DATA MINING CA-ONE

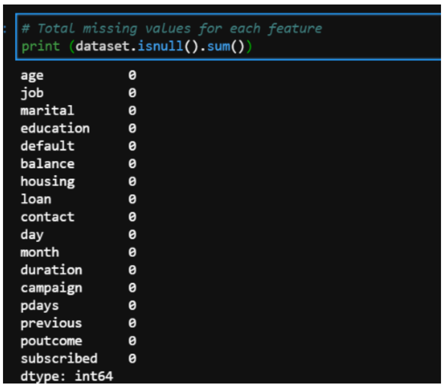
Random Forest & T-SNE – Data Analysis Report

Ishan Das

Banking Marketing Campaign

Understanding the data and pre-processing

The marketing dataset has 1 target variable ‘Subscribed’ and 16 independent variables. The dataset upon loading is scanned for empty/null values.



```
# Total missing values for each feature
print (dataset.isnull().sum())
```

```
age           0
job           0
marital       0
education     0
default       0
balance       0
housing       0
loan          0
contact       0
day           0
month         0
duration      0
campaign      0
pdays         0
previous      0
poutcome      0
subscribed    0
dtype: int64
```

With dataset info we can see the distribution of data across the dataset. As part of data pre-processing, the categorical features are converted / mapped to numerical.

List of categorical features in the dataset.

Subscribed
Month
Loan
Default
Housing
Contact
Material
Education
Job
Poutcome

- The categorical feature that is related / continuous like ‘MONTH’ or single value ‘YES/NO’ is mapped accordingly with mapper.
- The features that is not relatable or not supposed to be sequential – like education is decided for one hot encoding using pd dummies. It is recommended to **not map** these kinds of variables in sequence type mapper 1,2, 3... as that might lead to discrepancies in the model, ML models sometime
- PDays has values spread across -1 to +871 and a mean of 40. So, I have decided to segregate the same into three groups. 0 (never contacted), 1(contacted within 40 days) and 2 (contacted long back).
- Upon co-relation map analysis, it is also found that poutcome is corelated to pdays. However, there is not enough causation of - outcome to previous days contact. A customer can equally reject / accept the term deposit based on decision or other factors.

Poutcome_other, poutcome_unknown & **contact_unkown** do not give any significant weightage to the date. We are more interested in a success or failure story. Hence, these are removed from dataset.

Dataset split

The dataset is split to 75-25 Train-test set. The train set (X) is feature scaled. And applied with SMOTE in order to balance the train set with equal number of YES/NO subscribed data.

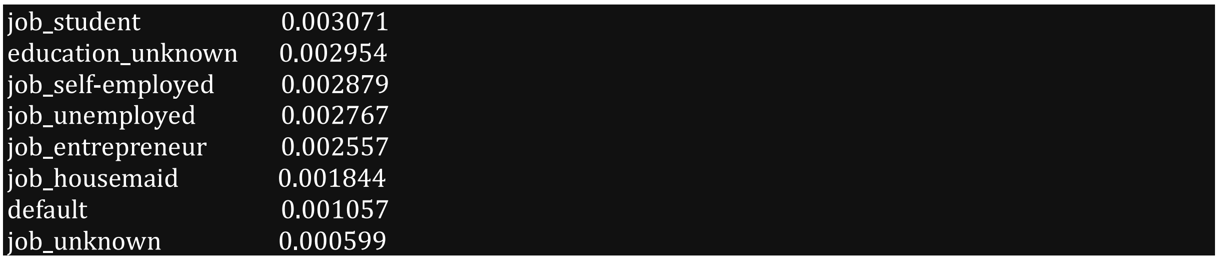
Reducing False Negative

- The GridSearchCV is decided to run with score as recall in order to reduce the false negative.
- **FN statement – "The user has subscribed; however, the model says NOT SUBSCRIBED"**
- **FP statement – "The user has not subscribed; the model says YES SUBSCRIBED"**
- From a business perspective, FP doesn't really do any financial damage. However, higher number of FN can damage the process and finance associated with it. We want our model to predict people. That will help us aligning marketing strategy to target customers. Hence, targeting reducing FN.
- A Cross-Validation of 5 is set for the grid search to do KFold cross validation and find the best result.
- Upon running the grid search, the best fit n_estimator is determined to be **450**.
- Random Forest Classifier is fitted & the result of confusion matrix is as below.

```
TP:  767
TN:  9342
FP:  652
FN:  542
```

Feature Important columns

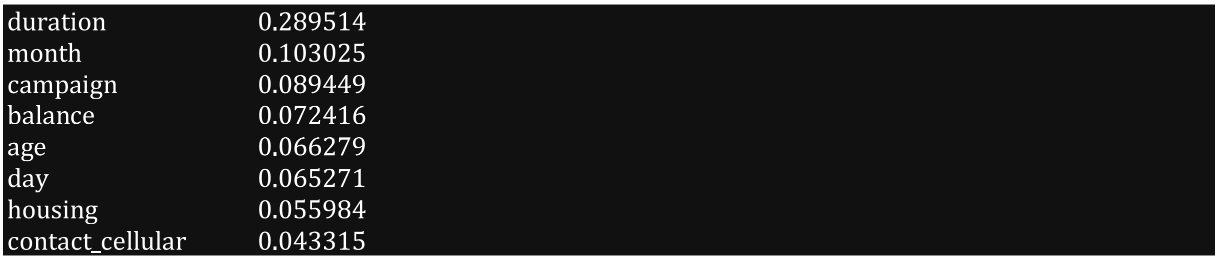
```
duration             0.289514
month                0.103025
campaign             0.089449
balance              0.072416
age                  0.066279
day                  0.065271
housing              0.055984
contact_cellular     0.043315
poutcome_success     0.035062
previous             0.030257
loan                 0.016416
pdays                0.014464
education_tertiary   0.009841
marital_married      0.009643
education_secondary  0.009468
marital_single       0.008810
contact_telephone    0.008112
job_blue-collar      0.007827
job_technician       0.007281
job_management       0.007011
poutcome_failure     0.006740
education_primary    0.006246
job_admin.           0.006118
marital_divorced     0.005056
job_services         0.004607
job_retired          0.004058
```



```
job_student          0.003071
education_unknown    0.002954
job_self-employed    0.002879
job_unemployed       0.002767
job_entrepreneur     0.002557
job_housemaid        0.001844
default              0.001057
job_unknown          0.000599
```

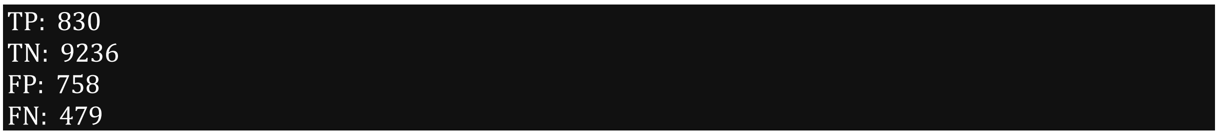
From the feature important columns, the below are selected to form a

SUBSET 1



```
duration           0.289514
month              0.103025
campaign           0.089449
balance            0.072416
age                0.066279
day                0.065271
housing            0.055984
contact_cellular   0.043315
```

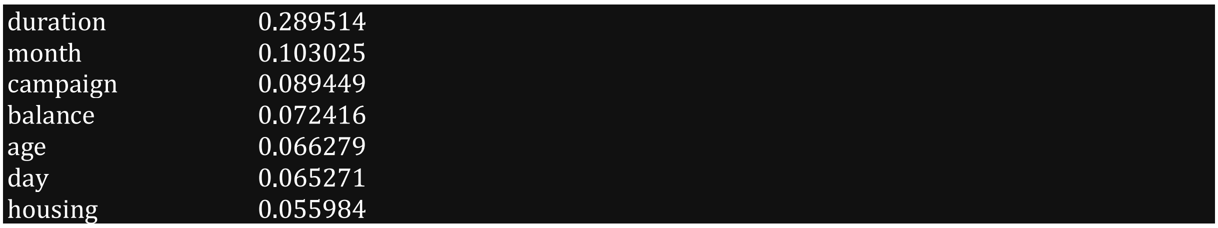
with RandomForest fitting the feature important columns, below is the confusion matrix results.



```
TP:  830
TN:  9236
FP:  758
FN:  479
```

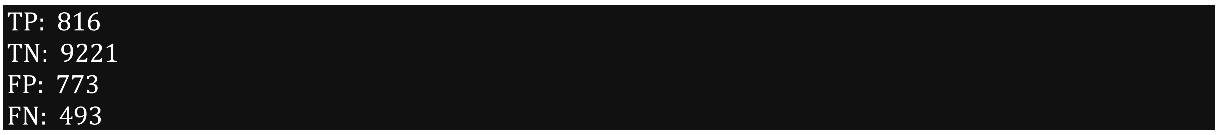
There is a significant decrease in FN and increase in TP. However, Lets try with decreasing one param

SUBSET2



```
duration           0.289514
month              0.103025
campaign           0.089449
balance            0.072416
age                0.066279
day                0.065271
housing            0.055984
```

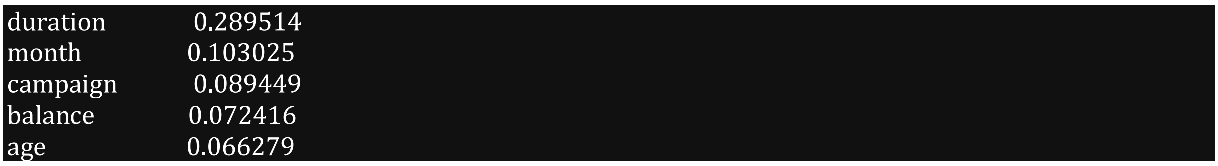
Result



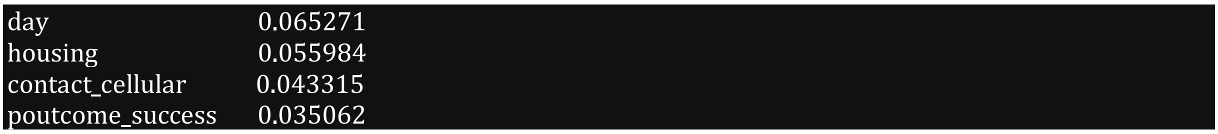
```
TP:  816
TN:  9221
FP:  773
FN:  493
```

The next subset is decided to add one param to subset1. The next feature important param- previous outcomes which are a success.

SUBSET3

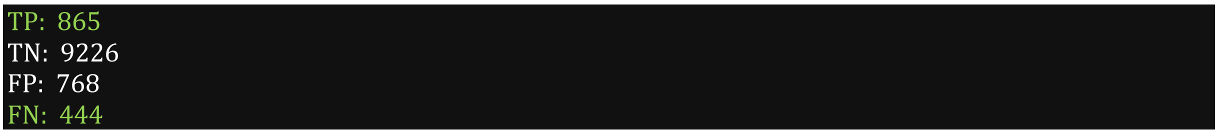


```
duration      0.289514
month         0.103025
campaign      0.089449
balance       0.072416
age           0.066279
```



```
day               0.065271
housing           0.055984
contact_cellular  0.043315
poutcome_success  0.035062
```

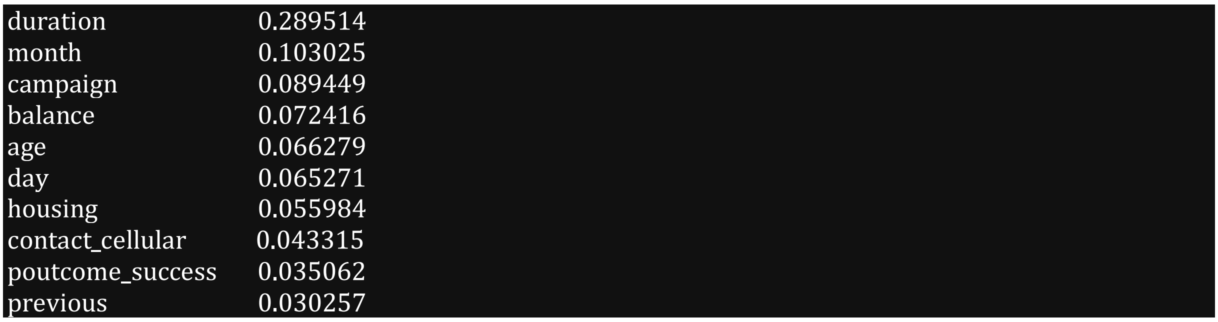
Result



```
TP:  865
TN:  9226
FP:  768
FN:  444
```

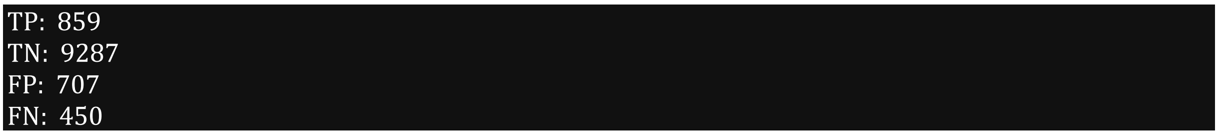
Just to be sure, if we add the next feature important param and create a subset 4-

SUBSET4

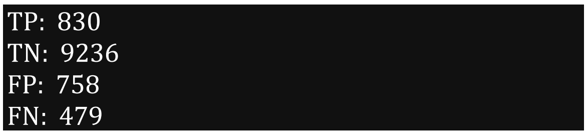
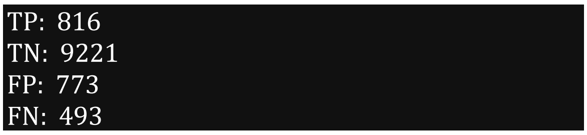
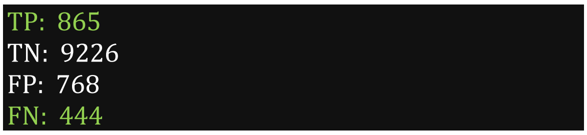
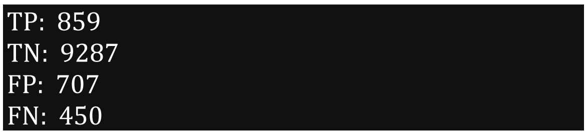


```
duration          0.289514
month             0.103025
campaign          0.089449
balance           0.072416
age               0.066279
day               0.065271
housing           0.055984
contact_cellular  0.043315
poutcome_success  0.035062
previous          0.030257
```

RESULT



```
TP:  859
TN:  9287
FP:  707
FN:  450
```

SUBSET1  TP: 830 TN: 9236 FP: 758 FN: 479	**SUBSET2**  TP: 816 TN: 9221 FP: 773 FN: 493
SUBSET3  TP: 865 TN: 9226 FP: 768 FN: 444	**SUBSET4**  TP: 859 TN: 9287 FP: 707 FN: 450

SUBSET3 is the best subset which gives a substantial decrease in FN and increase in TP. The feature set considered in this subset are –

- Duration
- Month
- Campaign
- Balance
- Age

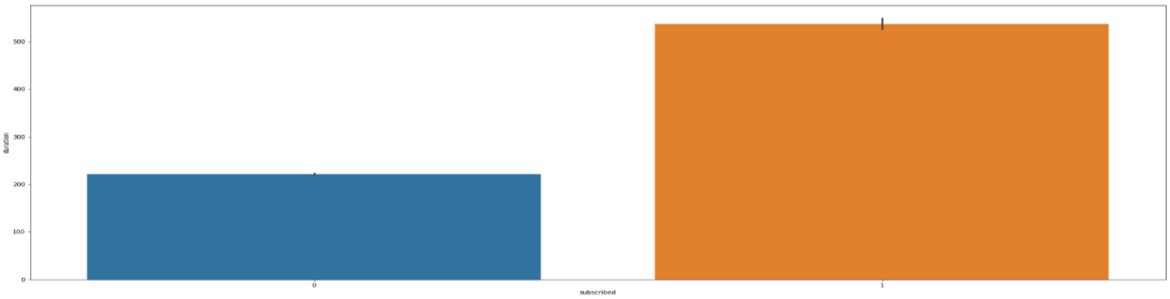
- Day
- Housing
- Contact_cellular
- Poutcome_success

From the initial run of the model which gave 542 False negatives. Subset 3 feature set model reduced FN to 444 that is around ~20% decrease.

Recommendations to marketing team

Below are the recommendations to marketing team and the features they should focus on. This is based on the data distribution observations as well as the consideration of feature important columns which gives the best result.

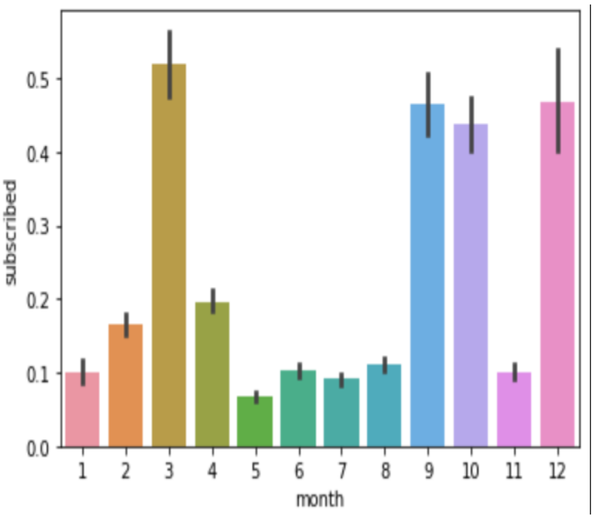
Duration of call



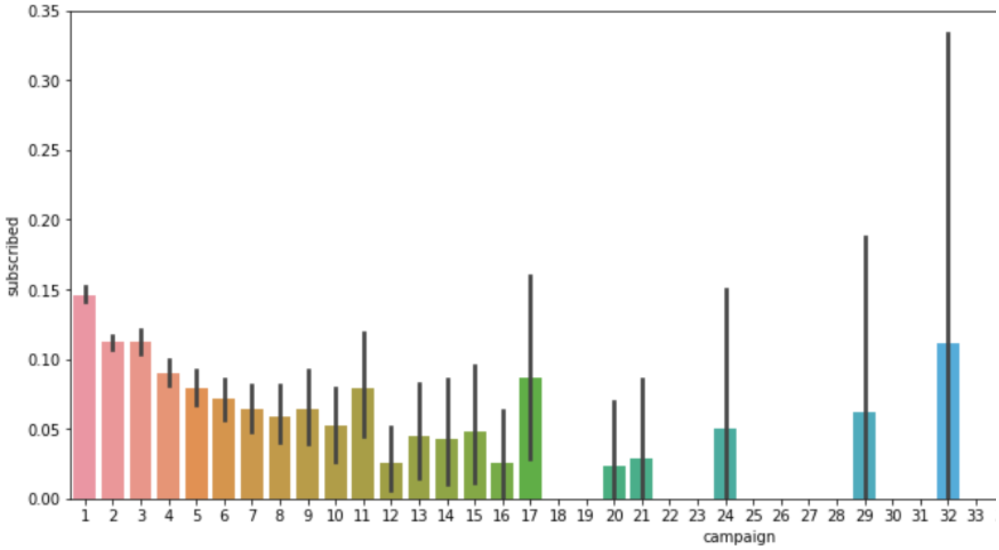
It is recommended that the duration of the call should be at least 200 secs, which should include enough information about the benefits and conditions of the term deposit.

Month

It is recommended that the campaign produces better results if it is done in the financial year closing i.e March or towards the year end (Dec). Various reasons can lead to this including yearend bonus and tax saving actions.

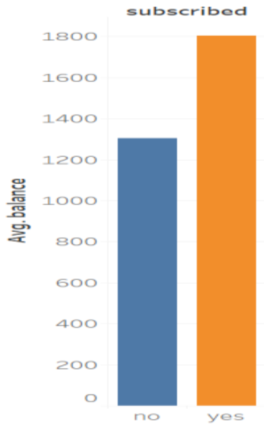


Campaign



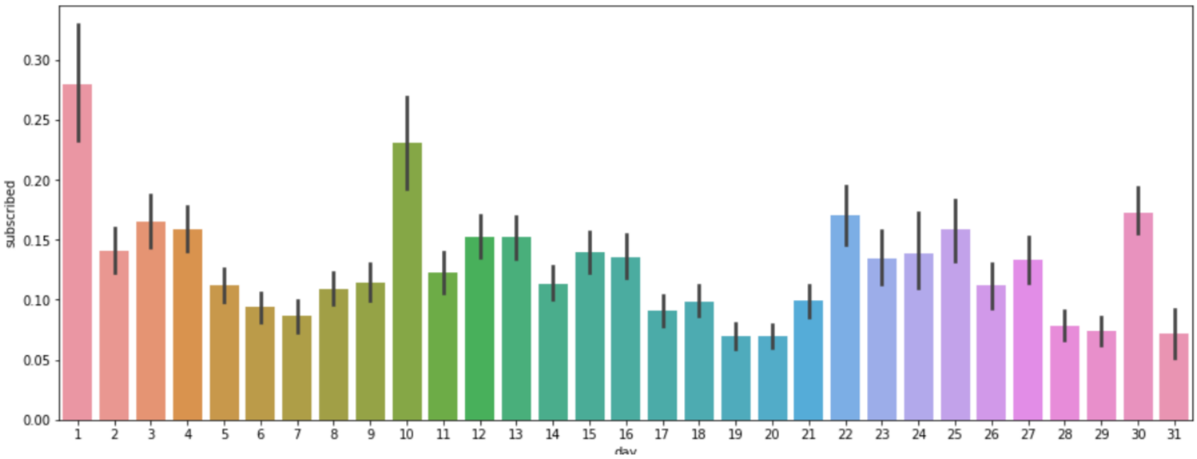
Although we see similar trend for 1-3 campaigns and 15+ campaigns, being a feature important parameter, it is recommended to keep it to a healthy count.

Balance



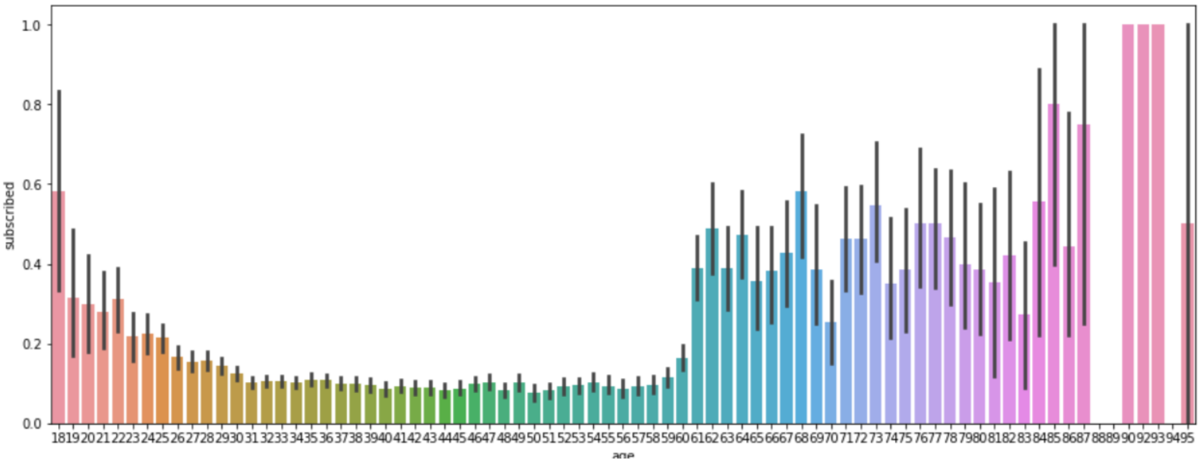
It is also recommended to contact account holders with average of 1200/- Logically the higher the balance the greater the chance of investment.

Day

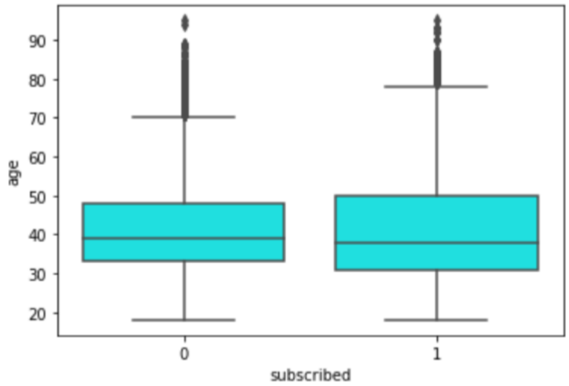


It is recommended to contact people during the start of the month, which seems logical considering salaries get credited then. People will be more inclined towards

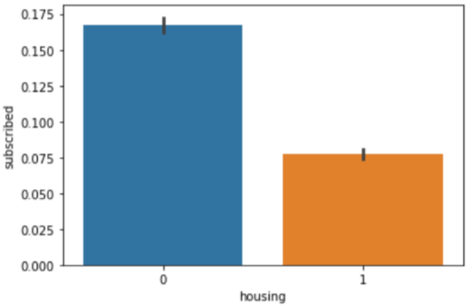
Age



The trend here seems to be quite scattered from bar plot. Hence, upon analysing the boxplot below we can see that the number of acceptances is high in mid-age 30-50 who can consider this term deposit as a retirement benefit. Targeting this age group will be beneficial.



Housing

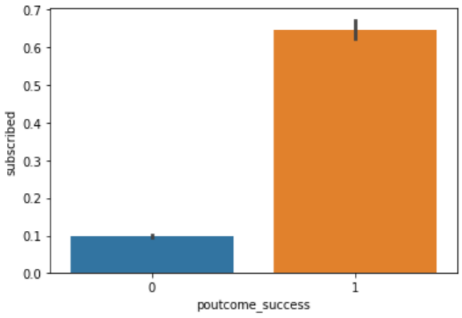


It is recommended to target customers with no housing loan. With less financial burden they should be more interested in investing in term deposit.

Contact_Cellular

It is recommended to contact customer over their cell phone always.

Poutcome_Success



Lastly people with previously bought term deposit are most likely to be interested in getting another one. It is recommended to include the customers in next marketing campaign.

The marketing team is suggested to look in the feature set discussed above and focus on a combination of these feature set.

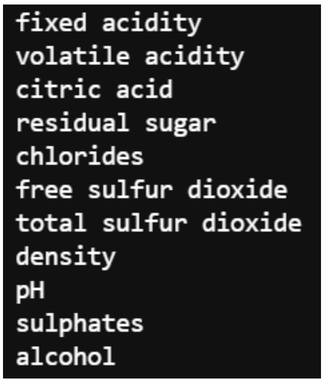
Supporting Documents

- CA1Part1_Jupyter.ipynb - with segmented code and outputs.
- CA1Part1.py – The set of code extracted from Jupyter notebook.

T-SNE

Data Preparation

Drink data is a non-linear dataset which as the following columns.



```
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
```

There is no feature or categorical / output column present in the dataset.

Upon plotting the correlation of the dataset [attached in zip] it is noticed that Density is highly corelated to residual sugar, which can be justified with causation. Hence, it is decided to removed density from the dataset. Apart from that, there is no missing values that is noticed.

Descriptive analysis on the dataset is performed, few things that is noticed out of that.

- Properties like chlorides, pH, sulphates, citric acid have very low standard deviation and hence it is expected that these data points will be very near to one another. That makes them a less of a candidate to look for pattern in the dataset.
- Total Sulfur Dioxide, Alcohol, Residual sugar have diverse data points.

t-SNE

For observing / visualizing data pattern, we will be using t-SNE on the dataset. We will try to segregate and get data insight based on cluster and sub-clusters.

t-SNE helps in visualizing high dimensional data by projecting it in lower dimensions. Basically, a dimensionality reduction technique.

PCA is also a dimensionality reduction technique, however it works on linear projection which is not efficient in capturing non-linear dependencies. PCA with its linear projections leads to crowding problem, which makes it difficult to determine the clusters in the dataset. There are chances that the data points are projected on top of one another which can't be distinguished.

The initial idea is to run t_SNE on the full dataset and see any natural occurring patterns.

Following general steps are taken to apply t-SNE.

- Feature scaling of the dataset.
- Determining the elbow-plot.
- Getting the kMeans labels with the cluster we determined from elbow plot.
- Fitting t-SNE and plotting the data points.

Elbow Plot with K-Means

The elbow method helps in determining the optimal number of clusters in k-means clustering. The elbow method plots the value of the cost function produced by different values of *k*. We use 1 to 11 in our project here.

The value of *k* where distortion declines the most is where we should determine the number of clusters.

Once that is determined, Kmeans is fitted to generate the labels and these labels for the data is used in t-SNE for visualization and colour.

Cluster Interpretations

The initial t-SNE run is made on the full dataset just to check how the natural clusters look like. Although there is some cluster separation, it is not clear on what basis. As mentioned earlier, upon looking at descriptive analysis of dataset, I have chosen to go ahead with converting 'total sulfur dioxide' to 1-0 label column based on its mean.

Subset – **['alcohol','total sulfur dioxide','residual sugar']**

The elbow plot suggests 2 clusters. The perplexity is adjusted from 40 – 150 for better result visibility.

The clusters are well separated based on **Total Sulfur dioxide** high and low. We also notice sub clusters, with high and low **residual sugar**. [Fig 2]

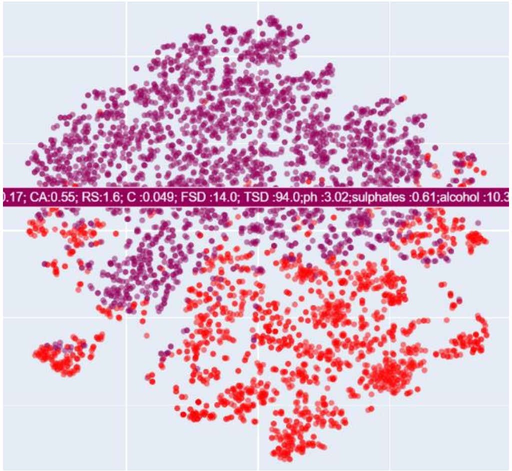


Figure 1 – full dataset

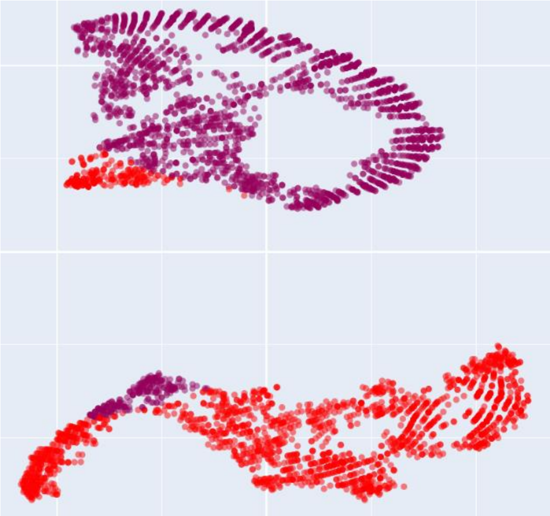


Figure 2 -SUBSET 1

Adding few more parameters to the subset.

[['alcohol','total sulfur dioxide','residual sugar','free sulfur dioxide']]

Based on the elbow plot the cluster is taken as 3. With perplexity and iteration adjustment. Observation from below visualization is as - Total sulfur dioxide and Residual sugar like first subset. However, we also notice a segregation based on alcohol below value 10 and above. This can also be considered as a sub-cluster. [Fig 3]

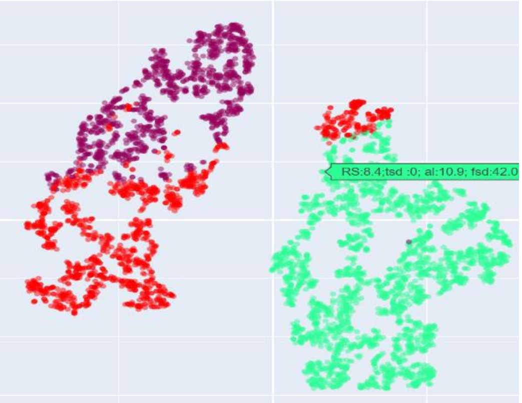


Figure 3 – SUBSET2

Adding few more parameters to the subset.

[['alcohol','total sulfur dioxide','residual sugar','pH','citric acid']]

The observation for subset-3 is quite like subset 2 and addition of citric acid or pH doesn't bring up new sub-clusters as these properties are evenly spread. [Fig 4]

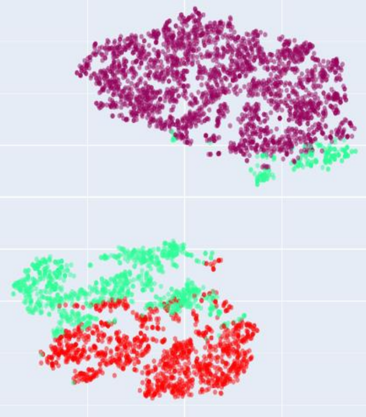


Figure 4 – SUBSET 3

There are number of other subset runs that can be found in attached zip **'.pyinb '**file. Which essentially shows the cluster break by **Total sulfur dioxide or Residual sugar**.

Generic Cluster conclusion

It can be interpreted that the beverages in the dataset can be divided by **high and low** total sulfur dioxide level. From an alcohol beverage perspective, given a dataset we can classify it into high or low total sulfur dioxide content. Basically, it is the sulphur dioxide preservative in the beverage.

Figure Reference Table

Fig 1	t-SNE__fullset_per40_3000.html
Fig 2	t-SNE__subset8_per150_2000.html
Fig 3	t-SNE__subset10_per50_2000.html
Fig 4	t-SNE__subset9_per100_2000.html