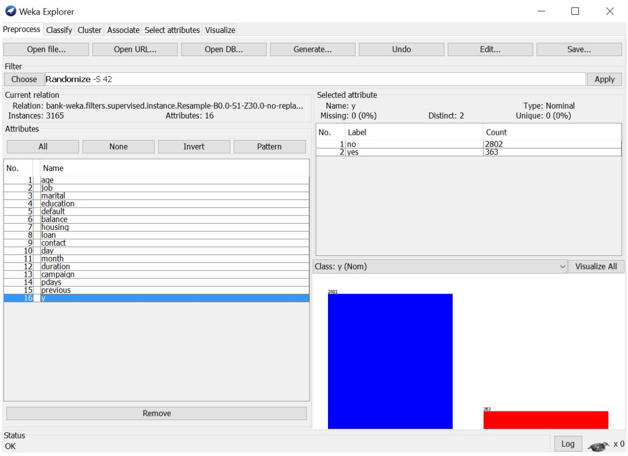
CS699 Term Project Spring 2016

Sai Abhilash Ghanta

4/14/2016

1 Unbalanced Dataset:

The dataset is subjected to randomization to ensure the instances are not grouped. This will help in better model testing.



Filter: Randomize -S 42

Current relation
Relation: bank-weka.filters.supervised.instance.Resample-B0.0-S1-Z30.0-no-repla...
Instances: 3165 Attributes: 16

Selected attribute
Name: y Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count
1	no	2802
2	yes	363

Attributes

No.	Name
1	age
2	job
3	marital
4	education
5	default
6	balance
7	housing
8	loan
9	contact
10	day
11	month
12	duration
13	campaign
14	pdays
15	previous
16	y

Class: y (Nom)

Status
OK

1. Naïve Bayes:

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.924	0.562	0.927	0.924	0.925	0.845	No
	0.438	0.076	0.426	0.438	0.432	0.845	Yes
Weighted Avg.	0.868	0.506	0.87	0.868	0.869	0.845	

2. J48 (Decision Tree):

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.961	0.59	0.926	0.961	0.944	0.759	No
	0.41	0.039	0.58	0.41	0.481	0.759	Yes
Weighted Avg.	0.898	0.526	0.887	0.898	0.891	0.759	

3. Multilayer Perceptron:

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.951	0.628	0.921	0.951	0.936	0.818	No
	0.372	0.049	0.498	0.372	0.426	0.818	Yes
Weighted Avg.	0.885	0.562	0.873	0.885	0.878	0.818	

4. Logistic:

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.975	0.719	0.913	0.975	0.943	0.878	No
	0.281	0.025	0.593	0.281	0.381	0.878	Yes
Weighted Avg.	0.895	0.639	0.876	0.895	0.878	0.878	

2 Oversampled Dataset:

The dataset is subjected to SMOTE function which is available in Supervised Instance of Weka explorer. This is will increase the number of values/instances in the specified class. The settings used for SMOTE function are

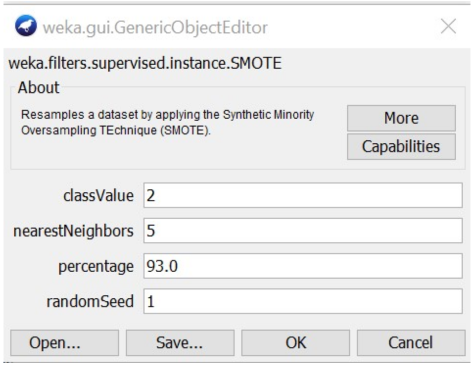
Class Value: 2

nearestNeighbors: 5

percentage: 100 for 2 times and 93 for third time. (To achieve equal number of instances)

random seed: 1

When SMOTE function is used, all the new instances of the class will be added at the end of the data set. So randomization of data is important as the models are tested using Cross Validation with 10folds. Randomization is done using the randomSeed of 42(same as in Oversampled).



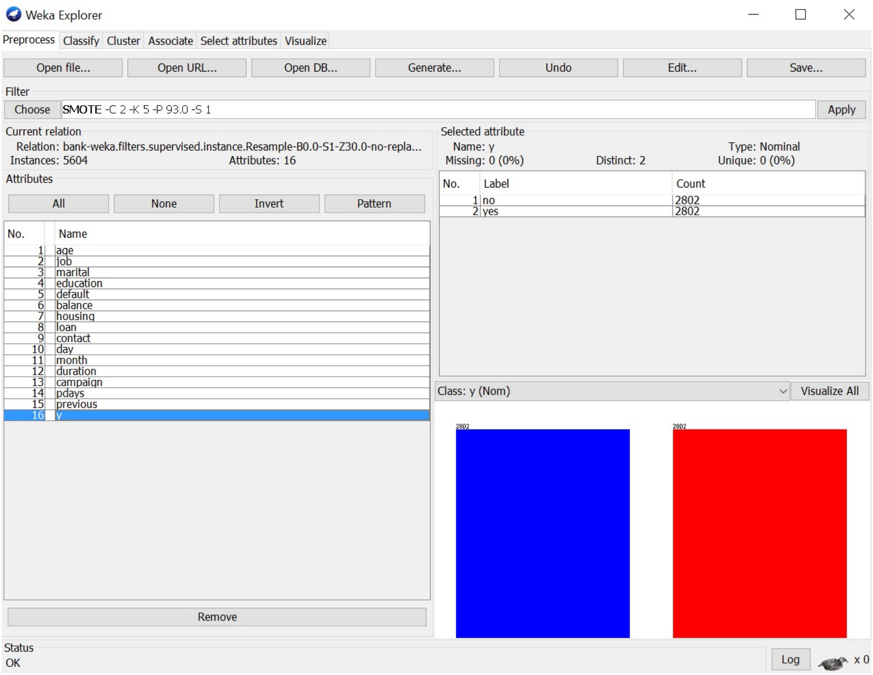
weka.gui.GenericObjectEditor

weka.filters.supervised.instance.SMOTE

About

Resamples a dataset by applying the Synthetic Minority Oversampling TEchnique (SMOTE).

classValue	2
nearestNeighbors	5
percentage	93.0
randomSeed	1



Weka Explorer

Filter: SMOTE -C 2 -K 5 -P 93.0 -S 1

Current relation
Relation: bank-weka.filters.supervised.instance.Resample-B0.0-S1-Z30.0-no-repla...
Instances: 5604 Attributes: 16

Selected attribute
Name: y Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count
1	no	2802
2	yes	2802

Attributes

No.	Name
1	age
2	job
3	marital
4	education
5	default
6	balance
7	housing
8	loan
9	contact
10	day
11	month
12	duration
13	campaign
14	pdays
15	previous
16	y

Class: y (Nom)

Status
OK

1. Naïve Bayes:

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.805	0.078	0.912	0.804	0.855	0.925	No
	0.922	0.196	0.825	0.922	0.871	0.925	Yes
Weighted Avg.	0.863	0.137	0.868	0.863	0.863	0.925	

2. J48 (Decision Tree):

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.914	0.079	0.92	0.914	0.917	0.934	No
	0.921	0.086	0.915	0.921	0.918	0.934	Yes
Weighted Avg.	0.918	0.082	0.918	0.918	0.918	0.934	

3. Multilayer Perceptron:

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.917	0.077	0.922	0.917	0.92	0.96	No
	0.923	0.083	0.918	0.923	0.92	0.96	Yes
Weighted Avg.	0.92	0.08	0.92	0.92	0.92	0.96	

4. Logistic:

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.898	0.082	0.917	0.898	0.907	0.953	No
	0.918	0.102	0.9	0.918	0.909	0.953	Yes
Weighted Avg.	0.908	0.092	0.908	0.908	0.908	0.953	

As the tables shows, the TP rate for Yes class of all the models are increased drastically with oversampled data. The oversampled data set after randomization is submitted along with the other supporting materials.

3 Undersampled Dataset:

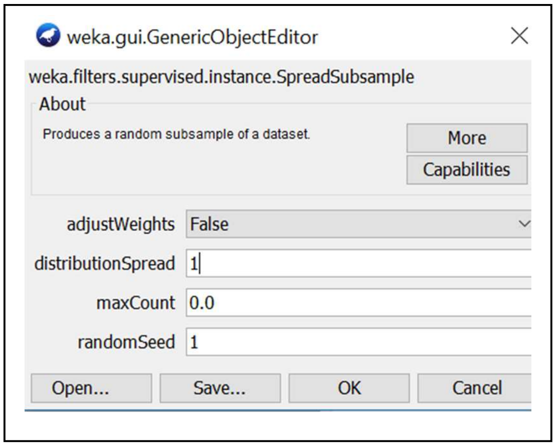
The original dataset is loaded again to Weka explorer and is subjected to "SpreadSubSample" function which is available in Supervised Instance of Weka explorer. This is will decrease the number of values/instances in the specified class "No" (distributionSpread=1). The settings used for "SpreadSubSample" function are

adjustWeights: False

distributionSpread: 1

maxCount: 0.0

random seed: 1


weka.gui.GenericObjectEditor

weka.filters.supervised.instance.SpreadSubsample

About

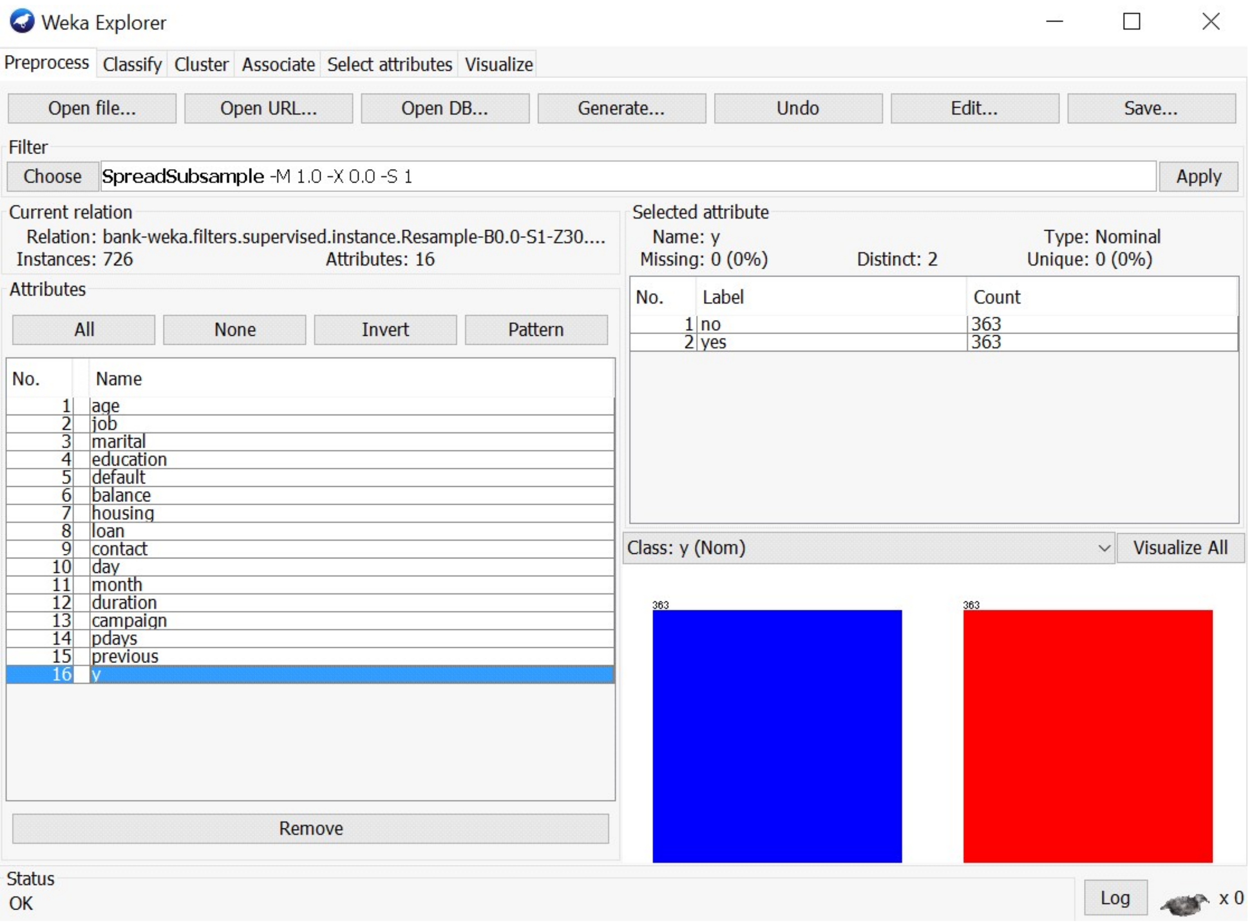
Produces a random subsample of a dataset.

adjustWeights: False

distributionSpread: 1

maxCount: 0.0

randomSeed: 1


Weka Explorer

Filter: SpreadSubsample -M 1.0 -X 0.0 -S 1

Current relation
Relation: bank-weka.filters.supervised.instance.Resample-B0.0-S1-Z30....
Instances: 726
Attributes: 16

Selected attribute
Name: y
Type: Nominal
Missing: 0 (0%)
Distinct: 2
Unique: 0 (0%)

No.	Label	Count
1	no	363
2	yes	363

Attributes

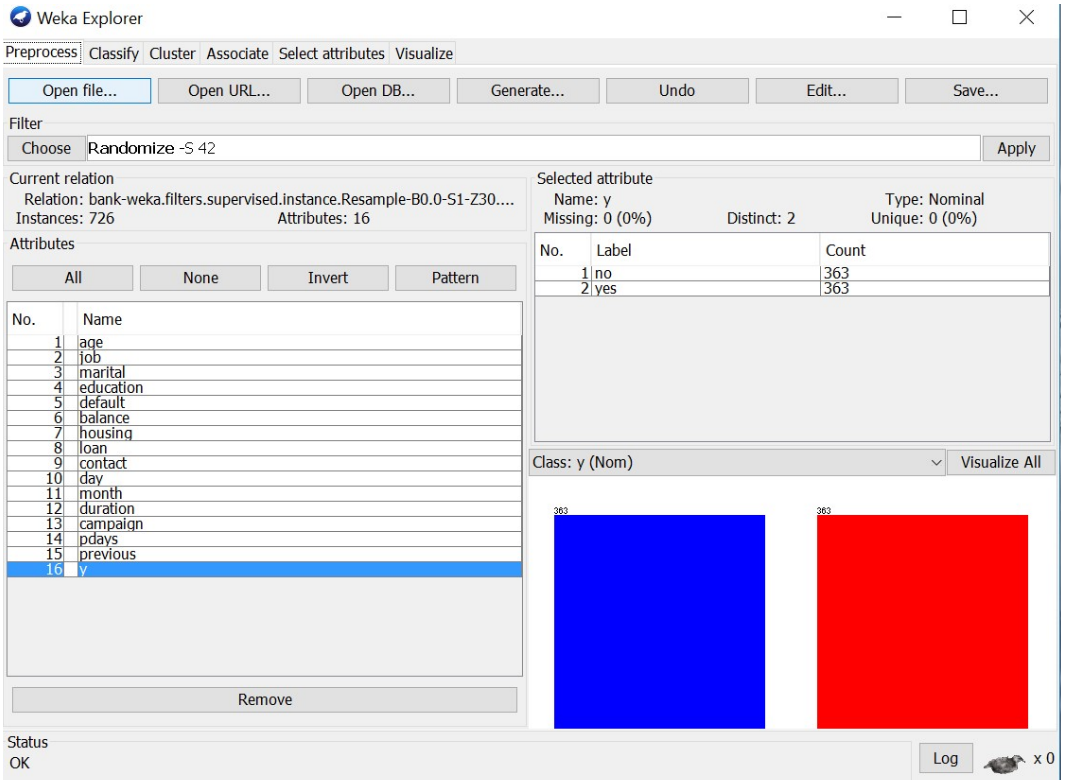
No.	Name
1	age
2	job
3	marital
4	education
5	default
6	balance
7	housing
8	loan
9	contact
10	day
11	month
12	duration
13	campaign
14	pdays
15	previous
16	y

Class: y (Nom)

Status
OK

When “SpreadSubSample” function is used, all the instances of the specified class will be deleted and the classes were arranged in two groups. So randomization of data is important as the models are tested using Cross Validation with 10folds.

Randomization is done using the randomSeed of 42(same as in Oversampled).



Filter
Randomize -S 42

Current relation
Relation: bank-weka.filters.supervised.instance.Resample-B0.0-S1-Z30....
Instances: 726 Attributes: 16

Selected attribute
Name: y Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count
1	no	363
2	yes	363

Attributes

No.	Name
1	age
2	job
3	marital
4	education
5	default
6	balance
7	housing
8	loan
9	contact
10	day
11	month
12	duration
13	campaign
14	pdays
15	previous
16	y

Class: y (Nom)

Status
OK

1. Naïve Bayes:

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.76	0.242	0.758	0.76	0.759	0.827	No
	0.758	0.24	0.76	0.758	0.759	0.827	Yes
Weighted Avg.	0.759	0.241	0.759	0.759	0.759	0.827	

2. J48 (Decision Tree):

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.793	0.182	0.814	0.793	0.803	0.853	No
	0.818	0.207	0.798	0.818	0.808	0.853	Yes
Weighted Avg.	0.806	0.194	0.806	0.806	0.806	0.853	

3. Multilayer Perceptron:

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.782	0.237	0.768	0.782	0.775	0.854	No
	0.763	0.218	0.778	0.763	0.771	0.854	Yes
Weighted Avg.	0.773	0.227	0.773	0.773	0.773	0.854	

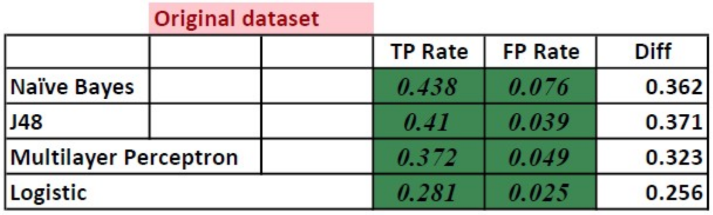
4. Logistic:

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.796	0.223	0.781	0.796	0.789	0.88	No
	0.777	0.204	0.792	0.777	0.784	0.88	Yes
Weighted Avg.	0.787	0.213	0.787	0.787	0.786	0.88	

4 Chosen Model:

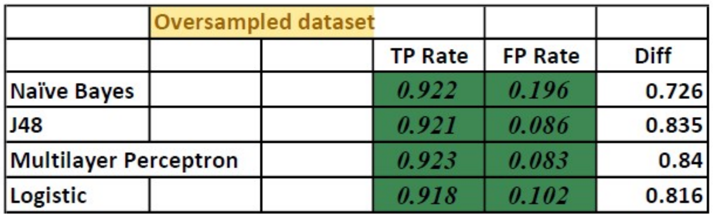
The best model for predicting the instances is the one which has high TP rate and low FP Rate for “Yes” class.

So, I calculated the differences between TP Rates and FP rates of all the performed classifiers on all the data sets. The one which has highest difference is the best model. Based on the results of calculation, I chose Multilayer Perceptron for Oversampled data set is the best classification model.



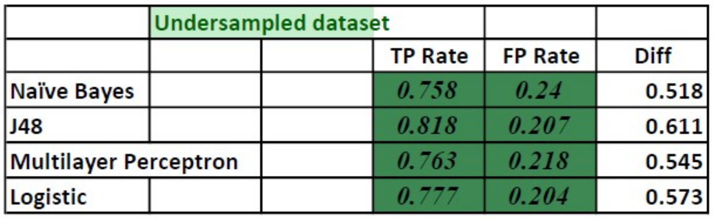
Original dataset

	TP Rate	FP Rate	Diff
Naïve Bayes	0.438	0.076	0.362
J48	0.41	0.039	0.371
Multilayer Perceptron	0.372	0.049	0.323
Logistic	0.281	0.025	0.256



Oversampled dataset

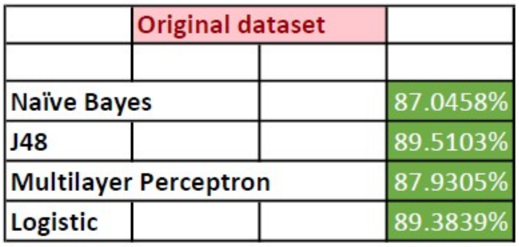
	TP Rate	FP Rate	Diff
Naïve Bayes	0.922	0.196	0.726
J48	0.921	0.086	0.835
Multilayer Perceptron	0.923	0.083	0.84
Logistic	0.918	0.102	0.816



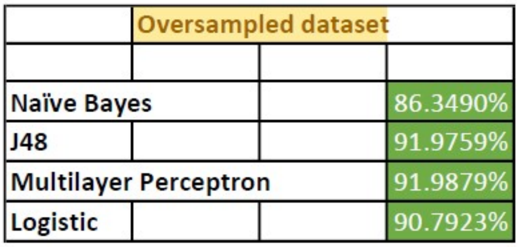
Undersampled dataset

	TP Rate	FP Rate	Diff
Naïve Bayes	0.758	0.24	0.518
J48	0.818	0.207	0.611
Multilayer Perceptron	0.763	0.218	0.545
Logistic	0.777	0.204	0.573

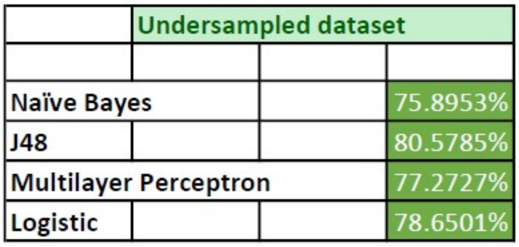
I also used the formula of Accuracy to cross-check the overall accuracy of the model. Multilayer Perceptron of Oversampled Dataset is having the best overall accuracy.



	Original dataset		
Naïve Bayes			87.0458%
J48			89.5103%
Multilayer Perceptron			87.9305%
Logistic			89.3839%



	Oversampled dataset		
Naïve Bayes			86.3490%
J48			91.9759%
Multilayer Perceptron			91.9879%
Logistic			90.7923%



	Undersampled dataset		
Naïve Bayes			75.8953%
J48			80.5785%
Multilayer Perceptron			77.2727%
Logistic			78.6501%

5 Discussion:

The given data set is having a having two classes with 2802 instances in “NO” class and 363 instances in “YES” class. In such a highly class imbalanced dataset, no classifier model can predict the “Yes” class instances with high accuracy. So I performed oversampling and undersampling to equal the number of instances in both the classes.

Because of this project, I really understood how to tackle class imbalanced datasets. I also see the necessity of randomization for getting consistently promising classifier model. After such tests, I feel Mulitlayer Perceptron on oversampled dataset is best classifier. All the screen snaps of performing models were included in the ZIP folder.

Obeservation: I also observed that the accuracy of classifiers may change on different weka sessions. I think we can avoid this using certain data pre-processing steps and changing the classifier settings.