Lab Exercise 2: Using Amazon Machine Learning to Detect Phishing Websites

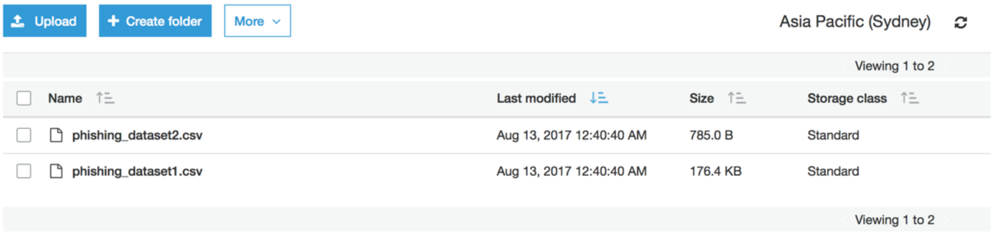
Group 1

A phishing website tricks you into believing you are on a legitimate website to try to steal your account password or other confidential information. You could land on a phishing site by mistyping a URL (web address).

In this Lab exercise, you will train and use a machine learning model to identify potential phishing websites. To complete this exercise, you will use publicly available phishing websites datasets from the University of California at Irvine (UCI) Machine Learning Repository. These datasets contain general information about attributes of phishing websites. You will use this data to identify which websites are most likely to be phishing websites.

Step 1: prepare your data

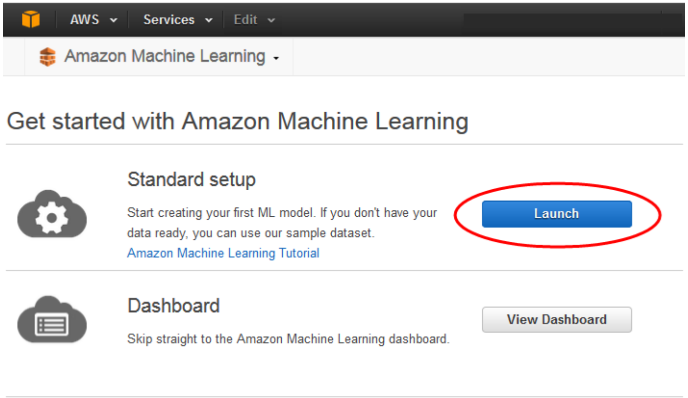
1. Download the file that contains the important features for predicting phishing websites, which is '*phishing_dataset1.csv*'; download the file '*phishing_dataset2.csv*' that you will use to predict whether the given websites are phishing or not. Save them in your desktop.
2. Sign in to the AWS Management Console https://cits5503.signin.aws.amazon.com/console and open the Amazon S3 console. In the **All Buckets** list, create a bucket or choose the location where you want to upload the files. In the navigation bar, choose **Upload**. After that, Choose **Add Files**. In the dialog box, navigate to your desktop, choose *phishing_dataset1.csv* and *phishing_dataset2.csv*, and then choose **Open**.



Name	Last modified	Size	Storage class
phishing_dataset2.csv	Aug 13, 2017 12:40:40 AM	785.0 B	Standard
phishing_dataset1.csv	Aug 13, 2017 12:40:40 AM	176.4 KB	Standard

Step 2: create a training datasource

Open the Amazon Machine Learning console, choose **Get started**. After that, on the **Get started with Amazon Machine Learning** page, choose **Launch**.



Get started with Amazon Machine Learning

Standard setup

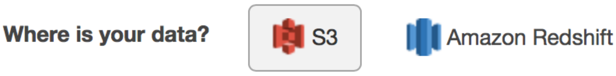
Start creating your first ML model. If you don't have your data ready, you can use our sample dataset.

Amazon Machine Learning Tutorial

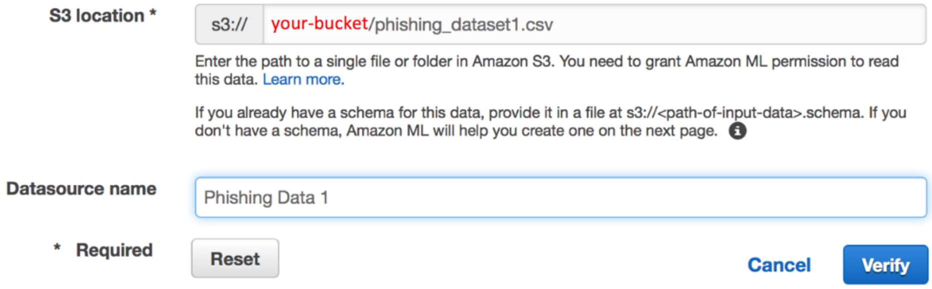
Dashboard

Skip straight to the Amazon Machine Learning dashboard.

1. Click **Create new…** and then select '**Datasource and ML model**', after that, on the Input Data page, for **Where is your data?**, make sure that **S3** is selected.



2. For **S3 Location**, type the full location of the *phishing_dataset1.csv* file from Step 1: prepare your data. For example: your-bucket/ *phishing_dataset1.csv*. Amazon ML prepends s3:// to your bucket name for you.
3. For **Datasource** name, type your ID Phishing Data 1. After that, choose **Verify**. In the **S3 permissions** dialog box, choose **Yes**.



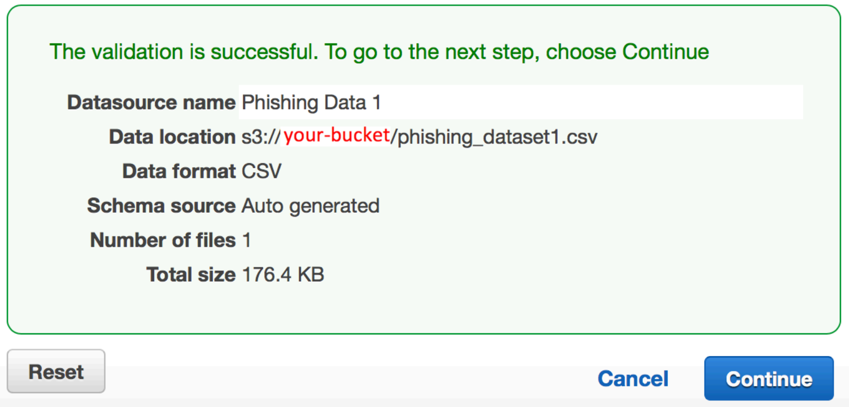
S3 location * s3:// your-bucket/phishing_dataset1.csv

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: Phishing Data 1

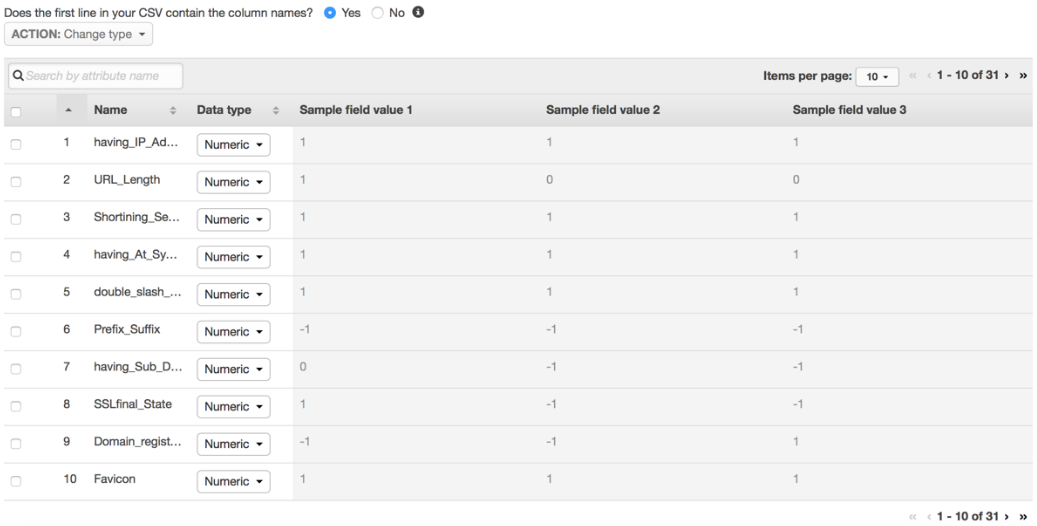
4. If Amazon ML can access and read the data file at the S3 location, you will see a page similar to the following. Review the properties, and then choose **Continue**.



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

Datasource name	Phishing Data 1
Data location	s3://your-bucket/phishing_dataset1.csv
Data format	CSV
Schema source	Auto generated
Number of files	1
Total size	176.4 KB

5. For '**Does the first line in your CSV contain the column names?**', make sure you choose **Yes**. After that, select **Continue**.



Does the first line in your CSV contain the column names? ● Yes ○ No

ACTION: Change type

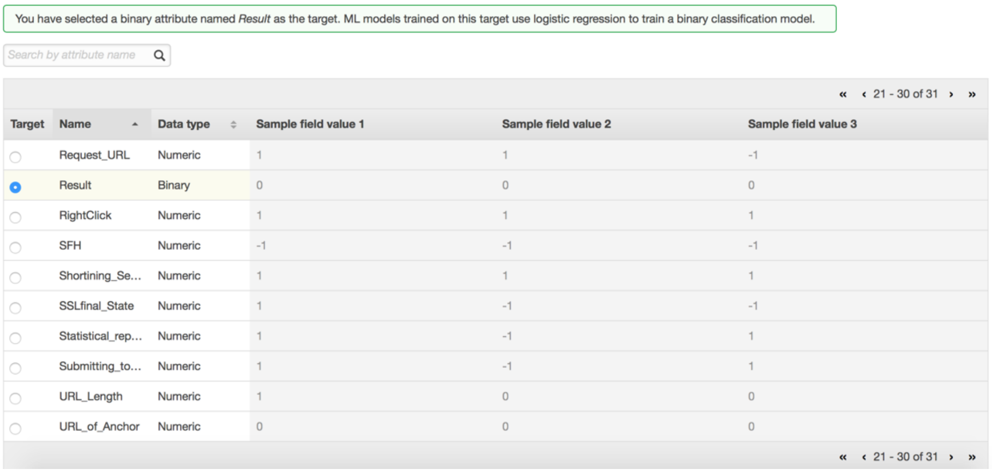
Search by attribute name

Items per page: 10 « ‹ 1 - 10 of 31 › »

□		Name	Data type	Sample field value 1	Sample field value 2	Sample field value 3
□	1	having_IP_Ad...	Numeric	1	1	1
□	2	URL_Length	Numeric	1	0	0
□	3	Shortining_Se...	Numeric	1	1	1
□	4	having_At_Sy...	Numeric	1	1	1
□	5	double_slash_...	Numeric	1	1	1
□	6	Prefix_Suffix	Numeric	-1	-1	-1
□	7	having_Sub_D...	Numeric	0	-1	-1
□	8	SSLfinal_State	Numeric	1	-1	-1
□	9	Domain_regist...	Numeric	-1	-1	1
□	10	Favicon	Numeric	1	1	1

« ‹ 1 - 10 of 31 › »

6. Next, select a target attribute, which the ML model must learn to predict. Here select ‘Result’ as the target, then choose **Continue**.



You have selected a binary attribute named *Result* as the target. ML models trained on this target use logistic regression to train a binary classification model.

Search by attribute name

« ‹ 21 - 30 of 31 › »

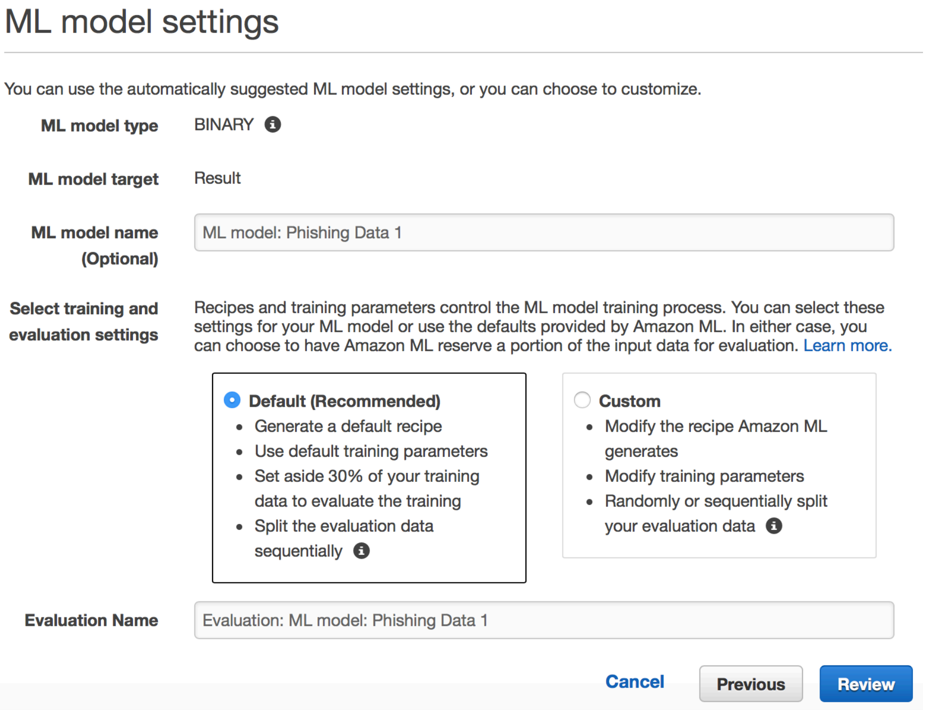
Target	Name	Data type	Sample field value 1	Sample field value 2	Sample field value 3
○	Request_URL	Numeric	1	1	-1
●	Result	Binary	0	0	0
○	RightClick	Numeric	1	1	1
○	SFH	Numeric	-1	-1	-1
○	Shortining_Se...	Numeric	1	1	1
○	SSLfinal_State	Numeric	1	-1	-1
○	Statistical_rep...	Numeric	1	-1	1
○	Submitting_to...	Numeric	1	-1	1
○	URL_Length	Numeric	1	0	0
○	URL_of_Anchor	Numeric	0	0	0

« ‹ 21 - 30 of 31 › »

7. On the **Row ID** page, for **Does your data contain an identifier?**, make sure that **No**, the default, is selected. After that, choose **Review**, and then choose **continue**.

Step 3: create an ML model

1. For **ML model settings**, name it as follow: ML model: your ID Phishing data 1 , then ensure that **Default** is selected.



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 Result

ML model name (Optional) ML model: Phishing Data 1

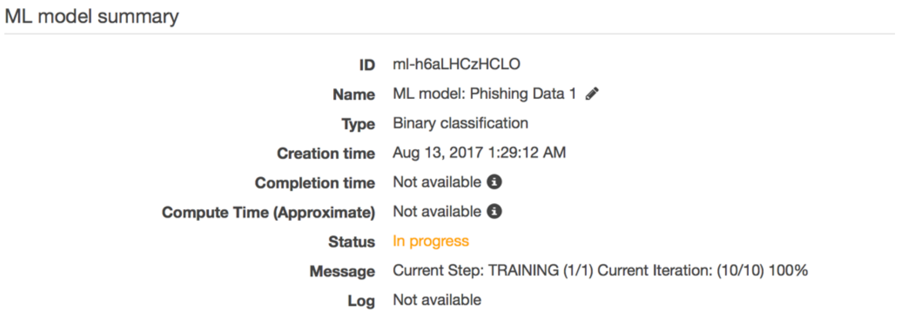
Select training and evaluation settings Recipes and training parameters control the ML model training process. You can select these settings for your ML model or use the defaults provided by Amazon ML. In either case, you can choose to have Amazon ML reserve a portion of the input data for evaluation.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

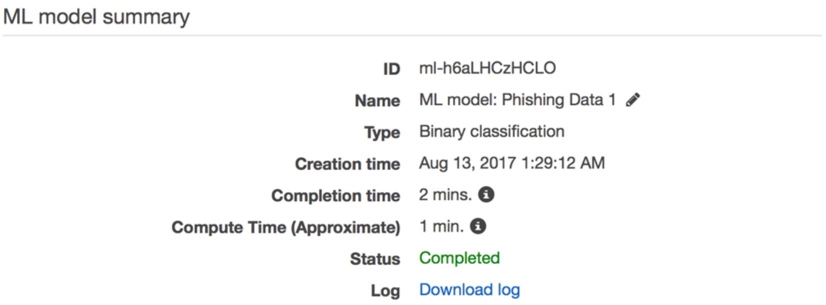
Evaluation Name Evaluation: ML model: Phishing Data 1

2. Choose **Review** to review your settings, and then choose **Finish**.
3. While your model has completed all actions, it reports the status as **Completed**. Wait for the evaluation to complete before proceeding.



ML model summary

ID	ml-h6aLHCzHCLO
Name	ML model: Phishing Data 1
Type	Binary classification
Creation time	Aug 13, 2017 1:29:12 AM
Completion time	Not available
Compute Time (Approximate)	Not available
Status	In progress
Message	Current Step: TRAINING (1/1) Current Iteration: (10/10) 100%
Log	Not available

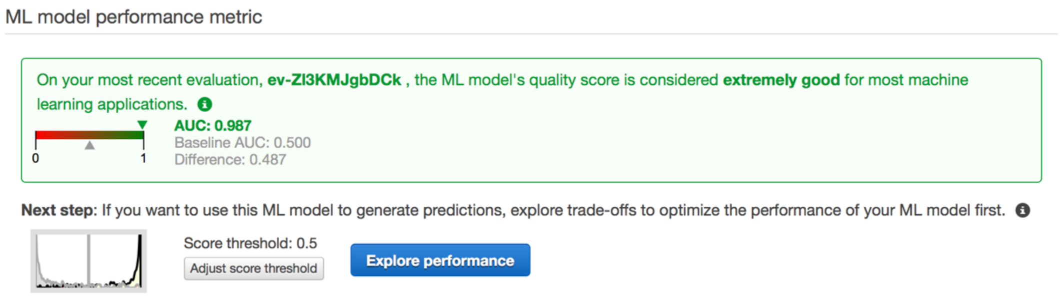


ML model summary

ID	ml-h6aLHCzHCLO
Name	ML model: Phishing Data 1
Type	Binary classification
Creation time	Aug 13, 2017 1:29:12 AM
Completion time	2 mins.
Compute Time (Approximate)	1 min.
Status	Completed
Log	Download log

Step 4: review the ML model's predictive performance and set a score threshold

1. On the **ML model summary** page, in the **ML model report** navigation pane, choose **Evaluation**, choose **Evaluation: ML model: Phishing Data 1**, and then choose **summary**.
2. On the **Evaluation summary** page, review the evaluation summary, including the model's AUC performance metric.



ML model performance metric

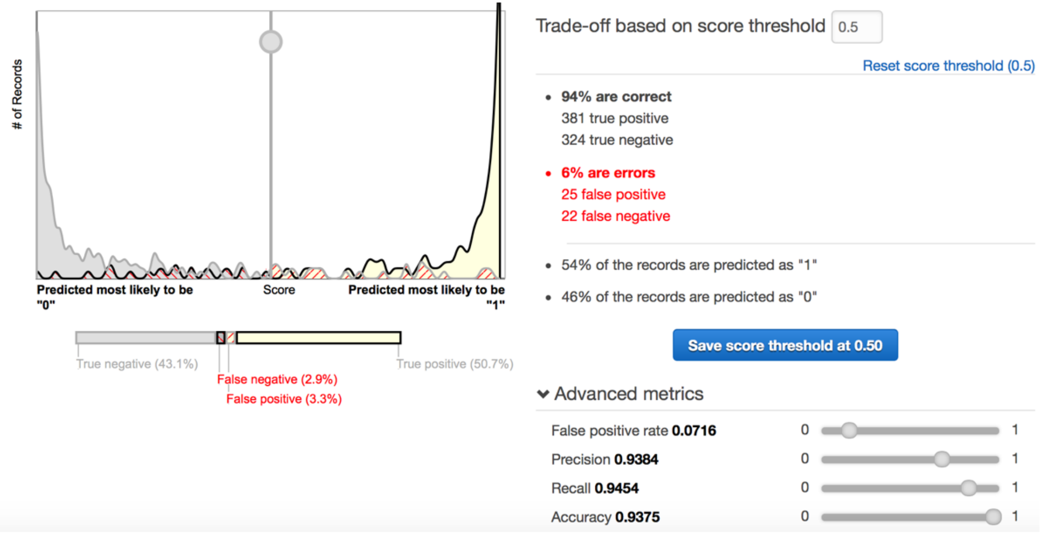
On your most recent evaluation, **ev-ZI3KMJgbDCk** , the ML model's quality score is considered **extremely good** for most machine learning applications.

AUC: 0.987
Baseline AUC: 0.500
Difference: 0.487

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.

Score threshold: 0.5

3. To select a score threshold for your ML model, on the **Evaluation summary** page, choose **Explore performance**.



# of Records
Predicted most likely to be "0"
Score
Predicted most likely to be "1"
True negative (43.1%)
False negative (2.9%)
False positive (3.3%)
True positive (50.7%)
Reset score threshold (0.5)
Save score threshold at 0.50

Trade-off based on score threshold 0.5

- **94% are correct**
  381 true positive
  324 true negative
- **6% are errors**
  25 false positive
  22 false negative

- 54% of the records are predicted as "1"
- 46% of the records are predicted as "0"

Advanced metrics

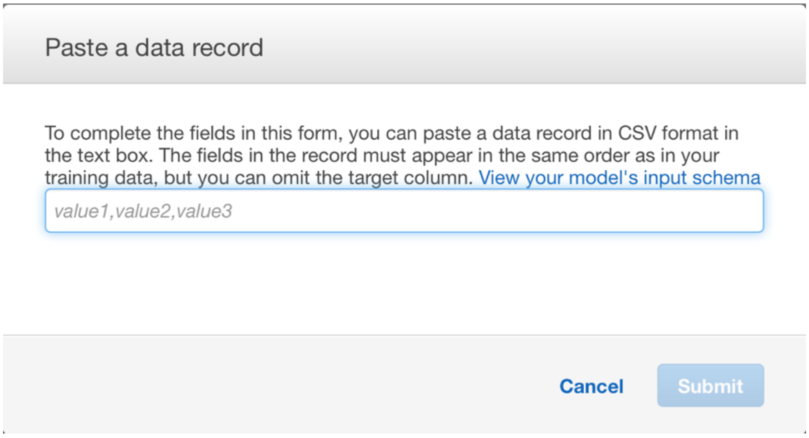
False positive rate **0.0716**	0		1
Precision **0.9384**	0		1
Recall **0.9454**	0		1
Accuracy **0.9375**	0		1

4. Let's say you want to target the lower 45% of the websites that will be identified as phishing (0). Slide the vertical selector to set the score threshold to a value that corresponds to **45% of the records are predicted as "0"**. Save your score threshold and take a screenshot of your model performance chart with your fine-tuned score threshold.

Step 5: use the ML model to generate predictions

A- Real-time prediction:

1. To try a real-time prediction, in the **ML model report** navigation pane, choose **Try real-time predictions**.
2. Choose **Paste a record**, then copy the following observation and paste it in the dialog box:
   -1,-1,1,1,1,-1,-1,-1,-1,1,1,1,-1,-1,0,0,1,1,0,1,1,1,1,1,1,1,-1,1,0,1



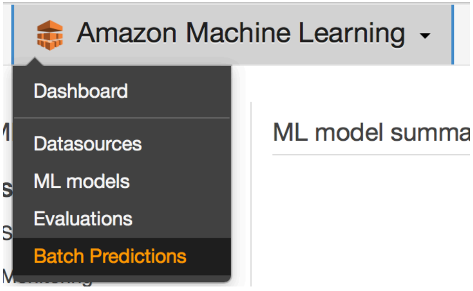
Paste a data record

To complete the fields in this form, you can paste a data record in CSV format in the text box. The fields in the record must appear in the same order as in your training data, but you can omit the target column. View your model's input schema

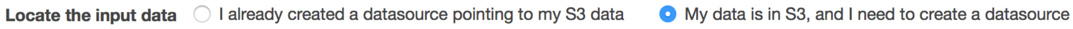
3. Choose **Submit.** Is this website a phishing website? Answer this question and take a screenshot of the prediction results with the predicted score.

B- Batch prediction:

1. To create a batch prediction, choose **Amazon Machine Learning**, and then choose **Batch Predictions**.



2. Choose **Create new batch prediction**.
3. On the **ML model for batch prediction** page, choose **ML model: your ID Phishing Data 1**, then choose **Continue**.
4. For **Locate the input data**, choose **My data is in S3, and I need to create a datasource**.



Locate the input data ○ I already created a datasource pointing to my S3 data ● My data is in S3, and I need to create a datasource

5. For **Datasource name**, type your ID Phishing Data 2; for S3 Location, type the full location of the *phishing_dataset2.csv* file: your-bucket/ *phishing_dataset2.csv.*
6. For **Does the first line in your CSV contain the column names?**, choose **Yes**. After that, choose **Verify**, then choose **Continue**.
7. For **S3 destination**, type the name of the Amazon S3 location where you uploaded the files in Step 1: Prepare Your Data. Amazon ML uploads the prediction results there.

8. For **Batch prediction name,** accept the default, **Batch prediction: MLmodel: Banking Data1**. Amazon ML chooses the default name based on the model it will use to create predictions. In this tutorial, the model and the predictions are named after the training datasource, Phishing Data 1.

9. Choose **Review**.

10. In the **S3 permissions** dialog box, choose **Yes**.

11. On the **Review** page, choose **Finish**. While Amazon ML processes the request, it reports a status of **In Progress**. After the batch prediction has completed, the request's status changes to **Completed**. Now, you can view the results.

To view the predictions

1. Choose **Amazon Machine Learning**, and then choose **Batch Predictions**.

2. In the list of predictions, choose **Batch prediction: ML model: Banking Data 1**. The **Batch prediction info** page appears.

3. To view the results of the batch prediction, go to the Amazon S3 console at https://console.aws.amazon.com/s3/ and navigate to the Amazon S3 location referenced in

the **Output S3 URL** field. From there, navigate to the results folder, which will have a name similar to s3://aml- data/batch-prediction/result.

4. The prediction is stored in a compressed .gzip file with the .gz extension.

5. Download the prediction file to your desktop, uncompress it, and open it.

6. The file has two columns, **bestAnswer** and **score**, and a row for each observation in your datasource. The results in the **bestAnswer** column are based on the score threshold of 0.45 that you set in Step 4: Review the ML Model's Predictive Performance and Set a Score Threshold. A **score** greater than 0.45 results in a **bestAnswer** of 1 (not phishing), which is a positive response or prediction, and a **score** less than 0.45 results in a **bestAnswer** of 0, which is a negative response or prediction (phishing).
7. Take a screenshot of the batch prediction results with the predicted scores.

Now that you have created, reviewed, and used your model, clean up the data and AWS resources you created to avoid incurring unnecessary charges and to keep your workspace uncluttered.

Step 6: Clean Up

To delete the input data stored in Amazon S3

1. Open the Amazon S3 console at https://console.aws.amazon.com/s3/.

2. Navigate to the Amazon S3 location where you stored the *phishing_dataset1.csv* and *phishing_dataset2.csv* files

3. Select the *phishing_dataset1.csv, phishing_dataset2.csv*, and writePermissionCheck.tmp files.

4. Choose **Actions**, and then choose **Delete**.

5. When prompted for confirmation, choose **OK**.

Although you aren't charged for keeping the record of the batch prediction that Amazon ML ran or the datasources, model, and evaluation that you created during the Lab, we recommend that you delete them to prevent cluttering your workspace.

To delete the batch predictions

1. Navigate to the Amazon S3 location where you stored the output of the batch prediction.

6. Choose the batch-prediction folder.

7. Choose **Actions**, and then choose **Delete**.

8. When prompted for confirmation, choose **OK**.

To delete the Amazon ML resources

1. On the Amazon ML dashboard, select the following resources.

- Your ID Phishing Data1 datasource
- Your ID Phishing Data1_[percentBegin=0,percentEnd=70,strategy=sequential]datasource
- Your ID Phishing Data2_[percentBegin=70,percentEnd=100,strategy=sequential] datasource
- Your ID Phishing Data2 datasource
- The ML model: Your ID Phishing Data 1 ML model
- The Evaluation: ML model: Your ID Phishing Data1 evaluation

2. Choose **Actions**, and then choose **Delete**.

3. In the dialog box, choose **Delete** to delete all selected resources.