Northeastern University

Analyzing Marketing campaigns of a Portuguese banking Institution

Team X:

Anuj Khanna
Hang Wu
Ivan Todorov
Fatima Nurmakhamadova
Supreeth Murugesh

Submitted to:

Prof. Justin Grosz,
College of Professional Studies,
Northeastern University

Course:

ALY6040 - Data Mining

Agenda

1. Dataset Introduction
2. Data Cleaning
3. EDA & Visualization
4. Modeling
5. Benchmarking
6. Recommendations

Dataset Overview

- Derived from a bank in Portugal that describes the results from the organization's marketing campaigns, proposed by the phone calls.
- Dataset Info:
  - **41,188** data points - Phone calls made from May 2008 to November 2010.
  - **20** features including Social and Economic attributes such as Age, Gender, Education, employment variation rate, consumer confidence index, outcome of the previous marketing campaign and more.
- Our target variables is whether **client subscribed a term deposit** (Yes/No).

Project Goal: Providing a recommendation to the Portuguese Bank about campaign improvements to ultimately bring additional revenue to the bank.

Data Cleaning

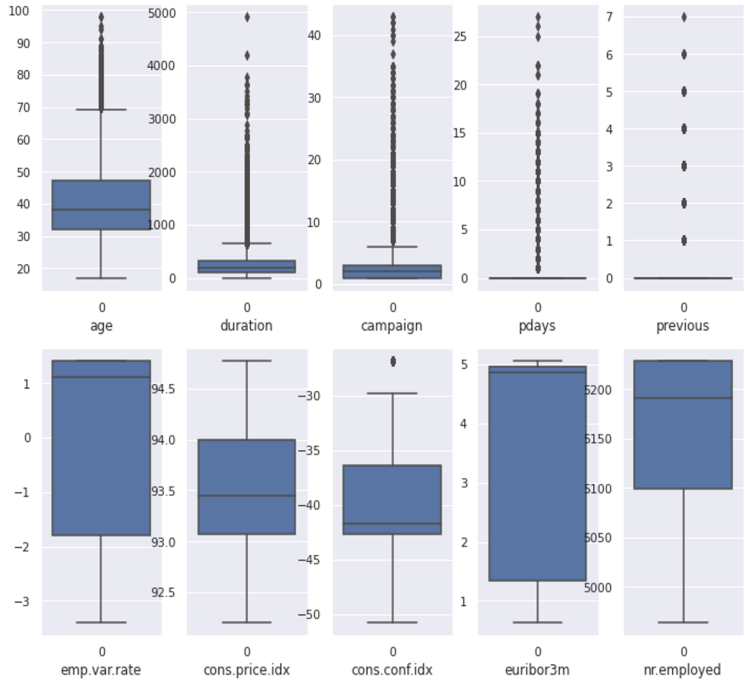
Data type:

```
age                int64
job             category
marital         category
education       category
default         category
housing         category
loan            category
contact         category
month           category
day_of_week     category
duration           int64
campaign           int64
pdays              int64
previous           int64
poutcome        category
emp.var.rate     float64
cons.price.idx   float64
cons.conf.idx    float64
euribor3m        float64
nr.employed      float64
client_decision category
dtype: object
```

Missing values:

```
Column Name: job 0.8%
Column Name: marital 0.19%
Column Name: education 4.2%
Column Name: default 20.87%
Column Name: housing 2.4%
Column Name: loan 2.4%
```

Boxplot for the numerical variables

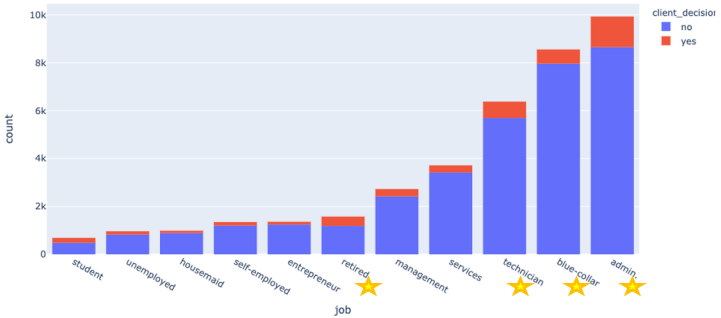


- Converting objects to categorical type
- Deleting ‘unknown’ values ≤ 5%, replace with mode values ≤30%
- The outliers are in total 12,875 rows

Data Visualization

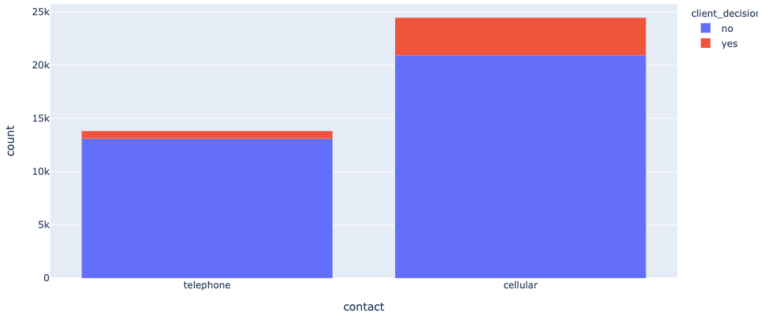
Graph 1:

Job of the Customers by Client Decision

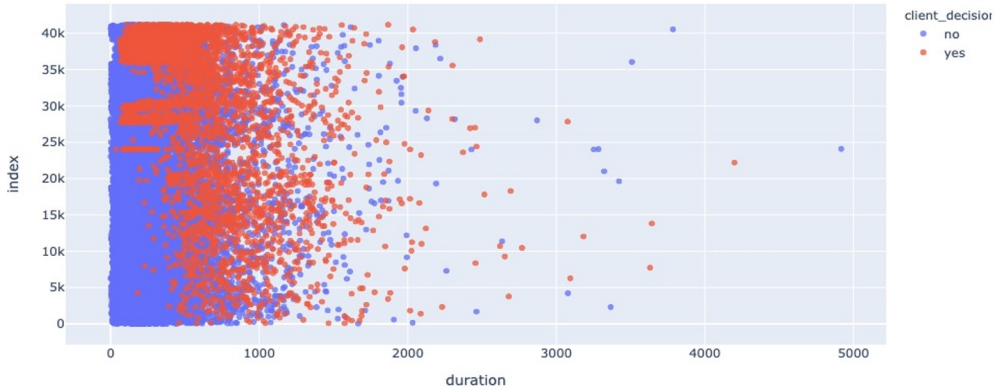


Graph 2:

The Contact Communication Types



Graph 3: Effects on client decision based on call duration



Logistic Regression

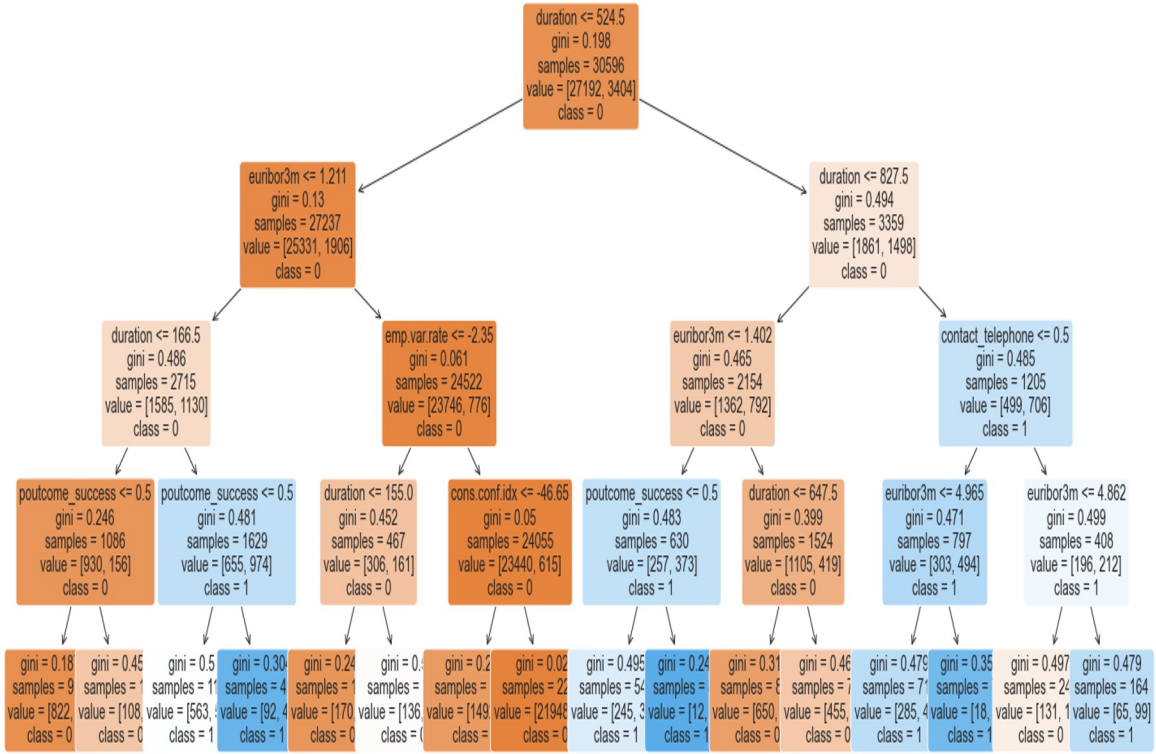
- **Poutcome_success**, it is highly likely that the client is likely to open a deposit account with the bank again.
- Model Overall Accuracy : **91.16%**
  - Precision: **98%**
  - Recall: **38%**

Logit Regression Results

Dep. Variable:	client_decision	No. Observations:	30596
Model:	Logit	Df Residuals:	30551
Method:	MLE	Df Model:	44
Date:	Sat, 07 May 2022	Pseudo R-squ.:	0.3956
Time:	19:58:14	Log-Likelihood:	-6456.2
converged:	False	LL-Null:	-10682.
Covariance Type:	nonrobust	LLR p-value:	0.000

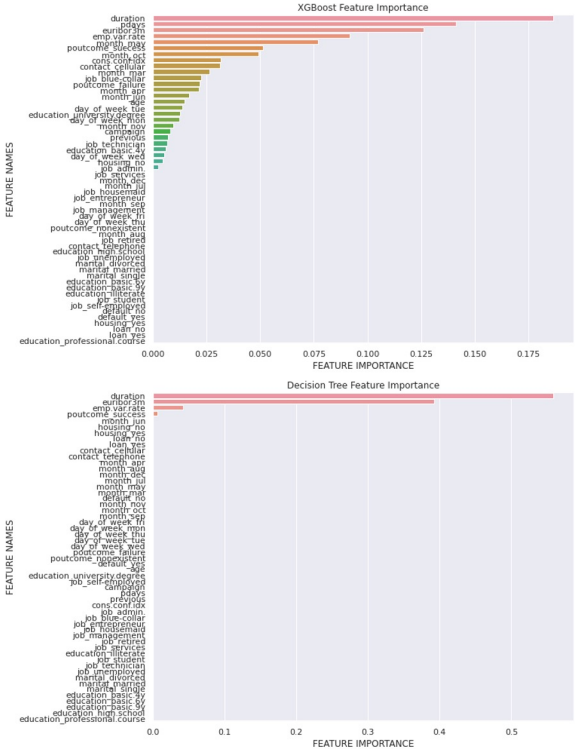
	coef	std err	z	P>|z|	[0.025	0.975]
age	-0.0036	0.003	-1.362	0.173	-0.009	0.002
duration	0.0047	8.6e-05	54.968	0.000	0.005	0.005
campaign	-0.0430	0.013	-3.236	0.001	-0.069	-0.017
pdays	0.0138	0.016	0.860	0.390	-0.018	0.045
previous	0.1245	0.063	1.969	0.049	0.001	0.248
emp.var.rate	0.1665	0.043	3.890	0.000	0.083	0.250
cons.conf.idx	0.0420	0.004	11.017	0.000	0.035	0.049
euribor3m	-0.7785	0.044	-17.637	0.000	-0.865	-0.692
job_blue-collar	-0.3155	0.092	-3.419	0.001	-0.496	-0.135
job_entrepreneur	-0.2699	0.146	-1.844	0.065	-0.557	0.017
job_housemaid	-0.0476	0.168	-0.283	0.777	-0.377	0.282
job_management	-0.1011	0.098	-1.035	0.301	-0.293	0.090
job_retired	0.3137	0.124	2.539	0.011	0.072	0.556
job_self-employed	-0.1722	0.136	-1.269	0.204	-0.438	0.094
job_services	-0.1437	0.098	-1.463	0.144	-0.336	0.049
job_student	0.1968	0.135	1.462	0.144	-0.067	0.461
job_technician	-0.0167	0.081	-0.206	0.837	-0.176	0.142
job_unemployed	-0.0249	0.144	-0.174	0.862	-0.306	0.257
marital_married	-0.0409	0.077	-0.531	0.596	-0.192	0.110
marital_single	0.0094	0.087	0.108	0.914	-0.161	0.180
education_basic.6y	0.0813	0.135	0.602	0.547	-0.183	0.346
education_basic.9y	-0.0092	0.106	-0.087	0.931	-0.217	0.199
education_high.school	-0.0279	0.102	-0.272	0.786	-0.229	0.173
education_illiterate	0.5136	0.844	0.608	0.543	-1.141	2.169
education_professional.course	0.0901	0.113	0.799	0.424	-0.131	0.311
education_university.degree	0.1797	0.102	1.762	0.078	-0.020	0.379
default_yes	-15.6749	1.2e+04	-0.001	0.999	-2.35e+04	2.35e+04
housing_yes	-0.0179	0.047	-0.381	0.703	-0.110	0.074
loan_yes	-0.0752	0.065	-1.151	0.250	-0.203	0.053
contact_telephone	-0.1996	0.073	-2.729	0.006	-0.343	-0.056
month_aug	0.0550	0.107	0.512	0.608	-0.155	0.266
month_dec	0.2781	0.228	1.220	0.223	-0.169	0.725
month_jul	0.3838	0.103	3.722	0.000	0.182	0.586
month_jun	0.5127	0.104	4.943	0.000	0.309	0.716
month_mar	1.5138	0.136	11.167	0.000	1.248	1.779
month_may	-0.7784	0.083	-9.327	0.000	-0.942	-0.615
month_nov	0.0557	0.116	0.480	0.631	-0.172	0.283
month_oct	0.4290	0.136	3.153	0.002	0.162	0.696
month_sep	0.0431	0.141	0.306	0.760	-0.233	0.319
day_of_week_mon	-0.1040	0.076	-1.367	0.172	-0.253	0.045
day_of_week_thu	0.0420	0.073	0.573	0.567	-0.102	0.186
day_of_week_tue	0.1017	0.076	1.343	0.179	-0.047	0.250
day_of_week_wed	0.1706	0.076	2.260	0.024	0.023	0.319
poutcome_nonexistent	0.6144	0.107	5.746	0.000	0.405	0.824
poutcome_success	1.8206	0.125	14.540	0.000	1.575	2.066

Decision Tree



- **Duration**, **Poutcome_success** and **euribor3m**, are the important features which can be the deciding factor for customers to open a term deposit with the bank.
- Model Overall Accuracy: **91.18%**
  - Precision: **95%**
  - Recall: **56%**

Benchmarking: Feature Importance

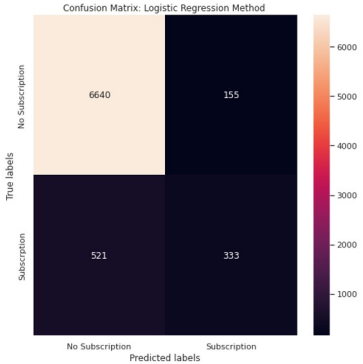


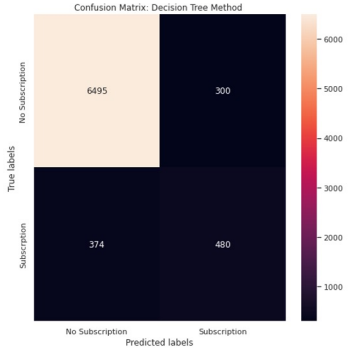
More Features have weights than the decision tree FI, this means that xgBoost can effectively reduce **overfit and high sensitivity**, and indicating the exact importance ranking between features.

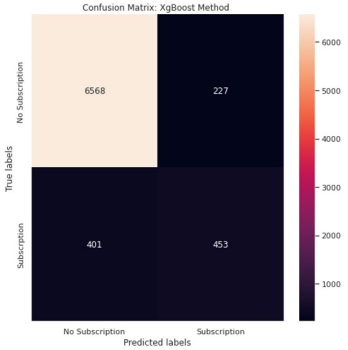
LR, as mentioned before, still introduce huge **sensitivity**.

Duration has been the most importance over all other predictors despite being the most volatile. Social-economic predictors ranks lower than duration (Eribor interest rate for 3 months, employment variance rate for 1 month), the best non social economic predictor is **month May**.

Benchmarking: Confusion Matrix and Benchmark Table







Models	Accuracy (%)	MSE	Precision	Recall	Time (s)
Logistic Regression	91	0.088	93	38	1.2
Decision Tree	91	0.088	95	56	0.1
XgBoost	92	0.082	94	53	2.04

- Decision Tree works well on minimizing FN (**374**) and maximizing TP (**480**).
- Logistic Regression works well on minimizing FP (**155**) and maximizing TN (**6640**).
- XgBoost (30 trees, 0.5 learning rate) has the best Overall Accuracy(**92%) and MSE(0.082)** , making it the best model of the 3, especially when both **FP** and **FN** should be minimized, neither **LR** nor **DT** were doing better.

Conclusion

Q: What factors affect the customers decision?

- Duration
- Euribor rate
- Contact mode (telephone)
- Previously contact individuals

Recommendations

- Change the contact communication type solely to cellular
- Set a call duration limit from 3 to 10 minutes
- Build the relationship and discover the customers that agreed to place deposits in the previous campaigns
- Both the European interest rate and financial confidence are the most economic-social factors that affect the personal decisions