Introduction

The mailing campaign for potential donors dataset has 19372 rows and 25 columns. Amongst the 25 columns, there are 2 target variables (TARGET_D and TARGET_B) and 23 predictors. Out of the 23 predictors, there 12 continuous variables and 11 categorical variables. Table 1 shows the Continuous variables in the dataset while Table 2 shows the categorical variables in the dataset.

Table 1: Continuous Variables

Item	Continuous Variables
1	TARGET_D
2	CONTROL_NUMBER
3	MONTHS_SINCE_ORIGIN
4	DONOR_AGE
5	PCT_OWNER_OCCUPIED
6	RECENT_CARD_RESPONSE_PROP
7	MONTHS_SINCE_LAST_PROM_RESP
8	LAST_GIFT_AMT
9	NUMBER_PROM_12
10	MONTHS_SINCE_LAST_GIFT
11	MONTHS_SINCE_FIRST_GIFT
12	MEDIAN_HOUSEHOLD_INCOME_IN_100
13	MEDIAN_HOUSEHOLD_VALUE_IN_100

Question 1 Create a Histogram for a continuous variable

Question 1 Solution

After analyzing the data features, I created a histogram for 'LAST_GIFT_AMT' feature. The 'LAST_GIFT_AMT' is the amount of most recent donation from an individual to the charitable organization. A Histogram visualizes frequency of score occurrences in a continuous dataset. Since the dataset is mainly concerned with mailing campaign of potential donors, I decided to visualize the distribution of 'LAST_GIFT_AMT'. From Figure 1 it can be seen that distribution of 'LAST_GIFT_AMT' is right-skewed. The median of this distribution shows that 50% of the individuals donated above $15.0 while 50% donated below $15.0 to the charitable organization. Only one individual donated $450.0

1. **Question 1.1** Provide the mean, median, standard deviation, and confidence intervals.
   Question 1.1 Solution
   The mean of the distribution of the 'LAST_GIFT_AMT' feature is is: 16.5845

Table 2: Continuous Variables

Item	Continuous Variables
1	TARGET_B
2	IN_HOUSE
3	URBANICITY
4	CLUSTER_CODE
5	HOME_OWNER
6	DONOR_GENDER
7	INCOME_GROUP
8	PUBLISHED_PHONE
9	WEALTH_RATING
10	PEP_STAR
11	RECENT_STAR_STATUS
12	RECENCY_FREQ_STATUS

The Median of the distribution of 'LAST_GIFT_AMT' feature is: 15.0
The Standard Deviation of the distribution of 'LAST_GIFT_AMT' feature is: 11.9777
The Confidence Intervals of the 'LAST_GIFT_AMT' are: (16.4155, 16.7534)

2. **Question 1.2** Explain what these descriptive statistics tell us about the variable distribution

Question 1.2 Solution

The mean describes the average amount of donation from individuals to the charitable organization. This means that the average amount the individuals donated to the charitable organization is $16.58. The mean works well with a normal distribution. For a skewed distribution, the mean could be misleading because of the presence of outliers.

The median is the middle score of a dataset. The median of this distribution shows that 50% of the individuals donated above $15.0 while 50% donated below $15.0 to the charitable organization.

The standard deviation describes how much the data defers from the mean. For the 'LAST_GIFT_AMT', the standard deviation of $11.9 shows the spread of the data from the mean.

Confidence interval explains the range of values that can be used to tell the true mean of a population. The mean of the'LAST_GIFT_AMT' lies between (16.4155, 16.7534).

Question 1.3

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

Question 1.3 Solution
The variable does not follow a normal distribution because the majority of the data as we can see from the distribution plot in Figure 1 is not in the middle with equal decreasing amounts to the tail. Rather, the distribution is right-skewed because majority of the data (gift donations) is close to zero with a single donation of $450 which is the maximum amount. A normal distribution has a bell curve and has about 68% of the dataset on the first mean. A right-skewed distribution has a bell curve and a long tail pointing to the right direction.

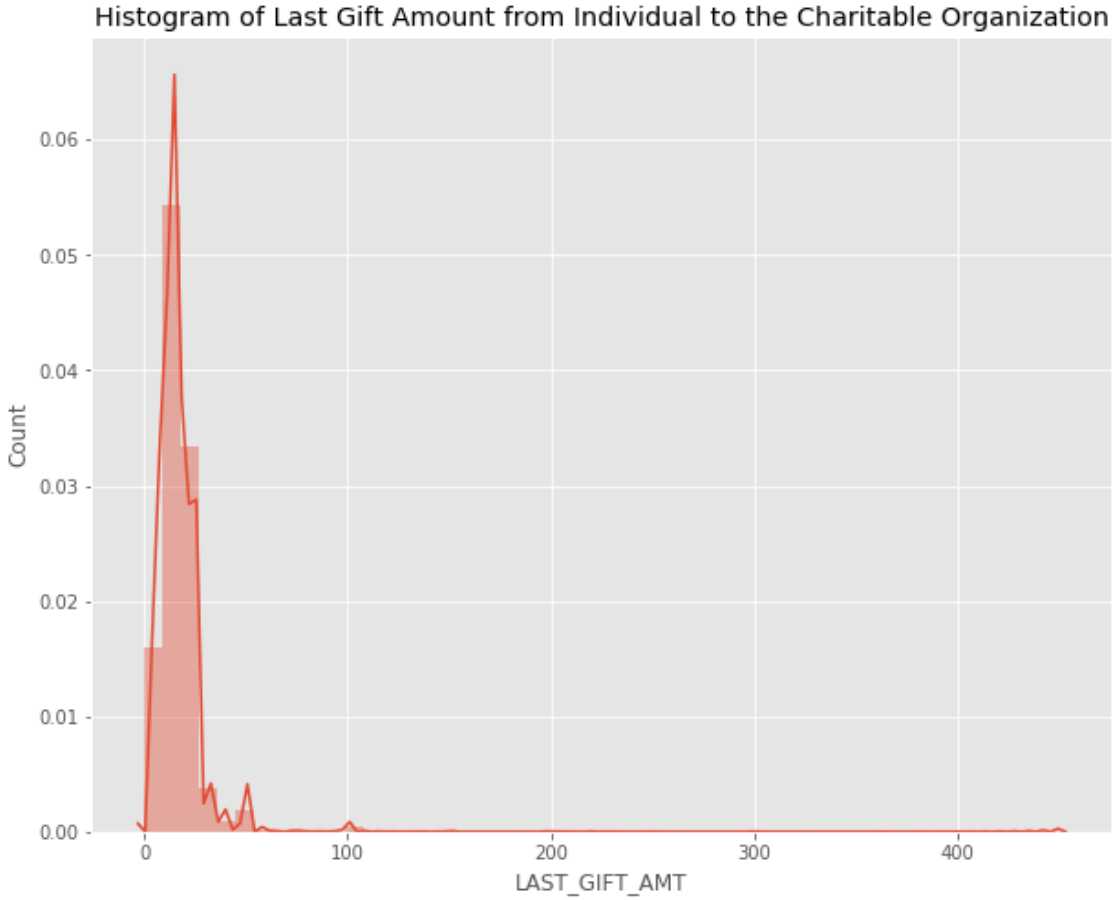

Figure 1: Distribution of Last Gift Amount from Individual to Charitable Organization

3.

Question 2: Create a Correlation matrix for all continuous variables and Chi-square test of association for all categorical variables
Question 2 Solution

In statistics, correlation tells the relationship between two variables. I used a heatmap correlation matrix to visualize the relationship between the continuous variables. The lighter colors show a strong positive correlation, the darker colors show a strong negative correlation while the red colors just show a no correlation between two variables.

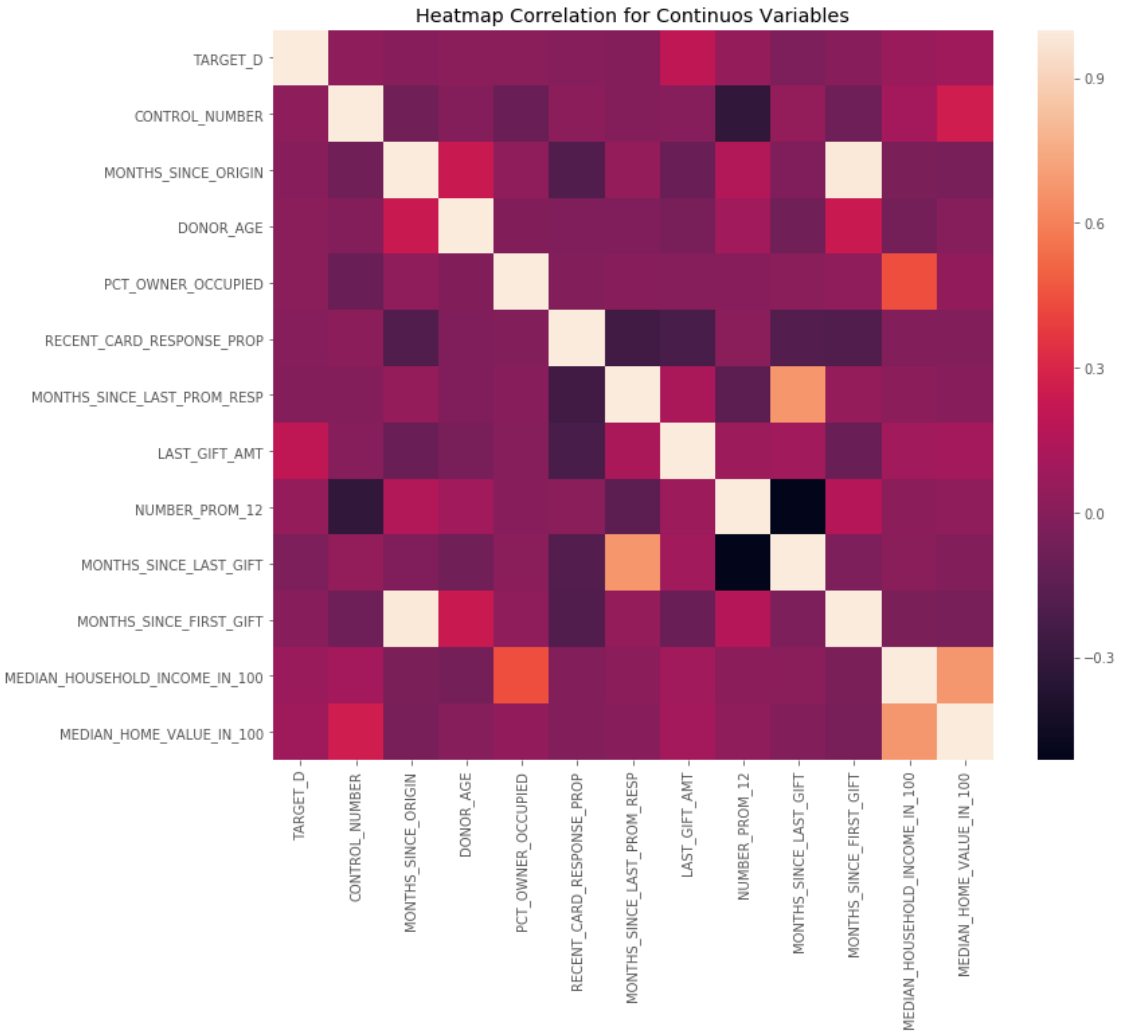

Figure 2: Heatmap Correlation Matrix Showing the Relationship Between Continuous Variables

Question 2.1 Provide examples of the different ways variables may or may not be correlated, and explain

Question 2.1 Solution

Correlation coefficient ranges between -1(strong negative correlation) and +1(strong positive correlation). From the correlation matrix, the lighter colors represents +1(strong positive correlation), darker colors represents -1(strong negative correlation), while the

red colors represent 0 (no correlation).
'MONTHS_SINCE_LAST_GIFT' and 'CONTROL_NUMBER' have a correlation value of 0.046103. This shows that there is a small correlation between MONTHS_SINCE_LAST_GIFT and CONTROL_NUMBER.
There is a negative correlation between 'RECENT_CARD_RESPONSE_PROP' and MONTHS_SINCE_ORIGIN'. The correlation value is -0.197896.

Chi-square test of association for all categorical variables The Chi-Square test for association or Chi-Square test for independence is used to check if there is a relationship between two nominal (categorical) variables. During the test, the frequency of each category for one nominal variable is compared across the categories of the second nominal variable. The data is displayed in a contingency table where each row represents a category for one variable and each column represents a category for the other variable.

In other to determine the Chi-square test for association, the null hypothesis, alternative hypothesis, and p-values are considered:
The H_0 (Null Hypothesis): There is no relationship between variable 1 and variable 2
The H_1 (Alternative Hypothesis): There is a relationship between variable 1 and variable 2

The p-value is used to decided whether to accept or reject a hypothesis. If p-value is significant, you can reject a hypothesis and claim that the findings support the alternative hypothesis.

In other to determine the relationship between two categorical variable, I set my p-value to 0.05. A small p-value indicates strong evidence against the null hypothesis, the null hypothesis will be rejected. A large p-value indicates a weak evidence against a null hypothesis, so you fail to reject the null hypothesis.

p-value less than 0.05 indicates that there is a relationship between two categorical variables

The p-value for 'URBANICITY' and 'TARGET_B' is 0.000216. This is less than 0.05. The null hypothesis can be rejected. This result indicates that there is a relationship between 'URBANICITY' and 'TARGET_B'

p-value greater than 0.05 indicates that there is a relationship between two categorical variables

The different ways variables may not be associated

p-value for 'DONOR_GENDER' and 'TARGET_B' is 0.17380. This is greater than 0.05.

The null hypothesis failed to be rejected. This result indicates that there is no relationship between 'DONOR_GENDER' and 'TARGET_B'

Question 3: Build a Linear Regression Model using a target and predictor variables
Question 3 Solution
Exploratory Data Analysis
During the exploratory analysis, I visualized the correlation matrix of the 'SampleDonor' dataset to understand the relationship between different variables. Figure 3 shows the heatmap visualization of the sample donor dataset. The light colors show a strong positive correlation, the dark colors show a strong negative correlation while the read color just shows no correlation between two variables.

The next variable I explored was the target variable distribution. For the 'SampleDonor' dataset, there are two target variables. Figure 4 shows the 'TARGET_D' variable distribution. 'TARGET_D' is the amount of donation in dollars from the individual in response to last year's 97NK mail solicitation from the charitable organization. The distribution plot shows the amount donated by individuals based on the last solicitation email received from the organization. The maximum amount donated was $200. The average amount donated by individuals was $3. The distribution is right-skewed. It is a bell curve with the long tail to the right.

The target variable for the classification model is a binary target, 'TARGET_D'. a value of 1 denotes individual who donated in response to last year's 97NK mail solicitation from the charitable organization, zero if an individual did not donate. Figure 5 there were individuals did not make any donation based on the last year's 97NK mail solicitation from the charitable organization. There could be multiple reasons for this. One of the reasons could be based on individual's preferred means to communication. Some individual might respond better to an email, some a personalized letter, some text messages addressed to them. Exploring the donor age and the target variable can help understand the categories of donors and come up with a better communication strategy based on their age.

Since the donation is based on mailing response, I would recommend that the organization study their clients and understand what communication platforms work best.

Data Cleaning and Preprocessing
Before building a machine leaning model, the dataset has to be prepared for the model. After exploring the dataset I discovered that there were missing values in the dataset. This is seen in Table 3. Missing values are cells with missing entries and cells with character signs. 'URBANICITY' variable had 454 cells with '?' symbol and was treated as a missing value.

I discovered that there were variables with zero entries. After understanding the variables, I found out that a zero was meaningful. Table 4 shows the number of variables with zero

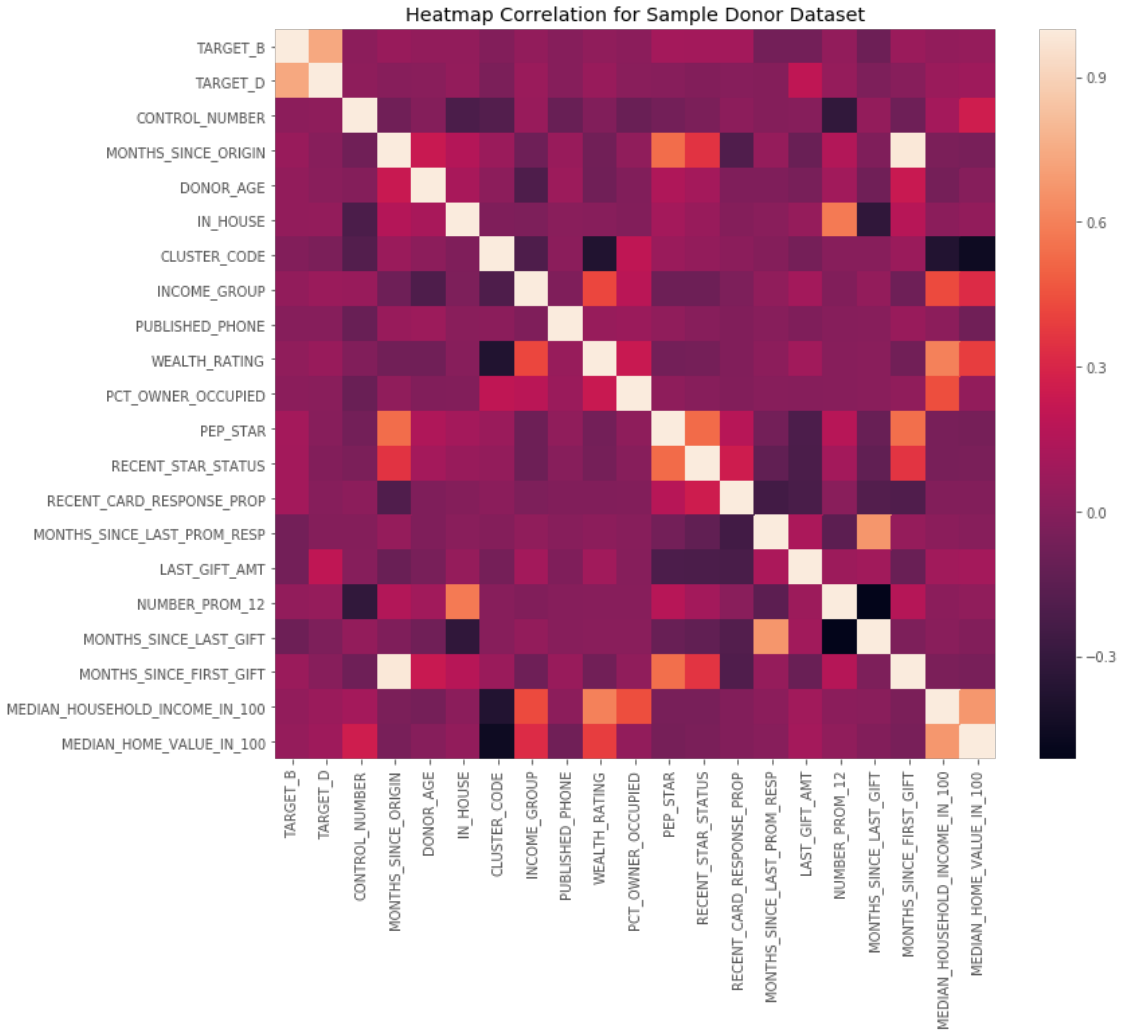

Figure 3: Heatmap Correlation Matrix Showing the Relationship Between the Variables in the Sample Donor Dataset

Table 3: Frequency of Missing Variables

Item	Variables	Missing Value Count
1	DONOR_AGE	4795
2	CLUSTER_CODE	454
3	INCOME_GROUP	4392
4	WEALTH_RATING	8810
5	MONTHS_SINCE_LAST_PROM_RESP	246
6	URBANICITY	454

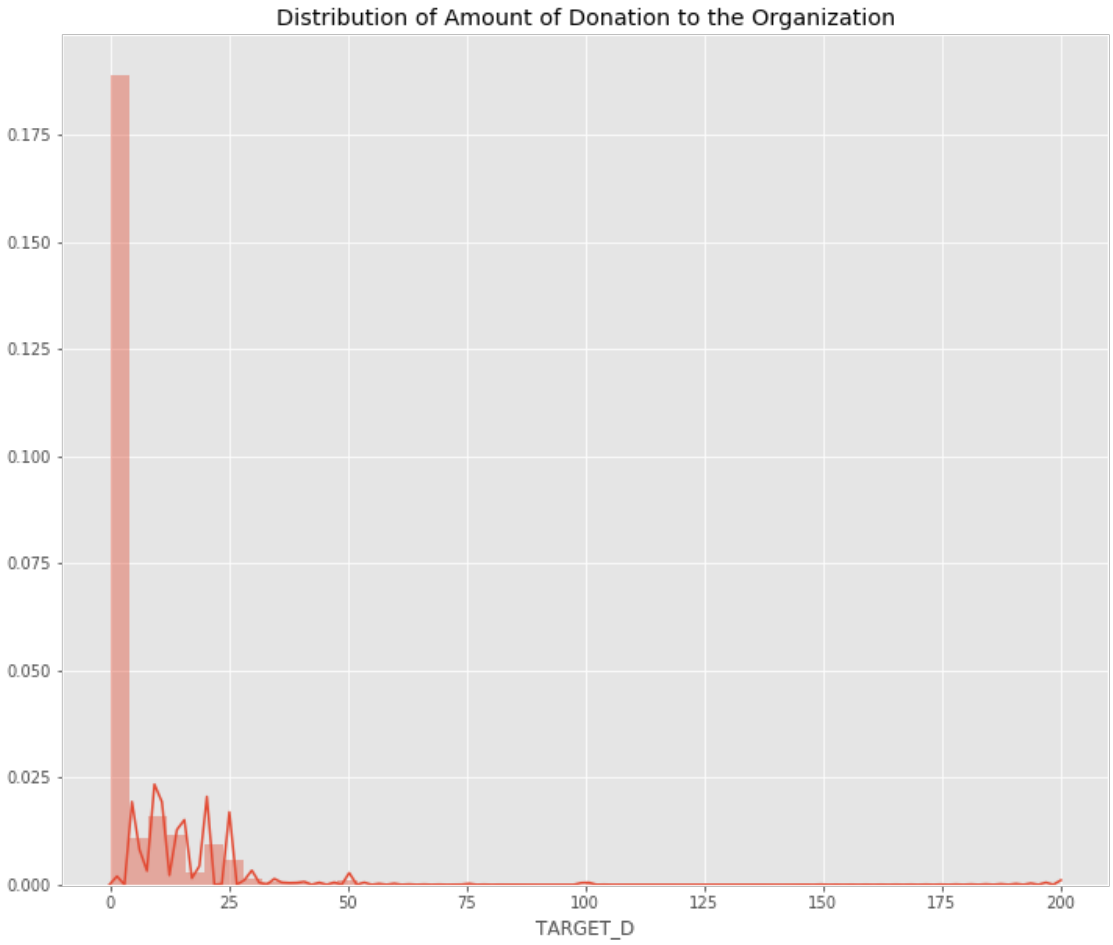

Figure 4: Distribution of Continuous Variable Target for Linear Model

entries and their counts. For example, zero in 'LAST_GIFT_AMT' means that there was no donation from an individual to the organization. Treating zero entry in 'DONOR_AGE' variable as a missing variable did not make sense for two reasons. First, only two entries had zero, secondly it could also be a donation from an individual's baby that is less than a year old. If the entries were numerous, I would have considered it a missing value.

- MEDIAN_HOME_VALUE_IN_100 218: Zero entry in this variable means that an individual doesn't own a home

- MEDIAN_HOUSEHOLD_INCOME_IN_100: Zero entry in this variable means that an individual has no income.

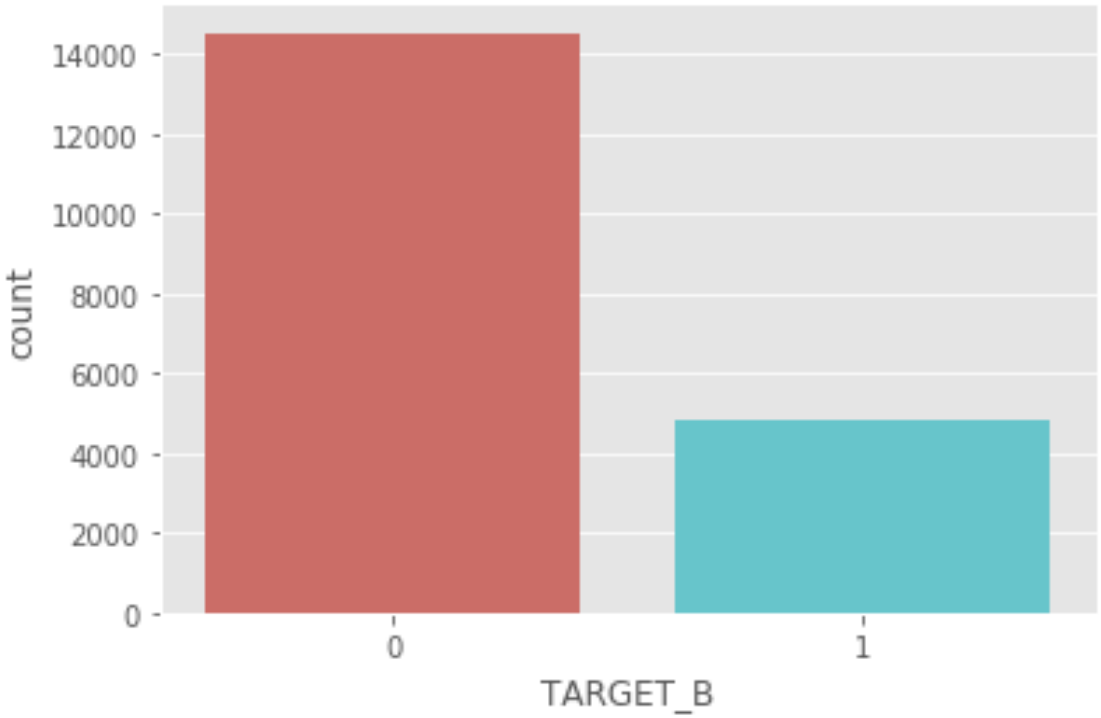

Figure 5: Distribution of Continuous Variable 'TARGET_D' for Linear Model

- PCT_OWNER_OCCUPIED: Zero in this entry means that there are occupants in a neighborhood that are not home owners. Those could be rented apartments of Government projects.
- RECENT_CARD_RESPONSE_PROP: Zero in this entry means that an individual has not responded to a promotion card or other solicitations from the charitable organization since four years ago.
- LAST_GIFT_AMT: Zero entry in this variable means that there was no recent donation from the individual to the charitable organization.

After understanding the missing value and zero entries, I focused on the missing values in Table 3. I used *Sklearn Imputer* with 'most-frequent' strategy to take transformed missing values in the categorical variable by replacing the the missing values with the mode of the variables and 'mean' strategy for replacing continuous missing variables with the mean of the variables.

The next step of preprocessing was transforming categorical variable into numerical variable. The reason for is because some machine learning algorithms can support categorical variables without further manipulation while other algorithms do not. For this project, I used *Pandas get_dummies* to convert categorical variables into numerical variable. This also took care of

Table 4: Frequency of Zero Entries

Item	Variables	Zero-Count
1	DONOR_AGE	2
2	MEDIAN_HOME_VALUE_IN_100	218
3	MEDIAN_HOUSEHOLD_INCOME_IN_100	174
4	PCT_OWNER_OCCUPIED	218
5	RECENT_CARD_RESPONSE_PROP	3936
6	LAST_GIFT_AMT	75

dummy variable trap by dropping the first entry of the categorical variables encoded.

The last step in my data preprocessing was data normalization. Considering the nature of the dataset and the algorithms I utilized, not standardizing the dataset could cause some algorithms that give weight to features (Logistic Regression) or other machine learning algorithms that rely on distance measures (KNN) to perform poorly. I used *sklearn StandardScalar* to standardize the features in the dataset.

Linear regression is a linear approach to modelling the relationship between a dependent variable and and independent variable. The dependent variable is the target variable while the independent variables are the predictor variables. When a dataset contains two or more independent variables, a multiple linear regression model can be used.
The linear regression model will be used to address the following:

1. Provide detailed explanation of the results as it relates to the model fit, statistical measure of the variables, and model assumptions

2. Provide suggestions to either improve the models prediction or provide a new approach to the prediction

Before executing a linear regression model, there are certain assumptions that have to be met. In other to provide the detailed results as it relates to model fit, statistical measure of the variable, I used *statsmodels* to generate a comprehensive table with statistical information about the variables in the 'SampleDonor' dataset. *statsmodels* is a Python module that provides classes and functions for the estimation of many different statistical models as well as for conducting statistical tests, and statistical data exploration. Linear regression works best with dataset that meets its assumptions.Certain models make an assumption about the data. These assumptions help to determine if a technique is suitable for analysis.

```
Omnibus:        28905.538        Durbin-Watson:         2.038
Prob(Omnibus):    0.000      Jarque-Bera (JB):   30355803.177
```

Skew:	8.828	Prob(JB):	0.00
Kurtosis:	196.122	Cond. No.	4.87e+07

Omnibus/ Prob(Omnibus): This is a test of skewness and kurtosis of the residual. The values are expected to be close to zero as zero value indicates normalcy. The Prob(Ominibus) performs a statistical test showing the probability that the residuals are normally distributed. For a normally distributed data, the value should be close to 1. 0.000 Prob(Ominibus) is relatively low, this explains that the residual is not normally distributed. The value for Ominubus is relatively high, this indicates that there is no normalcy.

Skew: This is a measure of data symmetry. This value should be close to zero, indicating that the residual is a normal distribution. 8.828 is a relatively high skew value. Hence, residual distribution is not normal.

Kurtosis: This is a measure of 'peakiness' or curvature of the data. Higher peaks lead to greater Kurtosis. Greater Kurtosis can be interpreted as a tighter clustering of residuals around zero, indicating a better model with few outliers. The value of Kurtosis is relatively high (196.122). This explain how bad model is with possible outliers

Durbin-Watson: This tests for homoscedasticity. Homoscedasticity is when the noise of a model can be described as random and same throughout all the independent variables. Two indicates that there is no autocorrelation, 0 to $\leq$ 2 indicates that there is a positive autocorrelation while 2 - 4 is a negative autocorrelation. The model is negatively autocorrelated with high Durbin-Watson value of 2.038 .

Jarque-Bera (JB): This is like the Ominibus test because it also tests for both skewness and Kurtosis. This test should should confirm the Ominibus test. The values are the same 0.00 .

Control number: this is a measure of the sensitivity of a function's output as compared to its input. When we have multicollinearity, we can expect much higher fluctuations to small changes in the data, hence, we hope to see a relatively small number, something below 30. We have a very high value (4.87e+07) in this case. There is no multicollinearity.

After observing the model output parameters, it can be concluded that the model output is not satisfactory. I recommend that the independent variables has to be revisited and further preprocesseing has to be done.

The statsmodel also provided model accuracy parameters:

- **Df Residuals/Df Model**: This is degrees of freedom i.e., the number of values in the

final calculation of a statistics that are free to vary

- **$R\hat{2}$/Adj $R\hat{2}$**: This is an estimate of the strength of the relationship between model and the response variable. We see very low values of $R\hat{2}$ and adjusted $R\hat{2}$ () which indicates that model is not able to explain much of a variation. Both these values are usually close to each other. We might have to check whether assumptions of linear model has been fulfilled or not and accordingly repeat preprocessing the model F-statistic/Prob (F-statistic)

- **AIC/BIC**: value indicates performance of the model and is usually around 200-400, the lower the better. In this case, we have very high values

- **coef** The coefficient of 0.0076 means that as the X[0] variable increases by 1, the predicted value of y_1 increases by 0.0076

- **std err**: it tells how wrong the regression model is on average using the units of the response variable. Smaller values are better because it indicates that the observations are closer to the fitted line.

- **t & P>** $|t|$ is the t scores and p-values for hypothesis test. The target variable has statistical significant p-value; there is a 97.5% confidence intervals for the target variable (meaning we predict at a 95% percent confidence that the value of RM is between -0.001 to 0.017)

Using a linear regression model depends on the dataset. If the assumptions are met, the model can be used. For model improvement, I will suggest the following steps below:

- . p-value as 0.05.

- After fitting the full model with all predictors, from the OLS Regression Summary table, consider the predictors with the small p-value. For predictors with p > 0.05, remove those predictors from the dataset

- Fit the model without those variables with high p-value

This approach will keep the features with P-value < 0.05. This could improve the model. It is not the best approach because other machine learning algorithm (Classification Models) will perform better with all predictors. However, linear models rely upon a lot of assumptions, if assumptions get violated, R-square and p-values are less reliable.

In other to improve the model, I used was the *Sklearn train-test-split*. I performed and 80/20 train-test split on the dataset. 80% is for training the multiple linear regression model while 20% was for model testing.

Question 4: Compare the Linear Regression Model versus other Machine Learning Methods

1. Train the same data and target with 3 other machine learning methods of your choice
2. Compare all models based on their results. Which is the best model?
3. Explain which model statistics support the best model and why?

Solution

Since my dataset was already preprocessed, I performed an 80/20 train-test split, splitting the dataset into two pieces. 80% of the dataset will be used to train the model while 20% will be used for model prediction.

Since the dataset was already standardized, I went ahead to fit the machine learning models into the training set. I compared the linear regression model with 3 other machine learning methods and they are :

- Linear Regression Model
- RandomForest Regressor Model
- K-Nearest-Neighbors Classifier
- Naive Bayes Classifer

Next I tested the performance of my models by making predictions with the test set results. I created the vector of predictions and used the predict method to predict the observation of the test set. I used an accuracy score metric to evaluate the performance of the models.

```
Accuracy of Linear_Regression classifier is 66.38
Accuracy of Random_Forest classifier is 79.05
Accuracy of k_Nearest_Neighbour classifier is 76.67
Accuracy of Naive_Bayes classifier is 74.5
```

The Random forest Regressor scored higher than the Linear Regression Model. Linear Regression had the least score. Given that the assumptions of linear regression were not met, it can be concluded that the 'SampleDono' dataset will work better with a non linear model. Random Forest is an ensemble learning which combines a lot of decision tree method. The major advantage of the Random Forest algorithm is that it leverages the power of the crowd. The individuals trees might not all be good but by averaging them, the result comes out better. This is one of the reasons Random Forest is popularly used. It is also easy to explain

and understand.

Question 5: Build the Best Classification Model using Machine Learning Methods

1. Train the same data using the binary target with at least 5 different machine learning methods of your choice
2. How would you go about comparing the accuracy of each?
3. Which method provides the best accuracy based on the comparison? What were the accuracy measures you used to support the best the model?
4. Explain why you think that method performed the best?

Solution For the classification model, I used a very common resampling technique of k-fold cross-validation. This simply means that I separated my data into k parts and then fit my models on k-1 folds before predictions for the kth hold-out. I repeated the process for every single fold and average the resulting predictions. I repeated this step because my target variable changed. Based on the target variable, the goal of this classification model is to predict the individuals that donated to the charitable organization as a result of the response to last year's 97NK mail solicitation from the organization to the individual. A values of 1 if they donated and Zero is if they did not.

After performing cross validation, since the dataset was already standardized, I went ahead to fit the machine learning models into the training set. I compared 5 different Classification models and they are :

- Logistic Regression Model
- RandomForest Classifier
- K-Nearest-Neighbors Classifier
- Naive Bayes Classifier
- Support Vector Machines

The next thing I did was to test the performance of my models by predicting the test set results. I created the vector of predictions and used the predict method to predict the observation of the test set.

b) How would you go about comparing the accuracy of each One of the important thing to consider when comparing a model with more than one algorithm; ensuring that the machine learning algorithms are compared on the same dataset is important. This evaluation is achieved by forcing each algorithm to be evaluated on a consistent test set. The result

of running each algorithm is printed with the mean accuracy and standard deviation accuracy.

The Random Forest Model has the best accuracy accuracy score with zero standard deviation error. Due to the high accuracy of the Random Forest model in the 'SampleDono' dataset, I used confusion matrix to further validate the Random Forest Model performance.

The Random forest Regress or scored higher than the Linear Regression Model. Linear Regression has the least score. Given that the assumptions of linear regression were not met, it can be concluded that the 'SampleDono' dataset will work better with a non linear model. Random Forest is an ensemble learning which combines a lot of decision tree method. The major advantage of the Random Forest algorithm is that it leverages the power of the crowd. The individuals trees might not all be good but by averaging them, the result comes out better. This is one of the reasons Random Forest is popularly used. It is also easy to explain and understand.

c

10* **actual value**

Prediction outcome

		p	**n**	**total**
p′		True Positive	False Negative	P′
n′		False Positive	True Negative	N′
total		P	N	

article array graphicx multirow

I used the confusion matrix because it was very easy to tell the model that made the the best prediction.
Confusion matrix is a table that is used to describe the performance of a classification model.
Definition of Confusion Matrix Terms:

- **True positives (TP)** : These are cases in which we predicted yes for example, they have the disease, and they do have the disease.
- **True negatives (TN)**: We predicted no, for example they don't have the disease. **False positives (FP)**: We predicted yes, but they don't actually have the disease.
- **False negatives (FN)**: We predicted no, for example, they actually do have the

disease.

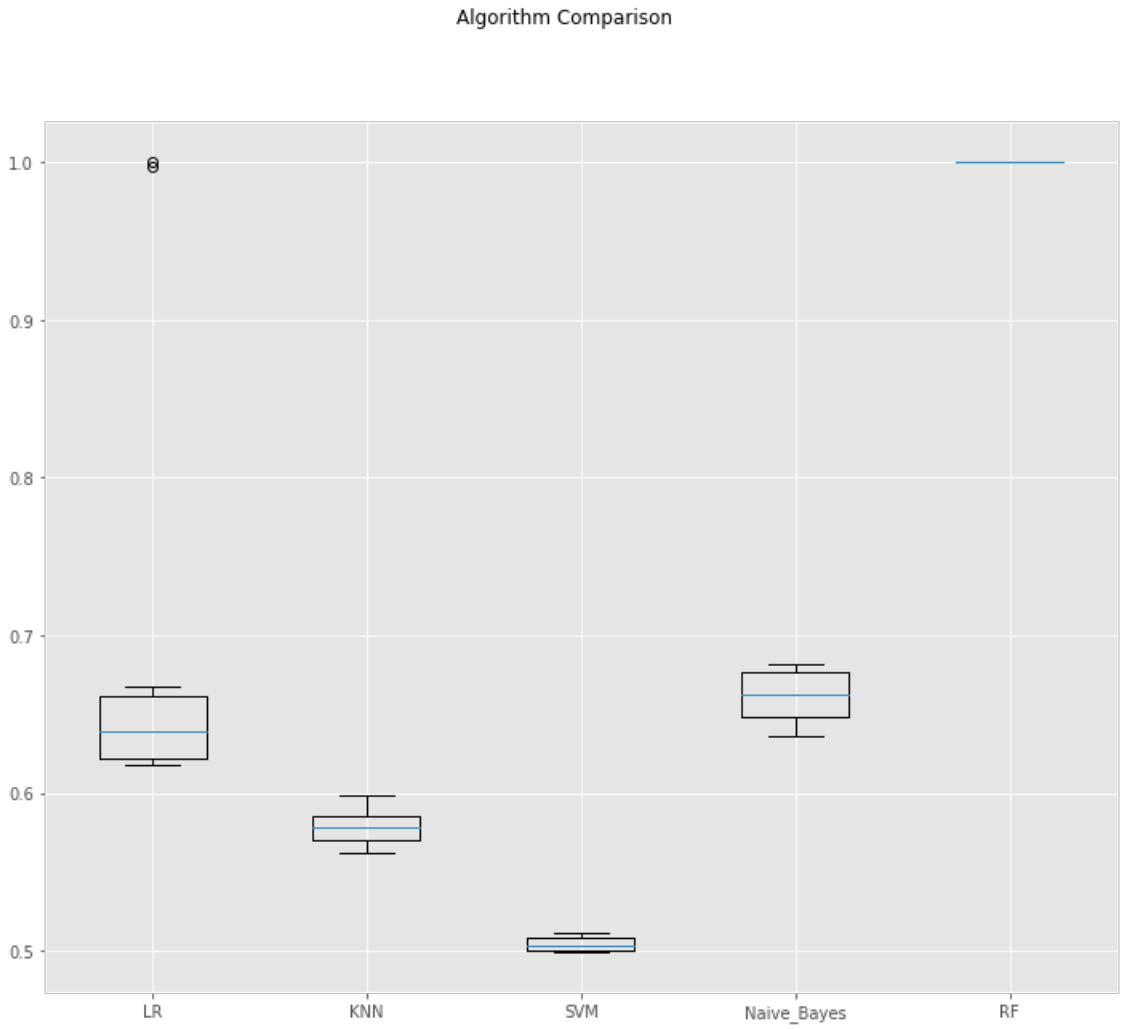

Figure 6: Classifier Algorithm Comparison Showing Mean Accuracy and Standard Deviation Accuracy

```
LR: 0.999809 (0.000550)
KNN: 0.834231 (0.013270)
SVM: 0.999576 (0.000685)
Naive_Bayes: 0.994848 (0.002363)
RF: 0.999999 (0.000003)

 Random Forest Confusion Matrix Result
 [2880,    0]
```

```
[ 0  ,   995]

K-Nearest Neighbors Confusion Matrix Result
[2759,  121]
[ 571,  424]

SVM Confusion Matrix Result
[2880,    0]
[1 ,     994]

Naive Bayes Confusion Matrix Result
[ 130, 2750]
[  0,  995]

Logistic Regression Confusion Matrix Result
[2880,    0]
[ 9,    986]
```

With confusion matrix evaluation, Random Forest classifier still performed better than the other classification models

```
Random Forest Model Predicted 100 percent accuracy
Logistic Regression made predicted 9 inaccurately
Naive Bayes predicted 2,750 inaccurately
KNN predicted 1,263 inaccurately
SVM predicted 1 inaccurately
```

Conclusion

This project utilized various machine learning algorithms to determine the monetary donation from individuals to a charitable organization based on the response to last year's 97KN mail solicitation from the charitable organization.

I used various statistical methods to understand the relationship between categorical variable and relationships between categorical variable. I observed that Linear Regression was not a good model for the 'SampleDono' dataset. Classification models performed well. Random Forest has the highest accuracy.

Further exploration of the dataset will provide a communication preference and a communication approach for the charitable organization to be more strategic about solicitation.

Further work: I will compare Random Forest against other classification models and see how well Random Forest model performs.

References

1. https://www.accelebrate.com/blog/interpreting-results-from-linear-regression-is-the-data-appropriate

2. https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients

3. https://www.bogotobogo.com/python/scikit-learn/scikit_machine_learning_Data_Preprocessing-Missing-Data-Categorical-Data.php 4. https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/