IDS 572
Assignment – 2
Target Marketing – Fundraising

Bala Rohit Yeruva

Fall 2016

1. Data Exploration & Cleaning

After observing the data, we could see that there are 480 independent variables and one dependent variable (TARGET_B). The first task before building a model is to reduce and clean the huge dataset. We have undertaken this daunting task starting with counting the total number of variables with missing values. It is observed that out of 480 variables, there were 161 variables with missing values. We have then followed the below step-wise approach in cleaning and reducing the data:

Generating New Variables:

Out of the 161 variables which have missing values, there are a few variables were the missing values have a meaningful notation based on the variable definition. We have found 34 such variables. For all these variables, we have generated new variables where the missing values were substituted by their meaningful notation and the old variables were removed. Additionally, there were four date variables which deemed useful and were converted to the number of days from promotion day. The following are a few such variables.

Original Attribute	New Attribute
NUMCHLD (Missing – No Chidren)	numChild if(NUMCHLD=="1"||NUMCHLD=="2"|| NUMCHLD=="3"||NUMCHLD=="4"|| NUMCHLD=="5", NUMCHLD, 0)
PETS (Y – Interested; Missing – Not Interested)	pets if (PETS == "Y", "1", "0")
FISTDATE (First Gift Date)	daysFromFirstGift date_diff(FISTDATE, ADATE_2)/(1000*60*60*24)

Table 1.1

Eliminating Useless Attributes:

After the above process, we still have all the 480 variables. We have then removed a few variables in the following process:

a. Eliminating based on other Variables: Out of the 480 variables, we have found that around 96 variables have their summary defined in another variables. Hence these variables were removed from the dataset. The following are a few such variables.

Attributes	Reason
ADATE2 – ADATE24, RFA2 – RFA24	Summary of these attributes is captures in CARDPROM, MAXADATE, NUMPROM, CARDPRM12, NUMPRM12
RDATE3 – RDATE24, RAMNT3 – RAMNT24	Summary of these attributes is captures in NGIFTALL, CARDGIFT, MINRAMNT, MINRDATE

Table 1.2

b. Eliminating based on usefulness and intuition: After the above process, we were left with 384 variables out of which 14 variables were removed which were deemed to be of less or no importance in predicting the donors. The following are a few such variables.

Attributes	Reason
DOB	This attributes was removed because we already have the AGE attribute. Moreover, we believed it wouldn't impact the outcome.
AGEFLAG	This attributes indicates if source of AGE. Even this variable is of no importance as it has no impact on the outcome.
GEOCODE	This variable indicates the level geography at which a record matches the census data which has no effect on whether a person would make a donation.
CONTROLN	This is a unique record identifier which has no impact on the outcome.
TARGET_D	This variable indicates the amount donated by the donor. This variable has values only for donor and hence including it would be inappropriate.

Table 1.3

Replacing Missing Values:

After elimination the above variables, we have 370 variables with 26 variables having missing values. Out of these 26 variables, there were 20 quantitative variables and 6 categorical variables. We have imputed these missing values in the following process:

a. Replacing Missing Values using Average: The missing values in all the 20 quantitative variables were replaced by the average value of that particular variable. The following are a few such variables.

Attributes
AGE
INCOME
MBCOLECT
TIMELAG
WEALTH1
PUBHLTH
MAGFEM
PUBPHOTO

Table 1.4

b. Replacing Missing Values using N: The missing values in all the 6 categorical variables were replaced by the "N". The following are the variables:

Attributes
GEOCODE
GEOCODE2
MAILCODE
OSOURCE
SOLIH
SOLP3

Table 1.5

Principle Component Analysis (PCA) for Variable Reduction:

After imputing the missing values, we have 370 variables with no variables have missing values. We then performed PCA on the following subsets of variables which are correlated between themselves.

- Variables related to Neighborhood: This subset contains a total of 283 variables. After performing PCA on these variables with retaining 80% of the variance, we were able to achieve 17 PC variables.

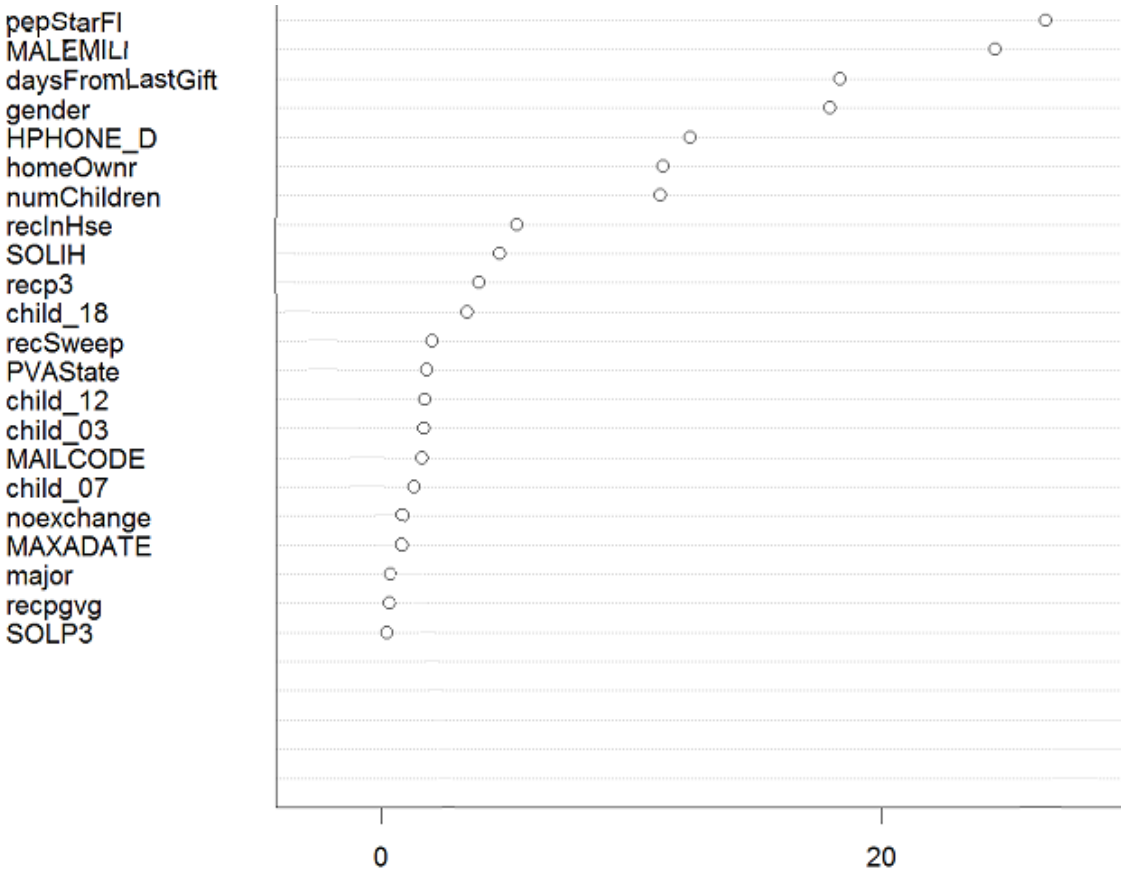
- Variables related to Donor Interests: This subset contains a total of 18 variables. After performing PCA on these variables with retaining 90% of the variance, we were able to achieve 12 PC variables.

- Variables related to Donor's Response on other Mails: This subset contains a total of 14 variables. After performing PCA on these variables with retaining 90% of the variance, we were able to achieve 11 PC variables.

Performing PCA on these subsets reduced the 315 variables to 40 PCs. Therefore, the total number of variables were reduced from 370 to 95.

Variable Reduction: Using Variable Importance from Random Forest:

For the remaining 95 variables, we have used a Random Forest model to calculate the variable importance. From the Variable Importance plot, we have removed the following 22 variables which have less then 25 mean decrease in Gini. For the complete plot, please refer to appendix A.



Final Variables:

After this entire process of data exploration, cleaning and reduction, we were able to reduce the number of variables from 480 to 73 variables with 40 PCs and 33 normal variables.

2. Modelling

After cleaning, reducing and finalizing the number of variables, we have partitioned the dataset into 60:40 for training and validation respectively. We have run each of the models that we have built on the subsets of variables as below:

- All PCAs: This subset has all the PCA variables included. It had 73 variables.
- W/O Neighborhood PCA: This subset does not contain the PCs obtained from running PCA on the neighborhood variables. Rather it includes all the neighborhood variables.
- W/O Mail Response PCA: This subset does not contain the PCs obtained from running PCA on the mail response variables.
- W/O Donor Interest PCA: This subset does not contain the PCs obtained from running PCA on the donor interest variables.
- No PCAs: This subset does not contain any PC variables but includes all the variables on which PCA was executed.

Additionally, while deciding on the best model we have considered both accuracy and the recall or the hit rate for class 1. We have followed this approach because we want a model that can capture or predict as many real donors as possible. Therefore, accuracy without a good recall (Class 1: Donors) wouldn't maximum the profit from the model.

Naïve Bayes:

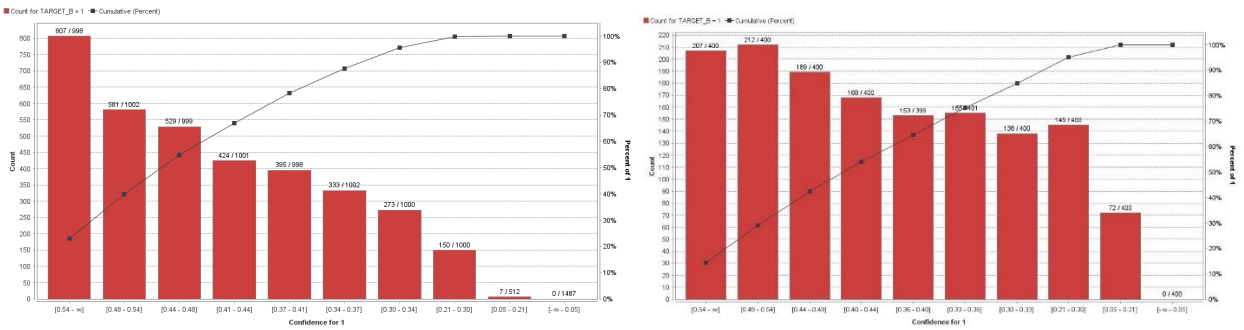
Our first model is a Naïve Bayes classifier which is a simple probabilistic classifier based on Bayes' theorem. When using a Naïve Bayes model, an unbiased classification is hard to achieve because the classification is dependent on the probabilities. We have however, tried to build several the models using different subsets of the data as observed from table 2.1.

Subset	***Accuracy on Training Data***	***Accuracy on Validation Data***	***Precision(1) on Training Data***	***Precision(1) on Validation Data***	***Recall(1) on Training Data***	***Recall(1) on Validation Data***
All PCAs	90.03%	54.27%	84.19%	39.41%	87.37%	49.90%
W/O Neighborhood PCA	**79.48%**	**52.00%**	**65.27%**	**38.89%**	**86.02%**	**58.51%**
W/O Mail Response PCA	90.02%	54.27%	83.84%	39.49%	87.86%	50.94%
W/O Donor Interests PCA	89.83%	53.70%	83.44%	38.95%	87.82%	50.59%
No PCAs	79.06%	51.68%	64.71%	38.70%	85.83%	58.79%

Table 2.1

From this table, we could observe that choosing different subsets of the data will make a difference in the model performance. Further, we could see that when we use all the variables related to the Neighborhood instead of using the corresponding PCA, we have achieved a better accuracy and recall for class 1. Therefore, among the above models, the second model would be the best model to choose in order to maximum the profit.

The following are the lift charts obtained when the above best model is run on the training and the validation dataset respectively. For an enhanced image, please refer to appendix B. For the lift charts & confusion matrices obtained from all the Naïve Bayes models, please refer to appendix B.



Decision Tree: W-J48

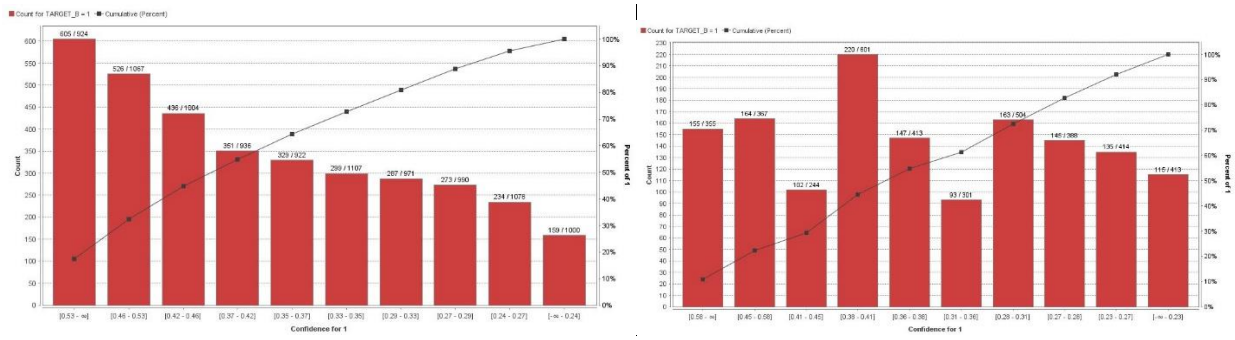
Our second model is a decision tree (W-J48: Weka Extension) which uses the C4.5 algorithm to build a tree. While building the models, the parameters that best worked for us were Laplace smoothing along with a confidence threshold of 0.75 for pruning and 100 as the minimum number of cases at the leaf node. For all the parameters used, please refer to appendix C.

Subset	***Accuracy on Training Data***	***Accuracy on Validation Data***	***Precision(1) on Training Data***	***Precision(1) on Validation Data***	***Recall(1) on Training Data***	***Recall(1) on Validation Data***
All PCAs	68.93%	62.58%	62.34%	44.75%	24.03%	17.16%
W/O Neighborhood PCA	69.03%	62.68%	62.02%	45.53%	25.29%	19.11%
W/O Mail Response PCA	68.93%	62.58%	62.34%	44.75%	24.03%	17.16%
W/O Donor Interests PCA	68.93%	62.58%	62.34%	44.75%	24.03%	17.16%
No PCAs	**69.04%**	**62.65%**	**61.29%**	**45.71%**	**26.75%**	**20.36%**

Table 2.2

From the table, we could observe that choosing different subsets of the data will have less impact on the model performance. However, we could see that the model with no PCAs has better Recall for Class 1 along with the accuracy. Therefore, the fifth model would be our best decision tree model.

The following are the lift charts obtained when the above best model is run on the training and the validation dataset respectively. For an enhanced image, please refer to appendix C. For the lift charts & confusion matrices obtained from all the Decision Tree models, please refer to appendix C.



Random Forest: w-RandomForest

The next model that we have built is a Random Forest Model (Weka Extension). We have achieved better results while setting the maximal tree depth value to 100 and number of trees to ten. For all the parameters used, please refer to appendix D.

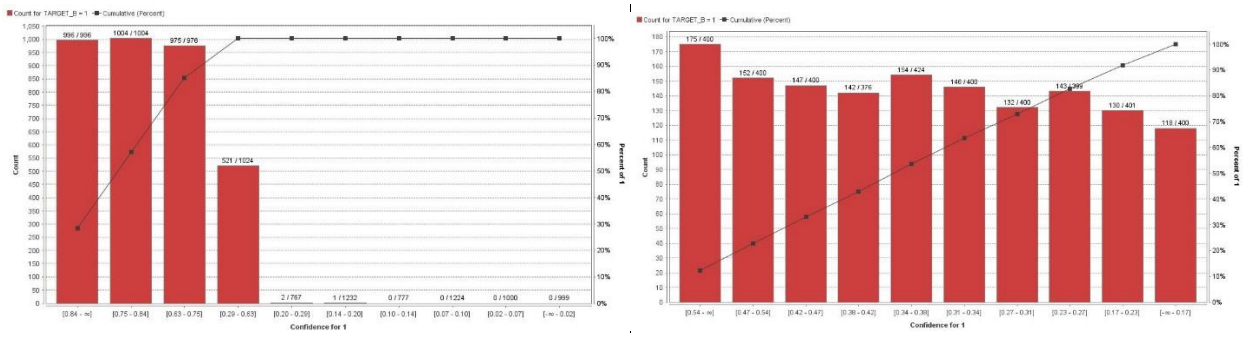
Subset	***Accuracy on Training Data***	***Accuracy on Validation Data***	***Precision(1) on Training Data***	***Precision(1) on Validation Data***	***Recall(1) on Training Data***	***Recall(1) on Validation Data***
All PCAs	99.20%	62.00%	99.95%	42.49%	97.72%	15.91%
W/O Neighborhood PCA	99.03%	61.03%	99.95%	40.68%	97.23%	16.54%
W/O Mail Response PCA	99.25%	61.60%	99.90%	41.29%	97.91%	15.98%
W/O Donor Interests PCA	99.22%	60.40%	99.80%	38.29%	97.91%	16.47%
NO PCAs	**98.92%**	**61.27%**	**99.75%**	**41.25%**	**97.09%**	**17.16%**

Table 2.3

From the above table, similar to decision tree models, we could observe that choosing different subsets of the data will have less impact on the model performance. We could see that the model with no PCAs had a Recall value of 17.16% which is not a preferable value.

However, when compared with the other models, it is relatively doing better. Therefore, if we had to choose a better model among the above models, then we would choose the fifth model.

The following are the lift charts obtained when the above better model is run on the training and the validation dataset respectively. For an enhanced image, please refer to appendix D. For the lift charts & confusion matrices obtained from all the Random Forest models, please refer to appendix D.



Booted trees: Gradient Boosted Trees

We have used Gradient Boosted Tree to build Boosted Tree models. We have achieved better results while setting the maximal tree depth value to five and minimum number of cases at a leaf node to ten and number of trees to 20. For all the parameters used, please refer to appendix E.

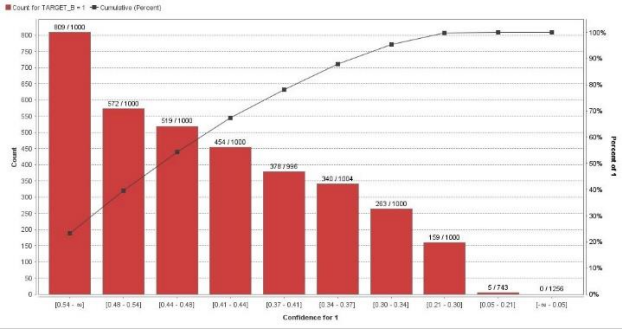
Subset	*Accuracy on Training Data*	*Accuracy on Validation Data*	*Precision(1) on Training Data*	*Precision(1) on Validation Data*	*Recall(1) on Training Data*	*Recall(1) on Validation Data*
All PCAs	**68.66%**	**60.30%**	**52.66%**	**46.70%**	**86.36%**	**73.25%**
W/O Neighborhood PCA	68.94%	59.15%	52.95%	45.69%	85.92%	71.86
W/O Mail Response PCA	69.76%	60.72%	53.93%	46.81%	81.99%	67.34%
W/O Donor Interests PCA	68.79%	59.40%	52.78%	45.90%	86.70%	71.92%
No PCAs	69.18%	59.52%	53.24%	45.91%	84.27%	70.12%

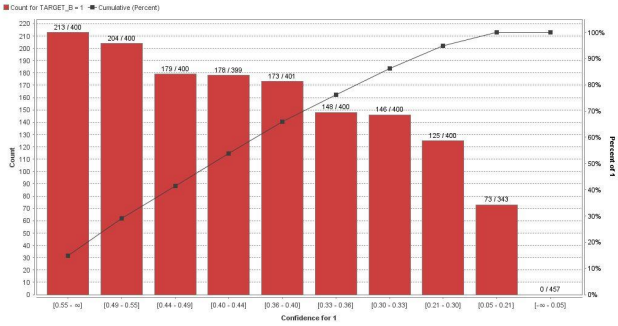
Table 2.4

From the above table, we could observe that choosing different subsets of the data will make a difference in the model performance. Further, we could see that the model with all the PCAs included has the best Recall value for Class 1 along with a reasonable accuracy. We could also

note that this model has the best recall value for Class 1 for all the models built till now. Therefore, the first model above would be our best Boosted Tree model.

The following are the lift charts obtained when the above better model is run on the training and the validation dataset respectively. For an enhanced image, please refer to appendix E. For the lift charts & confusion matrices obtained from all the Boosted Tree models, please refer to appendix E.





Logistic Regression: w-Logistic

Logistic Regression is a technique used for building and using a multinomial logistic regression model to predict a binary response, in this case, to distinguish a donor from non-donor. The same scenarios were tested for logistic regression with Laplace correction. For all the parameters used, please refer to appendix F.

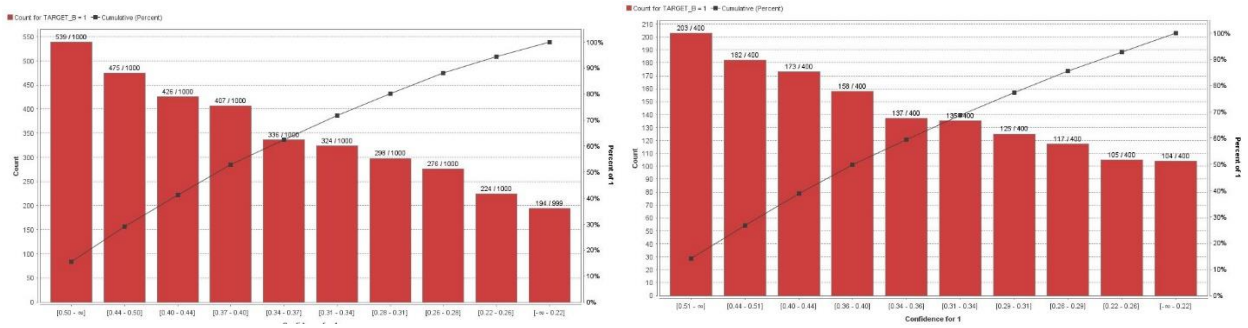
Subset	***Accuracy on Training Data***	***Accuracy on Validation Data***	***Precision(1) on Training Data***	***Precision(1) on Validation Data***	***Recall(1) on Training Data***	***Recall(1) on Validation Data***
All PCAs	**52.43%**	**50.12%**	**40.24%**	**39.91%**	**79.42%**	**76.44%**
W/O Neighborhood PCA	60.03%	53.57%	45.01%	40.78%	73.98%	64.28%
W/O Mail Response PCA	51.34%	48.95%	39.84%	39.52%	81.75%	79.01%
W/O Donor Interests PCA	52.11%	49.43%	40.14%	39.57%	80.29%	77.00%
No PCAs	63.38%	51.65%	48.03%	39.46%	81.02%	64.42%

Table 2.5

From the table, we could see that with every subset of the data, we were able to achieve almost equal accuracy on the training and the validation data. Additionally, we could see that the performance also varies with the variation in the subset of the data. Even here, we could

see that the model in which all the PCAs were included has a better Recall value for Class 1 and a reasonable accuracy. Therefore, the first model will be our best Logistic Regression model.

The following are the lift charts obtained when the above better model is run on the training and the validation dataset respectively. For an enhanced image, please refer to appendix F. For the lift charts & confusion matrices obtained from all the Logistic Regression models, please refer to appendix F.



Comparative Evaluation of all the best models from each technique:

After finalizing on the best models from each technique, we compared the performance of each model against the other model as seen in table 2.6. We then observe that among all the models, the model from Logistic Regression has the highest Recall value for Class 1 but the model from Boosted trees has a higher accuracy along with a good Recall value which is just slightly less than that of the Logistic Regression model.

Model	***Accuracy on Training Data***	***Accuracy on Validation Data***	***Recall(1) on Training Data***	***Recall(1) on Validation Data***
Naïve Bayes	79.48%	52.00%	86.02%	58.51%
Decision Trees	69.04%	62.65%	26.75%	20.36%
Random Forest	98.92%	61.27%	97.09%	17.16%
Boosted Trees	68.66%	**60.30%**	86.36%	**73.25%**
Logistic Regression	52.43%	50.12%	79.42%	76.44%

Table 2.6

Therefore, from the above analysis we have concluded that from all the models built, the model built using **Gradient Boosted Trees is our best model**.

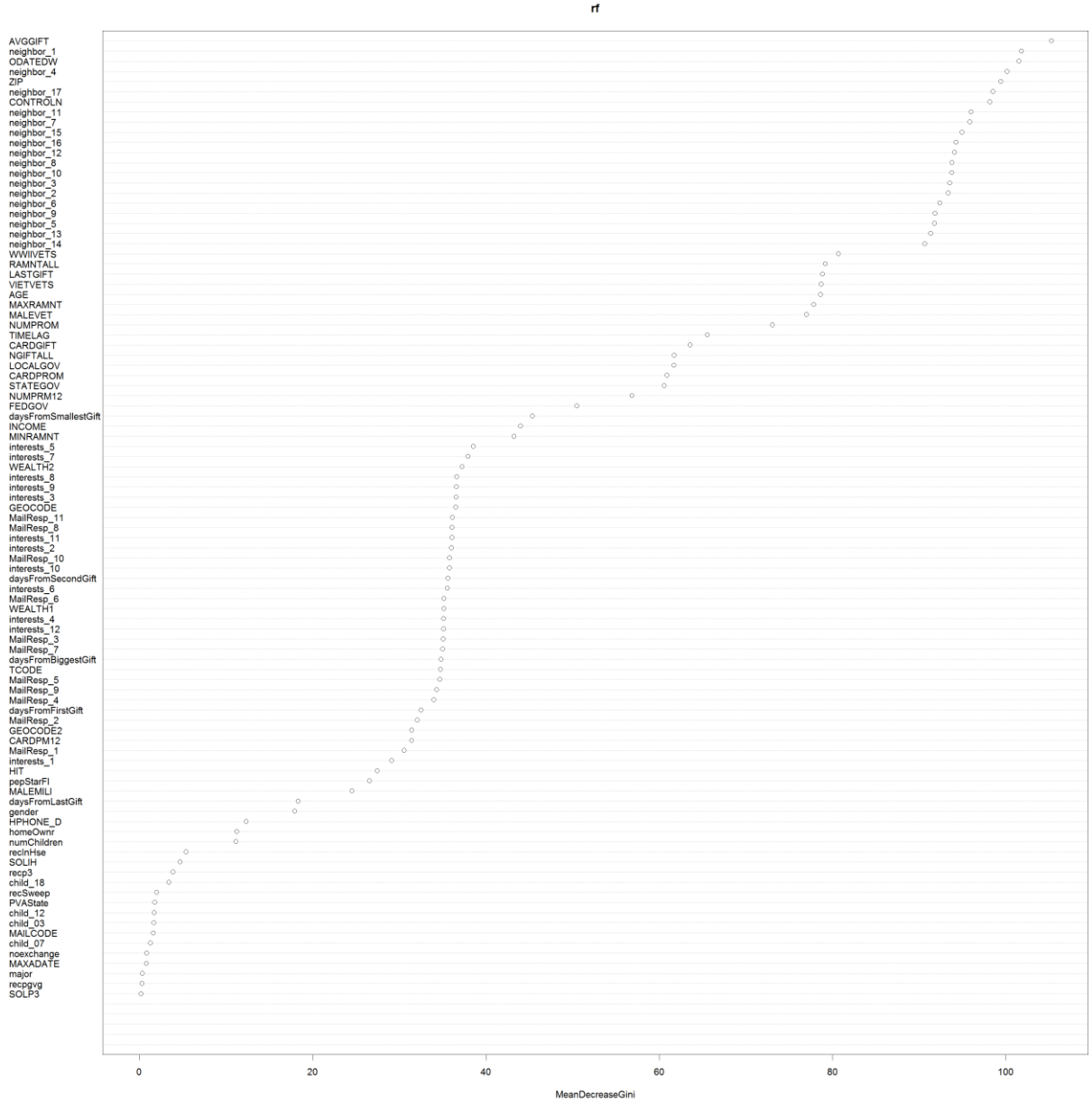
3. Classification under Asymmetric Response and Cost

We need to use weighted sampling such that we have appropriate proportion of each class so that we can obtain more information from the rare cases. In the original dataset, since the response rate is only 5.1% and if we use simple random sampling, we might end up in building a model on the data which has only non-responders or a very few responders.

In most of the cases, a model's performance is determined by the accuracy of the model. However, in specific scenarios as in our case, in order to achieve maximum net profit, a model must capture as many real donors as possible. Therefore, while selecting our best model, we would also concentrate on the class recall or the hit rate for the donors.

Appendix

Appendix A



Appendix B

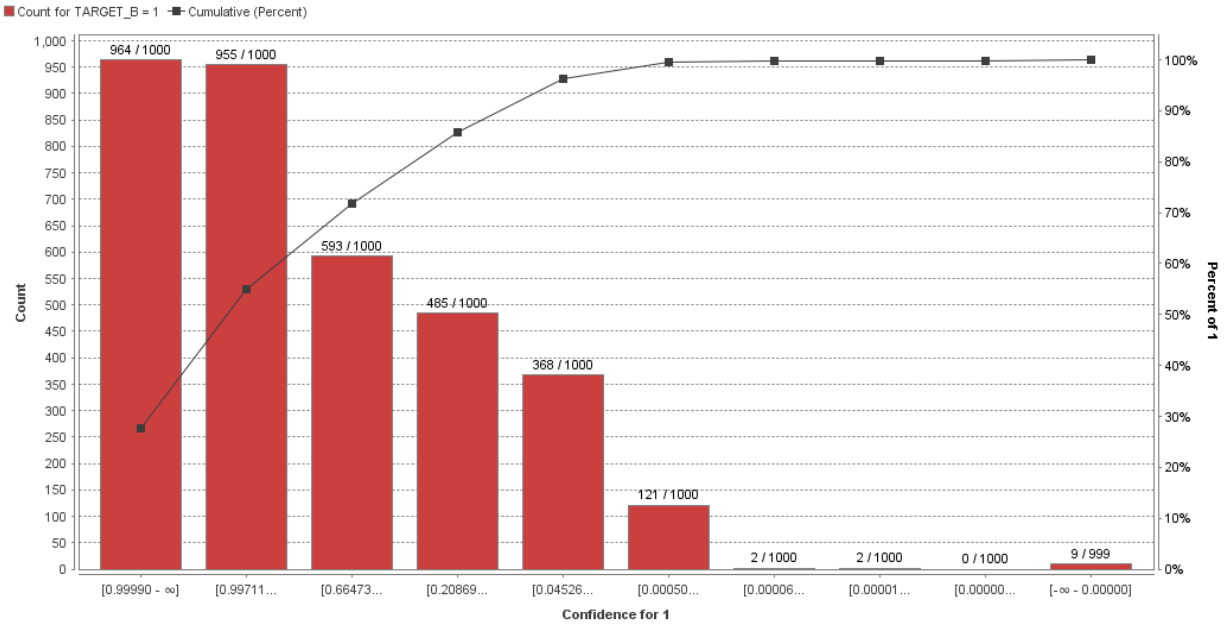
Naïve Bayes Models

All PCAs:

Training Data

accuracy: 90.03%

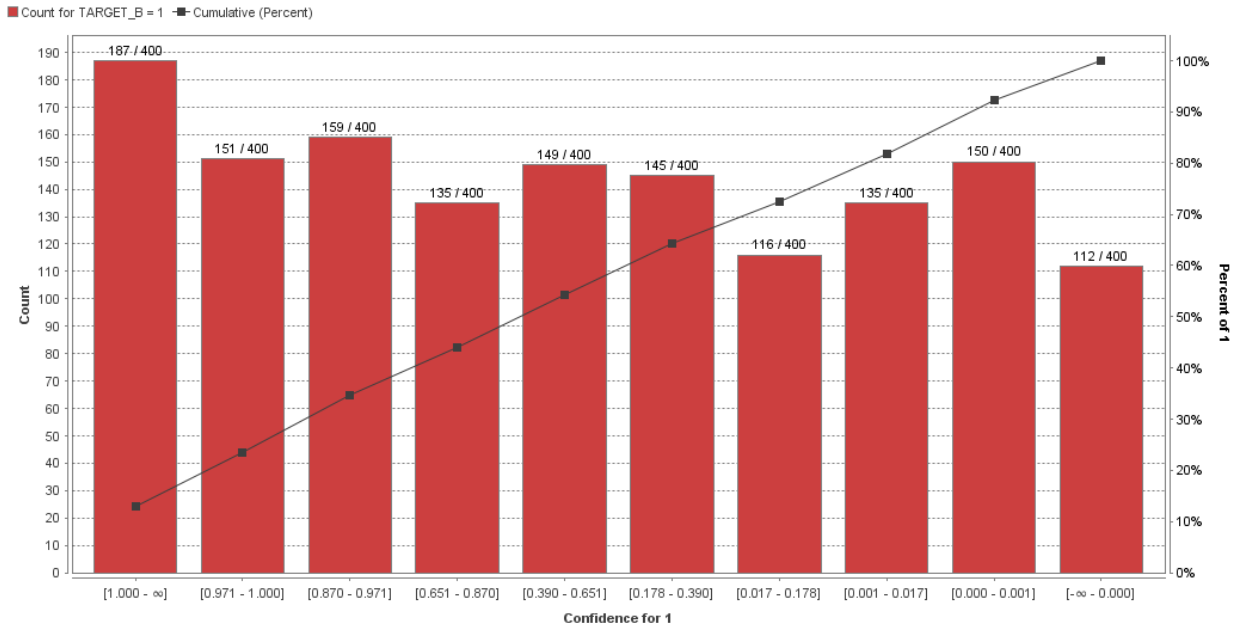
	true 0	true 1	class precision
pred. 0	3601	260	93.27%
pred. 1	338	1800	84.19%
class recall	91.42%	87.38%	



Validation Data

accuracy: 54.37%

	true 0	true 1	class precision
pred. 0	1457	721	66.90%
pred. 1	1104	718	39.41%
class recall	56.89%	49.90%	

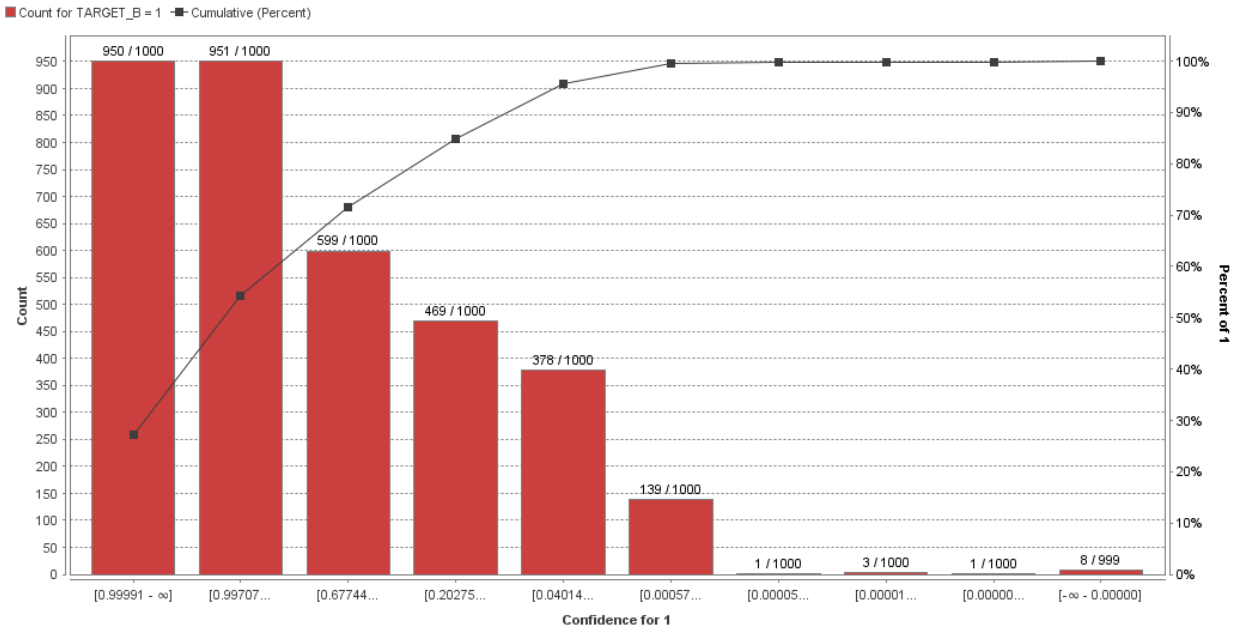


W/O Neighborhood PCA:

Training Data

accuracy: 79.48%

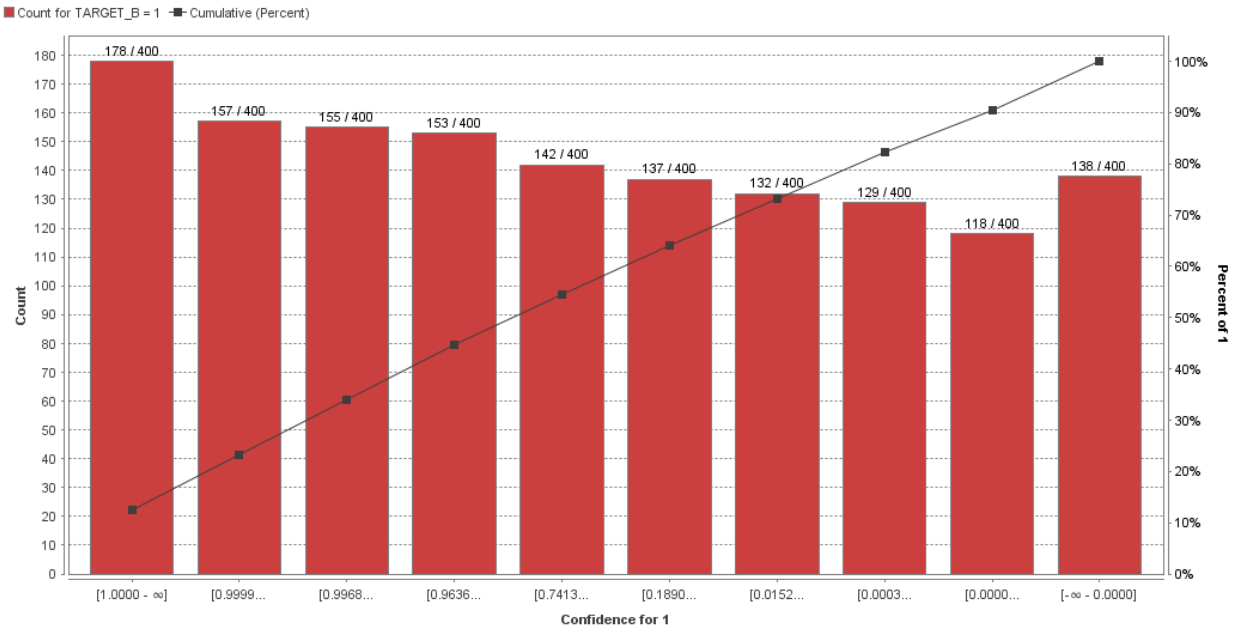
	true 0	true 1	class precision
pred. 0	2996	288	91.23%
pred. 1	943	1772	65.27%
class recall	76.06%	86.02%	



Validation Data

accuracy: 52.00%

	true 0	true 1	class precision
pred. 0	1238	597	67.47%
pred. 1	1323	842	38.89%
class recall	48.34%	58.51%	

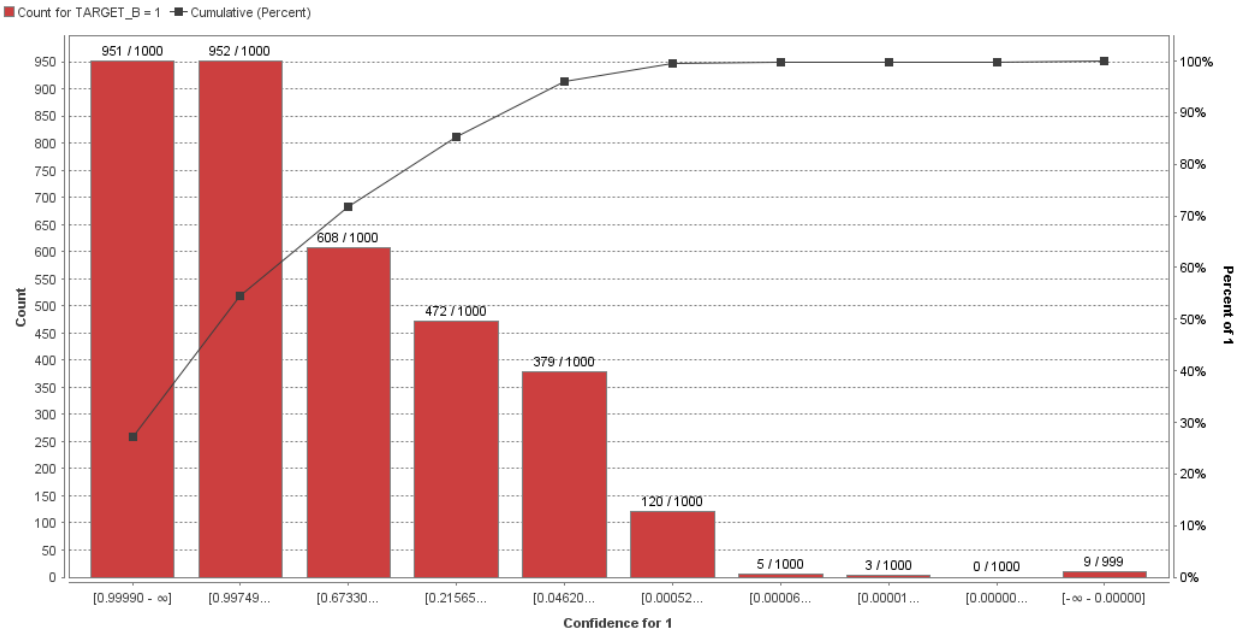


W/O Mail Responses PCA:

Training Data

accuracy: 90.02%

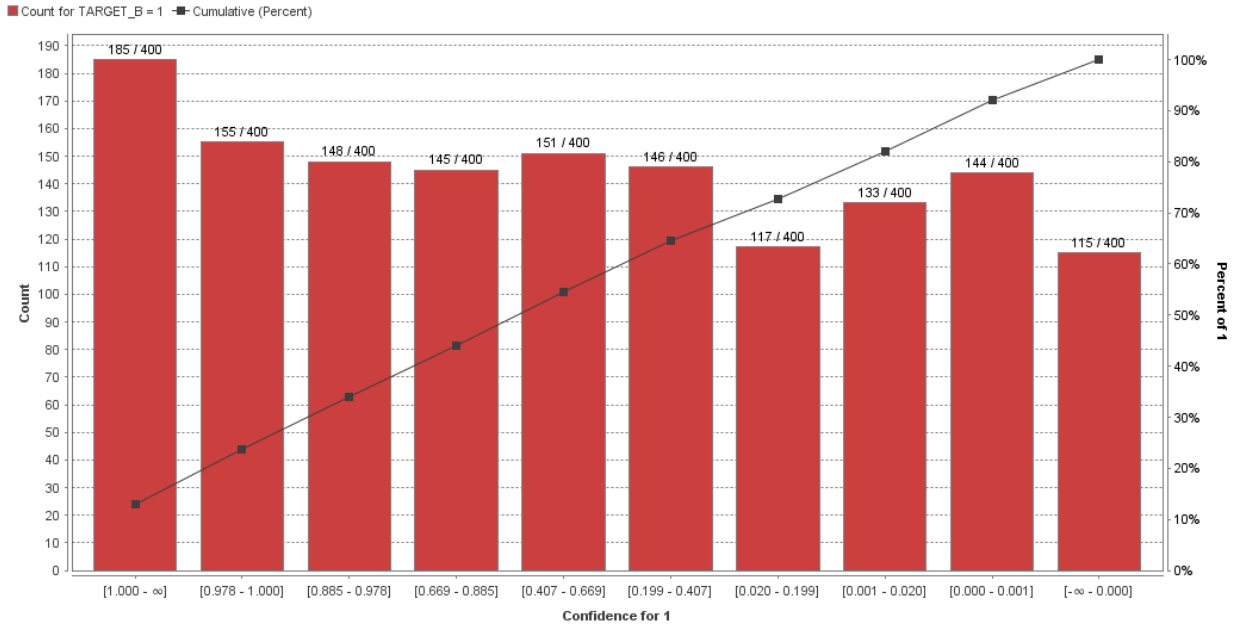
	true 0	true 1	class precision
pred. 0	3590	250	93.49%
pred. 1	349	1810	83.84%
class recall	91.14%	87.86%	



Validation Data

accuracy: 54.27%

	true 0	true 1	class precision
pred. 0	1438	706	67.07%
pred. 1	1123	733	39.49%
class recall	56.15%	50.94%	

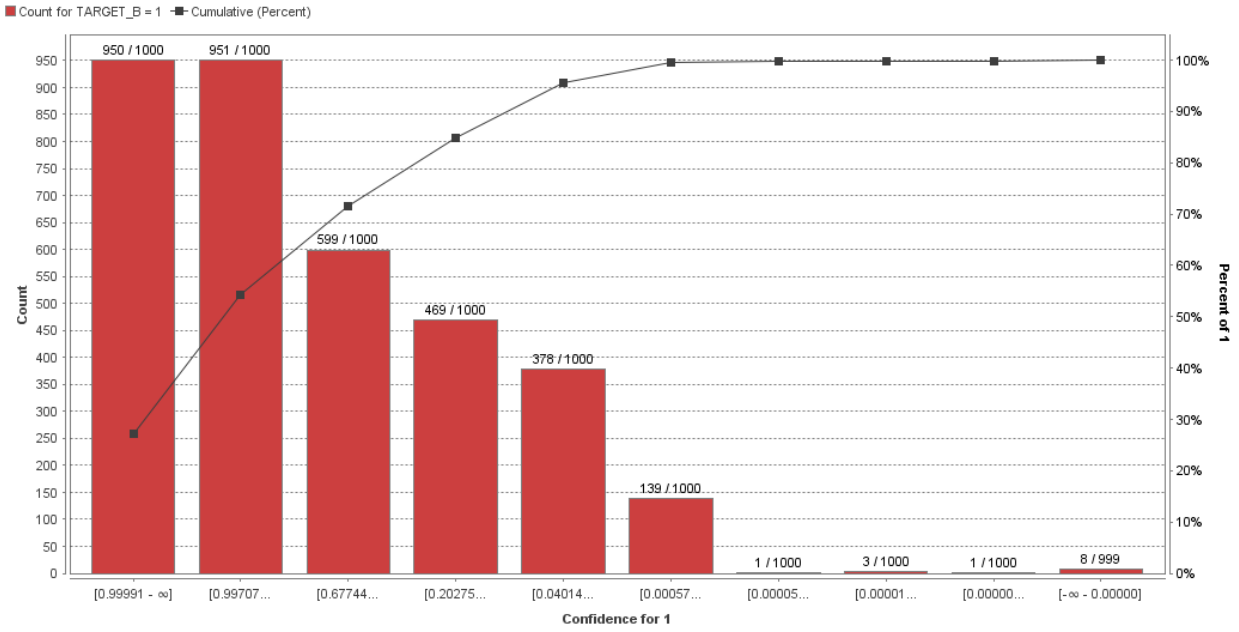


W/O Donor Interests PCA:

Training Data

accuracy: 89.83%

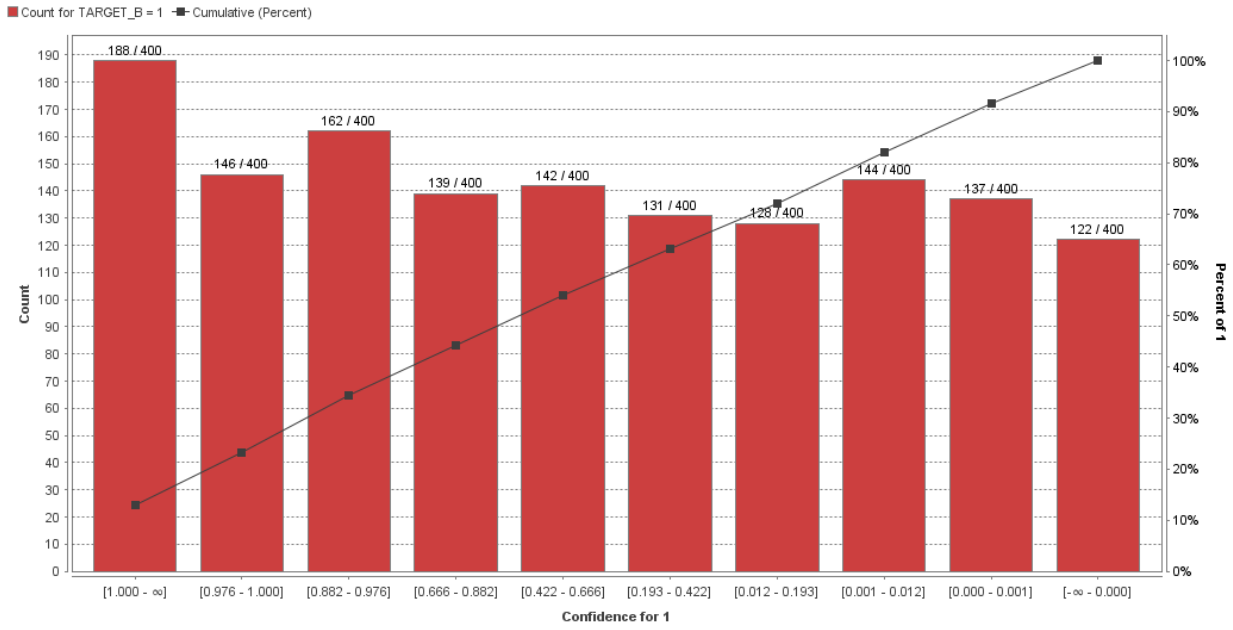
	true 0	true 1	class precision
pred. 0	3580	251	93.45%
pred. 1	359	1809	83.44%
class recall	90.89%	87.82%	



Validation Data

accuracy: 53.70%

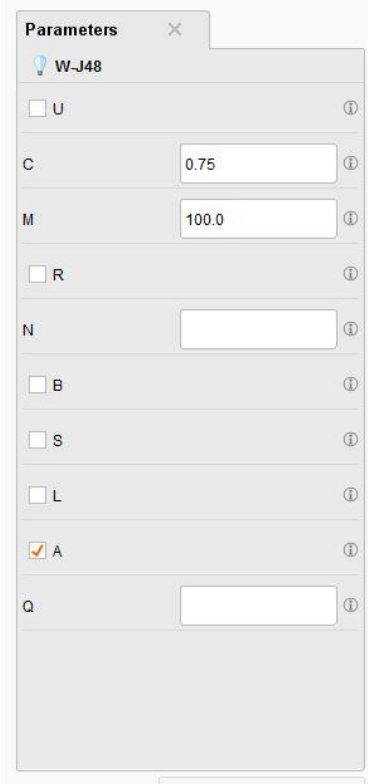
	true 0	true 1	class precision
pred. 0	1420	711	66.64%
pred. 1	1141	728	38.95%
class recall	55.45%	50.59%	



Appendix C

Decision Tree Models

Parameters:

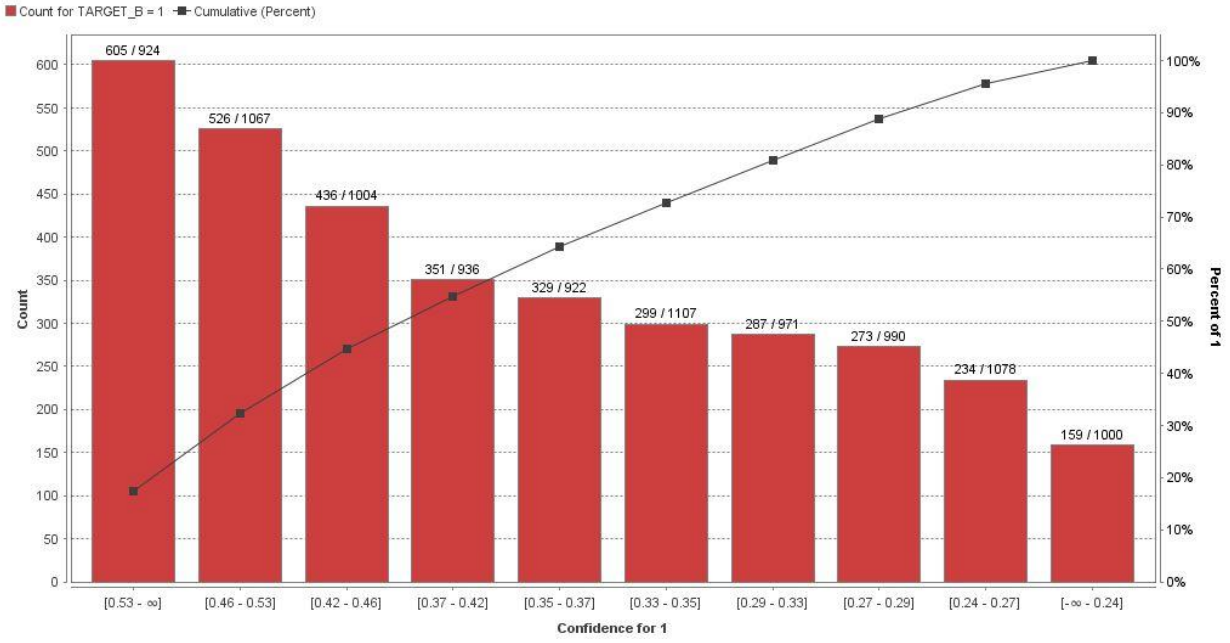


All PCAs:

Training Data

accuracy: 68.93%

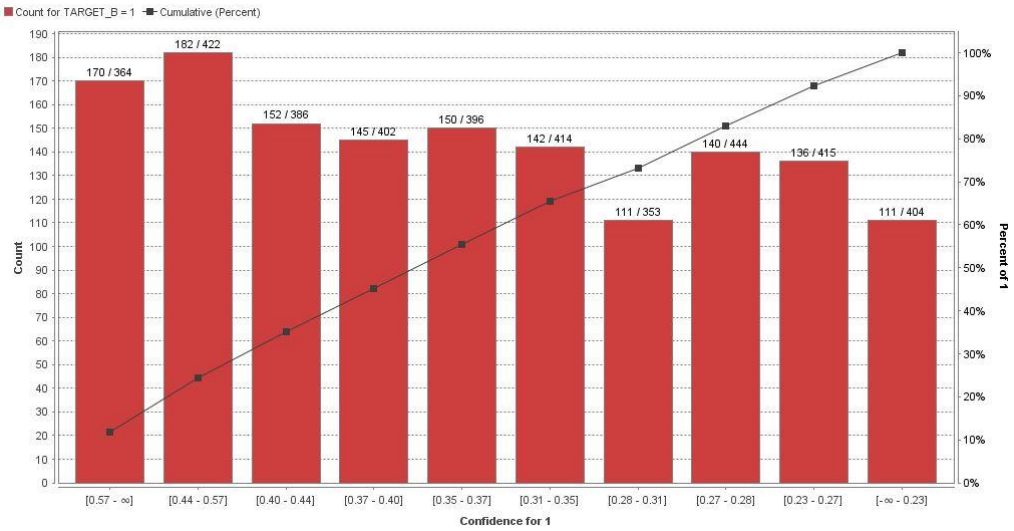
	true 0	true 1	class precision
pred. 0	3640	1565	69.93%
pred. 1	299	495	62.34%
class recall	92.41%	24.03%	



Validation Data

accuracy: 62.58%

	true 0	true 1	class precision
pred. 0	2256	1192	65.43%
pred. 1	305	247	44.75%
class recall	88.09%	17.16%	

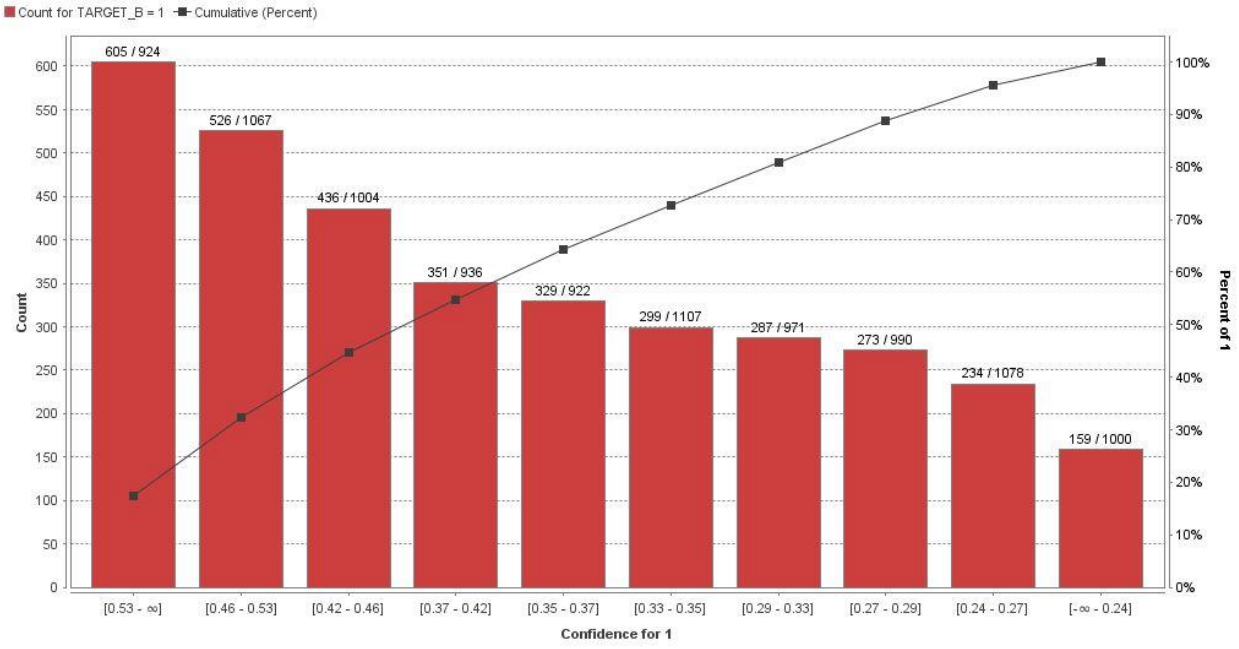


W/O Neighborhood PCA:

Training Data

accuracy: 69.03%

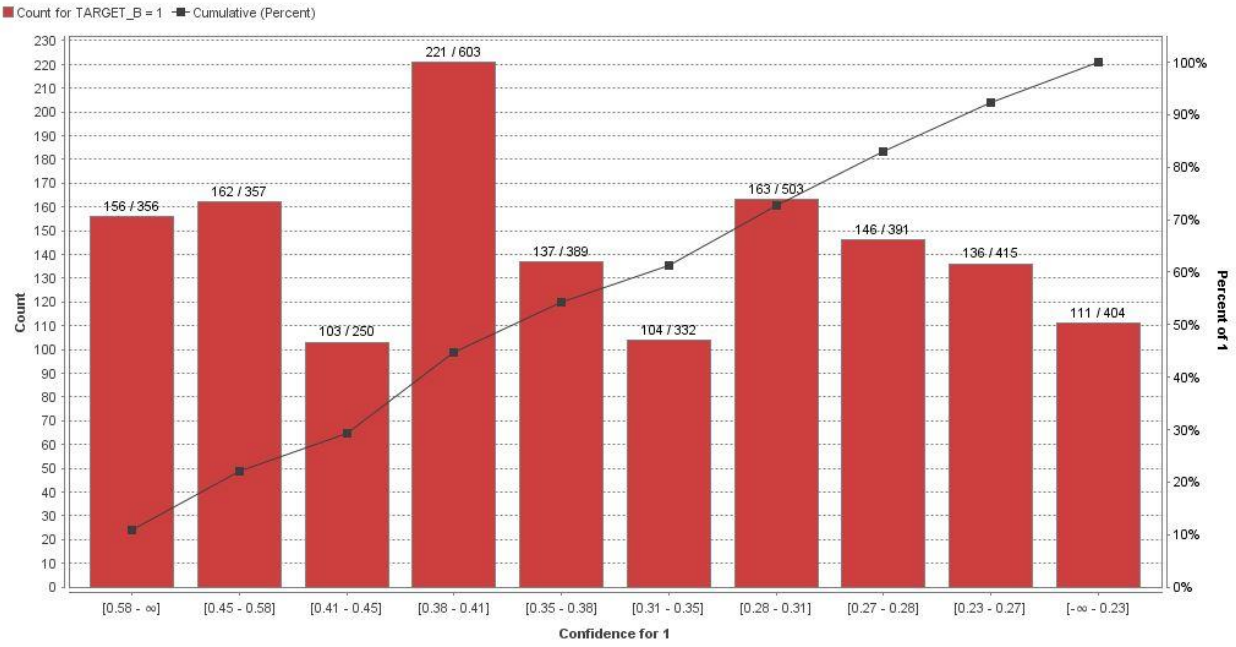
	true 0	true 1	class precision
pred. 0	3620	1539	70.17%
pred. 1	319	521	62.02%
class recall	91.90%	25.29%	



Validation Data

accuracy: 62.68%

	true 0	true 1	class precision
pred. 0	2232	1164	65.72%
pred. 1	329	275	45.53%
class recall	87.15%	19.11%	

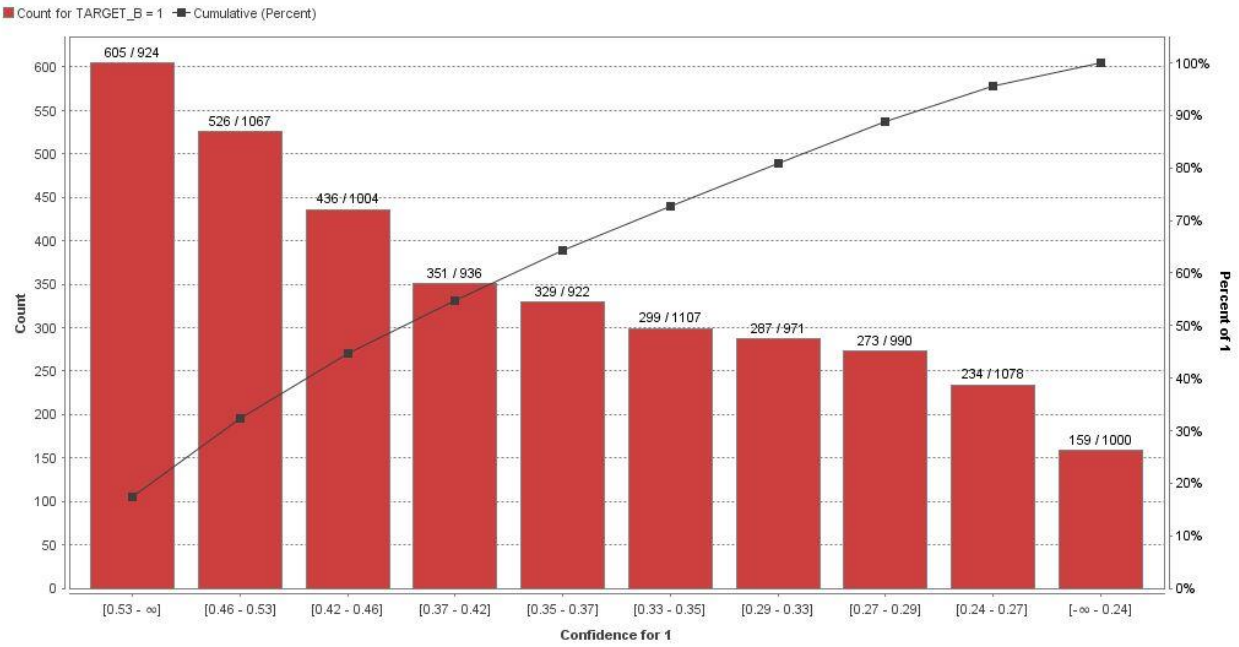


W/O Mail Responses PCA:

Training Data

accuracy: 68.93%

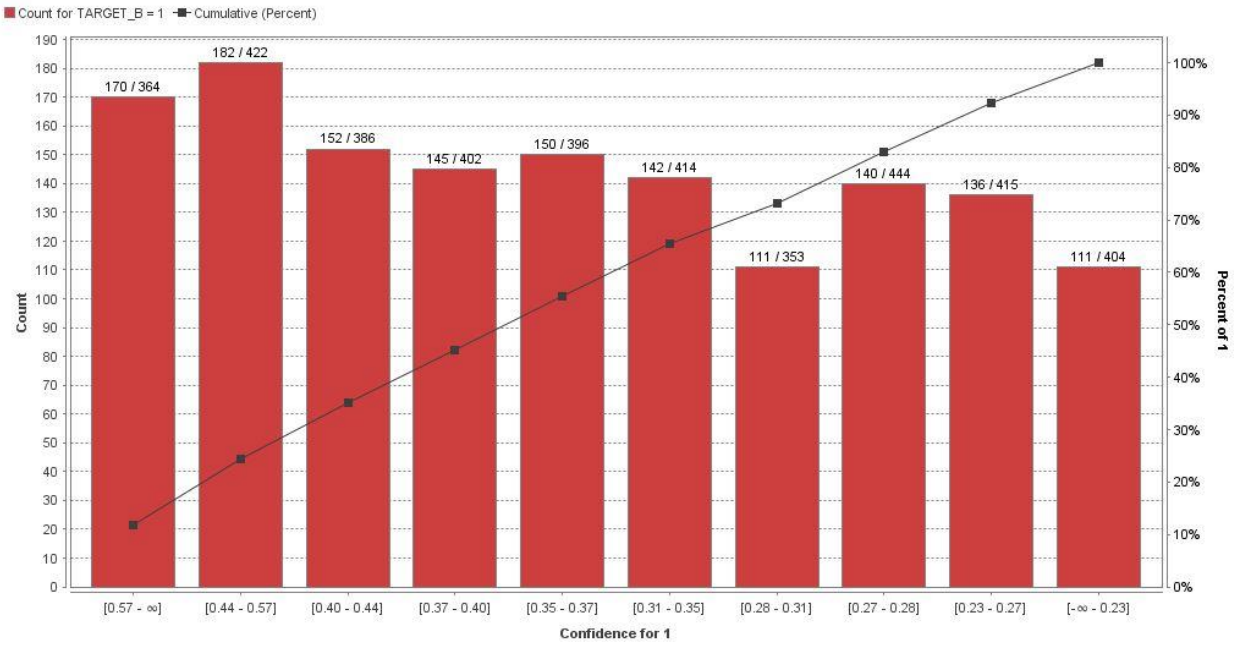
	true 0	true 1	class precision
pred. 0	3640	1565	69.93%
pred. 1	299	495	62.34%
class recall	92.41%	24.03%	



Validation Data

accuracy: 62.58%

	true 0	true 1	class precision
pred. 0	2256	1192	65.43%
pred. 1	305	247	44.75%
class recall	88.09%	17.16%	

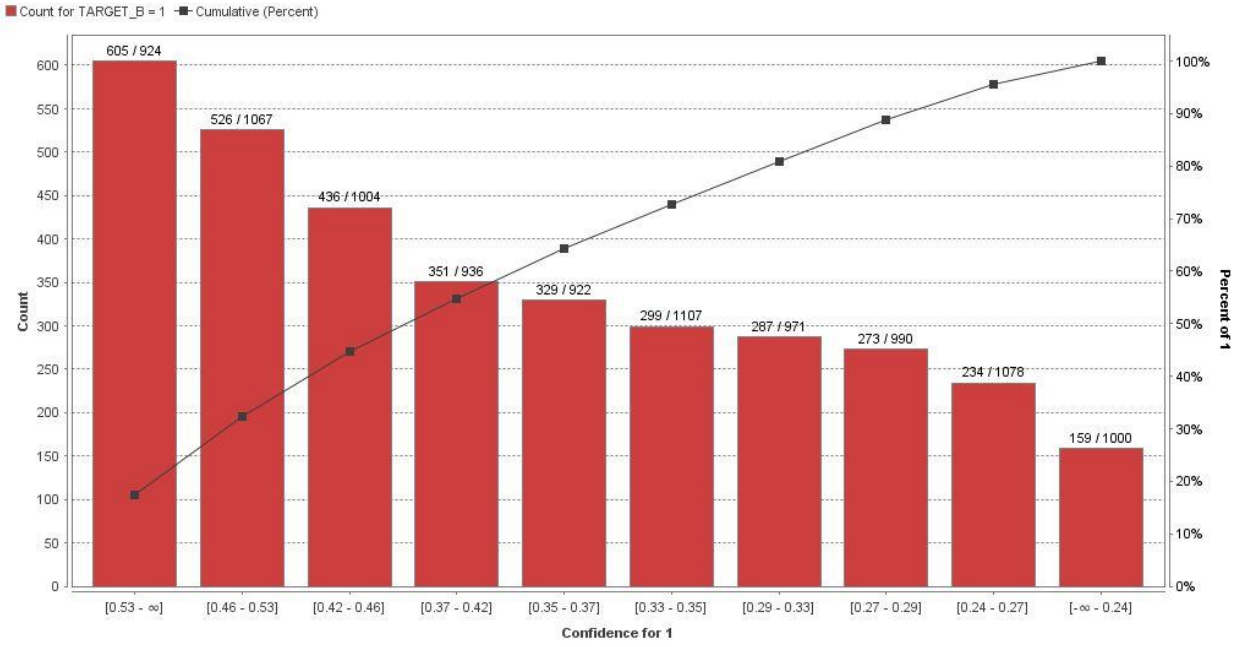


W/O Donor Interests PCA:

Training Data

accuracy: 68.93%

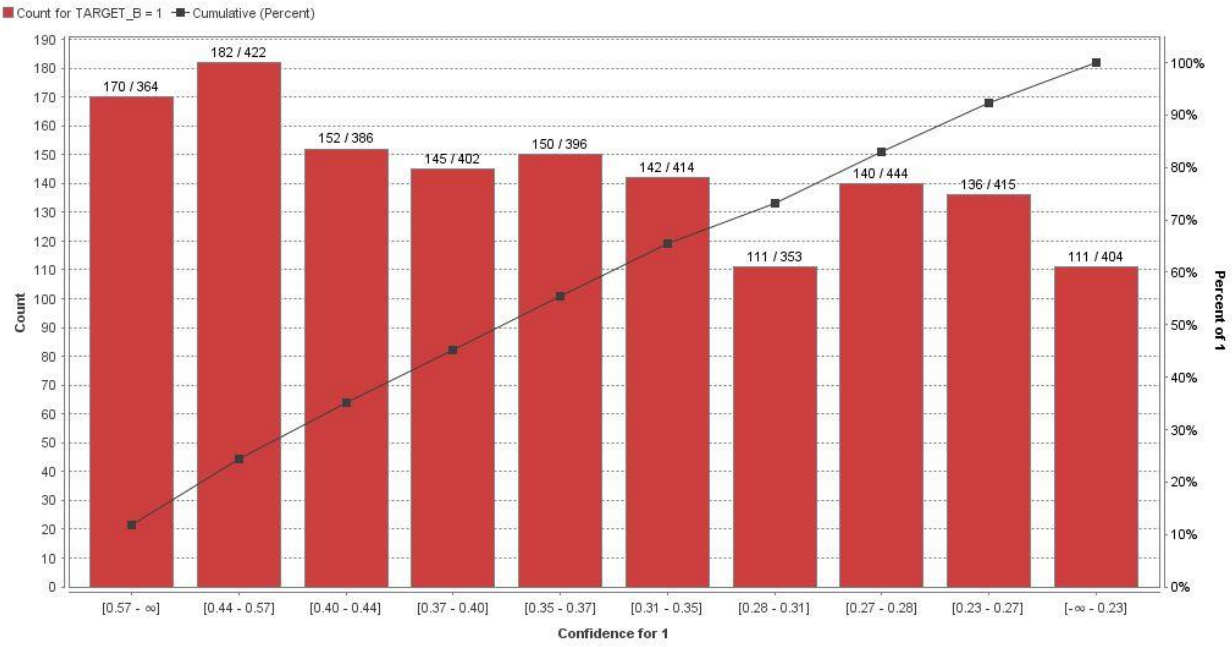
	true 0	true 1	class precision
pred. 0	3640	1565	69.93%
pred. 1	299	495	62.34%
class recall	92.41%	24.03%	



Validation Data

accuracy: 62.58%

	true 0	true 1	class precision
pred. 0	2256	1192	65.43%
pred. 1	305	247	44.75%
class recall	88.09%	17.16%	



Appendix D

Random Forest Models

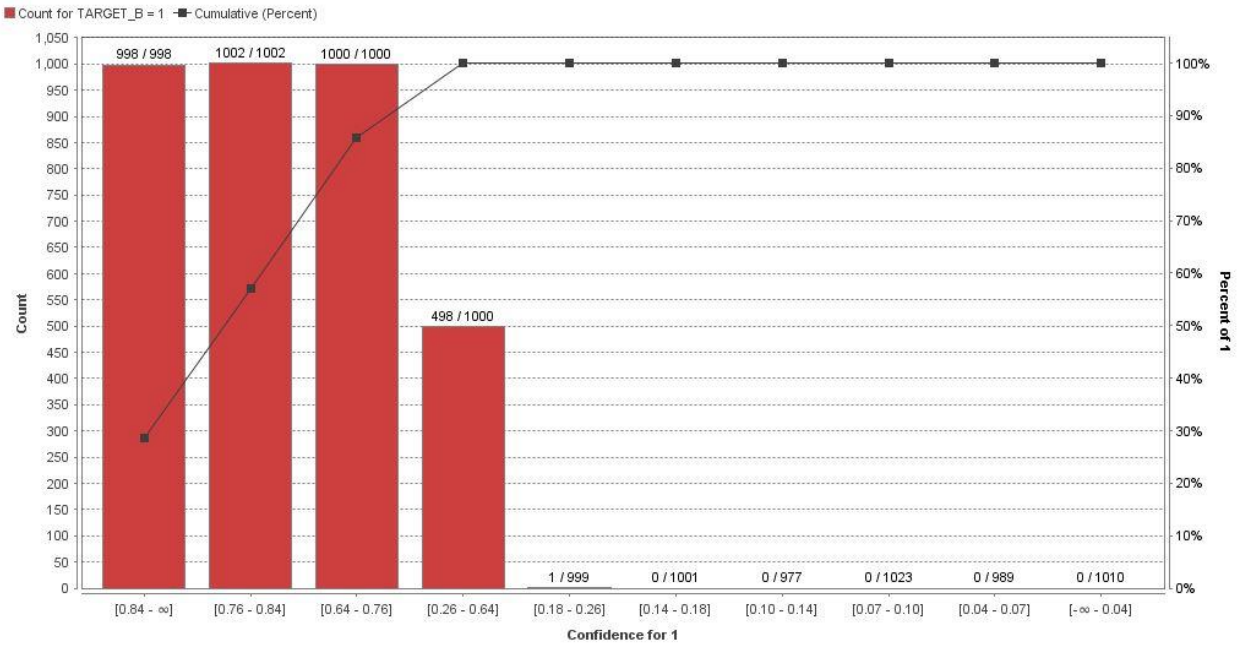
Parameters:

All PCAs:

Training Data

accuracy: 99.20%

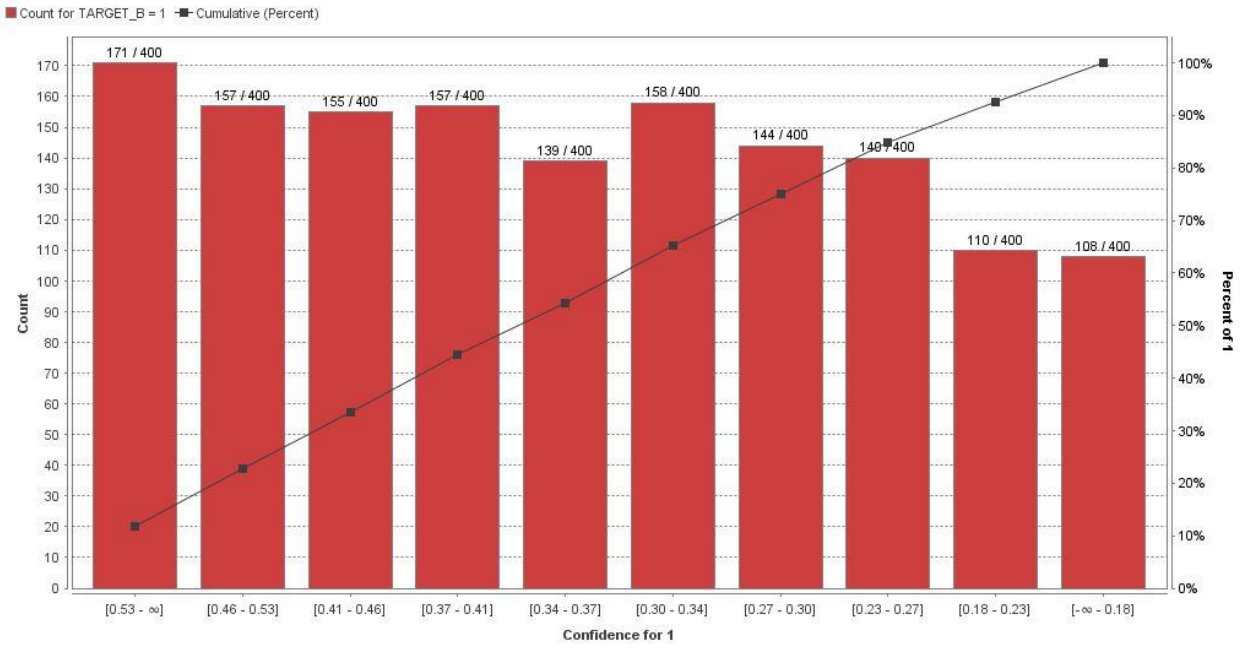
	true 0	true 1	class precision
pred. 0	3938	47	98.82%
pred. 1	1	2013	99.95%
class recall	99.97%	97.72%	



Validation Data

accuracy: 62.00%

	true 0	true 1	class precision
pred. 0	2251	1210	65.04%
pred. 1	310	229	42.49%
class recall	87.90%	15.91%	

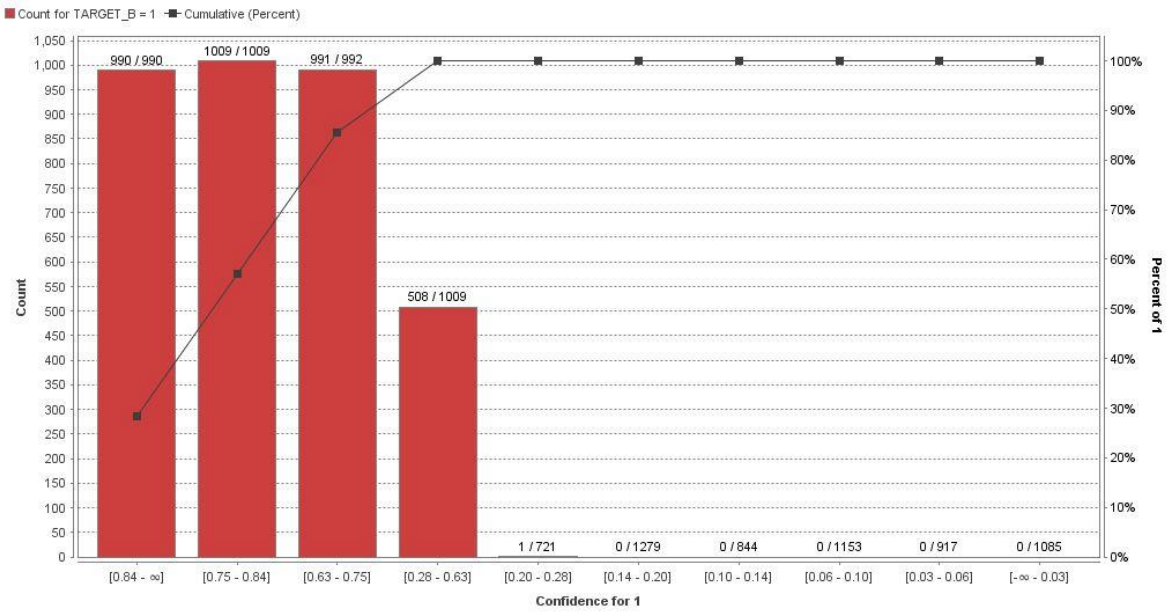


W/O Neighborhood PCA:

Training Data

accuracy: 99.03%

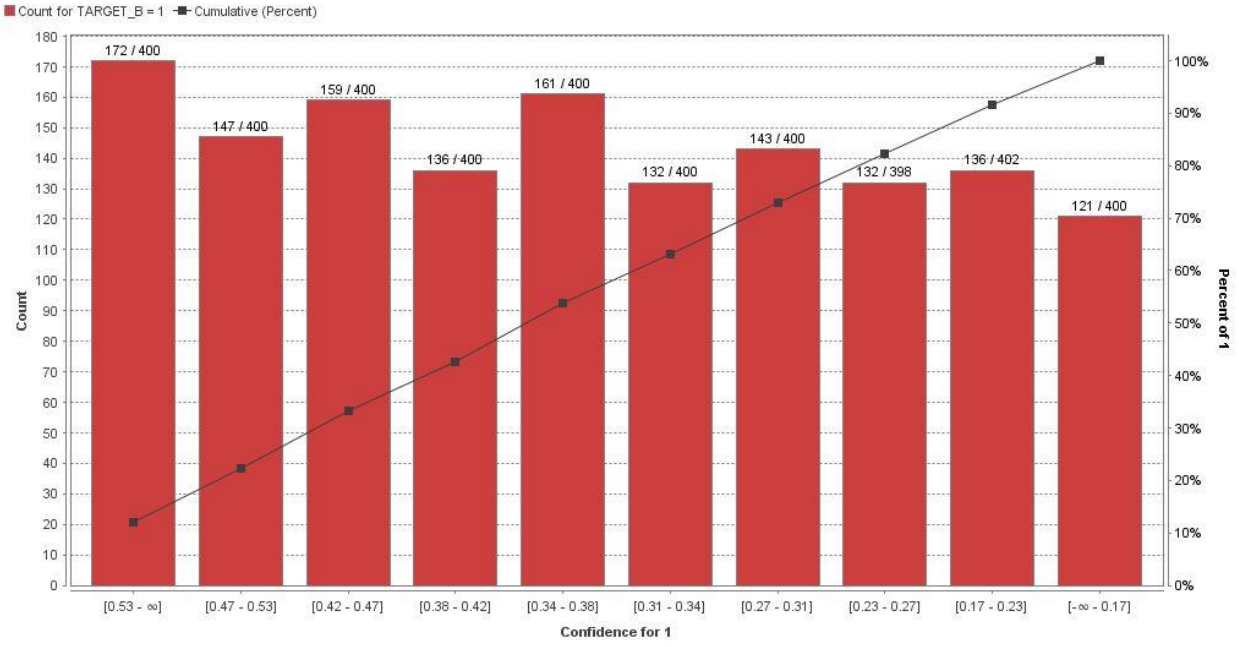
	true 0	true 1	class precision
pred. 0	3938	57	98.57%
pred. 1	1	2003	99.95%
class recall	99.97%	97.23%	



Validation Data

accuracy: 61.30%

	true 0	true 1	class precision
pred. 0	2214	1201	64.83%
pred. 1	347	238	40.68%
class recall	86.45%	16.54%	

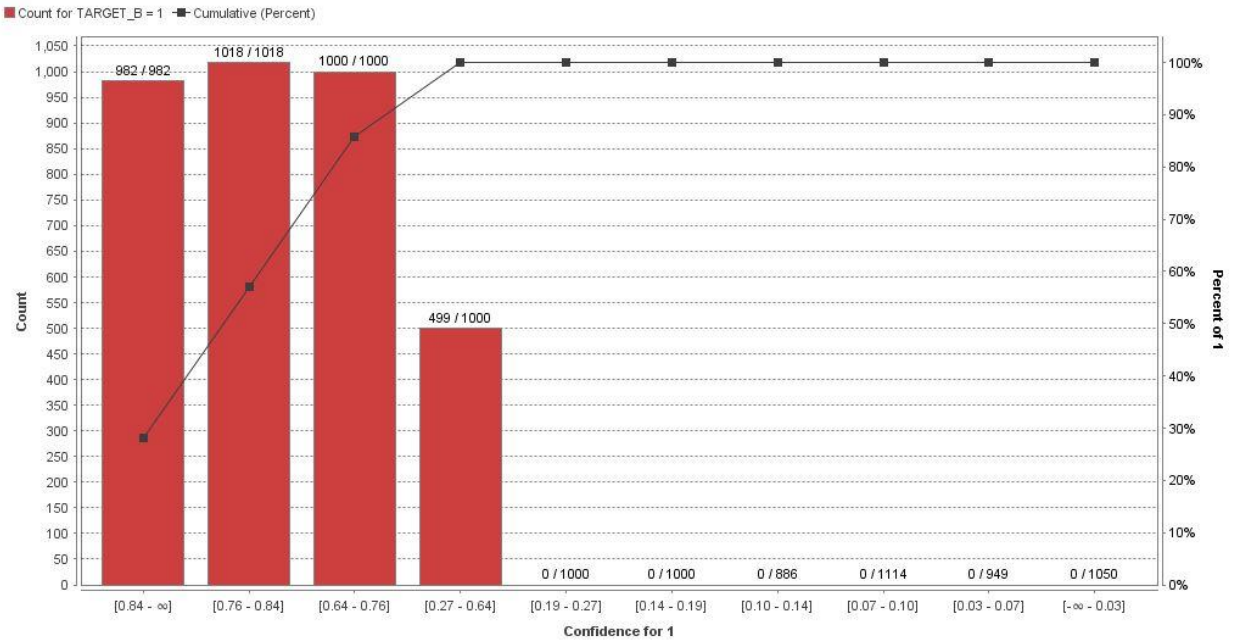


W/O Mail Responses PCA:

Training Data

accuracy: 99.25%

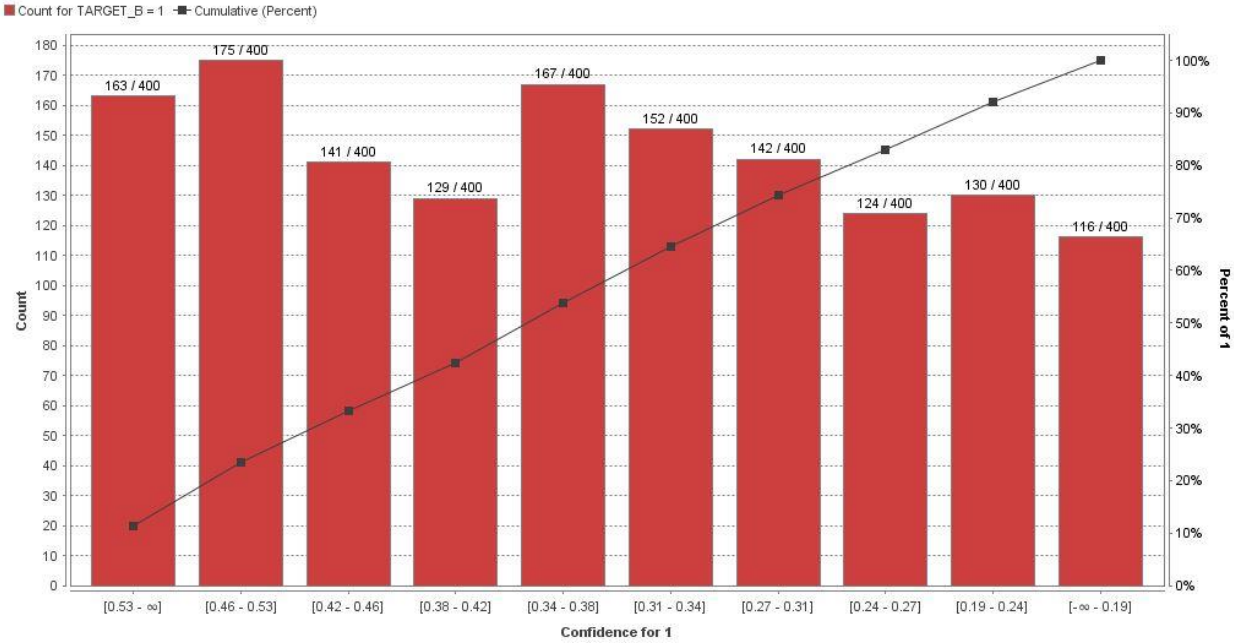
	true 0	true 1	class precision
pred. 0	3937	43	98.92%
pred. 1	2	2017	99.90%
class recall	99.95%	97.91%	



Validation Data

accuracy: 61.60%

	true 0	true 1	class precision
pred. 0	2234	1209	64.89%
pred. 1	327	230	41.29%
class recall	87.23%	15.98%	

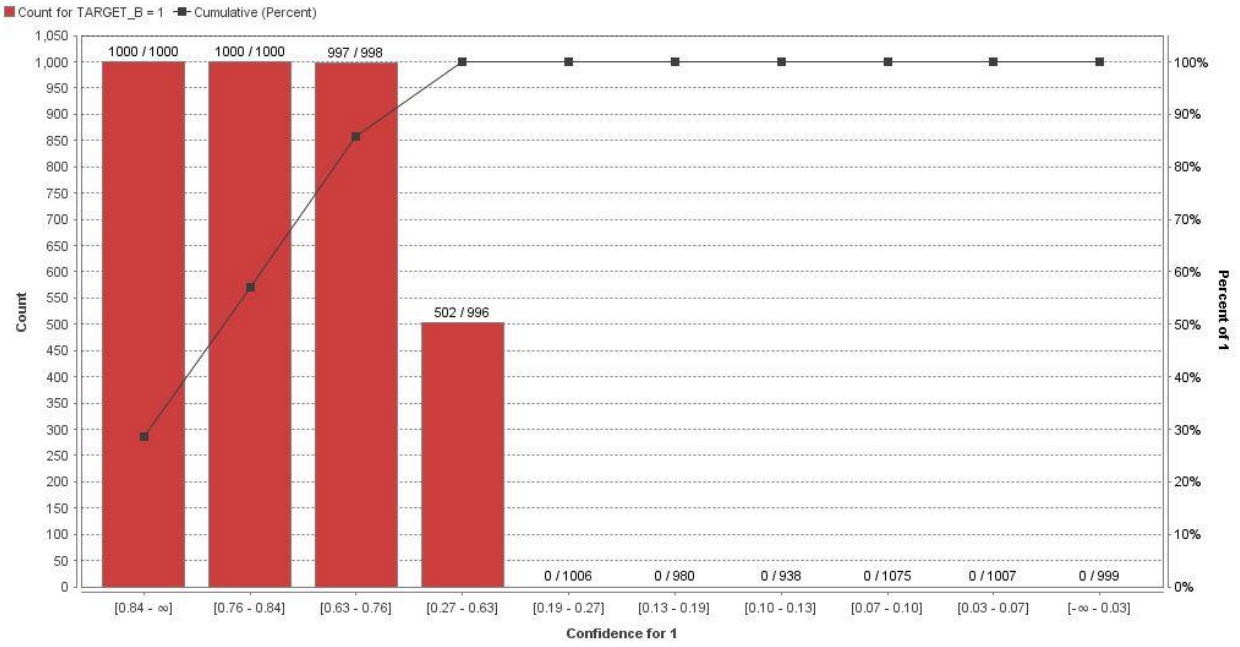


W/O Donor Interests PCA:

Training Data

accuracy: 99.22%

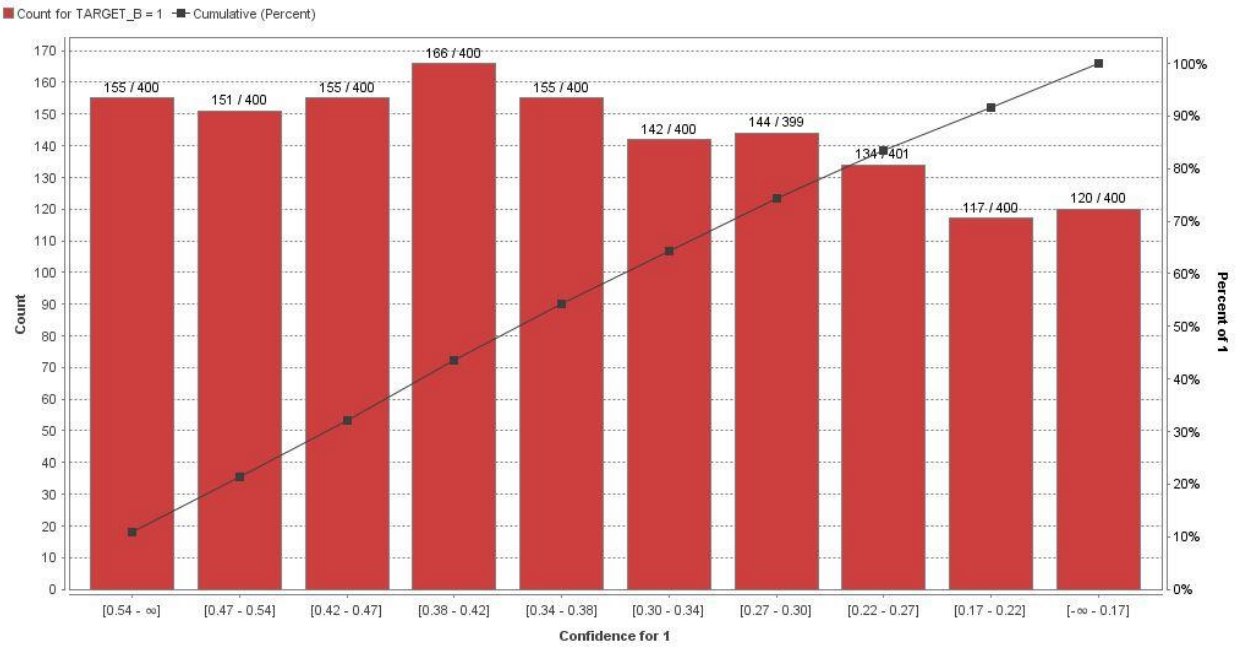
	true 0	true 1	class precision
pred. 0	3935	43	98.92%
pred. 1	4	2017	99.80%
class recall	99.90%	97.91%	



Validation Data

accuracy: 60.40%

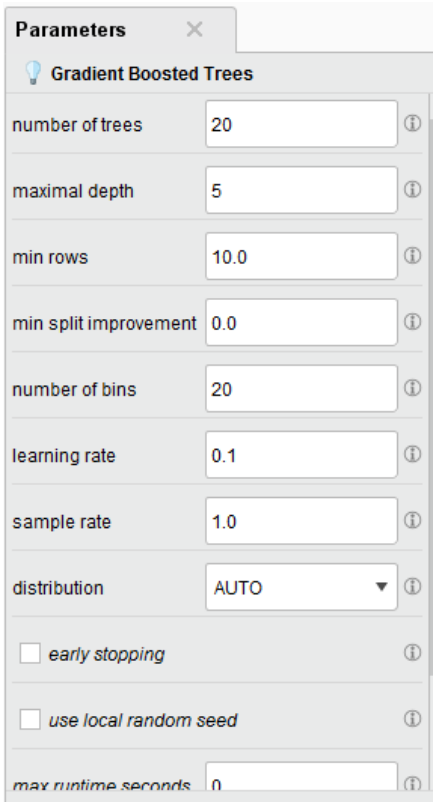
	true 0	true 1	class precision
pred. 0	2179	1202	64.45%
pred. 1	382	237	38.29%
class recall	85.08%	16.47%	



Appendix E

Boosted Tree Models

Parameters:

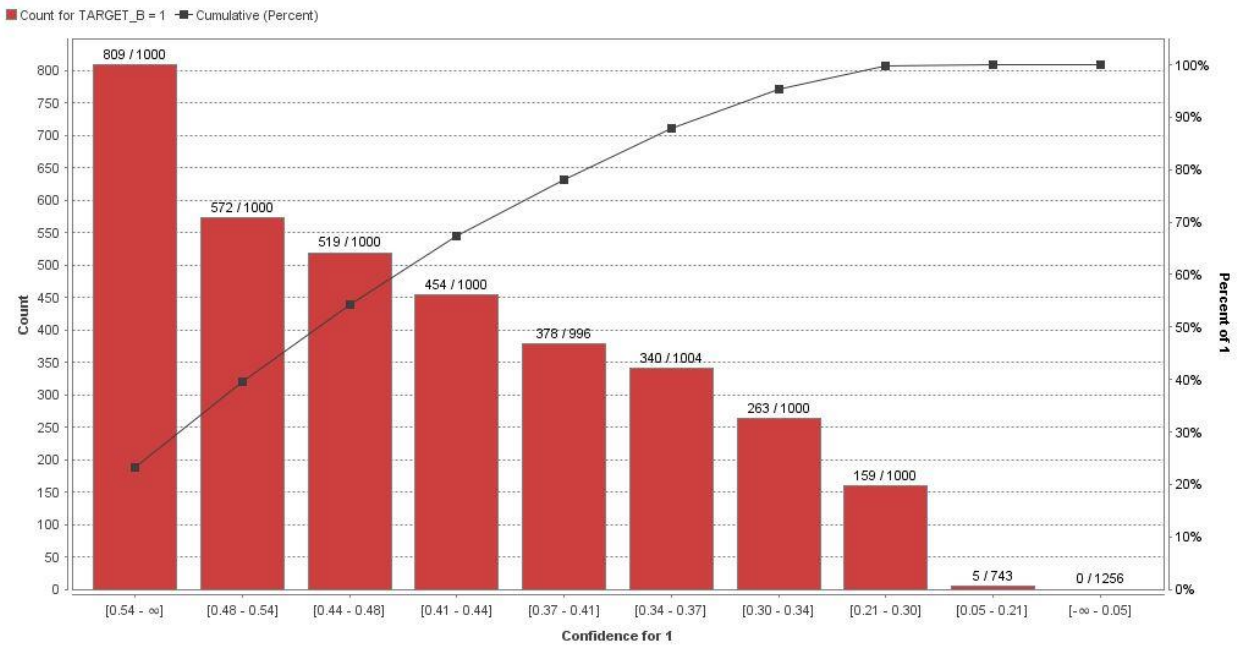


All PCAs:

Training Data

accuracy: 68.66%

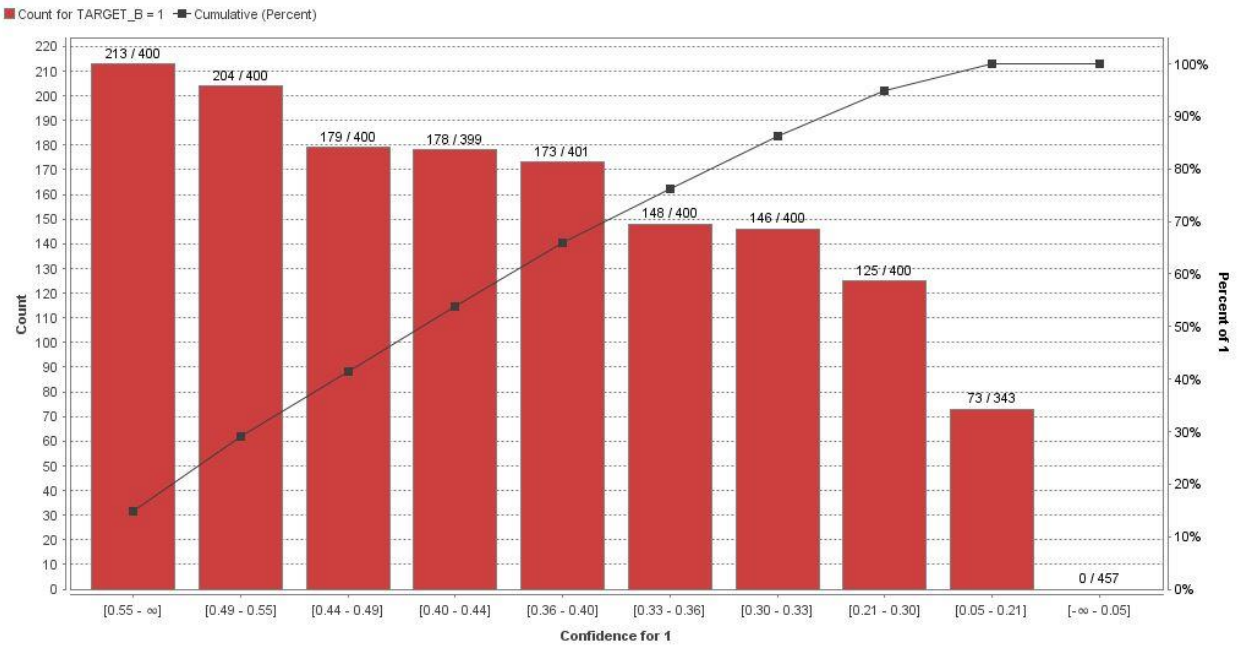
	true 0	true 1	class precision
pred. 0	2340	281	89.28%
pred. 1	1599	1779	52.66%
class recall	59.41%	86.36%	



Validation Data

accuracy: 60.30%

	true 0	true 1	class precision
pred. 0	1358	385	77.91%
pred. 1	1203	1054	46.70%
class recall	53.03%	73.25%	

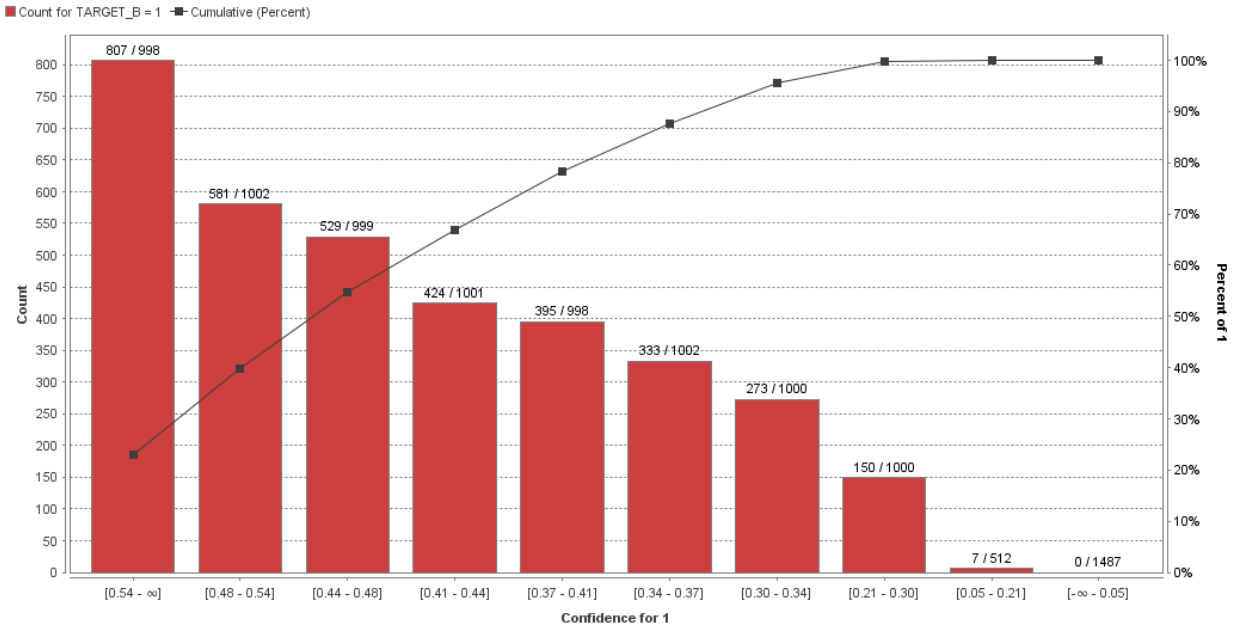


W/O Neighborhood PCA:

Training Data

accuracy: 68.94%

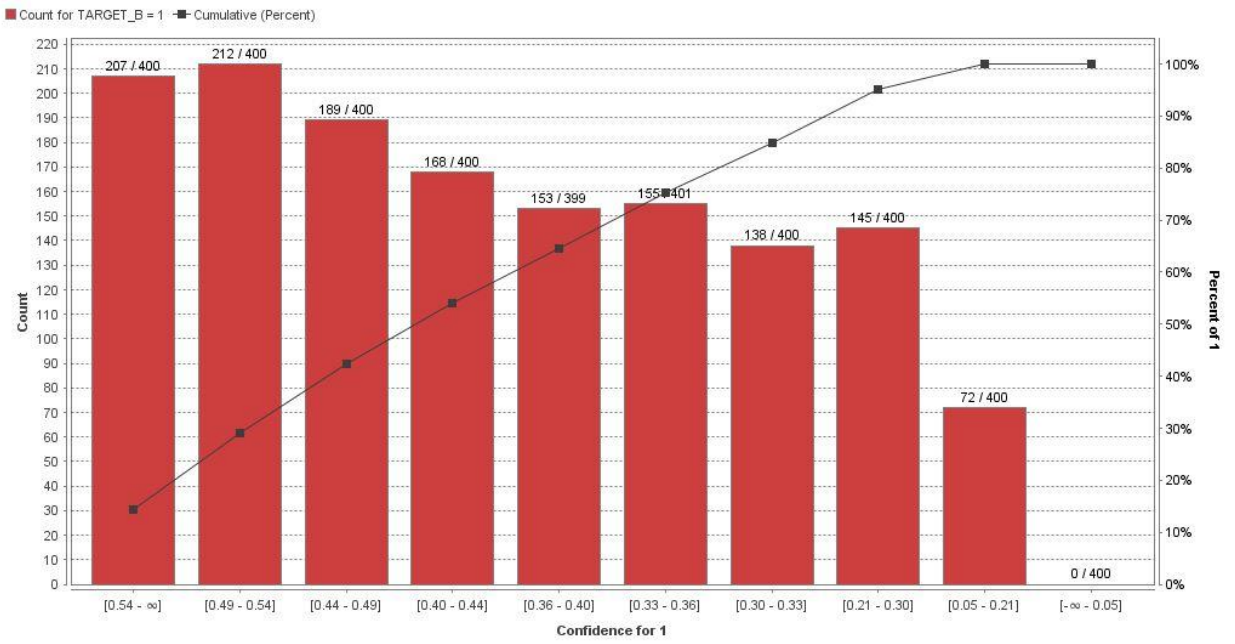
	true 0	true 1	class precision
pred. 0	2366	290	89.08%
pred. 1	1573	1770	52.95%
class recall	60.07%	85.92%	



Validation Data

accuracy: 59.15%

	true 0	true 1	class precision
pred. 0	1332	405	76.68%
pred. 1	1229	1034	45.69%
class recall	52.01%	71.86%	

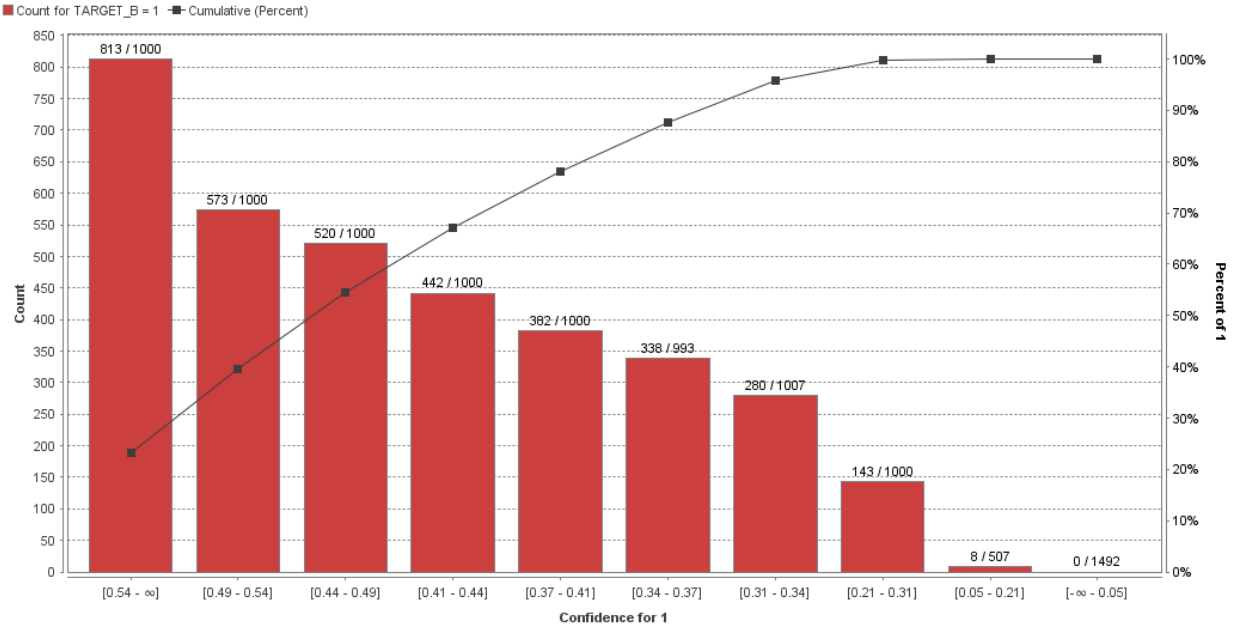


W/O Mail Responses PCA:

Training Data

accuracy: 69.76%

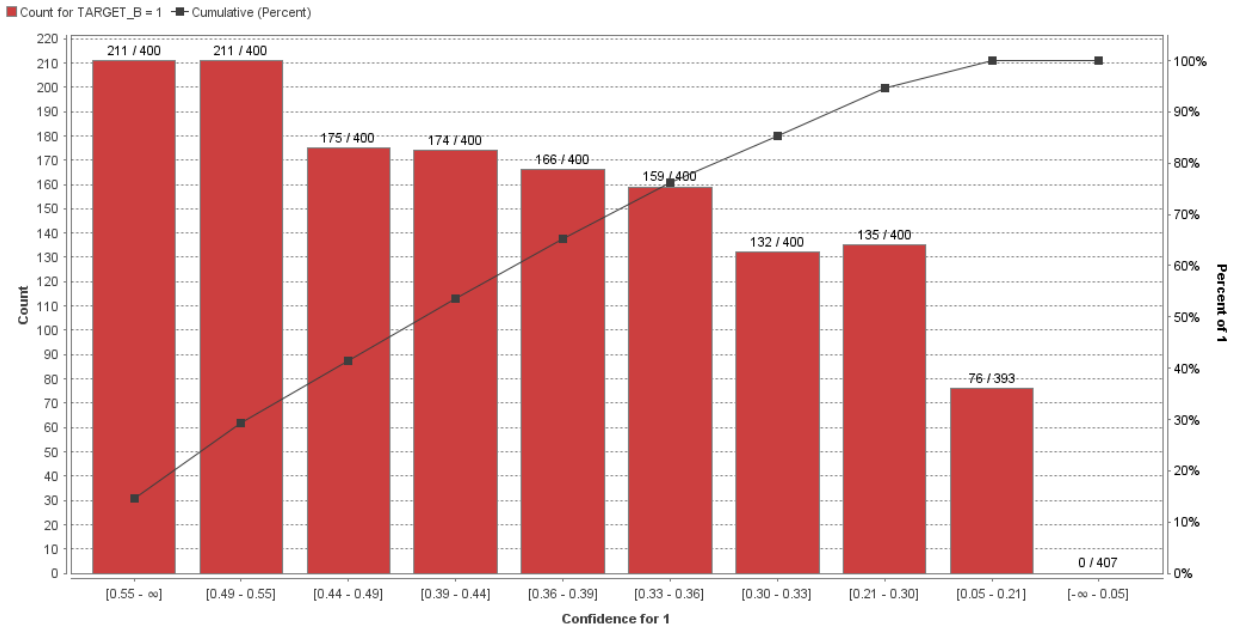
	true 0	true 1	class precision
pred. 0	2496	371	87.06%
pred. 1	1443	1689	53.93%
class recall	63.37%	81.99%	



Validation Data

accuracy: 60.72%

	true 0	true 1	class precision
pred. 0	1460	470	75.65%
pred. 1	1101	969	46.81%
class recall	57.01%	67.34%	

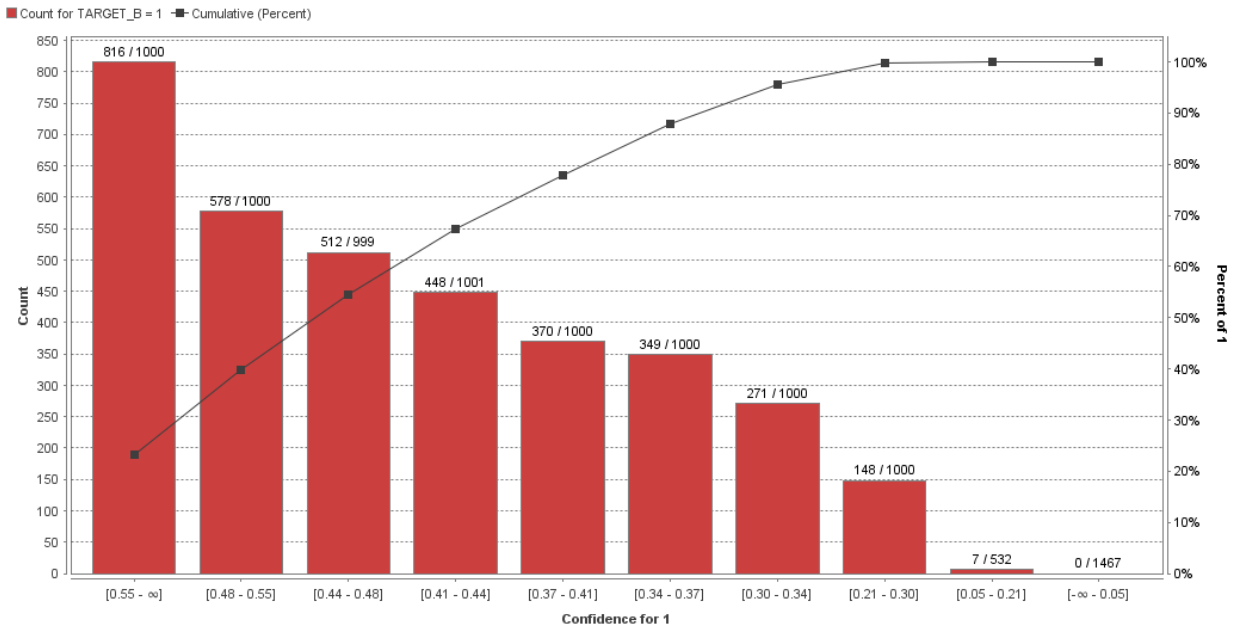


W/O Donor Interests PCA:

Training Data

accuracy: 68.79%

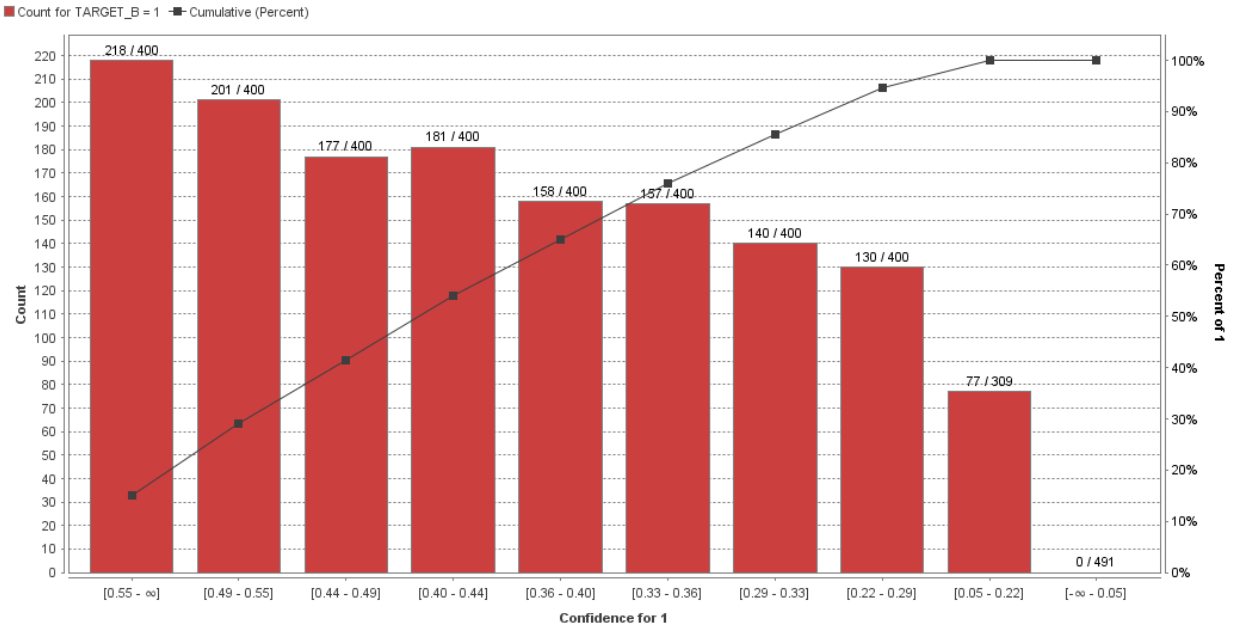
	true 0	true 1	class precision
pred. 0	2341	274	89.52%
pred. 1	1598	1786	52.78%
class recall	59.43%	86.70%	



Validation Data

accuracy: 59.40%

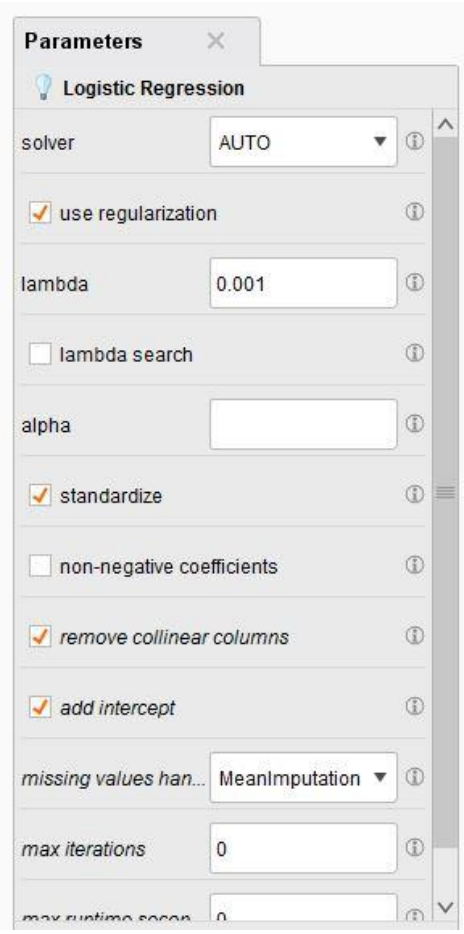
	true 0	true 1	class precision
pred. 0	1341	404	76.85%
pred. 1	1220	1035	45.90%
class recall	52.36%	71.92%	



Appendix F

Logistic Regression Models

Parameters:

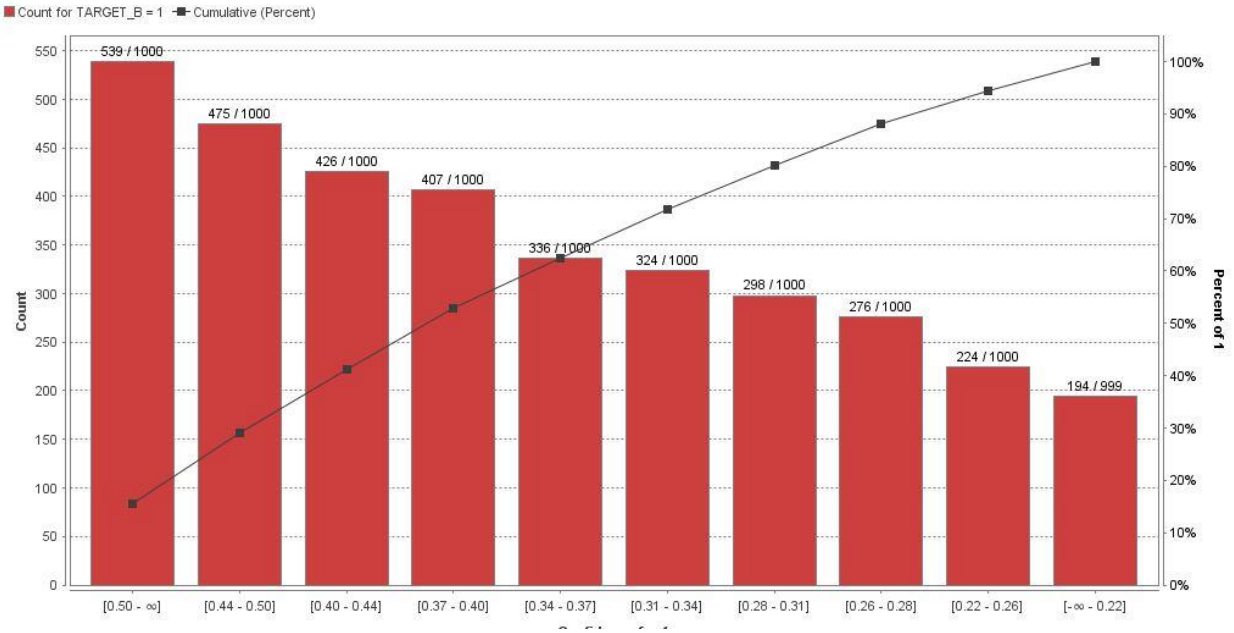


All PCAs:

Training Data

accuracy: 52.43%

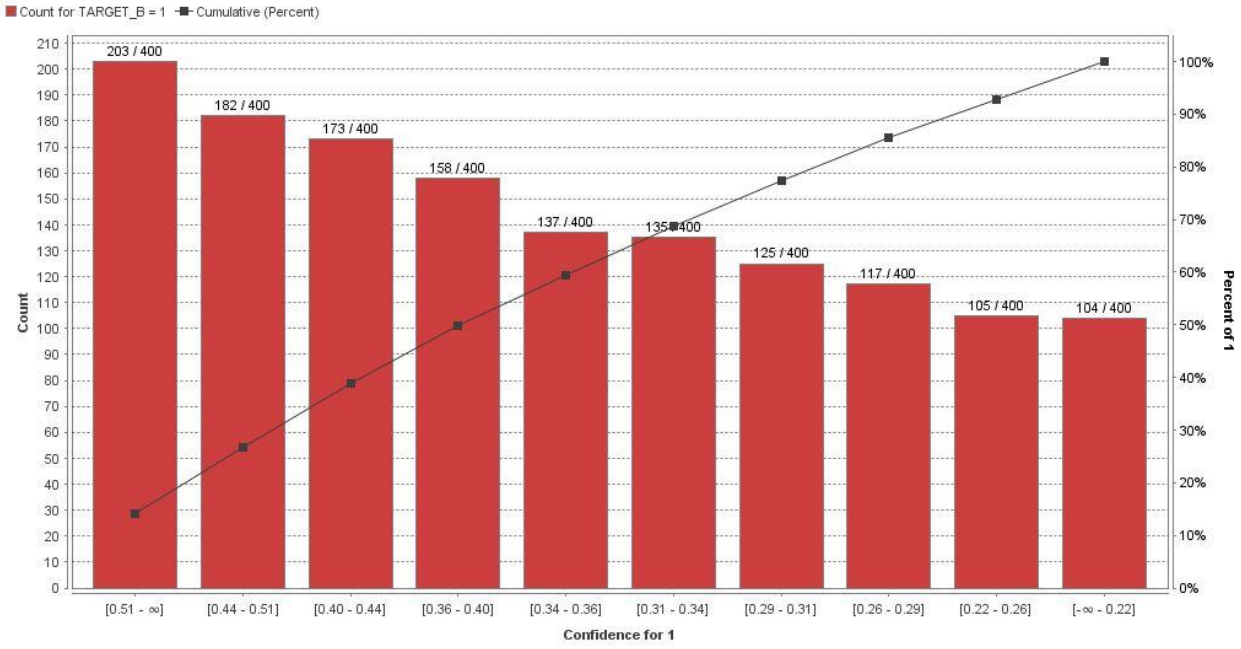
	true 0	true 1	class precision
pred. 0	1509	424	78.07%
pred. 1	2430	1636	40.24%
class recall	38.31%	79.42%	



Validation Data

accuracy: 50.12%

	true 0	true 1	class precision
pred. 0	905	339	72.75%
pred. 1	1656	1100	39.91%
class recall	35.34%	76.44%	

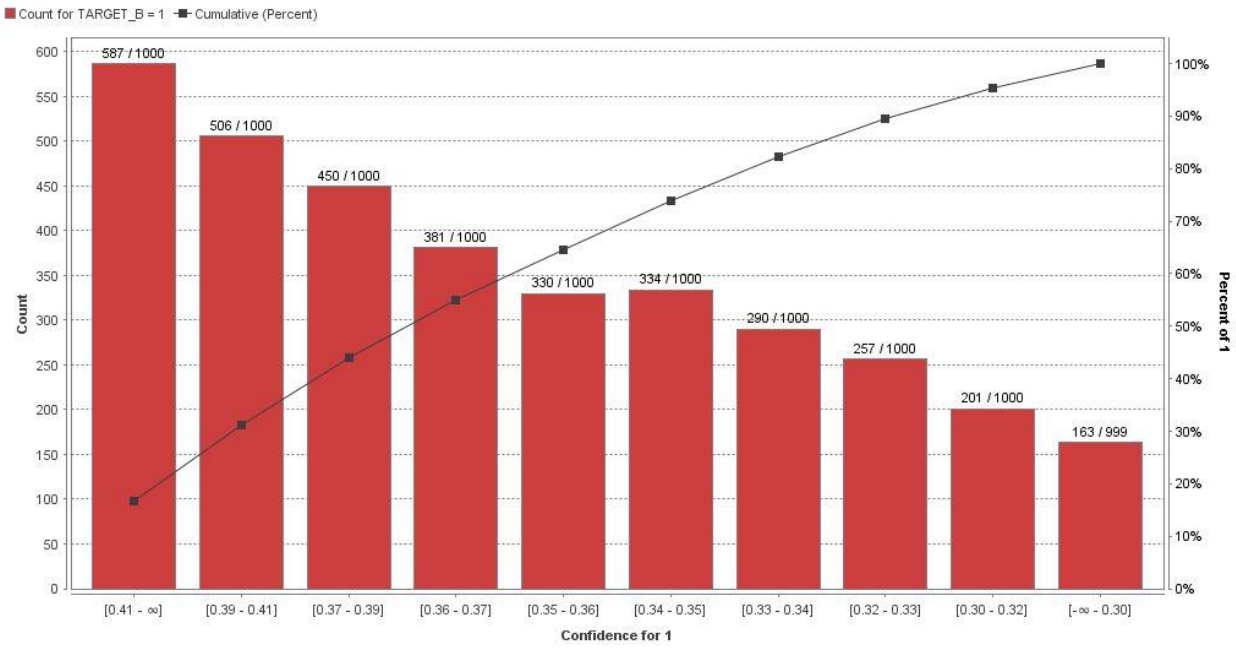


W/O Neighborhood PCA:

Training Data

accuracy: 60.03%

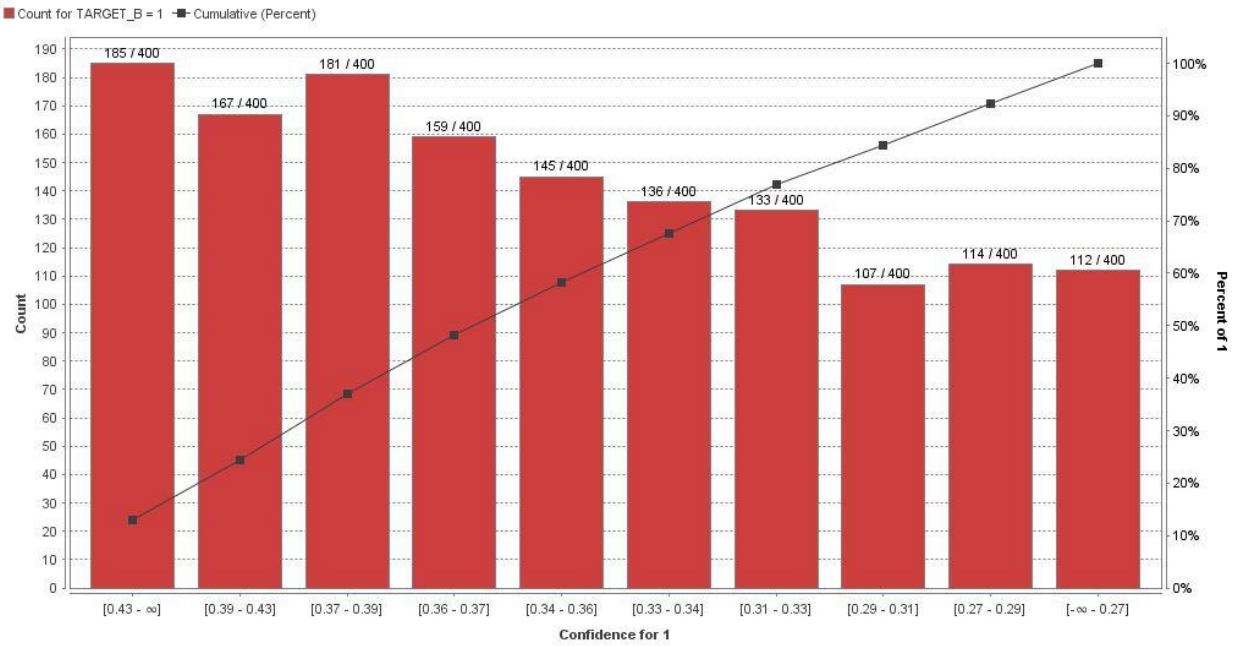
	true 0	true 1	class precision
pred. 0	2077	536	79.49%
pred. 1	1862	1524	45.01%
class recall	52.73%	73.98%	



Validation Data

accuracy: 53.57%

	true 0	true 1	class precision
pred. 0	1218	514	70.32%
pred. 1	1343	925	40.78%
class recall	47.56%	64.28%	

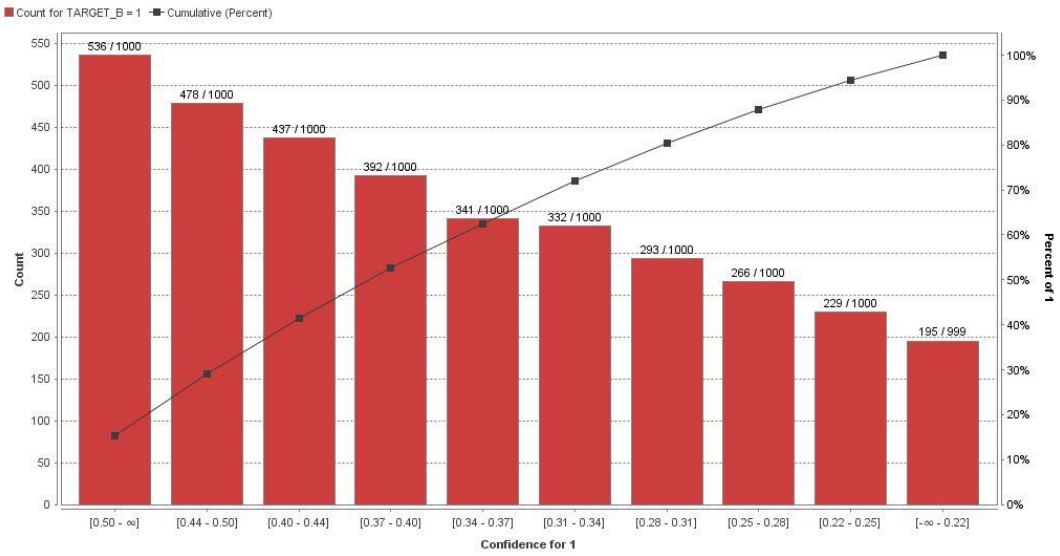


W/O Mail Responses PCA:

Training Data

accuracy: 51.34%

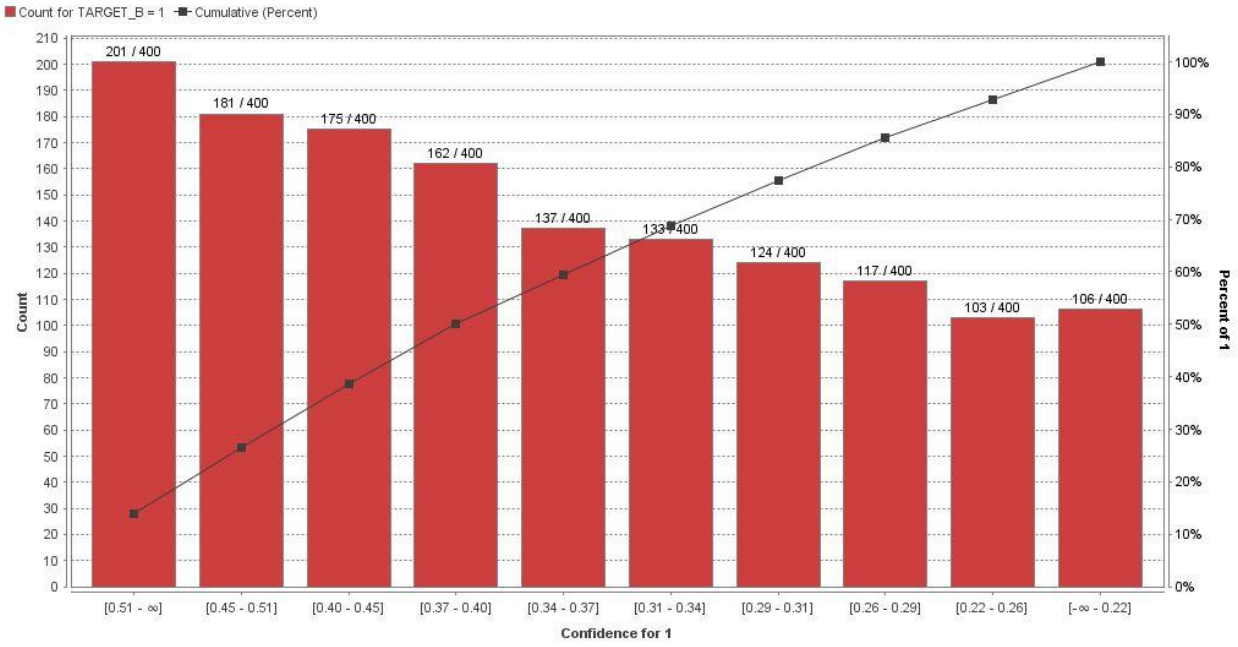
	true 0	true 1	class precision
pred. 0	1396	376	78.78%
pred. 1	2543	1684	39.84%
class recall	35.44%	81.75%	



Validation Data

accuracy: 48.95%

	true 0	true 1	class precision
pred. 0	821	302	73.11%
pred. 1	1740	1137	39.52%
class recall	32.06%	79.01%	

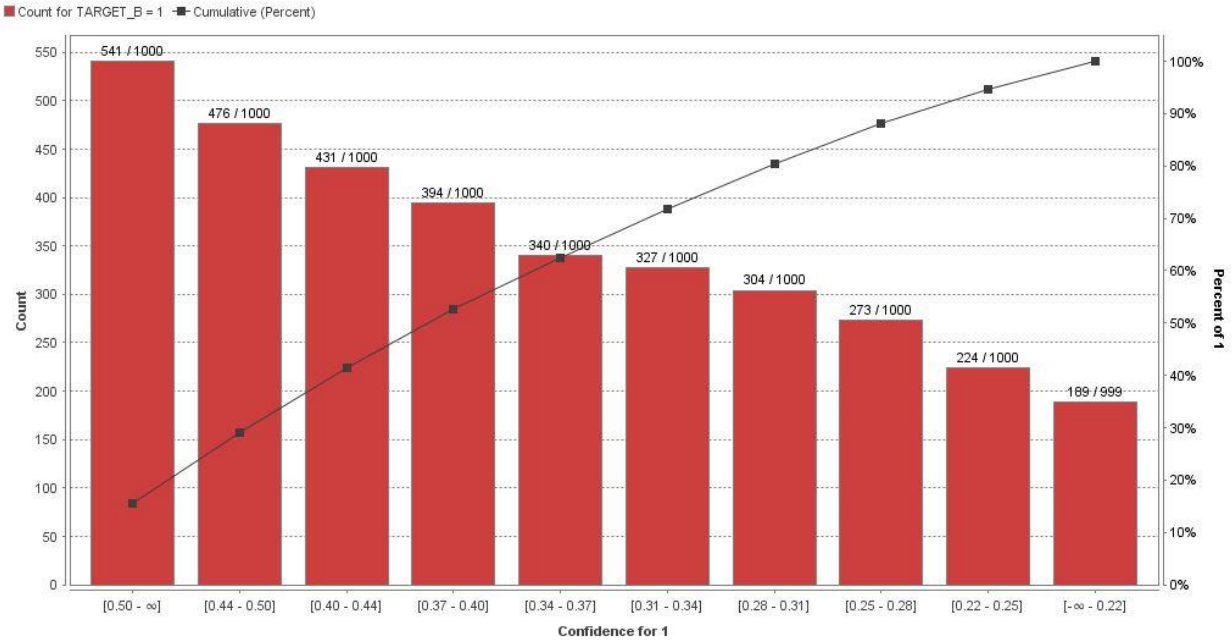


W/O Donor Interests PCA:

Training Data

accuracy: 52.11%

	true 0	true 1	class precision
pred. 0	1472	406	78.38%
pred. 1	2467	1654	40.14%
class recall	37.37%	80.29%	



Validation Data

accuracy: 49.43%

	true 0	true 1	class precision
pred. 0	869	331	72.42%
pred. 1	1692	1108	39.57%
class recall	33.93%	77.00%