JOIN

#JOINdata
@lookerdata

Hands-on Lab: Predictive Analytics with Looker and Amazon SageMaker Powered by AWS

Eric Carr
Alliances

looker.com/hol

Select the **Predictive Analytics with Looker and Amazon SageMaker Powered by AWS** lab in the drop-down

Eric Carr

Sales Engineer, Alliances

Agenda

1. Data science workflow
2. Exploring the data
3. Training a model
4. Testing a model and analyzing performance
5. Questions

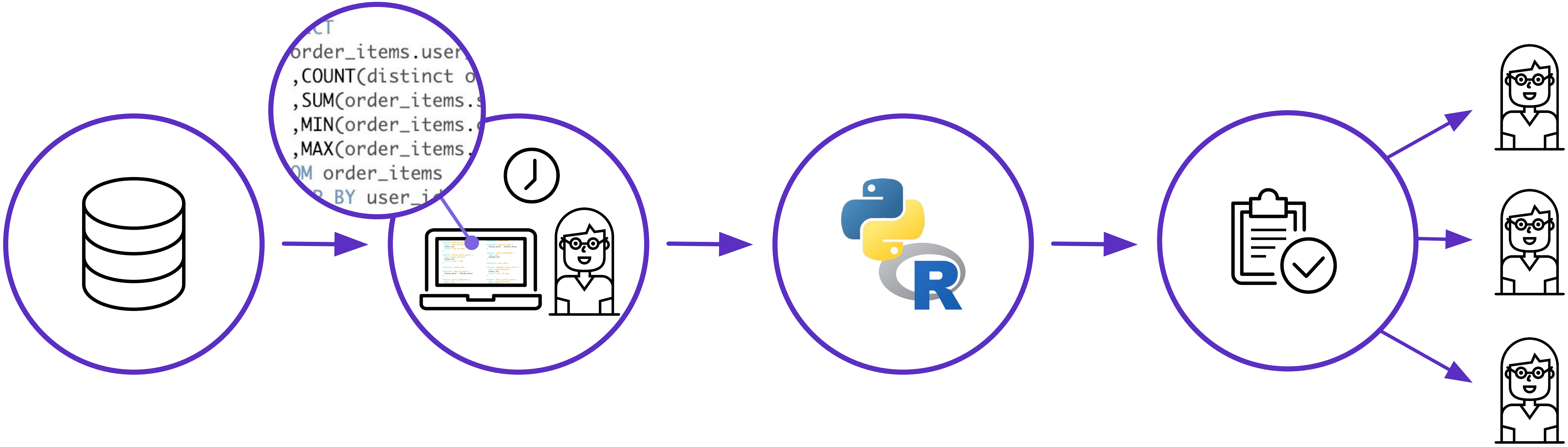
Data science workflow

Data science

Why is predictive analytics important?

- Forecasting, sales, events, volume
- Fraud detection
- Computing risk

Workflow pre-Looker



Workflow with Looker

What is SageMaker?

Machine learning for every developer and data scientist

- Fully-managed service that covers the entire machine learning workflow
- Quickly build, train, and deploy machine learning models
  - Build and optimize a ML algorithm from the built-in marketplace
  - Train the model to optimize performance
  - Deploy to a fully managed environment with auto-scaling
- Use Looker’s Action Hub integration with Amazon SageMaker to streamline the data science workflow by allowing model training and inference to be initiated directly from within the Looker Scheduler

Exploring the data

Will a customer enroll in a term deposit?

Make a prediction using bank client information

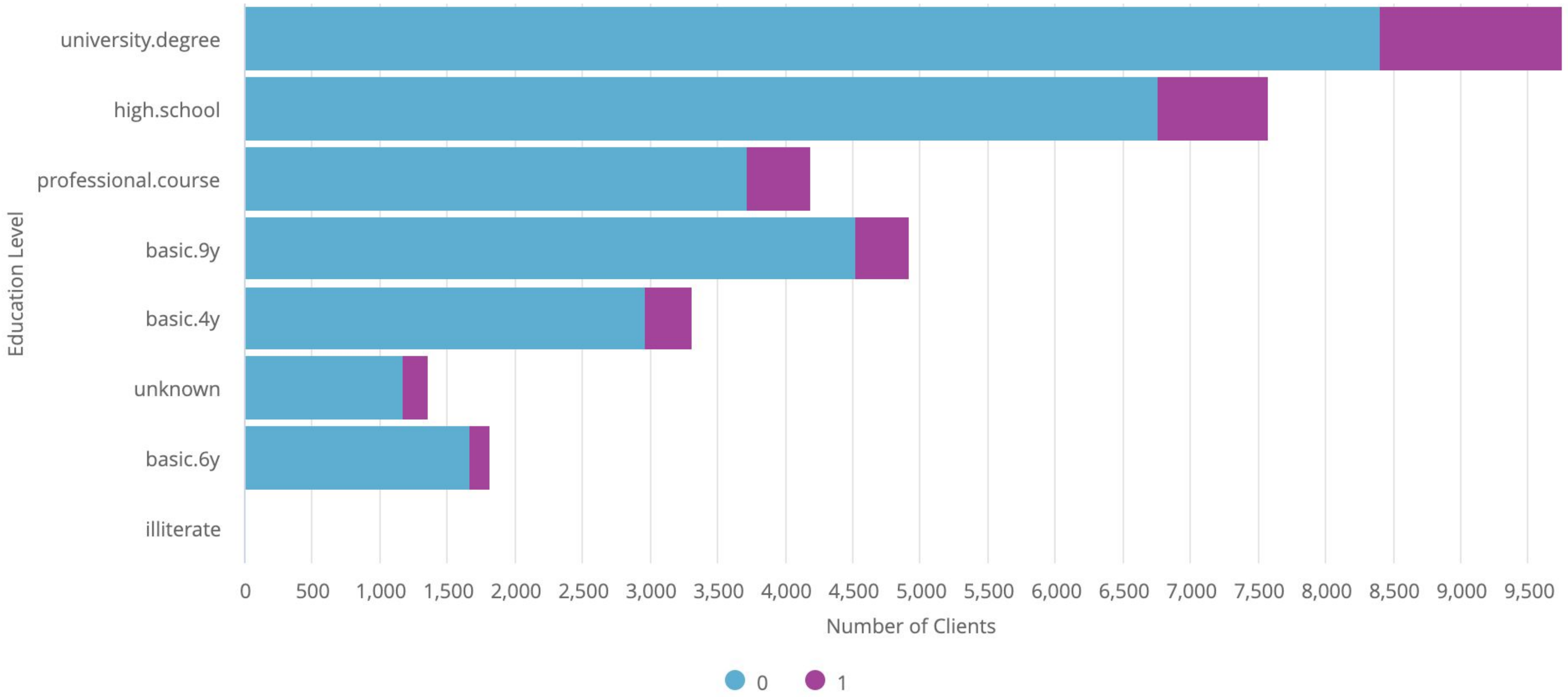
- The scenario: You work for the marketing department of a bank, and you need to predict if a customer will enroll in a term deposit using the client data that you have available.
    - Client demographics
    - Responses to prior marketing events
    - External environment factors

- Explore the data and identify client variables that you think will help predict whether or not a client enrolls in a term deposit.

Explore the data

What variables will influence a client's behavior the most?

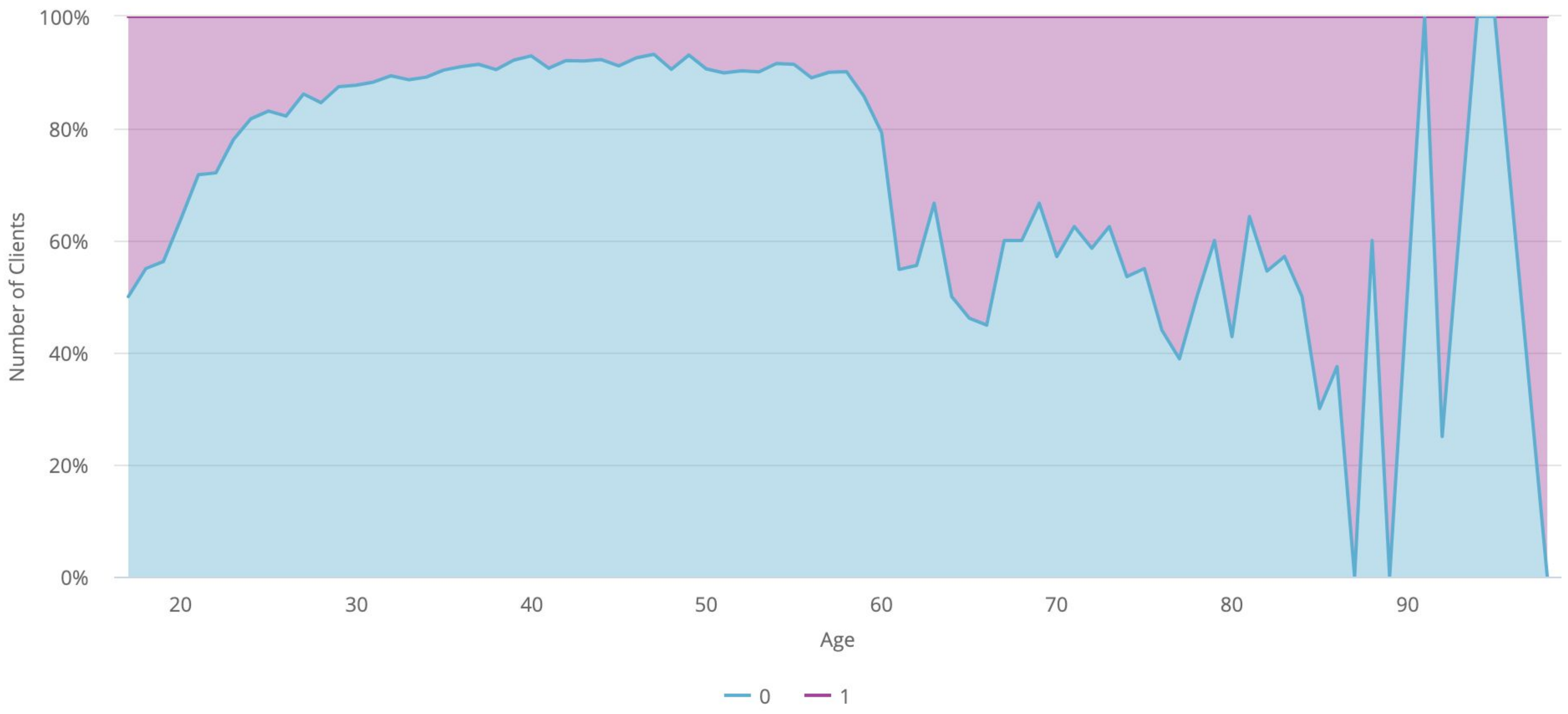
Education level:



Explore the data

What variables will influence a client's behavior the most?

Age:



Training a model

How to train your model

1. Create a dataset that includes the predictive variables and prediction target for a portion (commonly 70%) of your existing data using munged data
2. Make this data set available to an analytical tool
   a. Amazon SageMaker via direct Looker integration
   b. Python or R via the Looker SDK
3. Apply a training algorithm to the training data set to create a model that can be tested using the remaining data (and then applied to future data to make predictions)

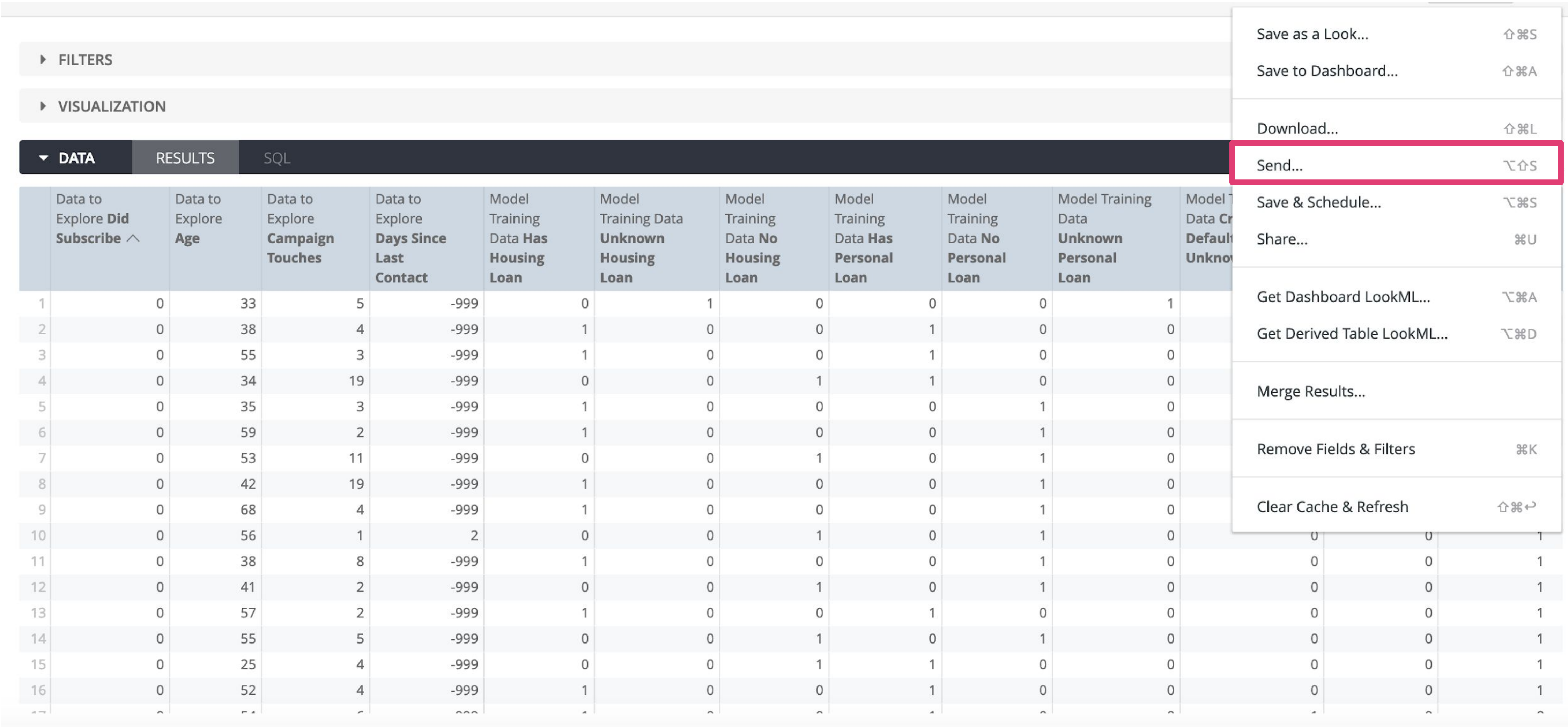
Create a training dataset

Include key variables identified via exploration

	Data to Explore **Did Subscribe** ^	Data to Explore **Age**	Data to Explore **Campaign Touches**	Data to Explore **Days Since Last Contact**	Model Training Data **Has Housing Loan**	Model Training Data **Unknown Housing Loan**	Model Training Data **No Housing Loan**	Model Training Data **Has Personal Loan**	Model Training Data **No Personal Loan**	Model Training Data **Unknown Personal Loan**	Model Training Data **Credit Default Status Unknown**	Model Training Data **Is In Credit Default**	Model Training Data **Is Not In Credit Default**
1	0	33	5	-999	0	1	0	0	0	1	0	0	1
2	0	38	4	-999	1	0	0	1	0	0	1	0	0
3	0	55	3	-999	1	0	0	1	0	0	1	0	0
4	0	34	19	-999	0	0	1	1	0	0	1	0	0
5	0	35	3	-999	1	0	0	0	1	0	0	0	1
6	0	59	2	-999	1	0	0	0	1	0	1	0	0
7	0	53	11	-999	0	0	1	0	1	0	0	0	1
8	0	42	19	-999	1	0	0	0	1	0	0	0	1
9	0	68	4	-999	1	0	0	0	1	0	0	0	1
10	0	56	1	2	0	0	1	0	1	0	0	0	1
11	0	38	8	-999	1	0	0	0	1	0	0	0	1
12	0	41	2	-999	0	0	1	0	1	0	0	0	1
13	0	57	2	-999	1	0	0	1	0	0	0	0	1
14	0	55	5	-999	0	0	1	0	1	0	0	0	1
15	0	25	4	-999	0	0	1	1	0	0	0	0	1
16	0	52	4	-999	1	0	0	1	0	0	0	0	1

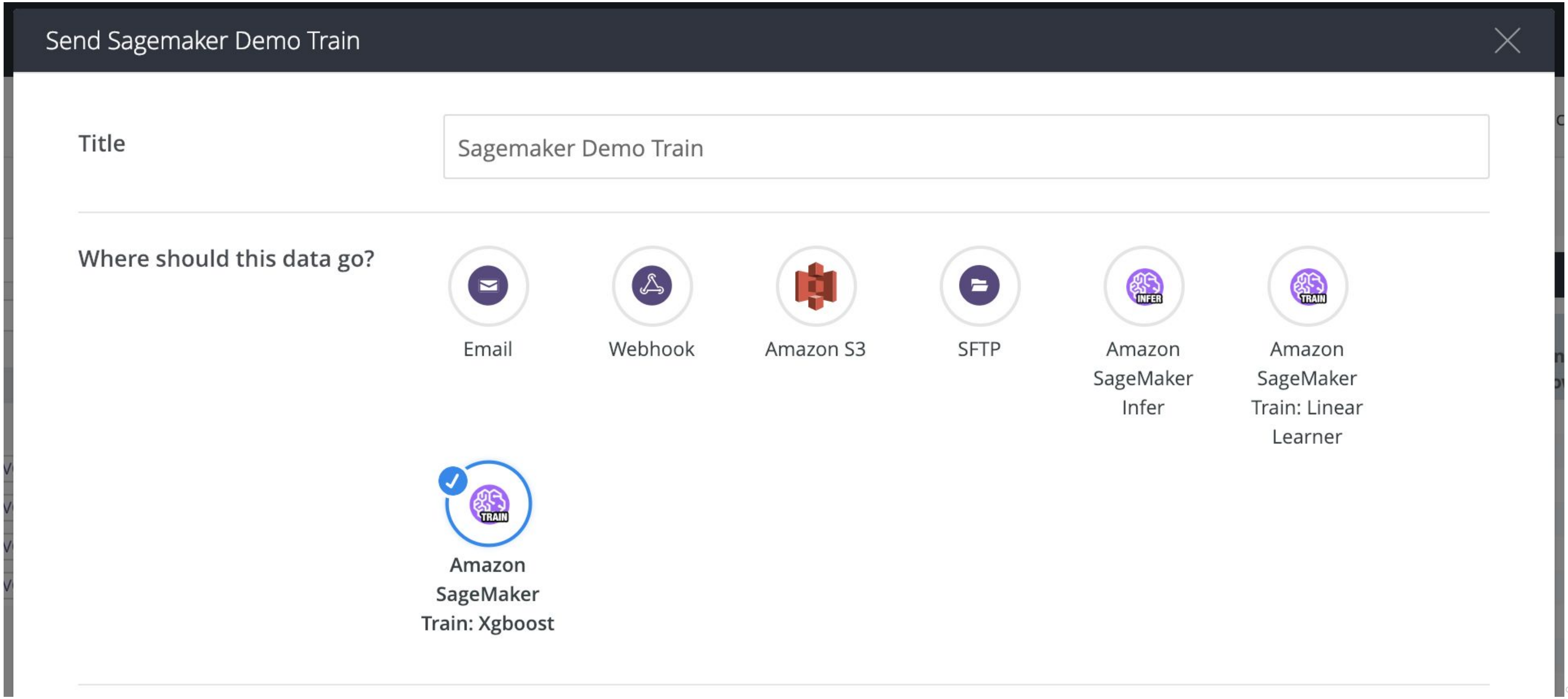
Send the training data to SageMaker

Use **Send** to send once, or schedule it for recurring processes



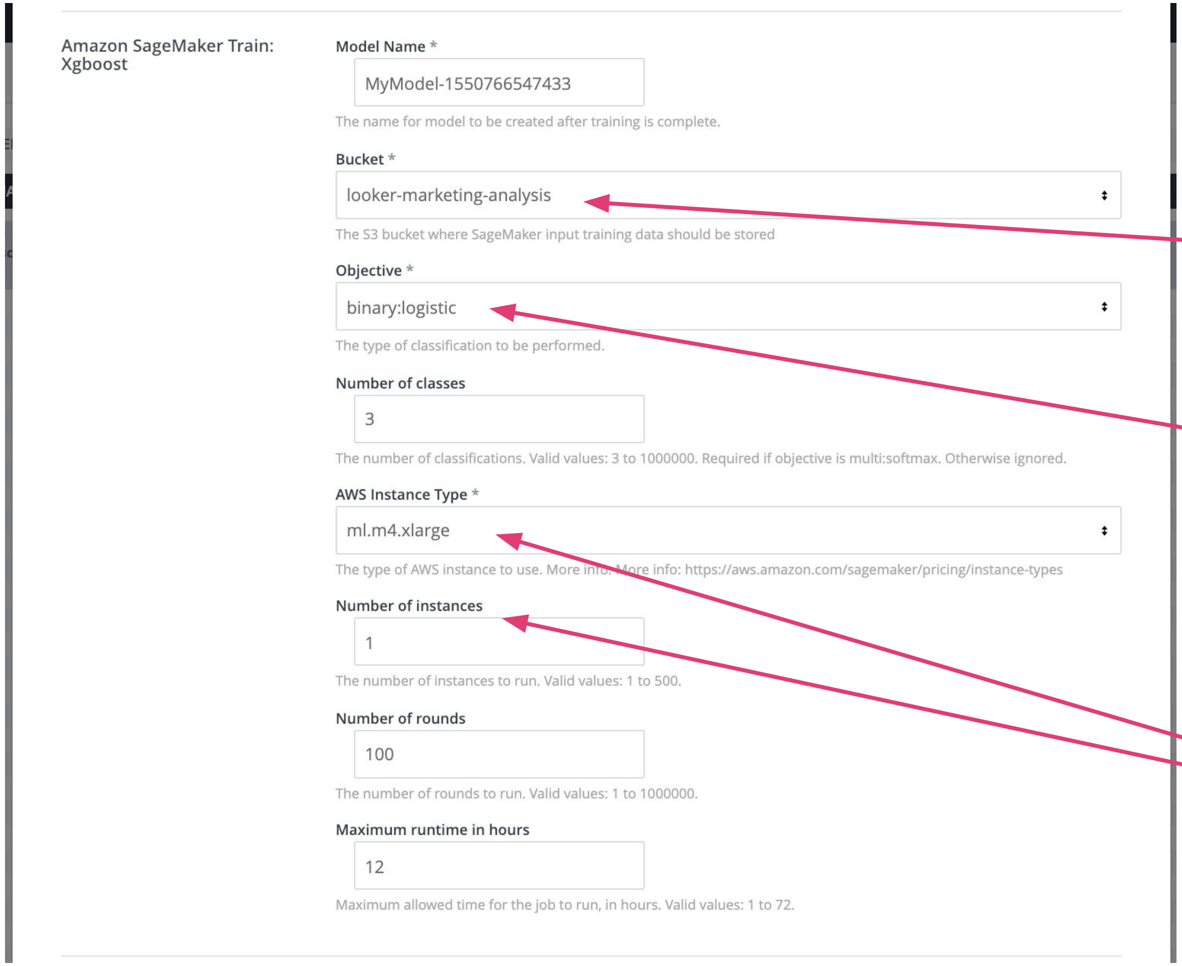
Send the training data to SageMaker

Choose a model training algorithm



Send the training data to SageMaker

Define parameters



The model name should be unique and specific, as you will need to call upon it when using your trained model in the future

Bucket is the S3 bucket where your model and data will be stored

Objective options:

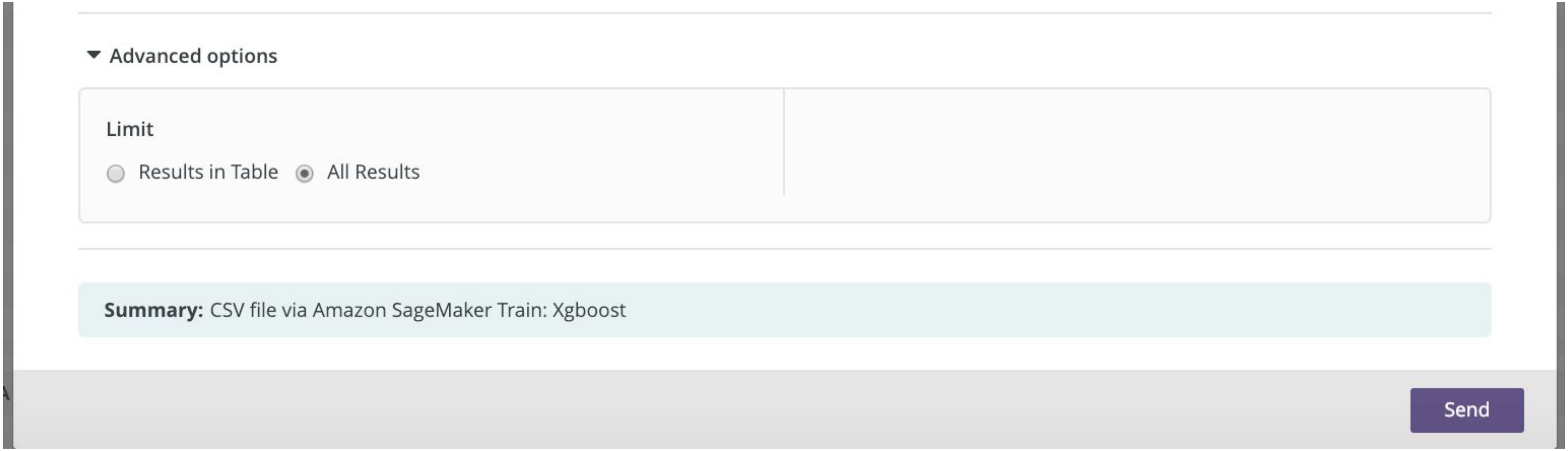
- `binary:logistic` = yes/no output (e.g., will a customer enroll or not?)
- `reg:linear` = predict a number (e.g., what a customer's lifetime value will be)
- `multi:softmax` = creating groupings

AWS Instance Type and **Number of instances** will influence the speed of number crunching

looker

Send the training data to SageMaker

Make sure to send **All Results**



looker

Testing a model and analyzing performance

How to test your model

1. Build the same data set that you used for training your model, but use your testing data (the remaining 30% of the data that was not used for model training)
2. Apply your model to the testing data
3. Measure how well your model predicts the desired target

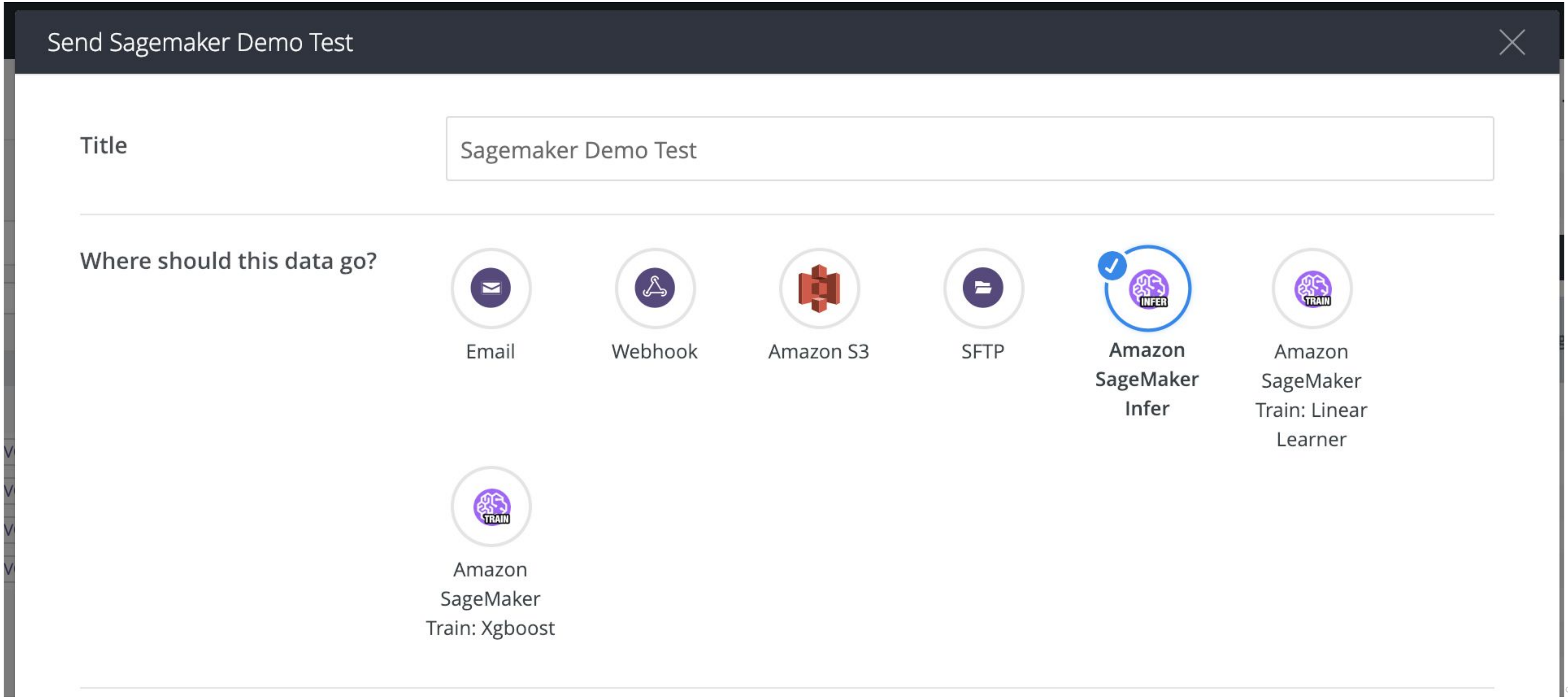
Create a testing dataset

Should be the same as your training dataset, but use the remaining test data

	Model Testing **Age**	Model Testing **Campaign Touches**	Model Testing **Days Since Last Contact**	Model Testing **Has Housing Loan**	Model Testing **No Housing Loan**	Model Testing **Unknown Housing Loan**	Model Testing **Has Personal Loan**	Model Testing **No Personal Loan**	Model Testing **Unknown Personal Loan**	Model Testing **Credit Default Status Unknown**	Model Testing **Is In Credit Default**	Model Testing **Is Not In Credit Default**
1	39	1	-999	0	0	1	0	0	1	1	0	0
2	41	2	-999	0	0	1	0	0	1	0	0	1
3	34	2	-999	0	1	0	0	1	0	1	0	0
4	33	23	-999	1	0	0	0	1	0	0	0	1
5	36	5	-999	0	1	0	0	1	0	1	0	0
6	58	2	4	0	1	0	1	0	0	0	0	1
7	55	4	-999	0	1	0	0	1	0	0	0	1
8	41	3	-999	0	1	0	0	1	0	0	0	1
9	56	5	-999	0	1	0	0	1	0	1	0	0
10	48	2	1	0	1	0	1	0	0	0	0	1
11	45	5	-999	1	0	0	0	1	0	0	0	1
12	30	26	-999	1	0	0	0	1	0	0	0	1
13	30	7	-999	1	0	0	0	1	0	0	0	1
14	57	1	-999	0	1	0	1	0	0	0	0	1
15	54	4	-999	1	0	0	0	1	0	0	0	1
16	39	3	-999	0	1	0	0	1	0	0	0	1
17	54	1	4	1	0	0	0	1	0	0	0	1

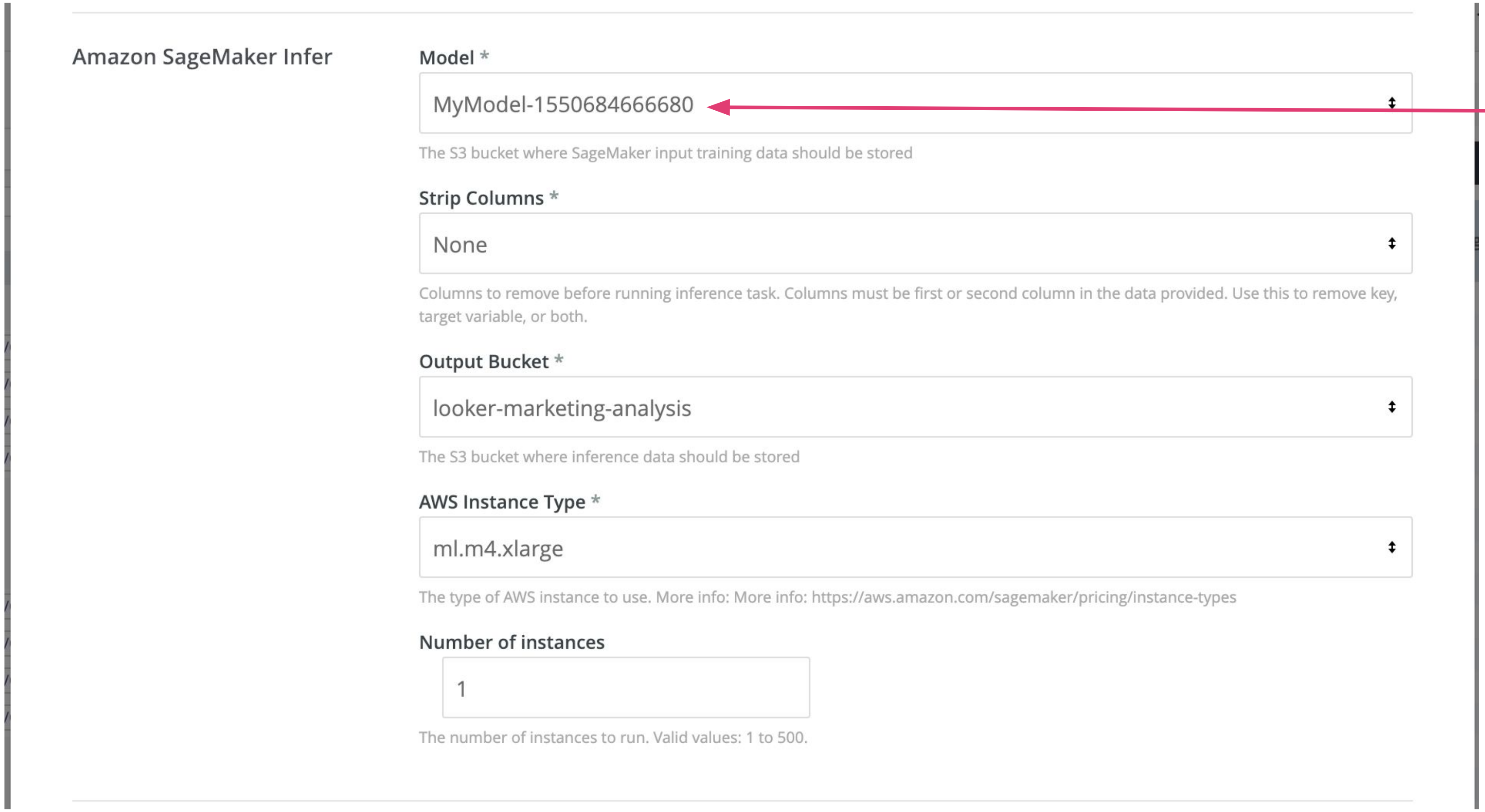
Send the test data to SageMaker

Use Amazon SageMaker Infer



Send the test data to SageMaker

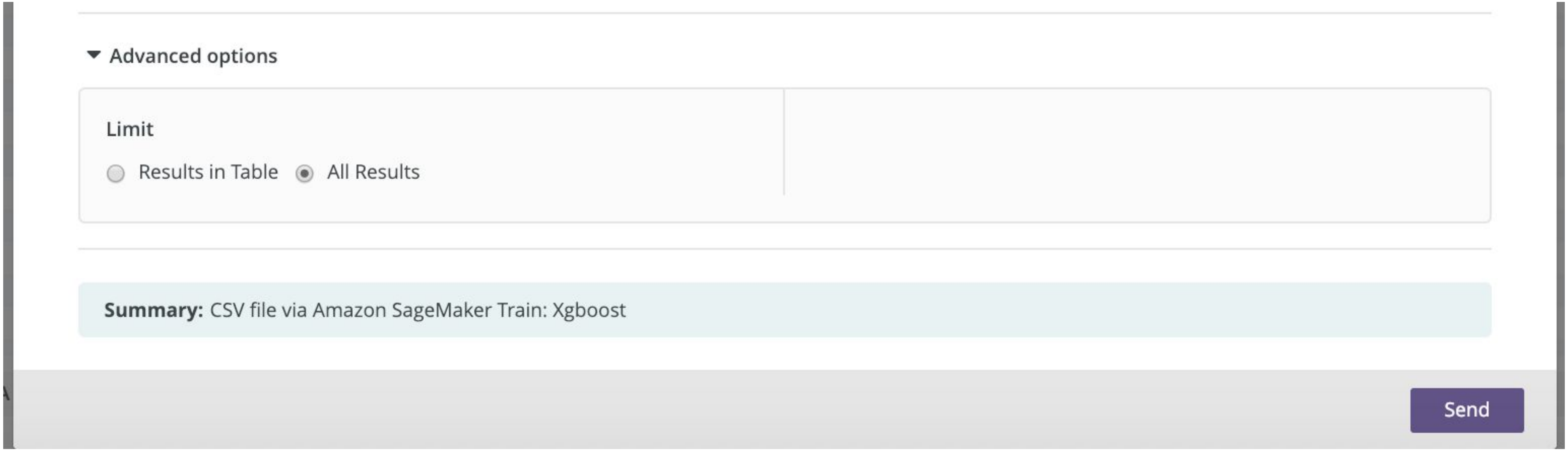
Apply the model you trained



Make sure to use the name of the model that you created when training your model!

Send the test data to SageMaker

Make sure to send **All Results**



Prediction data set

	Prediction Analysis **Client ID**	Prediction Analysis **Did Subscribe**	Prediction Analysis **Predicted Value**	Prediction Analysis **Prediction**
1	91e13bc1-5481-428d-8e97-45327c4a8398	0	0	34.81%
2	0a9da534-c207-4273-9037-f19aabf2a144	0	0	27.09%
3	71788ed6-d521-4cf9-8b0c-c15d7b9c8b8e	0	0	43.73%
4	9693b4d0-fbcf-4d0b-8d37-0f3dfdb4a264	0	0	8.11%
5	9e31f4c0-a5fd-485c-b35a-a010cf3295aa	1	0	12.70%
6	be8d5091-3c44-4df2-84cb-f854e186d850	0	0	21.74%
7	4a608cb7-0e2b-4899-a06c-9a4268326d2c	0	0	5.40%
8	944d6d77-acf6-43dd-8a47-4a103c79f8ca	0	0	9.46%
9	ccfefca1-8837-4307-90a4-dae1f2bd03b8	0	0	4.65%
10	3758fd1c-fc06-4beb-b61b-8ae8fd4e9d26	0	0	10.60%
11	1f07fcea-cd31-4f2a-bb6f-a668fd07c2d4	0	0	9.91%
12	695e4361-07ec-4dda-88ba-726877e11918	0	0	18.87%
13	c778adee-b543-452c-b7fc-f3bc4d856824	0	0	3.75%
14	22ebd36e-2950-4acf-a144-7626a681d7f7	0	0	30.36%
15	e65efa5e-a1c5-4ba8-aef0-3b80fc63f336	0	0	3.27%
16	d8b6e050-a563-4102-bf7e-675a388c9919	0	0	6.64%
17	f2c0ebcf-33aa-425b-b156-2efaa863ddfe	0	0	5.35%
18	a41618b2-a20c-45fa-96cb-ba292906223e	0	0	6.61%
19	fe546df0-e69f-423d-a318-b2c9c0789955	0	0	3.91%

Measuring success

A few key terms

- **True Positives:** Clients who enrolled in a term deposit that we predicted correctly
- **True Negatives:** Clients who did NOT enroll in a term deposit that we predicted correctly
- **False Positives:** Clients who did NOT enroll in a term deposit that we predicted WOULD enroll in the CD
- **False Negatives:** Clients who DID enroll in the term deposit that we predicted would NOT enroll in the CD

	Predicted Value >	0	1
	Subscribed	Predict	Predict
1	0	3,609	52
2	1	368	90

looker

Measuring success

Calculating sensitivity and specificity

- **Sensitivity:** What fraction did we predict would enroll in the term deposit out of all actual enrollments?

  90 predicted / (90 + 368 actual) = 90/458 = 0.197

- **Specificity:** What fraction did we predict would not enroll in the term deposit out of all clients who did not enroll?

  3,609 predicted / (3,609 + 52 actual) = 3,609/3,661 = 0.986

	Predicted Value >	0	1
	Subscribed	Predict	Predict
1	0	3,609	52
2	1	368	90

looker

Measuring success

Calculating sensitivity and specificity

▸ **FILTERS** Cutoff **is 0.5**

0.1655
MAE

0.8980
AUC

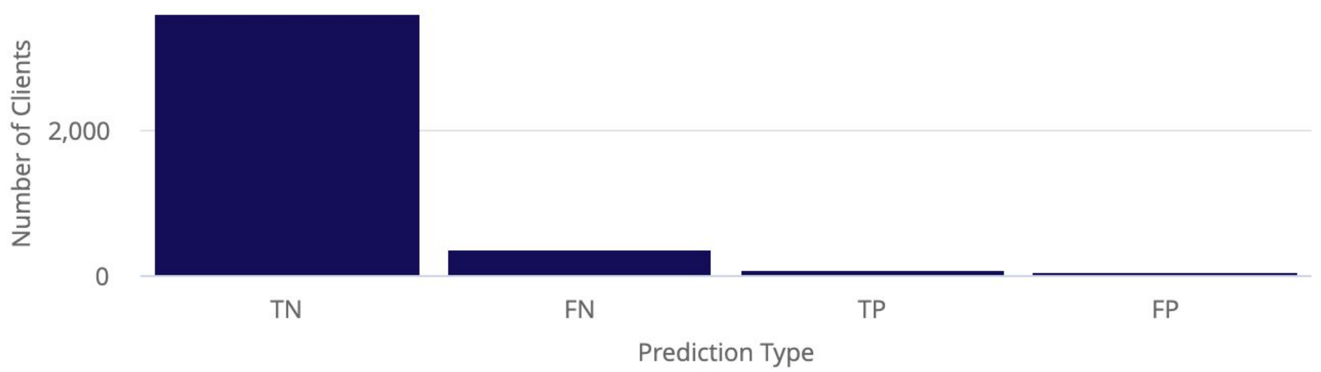
Predicted vs Actual Confusion Matrix

	Prediction Analysis **Predicted Value**	0	1
	Prediction Analysis **Did Subscribe**	Prediction Analysis **Number of Clients**	Prediction Analysis **Number of Clients**
1	0	3,609	52

0.6338
Sensitivity

0.9075
Specificity

Predictions by Type



What do we do with this information?

1. Run new client data through the model to identify candidates likely to enroll in a term deposit
2. Target these clients for outreach inviting them to subscribe
   a. Send an email
   b. Call them directly
3. Focus marketing campaigns for bringing in new clients on demographic groups most likely to subscribe to these types of additional programs

Questions?

looker

Thank you

Rate this session in
the JOIN mobile app

#JOINdata
@lookerdata