Overview

- The likelihood that someone will take a loan in their lifetime is almost a guarantee in today's society.

- What are the chances, that a customer will subscribe to a loan ?

Background

- Obrigado Bank is a Portuguese banking institution, with locations all over Portugal and Europe.

- Term deposits allow banks to hold onto a deposit for a specific amount of time, banks can invest in higher gain financial products to make a profit.

- Banks hold better chances to persuade term deposit clients into buying other products to increase their revenues.

Goals

- Create and implement a machine learning model that will classify customers that are more likely to subscribe to a loan.
- The model will have an AUC score higher than 90% and an FP score less than 10%.
- Find features any feature that may be significant with regards to the outcome of the model and the campaign.

Dataset

- Data comes from UCI Machine Learning Repository
- Instances : 45211
- Attributes/Features: 17

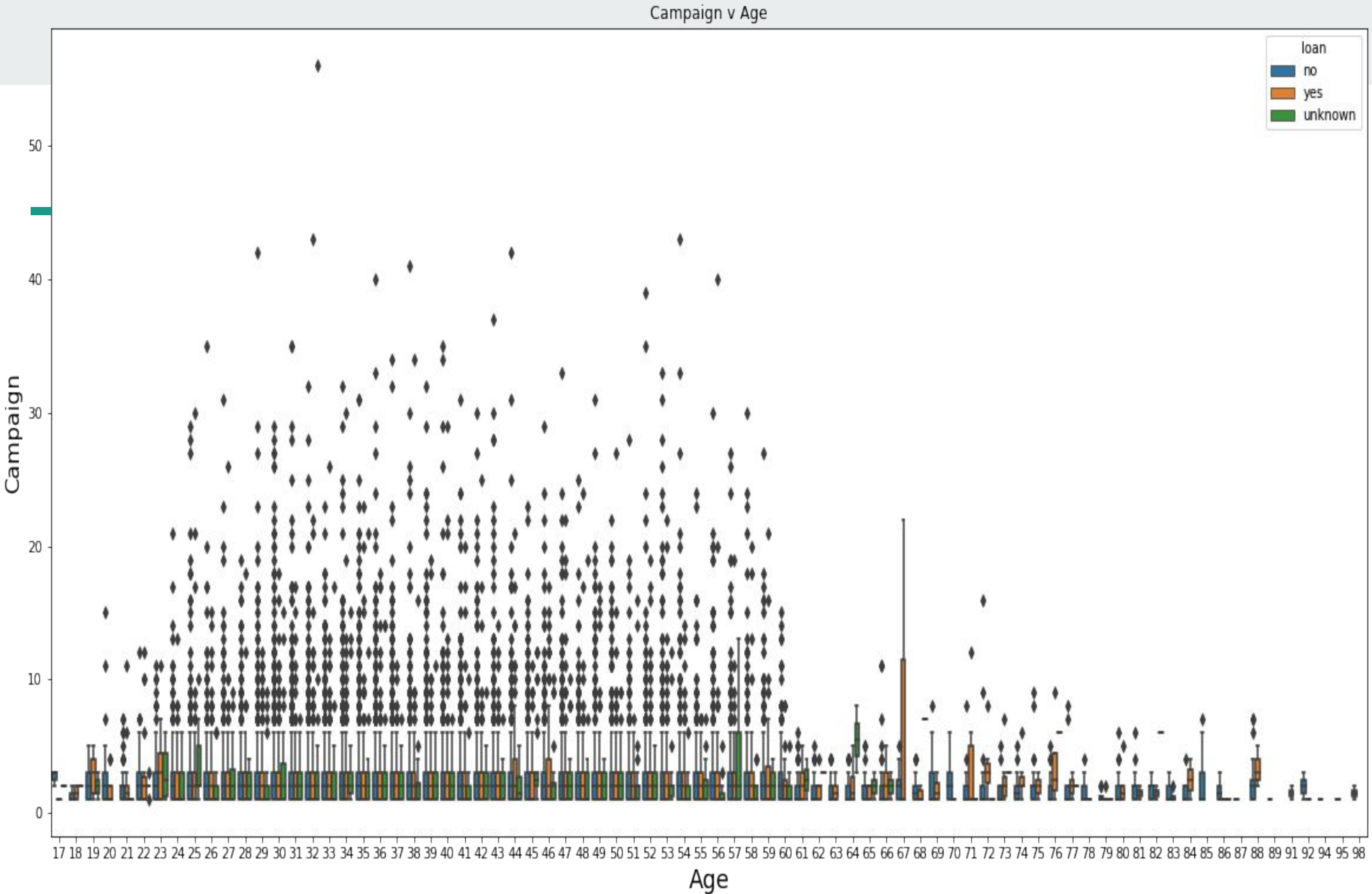
Data Wrangling

- The data came clean so there was no data wrangling needed.

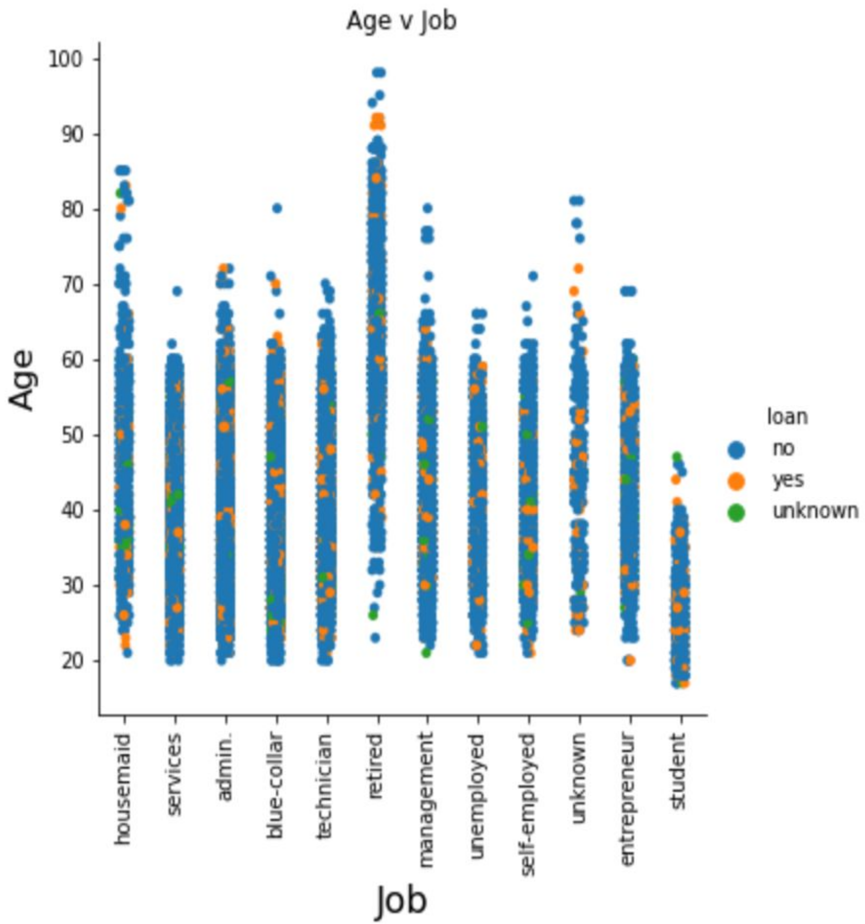
Exploratory Data Analysis

- Does being married play a factor in whether a person has a loan or not ?
- Does having a specific profession or career factor in whether someone has a loan?
- Does age play a big factor in who will take a loan ?

Age

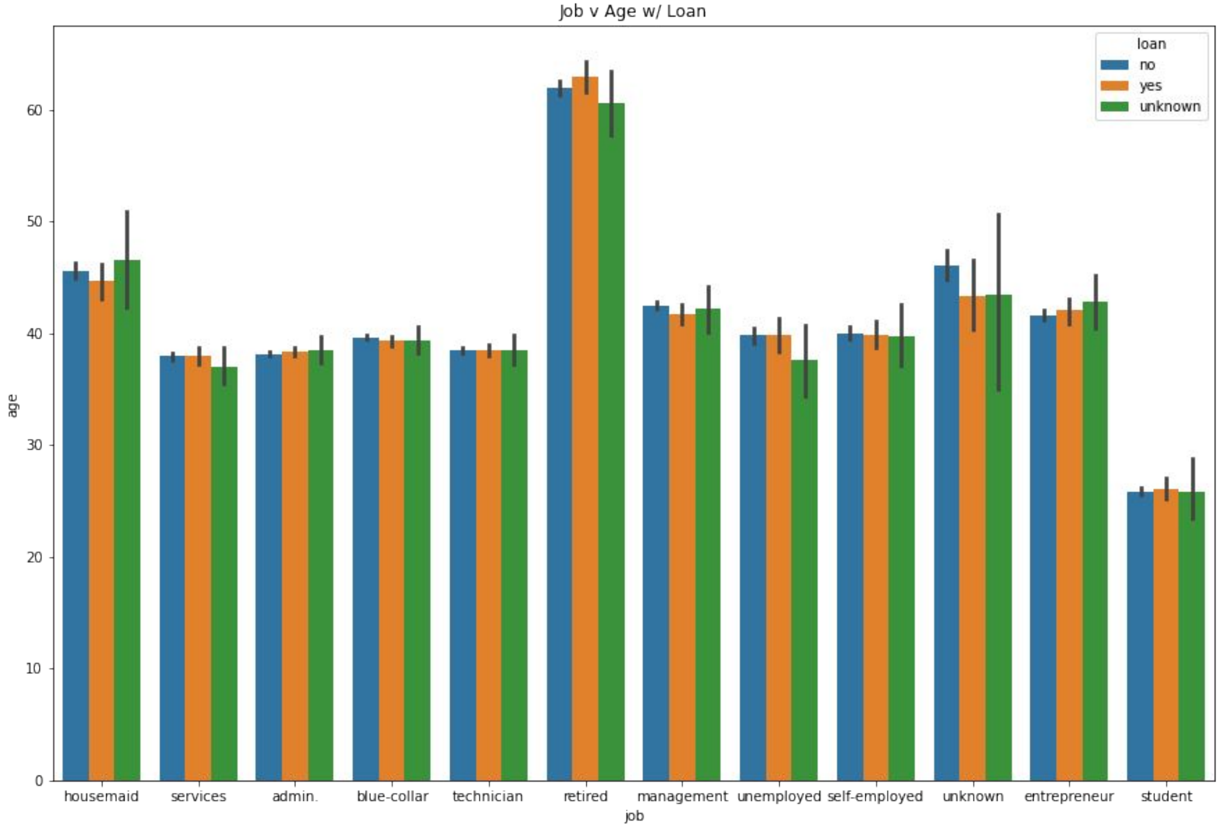

Campaign v Age
Campaign
Age
loan
no
yes
unknown
0
10
20
30
40
50
17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 91 92 94 95 98

Continued

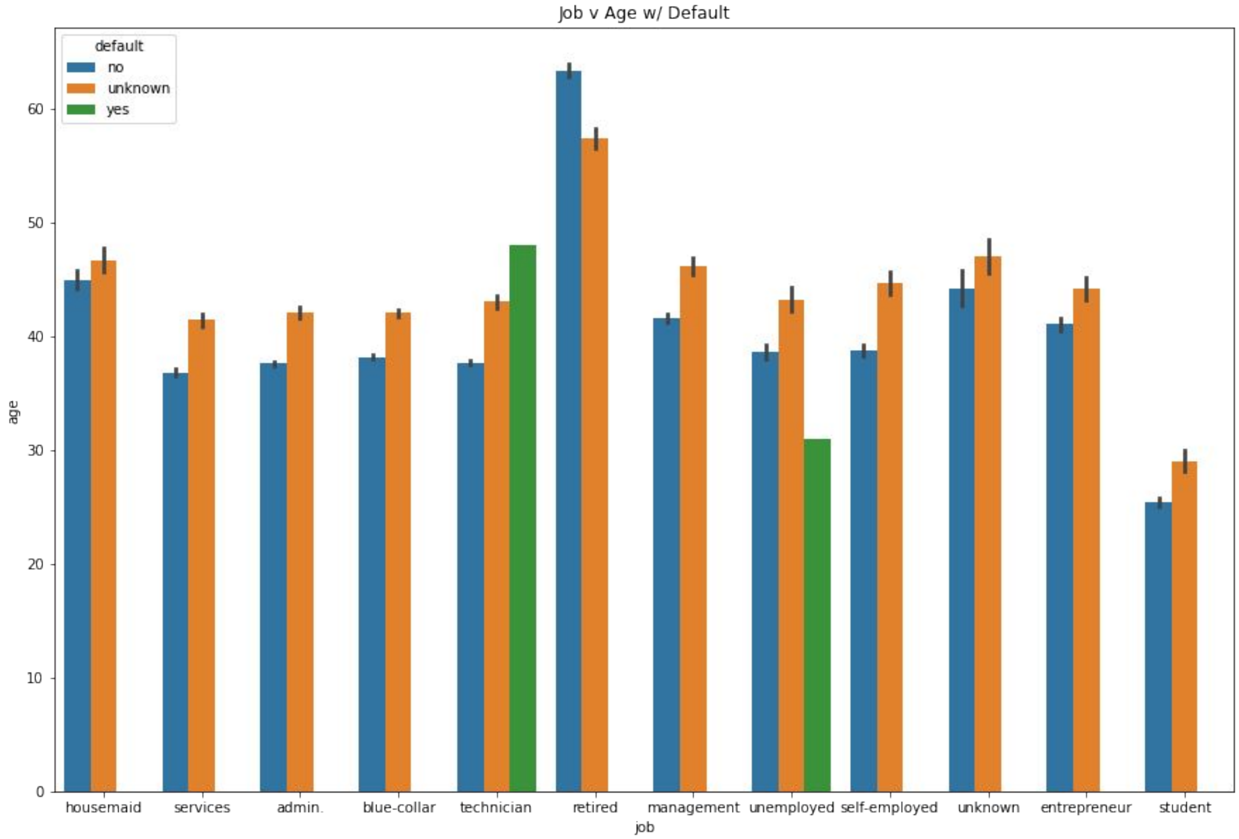


Age v Job
100
90
80
70
60
50
40
30
20
Age
housemaid
services
admin.
blue-collar
technician
retired
management
unemployed
self-employed
unknown
entrepreneur
student
Job
loan
no
yes
unknown

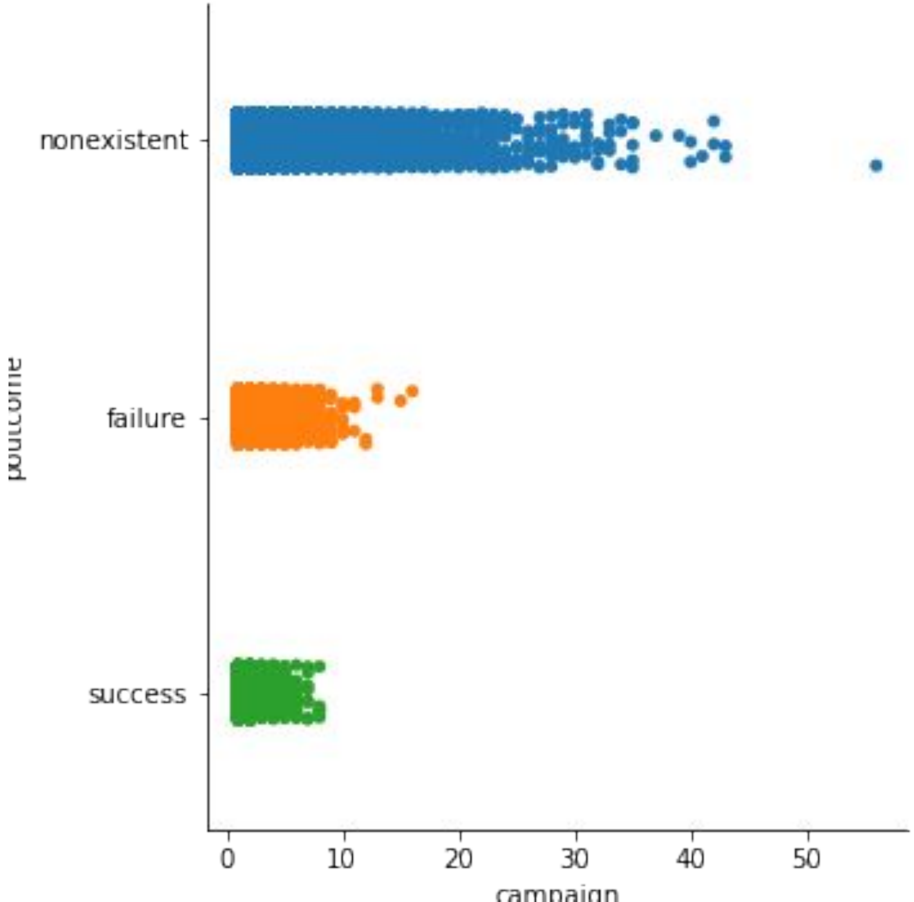
Occupation


Job v Age w/ Loan
loan
no
yes
unknown
0
10
20
30
40
50
60
age
housemaid
services
admin.
blue-collar
technician
retired
management
unemployed
self-employed
unknown
entrepreneur
student
job

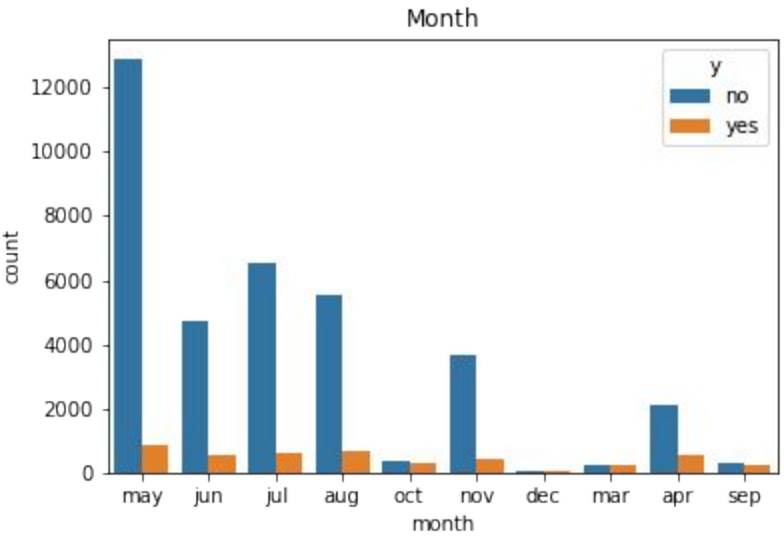
Continued


Job v Age w/ Default
default
no
unknown
yes
age
0
10
20
30
40
50
60
housemaid
services
admin.
blue-collar
technician
retired
management
unemployed
self-employed
unknown
entrepreneur
student
job

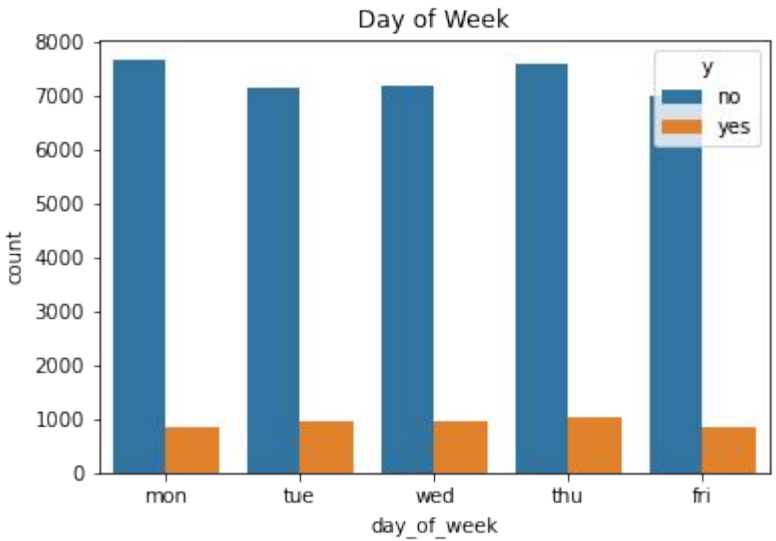
Campaign Performance



Continued

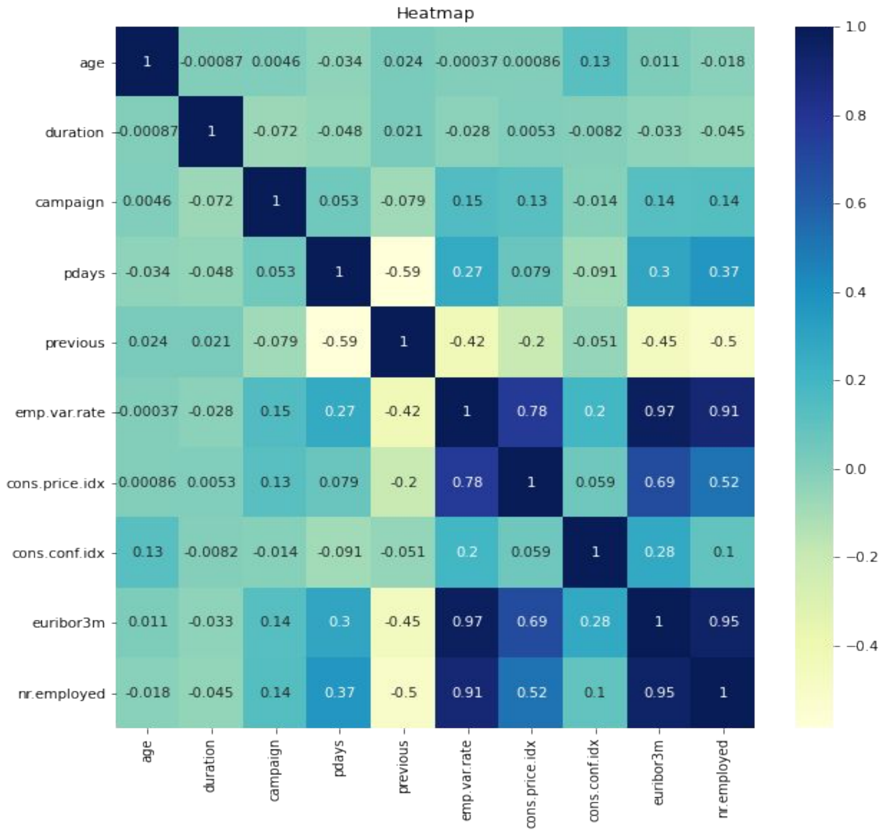


Month
y
no
yes
count
12000
10000
8000
6000
4000
2000
0
may
jun
jul
aug
oct
nov
dec
mar
apr
sep
month



Day of Week
y
no
yes
count
8000
7000
6000
5000
4000
3000
2000
1000
0
mon
tue
wed
thu
fri
day_of_week

Heatmap



Heatmap
age
duration
campaign
pdays
previous
emp.var.rate
cons.price.idx
cons.conf.idx
euribor3m
nr.employed
1 -0.00087 0.0046 -0.034 0.024 -0.00037 0.00086 0.13 0.011 -0.018
-0.00087 1 -0.072 -0.048 0.021 -0.028 0.0053 -0.0082 -0.033 -0.045
0.0046 -0.072 1 0.053 -0.079 0.15 0.13 -0.014 0.14 0.14
-0.034 -0.048 0.053 1 -0.59 0.27 0.079 -0.091 0.3 0.37
0.024 0.021 -0.079 -0.59 1 -0.42 -0.2 -0.051 -0.45 -0.5
-0.00037 -0.028 0.15 0.27 -0.42 1 0.78 0.2 0.97 0.91
0.00086 0.0053 0.13 0.079 -0.2 0.78 1 0.059 0.69 0.52
0.13 -0.0082 -0.014 -0.091 -0.051 0.2 0.059 1 0.28 0.1
0.011 -0.033 0.14 0.3 -0.45 0.97 0.69 0.28 1 0.95
-0.018 -0.045 0.14 0.37 -0.5 0.91 0.52 0.1 0.95 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Feature Engineering

- First, we removed any features with correlation > 80%
- Next we transformed our categorical data using Onehotencoding

Handling Categorical Data

Because most of the variables we are working with are categorical we will be using 1 hot encoding to satisfy the mathmatical equations that our model will be solving for

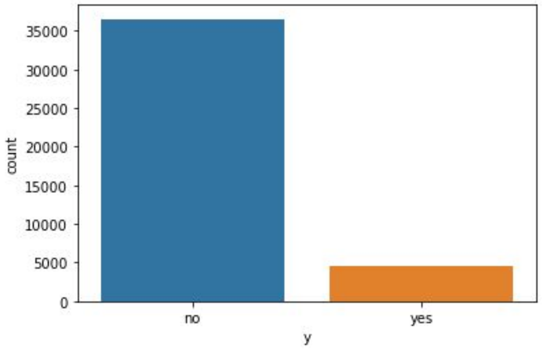
In [21]:
```
columns=df.select_dtypes(include=[object]).columns
df=pd.concat([df,pd.get_dummies(df[columns])],axis=1)
df=df.drop(['job','marital','education','default','housing','loan','contact','month','day_of_week','poutcome'],axis=
df.info()
df.head()
```

```
Int64Index: 41188 entries, 0 to 41187
Data columns (total 63 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   age                             41188 non-null  int64
 1   campaign                        41188 non-null  int64
 2   pdays                           41188 non-null  int64
 3   previous                        41188 non-null  int64
 4   cons.price.idx                  41188 non-null  float64
 5   cons.conf.idx                   41188 non-null  float64
 6   nr.employed                     41188 non-null  float64
 7   y                               41188 non-null  object
 8   job_admin.                      41188 non-null  uint8
 9   job_blue-collar                 41188 non-null  uint8
 10  job_entrepreneur                41188 non-null  uint8
 11  job_housemaid                   41188 non-null  uint8
 12  job_management                  41188 non-null  uint8
 13  job_retired                     41188 non-null  uint8
```

Continued

Taking care of imbalanced data

Before



After

```
Before oversampling:  Counter({'no': 36548, 'yes': 4640})
After oversampling:  Counter({'no': 36548, 'yes': 36548})
```

Machine Learning

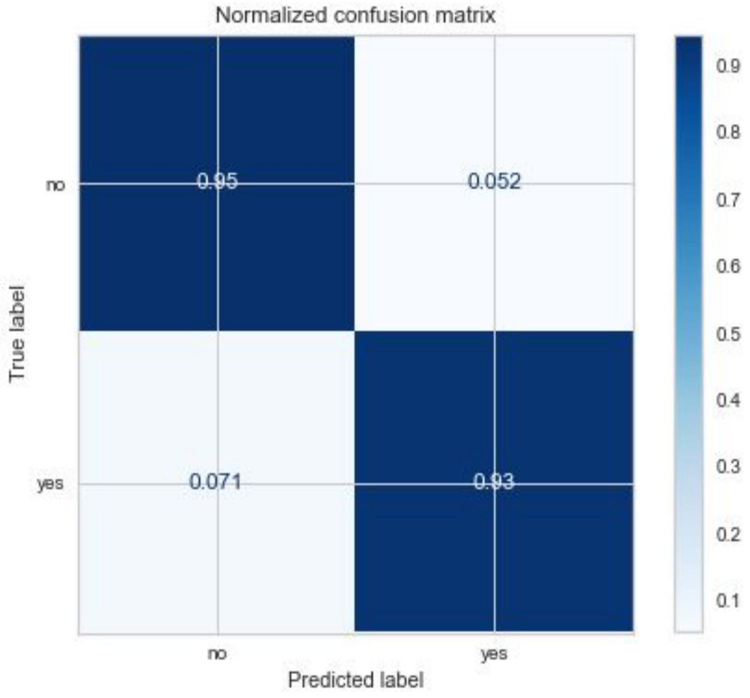
Models we are using are :

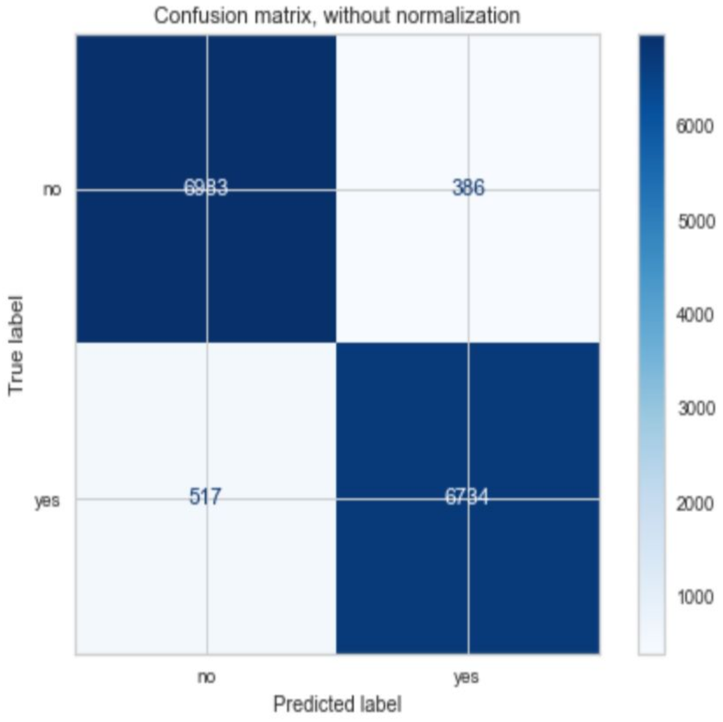
- KNN
- SVR
- Logistic Regression
- Random Forest
- XGBoost

Performance

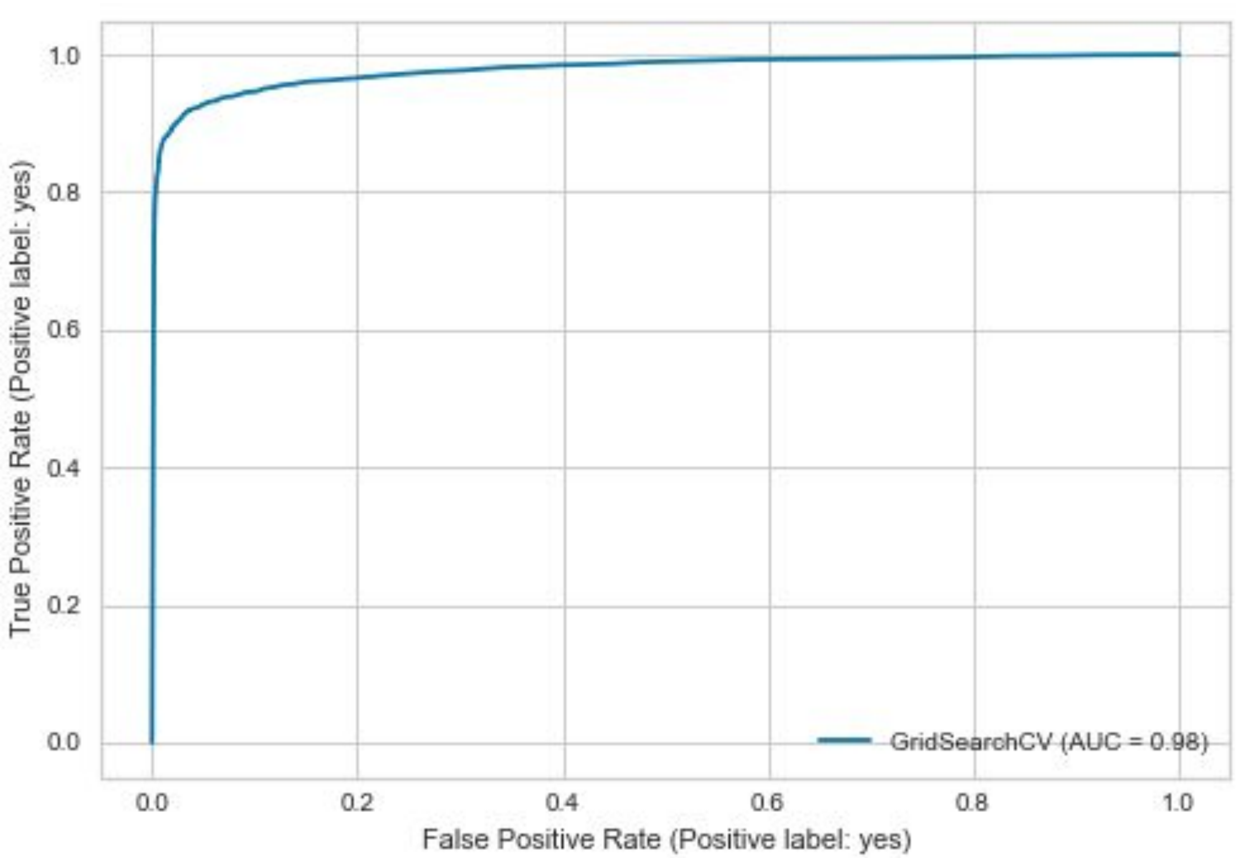
Model	FP%	FN%	AUC%
KNN	7.9	7.3	96
LR	11	1.9	97
SVR	11	2.3	97
RF	7.1	5.2	98
XGBoost	9.2	2.9	97

Random Forest Confusion Matrix



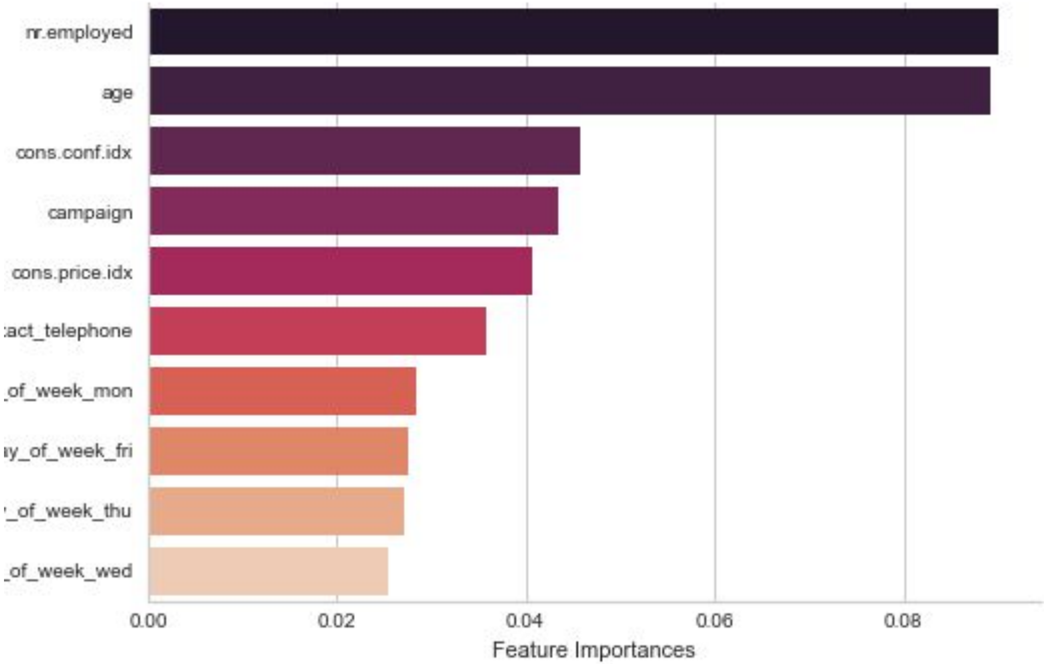


LR ROC



1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate (Positive label: yes)
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate (Positive label: yes)
GridSearchCV (AUC = 0.98)

Feature Importance



Recommendations

- Focus on calling customers Tues,Wed,Thur
- Drop any customer that hasn't subscribed in at least 15 days or 8 call attempts.
- Keep employees working and happy, any drop in workforce affects the campaign.
- Segment customers by age and incorporate lifestyle segmentation
- Focus on targeting entrepreneurs.Target the management occupation to help drive volume.

Takeaways

- April-August produces the highest results for the campaign.
- Age is a factor as we can see from the feature importances chart and heatmap
- Any chance of success will happen within 10 days of direct marketing to a customer; after 10 days there were no successful attempts.
- Employees carry high significance, with regards to the feature importances chart, with how many employees are there to make the calls

Future work

- Include time
- Include more personal data such as loan balance, account balance, debt to income ratio
- Include geographic location