Exploratory Data Analysis

<ABC Bank>

<14/05/2022>

Group information

Group name : Koro
Specialization : Data science

Name	Email	Country	College
Muhammed KURNASAN	mohamedkornasan@hotmail.com	Turkey	Turkish-German University

Data Glacier

Github repo : https://github.com/kurn3san/Koro_DG_Internship_Project

Case Study

Problem : ABC Bank wants to sell it's term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

Task : We are analyzing the Data related to banks to help ABC make the right decision before launching their product

After understanding the problem, The work was made through 4 steps :

1. Gaining insights from the Data
2. Data preparation
3. Data Visualization
4. Conclusion and ML model Recommendations

Insights

We have a total of 21 columns (5 of them float64 , 5 Int64 and the rest are objects) without any Null values.

#	Column	Non-Null Count	Dtype
0	age	41188 non-null	int64
1	job	41188 non-null	object
2	marital	41188 non-null	object
3	education	41188 non-null	object
4	default	41188 non-null	object
5	housing	41188 non-null	object
6	loan	41188 non-null	object
7	contact	41188 non-null	object
8	month	41188 non-null	object
9	day_of_week	41188 non-null	object
10	duration	41188 non-null	int64
11	campaign	41188 non-null	int64
12	pdays	41188 non-null	int64
13	previous	41188 non-null	int64
14	poutcome	41188 non-null	object
15	emp.var.rate	41188 non-null	float64
16	cons.price.idx	41188 non-null	float64
17	cons.conf.idx	41188 non-null	float64
18	euribor3m	41188 non-null	float64
19	nr.employed	41188 non-null	float64
20	y	41188 non-null	object

Data preparation

1- Data preparation phase:

- Checking for NAN values in Data and dealing with it (already checked)

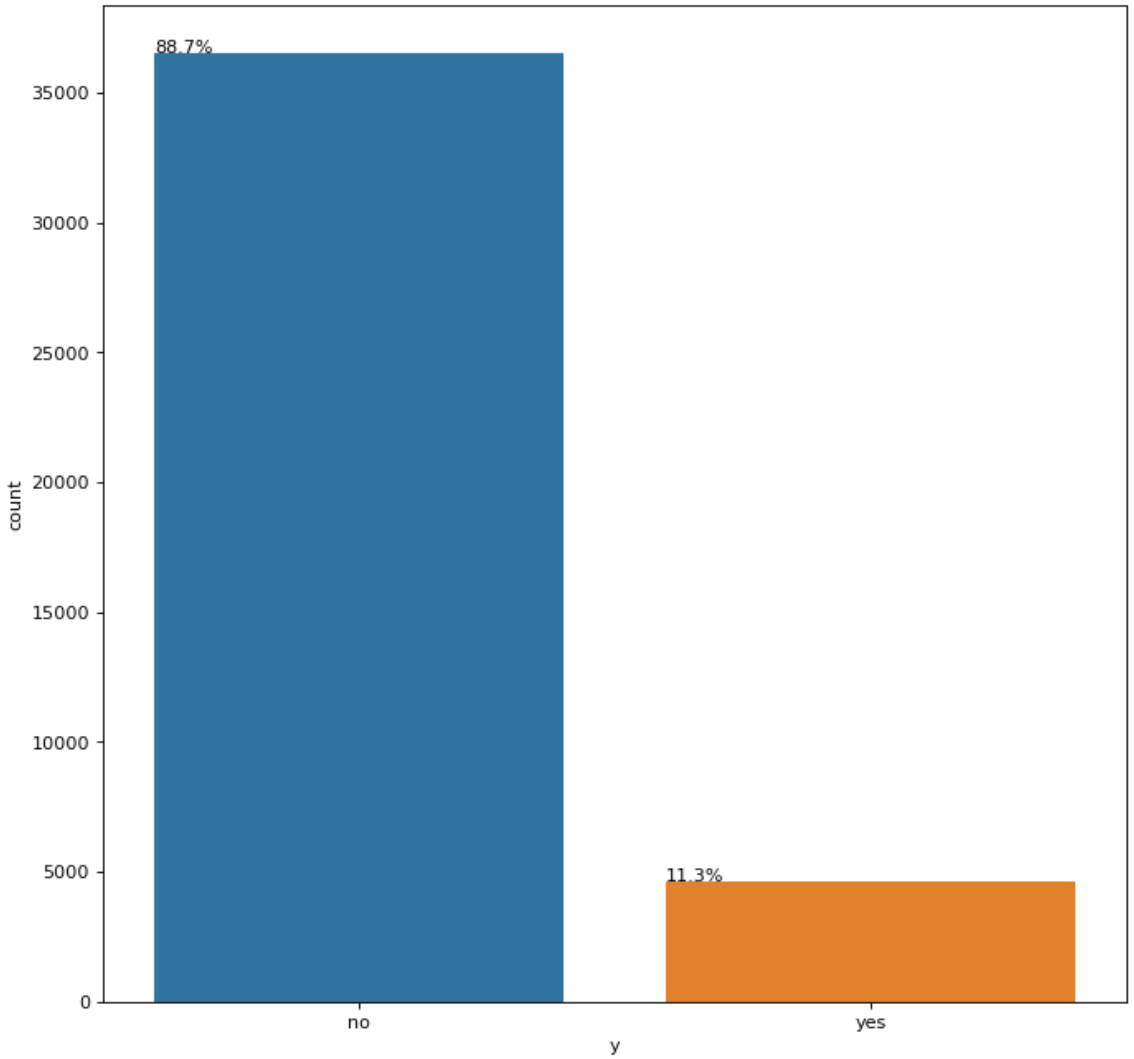
- Checking whether we have imbalanced data or not

- Dealing with imbalanced data (if found)

Data preparation

The Data we have contains 41188 Row and 21 Columns but the target data (y column) is unfortunately unequally distributed as shown :

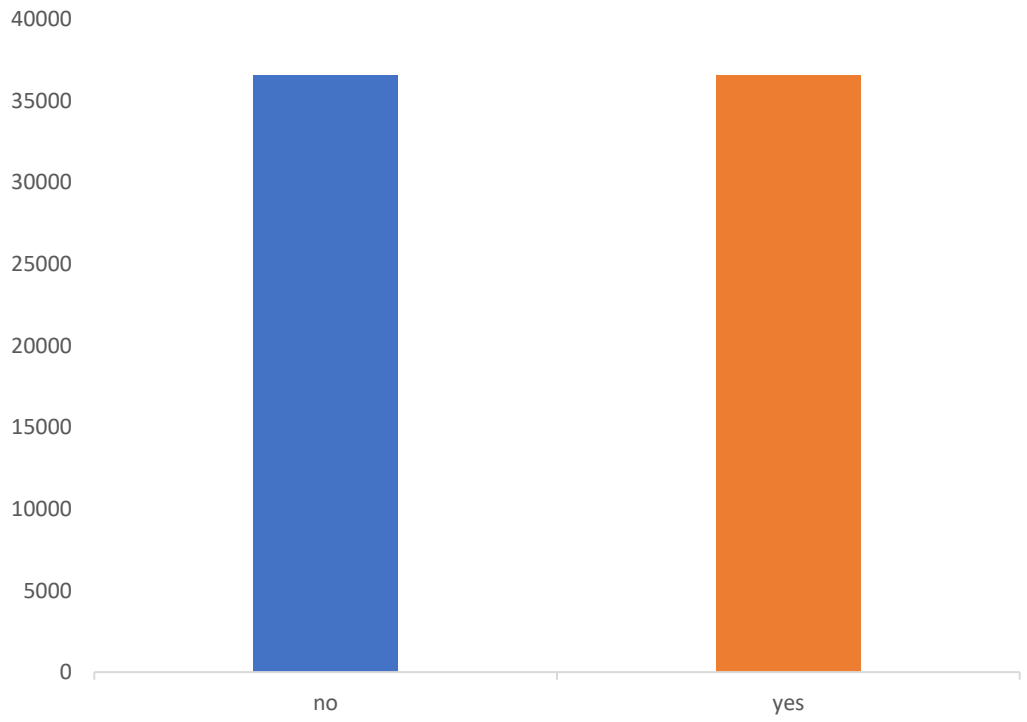
We can clearly see that more than 88% of the data is a No and only a small portion (10%) is yes



Data preparation

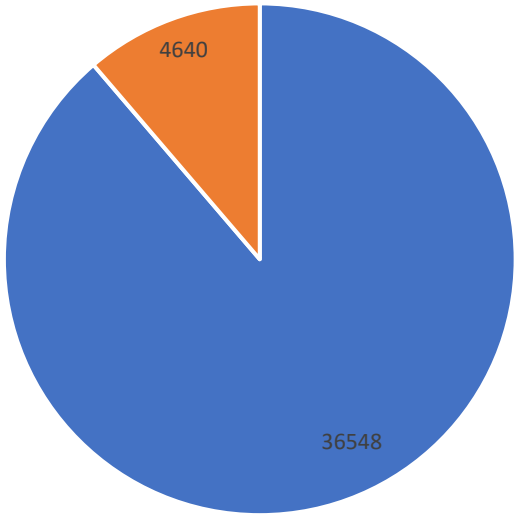
To overcome this problem we tried many methods including SMOTE and the results are phenomenal!

We managed to make the data 100% balanced by creating synthetic instances and now we have a total of 73096 rows (more data = better training)

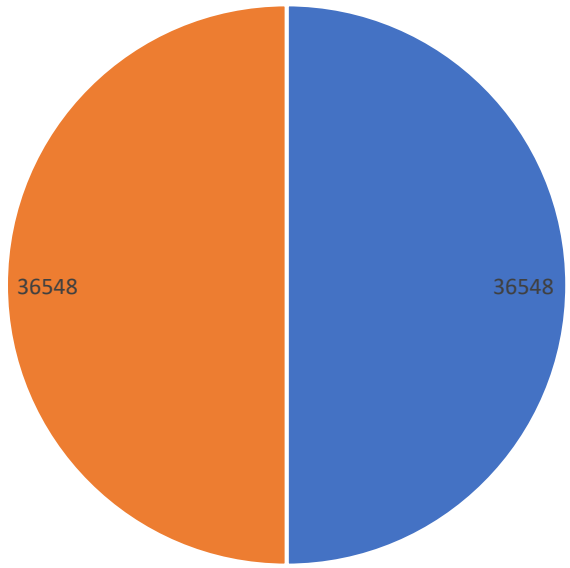


Data preparation

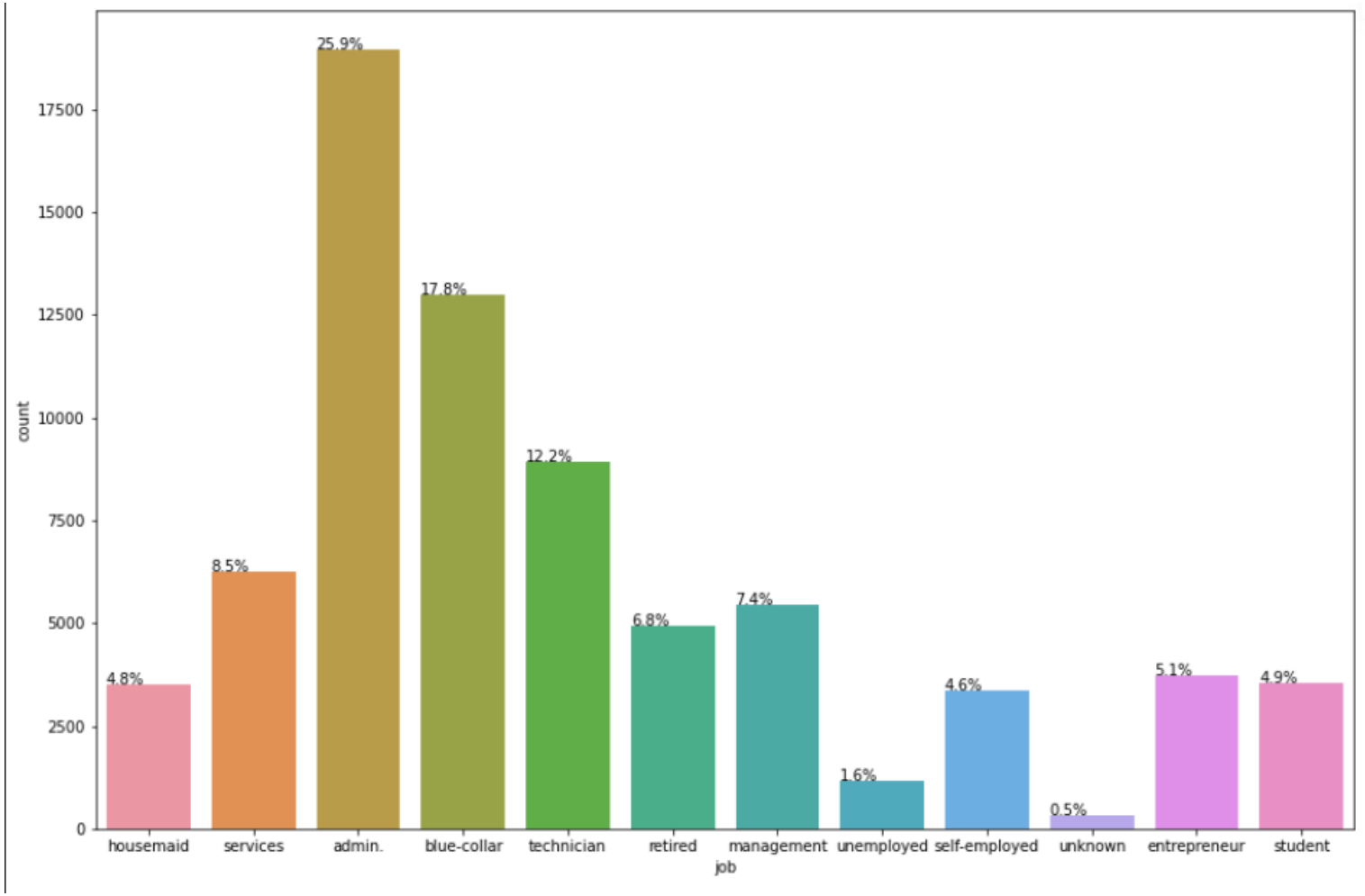
The data before applying SMOTE method :



The data after :



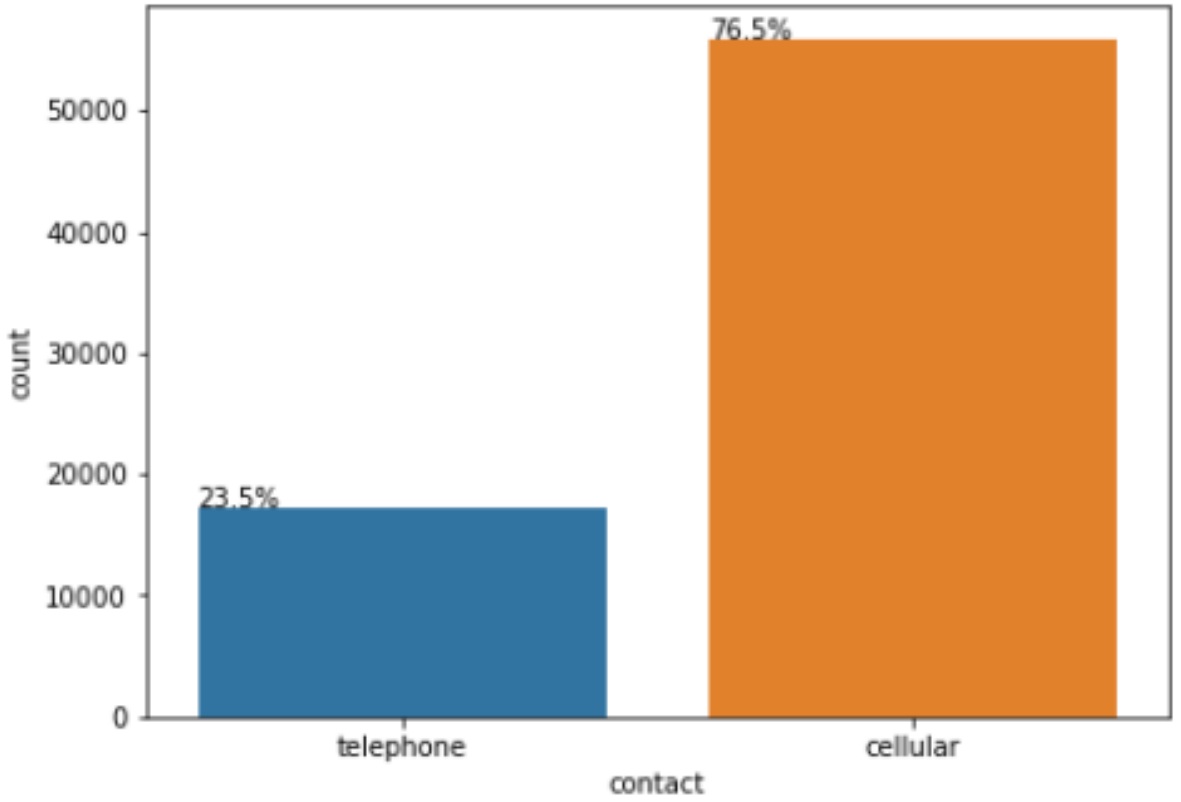
Data Visualization



The graph shows the jobs of the bank customers, we can clearly see that admin blue-collar and technician are the top 3 jobs with more than 50% of the whole bank costumers

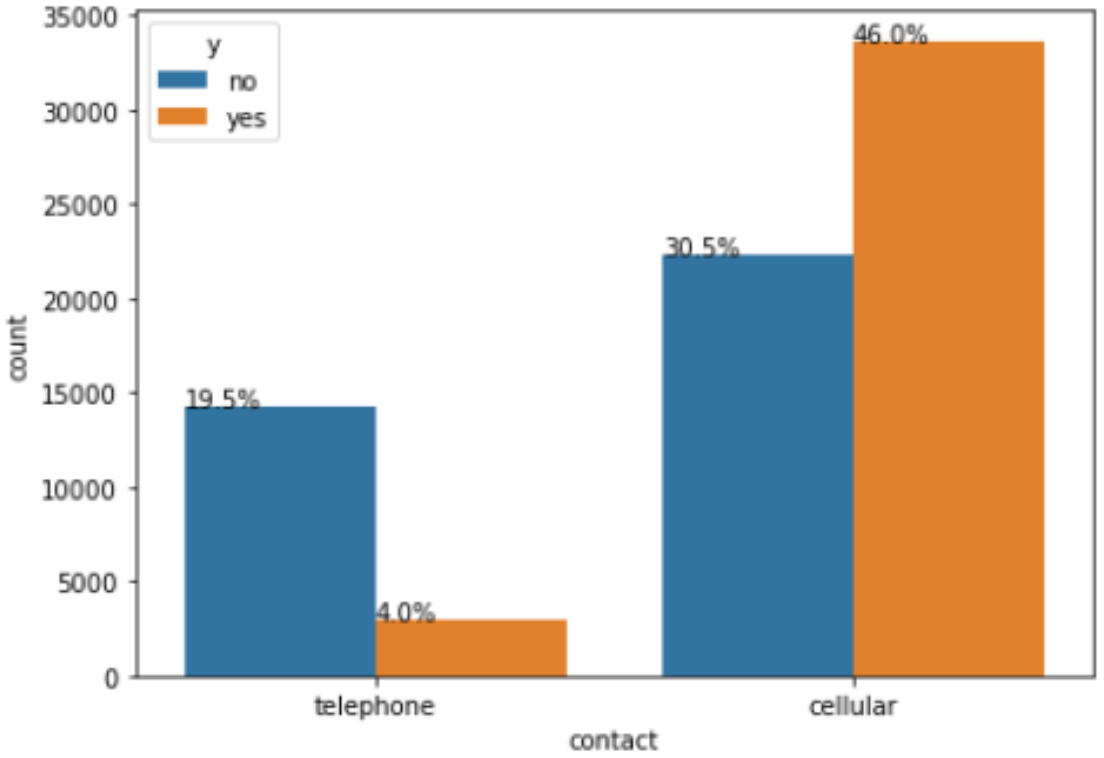
Data Visualization

the costumer that contact the bank by cellular are greater than telephone

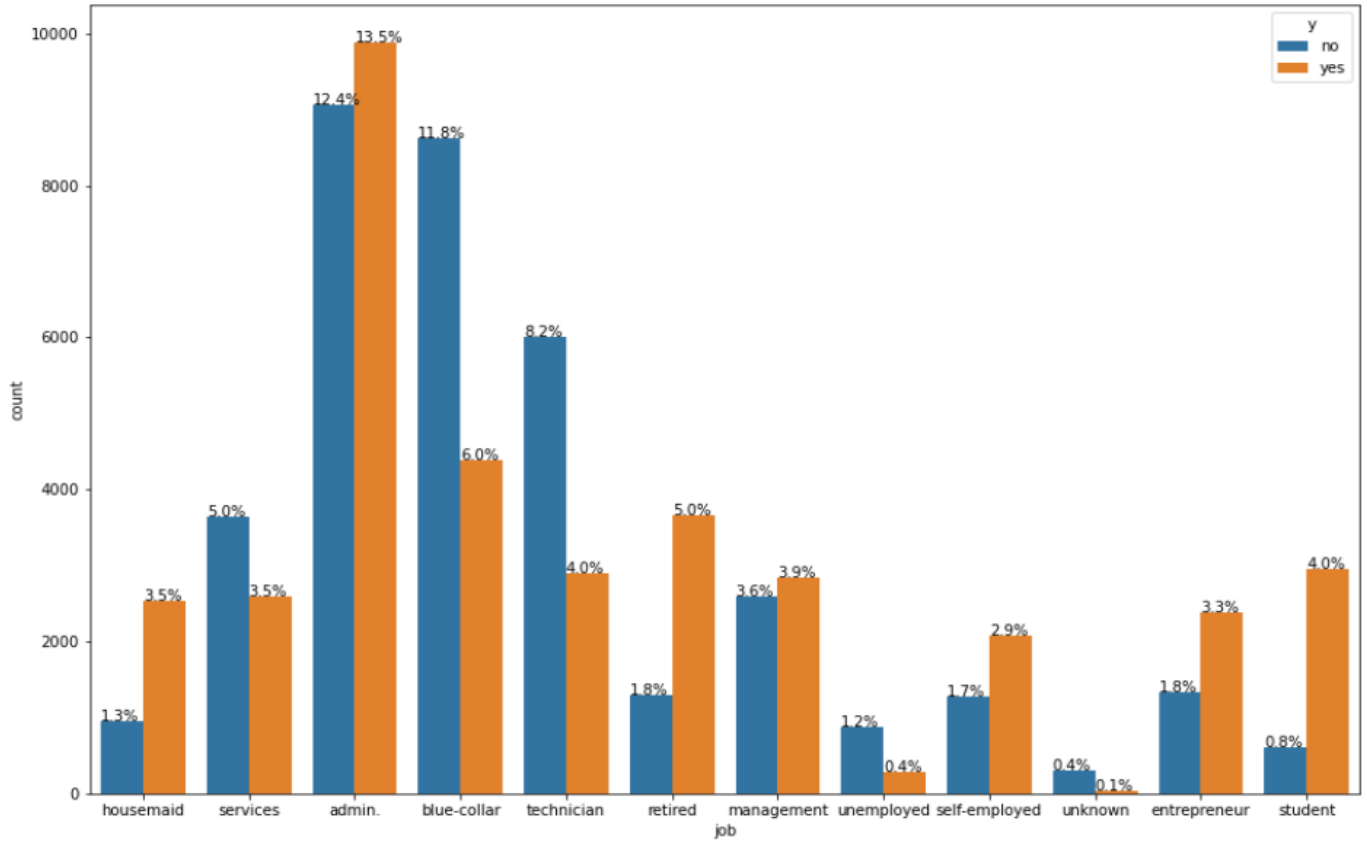


Data Visualization

And the contact's effect is clear! People who use cellular are almost 10 times more likely to subscribe to the product that people with telephone. (46% vs 4%)



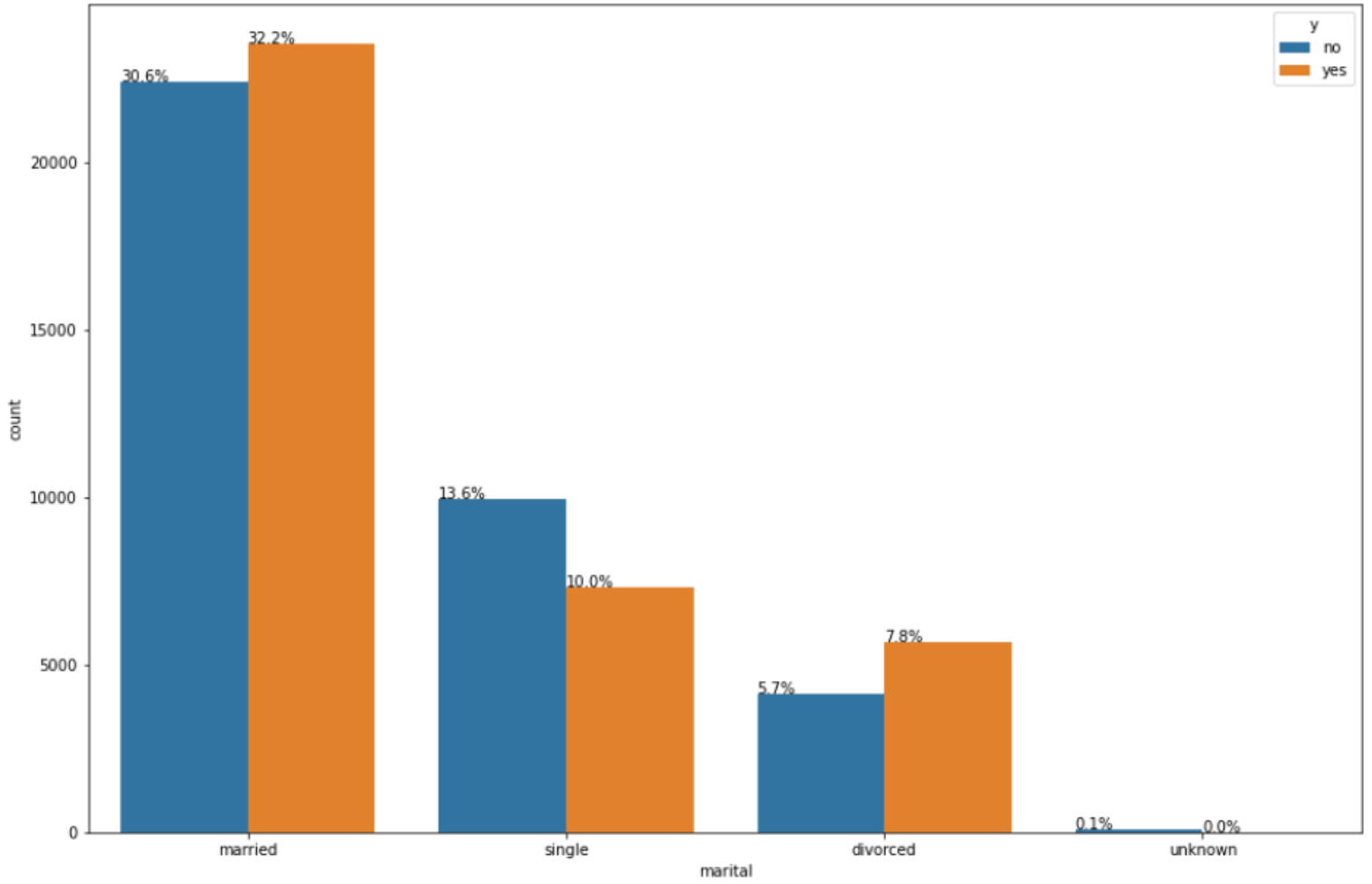
Data Visualization



Even though the admin, blue collar and technician are the top 3 jobs in the banks customers, the people who subscribe the most are students (for every 5 subscribers 1 does not subscribe)

Data Visualization

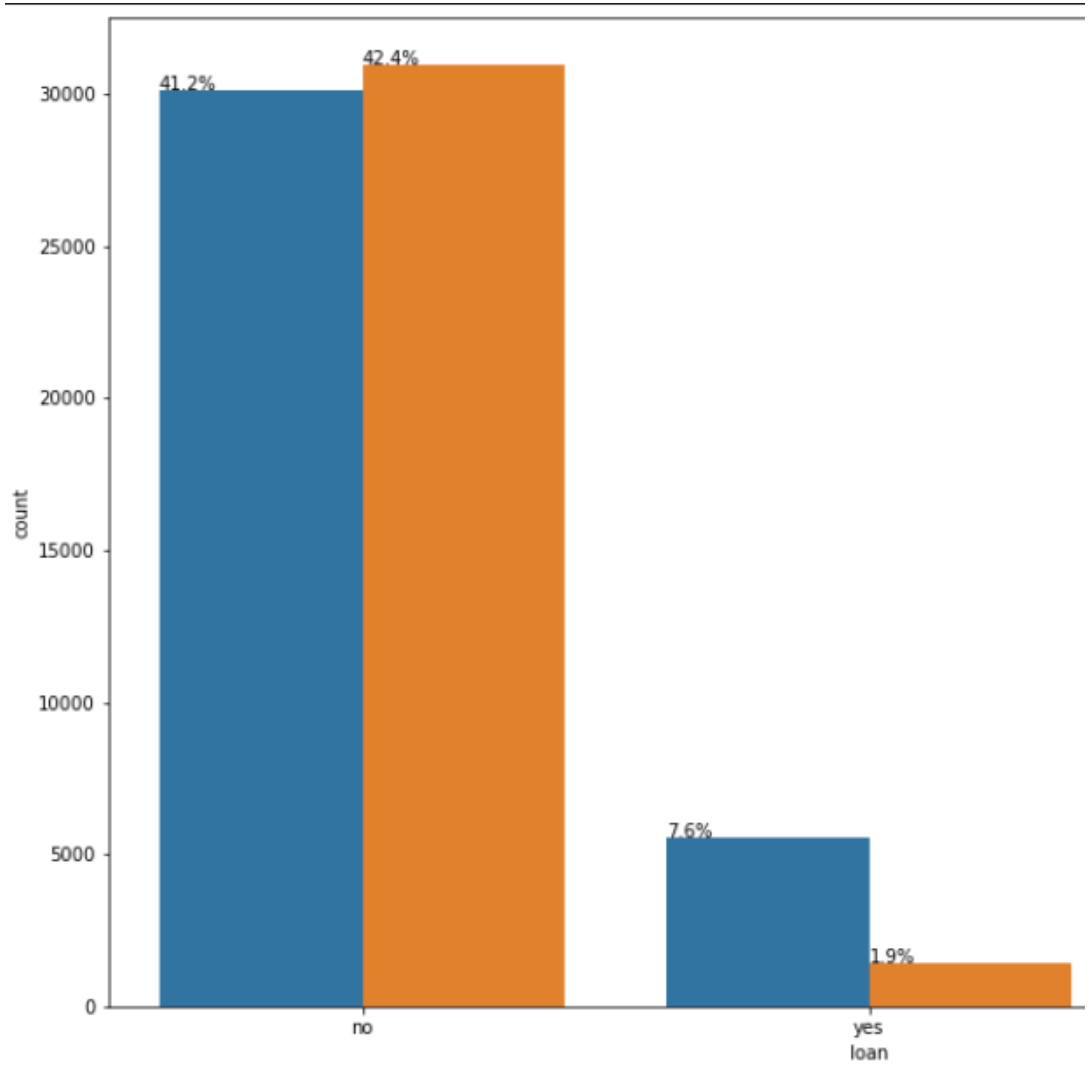
Social status seems to be somewhat important. Married and divorced people seem to be more likely to subscribe



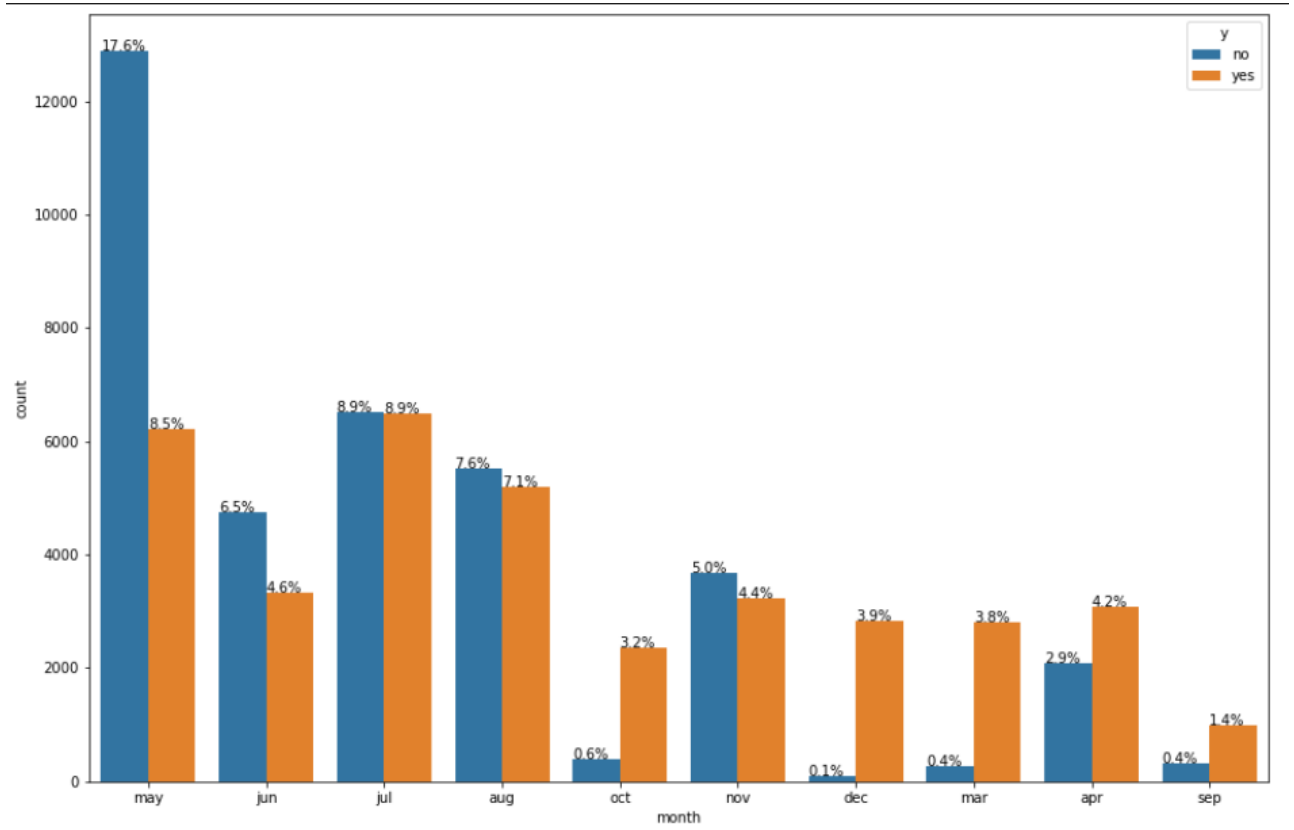
Data Glacier

Data Visualization

People with loans are less likely to subscribe



Data Visualization

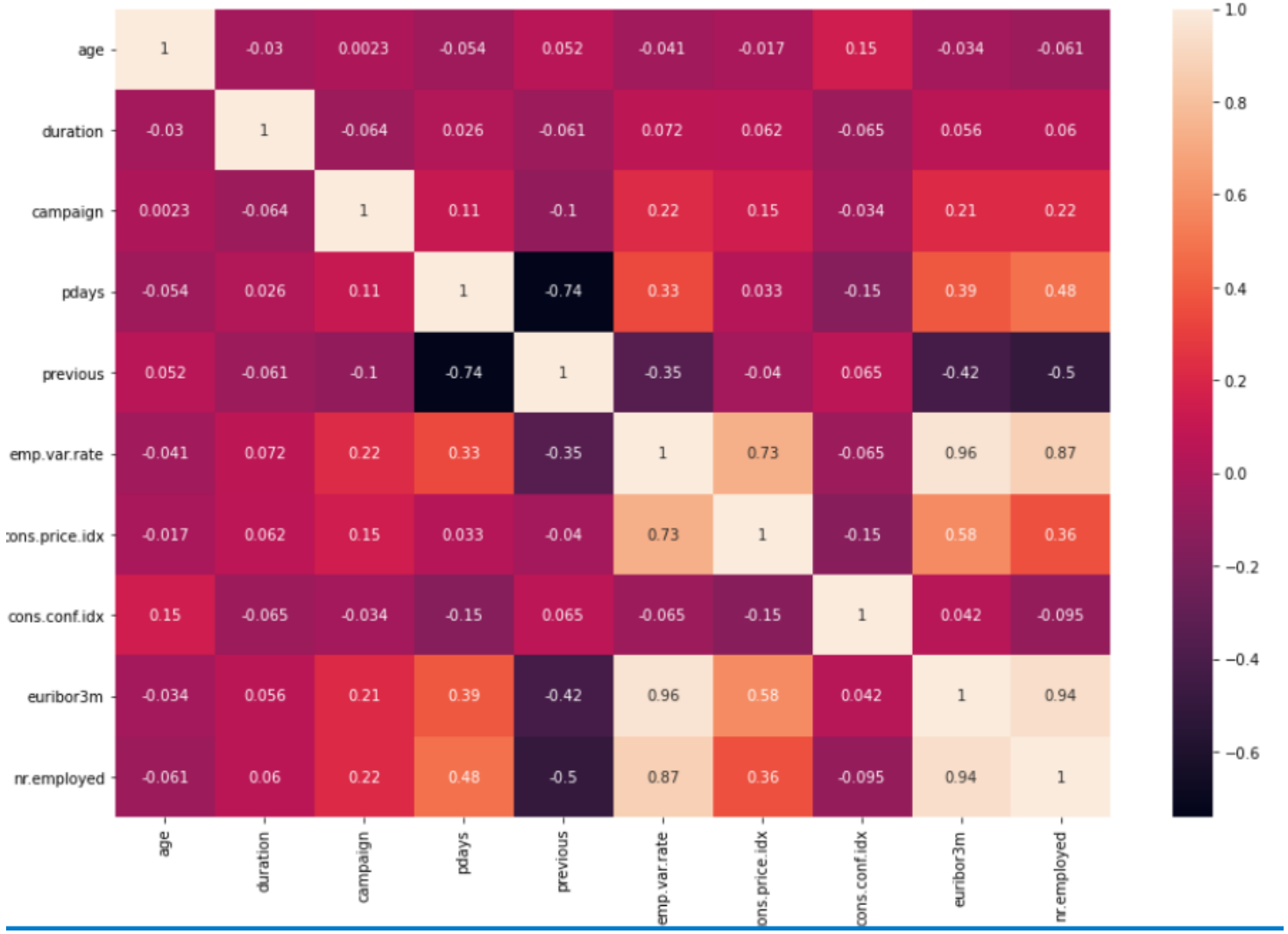


Subscriptions in Mart , April, October in December are the peak of the year, and may is the worst

Data Glacier

Data Visualization

The heat map:
To show the correlation between columns



Data Glacier

Recommendations

1.Students and home maids are most likely to subscribe to the product.

2.People who are married has been contacted more for the deposits by the bank.

3.The contact type (cellular vs telephone) plays a role.

4.People has been contacted more in the month of May than any other month. They have not been contacted in January and Februray at all.

5.Single people are less likely to subscribe.

6.People with no personal loan has been contacted more by the bank.

7.People who are in university has been contacted more by the bank.

8.Age,Duration,Campaign have outliers and are rightly skewed.

9.Pdays have more than 70% of data imputed so it is better either to impute or remove the column.

10.Euribor3m with nr.employed and emp.var.rate with nr.employed with the highest correlation

Model Recommendations

(technical user)

1.Since the data contains many columns with categorical data which are going to increase the dimensionality (e.g. after applying one-hot encoding) we recommend using PCA for dimensionality reduction.

2.We recommend starting with tree-based models like Decision Tree, Extra Tree and the Random forest classifier because they are simple yet effective models.

3.Lasso and Ridge classifiers can also be used in case PCA didn’t give any improvements in the results

4.Accuracy score can be used as an accuracy metric since the problem is classification problem.

5.Only choosing the model is not sufficient, Tuning the hyper parameters plays a huge rule.

6.Finally we recommend using more than one model, tuning the hyperparameters and settling for the best model (accuracy-wise)

Thank You