Customer Churn Prediction Project: AWS ML Pipeline

This project focuses on building a **customer churn prediction model** using **AWS services** for **data storage, processing, and machine learning**. The goal is to identify key factors influencing churn and develop a predictive model to assist businesses in **retaining customers**.

Step 1: Create an IAM User

To ensure **secure access** to AWS resources, an IAM user with the necessary permissions was created. Policies were assigned to allow interactions with **S3, Glue, and SageMaker**.

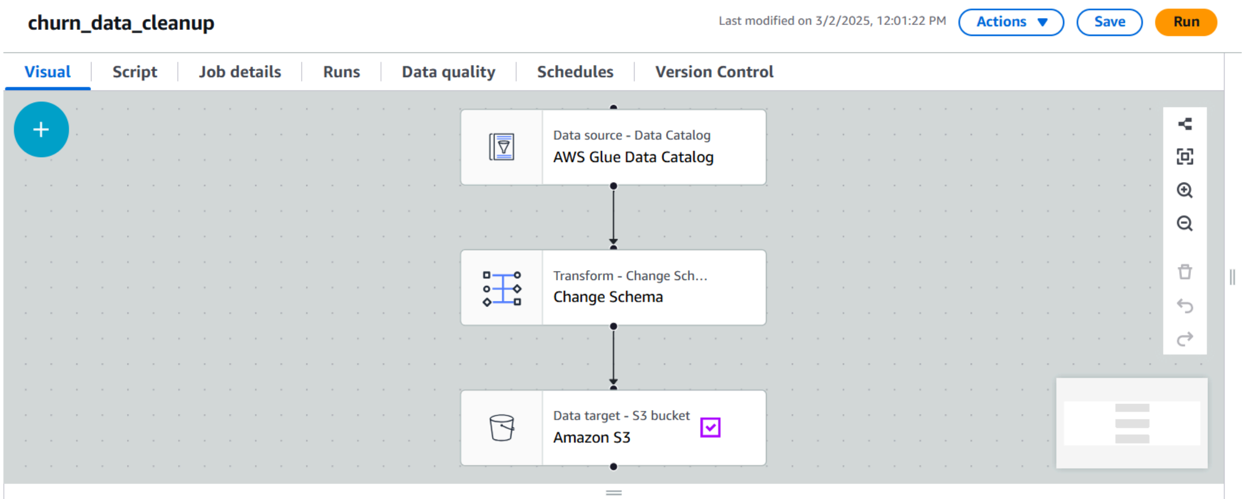
Step 2: Set Up S3 Storage

An **S3 bucket** was created to store the raw dataset (`WA_Fn-UseC_-Telco-Customer-Churn.csv`). This dataset contained **7,043 customer records** with **21 attributes** related to customer demographics, service subscriptions, and billing details.

Step 3: Data Cleaning & Transformation Using AWS Glue ETL

To preprocess the data, an **AWS Glue Crawler** was used to catalog the dataset. Then, an **AWS Glue ETL job** (Visual ETL) was created to:

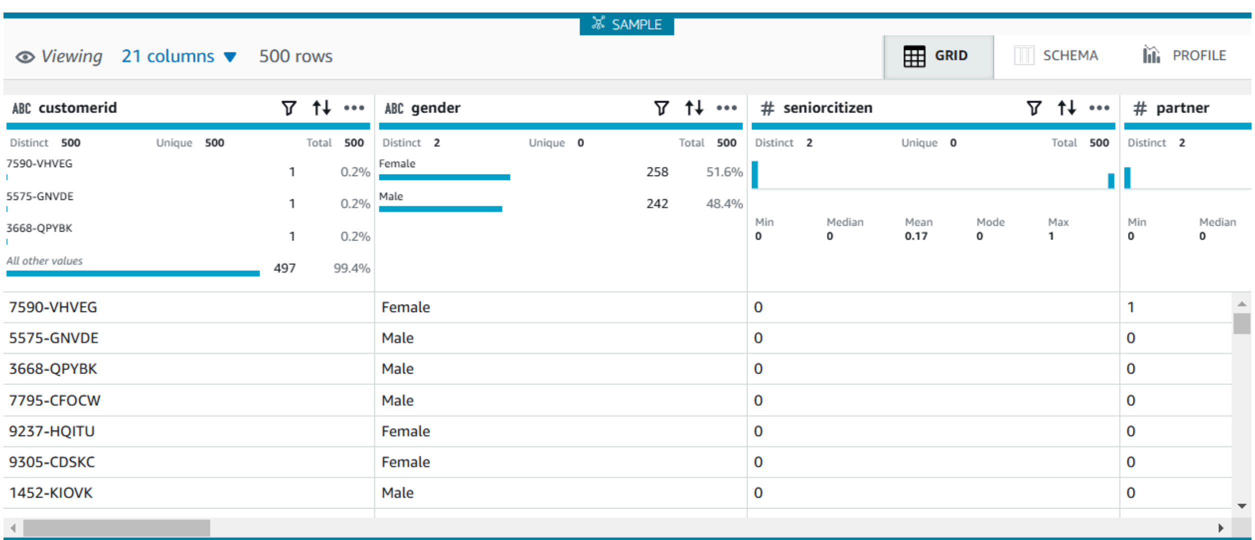
- Remove null values.
- Ensure consistent data formatting.
- Convert categorical features into structured formats.
- Save the cleaned dataset back to **S3** in **CSV format**.

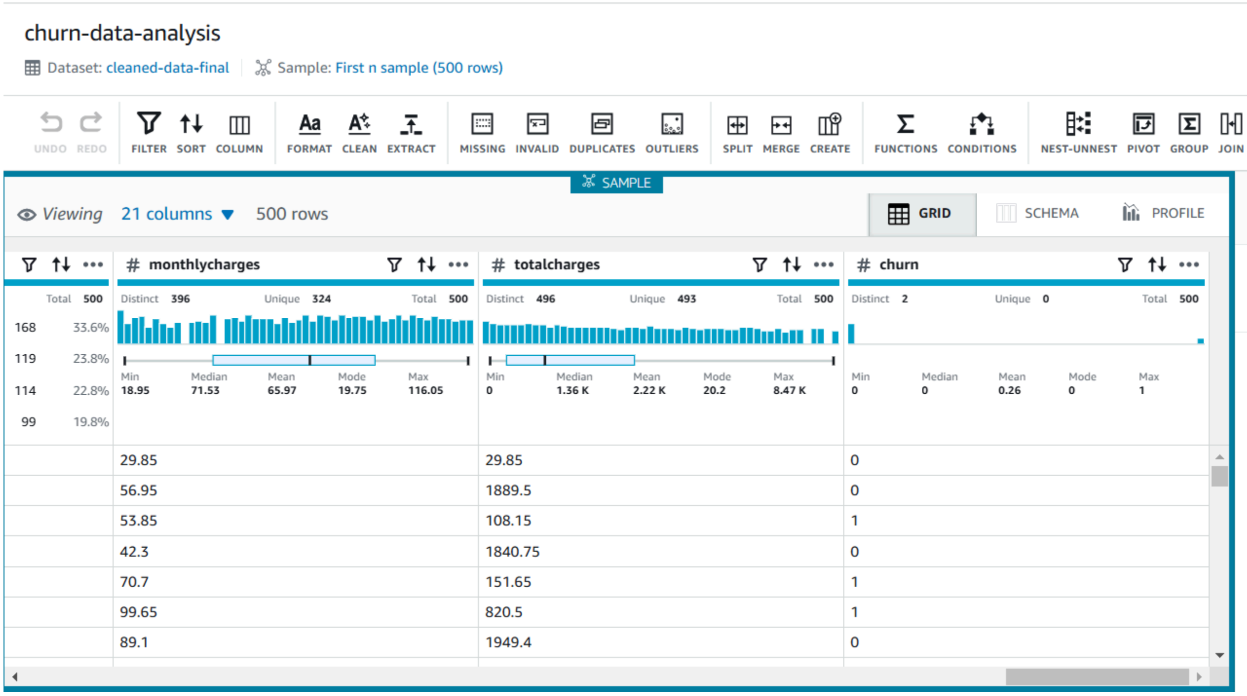


Step 4: Use AWS Glue DataBrew

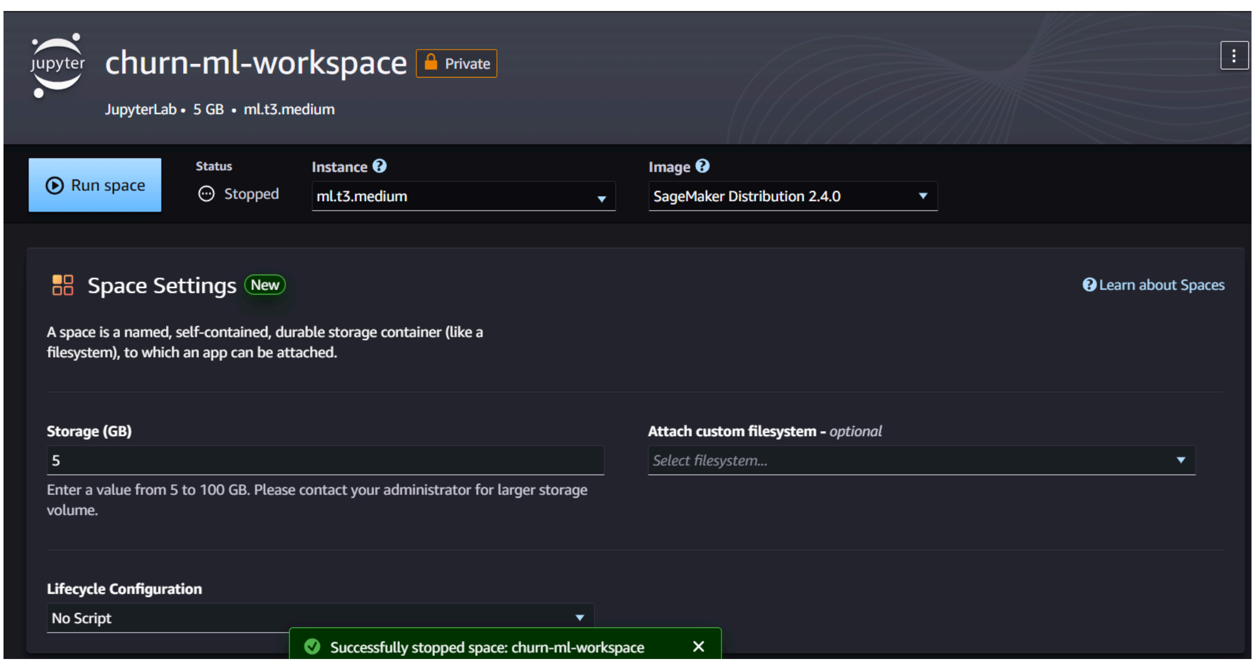
AWS Glue DataBrew was used for **data exploration and validation**, allowing for:

- Data profiling to detect **missing values, outliers, and inconsistencies**.
- Validation of data transformations applied in **AWS Glue ETL**.
- Ensuring the dataset was ready for machine learning.





Step 5: Use AWS SageMaker for Machine Learning



Model Training Process

AWS SageMaker was used to **develop and train churn prediction models**:

1. **Data Loading** – The cleaned CSV file was imported from **S3** into a SageMaker Jupyter Notebook.
2. **Data Preprocessing** – Categorical features were encoded using **one-hot encoding**.
3. **Feature Scaling** – Numeric features were **normalized** to improve model performance.
4. **Handling Class Imbalance** – **SMOTE (Synthetic Minority Over-sampling Technique)** was applied to balance churn labels.
5. **Model Selection & Training** – Four models were trained:
   - **Logistic Regression**
   - **Random Forest**
   - **XGBoost**
   - **SMOTE Logistic Regression**

The jupyter notebook can be found here:

Step 6: Test Model & Performance Analysis

Models were evaluated using **accuracy, precision, recall, and F1-score**.

Model	Accuracy	Churn Recall	Churn Precision	Churn F1-Score
Original Logistic Regression	**0.82**	0.60	0.69	0.64
Random Forest	0.79	0.47	0.64	0.54
SMOTE Logistic Regression	0.76	**0.78**	0.53	**0.63**
XGBoost	0.79	0.52	**0.64**	0.57

SMOTE Logistic Regression performed best for detecting churners with the highest **recall (0.78)**. This means it is best at identifying customers likely to churn.

XGBoost provided a balance between precision and recall, useful for reducing false alarms.

Logistic Regression (without SMOTE) had the highest accuracy (0.82) but missed many churners.

Confusion Matrix Analysis

The confusion matrix from the best model (SMOTE Logistic Regression) shows:

- **83 false negatives (missed churners)**
- **258 false positives (customers predicted to churn but stayed)**
- **290 true positives (correctly identified churners)**
- **778 true negatives (correctly identified non-churners)**

Step 7: Understanding Insights & Business Recommendations

Feature	Impact on Churn	Interpretation
Monthly Charges (10.29)	🔥 Most Influential	Higher monthly charges **increase churn risk**
Tenure (3.85)	⬇️ Lower churn	Customers with **longer tenure** are less likely to churn
Fiber Optic Internet (3.46)	⬆️ Higher churn	Customers with **fiber optic service** tend to **churn more**
Total Charges (1.71)	⬇️ Lower churn	Higher total charges = **less churn**, meaning **loyal customers stay**
Streaming Movies (1.40)	⬆️ Higher churn	Customers **with streaming services** may be **more likely to leave**
Streaming TV (1.23)	⬆️ Higher churn	Similar to Streaming Movies, users with **extra services might leave faster**
Phone Service (1.14)	⬆️ Higher churn	Customers **with phone service** might be more likely to leave
Two-Year Contract (1.11)	⬇️ Lower churn	**Longer contracts reduce churn**, since customers are locked in
Multiple Lines (0.92)	⬆️ Higher churn	**More phone lines = higher churn** (Maybe due to higher costs?)
Electronic Check (0.80)	⬆️ Higher churn	Customers paying with **electronic check** churn more (less commitment?)

Business Recommendations to Reduce Churn

Based on these insights, businesses can implement the following strategies:

1. Lower Monthly Charges for High-Risk Customers

- Offer discounts or loyalty programs for customers with high monthly bills.

2. Target Fiber Optic Internet Users for Retention

- Offer special promotions for fiber optic users who are at risk of leaving.
- Investigate why fiber optic customers churn more (pricing, service issues?).

3. Encourage Long-Term Contracts

- Customers on two-year contracts churn less, so provide:
- Discounts for yearly payments
- Free upgrades for long-term subscribers

4. **Address Payment Method Issues**

- Customers paying via electronic check churn more, so:
- Encourage auto-pay options (credit card, PayPal, etc.)
- Offer rewards for switching payment methods

5. **Optimize Streaming Services**

- Customers with streaming services churn more, so:
- Bundle streaming with internet plans
- Provide exclusive offers for streaming subscribers