**A close up of a logo

Description automatically generated**

**SEP 721 – Data Analytics, Machine Learning and AI on Cloud Platforms**

**Assignment 1: Qwiklabs- 1 and 2**

**Submitted by,**

**Greeshma Gopal(gopalg)**

**ID- 400245291**

# Lab 1: Creating Models with Amazon SageMaker

# I have used qwiklab credits to access the jupyter file and the dataset.

# A screenshot of a cell phone Description automatically generated

# The dataset here is about mobile operator customers. The prediction we will be doing is the churning of the customers or in short finding the less satisfied customers and offering them better plans just so they do not switch to a different operator.

# A screenshot of a social media post Description automatically generated

# Qwiklab creates a bucket which would help us store the files

# A screenshot of a cell phone Description automatically generated

# The required libraries for modelling and training are imported initially.

# A screenshot of a cell phone Description automatically generated

# In this step, we are unzipping the dataset file to extract the data

# A screenshot of a cell phone Description automatically generated

# Reading the dataset which is assigned as a data frame in the variable churn

# A picture containing screenshot Description automatically generated

# Data visualization

# A screenshot of a social media post Description automatically generated

# A screenshot of a cell phone Description automatically generated

# A picture containing text Description automatically generated

# Removing the target variable from the data frame churn

# A screenshot of a cell phone Description automatically generated

# Visualizing the relationship between each of the features and our target variable

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

# A screenshot of a cell phone Description automatically generatedA screenshot of a cell phone Description automatically generated

# A close up of a device Description automatically generated

* Converting categorical features into numeric features

# A close up of a logo Description automatically generated

# Splitting Train and Test data

# A screenshot of a cell phone Description automatically generated

* Uploading these files to S3 bucket which qwiklabs has created by default or the lab

# A screenshot of a cell phone Description automatically generated

* Specifying the locations of the XGBoost algorithm containers

# A screenshot of a cell phone Description automatically generated

# Defining hyperparameters such as max\_depth, subsample, num\_round, eta and gamma after which the training data model fitting is done.

# A screenshot of a social media post Description automatically generated

# Creating the model and deploying it on the end point

# A screenshot of a cell phone Description automatically generated

# Setting up serializers and de-serializers for passing the data in the model behind the end point

# A screenshot of a cell phone Description automatically generated

# Writing the function to loop the test data. The final predictions are converted into CSV format

# A screenshot of a cell phone Description automatically generated

# Predicting whether the customer churned 1 or 0, which is a confusion matrix

# A screenshot of a cell phone Description automatically generated

# Plotting the values of the predictions

# A screenshot of a cell phone Description automatically generated

# Minimizing the cost function. The chart indicates how picking a threshold too low results in costs skyrocketing as all customers are given a retention incentive. Meanwhile, setting the threshold too high results in too many lost customers, which ultimately grows to be nearly as costly.

# A picture containing screenshot Description automatically generated

# Lab 2: Applied Machine Learning: Building Models for an Amazon Use Case

# The data set here is related to IMDB movie ratings, and we are supposed to predict if a given movie will be nominated or winning on the upcoming award season.

# Importing the necessary libraries

# A screenshot of a cell phone Description automatically generated

# Loading data into the bucket which was created

# A screenshot of a cell phone Description automatically generated

# Visualizing the data related to genres, ratings, awards, releases etc.

# A screenshot of a social media post Description automatically generated

# Merging the data from three different data sets

# A close up of a logo Description automatically generated

# Dropping the target variable from the data frame

# A screenshot of a cell phone Description automatically generated

# Uploading the raw data into S3 bucket

# A screenshot of a cell phone Description automatically generated

# Viewing the first 30 records of the data frame

# A screenshot of a cell phone Description automatically generated

# Data visualization using describe function

# A screenshot of a cell phone Description automatically generated

# Loading the pickle file into padas data frame and dropping some features

# A screenshot of a cell phone Description automatically generated

# Display tables with runtimes as \N and also display tables with year as \N

# A screenshot of a cell phone Description automatically generated

# A separate column called nomination\_winner added which would be either 0 or 1

# A screenshot of a social media post Description automatically generated

# Filling the missing values using fillna

# A screenshot of a cell phone Description automatically generated

# Creating data frame based on run time minutes

# A screenshot of a social media post Description automatically generated

# Viewing the optimized data

# A screenshot of a cell phone Description automatically generated

# Setting flags for feature selection

# A screenshot of a social media post Description automatically generated

# Exploring the data by plotting

# A screenshot of a cell phone Description automatically generated

# Plot rating vs rating count & Histogram of Rating

# A picture containing crossword puzzle Description automatically generatedA screenshot of a cell phone Description automatically generated

# Feature selection

# A screenshot of a cell phone Description automatically generated

# PCA for reducing to 3 main features

# A screenshot of a social media post Description automatically generated

# 

# Train and Test data split

# A picture containing screenshot Description automatically generated

# Fitting Logistic regression model

# A screenshot of a cell phone Description automatically generated

# Fitting Support vector machine

# A screenshot of a cell phone Description automatically generated

# Fitting Decision Tree model

# A screenshot of a cell phone Description automatically generated

# Fitting Random forest model

# A screenshot of a cell phone Description automatically generated

# Fitting Gaussian model

# A screenshot of a cell phone Description automatically generated

# Fitting neural network

# A screenshot of a cell phone Description automatically generated

# Viewing the metrics of every model which was fitted to determine the right model

# A close up of text on a white surface Description automatically generated

# A close up of text on a white surface Description automatically generated

# A close up of text on a white surface Description automatically generated

# By Comparing the different models, I found that gradient boosting classifier was showing a better accuracy and roc value. Hence, I ran the same model against the 2005 data and below were the results of the same.

# A close up of a mans face Description automatically generatedA close up of a piece of paper Description automatically generated