Baixi Guo
11/12/2022

MATH 180 Final Project Report
(Student Performance)

Introduction:

Students' performance in high school is important to academic success and can determine if they can graduate on time to start their careers or pursue higher education. However, there are many factors that may have influences on student performance, such as age, gender, numbers of absences or even some of the socio-economic factors. Socio-economic status can determine whether a person or family have sufficient resources that might impact their quality of life or health, it gives us incentive to investigate the influence of socio-economic factors by using various machine learning models. Throughout the project, we were interested in what and how socio-economic factors would have effects on students' performance in math class. Specifically, we wanted to investigate how many predictors should be included to minimize the test errors, how including socio-economic factors would affect the model accuracy and test errors, and what models have the best prediction.

We used the *Student Performance* dataset from Kaggle which was published in paper "*Using Data Mining to Predict Secondary School Student Performance*.". The dataset was about approaching student achievement in secondary education of two Portuguese schools. Based on the description of the dataset, "The data was collected using school reports and questionnaires, and the data attributes include student grades, demographic, social and school related features. The target attribute G3 (final grade) has a strong correlation with attributes G1 (first period grade) and G2 (second period grade)." (Chauhan, 2022). The dataset contained 30 predictors, in which only 2 were quantitative and 28 were qualitative predictors, and 3 outcomes regarding the grade from first, second and final period which were continuous data. There were 395 rows in the dataset, which contained 13035 total data points. In addition, we have chosen some of the predictors as our socio-economic predictors, they were: "Pstatus" (parental status), "Fedu" (father education), "Medu" (mother education), "Fjob" (father job), "Mjob" (mother job), "famsup" (family support), "paid" (extra paid classes) and "internet"(internet access). Eventually, this dataset had complete data and could be used to be analyzed by multiple machine learning models.

We applied a forward subset selection model to perform feature selection which would yield the least test errors while improving the model interpretability and reducing the variance with negligible increase in bias. We decided to use supervised learning methods with given predictors and outcomes. Moreover, the problem would belong to the regression problem because of quantitative outcomes. Therefore, one parametric model such as linear regression and

non-parametric methods such as regression spline, decision tree, bagging tree, random forest and multilayer perceptron would be considered to predict the students' final grade outcome based on the 30 given predictors. In order to evaluate the models' performance, we used mean squared errors (MSE) to decide which models perform better as MSE was adequate for the regression problem, cross-validation would be used to find optimal parameters and hypothesis testing would be a tool for linear regression to discover which predictors were statistically significant.

According to the result from RMarkdown file, we realized that most of the socio-economic factors do not have significant effects on the prediction performance, whereas the most influential predictors were "failures' ' and "absences", which were not part of the socio-economic factor list. Through the subset selection, we determined that the more predictors were included, the less cross-validation test MSE that the model would have. This applied with the dataset consists of only socio-economic predictors and the dataset considering all factors. We also utilized models such as linear regression, spline, trees, forest and neural network to make predictions about the "G3" final grade outcome. The results came out that the random forest yielded the lowest test MSE while spline yielded the largest test MSE. We were surprised that the neural network did not even perform better than the linear regression model, which might be due to the size of our dataset.

Result:

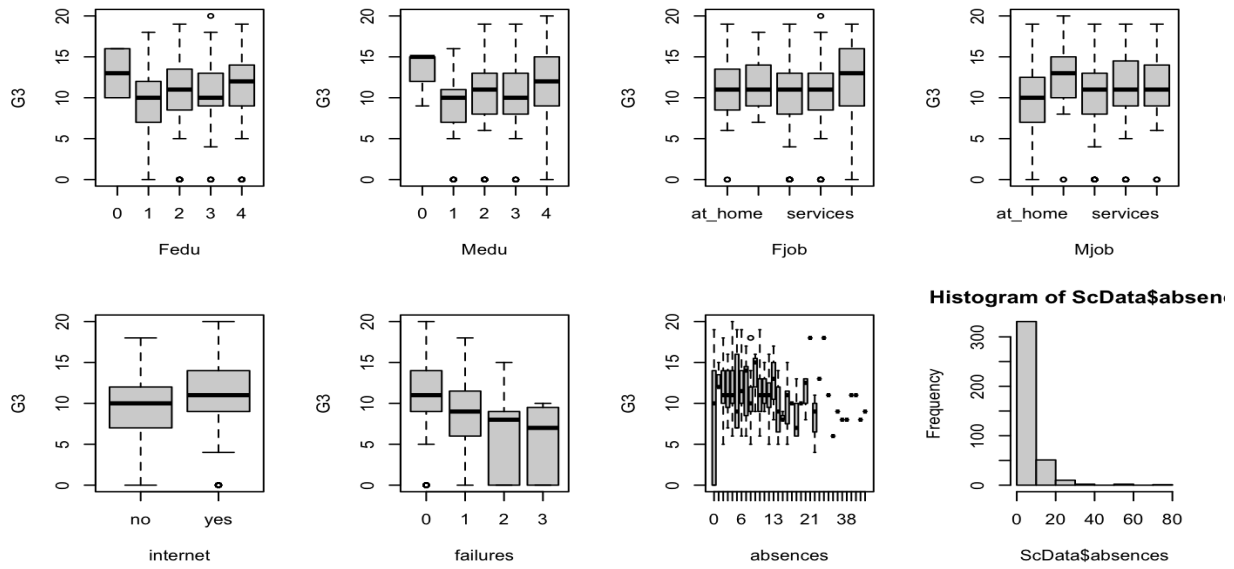
Preprocessing:

We have preprocessed our dataset in order to perform any further analysis. Our preprocessing consists of three parts: data cleaning, data reduction and data transformation.

In data cleaning, we have checked the existence of empty values in the dataset, which had been verified by Kaggle and us that there were 100% valid values, in other word, there was zero null value. However, we found data abnormalities of the predictors "Medu" and "Fedu" from Figure 1 because they only had one or two data points for a specific label which might affect the model prediction and predictors'' statistical significance in the hypothesis testing. As a result, we removed the rows which contained the specific label in "Medu" or "Fedu" predictors. In addition to the visualization which demonstrated the relationship between predictors and outcome, we found out that socio-economic predictors had different influences on the "G3" (final grade) outcome with different labels.

Relationship of Predictors and Outcome
FIGURE 1 (Section 1.1 in RMarkdown File): *Relationship between predictors (majorly socio-economic factors) and "G3" (final grade) outcome were shown in box and distribution plot. There were obvious differences in final grade which were influenced by predictors with*

different labels. For the "Fedu" and "Medu" predictor, the data in label "0" had been removed due to the data abnormalities.



In data reduction, we used different subset selection models from "*Leaps*" library and self-defined a manual forward subset selection model based on pseudo-code of forward subset selection algorithm introduced in "*An Introduction To Statistical Learning*" (James, Witten et al, 2013); therefore, we could achieve the goals of improving the model interpretability while preventing it from model overfit. The reason for defining our own version of subset selection model was that we ended up realizing that all built-in subset selection and shrinkage regression models from "*Leaps*" library did not support qualitative predictors according to "*Modern Applied Statistics with S Plus*", ""Leaps" library is not adequate for qualitative data" (B.D. Ripley, 2002, P. 176). Another reason was that the linear model would treat each label of a predicator as its individual predictors and would select a specific label as its best predictors during the feature selection process. Self-defined forward subset selection fixed this issue by not treating each label as its individual predictor like built-in methods did. The algorithm of the method was outlined below:

1. Started with an empty model.
2. Initialized formula list, which used to store the formula for best i-th predictor model.
3. For each i-th predictor model, chose the one that yielded the smallest RSS and stored the formula in the corresponding index.
    a. For the (i+1) predictor model, use step 3 but tested on each predictor based on the formula from the i-predictor model.
4. Repeated step 3 until all the i-th predictor models selected its best models.
5. Output the formula list

The example of the formula list's output from the forward subset selection model was shown below in Figure 2 where the formula list elements were shown from left to right.

Self-Defined Forward Subset Selection's Output

FIGURE 2 (Section 2.2 in RMarkdown File): *Output from self-defined forward subset selection model which showed the best model of each number of predictors.*

```
[1] "G3 ~+Medu"                                    "G3 ~+Medu+paid"                                    "G3 ~+Medu+paid+Mjob"
[4] "G3 ~+Medu+paid+Mjob+famsup"                   "G3 ~+Medu+paid+Mjob+famsup+Pstatus"                "G3 ~+Medu+paid+Mjob+famsup+Pstatus+Fjob"
[7] "G3 ~+Medu+paid+Mjob+famsup+Pstatus+Fjob+Fedu" "G3 ~+Medu+paid+Mjob+famsup+Pstatus+Fjob+Fedu+internet"
```

In order to evaluate which i-th predictor model had the best prediction, we performed the 10-folds cross validation. The results of cross-validation on self-defined forward subset selection models using all predictors (left) and only considering socio-economic predictors (right) were shown below in Figure 2. The trend in the graphs provided an answer to our investigation that "how many predictors should be included to minimize the test errors". Models tended to have less cross-validation test MSE if it included more predictors. For the left graph, the cross-validation error difference between the best and worst models was about 5 and there was a significant drop in MSE when 5 predictors were included. For the right graph, the difference between best and worst model was about 1 and there was a significant decrease in MSE when 4 predictors were included. The MSE of the best model from the left graph is smaller than the right one because more important predictors were introduced and they were not part of the socio-economic predictor list. In addition, the difference of cross-validation error on the right graphs is quite small compared to the left one, which implies, the answer we are looking for about how socio-economic factors affect the model and error, that more socio-economic factors included did not have substantial improvement and the the majority of the socio-economic factors had small influence on the model.

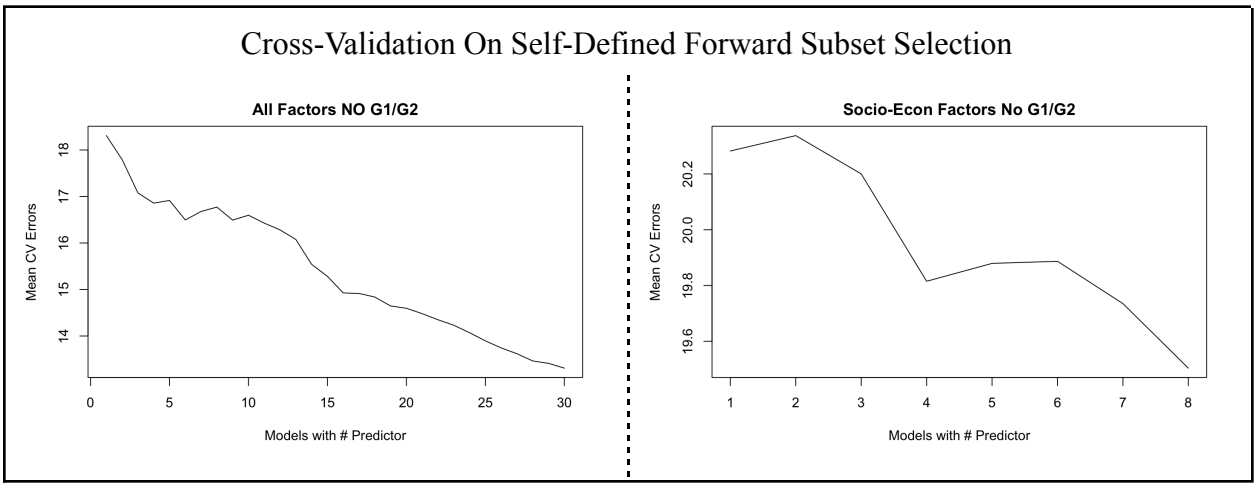


FIGURE 3 (Section 2.2 in RMarkdown File): *The cross validation was used on self-defined forward subset selection mode with dataset with all predictors (left) and dataset with only socio-economic predictors (right). Both models showed that the more predictors were included, the less cross-validation MSE the models had.*

In data transformation, we used both automatic and manual one-hot encoding methods on qualitative data because model prediction would be inaccurate as it was meaningless that the categorical data with labels were treated as some continuous numbers (i.e. taking derivative on the labels). By specifying the "stringAsFactor" parameter as "True" in "read.csv()" when loading the dataset to the script, the one-hot encoding would be applied automatically by R to create dummy variables for labels other than the baseline for each qualitative predictors. However, that only applies to character data type while still treating some numeric data type as integer. This needs to be handled manually because the integer values of the numeric predictors such as "Medu" and "Fedu" represented some categories but it was not represented in character type.

In the final part of the data preprocessing, the dataset was extracted and splitted into two different dataset. One included all predictors while another one included only socio-economic predictors, but they both included the "G3" response.

Model Prediction:

We considered using multiple machine learning and one deep learning model to discover the model that had the best prediction. They were linear regression, regression spline, decision tree, bagging tree, random forest and multilayer perceptron (MLP).

Linear regression was our model to come into our mind as this was the most common model to solve regression problems. We created three linear regression models with different sets of predictors or interaction terms between predictors, and compared them in terms of train and test MSE. The sets of predictors used in each model had been demonstrated in Table 4. The first set of predictors were only from the socio-economic predictor list, the second set of predictors were chosen among all the predictors while the third set was a set of predictors that caused the model to overfit. The predictors were chosen in a way that was based on the observations from the predictors' statistical significance in model fit summary, while the interaction terms between predictors were created based on the observation from "pairs()", which produced a matrix of scatterplots between predictors. The set of predictors had been tested so that its model yielded a currently smallest test MSE.

Set Number	Predictors
0	Pstatus*Mjob, Medu*Fedu, Medu*Mjob, famsup

1	age, health, failures, absences*romantic, sex, studytime, Medu*Mjob
2	age, health, failures, schoolsup, romantic, school, higher, sex*health, health*failures, studytime*failures, Dalc*failures, romantic*absences higher*absences, address*school, higher*studytime, higher*age

TABLE 4 (Section 3.1 in RMarkdown File): *There were three linear regression models with three sets of different predictors or terms of predictors' interactions. First one only had socio-economic predictors, second one considered all the predictors and third one were the predictor that caused the model to overfit*

The comparison was shown in Figure 5, which demonstrated that the first linear regression had higher train MSE than the second one whereas the third one had a much higher amount of errors compared to the prior one. However, model 2 had the lowest, model 1 had the middle and model 0 had the highest train MSE. That's why the model 2 was called "overfit model". Through the bias-variance tradeoff on all the linear regression models, we chose the second one as the optimal model. In addition, model 0, which contains only socio-economic predictors, had higher bias and variance compared to model 1, it implied that socio-economic predictors had smaller influence on predicting the outcomes.

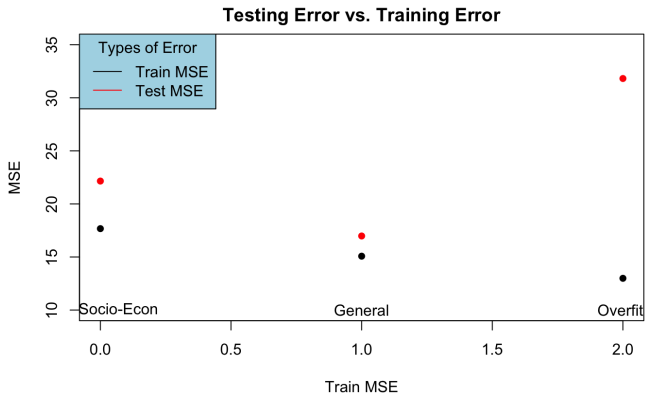


FIGURE 5 (Section 3.1 in RMarkdown File): *We compared three linear regression models with different sets of predictors specified in Table 4. Model 2 had the highest test MSE and smallest train MSE, model 0 had the highest train MSE. However, model 1 had relatively low train and test MSE among all models.*

Regression spline is considered because splines could solve regression problems and capture nonlinearities with more degrees of freedom. Cross-validation was used to prevent the model from overfitting when it became very flexible. Through the ten-folds cross-validation, the optimal degrees of freedom of the model would be selected from Figure 6. In the graph, when degree of freedom increased up to 5, the model yielded the lowest cross-validation test errors.

Additional Observation we made was that when degrees of freedom increased to 4, there seemed to be a substantial decrease in cross-validation error. However, the scale of the y-axis ranging from 20.00 to 20.30 implies that the improvement was significant after choosing an optimal degree of freedom. Finally, one limitation might exist in this model was that it only accepted quantitative predictors whereas we only had "absences" and "age" as quantitative factors. Therefore, there might be other useful qualitative predictors that could not be included in the regression spline due to the restriction and that the model might under-perform.

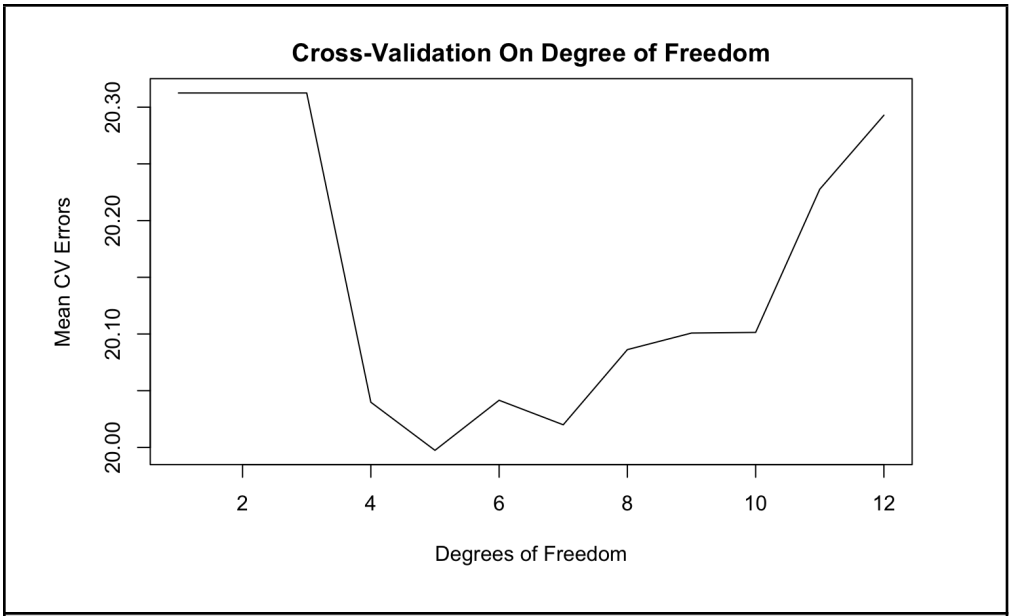


FIGURE 6 (Section 3.2 in RMarkdown File) : *The cross-validation was performed on a regression spline to select the optimal degrees of freedom that yielded the lowest cross-validation test error. Based on our observation, cross-validation error is the lowest when degrees of freedom = 5.*

Regression decision tree was the third model we considered as there was restriction on spline and it could be hard to interpret with high degrees of freedom. Another cross-validation was used on pruning the tree to certain amounts to ensure the model would not overfit and pruning improved the model's interpretability as size of tree decreased. Figure 7 demonstrated that through the cross validation, when size of tree equaled 3, the deviance reached its lowest and started to climb up when size of tree increased to about 8. Additionally, the best model and worst model had substantial differences in deviance which implied that pruning the tree was effective to improve the model's performance.

Bagging trees was also considered, because according to the lecture slide, "it reduces the variance and is more robust than the decision tree. However, it does not fix the problem of high bias. Random Forest is considered to reduce the uncorrelated errors." (Rube, 2022, P.13-18). We

adjusted the values of parameter "mtry" in bagging tree and random forest, which represented "number of variables to randomly sample as candidates at each split" (AFIT Data Science Lab R Programming Guide). We set "mtry" = 30 for bagging trees based on the explanation in "*An Introduction To Statistical Learning*" that "For instance, if a random forest is built using m = p [In our project, p = 30], then this amounts simply to bagging."(James, Witten et al, 2013, P. 353), and based on the lecture, "mtry" = 30/3 = 10 for random forest according to the lecture 18, "Typically $m = p^{\frac{1}{3}}$ (classification) or $m = \frac{1}{3}p$" (Rube, 2022, P. 18).

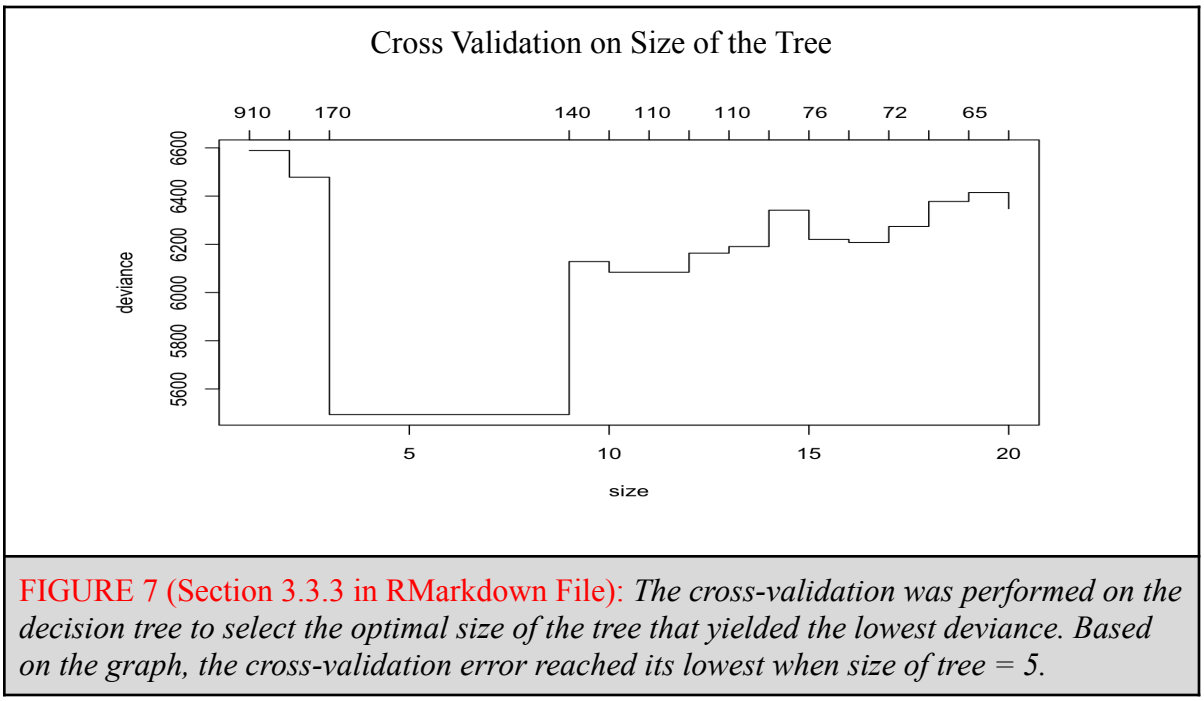
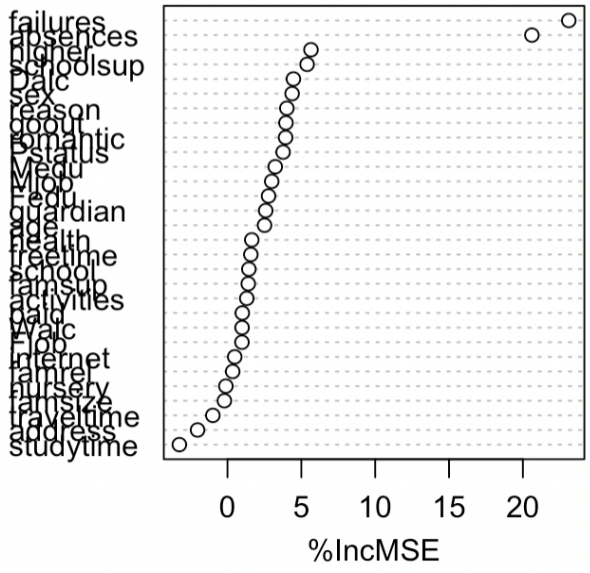


FIGURE 7 (Section 3.3.3 in RMarkdown File): *The cross-validation was performed on the decision tree to select the optimal size of the tree that yielded the lowest deviance. Based on the graph, the cross-validation error reached its lowest when size of tree = 5.*

Finally, we used "importance()" in the "randomForest" library to observe which predictors were important in the random forest model. In the next page, Figure 8 showed that two most important predictors were "failures" and "absences" whereas other predictors are not that useful for prediction. This plot also implied that none of the socio-economic factors had substantial influence on random forest models.

Last model we considered was multilayer perceptron (MLP) because we wanted to try a neural network model to observe if there was any improvement on prediction and how well the neural network can capture the nonlinearity if it existed. We performed hyper-parameter tuning on several parameters, which were layer dropout rates, batch sizes, epochs, neuron units and hidden layers' structure to yield an optimal model through ten fold cross validation while keeping other parameters constant. In the next page, Table 9 showed the result of cross validation, which specified what parameters were used, what lists of values were used for tuning, what was the resulting optimal value and how did it compare to the original parameter value. We observed that

the parameters with difference in values were “batch size”, “epoch” and “neuron units”. We compared the original and tuned model and plotted the difference on the train and test MSE between them in Figure 10. We observed that the tuned model yielded slightly less test MSE than the original model, which implied that the hyperparameter tuning did not have an obvious effect in the MLP model prediction.



Variable Importance For Random Forest

FIGURE 8 (Section 3.3.6 in RMarkdown File): *The plot shows which variables were important to make predictions. It turned out that “failures” and “absences” were the two most important predictors among all predictors*

Hyperparameter Tuning for MLP

Parameter Names	List of Value For Tuning	Optimal Value	Original Value
Dropout Rate	0.2, 0.4, 0.6, 0.8	0.4	0.4
Batch Size	8, 16, 32, 64, 128, 189 (All Rows)	64	32
Epoch	10, 50, 100, 150, 200	100	50
Neuron Unit	25, 50, 100,150, 200	150	50
Layout of Hidden Layer	- Original: 2 Dense & 1 Dropout layers - New A: 3 Dense & 2 Dropout layers - New B: 4 Dense & 1 Dropout layers - New C: 4 Dense layers	Original	Original

TABLE 9 (Section 3.4.3 - 3.4.7 in RMarkdown File): *The table demonstrated what parameters were tuned (1st column), what values were used to tune in cross validation (2nd column), what values were the optimal values (3rd column) and what values were the original values (4th column). If original and optimal values were different, values were highlighted in blue.*

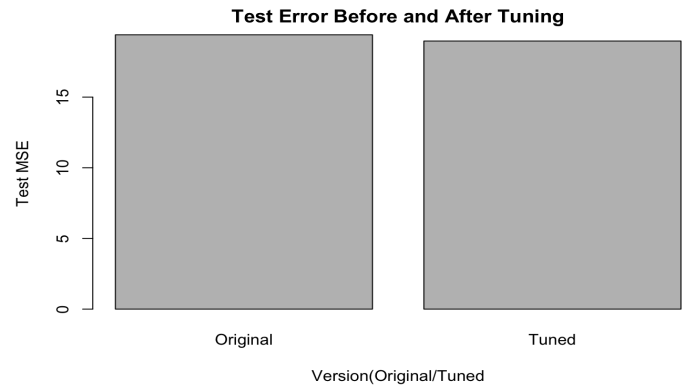


FIGURE 10 (Section 3.4.2 in RMarkdown File): *The plot shows the comparison between the model with and without the hyperparameter tuning. We observed that there is not much difference between the two models.*

In summary to the model prediction, we had compared all the models' test MSE with the same train and test data split. According to Figure 11, random forest performed the best while regression spline performed the worst. Surprisingly, MLP was the second worst model among all the models as we would expect deep learning models to have excellent performance. Since deep neural networks require large amounts of data to have accurate predictions and we only had less than 200 rows of data, this might be the reason that MLP was at its disadvantage.

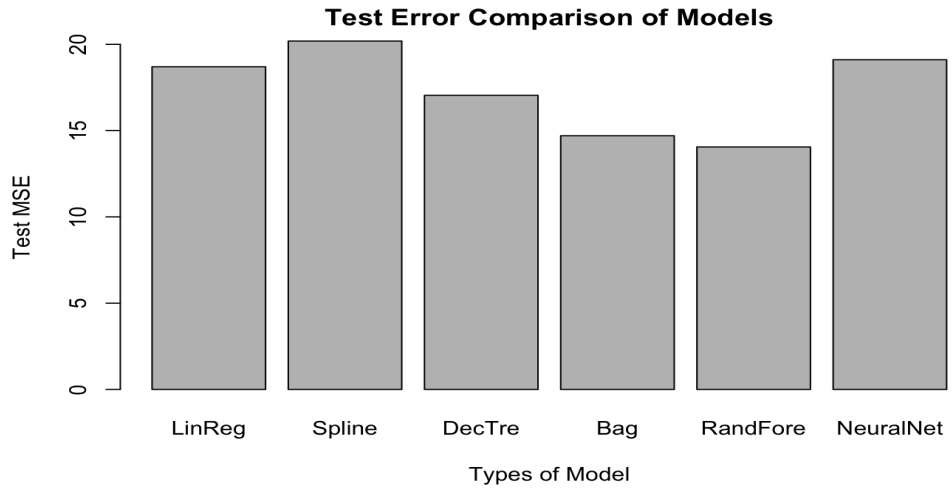


FIGURE 11 (Section 3.5 in RMarkdown File): *The plot showed the comparison of all the models we considered, which "LinReg" referred to linear regression, "Spline" referred to regression spline, "DecTre" referred to pruned decision tree,*

"Bag" referred to bagging trees, "RandFore" referred to random forest and "NeuralNet" referred to tuned MLP. Random forest yielded the smallest test MSE while regression spline yielded the largest test MSE.

Conclusion:

We were answering our questions throughout our analysis. In the plot that showed the relationship of predictors and outcome in the preprocessing section, we could observe that there were different influences on the outcomes with different labels from each socio-economic predictor. However, influence to outcome was not obvious when more socio-economic predictors were included in self-defined forward subset selection, linear regression models with only socio-economic predictors tended to have higher bias, and none of the important variables were socio-economic factors in random forest. Therefore, we could conclude that socio-economic problems had little effects on model prediction and errors. To answer how many predictors were included in the model to minimize the test MSE, we utilized forward subset selection and realized that if we included all predictors, the model would have the lowest test MSE. However, the linear regression model that had overfit issues had much higher test MSE compared to others, so that these two findings contradicted each other. We doubted that some predictors were highly correlated and hidden variables that were useful to the model prediction were not present in the dataset. Therefore, the predictors we had were not enough. Finally, we tried out multiple models and realized that random forest performed the best whereas the regression spline performed the worst. In addition, there were limitations and sources of noise in the project. The sources of noise could come from the data collection, in which they were collected through questionnaires. Therefore, questionnaire participants might give falsified information due to incentive. Limitation could be the fact that LASSO and spline did not support qualitative data and the dataset didn't have a large amount of data for MLP to perform well. In the future, we would consider using Group LASSO and Generative Additive Model (GAM) to overcome the restrictions of not supporting qualitative data and discover how the new models could have influence on the feature selection and mode prediction respectively.

Reference

James, G., Witten, D., Hastie, T. & Tibshirami, R. (2013, June 24). *ISLR2: Introduction to statistical learning, Second edition*. Retrieved December 9, 2022, from https://mran.microsoft.com/web/packages/ISLR2/ISLR2.pdf

Chauhan, A. (2022, October 7). *Student performance*. Kaggle. Retrieved December 9, 2022, from https://www.kaggle.com/datasets/whenamancodes/student-performance

Venables, W.N., & Ripley, B.D.(2002, March 15). *Modern Applied Statistics With S.* Retrieved December 9, 2022, from http://staff.ustc.edu.cn/~houbo/course/Modern%20Applied%20Statistics%20with%20Splus%20(Fourth%20edition).pdf

Rube, T. (2022). *Bagging, Random Forest & Boosting.* Retrieved December 9, 2022, from https://catcourses.ucmerced.edu/courses/25258/files/folder/Lecture%20Slides?preview=5449240,

AFIT Data Science Lab R Programming Guide. *Random Forests*. Retrieved December 9, 2022 from https://afit-r.github.io/random_forests

Cortez, P., Silva, A. (2008, January). *USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE.* Retrieved December 9, 2022, from http://www3.dsi.uminho.pt/pcortez/student.pdf