was selected due to its robustness to imbalanced datasets and ability to capture complex relationships. Hyperparameter tuning was performed using Grid Search with 3-fold cross-validation. The best parameters identified were:
- `n_estimators`: [X]
- `max_depth`: [X]
- `min_samples_split`: [X]

The model was trained on the preprocessed dataset and achieved the following metrics on the test set:
- Accuracy: [X%]
- Precision: [X%]
- Recall: [X%]
- F1-Score: [X%]
- ROC-AUC: [X%]

These metrics indicate that the model effectively balances precision and recall, ensuring minimal false positives while capturing most churn cases.

Model Evaluation

The model was evaluated using the following key metrics:
- **Accuracy**: Measures overall correctness of predictions.
- **Precision**: Proportion of predicted churns that are actually churns.
- **Recall**: Proportion of actual churns correctly identified.
- **F1-Score**: Harmonic mean of precision and recall.
- **ROC-AUC**: Measures the ability of the model to distinguish between classes.

The final model achieved:
- **F1-Score**: [X%]
- **ROC-AUC**: [X%]

These results demonstrate the model's effectiveness in predicting customer churn.

Model Deployment

The trained model was deployed as a RESTful API using FastAPI. The API accepts JSON input containing customer details and returns the prediction along with the churn probability. Additionally, a user interface was created using HTML/JavaScript for easy interaction with the API. The application was deployed on Render, ensuring it is accessible online.

Conclusion and Future Work

This project successfully developed and deployed a machine learning model for predicting customer churn. The Random Forest classifier demonstrated strong performance, achieving an F1-score of [X%]. Future work could involve:
- Exploring other algorithms like XGBoost or Gradient Boosting.
- Incorporating more features or external data sources.
- Deploying the model on a larger scale for real-time predictions.

References

- Dataset Source: [Telco Customer Churn Dataset](https://www.kaggle.com/datasets/blastchar/telco-customer-churn)
- Libraries Used: scikit-learn, pandas, numpy, FastAPI, joblib, Render
- Documentation: [FastAPI Docs](https://fastapi.tiangolo.com/), [Render Docs](https://render.com/docs) change this to pdf file with the name report.pdf