Problem Definition

- Bank marketing campaign by one of the Portuguese banking institution
- Based on phone calls from 2008 to 2010
- 11162 data sample
- 17 Features
- Objective is to predict whether the client will subscribe to a term deposit or not

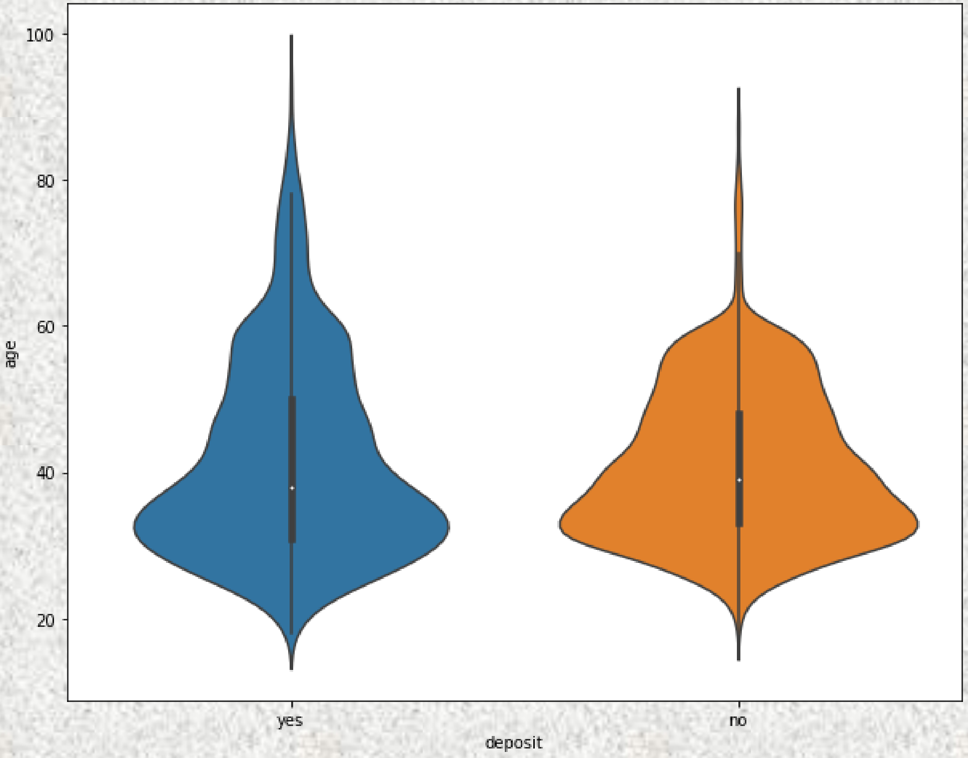
Methodology

Tools Used

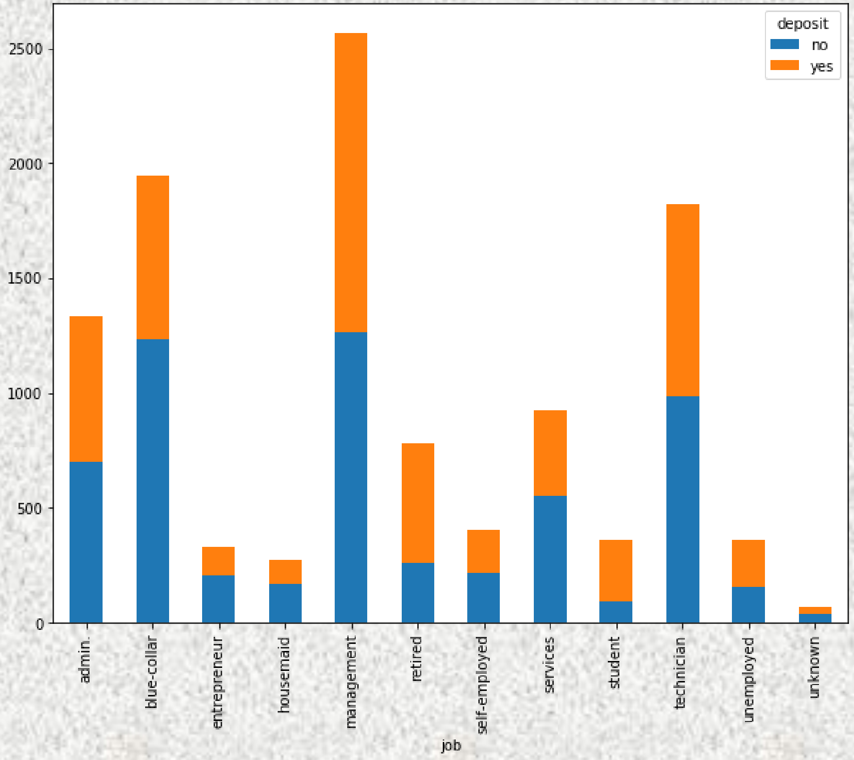
SN	Tool	Description	Version
1	Programming Language	Python	3.7.3
2	Distribution Platform	Anaconda	1.9.7
3	Environment	Jupyter Notebook	5.7.8

Age , Job (Feature Description and EDA)

Age – Age of Client

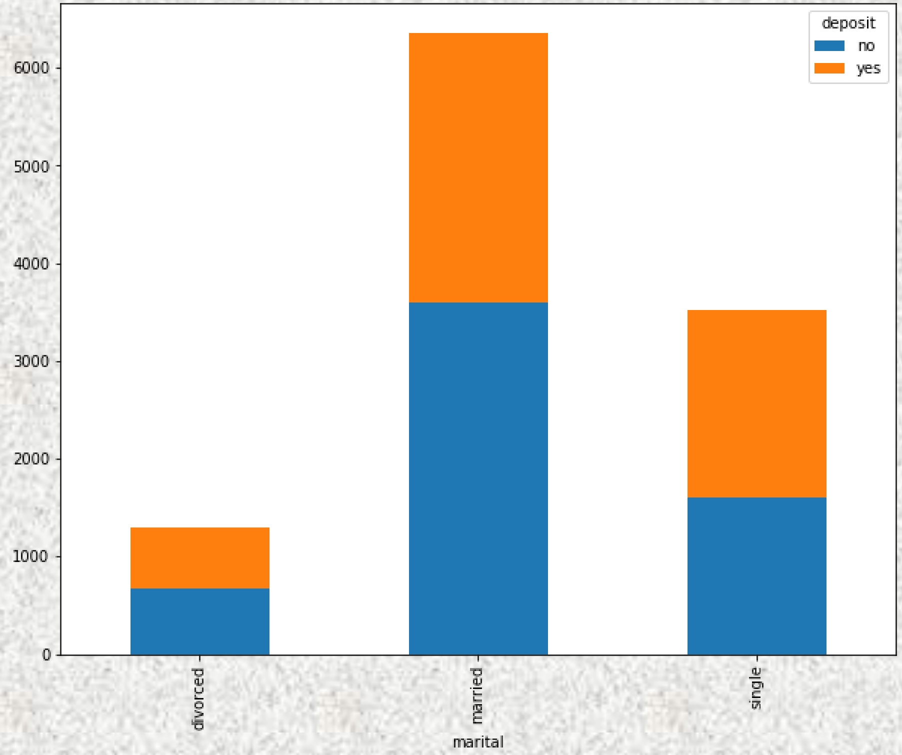


Job – Job of client

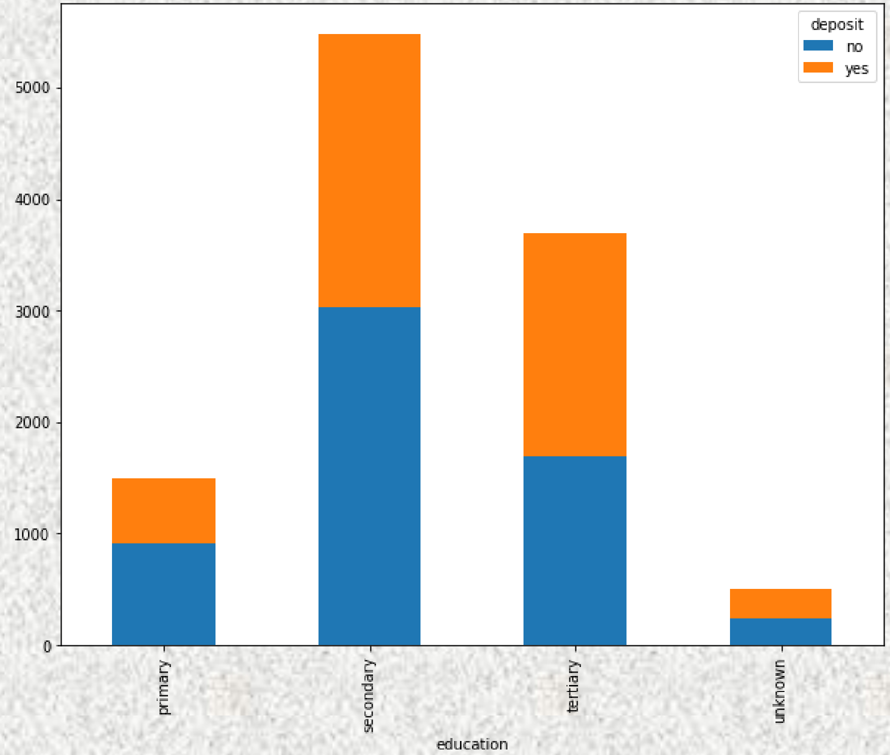


Marital , Education

Marital – Marital Status of Client

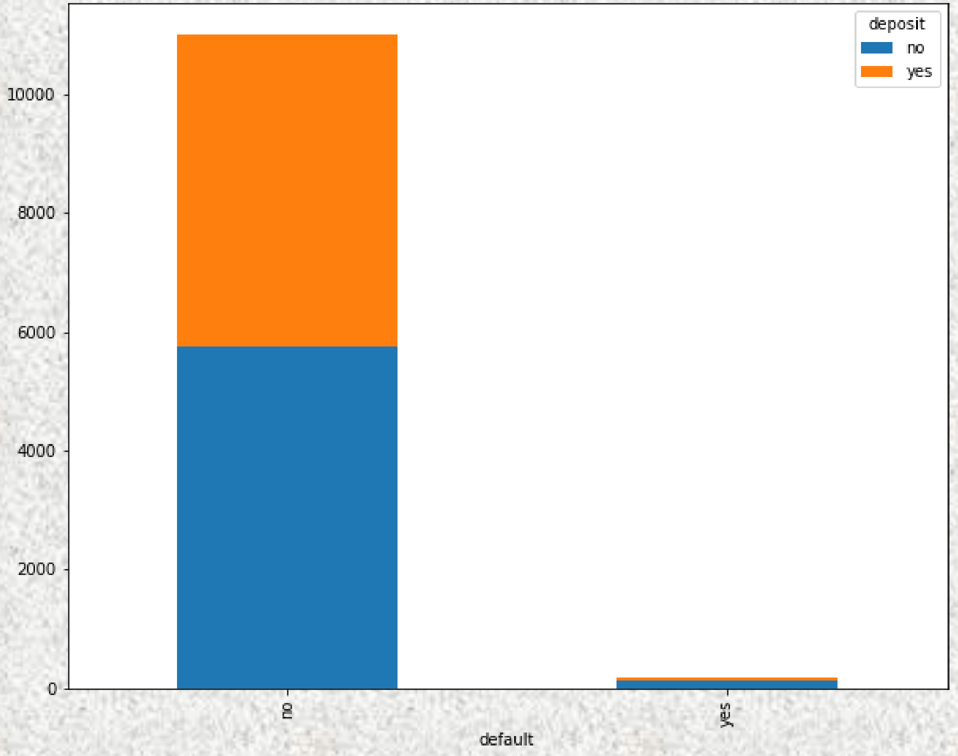


Education – Education type of Client

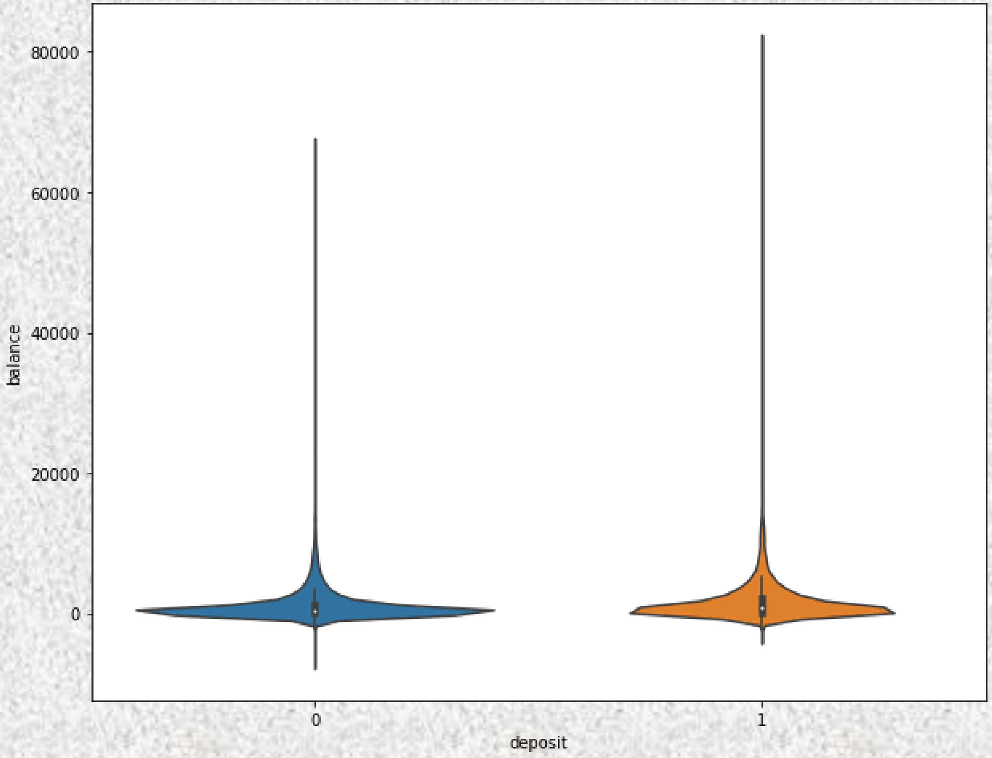


Default , Balance

Default - whether the client has credit in bank or not?

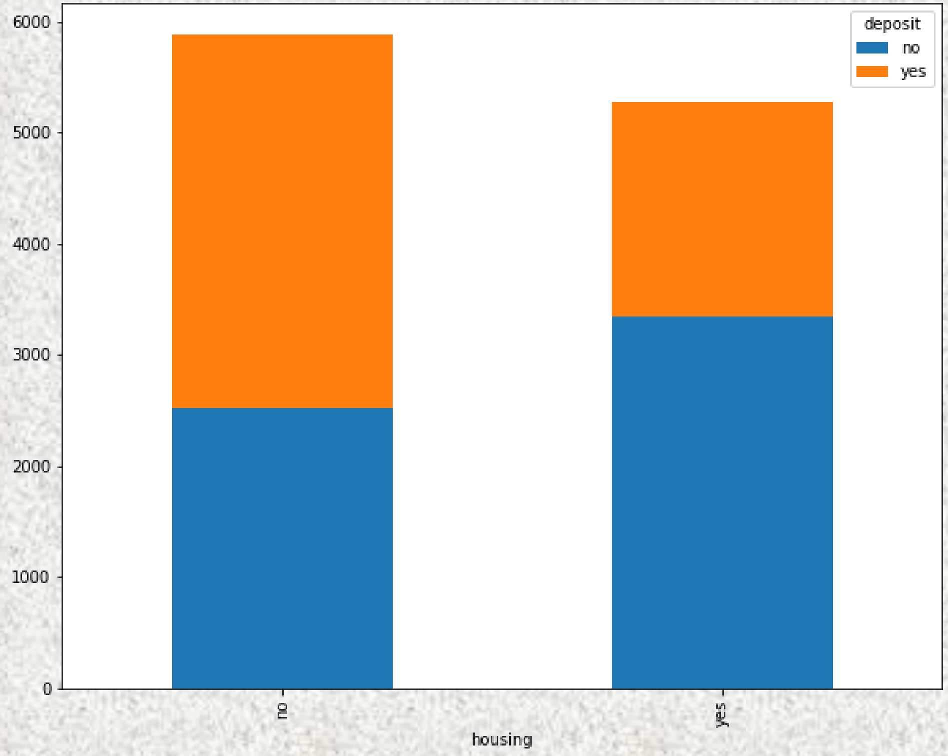


Balance – balance in bank account

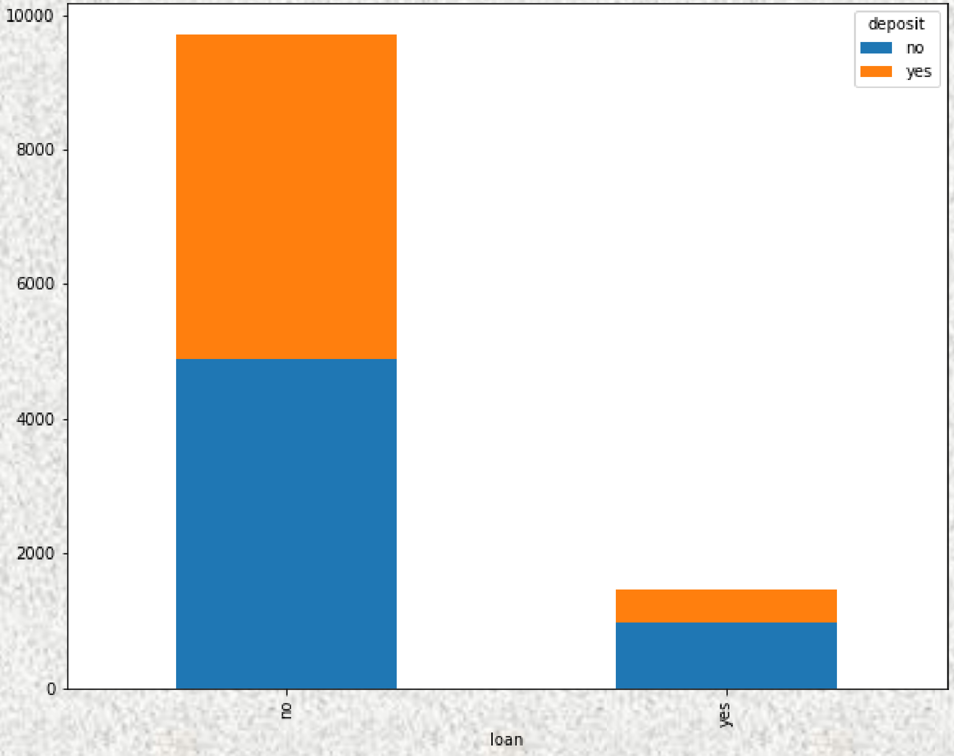


Housing , Loan

Housing - Whether the client has got any housing loan from bank?

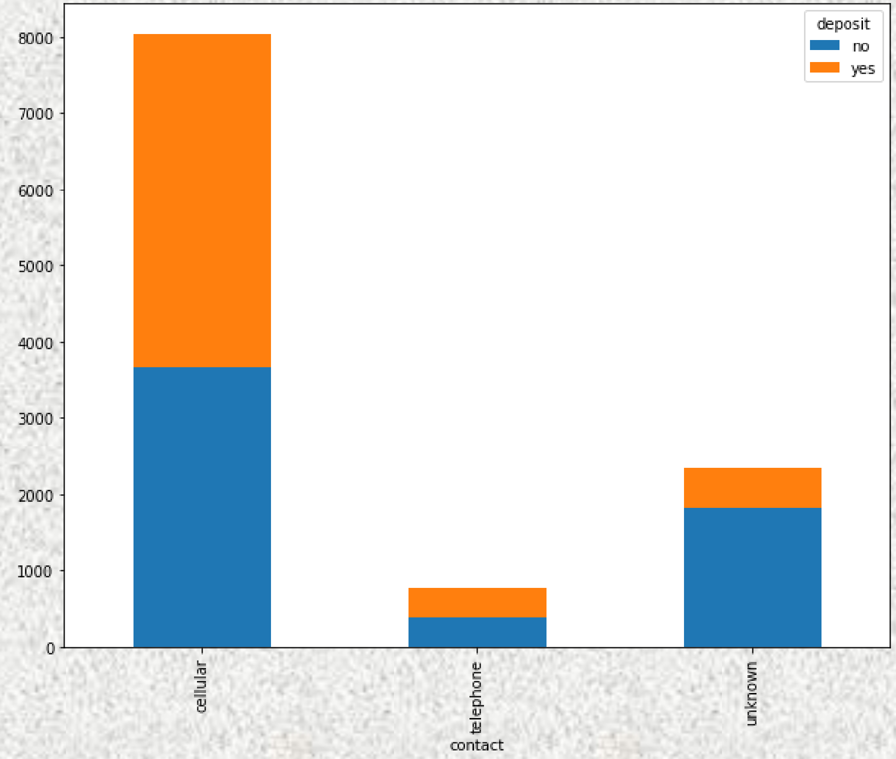


Loan - Whether the client has got any personal loan from bank?

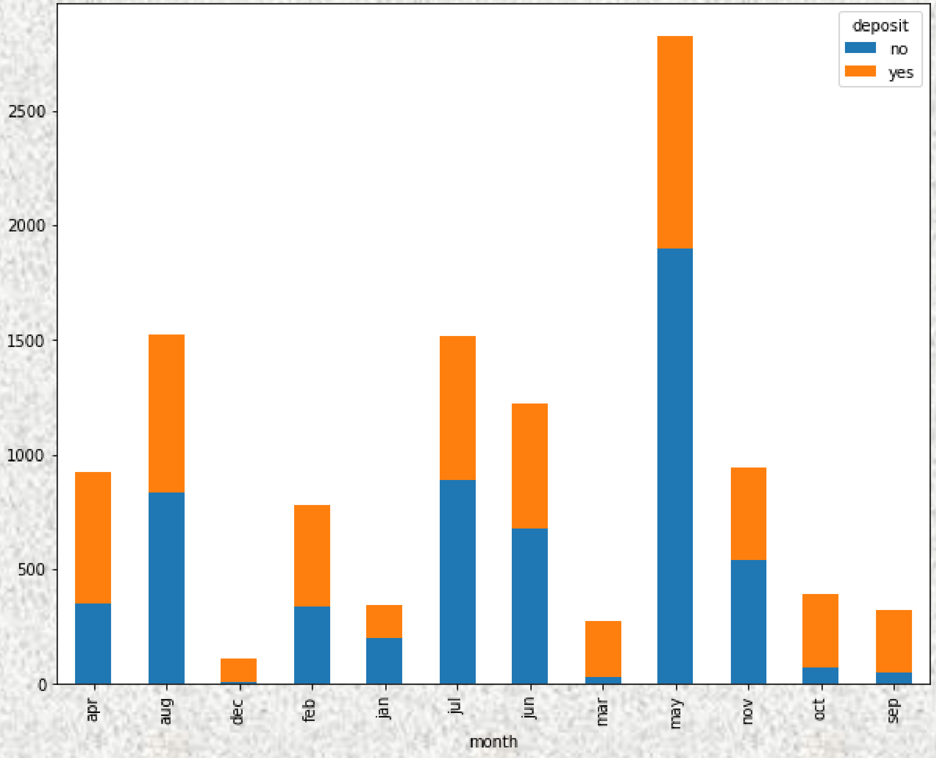


Contact , Month

Contact - way of communication

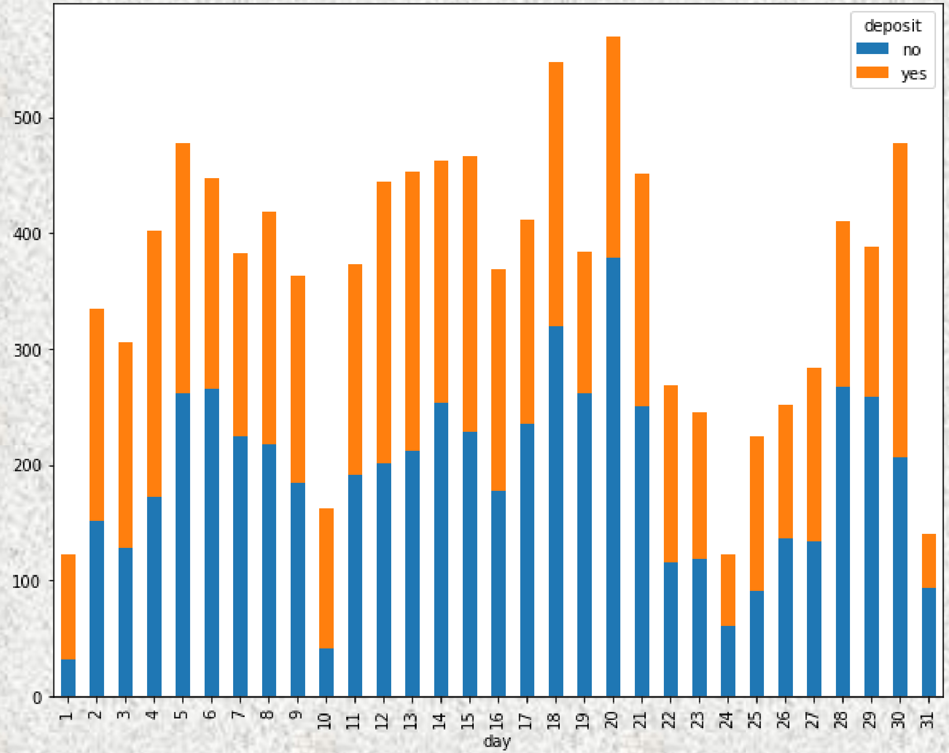


Month – month of communication

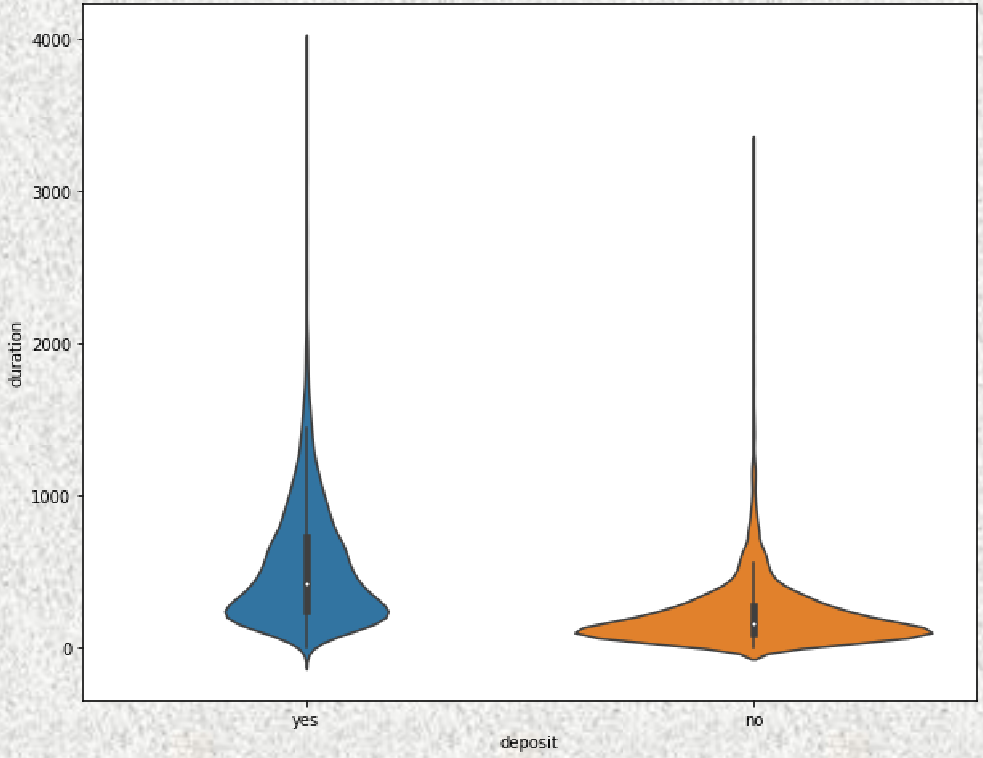


Day , duration

Day – day of the month for contact

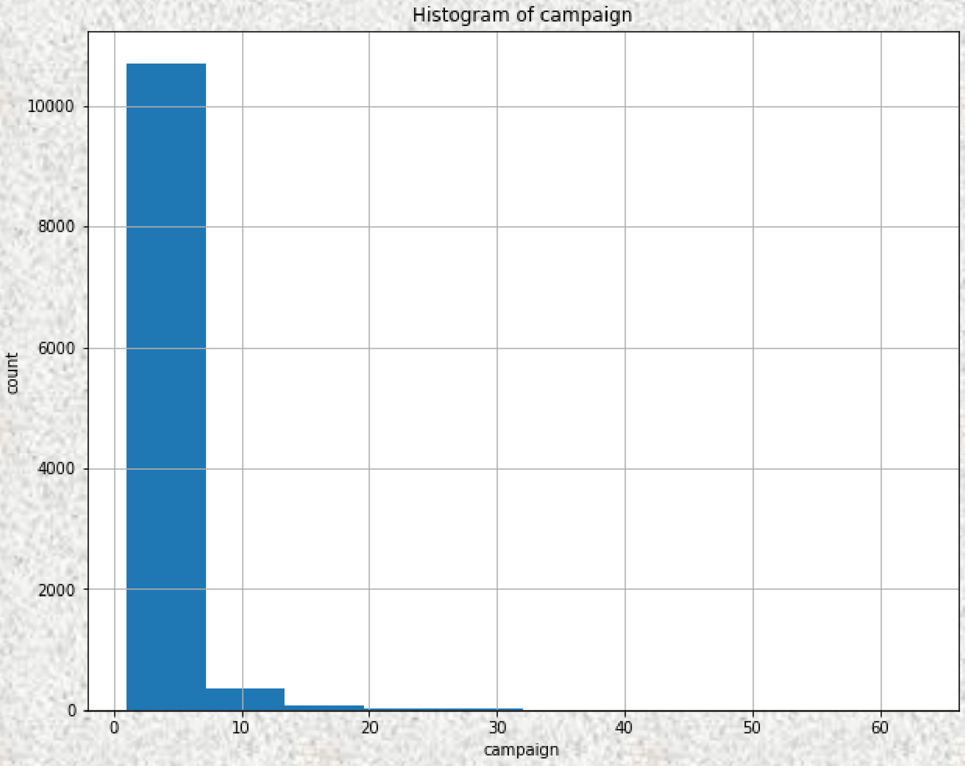


Duration – duration of last call

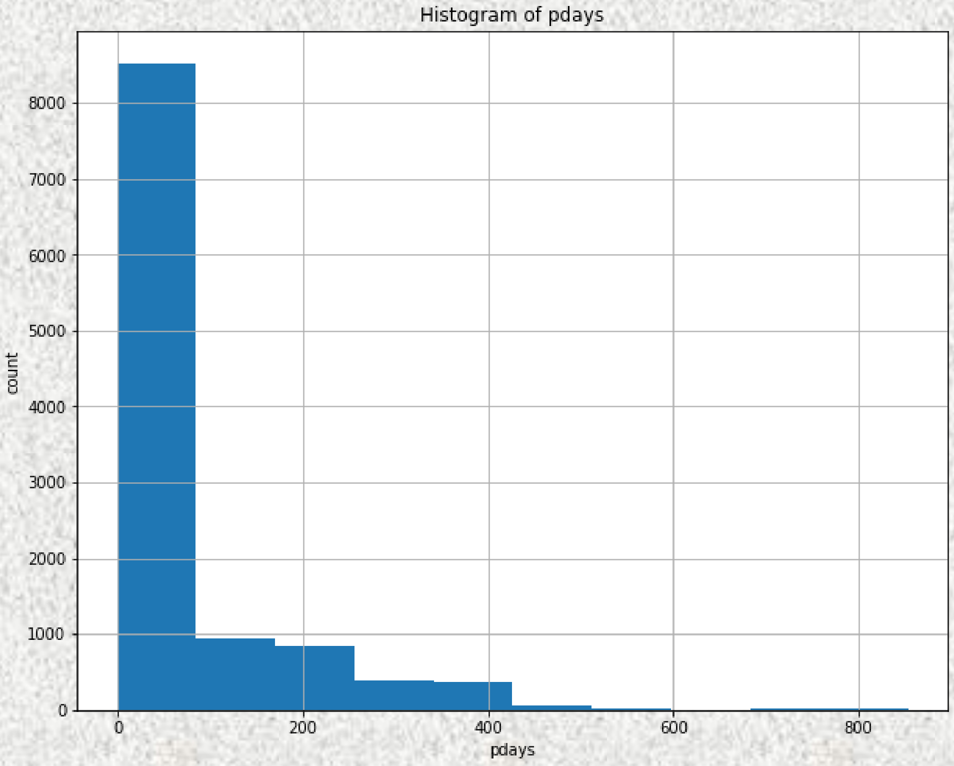


Campaign , pdays

Campaign - Number of times this client was contacted during this campaign

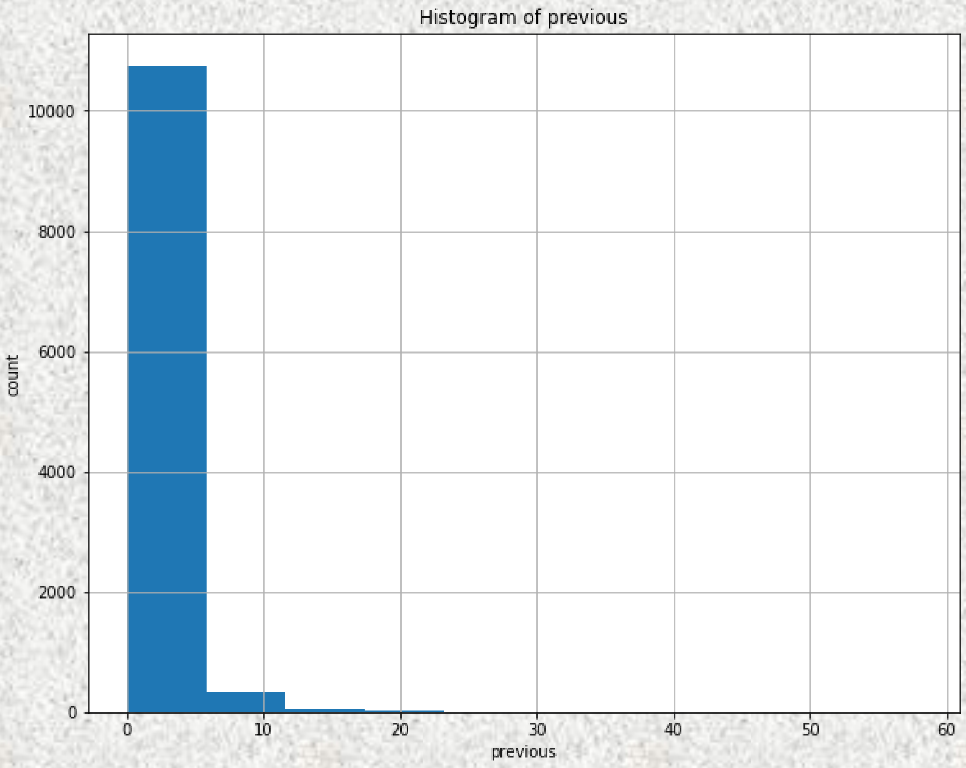


Pdays- number of days that passed after the client was last contacted in previous campaign.

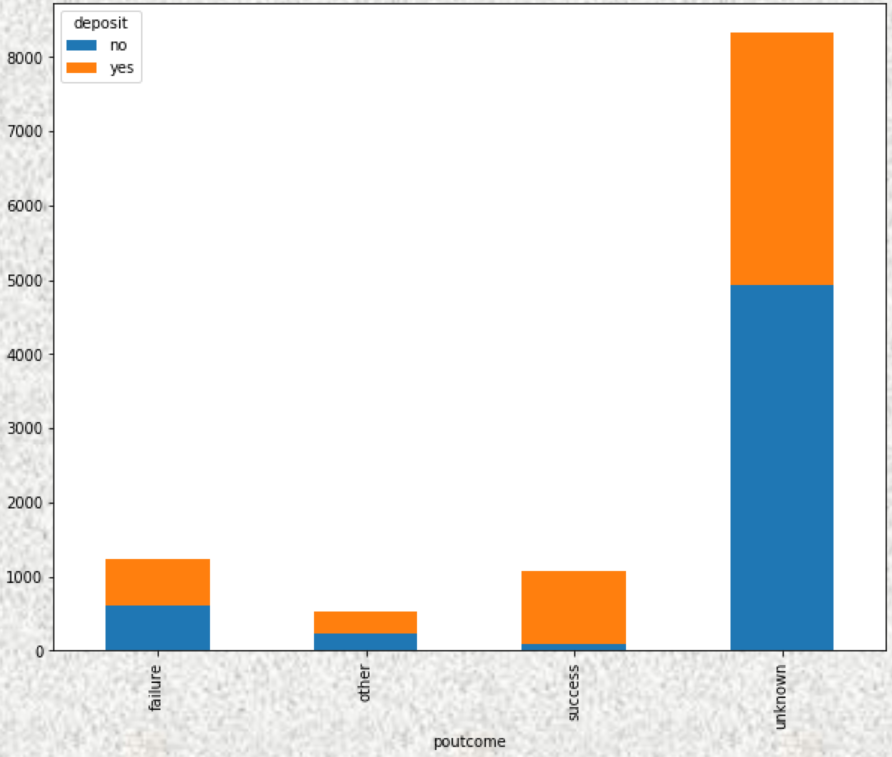


Previous , poutcome

Previous - Number of times this client was contacted before this campaign

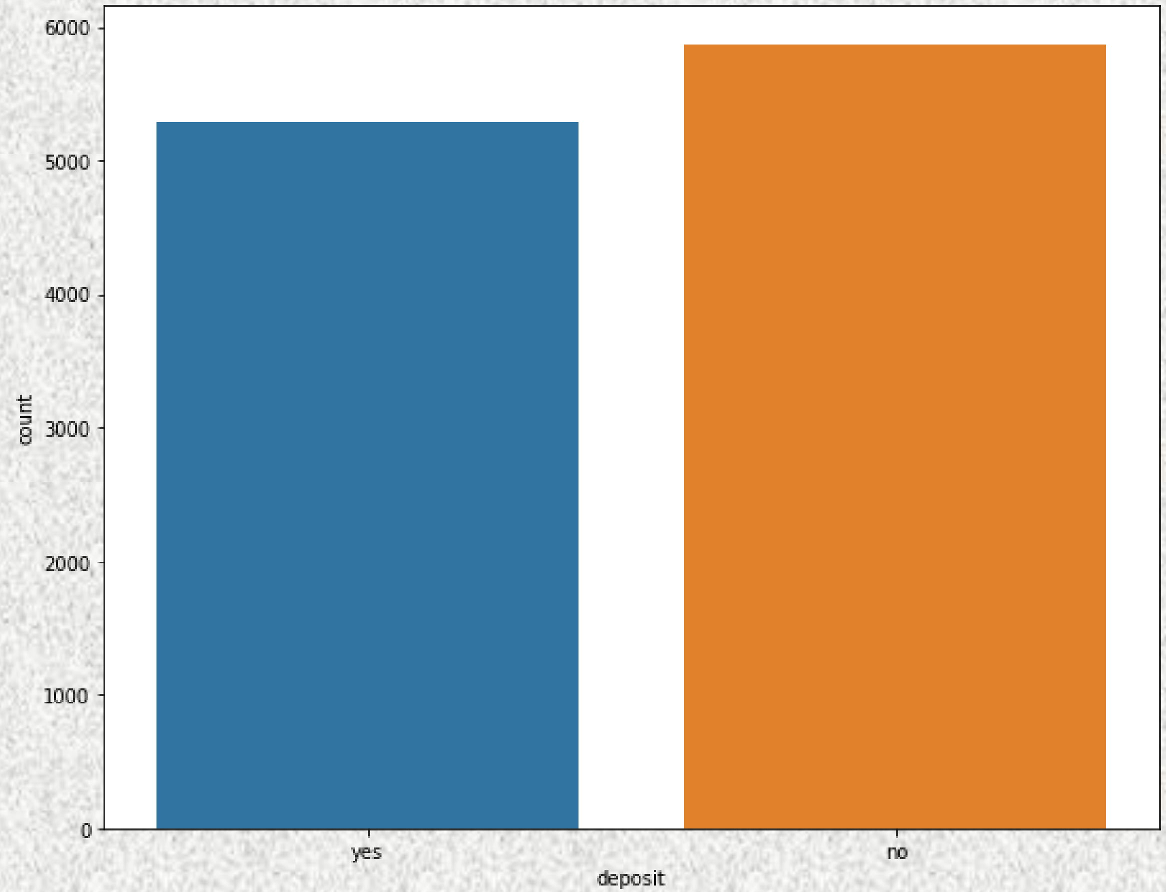


Poutcome - The outcome of previous marketing campaign

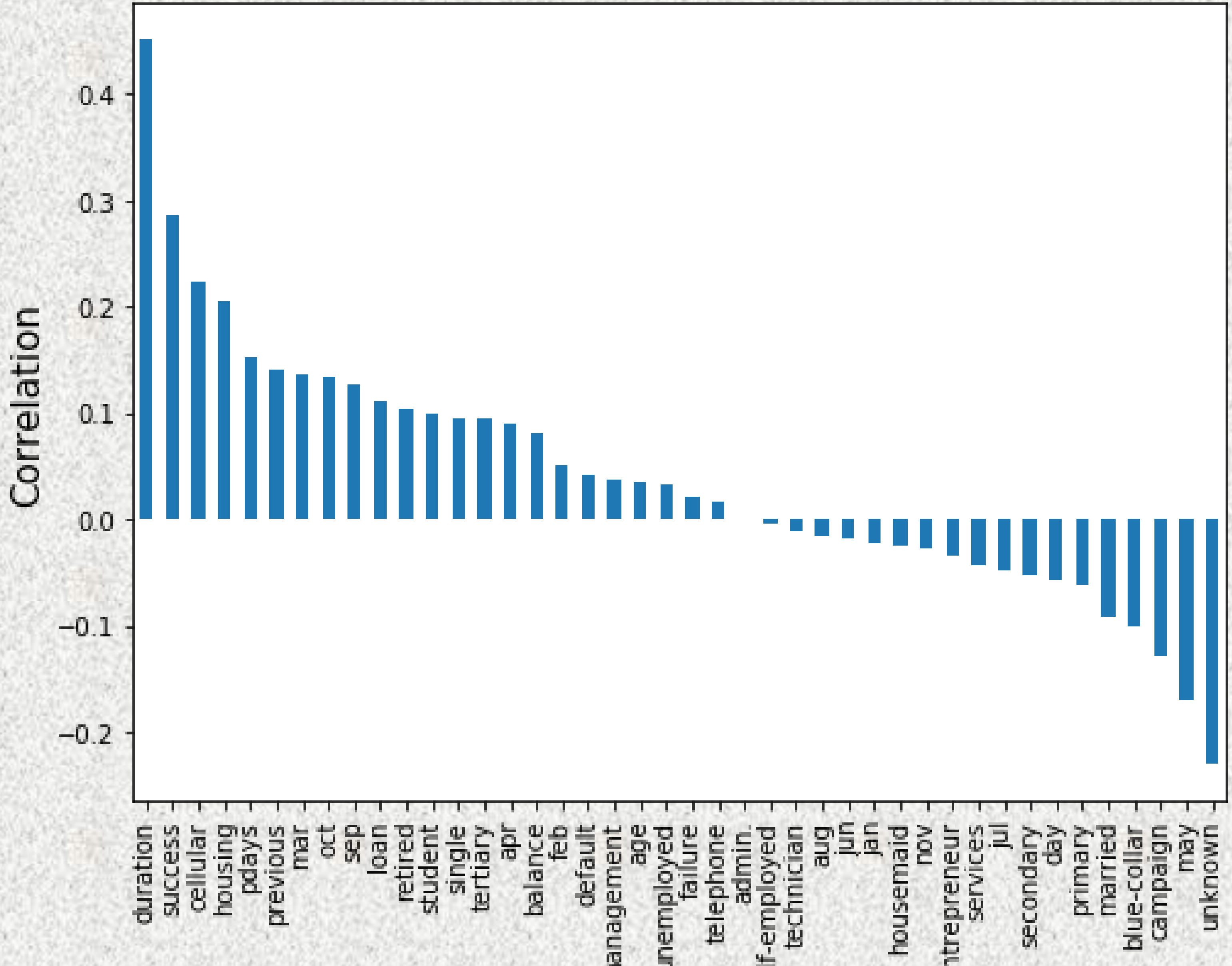


Deposit (target variable)

- Deposit - Whether the clients said yes to subscribe for a term deposit



Correlation With Target Variable



Correlation With Target Variable

SN	Top features with +ve correlation		Top features with -ve correlation	
	Feature	Correlation	Feature	Correlation
1	duration	0.451919	Unknown	-0.230470
2	Success	0.286642	May	-0.170507
3	Cellular	0.223252	Campaign	-0.128081
4	housing	0.203888	blue-collar	-0.100840
5	Pdays	0.151593	married	-0.092157

Preprocessing of Categorical Variables

With two categories

- Replace two categories with 0 and 1

```
bank['housing'].replace(to_replace='no', value=1, inplace=True)
bank['housing'].replace(to_replace='yes', value=0, inplace=True)
```

With multiple categories

- One hot encoding

```
one_hot = pd.get_dummies(bank['month'])
bank = bank.drop('month',axis = 1)
bank = bank.join(one_hot)
bank = bank.drop('dec',axis = 1)
```

Algorithms Used

- K Nearest Neighbours
- Naïve Bayes
- Logistic Regression
- Logistic Regression with polynomial Features
- Support vector machines
- Decision Trees
- Random Forest
- XGBOOST (GBDT)

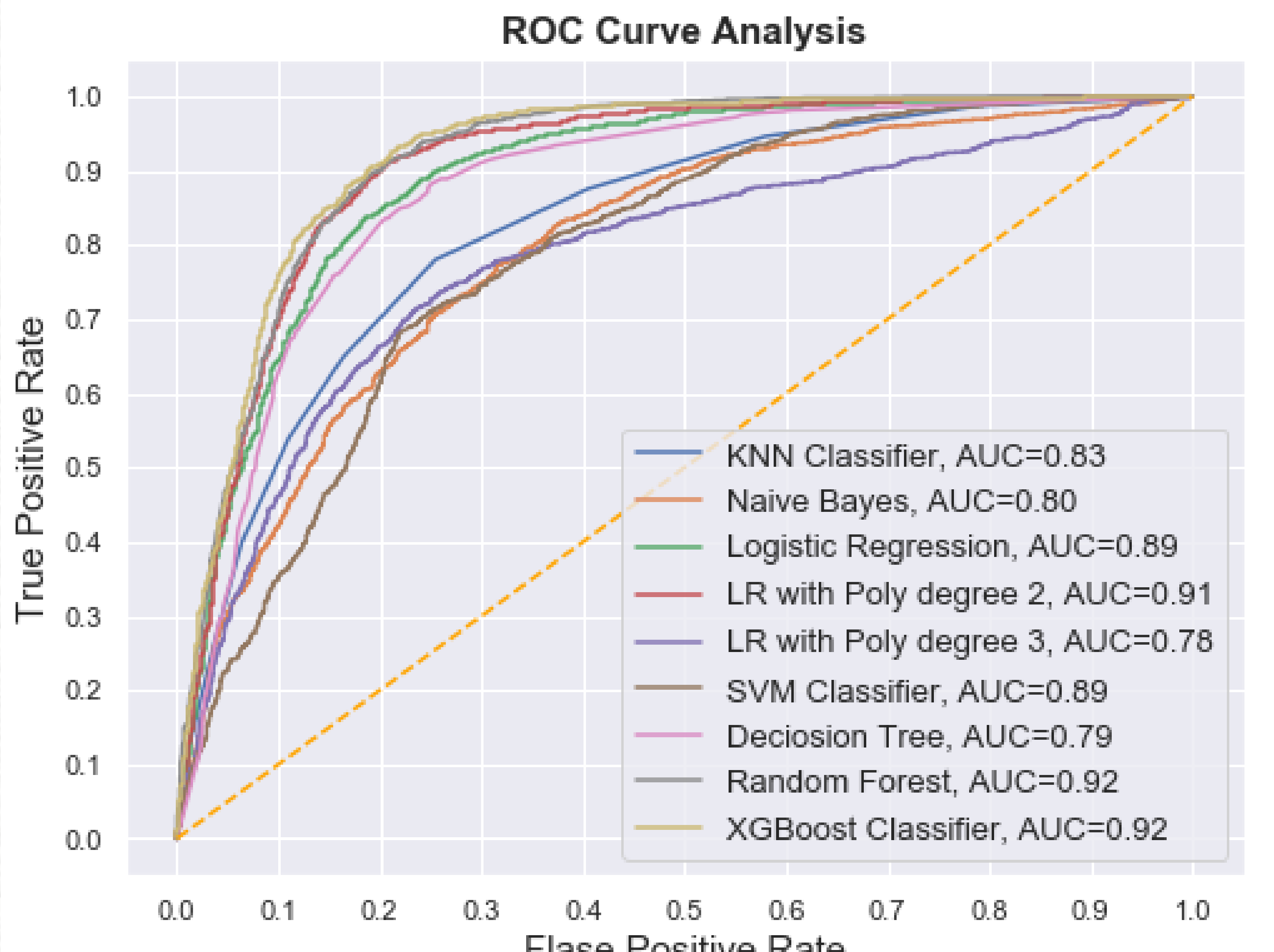
Best Parameters with gridsearchcv

SN	Algorithm	Best parameters
1	KNN	K=9
2	Naïve Bayes	Does it internally
3	Logistic Regression	'C': 1.623776739188721, 'max_iter': 100, 'penalty': 'l2', 'solver': 'saga'
4	LR with polynomial Feature degree 2	'C': 1, 'max_iter': 5000, 'penalty': 'l1', 'solver': 'saga'
5	LR with polynomial Feature degree 3	'C': 1, 'penalty': 'l2'
6	SVM	'C': 1, 'kernel': 'linear'
7	Decision Tree	'criterion': 'gini', 'max_depth': 8, 'min_samples_leaf': 5, 'min_samples_split': 3
8	Random Forest	'bootstrap': False, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 2000
9	XGBOOST	'colsample_bytree': 0.83, 'eta': 0.20000000000000004, 'gamma': 5, 'min_child_weight': 4.0, 'subsample': 0.94

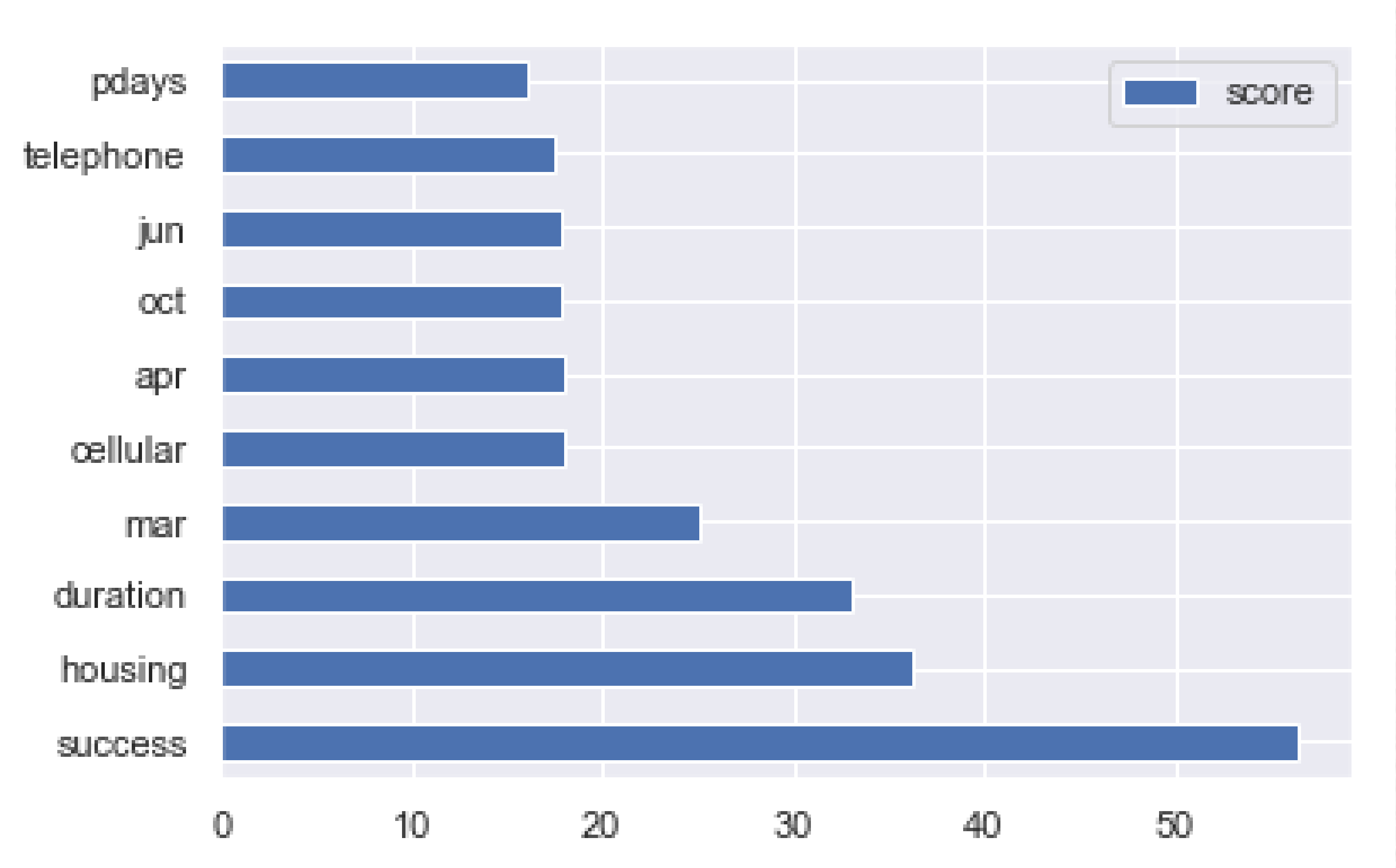
Comparison of Results

SN	Algorithm	Recall	Accuracy	AUC
1	KNN	0.68	0.74	0.83
2	Naïve Bayes	0.53	0.70	0.80
3	Logistic Regression	0.79	0.81	0.89
4	LR with polynomial Feature degree 2	0.83	0.84	0.91
5	LR with polynomial Feature degree 3	0.64	0.83	0.78
6	SVM	0.89	0.82	0.89
7	Decision Tree	0.82	0.81	0.79
8	Random Forest	0.88	0.84	0.92
9	XGBOOST	0.88	0.85	0.92

ROC Curve


ROC Curve Analysis
True Positive Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Flase Positive Rate
KNN Classifier, AUC=0.83
Naive Bayes, AUC=0.80
Logistic Regression, AUC=0.89
LR with Poly degree 2, AUC=0.91
LR with Poly degree 3, AUC=0.78
SVM Classifier, AUC=0.89
Deciosion Tree, AUC=0.79
Random Forest, AUC=0.92
XGBoost Classifier, AUC=0.92

Top 10 Important Features



Conclusions

This sums up for the classification task of bank marketing dataset. We find that XGBOOST gives us the best value for accuracy which is 0.85 while SVM gives us the best value for recall which is 0.89. The best AUC score of 0.92 comes from Random Forest as well as XGBOOST. The results of K nearest neighbours and Naïve Bayes are less while rest of the algorithms are giving more or less same result with minor differences.

As per the feature importance of XGBOOST it is clear that bank need to focus more on clients with success in previous campaign. Whether client uses cellular phone or not and the month in which client is being called play a vital role and the strategies of marketing campaign should be decided accordingly.