Machine Learning Engineer Nanodegree

CAPSTONE PROJECT REPORT

To predict the student performance in secondary education.

Akhila Chitluri

June 27Th 2018

I. Definition

Project Overview:

This project is about prediction of student's performance based on their past results. The data is contains the features students in of two schools in Portuguese. The Aim of collecting this data is based on the results ,more efficient student prediction tools can be developed to enhance the school resource management .Our Aim in this project is to predict the performance of students based on their primary education this will definitely help them to enhance their future. In this problem we need prediction .We can only predict the by analyzing the whole data .We cannot just do it by seeing the data we need to do the data analysis face and then we can predict the data by using some algorithms. For this type of process data Analyzing and visualizing and data transforming and predicting we need to use machine learning.

The data was collected from the UCI machine Learning .It was published by Paulo Cortez from University of Minho.

Dataset Link: https://archive.ics.uci.edu/ml/datasets/student+performance
Papers Referred: http://www3.dsi.uminho.pt/pcortez/student.pdf

Problem Statement:

Our Aim is to classify the student's performance for secondary education based on the primary education. Our data contains the details of the students in the primary education and some data related to their performance. We need to classify the data to predict the performance of the students in their secondary education.

Given the details of 649 students related to their primary education and some features related to their family and there economical status based on that now we need to predict the how a student secondary education would be.

Features and Description :

These are the features related to dataset:

→ *School*
→ *Sex*
→ *Age*
→ *Family size*
→ *Father's Education*
→ *Mother's Education*
→ *Mother's job*
→ *Father's job*
→ *Internet*
→ *travel time to school to home*
→ *Family educational support*
→ *wants to go to higher educational*
→ *Quality of family relationships*
→*failures*
→ *Absences*
→ *1st time grade*
→ *2nd term grade*

By considering the above factors we will classify the data and then we will predict their future performance in their secondary education.

Our Project is related to supervised learning as we know all the labels and we will use classification techniques to classify the data like logistic regression, Ensemble methods
And then we will select the one which performs well and some optimization techniques and finally we will predict the target Variable.

Performance Metrics:

In this we will use the, F beta score, recall and precision .

We define them as :

Precision = TP/(TP+FP)
Recall = TP(TP+FN)
F beta score = (1+beta^2)(precision*recall)/(beta^2*precision)+recall*
Beta =0.5

Generally we use accuracy as the performance matrices as Accuracy may be biased in some cases which we may end up with wrong results for that case we will use the Recall and precision

As fbeta score is the combination of both we can assure that the data may not be biased and we will get the appropriate results.

II. Analysis

Data Exploration:

In this Data Exploration, I explored who may Attributes and how many instances the data set contains. In this face I found that, there are missing values in the data. A also explored that the data consists of mixed variants like categorical, numerical and some binary data. (Values Yes/No). I also found some attributes which are not necessary for our classification like address, romantic, guardian.

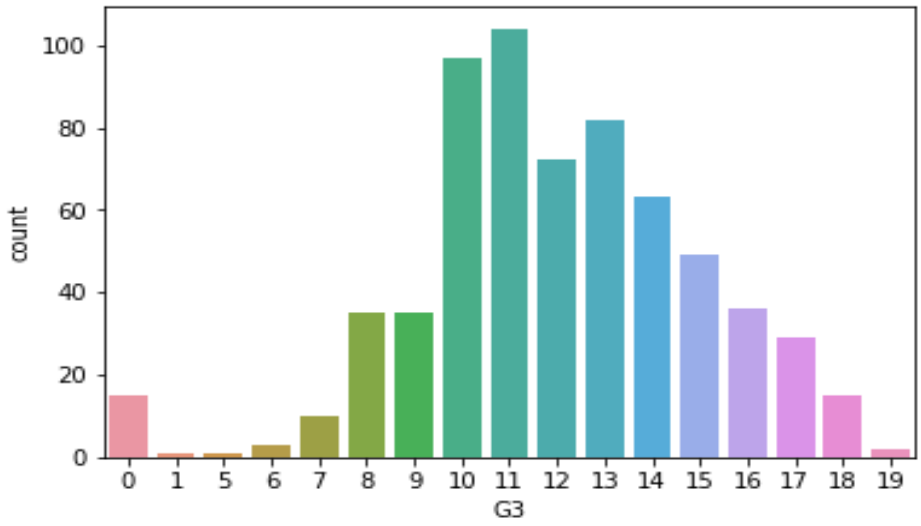
Data Visualization :

→ In this part, I found some data related to the data like its mean, standard deviation, min, max values.

```
: display(data.describe())
```

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	abse
count	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	649.000000	649.00
mean	16.744222	2.514638	2.306626	1.568567	1.930663	0.221880	3.930663	3.180277	3.184900	1.502311	2.280431	3.536210	3.65
std	1.218138	1.134552	1.099931	0.748660	0.829510	0.593235	0.955717	1.051093	1.175766	0.924834	1.284380	1.446259	4.64
min	15.000000	0.000000	0.000000	1.000000	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.00
25%	16.000000	2.000000	1.000000	1.000000	1.000000	0.000000	4.000000	3.000000	2.000000	1.000000	1.000000	2.000000	0.00
50%	17.000000	2.000000	2.000000	1.000000	2.000000	0.000000	4.000000	3.000000	3.000000	1.000000	2.000000	4.000000	2.00
75%	18.000000	4.000000	3.000000	2.000000	2.000000	0.000000	5.000000	4.000000	4.000000	2.000000	3.000000	5.000000	6.00
max	22.000000	4.000000	4.000000	4.000000	4.000000	3.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	32.00

→As my problem is about multi classification I detected who many class will the output contain.



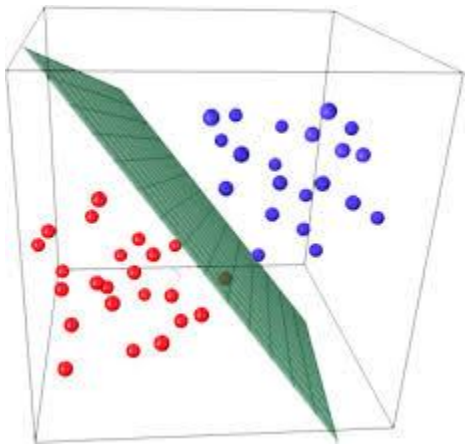
Algorithms and Techniques:

Algorithms:

Our Aim is to classify the data as we have the data which is labeled we use supervised classification and I used following Techniques and Algorithms.

1. Logistic Regression:

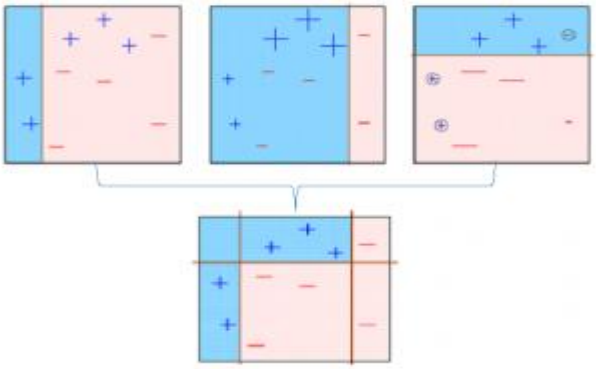
Logistic function is used to predict the probability of the occurrence of an event by fitting data to a mathematical logit function (inverse of sigmoidal logistic function).Logistic Regression because it is simple to fast .But there are some disadvantages they are it may sometimes leads to over fitting .I the case of more features this Logistic Regression may not work well in all cases. Logistic regression classifies in this way.



2. Adaboost Classifier:

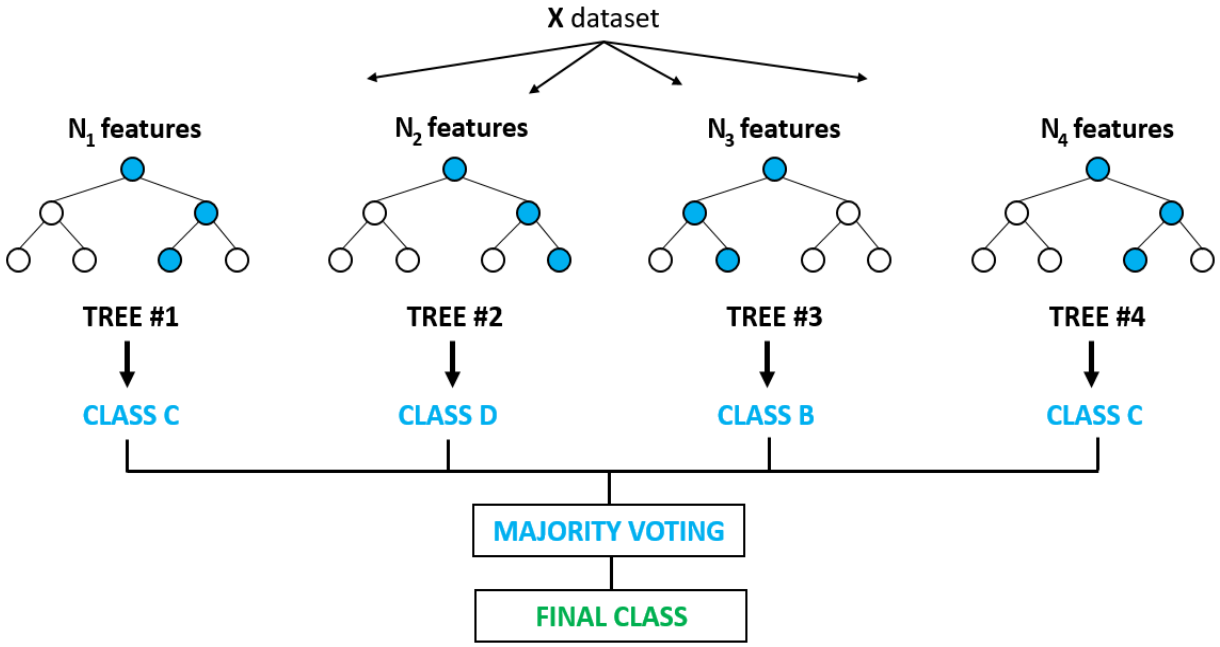
This is one type of Boosting classifiers in the ensemble methods. In Adaboost classifier the "data will be divided into subsets and then it will be evaluated according" .This shows that it is not biased and the data may not over fit.

Adaboost classifies the data in this way.



3. Random Forest Classifier:

This Random forest Classifier uses "decision tress by selecting some features and classifies" the data this process will be repeated no of times and the best one among them will be selected. This shows that the data is will not be biased and does not over fits.



We use some Techniques to optimize the model and to get best results and in this problem. I use Grid search cv which helps to get the better results based on our parameters tuning .By tuning parameters the data may give more accurate results and its performance increases.

Benchmark Model:

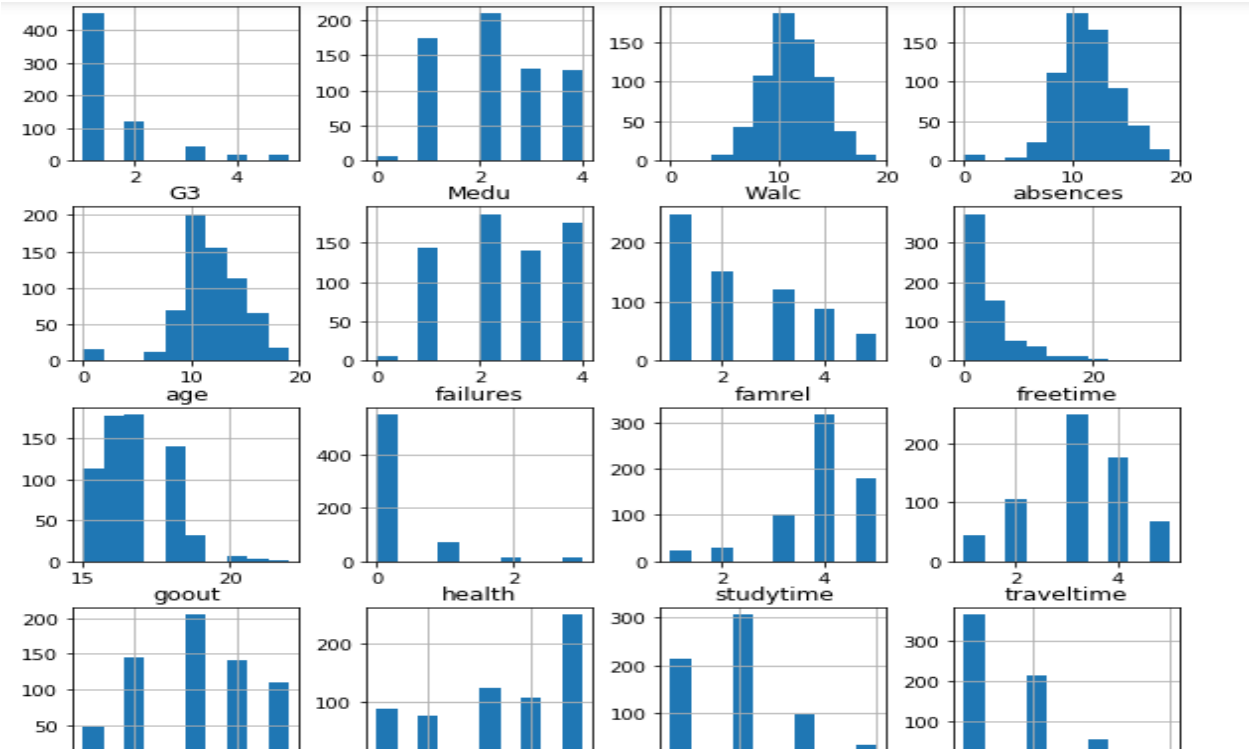
I used Logistic regression as the benchmark model as it is fast to implement and simple classifier. In this I calculated the accuracy and fbeta score as performance matrices to improve the results the next ensemble methods like ada boost classifier and the random forest classifiers will be used and the best f them i.e.: that will give the best fbeta and accuracy results will be selected.

III Methodology:

Data Preprocessing:

i. *In this stage I removed all the missing values and converted all the binary values yes/no to 1/0 respectively. And also I changed one attribute which has its LE3(less than 3) and GT3 (greater than 3) .so I changed this data into 0/1 respectively as it is also a binary data.*

ii. *I also visualized the attributes the by using histograms whether it is skewed or normally distributed. By this I noticed that some of the data is skewed but does not vary to large values .so I concluded that the data can be normalized without using log transform.*



iii. *Data transformation: In this the will be transformed by using min max scalar and then I used one hot encoding to convert the categorical data into numerical .*

	age	famsize	Medu	Fedu	traveltime	studytime	failures	schoolsup	famsup	paid	...	Mjob_teacher	Fjob_at_home	Fjob_health	Fjob_other	Fjob_ser
0	0.428571	1.0	1.00	1.00	0.333333	0.333333	0.0	1.0	0.0	0.0	...	0	0	0	0	
1	0.285714	1.0	0.25	0.25	0.000000	0.333333	0.0	0.0	1.0	0.0	...	0	0	0	1	
2	0.000000	0.0	0.25	0.25	0.000000	0.333333	0.0	1.0	0.0	0.0	...	0	0	0	1	
3	0.000000	1.0	1.00	0.50	0.000000	0.666667	0.0	0.0	1.0	0.0	...	0	0	0	0	
4	0.142857	1.0	0.75	0.75	0.000000	0.333333	0.0	0.0	1.0	0.0	...	0	0	0	1	

5 rows × 41 columns

iv. *Data splitting : we will split the data into training data and testing data in which we train the model using training data and test using testing data I kept 20% of data for testing.*

Implementation.

Data classification and fitting and Predicting with different models.

a) *For every model we need to import the classifier for the respective packages in the sklearn .*
b) *We will fit the data to the model with training data and then we will predict with the by using testing data after that we will find the accuracy and fbeta score by using the true values and predicted values.*

Metric	*Logistic regression*	*Ada boost Classifier*	*Random Forest classifer*
Fbeta score	*0.1788*	*0.1044*	*0.3150*

Among all above I used random forest classifier as it gave me the best when compared to the other two algorithms.

v. *I got a problem with feature selection process .As there are 33 attributes it is not necessary to consider all the features we need to indentify the which feature contributes more weightage on the output to overcome this problem I used "feature_importance" attribute in the random forest to overcome this problem which clearly gives the weightages of each feature. I took top 5 features and modeled the data and the results are as follows.*

Fbeta score before feature selection	*Fbeta score after feature selcetion*
0.3150	*0.4480*

Refinement :

Optimization:

For optimization as the above fbeta scores suggests I optimized the model by using Random forests for this I used Grid search as optimization technique.

I did parameter tuning in the random forest like max_depth , criterion, n_estimators .

I used Fbeta score as performance matrices

Score before Grid search cv	*Score after Grid search cv*
0.4480	*0.6514*

IV Results

Model Evaluation and Validation:

Finally my model selection is random forest classifier .At first it gave the f_beta score result as 0.44 and accuracy score 0.44 and finally by using grid search cv and tuning parameters our fbeta score improved to 0.65 .

We use f_beta score as performance metrics which is the weighted harmonic mean of precision and recall. I used shuffle_split which is Random Permutation cross-validator.Which made our model robust.

Fbeta score before cross validation and Grid search	*Fbeta score after cross validation and Grid search*
0.4480	*0.6514*

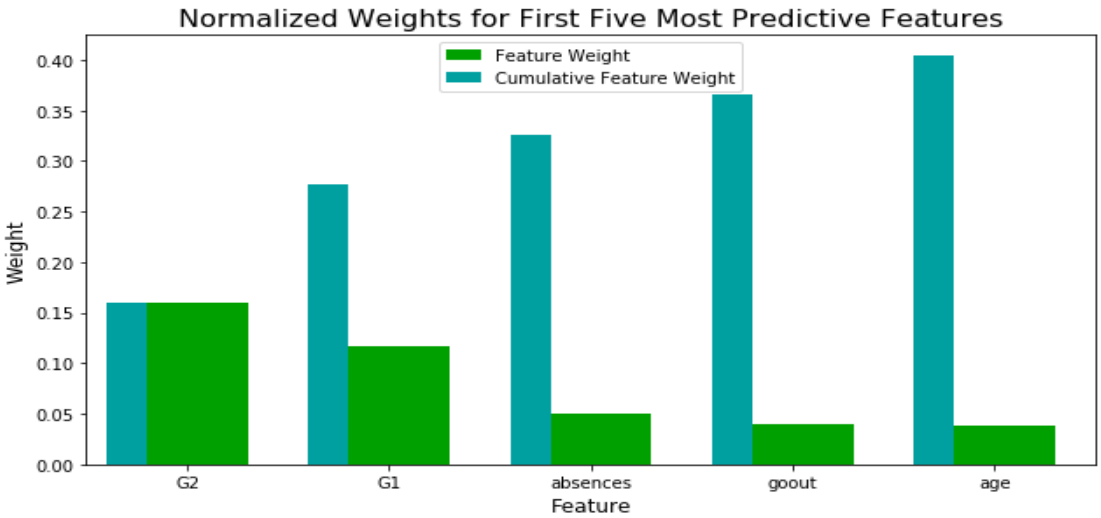
By this I can say that the our model is robust to outliers and we can use it for any type of applications in our day to day life. By this a can say that our model is more accurate and good when compared to other models I used before.

Justification :

Our benchmark model is logistic regression and final model is random forest classifier. In the case of logistic regression the fbeta score is very low its value is 0.17 and in our final model at initial stage our fbeta score is 0.31 which is far better than logistics' score .By this analysis I selected the final model is random forest and after using our optimization techniques and tuning the parameters we achieved our fbeta score to 0.65 % .Hence I can justify that our model gives best solution..

V. conclusion:

Our final conclusion is the data is multi class classified and all the features does not define the output .out of 33 attributes present in my data set after removing unnecessary attributes we get 29 and after using feature importance that there are five features which mostly contribute to the output .They are



By this I can conclude that when we use many features it will be better to ensemble methods to classify and we can easily classify the data .In this project after finding the different models the best results that was provided by random forest classifier and I was satisfied as my fbeta score improved from initial to final from 0.44 to 0.65.

Reflection :

1. *I collected my data from the UCI machine learning and my data set name is student performance.*
2. *After collecting the dataset I observed the data and its attributes which contains 649 instances and 33 attributes.*
3. *The data is inconsistent as my data is contains some of the missing values and my is multi variant data which contains the numerical, categorical and binary values also.*
4. *In the pre processing of the data I removed all the missing values and then separated target variables and input variables and plotted curves to visualize the data.*
5. *The I choice one bench mark model logistic regression and classify the data with that classification and then I also tried different types of the classifications and predicted the values and fbeta score of all the classifications and picked the which has highest fbeta score.*
6. *Among all I used the best one is Random forest classifier and then I check for important features .by using that features I tried to fit the data for the same classifier and then checked the fbeta score which gave me approximately the same answer .*
7. *So by using those features of data I used a technique called grid search cv I optimized the data by tuning the parameters and I finally successes in getting the fbeta score as 65%.*
8. *By this I learned, that we reach to final aim only when we solve the problem by breaking it into pieces. As a part of this project I learned how to analyze the data and to visualize it what are the problems we face if don't visualize the data and how to tune the parameters without over fitting any model and how to make analysis form the plots. Over all, this entire project made me improve my problem solving skills.*

Improvements:

I think by using another ensemble methods like Xboost classifier we may get the better results.

We can also using boosting and bagging algorithms that may also gives good results.

I think my choice of algorithm is good and performed well.

References:

I used visuals.py from the previous project for feature selection.

Ref: https://machinelearningmastery.com/boosting-and-adaboost-for-machine-learning/
Ref: https://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/
Techinques:
https://archive.ics.uci.edu/ml/datasets/student+performance
http://scikit-learn.org/stable/modules/generated/sklearn.metrics.fbeta_score.html