**DATA/MSML 650**

**Final Project: Second Progress**

**Title:** E-Commerce Predictive Analysis using AWS

**Team Members:** Tausif Khan, Varun Jain, Hsi-Chun Wang, Sidhant Chanana, Allen Mathews

**Problem Statement**

Using past sales and consumer data, we plan to predict and forecast the future consumer trends by leveraging the use of predictive machine learning regression models. To deploy these machine learning models effectively we plan on using AWS functionalities like Amazon SageMaker and AWS Lambda to set up the data pipeline for our model.

**Architecture**

Our project's architecture makes use of AWS’ managed and scalable capabilities for dashboard development and predictive analytics. An Amazon S3 bucket serves as the primary data repository, where the raw data is first stored before the data pipeline starts. S3 is perfect for both raw and processed data as it guarantees accessibility and durability. The computational resources were securely isolated using a Virtual Private Cloud (VPC), guaranteeing that all activities take place in a private and regulated network environment. Data is processed, scaled, and fed into an ML algorithm model for time-series forecasting in order to do predictive analysis using Amazon SageMaker. SageMaker makes machine learning easier by offering a stable environment that streamlines model setup, tweaking, and training.

IAM roles and rules were set up for each of the five team members in order to provide security and role-based access. Following the least privilege concept, these jobs provide certain rights to access services such as S3, SageMaker, and QuickSight. The predictive analysis's findings are saved back into S3 for further use. Finally, a thorough dashboard is created using Amazon QuickSight. Interactive visualizations are made possible by QuickSight's immediate retrieval of the processed data from S3. This modular design is appropriate for dynamic commercial contexts since it guarantees scalability, security, and ease of maintenance.

**Development**

The dataset is retrieved from the S3 bucket using an AWS SDK (boto3), into the jupyter notebook in SageMaker.We forecasted hourly sales data using an LSTM model. It begins by converting the order date to extract date and time components, creating a continuous time series indexed by order date and time. Hourly sales are calculated, scaled, and split into training and test sets. Data sequences of 24-hour intervals are created to predict the next hour’s sales, which is suitable for LSTM input. We trained the LSTM model with 50 units on the data for 20 epochs. The model forecasted sales between 1 to 5 days using a rolling window approach. The scaled forecasted sales are inverse-scaled for accurate validation metrics. Finally, we evaluated the performance from our model with error metrics, and visualized the predicted sales to assess the accuracy.

We also implement a Seasonal Autoregressive Integrated Moving Average (SARIMA) model for time-series forecasting, which focuses on predicting sales over a period of next 24 hours (multi-step ahead prediction in terms of prediction unit hour). We began the process with robust data preprocessing like creating the lagged features and scaling the data to improve model performance. Our architecture was designed to effectively extract patterns from sequential data using convolutional layers and dense layers.

**Deployment**

We deployed the machine learning models and data pipeline on AWS with a combination of services. The models, including LSTM and SARIMA, were trained on Amazon SageMaker. We have created IAM roles for each of our team members for security and access management. Our features - Lambda, SageMaker and S3 - have the permission to access the resources required for their tasks. This approach ensures security and maintains compliance by limiting access to our data. The dataset is loaded into the Amazon Aurora to run queries for different analytics that we can do in terms of several fields like region wise sales, analyzing monthly sales trends, indexing and retrieving addresses etc.

**Problems Encountered:**

One of the primary issues was establishing a smooth interface across AWS services, namely making sure that Amazon S3, SageMaker, and QuickSight were all communicating with each other under a Virtual Private Cloud (VPC). To prevent permission failures or service access problems, it was necessary to pay close attention to detail and conduct thorough testing while configuring security groups and IAM roles to provide access while upholding stringent security regulations.

Managing Consistent Data Preprocessing: It was difficult to make sure that the unprocessed data that was posted to S3 was flawlessly structured. More preprocessing procedures were needed in SageMaker due to missing data entries and inconsistent time formats. This increased the data pipeline's complexity and highlighted how crucial strong error-handling and data validation procedures are.