LINEAR REGRESSION WITH AMAZON AWS MACHINE LEARNING



Here we show how to use Amazon AWS Machine Learning to do linear regression. In a previous post, we explored using <u>Amazon AWS Machine Learning for Logistic Regression</u>.

To review, **linear regression** is used to predict some value y given the values x1, x2, x3, ..., xn. in other words it finds the coefficients b1, b2, b3, ... , bn plus an offset c to yield this formula:

y = b1x1 + b2x2 + b3x3 + + c.

It uses the **least squares error** approach to find this formula. In other words, think of all these values x1, x2, ... existing in some N-dimensional space. The line y is the line that minimizes the distance between the observed and predicted values for all these values. So it is the line that most nearly split right down the middle of the data observed in the training set. Since we know what that line looks like, we can take any new data, plug those into the formula, and then make a prediction.

As always models are built like this:

- Take an input set of data that you think it correlated. Such as hours of exercise and weight reduction.
- Split that data into a training set and testing set. Amazon does that splitting for you.
- Run the linear regression algorithm to find the formula for y. Amazon picks linear regression based upon the characteristics of the data. It would pick another type of regression or classification model is we picked a data set that for which that was a better fit.

- Check how accurate the model is by taking the square root of the differences between the observed and predicted values. Amazon actually uses the mean of this difference.
- Then take new data and apply the formula y to make a prediction.

Get Some Data

We will use this data of student test scores from the UCI Machine Learning repository.

I copied this data into Google Sheets <u>here</u> so that you can more easily read it. Plus I show the training data set and the one used for prediction.

You download <u>this data in raw format</u> and upload it to Amazon S3. But first, we have to delete the column headings and change the semicolon (;) separators to commas (,) as shown below. We take the first 400 rows as our model training data and the last 249 for prediction. Use vi to delete the first from the data as Amazon will not read the schema automatically (Too bad it does not).

```
vi student-por.csv
sed -i 's/;/,/g' student-por.csv
head -400 student-por.csv > grades400.csv
tail -249 student-por.csv > grades249.csv
```

Now create a bucket in S3. I called it gradesml. Call yours some different name as it appears bucket names have to be unique across all of S3.



3 files.

	Amazon S3	> gradesml			
\bigcirc	Over	view	Propert	ies	Per
	🔔 Upload	+ Create folder	More ~		
				This buck	et is empt
			$ \geq $		

and make sure the permissions are set to **read**.



	<u> </u>		🗘 Walker Rowe 👻 Glob	al • Support •
Amazon S3 > gradesml				
student-por.csv Latest version 👻				
Overview Properties Permissio	15			
Access for your AWS account				
Account ()	Read object 0	Read object permissions ()	Write object permissions ()	
Critique_americain	Yes	Yes		
Access for other AWS accounts			Everyone	×
+ Add account Delete			Everyone will have access to one or all of	the following:
Account 0	Read object 0	Read object permissions 0	read this object, read and write permissio	ns.
Public access			Read object	
Group ()	Read object 0	Read object permissions 0	Access to this object's ACL	
O Everyone		1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1	Dead object nemissions Ca	ncel Save

🗨 Feedback \, Q English (US)

Amazon Machine Learning and then Create New Data Source/ML Model. If you have not used ML before it will ask you to sign up. Creating and evaluating models is free. Amazon charges you for using them to make prediction on a per 1,000 record basis.

Preserv Postery Terms of User Click on

orage).				Group	A
	\$	Machine Learning		AR & VR	
		Amazon SageMaker Amazon Comprehend		Amazon Sumerian 🗷	
		AWS DeepLens Amazon Lex	25 2 2 2 2 2	Application Integration	
		Machine Learning Amazon Polly		Step Functions	
		Rekognition		Simple Notification Service	
		Amazon Transcribe		Simple Queue Service	
5		Anazon Hanslate		SWF	
	<u></u>	Analytics		Customer Engagement	
		Athena		Amazon Connect	
		EMR		Pinpoint	
		CloudSearch		Simple Email Service	Click create new
Datasour	ce and I	ML model.			

🚓 Amazon Machine Learning 🗸

Objects

Create new Actions	-				
Datasource and ML model Datasource	t name or ID				
ML model		¢	Туре	¢	ID
Evaluation Batch prediction	model: banking		Batch predictio	n	bp-z56x
Evaluation: ML mod	el: banking		Evaluation		ev-o5yt6
ML model: banking			ML model		ml-BwZ5

banking [percentBegin=70, percentEnd=10... Datasource ds-TWFEFill in the **S3 location** below. Notice that you do not use the URL. Intead, put the bucket name and file name:

Click verify and Grant Permissions on the screen that pops up next.

now to access vo	pur data and give it permission to access it.
C2 location *	
53 location	s3:// gradesml/grades.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.</path-of-input-data>
source name	Student-por.csv
	The validation is successful. To go to the next step, choose Continue
	Datasource name Student-por.csv
	Data location s3://gradesml/grades.csv
	Data format CSV
	Schema source Auto generated
	Number of files 1
	Total size 67.6 KB
* Required	

Give the data

source some name then click through the screens. It fill make up field names (we actually don't care what names it uses since we know what each column means from the original data set). It will also determine whether each value is categorical (drawn from a finite set) or just a number. What is important for you to do is to pick the **target**. That is the dependant value you want it to predict, i.e., **y**. From the input data **student-por.csv** pick **G3**, as that is the student's final grade. These grades are from the Portuguese grammar school system and 13 is the highest value.

Below <u>don't use</u> students-por.csv as the input data. Instead use grades400.csv.

nport	t you	ir data to cre	ate an Amazor	ML datasource. Amazon	ML can use your datasource to crea	ate and evaluate an ML model, and vo	ou can u
		Where is	your data?	👔 S3	on Redshift		
<u> </u>	date			•			
530	Jata	access		data and she it seesing	»		
Tell A	Amaz	zon ML now	to access your	data and give it permission	n to access it.		_
		53	location *	s3:// gradesml/studen	t-por.csv		
				Enter the path to a single file or Learn more.	r folder in Amazon S3. You need to grant /	Amazon ML permission to read this data.	
				If you already have a schema fe	or this data, provide it in a file at s3:// <path you create one on the next name</path 	h-of-input-data>.schema. If you don't have a	
				second, remaining the time (100p)	vee er one on ere new paye. O		
		Datasou	irce name				
			Required	Decet			
eve put D her	Ama ata ma	Services v azon Machin 2. Schema 3	Resource Group e Learning . Target 4. Row ID	Datasources > Create da 5. Review	itasource	ूी Walker Rowe ┵ Ireland ┵ Suppor	rt -
av input D chel sazon M e data. T es the f CTION:	Ama hata ML scar This en first line	Services azon Machin 2. Schema 3 nned your input d nables Amazon M e in your CSV con pe type *	Resource Group e Learning • . Target 4. Row ID ata and inferred the co L to read the input dat tain the column name	Datasources > Create da 5. Review blamn names and data type for each of a correctly and to produce accurate pre s ² © Yes ® No O	ItaSOUICE the columns in your dataset. Review and edit the d edictions. Learn more.	Walker Rowe Ireland Suppor	t •
ev put D chel chel con M data. 1 s the f TION: Search	Ama Ama Mata HL scar This en first line Chang	Services azon Machinu 2. Schema 3 nned your input d nables Amazon M e in your CSV con ge type * wbute name	Resource Group e Learning • . Target 4. Row ID ata and inferred the cc to read the input dat tain the column name	Datasources > Create da 5. Review Sumn names and data type for each of a correctly and to produce accurate pre s ² • Yes * No •	ItaSOUICE the columns in your dataset. Review and edit the d dictions. Learn more.	Walker Rowe Treland Support tata type for each column to ensure that it accurately replace to the second	rt • Presents
av sput D chel uzon M data. 1 Tion: Search	Ama Ama Mata ML scar This en first line Chang b by ett	Services azon Machin 2. Schema 3 nned your input d hables Amazon M e in your CSV con ge type = Woute name Name	Resource Group e Learning Larget 4. Row ID ata and inferred the co L to read the input dat tain the column name Data type	Datasources > Create da S. Review Mamm names and data type for each of a correctly and to produce accurate pro s? Yes No Sample field value 1	It&SOUICE the columns in your dataset. Review and edit the di edictions. Learn more.	Walker Rowe • Ireland • Suppor ata type for each column to ensure that it accurately rep Items per page: 10 • < 1 - 10 o Sample field value 3	t • presents of 33 > »
av sput D sput D sthef TioN:	Ama Ama Mata ML scar This en first line Chang Chang A by art 1	Services azon Machin 2. Schema 3 nneed your input d hables Amazon M e in your CSV con ge type * vbute name Var01	Resource Group Learning Larget 4. Row ID target 4. Row ID tain the column name Data type Categorical =	Datasources > Create da S. Review	Itasource the columns in your dataset. Review and edit the d dictions. Learn more.	Waker Rove Tretand Support Lata type for each column to ensure that it accurately rep Items per page: 10 - 0 Sample field value 3 GP	t • presents
even D chel chel chel s the f TTION:	Ama Ama ma Missen first link chang h by aff 1 2	Services azon Machinu 2. Schema 3 med your input d hables Amazon M e in your CSV con ge type * troute name Var01 Var02	Resource Group e Learning Target 4. Row ID ata and inferred the co to read the input dat tain the column name Data type Categorical Categorical =	Datasources > Create da S. Review Dumn names and data type for each of a correctly and to produce accurate pro s? Yes No Sample field value 1 OP F	Itasource Ithe columns in your dataset. Review and edit the di edictions. Learn more. Sample field value 2 OP F	Walker Rowe → Ireland → Support atta type for each column to ensure that it accurately rep tata type for each column t	rt -
av aron M aron M ar	Ama Ama Mata Mata Mata Mata Mata Mata Mata Ma	Services azon Machin 2. Schema 3 nned your input d hables Amazon M e in your CSV con ge type • Voute name Var01 Var02 Var03	Resource Group e Learning • . Target 4. Row ID ata and inferred the co L to read the input dat tain the column name C Data type Categorical • Categorical • Numeric •	Datasources > Create da S. Review Determin names and data type for each of a correctly and to produce accurate pre s? Yes No Sample field value 1 GP F 18	ttasource the columns in your dataset. Review and edit the d edictions. Learn more.	Walker Rowe Ireland Support tata type for each column to ensure that it accurately rep tata type	rt -
av aput D cher xzon M data. 1 s the f TTION: Search	Ama hata maa fil, scar This en Chang Chang 1 1 2 3 4	Services azon Machin 2. Schema 3 need your input d ables Amazon M e in your CSV con ge type bybute name Var01 Var01 Var02 Var03 Var04	Resource Group e Learning A. Row ID Target 4. Row ID ta and inferred the co to read the input dat tain the column name Categorical • Categorical • Numeric • Categorical •	Datasources > Create da S. Review Dumn names and data type for each of a correctly and to produce accurate pro s? Yes No C Sample field value 1 C D F 18 U U C C C C C C C C C C C C C C C C C	Itasource Ithe columns in your dataset. Review and edit the di dictions. Learn more.	Walker Rowe Ireland Support Items per page: 10+ (1-10 o Sample field value 3 GP F 15 U	rt -
av aput D che azon N data. 1 s the f TTION: Search	Ama Inta ML scar This en first line chang b by art 1 2 3 4 5 5	Services azon Machin 2. Schema 3 nned your input d hables Amazon M e in your CSV con pe type trobute name Var01 Var02 Var03 Var04 Var05 Nam5	Resource Group e Learning A. Row ID Target 4. Row ID to read the input dat tain the column name Data type Categorical Categorical • Categorical • Categorical • Categorical • Categorical •	Datasources > Create da 5. Review burn names and data type for each of a correctly and to produce accurate pro- s? Yes No ● Sample field value 1 GP F 18 U GT3 A	ttasource the columns in your dataset. Review and edit the de dictions. Learn more. Sample field value 2 OP F 17 U GT3 T	Walker Rose Inteland Support A Walker Rose Inteland Support A Support A Support A Support A Support A Support A	resents
event D hput D hcon M data. 1 s the f TTION: Search	Ama atta ML scar first find first find first find a 1 2 3 4 5 6 7	Services <	Resource Group e Learning Larget 4. Row ID ata and inferred the co to read the input dat tain the column name Data type Categorical Categorical Categorical Categorical Categorical	Datasources > Create da S. Review Dumn names and data type for each of a correctly and to produce accurate pre P Sample field value 1 GP F 18 U GT3 A 4	Itasource Ithe columns in your dataset. Review and edit the di edictions. Learn more.	Weaker Rowe Tretand Support A Waker Rowe Tretand Support A Support A Support A Support A Support A Support A Support A Support A Support A Support A Support A Support A Support A Support A Support	rt -
av sput D che tron M s the f TTON: Search	Ama hata maa tL scar first link chang first link chang 1 2 3 4 5 6 7 8	Services azon Machin 2. Schema 3 nned your input d hables Amazon M e in your CSV con pe type Woole name Var01 Var02 Var03 Var04 Var05 Var05 Var05 Var07 Var08	Resource Group e Learning Learning Target 4. Row ID ata and inferred the co L to read the input dat tain the column name Categorical	Datasources > Create da S. Review Datasources > Create da S. Review Datasources > Create da S. Review Datasources and data type for each of a correctly and to produce accurate pro s? Yes No Sample field value 1 GP F 18 U GT3 A 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4	tasource the columns in your dataset. Review and edit the deficitors. Learn more. Sample field value 2 GP GP GP GT GT GT GT GT I I I I I I I I I I I I	Walker Rose • Ireland • Support Lata type for each column to ensure that it accurately reg Bample field value 3 GP IS U LE3 T 1 1	presents
ev av av av av azon M data. 1 s the fi TTION: Search	Ama atta ma fil, scaar first links chang c	Services azon Machin 2. Schema 3 ables Amazon M e in your CSV con ge type Var01 Var02 Var03 Var04 Var05 Var05 Var05 Var05 Var08 Var09 Var0	Resource Group e Learning A. Row ID Target 4. Row ID tata and inferred the co L to read the input dat tain the column name Categorical • Categorical •	 Datasources > Create data 5. Review Sample field value 1 GP β 18 U GT3 A 4 at_home 	ttasource the columns in your dataset. Review and edit the d defections. Learn more. Sample field value 2 GP GP GP G G G G G G G G G	Waker Rove	rt -



-	1 0F00117 TV
ID	ml-Q5G6ld7g7Xj
Name	ML model: training <i>F</i>
Type Creation time	Numerical regression
Completion time	Not available
Compute Time (Approximate)	Not available O
Status	Pending
Log	Not available
Datasource (training)	
Datasource ID	ds-E9YJUuZ0NWU
Target	_Target_
Input schema	View input schema
Evaluations	
Evaluations created	1
Latest evaluation result	Not available
	Perform another Evaluation
Predictions	
CloudWatch metrics	Ciew in CloudWatch
	A single dataset
	Generate one-time predictions for a single dataset.
	Generate batch predictions
	Generate batch predictions Try real-time predictions
	Generate batch predictions Try real-time predictions Generate real-time predictions in your browser.
	Generate batch predictions Try real-time predictions in your browser. Try real-time predictions
	Generate batch predictions Try real-time predictions in your browser. Try real-time predictions Enable real-time predictions
	Generate batch predictions Try real-time predictions Generate real-time predictions in your browser. Try real-time predictions Enable real-time predictions To enable real-time predictions now, create a real-time prediction endpoint.

While waiting

are create another **data set**. This is not a model so it will not ask you for a target. Use the **grades249.csv** file in S3, which we will use in the **batch prediction** step.

Ohiasta

Objects						
Create new Actions -						
Filter: All types V Q Object name or ID					Items per	r p
Name 🗘	Туре	¢	ID \$	•	Status 🗘	;
predicition	Datasource		ds-yBotR7rXRo5		In progress	
Evaluation: ML model: training	Evaluation		ev-1XUCxHi1MzG		Completed	
ML model: training	ML model		ml-Q5G6ld7g7Xj		Completed	
training IngroantRagin=70 parcontEnd=1	Datacourco		de Sul0e76DBSr		Completed	

evaluation is done. We can see which one it is from the list above as it says **evaluation**. Click on it. We explain what it means below.

Name Evaluation: ML model: training ♪ Datasource ID ds-8yI0c76R88r Output location Not available Creation time Mar 15, 2018 12:17:24 PM Completion time 3 mins. ③ Compute Time (Approximate) 2 mins. ④ Status Completed Log Download log	Name Datasource ID Output location Creation time	Evaluation: ML model: training 🖋 ds-8yI0c76RB8r Not available Mar 15, 2018 12:17:24 PM
Datasource ID ds-8yl0c76RB8r Output location Not available Creation time Mar 15, 2018 12:17:24 PM Completion time 3 mins. I Compute Time (Approximate) 2 mins. I Status Completed Log Download log	Datasource ID Output location Creation time	ds-8yI0c76RB8r Not available Mar 15, 2018 12:17:24 PM
Output location Not available Creation time Mar 15, 2018 12:17:24 PM Completion time 3 mins. I Compute Time (Approximate) 2 mins. I Status Completed Log Download log model performance Vertical status score is better than the baseline. I Status Status score is better than the baseline. I	Output location Creation time	Not available Mar 15, 2018 12:17:24 PM
Creation time Mar 15, 2018 12:17:24 PM Completion time 3 mins. Compute Time (Approximate) 2 mins. Status Completed Log Download log model performance	Creation time	Mar 15, 2018 12:17:24 PM
Completion time 3 mins. Compute Time (Approximate) 2 mins. Status Completed Log Download log model performance On your most recent evaluation, ev-1XUCxHi1MzG , the ML model's quality score is better than the baseline.	Completion time	mai 15, 2010 12.11.24 Fm
Compute Time (Approximate) 2 mins. Completed Status Completed Log Download log model performance On your most recent evaluation, ev-1XUCxHi1MzG , the ML model's quality score is better than the baseline. Complete that the baseline is	Completion time	3 mins. 0
Status Completed Log Download log Log Download log On your most recent evaluation, ev-1XUCxHi1MzG , the ML model's quality score is better than the baseline. Image: Completed complete	Compute Time (Approximate)	2 mins. 0
Log Download log L model performance On your most recent evaluation, ev-1XUCxHi1MzG , the ML model's quality score is better than the baseline. RMSE: 1.7457 RMSE baseline: 2.933	Status	Completed
Con your most recent evaluation, ev-1XUCxHi1MzG , the ML model's quality score is better than the baseline.	Log	Download log
	On your most recent evaluation, ev-1XUCxHi1MzG , RMSE: 1.7457 RMSE baseline: 2.933	the ML model's quality score is better than the baseline. ()

Amazon shows

the RMSE. This is the square root of the sum of the squared differences of the observed and predicted values. We square and then take the square root so that all the numbers are positive, so they do not cancel each other out. Amazon also uses the mean, meaning average, by multiplying this sum by 1 / n, where n is the sample size.

If the model and the evaluations were the same, this number would be 0. So the closer to o zero we get the more accurate is our model. If the number is large, then the problem is not the algorithm, it is the data. So we could not pick another algorithm to make it much better. There is really only one algorithm used for LR, finding the least squares error. (There are more esoteric ones.) If MSE number is large then either the data is not correlated or, more like, most of the data is correlated, but some

of it is not and is thus messing up our model. What we would do is drop some columns out and rebuild out model to get a more accurate model.

What value means the model is good? The model is good when the distribution of errors is a normal distribution, i.e., the bell curve.

Put another way, click Explore Model Performance.



No taos

See the

histogram above. Numbers to the left of the dotted line are where the predicted values were less than the observed ones. Numbers to the right are where they are higher. If this distribution were entered on the number 0 then we would have a completely random distribution. That is the idea situation where our errors are distributed randomly. But since it is shifted there is something in our data that we should leave out. For example, family size might not be correlated to grades.

Above Amazon showed the **RMSE baseline**. This is what the RMSE would be if we could have an input data set in which there was this perfect distribution of errors.

Also here we see the limitations of doing this kind of analysis in the cloud. If we have written our own program we could have calculated other statistics that showed exactly which column was messing up our model. Also we could try different algorithms to get rid of the bias caused by **outliers**, meaning numbers far from the mean that distort the final results.

Run the Prediction

Now that the model is saved, we can use it to make predictions. In other words we want to say given these student characteristics what are their likely final grades going to be.

Select the **prediction** datasource you created above then select **Generate Batch Predictions**. Then click through the following screens.

ion 💡

ID	ds-yBotR7rXRo5	
Name	predicition 🖋	
Creation time	Mar 15, 2018 12:25:39 PM	
Completion time	4 mins. 💿	
Compute Time (Approximate)	13 mins. 📵	
Status	Completed	
Message	Not available	
Input schema	View input schema	
Log	Download log	
	Use this datasource to 🕶	
	Copy settings to create a new datasource Create (train) an ML model	
001	Evaluate an ML model	
S3 location	Generate batch predictions	
Number of files		
Data format	24.6 // P	
Data magnetical size	34.0 ND	
Data rearrangement	None	
1. ML model for batch prediction 2. Data for batch prediction	3. Batch prediction results 4. Review	
ML model for batch prediction		
Choose the ML model to use for generating batch predictions. Batch Select an ML model	predictions generate predictions all at once for a large number of data records	
Q Search All ML models by name or ID		
ML model name ML model: training		Change ML model
ML model ID mL0566M2n706	Innut schema	

del ID ml-Q5G6ld7g7Xj ML model type Numerical regression Target attribute _Target_ Target type NUMERIC Creation time Mar 15, 2018 12:17:24 PM Status Completed Number of attributes 33 Datasource ID ds-E9YJUuZ0NWU Evaluations created 1 Latest evaluation 1.746 (RMSE) result Log Download log Batch predictions 0 created Tags No tags • You selected ML model mI-Q5G6Id7g7Xj. To go to the next step, choose Continue Cancel Continue

ML model settings

You can use the automa	atically suggested ML model settings, or you can ch	oose to customize.	
ML model type	REGRESSION 1		
ML model target	_Target_		
ML model name (Optional)	ML model: training		
Select training and evaluation settings	Recipes and training parameters control the ML r for your ML model or use the defaults provided by Amazon ML reserve a portion of the input data fo	nodel training process. You can select th y Amazon ML. In either case, you can ch r evaluation. Learn more.	ese settings oose to have
	 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 ⁽¹⁾ 	 Custom Modify the recipe Amazon ML generates Modify training parameters Randomly or sequentially split evaluation data ^(*) 	your
Evaluation Name	Evaluation: ML model: training		
then create ML	. model.	Cancel Previous	Click review
1. ML model for batch predic	tion 2. Data for batch prediction 3. Batch prediction re	esults 4. Review	
Data for batch	prediction		
Locate the input data to use the Locate the input data (*) You selected ML model million Enter the datasource na	or the batch prediction. Learn more about S3 permissions. already created a datasource pointing to my S3 data O M -Q5G6Id7g7Xj	y data is in S3, and I need to create a datasource	
Datasource na	me predicition		Cha
Datasouro Creation t Sta Datasource t S3 local	e ID ds-yBotR7rXRo5 ime Mar 15, 2018 12:25:39 PM itus Completed ype S3 ition s3://gradesml/grades249.csv	Input sohema Target attribute Target type Number of attributes Models trained	View input schema 33 0
Data rearrangen Tags	ent None	Batch predictions created	0

No tags

Here we tell it

where to save the results in S3. There it will save several files. The one we are interested in is the one where it calculates the **score**. It should tack it onto the input data to make it easier to read. But it does not. So I have pasted it into <u>this spreadsheet</u> for you on the sheet called prediction and added back the column headings. I also then added a column to show how the MSE mean squared error is calculated.

1. ML model for batch prediction	Data for batch prediction	3. Batch prediction results	4. Review
----------------------------------	---	-----------------------------	-----------

Batch prediction results

The estimated cost for generating your predictions is \$0.10. This estimate is based on the 249 data records included in your prediction rec The Amazon ML fee for batch predictions is \$0.10 per 1,000 predictions, rounded up to the next 1,000. Learn more. Type the path to the S3 location in which the prediction results will be saved. S3 destination s3:// gradesml/predictions.csv Batch prediction name Batch prediction: ML model: training (Optional) Cancel Previous Review aws Services 👻 Resource Groups 👻 象 Amazon Machine Learning - Batch Predictions > Create batch prediction 1. ML model for batch prediction 2. Data for batch prediction 3. Batch prediction results 4. Review Review Review and make any changes, and then click Finish. Edit ML model for batch prediction ML model Name ML model: training ML model ID mI-O5G6Id7g7Xi Edit Data for batch prediction Datasource name prediction Data location s3://gradesml/grades249.csv Edit Batch prediction results Output location s3://gradesm//predictions.csv Batch prediction name Batch prediction: ML model: training Cost Estimate The estimated cost for generating your predictions is \$0.10. This estimate is based on the 249 data records included in your prediction request The Amazon ML fee for batch predictions is \$0.10 per 1,000 predictions, rounded up to the next 1,000. Learn more Tags 0 Amazon ML copies a maximum of 10 tags from parent objects. Edit the list to keep the tags you need. No tags Cancel Previous Create batch predicts 🗬 Feedback 🛛 😧 English (US)

ID	bp-ebhjggKYchO
Name	Batch prediction: ML model: training 🥒
Creation time	Mar 15, 2018 12:43:38 PM
Completion time	Not available 🕚
Compute Time (Approximate)	Not available 🟮
Status	In progress
Datasource ID	ds-yBotR7rXRo5
ML model ID	ml-Q5G6ld7g7Xj
Input S3 URL	s3://gradesml/grades249.csv
Output S3 URL	s3://gradesml/predictions.csv/
Log	Not available
Processing information	
Number of records seen	Not available
Records that failed to process	Not available

As you can see,

it saves the data in S3 in a folder called **predictions.csv**. In this case it gave the prediction values in a file with this long name **bp-ebhjggKYchO-grades249.csv.gz**. You cannot view that online in S3. So download it showing the URL below and look at it with another tool. In this case I pasted the data into Google Sheets.

Overview	Properties	Permissions		
Q Type a prefix and press Enter	to search. Press ESC to clear.			
🚣 Upload 🕂 Create folder	More ~			
□ Name ↑=		Last		
predictions.csv		-		
grades249.csv		Mar 0400		
grades400.csv		Mar 0400		

Overview	Properties	Permissions	
Open Downlo	ad Download as	Make public	Copy path
Owner amazon-machine-lea	rning-admin+dub		
Last modified Mar 15, 2018 12:44::	38 PM GMT-0400		
Etag 6e04d3aa1bb4d0f2a	af9131e08dedb45e		
Storage class Standard			
Server side encrypti None	on		
Size			

Download the

data like this:

wget

https://s3-eu-west-1.amazonaws.com/gradesml/predictions.csv/batch-prediction/ result/bp-ebhjggKYch0-grades249.csv.gz

Here is that the data looks like with the prediction added to the right to make it easy to see. Column AG is the student's actual grade. AH is the predicted value. AI is the square of the difference. And then at the bottom is MSE.

AC	AD		AE	AF	AG	AH	AI
health	absences	G1		G2	G3	score	(obs-pred) sqrd
	4	4	15	14	17	13.30	13.70998729
	5	0	14	13	14	12.18	3.311817626
	4	0	11	12	13	12.68	0.1035101929
	1	10	12	15	15	14.13	0.7519317796
	4	4	12	16	16	13.11	8.379114409
	5	16	10	11	11	8.30	7.28465498
	3	6	10	13	13	11.27	2.991585344
	5	0	9	12	12	10.49	2.270657197
	3	11	9	11	12	9.44	6.53817229
	2	9	13	14	15	12.14	8.19110124
	4	0	13	17	17	14.61	5.729368832
	4	2	12	15	15	13.06	3.750729422
	3	0	14	17	17	15.52	2.18750016
	A	21	0	10	10	7 26	7 5001083/5