Bank Marketing Dataset Analysis

- of a portugal bank

DATA EXPLORERS

THE STAGES OF RESEARCH

Bank Marketing Dataset

Problem Statement

Analyse the problem

Fundamental Analysis

Review related literature & EDA

Pre-processing & Approach

Divide & conquer the problem

Modeling & Optimization

Use machine learning algorithms & optimize it

Conclusion

Lessons learnt - what worked & what didn't

PROBLEM STATEMENT

Over three years, (2008- 2010) a Portugal bank has reached out to 41,000+ customers to open a Term deposit account

Bank has enjoyed mixed-success with failure and success outcomes.

Help the bank in understanding the tele-marketing campaign outcome

Bank can use this to better profile potential customers

FUNDAMENTAL ANALYSIS

2008 - 2010 - A brief economic history of Portugal

2008

Portugal officially declared a recession in Jan'08

2009

Two banks went bankrupt.

Economic growth rate is -ve

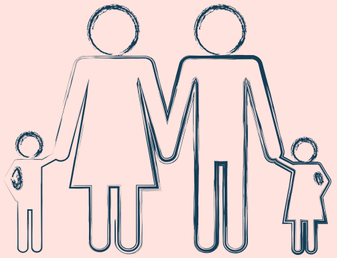
2010

Bank bailout happened.
Economic growth numbers +ve

FUNDAMENTAL ANALYSIS

a socio-economic profile

Portugual is a romantic nation



Fertility rate: 1.5

Bargain a lot

They don't go into a restaurant without checking the menu price first.

They wait for a better deal always. Portuguese don't hurry.

Average size of households - 3 people

Developed nation

Advanced Economy.

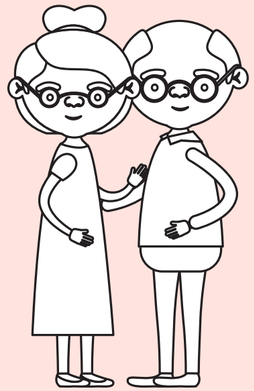
Taxes range from 14.5% to 44%

Portuguese live longer

Life expectancy ratio for male is upto 78 years.

For females it is 82 years.

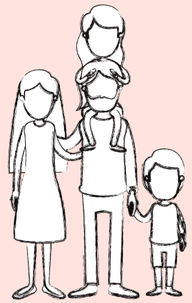
Average age of the population is 42 years



Use cash / debit cards.

They don't go for loans unnecessarily. Use of credit cards is mostly a No

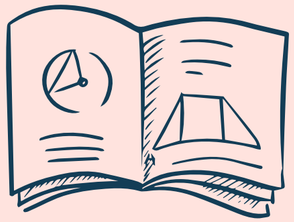
Kids live with their parents till 24 years



Bonus paid - 2 times a year. June and Dec

It's compulsory

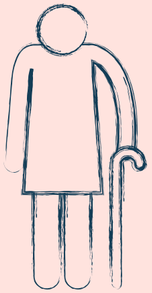
Get married at 30 years on an average.



Education till schooling is free & compulsory.
Literacy rate 99.4%

Minumum Wage is 683 euros/mo

Retirement age is 65



DATASET COLUMNS

- Age
- Job
- Marital
- Education
- Contact mode, month, day
- duration of the call
- campaign - no of calls present campaign
- p-days no of days previously called
- defaulted or not
- housing, personal loan
- duration of call
- previous campaigns and previous outcome

Consumer Demographics

- consumer price index
- emp variation rate
- consumer confidence index
- euribor 3m
- number employed
- y - dependent variable

Socio Economic Factors

DOMAIN KNOWLEDGE

Term deposit or Time deposit

Term Deposit generate much higher interest rate than a savings deposit

Money is locked for a certain amount of time

If money is withdrawn before maturity, significant penalty will be levied

Usually, people having idle money laying around will go for term deposit

Euribor Rate

Euribor rate - base rate at which one bank lends to another bank

Interest rate is dependent on euribor rate. More euribor, more interest

Consumer confidence index

Consumer confidence on the economy. More confidence, better the economy.
More confidence, more spending of money

Consumer price index

Consumer price denote inflation. More inflation, less money to save

Emp variation rate

Denotes employment rate. If positive, people will be employed more.
Therefore, spend more.

EDA - INSIGHTS

Year included based on Euribor rates, price index and confidence index

2008 - year of recession
Unemployment rate was around 40% for young people

Lot of calls made - only 5% said yes

By Jun'09, less calls went and almost 30% said yes

By Jan'10, less calls went and almost 50% said yes

Count of y	Y2008		Y2008 Total	Y2009		Y2009 Total	Y2010		Y2010 Total	Grand Total
	no	yes		no	yes		no	yes		
mar				156	126	282	114	150	264	546
apr				2016	442	2458	77	97	174	2632
may	7523	240	7763	5270	524	5794	90	122	212	13769
jun	4186	188	4374	451	264	715	122	107	229	5318
jul	6278	407	6685	112	66	178	135	176	311	7174
aug	4904	271	5175	506	264	770	113	120	233	6178
sep				161	106	267	153	150	303	570
oct	25	42	67	267	180	447	111	93	204	718
nov	3426	190	3616	189	168	357	70	58	128	4101
dec	9	1	10	84	88	172				182
Grand Total	26351	1339	27690	9212	2228	11440	985	1073	2058	41188

No of consumer saying 'no' and 'yes' by year

EDA - INSIGHTS

2009 - Banks were offering - max 2% interest

2010 - year of recession Banks were even lesser interest

min: 0.6%; max: 1.5%

2008 - year of recession Banks were offering higher rate of interest - atleast 3%

Euribor rates	Y2008		Y2008 Total	Y2009		Y2009 Total	Y2010		Y2010 Total	Grand Total
	no	yes		no	yes		no	yes		
0.5-1				1235	814	2049	877	964	1841	3890
1-1.5				7821	1288	9109	108	109	217	9326
1.5-2				156	126	282				282
3-3.5	4	1	5							5
3.5-4	9		9							9
4-4.5	3408	181	3589							3589
4.5-5	22922	1149	24071							24071
5-5.5	8	8	16							16
Grand Total	26351	1339	27690	9212	2228	11440	985	1073	2058	41188

Euribor rates	Y2008		Y2008 Total	Y2009		Y2009 Total	Y2010		Y2010 Total	Grand Total
	no	yes		no	yes		no	yes		
0.625-0.65							136	149	285	285
0.65-0.675							73	114	187	187
0.675-0.7							64	103	167	167
0.7-0.725				318	295	613	57	50	107	720
0.725-0.75				219	138	357	47	28	75	432
0.75-0.775				70	49	119	26	32	58	177
0.775-0.8				38	36	74	23	20	43	117

1

1

Even with interest lower than 0.7%, more than 50% people called opted for a TD account in 2010

No of consumer saying 'no' and 'yes' by year & interest rates

EDA - INSIGHTS

Reason for low interest rates by the bank in 2009 and 2010:

During the course of recession, government wants people to spend the money or invest in alternative sources like shares that would be beneficial for the country

To discourage people from holding money in bank accounts, less interest rate is issued

This is why the bank has reduced the number of calls made from mid - 2009 / end - 2010

Reason for less people opting for Term Deposit in 2008

Country is in recession; unemployment rate shot up to 18% on an average.
They don't want money locked up in the bank.

Reason for people holding accounts irrespective of low interest rates in 2009 & 2010:

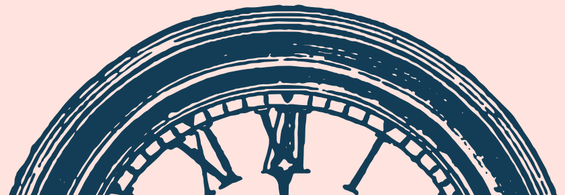
They must be having some idle money lying around.

They are not financially educated to invest in alternative sources / don't have the budget / looking for safe investments

Reason for high interest rates by the bank: in 2008

Banks were in need of money as the cash flow to the bank had reduced due to recession.

They frantically reached a lot of people to get deposits in the bank.



EDA - INSIGHTS

Emp rate vs year

When emp rates are positive in 2008, a lot of calls were made by the bank yet only 5% of the consumers opted for a TD

Negative rates since late 2008's.

2009: -1.8 to -3.4
2010: -1.1 to -1.8

Emp var rates	Y2008		Y2008 Total	Y2009		Y2009 Total	Y2010		Y2010 Total	Grand Total
	no	yes		no	yes		no	yes		
-3.4				617	454	1071				1071
-3				84	88	172				172
-2.9				1069	594	1663				1663
-1.8				7442	1092	8534	281	369	650	9184
-1.7							370	403	773	773
-1.1							334	301	635	635
-0.2	9	1	10							10
-0.1	3451	232	3683							3683
1.1	7523	240	7763							7763
1.4	15368	866	16234							16234
Grand Total	26351	1339	27690	9212	2228	11440	985	1073	2058	41188

2

1

Positive rates in may, jun, jul 2008.
2008: 1.4 to -0.2

Rates were more negative in the 2009's

20% people who opted for a TD might had a secure employment

Rates improved in 2010. Yet less calls were made by the bank

PRE-PROCESSING & APPROACH

Columns that needed pre-processing

RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):

#	Column	Not Null	Dtype	Null values	Imputed with
0	age	41188 non-null	int64		
1	job	41188 non-null	object	330	Mode
2	marital	41188 non-null	object	80	Mode
3	education	41188 non-null	object	1731	Mode
4	default	41188 non-null	object	8597	Mode
5	housing	41188 non-null	object	990	Mode
6	loan	41188 non-null	object	990	Mode
7	contact	41188 non-null	object		
8	month	41188 non-null	object		
9	day_of_week	41188 non-null	object		
10	duration	41188 non-null	int64		removed
11	campaign	41188 non-null	int64		
12	pdays	41188 non-null	int64	39673	Mode
13	previous	41188 non-null	int64		
14	poutcome	41188 non-null	object		
15	emp.var.rate	41188 non-null	float64		
16	cons.price.idx	41188 non-null	float64		
17	cons.conf.idx	41188 non-null	float64		
18	euribor3m	41188 non-null	float64		
19	nr.employed	41188 non-null	int64		
20	y	41188 non-null	object		

dtypes: float64(4), int64(6), object(11)

Categorical columns - imputed with mode

Duration column is removed as it is highly correlated with the dependent variable

Numerical columns imputed with Mode for pdays - number of days before a consumer was contacted

SMOTE also used

APPROACH

Activites and schedule

Tasks	Week 1	Week 2	Week 3
Fundamental analysis, Pre-processing			
Run baseline accuracy tests with LR			
Feature Engineering & Modeling			
Model Optimization			
Presenting findings			

MODELING & OPTIMIZATION

Building the dataset for ML algorithms

- Ran logistic regression with plain vanilla dataset

  Check baseline accuracy; Around 89.5%

- Feature Engineering

  - Brought external data like year, interest rates, unemployment rates into the dataset
  - Enhanced internal data - complex numerical columns with multiple values like age, euribor binned
  - Removed highly co-related columns

- Modeling

  Different algorithms used
  - Logistic Regression
  - Decision Trees & Random Forest
  - Boosting algorithms - Adaboost, Gradient, XGBoost, LGBM

OPTIMIZATION

Optimization techniques used

Feature Engineering

Removed features using random forest feature importance recommendations

Broke down single features into multiple features

e.g. euribor 3m was broadly classfied into less than 1, greater than 1 etc. broke into small chunks as they had better pattern.

Grid Search

Grid search was used for all the algorithms.

Best Parameters were found and then re-learnt the model

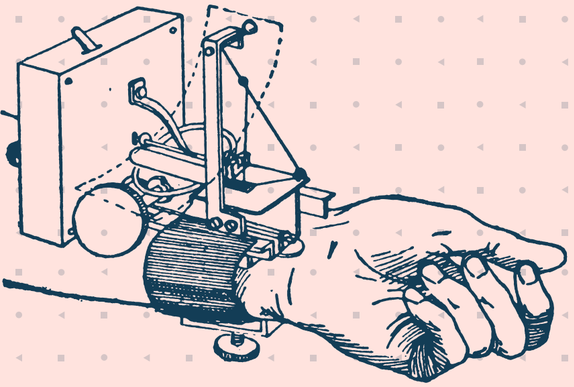
Parameter configuration was mostly trial and error

Voting classifiers

Voting classifiers used with Random forest, LGBM and XGBoost.

Voting classifiers needed tuning. Assigned weights using voting='soft' and 'weights' parameter

MODELING & OPTIMIZATION



test train split: 0.6

SMOTE not used;
featured variables were
categorical variables

Algorithm	Accuracy	Confusion Matrix
Logistic Regression (LR)	0.893724205	[[14402 244] [1507 323]]
Decision Tree	0.850691916	[[13438 1208] [1252 578]]
DT with Hyper Parameter Tuning	0.90064336	[[14391 255] [1382 448]]
Random Forest (RF)	0.889536295	[[14170 476] [1344 486]]
RF with Hyper Parameter Tuning	0.901978636	[[14454 192] [1423 407]]
Ada Boost (AB)	0.888929352	[[14645 1] [1829 1]]
AB with Hyper Parameter Tuning	0.895362952	[[14501 145] [1579 251]]
Gradient Boost (GB)	0.899065307	[[14456 190] [1473 357]]
GB with Hyper Parameter Tuning	0.90106822	[[14459 187] [1443 387]]
LGBM	0.900764749	[[14370 276] [1359 471]]
LGBM with Hyper Parameter Tuning	0.900886137	[[14383 263] [1370 460]]
XG Boost	0.896576839	[[14288 358] [1346 484]]
XG Boost with Hyper Parameter Tuning	0.900825443	[14427 219] [1415 415]

Random forest fared better

adaboost is crazy; TP is 1

Voting classifiers yielded
less accuracy

CONCLUSION

What we learned!

Spend much time analyze the data from all angles. It will reward in the long run. EDA is the key.

Applying an algorithm without understanding how it works may cause frustration when accuracy doesn't improve

Accuracy is not everything. Co-relate with confusion matrix

Check the model with different splits. Accuracy has to be consistent for a robust model

If you have engineered a lot of categorical features, using SMOTE will reduce your accuracy

Don't find patterns in the data without feature engineering. It might be a waste of time.

Feature engineering is all about creativity.

When coming to finding patterns in a dataset, use excel pivot tables

Use git for version control - using branches + multiple files

DATA EXPLORERS

THANK YOU

Meet our team

srikanth (c)

kannan (vc)

hafiz

vimal

jeeva

venky

solomon

raja

ram

antony