**Milestone Report for Capstone 1 project**

**Context:**

Commercial banks spend considerable time and energy in marketing various products/offerings to their clients. In terms of their marketing ROI (return on investment), it makes sense to only focus their marketing efforts on those clients who they believe will be most responsive to a particular product offering(s) vs the whole universe of banking clients.

Business stakeholders that would benefit most from this project would be marketing managers at commercial banks who want to be efficient in their use of time and energy. Using results from this project, they would be able to classify clients expected to be most responsive to a particular product/offering and hence focus their effort primarily or only on those clients vs whole universe of banking clients they could have potentially targeted.

**Specific problem being solved:**

Specific problem I will be addressing for commercial banks is that of increasing subscription to term deposit accounts among bank’s clients who have checking accounts but don’t have term deposit accounts. Increasing subscription to term deposit accounts among clients is important for commercial banks because it ensures availability of funds for a defined term that the commercial banks can then lend out to corporations/individuals for profit.

As a result of this project, marketing managers at commercial banks would be able to identify those clients (that already have current accounts but don’t have term deposit accounts) that are expected to be most responsive to an offer of term deposit account. Marketing managers would then focus their efforts only on those clients vs all the clients, resulting in high ROI on their marketing efforts.

**Methodology used:**

Following steps were performed for this project:

1. Dataset was obtained from Kaggle
2. Data was wrangled as necessary.
3. Dataset was explored to get better understanding of the data and relevant business. Dataset was further processed to prepare it for training.
4. Following binary classification models were chosen to experiment with:
   1. Logistic regression
   2. Random forest
   3. Gradient boost
   4. SVM (Support vector machine)

Each of the model above was trained on the training dataset and then its performance (i.e. its quality of prediction of banking clients most open to offer of term deposit) evaluated on the test dataset using following metrics:

1. Accuracy
2. Confusion matrix
3. Classification report
4. Area under ROC and Precision Recall curves
5. Best model was chosen using area under ROC curve metric. Area under ROC curve was chosen as a metric because it gives one number that takes into consideration both the false positive rate and true positive rate of binary classifiers, can be used to compare binary classifiers effectively and is relevant to a binary classification outcome project such as this one.

Following sections describe the steps that were performed and results obtained for this project.

**Data collection:**

Data source for this project was Kaggle:

<https://www.kaggle.com/prakharrathi25/banking-dataset-marketing-targets>

**Data wrangling**

Dataset had the following variables:

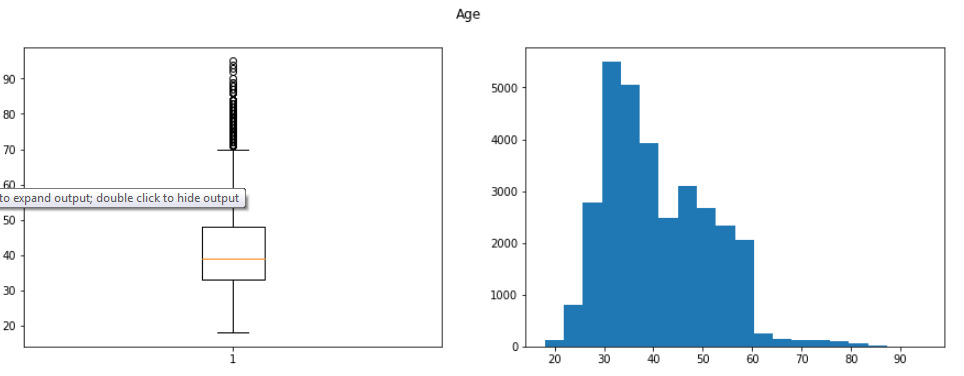
|  |  |  |
| --- | --- | --- |
| Variable name | Variable description | Variable type |
| ID | Unique client ID | Numerical |
| age | Age of the client | Numerical |
| job | Type of job | Categorical |
| marital | Marital status of client | Categorical |
| education | Education level | Categorical |
| default | Credit in default | Categorical |
| balance | Outstanding balance | Categorical |
| housing | Housing loan | Categorical |
| loan | Personal loan | Categorical |
| contact | Type of communication | Categorical |
| day | Day of month of contact | Numerical |
| month | Contact month | Numerical |
| duration | Contact duration | Numerical |
| campaign | Number of contacts performed during this campaign to the client | Numerical |
| pdays | Number of days that passed by after the client was last contacted | Numerical |
| previous | Number of contacts performed before this campaign | Numerical |
| poutcome | Has the client subscribed to term deposit in previous campaign? | Categorica  l |
| subscribed | Has the client subscribed to term deposit in this campaign? | Categorical |

The dataset had 18 variables/columns (8 of them numerical, 10 of them categorical) and about 31,600 rows in total. The dependent variable is “subscribed” which is of categorical type. Rest are independent variables which influence/impact the dependent variable in some way.

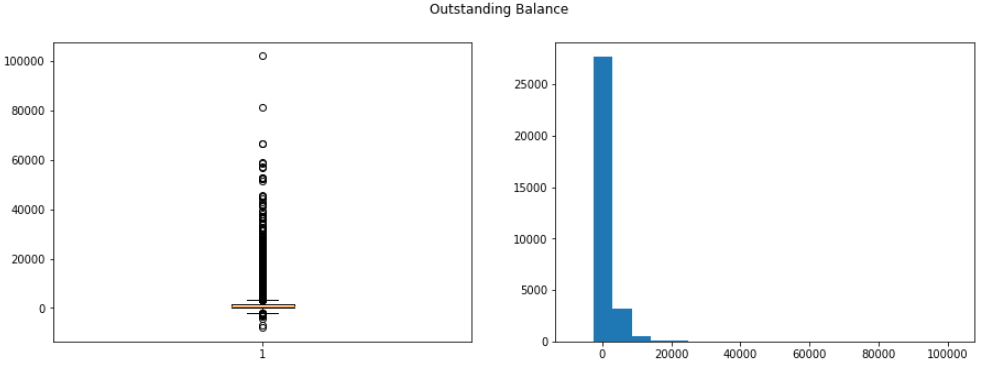
Since the dataset was obtained from Kaggle, it was relatively clean i.e. there were no missing or null values in any of the columns and hence there was not much to do in terms of treating missing values or null values.

Looking at the box plots and histograms for numerical columns gave no indication as to any variable with seemingly incorrect/wrong values which could push the mean values to one extreme or the other:

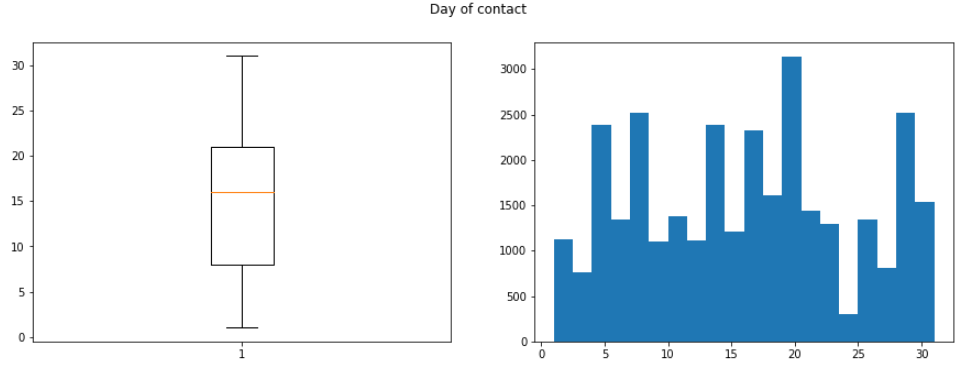
1. age (i.e. age of client) varies from 18 years to 95 years which is reasonable since clients can live upto 95years.



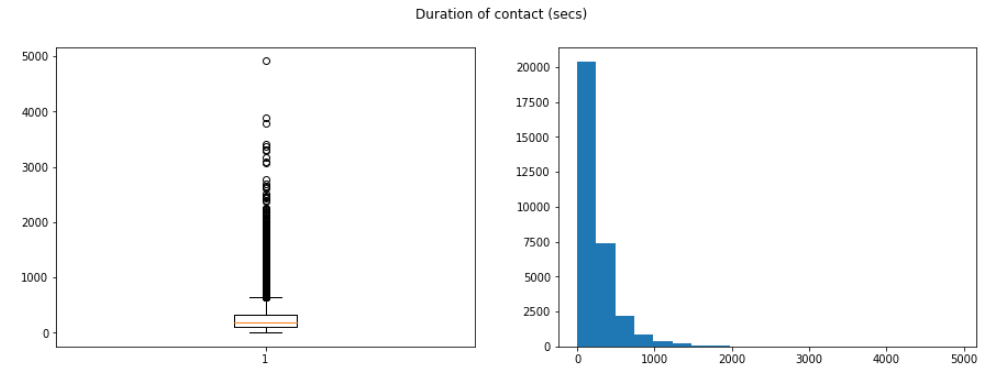
1. balance (i.e. outstanding balance) varies from -8019 to 102127 which is reasonable. It is possible for some clients to have negative balance



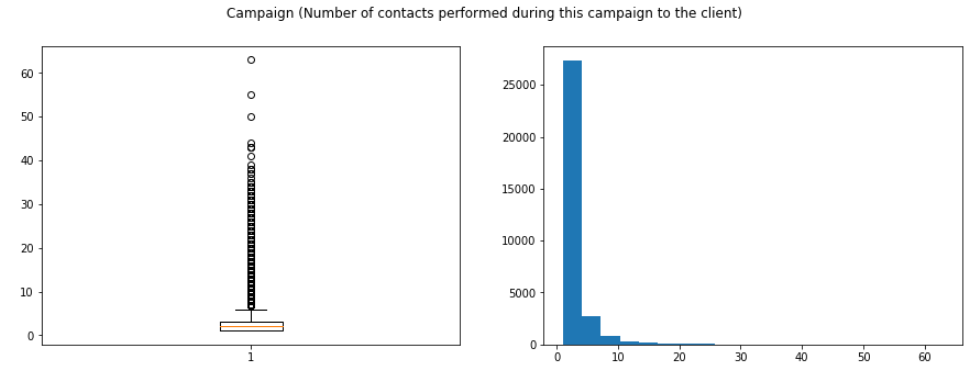
1. day (i.e. day of month of contact) varies from 1 to 31 which is reasonable since maximum number of days in a month is 31



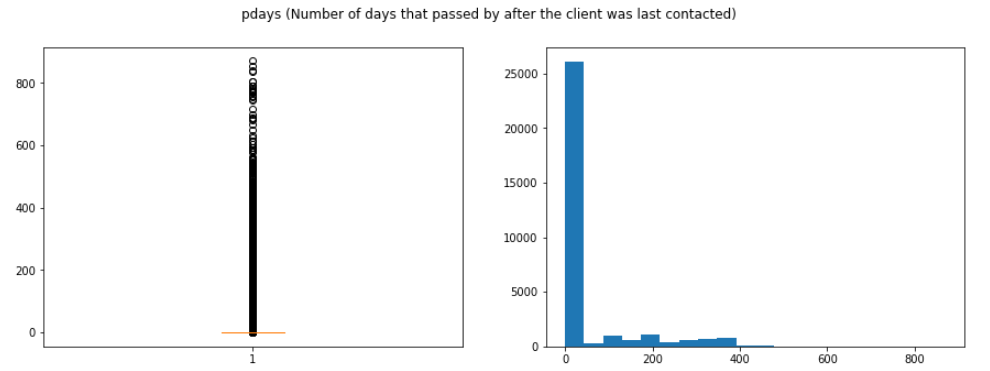
1. duration (i.e. contact duration) varies from 0 to 4918 secs which is reasonable since 4918/3600 secs is 1.36hrs. One to two hours is reasonable timeframe for contact duration.



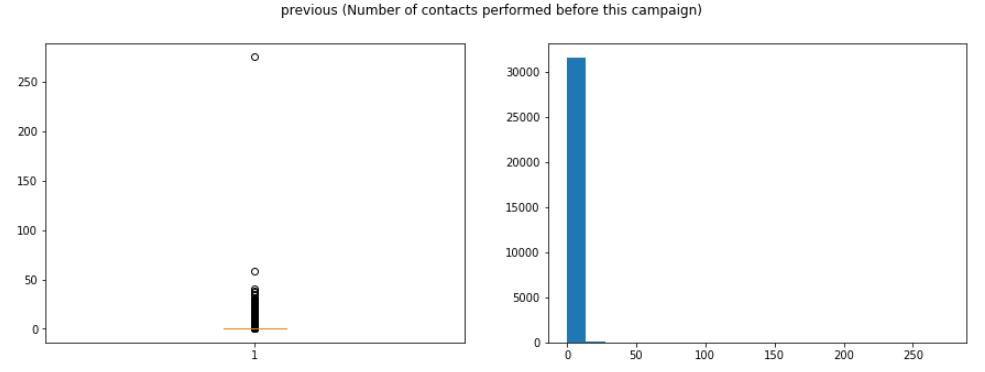
1. campaign( i.e. number of contacts performed during this campaign to the client) varies from 1 to 63. 63 seems to lie on the continum of the values. So, there was no need to treat it as an incorrect/wrong value.



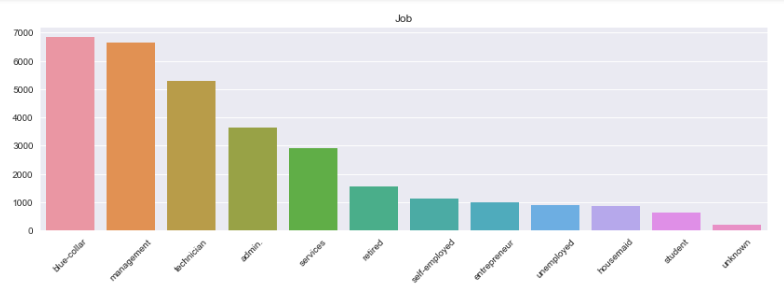
1. pdays (i.e. number of days that passed by after the client was last contacted) varies from -1 to 871 days. 871/364 is about 2.39 years which is possible since banks may not be in touch with the clients for few years.

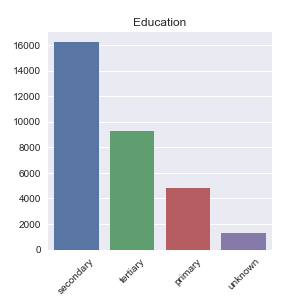
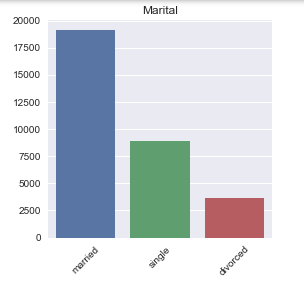


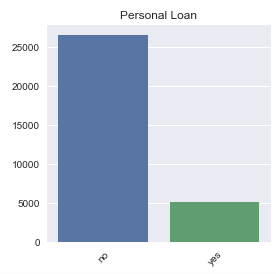
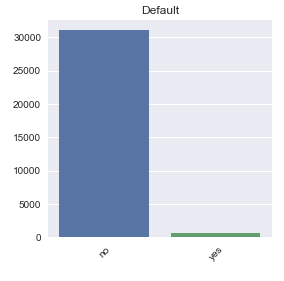
1. previous (i.e. number of contacts performed before this campaign) varies from 0 to 275. 275 lies way away from other values. So it can potentially be considered an outlier. However, for the purpose of this project, it was left as it is since it is possible for a bank client to have been contacted that many times in previous campaigns.

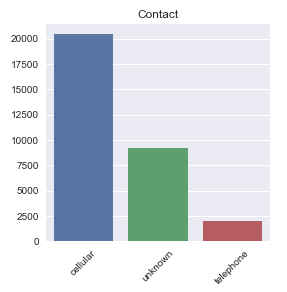
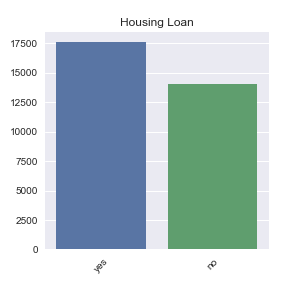


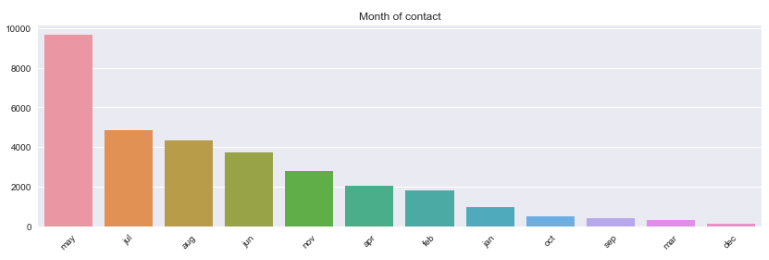
Based on results above, there was not much data wrangling done with numerical variables.

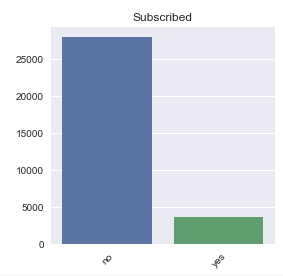
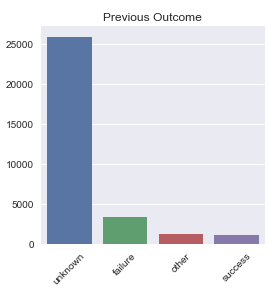
Looking at barplots for categorical columns:











The categories for all the categorical columns seemed well defined. Only thing that stood out was that the “subscribed” dependent variable has more no’s than yes’s i.e. an imbalanced variable.

Overall, since the dataset was relatively clean, there wasn’t much data wrangling to perform either on numerical columns or categorical columns.

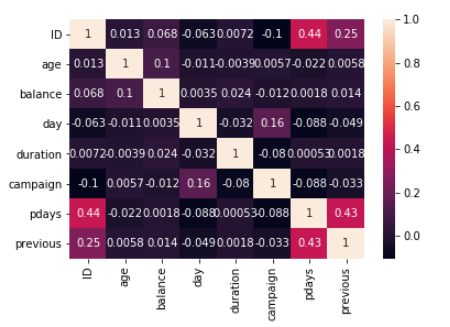
**Data exploration, Data processing and Statistical Data analysis:**

Distribution of numerical variables

As can be seen from histograms of numerical columns above, most of them are not evenly distributed. Except for "age" and "day", they are skewed i.e. distributed at one end of the x-axis or the other. Typically for skewed distributions, log transformations tend to be done to make them more normal-like before feeding them to linear models. In this case, I leave them as they are since I will be feeding them primarily into non-linear models where there is no necessity for normal looking/like distributions for numerical variables.

Correlation of numerical varibles

Pairwise correlation between numerical variables was explored via pairwise correlation plot and heatmap (shown below) to see if there was any meaningful correlation amongst them. Since there was no meaningful correlation, all the numerical columns were fed to the models.

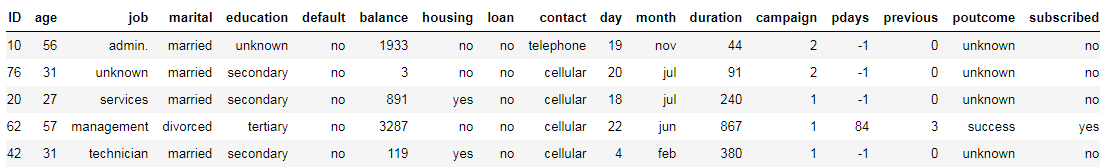


Data processiong and independent variables impacting dependent variable the most

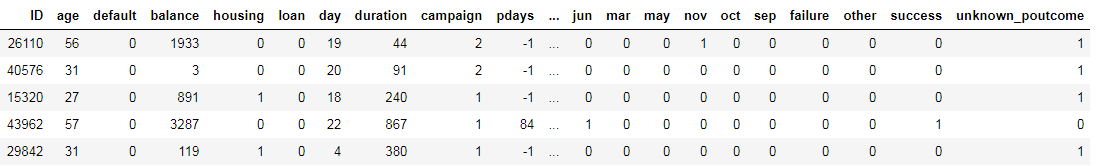
In order to identify independent variables impacting the depending variable “subscribed” the most, logistic regression can be performed to identify independent variables with the largest coefficients. However, before doing logistic regression, data would need to processed in order to make it ready for logistic regression.

Following data processing steps were performed:

1. Leave numerical variables as they are
2. Do onehot encoding for categorical variables with more than 2 labels/classes – this step increases the total number of columns from 18 to 50 for the dataset since for example if a categorical variable has 4 labels/classes, then this step produces 4 derived independent variables for it.
3. Initial dataset would look something like this:



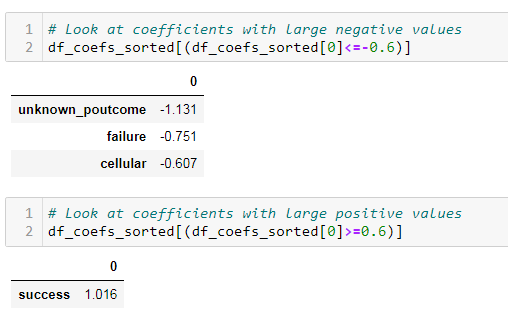
1. Dataset after onehot encoding would look something like this:



1. Create a user defined replace map for categorical variables with exactly 2 labels/classes
2. Divide the dataset into independent variables i.e. x and dependent variables i.e. y and then each of them into training set and test set.

Training dataset would be used to train the logistic regression model. During training phase, coefficients for model that best fit the training dataset would be learned.

Running the logistic regression resulted in following 4 derived independent variables having the most impact on the dependent variable "subscribed":



Unknown\_poutcome, success and failure are derived from independent variable “poutcome” and are essentially its classes/labels. Similarly, cellular is derived independent variable “contact” and is essentially its class/label. Based on coefficient values:

1. if “poutcome” was a "success", then there is a strong likelihood that clients will subscribe to term deposit in this campaign.
2. If poutcome was "unknown\_poutcome" or "failure", then there is a strong likelihood that clients will NOT subscribe to term deposit in this campaign.
3. Similarly, if “contact” was "cellular", then there is a moderate likelihood that clients will NOT subscribe to term deposit in this campaign.

Running logistic regression with statsmodels provides p-values for derived independent variables:

Derived Independent variable p-value i.e P>|z|:

1. unknown\_poutcome 0.00
2. success 0.00
3. failure 0.00
4. cellular 0.00

Since P>|z| < 0.05 at 95% confidence level for above 4 derived independent variables, this implies that the null hypothesis (i.e. there is no statistically significant association between them and the dependent variable) can be rejected which implies they do have statistically significant association with the dependent variable "subscribed".