Report

Final Report: Customer Churn Prediction and Analysis Project

Project Overview

This project aimed to develop a machine learning model to predict customer churn for a telecommunications company. The dataset used included customer demographics, service usage patterns, and billing information. The key phases of the project involved data collection, data analysis, model development, and final deployment. The project concluded with the deployment of the model as a web application to predict churn in real time.

Week 1: Data Collection and Exploration

1. Data Collection

The dataset, consisting of 7,043 rows and 21 columns, included the following key features:

- Demographics: Gender, SeniorCitizen, Partner, Dependents.
- Service features: InternetService, StreamingTV, OnlineSecurity.
- Billing features: MonthlyCharges, TotalCharges, Contract type.
- Target variable: Churn (Yes/No).

2. Data Exploration

- Missing values: Detected and removed 11 missing values in the TotalCharges column.
- Imbalance: Noted class imbalance with a churn rate of ~26.5%.
- Data types: Identified categorical variables for encoding, such as Gender and InternetService.

Key Insights

- The imbalance in the Churn variable requires careful model handling to avoid biased predictions.
- Missing values were minimal and did not impact the overall dataset.

Tools Used

- Python: Pandas, NumPy.

Deliverables

- Exploratory Data Analysis (EDA): Insights into the dataset.
- Initial Report: Summary of patterns related to churn.

Week 2: Data Cleaning, Analysis, and Visualization

1. Data Cleaning

- Removed missing values in the TotalCharges column.
- Retained outliers in MonthlyCharges, as they likely represented valuable customers.
- Applied one-hot encoding for categorical variables such as Contract and InternetService.

2. Data Analysis

- Contract type: Customers with month-to-month contracts were more likely to churn.
- Payment method: Electronic check users showed the highest churn rates.
- Tenure: Customers with shorter tenure were at higher risk of churn.
- Created a new feature, tenure_group, to better capture patterns in customer behavior.

3. Data Visualization

- Visualized relationships between Contract type, Payment method, and churn using graphs.
- Created a correlation heatmap that showed moderate correlations between features like tenure and MonthlyCharges.

Key Insights

- Customers with month-to-month contracts and those paying via electronic check should be targeted for retention efforts.
- Tenure is a significant predictor of churn.

Tools Used

- Python: Matplotlib, Seaborn, Plotly.

Deliverables

- Cleaned Dataset: Prepared for model development with key features created.
- Analysis Report: Detailed insights into churn behavior.

Week 3: Machine Learning Model Development

1. Model Selection

Several models were considered, including:

- Logistic Regression: Baseline model with ~78% accuracy.
- Random Forest: Performed well with ~80% accuracy.
- XGBoost (XGBClassifier): The best-performing model with 84% accuracy.

2. Model Training and Evaluation

- XGBClassifier: Delivered the highest performance with an accuracy of 84%, a precision of 0.87, a recall of 0.78, and an F1 score of 0.82.

3. Model Optimization

- Applied GridSearchCV for hyperparameter tuning.
- Handled class imbalance using SMOTE, improving recall and ensuring the model captures more churn cases.

Key Insights

- XGBClassifier outperformed other models and effectively handled imbalanced data.

Tools Used

- Python: Scikit-learn, XGBoost, SMOTE.

Deliverables

- Model Performance Report: Highlighting the XGBClassifier as the best candidate for deployment.

Week 4: MLOps, Deployment, and Final Presentation

1. MLOps Implementation

- Logged model parameters and metrics using MLflow for reproducibility and comparison.

2. Model Deployment

- The final XGBClassifier model was deployed as a web app using Streamlit for real-time churn predictions.
- The app allows users to input customer data and receive predictions instantly.

3. Final Report and Presentation

- Delivered a comprehensive final report summarizing the entire project.
- Prepared a business-focused presentation for stakeholders, demonstrating the model's utility in reducing churn.

Key Insights

- The deployed Streamlit app provides a user-friendly interface for stakeholders to make data-driven decisions.
- The XGBClassifier model with 84% accuracy offers a practical solution for customer churn prediction.

Tools Used

- MLflow for tracking, Streamlit for app deployment, Flask (optional for API creation).

Deliverables

- Deployed Web Application: Functional web app for real-time customer churn prediction.
- Final Report: Detailed summary of the entire workflow.
- Business Presentation: Communicating key findings and recommendations.

Key Insights from Each Phase

- 1. Data Exploration: The churn imbalance (~26%) highlighted the need for proper handling of class imbalance during model development.
- 2. Data Analysis: Customers with month-to-month contracts and those paying via electronic check are key targets for retention.
- 3. Modeling: The XGBClassifier achieved an accuracy of 84%, demonstrating its effectiveness in real-world churn prediction.
- 4. Deployment: The Streamlit app provides telecom companies with a practical tool for identifying at-risk customers and improving retention strategies.

Conclusion

The XGBClassifier model's high accuracy and robust performance make it an ideal solution for predicting customer churn. This tool equips businesses with the insights needed to enhance customer retention strategies and reduce churn effectively.