ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse287942020.pdf
https://doi.org/10.30534/ijatcse/2020/287942020

Improving classification accuracy :The KNN approach

Dr.Dhimant Ganatra[1] Dr.Dinesh Nilkant[2]
[1]CMS Business School, JAIN (Deemed-to-be University), Bangalore-560009, India
dr.dhimantganatra@cms.ac.in
[2]CMS Business School, JAIN (Deemed-to-be University), Bangalore-560009, India
dineshnilkant@cms.ac.in

ABSTRACT

The non-parametric tree-based methods are the go-to choice in a classification setting. They are simple and very useful for interpretation especially in contexts that require a business rule. They are however not very competitive in terms of predictive power when compared to other supervised learning approaches. Combing multiple trees in order to improve prediction accuracy is the next best option. Ensemble methods are powerful prediction models but they come at a cost. The loss of interpretability on account of aggregating trees may not be feasible in every decision making scenario. Also, depending on the business goal, the class-specific performance may be crucial. The true positive rate and positive predicted value may be more important than the overall accuracy.A non-parametric approach like K-Nearest Neighbours (KNN) can be superior when we have a complicated decision boundary,. Though simple, KNN often produces a classifier that is closer to the optimal Bayes classifier. For the B-School in question which needs to classify applicants into placeableor non-placeablebased on their past academic performance, a comparison of both the approaches is made to identify a superior performer on the positive class. A model's sensitivity is more crucial than the reduction in the overall error in the given scenario.

Key words: Decision Tree, K-Nearest Neighbours, Machine Learning, Sensitivity.

1. INTRODUCTION

1.1 The Classification Setting

A business rule has been developed by [1] for a B-School using which it was possible to divide applicants for admission into two classes - Placeable and Not Placeable. Owing to the impact placement has on a B-School's brand image, the management wanted to devise a strategy which can help differentiate the pool of applicants. For many Schools like the one in question, there is never a dearth of eligible candidates. The only question is how to ensure the right candidate is offered admission. The right here being a placeable candidate. The admission team was keen to use the power of analytics in arriving at this decision and also wanted to keep the model simple and interpretable. Given this context, using the classification tree algorithm, a model was developed by [1]. On account of the recent developments across the world owing to the Covid-19 pandemic, there has been a spurt in the number of applicants for admission. These numbers are comparable to levels last seen during the slowdown in 2008. This is because an MBA is seen as an investment during recession and slowdown [2]. According to the international non-profit organisation of business schools, the Graduate Management Admission Council (GMAC), which owns and conducts the standardised Graduate Management Aptitude Test (GMAT), MBA applications have always increased in recessions. Owing to the change in the current scenario, it was felt that there needs to a model which can improve upon the accuracy of the model currently being used without having to comprise on simplicity and interpretability. The objective of this research paper is to develop a model using the simple yet effective K-Nearest Neighbours (KNN) model. In the process, a comparison is made between the classification tree and KNN algorithm.

1.2 K-Nearest Neighbours

KNN is one of the simplest and accurate algorithms for classification and regression models. The algorithm was proposed in 1951 by [3].This was modified in 1967 by [4]. It is a non-parametric algorithm, meaning that it does not make any assumptions of the underlying data distribution. It is also a lazy algorithm, that is, it takes little or no training time because it memorises the training dataset instead of learning a discriminative function from the training set. The prediction step in KNN is expensive. For every prediction, KNN searches for the nearest neighbours in the entire training set. In spite of its slow characteristic, KNN is used extensively due its good characteristics of simplicity and reasonable accuracy. KNN predicts the class of a given test observation by identifying the K observations in the training data that are nearest to it. The test observation is then assigned to a class containing majority of its neighbours. The distance between the test observation and each of the training set observation is determined by an explicit distance measure the most common being the Euclidean distance. Selecting the value of K is the most critical problem and has a severe effect on the classifier obtained. The decision boundary is exceedingly flexible with K=1 and model has low bias but high variance. With flexibility, the training error rate declines, but the test error may not. As K increases, the method becomes less flexible [5].The KNN algorithm can be tested for different values of K using the cross-validation technique, thereby making it possible to choose the optimal level of flexibility. An important requirement when using the KNN classifier is to standardise the features such that they have a mean of zero and standard deviation of one. This is necessary because the class of a given test observation is predicted by identifying the observations that are nearest to it. In doing so, the scale of the feature matters. The scale effect caused by the

use of features with different measurement scales is removed by standardisation.

2. LITERATURE REVIEW

The decision tree and KNN algorithms have been used in many studies in the educational setup. Most of these studies are focused on data after a student has been admitted into a course of study. We did not come across a study which uses the classification algorithms to connect admission with placement. It is necessary that the model helps take decision before a student is offered a seat into the course.

The decision tree classification model for university admission system has been shown in [6]. Four attributes were used to build the model and the class attribute with two values: rejected and accepted was the response variable. To analyse the relationships between the measures of high school achievement and successful completion of students' first math and English courses in community college, decision tree method was used by [7].Reference [8] applied the decision tree algorithm on past performance data of engineering students to generate a model to predict their performance in an examination. The model had an accuracy of close to 60% and used sixteen predictor variables so as to predict their performance in the examination of first semester. Placement rules that the college can apply directly in their placement process were developed and validated. A new perspective to student retention has been provided by [9] using the classification trees. For policy makers at the university level, it is an important issue on account of the potential negative impact on the University's image and the career path of the student dropouts. Supervised learning algorithms have been used by [10] to identify the academic characteristics of students which enhance the probability of placement. The proposed CT-ANN algorithm achieved higher accuracy in predicting placement than other conventional techniques. To assist faculty and management in taking an informed decision about a student's performance, reference [11] concluded that the decision tree algorithm can be incorporated in the academic environment. Complex structures were easily illustrated by classification trees and random forests in [12] which otherwise would have taken many interaction terms to find using the common regression techniques. Simple linear regression gave the best accuracy among five algorithms compared in [13] to predict student's performance. Reference [14] applied the KNN algorithm among others to predict the performance of students in end semester university examinations. KNN in this case performed poorly as compared to others. Five classification algorithms were compared by [15] for predicting student's grades in a multiclass classification case. KNN was found to match the accuracy of other sophisticated classifiers.

3. METHODOLOGY

The objective here is to improve the accuracy especially of the positive class of the classifier model developed by [1]. The classifier model was built following the requirement of the institute to differentiate applicants into two categories – placeable and non-placeable. Since the cost of misclassifying a non-placeable category is high, it is felt that three is a further need to see the possibility of increasing the model's sensitivity but not at the cost of interpretability. R language and environment has been used for all the computing and graphics [16].

As per [1], 215 students who completed their MBA from the business school have been selected for the study. Here the response variable is a two-class categorical variable – Placement with labels as Placed and Not Placed. There are10 predictor variables.

The 75–25 technique is used [17] to split the data set. The crosstab in respect of the response variable is as shown in Table 1.

Table 1: Count of Train and Test dataset

	Not Placed	Placed	Total
Train data set	50	111	161
Test data set	17	37	54
Total	67	148	215

4. CLASSIFIER MODEL

4.1 Decision Tree Classifier

The algorithm to create the decision tree is provided by Recursive Partitioning and Regression Trees (rpart) library [18].

The classification tree for the train set is shown inFigure1 shows. Nodes that split to the right do not meet the criteria while the split towards the leftmeet the criteria. Every node is labelled by the predicted class, either Not Placed or Placed. The percentages have to be read from left to right, the probability of Not Placed is shown on the left.

From Figure1, the interpretation of node 7 of the tree is that if an applicant has a score of more than 56% in SSC and more than 66% in Degree, then there is more than 94% chance that the applicant is likely to be Placed and the support is 49%.

Table 2 shows the confusion matrix for the test dataset. The accuracy of classifying Not Placed (sensitivity) is 64.71%, while the accuracy of classifying Placed (specificity) is 91.89%. The overall accuracy is 83.33%. The positive and negative predicted values are 0.7857 (precision) and 0.8500 respectively. The classification accuracy of the test dataset is within 10% of the training dataset which provides evidence of the utility of the model [19].Given the problem on hand for the B-School, we need a higher accuracy in predicting positive class than the negative class. The F-Measure which combines both recall and precision is 0.7097. Values closer to 1.0 are considered the best. The overall worth of a classification treecan be understood from the Receiver Operating Characteristic curve (ROC curve). Figure 2 shows the ROC curve. The area under the curve (AUC) which indicates the proportion of concordance pairs in the data is 0.8959. A model with higher AUC is preferred.

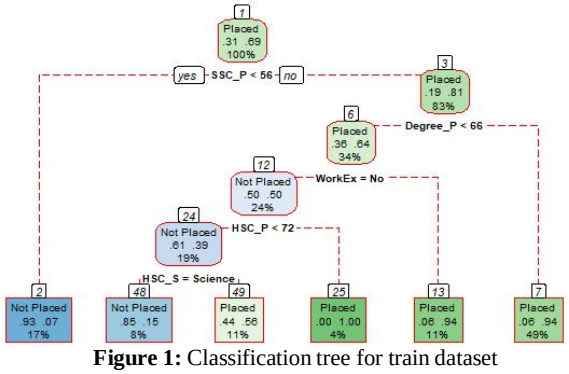

Figure 1: Classification tree for train dataset

Table 2: Confusion Matrix based on Classification Tree

Predicted	Actual		Overall %
	Not Placed	Placed	
Not Placed	11	3	
Placed	6	34	
% Correct	64.71	91.89	83.33

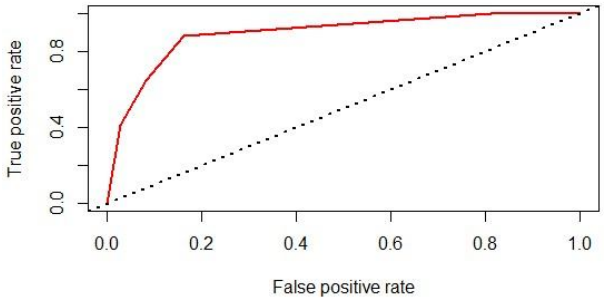

Figure 2: ROC Curve for classification tree

4.2 KNN Classifier

KNN is very sensitive to differences in the value range of predictor variables. It is advisable to rescale all the numeric variables. Although all numeric variables in the dataset correspond to percentages, we will rescale the numeric variables by standardizing them. We use the Shapiro-Wilk test to check for normality in order to decide how to standardise the numeric variables. For those variables which are normally distributed, we standardize using mean and standard deviation, and for all other variables, we have used median and inter-quartile range [20]. Also, KNN algorithm primarily works with numerical data. In our dataset we have categorical features and hence we have to transform them into numerical variables. Once the data has been pre-processed, we use the 75–25 technique [17] to split the dataset into train and test.

To build the KNN classifier we use the class package[21]. One of the inputs to be provided for the KNN algorithm is the number of neighbours, K to consider. The value of K will decide the flexibility of the model. For an arbitrarily chosen K=5, the model's accuracy on the test dataset is 0.8889 with a sensitivity of 0.7647 and specificity of 0.9459. The KNNoutperforms the decision tree classifier not just on the overall accuracy but on class-specific performance as well.

We can further attempt to optimise the results using the cross-validation technique with 13 different values of K. Figure 3 shows the results of the cross-validation plotting the model accuracy for different values of K.

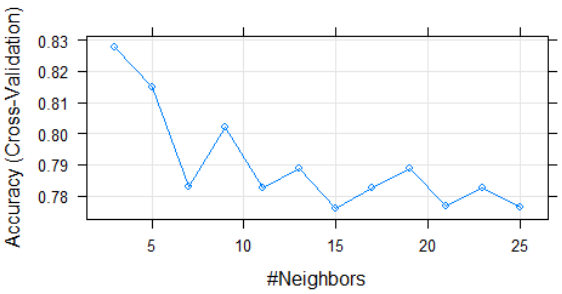

Figure 3: Accuracy v/s No. of Neighbors plot

Table 3: Confusion Matrix based on K=3 for KNN

Predicted	Actual		Overall %
	Not Placed	Placed	
Not Placed	14	4	
Placed	3	33	
% Correct	82.35	89.19	87.04

We build a model with the best value of K=3 and make predictions on the test set. Table 3 shows the confusion matrix. Compared to the model with K=5, we observe that though there is a marginal decrease in the overall accuracy, the sensitivity has improved from 0.7647 to 0.8235 which in fact is the objective of this research.

5. CONCLUSION

Trees fit the data nicely and are easy to interpret but they are plagued from high variance. The result that we get could be significantly different on account of a small change in the training data [22]. But since they provide the advantage of creating a business strategy based on the result and also since the trees can be visualized, they are preferred in many decision-making scenarios. With 83.33% accuracy, the classifier based on tree is an aid to the admission team of the B-School. There is always a high cost attached to misclassifying a non-placeable candidate as compared to losing out on a placeable candidate. In such a scenario, sensitivity is more important than overall accuracy. With a sensitivity of only 64.71%, the admission team is not very convinced of applying the business rule as generated by the decision tree. An alternative approach was provided by the KNN algorithm.

KNN works by directly measuring the distance between observations and inferring the class of test data from the class of its nearest neighbours.With a sensitivity of 82.35%, the approach has found favor with the admission team though the algorithm lacks interpretability as compared to the decision tree. The KNN approach has led to approximately 27% increase in the model sensitivity and close to 5% increase in that of overall accuracy. Though the increase in sensitivity has come at the cost of a marginal decrease in specificity, the KNN approach is more suited for the case at hand. The results here are in line to a similar study by [23].

Future work: Future studies can examine the use of Neural Networks or Naïve Bayes as learning machines.

REFERENCES

1. Ganatra, D.,& Nilkant, D. (2019).**A Business Rule for a B-School using Machine Learning**. *International Journal of Advanced Trends in Computer Science and Engineering*, vol 8, no. 6, pp. 3621-3627, Dec 2019. doi:10.30534/ijatcse/2019/145862019
2. Economic Times (2020, February 24). **It pays to do an MBA in a Slowdown**. https://economictimes.indiatimes.com/jobs/placements-for-the-class-of-2020-non-bluechip-b-schools-beat-slowdown-blues/articleshow/74234601.cms
3. Fix, E.,& Hodges, J.L. (1951). Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties. *Technical Report 4*, USAF School of Aviation Medicine, Randolph Field.

4. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, vol. 13, no. 1,pp. 21–27.
5. James, G., Witten,D., Hastie, T., & Tibshirani, R. **An Introduction to Statistical Learning with Applications in R**. Springer, 2017, ch.2.
6. Mashat, A. F., Fouad,M. M., Yu,P. S.,& Gharib, T. F. (2012).**A Decision Tree Classification Model for University Admission System**. *International Journal of Advanced Computer Science and Applications*, vol. 3(10), pp. 17-21.
7. Bahr, P. R. et al. (2019). **Improving Placement Accuracy in California's Community Colleges Using Multiple Measures of High School Achievement**. *Sage Journal*, vol. 47(2), pp. 178-211.
8. Kabra,R. R.,& Bichkar,R. S. (2011).**Performance Prediction of Engineering Students using Decision Trees**. *International Journal of Computer Applications*, vol. 36(11), pp. 8-12.
9. Yu,C. H.,DiGangi, S., Jannasch-Pennell, A.,& Kaprolet, C. (2010).**A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year**. *Journal of Data Science*, vol. 8, pp. 307-325.
10. Chakraborty,T., Chattopadhyay,S., & Chakraborty,A. K. (2018).**A novel hybridization of classification trees and artificial neural networks for selection of students in a business school**. *OPSEARCH*, Springer, vol. 55(2), pp. 434-446.
11. Musawenkosi, L. H., &Bhekisipho, T. (2017).**Development of the Academic Model to Predict Student Success at VUT-FSASEC Using Decision Trees**. *International Journal of Computer and Information Engineering*, vol. 11(11), pp. 1188-1191.
12. Mendez,G.,Buskirk,T., Lohr,S., &Haag,S. (2013).**Factors associated with persistence in Science and Engineering majors: An exploratory study using classification trees and random forests**. *Journal of Engineering Education*, vol. 97(1).
13. Dhankhar,A.,Solanki K., RatheeA., &Ashish. (2019) **Predicting Student's Performance by using Classification Methods**. *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8(4), pp. 1532-1536.
14. Anuradha, C., & Velmurugan, T. (2015). **A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance.***Indian Journal of Science and Technology*, vol 8(15).
15. Taruna, S.,& Pandey, M. (2014). **An empirical analysis of classification techniques for predicting academic performance**. *IEEE International Advance Computing Conference*, pp. 523-528.
16. **The R project for statistical computing**. https://www.r-project.org/
17. Forman, G., & Scholz, M. (2010)**Apples to apples in cross-validation studies: Pitfalls in classifier performance measurement**. *ACM SIGKDD Explorations*, vol.12(1), pp. 49–57.
18. **Package rpart**. https://cran.r-project.org/web/packages/rpart/rpart.pdf
19. Holden, J. E., Finch, W. H., &Kelley, K. (2011).**A Comparison of Two-Group Classification Methods, Educational and Psychological Measurement**. *Educational and Psychological Measurement*, vol71(5), pp. 870–901.
20. Brownlee, J. (2020). **How to Scale Data with outliers for Machine Learning.** *Machine Learning Mastery***.**
21. **Package class** https://cran.r-project.org/web/packages/class/class.pdf
22. Lander, J. P. **R for Everyone – Advanced analytics and graphics.** Pearson Education Inc, 2016, pp.310-312.
23. Solichin,A. (2019). **Comparison of Decision Tree, Naïve Bayes and K-Nearest Neighbours for Predicting Thesis Graduation**, *6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, Bandung, Indonesia.pp. 217-222. doi: 10.23919/EECSI48112.2019.8977081.