YDS 2018 - Datathon
MEDIPLEXIS

Version 1.1

Pune | +91 20 6606 5000

Problem Scenario

An pharma company Mediplexis™, based in the United States, has several products in the various therapy areas. It markets these products to physicians all over the country through various channels -

1. Sales representative (Medical rep - *Rep_Live*)
2. Remote representative (Medical rep via video conference – *Rep_Remote*)
3. Peer to Peer marketing programs (such as conferences, etc. – *P2P*)
4. Digital message (such as mobile messages – *Digital_Push*)
5. Online video (such as online video, etc – *Digital_Pull*)
6. Direct mail (Snail mail – *Direct_Mail*)
7. Online advertisement (Such as website banner – *Digital_Pull*)
8. Digital Email (such as Email – *Digital_Push*)

However, it's current marketing efforts are not proving to be very successful – Physicians are not engaging with the content delivered to them through these various channels. Mediplexis™ now wants to improve its targeting strategy by analyzing the "***affinity***" of each physician to the above channels.

"Affinity" data has been collected for various physicians through historical channel interactions with doctors and every doctor has been given an affinity rating for each of the above 8 channels on a scale of 0 – 1 however lot of HCPs do not have all channel information.

Develop a approach to assign affinity for every tactic of HCP in the "*SUBMISSION*" file

Dataset

HCP_ID	RL	P2P	OLV	RR	DRT	DMS	OLA	DEM	Region	Value	Speciality_ID
9304	0.645833	0.475755	NA	NA	0	0.169355	NA	0.004891	urban	H	307999
9305	0.419355	0.190302	NA	0.115167	0	[illegible]	NA	0.00698	urban	H	307999
9306	1	NA	NA	0.134362	0	0.149254	0.143655	[illegible]	urban	L	307999
9307	0.564102	NA	0.054703	0.076778	NA	0.138889	0.084997	0.032602	urban		
9308	0.584906	NA	0.145874	0.383891	0	0.198198	0.133489	0.012163	urban	H	
9309	0.611111	NA	NA	0.076778	0	0.052632	0.138357	0.003099	urban	M	
9310	0.666667	NA	0.018234	0.191945	0	NA	0.096523	0	urban	H	307999

Missing Value

Dataset Features:

HCP_ID	Unique ID for each Doctor
Specialty_ID	Doctor's specialty code {307999, 92938, 117101, 569454, 371651, 18766, 321702, 462171, 295159, 306909, 333055, 155443, 334432, 96835, 483435, 262386}
Region	Doctor office location {Urban, Rural}
Value	Value of the HCP to Mediplexis (High (H), Medium (M), Low (L), Undefined(U))
Rep_Live (RL)	Doctor's Affinity (aggregated at channel level) to each of the eight channels on a scale of 0 – 1; 1 being the least affinity and 0 being highest
Peer-to-Peer (P2P)	
Online Video (OLV)	
Rep_Remote (RR)	
Direct_Mail (DRT)	
Direct_Message (DMS)	
Online_Advertisement (OLA)	
Direct_Email (DEM)	

Supporting files – Demographic data

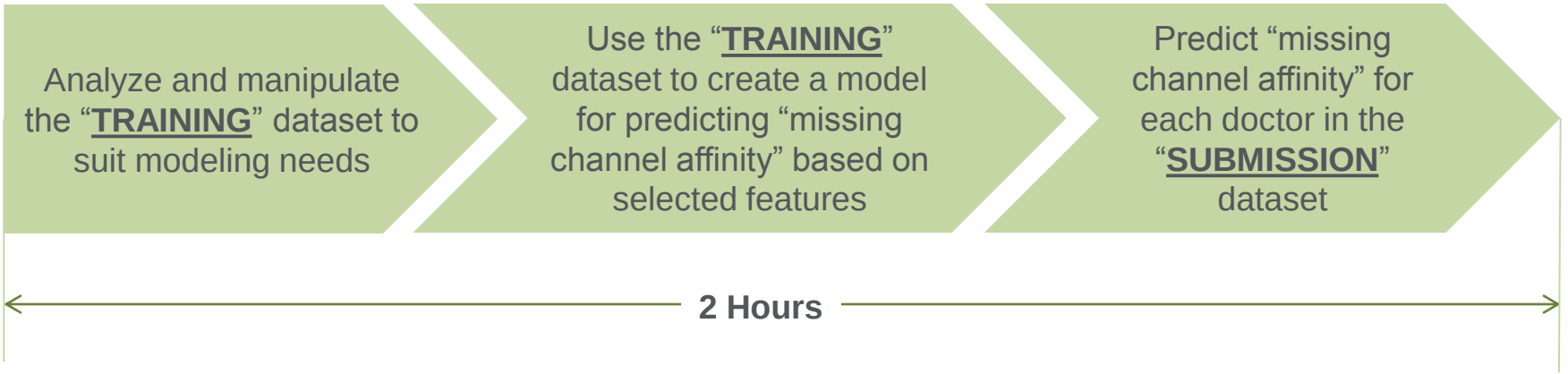
The client has also able to provide demographic information for each HCP

HCP_ID	gender	Age
14586	FEMALE	34
46157	MALE	34
16600	MALE	34
16601	MALE	33
16602	FEMALE	33
16603	MALE	32
16604	FEMALE	32
16605	FEMALE	31
16606	MALE	31
16607	MALE	31
3171	MALE	30
16608	MALE	25
16609	FEMALE	25
16610	MALE	25
11321	FEMALE	24
4908	MALE	24
11322	MALE	24
16611	FEMALE	24
16612	FEMALE	24
16613	FEMALE	24
2279	FEMALE	23
43723	FEMALE	23
14091	FEMALE	23

Demographic File (demog.csv):

- ***HCP_ID*** – Primary Key
- ***Gender –*** Physician gender
- **Age** – Physician age

You will have two hours to predict "Affinity" for missing channels in the "SUBMISSION" dataset



We will evaluate your solution using Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[(p_i - a_i)^2]}$$

Legend

p_i : *predicted value of "missing channel affinity" for each doctor*
a_i : *Actual value of "missing channel affinity" for each doctor (not provided to you)*
n : *Total number of doctors in the test dataset*

Your solution should be submitted to us in two sections - Section A (Data Analysis) & Section B (Model results)

Submitting “Section A”

In Section A, please provide your findings and justification on the below topics -

1) ***Quality checks performed / Errors found***: Please report any potential errors / inaccuracies in the training dataset

2) ***Data preprocessing steps:*** Please provide potential data aggregations / transformations performed including how the quality checks from section 1 are dealt with

3) **Key observations / Trends:** Please provide any information on your key observations / insights from the datasets

4) **Model choice explanation:** Please justify the usage of a particular model for the dataset provided

Please submit your thoughts in word/PPT/Excel on the following via email details provided -

Ensure that you follow the file naming convention (Page 7) for submission of your document

Submitting “Section B”

- Update the “SUBMISSION” dataset with your predicted “AFFINITY” for missing channel for each doctor, in the column where channel is missing (see below table for details)
- Predictions for the submission dataset are to be submitted via email – please follow file naming convention given in Page 7

“Submission.csv” File Structure

HCP_ID	RL	P2P	OLV	RR	DRT	DMS	OLA	DEM
9304	0.645833	0.475755	NA	NA	0	0.169355	NA	0.00489
9305	0.419355	0.190302	NA	0.115167	0	0.012195	NA	0.0069
9306	1	NA	NA	0.134362	0	0.149254	0.143655	0.06029
9307	0.564102	NA	0.054703	0.076778	NA	0.138889	0.084997	0.03260
9308	0.584906	NA	0.145874	0.383891	0	0.198198	0.133489	0.01216
9309	0.611111	NA	NA	0.076778	0	0.052632	0.138357	0.00309
9310	0.666667	NA	0.018234	0.191945	0	NA	0.096523	

- Update cells containing NA with your prediction
- Update file name as given in the previous slide

Ensure that you follow the file naming convention (Page 7) for submission of your document

File Naming convention for Sections A and B

File naming Instructions

- Update your response files for both sections to the following naming convention before submitting
- File naming convention: ***FirstName_LastName_ddmmyyyy***; For example, for Prakash Kumar with date of birth 10 Feb. 1987
    - Section A submission should be named Prakash_Kumar_10021987.pptx (or ppt/doc/docx/xls/xlsx)
    - Section B submission should be named Prakash_Kumar_10021987.csv
    - In case of single name use FNU in place of first name FNU_Prakash_10021987.csv

Please recheck FILE NAME before submitting!!!

Good Luck!