



---

# STAT 656 MIDTERM EXAM

---

Sara Foster



JULY 8, 2021  
TEXAS A&M UNIVERSITY

## Contents

Data Exploration .....	2
Data Exploration .....	2
Logistic Regression Model .....	4
Choosing Parameters with Cross Fold Validation .....	4
Evaluating Chosen Parameters of Logistic Regression Model with 70/30 Validation .....	4
Decision Tree Model .....	5
Choosing Parameters with Cross Fold Validation .....	5
Evaluating Chosen Parameters of Decision Tree Model with 70/30 Validation .....	6
Random Forest Model .....	8
Choosing Parameters with Cross Fold Validation .....	8
Evaluating Chosen Parameters of Random Forest Model with 70/30 Validation .....	9
Recommendations .....	10

Sara Foster  
STAT 656 MIDTERM  
Student ID # 821007101

## Data Exploration

### Data Exploration

We have a fairly large data set with a lot of missing and outliers for several variables as seen in **Figure 1**. The variable with the most missing features is WEALTH\_RATING, but it is not more than 24% of the dataset, so will not need to be excluded from analysis. The data dictionary was provided, however some of the limits did not match what was generated by python (see code). Defaulted to the data map provided. All values with missing variables will be imputed, but having so many outliers will need to be taken into consideration when conducting analysis. Some data exploration was done with histogram plots for columns with missing features to gain a better understanding of the dataset. Because this data is full of categorical variables, it is harder to model in some cases. However, **Figure 2** shows histograms that were developed after running the RIE function. Despite having outliers or missing parameters, most of these columns seem to be slightly skewed. Transforming these variables could be done to improve overall model fit. There is no guarantee variable transformation would improve overall model fit, so the variables were left untransformed. The outliers identified could have some leverage or influence on our logistic regression model, however this should be minimized within the decision tree and random forest.

```
***** Data Preprocessing *****
Features Dictionary Contains:
30 Interval,
6 Binary,
6 Nominal, and
2 Excluded Attribute(s).

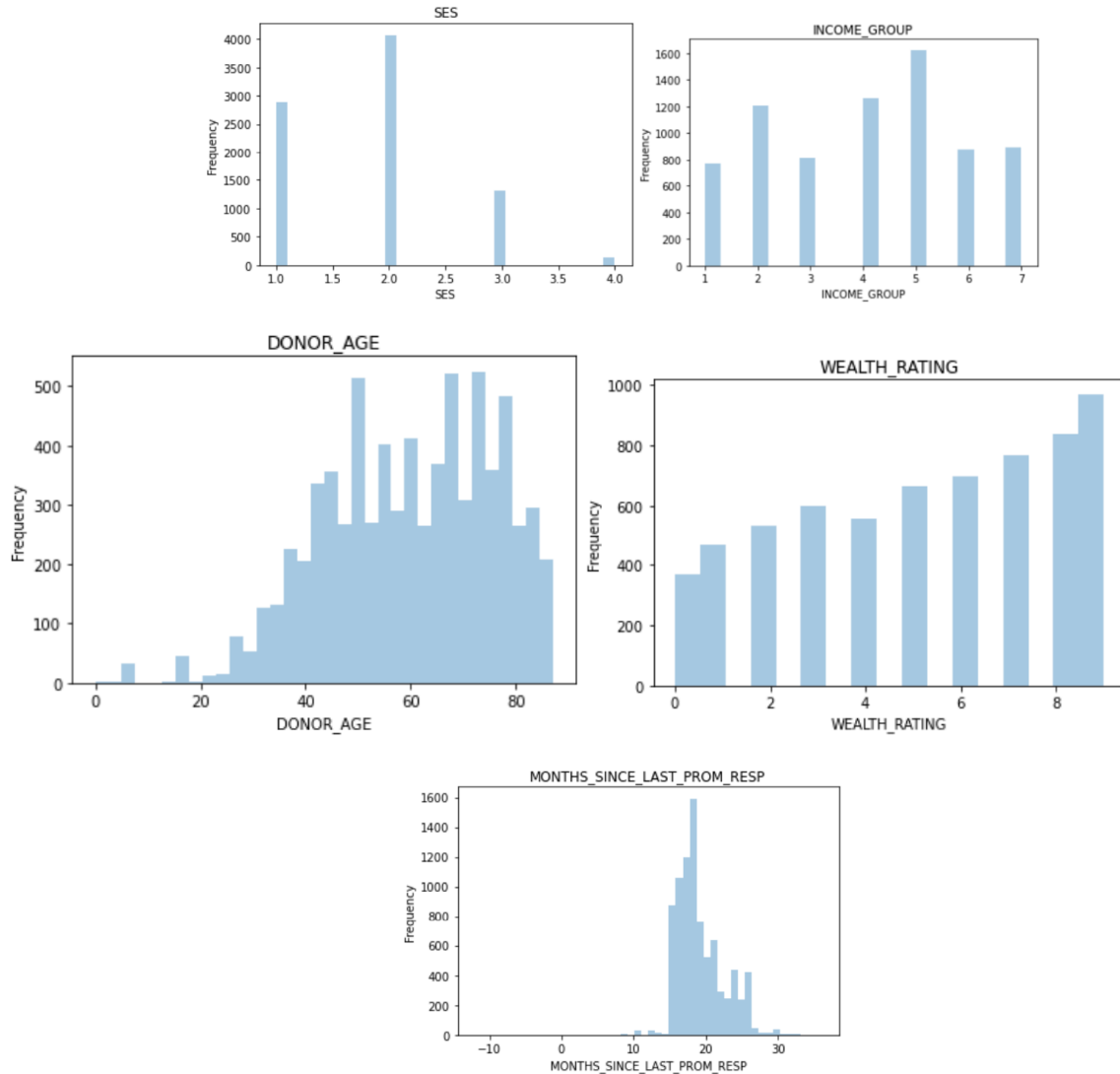
Data contains 8546 observations & 44 columns.

Attribute Counts
..... Missing Outliers
TARGET_B..... 0 0
CONTROL_NUMBER..... 0 0
SES..... 140 0
URBAN_CITY..... 717 462
IN_HOUSE..... 0 0
HOME_OWNER..... 0 0
DONOR_GENDER..... 272 0
INCOME_GROUP..... 1103 0
PUBLISHED_PHONE..... 0 0
OVERLAY_SOURCE..... 0 1103
PEP_STAR..... 0 0
RECENT_STAR_STATUS..... 0 0
REGENCY_STATUS_96NK..... 0 0
FREQUENCY_STATUS_97NK..... 0 0
WEALTH_RATING..... 2079 0
MONTHS_SINCE_ORIGIN..... 0 0
DONOR_AGE..... 1164 0
MOR_HIT_RATE..... 0 0
MEDIAN_HOME_VALUE..... 0 0
MEDIAN_HOUSEHOLD_INCOME..... 0 0
PCT_OWNER_OCCUPIED..... 0 0
PER_CAPITA_INCOME..... 0 0
RECENT_RESPONSE_PROP..... 0 0
RECENT_AVG_GIFT_AMT..... 0 0
RECENT_CARD_RESPONSE_PROP..... 0 0
RECENT_AVG_CARD_GIFT_AMT..... 0 0
RECENT_RESPONSE_COUNT..... 0 0
RECENT_CARD_RESPONSE_COUNT..... 0 0
MONTHS_SINCE_LAST_PROM_RESP..... 39 0
FILE_CARD_GIFT..... 0 0
LIFETIME_CARD_PROM..... 0 0
```

Sara Foster  
 STAT 656 MIDTERM  
 Student ID # 821007101

LIFETIME_PROM.....	0	0
LIFETIME_GIFT_AMOUNT.....	0	0
LIFETIME_GIFT_COUNT.....	0	0
LIFETIME_AVG_GIFT_AMT.....	0	0
LIFETIME_GIFT_RANGE.....	0	0
LIFETIME_MAX_GIFT_AMT.....	0	0
LIFETIME_MIN_GIFT_AMT.....	0	0
LAST_GIFT_AMT.....	0	0
CARD_PROM_12.....	0	0
NUMBER_PROM_12.....	0	0
MONTHS_SINCE_LAST_GIFT.....	0	0
MONTHS_SINCE_FIRST_GIFT.....	0	0

**Figure 1: Data Preprocessing for Entire Dataset.**



**Figure 2: Missing Data Distributions**

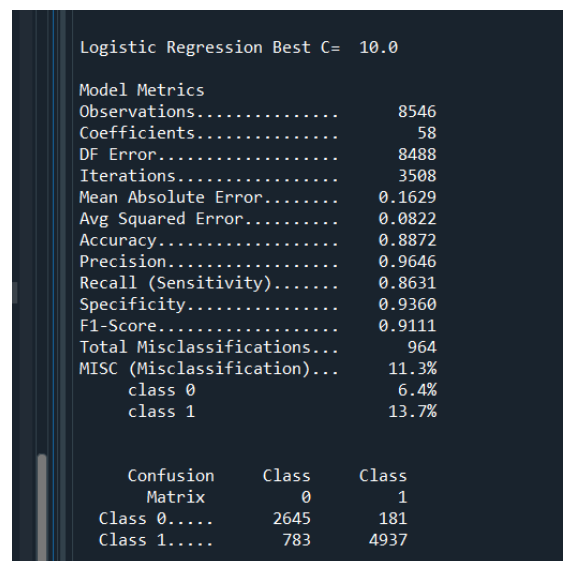
## Logistic Regression Model

### Choosing Parameters with Cross Fold Validation

The model for logistic regression was built by using hyper parameterization with ten fold cross validation then validating the selected parameters with 70/30 split of the dataset. Originally four fold validation was applied to select parameters, but there isn't a major difference in model fit when examining the MISC and VMISC. The only variability is the amount of time it takes to process.

Changing the cross fold validation from 10 to 4 fold didn't change the misclassification rate by much however it did affect which best C was chosen. For cross fold validation of 10, the hyper parameterization chosen was 10 whereas for cross fold validation of 4 it was 0.0001.

As seen in **Figure 3**, the specificity, sensitivity, and F-1 score all indicate a good model fit. The only downfall is the model isn't quite as good at determining when people did donate money. However, the model is good at knowing when people did not donate.



```
Logistic Regression Best C= 10.0

Model Metrics
Observations..... 8546
Coefficients..... 58
DF Error..... 8488
Iterations..... 3508
Mean Absolute Error..... 0.1629
Avg Squared Error..... 0.0822
Accuracy..... 0.8872
Precision..... 0.9646
Recall (Sensitivity)..... 0.8631
Specificity..... 0.9360
F1-Score..... 0.9111
Total Misclassifications... 964
MISC (Misclassification)... 11.3%
  class 0      6.4%
  class 1     13.7%

Confusion Matrix
Class 0..... 2645 181
Class 1..... 783 4937
```

**Figure 3: Results of Logistic Regression with Cross Validation.**

### Evaluating Chosen Parameters of Logistic Regression Model with 70/30 Validation

**Figure 4** gives us an idea of how our model performs against a holdout set. Looking at the results, it seems as if we are not in danger of overfitting as the MISC and VMISC are reasonably close to one another. It is odd to see the VMISC be lower than the MISC, normally VMISC is higher than MISC.

The model performs well when including all 58 coefficients, however including all the variables means the model will be affected by the outliers. These outlier data points could influence the overall prediction of the model by "pulling" to one extreme or the other. This could inhibit overall model performance, but the model validates well which means these outliers may not exert a large level of influence in general. Even though the logistic regression model performs well it is still affected by the relationships between variables, or collinearity. This may account for the model's inability to accurately predict individuals who will donate. The collinearity will be accounted for in the decision tree and random forest model, so these may perform better than the logistic regression.

```

***** Logistic Regression 70/30 Validation *****

```

Model Metrics.....	Training	Validation
Observations.....	5982	2564
Coefficients.....	58	58
DF Error.....	5924	2506
Iterations.....	1532	1532
Mean Absolute Error....	0.1648	0.1630
Avg Squared Error.....	0.0841	0.0820
Accuracy.....	0.8847	0.8896
Precision.....	0.9633	0.9651
Recall (Sensitivity).....	0.8614	0.8643
Specificity.....	0.9325	0.9390
F1-score.....	0.9095	0.9119
Total Misclassifications...	690	283
MISC (Misclassification)...	11.5%	11.0%
class 0.....	6.7%	6.1%
class 1.....	13.9%	13.6%

Training	Class	Class
Confusion Matrix	0	1
Class 0.....	1825	132
Class 1.....	558	3467

Validation	Class	Class
Confusion Matrix	0	1
Class 0.....	816	53
Class 1.....	230	1465

Figure 4: Logistic Regression with 70/30 Validation.

## Decision Tree Model

### Choosing Parameters with Cross Fold Validation

Testing different parameters with cross fold validation enabled us to determine the optimal tree depth of 7 as seen in **Figures 5-7**. The MISC indicates a better predictive performance than the logistic regression model. The model still has trouble with identifying the people who did donate, but still had a decently high sensitivity at 85%. When the tree split, leaf size, and even the amount of cross folding were altered, it had little to no affect on the model results. 63 features were chosen by the model which indicates almost all of the variables provided in the data were utilized in order to build the model. The chosen model depth of 7 indicates the variables seen in **Figure 6** as the most important ones for model building.

```

Decision Tree with Best Depth= 7

```

FEATURE.....	IMPORTANCE
MONTHS_SINCE_ORIGIN.....	0.4809
MONTHS_SINCE_FIRST_GIFT....	0.2863
RECENT_STAR_STATUS.....	0.2068
RECENT_RESPONSE_PROP.....	0.0046
PER_CAPITA_INCOME.....	0.0033
FREQUENCY_STATUS_97NK1....	0.0033
NUMBER_PROM_12.....	0.0023
PEP_STAR.....	0.0022
MEDIAN_HOME_VALUE.....	0.0019
DONOR_AGE.....	0.0013

Figure 5: Chosen Tree Depth = 7 with Feature Importance.

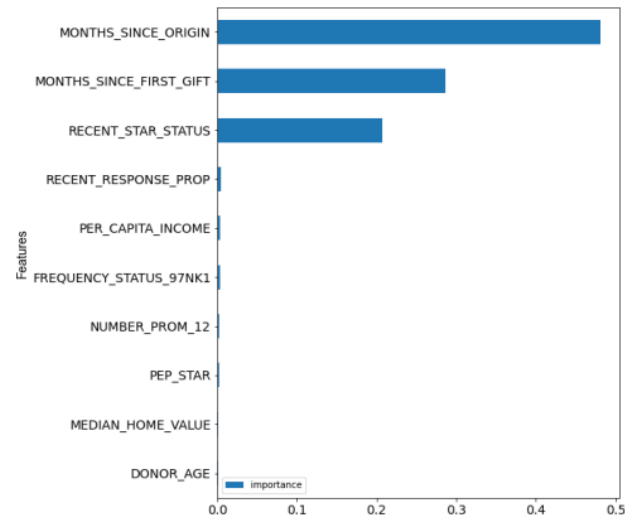


Figure 6: Graph of Feature Importance.

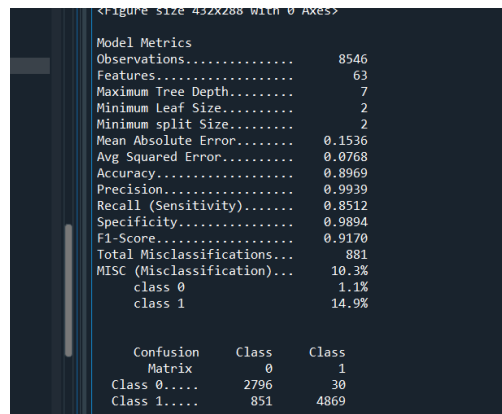


Figure 7: Model Metrics for Decision Tree Depth = 7 with Cross Fold Validation.

### Evaluating Chosen Parameters of Decision Tree Model with 70/30 Validation

When doing validation testing with the chosen model, the results seen in **Figures 8-10** were generated. The MISC and VMISC were reasonably close which means overfitting is not a concern. The model still performed with a reasonable amount of accuracy. The decision tree model is able to identify people who did not donate well, but model performance drops when attempting to identify those who did donate. This is reflected in the specificity and sensitivity ratings of the validation model.

The 70/30 validation indicated the most important features in **Figure 9**. These features imply individuals who are solicited frequently for donations and are known to be active donors are the same ones who donated in response to last year's promotion. They are different in order of importance as compared to those listed in **Figure 6**. The top 3 features maintain their level of importance during model validation which tells us those are the main drivers of this dataset. Essentially, keeping an up to date database of who your donors are increases the likelihood they'll give. Donors are more likely to give based on the frequency of solicitation for donations.

```

***** Best Decision Tree 70/30 Validation *****

FEATURE..... IMPORTANCE
MONTHS_SINCE_ORIGIN..... 0.3650
MONTHS_SINCE_FIRST_GIFT..... 0.2247
RECENT_STAR_STATUS..... 0.2069
LIFETIME_PROM..... 0.1788
RECENT_RESPONSE_PROP..... 0.0067
FILE_CARD_GIFT..... 0.0039
LIFETIME_GIFT_RANGE..... 0.0033
PER_CAPITA_INCOME..... 0.0031
LAST_GIFT_AMT..... 0.0022
INCOME_GROUP6..... 0.0019

Feature Importances:
<Figure size 432x288 with 0 Axes>
<Figure size 432x288 with 0 Axes>

```

Figure 8: Feature Importance for Decision Tree with Chosen depth = 7.

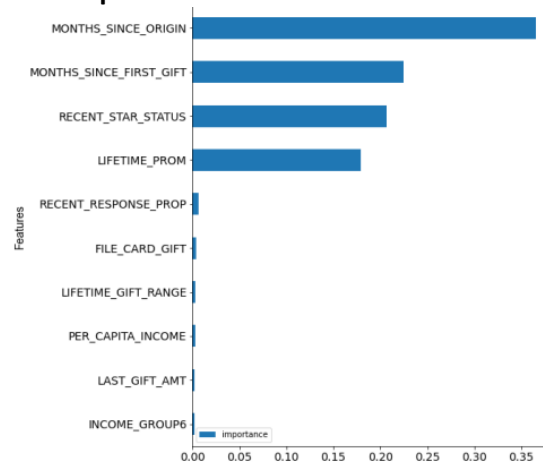


Figure 9: Feature Importance Graph for Decision Tree with 70/30 Split.

Model Metrics.....	Training	Validation
Observations.....	5982	2564
Features.....	63	63
Maximum Tree Depth.....	7	7
Minimum Leaf Size.....	2	2
Minimum split Size.....	2	2
Mean Absolute Error....	0.1552	0.1609
Avg Squared Error.....	0.0776	0.0819
Accuracy.....	0.8955	0.8908
Precision.....	0.9888	0.9816
Recall (Sensitivity).....	0.8544	0.8507
Specificity.....	0.9801	0.9689
F1-score.....	0.9167	0.9115
Total Misclassifications...	625	280
MISC (Misclassification)...	10.4%	10.9%
class 0.....	2.0%	3.1%
class 1.....	14.6%	14.9%

Training	Class	Class
Confusion Matrix	0	1
Class 0.....	1918	39
Class 1.....	586	3439

Validation	Class	Class
Confusion Matrix	0	1
Class 0.....	842	27
Class 1.....	253	1442

Figure 10: 70/30 Split of Decision Tree with depth = 7.



## Random Forest Model

### Choosing Parameters with Cross Fold Validation

After much trial and error the model generated was the one as seen in **Figures 11 and 12**. Originally, nothing was done to inhibit tree growth by the way of leaf or split size which resulted in overfitting. By enforcing a leaf and split size, model overfitting was minimized but not eliminated. Adjusting the number of trees and depth finally results in an ideal model with 395 trees and a depth of 7. The model does not seem to be overexplaining (or overfitting) our dataset, and the best depth of 7 validates the decision tree model selection conducted previously. The difference between the random forest and decision tree models was the decision to utilize a cross validation of 4 instead of 10 for random forest. The only thing this affected was the speed at which the model parameters were determined.

```
Evaluate Using Entire Dataset with Best Parameters
Best Number of Trees (estimators) = 395
Best Depth = 7
Best Leaf Size = 9
Best Split Size = 18
Best Max Features = None

Model Metrics
Observations..... 8546
Features..... 63
Maximum Tree Depth..... 7
Minimum Leaf Size..... 9
Minimum split Size..... 18
Mean Absolute Error..... 0.1575
Avg Squared Error..... 0.0753
Accuracy..... 0.8963
Precision..... 0.9959
Recall (Sensitivity)..... 0.8486
Specificity..... 0.9929
F1-Score..... 0.9164
Total Misclassifications... 886
MISC (Misclassification)... 10.4%
  class 0      0.7%
  class 1     15.1%

Confusion Matrix
Class 0..... 2806
Class 1..... 866 4854

FEATURE..... IMPORTANCE
MONTHS_SINCE_ORIGIN..... 0.3585
MONTHS_SINCE_FIRST_GIFT... 0.3118
RECENT_STAR_STATUS..... 0.1864
LIFETIME_PROM..... 0.0897
RECENCY_STATUS_96NK0:A..... 0.0211
PER_CAPITA_INCOME..... 0.0041
RECENT_CARD_RESPONSE_PROP.. 0.0037
RECENT_RESPONSE_PROP..... 0.0033
MEDIAN_HOME_VALUE..... 0.0023
WEALTH_RATING..... 0.0017
```

Figure 11: Chosen Metrics for the Best Model.

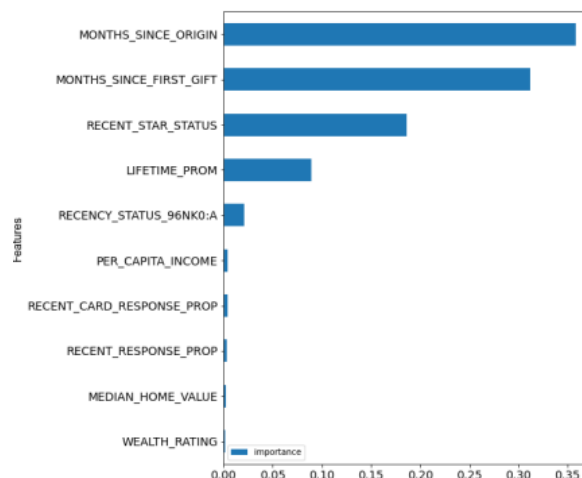


Figure 12: Feature Importance with Chosen Metrics for Best Model.

Sara Foster  
STAT 656 MIDTERM  
Student ID # 821007101

### Evaluating Chosen Parameters of Random Forest Model with 70/30 Validation

**Figures 13 and 14** show the model validation. The model performed well within the holdout set, meaning it did not overfit our data and still performed reasonably well. The nine ten features identified as important within the model were MONTHS\_SINCE\_FIRST\_GIFT, MONTHS\_SINCE\_ORIGIN, RECENT\_STAR\_STATUS, LIFETIME\_PROM, RECENCY\_STATUS\_96, RECENT\_CARD\_RESPONSE\_PROP, RECENT\_RESP\_PROP, WEALTH\_RATING, and PER\_CAPITA\_INCOME. These varied in importance within the full dataset as well as the validation, however seem to be the driving predictors of the model. Random forest reinforces how all the models have difficulty in predicting if someone donated in response to the 97NK mail solicitation from the organization.

```
Evaluating Using 70/30 Partition
Evaluating Best Random Forest
Best Trees= 395
Best Depth= 7
Best Leaf Size = 9
Best Split Size = 18
Best Max Features = None

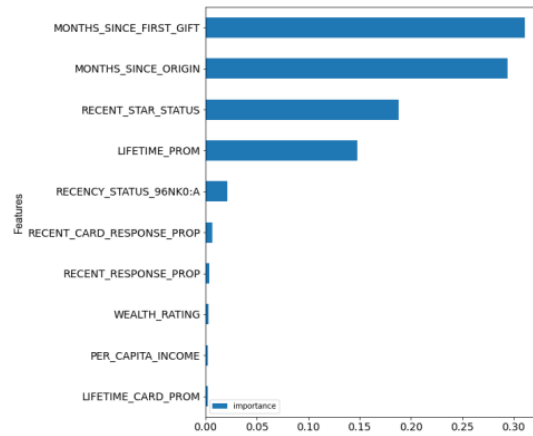
Model Metrics..... Training Validation
Observations..... 5982 2564
Features..... 63 63
Maximum Tree Depth.... 7 7
Minimum Leaf Size..... 9 9
Minimum split Size.... 18 18
Mean Absolute Error.... 0.1615 0.1665
Avg Squared Error..... 0.0754 0.0787
Accuracy..... 0.8962 0.8959
Precision..... 0.9956 0.9924
Recall (Sensitivity)..... 0.8494 0.8490
Specificity..... 0.9923 0.9873
F1-score..... 0.9167 0.9151
Total Misclassifications... 621 267
MISC (Misclassification)... 10.4% 10.4%
  class 0..... 0.8% 1.3%
  class 1..... 15.1% 15.1%

Training Class Class
Confusion Matrix 0 1
Class 0..... 1942 15
Class 1..... 606 3419

Validation Class Class
Confusion Matrix 0 1
Class 0..... 858 11
Class 1..... 256 1439
```

Figure 13: Evaluating Model with 70/30 Validation.

```
FEATURE..... IMPORTANCE
MONTHS_SINCE_FIRST_GIFT.... 0.3111
MONTHS_SINCE_ORIGIN..... 0.2944
RECENT_STAR_STATUS..... 0.1877
LIFETIME_PROM..... 0.1477
RECENCY_STATUS_96NK0:A.... 0.0214
RECENT_CARD_RESPONSE_PROP.. 0.0067
RECENT_RESPONSE_PROP..... 0.0033
WEALTH_RATING..... 0.0032
PER_CAPITA_INCOME..... 0.0022
LIFETIME_CARD_PROM..... 0.0022
```



**Figure 14: Feature Importance for the 70/30 Validation.**

## Recommendations

Comparing all three models to each other, the random forest model would be the most effective model for this dataset because it has the lowest VMISC without overfitting the data at 10.4% with high recall and specificity. The VMISC and MISC are nearly exactly the same which means the model has not been overfitted to the dataset, whereas the decision tree and logistic regression both had about 0.5% discrepancies. All three models had trouble with predicting whether an individual donated in response to the 97NK mail solicitation, but good at knowing when a person did not donate in response to the 97NK mail solicitation.

There were five common factors within the decision tree and random forest model: lifetime prom, months since first gift, months since origin, recent star status, and one of the wealth indicators. The wealth indicator was generally either the wealth rating or the per capita income as a driving predictor. By examining these variables and the importance they play in the random forest model, it is important for the organization to keep an up to date records on their donors and to consistently send them promotions to donate. An individuals income level is clearly a factor in whether or not they donate which is another component to keep in mind. If this organization is trying to increase or maintain their level of donors, then it would be prudent for them to target high income individuals consistently in addition to maintaining accurate records of these high income donors. Being a star status donor also implies a person will be more likely to donate, so ensuring donors attain star status also increases their likelihood of donating.

Overall, the random forest model does a good job at predicting who will not donate, and still performs reasonably well at predicting who will donate. This model can be utilized to determine who will not donate consistently. This is value added information regardless because it can drive the organization to steer clear of undesirable donors and target those who will consistently donate money to their cause.