Maximizing Bank's Marketing campaign profitability through Machine Learning

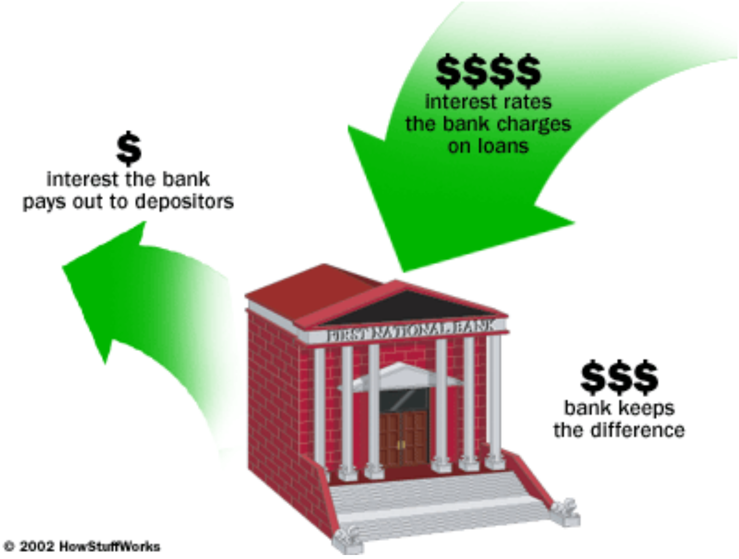
1. ABSTRACT

In this analysis, we examine call data from the marketing team selling term deposits to consumers for a Portuguese Bank. We then build a predictive model to determine which calls are most likely to result in a term deposit being taken, and ultimately provide a cost-benefit analysis to choose a machine learning algorithm that will maximize campaign profitability.

2. BACKGROUND

Banks make money by selling and borrowing money. They sell money in the form of a variety of loans and borrow money from investors and customers by providing term deposits and other financial products.

The interest rates that banks take for loans is usually much higher than the interest rate provided to customers while borrowing money from them. This is where and how profits are generated for banks and this difference in amount is called the Net interest Income.



Banks have marketing teams that are responsible for selling loans and deposits. This team makes use of multiple online and direct marketing channels to contact existing account holders and sell these products to them.

Campaign managers decide on which channel is best for the kind of products they have and it's their responsibility to maximize customer subscription with minimum budgets.

For this case study, our client is a Portuguese retail bank utilizing a direct marketing technique widely known as Telemarketing, in which employees made calls to customers directly to sell their long-term deposits plan.

Long term deposits are a medium through which banks borrow money from customers and in return provide them with interest earnings. These have a fixed term anywhere from 1 year to 5 years during which they will not be able to cash the amount. These are extremely safe investments and are therefore very appealing to conservative, low-risk investors

By making use of business data related to such activity, we will use Machine Learning to attempt to predict if a client with subscribe to a term deposit or not. We will then optimize this model to maximize profitability via a cost-benefit analysis.

This will help the Portuguese Campaign Manager to weed out non-prospective customers in initial stages and direct resources/efforts to only those customers that are most likely to subscribe to a product - thereby helping them achieve maximum ROI with minimum expenditure.

3. DATA SOURCES

The data related with direct marketing campaigns of Portuguese banking institution is made publicly available on UCI website[1] in CSV format.
In addition to this, information from **external sources** have been used in order to analyze the data and arrive at best suitable model for this business scenario.

Comment [1]: If this is true, there must be citations elsewhere in the report. Otherwise it should be removed.

4. DATA AT A GLANCE

Data collected is from May 2008 to Nov 2010 with 21 attributes and 41,188 observations.
Attributes are broadly classified into 4 heads;

- Client data,
- data related to last contact of the current Campaign,
- Social-Economic and
- other attributes as seen in the below table.

Out of the 21 attributes, 20 are independent variables (X) and one is a dependent variable (Y).

S.NO	ATTRIBUTE	DESCRIPTION	TYPE
1	Age	-	Numeric
2	Job	Type of job	Categorical
3	Marital	Marital status	Categorical
4	Education	-	Categorical
5	Default	Has credit in default?	Categorical
6	Housing	Has housing loan?	Categorical
7	Loan	Has personal Loan?	Categorical
8	Contact	Contact communication type	Categorical
9	Month:	Last contact month of year	Categorical
10	Day of week	Last contact day of the week	Categorical
11	Duration*	Last contact duration, in seconds	Numeric
12	Campaign	Number of contacts performed during this campaign and for this client includes last contact	Numeric
13	pdays	Number of days that passed by after the client was last contacted from a previous campaign	Numeric
14	previous	Number of contacts performed before this campaign and for this client	Numeric
15	poutcome	Outcome of the previous marketing campaign	Categorical

16	**emp.var.rate**	**Employment variation rate: quarterly indicator**	**Numeric**
17	**cons.price.idx**	**Consumer price index - monthly indicator**	**Numeric**
18	**cons.conf.idx**	**Consumer confidence index - monthly indicator**	**Numeric**
19	**euribor3m**	**Euribor 3 month rate - daily indicator**	**Numeric**
20	**nr.employed**	**Number of employees - quarterly indicator**	**Numeric**
21	**y**	**Has the client subscribed a term deposit?**	**Categorical**

5. PREDICTION GOAL

The end goal is to make predictions about existing customers and whether they will subscribe to a long-term deposit scheme or not based on the attributes. We have used Logistic regression, Support Vector Machines and Random Forests to classify customers into prospective and non-prospective categories. New campaign Profitability has been considered as the metric for evaluating these models and the one with highest profitability has been shortlisted as the final model.

6. LIMITATIONS

- Interest rates, principal amounts and duration of the term deposit are not available. Calculated assumptions have been made to arrive at Net interest income.
- Client demographic data is limited. Key client data such as salary range, bank balances that are strong indicators of a customer's financial strength are unknown. These in conjunction with education and job affect standard of living/savings ability and hence could have help in predicting the outcome more accurately.
- Marketing spend data is unavailable. As such we will be making educated assumptions in order to determine marketing ROI.

7. DATA WRANGLING

- While checking for missing values it was found that 5 of the 21 variables had them at varying proportions

VARIABLE	% MISSING VALUES
job	0.801204
marital	0.194231
education	4.20268
default	20.872584
housing	2.403613
loan	2.403613
ALL OTHER VARIABLES	0

All variables including those with missing values have been investigated during EDA phase and dealt with accordingly.

- The variable 'default' is grouped as follows:

default	no	yes
outcome		
0	26626	3
1	3859	NaN

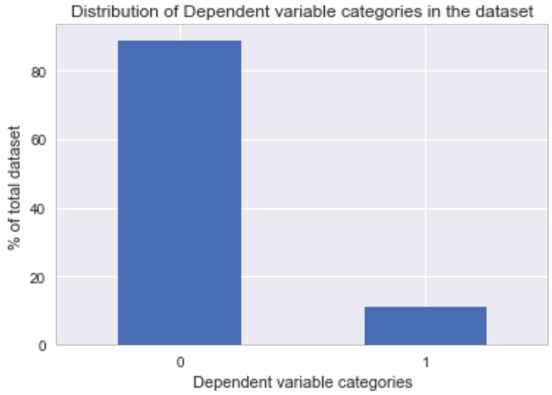
According to this, only 3 customers with credit defaults have been contacted and the remaining contacted were those with either no credit amounts or with cleared monthly payments. This attribute will therefore be of no use for predicting the outcome and hence has been removed.

- The variable 'duration' highly affects the output target; i.e. if a call has not been made at all or not been picked by the customer it results in a failed outcome. Yet, the duration is not known before a call is performed. Also, after the end of the call outcome is obviously known (success or a failure). Hence this variable has been removed from future analysis.

8. *EXPLORATORY DATA ANALYSIS*

8.1 CONVERSION RATE:

Out of all the customers contacted, 11.2% of them have accepted the offer and the remaining 88.8% of them have rejected it. In term of resource utilization we could view this as a non-productive marketing effort towards 88.8% of contacted customers who did not take up the offer.
The intention of this study is to predict the most probable customers much before making calls and help campaign managers channel employee efforts only towards them.

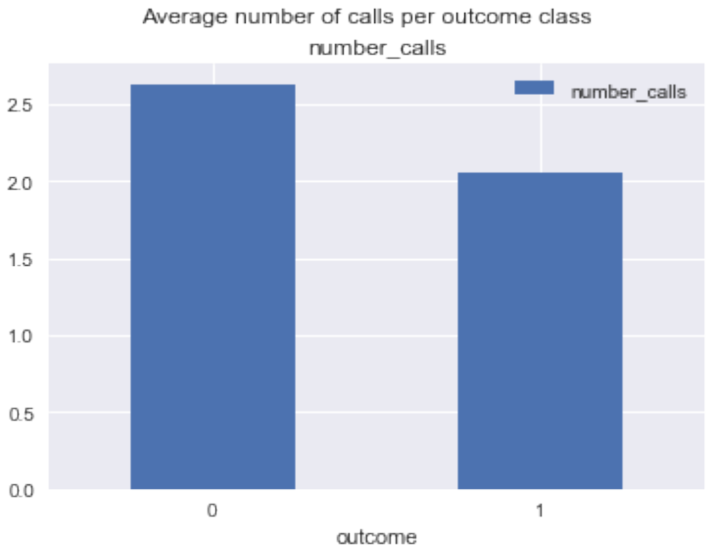


From a statistical perspective, this dataset is highly imbalanced. Unfavourable outcomes make up to about 88.8% of the total observations while the favourable ones make up the remaining 11.2% making the data vulnerable to prediction biases while trying to fit machine-learning models. We will be dealing with this issue after performing EDA in order to understand the dynamics of unmanipulated data better.

Comment [2]: Addition

8.2 NUMBER_CALLS

Customers on an average made their decision by 3rd call with a bank representative. Those who agreed to the offer made their decision within the 2nd call.



For the purpose of modelling, we have not taken into account the call distribution and have assumed that customers decide to accept or deny by the 3rd call. The overall average calls aspect of this attribute has been used in cost-benefit analysis to calculate the cost for calling a customer during the campaign and individual outcome wise data has been ignored.

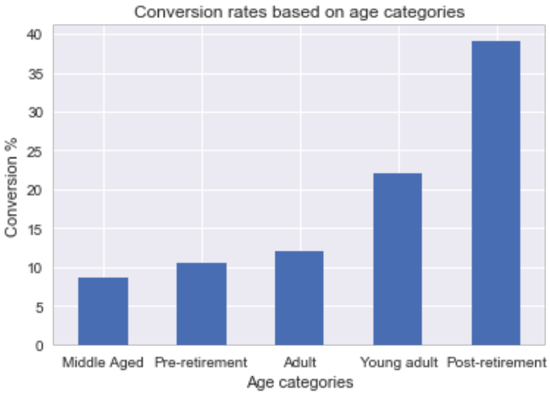
8.3 Age

Investment decisions can highly vary based on the age of a person. Research conducted by Dale Kintzel (Portfolio Theory and life-cycle investment)[1] confirms that at a young age, individuals tend to be more aggressive with respect to returns and hence choose to invest a major portion of their saving into stocks. As they age, they tend to move towards investments that have lower risk and can give them guaranteed returns such as government Bonds and Term-deposits provided by banks.

It would be interesting to get a picture of the age groups of customers contacted and how the outcome is spread across them. Customers between 17yrs to 98yrs, with an average age of 40yrs were contacted.

For the purpose of analysis following age groups have been created

- Young adult – less than 25 yrs
- Adult - 25 yrs to 35 yrs
- Middle aged – 36 yrs to 54 yrs
- Pre-retirement - 55 yrs to 60 yrs
- Old – 60 yrs and above



A huge number of customers, approx. 45% aged over 60 have reacted positively to the offer. This is in line with the observations made by Dale Kintzel making this category very promising while building our predictive model. However, Young adults having high conversion rates is counter acting to the study made by Dale. Since not much data is available at the moment, as part of future study, interviews can be conducted with individuals in this age group to understand the reason better.

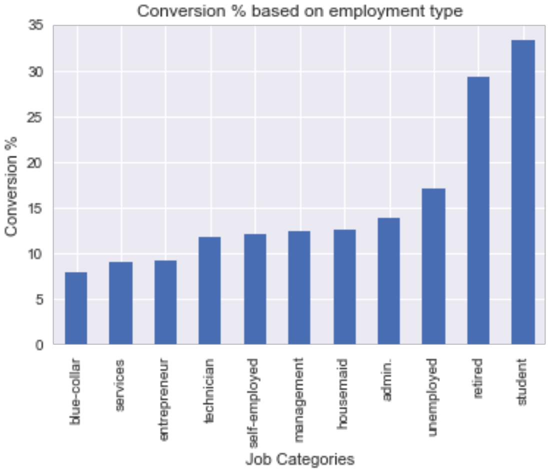
In addition to this, Chi-square test results have confirmed that there is indeed a significant difference in how individuals with differing age groups have responded to Term deposit offers making this variable very important during modelling.

For modelling purpose, Adult, middle Aged and Pre-retirement groups with similar conversion rates have been combined into Working adults category.

Comment [3]: Addition

8.4 JOBS

According to the dataset, Individuals who are retired, are students, unemployed or have administrative roles have higher conversion rates compared to those working in other roles.



Retired individuals do not draw regular incomes. Their main interest would be to invest all the savings that have been accumulated so far into low-risk safe return generating products such as term deposits. Also, the average age of retired individuals is 62 which supports the analysis derived from age variable.

Students do not have a consistent source of income and are most likely to look for avenues that can grow their savings without having inherent risks until they reach their prime earning period. Blue-collar, housemaids and technicians have very low disposable incomes and tend to have minimum to no savings for them to be able to invest.

On the other hand, risk taking, whether financial or social is a distinguishing characteristic of entrepreneurs. High-risk avenues like stocks and other equity based instruments are more rewarding for them than term deposits.

It is interesting to see unemployed in the high conversion category. External data can be combined in future to arrive at a consensus.

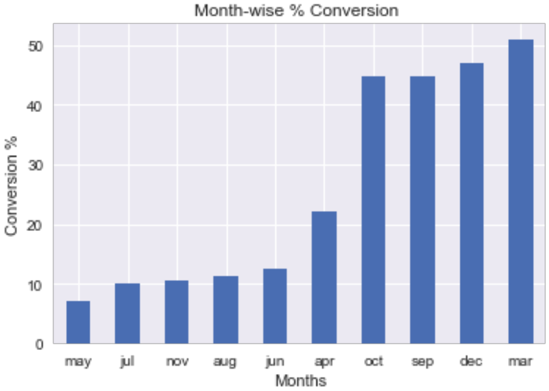
Mean age of the top two categories has been calculated to see if the results support what has been observed while analyzing the Age category and there indeed do. Young adults represent 25 yrs age group and retired individuals represent individuals aged 60 yrs and more which in in-line with our previous observation.

Chi-square analysis showed a statistically significant difference in deciding whether they wanted to take term deposits based on the kind of job that they were doing. An underlying factor for this could be the difference in salaries which could be considered for future analysis.

8.5 MONTHS OF YEAR

The Portuguese tax year runs concurrently with the calendar year from 1 January to 31 December. Individuals hold liquid cash until year-end in anticipation of unexpected expenditures over the course of that year.

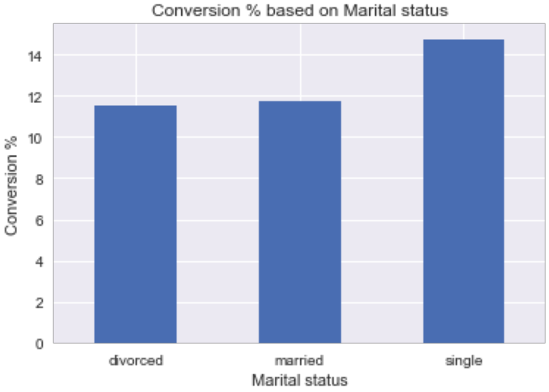
While locking funds in low-return investments such as term deposits in initial months itself might not be a good decision, year-end could be a good time to invest in them in order maximize tax benefits.



Data ranges from May 2008 to Nov 2010, thus reducing the possibility of random occurrences to a good extent.

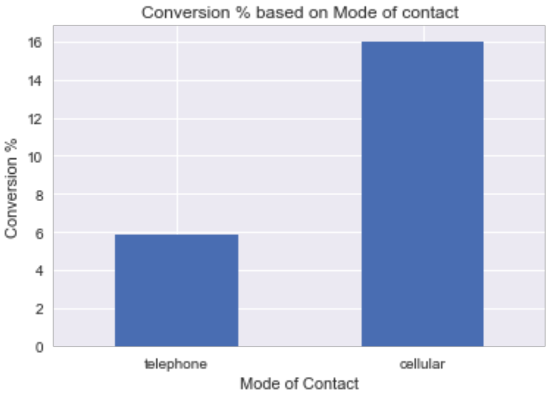
8.6 MARITAL STATUS

Not much can be inferred from marital status and its effect on the outcome in isolation. Each of the categories has close and similar conversion rates.



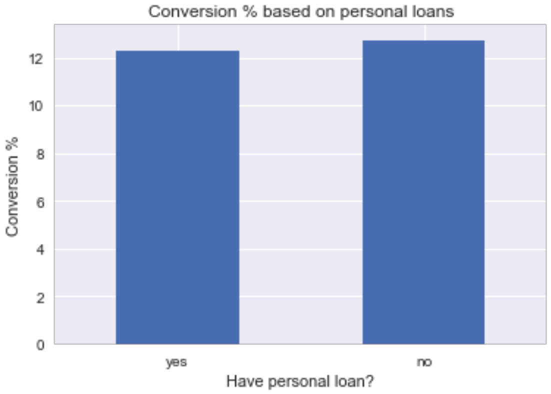
8.7 Contact

It has been observed that the conversion rate for customers contacted via mobile phones is 3 times higher than those contacted through landlines. Mobile bank transactions are slowly gaining popularity among individuals with varying socio-economic backgrounds due to comfort and ease of operation and this would provide banks with customer's mobile information.



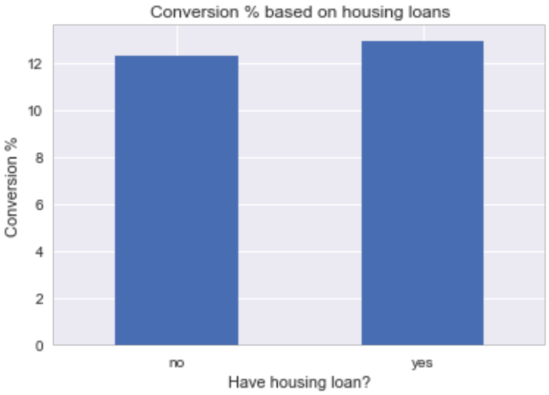
This shift could directly impact outcomes of marketing campaigns in study and thereby improving campaign profits to a great extend; based on above observation if everything else remains constant, just by contacting all customers via mobiles could increase profits by 3 times.

8.8 Loan



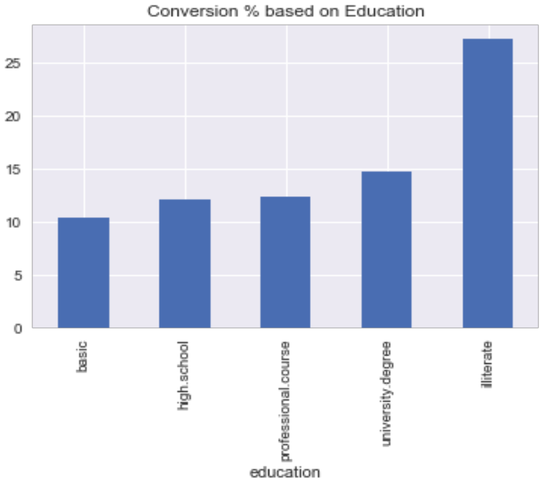
The above graph represents how customers having prior loans have reacted towards the campaign compared to those who didn't. Conversion rates for both the categories are similar hence not much can be deduced from this variable. This variable is being dropped from further analysis.

8.9 Housing(loan)



In the case of customers with/without housing loans, conversion rates are similar hence not much can be deduced from this variable. This variable is being dropped from further analysis.

8.10 Education



Similar conversion rates have been observed across all education categories other than illiterate category. However, sample size of this category is negligible compared to that of customers with some kind of education background.

8.11 PDAYS

Approx. 90% of the customers have not been contacted for any marketing campaigns in the past. This variable will not help in predicting outcome of the campaign and therefore will be dropped from further analysis.

8.12 POUTCOME

Previous campaign outcome is non-existent for 90% of the customers. This variable will not add any value while trying to predict current campaigns outcome and hence will be dropped from further analysis.

8.13 PREVIOUS

This variable is same as poutcome. It is the number of contacts performed before this campaign and for this client. This variable shall be dropped as well.

8.14 MACRO-ECONOMIC FACTORS

Following macro-economic factors have been considered in this study. A correlation matrix has been created to see if these factors had any relation among themselves

CORRELATION MATRIX	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
emp.var.rate	1	0.775	0.211	0.972	0.907
cons.price.idx	0.775	1	0.070	0.689	0.524
cons.conf.idx	0.211	0.070	1	0.292	0.115
euribor3m	0.972	0.689	0.292	1	0.945
nr.employed	0.907	0.524	0.115	0.945	1

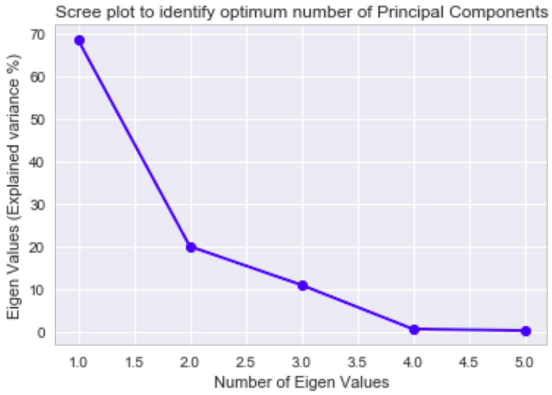
There is a high correlation among employee variable rate, euribor3m, nr.employed and cons.price.idx.

NORMALIZATION & PRINCIPAL COMPONENT ANALYSIS

Principal component analysis has been incorporated on the dataset to overcome the issue of multicollinearity but before doing that data has been normalized to bring all continuous values onto a common scale in order to reduce biases.

Scree plot shows the fraction of variance explained by each PC and helps with identifying maximum number of components that are sufficient to represent all the variables that are considered for PCA.

From the below plot, for Eigen values greater than 3, slope seems to flattened out.



Explained variance %	Number of Eigen values
68.463879	1
19.954623	2
10.871448	3
0.493359	4
0.216691	5

Top 3 PCs cater to 97% of variance and hence these have been considered for further analysis. These PCs have been merged with the variables remaining after performing EDA, to arrive at our final dataset.

Dummy variables have been created for all the categorical variables to make the data suitable to work with Machine-Learning algorithms.

9. PROFITABILITY - EVALUATION METRIC

Net campaign profitability is the amount that is left after deducting marketing expenditure incurred for making customer calls from the net interest income. Our end goal of this analysis is to present to the marketing manager a predictive model that can return highest possible profitability and from a modelling perspective the goal is to identify optimum model parameters that can maximize this chosen metric.

A cost benefit analysis has been performed to arrive at an equation that can determine Net interest income and marketing expenditure.

COST BENEFIT ANALYSIS

Net Profitability = Net Interest Income - Total Marketing Expenditure

1) **Net Interest Income**

This is the difference between the interest income generated by banks by giving loans and the interest paid to customers for borrowing their money in the form of Long-term deposits and like.

Since this information is not available in the dataset, information gathered from World bank data* has been considered. Portuguese banks have on an average made 4.3% points of Interest margin during the campaign period.
This when multiplied with the principal loan amount will give us the net interest income in currency terms per converted customer.

Assumptions for calculating Net Interest Income
For calculation purposes,

- principal amount for long-term deposit has been kept at $1000
- term of this deposit has been limited to 1 year with a Simple Interest rate of return. Both the values can be changed based on future campaign provisions.

NET INTEREST INCOME FROM CONVERTING ONE CUSTOMER	
Long-term deposit amount per customer	1000
Net interest margin[2]	4.3 % of Term deposit amount
Net Interest Income per converted customer	**$43**

2) **Marketing expenditure**

Marketing expenditure is the amount spent on employees for making calls to customers regarding the offer until they make their final decision of either accepting the offer or rejecting it.

Assumptions for calculating Net marketing expenditure

- Hourly salary for a similar role has been obtained from Glassdoor.com
- Total time spent on each customer includes time for pre-work and post-work.
- Productive time has been approximated to 6hrs per day which includes breaks.
- Number of working days for bank employees has been assumed to be 5 days/week with 20 official holidays & vacation days per annum.

TOTAL MARKETING EXPENDITURE PER CUSTOMER		
Tt	Average time spent on each customer during the campaign	645.7 seconds + pre & post call work = ~ 30 minutes
Tc	Total Number of customers (From the dataset)	41188 customers
Dw	Number of working days in the campaign period	2.5 years *(5 working days *52 weeks) - national holidays & vacation/sick days = 630 days
Cc = Tc/Dw	Number of customers called per day	41188 customers/630 working days = 65.4 customers /day
S	Salary per employee per day[3]	$128

H	Number of actual working hours per day considering breaks	6 hrs
Ce= H * 60 mins / Tt	Number of customers called per day per employee	6 hrs * 60 mins/ 30 mins = 12 customers/day/employee
Ne = Cc / Ce	Total number of Employees that would have been required to fulfill the requirement of 65.4 customers/day	65.4 customers / 12 = ~ 5.45 employees/day
Cm = S* Ne	Cost of marketing per day	$128 * 5.45 employees = $697.37
Cm / Cc	Cost of marketing per customer	$697.37 / 65.4 customers = ~ $11

Net Profitability = Net Interest Income - Total Marketing Expenditure
= $43*(Number of converted customers) - $11*(Total number of customers contacted)

Confusion matrix has been used to find True Positives, False positives, True Negatives and False negatives for every model and at varying probability tresholds. This metric in terms of confusion matrix attributes would be

Net Profitability= $43*(True Positives) - $11(True Positives + False Positives)

10. MACHINE LEARNING ALGORITHMS AND RESULTS

BASELINE PERFORMANCE:

11.2% has been the conversion rate for this campaign. Out of the total 41188 customers contacted, 4613 of them have accepted the offer and the remaining 36575 customers have rejected it.

Profitability = $43*(Total converts)-$11*(Total customers contacted)
= $43*(4613)-$11*(41188)
= $ 198,359 - $ 453,068
= (- $ 2,54,709)

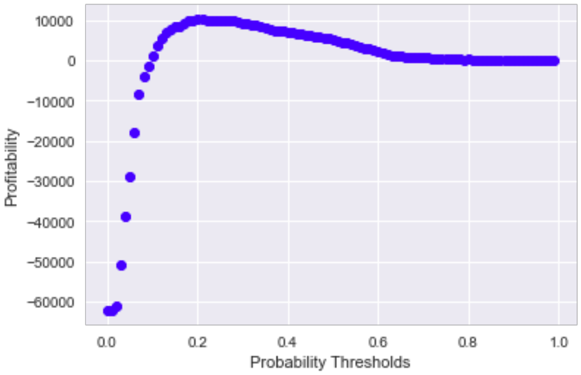
Overall, this campaign made a loss of $2,54,709 hence our goal is not only to obtain profits but also to develop a model that can maximize it.
Logistic regression, Random Forest and Support Vector Machine algorithms have been trained and tested to arrive at a final model that returned highest profitability.

Final bank dataset has been divided into Train, Validate and Test sets. Train set was used to fit the data to ML model, validate set was used to determine the tuning parameters and finally the test data was used for testing the model and its outcome.

10.1 Logistic Regression:

10.1.1 Logistic Regression on Original imbalanced Data with regularization

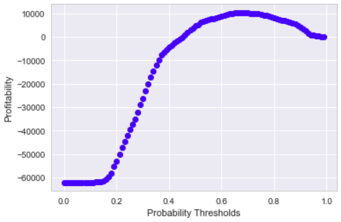
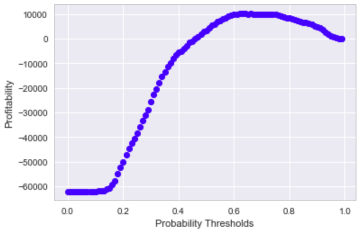
Sklearn's Logistic Regression was used to fit and test our bank dataset. Highest profitability of $10,159 was achieved with C = 100 at threshold of 0.21 with ROC AUC of 0.7746. Below is a graph of profitability at varying thresholds with the finalized regularization parameter



10.1.2 Logistic Regression on Resampled data

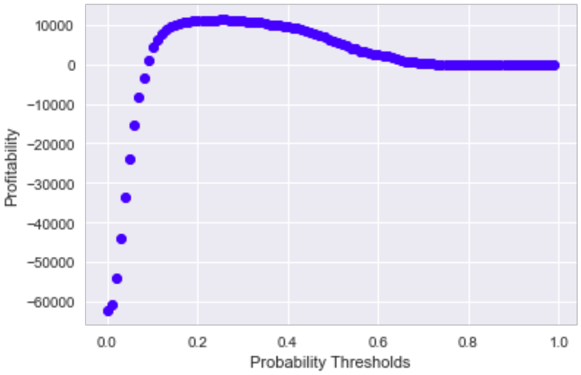
Dataset rebalancing has not proven to be effective in our business case. There was no improvement in profitability and ROC AUC scores.

	LOGISTIC REGRESSION	
	Upsampled data	Downsampled data
Final Regularization Parameters	C = 100	

Threshold	0.67	0.63
Profitability	$ 10,157	$ 10,114
ROC AUC	0.7783	0.7792
Profitability with varying thresholds		

10.2 Random Forest Classifier

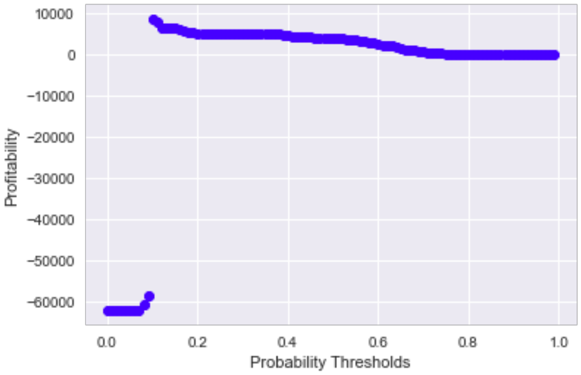
With minimum sample leaves at 5, minimum sample split at 10 and 50 estimators, profitability stood at $ 11,359 at a threshold of 0.26 with ROC AUC of 0.7955
Random Forest classifier performed better compared to Logistic regression and Support Vector Machines.



10.3 Support Vector Machines

Following are the profitability scores with SVM models

	Kernel	Gamma	C	ROC AUC	Profitability
SVM	RBF	auto	1	0.7073	$ 8,296
		auto	0.1	0.7095	$ 8,536
		10	1	0.6875	$ 2,804
	Linear	Not Applicable	1	0.6307	$ 1,076
			0.1	0.6586	$ 7,227
			0.01	0.7063	$ 6,551



It can be observed that SVM was not able to perform as good as either of the other two models.

11. CONCLUSION

Baseline profitability i.e performance of the original campaign stood at (- $ 2,54,709). By incorporating predictive analysis through machine learning algorithms there has been a drastic improvement in outcome predictions thereby providing a multi-fold increase in profitability. Summary of profitability with multiple models can be seen in the below table:

	Profitability	**ROC AUC**
Logistic Regression	$10,159	0.7746
Random Forest Classifier	$11,359	0.7955
Support Vector Machine	$8,536	0.7095

Random Forests turned out to be the best model of the three machine learning models used in our analysis. It provided an increase in profitability by 104.5 % over baseline model providing campaign profits of $ 11,359 to the marketing team.

12. FUTURE WORK

The focus of this work was to arrive at a predictive model that could maximize net Profitability for campaign managers. From the analysis done during this study, few variables had potential to help with other problematic areas faced by managers such as resource allocation and planning.

Campaign variable can be analyzed in detail to arrive at a probable cut-off for the number of calls that could be made to a customer before making the efforts redundant. This could help cut down on campaign costs and improve the overall efficiency of the team by helping them direct their efforts towards most probable customers.

When economy is flourishing, more and more individuals are willing to invest. This requires a need for hiring additional employees to fulfill those temporary needs which could be a good input for campaign managers during resource allocation step. The same could be achieved from an in-depth analysis of peak months of the year when customers are most likely to accept term-deposits if offered during that time frame.

13. ACKNOWLEDGEMENTS

I would like to thank my mentor Benjamin Bell, for guiding me through the project, especially with suggestions related to developing profitability metric to evaluate machine learning models.

14. APPENDIX

Consumer confidence index is defined as the degree of optimism that consumers have on the state of the economy, which is expressed through their activities of savings and spending. It measures how confident people feel about their income stability. Consumer confidence usually increases when the economy expands.

Consumer price index is a measure that examines the weighted average of prices of a basket of consumer goods and services, such as transportation, food and medical care. It is one of the most frequently used statistics for identifying periods of inflation or deflation.

Euribor3m refers to the interest rate banks charge each other on overnight loans. It is a benchmark rate banks use when setting interest rates on term deposits.

Employee Variable rate gives us a picture of how the economy is doing based on the number of employed people.

Nr.Employed is a quarterly indicator of the number of employed individuals during that period.

15. REFERENCES

- [1]https://www.ssa.gov/policy/docs/policybriefs/pb2007-02.html
- [2]http://data.worldbank.org/indicator/FR.INR.LNDP?end=2010&locations=PT&name_desc=true&start=2007
- [3]https://www.glassdoor.com/Salaries/sales-and-service-representative-salary-SRCH_KO0,32.htm