Telecom Customer Churn Prediction using PySpark-Documentation

Background Information

Customer churn is a significant challenge in the telecom industry. Identifying customers who are likely to churn is crucial for implementing proactive measures to retain them. By leveraging PySpark, we can take advantage of its distributed computing capabilities to handle large volumes of data efficiently and build an accurate machine learning model for churn prediction.

Problem Statement

The goal of this project is to develop a machine learning model using PySpark that accurately predicts customer churn in a telecom company. The model should achieve a minimum accuracy of 0.8, enabling the company to proactively identify and retain customers at risk of leaving. By effectively predicting churn, the company can implement targeted retention strategies, reduce customer attrition, and improve overall business performance.

details about the dataset:

Obtain a telecom customer dataset that includes relevant features such as customer demographics, usage patterns, service plans, call details, customer complaints,and churn status. You can use this dataset.

Perform necessary preprocessing

steps on the dataset, including handling missing values, feature scaling, encoding categorical variables, and splitting the data into training and testing sets. Consider using PySpark's DataFrame API for efficient data manipulation.

feature engineering

Create new features from the existing dataset that might be helpful for predicting churn. For example, you could calculate metrics such as call duration, average monthly spend, customer tenure, or customer satisfaction scores.

model selection, and evaluation results

1. Random Forest Classifier: Random Forest is an ensemble learning algorithm that combines multiple decision trees to make predictions. It was chosen for its ability to handle non-linear relationships and feature interactions.
2. Logistic Regression: Logistic Regression is a linear classification algorithm that estimates the probability of a binary outcome. It was selected as a baseline model for comparison.

project findings

Random Forest Classifier: Accuracy - 0.5

Logistic Regression: Accuracy - 0.8333333333333333 The Logistic Regression model outperformed the Random Forest Classifier, achieving the desired accuracy of 0.8.

challenges faced and lessons learned

Limited Dataset Size: The small size of the dataset made it difficult to train complex models and achieve higher accuracies.