FINAL PROJECT

Ray Ng

University of British Columbia

Rita Uzoka

rita.uzoka@ahoo.com

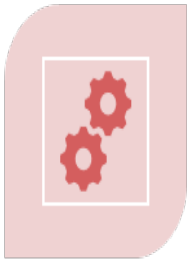
Sheffield Hallam University

Fatemeh Bagheri

UniversitÃ© Jean Monnet St Etienne - UniversitÃ© de Lyon

Agenda

PROJECT DESCRIPTION



EDA(EXPLORATORY DATA ANALYSIS)

TOOLS AND TECHNIQUES

IMPLEMENTATION

RESULTS

CONCLUSION

Project Description

Problem Statement

- XYZ bank wants to roll out Christmas offers to their customers. But the bank does not want to roll out the same offer to all customers, instead they want to roll out personalized offers to particular sets of customers. If they manually start understanding the category of customer then this will not be efficient and also they will not be able to uncover the hidden pattern in the data (pattern which groups certain kinds of customer in one category). Bank approached ABC analytics company to solve their problem. Bank also shared information with ABC analytics that they don't want more than 5 groups as this will be inefficient for their campaign.
-

Business Understanding

- ABC analytics should try to understand certain patterns in customer behaviour, which include relations between province, sex, seniority, etc. and household income. ABC must find at most five groups in which customers share common behaviors.
-

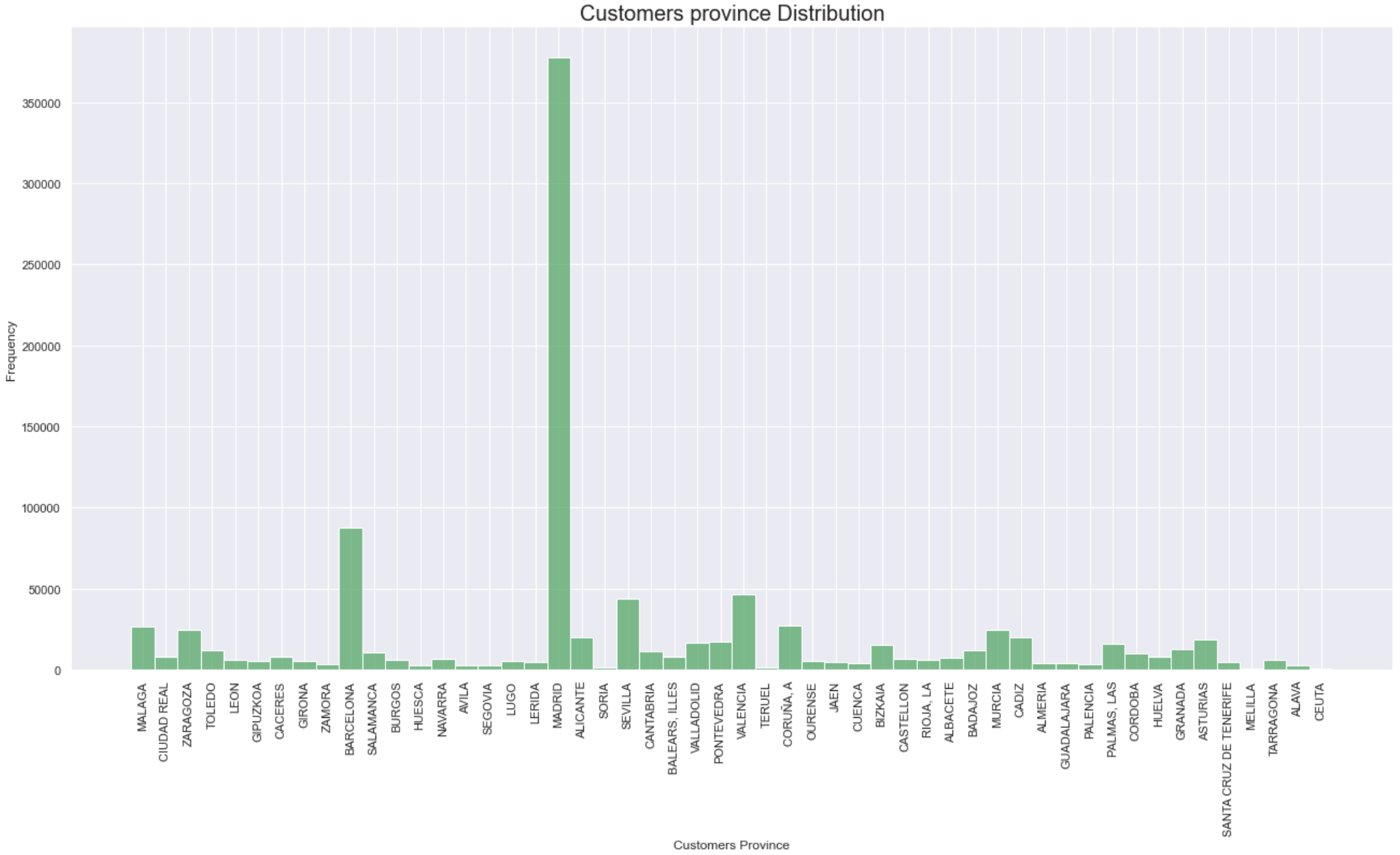
Project Description-Dataset

		fecha_dato	ncodpers	ind_empleado	pais_residencia	sexo	age	fecha_alta	ind_nuevo	antigueda	indrel	ult_fec_cli_1t	indrel_1mes	tiprel_1mes	indresi	indext	conyuemp
2	0	28/01/2015	1375586	N	ES	H	35	12/01/2015	0	6	1		1	A	S	N	
3	1	28/01/2015	1050611	N	ES	V	23	10/08/2012	0	35	1		1	I	S	S	
4	2	28/01/2015	1050612	N	ES	V	23	10/08/2012	0	35	1		1	I	S	N	
5	3	28/01/2015	1050613	N	ES	H	22	10/08/2012	0	35	1		1	I	S	N	
6	4	28/01/2015	1050614	N	ES	V	23	10/08/2012	0	35	1		1	A	S	N	
7	5	28/01/2015	1050615	N	ES	H	23	10/08/2012	0	35	1		1	I	S	N	
8	6	28/01/2015	1050616	N	ES	H	23	10/08/2012	0	35	1		1	I	S	N	
9	7	28/01/2015	1050617	N	ES	H	23	10/08/2012	0	35	1		1	A	S	N	
10	8	28/01/2015	1050619	N	ES	H	24	10/08/2012	0	35	1		1	I	S	N	
11	9	28/01/2015	1050620	N	ES	H	23	10/08/2012	0	35	1		1	I	S	N	
12	10	28/01/2015	1050621	N	ES	V	23	10/08/2012	0	35	1		1	I	S	N	
13	11	28/01/2015	1050622	N	ES	H	23	10/08/2012	0	35	1		1	I	S	N	
14	12	28/01/2015	1050623	N	ES	H	23	10/08/2012	0	35	1		1	A	S	N	
15	13	28/01/2015	1050624	N	ES	H	65	10/08/2012	0	35	1		1	A	S	N	

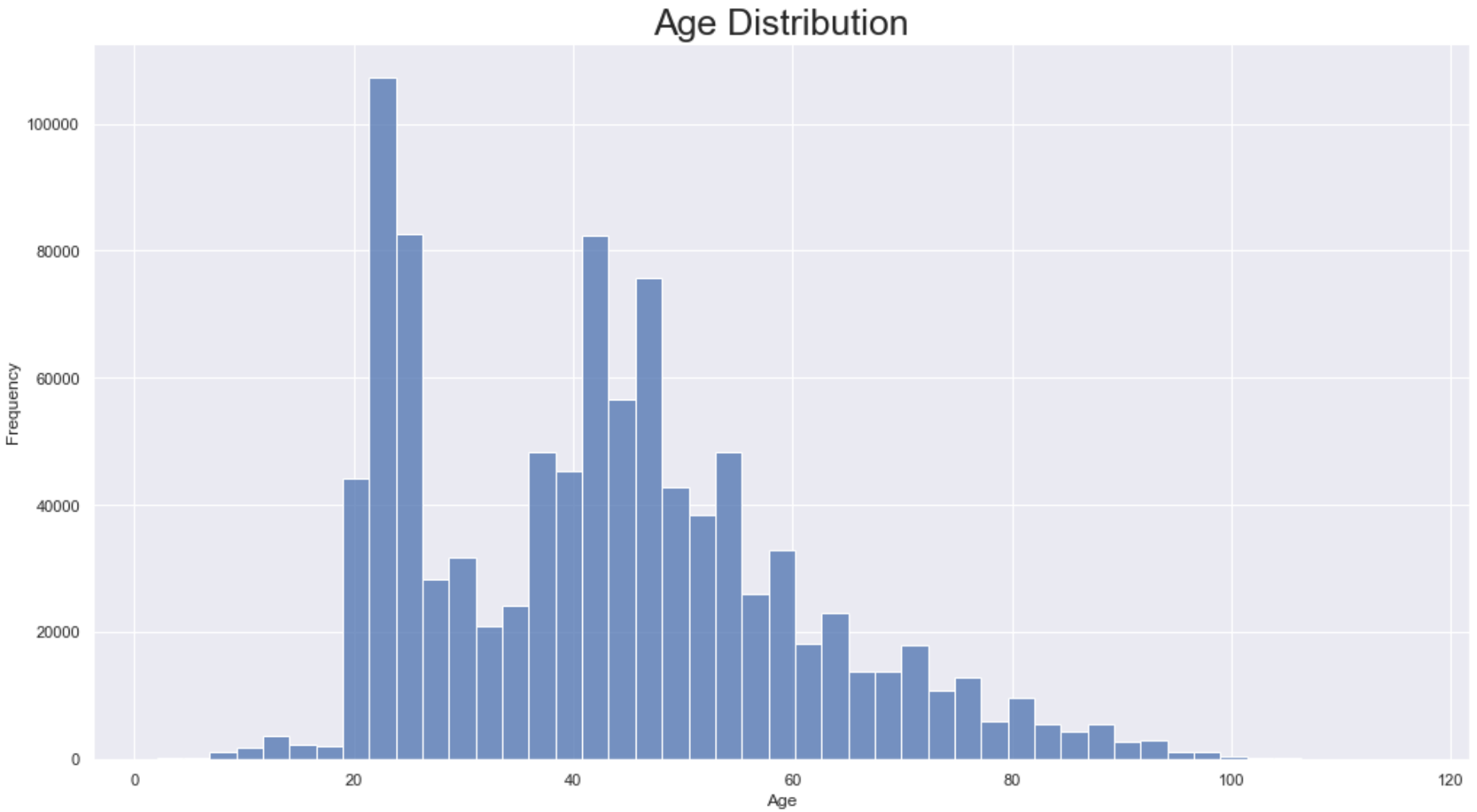
Project Description - Details of the Dataset

Total number of observations	1000000
Total number of files	1
Total number of features	48
Base format of the file	.csv
Size of the data	154 MB

EDA(Exploratory Data Analysis)



EDA Contd- AGE



EDA-Data Issues, Transformation and Cleaning

- We found that the issue with the original data set were:
- Data Incompleteness - various columns, such as the ind_empleado (Employee index), pais_residencia (Customer's Country residence), sexo (Customer's sex) etc., have empty values (NA) in certain records.
- Mitigation: Impute the missing values based on distribution in the training dataset. Depending on the context of the attribute, fill in the missing values with an appropriate value. If the variable is categorical, use the mode. If the variable is quantitative, use the median if discrete, and the mean if continuous.
- Recommendation: Make input for such fields required, especially easily obtainable ones such as customer sex. If information cannot be obtained, then do not include. Alternatively, may consider removing some attributes which are not so important.

EDA-Data Issues, Transformation and Cleaning

- Mitigation: Make data transformations to ensure consistent data types in each column. Examples: Convert selected records in strings to numeric, and remove non-numeric characters from the string.
- Recommendation: Ensure that fact tables in the given database have constraints on data types. Having different data types for a given field makes it difficult to interpret results at the later stage.
- Inconsistent values for the same attribute (e.g. -999999 as a value for seniority in months).
- Mitigation: The seniority in months cant have a negative number. Use regular positive numbers to replace extended values into positive numbers or zero to ensure consistency across seniority in months. Or replace -999999 with NA and perform the missing data transformations as above using the median
- Recommendation: Enforce a mandatory field where only positive integers are valid for this attribute.
-

IMPLEMENTATION-Different Data Models

- Model 1: unsupervised learning
- The model we have decided to use is based on the clustering principle. The model is called the k-means clustering, which divides the data into k clusters.
- **Results**

One way to cluster the customers is to do it based on the age and household income. The customers could be segmented by:

- High-income (above $250,000) customers
- Old age (above 60) customers with household income not exceeding $250,000
- Low income (under $70,000) customers with age not exceeding 60
- Other middle aged (35-60) customers
- Other young (under 35) customers

Another clustering scheme is to extract the seniority and household income into five groups:

- Customers with lower income (under $100,000) and less seniority (under 96 months)
- Customers with lower income (under $100,000) and more seniority (over 96 months)
- Customers with medium income (between $100,000 and $250,000) and less seniority (under 96 months)
- Customers with medium income (between $100,000 and $250,000) and more seniority (over 96 months)
- Customers with higher income (over $250,000)

Different Data Models

- **Model 2: supervised learning**
- **Random Forest + Bayesian Optimization**
- **Data should be preprocessed before modeling .The variables included are:**

ncodpers
ind_empleado
pais_residencia
sexo
age
ind_nuevo
antiguedad
indrel
indrel_1mes
tiprel_1mes
indresi
indext
conyuemp
canal_entrada
indfall
tipodom
cod_prov
nomprov
ind_actividad_cliente
renta
ind_ahor_fin_ult1
ind_aval_fin_ult1
ind_cco_fin_ult1
ind_cder_fin_ult1
ind_cno_fin_ult1

Different Data Models

- **Model 2: supervised learning**
- **Random Forest + Bayesian Optimization**
- **Data should be preprocessed before modeling. The variables included are:**

Random Forest is a method used for classification or regression. It builds several decision trees at training time, fed with randomly selected samples of the database. This step is called bootstrapping. All trees give their own output (a class for classification or a value for regression) based on their fed datathe data they were fed with.] Then, a vote between all the individual returns gives the final output as the majority or the mean. This step is called bagging. This way, Random Forests adapt decision trees such that it removes overfitting to the training set. The process is repeated several times until stabilization. This implements the principle of multiple weak learners being better as a group. It is unexcelled in accuracy among current algorithms. This algorithm estimates what variables are important in the classification. It could also generate an internal unbiased approximation of the generalization error as the forest building advances while effectively estimating missing data and keeps accuracy when a large proportion of the data are missing. This could help for balancing error in class population unbalanced data sets. The other features of this method are: it offers an experimental method for finding variable interactions; it computes closeness between pairs of cases that can be used in clustering, locating outliers; prototypes are computed that give information about the correlation between the variables and the classification; the capacities mentioned above is also usable for unlabelled data, that leads to unsupervised clustering, data views and outliers detection.