Internet at Home

ML-2017-FALL-GROUP8-PROJECT

Akshay Suresh, MS in Data Science, Fall 2017
Noul Singla, MS in Data Science, Fall 2017
Sai Charan Konanki, MS in Data Science, Fall 2017

I. INTRODUCTION

The problem statement is prediction of internet connection status at student's home. The data was obtained in a survey of students taking Portuguese language courses in secondary school. Source of the data being used for the analysis is Kaggle. It is the result of a survey done on students taking Portuguese language courses in secondary school. The data has a lot of interesting social, gender and academic information about the students.

The kind of questions to be answered using this dataset are:

A. Which predictors can be most helpful in predicting the internet connection status at the student's home?

B. Using various supervised machine learning techniques, determine which model performs the best?

C. How could the models be improved to make them more accurate?

D. Is the information from the survey sufficient to solve the problem statement effectively?

E. Could this output be used for identifying customers with higher conversion potential?

The given problem statement is a classification problem. Various models have been implemented to predict internet connectivity at home. Efficiency of the model is determined by the best true negative rate (number of correctly predicted negatives). Selected model can be used to pitch a solution to the client on how to identify potential customers for internet connection.

I.A. DATA PREPROCESSING

The dataset contains 649 rows and 33 columns. The data is bifurcated into 3 sets namely training, selection and test in the ratio 3:1:1. An analysis of the training data provided the following findings.

Single Variable Statistics

There are 11 continuous, 5 ordered categorical and 17 categorical variables.

Outlier information

There are no missing values or obvious outliers in the dataset.

Pairwise information

There is no strong correlation between the predictors.

II. METHODS

1. Logistic regression: Since this is a 2-class problem, logistic regression may be a good interpretable fit. Logistic regression is implemented with subset selection using forward, backward and LASSO.

2. Linear Discriminant Analysis: Implementing LDA to see if the assumption of normality of predictors for each class would provide a good fit on the dataset.

3. Random forest: As there are a lot of categorical predictors, trees can be a good fit for segregating classes. RF with cross validation is further implemented with training control to find the best fit.

4. Bagging: Since data is limited, bagging is a way to decrease the variance of predictions by generating additional data for training from original dataset using repetitions.

5. Boosting: Subsets of the original data are used to produce a series of average performing models and then boost their performance by combining them together.

6. Tree with added data for 1 category: Added more repetitive data to "no" category to generate the tree and see if classification of "no" increases and how it affects the model.

A separate approach is considered to vary the threshold for classification and see if it makes more sense from a business perspective.

7. Introduced Naïve Bayes as it works well for independent predictors which seems to be the case in the dataset.

8. Implement threshold variations for Logistic regression, LDA and Naïve Bayes.

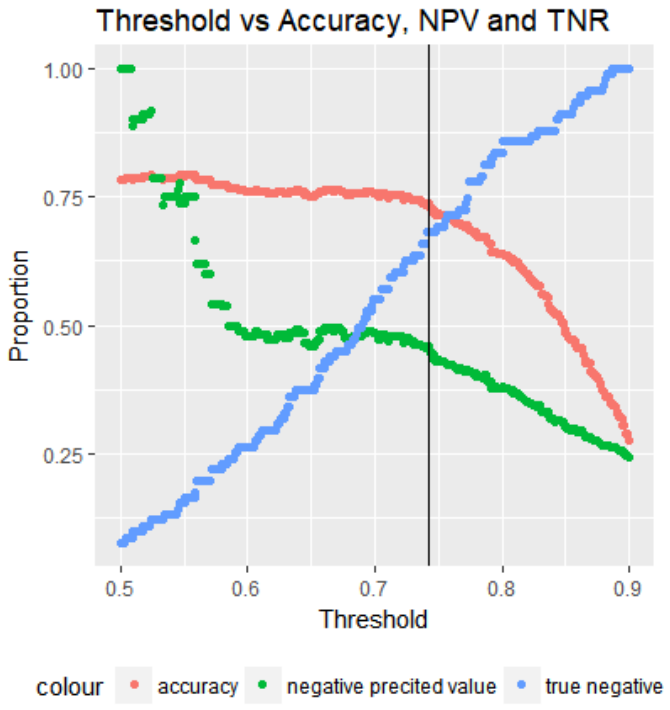
III. RESULTS

III.A. LOGISTIC REGRESSION

The decision to use logistic regression for analysis is based on the response variable being dichotomous. The threshold which is used for classifying plays a major role here as predicting better negatives is significant for the business use case. Hence, threshold is varied to obtain better true negative rate without significantly affecting negative predictive value and accuracy. Instead of using all the predictors to implement logistic regression, only a subset of the predictors is used for building the models.

*III.*A.1. SUBSET SELECTION WITH LASSO

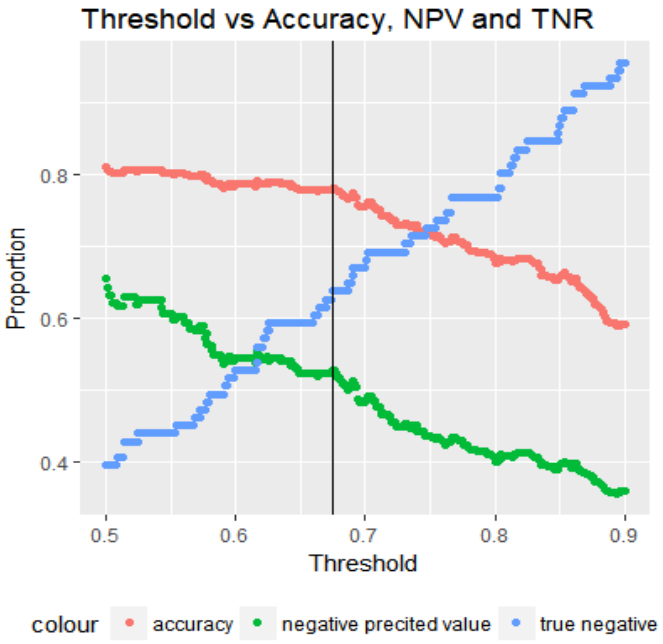
In LASSO regularization, coefficients of various predictors are reduced in magnitude. The predictors having a coefficient of 0 are not used to build the model. Hence, a model is built using the obtained predictors and coefficients. The accuracy, negative predictive value and true negative rate for different thresholds can be seen here:



From the above graph, optimum threshold is found out to be 0.742. True negative rate(TNR) obtained is 0.68, negative predictive value(NPV) and accuracy are 0.46 and 0.73 respectively. There is a tradeoff between TNR and NPV. It can be said that the logistic regression model using predictors obtained from LASSO has not performed well as the TNR obtained is not satisfactory. Hence, subset selection using forward and backward selection is implemented.

*III.*A.2. FORWARD SUBSET SELECTION

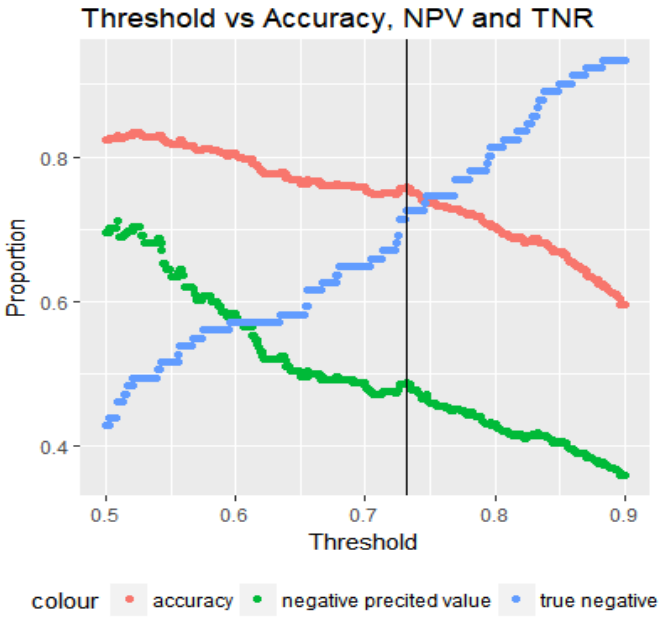
Model is built using the predictors obtained from forward subset selection.



The optimum threshold here is found out to be 0.74. The TNR obtained for the model is 0.63 and the NPV is 0.53. There is a significant increase in NPV while TNR has decreased.

*III.*A.3. BACKWARD SUBSET SELECTION

As the above subset selection methods have not performed well, backward subset selection is implemented. The accuracy, negative predictive value and true negative rate for different thresholds can be seen here:



Using the predictors selected from backward subset selection methods as inputs, a model is built. From the above graph, optimum threshold for this model is 0.73. Accuracy and Negative Predictive Value(NPV) of the model are 0.73 and 0.45 respectively. True Negative Rate (TNR) is 0.72 which is very high when compared to the above models.

Below table provides the performance statistics obtained for models that have been created using different logistic subset selection methods.

Model	Threshold	Accuracy	NPV	TNR
LASSO	0.742	0.74	0.46	0.68
Forward	0.675	0.78	0.52	0.63
Backward	0.732	0.76	0.49	0.725

From the above table, it could be seen that backward subset selection model has the best True Negative Rate. Whereas Forward has the best Negative Predictive Value. Considering the business use case, it can be concluded that the model with the best TNR i.e., the model that predicts better negatives can be taken forward to be analyzed on selection dataset. The predictors that were used to build backward subset selection model are school, mother's education, father's education, mother's job, father's job, reason, nursery, romantic, free time and G3.

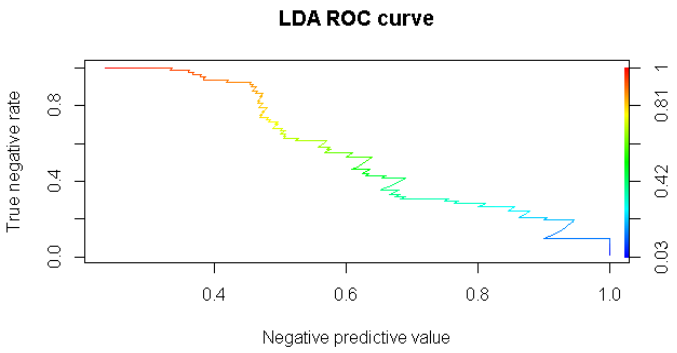
III.B. LINEAR DISCRIMINANT ANALYSIS

Linear Discriminant Analysis assumes that for each class, the predictors exhibit a normal distribution. This model would perform well on the dataset if this assumption holds true. The model implementation results are as follows:

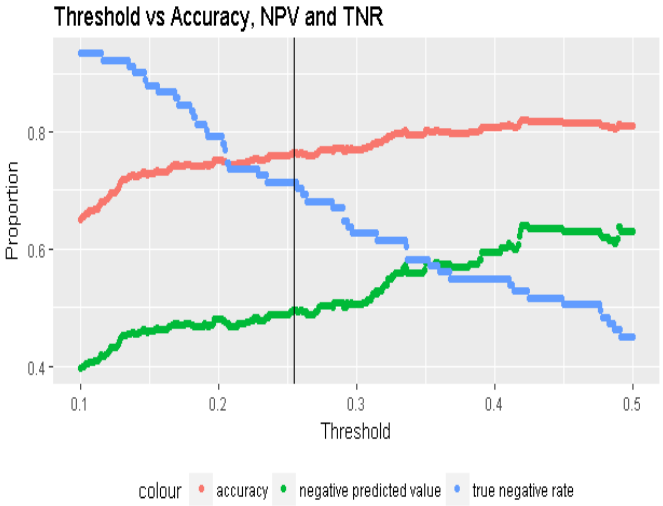
		Actual	
		no	yes
Predicted	no	41	24
	yes	50	274

Confusion Matrix for LDA

The ROC curve is as follows:



The area under the curve is 0.8546353. The above results of the confusion matrix are based on a threshold of 0.5. The threshold is varied to increase the true negative rate while limiting the effects on the accuracy and negative predictive value of the model. The accuracy, negative predictive value and true negative rate for different thresholds can be seen here:



From the given range of thresholds, the one that provides the best true negative rate while limiting the reduction in accuracy and negative predictive value is chosen. A threshold of 0.255 provides the following results and improvements:

		Actual	
		no	yes
Predicted	no	65	66
	yes	26	232

Confusion Matrix for LDA with threshold = 0.255

Threshold	Accuracy	Negative Predicted Value	True Negative Rate
0.255	0.7634961	0.4961832	0.7142857

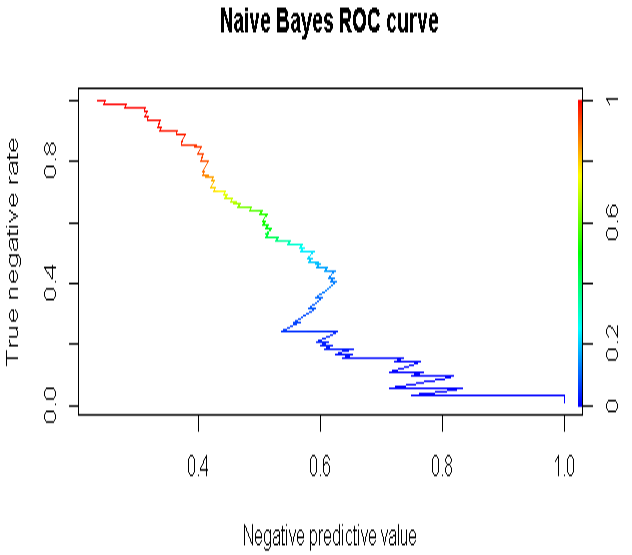
III.C. NAIVE BAYES

Naïve Bayes could be potential fit given that there was no significant correlation between the predictors. The model implementation results are as follows:

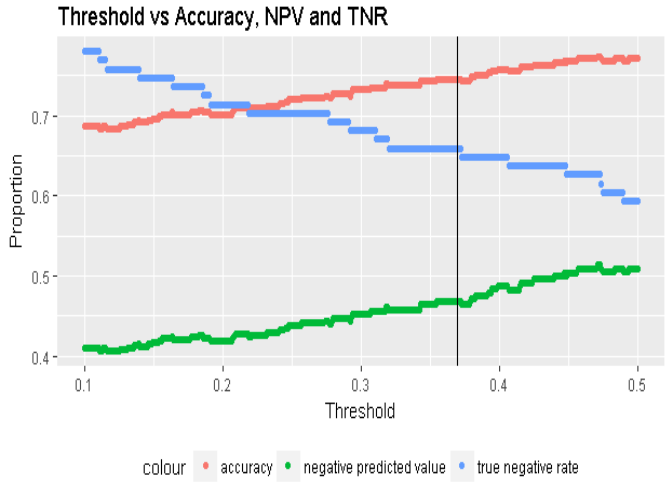
		Actual	
		no	yes
Predicted	no	54	52
	yes	37	246

Confusion Matrix for Naïve Bayes

The ROC curve is as follows:



The area under the curve is 0.8005384. The above results of the confusion matrix are based on a threshold of 0.5. The accuracy, negative predictive value and true negative rate for different thresholds can be seen here:



From the range of thresholds, one that provides the best true negative rate while limiting the reduction in accuracy and negative predictive value is chosen. A threshold of 0.37 provides the following results and improvements:

		Actual	
		no	yes
Predicted	no	60	68
	yes	31	230

Confusion Matrix for NB with threshold = 0.37

Threshold	Accuracy	Negative Predicted Value	True Negative Rate
0.37	0.7455013	0.46875	0.6593407

III.D. **Trees**

Trees provide us a model which can segregate one class from another and provide rules which can be easily interpreted. The implemented Tree model made the first segregation on School and considered Mother's job (Mjob) and Mother's education(Medu) for further segregations.

III.D.1. **MAXIMAL TREE**

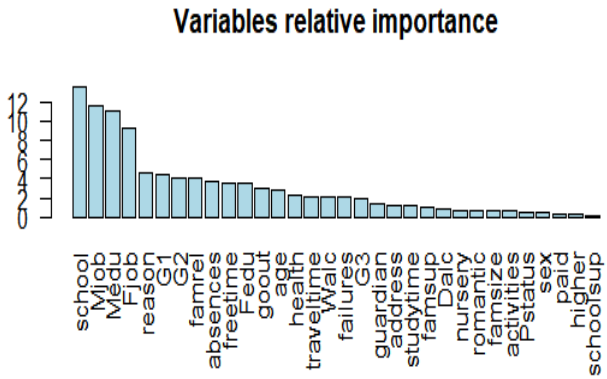
A maximal tree is trained and provides the following result:

		Actual	
		No	Yes
Predicted	No	60	39
	Yes	31	259

Confusion Matrix for Tree

III.D.2. **TREE WITH BAGGING DATA**

As it is a limited dataset, bagging trees are explored. Number of samples are varied in 10,50,100,500 and from 10 to 50 initially there is no change, with better result at 100 trees which downgrades at 500.



		Actual	
		no	yes
Predicted	no	52	1
	yes	39	297

Confusion Matrix for Bagging

There is reduction in the number of predicted negatives but the Negative Predictive Value is almost perfect. Thus,

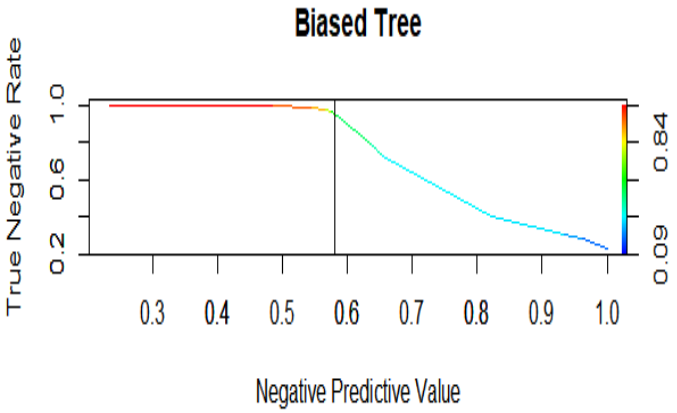
using multiple samples is a better option but the issue of low True Negative Rate is still persistent. To counter this, trees with Boosting and biased tree are implemented.

III.D.3. TREE ON RIGGED/BIASED DATA

Training data is added with records which have 'no' as the output thereby increasing the negative outcomes to twice. A maximal tree is implemented on this dataset resulting in better performance.

		Actual	
		no	yes
Predicted	no	87	63
	yes	4	235

Confusion Matrix for Biased Tree



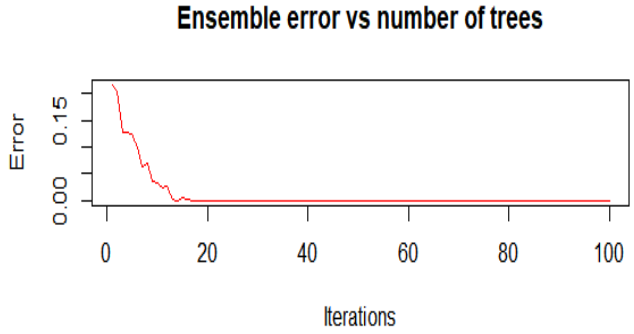
TNR vs NPV curve for Biased Tree

The True Negative rate increases to a significant level with a hit on the Negative Predictive Value which is considered acceptable as including more students without internet is the priority.

Hence tree trained on the biased dataset is considered as a good model to consider fitting on the selection set.

III.D.4. TREE WITH BOOSTING

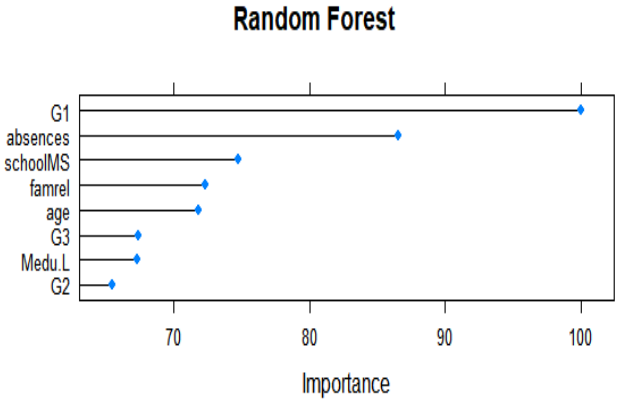
Boosting with Adaboost is considered as it adds additional weight to the category which is not having enough ratio. Boosting with 500 trees is considered, which is reduced to 100 and even further on studying the error. Even though the model is perfect, the error reaches to 0 very early, raising the concern of overfitting. Random Forest is further considered, as a similar result with cross validation would remove the concern of over fitting.



III.D.5. RANDOM FOREST

Random Forest is a model which would provide the benefit of using cross validations and limit predictors resulting in a stable output. It provides us a lot of flexibility in terms of parameter optimization. After training control with repeated cross validations and searching using random search and grid search methods, Random forest with 20 trees and 10 repeated cross validations with random search provides the best performance.

Top predictors with their relative importance for Random Forest implementation are: -



Variable Importance in Random Forest

The resultant model provides an accuracy of 1 like Adaboost. Since a number of cross validations are considered, the model may not over fit to the data set.

		Actual	
		no	yes
Predicted	no	91	0
	yes	0	298

Confusion Matrix for Random Forest

Hence, Random forest with above result on training set is chosen to be tested on the selection set.

III.E. MODEL OUTPUTS ON SELECTION DATASET

From the training set, 5 models are selected – Logistic Regression using backward subset selection, Linear Discriminant Analysis, Naïve Bayes, Tree and Random Forest. As described earlier, selection dataset constitutes 20% of the original dataset with 130 records. This data has same structure as the training dataset.

Internet	no	yes
Count	29	101

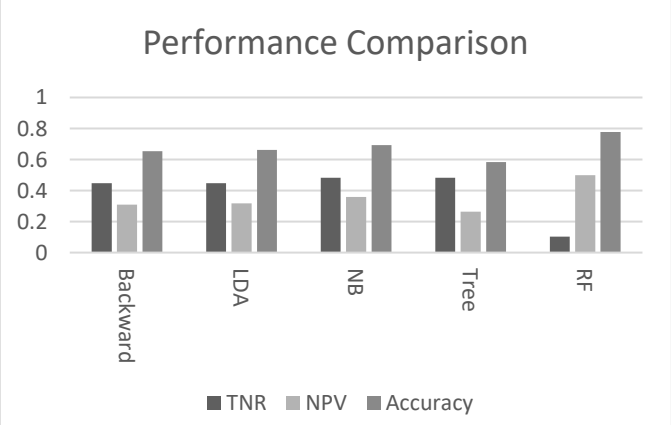
Distribution of result in Output set

The ratio of no to yes in the selection set is similar to training set.

			Actual	
			no (29)	yes (101)
Predicted	Logistic With backward subset	no	13	29
		yes	16	72
	LDA	no	13	28
		yes	16	73
	Naïve Bayes	no	14	25
		yes	15	76
	Tree	no	14	39
		yes	15	62
	Random Forest	no	3	3
		yes	26	98

Confusion Matrix on Selection set

Number of correctly predicted negatives are mostly 13 or 14 with one exception of 3. Number of correctly predicted positives are in the range of 62-98.



Accuracy, NPV and TPR comparison for all models

It is clearly visible that the True Negative Rate of Random Forest is the worst. The reason for the same can be that due to limited data, it adjusted to the training dataset which seems to be like a case of overfitting even after multiple repeated Cross Validations.

For True Negative rate, LDA, Logistic, Naïve Bayes and Tree perform similarly while Naïve Bayes and Tree outperform others by a margin of about 3 percent. Negative Predictive Value is the deciding factor when comparing Naïve Bayes and Tree model. Naïve Bayes with 35.89% clearly outperforms Tree with 26.41%. Even though Negative Predictive Value holds second importance to True Negative rate, in this case it is better for comparison as both models have same True Negative Rate.

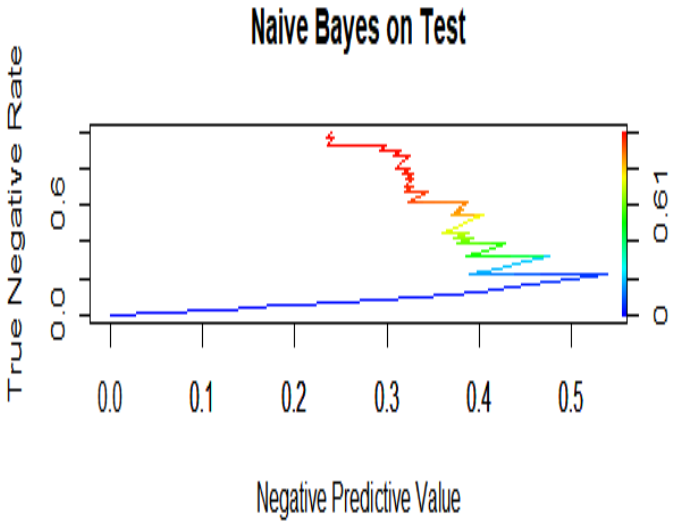
III.F. MODEL OUTPUT ON TEST DATASET

Naïve Bayes is selected as the best model based on the results from the selection set. The results on the test set are as follows:

		Actual	
		no	yes
Predicted	no	10	24
	yes	19	77

Confusion Matrix on Test Data with Naïve Bayes

True Negative Rate has further degraded. It seems to be the case that Naïve Bayes is not an optimal solution.

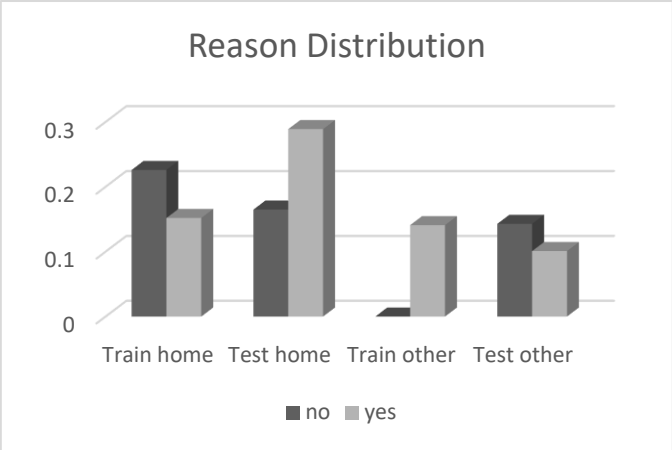


TNR vs NPV curve for Naïve Bayes

III.G. ANALYZING THE NAÏVE BAYES PERFORMANCE

Results on the selection and test sets prove that Naïve Bayes is not able to clearly differentiate the output category based on its learning from the training set. Even though it is performing reasonably well on selection set the performance on the test set is not as expected. Further analysis for this failure is done and the major findings are detailed here.

On comparing the distribution of training with the test data, some of the changes are having a negative effect. For the categorical predictors, major differences in distribution of 'Reason' and 'Traveltime' are observed.

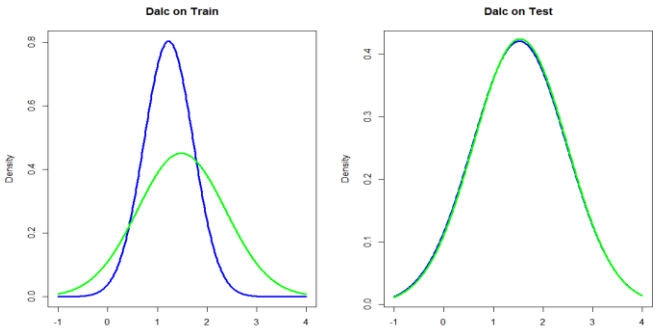


For 'Reason', the distribution of 'home' as category reduces for 'no', while at the same time increases for 'yes'. There are zero cases of 'other' with 'no' output in the training while it is relatively high in the test set. For the second predictor 'Traveltime' with 'no' as response the ratios are highly altered for category '1' and '2'.

	Train		Test	
Reason	home	other	home	other
no	0.226	0.000	0.165	0.143
yes	0.152	0.141	0.289	0.101
Traveltime	1	2	1	2
no	0.387	0.484	0.473	0.352
yes	0.535	0.424	0.614	0.299

Distribution for predictors in train and test set

While considering the continuous predictors, 3 predictors have major changes in distribution. 'goout' which has low value of mean for 'no' as compared to 'yes' in the training set has almost the same mean on the test set and hence is not relevant anymore.



Density plot for value of Dalc

'Dalc which has low value of mean and SD for 'no' as compared to 'yes' in the training set has the same mean and SD on the test set and hence not relevant anymore.

	Train		Test	
goout	Mean	SD	Mean	SD
no	2.677	1.137	3.165	1.204
yes	3.121	1.180	3.218	1.179
Dalc	Mean	SD	Mean	SD
no	1.226	0.497	1.527	0.947
yes	1.485	0.885	1.540	0.939
G2	Mean	SD	Mean	SD
no	12.032	2.496	10.110	3.472
yes	11.828	2.861	11.822	2.769

Mean and SD for predictors in train and test set

'G2' which has almost similar mean and SD on training set and is not a relevant predictor has very different mean and SD on the test set making it a relevant predictor.

IV. DISCUSSION

Based on the analysis none of the models implemented in this project were able to find a strong association between the predictors and response.

One reason could be that the data used was not informative enough of the general relation. The solution to this problem would be gathering more data.

Furthermore, a limited set of models have been considered, each of which have a set of assumptions and it is possible that those were the reason for not providing a good fit on the dataset. Other models could be considered based on further analysis of predictors.

Another enhancement could be consideration of statistical interactions between predictors or transformation of predictors.

It could be another point of interest to see if using ordinal predictors as continuous will have an impact on the output.

V. **ACRONYMS**

LDA	Linear Discriminant Analysis
NB	Naïve Bayes
TNR	True Negative Rate
TPR	True Positive Rate
NPV	Negative Predictive Value
RF	Random Forest
LASSO	Least Absolute Shrinkage and Selection operator
SD	Standard Deviation

VI. STATEMENT OF CONTRIBUTIONS

Work	Contributor
Problem statement selection	All members
Preliminary Analysis	All members
Project proposal	All members
Logistic Regression	Sai Charan Konanki
Linear Discriminant Analysis	Akshay Suresh
Naïve Bayes	Akshay Suresh
Tree based methods	Noul Singla
Selection Model Analysis	All members
Test Analysis	All members
Result Interpretation	All members
PPT Creation	All members
Project report creation	All members