Attempted by: **Radhika Mittal**
radhika_mttl@icloud.com

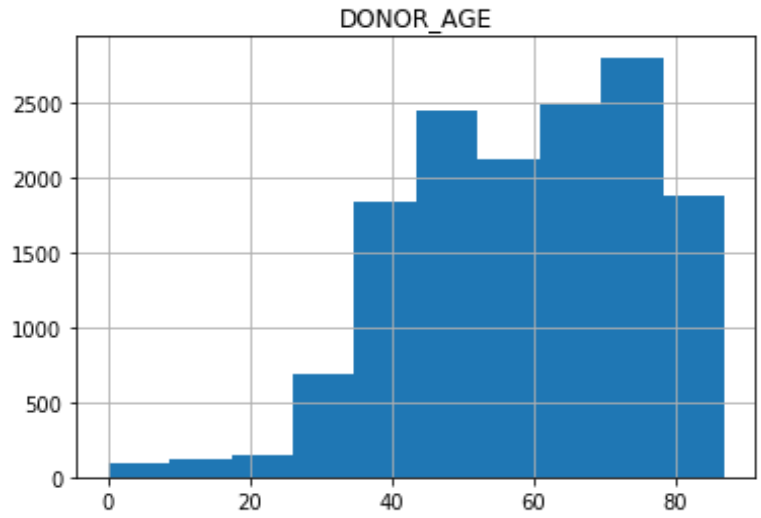
PREDICTIVE MODEL BUILDING PROJECT REPORT

Please perform the following statistical related tasks and be ready to explain the results using the provided 20K account data sample. The sample is related to a mailing campaign for potential donors. The sample data's file layout and definition of the fields will also be provided. Please utilize any of the known statistical / machine learning tools to complete these tasks.

1) Create a Histogram for a continuous variable (EDA_Predmodel.py)

a) Provide the mean, median, standard deviation, and confidence intervals.

Solution 1a:



Histogram for continuous variable (DONOR_AGE)

Age is a continuous variable as it takes values 0 and 100 and all the continuous values in between.

Descriptive stats for continuous variable DONOR_AGE

```
> summary(donor_data$DONOR_AGE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   47.00   60.00   58.92   73.00   87.00    4795
```

(Descriptive stats for other variables are included in the python notebook titled EDA question 1 and 2)

b) Explain what these descriptive statistics tell us about the variable distribution

Solution 1b:

We can see that the continuous variable age does not follow a normal distribution, neither does the TARGET_D which is the donation amount and the other continuous variables. The variable

Attempted by: **Radhika Mittal**
radhika_mttl@icloud.com

TARGET_D is more right skewed since maximum donation amount is 0. The descriptive stats also shows the NA/Null values in the data so we can decide how we would like to proceed with the data pre-processing. In this case we will not simply drop the null values because that will reduce our data significantly thus, we will fill these values which fills the missing value with the observed value in the previous row.

c) Does the variable follow a normal distribution? Explain your answer.

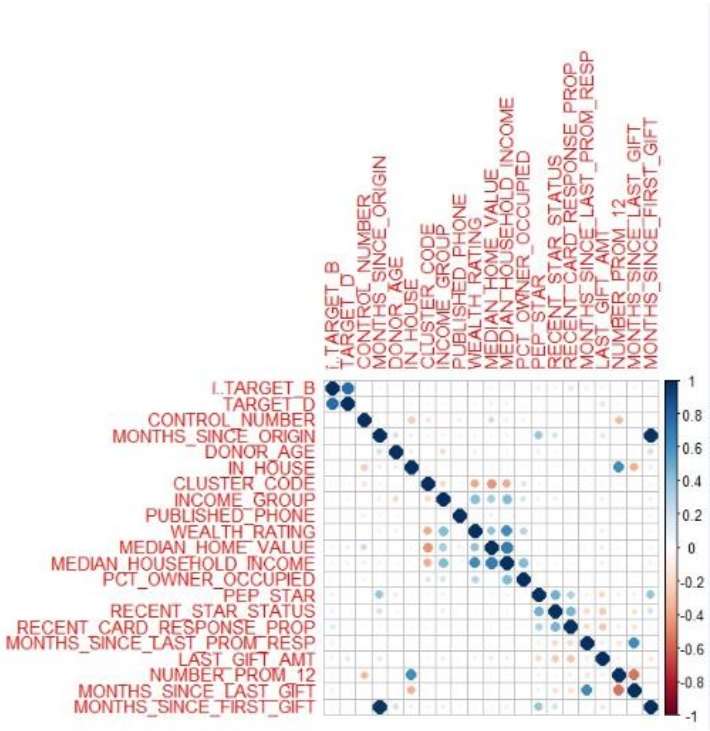
Solution 1c:

The variable does not follow a normal distribution, as we can see from the results there are more donors after the age of 25 and before the age of 87 which is the max age upto which there are donors in the data. There is no consistency observed as per the histogram plotted thus, it is not normally distributed it is a bit right skewed. To deal with this kind of data we can perform transformations like the box-cox transformation which allows to run a broader number of test, Centering, scaling the data or normalizing and standardizing the data depending on the kind of variables and imbalance we are dealing with.

2) **Create a Correlation matrix for all continuous variables and Chi-square test of association for all categorical variables**

a) Provide examples of the different ways variables may or may not be correlated, and explain.

Solution 2a:



From the correlation plot above we can see that variables like MONTHS_SINCE_FIRST_GIFT and MONTHS_SINCE_ORIGIN are highly correlated. TARGET_B and TARGET_D are highly correlated with a positive correlation of 1. It would be safe to drop one of these variables when building our regression and classification models for predictive analysis, as they bring the same information and affect some model results too. MEDIAN_HOUSEHOLD_INCOME AND MEDIAN_HOME_VALUE seem to have a correlation of positive 0.8 but we would not consider eliminating these for our analysis. (The above plot was produced in R).

b) Provide examples of the different ways variables may or may not be associated, and explain.

Solution 2b:

The chi-square calculation below represent the values for the categorical variables in our data. Chi-square measures the degree of association between categorical variables, the 'goodness of fit' aspect that measures how well the distribution of the data fits with the distribution that is expected. Here we observe the X-squared, df and p-value to measure the degree of freedom among the variables this enables us to accept or reject the null hypothesis(which states that the given observations occur only based on chance and are not influence by other observations). A p-value less than the significant level of 0.05 means that you cannot accept the null hypothesis. All our p-values are less than the significant level thus we reject the null hypothesis. (The chisquare test was implemented in R)

Chi-square results for the given categorical variables to check their level of association:

```
> chisq.test(ddclean$DONOR_GENDER, ddclean$URBANICITY)

        Pearson's Chi-squared test

data:  ddclean$DONOR_GENDER and ddclean$URBANICITY
X-squared = 21.247, df = 8, p-value = 0.006519

> chisq.test(ddclean$URBANICITY,ddclean$HOME_OWNER )

        Pearson's Chi-squared test

data:  ddclean$URBANICITY and ddclean$HOME_OWNER
X-squared = 301.82, df = 4, p-value < 2.2e-16

> chisq.test(ddclean$HOME_OWNER, ddclean$recency_freq_status)

        Pearson's Chi-squared test

data:  ddclean$HOME_OWNER and ddclean$recency_freq_status
X-squared = 42.369, df = 21, p-value = 0.003781
```

Attempted by: **Radhika Mittal**
radhika_mttl@icloud.com

3) **Build a Linear Regression Model using a target and predictor variables**
   a) Provide detailed explanation of the results as it relates to the model fit, statistical measure of the variables, and model assumptions.
   b) Provide suggestions to either improve the model's prediction or provide a new approach to the prediction.

Solution 3a and 3b: (Regression_model.py)

The Linear Regression Model was built using the sklearn package linear model for linear regression. Here we have multiple explanatory/predictor variables like DONOR_AGE, MEDIAN_HOUSEHOLD_INCOME, IN_HOUSE etc. thus this is multiple linear regression with our Target variable being TARGET_D which is the amount of donation. The table below shows the output where the predict column shows the values that the model has predicted and compared to the actual values in the data that were given.

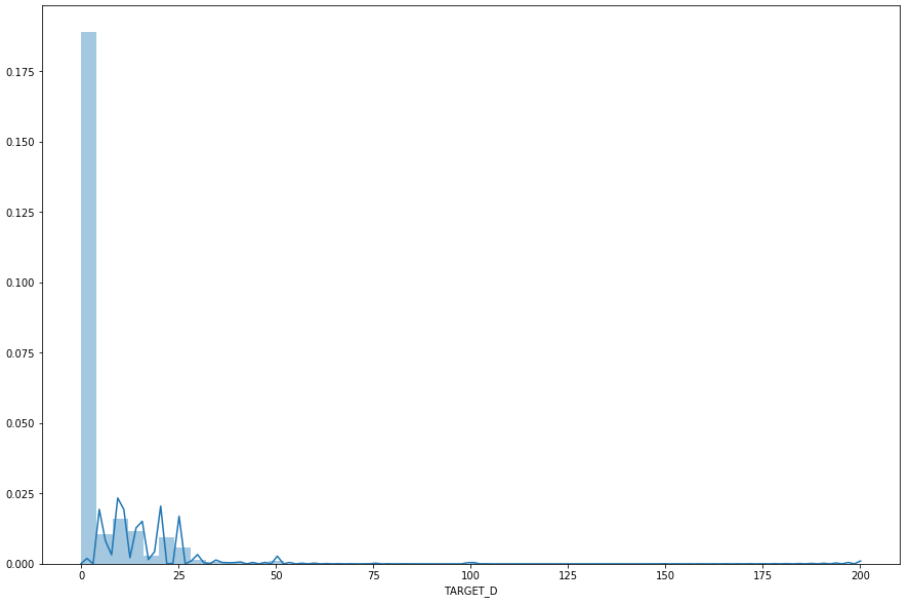
	Actual	Predicted
0	0	5.487645
1	0	3.618876
2	7	2.719763
3	0	3.036243
4	0	3.564535
5	0	3.100493
6	0	3.119325
7	0	4.685238
8	0	3.079533
9	11	4.936085
10	0	4.878154

Linear Regression Model performance accuracy measures are shown below:

```
print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_pred))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))

coefficient_of_determination = r2_score(y_test, y_pred)
print('R square value:',coefficient_of_determination)
```

```
Mean Absolute Error: 5.938006715689296
Mean Squared Error: 82.28774574819772
Root Mean Squared Error: 9.071259325374715
R square value: 0.031830170500590005
```



As per this plot it shows the maximum donations of amount 0 which makes our data skewed towards the left. To deal with this we can add more data for better precision. Scale our data with the available values and then re-run the model to check if we get better results.

4) **Compare the Linear Regression Model versus other Machine Learning Methods**
   a) Train the same data and target with 3 other machine learning methods of your choice
   b) Compare all models based on their results. Which is the best model?
   c) Explain which model statistics support the best model and why?

Solution 4a, 4b and 4c: (Regression_model.py)

The 3 other machine learning models built using the sklearn package in Python for regression are:

1)Random Forest

Attempted by: **Radhika Mittal**
radhika_mttl@icloud.com

2) Decision Tree

3) SVM

To check the best model the Mean Absolute Error(MAE), Mean Squared Error (MSE)and Root Mean Squared Error(RMSE) have been compared.

Name of the Model	Accuracy measure
Logistic Regression	MAE:5.9 MSE:82.28 RMSE:9.07
Random Forest	MAE:6.26 MSE:86.69 RMSE:9.31
Decision Trees	MAE:6.7 MSE:184.81 RMSE: 13.59
SVM	MAE: 4.14 MSE:97.18 RMSE:9.85

From the results above, Logistic Regression is the second best model after SVM. The MAE values closer to 0 signify a good model and better accuracy. Thus, in case of regression SVM regressor performs the best.

5) **Build the Best Classification Model using Machine Learning Methods**

a) Train the same data using the binary target with at least 5 different machine learning methods of your choice

b) How would you go about comparing the accuracy of each?

c) Which method provides the best accuracy based on the comparison? What were the accuracy measures you used to support the best the model?

d) Explain why you think that method performed the best?

Solution 5a, 5b, 5c and 5d: (Classification_model.py)

Attempted by: **Radhika Mittal**
radhika_mttl@icloud.com

6 machine learning classification models were built after splitting the data into train and test. The data required scaling and normalization as the 2 pre-processing steps to get significant results from our machine learning models.

The classifier machine learning models built were:

1) Random Forest

2) Naïve Bayes

3)KNN

4) SVM

5) Decision Tree

6) Multilayer Perceptron

Since our predictive model required multilayer classification the above models were the best choices for such a project. Random Forest did not yield binary results as expected thus, was not a good model as compared to the others. The accuracy of the 5 ML classifiers was as follows:

Name of the classification model	**Observed Accuracy**
Naïve Bayes	70.81%
KNN	68.12%
SVM	75.12%
Decision Tree	63.03%
Multilayer Perceptron (NN)	72.75%
Random Forest	75.33%

The data required scaling transformation for a better training model. After the transformation the Random Forest Classifier and SVM classifier had accuracies of 75.33% and 75.12% respectively, to predict the potential donors(0: Non-donor, 1: Donor) both exhibited a precision of 58% for Class 1: The potential donors. Second best model is the Multi-layer perceptron with an accuracy of 72.75% due to it's multilayer approach it is able to take the multiple features into consideration which is exactly what is required by our classifier. Then follows the Naïve Bayes classifier, KNN classifier and then the Decision Tree. The accuracy is nothing but the comparison of the actual values and the predicted values and how many the classifier was able to label correctly.

Random Forest Classifier is a good fit model for this predictive model followed by SVM.

The tuning parameters in the Random Forest Classifier: 'n_estimator' which is the number of trees in the forest has been set to 500 and 'random_state' to 0 which is the parameter responsible for generating random numbers. The accuracy of the model specifically the precision increased after setting these tuning parameters for the random forest model. The model basically fits a number of decision trees on samples of the dataset and uses averaging the values for better accuracy and to prevent model over-fitting.

Attempted by: **Radhika Mittal**
radhika_mttl@icloud.com

Support Vector Machine is one of the best models to choose for a supervised learning classification, when given labeled training data the algorithms shows an output that has an optimal hyperplane which categorizes new and possible examples. It linearly separates the given classes (In this case 2) and categorizes the values accordingly. Hence, it was successful in accurately classifying the data as compared to the other models.

```
#Calculating the Accuracy, Precision and Recall for the model.
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
print(confusion_matrix(y_test,y_pred))
print(classification_report(y_test,y_pred))
print(accuracy_score(y_test, y_pred))
```

```
[[4762   37]
 [1534   60]]
             precision    recall  f1-score   support

          0       0.76      0.99      0.86      4799
          1       0.62      0.04      0.07      1594

avg / total       0.72      0.75      0.66      6393

0.7542624745815736
```

Accuracy of any model is measure of: (No. of correct predictions)/(No. of total predictions).
The **Precision and Recall** values for the models signify the correctly identified relevant instances. Here, we look at the true positive value for Precision as that is the percentage of potential donors that the campaign should consider mailing. Random Forest provides a 62% precision which is the true positive value of 1534 potential donors that should be mailed and as per that is the best model for this purpose, as shown in the snapshot above. We would not decide the best model based on recall because recall only gives the total correct identified instances but we need the true positives which is given by precision for (Class2: which is value 1 for potential donors).
So to conclude Random Forest is the best model for this problem considering the precision which suggests the campaign to mail 1534 donors who are most likely to donate.