Comparison of prediction models for math achievement

Abstract

Data from two schools on math achievement was used to train a series of predictive models. Each model was tuned and error scores (MSE and RMSE) calculated. The results showed that most models were unable to predict math scores using demographic information more successfully than baseline comparators. The most accurate model used support vector machine regression to predict scores. Suggestions for further improvement in predicting student scores are given.

Introduction

The intention of this project was to predict the final math score of students using a range of demographic information about each student. This would enable predictions to be run on enrolment, and for the school to provide particular assistance to those student with very low predicted final scores. The data for this project came from a student performance data set collected in two Portuguese schools and avaliable online through the UCI Machine Learning Repository.

The models chosen for this task were:

- Support Vector Machine regression (SVM), a model which maps the data in a higher dimensional space and attempts to find the flattest function which describes the data, subject to minimisation of the error.
- Random Forest regression, a model which calculated the average prediction of a number of decision trees trained on the data.
- Polynomial regression, which models the relationship between the target (predicted) varialbe and the predictors using a polynomial expression.
- Artificial Neural Network (ANN), a model using a series of nodes in layers to process and pass forward predicatons with the aim of minimusing the difference between the final prediction and the target (predicted) variable.

Each model will be fitted and tuned on the same avaliable data, and the most accuracte and effective model will be selected for use going forward.

Materials and methods¶

The dataset to be used was already in good order, and although it was checked for missing or incoherent values, none were found. The data was checked for outliers using boxplot visualisation and the calculation of z-scores. A small number of outliers was removed. Many of the variables were converted to integers - the binary variables which were only 'yes' or 'no' converet to '0' or '1'. For the more complex variables such as the mother and father's occupation, dummy variables were coded for each level and set of '0' if the level was true for each observation.

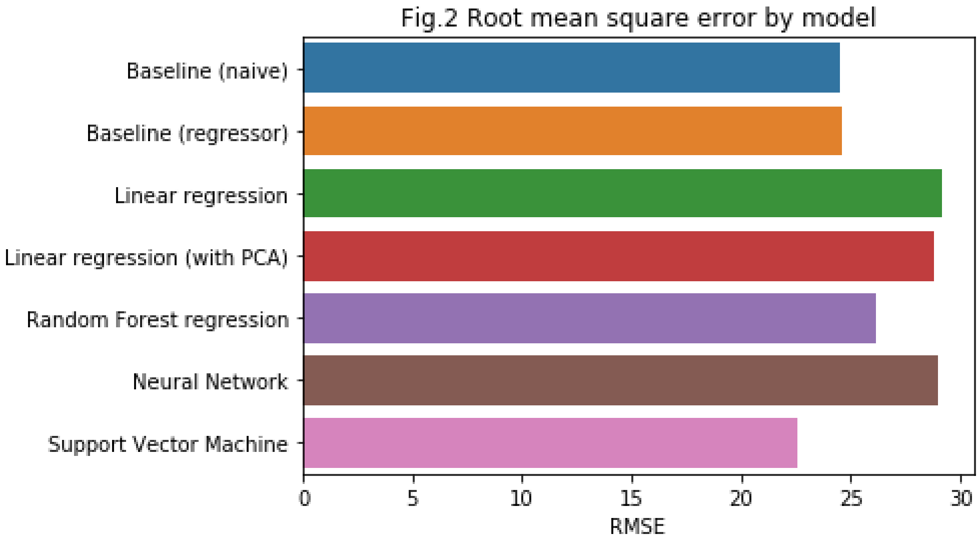
To determine which variables should be included in the model, all variables were tested for correlation and only those which have an absolute correlation of greater than 0.1 with any one of the math score variables were kept. This simplified the dataset, and only 15 of the intital possible 32 variables were kept.

The data was split for training and testing using a 5-fold cross-validation, where the dataset was divided into 5 parts and in each use the model was trained on 4 parts and tested on the third. A methodlogy of prediction was used so each prediction came from when the observation was in the sample set. Similar cross validation was used when tuning the model's hyper parameters. Each model was given an appropriate ranges of hyperparamters, which were iterated over in a grid. Each combination was tried to see which rpoduced the most accurate model. In many cases the tuning did not improve the model, likely because the inital hyperparametres chosen were closest to optimum.

Two primary measures of accuracy were used: mean square error and root mean square error (RMSE). Both are directionally neutral ways of measuring how far the predicited value for an observation was from the 'true' value of that observation. Ultimatly the model with the lowest root mean square error was chosen as the most desireable. In selecteding the models two kinds of baseline RMSE were used as comparison - a naive RMSE calculated using a model that always predicts the mean value of the target variable, and a dummy regressor's predictions.

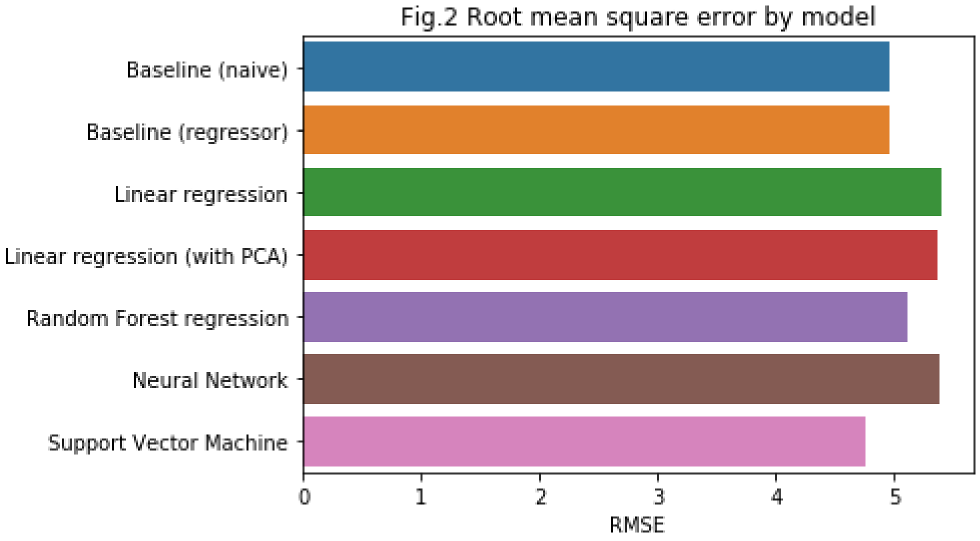
Results and discussion¶

The results showed a lack of success, with most models not being more accurate the the baseline measures that were computed. The mean square errors for the most accurate version of each model are presented in figure 1 below.



As can be seen, the support vector machine regression is the only model that have a lower error than the baselines. The difference in error rates between the linear regression with and without principal component analysis (PCA) indicates that the PCA was successful in improving model performance, though not enough to make it the most accurat model. In the analysis here the number of components was chosen to account for between 80% and 90% of variability. Ten components were derived and used in the model training and testing. Further work may find success in using different numbers of components or to run the PCA on the inital whole dataset.

Figure 2 below compares the RMSE of the various models.



Like the results for MSE, the RSME in figure 2 shows that the only model that was more accurate than the baseline was the suport vector machine regression. One reason for this may be that SVM is good at handling outliers in the data. Although the data was processed to remove extreme outliers, there was a number of student who had extremely low scores. The predictions across models often showed that it was more difficult to predict these very low scores. Additionally, SVM works well on medium or small dataset, whereas models such as Random Forest regression do better on larger datasets. Similarly, neural networks generaly function better on large volumes of data, and in this case the neural model has the highest error.

Conclusion

Overall, the models chosen did not have much success in predicting the final math score of a student better than the baseline. The one exception to this was the support vector machine regression model, which had the smallest error. This model should be used for any predictions neccessary.

Futher work could explore more deeply the reasons for the lack of success and undertake a comparison using alternative models. Additionally, more data on which to train the models is likely to assist in obtaining an accurate prediction. Data could be gathered from more scchols internationally, which would also give the results more reliability cross-culturally.