Machine Learning Summative Assignment 2019-20

rtjm84

Data Preparation and Analysis

I began by preparing and restructuring data from *studentInfo.csv*, some initial features used include a student's number of previous attempts and highest education level among others. Where data was not represented numerically, I used one-hot encoding using python pandas' dummies. This approach, when compared to replacing the original column with numerical values, greatly improved the methods' scores. I ensured that *NaN* values were replaced with the median of their columns where appropriate to reduce their affect on results. I then added data from other spreadsheets to produce a final dataframe of predictors using almost all of the data available, creating additional predictors where appropriate.

I have used *studentVle.csv* a little but would have preferred to explore the data further, for example, by using *vle.csv* to analyse the number of clicks per activity type, but this required too much processing for my machine to handle.

Additionally, care was taken when merging dataframes so that no new rows were added. This would have otherwise made it probable that at least one of a student's rows would be in each of the training and test sets, leading to a significant increase in accuracy but not reflecting the method's true performance.

I outputted the correlation, importance, and statistical description of each feature to understand their impact on prediction accuracy and whether further data manipulation should be considered.

Predictor	Purpose	Importance
dateConcat	Mean avg. of dates for site visits	0.27838
meanScore	Mean Score	0.10027
weightedTotal	Mean of: Score * Assessment Weight	0.08373

Table 1: Predictors of Highest Importance

Predictor	Purpose	Correlation
dateConcat	Mean avg. of dates for site visits	+0.6156
TMA	Assessment type: Tutor Marked Assessment	+0.4790
numRes	Number of scores/results available	+0.4272
numSites	Number of sites visited	+0.4102
sumConcat	Total no. clicks across all sites	+0.3761
weightedTotal	Mean of: Score * Assessment Weight	+0.3201
meanScore	Mean Score	+0.3183

Table 2: Strongest correlating predictors

As expected, there is some similarity between features of high importance and high correlation. *Figures 3-5* demonstrate the correlation for some of these features with the final result. Furthermore, a heat map comparing correlations of all predictors (post one-hot encoding) can be seen in *Figure 10*.

As for method-specific predictors, I only allowed the id_assessment and id_student features for classifier methods and not regressor methods due to their values lacking linear correlation.

Comparison of Methods

Several methods were tested on predicting all four final_results with basic cross validation to check validity and a 70/30 train/test set split:

Method	Accuracy
Random Forest	69.1
Decision Tree	62.3
Linear Regression	60.5
Logistic Regression	42.8
SVC	38.9

Table 3: Initial method scores

I chose the Random Forest Classifier as the first of my methods due to its high performance. For my second method, I wanted to choose a regression model and decided upon Logistic Regression despite its lower perfor-

mance. This decision was due to it having more potential parameters for tuning whilst also being more commonly used for classification problems.

Improving Chosen Methods

I began with hyper tuning parameters through randomized search before moving onto grid search for improved performance, both methods using cross validation. With this method I could run both methods on a variety of parameters to find the most appropriate. With both methods I chose a verbose of 2 and a 10-fold cross validation method. I also used validation curves to aid my range selection for parameter tuning (see *Figures 6-9*). I would have preferred to test a larger range of parameters to a greater extent but was restricted by the processing power of my machine which couldn't handle too much strain.

For the Random Forest classifier (RFC) I tested the following: number of estimators, maximum tree depth, minimum number of samples required for splitting, and minimum number of samples required to be at a leaf node. As for the Logistic Regression model (LR) I tuned: the norm used in the penalization process, tolerance of the stopping criteria, inverse of regularization strength, and the solver algorithm used. Additionally, for LR I normalized data which slightly improved performance and improved the numerical stability of the model as well as perhaps reducing the training time.

Conclusion

Overall, RFC has outperformed LR, my final scores for 4 categories were 71.7% for RFC and 67.3% for LR. When *Withdrawn* and *Fail* were combined and *Pass* and *Distinction* were combined to create 2-categories, RFC scored 91.0% whereas LR scored 87.8% (see Table 4). For a true indicator of performance, the weighted f1 score is perhaps the most accurate due to it being a weighted average of the precision and recall.

We can see from the normalized confusion matrices in *Figure 1* and *Figure 2* that when predicting 4 categories, RFC is much less likely to confuse similar categories such as predicting a *pass* grade for a *distinction*, a more common confusion than between a passing and a failing grade, this is supported by the lower *MSE* and *MAE*, and higher *EVS*.

	RFC (4 ctgry. / 2 ctgry.)	LR (4 ctgry. / 2 ctgry.)
f1 Score (Weighted)	0.701 / 0.910	0.648 / 0.878
R2 Score	0.576 / 0.640	0.527 / 0.510
MSE	0.414 / 0.090	0.468 / 0.122
MAE	0.320 / 0.090	0.370 / 0.120
Explained Variance Score	0.582 / 0.644	0.530 / 0.534
Accuracy	0.717 / 0.910	0.673 0.878

Table 4: Final Results for Chosen Methods

Appendix

	Precision	Recall	f1-score	support
Withdrawn	0.71	0.79	0.75	2959
Fail	0.66	0.41	0.50	2203
Pass	0.74	0.91	0.82	3688
Distinction	0.71	0.48	0.57	928
Accuracy			0.72	9778
Macro Avg.	0.70	0.64	0.66	9778
Weighted Avg.	0.71	0.72	0.70	9778

Table 5: RFC on 4 Categories (Withdrawn, Fail, Pass, Distinction)

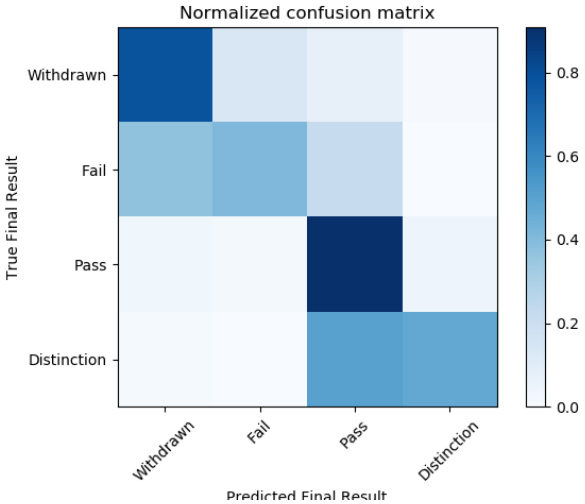

Figure 1: RFC on 4 Categories

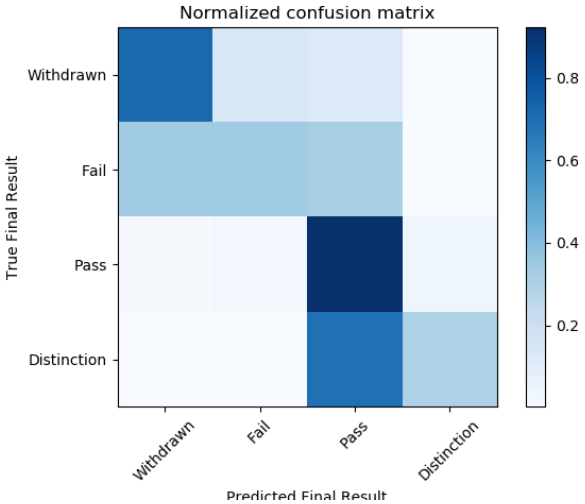

Figure 2: LR on 4 Categories

In the following: 1-Withdrawn, 2-Fail, 3-Pass, 4-Distinction

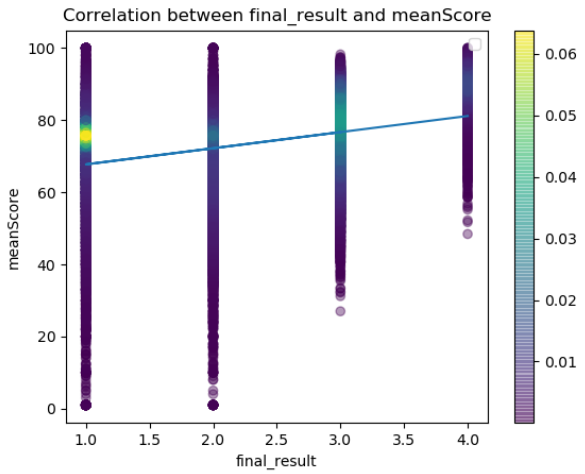

Figure 3: Linear Regression: meanScore vs final_result,

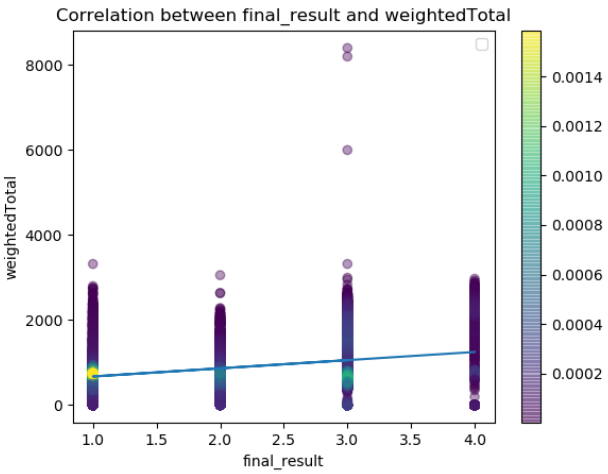

Figure 4: Linear Regression: weightedTotal vs final_result

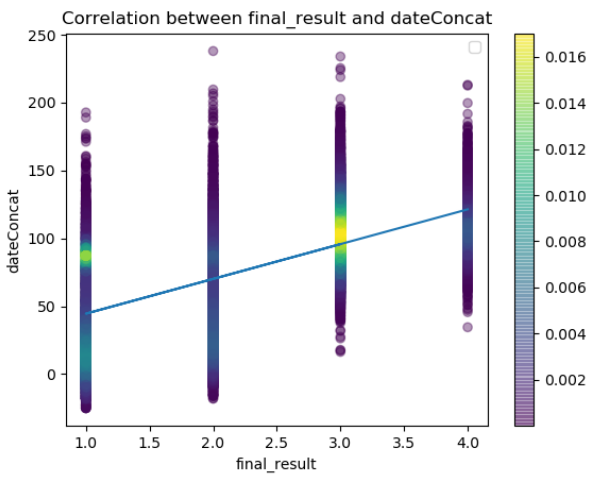

Figure 5: Linear Regression: dateConcat vs final_result

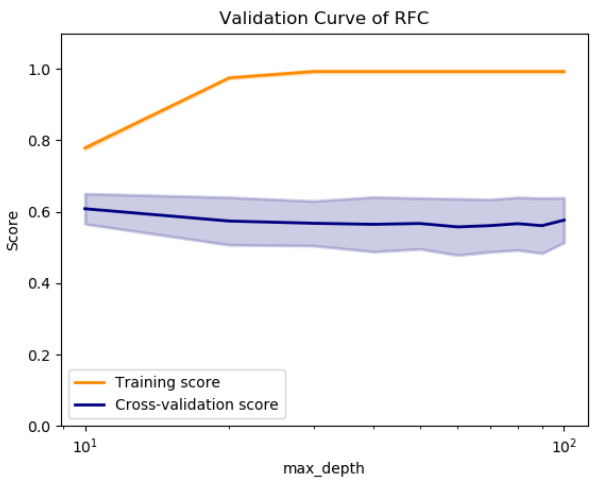

Figure 6: RFC max_depth Validation Curve

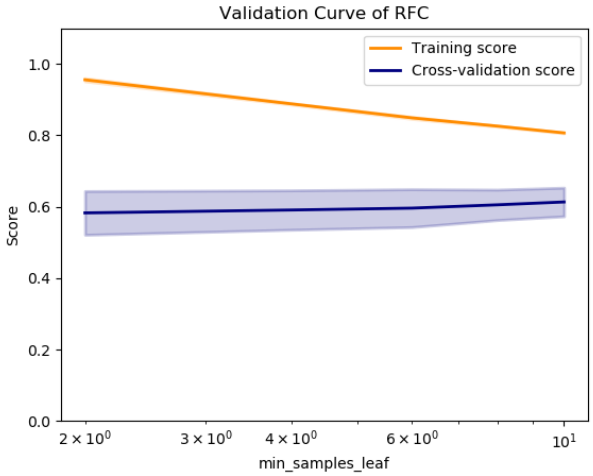

Figure 7: RFC min_samples_leaf Validation Curve

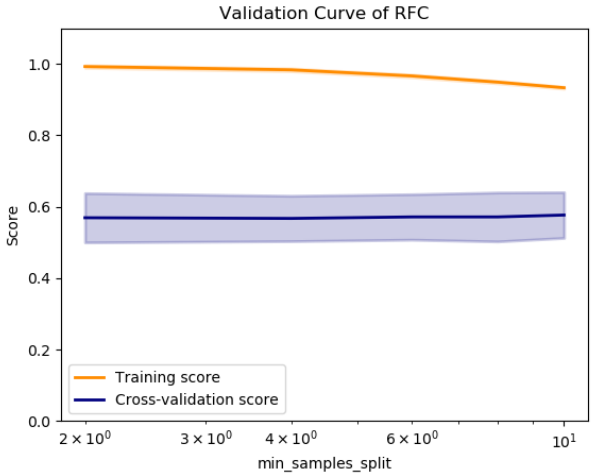

Figure 8: RFC min_samples_split Validation Curve

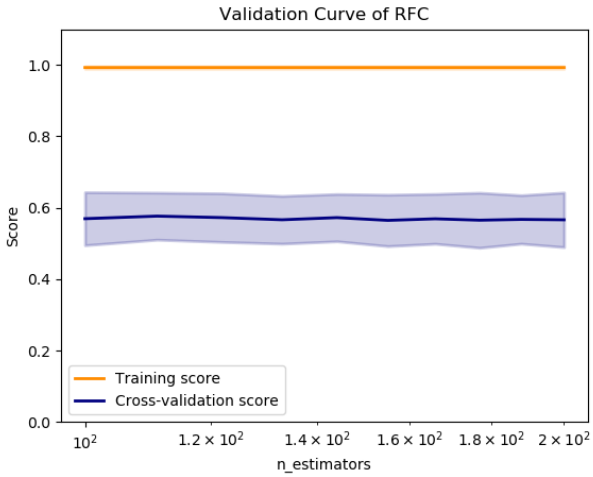

Figure 9: RFC n_estimators Validation Curve

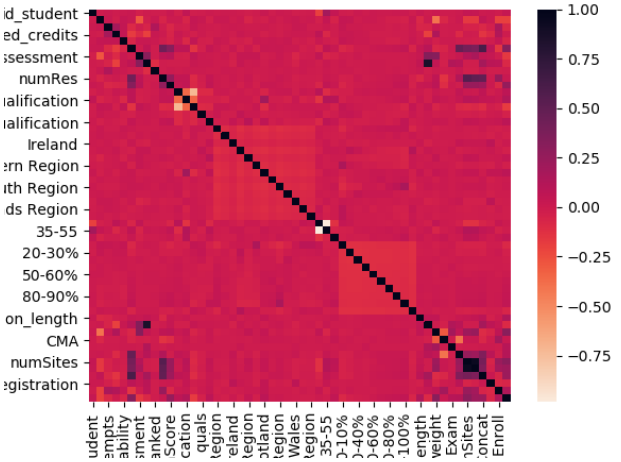

Figure 10: Heat map of Predictors' Correlation