**Python Case Study**

**BACKGROUND :**

The client is from banking sector and they conduct marketing campaign over phone to convert the liability customer (depositor) to Asset customer (borrower) to earn the interest to increase revenue. Sometimes regular follow-up is needed in order to subscribe to the bank's term deposit.

**OBJECTIVE :**

The client has rich customer records in their repository and they want leverage those to predict whether the customer will subscribe to fixed-term deposit products the bank is offering.

**DATA:**

The data has 17 customer attributes for 45211 customers. The data has customer's demographic information, relationship with the bank and customer response to the last personal loan campaign. Target column contains information regarding subscription to the term deposit.

**Questions :**

**1. Find the 5 point summary of the dataset.**

1. Data is Unbalanced. Having less successive events as compared to failure event

2. The data is more skewed. Some of them have more outliers.

3. According to analysis Target customers with the criteria of

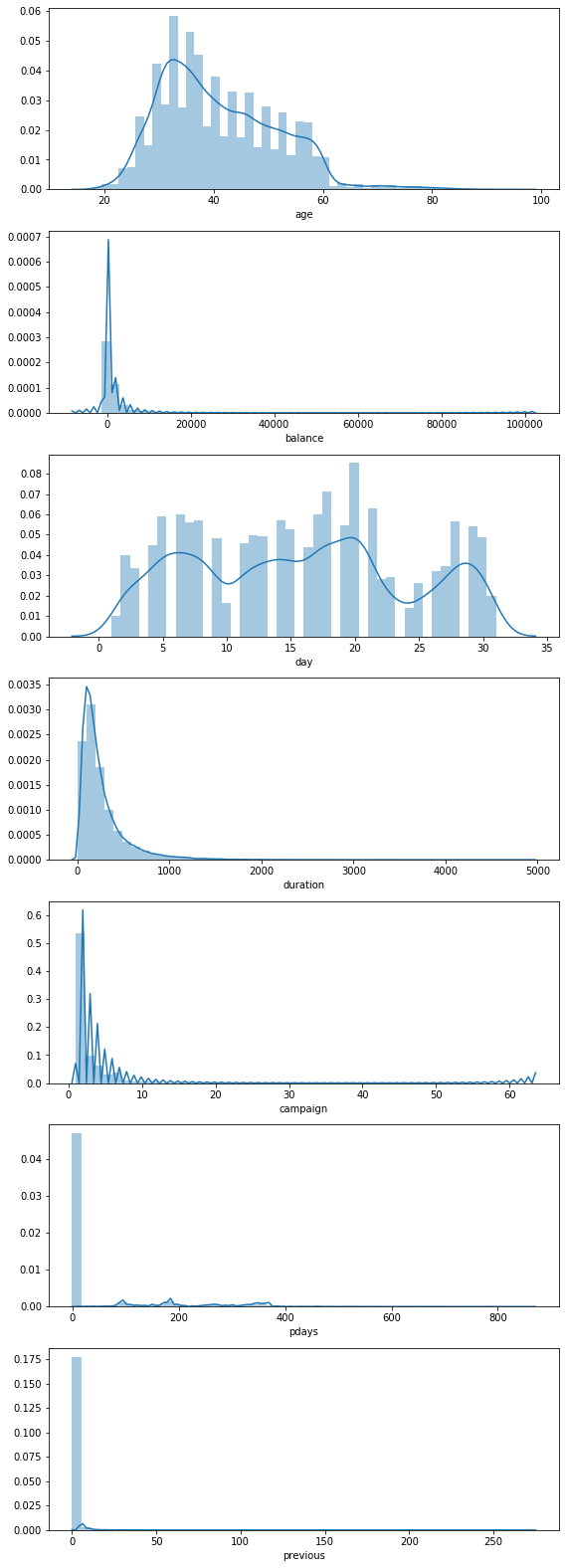
* Concentrate on young and old people instead of middle age
* More likely student and retied people have better results
* The persons who have spoken a more number of call has less chance of successive rate

4. The bank contacted most clients between May and August. The highest successive rate occurred in March, which is over 50% and successive rates in September, October, and December are over 40%.

5. The person who have have more than 1000 euros has more successive rate.

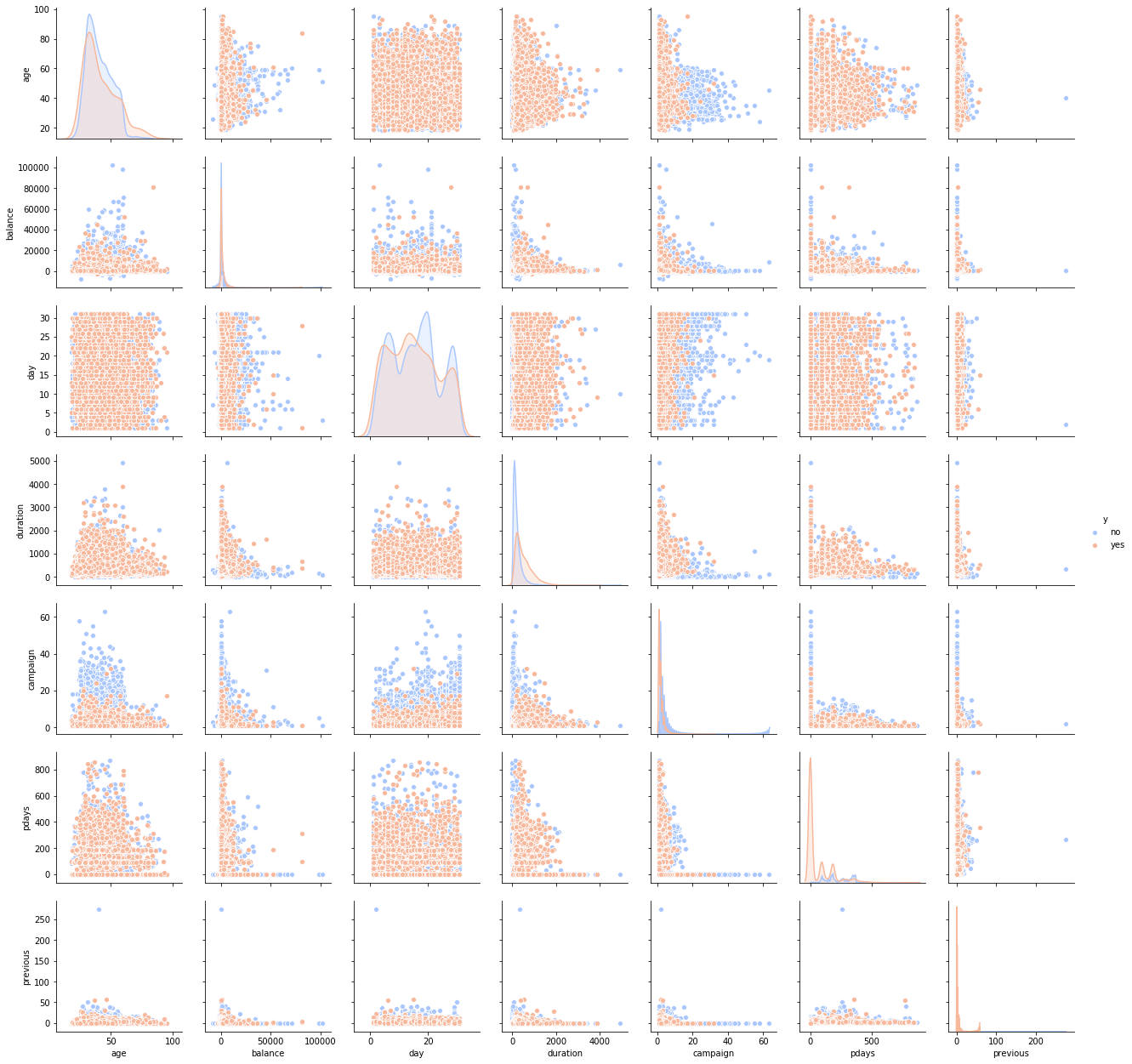
2. Plot to see how the data the dataset is distributed.

Numerical data



3. Determine how the variables are related to each other. If there is any relation explain it.

Scatter distribution with numerical variable

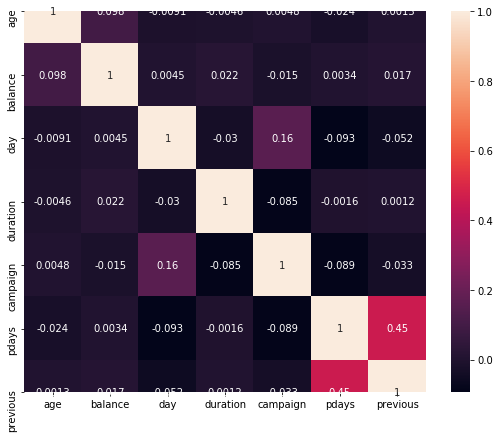


**4. Perform Data Preparation before classification model.**

For data preparation have used label encoder to prepare the data to use in modelling

**5. Visualized the relationship between variables**

**Correlation**



6. Create classification Models.

Models Used

* Logistic regression
* Decision tree
* SVM
* Random forest
* Neural network
* Random forest

7. Compare all classification models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Models | Accuracy in train | Accuracy in test | F1 score in Train | F1 score in TEST |
| Logistic regression | 88.08 | 88.20 | 0.15 | 0.14 |
| Decision tree | 100 | 86.54 | 100 | 43.65 |
| SVM | 98.44 | 88.44 | 92.7 | 0.00 |
| Neural network | 88.23 | 88.44 | 92.70 | 0.00 |
| Random forest | 99.07 | 89.26 | 0.00 | 0.37 |

**Conclusion:**

**Most of the time we use Accuracy as metric to determine the model performance. But,the successive rate in less in data so we have to concentrate on one more metric called F1 score.**

**Form the result we can use decision tree model because it has decent sum of F1 score nearly 44%.**