AWS Data Engineering

# Sage maker

# Used for creating jupyter notebook in AWS

# Data Gathering

* Public Repositories
* Scrapping the webpages

# Data Cleaning - Handling missing values

* Do nothing – (ignore observations )
* Remove the entire record
* Mode – (For Categorical Data)
* Median /average value replacement – ( For Numerical Data)
* Most frequent value
* Model – based imputation – ( K-Nearest Neighbors (feature similarity)  
  / Regression (using correlation matrix) / Deep Learning )
* Interpolation / Extrapolation – (estimate values from other observations within the range of a discreate set of known data points)
* Forward filling / Backward filling – ( filling the missing value from the preceding or succeeding values)
* Hot deck imputation – (randomly choosing the missing value from the set of related and similar variable)

# Feature Extraction

* **Dimensionality** – refers to the number of features in your dataset
* **Reducing** **dimensionality** – refers to reducing the number of features keeping only required ones
* **Feature selection**
* **Feature** **Extraction** –PCA- take features and give something less (reducing the number of features while still retaining as much information as possible
* **PCA**: **principle** **component** **analysis** : unsupervised algorithm that creates new features by linearly combining original features

# Encoding categorical data

Binarizer Encoding : for features of a binary nature -- true/false , on/off, male/female etc.

Label Encoding (non binary) : may imply ordinality, can use ordinal encoder.

One – hot -encoding: change nominal categorical values into numerical values

|  |  |  |
| --- | --- | --- |
|  | Is\_sunny | Is\_cloudy |
| 0 | 1 | 0 |
| 1 | 0 | 1 |
| 2 | 0 | 1 |
| 3 | 1 | 0 |

Example

|  |  |
| --- | --- |
|  | Weather |
| 0 | Sunny |
| 1 | Cloudy |
| 2 | Cloudy |
| 3 | sunny |

# Numeric feature engineering

Transforming numeric values so machine learning algorithms can better analyze them.

Change numeric values so all values are on the same scale.

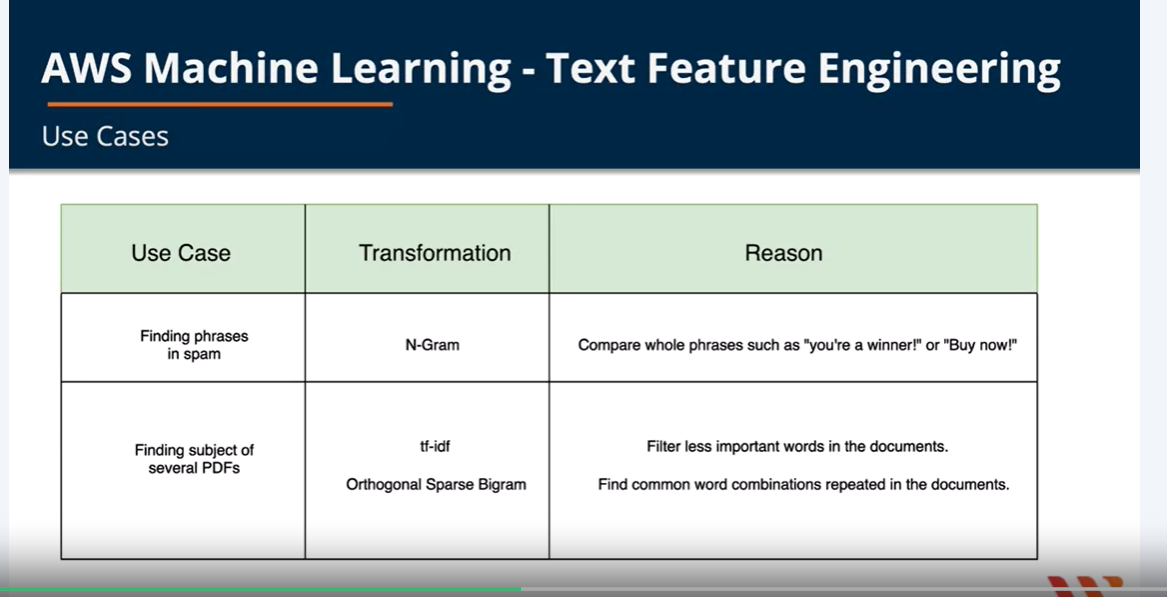
* **Normalization**: rescale the values into a range [0,1] .
* **Standardization**: Rescale data to have a mean of 0 & standard deviation of 1.
* **Binning**: re-organizing continuous data into bins – use qcut in python for it

# Text Feature Engineering

Text Feature Editing : transform text within data so machine learning algorithms can better analyze it.

Splitting text into smaller pieces

* Bag of words
  + tokenizes raw text and creates a statistical representation of the text.
  + Breaks up text by white spaces into single words
* N-Gram
  + Extension of bag of words which produces groups of words of size
  + Breaks up text by whitespace into groups pf words
* Orthogonal Sparse Bigram
  + Creates groups pf words of size n, returns every pair of words that includes the first word
  + Creates groups of words that always include the first word
* Term Frequency-Inverse Document Frequency (tf-idf).
  + Shows how important a word or words are o a given set of text by providing appropriate weights to terms that are common and less common.



# Amazon Mechanical Turk

Amazon Mechanical Turk service

is one of the on demand and

scalable human workforce service for completing the task

where the human capabilities

are generally outperforming the computers.

# AWS Migration Services

* Migrate data from source to machine learning repository
* Several aws services to help move data
  + **Amazon data pipeline :**
    - copy data using pipeline activities
    - schedule regular data movement
  + **Aws database migration services (DMS):**
    - Move data between databases
    - MySQL to MySQL
    - Aurora to DynamoDB
  + **AWS GLUE**
    - ETL SERVICE
  + **Amazon sage maker**
    - Use jupyter notebook
  + **Amazon Athena**
    - Run sql queries on S3 data

Exploratory data analysis in AWS

# Kinesis Data streams

* Gets data from data producers such as IoT, Social media.
* Uses shards to stream data to consumers such as EC2, lambda, Kinesis Data analytics.
* Consumers then send data to a data repository such as S3, DynamoDB, Redshift, or Business Intelligence Tools.

# Kinesis Data Firehose

* Receives data from producers such as IoT, social media
* Uses lambda functioning instead of shards to transmit producer data
* Lambda function puts data to data stores such as S3, Redshift, ElasticSearch, or splunk
* Can transmit directly from producers through Firehose to the data store (don’t have to use lambda intermediary)
* S3 events to store to DynamoDB

# Kinesis Video Streams

* Build video processing applications such as machine learning models
* Producers such as web cams, security cameras, audio feeds, images
* Data consumers – Kinesis video stream applications
* Stores to S3

# Kinesis Data Analytics

* Use SQL to process streaming data
* Sources: Kinesis Data Streams and Kinesis Data Firehose
* SQL queries put to S3, Redshift, or Visualization and Business Intelligence tools

# EMR – Amazon Elastic MapReduce

* Managed platform designed for cluster-centric workloads
* Amazon EMR is widely used for tasks like log analysis, data transformations, and large-scale machine learning.
* It allows the execution of prominent big data frameworks such as Apache Spark or Apache Hadoop on the AWS cloud, facilitating the processing and analysis of extensive data sets

Layers of Amazon EMR

* Storage -> (HDFS, EMRFS, Local file system)
* Cluster Resource Management -> (Responsible for Cluster Resource Management and scheduling tasks)
* Data Processing Frameworks -> (consists of Hadoop map reduce , Apache Spark)
* Applications and programs -> (helps processing and management of big data sets like HIVE, PGI, Streaming Libraries and ML Algorithms)