PRJ#01	Course Expiration Date: May 14, 2022 365 total days. 168 days remaining as on 27·11·2021
26·11·21 Friday (−169)	CS−1 started Topic: Healthcare Provider Fraud Detection Analysis
Done	Overview
ND	Research Papers ← Currently Working
ND	First Cut Approach
27·11·21 Saturday (−168)	Research Papers to be identified as of today. Total hours planned − 2(1/2 + 1/2 + 1/2 + 1/2) 8−9 am 9:30−10:30 am

Pg 25 imp.

① Random Oversampling → Non-Synthetic

② SMOTE → Synthetic
↳ KNN imblearn

Work Notes:

1. Oversampling and Outlier Detection techniques

→ ADASYN (Adaptive Synthetic Sampling Technique)
* Assign weights to minority class

→ Robust Rose (Also sensitive to outliers)
* Based on Kernel Density estimate around existing real minority cases.

Revision

M3 — 21.17 → Curse of dimensionality.
(Ch 4)

(Ch 9) → 2.6.1 → Geometric Intuition of Linear Regression.

(Ch 4) → 21.2 Imbalanced vs balanced dataset.

? 2. Can propensity score matching be anyhow used in binary classification for understanding the behaviour on of our ML model, or is it used in case of treatment of group of members vs. those of control in an our experiment?

~~I'm b~~

? 3. Doubts : Weighted Least Square.

* Confidence Interval.

* Oversampling w/o replacement and with replacement.

* Homogeneity

M2 : ch13. V24 → Confidence Interval Introduction

Q How can we use Confidence Interval in Analysis of Binary classification problem?

* Generative Adversarial Oversampling ~~method~~ Network oversampling algorithms.

In output layer of GANs :-

Generator
- Categorical → Softmax Activation
- Continuous → Sigmoid Activation

~~Defender~~ Discriminator
- Adam optimization algorithm with LR = 0.0002
- Beta 1 momentum = 0.5

Pending

→ New Techniques to be learnt:
GAN (from course)

References (weblinks) for GANs put across Notes (Apple) on CS-1 study references.

Important Words discussed in video:

→ Statistical Mechanics (Physics)

28.11.21 (Sunday) (-167) Holiday Break

29.11.21 (Monday) (-166) Reading GAN paper by Ian Goodf-ellow and Yoshua Bengio, et.al.

✓ Abstract
Question has been asked from the team.

- Introduction
Was ~~tasty~~ fairly simple to understand but still requires a quick reading.

- Related Work
Was quite complex to understand

- Adversarial Nets
  * Algorithm 1
  * Global Optimality of $P_g = P_{data}$
  * Theorem 1

* Convergence of Algorithm!
* Observation: A little difficult to understand.

~~* Conver~~

Mem. Notes: Remedy for Understanding Mathematics for ML & DL:

Mathematics for Machine Learning (MML)
– by Diesenroth, Faisal & Ong
Bookmarked for reading.

Book names for ML & DL Concepts reference present on Analytics Vidhya Blog.

– Experiments
– Advantages and Disadvantages
– Conclusions and Future Work

→ Imp References felt [To look at them]

[2], [10], [11], [25], [28]

Reading Generative Adversarial Minority Oversampling
– by Mullick, Datta, Das

Mem. Notes: Mail was sent to team for understanding about Generative Learning Algos.

Andrew Ng's Lecture Notes is important and is received as a reply.

Reading Andrew Ng's Lecture Notes from course CS229 Part IV

Generative Linear Algorithms

1. Gaussian Discriminant Analysis
   1.1 The Multivariate Normal Distribution.
   1.2 The Gaussian Discriminant Analysis
   1.3 Discussion: GDA and Logistic Regression
2. Naive Bayes
   2.1 Laplace Smoothing
   2.2 Event models for text classification

Glossary:

pg 1

$P(y|x;\theta)$; conditional distribution; logistic regression model $P(y|x;\theta)$ as $h_\theta(x) = g(\theta^T x)$; perceptron algorithm; decision boundary; inputs X; generative learning algorithms; class priors; Bayes rule; posterior distribution

pg 2

multivariate normal distribution; parameterized by mean vector;

covariance matrix; symmetric positive semi-definite; covariance

pg 3; pg 4

—NA—

pg 5

continuous-valued random variables; Bernoulli; log-likelihood

pg 6

contours

pg 7

discriminative algorithm; converse; asymptotically efficient; weaker assumptions; robust and less sensitive to incorrect modelling assumptions; Poisson

pg 8

deviations from modelling assumptions; discrete valued; unsolicited

pg 9

multinomial distribution; conditionally

independent

pg 10

joint liklihood; posterior probability; discretize

pg 11

-NA-

pg 12

parameterize our multinomial

pg 13

Bernoulli Event Model; Multi-variate Bernoulli event model; Multinomial event model; identity

Important Terms through CS229 video

To watch CS229 Summer 2019 Lecture on Youtube

Time Table for Tomorrow:

1. Explore Researches:
2. Revision of Deep Learning: [Applied AI Course]
3. Exploratory Data Analysis:

Q Suppose you are done with balancing out the 3 datasets, What is the next step that you will follow?

Exploratory Data Analysis to understand the data more in-depth.

Mistake #1: EDA is done as the FIRST step. It ~~is~~ will be then followed by Oversampling.

30.11.21 (Tuesday) (-165)

To start with Exploratory Data Analysis.

Q I'm hesitating to start with EDA task. Why?
Because it is a too much thought oriented task. And to do it with full concentration, I need complete focus on that task only.

This time I have decided to work on 3 task. So, I've decided to first finish the DL course video. It will take 2 hours.

Could not do due to sleep problem.

1.12.2021 (Wed) (-164)

Encoder - Decoder Models DL video for 2 hrs.

3 References weblinks to be read present in Apple Notes.

Applications also discussed along with References.

Code Example from Keras also discussed

Exploratory Data Analysis for Case Study 1

Haberman Dataset Visualization Assignment was checked on my github page.

1. Identifying datapoints, features and class
   - Conclusion
2. Univariate Analysis
   2.1 Probability Density function
   - Conclusion
   2.2 Cumulative Distribution Function
   - Conclusion
   2.3 Boxplot and whiskers
   - Conclusion
   2.4 Violin Plots
   - Conclusion

Pending (4)

3. Bivariate Analysis
   3.1 2-D Scatter Plot
      - Conclusion
   3.2 Pair Plot
      - Conclusion
Summary about the Entire Dataset

Apply tSNE on Donors Choose Dataset

Pie | Donut Chart
- Conclusion
Univariate Analysis on each feature
* Barplots
   - Conclusion
* kdeplot
   - conclusion
* Boxplot
   - Conclusion
Summary

Create Contents
Explain about the Case Study
Explain about each feature

P.T.O

Pending

Exploratory DATA A.

Steps to follow Initially to work with Google Colab.

1. Place the dataset in google drive which is downloaded from Kaggle.
2. Create a new colab sheet and start creating H1 headings and conclusion headings for each feature.
3. Also, start creating Contents Section to start referring to each heading. It should point to that heading when clicked on.

Imp

4. Each file of train and test resp. were placed in different train and test folders to avoid ambiguity. (1 train file got hidden on Mac) It got unhidden after uploading on gDrive.

5. In ipynb following was done
   - Intro
   - Dataset Description
   - Problem Statement

6. Facing problem in finding the descri-ption of each file and its features.
- Imports - Reading the csv
- Dataset columns
* Train Beneficiery/Inpatent/Opat
  → Columns
  → head
  → Shape
  → dtypes
- Graphs to understand the values in each feature.
  check using hue.
  * Univariate Analysis
    (1) Gender
      % & piechart
      *conclusion
    (2) State
      ~~Barplot with seaborn~~ hue.

~~Note~~ →

2-12-2021 Thurs (-163)

Attention Models in Deep Learning

DL → 2 hr video
* Notes in DL part 2.

Research paper for Neural Machine Translation by Jointly Learning to align and translate.
One-line implementation.
References | Shared.

6. EDA contd.

(2) State

~~Barplot with seaborn(hue)~~.

* Crosstab with pandas
* Crosstab bar plot

(Left Conclusion) Train / Test Beneficiery columns:-

Obj 1. BeneID ✓ → PK (UNQ)

Obj 2. DOB (Age can be calculated) — Age as of date of discharge

Obj ✓ 3. DODs (Date of Discharge) ✓

int ✓ 4. Gender ✓ Piechart

int ✓ 5. Race ✓ Piechart

Obj later 6 Renal Disease Indicator 65↑/Dis

int 7. State ✓ Crosstab barplot on Pandas

" 8. County ✓ —X—

Hospital Insurance 9. No of Months PartACov ✓ crosstab %

Medical Insurance 10 No of Months PartBCov ✓ crosstab

" ✓ 11. ChronicCond — Alzheimers

" ✓ 12. " " — Heartfailure

" ✓ 13. " " — KidneyDisease

" ✓ 14. " " — Cancer

" ✓ 15. " " — ObstrPulmonary

" ✓ 16. " " — Depression

" ✓ 17. " " — Diabetes

" ✓ 18. " " — IschemicHeart

" 19. " " — Osteoporasis

" 20. " " — rheumatoidarthritis

int 21. ChronicCond_Stroke
" 22. IPAnnualReimbursementAmt ✓ Box plot
" 23. " " DeductibleAmt ✓ Boxplot
" 24. OP " ReimbursementAmt ✓ Boxplot
" 25. OP " DeductibleAmt ✓ Boxplot
" 26.
" 27. Total Columns = 25
" 28.

3.12.2021 Train-Inpatient Columns :-
Friday

(-162) Obj 1. Bene-ID
" 2. ClaimID
claim Duration 3. ClaimStartDt } claim duration
" 4. ClaimEndDt
" 5. Provider — Just an ID.
int 6. InscClaimAmountReimbursed
Object 7. AttendingPhysician Just an ID.
Physicians 8. OperatingPhysician "
" 9. OtherPhysician "
Hospital Duration 10. AdmissionDt
" 11. ClmAdmitDiagnosisCode Hospital Stay Duration
int 12. DeductibleAmtPaid
Obj 13. DischargeDt
" 14. DiagnosisGroupCode
" 15. ClaimDiagnosisCode 1
" 16. " " " 2
" 17. " " " 3
" 18. " " " 4
" 19. " " " 5
" 20. " " " 6

object	21.	ClaimDiagnosisCode			7
"	22.	"	"	"	8
"	23.	"	"	"	9
"	24.	"	"	"	10
float	25.	"	Procedure	"	1
"	26.	"	"	"	2
"	27.	"	"	"	3
"	28.	"	"	"	4
"	29.	"	"	"	5
"	30.	"	"	"	6

Train — Outpatient Columns

object	1.	BeneID
"	2.	ClaimID
"	3.	ClaimStartDt
"	4.	ClaimEndDt
"	5.	Provider
int	6.	InscClaimAmtReimbursed
object	7.	AttendingPhysician
"	8.	OperatingPhysician
"	9.	OtherPhysician
"	10.	ClaimDiagnosisCode 1
"	11.	" " " 2
"	12.	" " " 3
"	13.	" " " 4
"	14.	" " " 5
"	15.	" " " 6
"	16.	" " " 7
"	17.	" " " 8
"	18.	" " " 9
"	19.	" " " 10

claim duration

Physicians

Pending

float	20.	Claim Procedure Code_1
"	21.	" " " _2
"	22.	" " " _3
"	23.	" " " _4
"	24.	" " " _5
"	25.	" " " _6
int	26.	DeductibleAmtPaid
object	27.	ClaimAdmitDiagnosisCode

Train Target

object	1	Provider
"	2	Potential Fraud

4.12.2021 Saturday (-161) No Work except GAN NIPS presentation by Goodfellow.

5.12.2021 Sunday (-160) GAN Main Research paper work contd. from 3.12.2021 Translation + Summarization which remained pending from 29.11.2021. 160-166. After 6 days.

* In summary, research highlights have been shared. Whereas on the task dated 29.11.21 original paper was published as reported.

6·12·2021 Monday (-159)	GAN overview paper Hindi translation.
7·12·2021 Tues (-158)	Doubts resolution and Reading
8·12·2021 Wed (-157)	Decision Trees χ^2 goodness-of-fit test Z-statistic Null-Hypothesis (H_0) ✓ Discrete & Continuous Distributions ✓ Binomial Distributions
9·12·2021 Thurs (-156)	Prob & Stats - Poisson Distribution.
10·12·2021 Friday (-155)	Read LilianWeng.github.io for Blogs to understand how to approach an ML problem → She is Applied Scientist at OpenAI.

Important Videos at the end of the Course.

Roles and Titles are open ended. Different companies call DS roles differently.

Data Scientist: DS spend a lot of time on end to end ML like data analysis and applying the analysis.

Data Analyst: SQL, Python, R for Data Analysis
- cleaning
- Outlier detection
- ~~Cler~~

Applied Scientist
- Mathematical rigour
- Cutting edge work modific-ation.

(DA + DS)

ML Engineer
- Bridge b/w SW Developer + ML & DS folks.
- Productionizes.

→ DL & ML - Topics from Live Sessions
- Transformers + BERT

→ Working on abstract
(1) Introduction:
(2) Research Section:
(3) First Cut Approach: How you'll solve the problem?

Pending (20)

Abstract ✓	done
	1) Introduction 1.1) Business Problem 1.1.1) ML formulation of the Business problem 1.1.2) Business Constraints 1.1.2.1) Data Column Analysis 1.1.2.2) All possible performance metrics
	2) Research 2.1) Min. 5 resources and explanation 2.2) How each one is useful
	3) First Cut Approach: How you'll solve the problem? 3.1) Observation 3.2) Takeaway 3.3) Weblinks 3.4) Diagrams 3.5) Advantages & Disadvantages
11.12.2021 Sat (-154)	- Working on abstract. ✓ - Working on data columns ✓
12.12.2021 Sun (-153)	- Transformers notes learnt till processes of all steps. - To start: positional encoding
13.12.2021 Mon (-152)	—X—

14.12.2021
Tuesday
(-151)

Reasons of Procrastination

A Unaccomplished Goals	Reasons
1. course completion	Didn't realize, I could have a better plan. Lack of experience in planning work.
2. Not learning the syllabus	Fear that I will forget and will have to revise again. Double work load.
3. Not bathing in the morning	Fear of water when it comes in contact with my hair after coming out of bathroom my hair gets cold and its too much work to dry them using a hair dryer.
4. Not praying on time in the morning	Just anger and too much time it takes.
5. folding clothes	Physical labour and efforts.
6. Utensils in dishwasher	Dirty work.
7. clothes in washing machine	Laziness
8. Watching videos of course work.	Boring lecture style.

How to Work for Improvement:-

1. Adv. (positive things) vs. Disadv. (-ve things) - Stress Reliever Exercise
2. overwhelming task - keep specific chunk of time.
3. Unclear & unvisualized flow of work. writedown steps for accomplishment.

4. Goal is unclear
5. fear of change. (Try different methods)
6. fear of failure.
7. overcommitment - I can do too many things
8. Addiction of working at last minute
(Break things in bits to complete)

* weekly planning (20-30 min exercise)
- Sunday night.
- make sure you understand the difference b/w imp & urgent.

* daily planning
- appointments
- emails
- organizer/software (No post it)
- Make sure it should be 65% time filled for your day. Mostly priorities.

* Art of saying 'no'
1. Give reason
2. Be diplomatic & don't make the other person hurt or angry.
3. Trade off - offer some other way to assist. Try do some other activity
4. Don't tell people "you'll think about it"
say.. I'm not ~~unsure~~ of ~~your~~ my calender.

* Planning tool
- Software planner or paper planner
- stick to one. Either/or!
- Make sure you don't fill it entirely.
- Only fill 60-65% of it and leave the rest. It should be flexible not rigid.

Self Help

- Don't use post its.

Eisenhower Decision Matrix	Urgent	Not urgent
Important	Do Do it now	Decide Schedule a time to do it later
Not important	Delegate who can do it for you (Tasks that need to be completed quickly)	Delete Eliminate it

Case Study 1 contd.

- describe about the train & test csv files.

Resourse Summary:

step 1. Analytics Vidhya Blog:

✓ Business constraints (5.) point 2 (imp.) - Interpretability point 3. (Latency requirements).

✓ How to handle oversampling.

✓ How to merge the data (preprocessing) → 10% percent.

1. Age 2. Category if Dead (Y/N) 3. claim duration 4. Hospital Duration.

✓ Creating Graphs

✓ Observations as well as takeaway

18-12-2021 Sat (-147)	Research Summary
Step 2:	Rohan Soni Medium:
	- How to raise suspicion on claim submitted for adjudication:
	1. Excessive prices charged on treatment / medicine
	2. Unusual high no. of invoices for a beneficiary in a short time frame.
	3. Billing for services which were not prescribed for the patient.
	4. Duplicate claim submission.
	5. Misrepresentation of the service provided
	6. charging for a more complex service than what was actually provided
	~~7.~~
Step 3.	NYC Data Science – very good EDA techniques
Step 4.	Names of models discussed
Step 5.	NCBI → RF, NB, K-fold CV.
First-Cut Approach	1. We need a good oversampling method. Not a SMOTE. SMOTE is based on KNN and is quite easy. It's result might not give us good results, although we can try.
	2. We need Feature Engg. to build features such as age from DOD - DOB, Y/N category if a beneficiary is ded or alive, claim & hospital stay duration.

3. combining the 3 different datasets to work on for getting best results.
4. We can see features which prove useful in predicting target by plotting graphs and with the help of that we can create more features.
5. For encoding categorical variables, mean/target encoding could be more useful than one-hot encoding. Although mean-encoding can lead to overfitting but use of regularizer can be beneficial.
6. We can use percentile methods to remove extreme outliers.
7. Models such as RF, GBDT-XGBoost will be tried first and then would go for SVM, LR and NB. & KNN

20-12-2021
Mon
(-145)
2021-W52

AlgoExpert.io
Coding Interview tips (5 Videos)
- How to best use algoexpert (30 mins)
- Coding Skills (23 mins)
- Problem Solving Ability (21 mins)
- Communication (37 mins)
- Culture Fit (19 mins)

Plan: 10 min → 2 min → (V. Watch, break, continue)

Time: 130 mins ⇓ 2 hrs 10 min

1. 10 → 2 → 10 → 2 → 10 → 2 | + 15 min break
2. 10 → 2 → 10 → 2 →
3. 3 → 2 | → 15 min break
4. 10 → 2 → 4 → 6 → 2 → 10 → 2 → 15 min break
5. 10 → 2 → 10 → 2 Left (1 min) → Break 15 min → 8:31 min → 1 + 9 → 2 → 10 → Break

Coding Interview Tips:

what will make you stand out?

* How to use AlgoExpert?

. How much time to prep for to interview?

→ 1-4 months. Also applies for more/less time

. What is your GOAL?

→ To get better at the interview.

→ Absorb the knowledge.

→ Try variations of questions.

→ As many questions as possible.

1. Tackle questions in order of difficulty.
   * Easy → Medium → Hard.
   * Mix them.
   * 10 Easy → 10 Medium → 10 Hard
2. Tackle them on Category.
   * Work on weaknesses
   ↓
   * Work on Strengths.
3. Do not take your strengths for granted.
4. Do not do an entire category.
5. Add timer to questions when you are done 50% of your prep.
6. Struggle through the problem before coming to the sol.
7. Use hints when you're really stuck, that too 1 at a time. Your goal should be to use as few hints as possible.
8. Running Code :→ Run as many times.

9. Use Run button very few times at the end of interview prep time.
10. Start few minutes of solving ques. to prep your sol. from scratch.
11. You're going to work through your own test cases. Take hints from test cases.
12. Optimal Space and Time complexity analysis after all the test cases have passed. Very Imp. If your optimality does not match or is worse than the one given, your sol. might not be optimal. Interview process → You would be given hints → "can you make it more optimal".
13. For sol. compare both your sol. and the correct sol.
14. Go through the video given to make you understand the correct sol. You will get wisdom.

* Coding Skills
- Descriptive variable naming
→ Name your " that your interviewers understand your sol^n. Use abbreviation in confirmation with your interviewer.
- Abstraction
→ Separate your code into logical chunks.
- Documentation
→ Esp. for difficult steps.
- Descriptive Code
→ for chunks of code that are difficult to read & understand.

- Idiomatic Coding Style.

→

- Testing

→ Create solid tests/edge/normal cases.

1 or 2 mistakes are not the end of the world.

* Problem Solving Ability

1. Eliminate ambiguity
   - If you are given a ques.
   - → Ask clarifying question.
   - → Ask edge cases
2. Are these a single problem or problem within a problem? Is there a repeated step?
3. Discussing tradeoffs or pros/cons.
   - I'm going to solve this using Recursion
     How is it beneficial?
   - Memoization or caching
     Benefit: faster Runtime
     Drawbacks: Uses Extra memory
4. Implementing in an organized manner.
   - May be you can by abstracting by writing the name of the helper method but not ~~writing the~~ filling the method really.
   - Ask the interviewer if thats fine.
5. Testing proactively
   - Testing non-trivial/non obvious.

Pending

* Communication
  1. Eliminating Ambiguity
     - Purposefully leave out details.
  2. Voicing your thought process → Don't let interviewer guess wrong things.
  3. Discuss Tradeoffs.
  4. Proactively asking for approval, such as, you might be using abbr. for certain long variables names.

→ You cannot overdo them. You need to maintain a balance b/w all the 5 points.
→ Don't describe each and every line.
→ Don't ask help too often.

* Culture fit

160 Coding Interview Questions

E

1	3	5	7	9	11	13	15	17
2	4	6	8	10	12	14	16	18

19	21	23	25
20	22	24	

M

1	3	5	7	9	11	13	15
2	4	6	8	10	12	14	16

17	19	21	23	25	27	29	31
18	20	22	24	26	28	30	32

33	34	35	36	37	38	39	40

Pending

	41	43	45	47	49	51	53		
	42	44	46	48	50	52	54		
	55								
H	1	3	5	7	9	11	13	15	17
	2	4	6	8	10	12	14	16	18
	19	21	23	25	27	29	31	33	35
	20	22	24	26	28	30	32	34	36
	37	39	41	43	45				
	38	40	42	44	46				
VH	1	3	5	7	9	11	13	15	17
	2	4	6	8	10	12	14	16	18
	19	21	23	25	27	29	31	33	
	20	22	24	26	28	30	32	34	

Case Study 1

To start with EDA in stages:-

1. Data Preprocessing
2. Exploratory Data Analysis

Submit Abstract again

focus on → EDA & FE

Plots : ① Detailed Explanation
② Analysis
③ Conclusions

Pending

EDA → In-depth

Plots.
1) What are you plotting b/w what features
2) Detailed analysis of plots.
If anomalies are found, I need to further investigate it.
3) No one-liners.

Follow :-
① How does a particular feature(s) impact the target variable?
② Try feature selection (10-15) check how these features impact the target variable

Live session b4 starting with EDA

1)	2)	3)
M2L	M2L	M3L
10.1	10.2	12.3
(2:5min) (2hrs 5min)	(2:4min) hrs	(2hrs)

③ Use doc-strings & mark-down cells
④ → Explain Exact Business Problem and who all benefits from this solution.
→ why this problem has to be solved?
→ Use-case and how my model improves that
→ why this problem has to be solved
→ why ML model is needed to solve the problem.

Pending

15 min break

Start M2L 10.1 (120 mins)	10 → 2	10 → 2	10 → 2	10 → 2	10 → 2
	10 → 2	10 → 2	10 → 2	10 → 2	10 → 2
	10 → 2	15 min break			15 min break
	10 → 2				

21-12-2021 IGF ch 1	Tuesday → (-144) (W-52) → 2021 Linear Algebra and other milestones time estimate.

22-12-2021 (-143) Wed 2021-W52	23-12-2021 (-142) Thurs 2021-W52	24-12-2021 (-141) Fri 2021-W52	25-12-2021 (-140) Sat 2021-W52	BREAK

26-12-2021 (-139) Sunday 2021-W52	To prepare proper headings for EDA colab sheets for both the datasets.
27-12-2021 (-138) Monday 2022-W1	11-1pm Week 1 Revision. contd. throughout the day (not well)
28-Dec-2021 (-137) Tuesday 2022-W1	Yesterday's pending work completion Todays work started @ 11:30 AM

29-Dec-2021 (-136) Wednesday (2022-W1)	3 videos finalized from Decision Trees General Topic and Ensembles Common Topic.
30-Dec-2021 (-135) Thursday (2022-W1)	11-1pm case study 1 • To start next time from chronic cond. Ob Pulmonary
(31-1)-Dec'-21 (-134-133) Friday, Saturday (2022-W1)	No work. New Year's eve and New Year 2022 "Party!"
2,3,4-~~Dec~~ Jan'-22 -132,-131,-130, (-129) Sunday, Mon, Tuesday (2022-W1-2)	No work "Boredom."
5th Jan'22 (-128) Wednesday (2022-W2)	No work could be started due to space declutter.

P.T.O.

6.1.2022 Thursday (-127) (2022-W2)	Very Unproductive day. Still It was OK!
7.1.2022 Friday (-126) (2022-W2)	Univariate Analysis - EDA. 1. Gender — value count, Pie chart 2. Age (From DOB) 3. Race — Pie Chart 4. State — Crosstab, Barplot 5. County Finished watching the first Reference EDA Live Session. Look at the notes over there. Imp. discussion in that live session.
8.1.2022 Saturday (-126) 2022-W2-D7	Watching video reference 2 for Case Study 1 Link on Notes.
9.1.2022 Sunday (-125) 2022-W3-D1	• Completing ½ hour leftover video from Ref. 2. • Started working on Ref. 3 video • Reading for Academic Success, Grades 2-6: Differentiated Strategies for Struggling, Average, and Advanced Readers By: Richard W. Strong, Harvey F. Silver and Matthew J. Perini Publisher: Sage Publications Date: 2007-8-28

2022 - W3-D6 Jan 14th 5:25pm	Sent: 1. Mail to Applied AI for finding the combined csv files and start working on feature importances.

PROGRESS CHART

2022-W3-D3 Jan 17th 3:40 pm	CNN on CIFR	Fraud Detection Analysis
	Abstract Submitted after couple of questions.	Mail sent: 5/July/2021
	Reply: 10th Dec 2021	Need to understand. Certain links present in the google drive folder.
	Suggestions for improvement shared.	Didn't understand at first place, what needs to be done in the assignment.
	Mail Sent: 10th Dec 2021	
	Asked about the hierarchical flow of structure for pointwise ~~discussions on~~, how to solve the case study.	Mail Sent: 2nd Oct 2021 DenseNet paper approach.
	Reply: 11th Dec 2021	Reply: 3rd Oct 2021 Image classification problem. Experiment by changing the no. of units in the densenet architecture to see how the accuracy changes. Reference link shared.
	Hierarchy is correct.	
	Mail Sent: 11th Dec 2021	
	Asked about what to do for columns where metadata is unknown.	

Fraud Detection Analysis.	CNN on CIFR
Reply: 13th Dec 2021	Mail Sent: 20th Oct 2021
Suggested to leave unnecessary columns and focus on imp. ones	Didn't understand the first cut approach of this assignment.
Abstract Submitted: 18th Dec 2021	Reply: 21 Oct 2021
Abstract Accepted: 20th Dec 2021	To do image augmentation for the images present in the data. Reference link shared.
Reply: 20th Dec 2021	
Asked to start EDA and then will arrange for a call to discuss how to proceed further.	Certain parameter corrections were suggested, like, batch size, 1, num_filter, compression, conversion of X_train & X_test to float32 ImageDataGenerator shear range, zoom_range, height_shift range, horizontal flip. A suggestion: To train the model in stages for 10 epochs and save the model weights and train again on best score to reduce the learning rate accordingly.
Mail Sent: 21st Dec 2021	
Asked a few unrelated questions	
Reply: 22nd Dec 2021	
Asked to focus on EDA only.	
Mail Sent: 14th Jan 2022	
Asked to help on feature importances.	

Fraud Detection Analysis	CNN on CIFR
Reply: 15th Jan 2022	Mail Sent: 21st Oct 2021
Asked to use SelectK Best if features are more than 300 otherwise do it for all.	Needed to understand a few things: 1. Unable to save previous epoch staging wts. and reload them for the next epoch staging. 2. New Loss and accuracy overfitting as compared to validation loss and validation accuracy.
Mail Sent: 25th Jan 2022	
Asked team doubt on shape difference b/w claim duration and dataset shape.	Reply: 21st Oct 2021
Reply: 26th Jan 2022	Callback correction Suggested: Model checkpoint
Use a stackoverflow method.	Mail Sent: 23rd Oct 21
Mail Sent: 28th Jan 2022	Didn't still understand, how to load the saved weights.
Still a ~~wait~~ diffeger again.	
Corrected on self. ~~Reply~~:	Reply: Correction suggested
	Mail sent: 24th Oct 2021
Mail Sent: 29th Jan 2022	
Bivariate Analysis Doubts	Didn't understand how to improve the learning rate.

Fraud Detection Analysis	CNN on CIFR
Reply: 31st Jan 2022	Reply: 25th Oct 2021
1. Domain knowledge needs to be used here. - Do FE. Read Blogs & Kernels - After features are created, we can then use those features to do EDA wrt target.	Please use learning rate scheduler to reduce learning rate scheduler or use ReduceLROnPlateau
2. Read others solutions Do FE based on domain expertise and intuition	Mail Sent: 18-1-2022
3. We can mention person is dead is alive like features. Yes, we can put death as zero for finding age.	Asked team to share some references for understanding ReduceLROn Plateau.
	Reply: 18-1-2022
	Asked to read docs and reference link shared.

2022-W4-D3 Jan 17th 7:35pm	Fraud Detection Analysis Train inpatient. → Refer past pages for understanding datatypes of columns.
2022-W4-D6 Jan 21st 12:32 pm	CNN on CIFAR assignment in fresh file without using dropout.
2022-W4-D7 Jan 22nd 5:45pm	Finishing CNN
2022-W5-D3 Jan 25th 4:25pm	EDA finish/not
2022-W5-D6 Jan 28th 9:23am	EDA doubt sent to AppliedAI Team.
2022-W6-D2 Jan 31st 8:45am	Tableau Part - I (20.11) → 1:50 ~~min~~ Hr. Data Interpreter → cleans data in Tableau Case Study 1 → EDA (Bivariate Analysis & FE) Thoughts: • What to do when a person is_dead → Y/N if Y, we can find age. If not? Should we make DOD as zero other than NAN. Or should we take age from current date which is not right approach ofcourse

Q How to create isdead flag → Yes/No
iterate through ds and create new colum after ~~def~~ & setting if-else where if Age == NaN set isDead = Y else N.

Not working Despite many efforts

Q For which all categorical columns in the ds, can we create scatterplot?

Worked for 7 hours.

~~2022-W6-D~~
~~2022-W6-D4~~
~~Jan~~
2022-W6-D4
Feb 2nd
12:05am

FE → Healthcare Provider Fraud Detection Analysis

Physicians FE.

Case Study 1

Exploratory Data Analysis :-

2022-Feb-12 (1) **Beneficiary Data :-**

Day 1

✓ NOTNULL BeneID

✓ P ✓ DOB - DOD ⇒ Age
✓ P ✓ IsDead

✓ NOTNULL Gender

✓ NOTNULL Race

✓NOTNULL	chronic conditions: 12 in no → Patient Risk Score
✓NOTNULL	Reimbursement: IP, OP
✓NOTNULL	Deduction: IP, OP
	(II) Inpatient Data :-
	BeneID
	claimID
P ✓	claim Duration — Start, End
	Provider
	claim Amount Reimbursed
P ✓	Physicians — Attending, operating, other
P ✓	Hospital Stay Duration — Admission, Discharge
	(III) Outpatient Data :-
	BeneID
	claimID
P ✓	claim Duration — Start, End

Provider

claim Amt Reimbursed

P ✓ Physician — Attending
Operating
other

columns for Inpatient — Doubts

1. Diagnosis Group Code
2. claim Diagnosis Code 1-10
3. claim Procedure Code 1-6
4. Deductable Amount Paid
5. claim Admit Diagnosis code

columns for Outpatient — Doubts

1. claim Diagnosis Code 1-10
2. claim Procedure Code 1-6
3. Deductable Amount Paid
4. claim Admit Diagnosis Code.

2022-W8-D4 Feb 16th 12:40pm	2 hours for chapter 8 & 9 in Module 1.
2022-W8-D6 Feb 18th 8:00am	Revision contd. Module 1 chapter 8 :- Computational Complexity.
11:45am	Module 1 chapter 9 :- SQL.
	Start Tomorrow from 9.3.
1:30pm	Assignment Case Study 1.
	Diagnosis Group Code. Inpatient ~~Outpatient.~~
	NOT NULL type of null.
2022-W8-D7 Feb 19th 12:35pm	Revision Hours completed till now are :- 16.34/199
2	

M2	ch13	13.32	17.5	22.11
ch11	13.1	13.33	17.6	22.12
11.1	13.2	13.34	17.7	22.13
11.2	13.3	13.35	~~[illegible]~~	22.14
11.3	13.4	13.36	M3 ~~ch20~~	22.15
11.4	13.5	ch15	21.1	22.16
11.5	13.6	15.1	21.2	22.17
11.6	13.7	15.2	21.3	22.18
11.7	13.8	15.3	21.4	22.19
11.8	13.9	15.4	21.5	22.20
11.9	13.10	15.5	21.6	22.21
11.10	13.11	15.6	21.7	22.22
11.11	13.12	15.7	21.8	22.23
11.12	13.13	15.8	21.9	22.24
11.13	13.14	15.9	21.10	22.25
11.14	13.15	15.10	21.11	22.26
11.15	13.16	ch16	21.12	22.27
11.16	13.17	16.1	21.13	22.28
~~17~~	13.18	16.2	21.14	22.29
~~18~~	13.19	16.3	21.15	22.30
ch12	13.20	16.4	21.16	22.31
12.1	13.21	16.5	ch22	~~ch24~~
12.2	13.22	16.6	22.1	
12.3	13.23	16.7	22.2	
12.4	13.24	16.8	22.3	
12.5	13.25	16.9	22.4	
12.6	13.26	16.10	22.5	
12.7	13.27	ch17	22.6	
12.8	13.28	17.1	22.7	
12.9	13.29	17.2	22.8	
12.10	13.30	17.3	22.9	
	13.31	17.4	22.10	

NOTE:- when the mentor calls to discuss on how to do imputation, just check what is the use of those columns in Health Insurance.

Bene ID NO USE
Gender Binary ✓
Race
State ✓ done Scatter ✓
County
No. of Month PartA Cov
" PartB "
IP Annual Reimbursement Amt
IP Annual Deductible Amt
OP Annual Reimbursement Amt
OP Annual Deductible Amt
Age ✓ done Scatter done
is Dead Binary ✓ done
Risk Score ✓ done distplot
Claim ID NO USE
Provider
Insclaim Amt Reimbursed
Claim duration
Physicians ✓ done ✓
Potential Fraud. NO USE

Gender
Race
State ✓ I
County
No. of Month PART A
" " " B II
IPARA
IPADA III
OPARA
OPARA IV

PDF very good analysis

Provider — 5 ✓
Insc AR — 6
Claim duration — 7 ✓
Potential Fraud.

4/March 2022

last point in sum
↳ Regression probl
pdf & cdf

Discussion Notes

done

outer join issue

Correction Left Outer Join

Target → Provider (1)

To show %wise class division of +ve & -ve class.

% class-wise imbalance.

If class 1 is important

class 1 → Fraud +ve class.

Precision, Recall, F1

Swap — 1 with 0

Business.

F1-Score. & Confusion-Matrix

~~ROC~~

Null-Imputation

1) why is data missing?

Linear / Real / Num. → Mean-Median

Categorical → Mode.

If missing >15% → make another cat

Threshold → for dropping
Do above and → do the below.

* If 50% or half - Missing Value.

1) MICE → linear-regression.

Time Taken 2) KNN → imputer.
Time & Memory toomuch.

But, it depends upon performance, Time & latency.

* How to handle Categorical data

① Ordinal / Label Encoding
↓
order

* ② Non-categorical (Nominal)
OHE / Mean Encoding

* 6 - column
* Row → 2^6.

2hrs ~~120~~ 40 / 3

~~80~~ 23 / 3 (60)

12/4/2022 Timeline

	List of Tasks to accomplish	Status
10%	1. Left Outer Join implementation.	✓
10%	2. Show %wise class division of positive and negative classes	✓
	3. Important Metrics :-	✓
25% Homework	1. Accuracy, Since dataset is almost balance	
	2. Precision, Recall & f1-Score Imp. Business wise	
	3. Confusion Matrix	
	4. Null-Imputation - Imporvization. Reason :- Null values cannot be replaced in case of skip patterns. 1. why is data missing? → OHE → Feature Importance	
25% Homework	Use boxplot to check distribution and type of distribution. No outliers + when data is symmetric / when data is skewed	
outliers highly impact the Model	- Linear / - Real / - Numerical } Mean - Median (Symmetric / Skewed)	
60%	* Categorical } - Mode (Interpolation Techniques) when data is skewed.	

How?

- If missing data is >15%, then, make another category (Threshold for dropping)

Look for this.

* If ≥ 50% data is missing, then MICE algo using linear-regression. (Good choice)

But some people try KNN (time & Memory intensive)

Not required

X 5. How to handle categorical data?

25%

1. Ordinal/Label-Encoding (order-wise)

~~Nominal~~

2. Non-Categorical (Nominal) OHE/Mean Encoding.

60%

NOTE

If there are, for eg, 6 columns, then the total rows should be around 2^6.

2 ∵ If we expect for each column, there should be positive & negative Target label.

13/4/2022

Missing Values

≥15 %

		Missing Rows
1. BenefID 8%	object	3318 Rows
2. claimID 8%	"	"
3. Insc. Claim Amt Reimbursed 8%	float ①	"
~~4. clm Admit Diagnosis Code~~	~~object ①~~	~~"~~
~~5. Diagnosis Group Code~~	~~" ②~~	~~"~~
~~6. clm Diagnosis Code 1~~	~~" ③~~	~~"~~
~~7. clm Diagnosis Code 2~~	" ④	3544 Rows
8. claim_duration 8%	float ②	3318 Rows
~~9. Physicians~~	~~object ③~~	~~"~~
10. Hospital_Stay_Duration 8%	float ③	"
~~11. Gender~~	~~float ④~~	~~"~~
~~12. Race~~	~~float ⑤~~	~~"~~

categorical

~~13. State~~	~~float~~	~~8~~	~~3318 Rows~~
~~14. County~~	~~float~~	~~9~~	~~"~~
15. No. of Months_Part A Cov 8%	"	8	"
16. No. of Months_Part B Cov 8%	"	9	"
17. IPAnnualReimbursement Amt 8%	"	10	"
18. IPAnnualDeductible Amt. 8%	"	11	"
19. Age 8%	"	12	"
~~20. is Dead~~	~~object~~	~~10~~	~~"~~
~~21. RiskScore categorical~~	~~float~~	~~11~~	~~"~~

Total features where Null 8% 21

* OPAnnualReimbursementAmt	float	8%
* OPAnnualDeductibleAmt	float	8%

Univariate Imputation	Multivariate Imputation
Imputes values in the i-th feature dimension using only non-missing values in that feature	Imputes values using the entire set of available feature dimensions to estimate the missing values
eg. impute.SimpleImputer	eg. impute.IterativeImputer * experimental for now

7~8% — ≤15%. Interpolation Techniques for Missing value Replacement

Feature	dtype	Cuc
~~clmDiagnosis Code 3,4,5~~	object	For some valid reason it would not be okay, to go ahead with replacing missing value in the codes section. Instead go for making a category and splitting the features into Yes(1) No(0). Reason? * For some claims if code1 applies, ~~so~~ the other code might not apply. This is very clearly written in MICE algo's explanation on ncbi under paid... "Model specifications" The specific term associated is "Skip Pattern" & is left intentionally blank.

22/4/2022

Flowchart 1° code पाने columns

1	2	3	4	5	6

↓ null values

New category / Rename it

↓

Rest → OHE

Similarly with codes 1–10.

Basic

ensemble 1 Bagging 2 Boosting 3 Stacking 4. cascading

Resources

Modelling

Topic: Case Study 1.
Date: 26/4/2022

Goal Setting

Modeling for Binary classification

Target to achieve

Understanding Problem Statement First ⇒ Ask Important Doubts...

1.1 Doubt

Which model categories to use in present case study?

Important

① Whether to remove BeneID & claimID? ② ∵ they are no longer needed.

Answer

— NLP based:
1. KNN
2. Naive Bayes
3. Logistic Regression

— Supervised Learning
1. SVM
2. Decision Trees
3. Random Forests GBDT

2. Should we stack algorithms based on best results on all the metrics?

Yes

— How do we analyse? Top 2 Top 3, how many? what criteria to decide; how many

tSNE for each

3. Hierarchy of subtasks in each algorithm:—

P.T.O

NLP Based

3.1 KNN

Doubt - Important

1. Upto how many features after vectorization are to be used in Modeling phase?

Ans

1. Hyperparameter Tuning → (1) mean train & test score (2) std train & test score
2. Metrics Functions
   a. Accuracy
   b. Precision, Recall & F1-Score
   c. Confusion Matrix

Aim: (1) Reduce False Negative ∵ they are Non-Fraud.

(2) Improve/Increase False positives ∵ they are frauds.

Oversampling SMOTE

(3) Business Metric → F1-Score.

<300 ↓ No need

>300 ↓ Select K Best

3.2 Naive Bayes

1. Hyperparameter Tuning:-
   (1) alpha
2. Metric functions
   a. Accuracy
   b. Precision, Recall & F1-Score
   c. Confusion Matrix

Aim: (1) Reduce False Negatives ∵ they are Non-Frauds

(2) Improve/Increase False positives ∵ they are frauds.

2 columns ↓ rows ↓

(3) Business Metric → F1-Score.

random no.	0.6	0.4
	0	1 (P)

Base Model

Important Doubt: Can we also use clustering type of Unsupervised Learning algorithms? Only in case we don't have target

3.3 Logistic Regression.

1. Hyperparameter Tuning :- (1) C

2. Metric Functions
   a. Accuracy
   b. Precision, Recall & F1-Score
   c. Confusion Matrix

Tools: CM, F1 Score

Aim :- (1) Reduce False Negatives ∵ they are Non-Frauds.
(2) Improve/Increase False positives ∵ they are frauds.

Important Doubt: Do we need to remove Provider columns in each train & test, if they are repeating? Checked in value counts. remove vectorize group by

3. Business Metric → F1-Score.

3.4 SVM Supervised Based

1. Hyperparameter Tuning :- (1) Alpha

2. Metric Functions
   a. Accuracy
   b. Precision, Recall & F1-Score
   c. Confusion Matrix

Aim :- (1) Reduce False Negatives ∵ they are Non-Frauds
(2) Improve/Increase False positives ∵ they are frauds.

3.5 Decision Trees

① Hyperparameter Tuning
  ① max_depth
  ② min_sample_split

② Metric Functions
  a. Accuracy
  b. Precision, Recall, F1-score
  c. Confusion Matrix

③ Business Metric ⇒ F1-score.

3.6 Random Forest GBDT

① Hyperparameter Tuning
  ① max_depth
  ② n_estimators

② Metric Functions
  a. Accuracy
  b. Precision, Recall, F1-score.
  c. Confusion Matrix.

③ Business Metric → F1-Score.

4. Threshold for selecting best models based on selected metrics. Criteria of selecting the no. of best models. Should we used stacked models?

5. Final Selection of Model Criteria :-
  a. Basic Metrics b. Business Metrics

6. References -
  1. Sklearn Tutorials 2. Course Notes
  3. Course Videos 4. Solved Assignments

Planning Tasks

1. Team Discussion.
   a. Presenting Ideas
   b. Brain Storming
2. Revisit References
3. First Draft
4. ~~Sub~~ First Submission.
5. Second Draft
6. Second Submission
7. Final Draft
8. Final Submission.

Timeline

1. Team Discussion.	1 Hour	29th April
2. Revisit References.	1 Day	
3. First Draft	1 Day	30th April
4. First Submission	-do-	
5. Second Draft	1 Day	1st May
6. Second Submission	-do-	
7. Final Draft	1 Day	2nd May
8. Final Submission	-do-	

Deadline Expected Monday Night → 2nd May 2022

checking Progress Percentage

Day 1		✓ X	Day 2 ~~Flow~~		✓ X
1.	~~Data Vectorization – OHE, etc~~		1.	~~Task~~ Flow 1	
2.	~~Colab Index Creation~~		2.	Flow 2	
3.	~~checking past related assignments~~				
4.	~~Reading References and video/blogs~~				
Total	Not Estimated correctly 60%		Total		

Day 3		✓ X	Day 4.		✓ X
1.			1.		
Total			Total		

27/4/2022

Execution Phase — Day 1

Task-1. Data Vectorization.

Function — OHE-Vectorizing Categorical Data

```
def categorical (train_col, test_col):
obj = CountVectorizer()
obj.fit(traincolumn)
X_train_Val_ohe(var) = obj.transform(train_col)
                 hint
X_test_val_ohe(var) = obj.transform(test_col)
                 hint
ft_name = obj.get_feature_names()
return (X_train....., Xtest....., ft_name)
```

Example

```
tra1 = X_train['col_name'].values
tes1 = X_test['col_name'].values

~~X_test~~
X_train_col_ohe, X_test_state_ohe,
        name
feat1 = Categorical(tra1, ~~tra~~ tes1)
print("After Vectorization")
print(X_train_~~state~~_ohe.shape,
      y_train.shape)
print(X_test_col_ohe.shape,
      y_test.shape)
```

P.T.O

contd.

```
print(feat1)
print("#"*100).
```

Categorical Features

1. claim Admit Diagnosis Code
2. ~~Clai~~ Diagnosis Group Code
3. claim Diagnosis Code 1-10
4. Claim Procedure Code 1-3
5. Physicians
6. Gender
7. Race
8. State → } Num
9. County → } Num
10. isDead
11. Risk Score. → Num

15 mins

@ 10 minutes

Numerical Features

Normalizer()

1. Insc Claim Amt Reimbursed
2. Deductible Amt Paid
3. claim Duration
4. Hospital Stay duration
5. No. of Months of Part A coverage
6. No. of Months of Part B Coverage
7. IP Annual Reimbursement Amt
8. IP Annual Deductible Amt
9. Age
10. OP Annual Reimbursement Amt.
11. OP Annual Deductible Amt

24 mins done

Note When you will be doing vectorization of features, you then need feature importance.

Flowchart Diagram.

Flow 1

Data Vectorization

↓

✓ Numerical Data

✓ Categorical Data

hstacking except Provider cols.

↓

Group-by

↓

hstack provider col

Flow 2

Functions

- Algorithms (6 types) including Hyperparameter Tuning + Scores
- Metrics
  - Accuracy
  - Precision + Recall
  - Confusion Matrix
  - Business Metric → F1-Score

Flow 3

↓

Base Model

↓

Random no. generation in the ratio of 50:50 in total size of no. of train data.

↓

~~Rom~~ Decision Tree

↓

Random Data + Actual Target values

↓

Metric

↓

Confusion Matrix

↓

How is the Base Model performing?

→ check on the basis of:

① FP ↑

② FN ↓

↓

Discussion with Mentor.

Flow 4

Once the Base Model performance is checked.

- KNN
- Naive Bayes
- Logistic Regression
- SVM
- Decision Trees
- Random Forest GBDT

checking performance on the basis of Metric Score

- Accuracy
- Precision + Recall
- Confusion Matrix
- Business Metric → F1-Score.

Ensembling for top 2 to 3 best models

Final Model → Test Predictions

6th May 2022

Flow 1

- Groupby → hstack → pickle files.
- Group by.

* You need to group providers
* Provider converted to category type.
* So no need of grouping.

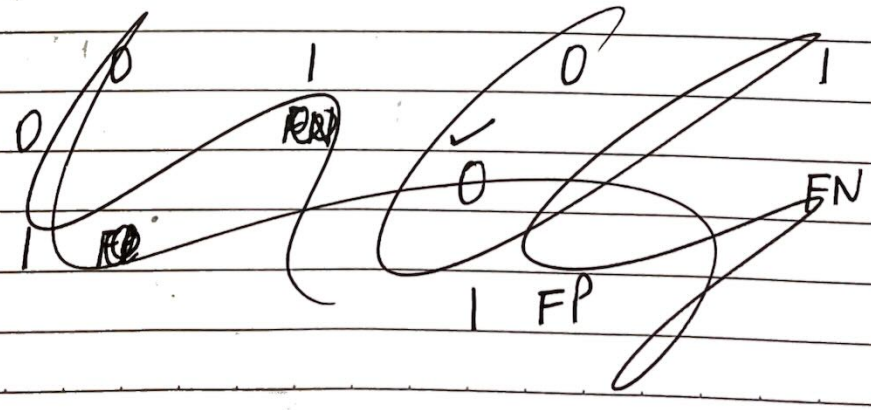
Simply do OHE.

- hstack.

11th May 2022.

Data Leakage.

- Vectorization. association with data leakage.



12 May 2022

$C_{r,c}$

FP $C_{0,1}$

FN $C_{1,0}$

TN $C_{0,0}$

TP $C_{1,1}$

F	P	C_{01}
F	N	C_{10}
T	N	C_{00}
T	P	C_{11}

0 1

0 TN FP ↑ P

1 FN ↓ TP

R

0	TN	FP
1	FN	TP

logistic Regression → C

Naive Bayes → alpha

SVM classi-fier ⇒ SGD + hinge loss → alpha

13 May 2022

	Metrics.	Logistic Regression	Decision Trees	Random Forests
1.	Best Hyperparameter	C = 1000	max_depth = ~~1000~~ 500 min sample split = ~~10~~ 5	max_depth = 10 n_estimators = 1000
2.	Train Accuracy	1.0	~~1.0~~ 0.49	~~0.788~~ 0.805
3.	CV Accuracy	0.996	~~0.997~~ 0.50	~~0.997~~ 0.828 0.839
4.	Train FN	0	~~0~~ 767	~~1685~~ 2093
5.	BCV FN	~~54~~ 44	~~34~~ 3273	~~1492~~ 1484
6.	Train FP	0	~~0~~ 7676	~~4826~~ 3869
7.	CV FP	4	~~0~~ 3247	~~767~~ 825
8.	F1-Score	1.0	~~1.0~~ 0.535	~~0.819~~ 0.82
9.	Train Precision	1.0	~~1.0~~ 0.532	~~0.753~~ 0.787
10.	CV Precision	0.999	~~1.0~~ 0.537	~~0.878~~ 0.898
11.	Train Recall	1.0	~~1.0~~ 0.532	~~0.897~~ 0.872
12.	CV Recall	0.992	~~0.995~~ 0.535	~~0.788~~ 0.789

14th May 2022.

Revision | classification Metrics

1. Accuracy

15th May 2022 SCS-1 Final submission Suggestion Queries.

(2) 1. Watch the instruction video from the beginning and make notes:-.

~~Notes~~: I II

(1) 2. Live Session → classroom Instruction → Abstract document.

III

16th May 2022 Monday | Final Notebook Submission. | Mentor Amarnath Sharma

1. a) Go through live Session ✓
   b) Then instructions in the classroom
   c) Then additional instruction @ end of abstract doc.

2. Live Session
   - Link in mail ✓
   - 125 minutes ✓
   - Watch in parts ✓
   - Make notes ✓
   - Save links ✓

3. Final Notebook instructions are @ end of abstract
   a. Don't train any model but ML best model.
   b. Don't read train data files
   c. Instructions :-

```
1. function1(x)        Id feature
   {
     preprocess data while featurizing
   }

eg → def final(x):
     {
       preprocess {
                    data cleaning;
                    filling missing value;
                    etc;
                  }
       compute features based on X
       use pre-trained model
       return predicted outputs
     }
```

Suppose, final (x) takes argument time & location parameters, then predict no. of pickups that can happen.

d?

e. final([time, location])
{
return predictions & No metric.
}

f.?

4. - After preprocessing data point X above, featurize it.
- Use Trained vectorizers/methods & using method followed for train data.

5. - Calculate time for predictions & report it
- keep time to be as low as possible.

Doubts 3 in Abstract end.

6. ? - Please predict for single datapoint, not the whole dataset.
- Use best model for prediction
- Replicate as if you're doing real time prediction.
- Take 1 input in final method.
- Use any other file.
- Do same for deployment.

7. Deploy model in :-
- heroku /
- AWS /
- GCP /
- Local

8. Record video of deployment making predictions & share link.
- Use same video in your blog.

9. Create either html (if features are less)
- or html page taking .txt file as i/p
- .txt file can contain values in a comma separated.
- Deployment is compulsory.

10. Start blog also.

11. <u>Note</u>.

1. <u>2 functions</u>.

<u>Syntax</u> :-

def final_1 (X):
{
Use best model from experiments
X (Raw data)
}

- Function 1
  - Entire pipeline
    - data preprocessing → Final predictions
  - Raw data input
  - Return predictions
  - i/p single point / set of points

Syntax :-

```
def final2(X,Y)
{
Use best
model from
experiments
&
return
final-metric
on X(Raw data)
& Y(target variable)
}
```

- Function 2
  - Entire pipeline
    - data preprocessing
    - Final predictions.
  - Take raw input along with target
  - Return metric to judge your model.

12. After submission, mentor might connect on call for evaluation of SCS-1. Explain every single line of code written.

Note 1

Note 2 Before acceptance of final submission with all modifications, call/anydesk would be done by mentor to execute the final.ipynb.

So save model parameters

Note 3 | Start blog only after acceptance of SCS-1.

17th May '22 | Solving Final Deployment Task.

1. a) Live Session.
- 25 mins ✓
- 25 mins 26:46
- 25 mins 39:47
- 25 mins
- 25 mins

Notes

Hands on Live Session:-

Deploy an ML model using Flask API on AWS.

Suggestions:-

1. R. paper/week after finishing the course.

Steps:

1. Build model on local box
2. Store model & other key model related variables in .pkl files.
   ↓
   pickle
3. Launch micro instance on AWS.
4. Connect to the AWS box(.ssh)
5. Move the files to AWS EC2 instance/box [scp]
6. Install packages on AWS box
   a. pip3
   b. pandas
   c. numpy
   d. sklearn
   e. beautifulsoup 4
   f. lxml
   g. flask
   h. re

   #copy above packages and do the above 6th point.
7. Run app.py on AWS box
8. check output in the browser.

check Session Notes. It has Screenshots along side.

19 May 2022

55:37 → 1:10 hr → 1:27 hr → 1:34 hr

20th May 2022 Done.

Work flow

1. Requirements
2. How to implement
3. What all things you'll do. Well structured plan.
4. Final deployment.

1. Requirements — 4 min

root /AFR/{a-e}

a. Database — SQL lite
b. model.py
c. model.pkl
d. count_vect.pkl
e. app.py

.../AFR/template/index.html

2. How to implement? — 2 min

Packages needed

- pip 3
- pandas
- numpy
- sklearn
- flask

3. Well structured Plan: 2min

- Install all the packages.
- Understand how ~~are~~ ~~at~~ is the flow going on in the project.

4. Final work and deployment

3:30 — 4:15 pm. = 45mins

Train

Model = pickle.

✓ Decision Tree.

Test.

✓

R preprocess

Bene Inpati Out

✓

provider

Model

↓

Pandas

Downloads

* Weka.

View

↓

controller

— X —

Dataset Read → Data Preprocessing

[Keep ipynb notebook on left and Sublime on right → start working]