upGrad Campus

upGrad

Topic - Capstone Project

INSTRUCTOR:

Today’s Agenda

Briefing on the capstone project to help you in better understanding the problem statement

Problem statement

- This project is about a mobile company 'XYZ Mobiles', a fictional China-based mobile company (manufacturer and supplier). The company caters to customers across all the market segments. It has phones that target all the customer classes and is quite popular in the Chinese market. The company sees India as a key opportunity to expand its sales. It has been tracking the Indian market for more than a couple of years.

- Let's see the reasons why the company is interested in the Indian market.

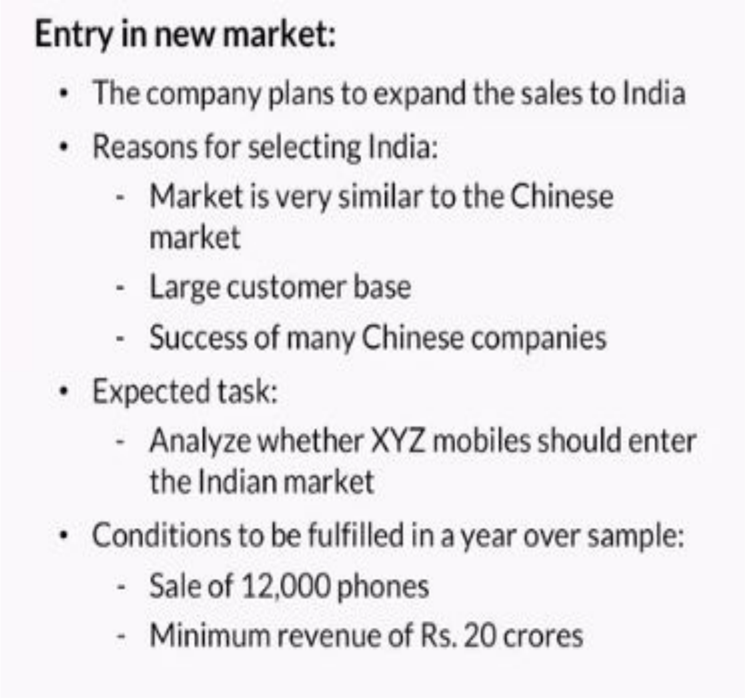
COMPANY DESCRIPTION

XYZ Mobiles:

- Mobile phone manufacturing and selling company in the Chinese market
- Provides mobile phones under different customer segments
- Wants to extend to overseas countries

Problem statement

Based on the data of both the Chinese and the Indian markets, your task is to help the company decide whether to enter the Indian market. Since you can't analyse the entire Indian and the Chinese market, the company has asked you to analyse sample data over two major cities, one from each country, to understand the sales pattern. Therefore, you are provided with the data for Shanghai and Mumbai.

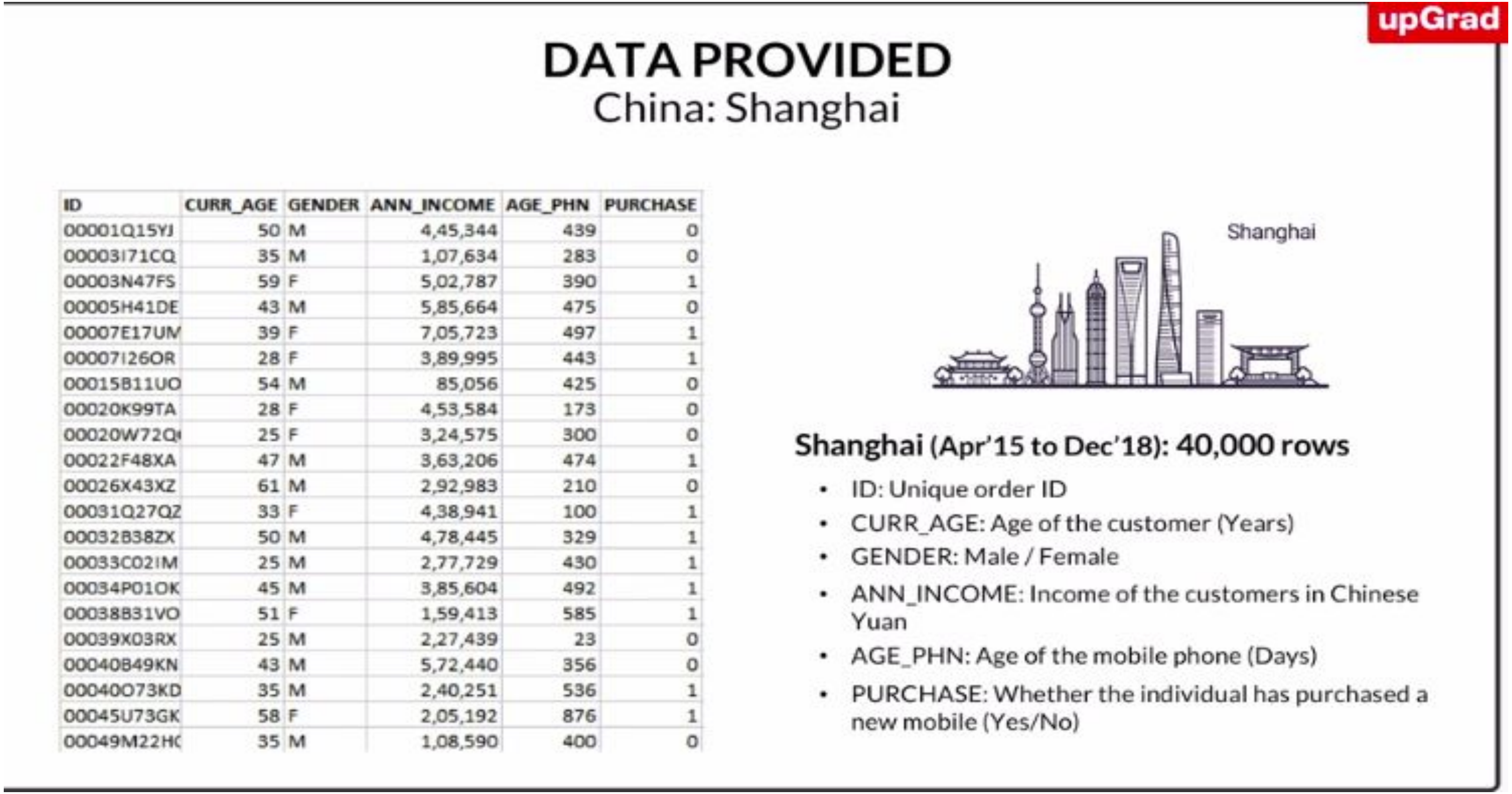


Entry in new market:

- The company plans to expand the sales to India
- Reasons for selecting India:
  - Market is very similar to the Chinese market
  - Large customer base
  - Success of many Chinese companies
- Expected task:
  - Analyze whether XYZ mobiles should enter the Indian market
- Conditions to be fulfilled in a year over sample:
  - Sale of 12,000 phones
  - Minimum revenue of Rs. 20 crores

Dataset

upGrad


upGrad
Shanghai

DATA PROVIDED

China: Shanghai

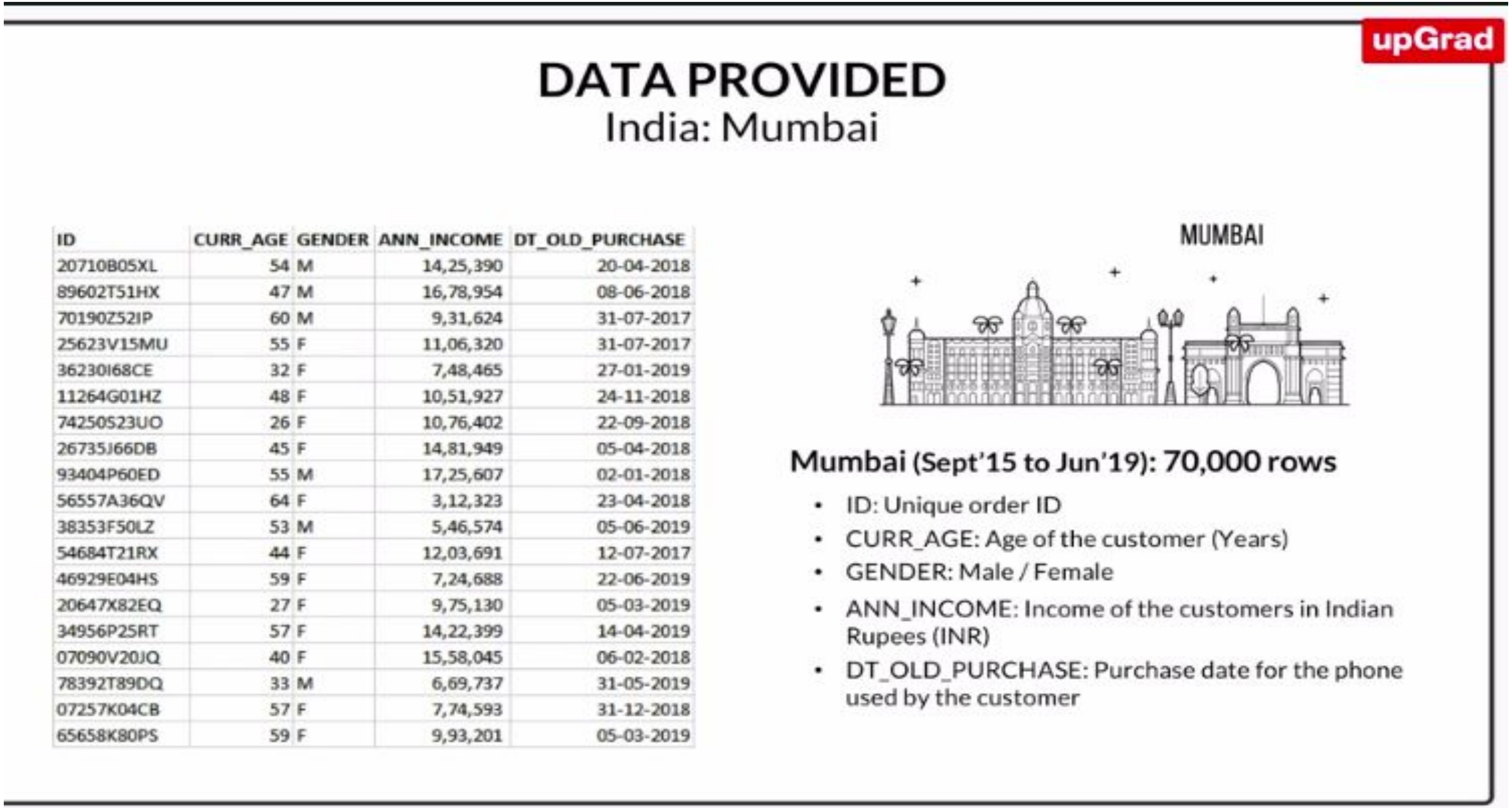
ID	CURR_AGE	GENDER	ANN_INCOME	AGE_PHN	PURCHASE
00001Q15YJ	50	M	4,45,344	439	0
00003I71CQ	35	M	1,07,634	283	0
00003N47FS	59	F	5,02,787	390	1
00005H41DE	43	M	5,85,664	475	0
00007E17UM	39	F	7,05,723	497	1
00007I26OR	28	F	3,89,995	443	1
00015B11UO	54	M	85,056	425	0
00020K99TA	28	F	4,53,584	173	0
00020W72Q	25	F	3,24,575	300	0
00022F48XA	47	M	3,63,206	474	1
00026X43XZ	61	M	2,92,983	210	0
00031Q27QZ	33	F	4,38,941	100	1
00032B38ZX	50	M	4,78,445	329	1
00033C02IM	25	M	2,77,729	430	1
00034P01OK	45	M	3,85,604	492	1
00038B31VO	51	F	1,59,413	585	1
00039X03RX	25	M	2,27,439	23	0
00040B49KN	43	M	5,72,440	356	0
00040O73KD	35	M	2,40,251	536	1
00045U73GK	58	F	2,05,192	876	1
00049M22H	35	M	1,08,590	400	0

Shanghai (Apr'15 to Dec'18): 40,000 rows

- ID: Unique order ID
- CURR_AGE: Age of the customer (Years)
- GENDER: Male / Female
- ANN_INCOME: Income of the customers in Chinese Yuan
- AGE_PHN: Age of the mobile phone (Days)
- PURCHASE: Whether the individual has purchased a new mobile (Yes/No)

Dataset

upGrad



upGrad

DATA PROVIDED

India: Mumbai

ID	CURR_AGE	GENDER	ANN_INCOME	DT_OLD_PURCHASE
20710B05XL	54	M	14,25,390	20-04-2018
89602T51HX	47	M	16,78,954	08-06-2018
70190Z52IP	60	M	9,31,624	31-07-2017
25623V15MU	55	F	11,06,320	31-07-2017
36230I68CE	32	F	7,48,465	27-01-2019
11264G01HZ	48	F	10,51,927	24-11-2018
74250S23UO	26	F	10,76,402	22-09-2018
26735J66DB	45	F	14,81,949	05-04-2018
93404P60ED	55	M	17,25,607	02-01-2018
56557A36QV	64	F	3,12,323	23-04-2018
38353F50LZ	53	M	5,46,574	05-06-2019
54684T21RX	44	F	12,03,691	12-07-2017
46929E04HS	59	F	7,24,688	22-06-2019
20647X82EQ	27	F	9,75,130	05-03-2019
34956P25RT	57	F	14,22,399	14-04-2019
07090V20JQ	40	F	15,58,045	06-02-2018
78392T89DQ	33	M	6,69,737	31-05-2019
07257K04CB	57	F	7,74,593	31-12-2018
65658K80PS	59	F	9,93,201	05-03-2019

Mumbai (Sept'15 to Jun'19): 70,000 rows

- ID: Unique order ID
- CURR_AGE: Age of the customer (Years)
- GENDER: Male / Female
- ANN_INCOME: Income of the customers in Indian Rupees (INR)
- DT_OLD_PURCHASE: Purchase date for the phone used by the customer

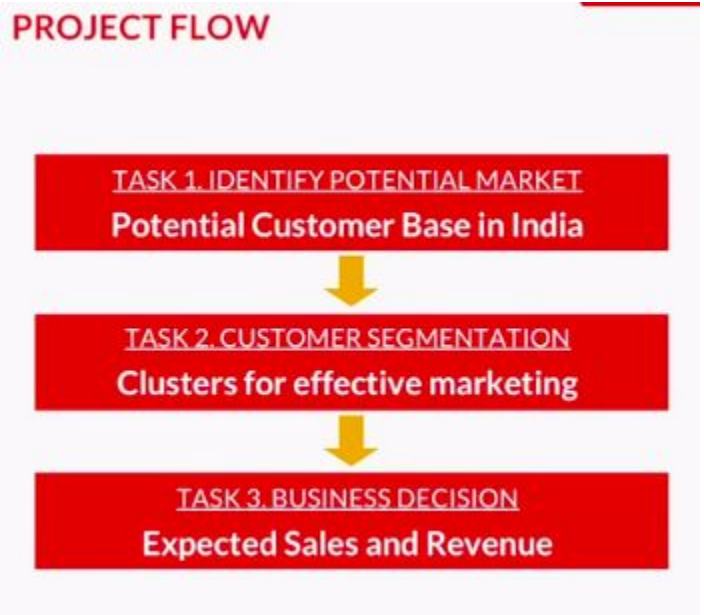
Company Expectation

The company has calculated the cost over one year after entering the Indian market as 16 crore rupees if it caters to customers in the sample city, Mumbai. So, the company has now asked you to analyse the case and decide whether they should go ahead with entering the new market.

They have provided you with the following criteria to reach the final decision:

1. Estimated revenue of 20 crore rupees from the total sales over one year in Mumbai over the sample data.
2. Minimum 12,000 units are to be sold in the sample data.
3. The above two conditions must be satisfied by the company to enter the market

Project flow



PROJECT FLOW

TASK 1. IDENTIFY POTENTIAL MARKET

Potential Customer Base in India

TASK 2. CUSTOMER SEGMENTATION

Clusters for effective marketing

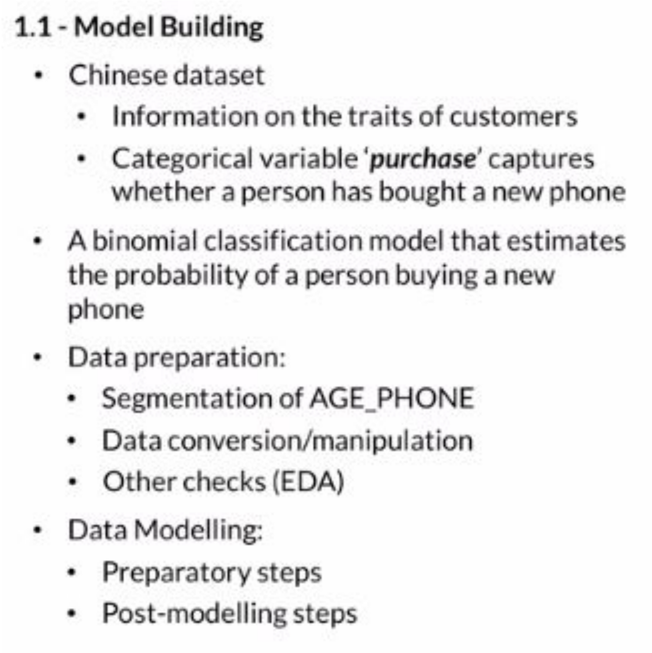
TASK 3. BUSINESS DECISION

Expected Sales and Revenue

The company would want you to first analyse the data for the customers in China. Then, based on the results and findings, you will try to predict the scenario in the Indian market.

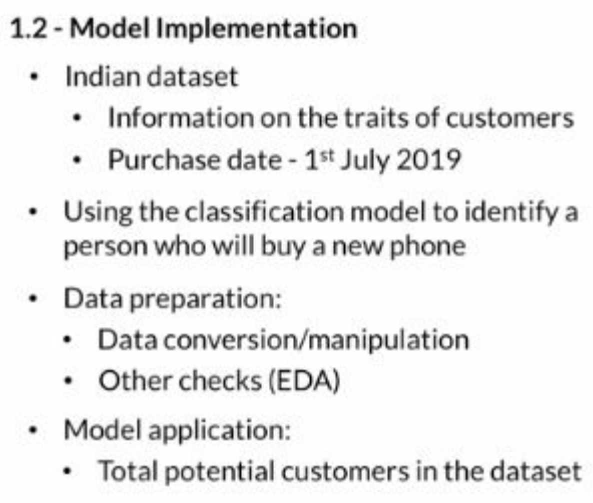
Task 1

In Task 1, XYZ Mobiles wants you to identify the individuals who would be interested in buying a new phone within a year. So you have two steps in this task that we can see image below.



1.1 - Model Building

- Chinese dataset
  - Information on the traits of customers
  - Categorical variable '***purchase***' captures whether a person has bought a new phone
- A binomial classification model that estimates the probability of a person buying a new phone
- Data preparation:
  - Segmentation of AGE_PHONE
  - Data conversion/manipulation
  - Other checks (EDA)
- Data Modelling:
  - Preparatory steps
  - Post-modelling steps

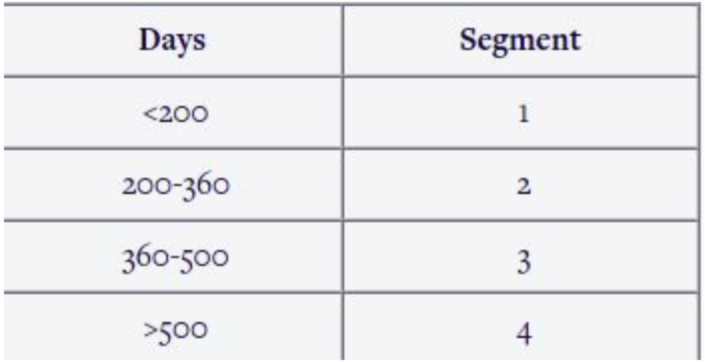


1.2 - Model Implementation

- Indian dataset
  - Information on the traits of customers
  - Purchase date - 1st July 2019
- Using the classification model to identify a person who will buy a new phone
- Data preparation:
  - Data conversion/manipulation
  - Other checks (EDA)
- Model application:
  - Total potential customers in the dataset

Step 1

Note: The company expects you to treat the variable ‘AGE_PHN’ as a categorical variable if included in the analysis. They would want you to create four segments for the variable and then begin building a model. Here is how the segments need to be divided according to the age of the phone:



Days	Segment
<200	1
200-360	2
360-500	3
>500	4

- Also, since the aim is to assess the performance over the year, it would not be very easy to analyse as the age of the phone will keep on changing every day. Therefore, to simplify the analysis, you can consider the purchase date for everyone in the Indian data set as 1st July 2019. The ‘purchase’ variable in the data set tells if an individual has purchased a new mobile phone. Your job is to build a binomial classification model using the data set that suggests a person is likely to buy a new phone based on the given attributes.

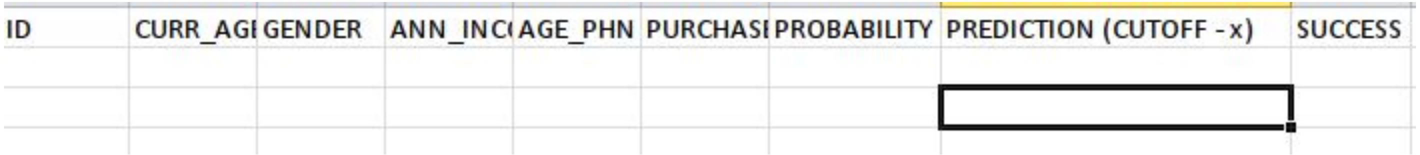
Results

Following results are expected at the end of task 1:

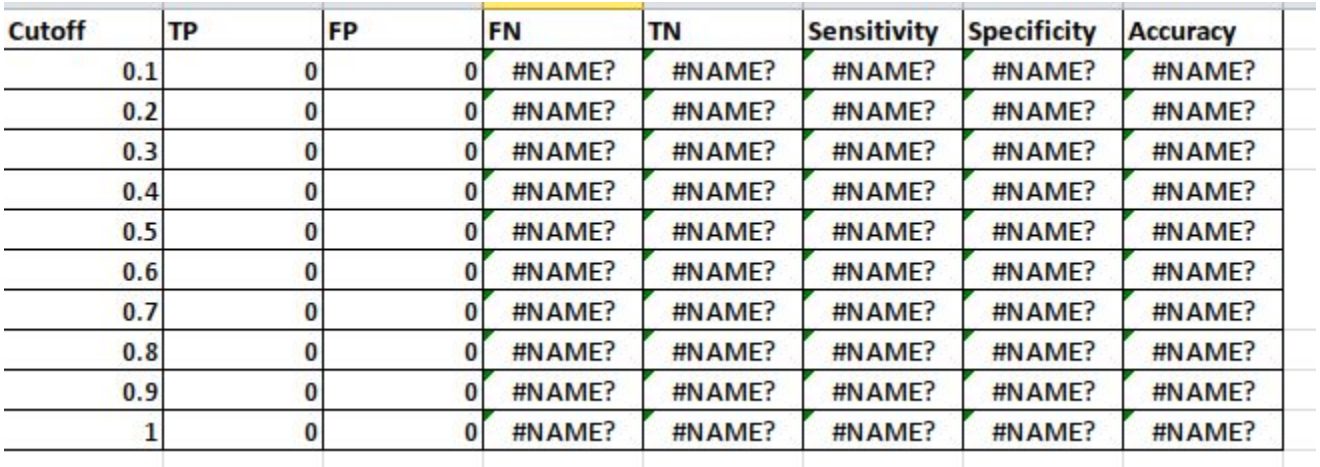
1. A classification model over the Chinese data set estimates if an individual is likely to buy a new phone based on the provided attributes.
2. Based on the learning in the module, justification should be provided for all the decisions made while building the model.
3. Business interpretation of the coefficients obtained for variables in the model
4. Metrics associated with the validation, performance, and evaluation of the model
5. Count of potential customers in the Indian market based on the model
6. Also, show some EDA(visualisation) on the column

Results

You have to create a prediction success column, in which you have to store prediction value based on cutoff value.



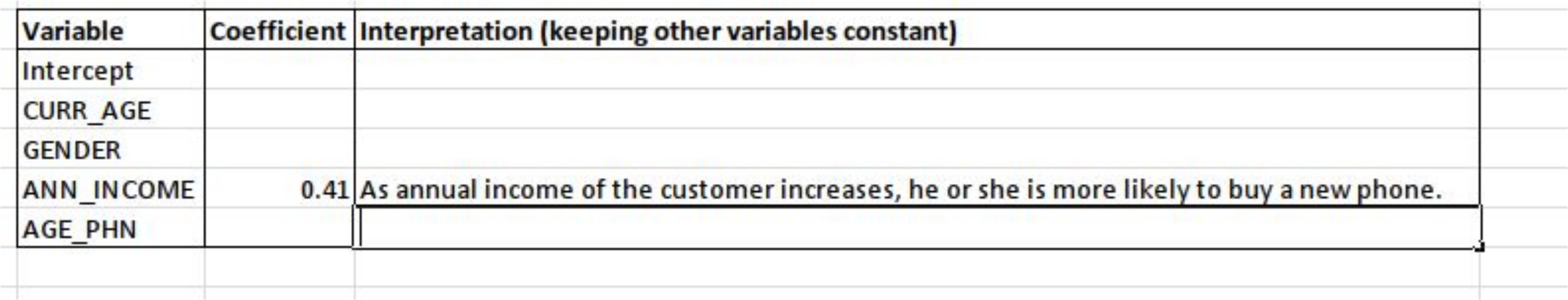
ID	CURR_AGI	GENDER	ANN_INC(	AGE_PHN	PURCHASI	PROBABILITY	PREDICTION (CUTOFF - x)	SUCCESS



Cutoff	TP	FP	FN	TN	Sensitivity	Specificity	Accuracy
0.1	0	0	#NAME?	#NAME?	#NAME?	#NAME?	#NAME?
0.2	0	0	#NAME?	#NAME?	#NAME?	#NAME?	#NAME?
0.3	0	0	#NAME?	#NAME?	#NAME?	#NAME?	#NAME?
0.4	0	0	#NAME?	#NAME?	#NAME?	#NAME?	#NAME?
0.5	0	0	#NAME?	#NAME?	#NAME?	#NAME?	#NAME?
0.6	0	0	#NAME?	#NAME?	#NAME?	#NAME?	#NAME?
0.7	0	0	#NAME?	#NAME?	#NAME?	#NAME?	#NAME?
0.8	0	0	#NAME?	#NAME?	#NAME?	#NAME?	#NAME?
0.9	0	0	#NAME?	#NAME?	#NAME?	#NAME?	#NAME?
1	0	0	#NAME?	#NAME?	#NAME?	#NAME?	#NAME?

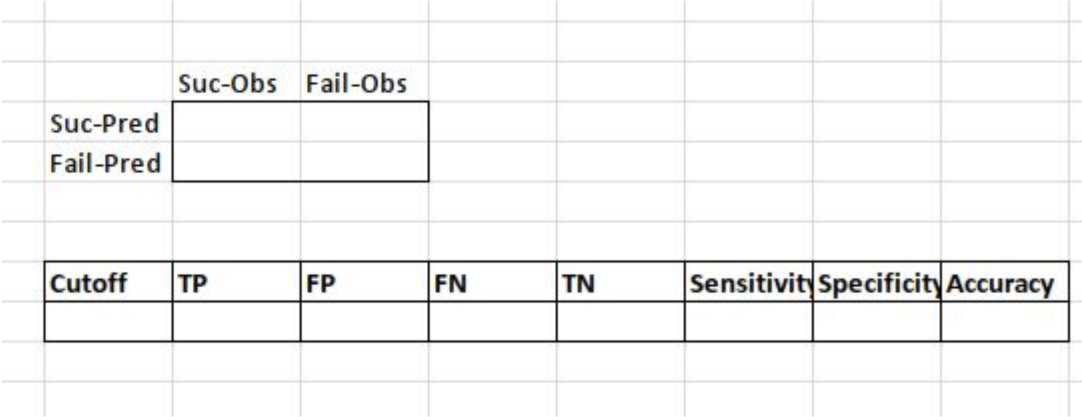
Results

Write an interpretation of the coefficient that you will get from the model.



Variable	Coefficient	Interpretation (keeping other variables constant)
Intercept		
CURR_AGE		
GENDER		
ANN_INCOME	0.41	As annual income of the customer increases, he or she is more likely to buy a new phone.
AGE_PHN		

1. Print confusion matrix from the prediction made on the test data set.
2. Also, Write the cutoff value chosen for prediction and why and what sensitivity, specificity, the accuracy?



	Suc-Obs	Fail-Obs
Suc-Pred		
Fail-Pred		

Cutoff	TP	FP	FN	TN	Sensitivity	Specificity	Accuracy

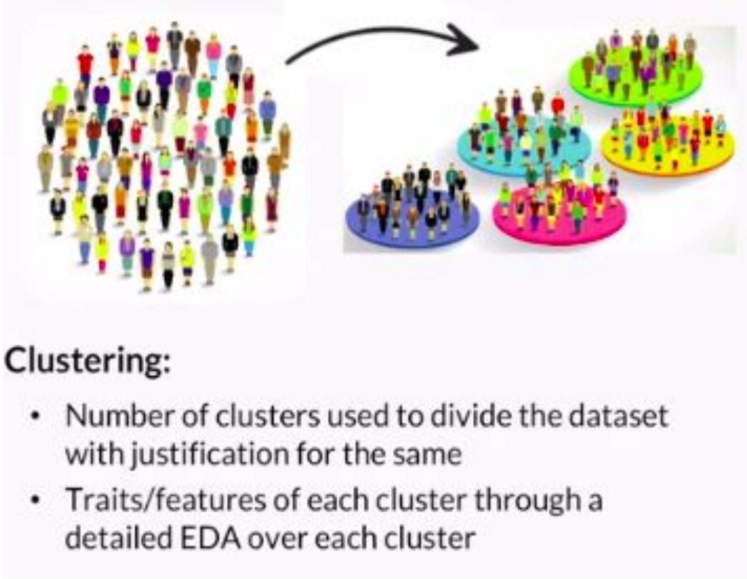
Task 2

Customer Segmentation

In task 1, the classification model gave us insights into India's customers who are likely to buy a new phone. However, the individual has the freedom to buy a phone of any brand which they feel will cater to their demands. They could go for an XYZ phone or another brand based on their utility. The company wants potential customers to prefer their phones over the other brands.

Segmentation:

- The company wants to capture the maximum market share possible in the new market
- A targeted marketing strategy for each group to lure different segments in the market
- Resulted in 40% market capture in China (expected in India too)



Clustering:

- Number of clusters used to divide the dataset with justification for the same
- Traits/features of each cluster through a detailed EDA over each cluster

Task 2 Result Expectation

In task 2, we have to cluster the potential customer base and identify their associated traits. These clusters will help the company understand the market and devise marketing techniques for each segment to attract maximum customers. The marketing team suggests that based on the similarities between the Indian and Chinese markets and the past trends, the company will capture a minimum 40% market share in each cluster after the devised strategies have been implemented. In the end, this will help in estimating the potential customers who are likely to buy a new phone manufactured by XYZ Mobiles.

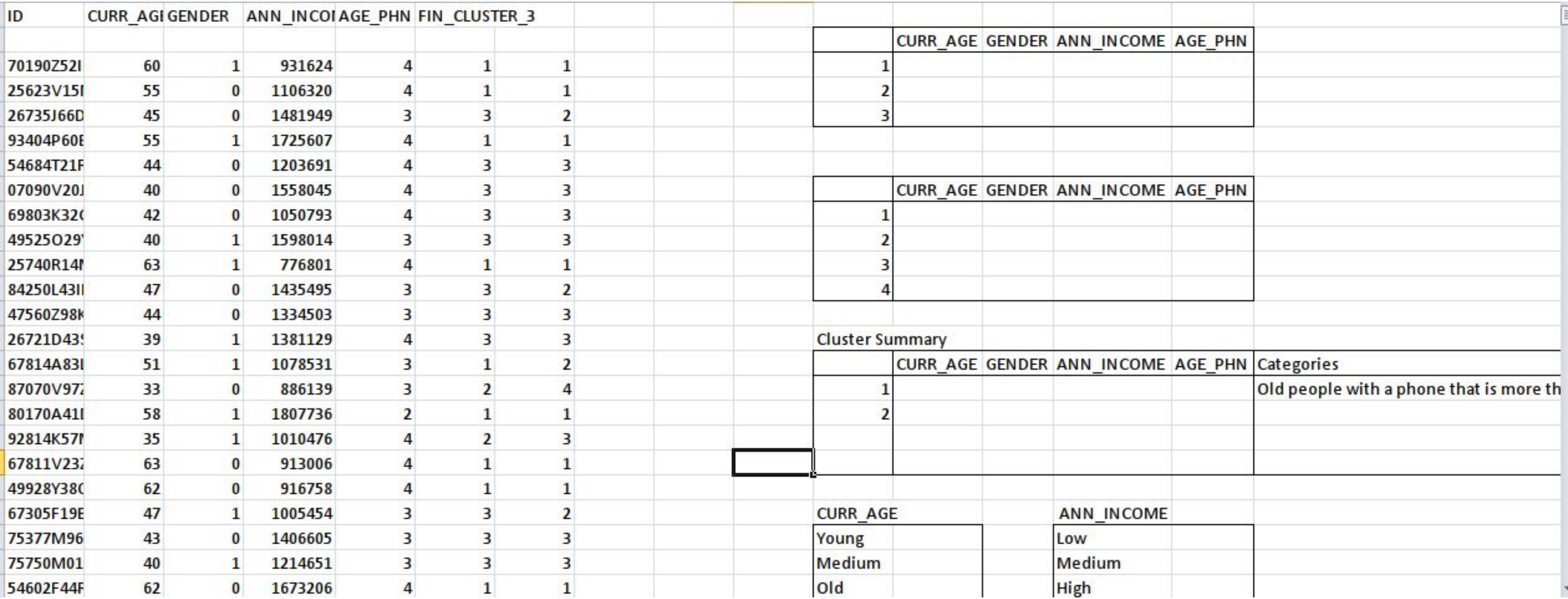
The expected deliverables in task 2 are as follows:

1. Number of clusters used to divide the dataset and justification for the same (keep in mind that the ideal case here would be to have 3 or 4 clusters, not more than that)
2. Traits/features of each cluster and
3. A detailed EDA over each cluster

Task 2 Result Expectation

upGrad

Below are some examples of EDA; you need to plot some charts using a column. For example, after finding clusters, try to plot between annual income and cluster column, similarly for other columns. it will give some insight and write that insight in the category column that you can see below



ID	CURR_AGI	GENDER	ANN_INCOI	AGE_PHN	FIN_CLUSTER_3	
70190Z52I	60	1	931624	4	1	1
25623V15I	55	0	1106320	4	1	1
26735J66D	45	0	1481949	3	3	2
93404P60E	55	1	1725607	4	1	1
54684T21F	44	0	1203691	4	3	3
07090V20J	40	0	1558045	4	3	3
69803K32(	42	0	1050793	4	3	3
49525O29'	40	1	1598014	3	3	3
25740R14I	63	1	776801	4	1	1
84250L43II	47	0	1435495	3	3	2
47560Z98K	44	0	1334503	3	3	3
26721D43!	39	1	1381129	4	3	3
67814A83I	51	1	1078531	3	1	2
87070V97Z	33	0	886139	3	2	4
80170A41I	58	1	1807736	2	1	1
92814K57I	35	1	1010476	4	2	3
67811V23Z	63	0	913006	4	1	1
49928Y38(	62	0	916758	4	1	1
67305F19E	47	1	1005454	3	3	2
75377M96	43	0	1406605	3	3	3
75750M01	40	1	1214651	3	3	3
54602F44F	62	0	1673206	4	1	1

	CURR_AGE	GENDER	ANN_INCOME	AGE_PHN
1				
2				
3				

	CURR_AGE	GENDER	ANN_INCOME	AGE_PHN
1				
2				
3				
4				

Cluster Summary

	CURR_AGE	GENDER	ANN_INCOME	AGE_PHN	Categories
1					Old people with a phone that is more th
2					

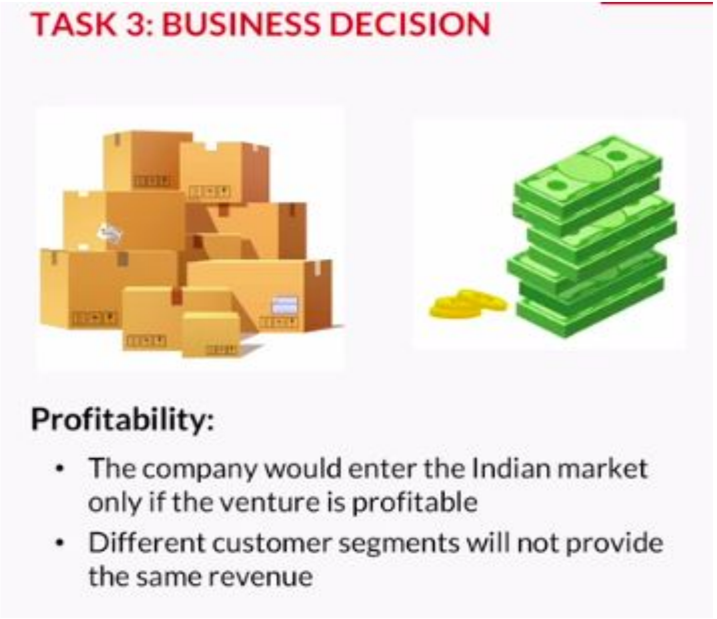
CURR_AGE	ANN_INCOME
Young	Low
Medium	Medium
Old	High

Task 3

upGrad

- After task 2, we segmented the potential customers for a new mobile phone into clusters. The purpose of clustering was to distribute the potential customers into buckets and generate an effective marketing strategy that would help to maximise the market capture. This strategy may help XYZ Mobiles to capture 40% market share in India.

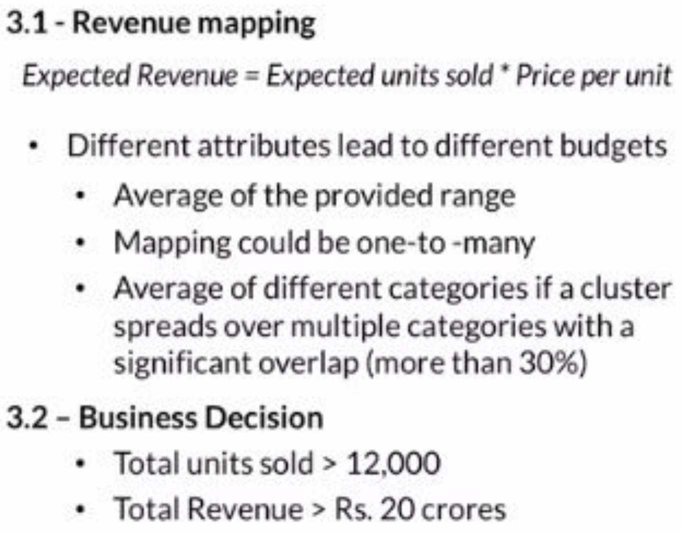
- The next task is to calculate the revenue for each cluster.



TASK 3: BUSINESS DECISION

Profitability:

- The company would enter the Indian market only if the venture is profitable
- Different customer segments will not provide the same revenue



3.1 - Revenue mapping

*Expected Revenue = Expected units sold * Price per unit*

- Different attributes lead to different budgets
  - Average of the provided range
  - Mapping could be one-to -many
  - Average of different categories if a cluster spreads over multiple categories with a significant overlap (more than 30%)

3.2 – Business Decision

- Total units sold > 12,000
- Total Revenue > Rs. 20 crores

Task 3 Result expectation

Each cluster has customers with different traits, and hence, you cannot expect all of them to spend the same amount on a mobile phone. For example, a person with a high income would be more likely to purchase an expensive phone than someone who is earning less. This variation could also come based on the customer's last phone's age, gender, or age. To understand the traits of each cluster, perform EDA and then map the cluster to its spending pattern. This will help estimate the total revenue expected over the sample in the Indian market.

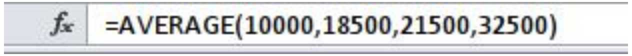
Take average over the range since there is a price range associated with the different categories. Also, if a cluster has over multiple categories (a significant overlap (more than 30%)), you have to take the average values under the two categories. This must be stated in your submission (images are provided on the platform for assumption regarding making a business decision).

The expected deliverables in task 3 are as follows:

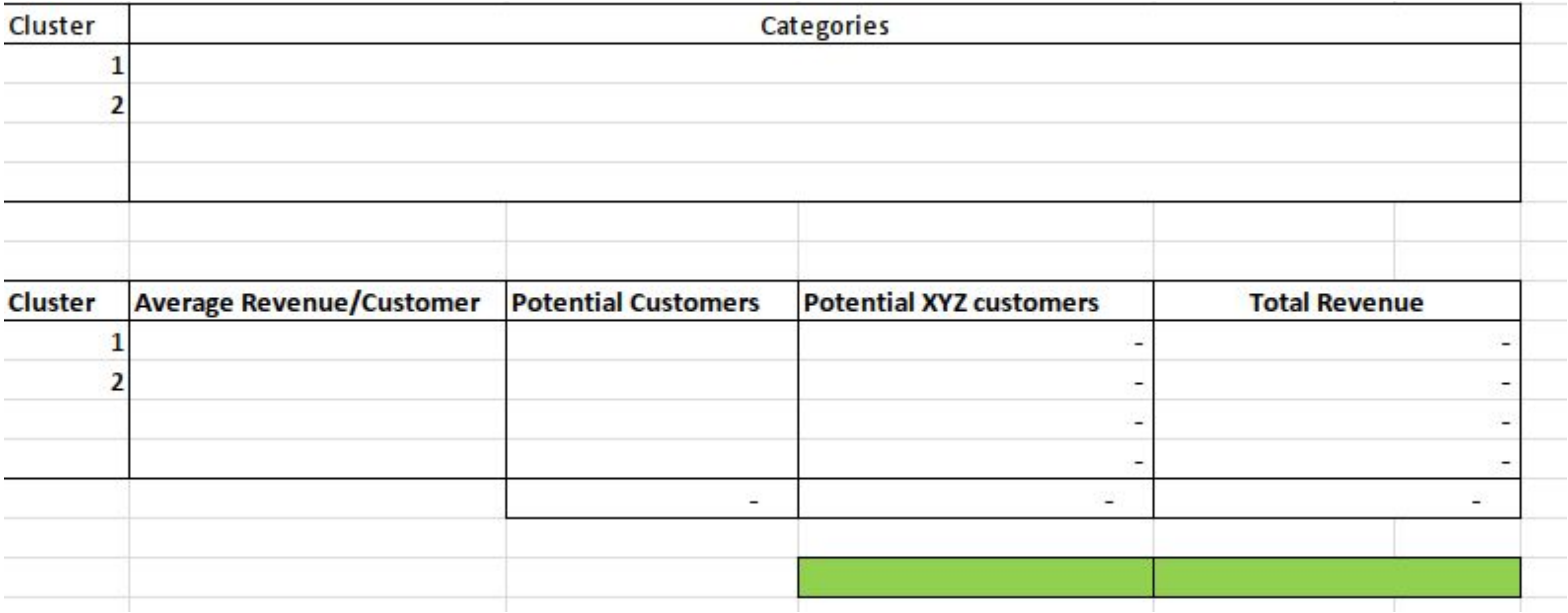
1. The expected revenue that the company can generate under each cluster
2. The final decision whether the company should enter the Indian market with business justification

Task 3 Result expectation

You have to find these columns values for getting your decision.



=AVERAGE(10000,18500,21500,32500)



Cluster	Categories
1	
2	

Cluster	Average Revenue/Customer	Potential Customers	Potential XYZ customers	Total Revenue
1			-	-
2			-	-
			-	-
			-	-
		-	-	-
			🟩	🟩

Any Questions!

upGrad
#LifeKoKaroLift

Thank You!