DGA or Benign?

Aayush S.
Prachi S.
Shikhar S.
Subhiksha M.
Vaibhav R.

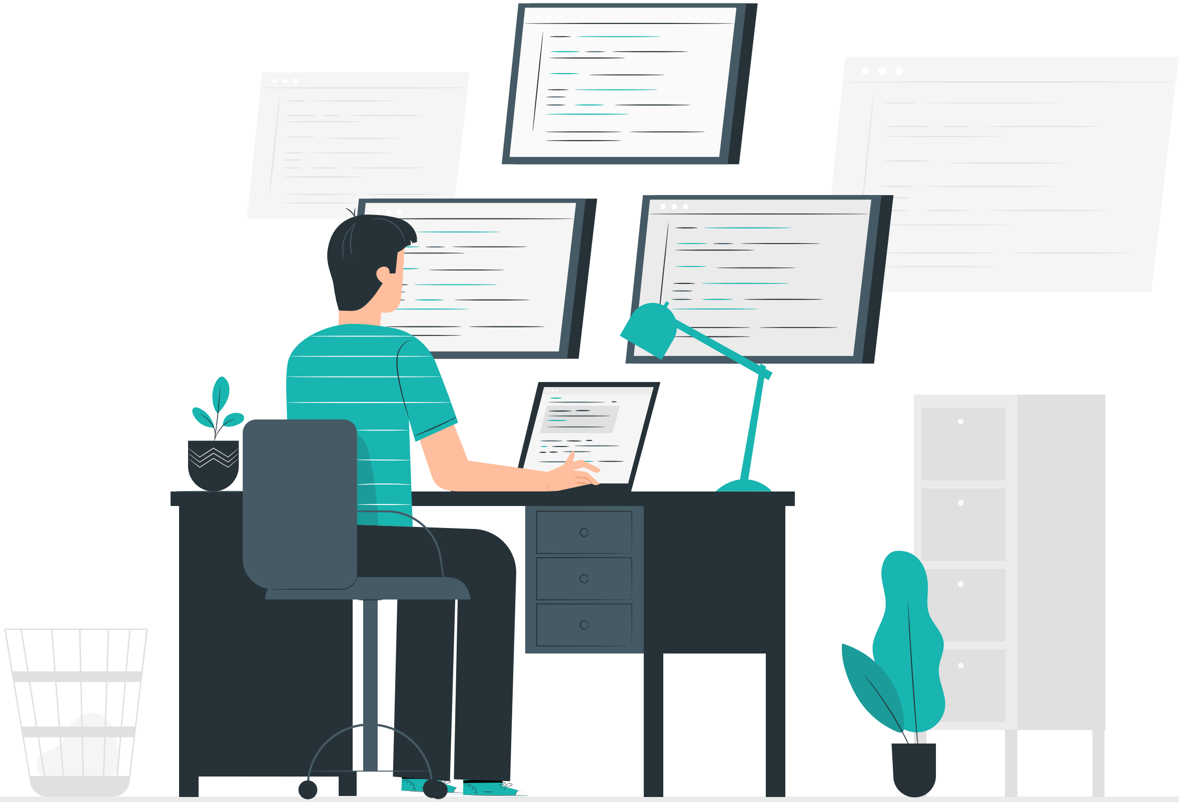


TABLE OF CONTENTS

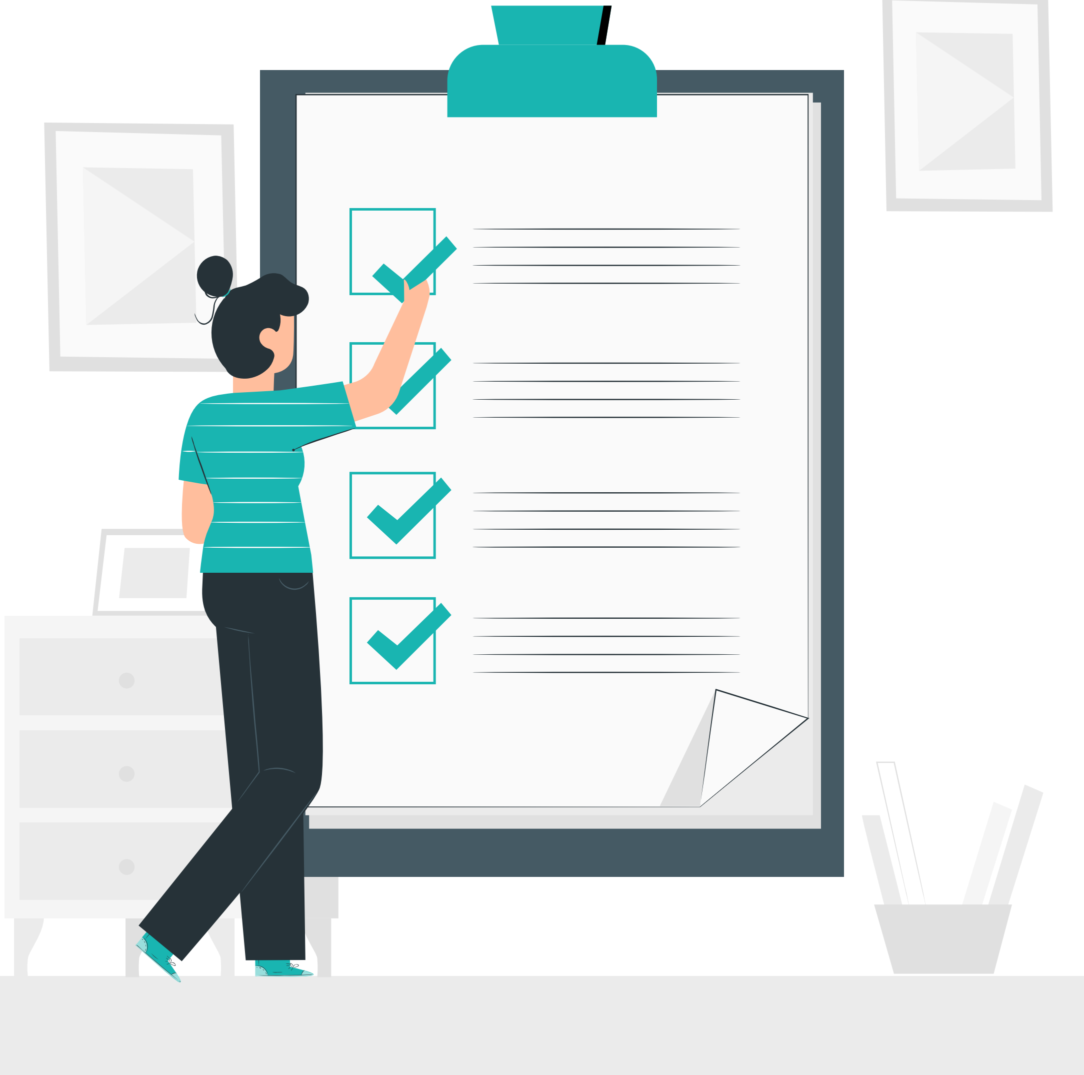
PROBLEM STATEMENT

Create a product that predicts if a FQDN or URL (e.g. www.google.com) is DGA or not. The product must go live in 10 weeks, and must be presented to your customers as a RESTful web service API. Some features to include in your product:

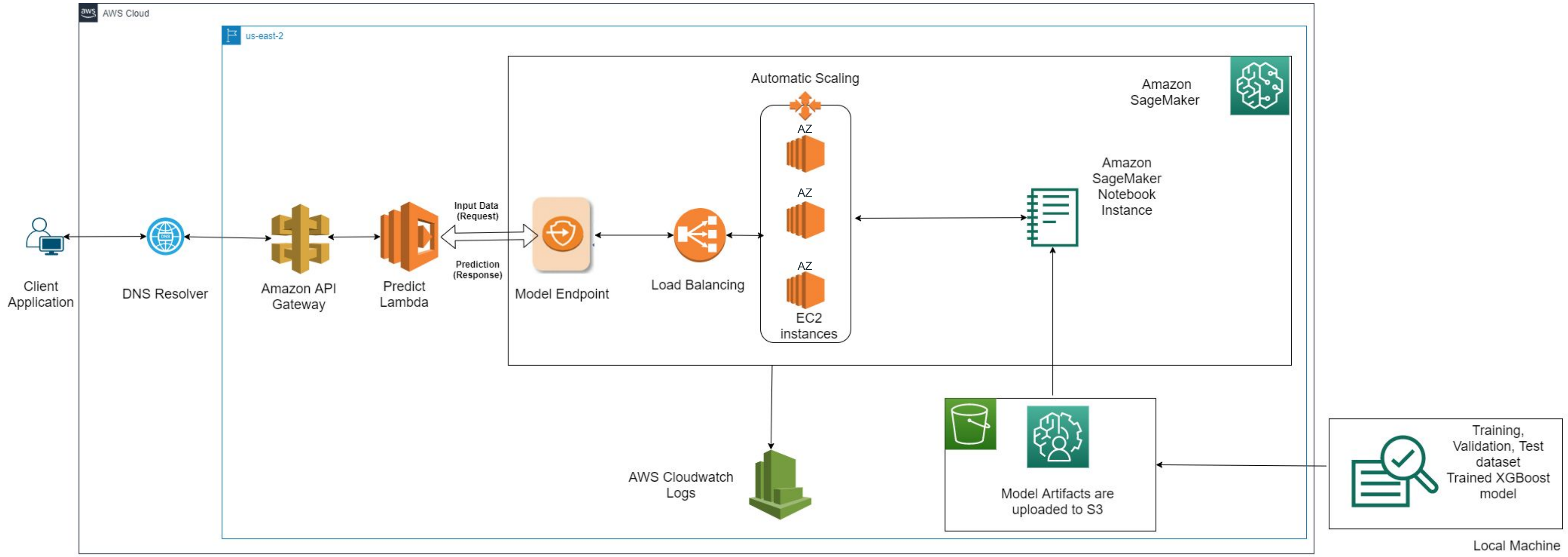
- Should be able to process 1,000,000 predictions per minute
- System uptime must be 99.999% or greater
- When passed an FQDN, your product should be able to return a TRUE/FALSE corresponding to whether the FQDN is DGA or NOT

You must collect and label your own training dataset, select your model, train and deploy it on AWS.

SYSTEM ARCHITECTURE



SYSTEM ARCHITECTURE



TOOLS & TECHNOLOGIES USED

API Gateway

AWS Lambda

Endpoint

XGBoost

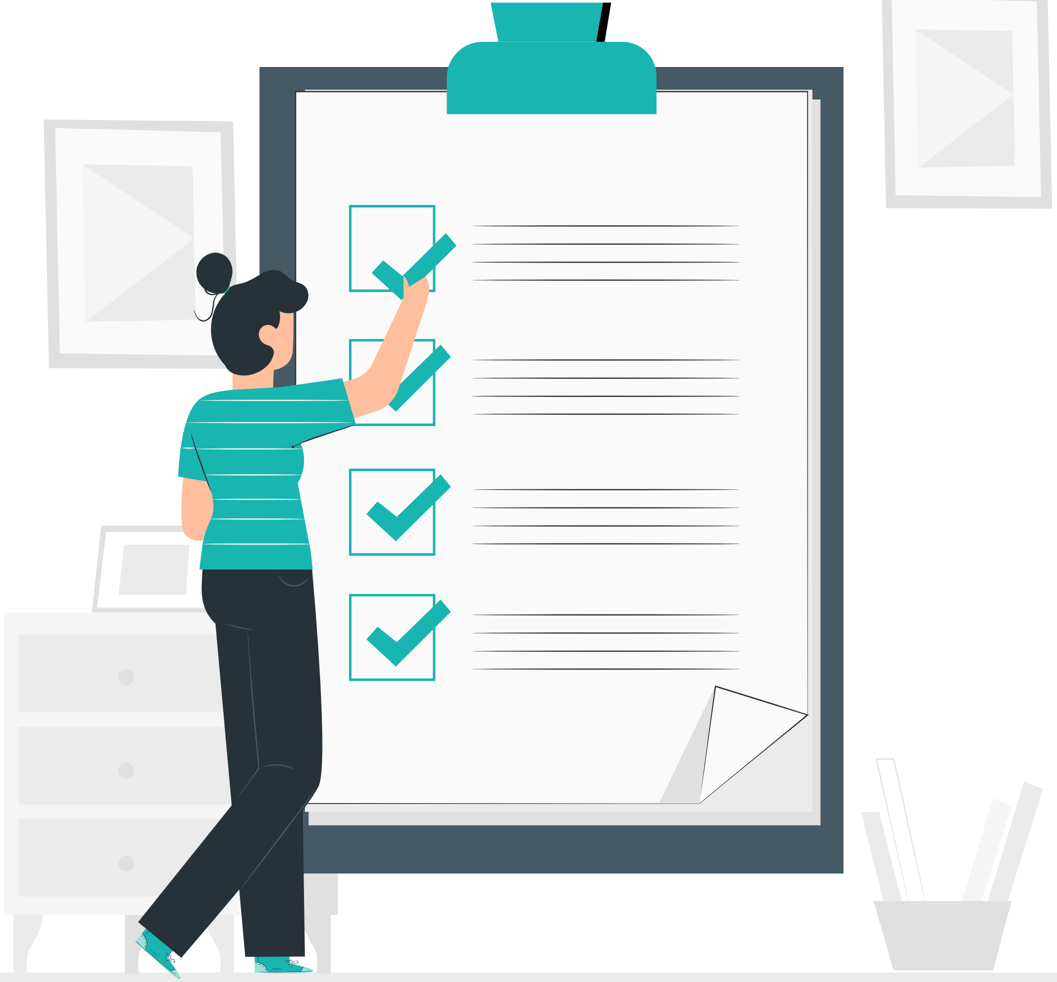
Machine Learning Model

Load Balancer

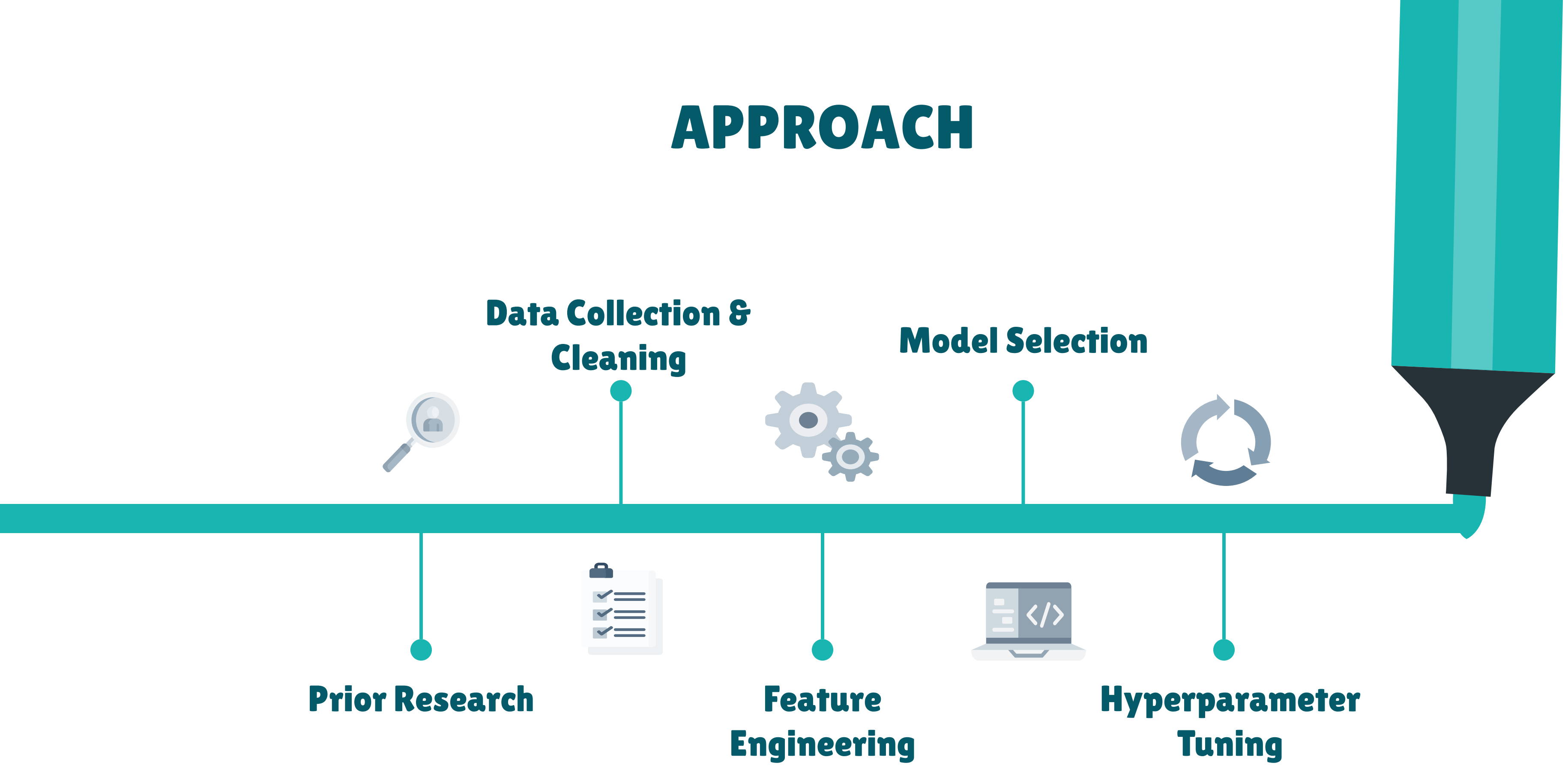
Sagemaker

Cloudwatch Logs

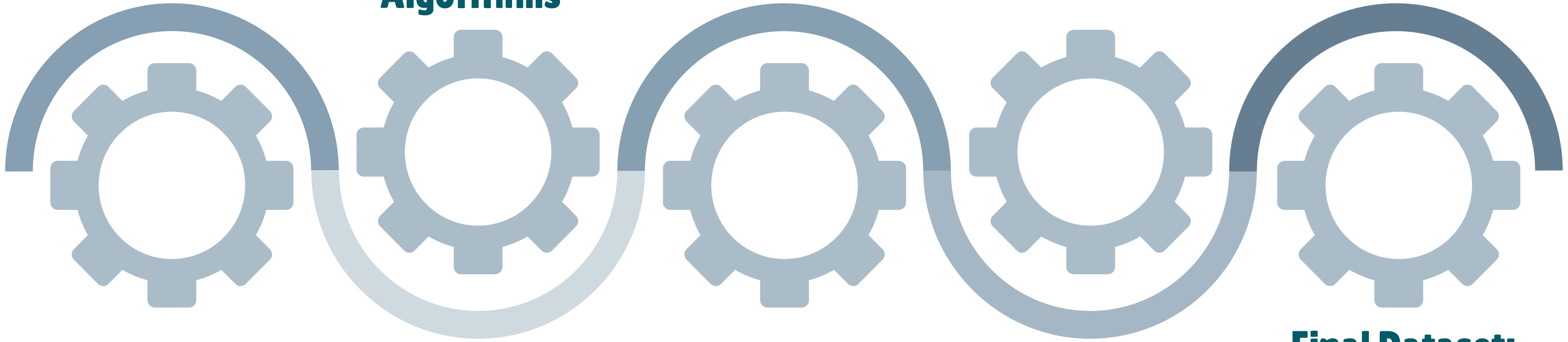
APPROACH



APPROACH


Prior Research
Data Collection & Cleaning
Feature Engineering
Model Selection
Hyperparameter Tuning

DATA COLLECTION/CREATION AND CLEANING



Benign Data

Alexa 1M,
Majestic Million, etc.

40+ DGA Algorithms

Banjori, Krakenv2, Locky, Monero,
Mydoom, Nymaim, Padcrypt, etc.

TLDExtract

Extracted Domain name

Removed Duplicates

dataset.drop_duplicates()

Final Dataset: 4M +

FEATURE ENGINEERING

Count of capitals

Convert domains to lower-case

Length of domains

Count of digits

Count of consecutive consonants

Ratio of count of vowels to length

Calculate Entropy

Count of unique characters

Convert domain characters to numbers

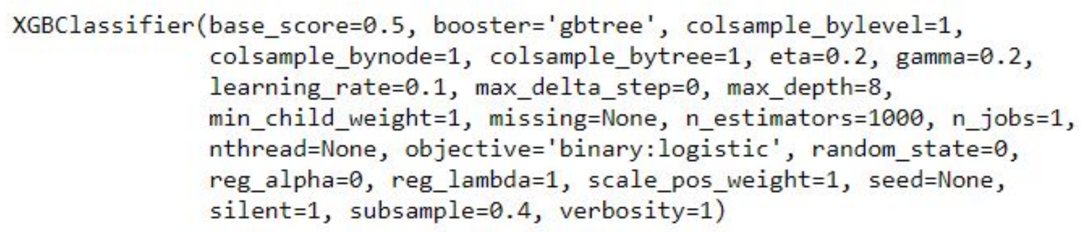
MODEL SELECTION

- We implemented the following classification algorithms:
  - Logistic Regression
  - Support Vector Machines
  - Decision Trees
  - Random Forest
  - Long Short Term Memory (LSTM)
  - XGBoost
- Amongst these, we decided to move forward with XGBoost for the following reasons:
  - Easier to deploy on AWS
  - Greater speed and efficiency compared to other algorithms
  - Better accuracy than LSTM and Random Forest

HYPERPARAMETER TUNING

Following are the parameters which we tuned (values for best model **bolded**):

- Booster - **gbtree**, gblinear
- Eta - 0.3, **0.2**
- Gamma - 0, **0.2**
- Max Depth - 6, **8**
- Reg_lambda - **1**
- Reg_alpha - **0**
- Objective - **binary: logistic**, binary: hinge, binary: logitraw
- N-estimators - 100, 500, **1000**



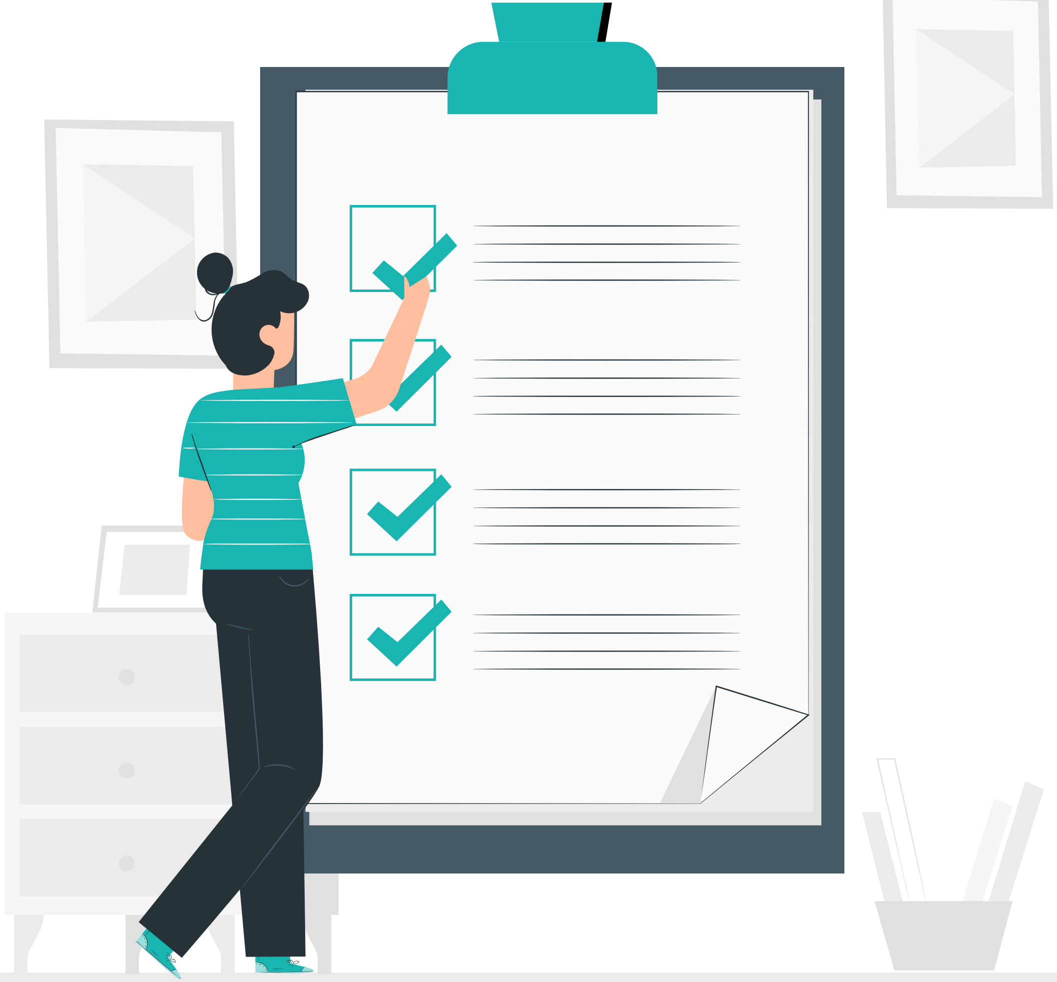
```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=1, eta=0.2, gamma=0.2,
              learning_rate=0.1, max_delta_step=0, max_depth=8,
              min_child_weight=1, missing=None, n_estimators=1000, n_jobs=1,
              nthread=None, objective='binary:logistic', random_state=0,
              reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
              silent=1, subsample=0.4, verbosity=1)
```

Accuracy of Best Model

Training: **95.12%**

Test: **93.16%**

LIKES/DISLIKES FOR EACH TECHNOLOGY



AWS S3

Likes

Per-file permission system (make a file public or not)

Centralized location for all the data

Cost effective if properly monitored and maintained

Dislikes

Renaming bucket

No direct way to upload .zip file

Drop in upload speed for large files

AWS SAGEMAKER

Likes

Provides Jupyter notebook instances for development

Charged only for what is used

Rich marketplace for existing models

Multiple servers for Training

Dislikes

Expensive

Difficult to customize

AWS LAMBDA

Likes

Flexible, supports many programming languages.

It's all configurable and easy to maintain

Rapid execution of code

Integration with other AWS tools

Dislikes

Convoluted Documentation

Importing of libraries is complex

AWS API GATEWAY

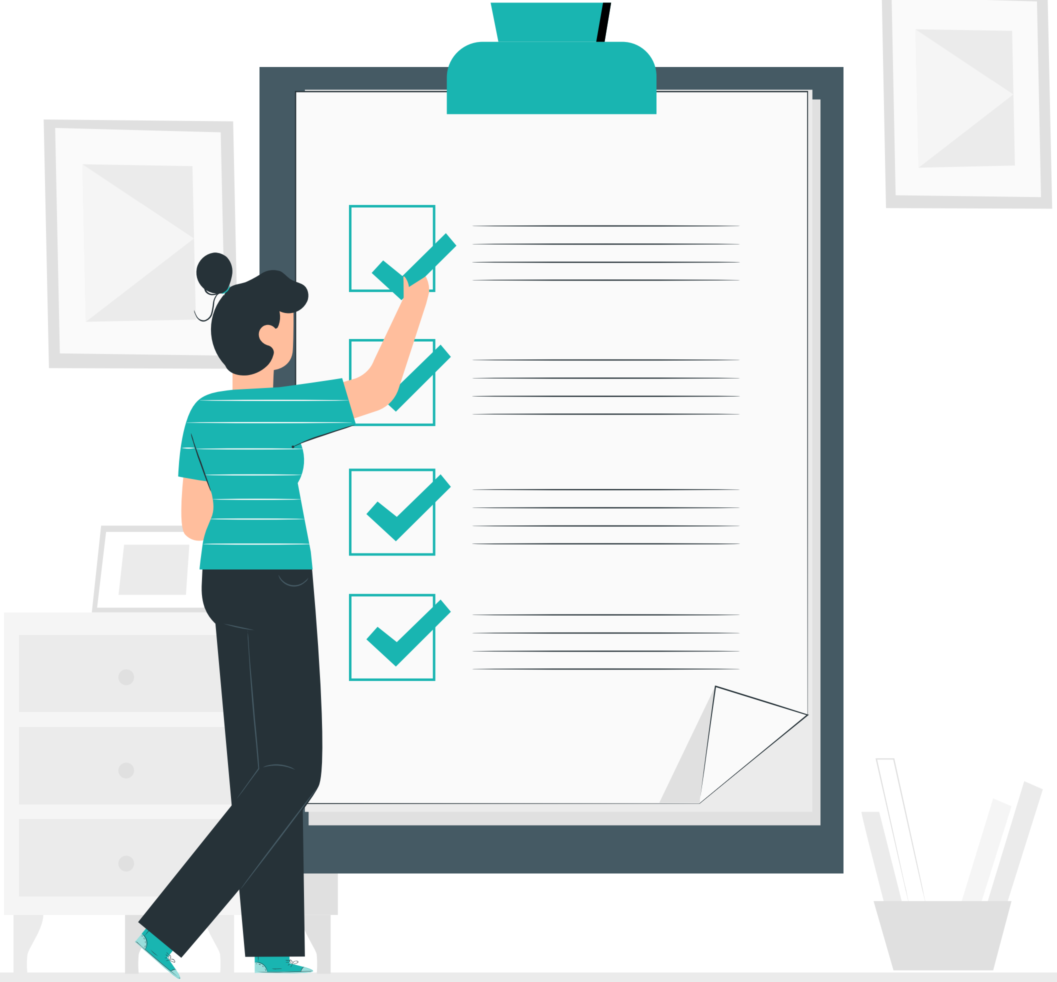
Likes

- Easy to change headers
- Very fast deployment
- Integrates well with AWS Lambda

Dislikes

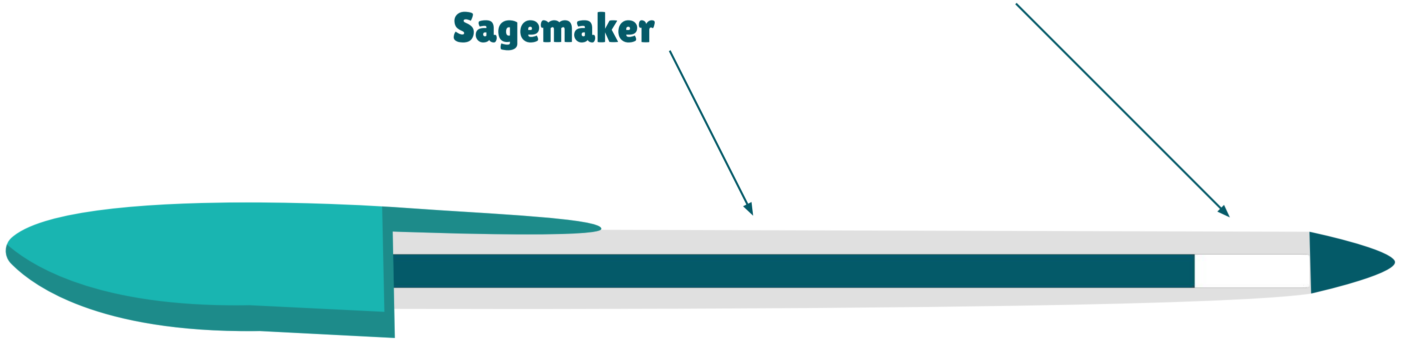
- No reminder to deploy the API
- Unable to avail through AWS Educate

BUDGET



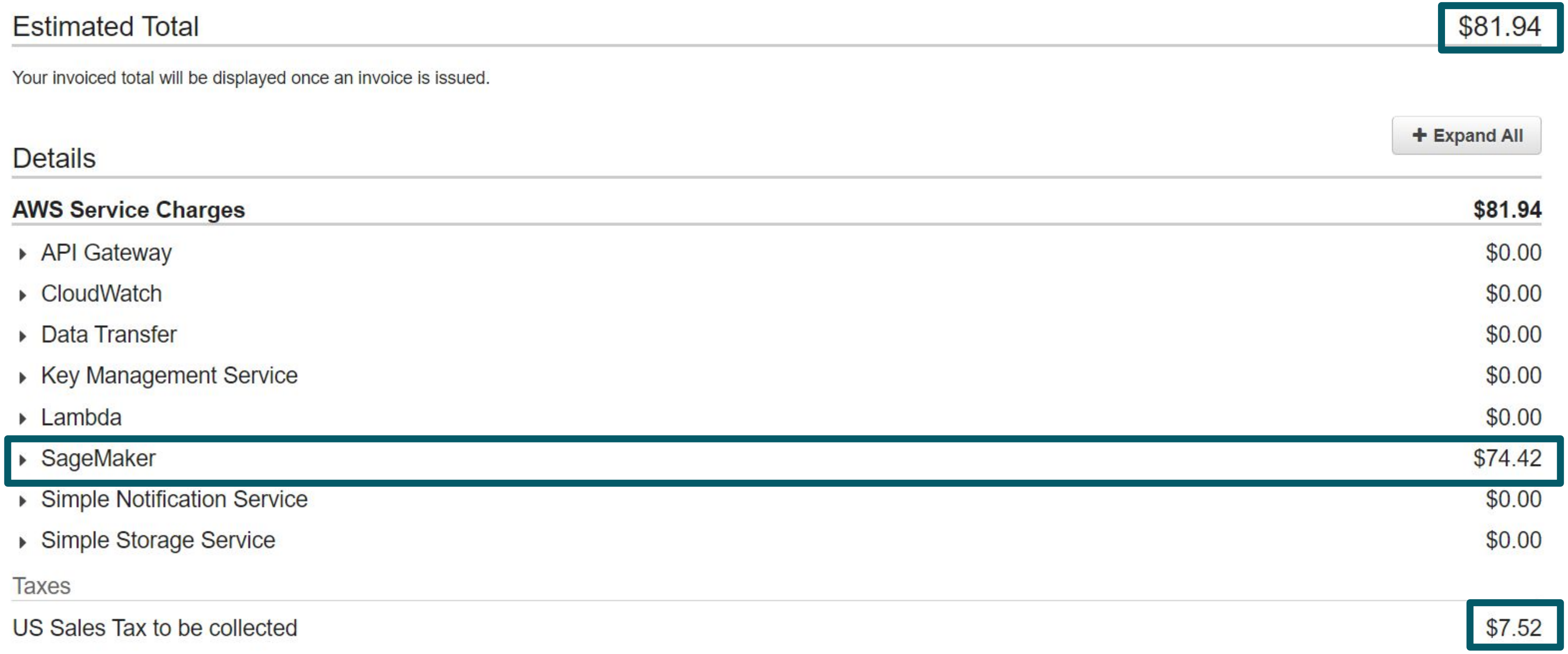
DEVELOPMENT COSTS

90%	10%
■	□
AWS Sagemaker	Sales Tax



Total: $81.94

DEVELOPMENT COSTS



Estimated Total	$81.94

Your invoiced total will be displayed once an invoice is issued.

+ Expand All

Details

AWS Service Charges	**$81.94**
API Gateway	$0.00
CloudWatch	$0.00
Data Transfer	$0.00
Key Management Service	$0.00
Lambda	$0.00
SageMaker	$74.42
Simple Notification Service	$0.00
Simple Storage Service	$0.00

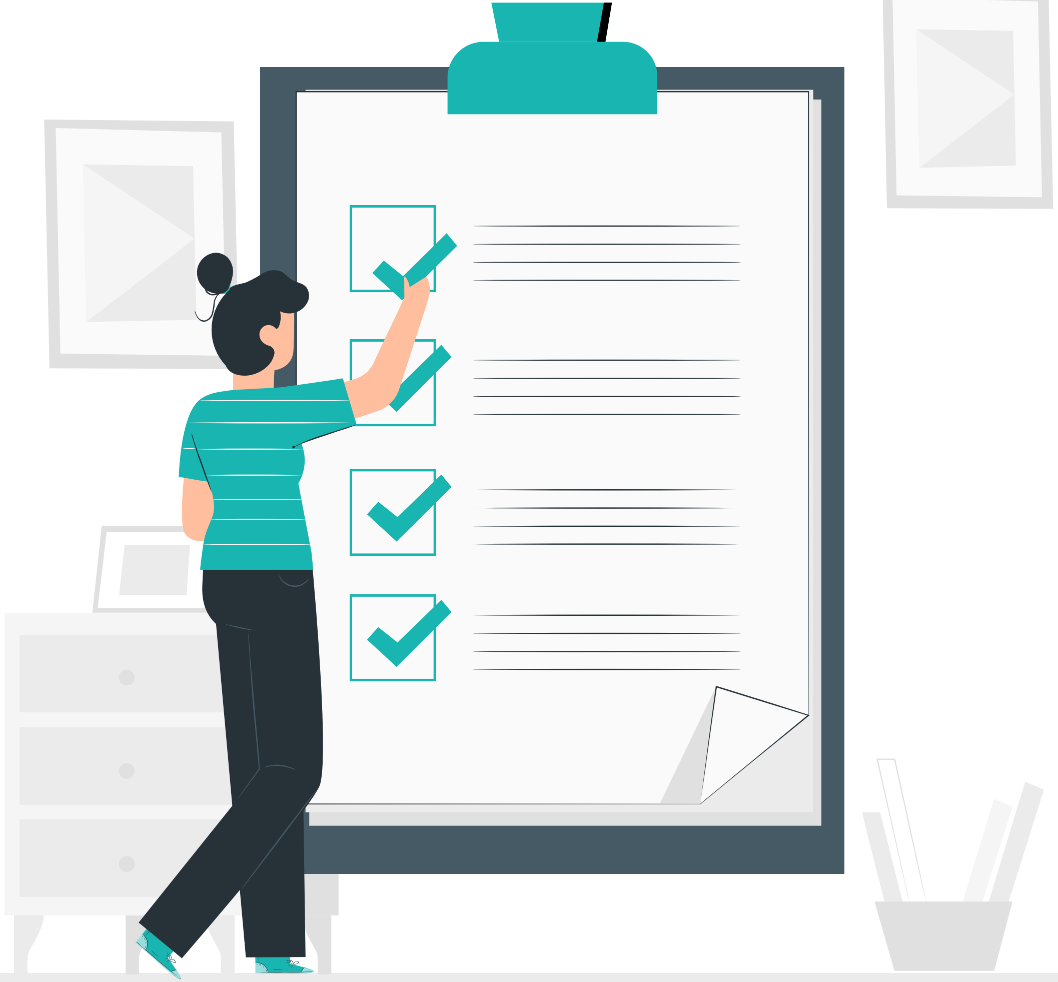
Taxes

US Sales Tax to be collected	$7.52

OPERATION COSTS FOR 24 HRS

AWS Service	Estimated Usage/Size	Cost/hr	Cost/24 hrs
Lambda	128 MB	$0.0000822351	$0.0019736425
API Gateway	10 M calls/day	1 $/million calls	$10.00
S3	<1 GB	$0.0000031944	$0.0007666666
Sagemaker	ml.m4.xlarge	$0.28	$6.72
Grand Total:			$16.7227423091

WHAT WOULD YOU DO DIFFERENTLY?



WHAT WOULD WE DO DIFFERENTLY NEXT TIME AROUND?

Training Dataset

Scale dataset. Include more DGA algorithms.

Features

Use feature scaling. Try n-gram features.

Model Selection

Apply LSTM in AWS. Use SageMaker Autopilot.

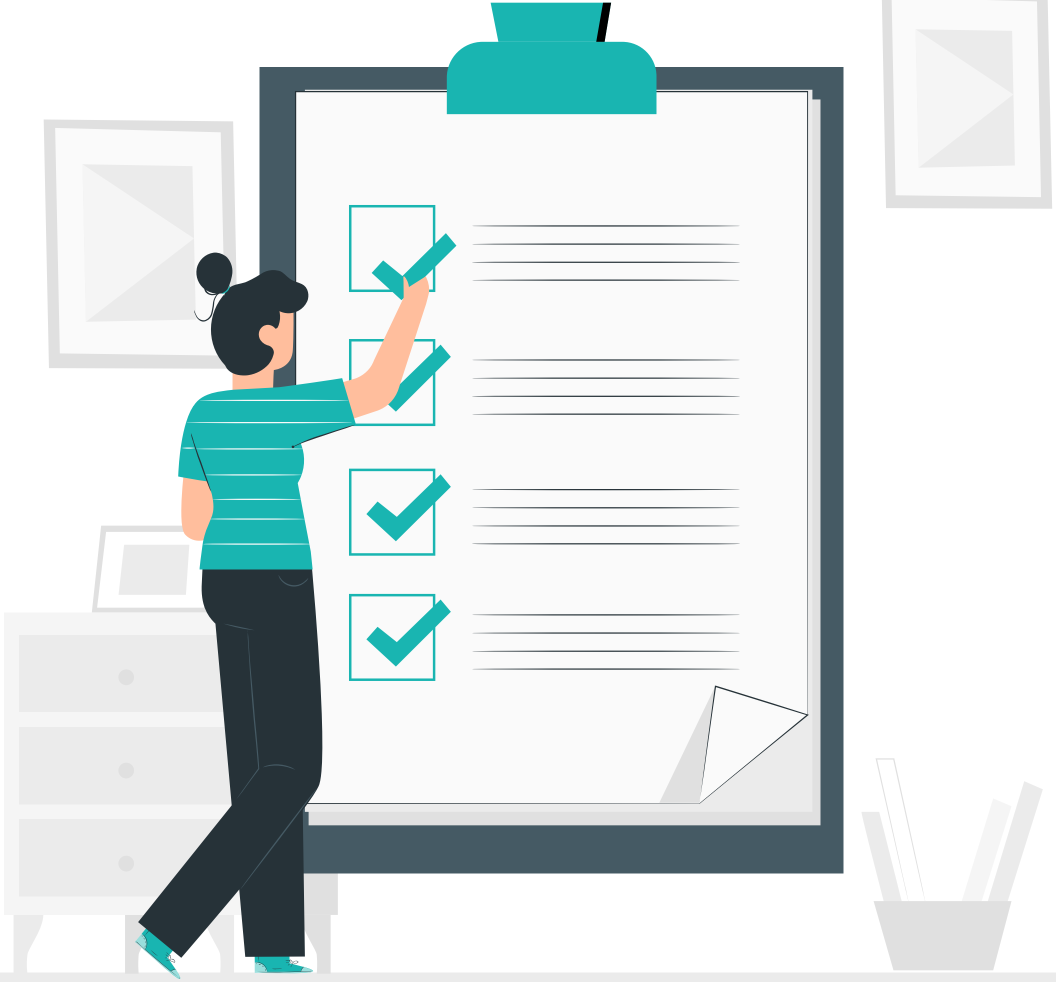
Hyperparameter Tuning

Use GridSearch to find optimum model parameters.

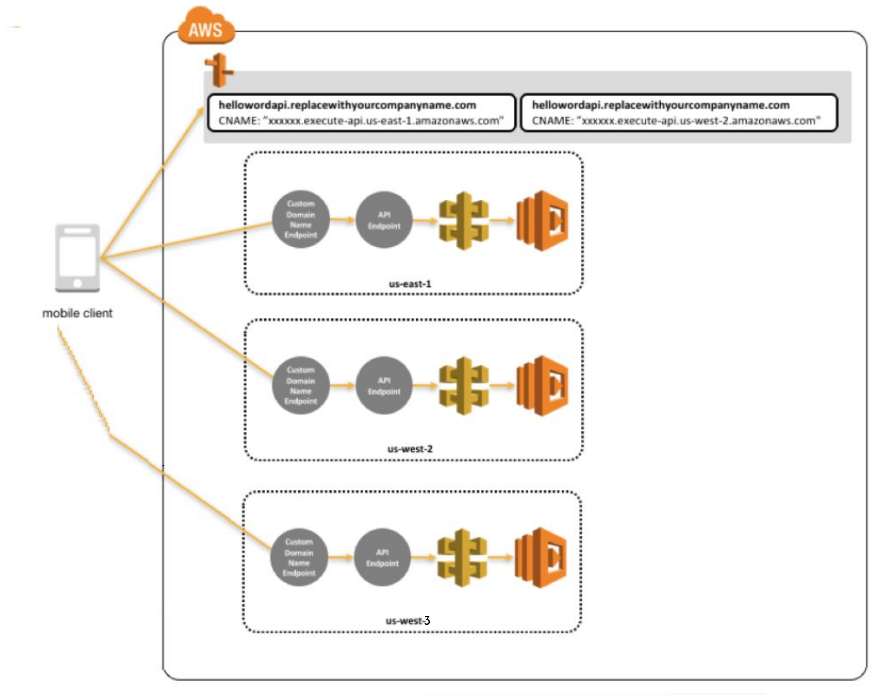
Analytics and Reporting

Use AWS EMR, Kinesis for large scale analytics.

SCALING



SCALING TO OTHER REGIONS


AWS
hellowordapi.replacewithyourcompanyname.com
CNAME: "xxxxxx.execute-api.us-east-1.amazonaws.com"
hellowordapi.replacewithyourcompanyname.com
CNAME: "xxxxxx.execute-api.us-west-2.amazonaws.com"
mobile client
Custom Domain Name Endpoint
API Endpoint
us-east-1
Custom Domain Name Endpoint
API Endpoint
us-west-2
Custom Domain Name Endpoint
API Endpoint
us-west-3

Route53

THANK YOU