Day 2

12 May 2024 18:52

Sample dataset (loan applications)

Loan_ID	Gender	Married	Number of Dependents	Education	Self Employed ?	Applicant Income	Co-applicant Income	Loan Amount	Loan Amount Term	Credit History	Property Area
LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban
LP001824	Male	Yes	1	Graduate	No	2882	1843	123	480	1	Semiurban
LP002928	Male	Yes	0	Graduate	No	3000	3416	56	180	1	Semiurban
LP001814	Male	Yes	2	Graduate	No	9703	0	112	360	1	Urban
LP002244	Male	Yes	0	Graduate	No	2333	2417	136	360	1	Urban
LP001854	Male	Yes	3+	Graduate	No	5250	0	94	360	1	Urban
...	Male	Yes	0	Graduate	No	3500	1667	114	360	1	Semiurban
LP001647	Male	Yes	0	Graduate	No	9328	0	188	180	1	Rural
LP001871	Female	No	0	Graduate	No	7200	0	120	360	1	Rural
LP001379	Male	Yes	2	Graduate	No	3800	3600	216	360	0	Urban
LP002789	Male	Yes	0	Graduate	No	3593	4266	132	180	0	Rural
LP001578	Male	Yes	0	Graduate	No	2439	3333	129	360	1	Rural
LP001318	Male	Yes	2	Graduate	No	6250	5654	188	180	1	Semiurban
LP001259	Male	Yes	1	Graduate	Yes	1000	3022	110	360	1	Urban
LP002804	Female	Yes	0	Graduate	No	4180	2306	182	360	1	Semiurban

Independent variables

Dependent variable ?

SAMPLE DATASET (automobiles)

KMs per liter	cylinders	displacement	horsepower	weight	acceleration	year	origin	Model name
18	8	307	130	3504	12	70	1	chevrolet chevelle malibu
15	8	350	165	3693	11.5	70	1	buick skylark 320
18	8	318	150	3436	11	70	1	plymouth satellite
16	8	304	150	3433	12	70	1	amc rebel sst
17	8	302	140	3449	10.5	70	1	ford torino
15	8	429	198	4341	10	70	1	ford galaxie 500
...	***	...	...	...	...	...	...	...
...	***	...	...	...	...	...	...	...
24	4	113	95	2372	15	70	3	toyota corona mark ii
22	6	198	95	2833	15.5	70	1	plymouth duster
18	6	199	97	2774	15.5	70	1	amc hornet
21	6	200	85	2587	16	70	1	ford maverick
27	4	97	88	2130	14.5	70	3	datsun pl510

Dependent variable ?

Sample data - jobs

Independent variables ?

Dependent variable ?

Month	Total Filled Jobs
2004M07	1795610
2004M08	1792770
2004M09	1809590
2004M10	1815580
2004M11	1856360
2005M04	1871630
2005M05	1867870
2005M06	1857260
2005M07	1858360
2005M08	1856320
2005M09	1876270
2005M10	1866920
2011M10	1903630
2011M11	1940200
2011M12	1983070
2012M01	1865540
2012M02	1932380

Stock prices

Date/time	$$	
8/7/2024 10am	10	
	11	
	12	
	10	

Stock price -> both dependent and independent!

Price at a moment is dependent on few other prices in the series (sequence)

11am price -> {9am, 10am, 10:30am}

TIME SERIES data

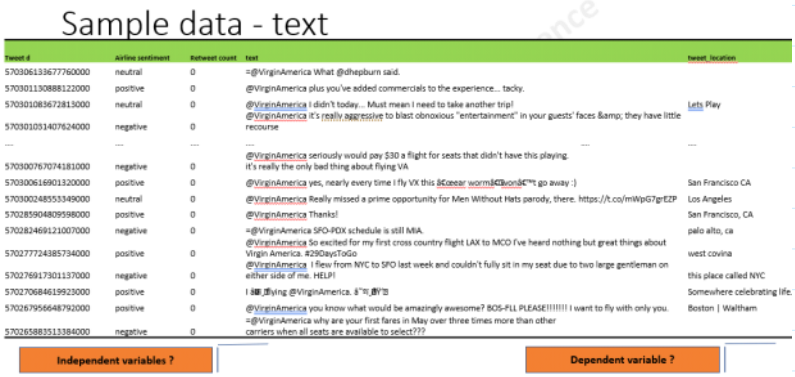
Examples : weather, all economic data, sensors, IOT, medical devices

Handling -->

- Traditional statistical methods (ARIMA, AR, MA, ARMA …. Variations of ARIMA)
- ML methods
    - Classical methods
        - Sequence models

- Advanced
  - Transformer

- ML/DL methods are better than classical statistical method (ARIMA)



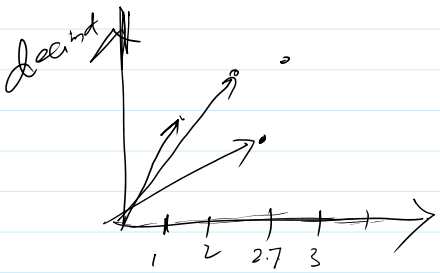
Processing text data (challenges)
- Unstructured
- Mixed cases
- Volume
- Language
- Typos, errors, grammatical errors
- Context of the text --> Semantics

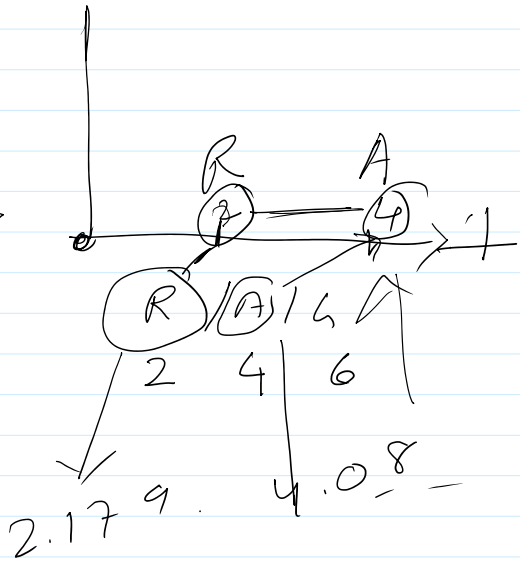
"My flight is delayed. Brilliant!' -> Happy/Sad

Feature types

1. Numeric
   a. Plain numeric
      i. Float
         1) Ratio (we will encounter 95% of the time)
         2) Interval
            a) Lacks a true zero
               i) Temp in C
               One. 50 C and 55C
   b. Discrete
      i. Count
         1) Number of children
         2) Number of cars
         3) Number of claims
         4) Score?
         5) No. of customers
2. Categorical (Qualitative variables)
   a. Nominal
      i. No order
      ii. Can we apply any math op?
         1) No
   b. Ordinal
      i. Which has order
         1) SML
         2) Rating scale
         3) Score
      ii. Can we apply any math op?
         1) Limited (sorting)
   c. Binary
      i. Two categories

   d. Cat data are BIG problem in AI
      i. ML/DL - are math wise - work well with decimals or cont data
         1) Assumption is that there is origin within the data (0,0)
3. Text
4. Date & time





5. Boolean
6. Spatial features (geo , locations)
7. Image/ videos
8. Audio

10. Derived form of features

Business scenario (from slide # 5) Lesson 3

Healthcare org

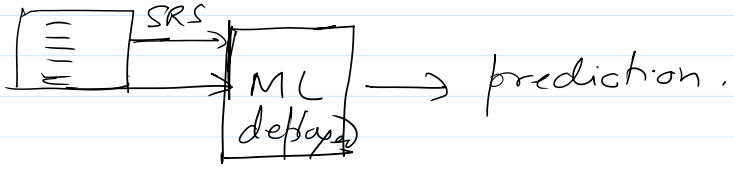
- Manage
  - Hospitals
  - Colleges
  - Clinics
  - Labs
  - Pharmacy
  - Medicines
  - staff
- Vast amount of data and huge number of data types
  - Volume
  - Variety
  - Format
  - dynamic
- 3 buckets for the data
  - Purpose
  - Relations
  - Storage
- Dat sources
  - Patient records
  - Clinical systems
  - Financial mgmt
  - Operational/administrative
  - Billing
  - Compliance system

- **Bucket 1 (Patient data)**
  - **Purpose**
    - Demographics (age …
    - History of illness
    - Insurance info
    - Diagnosis
    - Lab results
    - Treatment/ regimen
    - Prescriptions
    - …
    - …
    - …
    - EHRs , FHIR
      - Sensitivity
      - PHI
      - PII
      - Data secured
        - Access controls
        - Encryption
        - Backup and rollback …(failsafe)
  - Relations
    - Clinical info
    - Lab
    - Medication
    - Orders
    - Billing
    - ,,,
    - ,,,
  - Storage methods
    - Centralized location

- Bucket 2 (clinical records)
  - Decision making
  - Accurate diagnosis
  - Treatment plans
  - Monitoring the patients
  - Legal
  - Fin
  - Billing
  - Insu

Population and samples

- Sampling techniques [...,..,..]
  - Simple random sampling (SRS)
    - Every element has equal chance of being picked
    - [10, 13, 15, 16, 18, 19.... 1000 numbers]
    - Implement
      - Length of array = len(arr) = 1000
      - **Rand**(len=1000) - rand - function in Python/R
        - Return any number from 1 ... 1000
          - 77
      - Arr[77]
    - Use cases
      - Large dataset
      - Prelim analysis or understanding
      - selecting a list of transactions to review for financial audit



      - Ensure FAIRNESS of the evaluation
        - SRS
      - If the dataset is HOMOGENOUS
        - No groups in the data (not heterogenous)
          - SRS
    - Flip side
      - Outliers
      - Chance based -> selections may not fully representative
      - Samples being small

  - Stratified random sampling
    - Subgroups in the data
      - Randomly pick samples from each of the strata (subgroups)
    - **Variability** = subgroups (ensured)

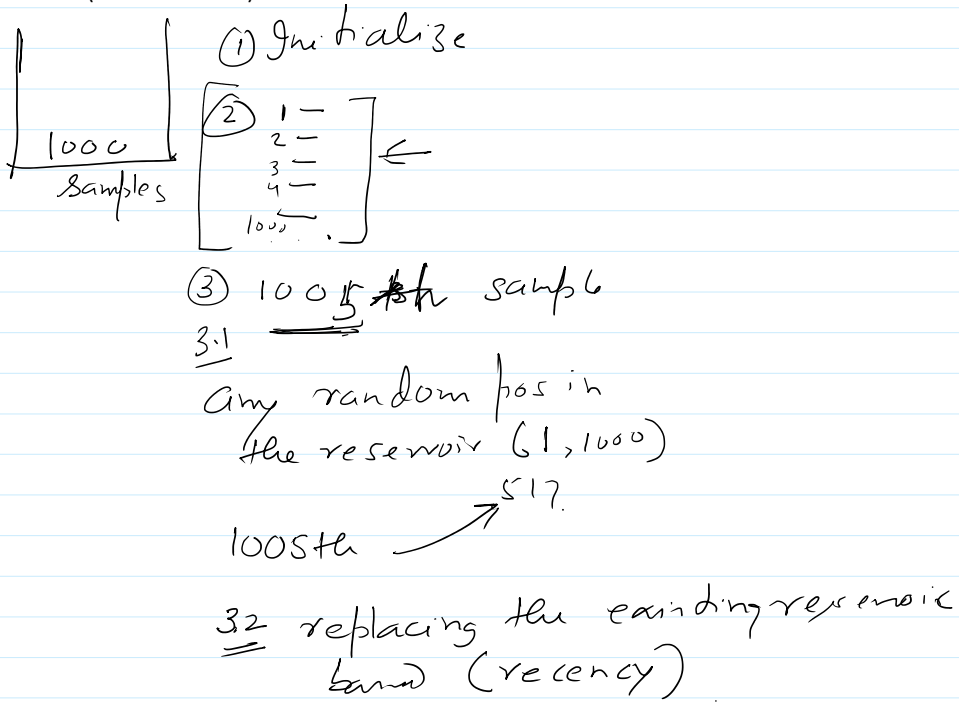
    - Use case
      - Clinical trial (new drug)
        - Effect on age groups, genders, cultures...
      - AI - all ML and DL model (evaluation)

  - Systematic sampling
    - Select every **nth** sample
    - Sorted, starting point
    - Maintain randomness in picking samples ???
    - May introduce some bias??? (periodicity)
      - Careful starting point , the interval (n) determination

    - Use case
      - Time series data

- Cluster sampling
  - Dividing the population into clusters or groups
    - Understanding cluster - Easy or difficult
      - Defining cluster
        - Using 1 col or multiple columns
    - 15 clusters
  - Subset of these 15 clusters
    - 5 of them
  - Use cases
    - Healthcare
      - Hospital patient records
      - Group of hospitals by location
      - Randomly select a subset of hospitals
        - Collect all patients from these subsets
    - Efficient
      - Large datasets
      - Spatial characteristics
      - CLUSTERS SHOULD BE NATURAL GROUPINGS
  - Flip side
    - Biased estimates
      - Some group
        - More variability than others
    - Miss out on other groups (completely)
    - Cluster size sufficient

- **Reservoir sampling**
  - Randomized algo
  - Pick FIXED number of samples --> model/analysis
  - Useful -
    - Large datasets
    - Perform random sampling
    - Data too large to fit into memory
    - When we deal with streaming data (continuous data)
  - Reservoir = bucket = fixed number of samples
  - Process the incoming stream
  - Random replacement (in the reservoir)



(1) Initialize

(2)

(3) 100[5] ~~5th~~ sample

3.1

any random pos in the reservoir (1, 1000)

517

100[5]th

3.2 replacing the existing reservoir based (recency)

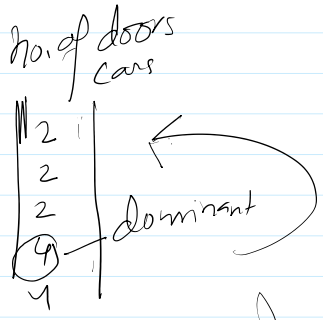
-
  - **Flip side**
    - Representativeness
      - No

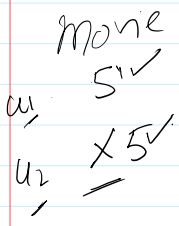
Measures of central tendencies

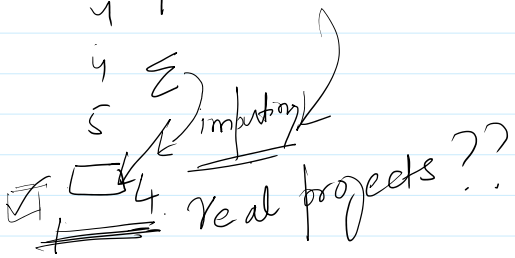
- Why we need understanding of CT
  - Know the typical range of values for features

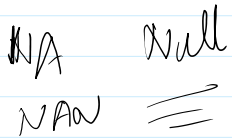
- Mean
  - Avg
  - Sum/n
  - Decimal (discrete)
  - Measure of the central value
  - Affected by outliers

  - Use cases
    - Aggregating
    - Imputing
      - Fill in the missing spots

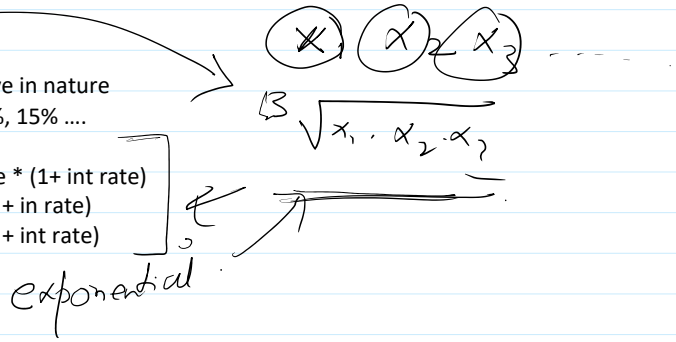








Geometric mean
- Column
  - Values are multiplicative in nature
    - Return rate 5%, 10%, 15% ....
    - Growth rate
    - Year 1 = initial value * (1+ int rate)
    - Year 2 = year 1 * (1 + in rate)
    - Year 3 = year 2 * (1 + int rate)



How do we find out if a feature exhibits exp nature?
- Plot it !!

Harmonic mean
- Features
  - Rates, ratios, speed

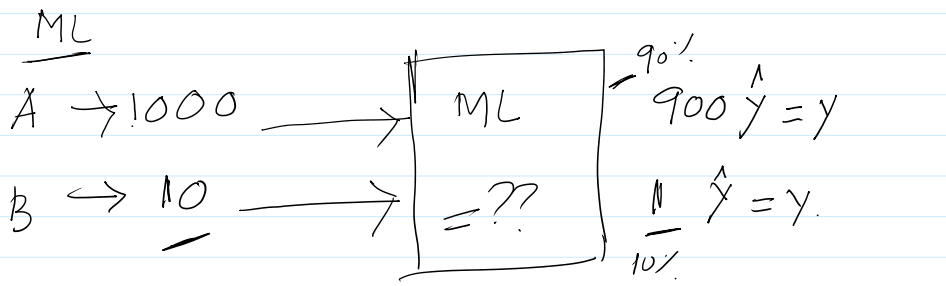
Harmonic Mean = n / (1/x1 + 1/x2 + 1/x3 + ... + 1/xn)

- Where:
  - x1, x2, x3, ..., xn are the individual values in the dataset.
  - n is the total number of values in the dataset.

The harmonic mean places more weight on smaller values in the dataset.

Math intuition behind this formula
- Makes the small values more weighty

Use case
-



.9
.1 2

$\frac{1}{.9} + \frac{1}{.1}$

$\frac{900+1}{1000+10} = \frac{901}{1010} \neq 90\%$