Data Glacier

Your Deep Learning Partner

Final Model Recommendation

Virtual Internship

17th Oct 2021

Background

- XYZ bank wants to roll out Christmas offers to their customers. But Bank does not want to roll out same offer to all customers instead they want to roll out personalized offer to particular set of customers. If they manually start understanding the category of customer then this will be not efficient and also they will not be able to uncover the hidden pattern in the data (pattern which group certain kind 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 group** as this will be inefficient for their campaign.

- Objective : Provide actionable insights to help the bank in segregation customers in such a way that the bank can focus on clients who are more likely to agree to the offer

The analysis has been divided into three parts:

- Data Description
- EDA
- Model Recommendation

Data Description

- age (numeric)
- job : type of job (categorical: "admin.","blue-collar","entrepreneur","housemaid","management","retired","self-employed","services","student","technician","unemployed","unknown")
- marital : marital status (categorical: "divorced","married","single","unknown"; note: "divorced" means divorced or widowed)
- education (categorical: "basic.4y","basic.6y","basic.9y","high.school","illiterate","professional.course","university.degree","unknown")
- default: has credit in default? (categorical: "no","yes","unknown")
- housing: has housing loan? (categorical: "no","yes","unknown")
- loan: has personal loan? (categorical: "no","yes","unknown")
- related with the last contact of the current campaign
- contact: contact communication type (categorical: "cellular","telephone")

Data Description

- month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")

- day_of_week: last contact day of the week (categorical: "mon","tue","wed","thu","fri")

- duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

- campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

- pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

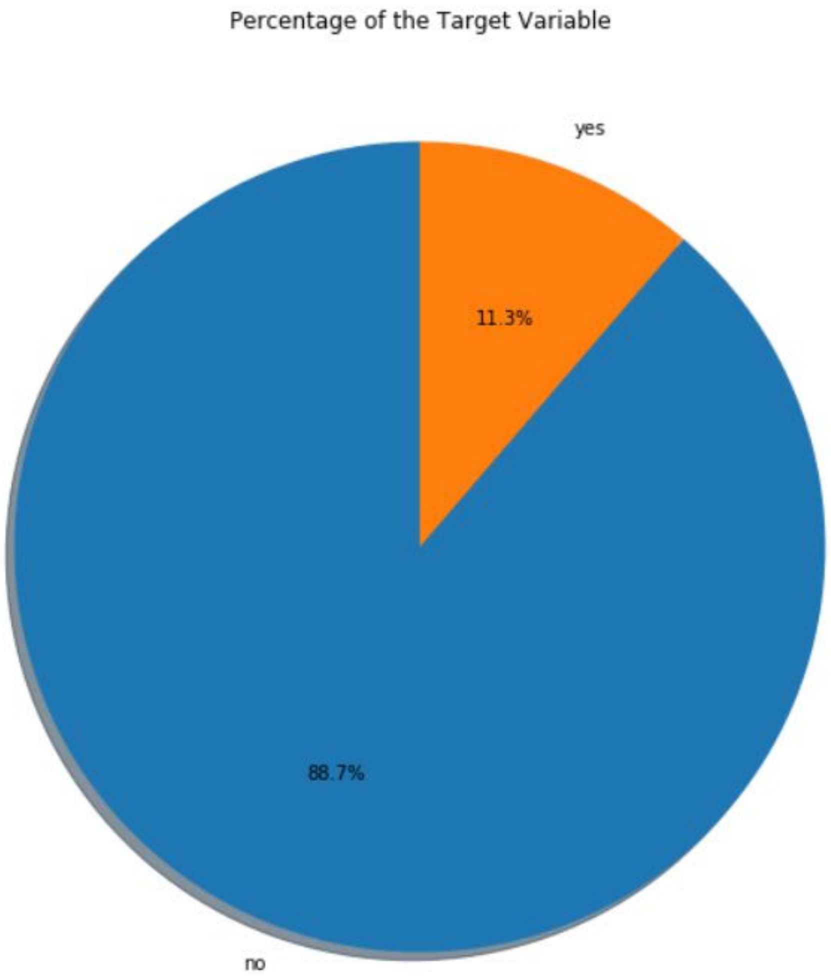
- previous: number of contacts performed before this campaign and for this client (numeric)

- outcome: outcome of the previous marketing campaign (categorical: "failure","nonexistent","success")

Data Description

- emp.var.rate (EVR): employment variation rate - quarterly indicator (numeric)
- cons.price.idx (CPI): consumer price index - monthly indicator (numeric)
- cons.conf.idx (CCI): consumer confidence index - monthly indicator (numeric)
- euribor 3m: euribor 3 month rate - daily indicator (numeric)
- nr.employed: number of employees - quarterly indicator (numeric)
- y - has the client subscribed a term deposit? (binary: "yes","no")

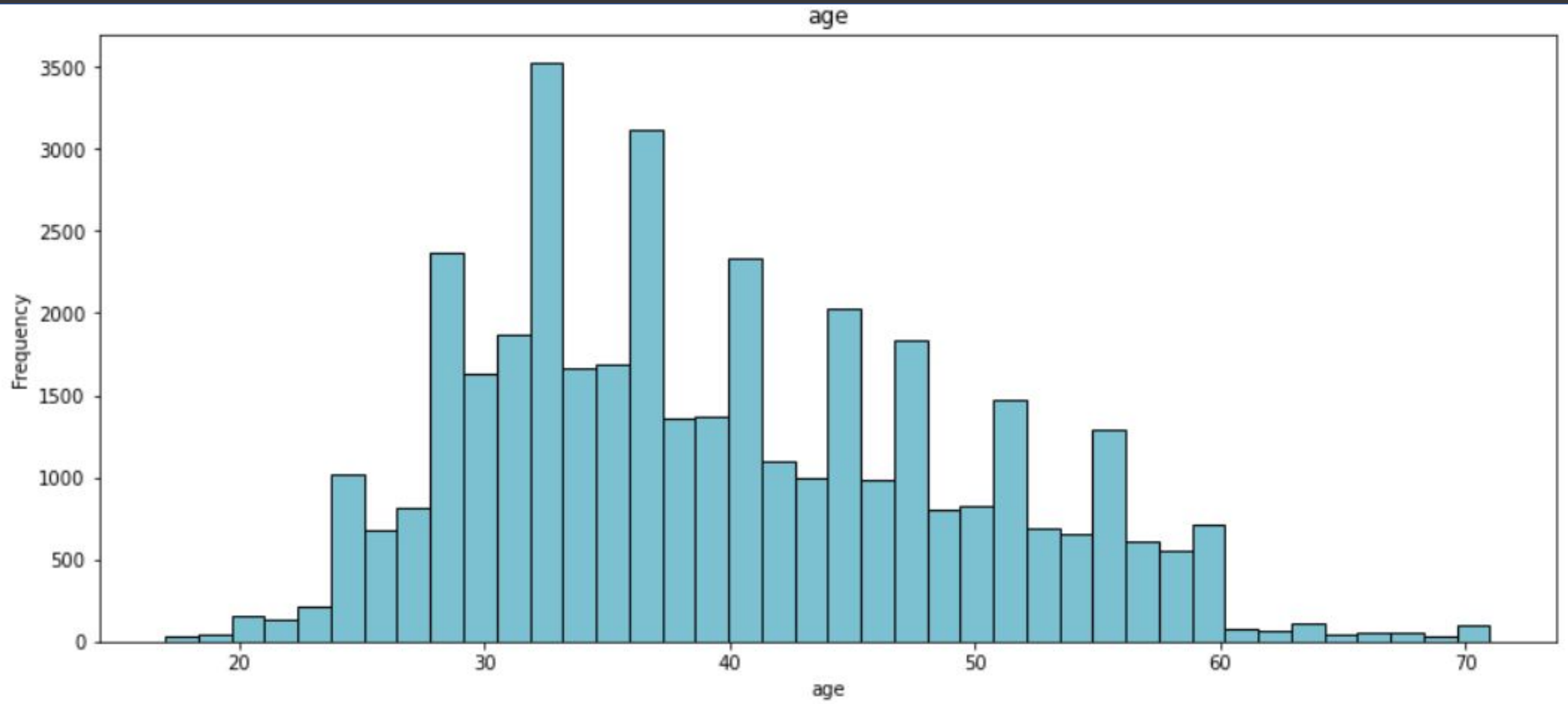
Target Variable



- The target column, i.e. ‘y’ tells whether the client had subscribed to term deposit or not

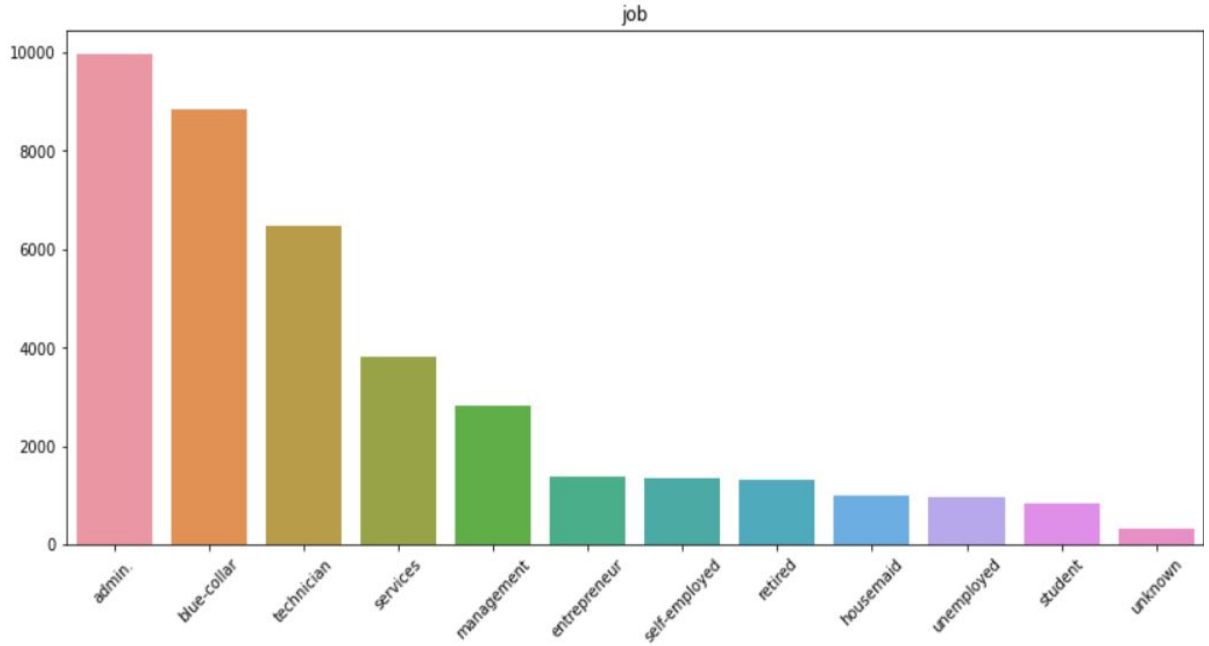
- From the above piechart, we notice that 11.3% of clients have subscribed a term deposit. In other words, 11.3% of the campaign calls are successful.

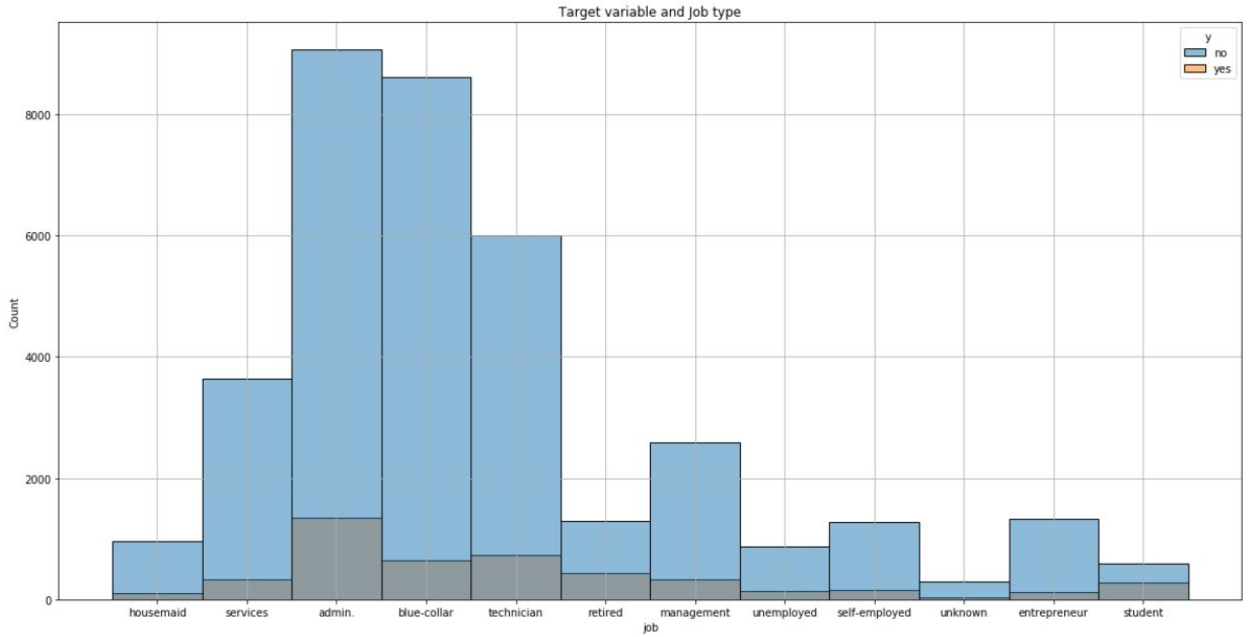
Age range of Clients



- 97.5% of the clients fall between ages 20 and 60

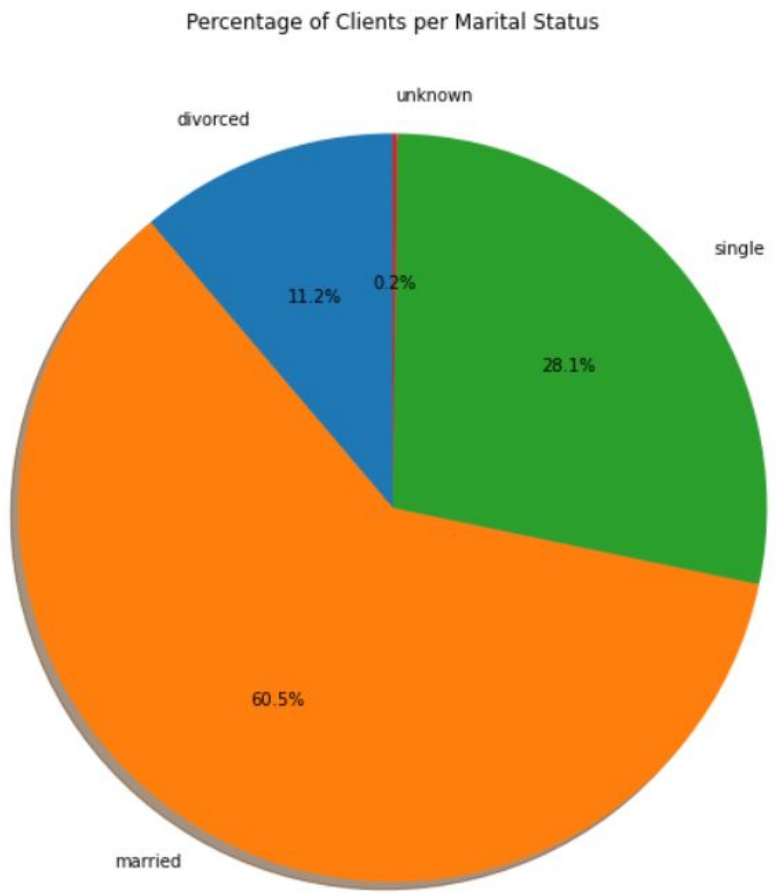
Job Description and its effect on Target Variable

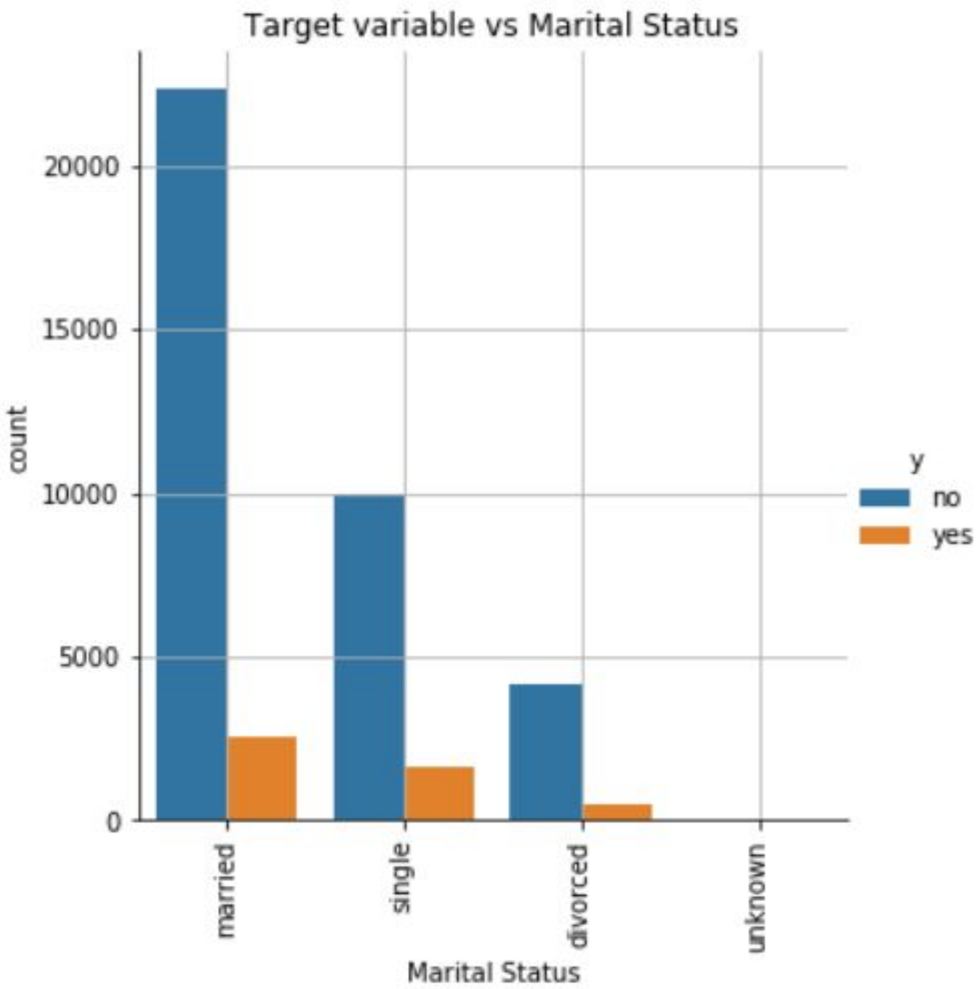




- The clients best 4 job types are admin, blue-collar, technician, and services.
- Customers from admin,blue-collar and technician job types open a deposit account

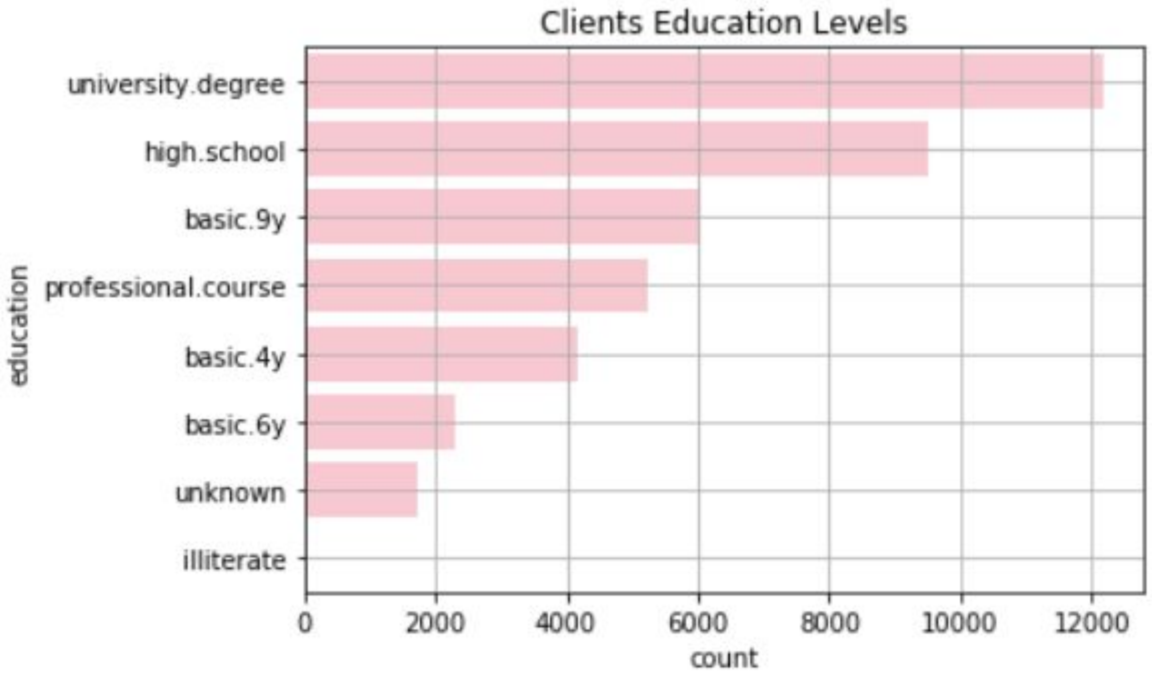
Customers Marital Status and its Effect

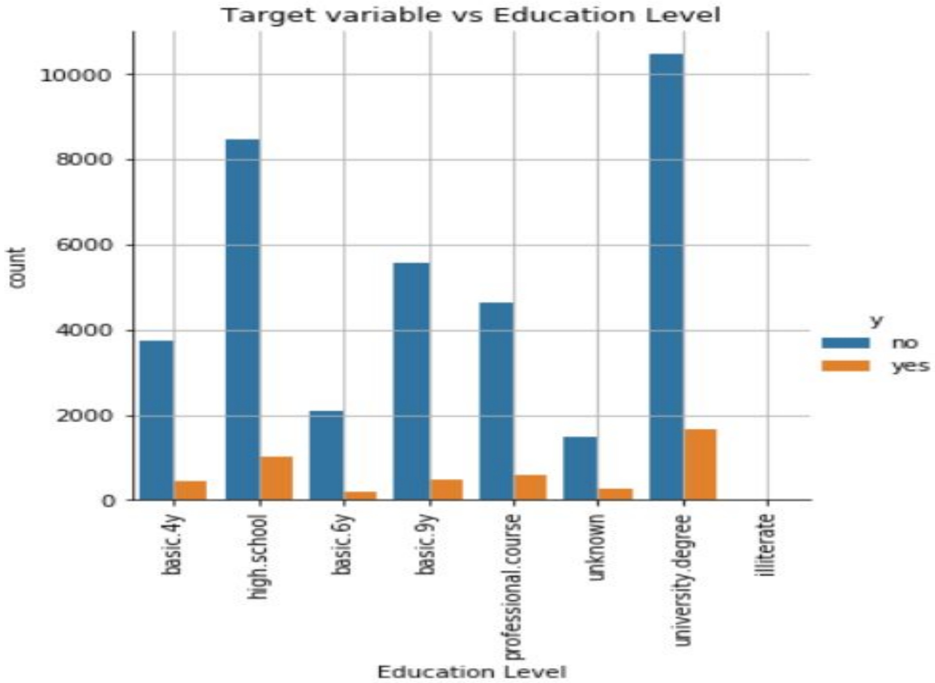




- 60.5% of the clients are married Customers who are married more open to a deposit account compared to other marital status

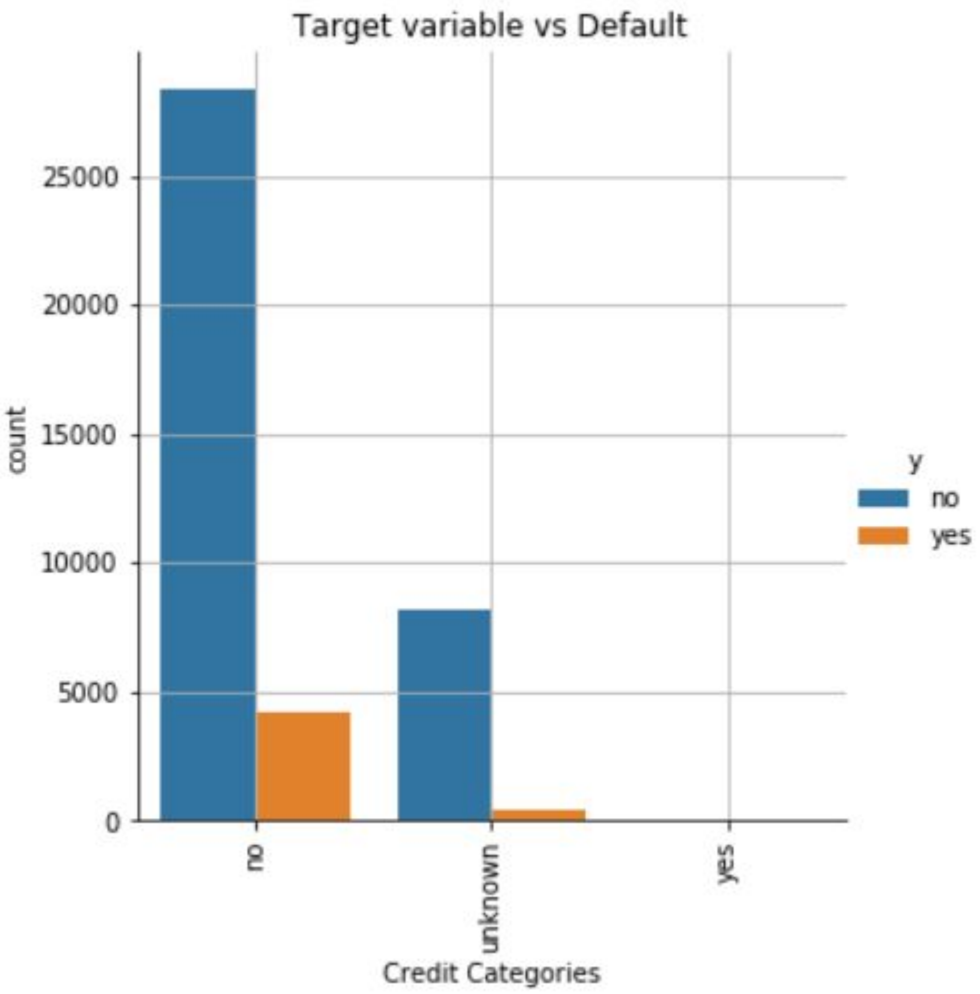
Customers Education Level and its Effect

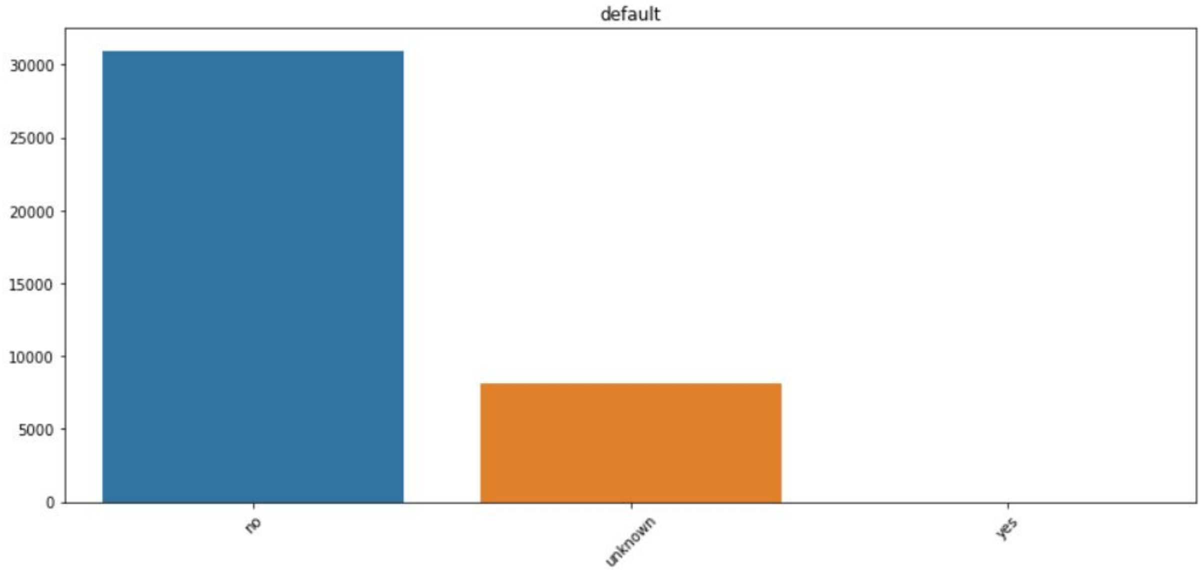




- The majority of clients have completed the university degree followed by the high school degree
- While 1730 clients have unknown education level
- Customers with university degree education level are more likely to sign up for schemes

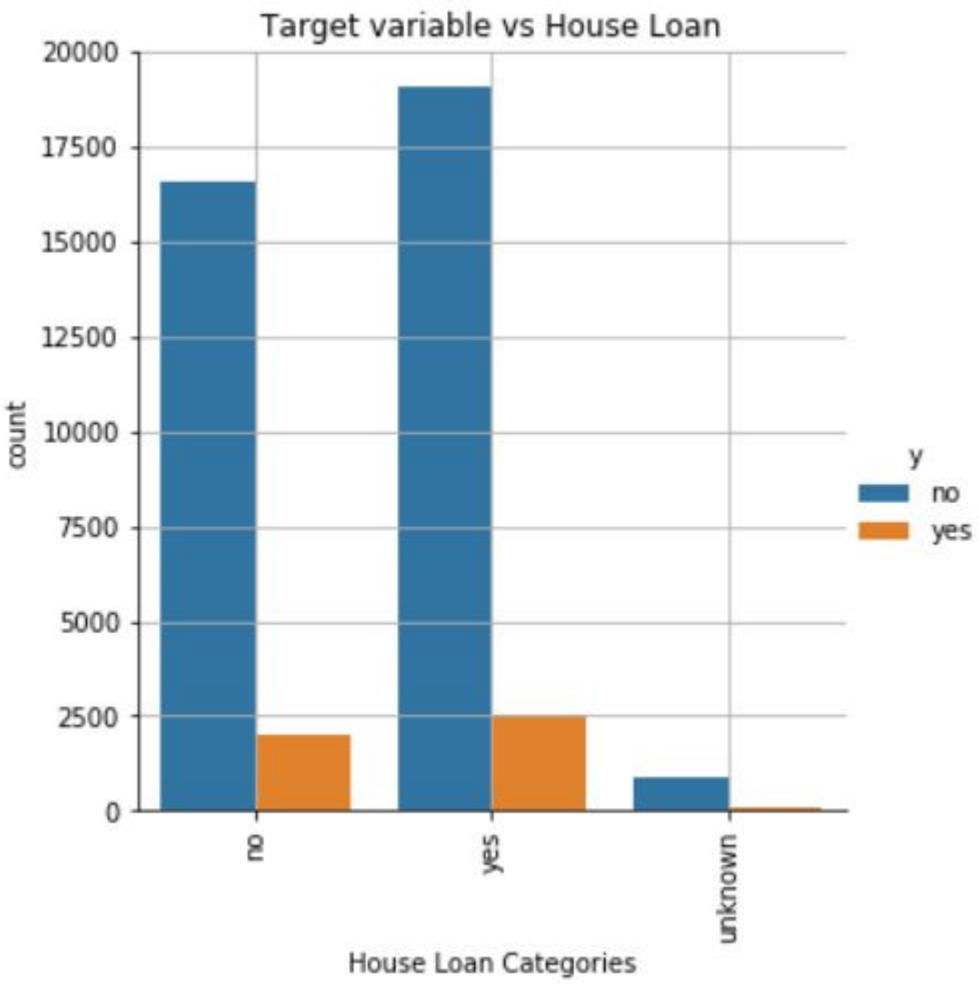
Customers Default Credit Category and its Effect

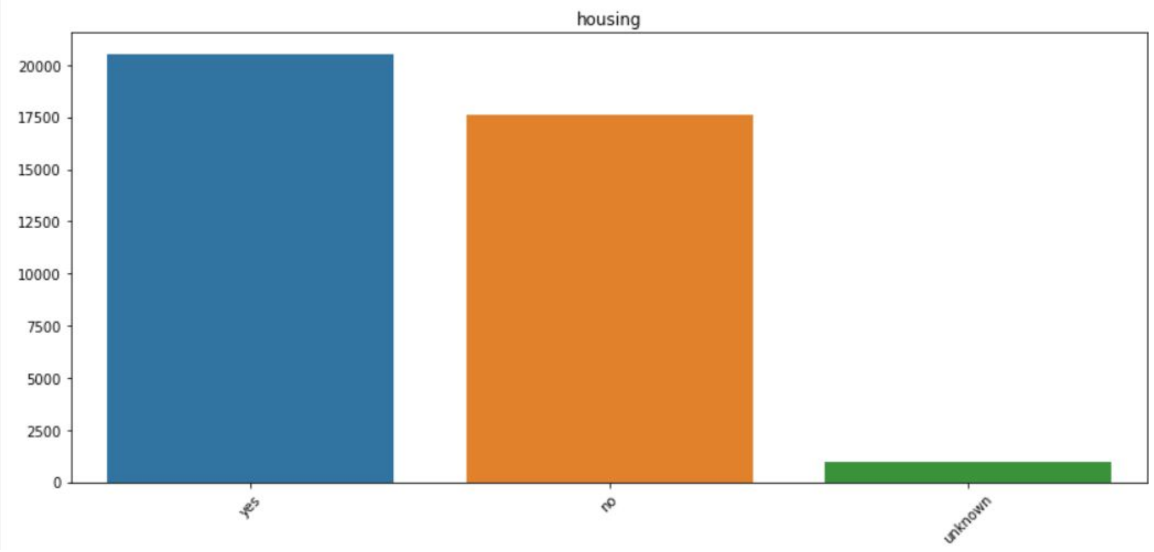




- Customers who have no credits in default open more deposit account compared to other categories

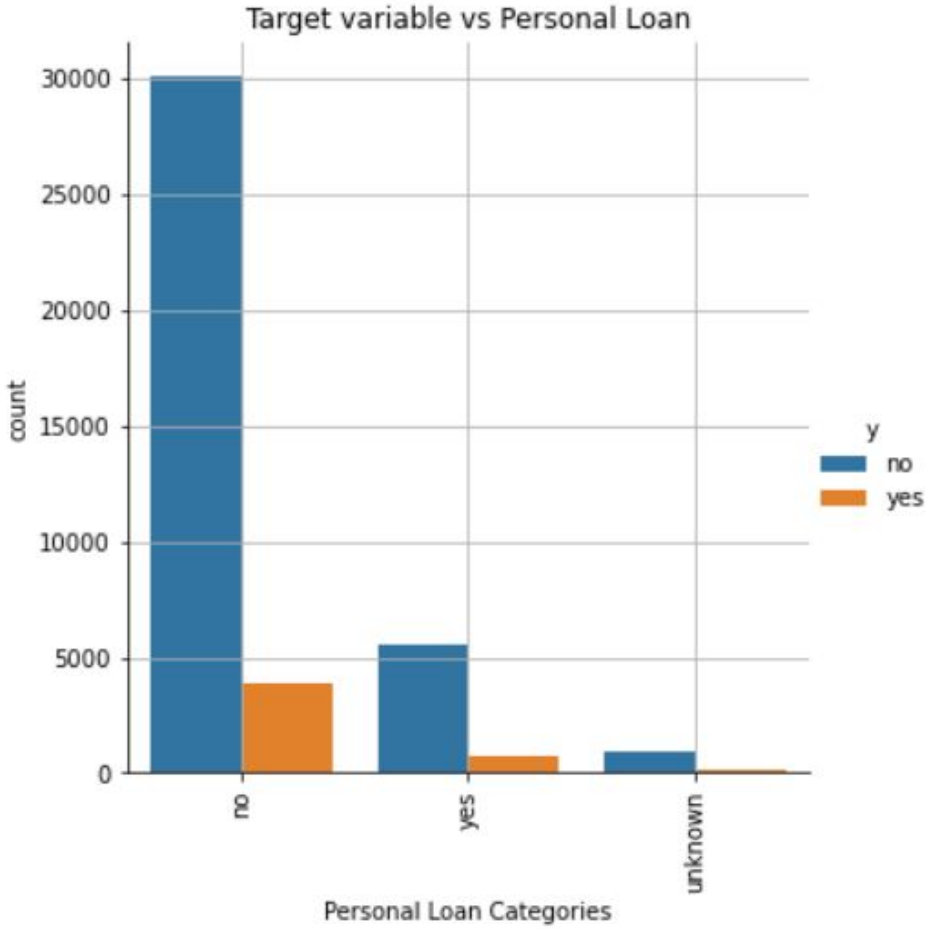
Customers Housing Loan Status and its Effect

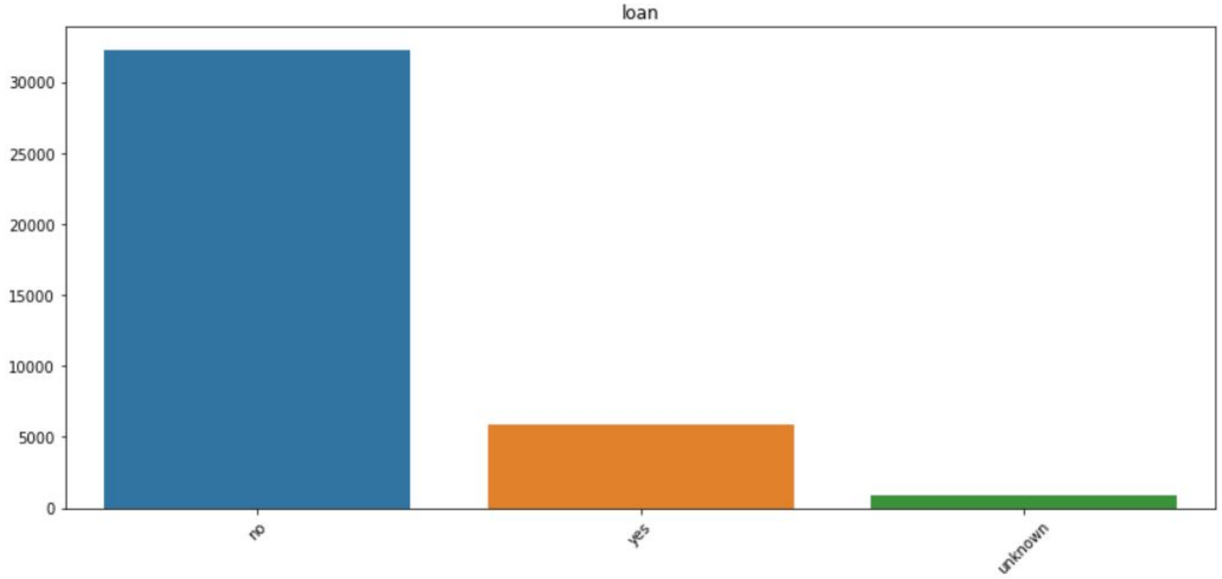




- Customers who have house loan open more deposit account compared to others categories.

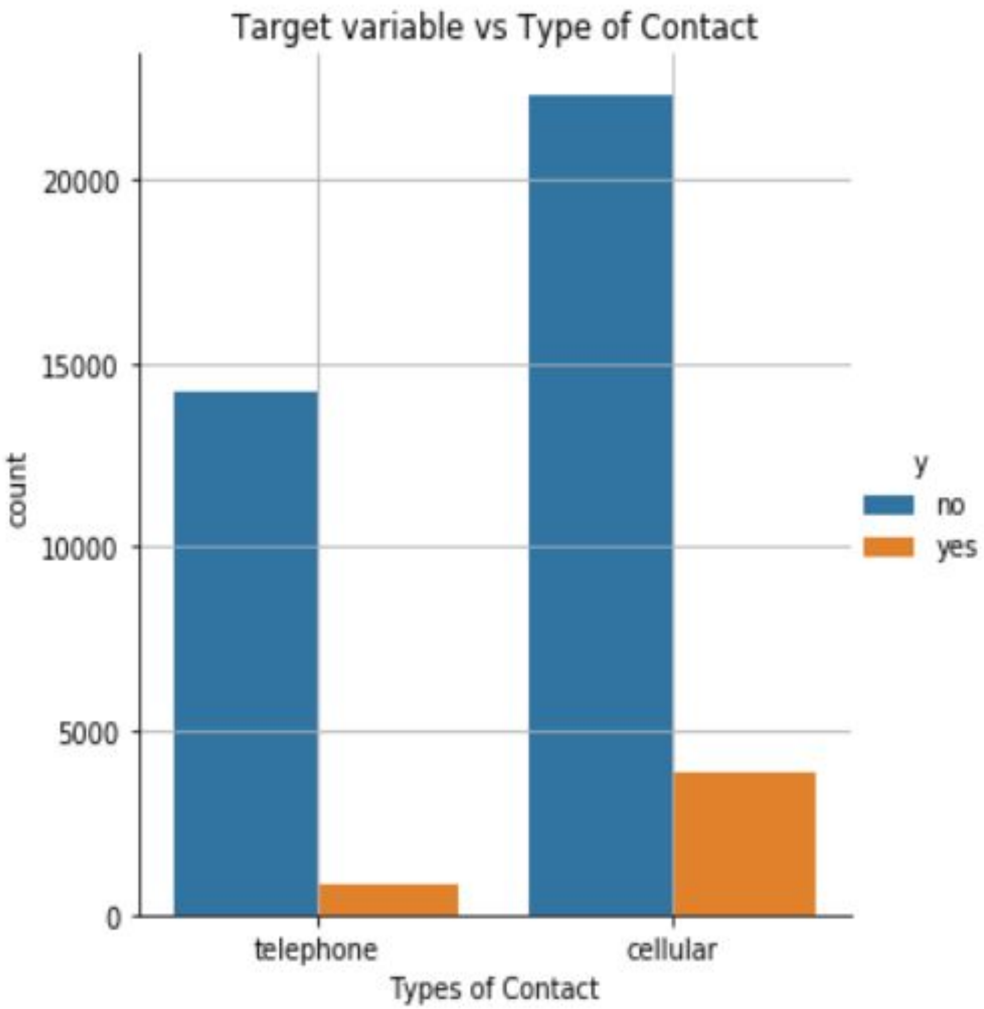
Customers Personal Loan Status and its Effect

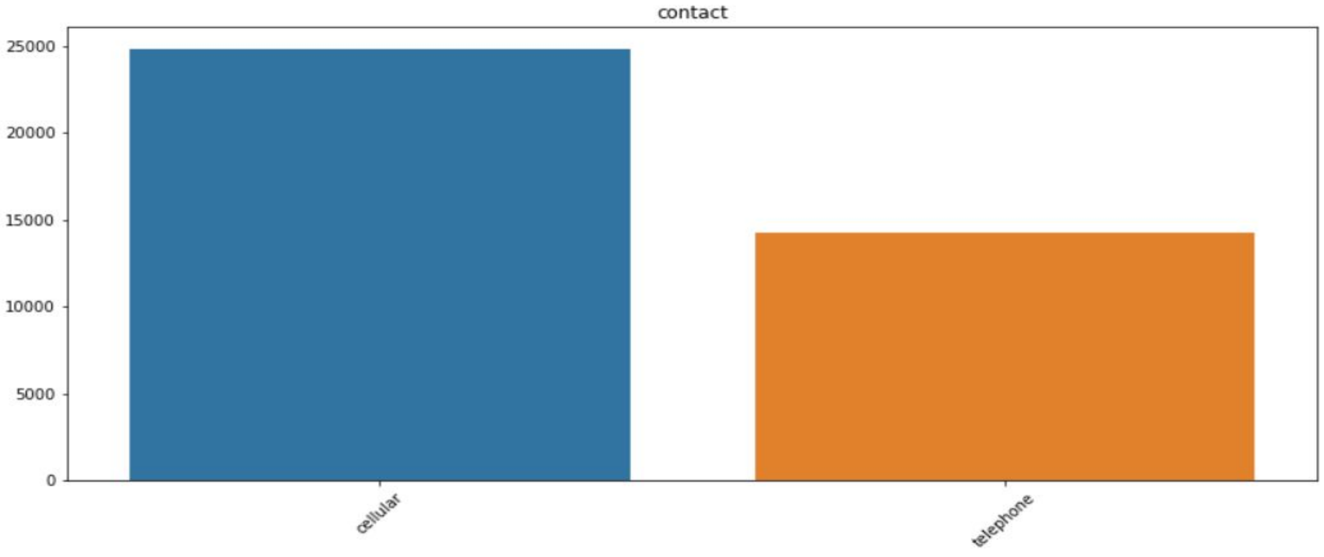




- Customers who have no personal loan open more deposit accounts compared to others categories.

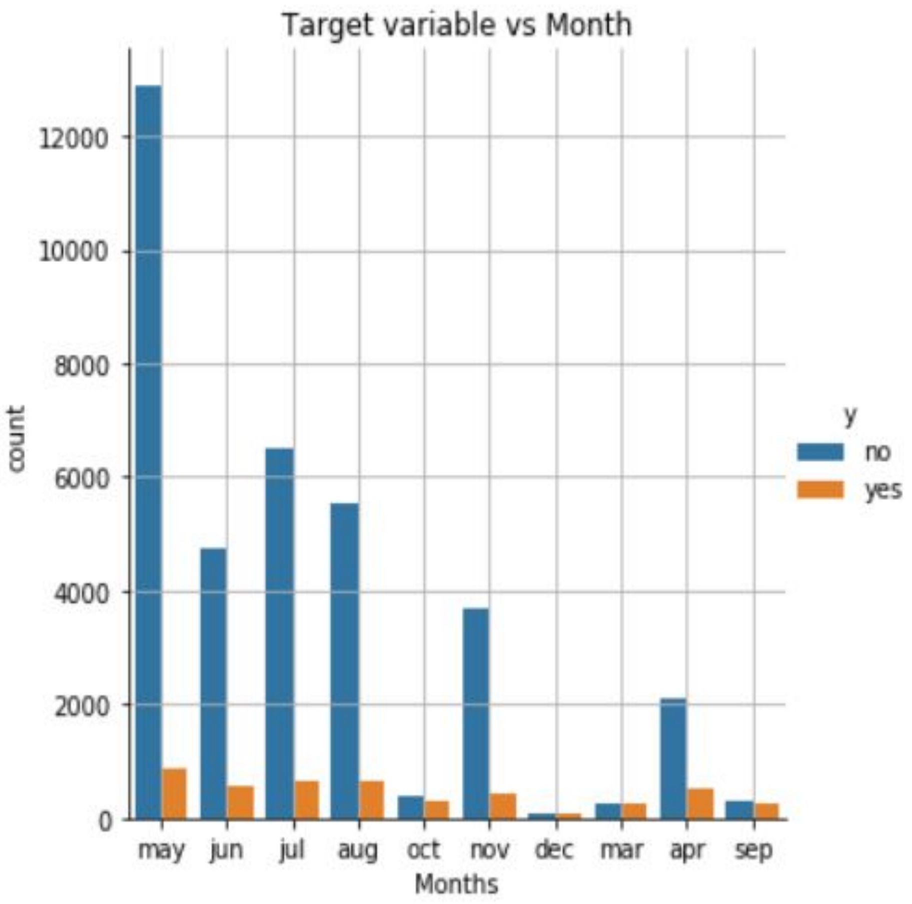
Customers Contact and its Effect

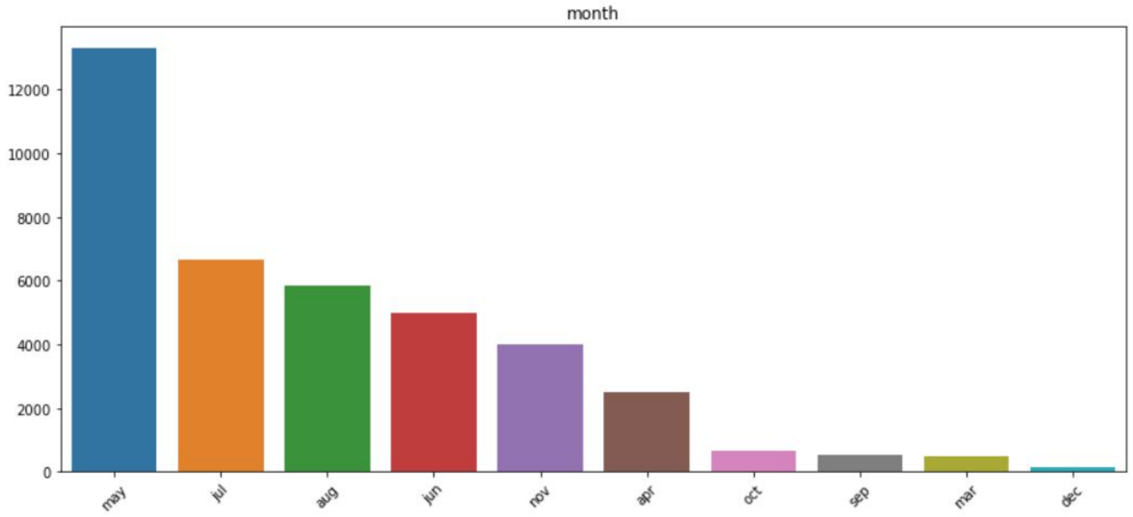




- The majority of contacts is of cellular type. And the clients that open a deposit accounts are almost all contacted by cellular.

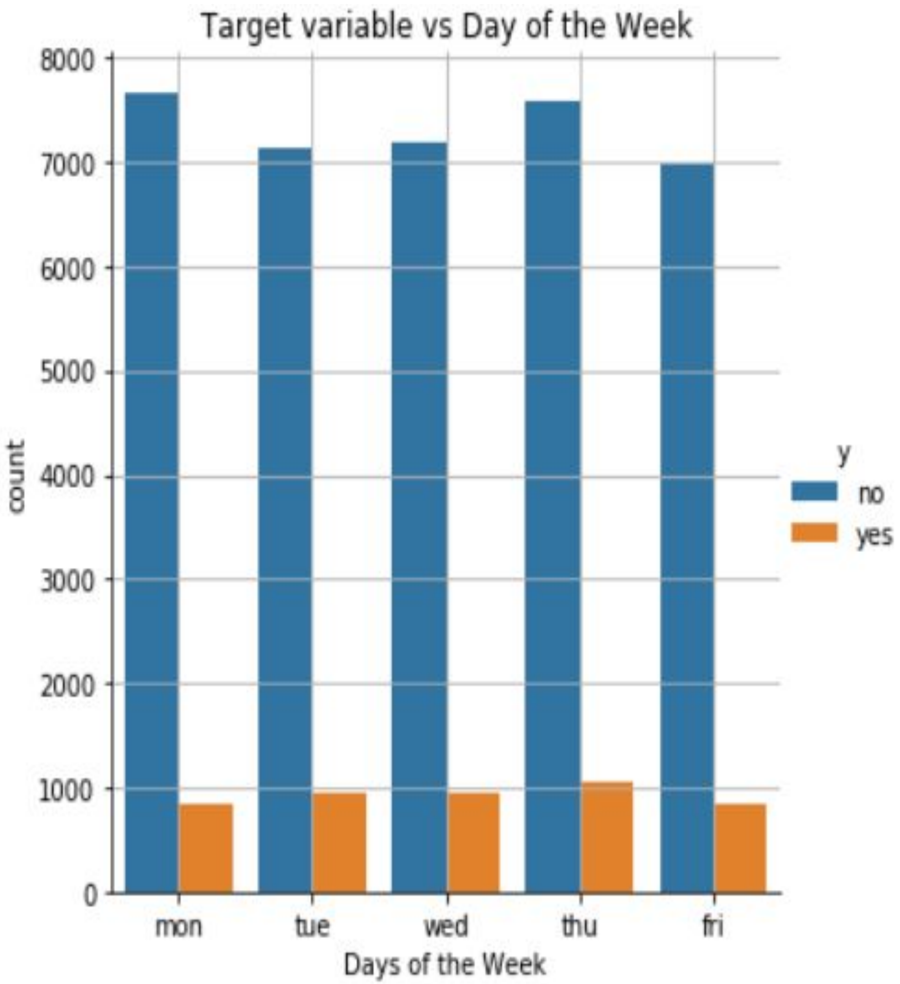
Month and its Effect

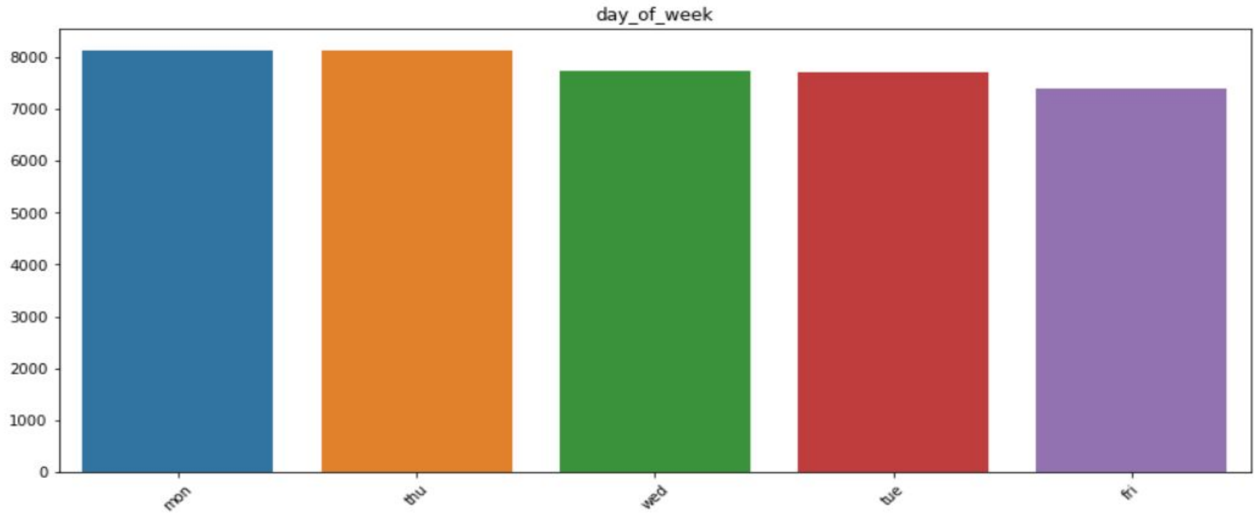




We can notice that the majority of calls were in the month of May. Moreover, the majority of clients that open a deposit account were contacted in the month of May. But comparing the number of call to the number of people signing up, the rate is quite poor for May.

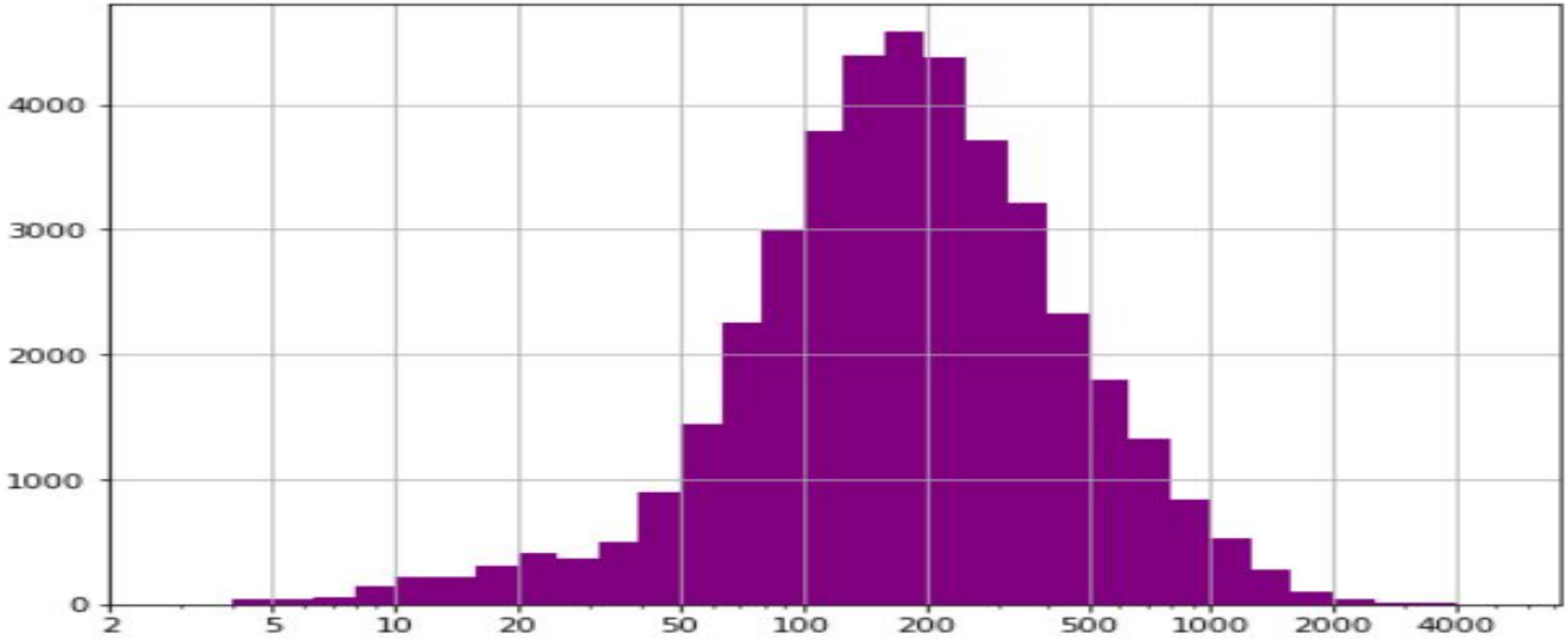
Days of week and its Effect





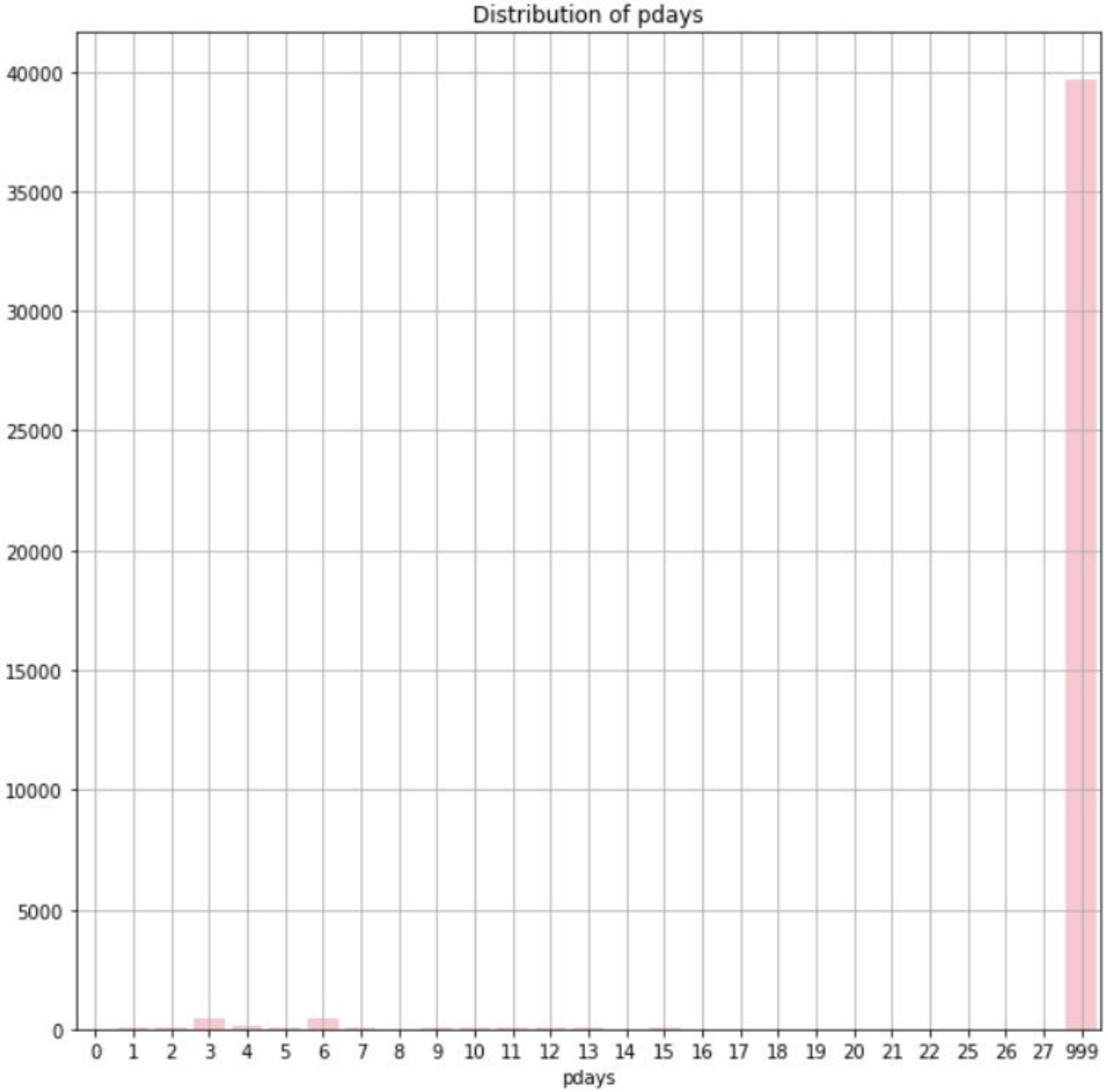
- We can notice that the majority of calls were in Thursday of every week..

Duration of Phone Call



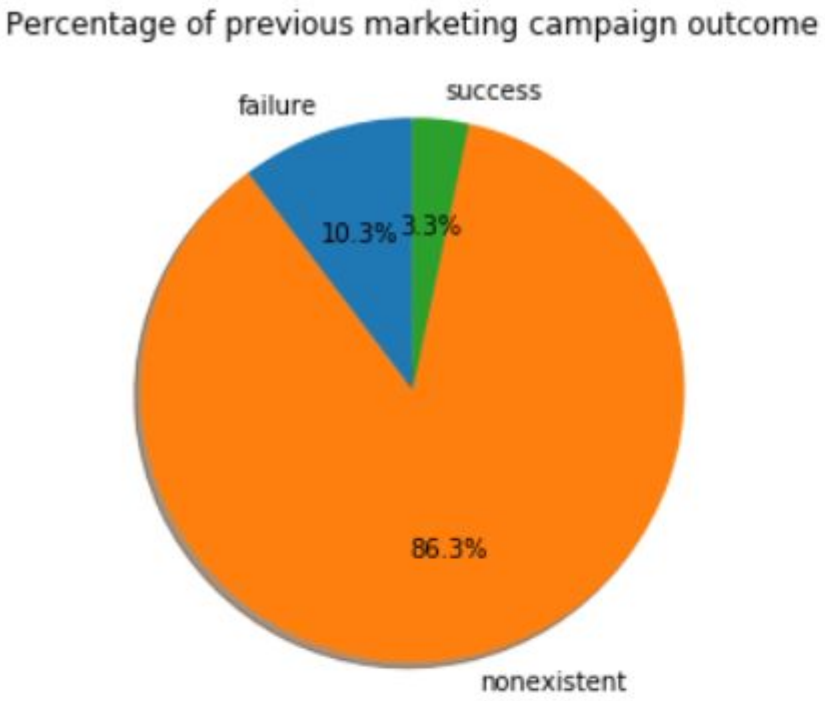
- The majority of calls duration lies in the range between 50 and 800 secs

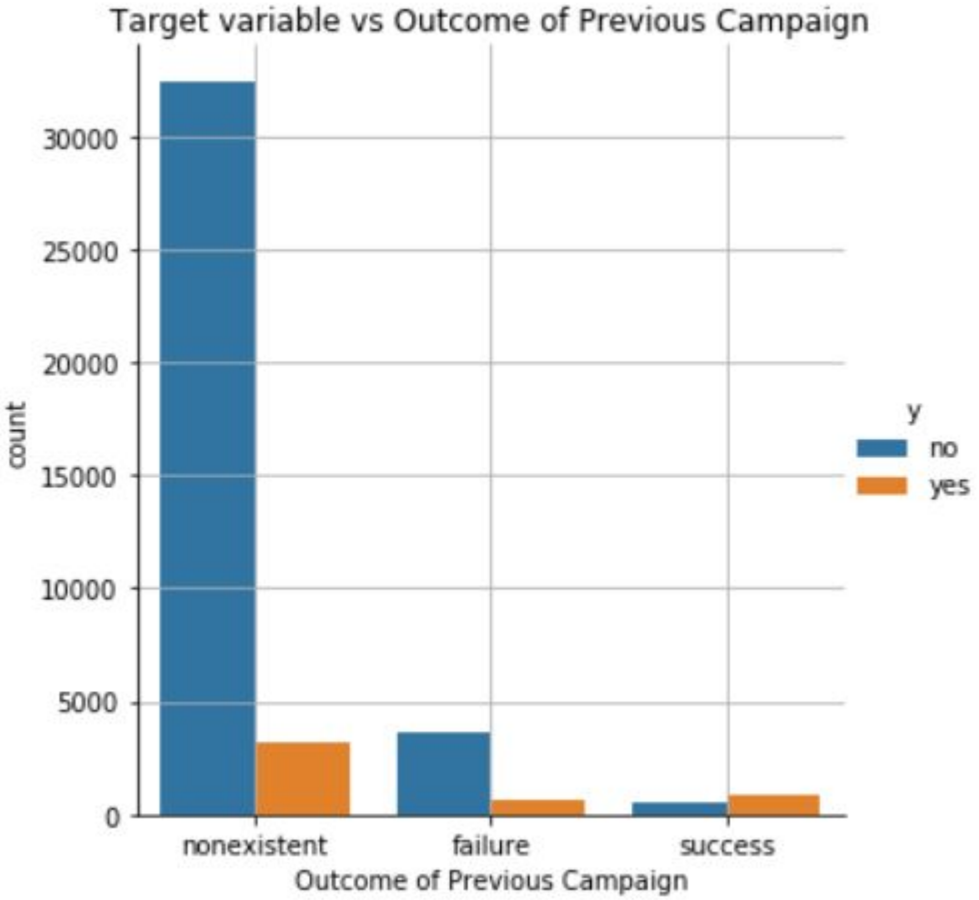
Gap between contacted days



- 999 means they weren't contacted before
- We can notice that the majority of clients are not previously contacted.
- The data has outliers at 3 days and 6 days.

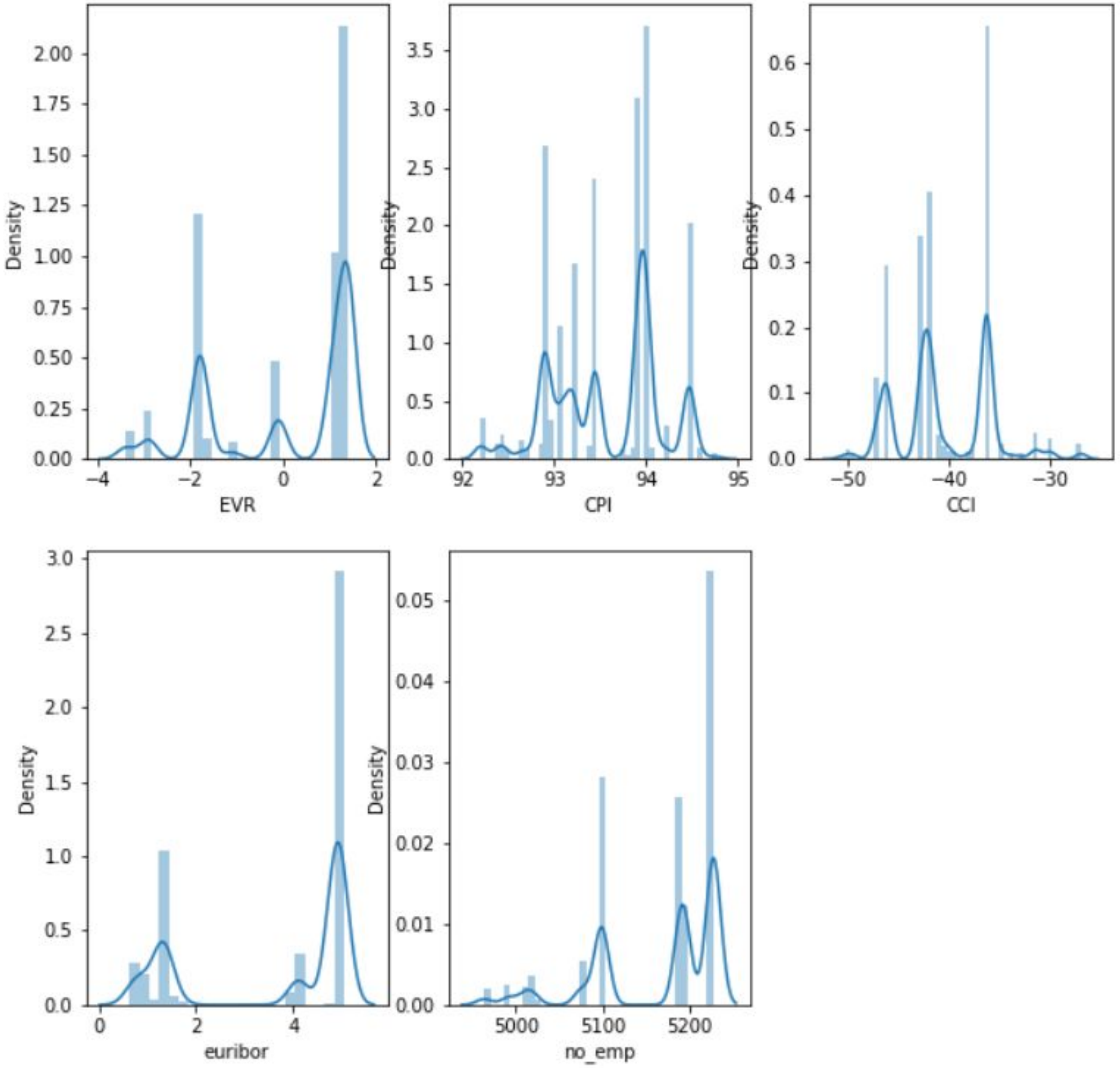
Previous Campaign Outcomes





- Majority of clients open deposit accounts in campaign of non previous campaign outcome.

EVR, CPI, CCI, euribor and no_emp



- We can see there is a high employee variation rate which signifies that they have made the campaign when there were high shifts in job due to conditions of economy
- The Consumer price index is also good which shows the leads where having good price to pay for goods and services may be that could be the reason to stimulate these leads into making a deposit and plant the idea of savings
- Consumer confidence index is pretty low as they don't have much confidence on the fluctuating economy
- The 3 month Euribor interest rate is the interest rate at which a selection of European banks lend one another funds denominated in euros whereby the loans have a maturity of 3 months. In our case the interest rates are high for lending their loans
- The number of employees were also at peak which can increase their income index that could be the reason the campaign targeted the leads who were employed to make a deposit

Model Implementation

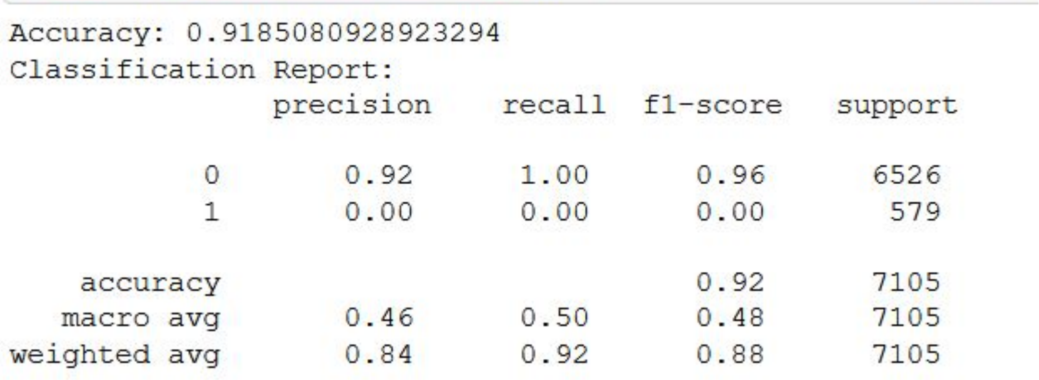
- We have implemented many models like- bagging, boosting, k nearest neighbours etc.

- When we remove the duration feature, the accuracy of the models goes down automatically. The models still predict 0 values quite well but get significantly worse at predicting 1 values which is what the clients need. Looking at the performance of the classifiers, Random Forest is still the best model for the job but it's not a very convincing choice. There's definite scope for improvement here. That could be through gathering more data, improving the methodology to collect the data and even applying advanced feature transformation methods to the same. The next level objective of the modeling should be to try and isolate the 1 values to find more useful insights.

- Let's have a look towards the models with some of the best results

Final Model Recommendation

Support Vector Machine

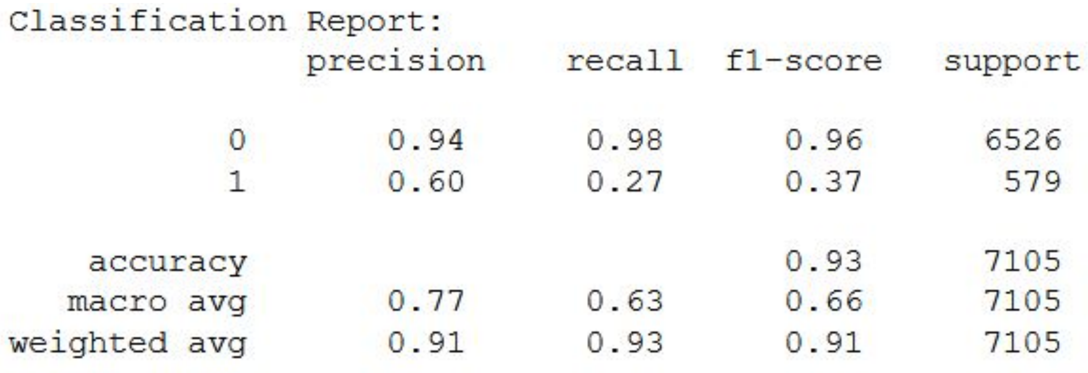


```
Accuracy: 0.9185080928923294
Classification Report:
              precision    recall  f1-score   support

           0       0.92      1.00      0.96      6526
           1       0.00      0.00      0.00       579

    accuracy                           0.92      7105
   macro avg       0.46      0.50      0.48      7105
weighted avg       0.84      0.92      0.88      7105
```

SVC looked like the best estimator in our initial modelling, but looking deeper into the metrics we can clearly see that there is a problem. The SVC is reasonably accurate but completely useless for our purposes. The classifier has really poor recall value and f1-score for the 1 value, making SVC completely inept for what we're trying to do.

Final Model Recommendation

Logistic Regression

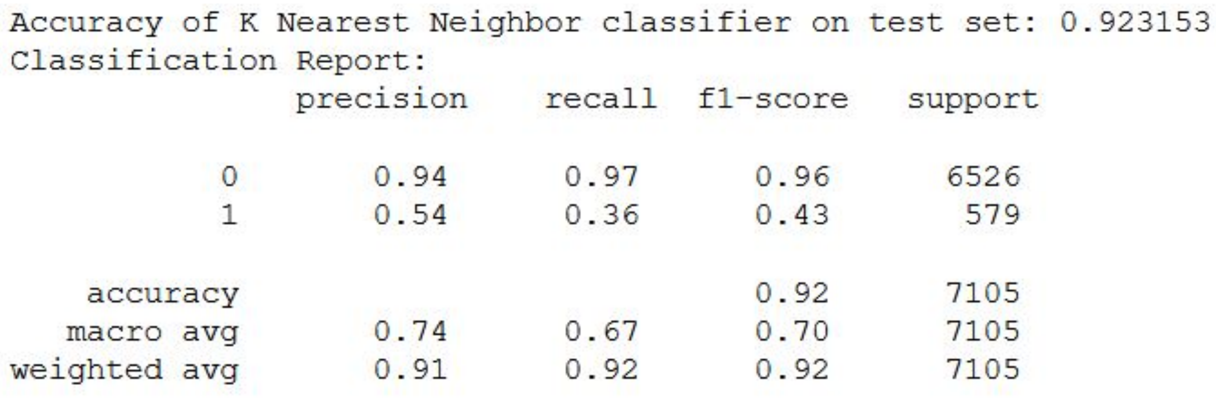


```
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.98      0.96      6526
           1       0.60      0.27      0.37       579

    accuracy                           0.93      7105
   macro avg       0.77      0.63      0.66      7105
weighted avg       0.91      0.93      0.91      7105
```

The Confusion matrix result is telling us that we have 6434+139 correct predictions and 440+92 incorrect predictions. The classification report tells us that the metrics for 0 values are very high but not so good for 1 values. In our case study, the 1 values are the ones that matter because we want to target customers who will be willing to take up the promotions offered. In the case of Logistic Regression here, the precision and the recall values are quite low compared to the accuracy of the overall model.

Final Model Recommendation

KNN

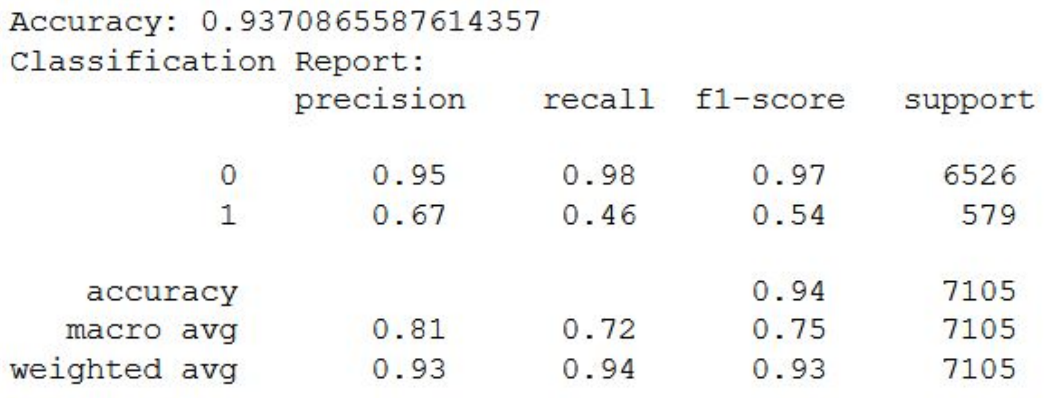


```
Accuracy of K Nearest Neighbor classifier on test set: 0.923153
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.97      0.96      6526
           1       0.54      0.36      0.43       579

    accuracy                           0.92      7105
   macro avg       0.74      0.67      0.70      7105
weighted avg       0.91      0.92      0.92      7105
```

Accuracy is similar to Logistic Regression. The precision value for 1 is lower than logistic regression but the recall and f1-score values are better. As the f1-score is better, we can say that KNN is more useful than Logistic Regression for our modelling purposes.

Final Model Recommendation

AdaBoosting

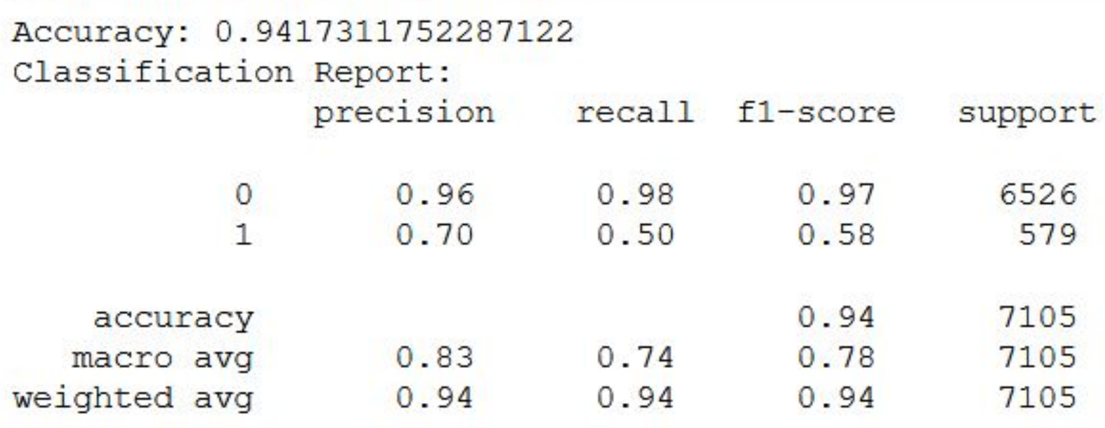


```
Accuracy: 0.9370865587614357
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.98      0.97      6526
           1       0.67      0.46      0.54       579

    accuracy                           0.94      7105
   macro avg       0.81      0.72      0.75      7105
weighted avg       0.93      0.94      0.93      7105
```

AdaBoost does a pretty good job. It's more accurate than Logistic Regression, KNN and SVC with better precision, recall and F1-scores.

Final Model Recommendation

Bagging: Random Forest



```
Accuracy: 0.9417311752287122
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.98      0.97      6526
           1       0.70      0.50      0.58       579

    accuracy                           0.94      7105
   macro avg       0.83      0.74      0.78      7105
weighted avg       0.94      0.94      0.94      7105
```

A glance at the classification report here tells us that Random Forest is much more useful than any model we've encountered so far. It's more accurate and predicts 1 values much more precisely than any model we've used before.

Thank You

Data Glacier

Your Deep Learning Partner