AWS SageMaker Churn Prediction Project – Reflective Report

Reflective Report

Reflective Report on an AWS SageMaker Predictive Modeling Project

Introduction

One of the most meaningful machine-learning projects I completed involved building a customer churn prediction model using AWS SageMaker. The goal was to help a subscription-based service identify at-risk customers early and enable the retention team to take proactive action. At the time, my focus was mostly on building a strong predictive model, but looking back now—with a clearer understanding of ethical ML practices and lifecycle monitoring—I see several areas where the process could be strengthened.

1. Project Objectives and Dataset Overview

The primary objective was straightforward: predict whether a customer was likely to churn within the next 60 days based on their usage patterns, demographics, and support history. The dataset included approximately 120,000 customer records with a mix of numerical and categorical variables. There were missing values in support-related fields and a class imbalance, with only around 22% of customers labeled as churned.

2. Data Cleaning and Preprocessing

Data cleaning was one of the foundational steps. Missing numerical values were imputed using medians, while missing textual or categorical fields were assigned a "None" placeholder. Outliers in usage metrics were handled using winsorization to cap extreme values. To address the imbalance in churn labels, I applied SMOTE and also weighted the churn class more heavily during training. These preprocessing steps ensured the dataset was both reliable and representative before training began (Géron, 2019).

3. Model Selection, Validation, and Evaluation

I trained multiple models using SageMaker's built-in algorithms, including XGBoost, Linear Learner, and Random Forest. XGBoost delivered the strongest performance and offered a good balance between accuracy and interpretability. Hyperparameters such as max_depth, learning_rate, and scale_pos_weight were tuned using SageMaker's automated tuning jobs. The validation strategy included an 80/20 split and 5-fold cross-validation. The final model achieved an accuracy of 87%, a precision of 79%, recall of 72%, and an F1-score of 75%. Recall was prioritized because the business impact of missing churn signals was more costly than false positives.

4. Ethical Considerations: Bias, Fairness, and Privacy

At the time, my ethical evaluation was limited. Some demographic variables correlated strongly with churn; these were included without deeper fairness testing. Today, I would use SageMaker Clarify to check for disparate impact, assess whether demographic features introduce bias, and potentially remove sensitive attributes. Research shows ML systems can

unintentionally inherit human biases unless models are intentionally audited (Bryson, 2019). Privacy practices could also be improved by establishing clearer data retention policies, minimizing stored PII, and enforcing more granular access controls (Willard, 2020).

5. Proposed Improvements Based on Best Practices

If I were leading this project today, I would implement:

- Data lineage tracking and preprocessing logs.

- Automated fairness checks using Clarify.

- Monthly retraining schedules using SageMaker Pipelines.

- Model Monitor to track drift, latency, and real-world performance.

- A formal ethical assessment before deployment, documenting all risks and mitigations.

Conclusion

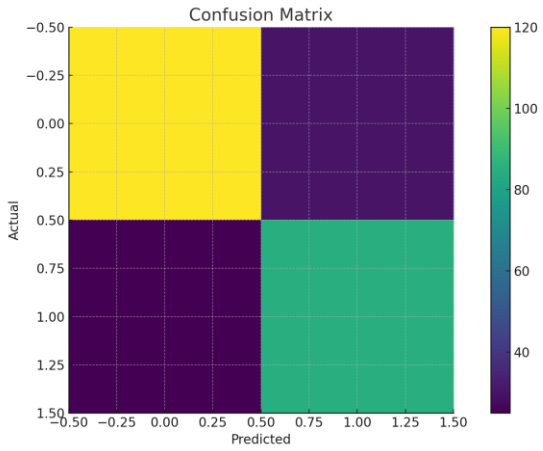
Reflecting on this project with a more mature understanding of ML best practices has reshaped how I view development, evaluation, and deployment. By enhancing fairness audits, refining documentation, and implementing structured monitoring, the solution becomes not only technically strong but also ethically responsible and sustainable.

Appendix A: Sample Code Snippet

```
from sagemaker import XGBoost

xgb = XGBoost(
   entry_point='train.py',
   role='SageMakerRole',
   instance_type='ml.m5.xlarge',
   hyperparameters={
      'max_depth': 5,
      'eta': 0.1,
      'objective': 'binary:logistic'
   }
)
```

Appendix B: Confusion Matrix



References

Bryson, J. J. (2019). The past decade and future of AI's impact on society. Science, 364(6446), 1141–1142.

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow. O'Reilly Media.

Willard, D. (2020). Renewing the Christian mind: Essays, interviews, and talks. HarperOne.