AWS ML

Todays’ Agenda

- Amazon Lex
- Amazon Polly
- Amazon Q Developer (Chatbot)
- Amazon Rekognition
- Amazon SageMaker

Amazon Polly

- Turn text into lifelike speech using deep learning
- Allowing you to create applications that talk
- SSML - Speech Synthesis Markup Language
- Voice Engine: Generative, long-form, neural and Standard

Amazon Lex

- Build chatbots quickly for your applications using voice and text
  - Example: a chatbot that allows your customers to order pizzas or book a hotel
- Supports multiple languages
- Integration with AWS Lambda, Connect, Comprehend, Kendra
- The bot automatically understands the user intent to invoke the correct
- Lambda function to “fulfill the intent”
- The bot will ask for ”Slots” (input parameters) if necessary

Amazon Q Developer (Chatbot)

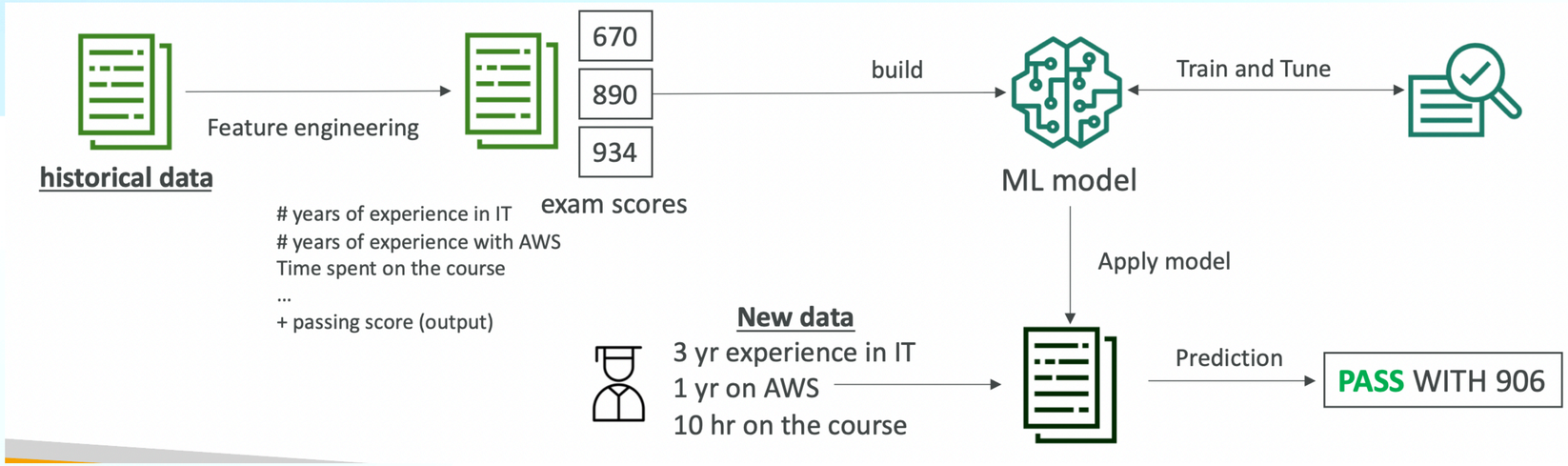
- Build chatbots quickly for your applications using voice and text
- Answer questions about the AWS documentation and AWS service selection
- Answer questions about resources in your AWS account
- Suggest CLI (Command Line Interface) to run to make changes to your account
- Helps you do bill analysis, resolve errors, troubleshooting
- AWS Chatbot is a way for you to deploy an AWS Chatbot in a Slack or Microsoft Teams channel that

Amazon Rekognition

- Find objects, people, text, scenes in images and videos using ML
- Facial Analysis and Facial Search to do user verification, people counting
- Create a database of “familiar faces” or compare against celebrities
- Use cases:
  - Labeling
  - Content Moderation
  - Text Detection
  - Face Detection and Analysis (gender, age range, emotions…)
  - Face Search and Verification
  - Celebrity Recognition

Amazon SageMaker

- Fully managed service for developers / data scientists to build ML models
- Typically, difficult to do all the processes in one place + provision servers
- Example: predicting your AWS exam score
-



Amazon SageMaker

- End to End ML service
- Collect and prepare data
- Build and train machine learning models
- Deploy the models and monitor the performance of the predictions
- Supervised Algorithms
- Unsupervised Algorithms
- Textual and Image processing algorithms

Amazon SageMaker Model Deployment comparison

Inference Type	Latency	Payload Size	Processing Time	Use Case
Real-time Inference	Low (milliseconds to seconds)	Up to 6 MB (one record)	Max 60 seconds	Fast, near-instant predictions for web/mobile apps
Serverless Inference	Low (milliseconds to seconds)	Up to 4 MB (one record)	Max 60 seconds	Sporadic, short-term inference without infrastructure, can tolerate cold starts
Asynchronous Inference	Medium to High “near real-time”	Up to 1 GB (one record)	Max 1 hour	Large payloads and workloads requiring longer processing times
Batch Transform	High (minutes to hours)	Up to 100 MB per invocation (per mini batch)	Max 1 hour	Bulk processing for large datasets Concurrent processing

Amazon SageMaker Studio

- End-to-end ML development from a unified interface
- Team collaboration
- Tune and debug ML models
- Deploy ML models
- Automated workflows

Amazon SageMaker Other Features

- Data Wrangler - Prepare Tabular data and image data for machine learning
- Feature Store - Can ingest features from a variety of sources
- SageMaker Clarify
- SageMaker GroundTruth
- ML Governance
- SageMaker Pipelines
- SageMaker Jumpstart

QUIZ TIME

Q & A