IDS 572 Assignment 3

Group Details:

Name	UIN
NagaShrikanth Ammanabrolu	676837954
Suresh Sappa	667192596
Sagar Kanchi	669850639

Table of Contents

Introduction

In the second assignment, the performance of a model is measured on the overall training and testing accuracy. Now, when we consider the costs and benefits as the measure of performance of the model, we initially lay more emphasis on attaining high **true recall accuracy** in the process of selecting a model that **maximizes net profit**.

Question No. 1)

Solution.

Modeling on full dataset and chosen subsets of variables

As mentioned earlier, our main focus is on maximizing net profit, we will in this section concentrate on how our data performs on the below mentioned techniques (considering full dataset and chosen subset of variables). After running on the full set of variables, and individual subsets, we identified prime categories (**'DONOR Information', 'Mail order offers'**, **'Neighborhood' and 'RFA'**) by carrying out data cleaning and data reduction.

1. **Decision Trees (W-J48)**
2. **Logistic Regression**
3. **Naïve Bayes**
4. **K-Nearest Neighbors**
5. **Random Forests**
6. **Support Vector Machines**

1.1 Decision Trees (W-J48)

The W-J48 model which is a part of the WEKA extension creates a decision tree using the C4.5 algorithms. We also observed that pruning the tree doesn't have significant impact on the model we are testing here.

Using W-J48, we experimented on unpruned tree and reduced error pruning, varying the confidence threshold and minimum number of instances per leaf.

Below listed are the details of the Parameters (optimized) on the W-J48 algorithm:

Decision Tree Criterion Parameters	Associated Label or Value
U	Unchecked
C	0.5
M	10.0
R	Unchecked
N	Blank

B	Unchecked
S	Unchecked
L	Unchecked
A	Checked
Q	12345

The PCA analysis is performed on the different combinations of subsets as selected in the previous assignment. The results of the performance metrics on the different combinations of subsets is summarized below.

Scenario	**Accuracy (%) for the Training Dataset**	**Accuracy (%) for the Test Dataset**	**Testing Recall Accuracy (%)**
Full dataset considered	83.76%	56.90%	32.73%
All PCA Values considered	82.43%	57.85%	37.11%
PCA values of 'DONOR Information' subset removed	78.13%	59.13%	27.87%
PCA values of 'Mail order offers' subset removed	82.26%	58.98%	36.00%
PCA values of 'Neighborhood' subset removed	84.61%	56.67%	32.18%
PCA values of 'RFA' subset removed	81.96%	58.40%	35.79%

Summary: The above table summarizes the different combinations of subsets that we have considered for PCA Analysis and their respective performance metrics. As shown in the highlighted scenario above, we can see that considering all the PCA values yields a better True recall accuracy (True 1's). Although this model yields us a poor accuracy on the Testing dataset, we are more concerned with the accuracy of the True recalls. Hence, the highlighted scenario yields us a better model as compared to the other scenarios using the W-J48 model.

1.2 W-Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

We experimented with checking and unchecking the turn on debugging output along with various values of maximum number of iterations.

Below listed are the details of the Parameters (optimized) on the W-Logistic Regression algorithm:

W-Logistic Regression Parameters	Associated Label or Value
D	Checked
R	1.0E-10
M	-1.0

The PCA analysis is performed on the different combinations of subsets as selected in the previous assignment. The results of the performance metrics on the different combinations of subsets is summarized below.

Scenario	Accuracy (%) for the Training Dataset	Accuracy (%) for the Test Dataset	Testing Recall Accuracy (%)
Full dataset considered	62.35%	62.32%	20.34%
All PCA Values considered	64.10%	67.68%	21.26%
PCA values of 'DONOR Information' subset removed	67.68%	63.50%	17.37%
PCA values of 'Mail order offers' subset removed	67.66%	63.98%	18.76%
PCA values of 'Neighborhood' subset removed	67.68%	64.00%	18.21%
PCA values of 'RFA' subset removed	67.74%	64.00%	18.42%

Summary: The above table summarizes the different combinations of subsets that we have considered for PCA Analysis and their respective performance metrics. As shown in the highlighted scenario above, we can see that considering all the PCA values yields a better True recall accuracy (True 1's). In this situation, considering all the PCA values for the analysis not only yields us the best testing recall accuracy, it also yields us a better overall test accuracy as compared to the other combination of subsets for PCA analysis and full dataset consideration

Hence, the highlighted scenario yields us a better model as compared to the other scenarios using the W-Logistic Regression model.

1.3 Naïve Bayes

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. [2]

We experimented by checking and unchecking the Laplace correction.

Below listed are the details of the Parameters (optimized)on the Naïve Bayes algorithm:

Naïve Bayes Parameters	Associated Label or Value
Laplace Correction	Checked

The PCA analysis is performed on the different combinations of subsets as selected in the previous assignment. The results of the performance metrics on the different combinations of subsets is summarized below.

Scenario	Accuracy (%) for the Training Dataset	Accuracy (%) for the Test Dataset	Testing Recall Accuracy (%)
Full dataset considered	54.21%	51.05%	63.86%
All PCA Values considered	60.99%	59.23%	50.87%
PCA values of 'DONOR Information' subset removed	61.41%	58.98%	49.2%
PCA values of 'Mail order offers' subset removed	60.81%	59.15%	51.29%
PCA values of 'Neighborhood' subset removed	54.29%	51.15%	64.35%
PCA values of 'RFA' subset removed	61.31%	58.75%	51.42%

Summary: The above table summarizes the different combinations of subsets that we have considered for PCA Analysis and their respective performance metrics. These combination of subsets yield us the minimum number of variables, which helps in reducing the complexity of the model while not compromising on the performance of the model. As shown in the highlighted scenario above, we can see that removing the PCA values of the 'Neighborhood' subset yields a better True recall accuracy (True 1's). Although this model yields us a poor accuracy on the Testing dataset, we are more concerned with the accuracy of the True

recalls. Since the true recalls indicate the actual donors who will maximize our profits, we are more interested in categorizing them as adequately as possible.

1.4 k-Nearest Neighbors

The optimization is carried out by experimenting the model with various values of 'k' and considering all the different measure types (MixedMeasures, NominalMeasures, etc.)

Below listed are the details of the Parameters(optimized) on the KNN algorithm:

KNN Parameters	Associated Label or Value
K	3
Weighted Vote	Unchecked
Measure types	Mixed Measures
Mixed measure	Mixed Euclidean Distance

The PCA analysis is performed on the different combinations of subsets as selected in the previous assignment. The results of the performance metrics on the different combinations of subsets is summarized below.

Scenario	Accuracy (%) for the Training Dataset	Accuracy (%) for the Test Dataset	Testing Recall Accuracy (%)
Full dataset considered	77.90%	56.50%	25.43%
All PCA Values considered	78.55%	57.83%	32.11%
PCA values of 'DONOR Information' subset removed	78.10%	58.10%	31.76%
PCA values of 'Mail order offers' subset removed	78.56%	57.10%	30.99%
PCA values of 'Neighborhood' subset removed	77.86%	56.12%	25.43%
PCA values of 'RFA' subset removed	78.48%	57.83%	32.24%

Summary: The above table summarizes the different combinations of subsets that we have considered for PCA Analysis and their respective performance metrics. These combination of subsets yield us the minimum number of variables, which helps in reducing the complexity of the model while not compromising on the performance of the model. As shown in the highlighted scenario above, we can see that removing the PCA values of the 'RFA subset'

yields a better True recall accuracy (True 1's). Although this model yields us a poor accuracy on the Testing dataset, we are more concerned with the accuracy of the True recall.

1.5 Random Forest

We experimented on various criterion (Gini index, information gain, gain ratio and accuracy) with various values of number of trees, maximal depth, subset ratios and checking and unchecking pruning and pre-pruning.

Below listed are the details of the Parameters(optimized) on the Random Forest algorithm:

Random Forest Parameters	Associated Label or Value
Criterion	Gini Index
Maximal Depth	50
Apply pruning	Checked
Confidence	0.1
Apply Pre-pruning	Unchecked
Guess subset ratio	Unchecked
Subset ratio	1.0
Voting strategy	Confidence vote
Use Local random seed	12345 (Checked)

The PCA analysis is performed on the different combinations of subsets as selected in the previous assignment. The results of the performance metrics on the different combinations of subsets is summarized below.

Scenario	Accuracy (%) for the Training Dataset	Accuracy (%) for the Test Dataset	Testing Recall Accuracy (%)
Full dataset considered	92.52%	56.73%	33.7%
PCA values of 'DONOR Information' subset removed	92.58%	57.57%	34.26%
PCA values of 'Mail order offers' subset removed	92.6%	59.03%	35.02%
PCA values of 'Neighborhood' subset removed	93.12%	57.53%	31.27%
All PCA Values considered	92.3%	59.6%	34.47%
PCA values of 'RFA' subset removed	92.3%	59.6%	34.47%

Summary: The above table summarizes the full dataset and different combinations of subsets that we have considered for PCA Analysis and their respective performance metrics. These combination of subsets yield us the minimum number of variables, which helps in reducing the complexity of the model while not compromising on the performance of the model. As shown in the highlighted scenario above, we can see that removing the PCA values of the 'Mail Order Offers' subset yields a better True recall accuracy (True 1's).

1.6 Support Vector Machine

Different types of kernel functions are used with various values of kernel gamma, kernel cache. We also checked for optimization by varying complexity constant 'C' and the Loss Function parameters.

Below listed are the details of the Parameters (optimized) on the Support Vector Machine algorithm:

Support Vector Machine Parameters	Associated Label or Value
Kernel Type	Radial
Kernel Gamma	0.01
Kernel Cache	200
C	1.0
Convergence Epsilon	0.001
Max iterations	100000
Scale	Checked
L pos.	1.0
L neg.	1.0
Epsilon	0
Epsilon plus	0
Epsilon minus	0
Balance cost	Checked
Quadratic loss pos.	Unchecked
Quadratic loss neg.	Unchecked

The PCA analysis is performed on the different combinations of subsets as selected in the previous assignment. The results of the performance metrics on the different combinations of subsets is summarized below.

Scenario	Accuracy (%) for the Training Dataset	Accuracy (%) for the Test Dataset	Testing Recall Accuracy (%)
Full dataset considered	98.63%	62.0%	20.5%
All PCA Values considered	85.03%	60.40%	37.11%

PCA values of 'DONOR Information' subset removed	84.56%	58.8%	44.48%
PCA values of 'Mail order offers' subset removed	84.25%	58.98%	45.17%
PCA values of 'Neighborhood' subset removed	97.45%	61.22%	22.24%
PCA values of 'RFA' subset removed	82.6%	58.9%	46.7%
All PCA Values Removed	98.63%	62.0%	20.5%

Summary: The above table summarizes the different combinations of subsets that we have considered for PCA Analysis and their respective performance metrics. These combination of subsets yield us the minimum number of variables, which helps in reducing the complexity of the model while not compromising on the performance of the model. As shown in the highlighted scenario above, we can see that removing the PCA values of the 'RFA' subset yields a better True recall accuracy (True 1's). Although this model yields us a poor accuracy on the Testing dataset, we are more concerned with the accuracy of the True recalls.

Hence, the highlighted scenario yields us a better model as compared to the other scenarios using the Random Forest algorithm.

Evaluative comparison of the different Best models

The best models from each of the six classification models are being summarized below for an overall picture of how these models perform against each other.

Classification Models	Accuracy (%) for the Training Dataset	Accuracy (%) for the Test Dataset	Testing Recall Accuracy (%)
Naïve Bayes	54.29%	51.15%	64.35%
W-Logistic Regression	64.10%	67.68%	21.26%
W-J48	82.43%	57.85%	37.11%
K-nearest neighbors	78.48%	57.83%	32.24%
Random Forest	92.6%	59.03%	35.02%
Support Vector Machine	82.6%	58.9%	46.7%

From the above table, we could conclude that Naïve Bayes is the classification model which yields us a better model as compared to the other models. Again, since the testing recall accuracy are what we care for the most in the particular dataset analysis, Naïve Bayes offers the best results for the given dataset and its subsets.

Conclusion

We started the model building process by implementing data mining techniques on the given full dataset. Later, we carried out data reduction, data exploration and implemented various individual subsets of variables in the data mining models and found four prime categories (as mentioned earlier) that gave a high true recall accuracy. The above summarized tables clearly portray that choosing a full dataset and various subsets yields different performance. Hence, we can conclude that choosing a different subset of variables yields us different performance parameters.

The best models obtained from each of the six data mining algorithms are used for further analysis as we try to find the best model that yields us a maximum net profit.

(Please refer Appendix for supporting information)

Question No. 2)

Solution.

Given that the dataset consists of 5.1% responders and the remaining 94.9% of them as non-responders, we have taken a weighted sample of the dataset. This weighted sampling of the dataset yields the dataset as 35% responders and 65% non-responders. Since a weighted sample has been considered for our analysis, the profit and loss values directly from a weighted sample dataset will yield inaccurate results and hence it should be accounted for.

Profit = $13 - $0.68 = $12.32

Cost = $0.68

To undo the effect of weighted sampling, the following process has been implemented by us and is reflected to adjust to the ratio of the original responders and non-responders.

Adjusted Profit: $12.32 X (0.051 / 0.35) = $1.7952
Adjusted Cost: $0.68 X (0.949 / 0.65) = $0.9928

Therefore, these scaled values of the Profit and Cost are now used to calculate the net profit in the further steps.

Hence, **Scaled Profit = $1.7952; Scaled Cost = -$0.9928**

The **Net Profit Formula** is now given as: If (TARGET_B==1, $1.7952, -$0.9928)

Below is a summarization of the Training and the Testing accuracy of all of the 6 Data mining models used throughout this assignment.

Classification Models	Accuracy (%) for the Training Dataset	Accuracy (%) for the Test Dataset	Testing Recall Accuracy (%)
Naïve Bayes	54.29%	51.15%	64.35%
W-Logistic Regression	64.10%	67.68%	21.26%
W-J48	82.43%	57.85%	37.11%
K-nearest neighbors	78.48%	57.83%	32.24%
Random Forest	92.6%	59.03%	35.02%
Support Vector Machine	82.6%	58.9%	46.7%

Lift of Net Profit

From the best models out of each of the 6 Data mining algorithms used for the analysis, we now calculate the maximum cumulative profit and the lift of the net profit. This information is summarized below.

Classification Models	Max Cumulative profit		Lift of Net Profit	
	Training	Validation	Training	Validation
Naïve Bayes	493.272	294.290	595.255	362.29
W-Logistic Regression	784.108	392.958	886.091	460.958
W-J48	2184.146	219.531	2286.129	287.531
K-nearest neighbors	2352.637	652.909	2454.62	720.909
Random Forest	2959.088	365.928	3061.071	433.928
Support Vector Machine	2435.202	390.034	2537.185	458.034

Note: For the calculation of the Lift of Net profit, ***(please refer Appendix)***.

Question No. 3)

Solution.

Since the lift curves describe how a model is performing against no-model, the next step is to obtain a lift curve of these maximum cumulative net profits. As per the calculations of the net profit, after considering the effect of a weighted dataset, the maximum cumulative profit has been described previously. Now, for each of the 6 data mining algorithms, we plot the cumulative net profit onto a single graph for a better idea of the model yielding us the best profit and to check if any particular algorithm is dominant.

The lift curve for the cumulative net profit is shown below.

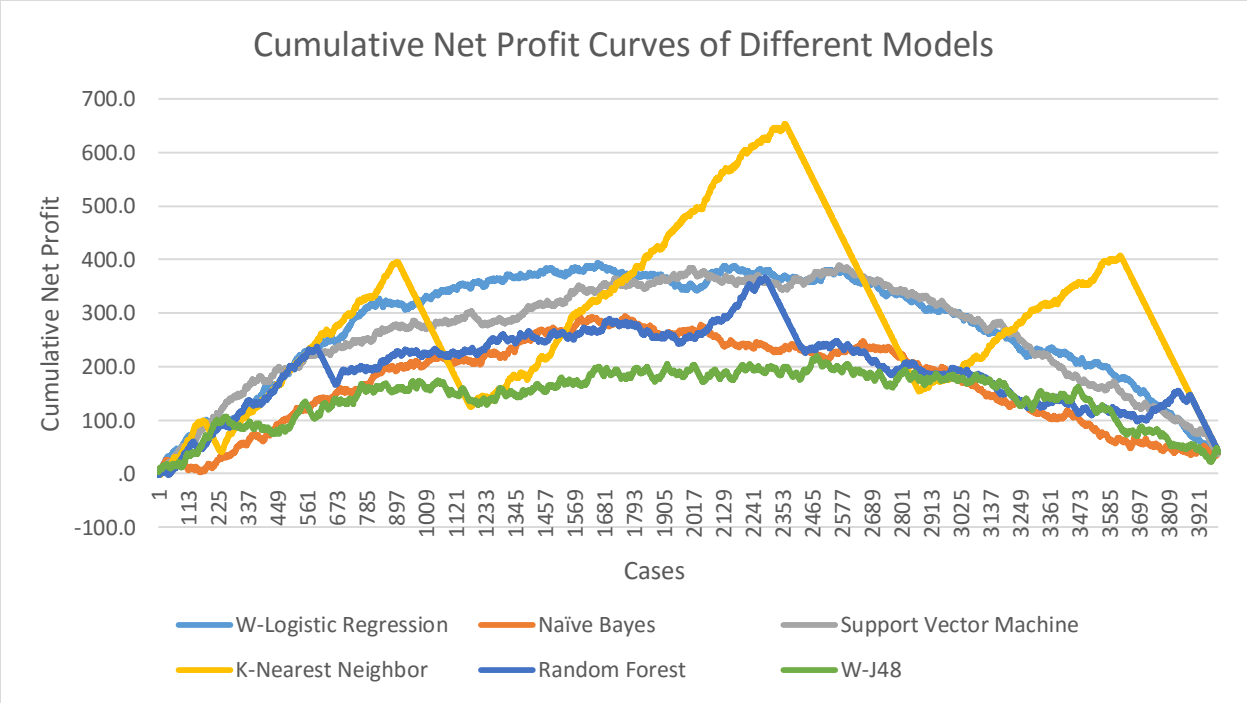

Fig 2.1. Plot of the cumulative net profit for all the six data mining models

The graph shows us that the model with the K-Nearest Neighbors algorithm yields us the maximum cumulative net profit of $652.909 and it clearly dominates the other algorithms on our chosen subset of variables.

(Please refer Appendix for supporting information)

Question No. 4)
Solution.

Now, our next step is to summarize the best model which yields us the maximum profit. Since in the given situation, profit is the most important aspect for the business to optimize their model upon, it is important that we prioritize the maximum profit considering each model. Here, the accuracy on the true recall cases is now less prioritized as compared to the maximum net profit of the model.

From Fig 2.1, a summarization of the maximum profit that we yield from different models is as described below:

1. **K-Nearest Neighbor**: This model could be considered as our best model, yielding a cumulative net profit of $652.909. This value of the cumulative net profit is obtained at a threshold of *0.3333*
2. **Logistic Regression**: This model could be considered as our second best model, yielding a cumulative net profit of $392.958. This value of the cumulative net profit is obtained at a threshold of *0.3556*
3. **Support Vector Machine**: This model could be considered as our third best model, yielding a cumulative net profit of $390.034. This value of the cumulative net profit is obtained at a threshold of *0.6058*
4. **Random Forest**: This model could be considered as our fourth best model, yielding a cumulative net profit of $365.928. This value of the cumulative net profit is obtained at a threshold of *0.6666*
5. **Naïve Bayes**: This model could be considered as our fifth best model, yielding a cumulative net profit of $294.290. This value of the cumulative net profit is obtained at a threshold of *0.9444*
6. **W-J48 (Decision Trees)**: This model could be considered as our sixth and last model to be considered, yielding a cumulative net profit of $219.531. This value of the cumulative net profit is obtained at a threshold of *0.1818*

Question No. 5)

Solution.

The final and the last step of the process is to test the model obtained on an unseen dataset. This process is known as 'Scoring' and is the final, but the most important step in validating the accuracy of the final model. To validate the model on a given dataset of 20,000 cases, we have to score and classify them as donor and non-donors as accurately as possible.

After having obtained a model using the K-Nearest Neighbors algorithm as our best model, we use the same to test on our validation dataset. The 'Write Model' operator is used to obtain a .mod file consisting the model's information and later it is applied to the unseen dataset labelled 'pva_futureData_forScoring.csv'.

We now summarize the scoring of the unseen dataset as described below.

Threshold	No. of Donors (Predicted 1)	No. of Non-Donors (Predicted 0)	Percentage of Donors (%)
0.65	6210	13,790	31.05

(Please refer Appendix for supporting information)

APPENDIX

Question No. 1

Naïve Bayes model

All PCA values considered

accuracy: 60.99%

	true 0	true 1	class precision
pred. 0	2606	1007	72.13%
pred. 1	1333	1053	44.13%
class recall	66.16%	51.12%	

Training Performance matrix

accuracy: 59.23%

	true 0	true 1	class precision
pred. 0	1637	707	69.84%
pred. 1	924	732	44.20%
class recall	63.92%	50.87%	

Testing Performance matrix

PCA values of 'DONOR Information' subset removed

accuracy: 61.41%

	true 0	true 1	class precision
pred. 0	2634	1010	72.28%
pred. 1	1305	1050	44.59%
class recall	66.87%	50.97%	

Training Performance matrix

accuracy: 58.98%

	true 0	true 1	class precision
pred. 0	1651	731	69.31%
pred. 1	910	708	43.76%
class recall	64.47%	49.20%	

Testing Performance matrix

PCA values of 'Mail order offers' subset removed

accuracy: 60.81%

	true 0	true 1	class precision
pred. 0	2601	1013	71.97%
pred. 1	1338	1047	43.90%
class recall	66.03%	50.83%	

Training Performance matrix

accuracy: 59.15%

	true 0	true 1	class precision
pred. 0	1628	701	69.90%
pred. 1	933	738	44.17%
class recall	63.57%	51.29%	

Testing Performance matrix

PCA values of 'Neighborhood' subset removed

accuracy: 54.29%

	true 0	true 1	class precision
pred. 0	1865	668	73.63%
pred. 1	2074	1392	40.16%
class recall	47.35%	67.57%	

Training Performance matrix

accuracy: 51.15%

	true 0	true 1	class precision
pred. 0	1120	513	68.59%
pred. 1	1441	926	39.12%
class recall	43.73%	64.35%	

Testing Performance matrix

PCA values of 'RFA' subset removed

accuracy: 61.31%

	true 0	true 1	class precision
pred. 0	2588	970	72.74%
pred. 1	1351	1090	44.65%
class recall	65.70%	52.91%	

Training Performance matrix

accuracy: 58.75%

	true 0	true 1	class precision
pred. 0	1610	699	69.73%
pred. 1	951	740	43.76%
class recall	62.87%	51.42%	

Testing Performance matrix

PCA values of all subsets removed

accuracy: 54.21%

	true 0	true 1	class precision
pred. 0	1867	675	73.45%
pred. 1	2072	1385	40.06%
class recall	47.40%	67.23%	

Training Performance matrix

accuracy: 51.05%

	true 0	true 1	class precision
pred. 0	1123	520	68.35%
pred. 1	1438	919	38.99%
class recall	43.85%	63.86%	

Testing Performance matrix

W-Logistic Regression model

All PCA values considered

accuracy: 64.10%

	true 0	true 1	class precision
pred. 0	2298	1173	66.21%
pred. 1	263	266	50.28%
class recall	89.73%	18.49%	

Training Performance matrix

accuracy: 67.68%

	true 0	true 1	class precision
pred. 0	3622	1622	69.07%
pred. 1	317	438	58.01%
class recall	91.95%	21.26%	

Testing Performance matrix

PCA values of 'DONOR Information' subset removed

accuracy: 67.68%

	true 0	true 1	class precision
pred. 0	3627	1627	69.03%
pred. 1	312	433	58.12%
class recall	92.08%	21.02%	

Training Performance matrix

accuracy: 63.50%

	true 0	true 1	class precision
pred. 0	2290	1189	65.82%
pred. 1	271	250	47.98%
class recall	89.42%	17.37%	

Testing Performance matrix

PCA values of 'Mail order offers' subset removed

accuracy: 67.66%

	true 0	true 1	class precision
pred. 0	3623	1624	69.05%
pred. 1	316	436	57.98%
class recall	91.98%	21.17%	

Training Performance matrix

accuracy: 63.98%

	true 0	true 1	class precision
pred. 0	2289	1169	66.19%
pred. 1	272	270	49.82%
class recall	89.38%	18.76%	

Testing Performance matrix

PCA values of 'Neighborhood' subset removed

accuracy: 67.68%

	true 0	true 1	class precision
pred. 0	3626	1626	69.04%
pred. 1	313	434	58.10%
class recall	92.05%	21.07%	

Training Performance matrix

accuracy: 64.00%

	true 0	true 1	class precision
pred. 0	2298	1177	66.13%
pred. 1	263	262	49.90%
class recall	89.73%	18.21%	

Testing Performance matrix

PCA values of 'RFA' subset removed

accuracy: 67.74%

	true 0	true 1	class precision
pred. 0	3622	1618	69.12%
pred. 1	317	442	58.23%
class recall	91.95%	21.46%	

Training Performance matrix

accuracy: 64.00%

	true 0	true 1	class precision
pred. 0	2295	1174	66.16%
pred. 1	266	265	49.91%
class recall	89.61%	18.42%	

Testing Performance matrix

PCA values of all subsets removed

accuracy: 62.35%

	true 0	true 1	class precision
pred. 0	3221	1876	63.19%
pred. 1	363	487	57.29%
class recall	89.87%	20.61%	

Training Performance matrix

	true 0	true 1	class precision
pred. 0	2156	1234	63.60%
pred. 1	260	315	54.78%
class recall	89.24%	20.34%	

Testing Performance matrix

W-J48 model

All PCA values considered

accuracy: 82.43%

	true 0	true 1	class precision
pred. 0	3486	601	85.29%
pred. 1	453	1459	76.31%
class recall	88.50%	70.83%	

Training Performance matrix

accuracy: 57.85%

	true 0	true 1	class precision
pred. 0	1780	905	66.29%
pred. 1	781	534	40.61%
class recall	69.50%	37.11%	

Testing Performance matrix

PCA values of 'DONOR Information' subset removed

accuracy: 78.13%

	true 0	true 1	class precision
pred. 0	3521	894	79.75%
pred. 1	418	1166	73.61%
class recall	89.39%	56.60%	

Training Performance matrix

accuracy: 59.13%

	true 0	true 1	class precision
pred. 0	1964	1038	65.42%
pred. 1	597	401	40.18%
class recall	76.69%	27.87%	

Testing Performance matrix

PCA values of 'Mail order offers' subset removed

accuracy: 82.26%

	true 0	true 1	class precision
pred. 0	3499	624	84.87%
pred. 1	440	1436	76.55%
class recall	88.83%	69.71%	

Training Performance matrix

accuracy: 58.98%

	true 0	true 1	class precision
pred. 0	1841	921	66.65%
pred. 1	720	518	41.84%
class recall	71.89%	36.00%	

Testing Performance matrix

PCA values of 'Neighborhood' subset removed

accuracy: 84.61%

	true 0	true 1	class precision
pred. 0	3579	563	86.41%
pred. 1	360	1497	80.61%
class recall	90.86%	72.67%	

Training Performance matrix

accuracy: 56.67%

	true 0	true 1	class precision
pred. 0	1804	976	64.89%
pred. 1	757	463	37.95%
class recall	70.44%	32.18%	

Testing Performance matrix

PCA values of 'RFA' subset removed

accuracy: 81.96%

	true 0	true 1	class precision
pred. 0	3525	668	84.07%
pred. 1	414	1392	77.08%
class recall	89.49%	67.57%	

Training Performance matrix

accuracy: 58.40%

	true 0	true 1	class precision
pred. 0	1821	924	66.34%
pred. 1	740	515	41.04%
class recall	71.11%	35.79%	

Testing Performance matrix

PCA values of all subsets removed

accuracy: 83.76%

	true 0	true 1	class precision
pred. 0	3546	581	85.92%
pred. 1	393	1479	79.01%
class recall	90.02%	71.80%	

Training Performance matrix

accuracy: 56.90%

	true 0	true 1	class precision
pred. 0	1805	968	65.09%
pred. 1	756	471	38.39%
class recall	70.48%	32.73%	

Testing Performance matrix

k-Nearest Neighbors model

All PCA values considered

accuracy: 78.55%

	true 0	true 1	class precision
pred. 0	3496	844	80.55%
pred. 1	443	1216	73.30%
class recall	88.75%	59.03%	

Training Performance matrix

accuracy: 57.83%

	true 0	true 1	class precision
pred. 0	1851	977	65.45%
pred. 1	710	462	39.42%
class recall	72.28%	32.11%	

Testing Performance matrix

PCA values of 'DONOR Information' subset removed

accuracy: 78.10%

	true 0	true 1	class precision
pred. 0	3477	852	80.32%
pred. 1	462	1208	72.34%
class recall	88.27%	58.64%	

Training Performance matrix

accuracy: 58.10%

	true 0	true 1	class precision
pred. 0	1867	982	65.53%
pred. 1	694	457	39.70%
class recall	72.90%	31.76%	

Testing Performance matrix

PCA values of 'Mail order offers' subset removed

accuracy: 78.56%

	true 0	true 1	class precision
pred. 0	3485	832	80.73%
pred. 1	454	1228	73.01%
class recall	88.47%	59.61%	

Training Performance matrix

accuracy: 57.10%

	true 0	true 1	class precision
pred. 0	1838	993	64.92%
pred. 1	723	446	38.15%
class recall	71.77%	30.99%	

Testing Performance matrix

PCA values of 'Neighborhood' subset removed

accuracy: 77.86%

	true 0	true 1	class precision
pred. 0	3497	886	79.79%
pred. 1	442	1174	72.65%
class recall	88.78%	56.99%	

Training Performance matrix

accuracy: 56.12%

	true 0	true 1	class precision
pred. 0	1879	1073	63.65%
pred. 1	682	366	34.92%
class recall	73.37%	25.43%	

Testing Performance matrix

PCA values of 'RFA' subset removed

accuracy: 78.48%

	true 0	true 1	class precision
pred. 0	3495	847	80.49%
pred. 1	444	1213	73.20%
class recall	88.73%	58.88%	

Training Performance matrix

accuracy: 57.83%

	true 0	true 1	class precision
pred. 0	1849	975	65.47%
pred. 1	712	464	39.46%
class recall	72.20%	32.24%	

Testing Performance matrix

PCA values of all subsets removed

accuracy: 77.90%

	true 0	true 1	class precision
pred. 0	3499	886	79.79%
pred. 1	440	1174	72.74%
class recall	88.83%	56.99%	

Training Performance matrix

accuracy: 56.10%

	true 0	true 1	class precision
pred. 0	1878	1073	63.64%
pred. 1	683	366	34.89%
class recall	73.33%	25.43%	

Testing Performance matrix

Random Forest model

All PCA values considered

accuracy: 92.30%

	true 0	true 1	class precision
pred. 0	3739	262	93.45%
pred. 1	200	1798	89.99%
class recall	94.92%	87.28%	

Training Performance matrix

accuracy: 59.60%

	true 0	true 1	class precision
pred. 0	1888	943	66.69%
pred. 1	673	496	42.43%
class recall	73.72%	34.47%	

Testing Performance matrix

PCA values of 'DONOR Information' subset removed

accuracy: 92.58%

	true 0	true 1	class precision
pred. 0	3751	257	93.59%
pred. 1	188	1803	90.56%
class recall	95.23%	87.52%	

Training Performance matrix

accuracy: 57.57%

	true 0	true 1	class precision
pred. 0	1810	946	65.67%
pred. 1	751	493	39.63%
class recall	70.68%	34.26%	

Testing Performance matrix

PCA values of 'Mail order offers' subset removed

accuracy: 92.60%

	true 0	true 1	class precision
pred. 0	3741	246	93.83%
pred. 1	198	1814	90.16%
class recall	94.97%	88.06%	

Training Performance matrix

accuracy: 59.03%

	true 0	true 1	class precision
pred. 0	1857	935	66.51%
pred. 1	704	504	41.72%
class recall	72.51%	35.02%	

Testing Performance matrix

PCA values of 'Neighborhood' subset removed

accuracy: 93.12%

	true 0	true 1	class precision
pred. 0	3764	238	94.05%
pred. 1	175	1822	91.24%
class recall	95.56%	88.45%	

Training Performance matrix

accuracy: 57.53%

	true 0	true 1	class precision
pred. 0	1851	989	65.18%
pred. 1	710	450	38.79%
class recall	72.28%	31.27%	

Testing Performance matrix

PCA values of 'RFA' subset removed

accuracy: 92.30%

	true 0	true 1	class precision
pred. 0	3739	262	93.45%
pred. 1	200	1798	89.99%
class recall	94.92%	87.28%	

Training Performance matrix

accuracy: 59.60%

	true 0	true 1	class precision
pred. 0	1888	943	66.69%
pred. 1	673	496	42.43%
class recall	73.72%	34.47%	

Testing Performance matrix

PCA values of all subsets removed

accuracy: 92.52%

	true 0	true 1	class precision
pred. 0	3759	269	93.32%
pred. 1	180	1791	90.87%
class recall	95.43%	86.94%	

Training Performance matrix

accuracy: 56.73%

	true 0	true 1	class precision
pred. 0	1784	954	65.16%
pred. 1	777	485	38.43%
class recall	69.66%	33.70%	

Testing Performance matrix

Support Vector Machine model

All PCA values considered

accuracy: 85.03%

	true 0	true 1	class precision
pred. 0	3444	403	89.52%
pred. 1	495	1657	77.00%
class recall	87.43%	80.44%	

Training Performance matrix

accuracy: 60.40%

	true 0	true 1	class precision
pred. 0	1882	905	67.53%
pred. 1	679	534	44.02%
class recall	73.49%	37.11%	

Testing Performance matrix

PCA values of 'DONOR Information' subset removed

accuracy: 84.56%

	true 0	true 1	class precision
pred. 0	3309	296	91.79%
pred. 1	630	1764	73.68%
class recall	84.01%	85.63%	

Training Performance matrix

accuracy: 58.80%

	true 0	true 1	class precision
pred. 0	1712	799	68.18%
pred. 1	849	640	42.98%
class recall	66.85%	44.48%	

Testing Performance matrix

PCA values of 'Mail order offers' subset removed

accuracy: 84.25%

	true 0	true 1	class precision
pred. 0	3298	304	91.56%
pred. 1	641	1756	73.26%
class recall	83.73%	85.24%	

Training Performance matrix

accuracy: 58.98%

	true 0	true 1	class precision
pred. 0	1709	789	68.41%
pred. 1	852	650	43.28%
class recall	66.73%	45.17%	

Testing Performance matrix

PCA values of 'Neighborhood' subset removed

accuracy: 97.45%

	true 0	true 1	class precision
pred. 0	3800	14	99.63%
pred. 1	139	2046	93.64%
class recall	96.47%	99.32%	

Training Performance matrix

accuracy: 61.22%

	true 0	true 1	class precision
pred. 0	2129	1119	65.55%
pred. 1	432	320	42.55%
class recall	83.13%	22.24%	

Testing Performance matrix

PCA values of 'RFA' subset removed

accuracy: 82.60%

	true 0	true 1	class precision
pred. 0	3224	329	90.74%
pred. 1	715	1731	70.77%
class recall	81.85%	84.03%	

Training Performance matrix

accuracy: 58.90%

	true 0	true 1	class precision
pred. 0	1684	767	68.71%
pred. 1	877	672	43.38%
class recall	65.76%	46.70%	

Testing Performance matrix

PCA values of all subsets removed

accuracy: 98.63%

	true 0	true 1	class precision
pred. 0	3890	33	99.16%
pred. 1	49	2027	97.64%
class recall	98.76%	98.40%	

Training Performance matrix

accuracy: 62.00%

	true 0	true 1	class precision
pred. 0	2185	1144	65.64%
pred. 1	376	295	43.96%
class recall	85.32%	20.50%	

Testing Performance matrix

Question No. 2

Steps to calculate the Net Profit:

1. Sort the Confidence (1) by decreasing order
2. The actual and the predicted donors are compared against each other
3. If an accurate classification is made, the profit associated with it is set to $1.7952
4. If an inaccurate classification is made, the cost associated with it is set to -$0.9928

We now know that the lift of a model describes how well a model performs as compared to the no-model scenario, it is important to compute the lift for each model.

As described below, we compute the lift for the training and testing datasets across all of the six data mining models.

Training Dataset

Total number of data points = 5999

Therefore**,** maximum cumulative profit = (5999*0.35) * 1.7952 – (5999*0.65) * 0.9928 = -$101.983 (negative cumulative profit)

Validation/Testing Dataset

Total number of data points = 4000

Therefore**,** maximum cumulative profit = (4000*0.35) * 1.7952 – (4000*0.65) * 0.9928 = -$68.00 (negative cumulative profit)

Lift for each model is given as = Maximum Cumulative Profit – (-68.00)

Using the above described computations, we find the lift of the net profit for the training and the testing datasets on all of the six models and it was summarized above.

Question No. 3

The lift of the cumulative net profit for each of the six different data mining models have been summarized into a single graph as mentioned earlier. The further details of this graph that includes the lift of net cumulative profit is included in the Excel spreadsheet linked below.

Lift of Cumulative Net Profit.xlsx

Question No. 5

For scoring the unseen validation dataset, we first write the optimal model using the K-Nearest Neighbor algorithm and then this model is applied to the unseen data. All of the files used in scoring this new unseen dataset are attached below.

Best model (KNN).mod

Output of Scoring Unseen dataset.xlsx

Final Scoring of Unseen Data.rmp

Best Model (KNN) - Writing the model.rm