Predict Test Scores of Students

Name:	**Sourin Das**
Registration No./Roll No.:	20273
Institute/University Name:	IISER Bhopal
Program/Stream:	EECS
Problem Release date:	January 12, 2023
Date of Submission:	April 16, 2023

1 Introduction

The objective of this project is to predict the score of individual students in a particular test given a data set which includes features such as school setting, school type, gender, pretest scores etc. In the given data set, we have eleven features and one target variable, that is the posttest score of the students.To predict the student score, firstly, I have done some preliminary data analysis(like which feature have how many categories). After choosing the best features out of all features,we have done model selection and training process, then finally, the model evaluation.

In the given data we have both categorical features(such as school setting, school type, gender etc) and numerical features (pretest,no of student). Most of the categorical feature in the data set has two or three categories, except school and classroom, which have twenty-two and sixty-seven categories respectively.In the Figure 1, we have plotted some feature with categories.

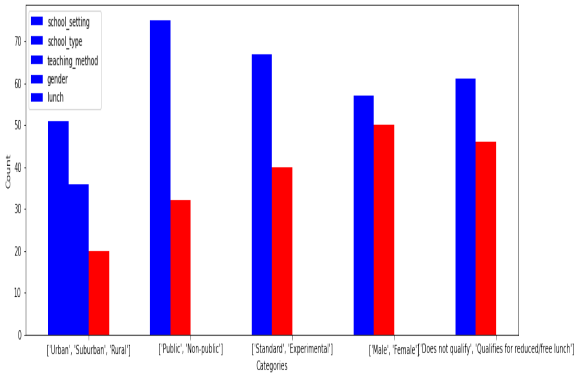

Figure 1: Overview of Features

2 Methods

For the problem, I have Used Random Forest Regression, Linear Regression,Decision Tree Regression, and AdaBoost Regressor model since the target variable is continuous.we can not use classifier for the problem because Classifiers are typically used for predicting categorical outcomes, where the output is a discrete class label.firstly, for the Train data set, we have encoded the categorical data to numerical data with the help of one hot encoding, then we have chosen those features which have a correlation greater than 0.3 with the target variable. after that, we dropped the other [1] [2] features from the train data. then we run each Regression model on the updated train data and got the output. [1, 2] the

[1] https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/

[2] https://towardsdatascience.com/random-forest-regression-5f605132d19d

regression model gives some float values to convert it to nearest integer I have used the round function. For this problem I have explored some other regression model such as Logistic regression,Random forest regression.But find the Random Forest Regression model very efficient.[3]

2.1 Cross-Validation technique

We splitted our given train data set into 2 different data set in the following proportion: 80%-20%: one to train our model(training set) and the other to test our model against the data points with a (validation set) with a random state of 62 with a 3 fold cross validation for each regression model. so as to make the complete use of our data. The stratify parameter will preserve the proportion of target/class label as in original data set, in both the training and validation data sets. We want to preserve the data set proportions for better prediction and reproducibility of results.

2.2 Feature selection

Feature selection aims to identify the most significant features of data to improve the performance of a model. One way to accomplish this is by computing the correlation coefficient between the target variable and different features, which measures the linear relationship between them. The intuition behind using correlation for feature selection is that good features are highly correlated with the target variable, indicating that they are informative for predicting the target variable. Additionally, it is desirable for features to be uncorrelated with each other, as correlated features provide redundant information to the model. In other words, if two features are highly correlated, one of them may be redundant and can be removed without adversely affecting the model's performance. Therefore, by selecting the features with the highest correlation coefficient with the target variable while also minimizing correlation between features, we can improve the accuracy and efficiency of our model.

2.3 Hyperparameter Tuning

Here I have used GridSearchCV to search through the best parameter values from the given set of the grid of parameters. For the linear regression there are 4 parameters and all the parameters are true, which is default setting for linear regression model.so no need to do grid-search in this case.for the Random forest regressionmodel i have optimized the six parameter(bootstrap, criterion, max_depth,max_features,min_samples_split,n_estimators)and found the optimized parameter.for the DecisionTree Regression i have optimized the parameter(criterion,max_depth,max_features, min_samples_split). and for the AdaBoost Regression i have took the parameter(learning_rate, loss, n_estimators) to optimized. then we find the best parameters for each regression model by using best_params_function.

2.4 Github file link

Here is the link to the github file. `https://github.com/SouriniiserB/DSML_project.git`

3 Evaluation Criteria

Precision, recall, and F-measure are evaluation metrics commonly used in classification tasks to measure the performance of a predictive model. In regression tasks, alternative metrics like mean absolute error (MAE) and root mean squared error (RMSE) are more commonly used. Here is a brief explanation of each of these metrics. [4] [3] [4]

Precision: Precision is the fraction of true positive predictions among all the positive predictions made by the model. It is a measure of the model's ability to avoid false positives. The formula for precision is:

[3] https://towardsdatascience.com/cross-validation-and-hyperparameter-tuning-how-to-optimise-your-machine-learning-model-13f005af9d7d

[4] https://www.analyticsvidhya.com/blog/2022/02/a-comprehensive-guide-on-hyperparameter-tuning-and-its-techniques/

precision = true positives / (true positives + false positives)

Recall: Recall is the fraction of true positive predictions among all the actual positive instances in the data. It is a measure of the model's ability to avoid false negatives. The formula for recall is:

recall = true positives / (true positives + false negatives)

F-measure: The F-measure is the harmonic mean of precision and recall. It is used to balance the importance of precision and recall when evaluating a model's performance. The formula for F-measure is:

F-measure = 2 * (precision * recall) / (precision + recall)

Micro-averaging: Micro-averaging is a technique used to combine the performance metrics of individual instances in the dataset. In micro-averaging, the overall performance is computed by summing up the true positives, false positives, and false negatives over all instances in the dataset. Micro-averaging gives more weight to the performance on the instances with larger counts.

Macro-averaging: Macro-averaging is a technique used to combine the performance metrics of individual instances in the dataset, but unlike micro-averaging, each instance is given equal weight. In macro-averaging, the performance metric is calculated for each instance, and then the average is computed over all instances in the dataset.

For regression tasks, precision, recall, and F-measure are not commonly used, since the predictions are continuous rather than discrete. Instead, mean absolute error (MAE) and root mean squared error (RMSE) are often used. MAE is the average absolute difference between the predicted and actual values, while RMSE is the square root of the average squared difference between the predicted and actual values. These metrics are used to evaluate the accuracy of a regression model.

Mean absolute error (MAE) is an evaluation metric used in regression tasks to measure the average absolute difference between the predicted and actual values. MAE is a popular metric because it is easy to understand and interpret, and it is less sensitive to outliers compared to other metrics like mean squared error (MSE) or root mean squared error (RMSE).

The formula for MAE is:

MAE = $(1/n) * |y_i - y_i'|$

where n is the number of data points, y_i is the actual value, y_i' is the predicted value, and $|y_i - y_i'|$ denotes absolute value.

To calculate the MAE, we first make predictions on a set of data points using a regression model. Then, we calculate the absolute difference between the predicted and actual values for each data point. Finally, we take the average of all the absolute differences to get the MAE.

MAE is a measure of the average magnitude of errors in the predictions made by the model. The lower the MAE, the better the model's performance. Another way to do so is by squaring the distance, so that the results are positive. This is done by the MSE, and higher errors (or distances) weigh more in the metric than lower ones, due to the nature of the power function.

MSE = $(1/n) * |y_i - y_i'|^2$

A backlash in MSE is the fact that the unit of the metric is also squared, so if the model tries to predict price in dollar, the MSE will yield a number with unit (dollar)^2 which does not make sense. RMSE is used then to return the MSE error to the original unit by taking the square root of it, while maintaining the property of penalizing higher errors. R-Squared (R^2 or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit). we generally explain the performance of a regression model by the MSE,RMSE and R-Squared values.

4 Analysis of Results

(a)MSE and RMSE are both measures of the error between the predicted values and the actual values. Lower values indicate better performance, as they suggest that the model's predictions are closer to the actual values. In this case, from Table 1 by looking at MSE and RMSE of each model we can suggest which model's predictions have a relatively small amount of error.

Table 1: Performance Of Different Classifiers Using All Terms

Classifier	Precision	Recall	F-measure
MNB	0.137348	0.123153	0.117819
LinearSVM	0.087294	0.113300	0.090413
DecisionTree	0.100314	0.086207	0.088892
RandomForest	0.112584	0.086207	0.087313
KNeighbors	0.082813	0.083744	0.078513

Table 2: Performance Of Different regression Using All Terms

Regression Model	MSE	RMSE	R-squared score
RandomForest Regressor	9.888	3.161	0.9512
Linear Regression	10.37104	3.1611	0.9489
AdaBoost Regressor	12.74876	3.5705	0.9364

(b) In a regression model, accuracy typically refers to the proportion of variance in the dependent variable that is explained by the independent variables. A higher R-Squared value suggests that the model is better at predicting the target variable. an from accuracy score one can indicates which model is performing well in predicting the values of the target variable.

(c) On the contrary,we can see in Table 2 the classifiers are not working properly on predicting the student scores ,because the target variable is continuous not discrete or a categorical type.

5 Discussions and Conclusion

1. Working on the problem we find that the features n_student,pretest,school_GOOBU,school_IDGFP,school _KZKKE,school_UKPGS, school_VVTVA, school_setting_Suburban, school_setting_Urban, school_type_Non _public, school_type_Public,teaching_method_Experimental,teaching_method_Standard, lunch_Does not qualify,lunch_Qualifies for reduced/free lunch are the best features to predict the student scores.we also see how random state of the train-test split effect the MSE and Accuracy of the regression model.

2. from our work we can see that though the regression models has predicting the with a very good accuracy score but strong correlation does not imply cause and effect relationship. moreover, the regression models have more than 5 MSE score which can be reduced.In this case we have used one hot encoding but for a large number of data this encoding technique may not be very usefull beacuse of high cardinality.

References

[1] Michael Fire, Gilad Katz, Yuval Elovici, Bracha Shapira, and Lior Rokach. Predicting student exam's scores by analyzing social network data. pages 584–595, 2012.

[2] Afolabi Ibukun.T. Ojo Grace Funmilayo. Student's performance predictionusing multiple linear regression anddecision tree. *International Journal of Advanced Research in Computer Engineering Technology*, 8, 2019.

[3] Noah H. Gilbert. Predicting success: An application of random forests to student outcomes. 2017.

[4] Frank Emmert-Streib and Matthias Dehmer. Evaluation of regression models: Model assessment, model selection and generalization error. *Machine Learning and Knowledge Extraction*, 1:521–551, 03 2019.