Python and AWS

CSCI 447/547 MACHINE LEARNING
Outline

- Data
  - Buckets
- Processing
  - Running an ML Algorithm
  - Creating a Notebook
  - Editing the Notebook
    - Python
- Output
  - Evaluating Results
Getting Started in AWS
Services

- Compute
  - EC2
  - Lightsail
  - ECR
  - ECS
  - EKS
  - Lambda
  - Batch
  - Elastic Beanstalk

- Storage
  - S3
  - EFS
  - FSx
  - S3 Glacier
  - Storage Gateway
  - AWS Backup

- Machine Learning
  - Amazon SageMaker
  - Amazon Comprehend
  - AWS DeepLens
  - Amazon Lex
  - Machine Learning
  - Amazon Polly
  - Rekognition
  - Amazon Transcribe
  - Amazon Translate
  - Amazon Personalize
  - Amazon Forecast
  - Amazon Textract

- Analytics
  - Athena
  - EMR
Amazon Machine Learning

Amazon Machine Learning makes it easy for developers of all skill levels to use machine learning (ML) technology. Amazon Machine Learning is a managed service for building ML models and generating predictions that enable the development of robust, scalable smart applications.
Get started with Amazon Machine Learning

Standard setup
Start creating your first ML model. If you don't have your data ready, you can use our sample dataset.
Amazon Machine Learning Tutorial

Dashboard
Skip straight to the Amazon Machine Learning dashboard.
Machine Learning – Easy Path

Input data

The first step to create an ML model is to show Amazon ML your historical data. This data must include the correct answers to the questions that you want the ML model to answer. Amazon ML will create a training datasource object containing statistics about your training data.

Just trying out Amazon ML and don’t have your data ready? Use s3://aml-sample-data/banking.csv. This dataset contains information about customers as well as descriptions of their behavior in response to previous marketing contacts. You use this data to identify which customers are most likely to subscribe to your new product.

You can preview the file here: banking.csv

Want a more guided experience? Start with the Amazon Machine Learning Tutorial.

Import your data to create an Amazon ML datasource. Amazon ML can use your datasource to create and evaluate an ML model, and you can use the datasource to review your data.

Where is your data? S3

S3 data access

Tell Amazon ML how to access your data and give it permission to access it.

S3 location

Enter the path to a single file or folder in Amazon S3. You need to grant Amazon ML permission to read this data. Learn more.

If you already have a schema for this data, provide it in a file at s3://path-of-input-data-schema. If you don’t have a schema, Amazon ML will help you create one on the next page.

Datasource name

Required: Reset

Cancel Verify
Machine Learning – Easy Path

Datasource name: Banking.csv

The validation is successful. To go to the next step, choose Continue.

Datasource name: Banking.csv
Data location: s3://aml-sample-data/banking.csv
Data format: CSV
Schema source: s3://aml-sample-data/banking.csv.schema
Number of files: 1
Total size: 4.7 MB

* Required
Reset
Cancel
Continue
Machine Learning – Easy Path

### Schema

Amazon ML scanned your input data and inferred the column names and data type for each of the columns in your dataset. Review and edit the data type for each column to ensure that it accurately represents the data. This enables Amazon ML to read the input data correctly and to produce accurate predictions. [Learn more]

#### ACTION: Change type

<table>
<thead>
<tr>
<th>Name</th>
<th>Data type</th>
<th>Sample field value 1</th>
<th>Sample field value 2</th>
<th>Sample field value 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>Numeric</td>
<td>age</td>
<td>44</td>
<td>53</td>
</tr>
<tr>
<td>job</td>
<td>Categorical</td>
<td>job</td>
<td>blue-collar</td>
<td>technician</td>
</tr>
<tr>
<td>marital</td>
<td>Categorical</td>
<td>marital</td>
<td>married</td>
<td>married</td>
</tr>
<tr>
<td>education</td>
<td>Categorical</td>
<td>education</td>
<td>basic.4y</td>
<td>unknown</td>
</tr>
<tr>
<td>default</td>
<td>Categorical</td>
<td>default</td>
<td>unknown</td>
<td>no</td>
</tr>
<tr>
<td>housing</td>
<td>Categorical</td>
<td>housing</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>loan</td>
<td>Categorical</td>
<td>loan</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>contact</td>
<td>Categorical</td>
<td>contact</td>
<td>cellular</td>
<td>cellular</td>
</tr>
<tr>
<td>month</td>
<td>Categorical</td>
<td>month</td>
<td>aug</td>
<td>nov</td>
</tr>
<tr>
<td>day_of_week</td>
<td>Categorical</td>
<td>day_of_week</td>
<td>thu</td>
<td>fri</td>
</tr>
</tbody>
</table>
Machine Learning – Easy Path

Target

Machine learning works by finding patterns that connect your data to the value to be predicted. To create an ML model, Amazon ML analyzes examples of data records with correct values. The column that contains these values in the training dataset is called the target.

Because you're using our sample banking data, choose $y$ as your target. Later, when you generate predictions, we'll give you another dataset that doesn't have this information, and Amazon ML will predict it for you.

Select the row containing the value you want to predict.

You have selected a binary attribute named $y$ as the target. ML models trained on this target use logistic regression to train a binary classification model.
Machine Learning – Easy Path

Row identifier (optional)

An optional row identifier helps you understand how prediction rows correspond to observation rows from the input data. If you choose to make an attribute the row identifier, Amazon ML will add that column to the prediction output. A row identifier is intended for reference purposes only. Amazon ML does not include the row identifier when training ML models.

Does your data contain an identifier?  ○ Yes  ○ No

Cancel  Previous  Review
Machine Learning – Easy Path
Machine Learning – Easy Path

ML model settings

You can use the automatically suggested ML model settings, or you can choose to customize.

ML model type: BINARY

ML model target: y

ML model name (Optional):

ML model: Banking.csv

Select training and evaluation settings

Recipes and training parameters control the ML model training process. You can select these settings for your ML model or use the defaults provided by Amazon ML. In either case, you can choose to have Amazon ML reserve a portion of the input data for evaluation. Learn more.

- Default (Recommended)
  - Generate a default recipe
  - Use default training parameters
  - Set aside 30% of your training data to evaluate the training
  - Split the evaluation data sequentially

- Custom
  - Modify the recipe Amazon ML generates
  - Modify training parameters
  - Randomly or sequentially split your evaluation data

Evaluation Name: Evaluation: ML model: Banking.csv

Cancel  Previous  Review
Machine Learning – Easy Path
Machine Learning – Easy Path
Machine Learning – Easy Path

ML model report

<table>
<thead>
<tr>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Settings</td>
</tr>
<tr>
<td>Monitoring</td>
</tr>
</tbody>
</table>

Tools

Try real-time predictions

Evaluations

Evaluation: ML model

ML model summary

- **ID**: ml-8sP7jRnmNLV
- **Name**: ML model: Banking case
- **Type**: Binary classification
- **Creation time**: Jan 18, 2019 8:01:51 AM
- **Completion time**: 3 mins
- **Compute Time (Approximate)**: 2 mins

Status: Completed

Log: Download log

Datasource (training)

- **Datasource ID**: ds-1LVT1LdRo7B
- **Target**: y
- **Input schema**: View input schema

Evaluations

- **Evaluations created**: 1
- **Latest evaluation result**: 0.936 (AUC)

Predictions

- **CloudWatch metrics**: View in CloudWatch
- **Score threshold**: 0.5

A single dataset
Generate one-time predictions for a single dataset.

Generate batch predictions
Machine Learning – Easy Path

ML model report
Summary
Settings
Monitoring
Tools
Try real-time predictions
Evaluations
Summary
Alerts (0)
Explore performance

Evaluation Summary

ID: ev-UmuWyOThLgf
Name: Evaluation: ML model: Banking.csv
Datasource ID: ds-jptcowrYkiW
Output location: Not available
Creation time: Jan 18, 2019 8:01:51 AM
Completion time: 3 mins
Compute Time (Approximate): 2 mins
Status: Completed
Log: Download log

ML model performance metric

On your most recent evaluation, ev-UmuWyOThLgf, the ML model's quality score is considered extremely good for most machine learning applications.

AUC: 0.936
Baseline AUC: 0.500
Difference: 0.436

Next step: If you want to use this ML model to generate predictions, explore trade-offs to optimize the performance of your ML model first.

Tags: Add or edit tags
No tags
Machine Learning – Easy Path

ML model report
Summary
Settings
Monitoring
Tools
Try real-time predictions

Evaluations
Evaluation: ML model performance
Summary
Alerts (0)
Explore performance

ML model performance
This chart shows the distributions of your predicted answers for the actual "1" and "0" records in your evaluation data. Any overlap of the actual "1" & "0" is where your ML model guesses wrong. Learn more.

Adjust the slider to indicate how much error you can tolerate from your ML model based on your needs. Moving the score threshold to the right decreases the number of false positives and increases the number of false negatives.

Trade-off based on score threshold 0.5

- 91% are correct
  551 true positive
  10,678 true negative

- 9% are errors
  316 false positive
  782 false negative

- 7% of the records are predicted as "1"
- 93% of the records are predicted as "0"

Advanced metrics
- False positive rate 0.0287
- Precision 0.5477
- Recall 0.4283
- Accuracy 0.9111
Jupyter Notebook Tutorial

- Notebook Tutorial
- Anaconda Download
Creating a Notebook in AWS

- Go to SageMaker Console
  - Create notebook
  - You will need an IAM role

- Go to IAM Console
  - Mine shows all kinds of errors – ignore these
  - Click role on left
  - Accept defaults (for now)
  - Fine the “arn…” at top – select and copy this
    - This goes into the role in the notebook page
Python

- Python Language Reference
- Python Standard Library
- Python Tutorial
Summary

- Data
  - Buckets

- Processing
  - Creating a Notebook
  - Editing the Notebook
    - Python
  - Running an ML Algorithm

- Output
  - Evaluating Results