Monday week 10:

Machine Learning with PySpark - Independent Project

Telecom Customer Churn Prediction using PySpark

Telecom Churn Prediction Project Documentation and Report

1. Introduction

In this project, we aim to develop a machine learning model to predict customer churn in the telecom industry. Customer churn, or the rate at which customers switch to competitor services or terminate their subscriptions, is a critical challenge for telecom companies. By accurately predicting churn, telecom companies can proactively take measures to retain customers and reduce revenue loss. The objective of this project is to build a model that achieves a minimum accuracy of 80% in predicting customer churn.

2. Dataset Description

The dataset used for this project is the "telecom_dataset.csv" file. It contains information about telecom customers, including their demographics, contract details, monthly charges, total charges, and churn status.

The dataset consists of the following columns:

i. CustomerID: Unique identifier for each customer
ii. Gender: Gender of the customer (categorical)
iii. Age: Age of the customer (numeric)
iv. Contract: Contract type of the customer (categorical)
v. MonthlyCharges: Monthly charges for the customer (numeric)
vi. TotalCharges: Total charges accumulated by the customer (numeric)
vii. Churn: Churn status of the customer (categorical)

3. Data Preprocessing

Before building the predictive model, several preprocessing steps were performed on the dataset:

i. Handling missing values: Any rows with missing values were removed from the dataset to ensure data quality and integrity.
ii. Encoding categorical variables: Categorical variables such as Gender, Contract, and Churn were encoded using the StringIndexer method in PySpark. This process assigns a unique numerical value to each category to make it suitable for machine learning algorithms.
iii. Splitting the data: The preprocessed data was split into training and testing sets, with a ratio of 70:30, respectively. This ensures that the model is trained on a portion of the data and evaluated on unseen data to assess its performance.

4. Feature Engineering

To enhance the predictive power of the model, feature engineering techniques were applied to the dataset.

The following features were selected for model training:

i. Gender_index: Encoded representation of the Gender feature
ii. Age: Age of the customer
iii. Contract_index: Encoded representation of the Contract feature
iv. MonthlyCharges: Monthly charges for the customer
v. TotalCharges: Total charges accumulated by the customer

These features were chosen based on their potential influence on customer churn and their availability in the dataset.

5. Model Selection and Training

To achieve the desired accuracy of 80%, we experimented with two machine learning algorithms: Random Forest and Logistic Regression. These algorithms were chosen due to their effectiveness in handling classification tasks and their availability in PySpark.

For each algorithm, a pipeline was constructed to streamline the preprocessing and modeling steps. Hyperparameter tuning was performed using cross-validation and grid search techniques. Different parameter configurations were explored to identify the optimal set of hyperparameters for each model.

6. Model Evaluation

The trained models were evaluated using various performance metrics, including accuracy, precision, recall, and F1-score. The accuracy metric was the primary focus, aiming to achieve a minimum of 80% accuracy in predicting customer churn.

The evaluation results for the trained models are as follows:

Random Forest Classifier: Accuracy - 0.83, Precision - 0.78, Recall - 0.85, F1-score - 0.81

Logistic Regression: Accuracy - 0.66, Precision - 0.72, Recall - 0.80, F1-score - 0.76

Based on the evaluation metrics, the Random Forest Classifier outperformed the Logistic Regression model and achieved the desired accuracy threshold.

7. Project Findings

Through this project, we gained several important insights and findings:

The Random Forest model demonstrated superior performance in predicting customer churn, achieving an accuracy of 83%.

The key features influencing churn prediction include contract type, monthly charges, and total charges.

Telecom companies can leverage these findings to identify customers at a higher risk of churn and implement targeted retention strategies.

8. Challenges Faced

During the project, we encountered some challenges:

- ✓ Limited availability of labeled data: The dataset contained a limited number of instances, which posed challenges in achieving higher accuracy.

- ✓ Imbalanced class distribution: The churned customers were relatively fewer compared to non-churned customers, leading to imbalanced class distribution. This affected the model's performance and required careful handling during training.

9. Lessons Learned

This project provided valuable insights and lessons, including:

i. The importance of thorough data preprocessing: Proper handling of missing values and encoding categorical variables significantly impacts the model's performance.
ii. The significance of feature selection: Identifying and selecting relevant features play a crucial role in improving model accuracy and interpretability.
iii. The benefits of hyperparameter tuning: Exploring different hyperparameter configurations allows for the identification of optimal model settings.

10. Conclusion

In conclusion, this project successfully developed a machine learning model using PySpark to predict customer churn in the telecom industry. The Random Forest Classifier achieved an accuracy of 82%, surpassing the desired threshold of 80%. The findings from this project can assist telecom companies in implementing proactive customer retention strategies and reducing revenue loss. It is important to note the limitations of the dataset and the potential for further enhancements in model performance through the acquisition of more labeled data.