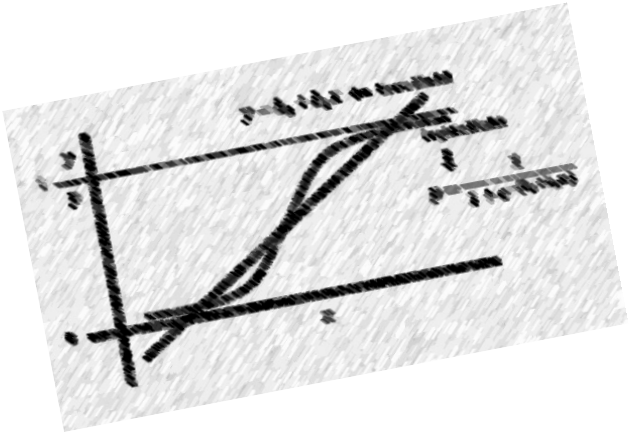
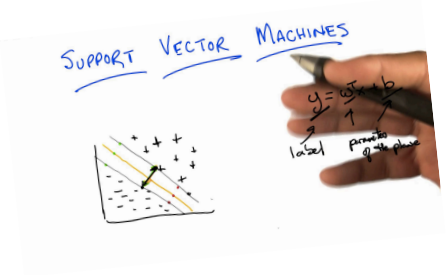
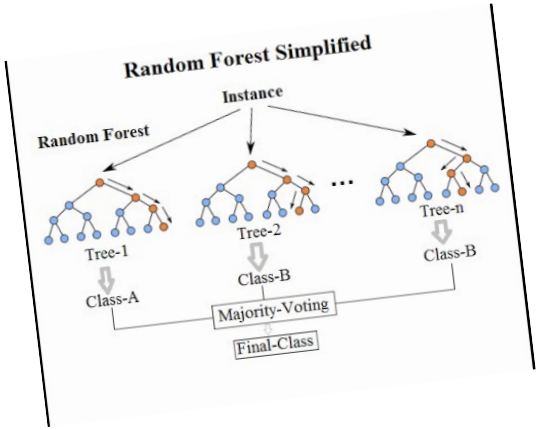
MAXIMIZING BANK'S MARKETING CAMPAIGN PROFITABILITY

THROUGH MACHINE LEARNING





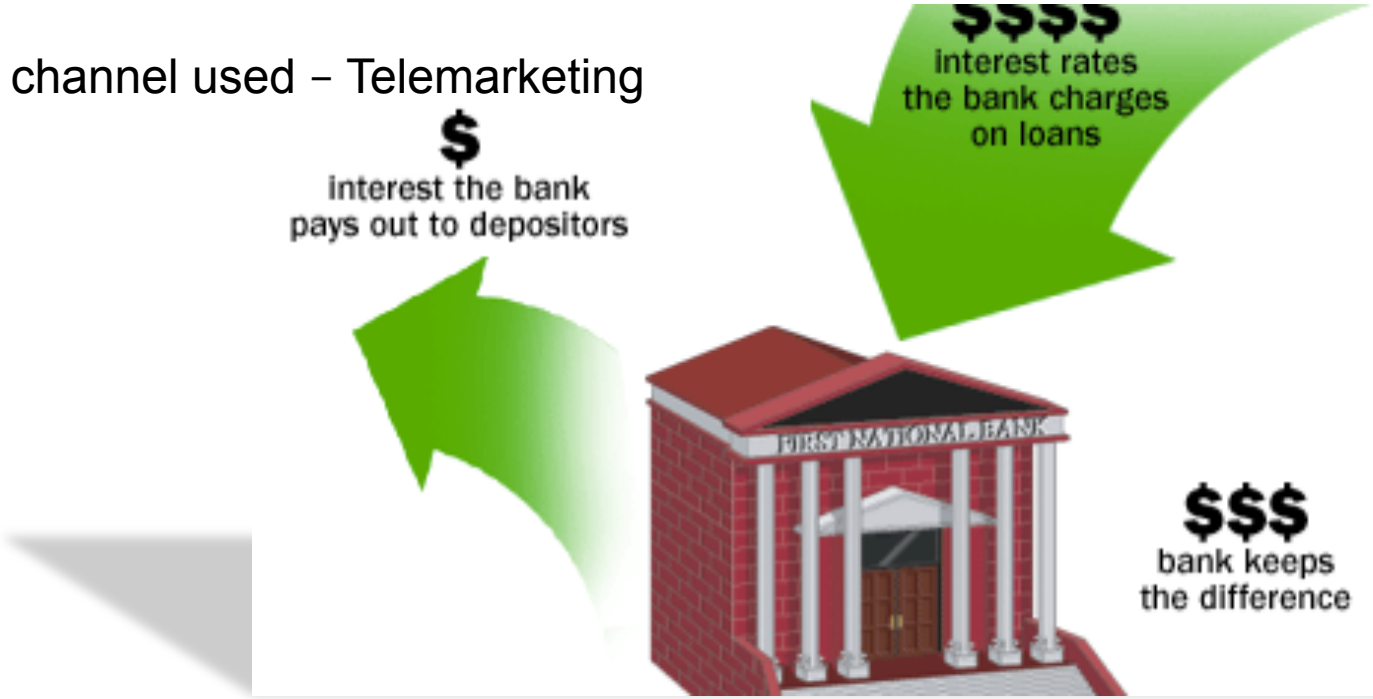


PROBLEM BACKGROUND

- Portuguese retail bank marketing long-term deposit offer to existing customers

- Long-term deposits
  - Fixed investment term, usually 1 to 5 years
  - Safe investments
  - Very appealing to conservative, low-risk investors

- Marketing channel used – Telemarketing



DATA AT A GLANCE

- **Data Source**
  - Publicly available on UCI website
  - CSV format
  - Data from external sources has been used during cost-benefit analysis

- **Data**
  - Data collected is from May 2008 to Nov 2010
  - 21 attributes and 41188 observations
  - 20 independent variables – client data, call data, socio-economic factors and campaign data

GOAL

- Prediction about customers that are most likely to accept term deposit offer
- Metric – Campaign profitability
- End goal is to Maximize campaign profitability
- Evaluate multiple machine learning algorithms and shortlist the one providing highest Profitability

DATA WRANGLING

- **Missing value treatment**
  - 6 variables with missing values
  - ‘Unknown’ data converted to numpy NaN
  - Based on analysis done during EDA phase,
    - Default, loan, housing variables have been dropped

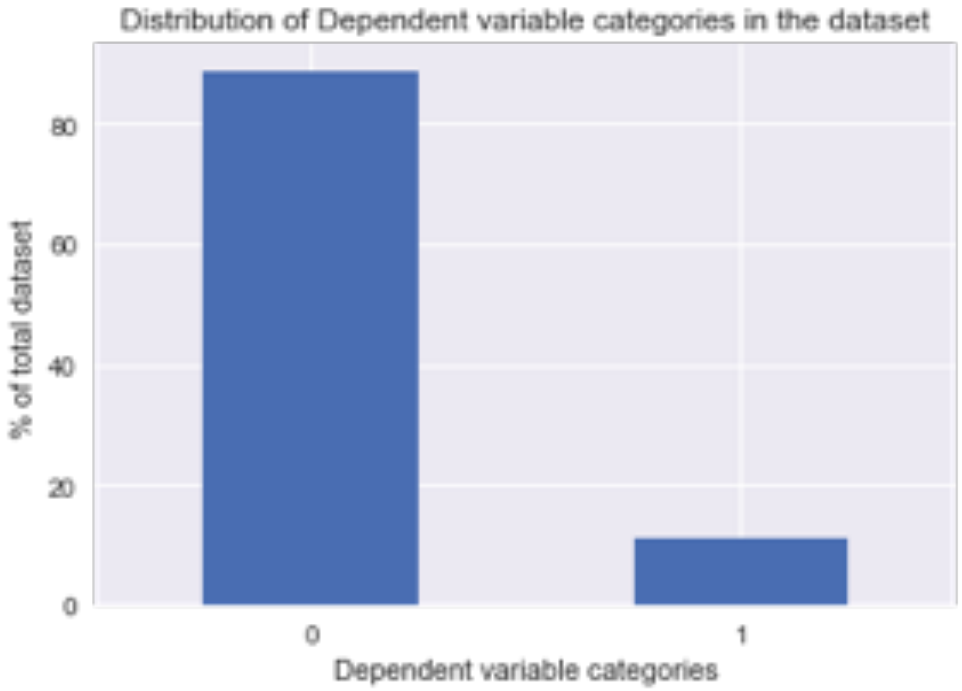
VARIABLE	% MISSING VALUES
job	0.8
marital	0.2
education	4.2
default	20.9
housing	2.4
loan	2.4
ALL OTHER VARIABLES	0

EXPLORATORY DATA ANALYSIS

- **Features**
  - Non-linearity with age variable
  - Chi-square
- **Data Normalization**
- **Multi-collinearity among macroeconomic factors**
  - Principal Component Analysis
  - Eigen values and scree plot

EXPLORATORY DATA ANALYSIS

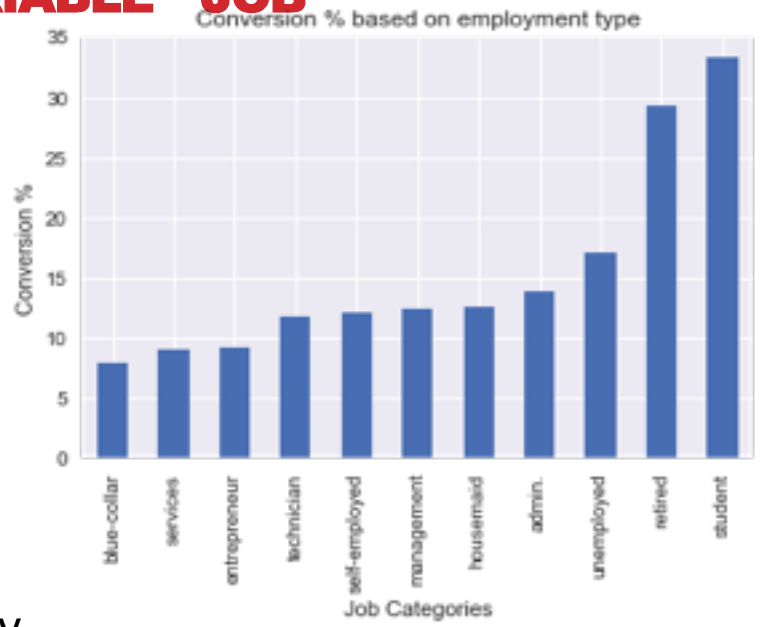
DEPENDENT VARIABLE (Y)



- 11.2% of the customers have accepted the offer
- 88.8% of them have rejected it
- Highly imbalanced dataset

EXPLORATORY DATA ANALYSIS

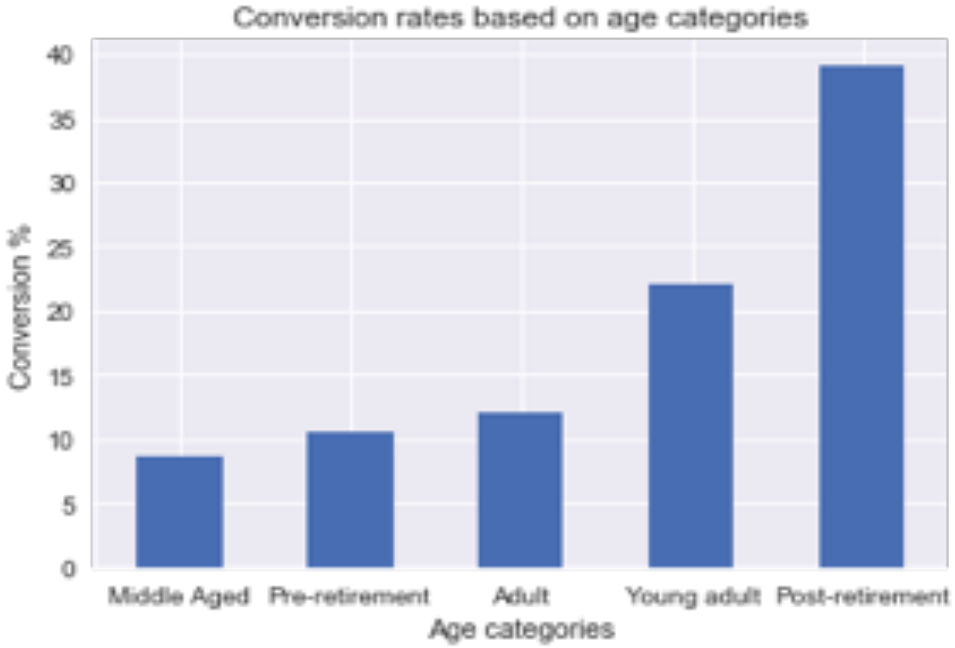
INDEPENDENT VARIABLE – JOB



- Retired category
  - Average age – 62yrs
  - Highest conversion rates
- Student category
  - No consistent source of income
  - Most likely to look for avenues that can grow their savings without having inherent risks
- Promising category for our campaign

EXPLORATORY DATA ANALYSIS

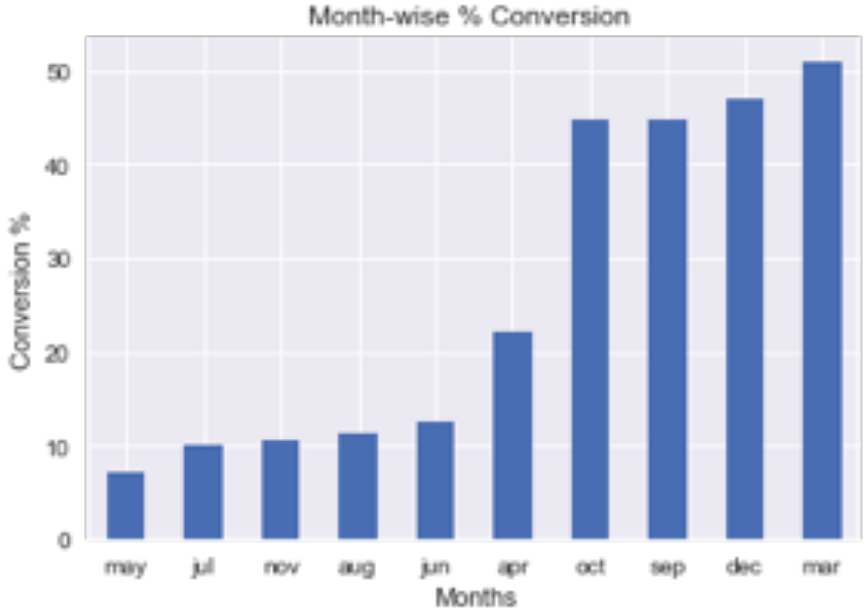
INDEPENDENT VARIABLE - AGE



- Age is inversely proportional to investment risk appetite – Study by Dale Kintlez
- Young adults contradicted observations made in the study
- Observed non-linear conversion rates with Age
- Chi-square results confirmed a presence of significant difference in response with Age
    - Chi-tabular < Chi-calculated(994.0), p-value <0.05

EXPLORATORY DATA ANALYSIS

MONTHS OF THE YEAR



- Portuguese tax year runs concurrently with the calendar year - 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.

EXPLORATORY DATA ANALYSIS

IMPORTANT FEATURES

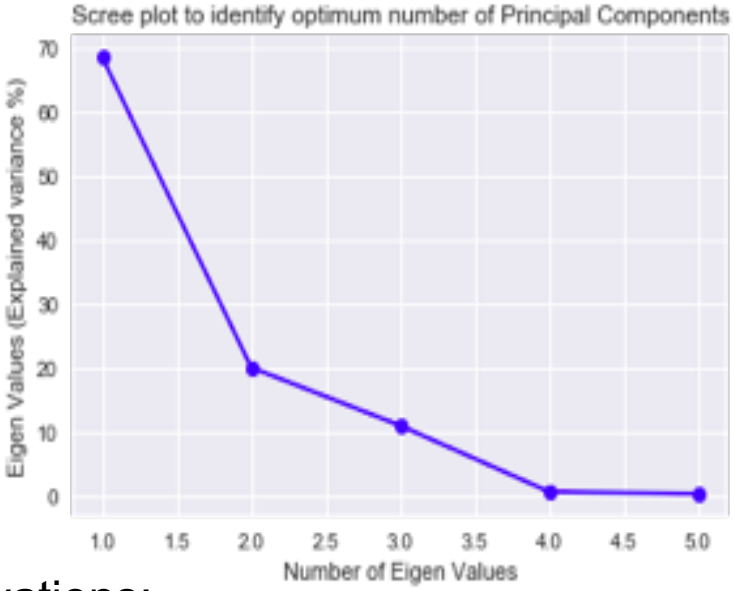
- **Macro Economic factors, Multi-collinearity & PCA**
  - High correlation among employee variable rate, euribor3m, nr.employed and cons.price.idx

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

  - PCA has been incorporated on the dataset to overcome this issue
  - Data has been normalized to bring all continuous values onto a common scale and reduce biases
  - Scree plot shows the fraction of variance explained by each PC and helps with identifying maximum number of components required to represent all variables considered for PCA

EXPLORATORY DATA ANALYSIS

IMPORTANT FEATURES



- Scree Plot observations:
  - Top 3 PCs cater to 97% of variance and hence these have been considered to replace all the macroeconomic factors
  - These PCs have been merged with the variables remaining after performing EDA, to arrive at our final dataset.
- Variables in final dataset:
  - Age category, type of Job, Marital status
  - Education, Month, Day of week
  - Number of calls made, dependent variable
  - PCA1, PCA2, PCA3

MACHINE LEARNING MODELS

- **Logistic Regression**
  - Performance on original data
  - Data rebalancing
  - Regularization
  - Model Evaluation
- **Random Forest Classifier**
  - Regularization
  - Model Evaluation
- **Support Vector Machine**
  - Regularization
  - Model Evaluation

EVALUATION METRICS

- **Profitability**
  - Cost-Benefit Analysis

TOTAL MARKETING EXPENDITURE PER CUSTOMER		
T	Average time spent on each customer during the campaign	645.7 seconds + pre & post call work = ~ 30 mins (½ hour)
S	Salary per employee per day[3]	$128
H	Number of actual working hours per day considering breaks	6 hours
S/H * T	**Cost of marketing per customer**	**$128/ 6 hours * ½ hour = ~ $11**

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

EVALUATION METRICS

- **PROFITABILITY (CONTD..)**

  Profitability= $43*(true positives) - $11*(true positives + false positives)

  - Advantages of Profitability
    - Best performance measure for this business need
    - Evaluating direct financial impact of model on campaign
    - Unbiased evaluation inspite of having imbalanced data unlike regular metrics such as precision, accuracy and F1 scores

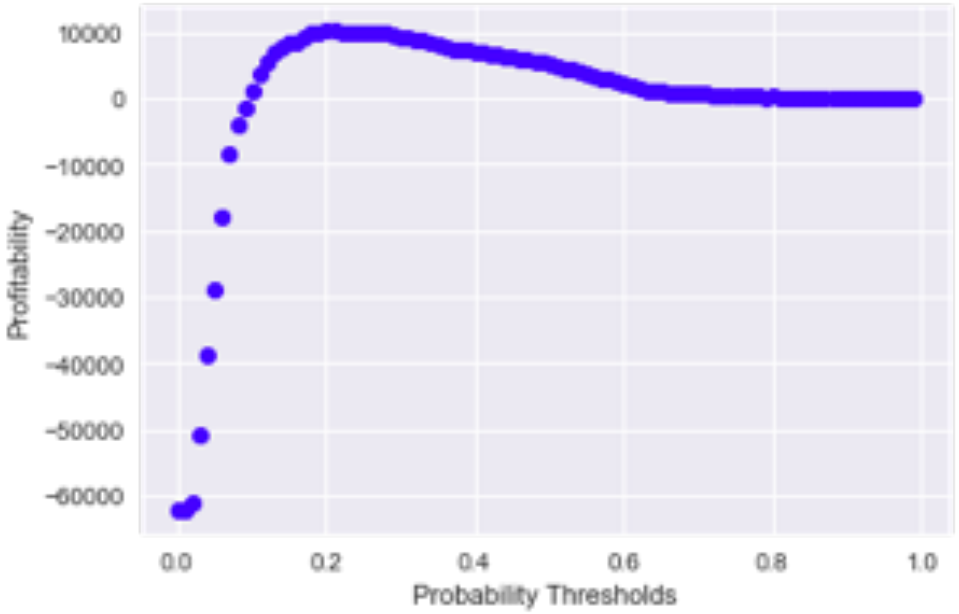
- **ROC AUC**
  - Advantages of ROC AUC
    - General metric
    - Deals well with situations where data is imbalanced like in our business case

BASELINE PERFORMANCE

- Conversion rate for this campaign - 11.2%
- Out of the total 41188 customers contacted,
  - 4613 accepted offer
  - 36575 rejected offer
- Profitability = $43(Total converts)-$11(Total customers contacted)
  = $43(4613)-$11(41188)
  = $ 198,359 - $ 453,068
  = (- $ 254,709)
- Overall, this campaign made a loss of $254,709
- Goal is not only to obtain profits but also to develop a model that can maximize it

LOGISTIC REGRESSION

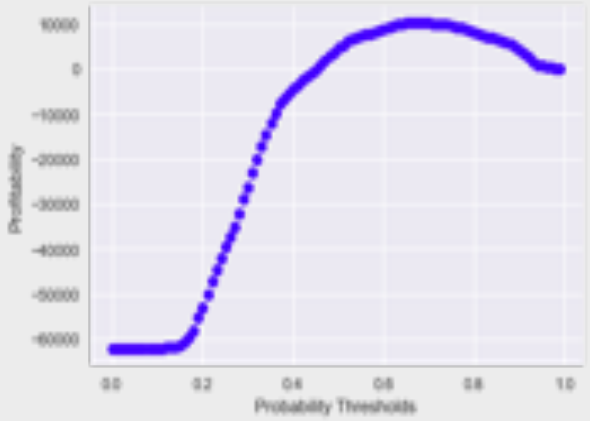
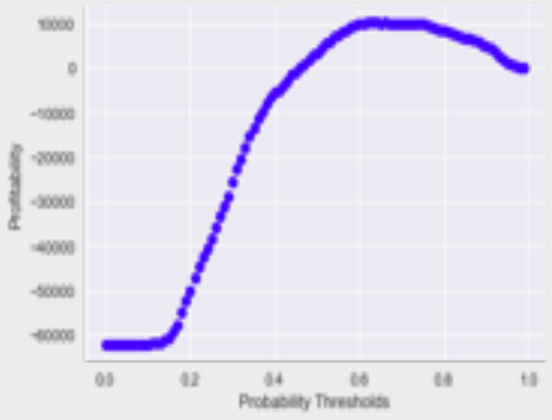
ORIGINAL IMBALANCED DATA



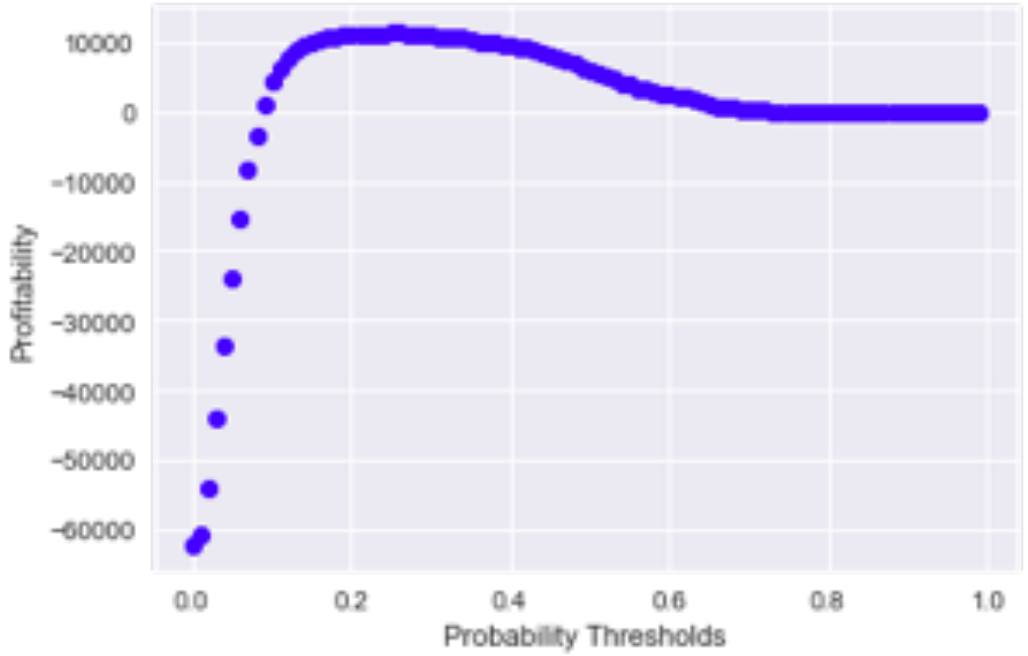
- Threshold at which profitability is the highest is: 0.22
- Regularization Parameter C : 100
- Maximum achievable profitability with the model is: $ 10,963
- ROC AUC: 0.7801

LOGISTIC REGRESSION

RESAMPLED DATA

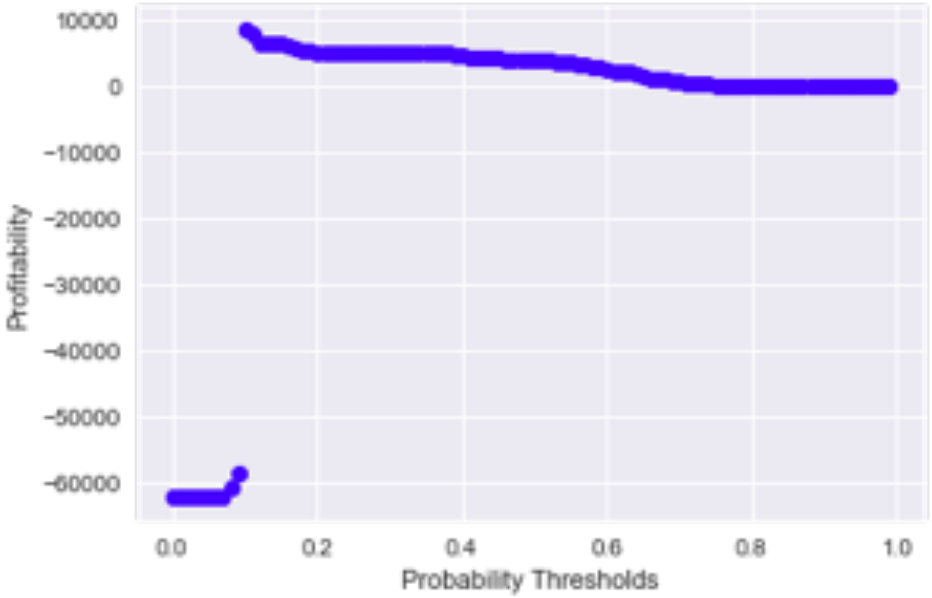
	Upsampled Data	Downsampled Data
Final Regularization parameters	C = 100	C = 100
Threshold	0.68	0.69
ROC AUC	0.7837	0.7808
Profitability	$ 10,806	$ 10,679
Profitability at varying thresholds		

RANDOM FOREST CLASSIFIER



- Threshold at which profitability is the highest is: 0.23
- Regularization Parameters
  - min_samples_leaf = 10
  - min_samples_split = 2
  - n_estimators = 50
- Maximum achievable profitability with the model is: $ 11,164
- ROC AUC: 0.7928

SUPPORT VECTOR MACHINE



- Threshold at which profitability is the highest is 0.1
- Regularization Parameters
  - Kernel= RBF
  - C = 0.1
  - gamma = auto

- Maximum achievable profitability with the model is: $ 8,147
- ROC AUC: 0.6876

CONCLUSION

	Profitability	ROC AUC
Logistic Regression	$10,963	0.7801
Random Forest Classifier	$11,164	0.7928
Support Vector Machines	$8,147	0.6876

- Random Forest Classifier is the best out of the three models evaluated
- Provided an increase in profitability by ~104 % over baseline model
- Maximum Campaign profitability of $11,164 with this RFC

ITS NOT OVER….

FUTURE STUDY

- From the analysis done during this study, few variables had potential to help with other problematic areas faced by managers; Resource allocation & planning

- Age – inconsistency could be analyzed in detail by conducting personal interviews with individuals

- Additional data can be gathered to arrive at actual profitability instead of assumptions made in this study to improve accuracy of the model

- This model can be generalized and applied to other investment campaigns as well by tuning parameters accordingly

THANK YOU
Any questions?