**Customer Churn Prediction Using PySpark**

**Introduction**

Predicting customer churn is a critical task for businesses to retain customers and reduce revenue loss. This project uses PySpark to develop a machine learning solution for predicting customer churn in a music app. The dataset captures user interactions with the service, such as playing songs, liking, sharing, and account-related actions. The target variable label indicates whether a user has churned.

This report summarizes the data exploration, feature engineering, and modeling steps, along with key findings and results.

**1. Project Goals**

* Predict customer churn using historical user activity data.
* Engineer features to distinguish between churned and retained users.
* Build and evaluate machine learning models to classify users as churned or retained.

**2. Data Overview**

The dataset comprises interactions from users of a music app. Key columns include:

* **userId**: Unique identifier for each user.
* **page**: Actions performed by users, such as "NextSong," "Thumbs Up," etc.
* **ts**: Timestamp of events.
* **label**: Target variable indicating churn (1) or retention (0).

**Data Exploration**

* **Total Users**: 226 unique users.
* **Unique Page Views**: Various user actions were categorized, including neutral, positive, negative, upgrade, and downgrade pages.

**3. Exploratory Data Analysis (EDA)**

**User Activity by Hour**

User activity was analyzed by aggregating events per hour:

* Events were counted per hour and visualized using scatter plots.
* Key Insight: User activity shows significant variation across hours.

**Churn vs. Retention Analysis**

1. **Average Songs Played**:
   * **Churned Users**: ~699 songs per user.
   * **Retained Users**: ~1108 songs per user.
   * Insight: Retained users are significantly more engaged.
2. **Gender Distribution**:
   * Churned Users: 20 females, 32 males.
   * Retained Users: 84 females, 89 males.
   * Insight: Gender does not show a strong correlation with churn.

**4. Data Preprocessing**

**Feature Engineering**

1. **Binary Gender Feature**:
   * Created GenderBinary (1 for Male, 0 for Female).
2. **Customer Level**:
   * Created LevelBinary (1 for free tier, 0 for paid tier).
3. **Page Interaction Features**:
   * Aggregated counts for neutral, negative, positive, downgrade, and upgrade actions.
4. **User Active Time**:
   * Calculated the number of days between the first and last activity.
5. **Churn Label**:
   * Used the label column to mark churned users.

**Feature Set Schema**

The final feature set includes:

* userId: Unique identifier.
* GenderBinary: Gender feature.
* LevelBinary: Customer tier.
* Interaction features for various page categories.
* UserActiveTime: Duration of user activity.
* label: Target variable.

**5. Machine Learning**

**Modeling Approach**

1. **Feature Assembly**:
   * Combined all features into a single dataframe suitable for machine learning.
2. **Model Selection**:
   * Chose appropriate machine learning algorithms using Spark ML.
3. **Model Tuning**:
   * Hyperparameters were optimized to improve prediction accuracy.

**6. Challenges and Insights**

* **Data Volume**:
  + PySpark effectively handled large datasets, leveraging distributed computation.
* **Feature Importance**:
  + User engagement metrics like song counts and active time strongly correlated with churn.
* **Categorical Features**:
  + Encoding techniques were applied to handle categorical variables such as gender and customer level.

**Results**

The project aimed to utilize Apache Spark's powerful analytics engine for processing large-scale data to identify customers likely to churn from Sparkify's music streaming platform. The following results were observed:

**Model Performance**

1. **Logistic Regression Model**
   * **F1-Score:** 0.66
   * **Recall:** 0.84
   * **Precision:** 0.54
2. **Key Performance Observations**
   * **High Recall (84%)**: The model effectively identifies 84% of customers at risk of churn, enabling the team to implement proactive retention strategies such as personalized offers.
   * **Moderate Precision (54%)**: The precision score indicates that 47% of flagged customers were unlikely to churn and may not require additional retention efforts.

**Implications**

While the high recall allows for early intervention, the moderate precision suggests a trade-off where some resources may be allocated to customers who are already satisfied with the service. This finding underscores the importance of continuously refining the model to optimize resource allocation while maintaining effective churn prevention strategies.