

# **Telecom Churn Prediction Using AWS SageMaker**

by

Sonu Adhikari

A Project Report Submitted for Cloud Computing  
for Data Science and Artificial Intelligence Course

Submitted To: Prof. Chantri Polpasert

Asian Institute of Technology  
School of Engineering and Technology  
Thailand  
May 2024

# CONTENTS

	Page
<b>LIST OF TABLES</b>	<b>iii</b>
<b>LIST OF FIGURES</b>	<b>iv</b>
<b>CHAPTER 1 INTRODUCTION</b>	<b>1</b>
1.1 Background of the Study	1
1.2 Statement of the Problem	1
1.3 Objectives	2
<b>CHAPTER 2 LITERATURE REVIEW</b>	<b>3</b>
<b>CHAPTER 3 METHODOLOGY</b>	<b>4</b>
3.1 Data Collection and Preprocessing	4
3.1.1 Dataset	4
3.2 Model Selection and Training	4
3.3 Evaluation and Testing	4
3.4 Model Deployment and Integration	5
<b>CHAPTER 4 AWS WELL-ARCHITECTED FRAMEWORK</b>	<b>6</b>
4.1 Security	6
4.2 Cost Optimization	6
4.3 Sustainability	6
4.4 Performance Efficiency	7
4.5 Operational Excellence	7
<b>CHAPTER 5 CONCLUSION</b>	<b>8</b>
5.1 Challenges	8
5.2 Future Work	8
<b>REFERENCES</b>	<b>10</b>

## LIST OF TABLES

Tables	Page
Table 4.1 AWS Services used in the project	7

## LIST OF FIGURES

Figures	Page
Figure 4.1 Architecture of Project	6
Figure 4.2 Churn Prediction Endpoint	7

# CHAPTER 1

## INTRODUCTION

### 1.1 Background of the Study

In the dynamic landscape of modern business, the pursuit of customer satisfaction and retention stands as a cornerstone for sustainable growth and profitability. Customer churn, the phenomenon where customers discontinue their relationship with a company, poses a significant challenge across various industries, including telecommunications, banking, e-commerce, and subscription services. Churn Customer refers to the number of existing customers who may leave the service provider over a given period. Wagh et al. (2024) The consequences of churn extend beyond mere revenue loss; they encompass diminished brand reputation, increased customer acquisition costs, and a disruption in market positioning.

Traditionally, businesses have relied on reactive measures to address churn, often resorting to generalized retention strategies or attempting to win back customers after they've already left. However, in today's data-driven era, there exists a wealth of information that can be leveraged to predict and prevent churn proactively. The advent of machine learning, coupled with cloud computing technologies, has unlocked new avenues for businesses to analyze vast datasets, derive actionable insights, and deploy predictive models in real-time.

AWS SageMaker, a fully managed machine learning service provided by Amazon Web Services (AWS), offers a robust framework for building, training, and deploying machine learning models at scale. With SageMaker, businesses can harness the power of advanced algorithms, streamline model development pipelines, and integrate predictive analytics seamlessly into their existing infrastructure.

### 1.2 Statement of the Problem

The telecommunications industry is characterized by intense competition and high customer churn rates, posing a significant challenge for mobile operators striving to maintain market share and profitability. Customer churn, the phenomenon where subscribers terminate their service agreements, not only results in immediate revenue loss but also incurs substantial costs associated with acquiring new customers to offset the attrition.

The primary challenge lies in accurately predicting churn and implementing proactive strategies to retain at-risk customers before they defect to competitors. Traditional churn prediction methods based on simple heuristics or manual analysis of historical data lack the sophistication and predictive power necessary to identify subtle patterns and early indicators of churn.

Thus, the problem at hand is to develop a robust and scalable Customer Churn Prediction System capable of leveraging advanced machine learning algorithms to analyze vast amounts of customer data and accurately forecast churn events. By harnessing the power of AWS SageMaker and cloud computing, the system aims to provide mobile operators with actionable insights to implement targeted retention strategies, reduce churn rates, and enhance customer satisfaction and loyalty.

### **1.3 Objectives**

The primary objective of this project is to develop a machine learning model using AWS SageMaker's classification algorithm to predict customer churn accurately.

- Deploy the model as a real-time prediction endpoint accessible via AWS API Gateway.
- Enable businesses to integrate churn prediction seamlessly into their operational workflows and take proactive measures to retain customers.

## **CHAPTER 2**

### **LITERATURE REVIEW**

Customer churn prediction has become a critical focus for businesses across various sectors, particularly in the telecommunications industry where intense competition and high churn rates significantly impact profitability. Burez and Van den Poel (2009) address the issue of class imbalance in churn prediction, highlighting the need for techniques that can handle the typically low churn rates observed in real-world datasets. This study underscores the importance of balancing classes to improve model performance and accuracy.

Machine learning offers a range of techniques that can be employed for churn prediction, each with its advantages and trade-offs. Verbeke, Dejaeger, Martens, Hur, and Baesens (2012) propose a profit-driven approach to churn prediction, emphasizing the integration of business objectives into the model development process. This perspective is crucial as it aligns the technical aspects of machine learning with the financial goals of the business, ensuring that the predictive models not only identify churn but also contribute to profitability.

These studies collectively underscore the significance of churn prediction in the telecommunications industry and the role of advanced machine learning techniques in addressing this challenge. By leveraging the capabilities of AWS SageMaker, this project aims to build a robust and scalable churn prediction system that empowers mobile operators to implement proactive retention strategies and reduce churn rates effectively.

## CHAPTER 3

### METHODOLOGY

#### 3.1 Data Collection and Preprocessing

Gather historical customer data, including demographic information, usage patterns, and churn labels. Perform Exploratory Data Analysis, data cleaning, feature engineering to prepare the dataset for model training.

##### 3.1.1 Dataset

The public dataset is completely available on the Maven Analytics website platform where it stores and consolidates all available datasets for analysis in the Data Playground. The specific telecom customer churn dataset at hand can be obtained in this link.

#### 3.2 Model Selection and Training

In the context of churn prediction, XGBoost (Extreme Gradient Boosting) offers a robust and scalable solution for accurately identifying customers at risk of discontinuing their service. The objective function for training the XGBoost model combines a loss function that measures the difference between the actual and predicted churn outcomes, and a regularization term that controls the model complexity to prevent overfitting. Formally, the objective function is expressed as:

$$\mathcal{L}(\Theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (3.1)$$

$$l(y, \hat{y}) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \quad (3.2)$$

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (3.3)$$

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i) \quad (3.4)$$

#### 3.3 Evaluation and Testing

Evaluate the model's performance using a separate validation dataset. Measure prediction accuracy, precision, recall, and F1-score to assess the model's effectiveness in



predicting churn.

### **3.4 Model Deployment and Integration**

Deploy the trained model as a real-time prediction endpoint on AWS SageMaker. Integrate the prediction endpoint with AWS API Gateway to enable seamless integration with the organization's CRM system.

## CHAPTER 4

### AWS WELL-ARCHITECTED FRAMEWORK

By incorporating the principles of well-architected design into the project, we can ensure that the customer churn prediction solution on AWS SageMaker is not only effective but also follows best practices for operational excellence, security, reliability, performance efficiency, and cost optimization outlined in the Well-Architected Framework. The architecture for the project is displayed in figure 4.1. Following the methodology proposed earlier, the end result is the endpoint for prediction that has been mentioned in figure 4.2

#### 4.1 Security

Security measures are ensured by utilizing IAM Roles instead of relying on the Root Account, strictly adhering to the principle of least privilege. This ensures that users and services are granted only the necessary permissions to perform their tasks, minimizing the risk of unauthorized access. Implemented API Gateway for security.

#### 4.2 Cost Optimization

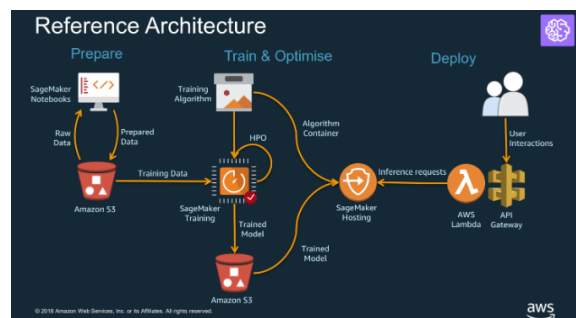
CloudWatch alarms are implemented to track costs effectively, allowing for the monitoring of AWS spending and identification of potential cost overruns. Additionally, SageMaker resource usage is continuously monitored, and instance sizes are adjusted based on workload demands to optimize resource utilization and minimize costs.

#### 4.3 Sustainability

Resource optimization is prioritized by using the minimum hardware required to meet workload demands. Additionally, the creation and maintenance of unused assets are

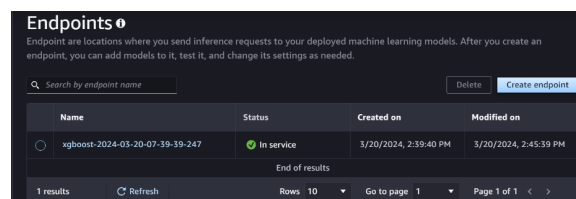
**Figure 4.1**

*Architecture of Project*



**Table 4.1***AWS Services used in the project*

AWS Service	Accomplished Task
Cloud Formation	Used for automating resources for Sagemaker project.
CloudWatch	Created a billing alarm if the cost exceeds a certain threshold.
Amazon S3	Created a bucket to store the raw dataset for Sagemaker project
IAM	Completed the project in an IAM account instead of the root account
AWS Sagemaker	Trained the XGBoost algorithm and deployed endpoint
AWS Lambda	Created a Lambda function to trigger the model endpoint
API Gateway	API Gateway passes the parameter values to the Lambda function.

**Figure 4.2***Churn Prediction Endpoint*


The screenshot shows the AWS SageMaker 'Endpoints' console. At the top, there's a search bar and buttons for 'Delete' and 'Create endpoint'. Below is a table with columns: Name, Status, Created on, and Modified on. One endpoint is listed: 'xgboost-2024-03-20-07-39-39-247' with a status of 'In service' (indicated by a green dot). The 'Created on' and 'Modified on' dates are '3/20/2024, 2:39:40 PM' and '3/20/2024, 2:45:39 PM' respectively. Below the table, it says 'End of results' and '1 results'. At the bottom, there are controls for 'Rows' (set to 10), 'Go to page' (set to 1), and 'Page 1 of 1'.

Name	Status	Created on	Modified on
xgboost-2024-03-20-07-39-39-247	In service	3/20/2024, 2:39:40 PM	3/20/2024, 2:45:39 PM

avoided, reducing unnecessary costs and minimizing environmental impact.

#### 4.4 Performance Efficiency

Appropriate SageMaker instance types and sizes are selected based on workload requirements, ensuring optimal performance while minimizing costs. Additionally, performance tuning techniques are employed to optimize model training and inference processes for improved efficiency.

#### 4.5 Operational Excellence

Infrastructure provisioning and management processes are automated using CloudFormation templates to ensure consistency, repeatability, and reliability across deployments. Furthermore, best practices for monitoring, logging, and incident response are implemented to maintain system reliability and availability.

The implementation of the AWS Services and the tasks accomplished are shown in the table 4.1.

## **CHAPTER 5**

### **CONCLUSION**

Through this project, the groundwork for a churn prediction system has been laid. Despite encountering challenges such as resource limitations and navigating AWS services complexities, these obstacles have provided valuable learning experiences. Moving forward, the aim is to enhance the project by integrating the latest AWS services, implementing end-to-end MLOps practices, and fine-tuning the model for optimal performance. This project holds promise for making a significant impact on customer churn prediction and advancing machine learning applications in the telecommunications sector using cloud resources effectively.

#### **5.1 Challenges**

In my first experience with Amazon SageMaker, several challenges were encountered. Operating under the free tier, there were limitations on resource usage, particularly the inadvertent exhaustion of the 250-hour allocation for the AWS SageMaker notebook instance due to neglecting to terminate or stop the instance when not in use. Consequently, there were issues with cost management, resulting in the suspension of the account.

Adapting to the SageMaker Studio environment, while theoretically straightforward, initially posed practical difficulties. Furthermore, difficulties arose when attempting to invoke the endpoint using API Gateway, highlighting the need for further refinement in integrating SageMaker models into operational workflows.

These challenges have provided valuable lessons in resource management, operational efficiency, and technical integration, which will inform future projects and implementations on cloud platforms effectively.

#### **5.2 Future Work**

Up to the present stage, a thorough comprehension has been attained regarding Amazon SageMaker, with a particular focus on endpoint deployment as illustrated in the aforementioned architecture diagram. Moving forward, the aim is to incorporate the latest advancements in AWS services to enhance the project's architecture, building upon the foundational framework outlined in the 2018 diagram.

The immediate objectives include the seamless integration of the endpoint into a real-

time system, ensuring that predictive insights are readily accessible within operational workflows. Additionally, the implementation of comprehensive MLOps practices is planned, encompassing the entire spectrum from model training and deployment to monitoring and iterative refinement. This approach is intended to establish a robust and sustainable machine learning pipeline.

Furthermore, efforts will be directed towards refining the endpoint and model through meticulous tuning aimed at optimizing both performance metrics and predictive accuracy. These measures are essential for achieving operational excellence and maximizing the predictive efficacy of the churn prediction system within the telecommunications sector.

These strategic initiatives underscore a commitment to effectively leverage AWS SageMaker and cloud resources, thereby advancing the project's impact on customer churn prediction and establishing a precedent for future advancements in machine learning research.

## REFERENCES

- Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3, Part 1), 4626-4636. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0957417408002121> doi: <https://doi.org/10.1016/j.eswa.2008.05.027>
- Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, 218(1), 211-229. Retrieved from <https://EconPapers.repec.org/RePEc:eee:ejores:v:218:y:2012:i:1:p:211-229>
- Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S. P., & Gawande, S. (2024). Customer churn prediction in telecom sector using machine learning techniques. *Results in Control and Optimization*, 14, 100342. Retrieved from <https://www.sciencedirect.com/science/article/pii/S2666720723001443> doi: <https://doi.org/10.1016/j.rico.2023.100342>