Bank Marketing

Which Customer to target?

By Jyothirmayee Nagireddy

Yes

Marketing Campaign

Investment

Statistics

Safety

Term Deposit

Finance

Banking

Rate of Interest

Client

No

Safe

The data is related to the direct marketing campaigns (phone calls) of a portuguese banking institution

The Problem

- Improve the marketing campaign by analyzing customer data & past marketing campaign and recommend which customer to target.
- Challenge : Skewed data.
  Classification problem with an imbalance Ratio of 10%.
- Built a predictive model that gives insights into which customer to target with low false positives and false negatives.

The Data

16 feature attributes divided into three groups

- Client Information
- Related to last contact of the previous campaign
- Related to current campaign

1 target attribute (y) - (yes/no)

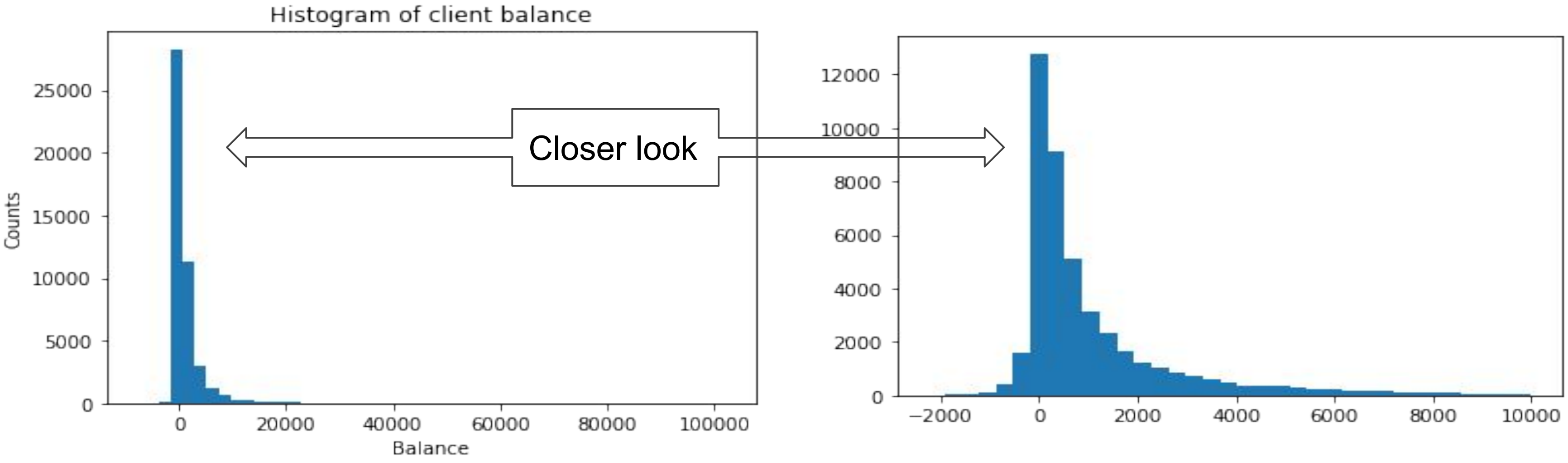
Data Wrangling

- No missing values
- Outliers were kept for further analysis
- Clean data that requires no pivoting or melting.
- Various statistics were computed on all the columns using .describe() method.

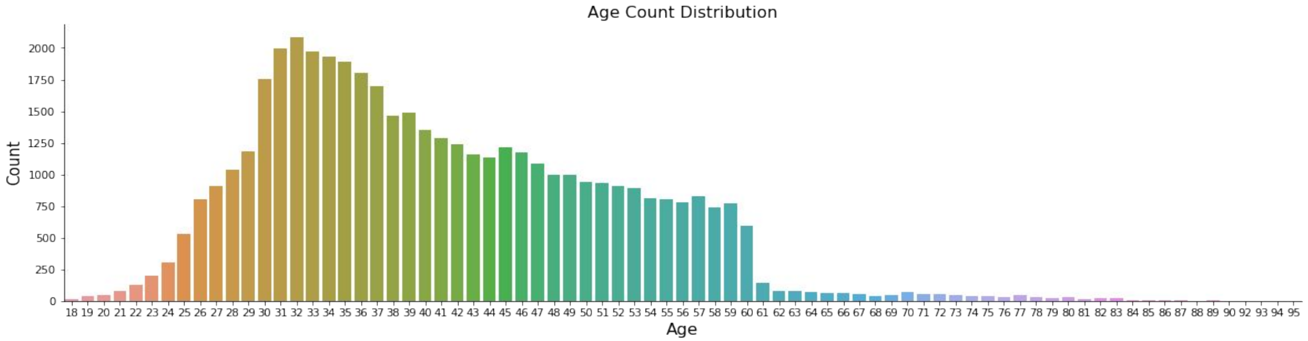
45,211 instances for training.
4521 instances of new unseen data for testing.

Data Storytelling

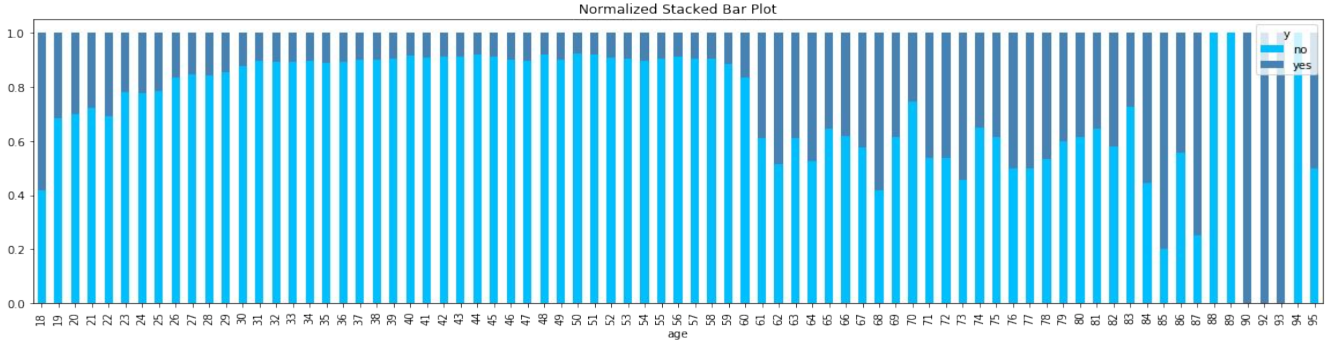
- Balance in a client's account ranges from a minimum of -8019 to a maximum of 102127. But 25th and 75th percentiles range between 72 to 1428 euros.
- Out of 7280 clients who had negative balance, 502 clients subscribed for a term deposit. Hence, even clients who had negative balances are statistically important. This attribute has a lot of outliers.



AGE



- Distributed between the ages of 18 and 95
- Mostly middle aged
- Mean age 40yrs
- STD - 10

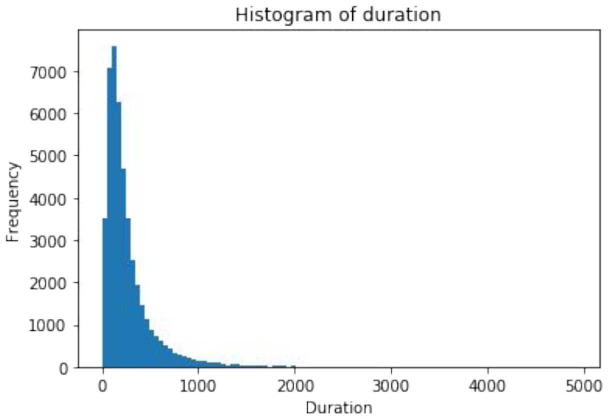


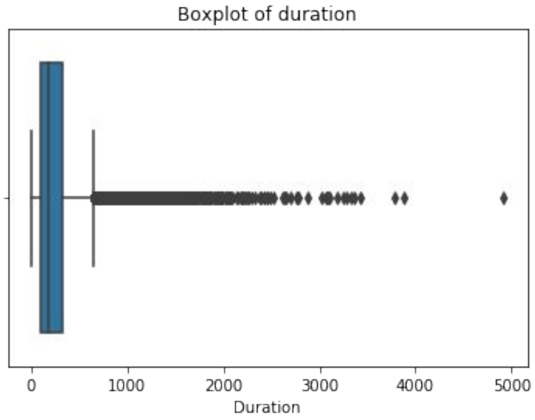
<u>Ages</u>

- Below 22
- Above 60

Have higher tendency to opt for a term deposit

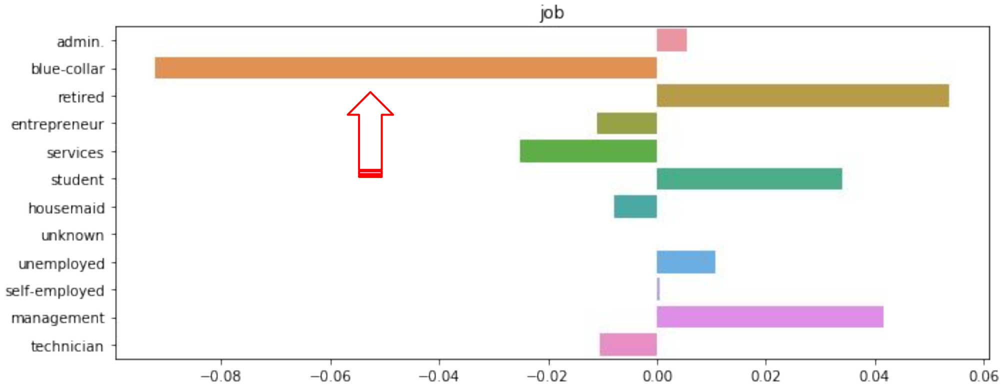
Duration of the call





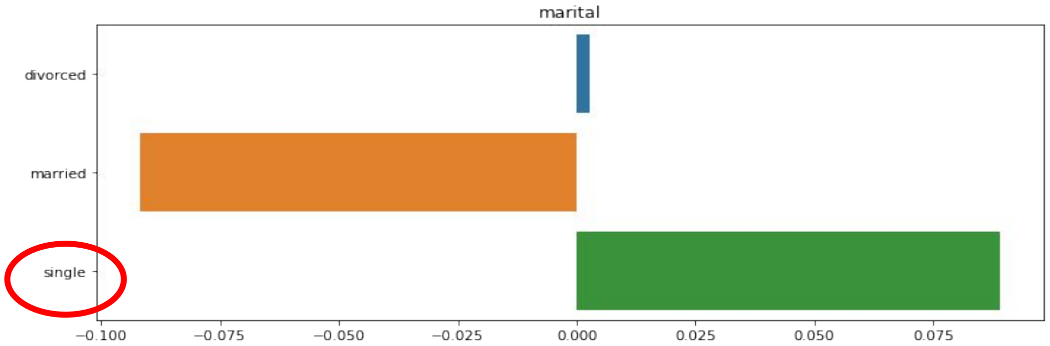
- Most of the data is skewed towards the left.
- Ranges between 0 and 4918 seconds.
- The data below the 75th percentile (319 seconds) gives a clearer picture.
- The call duration is usually about 5 minutes. But occasionally it got a little higher. And the maximum value of 4918 s could be an outlier.

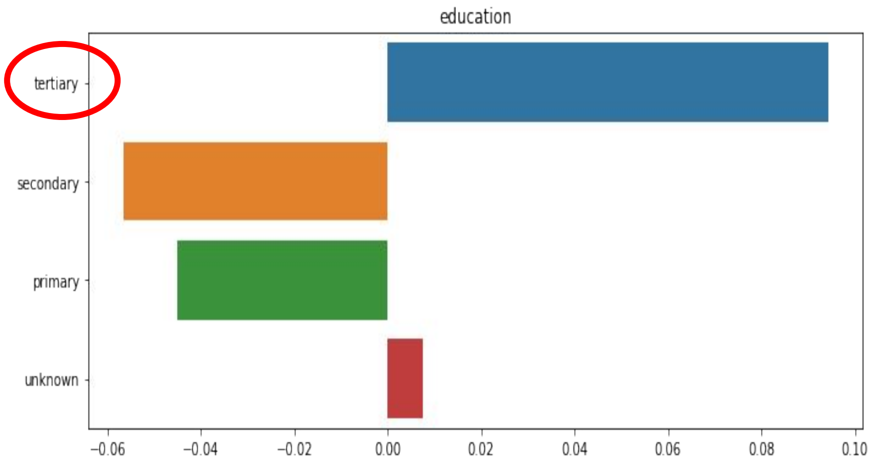
Recommendations based on EDA

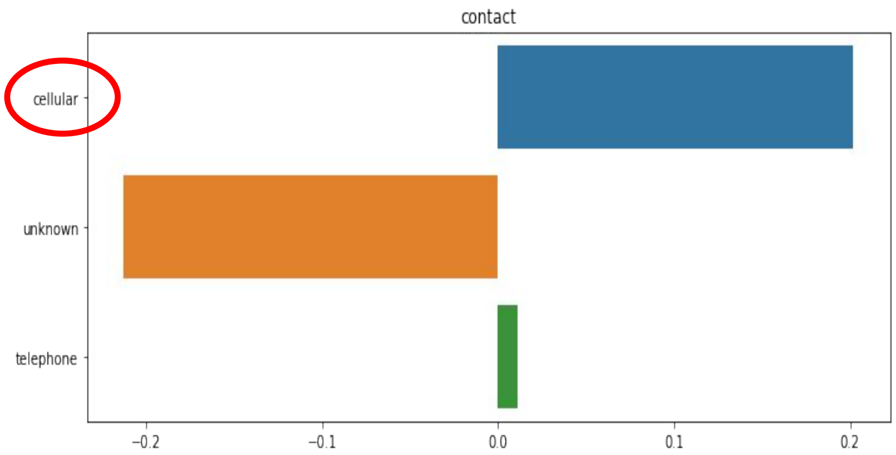


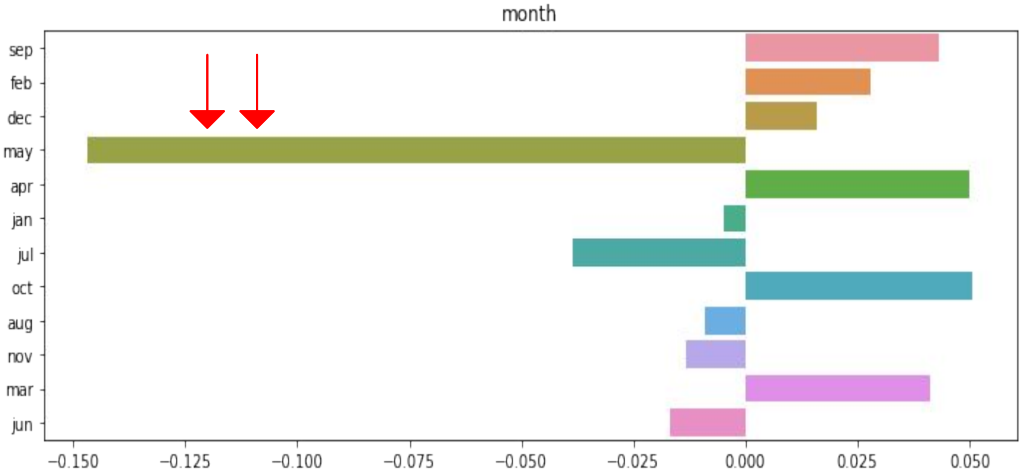
Target
- Retired
- Management
- students

Target single clients

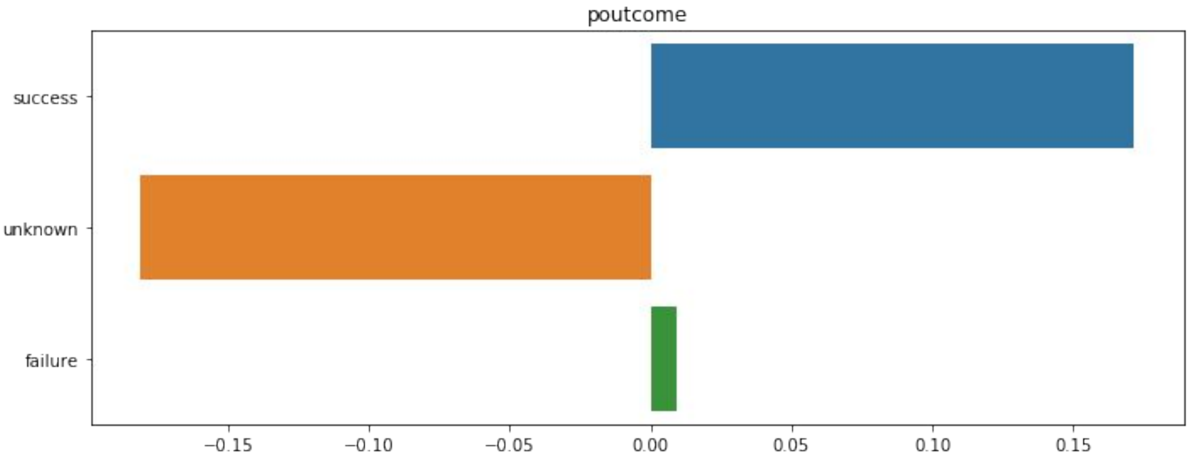






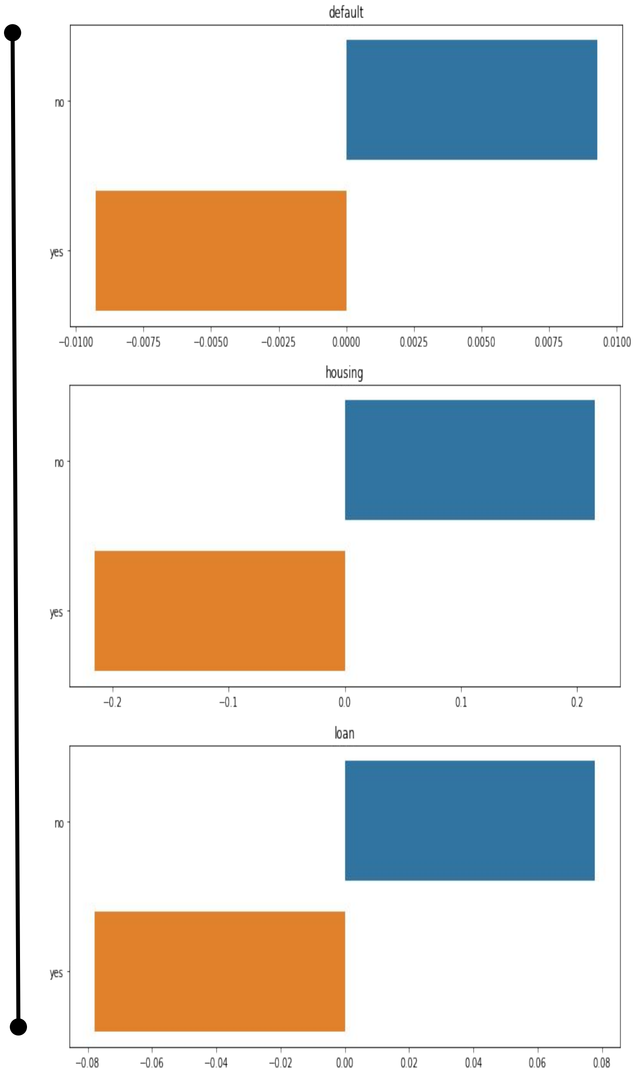


- Target more educated clients.
- Cellular contact is the best
- Try not to contact in the holiday season



Clients with success from previous campaign are more likely to subscribe to a term-deposit again.

The clients with no loans are more likely to subscribe to a term-deposit.



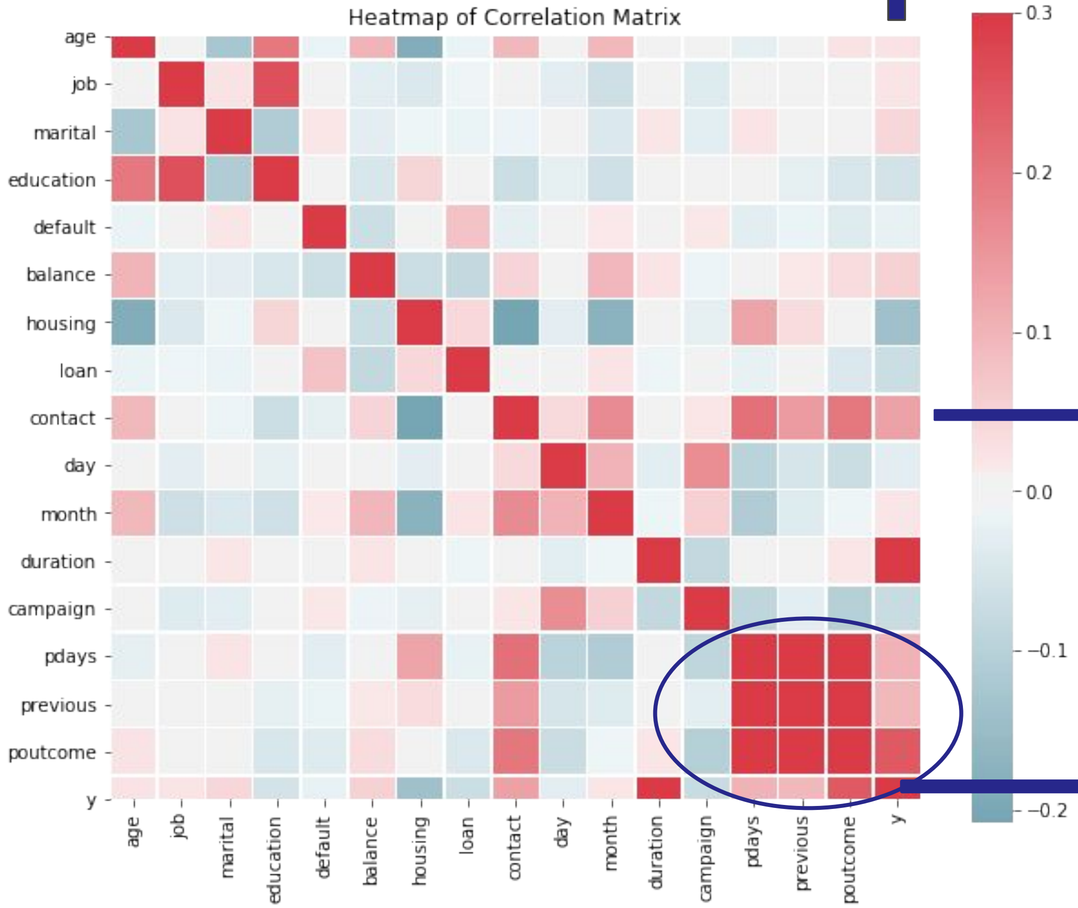
Inferential Statistics

T-test on three attributes helped us reject the Null Hypothesis with a p-value less than 10^-5.

The following attributes were tested with the group that caused ‘yes’ and the group that caused ‘no’ in the target variable:

- Age
- Balance
- Duration

One more Hypothesis test on ‘Balance’ to test the statistical significance between the group with positive balance and the group with negative balance. Null Hypothesis rejected with P-value less than 10^-300.



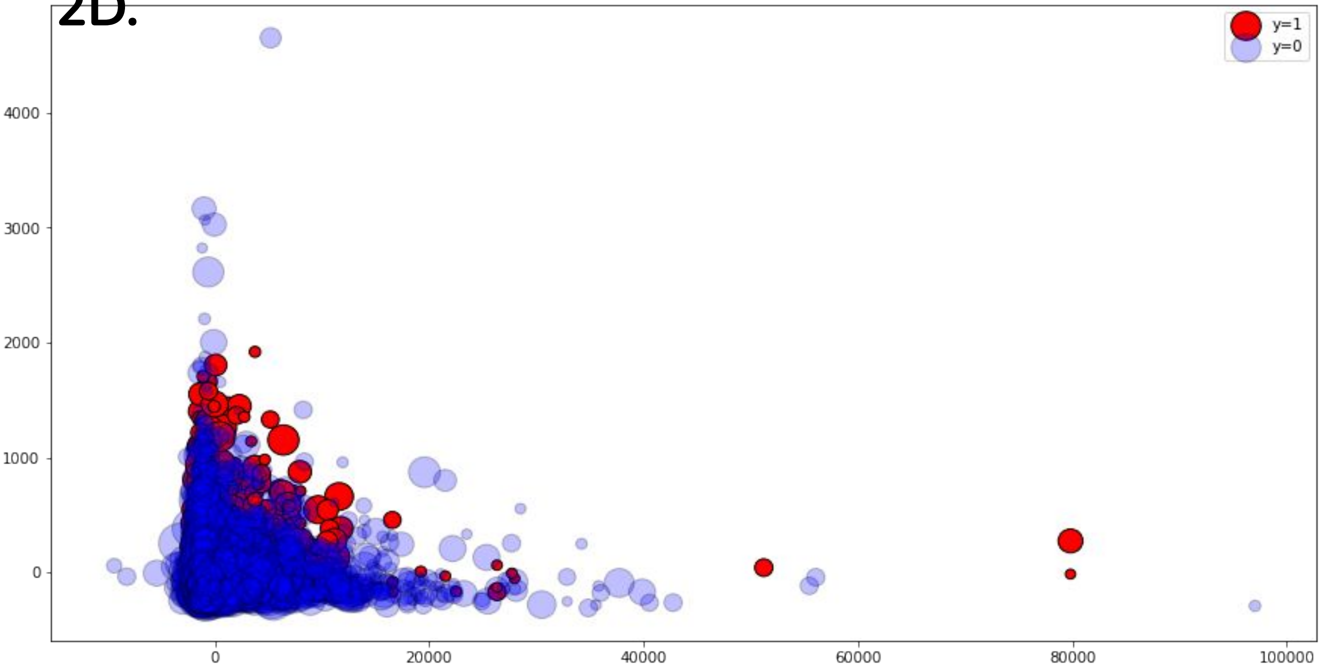
Target variable 'y' has strong positive correlation with

- Duration
- Poutcome
- Contact

The feature contact has correlations with a large number of other features

The features related to the previous campaign are highly correlated.

Principal Component Analysis (PCA) for Visualization of the data in 2D.



- Clients who did **not** subscribe for a term deposit.
- Clients who subscribed for a term deposit.

Visualization through PCA gives us a view about how skewed our data is. On a plot, we can understand our data much better.

Data Preprocessing

- Manually treating the categories in the categorical columns.
- PCA to explain variance-covariance structure of a set of variables.
- Upsampling the minority class
- StandardScalar() from sklearn
- Test-train split (1:4)
- K fold split for cross validation.

Models built:

- Linear regression
- Knn
- Support Vector Machine (SVM)
- Decision Tree
- Random Forest
- Extreme Gradient Boosting
- Gradient Boosting Classifier

Models	MCC
Random Forest Classifier	0.941
Decision Tree Classifier	0.927
K-Near Neighbors	0.863
Support Vector Machine	0.732
Gradient Boosting	0.722
XGBoost	0.718
Logistic Model	0.607

Best Model

Random Forest Classifier

Training Accuracy score : 100.0

Confusion matrix : [31995 0]
[0 31880]

Recall - train : 1.0

MCC : 1.0

Testing Accuracy score : 97.0

Confusion matrix : [7450 477]
[6 8036]

Recall - test : 0.999

MCC : 0.941

- Test Metric : Matthew's correlation coefficient (MCC)
- MCC returns a value between -1 (poorly fitted model) and 1 (best model)
- Highest MCC and lowest false negatives, false positives for Random Forest classifier

Hyper-Parameter tuning

Best Estimators for Parameters Tuned:

'n_estimators': 600,

'min_samples_split': 5,

'min_samples_leaf': 1,

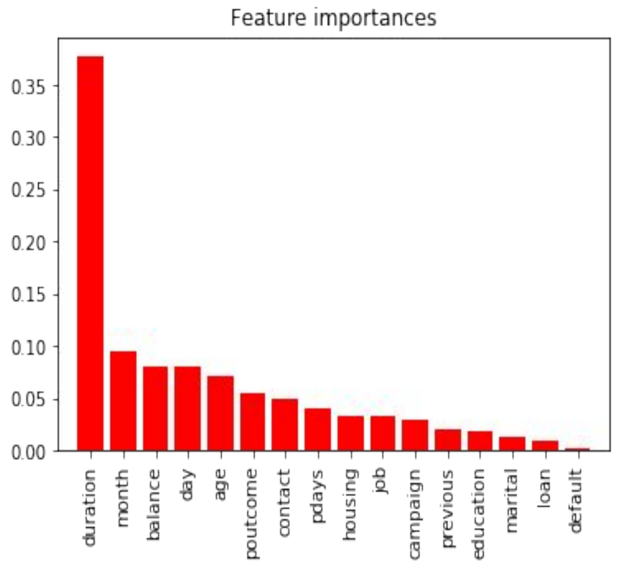
'max_features': 6,

'max_depth': 110,

'bootstrap': False

Training Accuracy score : 100.0
Confusion matrix : [31995 0]
[0 31880]
Recall - train : 1.0
MCC : 1.0

Testing Accuracy score : 97.0
Confusion matrix : [7534 393]
[11 8031]
Recall - test : 0.998
MCC : 0.957



From the best estimators from the hyper-parameter tuning, the max_features used for the model is 6. Hence the 6 most important features as interpreted from the feature importance plot are duration, month, balance, day, age, poutcome.

Evaluating The Model’s Performance with Unseen Data

Accuracy score : 0.989
Confusion matrix : [3952 48]
[1 520]
MCC : 0.95

- The new data that is imported is treated into categorical columns.
- StandardScaler is also applied to the new data to make the new data similar to the training data.
- The Random forest classifier performs very well in predicting new data.
- Matthew’s correlation coefficient is 0.95, which is as good as it works on test data.

Conclusion

- Duration of the call played a very important role.

  Engage the customer in the call long enough to understand the merits of the term deposit, the customer has a higher chance of subscribing to a term deposit.

- Most customers were contacted in May did not subscribe to a term deposit. This can be avoided in future. Avoid the holidays!
- Customers who subscribed to a term deposit in the previous campaign are more likely to go for it again.
- Targeting clients aged below 22 and above 60 yields better results.

The model has very low false negatives. It is a good sign since not many people will be missed by the marketing campaign.

Thank You