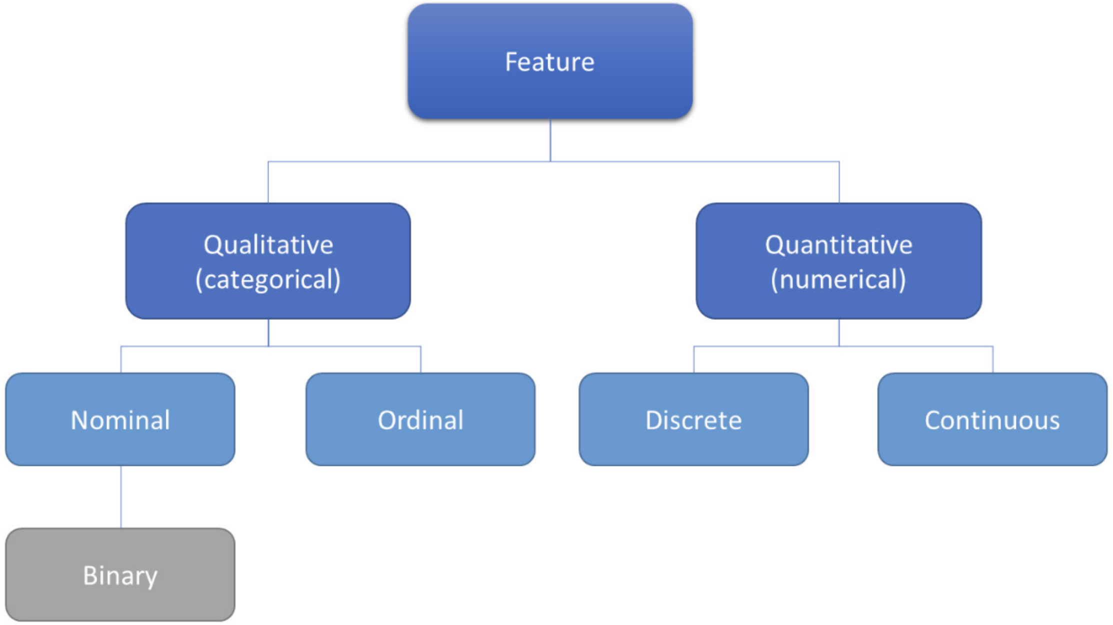
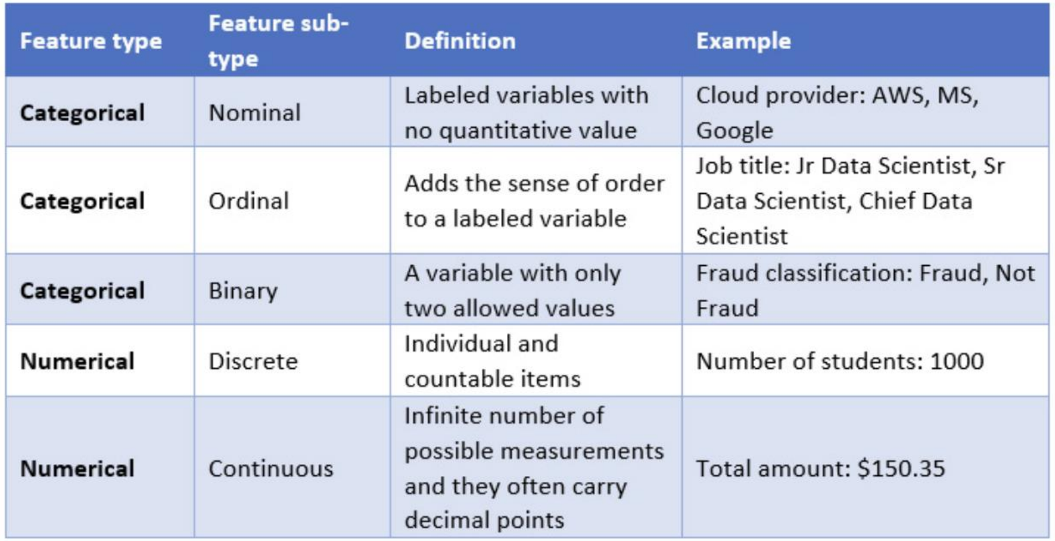
Tech experts

Semana 2

Tipos de dados





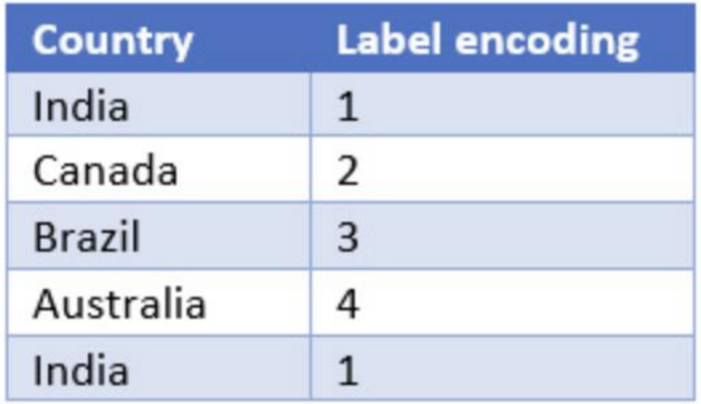
Feature type	Feature sub-type	Definition	Example
Categorical	Nominal	Labeled variables with no quantitative value	Cloud provider: AWS, MS, Google
Categorical	Ordinal	Adds the sense of order to a labeled variable	Job title: Jr Data Scientist, Sr Data Scientist, Chief Data Scientist
Categorical	Binary	A variable with only two allowed values	Fraud classification: Fraud, Not Fraud
Numerical	Discrete	Individual and countable items	Number of students: 1000
Numerical	Continuous	Infinite number of possible measurements and they often carry decimal points	Total amount: $150.35

Tipos de dados

- Some algorithm implementations, such as **`scikit-learn`**, may not accept string values on your categorical features.
- The data distribution of your variable may not be the most optimal distribution for your algorithm.
- Your ML algorithm may be impacted by the scale of your data.
- Some observations (rows) of your variable may be missing information and you will have to fix it. These are also known as missing values.
- You may find outlier values of your variable that can potentially add bias to your model.
- Your variable may be storing different types of information and you may only be interested in a few of them (for example, a date variable can store the day of the week or the week of the month).
- You might want to find a mathematical representation for a text variable.

Label Encoding

- Some algorithm implementations, such as **`scikit-learn`**, may not accept string values on your categorical features.

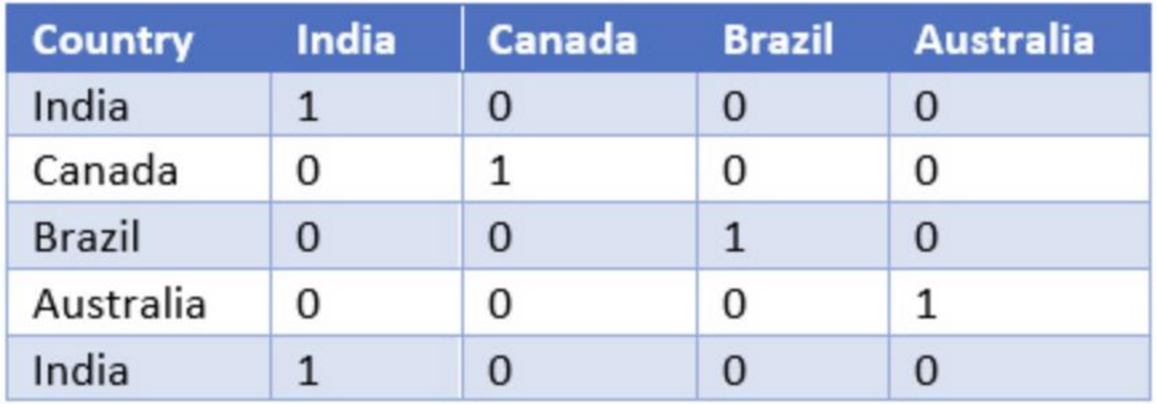


Country	Label encoding
India	1
Canada	2
Brazil	3
Australia	4
India	1

ANTECAO COM A ORDEM

One Hot Encoding

- Some algorithm implementations, such as **`scikit-learn`**, may not accept string values on your categorical features.

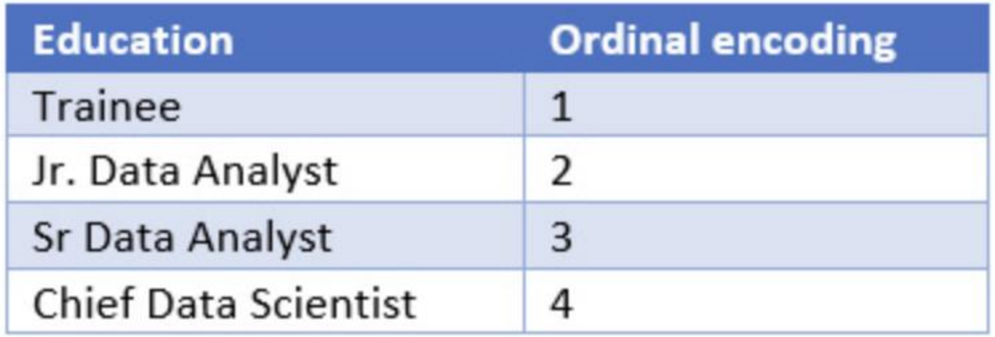


Country	India	Canada	Brazil	Australia
India	1	0	0	0
Canada	0	1	0	0
Brazil	0	0	1	0
Australia	0	0	0	1
India	1	0	0	0

BOM QUANDO TEMOS POUCAS CATEGORIAS, PARA MUITAS CATEGORIAS PODE SER UM PROBLEMA

Ordinal Encoding

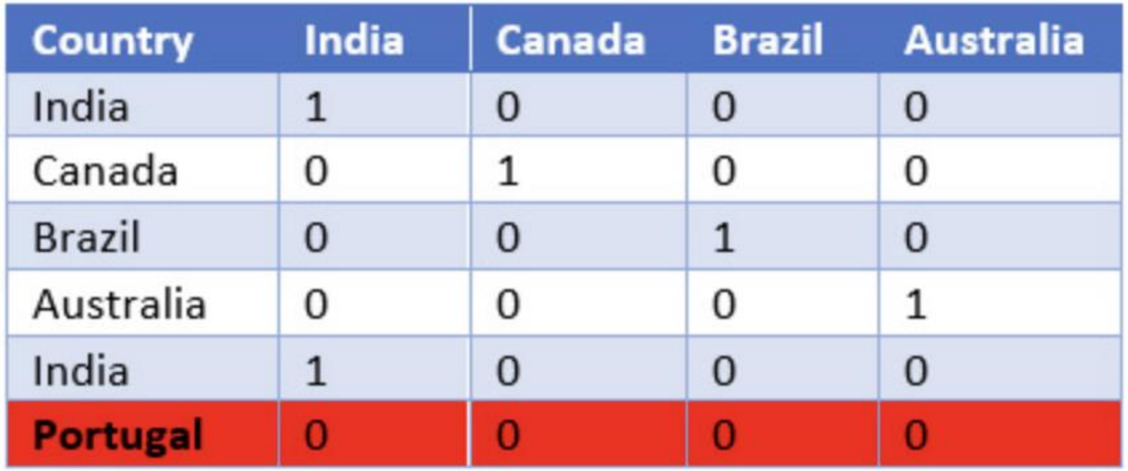
- Some algorithm implementations, such as **`scikit-learn`**, may not accept string values on your categorical features.



Education	Ordinal encoding
Trainee	1
Jr. Data Analyst	2
Sr Data Analyst	3
Chief Data Scientist	4

Dados de treino e teste

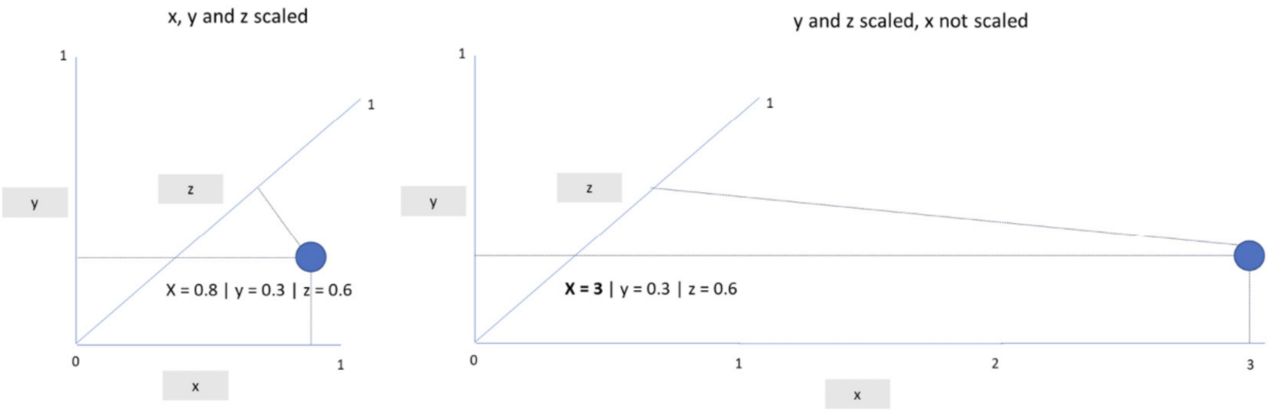
- Sempre aplicar mesmas transformações que foram executadas nos dados de treino nos dados de teste
- Os encoders devem ser “fitados” baseado nos dados de treino e aplicado nos dados de teste, ou seja, não se deve realizar o encoding em todos os dados e depois realizar o split.
- Os mesmos tratamentos devem ser aplicados na predição.
- Sempre se atentar com a distribuição dos dados de treino e teste.

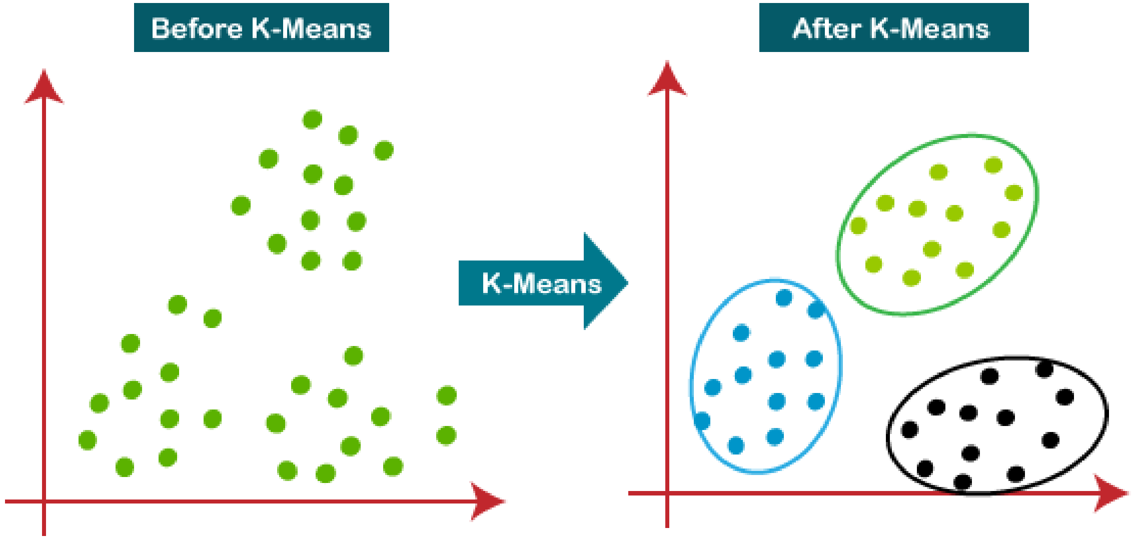


Country	India	Canada	Brazil	Australia
India	1	0	0	0
Canada	0	1	0	0
Brazil	0	0	1	0
Australia	0	0	0	1
India	1	0	0	0
Portugal	0	0	0	0

Normalização de dados

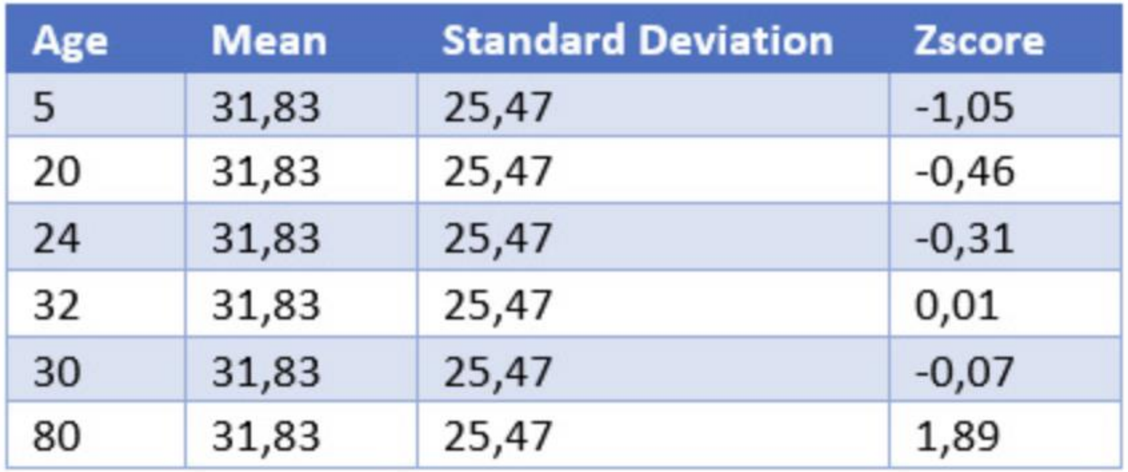
- Alterar a escala dos dados
  - Importante para algoritmos que são sensíveis à escala (ex: Knn, kmeans)


x, y and z scaled
1
1
y
z
X = 0.8 | y = 0.3 | z = 0.6
0
1
x
y and z scaled, x not scaled
1
1
y
z
X = 3 | y = 0.3 | z = 0.6
0
1
2
3
x



Standardization

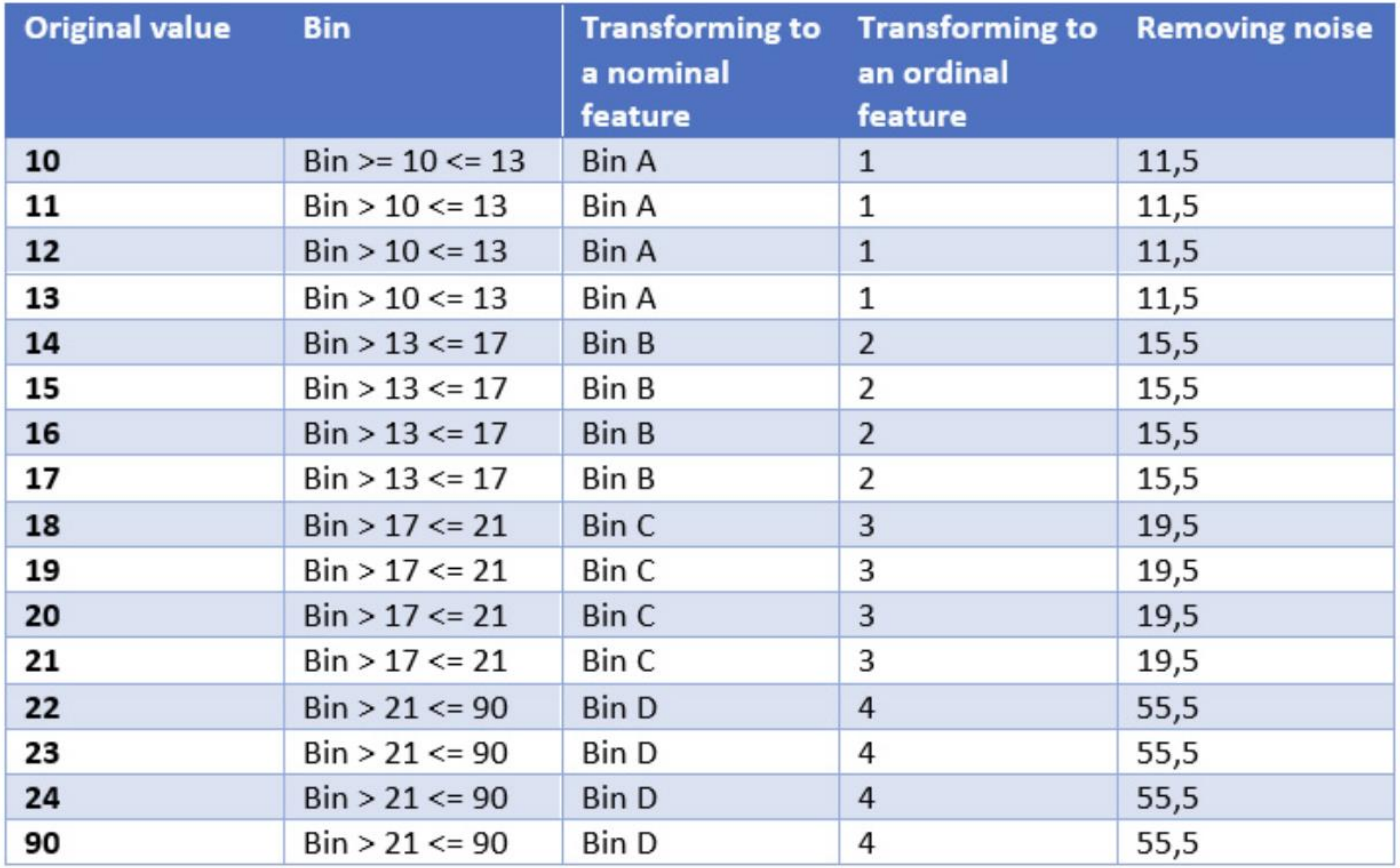
- Altera a distribuição dos dados, de maneira que a média sempre seja próxima de 0



Age	Mean	Standard Deviation	Zscore
5	31,83	25,47	-1,05
20	31,83	25,47	-0,46
24	31,83	25,47	-0,31
32	31,83	25,47	0,01
30	31,83	25,47	-0,07
80	31,83	25,47	1,89

$$X' = \frac{X - \mu}{\sigma}$$

Bining and discretization



Original value	Bin	Transforming to a nominal feature	Transforming to an ordinal feature	Removing noise
10	Bin >= 10 <= 13	Bin A	1	11,5
11	Bin > 10 <= 13	Bin A	1	11,5
12	Bin > 10 <= 13	Bin A	1	11,5
13	Bin > 10 <= 13	Bin A	1	11,5
14	Bin > 13 <= 17	Bin B	2	15,5
15	Bin > 13 <= 17	Bin B	2	15,5
16	Bin > 13 <= 17	Bin B	2	15,5
17	Bin > 13 <= 17	Bin B	2	15,5
18	Bin > 17 <= 21	Bin C	3	19,5
19	Bin > 17 <= 21	Bin C	3	19,5
20	Bin > 17 <= 21	Bin C	3	19,5
21	Bin > 17 <= 21	Bin C	3	19,5
22	Bin > 21 <= 90	Bin D	4	55,5
23	Bin > 21 <= 90	Bin D	4	55,5
24	Bin > 21 <= 90	Bin D	4	55,5
90	Bin > 21 <= 90	Bin D	4	55,5

Box Cox transformations

- Aplica funcoes matemáticas (power transformations) para normalizar os dados. Ex: Log, raiz quadrada, potencia, conjunto de funções.

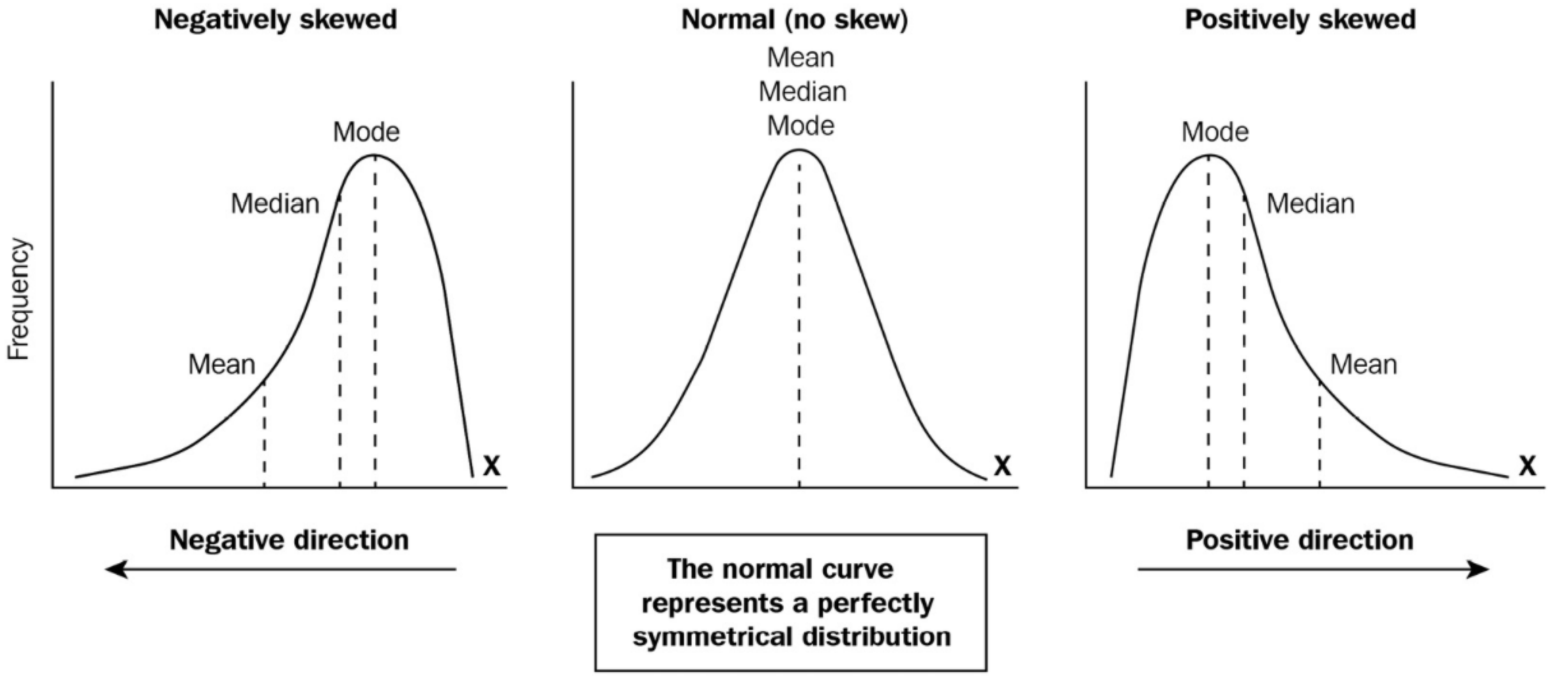


Figure 2.13 – Skewed distributions

Distribuicoes de dados

- Bernoulli: https://www.youtube.com/watch?v=j0HmgMNcOOQ
- Binomial:
- https://www.youtube.com/watch?v=O2WwMfC4Do4
- Poisson:
- https://www.youtube.com/watch?v=ZctEapUDGLk&t

Missing Values

- MCAR: Missing Completely at Random
- MNAR: Missing Not at Random
- listwise deletion: deletar linhas com valores faltantes ou colunas
- Imputation:
  - Média
  - Mediana
  - valor mais frequente
  - KNN
  - deep learning
  - Regressão
  - Multiple Imputation by Chained Equations (MICE)

Outliers

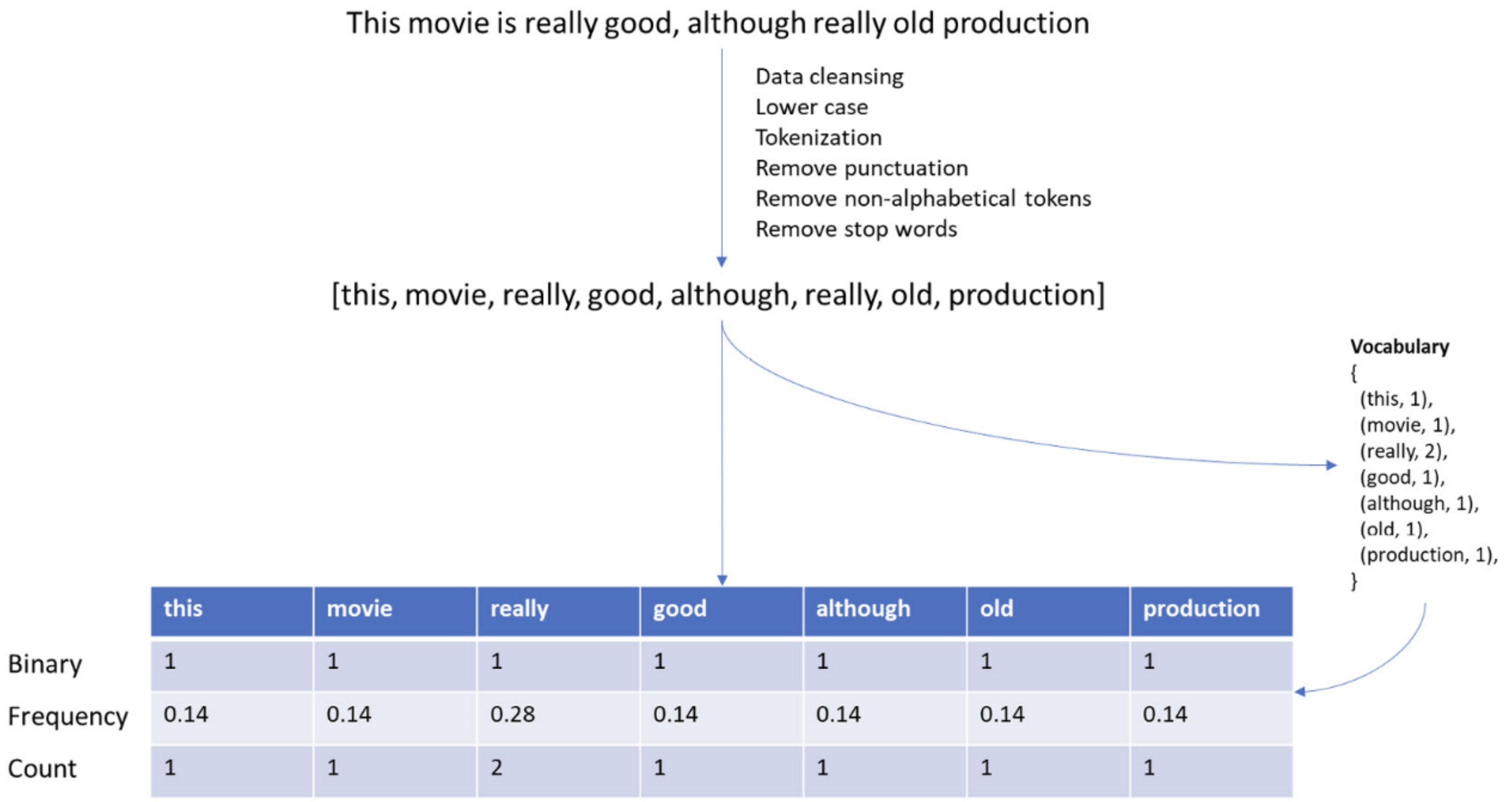
- AWS Cut Random forest
- Desvio padrão
- Pode ser interessante remover os valores, ou aplicar tratamentos parecidos com os de missing values

Bases Desbalanceadas

- Oversampling
- Udersampling
- SMOTE(KNN)
- Bootstraping (oversamping minor class)

Text trasnformations

- Bag of words (id para cada palavra)



This movie is really good, although really old production

Data cleansing
Lower case
Tokenization
Remove punctuation
Remove non-alphabetical tokens
Remove stop words

[this, movie, really, good, although, really, old, production]

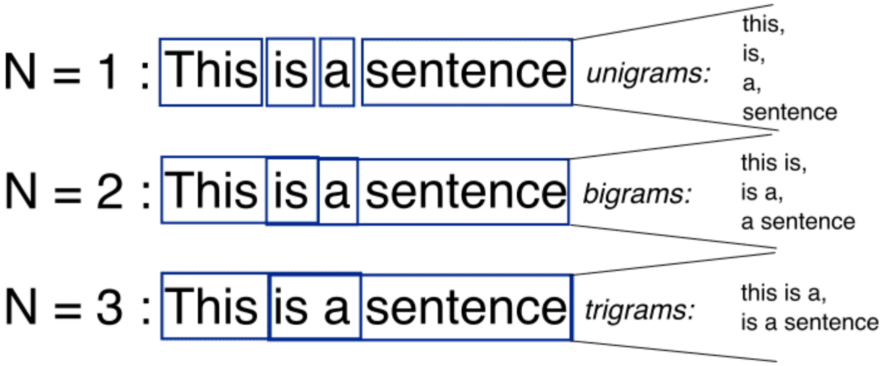
Vocabulary

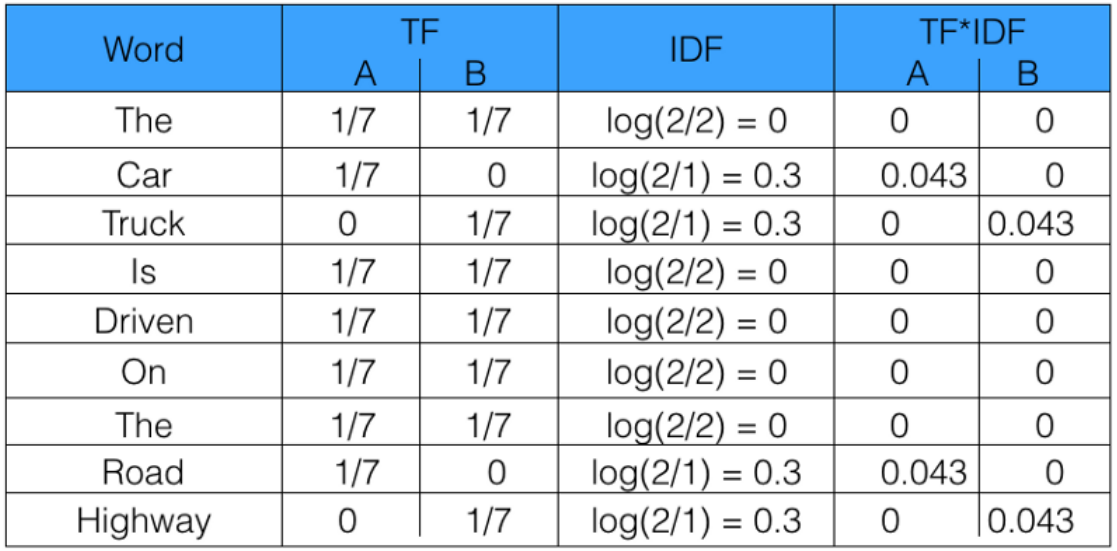
```
{
 (this, 1),
 (movie, 1),
 (really, 2),
 (good, 1),
 (although, 1),
 (old, 1),
 (production, 1),
}
```

	this	movie	really	good	although	old	production
Binary	1	1	1	1	1	1	1
Frequency	0.14	0.14	0.28	0.14	0.14	0.14	0.14
Count	1	1	2	1	1	1	1

Text trasnformations

- Bag of words (id para cada palavra)
- N gram
- TF-IDF
- Word embeddings





Word	TF		IDF	TF*IDF	
	A	B		A	B
The	1/7	1/7	log(2/2) = 0	0	0
Car	1/7	0	log(2/1) = 0.3	0.043	0
Truck	0	1/7	log(2/1) = 0.3	0	0.043
Is	1/7	1/7	log(2/2) = 0	0	0
Driven	1/7	1/7	log(2/2) = 0	0	0
On	1/7	1/7	log(2/2) = 0	0	0
The	1/7	1/7	log(2/2) = 0	0	0
Road	1/7	0	log(2/1) = 0.3	0.043	0
Highway	0	1/7	log(2/1) = 0.3	0	0.043

Question 1

- A company is building a fraud-detection model. Currently, the company does not have a sufficient amount of information because of a low number of fraud cases. Which method will improve the accuracy of the model?
  - A. Oversampling with bootstraping
  - B. Udersampling
  - C. Oversampling by SMOOT
  - D. Class weight adjustment

Question 1

- A company is building a fraud-detection model. Currently, the company does not have a sufficient amount of information because of a low number of fraud cases. Which method will improve the accuracy of the model?
  - A. Oversampling with bootstraping
  - B. Udersampling
  - C. Oversampling by SMOOT
  - D. Class weight adjustment

Question 2

- A machine learning (ML) specialist is optimizing a solution to define whether online payment transactions are fraudulent. The historical data of manually classified transactions includes the following data:
  - customer name (string)
  - customer type (integer)
  - transaction amount (float)
  - customer tenure (integer)
  - Transaction type (string) with values "normal" or "abnormal"

Which action should the ML specialist take to meet the requirements?

- A. Drop the customer name and enter the transaction type. Lauch the model training phase
- B. Drop the customer name and change the transaction type from string to numeric categories.
- C. Drop the transaction type and change the customer name categories
- D. Use the data without making changes to it and without any preprocessing tasks

Question 2

- A machine learning (ML) specialist is optimizing a solution to define whether online payment transactions are fraudulent. The historical data of manually classified transactions includes the following data:
  - customer name (string)
  - customer type (integer)
  - transaction amount (float)
  - customer tenure (integer)
  - Transaction type (string) with values "normal" or "abnormal"

Which action should the ML specialist take to meet the requirements?

- A. Drop the customer name and enter the transaction type. Lauch the model training phase
- B. Drop the customer name and change the transaction type from string to numeric categories.
- C. Drop the transaction type and change the customer name categories
- D. Use the data without making changes to it and without any preprocessing tasks

Question 3

- A Machine Learning Specialist is creating a new natural language processing application that processes a dataset comprised of 1 million sentences. The aim is to then run Word2Vec to generate embeddings of the sentences and enable different types of predictions.
  Here is an example from the dataset:
  "The quck BROWN FOX jumps over the lazy dog.`
  Which of the following are the operations the Specialist needs to perform to correctly sanitize and prepare the data in a repeatable manner? (Choose three.)

- A.Perform part-of-speech tagging and keep the action verb and the nouns only.
- B. Normalize all words by making the sentence lowercase.
- C. Remove stop words using an English stopword dictionary.
- D. Correct the typography on "quck" to "quick.€ג
- E. One-hot encode all words in the sentence.
- F. Tokenize the sentence into words.

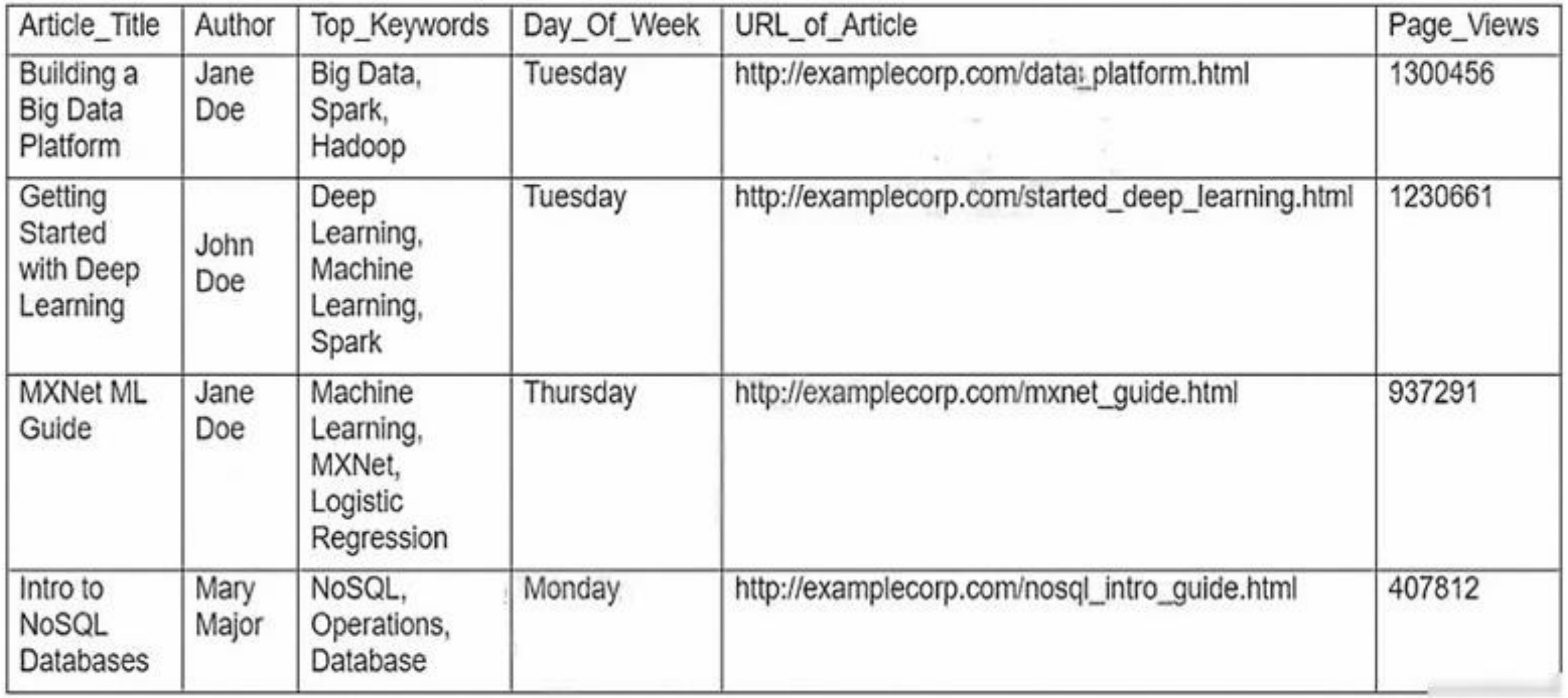
Question 3

- A Machine Learning Specialist is creating a new natural language processing application that processes a dataset comprised of 1 million sentences. The aim is to then run Word2Vec to generate embeddings of the sentences and enable different types of predictions.
  Here is an example from the dataset:
  "The quck BROWN FOX jumps over the lazy dog.`
  Which of the following are the operations the Specialist needs to perform to correctly sanitize and prepare the data in a repeatable manner? (Choose three.)

- A. Perform part-of-speech tagging and keep the action verb and the nouns only.
- B. Normalize all words by making the sentence lowercase.
- C. Remove stop words using an English stopword dictionary.
- D. Correct the typography on "quck" to "quick.€ג
- E. One-hot encode all words in the sentence.
- F. Tokenize the sentence into words.

Question 4

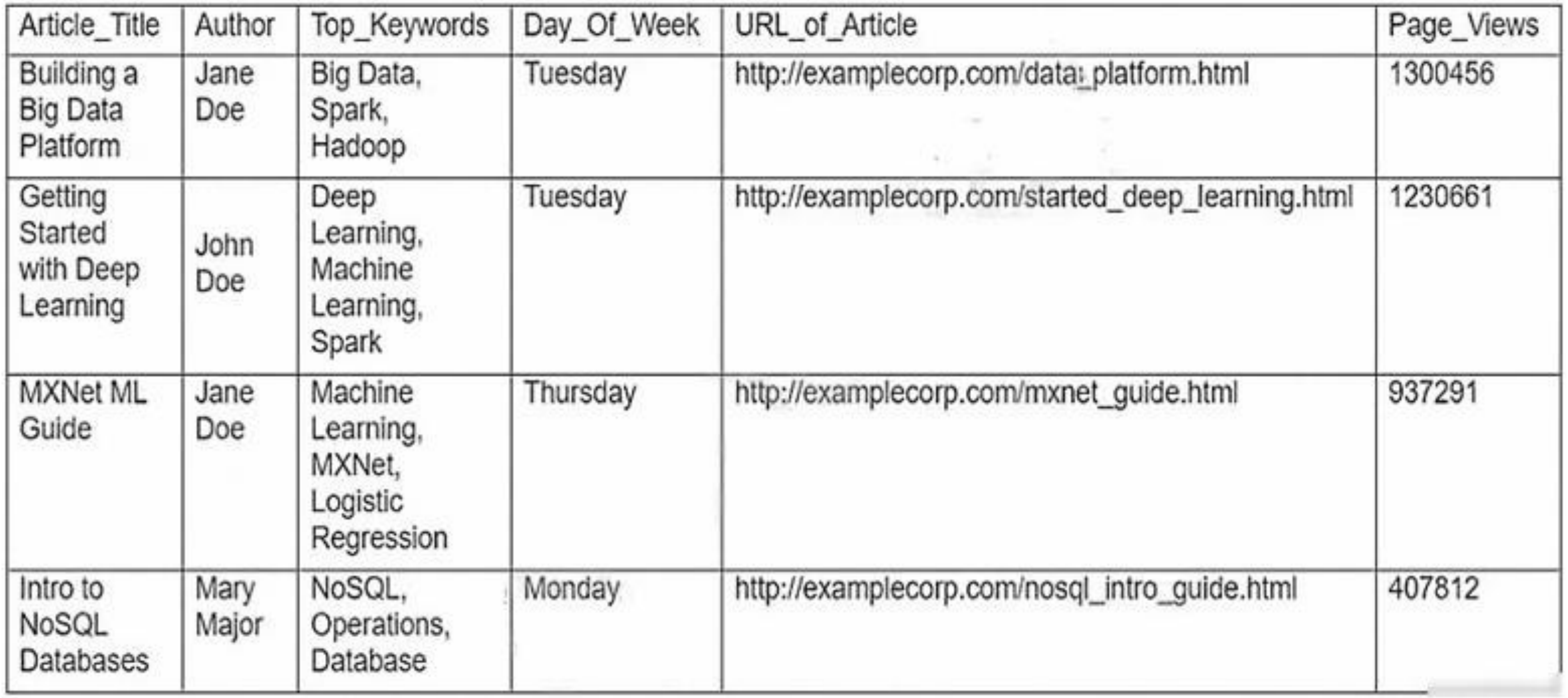
- Machine Learning Specialist is working with a media company to perform classification on popular articles from the company's website. The company is using random forests to classify how popular an article will be before it is published. A sample of the data being used is below. Given the dataset, the Specialist wants to convert the Day_Of_Week column to binary values.
  What technique should be used to convert this column to binary values?
  - A. Binarization
  - B. One-hot encoding
  - C. Tokenization
  - D. Normalization transformation



Article_Title	Author	Top_Keywords	Day_Of_Week	URL_of_Article	Page_Views
Building a Big Data Platform	Jane Doe	Big Data, Spark, Hadoop	Tuesday	http://examplecorp.com/data_platform.html	1300456
Getting Started with Deep Learning	John Doe	Deep Learning, Machine Learning, Spark	Tuesday	http://examplecorp.com/started_deep_learning.html	1230661
MXNet ML Guide	Jane Doe	Machine Learning, MXNet, Logistic Regression	Thursday	http://examplecorp.com/mxnet_guide.html	937291
Intro to NoSQL Databases	Mary Major	NoSQL, Operations, Database	Monday	http://examplecorp.com/nosql_intro_guide.html	407812

Question 4

- Machine Learning Specialist is working with a media company to perform classification on popular articles from the company's website. The company is using random forests to classify how popular an article will be before it is published. A sample of the data being used is below. Given the dataset, the Specialist wants to convert the Day_Of_Week column to binary values.
  What technique should be used to convert this column to binary values?
  - A. Binarization
  - B. One-hot encoding
  - C. Tokenization
  - D. Normalization transformation



Article_Title	Author	Top_Keywords	Day_Of_Week	URL_of_Article	Page_Views
Building a Big Data Platform	Jane Doe	Big Data, Spark, Hadoop	Tuesday	http://examplecorp.com/data_platform.html	1300456
Getting Started with Deep Learning	John Doe	Deep Learning, Machine Learning, Spark	Tuesday	http://examplecorp.com/started_deep_learning.html	1230661
MXNet ML Guide	Jane Doe	Machine Learning, MXNet, Logistic Regression	Thursday	http://examplecorp.com/mxnet_guide.html	937291
Intro to NoSQL Databases	Mary Major	NoSQL, Operations, Database	Monday	http://examplecorp.com/nosql_intro_guide.html	407812

Question 5

- A gaming company has launched an online game where people can start playing for free, but they need to pay if they choose to use certain features. The company needs to build an automated system to predict whether or not a new user will become a paid user within 1 year. The company has gathered a labeled dataset from 1 million users.
  The training dataset consists of 1,000 positive samples (from users who ended up paying within 1 year) and 999,000 negative samples (from users who did not use any paid features). Each data sample consists of 200 features including user age, device, location, and play patterns.
  Using this dataset for training, the Data Science team trained a random forest model that converged with over 99% accuracy on the training set.
  However, the prediction results on a test dataset were not satisfactory
  Which of the following approaches should the Data Science team take to mitigate this issue? (Choose two.)

- A. Add more deep trees to the random forest to enable the model to learn more features.
- B. Include a copy of the samples in the test dataset in the training dataset.
- C. Generate more positive samples by duplicating the positive samples and adding a small amount of noise to the duplicated data.
- D. Change the cost function so that false negatives have a higher impact on the cost value than false positives.
- E. Change the cost function so that false positives have a higher impact on the cost value than false negatives.

Question 5

- A gaming company has launched an online game where people can start playing for free, but they need to pay if they choose to use certain features. The company needs to build an automated system to predict whether or not a new user will become a paid user within 1 year. The company has gathered a labeled dataset from 1 million users.
  The training dataset consists of 1,000 positive samples (from users who ended up paying within 1 year) and 999,000 negative samples (from users who did not use any paid features). Each data sample consists of 200 features including user age, device, location, and play patterns.
  Using this dataset for training, the Data Science team trained a random forest model that converged with over 99% accuracy on the training set.
  However, the prediction results on a test dataset were not satisfactory
  Which of the following approaches should the Data Science team take to mitigate this issue? (Choose two.)

- A. Add more deep trees to the random forest to enable the model to learn more features.
- B. Include a copy of the samples in the test dataset in the training dataset.
- C. Generate more positive samples by duplicating the positive samples and adding a small amount of noise to the duplicated data.
- D. Change the cost function so that false negatives have a higher impact on the cost value than false positives.
- E. Change the cost function so that false positives have a higher impact on the cost value than false negatives.

Question 6

- A Data Scientist is developing a machine learning model to predict future patient outcomes based on information collected about each patient and their treatment plans. The model should output a continuous value as its prediction. The data available includes labeled outcomes for a set of 4,000 patients. The study was conducted on a group of individuals over the age of 65 who have a particular disease that is known to worsen with age.
  Initial models have performed poorly. While reviewing the underlying data, the Data Scientist notices that, out of 4,000 patient observations, there are 450 where the patient age has been input as 0. The other features for these observations appear normal compared to the rest of the sample population
  How should the Data Scientist correct this issue?

- A. Drop all records from the dataset where age has been set to 0.
- B. Replace the age field value for records with a value of 0 with the mean or median value from the dataset
- C. Drop the age feature from the dataset and train the model using the rest of the features.
- D. Use k-means clustering to handle missing features

Question 6

- A Data Scientist is developing a machine learning model to predict future patient outcomes based on information collected about each patient and their treatment plans. The model should output a continuous value as its prediction. The data available includes labeled outcomes for a set of 4,000 patients. The study was conducted on a group of individuals over the age of 65 who have a particular disease that is known to worsen with age.

  Initial models have performed poorly. While reviewing the underlying data, the Data Scientist notices that, out of 4,000 patient observations, there are 450 where the patient age has been input as 0. The other features for these observations appear normal compared to the rest of the sample population

  How should the Data Scientist correct this issue?

- A. Drop all records from the dataset where age has been set to 0.
- B. Replace the age field value for records with a value of 0 with the mean or median value from the dataset
- C. Drop the age feature from the dataset and train the model using the rest of the features.
- D. Use k-means clustering to handle missing features

Question 7

- An online reseller has a large, multi-column dataset with one column missing 30% of its data. A Machine Learning Specialist believes that certain columns in the dataset could be used to reconstruct the missing data.
  Which reconstruction approach should the Specialist use to preserve the integrity of the dataset?

- A. Listwise deletion
- B. Last observation carried forward
- C. Multiple imputation
- D. Mean substitution

Question 7

- An online reseller has a large, multi-column dataset with one column missing 30% of its data. A Machine Learning Specialist believes that certain columns in the dataset could be used to reconstruct the missing data.
  Which reconstruction approach should the Specialist use to preserve the integrity of the dataset?

- A. Listwise deletion
- B. Last observation carried forward
- C. Multiple imputation
- D. Mean substitution

Question 8

- A data scientist has explored and sanitized a dataset in preparation for the modeling phase of a supervised learning task. The statistical dispersion can vary widely between features, sometimes by several orders of magnitude. Before moving on to the modeling phase, the data scientist wants to ensure that the prediction performance on the production data is as accurate as possible.
  Which sequence of steps should the data scientist take to meet these requirements?

A. Apply random sampling to the dataset. Then split the dataset into training, validation, and test sets.

B. Split the dataset into training, validation, and test sets. Then rescale the training set and apply the same scaling to the validation and test sets.

C. Rescale the dataset. Then split the dataset into training, validation, and test sets.

D. Split the dataset into training, validation, and test sets. Then rescale the training set, the validation set, and the test set independently.

Question 8

- A data scientist has explored and sanitized a dataset in preparation for the modeling phase of a supervised learning task. The statistical dispersion can vary widely between features, sometimes by several orders of magnitude. Before moving on to the modeling phase, the data scientist wants to ensure that the prediction performance on the production data is as accurate as possible.
  Which sequence of steps should the data scientist take to meet these requirements?

A. Apply random sampling to the dataset. Then split the dataset into training, validation, and test sets.

B. Split the dataset into training, validation, and test sets. Then rescale the training set and apply the same scaling to the validation and test sets.

C. Rescale the dataset. Then split the dataset into training, validation, and test sets.

D. Split the dataset into training, validation, and test sets. Then rescale the training set, the validation set, and the test set independently.