Student ID : 210011371

A comparison of AdaboostM1 and Support Vector Machine classification on predicting subscription of term deposit

1. Description and Motivation of Problem:

We will be predicting the subscription of term deposit for a Portuguese banking institution by using binary classification methods support vector machine classification and Adaboost(M1) (using classification ensemble on tree as learner). We aim to access the model performance of these 2 models with results obtained by A. Lawi, A. A. Velayaty and Z. Zainuddin on Bank marketing for SVM[8][5] and Ensemble methods in Bank Marketing [6] .

2. Exploratory analysis

- Dataset: Bank Marketing dataset (bank-additional-full.csv) from UCI repository.[11]
- The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Sometimes the client is contacted more than one time in order to access if they subscribe bank term deposit or not [1].Original dataset chosen for the evaluation contains 41188 rows and 20 input features.
- There exist a class imbalance problem as shown in the figure 1.
- The following columns are dropped : duration, previous, month and day_of_week (as part of feature engineering) .Also columns nr.employed and euribor3m are dropped to avoid multicollinearity.
- The dataset was cleaned and outliers are removed and included the major criteria for accepting term deposit (rows and features reduced to 19016 and 15respectively). There is 6 numeric and 9 categorical features
- One hot encoding was done on column 'Y' (renamed as target) which check for term deposit subscription and 1 of the column after one hot encoding is removed to avoid multicollinearity

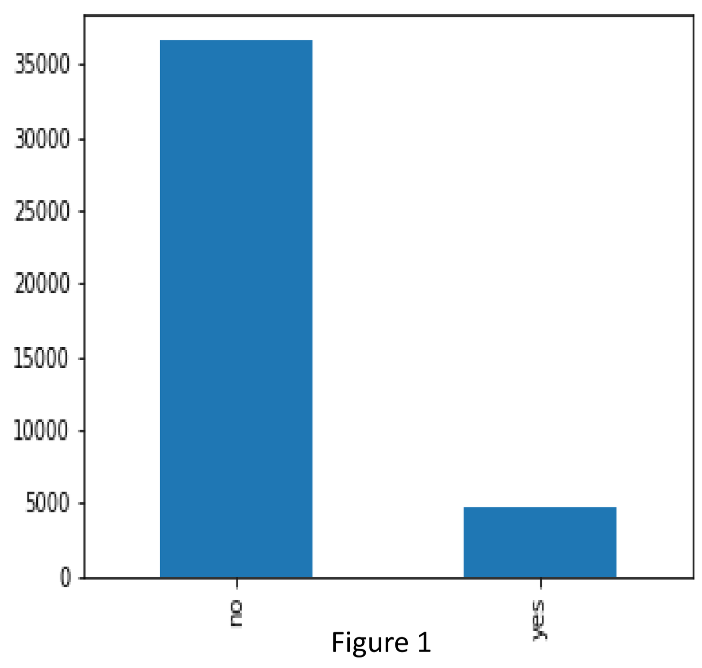
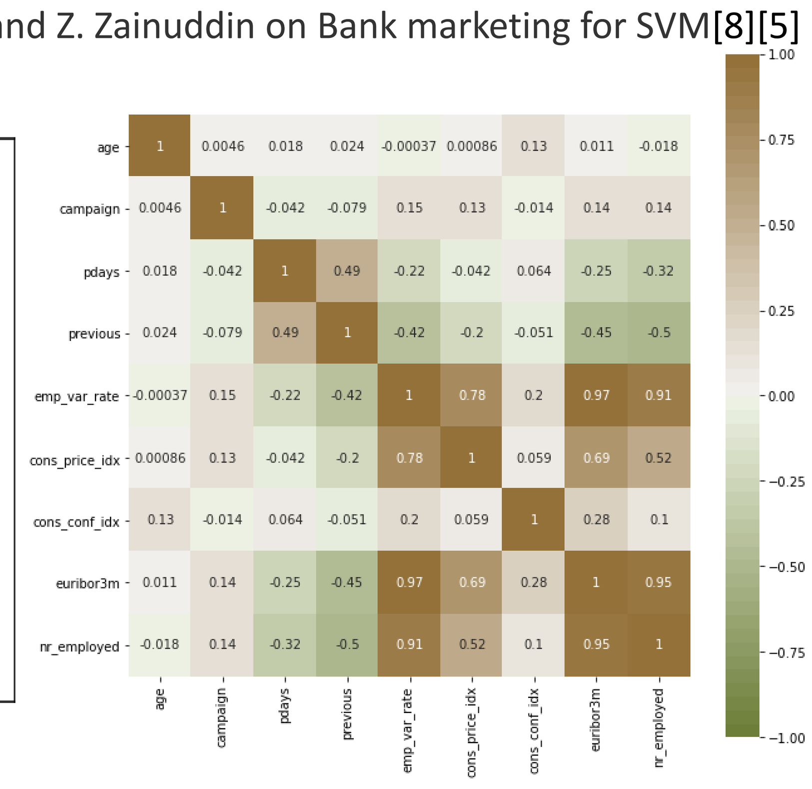


Figure 1


Correlation Matrix for all numeric values

3.1 Model 1 summary : AdaboostM1(tree as learner)

For handling the imbalanced data problem, one of the most frequently used classification learning is Ensemble. Freund and Schapire introduced AdaBoost (adaptive boosting) algorithm in 1995[1].It works on adjusting weight without the need of any priori knowledge of learner learning. The following algorithm of AdaBoost where developed later : AdaBoost.M1,AdaBoost.M2, Adaboost.R etc.[1] AdaBoost algorithm mainly focus on classification problem . Adaptive boosting named AdaBoostM1 is a boosting algorithm used for binary classification. Mostly AdaBoostM1 is used with shallow trees or decision stumps(default). [2].AdaBoost is considered to be a better performer when compared to models like Decision tee ,Naïve Bayes etc as it handles the imbalance in data. Hence its chosen as method for Ensemble learning with tree as weak leaner to improve the performance.

Advantages:

- Its widely used for imbalanced data. Since in our data the non- respondent greatly outnumber respondents in direct marketing[6] , it could be a good choice for improved prediction
- It's having high speed .Also its easy to programme as it have less input parameters.[1]
- Can be combined with any method to look for weak hypothesis without the need of any priori knowledge of WeakLearn[1]
- The accuracy of weak classifiers can be improved by using AdaBoost hence its better than Decision trees .

Disadvantage:

- Its performance depends on data and WeakLearn in a particular problem [1]
- When the base classifier is *too weak*, AdaBoost is not a good choice.*[3]*
- AdaBoost works well only on a quality dataset. Hence noisy data and outliers must be removed[1]

3.2 Model 2 summary: SVM

From 1990's onwards Support Vector Machine(SVM) is one of the most efficient machine learning algorithms, which is commonly used for pattern recognition [4]. SVM is applied to many pattern classification problems such as image recognition, speech recognition, text categorization, face detection, and faulty card detection.[4].

Advantages:

- It work well on binary classification problems and is expected to perform well with the prediction as the response variable is binary for the problem
- Its good in pattern classification as it performs well for classes when there is a clear margin of separation between them.
- It tends to be more effective in high dimensional spaces

Disadvantage:

- As there is huge imbalance in data SVM could perform poorly. SVM algorithm is sensitive to class imbalance problem[5]
- Since it works on high dimension , decision boundary plays an important role .It has the risk of neglecting the true positives in the case of imbalanced data if not considered properly.

4. Hypothesis

- I expect that AdboostM1 performs better using ensemble than SVM as it is not good for imbalanced data. Hence I assert that AdaboostM1 will have better AUC when compared to SVM and much better predictions as well
- I assert to have higher accuracy for SVM than the results obtained by A. Lawi, A. A. Velayaty and Z. Zainuddin.
- Ensemble methods in Bank Marketing [6] is using gradient boosting and AdaBoost SVM is used for analysis by A. Lawi, A. A. Velayaty and Z. Zainuddin instead of AdaboostM1. Hence the result could have variations as tree is used as weak learner

5. Methodology Description

1. Load data after preprocessing and perform a cvpartition with 70:30 for training and test respectively.
2. The test data set is not shown to the models.
3. Using tree as weak learners perform ensemble technique for method AdaboostM1.SVM uses kernel to transform the data and based on that it finds the optimal boundary. Since optimal hyperplane is influenced by scale of the input features, the data should be standardized and hence it is turned on.
4. Perform hyperparameter optimisation to improve the model performance and get better result. The models with lowest accuracy is chosen as the AUC curve is seen to improve when accuracy reduces (Table 1,Table2)
5. Calculate accuracy,f1_score,precision and recall for each model (Table 3 ,Table 6)
6. Compare the 2 models by plotting ROC curve.

6. Parameter selection :

AdaboostM1-,NumLearningCycles: 150, LearnRate: 0.05 ,Learner: templateTree (MaxNumSplit : 2000), Method :AdaBoostM1

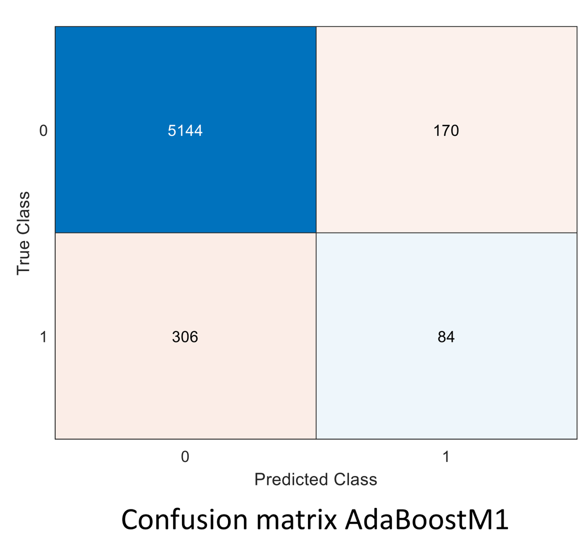
SVM - KernelFunction : gaussian, Standardize: on

Model	MaxNumSplits	NumsLearningCycle	LearnerRate	Accuracy	AUC
AdaBoostM1	500	200	0.1	92.4088	0.9933
	2000	150	0.05	90.2525	0.9957
	3000	150	0.05	90.2525	0.9956

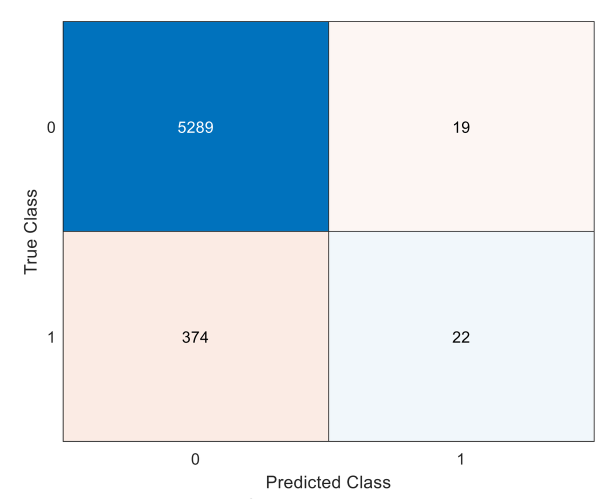
Table 1

Model	KernalFunction	Standardize	Accuracy	AUC
SVM	linear	on	93.4081	0.5791
	gaussian	on	93.7412	0.9043
	polynomial	on	92.6543	0.7989

Table 2



Confusion matrix AdaBoostM1



Confusion matrix SVM

7. Analysis and critical evaluation of experimental result:

- Accuracy for ordinary SVM is 93.11% where as in results obtained by A. Lawi, A. A. Velayaty and Z. Zainuddin it is 91.67%. This high accuracy could be due to high imbalance in data as well as the parameters chosen. The accuracy for AdaBoostM1 with tree(91.6550%) is lower than that of SVM (Refer Table 3). Also after 30 iterations AdaBoost SVM shows accuracy of 95.07% as in results obtained by A. Lawi, A. A. Velayaty and Z. Zainuddin[8]
- Accuracy is not a good measure for checking model performance for imbalanced data. Hence the model is evaluated based on AUC, F1 score, precision and recall. The models with low accuracy gives better AUC. The high AUC of AdaBoostM1 shows that it's a more accurate model when compared to SVM. The high F1 score and recall also convey that AdaBoostM1 dominates over SVM on model performance for imbalanced data.
- The TPR and FPR values of AdaBoostM1 is equally good as gradient boosting technique applied on ensemble methods in Bank Marketing[6]. Which in turn reflect that both are good models .
- The ROC curve plotted (fig 2) also shows that AdaBoostM1 has better performance when compared to SVM. SVM shows AUC of 0.9100 whereas AdaBoostM1 has 0.9954 which implies AdaBoostM1 performs better in distinguishing between the positive and negative classes and hence could give more accurate prediction when compared to SVM.(Table 7)

Best Model after Hyperparameter Tuning				
Model	Accuracy	Precision	Recall	F1- score
AdaBoostM1	91.6550	0.3307	0.2154	0.2609
SVM	93.1101	0.5366	0.0556	0.1007

Table 3

Before Hyperparameter Tuning				
Model	Accuracy	Precision	Recall	F1- score
AdaBoostM1	93.7938	0.6452	0.2051	0.3113
SVM	93.8464	0.7419	0.1742	0.2822

Table 6

Best Model	TPR	FPR
AdaBoostM1	0.2154	0.0083
SVM	0.0556	0.0045

Table 4

Best Model	AUC
AdaBoostM1	0.9954
SVM	0.9100

Table 7

8. Lesson learned and future work:

- AdaBoostM1 as an ensemble method is a good technique to improve model performance as it handles the imbalance in data. Also it can be combined with any method as a weak learner. From the results obtained by A. Lawi, A. A. Velayaty and Z. Zainuddin its evident that model performance depends on the weak learner chosen. Also choosing the right kernel function for SVM is very important for model performance.
- I will be applying SMOTE technique for handling imbalance data and would analyse on how it could perform on logistic regression, naïve bayes and random forest with and without balancing the data.
- Standard SVM doesn't perform well on imbalanced classification. In future, I will work on weighted SVM or cost-sensitive SVM to handle this .Also I will be trying to include more parameters on SVM to perform better analysis.

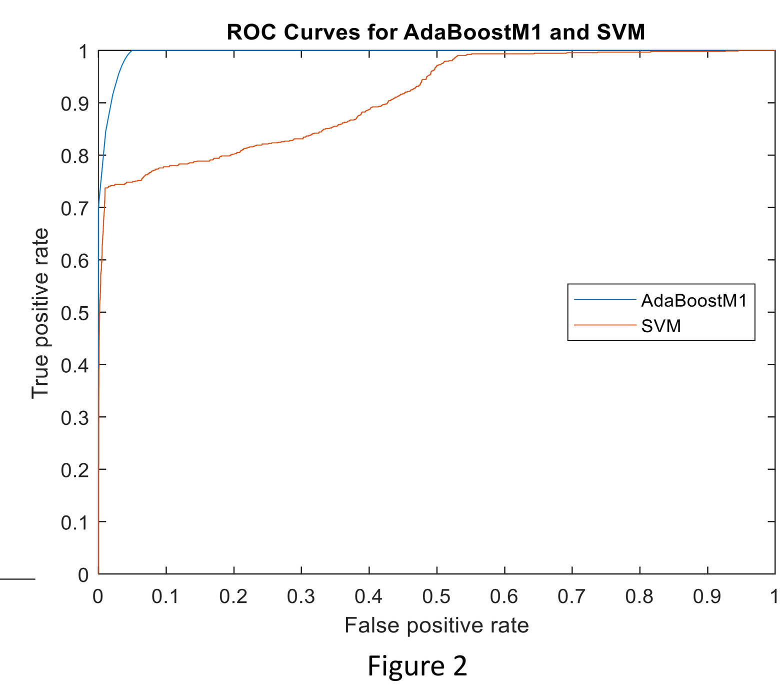


Figure 2

References:

[11]https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

[1] Chengsheng, Tu & Huacheng, Liu & Bing, Xu. (2017). AdaBoost typical Algorithm and its application research. MATEC Web of Conferences. 139. 00222. 10.1051/matecconf/201713900222.

[2] https://uk.mathworks.com/help/stats/ensemble-algorithms.html

[3] S. Karamizadeh, S. M. Abdullah, M. Halimi, J. Shayan and M. j. Rajabi, "Advantage and drawback of support vector machine functionality," *2014 International Conference on Computer, Communications, and Control Technology (I4CT)*, 2014, pp. 63-65, doi: 10.1109/I4CT.2014.6914146.

[4]V. K. Awasare and S. Gupta, "Classification of imbalanced datasets using partition method and support vector machine," *2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 2017, pp. 1-7, doi: 10.1109/ICECCT.2017.8117906.

[5] B. M. Shashidhara, S. Jain, V. D. Rao, N. Patil and G. S. Raghavendra, "Evaluation of Machine Learning Frameworks on Bank Marketing and Higgs Datasets," *2015 Second International Conference on Advances in Computing and Communication Engineering*, 2015, pp. 551-555, doi: 10.1109/ICACCE.2015.31.

[6] Pan, Youqin & Tang, Zaiyong. (2014). Ensemble methods in bank direct marketing. 11th International Conference on Service Systems and Service Management, ICSSSM 2014 - Proceeding. 1-5. 10.1109/ICSSSM.2014.6874056.

[7] https://www.sciencedirect.com/science/article/pii/S1569190X15000386

[8] A. Lawi, A. A. Velayaty and Z. Zainuddin, "On identifying potential direct marketing consumers using adaptive boosted support vector machine," *2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT)*, 2017, pp. 1-4, doi: 10.1109/CAIPT.2017.8320691.

[9] http://www.acadpubl.eu/jsi/2018-119-10/articles/10b/2.pdf

[10] https://link.springer.com/content/pdf/10.1007%2F3-540-36755-1_7.pdf