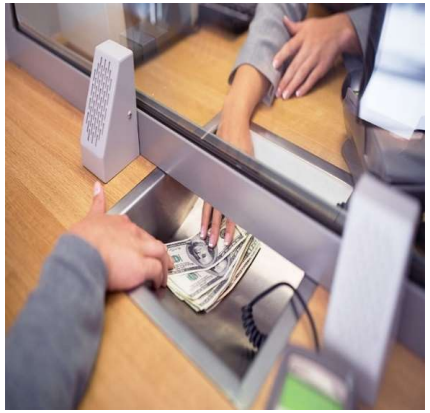
# Direct marketing campaigns (phone calls) of a Portuguese banking institution

## Executive Summary:


- The main focus of this project is to **predict whether the client will open a term deposit (variable y)** using **sociodemographic, social and economic context and account attributes**.

- Different supervised classification models were used and the most important features for the high score model (Ridge) explored are:
    - ❑Consumer price index, Communication type cellular
    - ❑Contact Month Jul, Consumers confidence Index
    - ❑Age, Job technician



- The results show that the Logistic regression model (using Ridge algorithm) is the best model

## Objective:

- To predict whether potential consumers will put deposits.
- To find out what machine learning model can predict the target variable better than others
- To explore what features affect consumer decision to put deposits

# Dataset Overview

| Feature Variables (21) | | |
|---|---|---|
| Age | communication type | outcome of the previous marketing campaign (poutcome) |
| Job | month | Employee variation rate |
| Marital | last_contact_day | consumer_price_index |
| Education | last_contact_duration | Consumer confidence index |
| Have_credit by default | Number of contacts with client | euribor 3 month rate |
| Housing loan | Number of days client was contacted in prev_campaign | employed staff rate |
| Personal loan | no_contact_bef_campaign_wth_salesperson | |

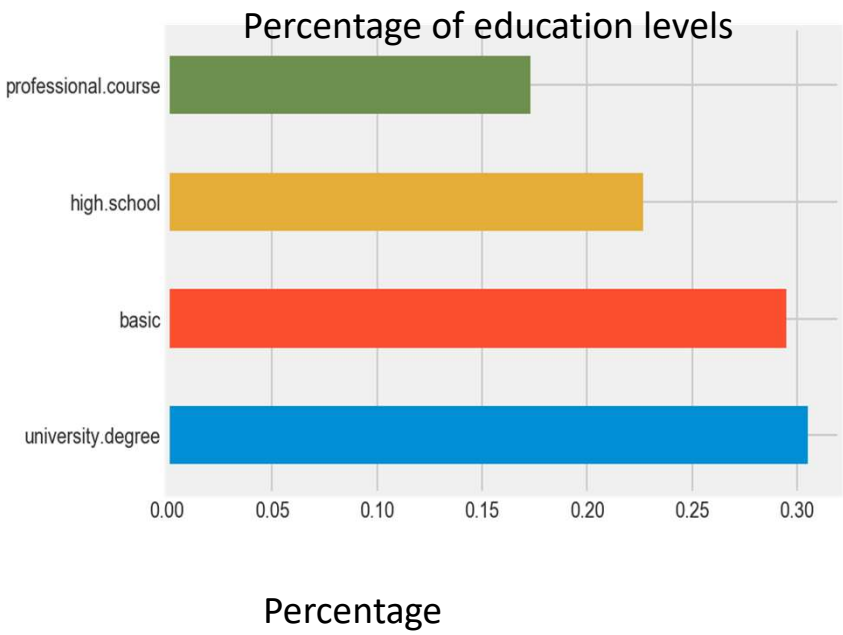
Dataset Size → 2999 observations with 21 feature.

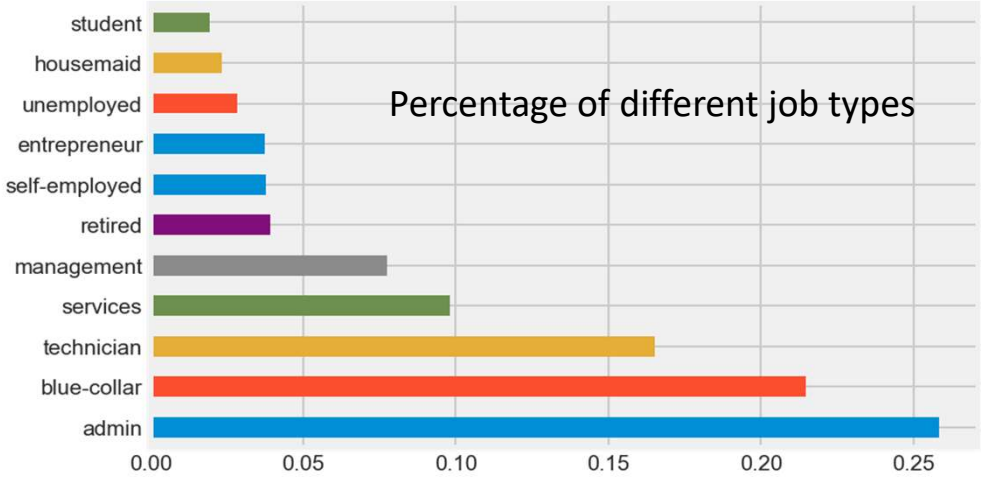
May 2008 to November 2010

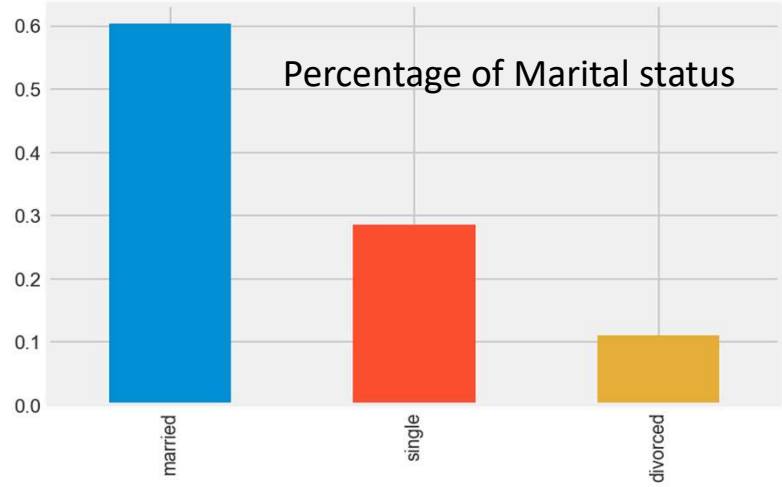

Background:
The Great Recession started to hit Portugal in 2008; that year the Portuguese economy did not grow (0.0%) and fell almost 3% in 2009

**Target is 'open deposit account or not'.**

# Exploratory Data Analysis  Part1

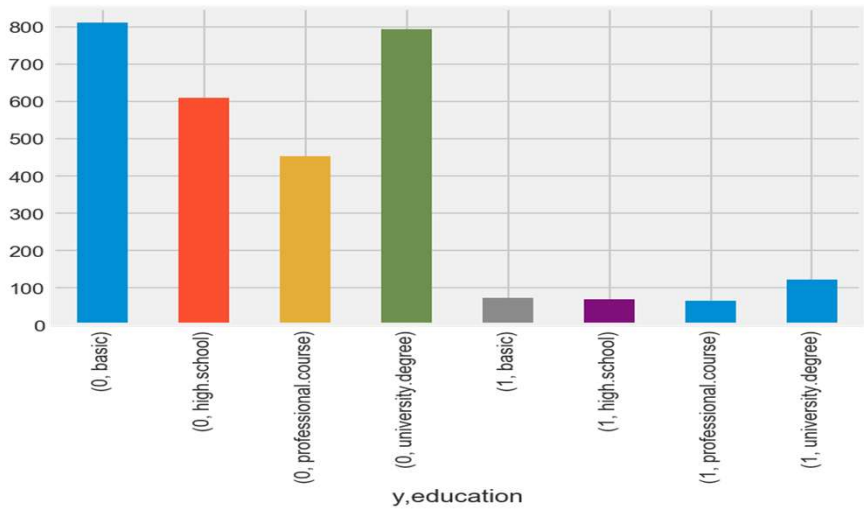


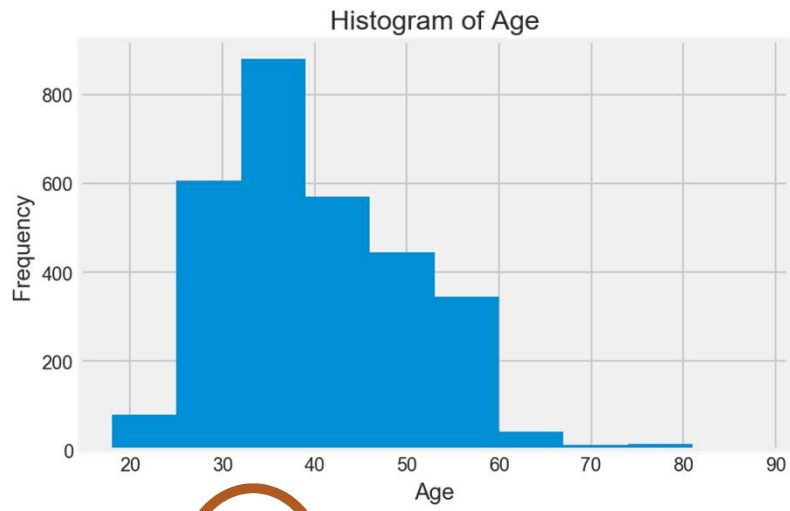



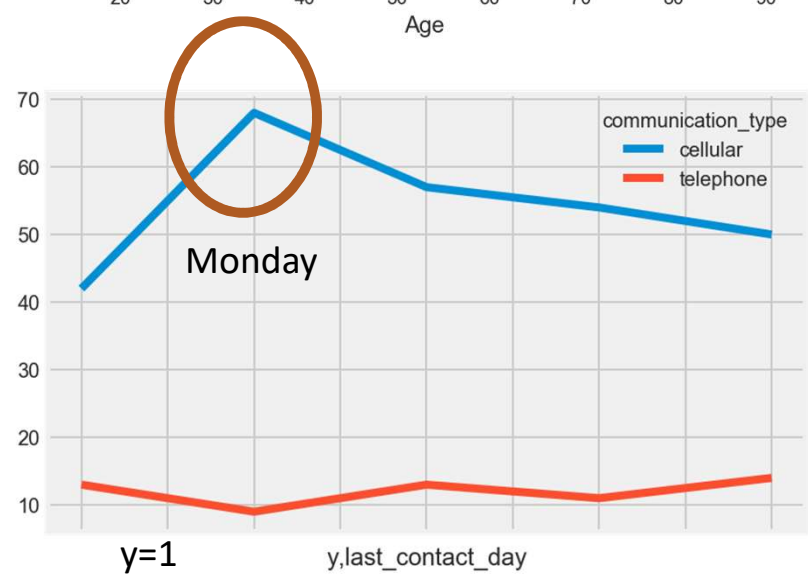




- Majority of potential customers are either highly educated or have basic level of education
- About 50% are either have admin (26%) or blue-collar jobs(22%).
- More than 60% are married

# Exploratory Data Analysis2




- Majority of customers were age bet ween 30-40 (Gen Y)

Char: Majority of them use the web and are mobile web user.
They seek out info and engage in two-way brand conversation.

No wonder that majority of them use cellular phones

Majority of contacts who put deposits for cellphone users were contacted on Mondays

# Feature Engineering Part 1 - Dealing with Multicollinearity

- Used VIF Variance Inflation Factor to detect multicollinearity.
  - The Variance Inflation Factor (VIF) is a measure of *collinearity among predictor variables* within a multiple regression.
  - Benchmark: If the VIF is between *5-10*, multicollinearity is likely **present** and you should consider dropping the variable.

- Columns deleted
  - nr_employees (Number of employees)
  - euribor3m (Euro rate for last 3 months)
  - Previous_campaign_outcome (previous campaign outcome)

# Dealing with Multicollinearity

- Multicollinearity is redundancy.
- When two or more predictors in a regression are highly related to one another

  - They do not provide unique and/or independent information about the impact of predictors on dependent variable.

# Feature Engineering Part 2 - Dealing with Imbalanced Dataset

Problem –

Imbalance dataset:

We already know!!! 88% didn't put a deposit.

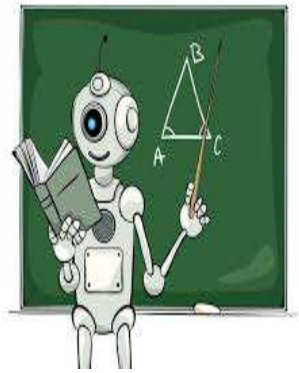
(Percentage of the larger group of the column (Putting deposit)

**Solution – The word is SMOTE !!!**

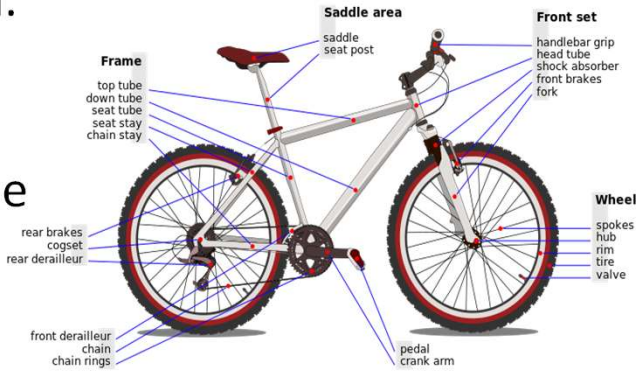
Need to have an equal number of instances(Y and N) of each classes.

# Modeling Workflow

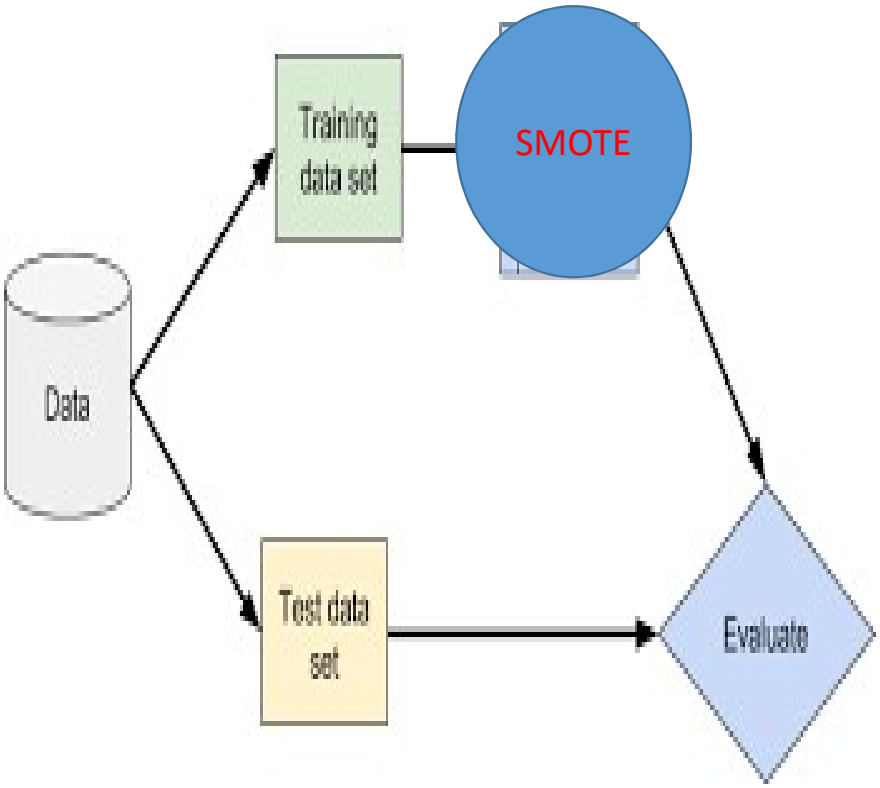
- Train-test split on your original data, for myself. 0.1:0.9.
- GridsearchCV on the training data set for all the learning algorithms (find best features).



- Check the baseline on the training set.
    - Baseline= 0.5
- Put in your test set on your best_estimator for each learning algorithm.
- Used the AUCROC score here to compare models



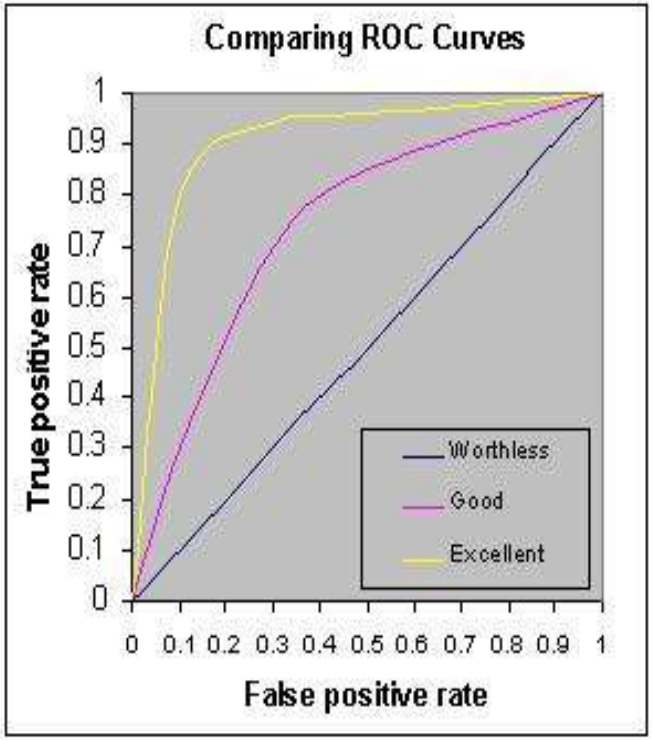

# Modeling Workflow




- Train-test split on your original data, for myself. 0.1:0.9.
- On the train data set, upsampling to make it a 50-50 split because of class imbalance.
- GridsearchCV on the training data set for all the learning algorithms.
- Check the baseline on the training set.
- Put in your test set on your best_estimator for each learning algorithm.
- Used the AUCROC score here to comapare models

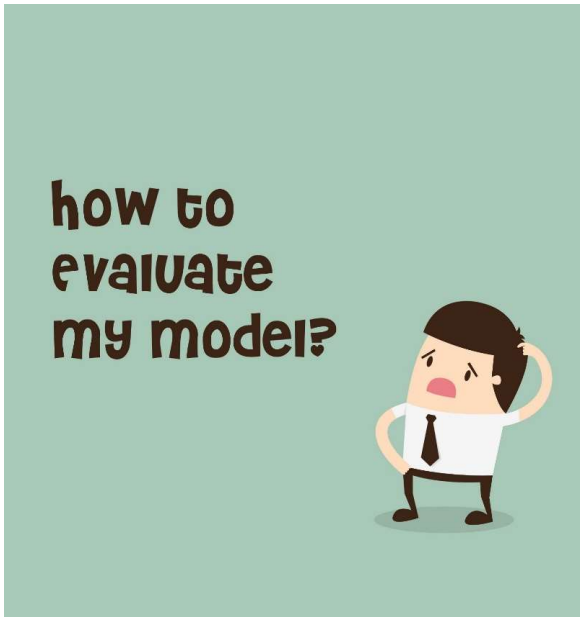
# Model Evaluation and Metrics

- Accuracy score is influenced by imbalanced dataset.
- Used AUROC
  - Not influenced by imbalance dataset.
  - The true positive rate is **plotted** in **function** of the false positive rate




*For Demo only, not model results*

# Model Evaluation and Metrics

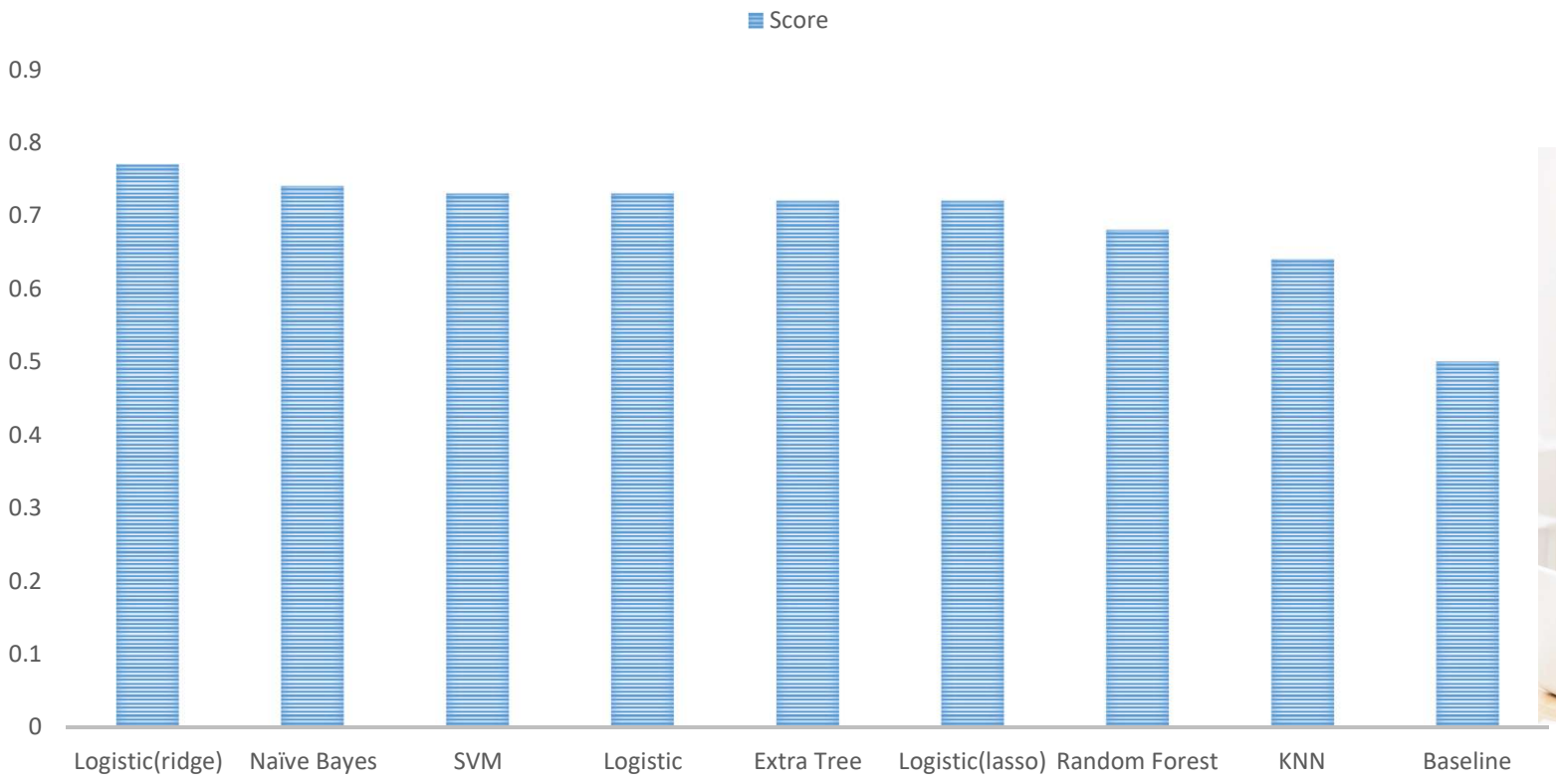



- Used AUROC: Performance measurement

  - Not influenced by imbalance dataset.

Higher the AUC, better the model is at distinguishing between putting and not putting deposit.

# Various Model Results




**Baseline=0.5**


Feature _Importance
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
consmr_price_indx
communication_type_cellular
month_jul
conmrs_confidnc_indx
age
job_technician
month_jun
month_mar
job_admin
n_days_clcontact_prev_campaign
ExtraTree
Logistic(Lasso)
Logistic(Ridge)
RandomForest
SVM

# Important features

| No | Ridge | SVM | ExtraTree | Lasso | Random Forest |
|---|---|---|---|---|---|
| 1 | Consumer price index | Contact Month jun | communication_type _cellular | Consumer price index | age |
| 2 | Communication type cellular | n_days_clcontact_ previous_campaign | n_days_clcontact_ previous_campaign | n_days_clcontact_ previous_campaign | conmrs_confidnc_indx |
| 3 | Contact Month Jul | Contact Month oct | conmrs_confidnc_ indx | month_mar | consmr_price_indx |
| 4 | Consumers confidence Index | Contact Month mar | housing_loan_no | conmrs_confidnc_indx | n_days_clcontact_ previous_campaign |
| 5 | age | communication_type_ cellular | marital_single | housing_loan_no | marital_single |
| 6 | Job technician | age | Contact Month Jun | age | housing_loan_no |
| | | | | | |

# Conclusion &Recommendation

- **Economic factor (Macro economical factor) :**
  - ❖The Consumer Price Index (CPI) is a measure that examines the weighted average of prices of a basket of consumer goods and services.
  - ❖ The Consumer confidence Index: The degree of **optimism** that consumers are expressing through their activities of savings and spending.

- **Social: Communication type: cellular**
  - ❖Target this group through social media in platforms they are active

❖**Demographic factor:**

- Age: offering different products or using different marketing approaches for different age and life-cycle groups

- **Month-July:**
  - ❖ Is the time people may receive the end of financial year

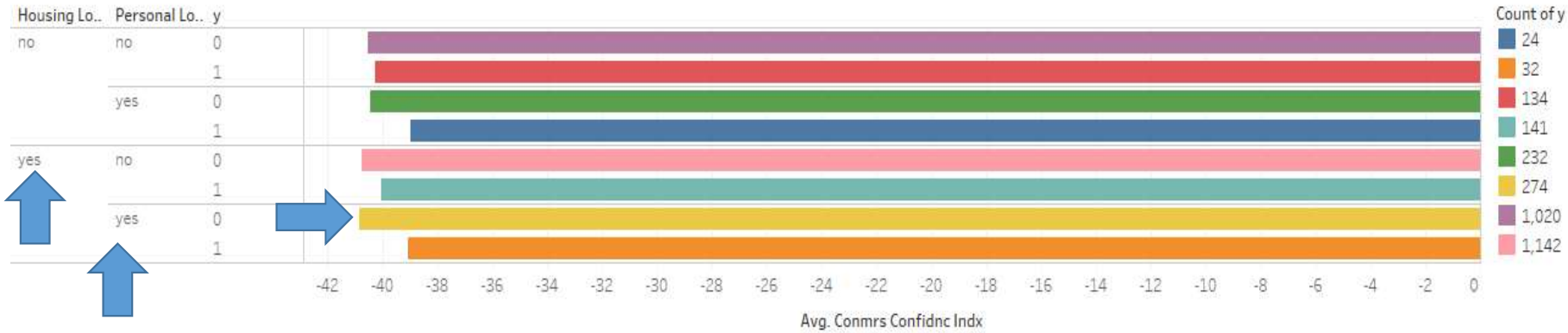
# A model of consumer behaviour

How do consumers respond to various marketing efforts the company might use?

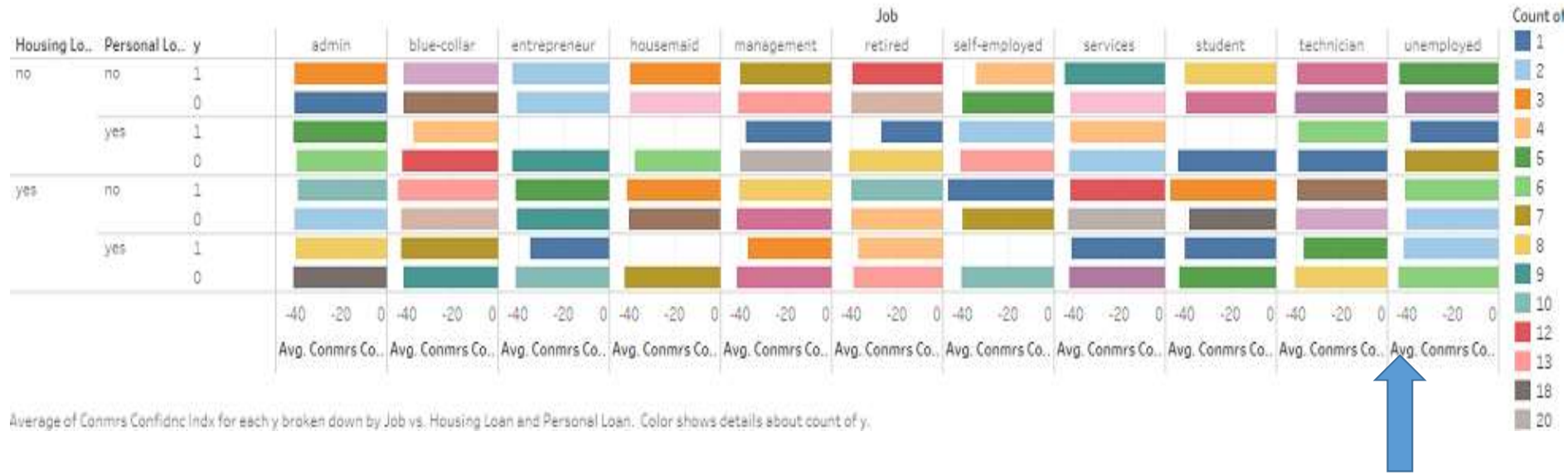
## The Black Box Model

| ENVIRONMENTAL FACTORS | | BUYER'S BLACK BOX | | BUYER'S RESPONSE |
|---|---|---|---|---|
| **Marketing Stimuli** | **Environmental Stimuli** | **Buyer Characteristics** | **Decision Process** | |
| | External ones | Internal ones | | |
| Product<br>Price<br>Place<br>Promotion | Economic<br>Technological<br>Political<br>Cultural<br>Demographic<br>Natural | Attitudes<br>Motivation<br>Perceptions<br>Personality<br>Lifestyle<br>Knowledge | Problem recognition<br>Information search<br>Alternative evaluation<br>Purchase decision<br>Post-purchase behaviour | Product choice<br>Brand choice<br>Dealer choice<br>Purchase timing<br>Purchase amount |

<house_personal loan>

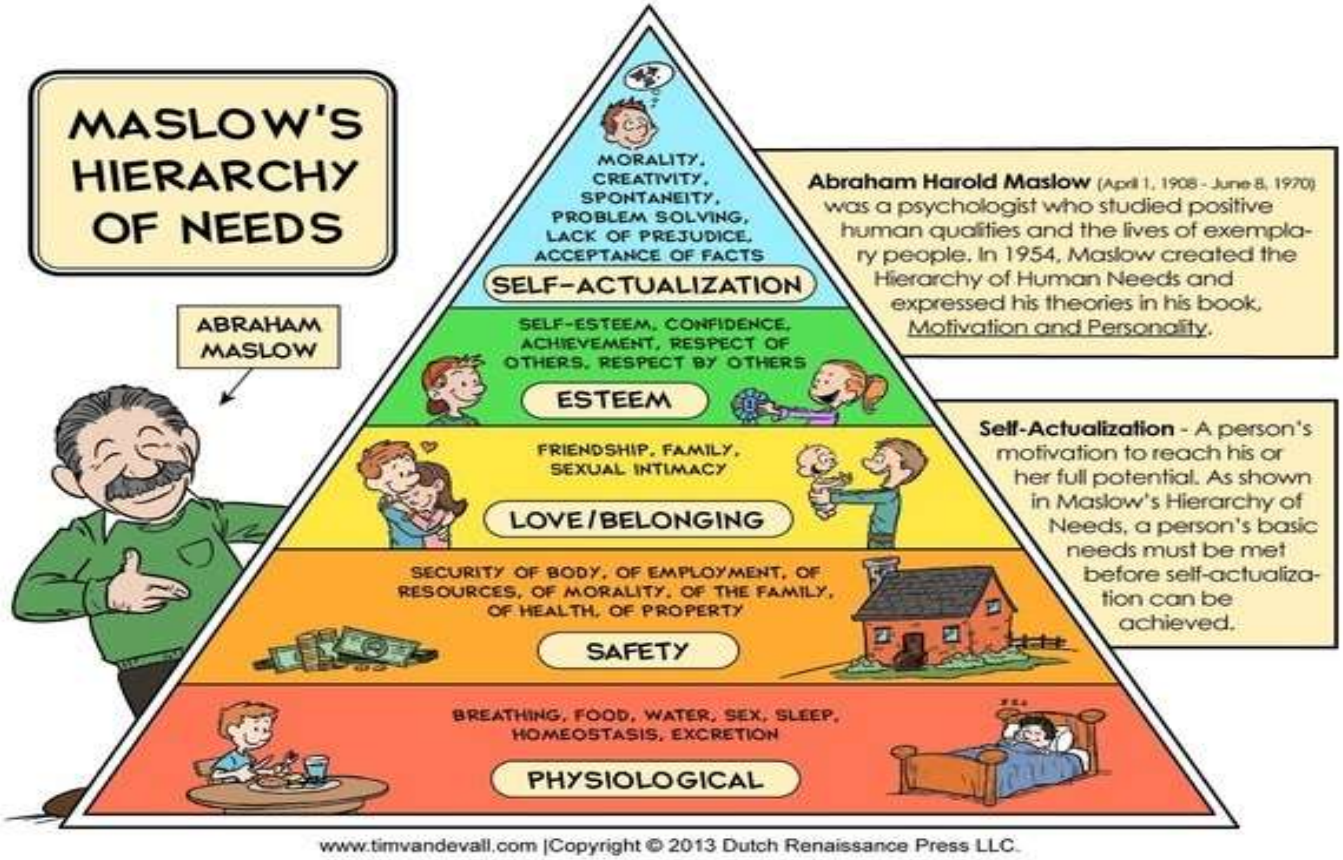



Average of Conmrs Confidnc Indx for each y broken down by Housing Loan and Personal Loan. Color shows details about count of y.




Average of Conmrs Confidnc Indx for each y broken down by Job vs. Housing Loan and Personal Loan. Color shows details about count of y.

# Influences of consumer behaviour: Psychological factors




Thank
you