Problem Statement:

The Portuguese Bank had run a telemarketing campaign in the past, making sales calls for a term-deposit product. Whether a prospect had bought the product or not is mentioned in the column named 'response'. The marketing team wants to launch another campaign, and they want to learn from the past one. They need to Reduce the marketing cost by X% and acquire Y% of the prospects (compared to random calling), where X and Y are to be maximized

Data and Attribute Information:

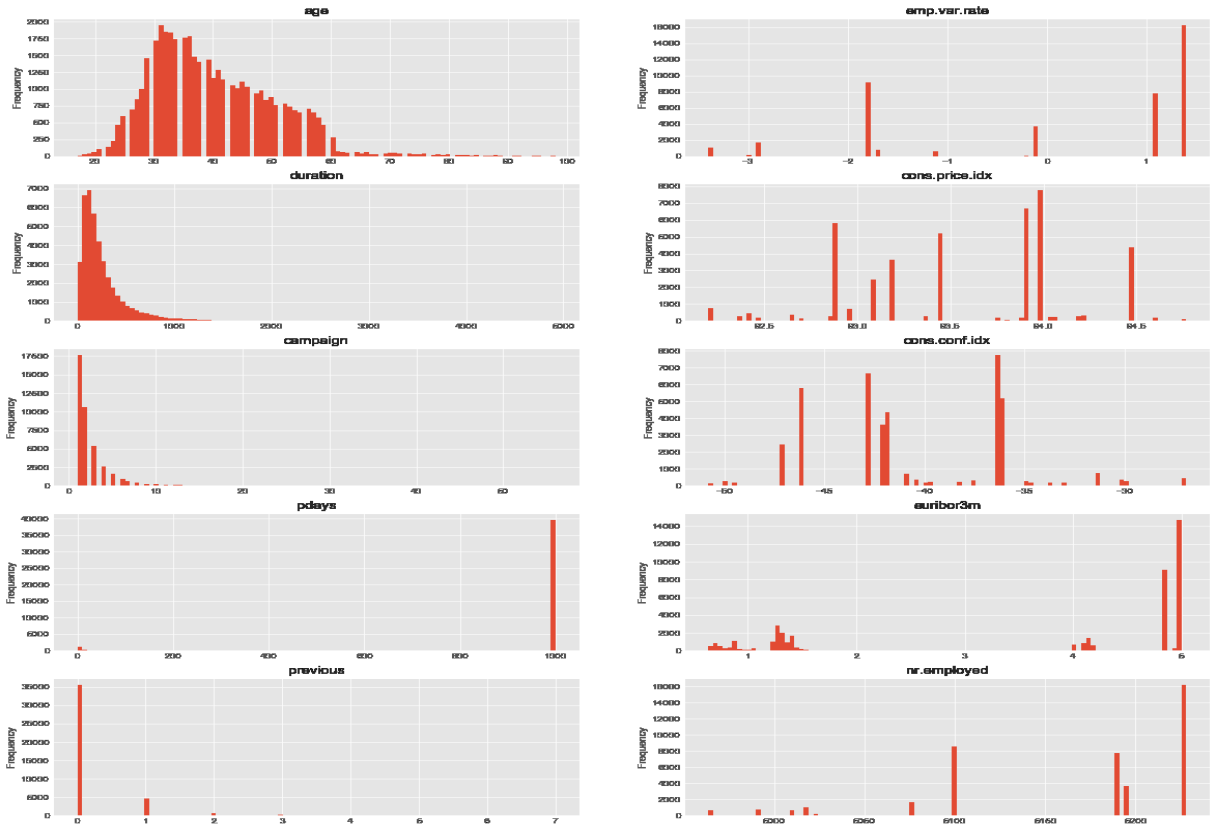
The data and the dictionary can be found here :
https://archive.ics.uci.edu/ml/datasets/Bank+Marketing.

Data Exploration Steps:

1) **Started with finding mean, median and Quantile information of Numerical features.**
2) Checked for Missing or null values in dataset and found that there are no missing values.
3) Checked if there are columns that have only one value or columns that have only one value for 90% of the observations.

   Observations: pdays variable consists of one value '999' for more than 90% of the data, which means that more than 90% of the customers were not contacted previously.

4) Checked the distribution of Numerical Variables.



Observations

- Distribution of duration, campaign should be transformed to log
- Create buckets for pdays, social and economic factors
- create flag for previous

5) Checked the relationship between the independent variables and target variable using bar and sns plots.

Observations

- Age seems redundant
- People not working like student, retired most likely to say yes
- Single people more likely for a yes compared to divorced or married
- Illiterate, unknown more likely for a yes
- People with no default more likely for yes
- Loan and Housing seem to have no impact
- more the duration more chances of yes
- Cellular more chance of yes
- More success in specific months and mid-weekdays
- More success in fewer campaign contacts
- More chances of success if success in previous campaigns
- Lesser Pdays more success
- higher previous higher success
- emp.var.rate is useful
- euribor3m is useful

Feature Engineering steps

1) Age, Duration, Campaign and leuribour3m are converted to logarithm scale.
2) Reduced number of categories for columns Pdays, job, Marital, education, day.
3) Removed variable of less importance as observed from Data exploration steps:

excl=['age','job','marital','education','default','housing','loan','day_of_week','duration',

'campaign','pdays','previous','poutcome','euribor3m','y','npdays']

4) Created dummy of categorical variables.

Model Development steps:

1) Train Test split with 30% of data used for testing.
2) Tried with logistic Regression and Random forest.
3) K fold cross validation with different values of K as 5.
4) Model was developed in three steps:

- Taking all features of importance.
- Selecting features of importance using Random forest variable importance methods.
- Handling class imbalance as only 10% of data contains “Yes” as target variable.

Model Validations and Results

Logistic Regression considering all features

```
Test Accuracy:  0.8289228777211297
Test AUC:  0.8519964570923375
              precision    recall  f1-score   support

           0       0.98      0.82      0.90     10961
           1       0.39      0.88      0.54      1396
```

Random Forest considering all features.

```
Test Accuracy:  0.9147042162337137
Test AUC:  0.7434496530941037
              precision    recall  f1-score   support

           0       0.94      0.96      0.95     10961
           1       0.65      0.52      0.58      1396
```

Logistic Regression after feature reduction

```
Test Accuracy:  0.9049121955167112
Test AUC:  0.6635432043643144
              precision    recall  f1-score   support

           0       0.92      0.98      0.95     10961
           1       0.65      0.35      0.46      1396
```

Random forest after feature reduction

```
Test Accuracy:  0.9136521809500688
Test AUC:  0.7466072404662637
              precision    recall  f1-score   support

           0       0.94      0.96      0.95     10961
           1       0.64      0.53      0.58      1396
```

Logistic Regression After Handling class Imbalance and Feature Reduction

```
Test Accuracy:  0.8289228777211297
Test AUC:  0.8519964570923375
              precision    recall  f1-score   support

           0       0.98      0.82      0.90     10961
           1       0.39      0.88      0.54      1396
```

Random Forest After Handling class Imbalance and Feature Reduction

```
Test Accuracy:  0.8355587925871976
Test AUC:  0.8719895872027655
              precision    recall  f1-score   support

           0       0.99      0.82      0.90     10961
           1       0.40      0.92      0.56      1396
```