


# Potential Customer Recognition

- Using R Data Analysis

Group5 - David Gu, Lukia Lu, Celine Zhao

4/23/19

A decorative light blue triangle is located in the bottom right corner of the slide, pointing towards the top right.

# Problem Statement

- **Company information:** European financial insurance company A
- **Current practice:** Providing all levels consumer insurance products
  - Auto insurance
  - Homeowners insurance
  - Commercial line insurance
- **Problem identification:** Marketing expense is over budget, exceeding the profit by **\$87,450**
- **Desired solution:**
  - Concentrate target customer group
  - Predict potential generated profit

# Executive Summary



Marketing expense overrun  
Customer response rate low



Use Machine Learning to  
predict precision marketing  
plan based on historical  
data



Error rate of prediction as  
low as possible while not  
overfitting

# Table of Contents

●	Background-----	iv
●	Methodology-----	vi
○	Data-----	vi
○	Descriptive Analysis-----	vii
○	Modeling Framework-----	ix
●	Findings-----	xiv
●	Summary & Conclusions-----	xviii
●	Recommendations-----	xix

# Background



# Variables

Type	Name	Description
Input Variables	custAge	The age of the customer (in years)
Input Variables	profession	Type of job
Input Variables	marital	Marital status
Input Variables	schooling	Education level
Input Variables	default	Has a previous defaulted account?
Input Variables	housing	Has a housing loan?
Input Variables	loan	Has a personal loan?
Input Variables	contact	Preferred contact type
Input Variables	month	Last contact month
Input Variables	day_of_week	Last contact day of the week
Input Variables	campaign	Number of times the customer was contacted
Input Variables	pdays	Number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
Input Variables	previous	Number of contacts performed before this campaign and for this client
Input Variables	poutcome	Outcome of the previous marketing campaign
Input Variables	emp.var.rate	Employment variation rate - quarterly indicator
Input Variables	cons.price.idx	Consumer price index - monthly indicator
Input Variables	cons.conf.idx	Consumer confidence index - monthly indicator
Input Variables	euribor3m	Euribor 3 month rate - daily indicator
Input Variables	nr.employed	Number of employees - quarterly indicator
Input Variables	pmonths	Number of months that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
Input Variables	pastEmail	Number of previous emails sent to this client
Target Variables	responded	Did the customer respond to the marketing campaign and purchase a policy?
Target Variables	profit	If the customer purchased a policy, how much profit (before marketing costs) did the company make on the policy?

- Customer background

## Information

- Basic economic indicator

- Company status

- Customer Liaison information

- Targets

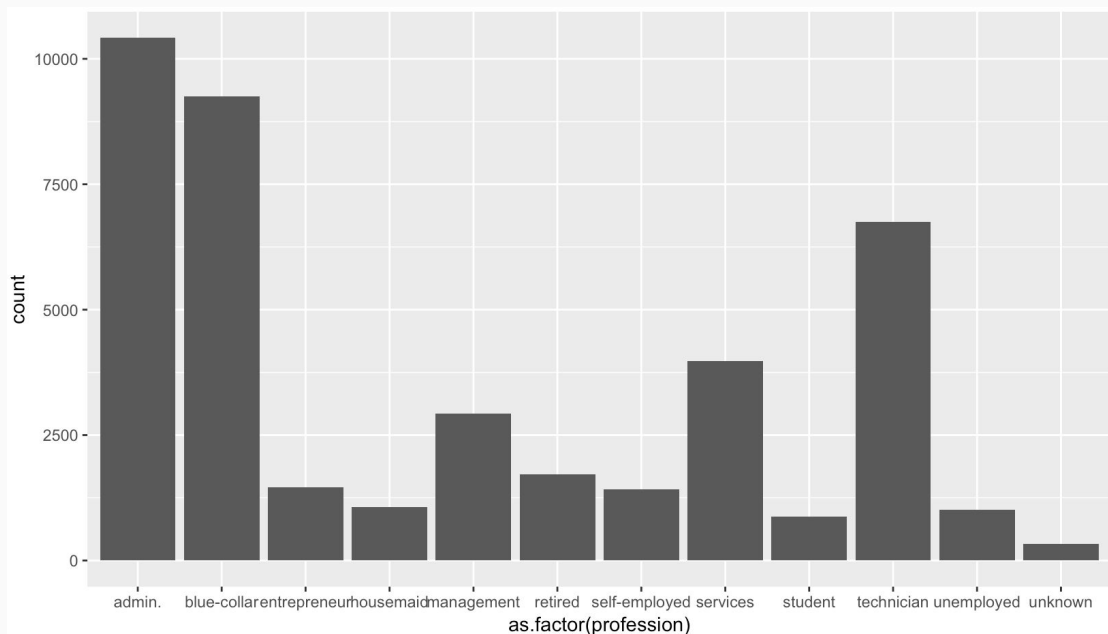
# Data

- **Sources:** Provided by insurance company using various methods such as surveys, e-mails and questionnaires.
- **Size & Structure:**

	Size	Classes	Source
Training dataset	8238 observations 24 variables	12 numeric 12 characters	Historical data set of all customers
Testing dataset	32950 observations 22 variables	12 numeric 20 characters	List of potential customers to whom to market

- **Unit of analysis:** individual client
- **Limitation & Delimitation:**
  - More than  $\frac{1}{4}$  of missing data regarding customer age and schooling variable
  - Serious class imbalance problem

# Descriptive Analysis - Profession



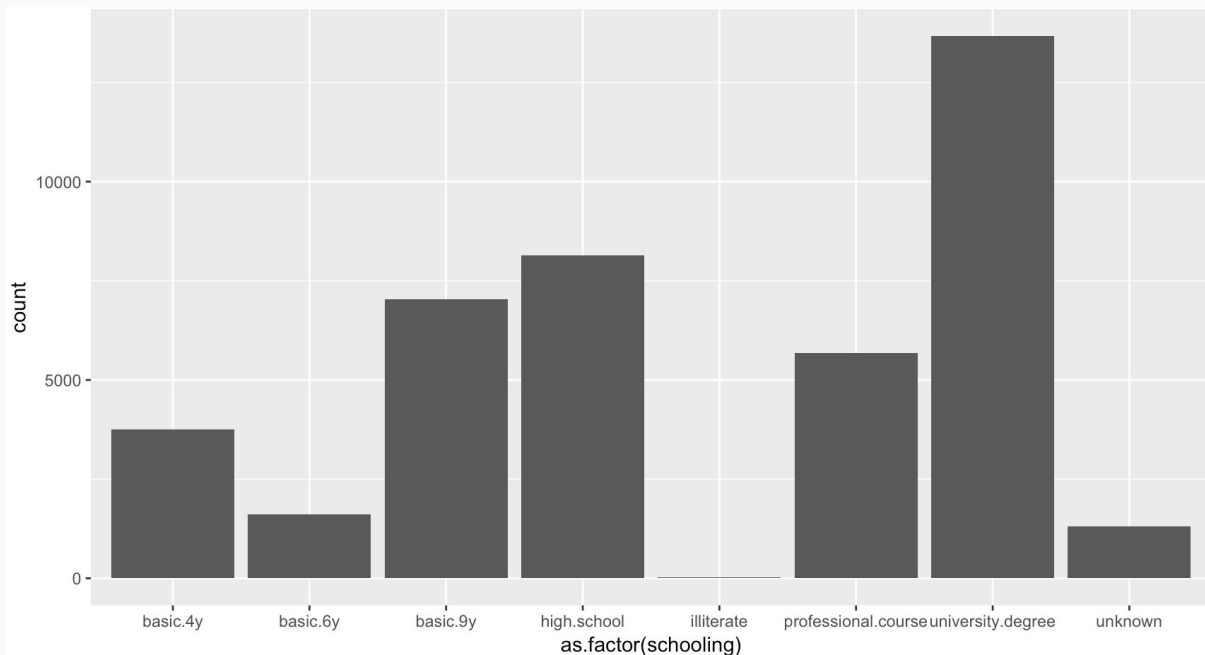
The profession of our clients fall mostly into 3 segments:

- Administrator
- Blue-collar
- Technician

Transform into numeric variable,  
eg: admin=0, blue-collar=1



# Descriptive Analysis - Schooling



The majority of our clients have a university degree, representing a high level of education background.

None of our client is illiterate.

# Modeling Framework

## – Data Cleaning

Age	Segment
0-30	Young
30-50	Middle
50-100	Old

- Imputing missing data (age & schooling)

- Assumption: customer age and schooling are related to profession
- Customer age: group the data and select the median age to fill NAs
- Schooling: group the data and select the mode of schooling to fill NAs

- Factoring the remaining character variables

- Other categories: find the mode to fill unknowns, transform character into numeric variable
- Customer age: divide age into three segments

- Analyzing the relationship between variables

- find variable correlations with “cor” function
- find variable importance with RandomForest

# Class Imbalance Issue

In classification problems, a **disparity in the frequencies of the observed classes** can have a significant negative impact on model fitting.

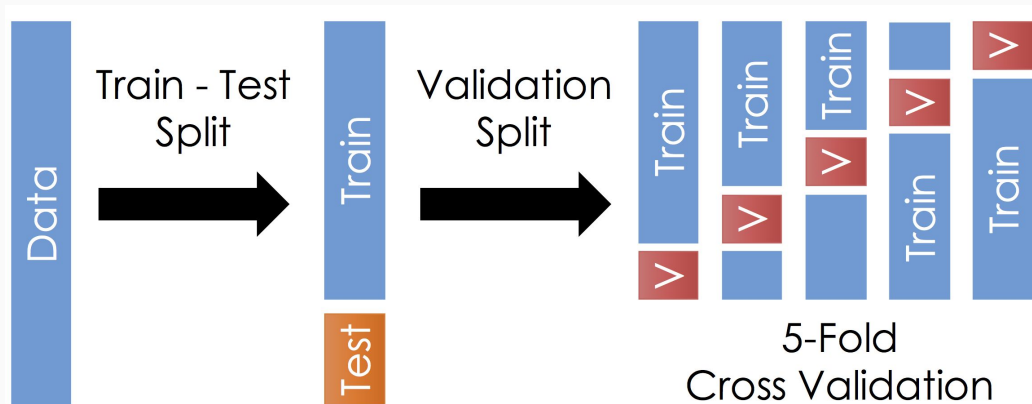
One technique for resolving such a class imbalance is to **subsample the training data** in a manner that mitigates the issues.

**Up-sampling:** randomly sample (with replacement) the minority class to be the same size as the majority class. We Use caret function `upSample()` to do this.

Not Responded	Responded	
5110	657	12.86%
5110	5110	50%

# Data Partition

- Training - 70% of train
- Testing - 30% of train
- UpSampledTraining - training with increased minority class (“responded=0”)



	Not Responded	Responded
Original Train	5110	657
UpSampled Train	5110	5110

# Modeling Framework– Classification

## Predicting “responded”

- method: Xgboost, GLM, LDA
- evaluation metrics: confusion matrix, error rate (with cross validation)
- best performer: GLM (with cross validation)

Error Rate	Original training data		Up sampling training data	
Xgboost	10.3%		24.89%	
	Important factors	All factors	Important factors	All factors
GLM	11.17%	7.8%	28.8%	17.94%
LDA	11.9%	10.40%	28.8%	22.91%

# Modeling Framework– Regression

Predicting “**profit**” for those who “responded”

- method: Xgboost, LM
- evaluation metrics: RMSE (with cross validation)
- best performer: LM (with cross validation)

RMSE	Original training data		Up sampling training data	
Xgboost	212.7209		223.4555	
	Important factors	All factors	Important factors	All factors
LM	126.8547	84.59152	128.3858	85.96366

# Findings - Correlation

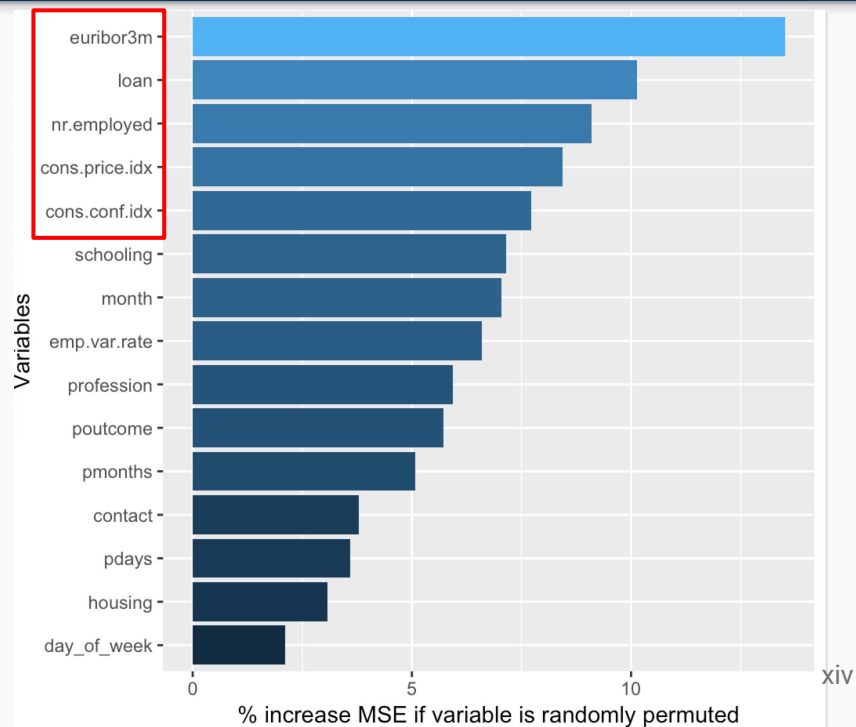
## Correlated factors

- a. **Euribor3m:** euribor 3 month interest rate

A selection of European banks lend one another funds whereby the loans have a maturity of 3 months.

Interest rate will decrease if client have more credibility.

- b. **Loan:** has a housing loan?  
c. **Nr.employed:** number of employees  
d. **Cons.price:** Consumer price index  
e. **Cons. conf:** Consumer confidence index

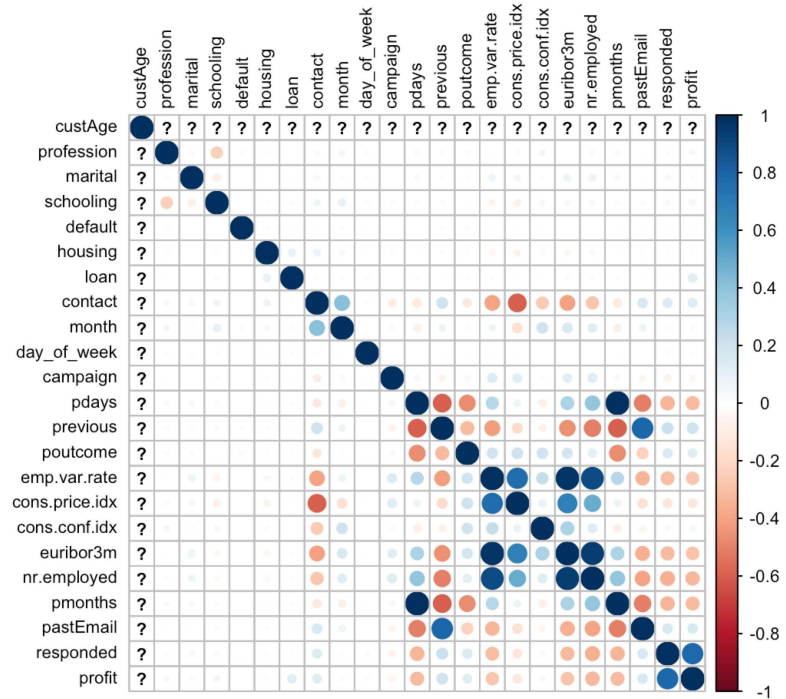


# Findings - Correlation

## •Positive Correlation:

- Employee variation rate - Number of employee
- - Euribor interest rate
- Emails been sent - Number of contact performed

Employee variation rate: the variation of how many people are being hired or fired due to the shifts.





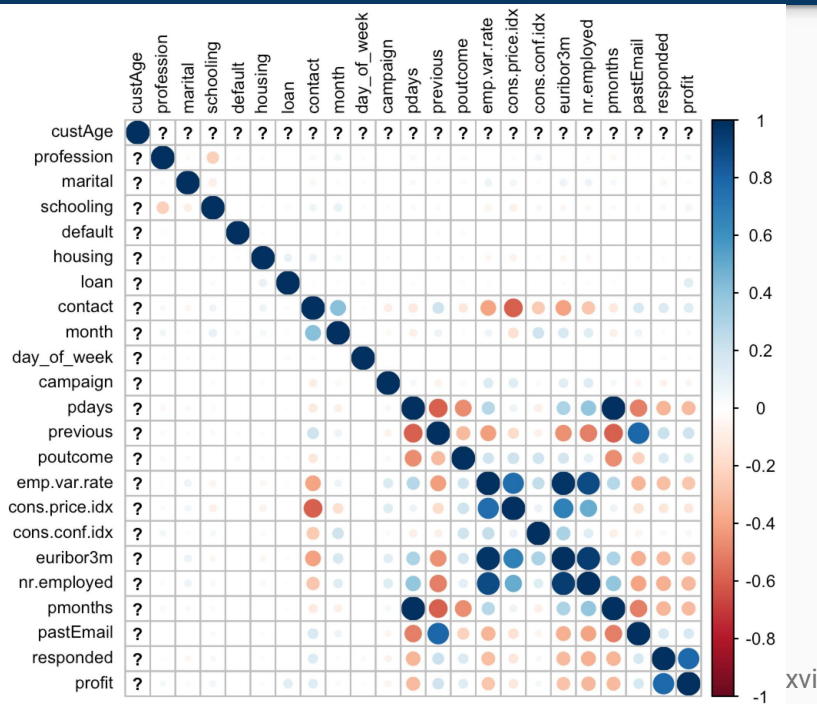
# Findings - Correlation

## • Negative Correlation:

- Consumer price index - Preferred contact type

	Increase	Decrease
Consumer Price Index	Inflation	Deflation
Preferred contact type	Telephone	Cellular

- Outcome - Number of day passed after contacted
- Euribor interest rate - Previous contact



# Findings - Model Performance

- Up Sampled Training data set does not improve model performance
  - incorrect use of Up Sampling function
  - the subsampling process probably induce more model uncertainty
  - the error rate is not a good metric in evaluating models handling class imbalance situation
- Using “Important factors” in predictive model does not create better performance
  - Possible over-fitting problems

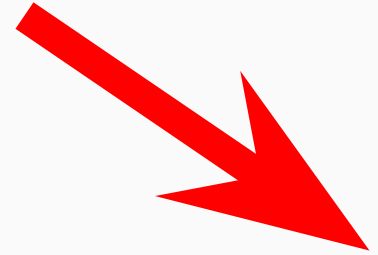
# Summary & Conclusion



1320



\$194.58



**\$-731745**

# Recommendation



## Contact Strategy

- First and fourth quarter
- Monday, Tuesday, Thursday
- Always follow-up: most pay at follow-up
- Employment rate is decreasing only
- Consumer confidence index around -40
- Euribor 3 month rate :below 1.5

## The ideal customer

- Old(50-100)
- "housemaid","management","retired","student","technician","unemployed"
- Single, Divorced
- avoid basic-educated
- Has loan
- Cellular only
- Less email contacts, less campaigns

# Thanks for listening

Q&A