**Content**

[1. What to learn 3](#_Toc21801135)

[2. AWS Provided Sample Questions 4](#_Toc21801136)

[1. TF-IDF (Term Frequency – Inverse Document Frequency) 4](#_Toc21801137)

[2.1.1. To learn more about it 4](#_Toc21801138)

[2. Fraud Detection 4](#_Toc21801139)

[3. Why Glue? 5](#_Toc21801140)

[4. Batch size and Learning rate in practice 5](#_Toc21801141)

[3. LinuxAcademy practice exam 6](#_Toc21801142)

[1. Difference between Kinesis Data Stream and Kinesis Firehose 6](#_Toc21801143)

[3.1.1. Kinesis Data Analytics 7](#_Toc21801144)

[3.1.2. Kinesis Firehose 7](#_Toc21801145)

[2. Characteristics of box plot 8](#_Toc21801146)

[3. Loss Functions (e.g. RMSE) 9](#_Toc21801147)

[4. Correlation 9](#_Toc21801148)

[5. CloudWatch Events – usecases 9](#_Toc21801149)

[6. SageMaker EI (Elastic Inference) 9](#_Toc21801150)

[4. Whizlabs – Free 15 questions 10](#_Toc21801151)

[1. targetAttributeName – data descriptors 10](#_Toc21801152)

[2. Hyperparameters 10](#_Toc21801153)

[3. Glue in more depth 10](#_Toc21801154)

[4. Some additional model evaluation methods 10](#_Toc21801155)

[5. KPI Charts 10](#_Toc21801156)

[6. Net Promoter Score (NPS) 10](#_Toc21801157)

[7. Data Imputation 11](#_Toc21801158)

[8. PR (Precision – Recall) Curve 11](#_Toc21801159)

[9. Binning 11](#_Toc21801160)

[5. AWS practice exam 12](#_Toc21801161)

[6. Quizlet and LinuxAcademy Flashcards 13](#_Toc21801162)

[1. SMOTE 13](#_Toc21801163)

[2. Apache Hive 13](#_Toc21801164)

[3. Learning Rate vs Batch Size 13](#_Toc21801165)

[4. Metrics 13](#_Toc21801166)

[6.4.1. Recall 14](#_Toc21801167)

[6.4.2. Precision 14](#_Toc21801168)

[5. Undersampling & Oversampling 14](#_Toc21801169)

[6. Threshold 14](#_Toc21801170)

[7. Others 16](#_Toc21801171)

[1. Collinearity 16](#_Toc21801172)

1. What to learn

Big data – kinesis and EMR.

1. learn 44 -> 35

2. whizlabs

3. quizlet

4. LA flashcards

5. LA flashcards

6. AWS Practice questions

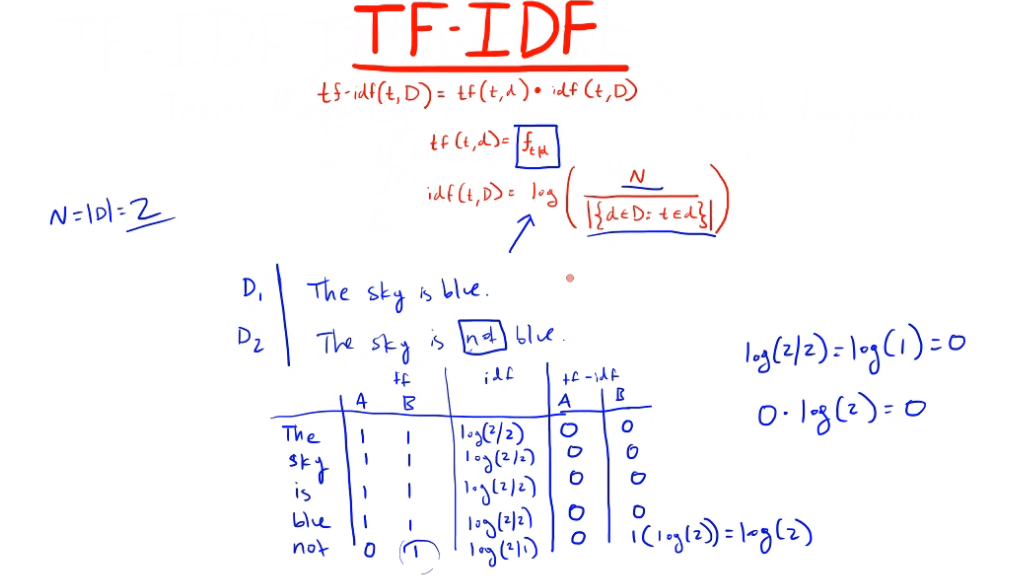
7. Practice Exam

8. Google Practice exams

9. Google ML course

1. AWS Provided Sample Questions
   1. TF-IDF (Term Frequency – Inverse Document Frequency)

The weight increases the more times a term occurs in the document, but that’s offset by the number of times the word appears in the entire dataset or corpus. This offset helps to remove the importance from really common words, like “the”, “or” “a”… Relevance ranking.



. It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling. The tf–idf value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general. tf–idf is one of the most popular term-weighting schemes today; 83% of text-based recommender systems in digital libraries use tf–idf.

* + 1. To learn more about it
* Dampening
* magnitude
  1. Fraud Detection

Fraud detection is a set of activities undertaken to prevent money or property from being obtained through false pretenses. Fraud detection is applied to many industries such as banking or insurance. In banking, fraud may include forging checks or using stolen credit cards. Other forms of fraud may involve exaggerating losses or causing an accident with the sole intent for the payout.

* 1. Why Glue?

AWS Glue is the correct answer because this option requires the least amount of setup and maintenance since it is serverless.

* 1. Batch size and Learning rate in practice

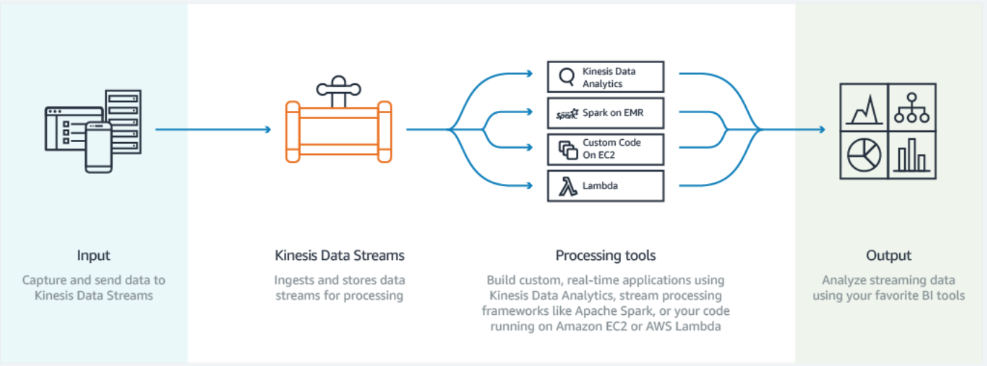
It is most likely that the **loss function is very curvy** and has **multiple local minima** where the training is getting stuck. **Decreasing** the **batch size** would help the Data Scientist stochastically **get out of the local minima saddles**. **Decreasing** the **learning rate** would **prevent overshooting the global loss function minimum**.

1. LinuxAcademy practice exam
   1. Difference between Kinesis Data Stream and Kinesis Firehose

*Firehose is send data to somewhere. Kinesis data streams on the other hand store data for up to 7 days for you to work with process.*

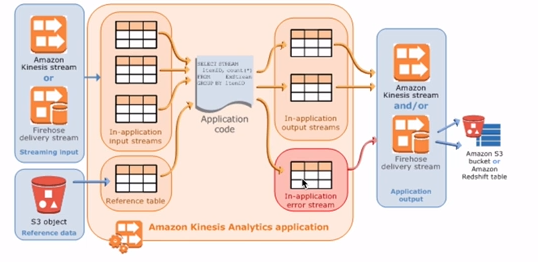
The main difference:

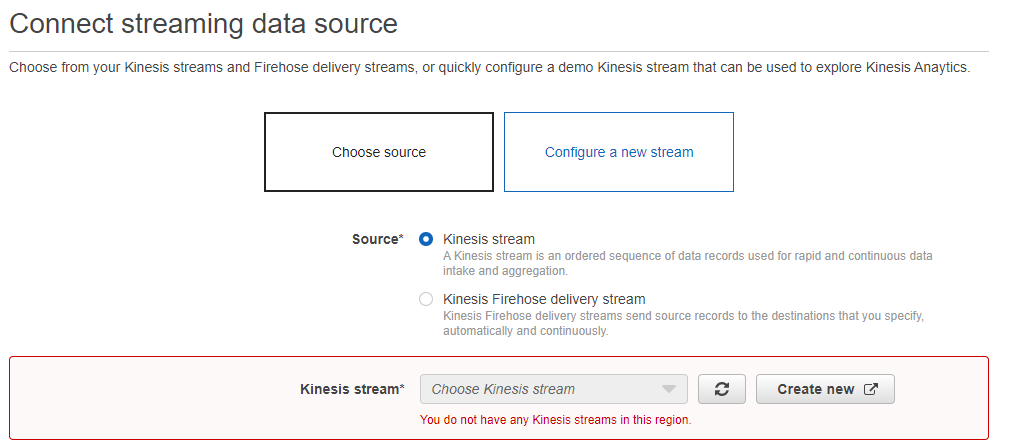
* Kinesis Streams just store the data in the stream. You should setup consumers to do something with it. So a lambda function and an EC2 instance can be reading from that stream at the same time and generating something or transforming data. Then an action can be taken, like send someone a message or store data in S3. Capture process and store data streams, **custom** application.
* Kinesis Firehose is meant to be used when you want to collect streaming data and store it in one of the 4 destinations. S3, Redshift, Splunk, Elasticsearch. Not used for real-time analytics. Can also transform data before delivering it. Load data streams into AWS data source. **Delivery stream.**





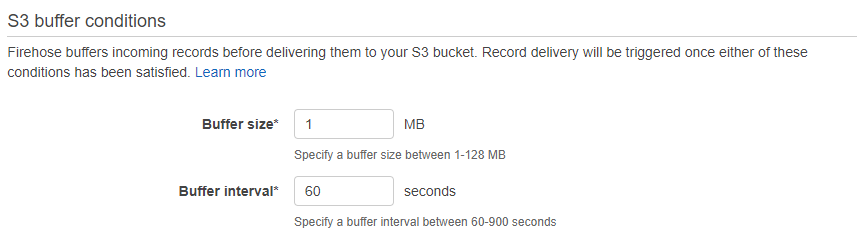
* + 1. Kinesis Data Analytics

Kinesis Data Analytics – Analyze data streams with SQS or Java. 



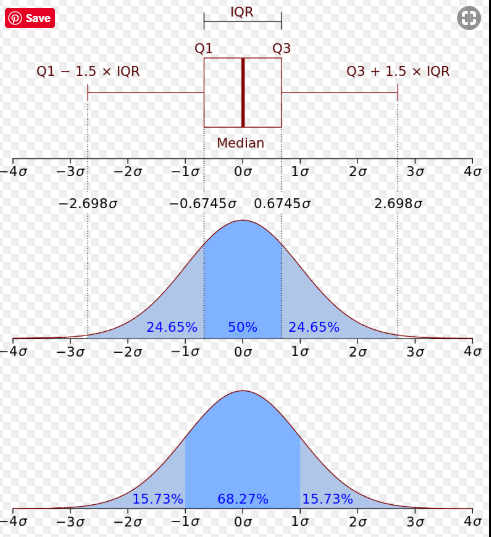
* + 1. Kinesis Firehose

As soon as one of these conditions become true, the buffer will be flushed and data will be written into S3. (it can be ES, RedShift or Splunk too).



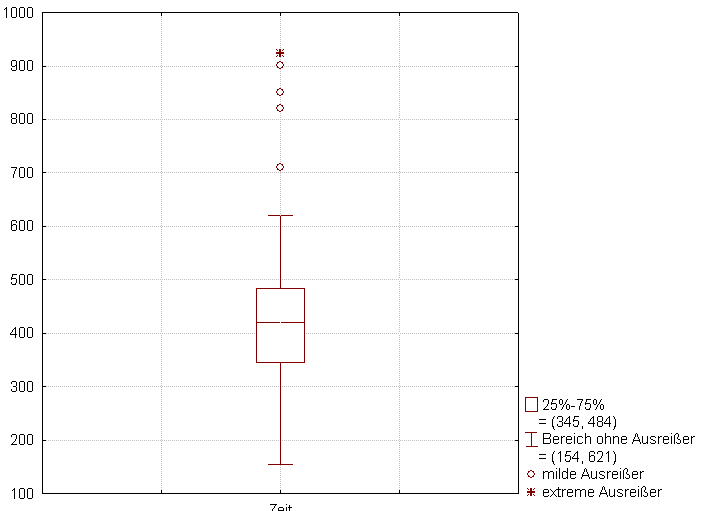
* 1. Characteristics of box plot

Boxplot and probability density function.



The points shown outside the ends of the whiskers are outliers or suspected outliers.

The ends of the lines (or whiskers) extending from the box indicate the minimum and maximum data values – unless outliers are present in which case the whiskers extend to a maximum of 1.5 times the inter-quartile range.



* 1. Loss Functions (e.g. RMSE)

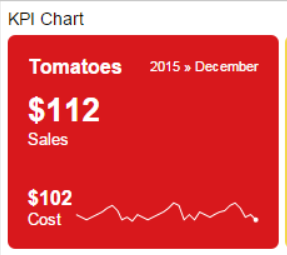
It's a method of evaluating how well your algorithm models your dataset. If your predictions are totally off, your loss function will output a higher number. If they're pretty good, it'll output a lower number. As you change pieces of your algorithm to try and improve your model, your loss function will tell you if you're getting anywhere.

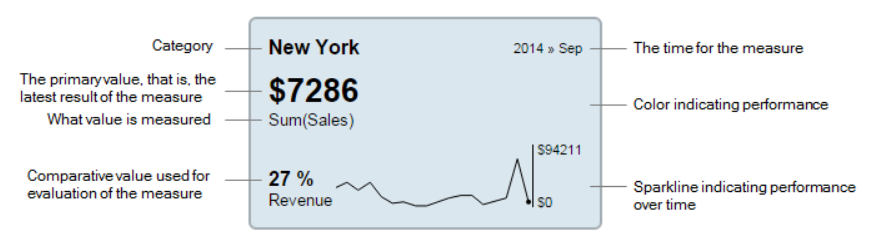
RMSE is a loss function for regression, but there are others for other types of problems.

* 1. Correlation
  2. CloudWatch Events – usecases
  3. SageMaker EI (Elastic Inference)

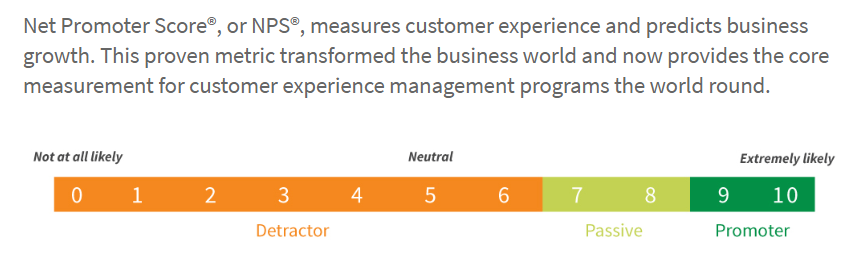
Elastic Inference (EI) is a resource you can attach to your EC2 CPU instances to accelerate your deep learning (DL) inference workloads. Amazon Elastic Inference allows you to attach GPU-powered acceleration to many amazon machine instances in order to reduce the cost of running deep learning inference. EI supports TF, MXNet, ONNX (Open Neural Network Exchange) thorugh MXNet.

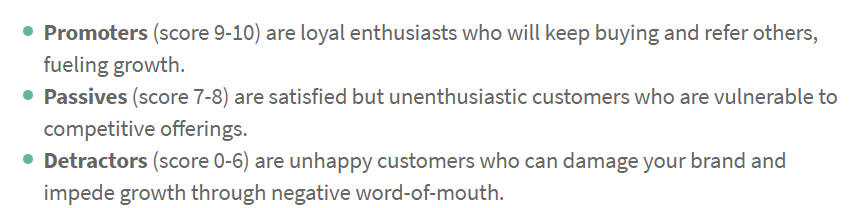
1. Whizlabs – Free 15 questions
   1. targetAttributeName – data descriptors
   2. Hyperparameters
   3. Glue in more depth
   4. Some additional model evaluation methods
   5. KPI Charts

The KPI chart is used to, at a quick glance, give information about the current performance of a company or organization. Factors, which are crucial for monitoring how the company performs, are measured and then presented in form of KPIs, Key Performance Indicators. The type of information that is shown varies. Examples of KPIs are information about net revenue, sales growth, and customer satisfaction.



* 1. Net Promoter Score (NPS)



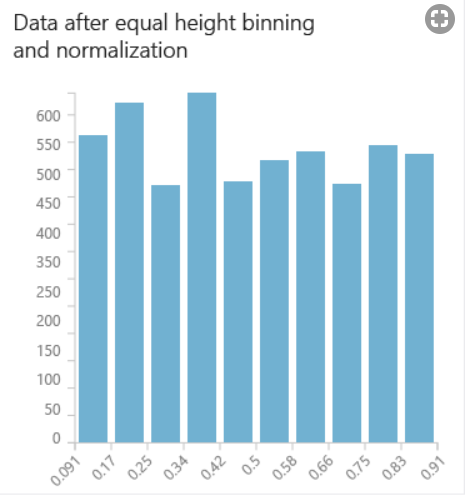


* 1. Data Imputation

How to replace the missing values in the dataset.

* 1. PR (Precision – Recall) Curve
  2. Binning

Binning or grouping data (sometimes called quantization) is an important tool in preparing numerical data for machine learning, and is useful in scenarios like these:

* A column of continuous numbers has too many unique values to model effectively, so you automatically or manually assign the values to groups, to create a smaller set of discrete ranges.
* For example, you could use entropy scores generated by Group Data into Bins to identify the optimal groupings of data values, and use those groups as features in your model.
* Replace a column of numbers with categorical values that represent specific ranges.
* For example, you might want to group values in an age column by specifying custom ranges, such as 1-15, 16-22, 23-30, and so forth for user demographics.
* A dataset has a few extreme values, all well outside the expected range, and these values have an outsized influence on the trained model. To mitigate the bias in the model, you might transform the data to a uniform distribution, using the quantiles (or equal-height) method.

1. AWS practice exam
2. Quizlet and LinuxAcademy Flashcards
   1. SMOTE

With datasets that are not fully populated, the Synthetic Minority Over-sampling Technique (SMOTE)

adds new information by adding synthetic data points to the minority class. This technique would be the

most effective in this scenario.

* 1. Apache Hive
  2. Learning Rate vs Batch Size
  3. Metrics

<https://www.mikulskibartosz.name/precision-vs-recall-explanation/>

Recall = TP / (TP + FN)

Precision = TP / (TP + FP)

Accuracy = TP + TN / Total

AUC - Area Under the ROC Curve (AUC) measures the ability of a binary ML model to predict a higher score for positive examples as compared to negative examples.

F1 - The macro-averaged F1-score is used to evaluate the predictive performance of multiclass ML models. This puts recall and precision together.

RMSE - The Root Mean Square Error (RMSE) is a metric used to evaluate the predictive performance of regression ML models.

Cut-off - ML models work by generating numeric prediction scores. By applying a cut-off value, the system converts these scores into 0 and 1 labels.

Accuracy - Accuracy measures the percentage of correct predictions.

Precision - Precision shows the percentage of actual positive instances (as opposed to false positives) among those instances that have been retrieved (those predicted to be positive). In other words, how many selected items are positive?

Recall - Recall shows the percentage of actual positives among the total number of relevant instances (actual positives). In other words, how many positive items are selected?

* + 1. Recall

TP / (TP + FN)

Recall answers What proportion of actual positives was identified correctly?

AWS: Recall shows the percentage of actual positives among the total number of relevant instances (actual positives). In other words, how many positive items are selected?

Think - If I was to recall what I did 10 years ago, I probably will guess something right, but there are other things that I will not guess because I forgot them. E.g. It is what I guessed right over (what i guess right + what I should have flagged as well)

* + 1. Precision

Precision answers What proportion of positive identifications was actually correct?

Precision = TP / (TP + FP)

AWS: Precision shows the percentage of actual positive instances (as opposed to false positives) among those instances that have been retrieved (those predicted to be positive). In other words, how many selected items are positive?

Think: Trying to be precise, but you ended up marking things as something that they aren't. E.g. What you guess right over (what you guess right + what we flagged, but we were wrong)

* 1. Undersampling & Oversampling

Resampling in data analysis are techniques used to adjust the class distribution of a data set (i.e. the ratio between the different classes/categories represented)

* 1. Threshold

what about an email message with a prediction score of 0.6? In order to map a logistic regression value to a binary category, you must define a classification threshold (also called the decision threshold). A value above that threshold indicates "spam"; a value below indicates "not spam." It is tempting to assume that the classification threshold should always be 0.5, but thresholds are problem-dependent, and are therefore values that you must tune.

1. Others
   1. Collinearity
   2. Deploy custom algorithm using SageMaker