Hackathon Report: Student Performance Prediction

Ajay Ramesh
IMT2017502

Amog Johri
IMT2017003

Arjun Verma
IMT2017008

***Abstract*—Education is extremely important for the evolution of society. The analysis of Educational Data provides a great opportunity to understand and improve the educational system and its policies. In addition it also represents the critical developmental period for both late adolescents and young adults (Chickering, 1969). We utilized several machine learning and data mining methodologies to predict student performance based on various social and academic factors. The results show that a good predictive accuracy can be achieved provided that the previous grades of the student is also considered. But, it is also important to understand the effects of social factors which can substantially influence a student's performance. Predicting a student's performance without previous grades is challenging but very useful[1].Hence, our current project focuses primarily on determining the performance of a student exclusively on the basis of social factors. The data-set was obtained from https://archive.ics.uci.edu/ml/datasets/Student+Performance which consists of Portuguese population. Further data was collected using google-forms in order to carry out similar study over Indian University Students.**

I. Introduction

To characterize education as merely important comes as a massive understatement. In the modern day, education has proven to be a paramount factor determining the success of an individual. However, there are a number of factors which determine the success of one's education itself. One of the most important however, often neglected factors include the social factors. In the current project we try to highlight and draw conclusions over the student's academic performance based on social factors and their daily academic life.

Our project pivots about two broad objectives:

1. Using the UCI Student Performance Data-set (USPD), we attempt at building a model to predict the Student's Academic Performance based on various social factors.
2. Collecting a similar data-set from Indian Students and applying the models trained on the USPD.

The USPD was collected from students under secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features and it was collected by using school reports and questionnaires. Even though the grades previously attained by the student show a very strong correlation (see Fig 1) to the student's future grades, a conscious decision was taken to remove it from the feature attributes (for the majority of the project) to better analyze the effects of social factors. The source paper also claims that prediction of final grades without the first and second period grades is more useful[2]. The data-set contained 1040 instances with 33 attributes. However, after a thorough data-analysis 15 (social and demographic related) attributes were used for training the model. We also tried to include previous grades as attributes and for obvious reasons, the predictions based on those showed a very small error. For the majority of the project, we worked with 15 social and demographic attributes. Following one of the most basic thumb rules in machine learning, we started building linear models for the data before switching to more complex models such as Support Vector Regression with Radial Basis Function kernel and Artificial Neural Networks.

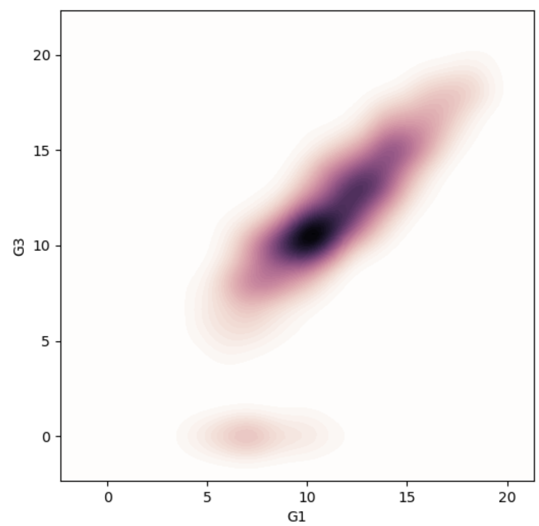

Fig. 1. Final Grade (G3) vs First-Period Grade (G1)

Our second objective involved collecting data from Indian University Students and making similar predictions for the same. Google-forms were used to collect the data and we obtained 151 instances. The data collected had a lot of junk values and thus a lot of data cleaning had to be done. Records with missing grade(G3) values were discarded. Any missing values in a column with categorical data, were imputed by using a random data generator, between the maximum and the minimum value in the column. The missing binary values were also imputed similarly using a random data generator.

II. Dataset

The dataset was obtained from the UCI Machine Learning Repository. The dataset contains 1040 records and contains attributes such as student grades, demographic, social and school related features. The data was collected using school reports and questionnaires. Few example social attributes included in the data are -

- Residence - Urban/Rural
- Relationship status
- Alcohol consumption - Weekly and Daily

Few academic attributes of the data are -

- Study time - number of hours per week
- Number of past failures
- Previous grades

Overall, the dataset contains 33 attributes. Detailed description of each attribute is given in the instructions.txt file. In addition, we collected data from fellow students by rolling out google forms on social media. This data was collected with the intent of using it as a test set.

III. Objectives

We have divided our work into the following two objectives:

A. Study the effect of (a) past academic performance and (b) social factors in predicting grades on the dataset obtained from the UCI Machine Learning Repository.

B. Use our model to infer characteristic relationships between various social factors and grades on the data we collected from Indian Universities.

IV. Data Exploration and Pre-processing

DATA EXPLORATION TECHNIQUES :

The data contained many categorical fields such as - sex, mother's and father's education, internet access etc. Most of the categorical fields were of binary nature (yes/no) and were encoded accordingly using Label Encoding technique. The decision to drop a number of attributes were taken by a careful study of the heatmap of the computed correlation matrix. The various dependencies of the attributes on each other was taken into consideration before making the decision of which attributes to keep and which attributes to drop. The Correlation Heatmap between the features is shown in Figure 2.

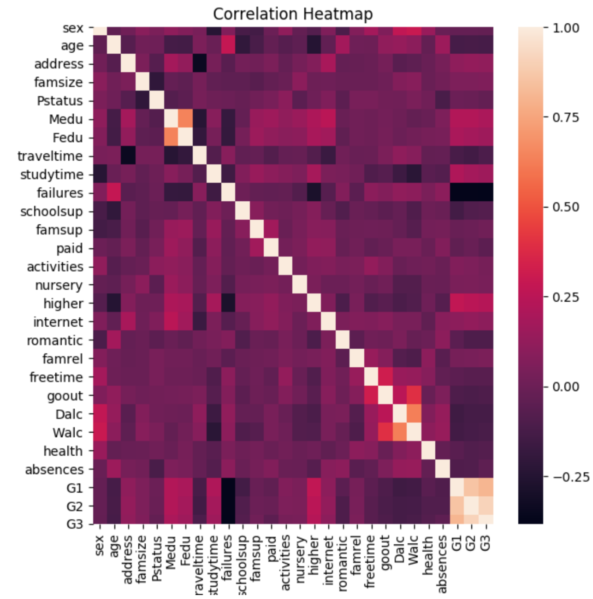

Fig. 2. Correlation heatmap

Fig. 3. Weekday alcohol consumption and final grade

We plotted each social factor against G3 (Final grade) to see their individual effects on a student's academic performances. Some of the inferences drawn are stated ahead. A KD scatterplot showing the relationship between amount of weekday alcohol consumption and final grade obtained out of 20 (G3) is shown in Figure 3. The relation between daily alcohol consumption and student performance is quite clear. The students who consume less alcohol daily have obtained higher scores.

The relation between student travel time and scores is shown in Figure 3 The students who take more than 3 hours to travel to school have obtained lower scores that their peers. This could be due to the decrease in study time because of their long travel times. Thus, various social and academic factors that affect a student's performance can be analysed from such raw data.

DATA PRE - PROCESSING TECHNIQUES :

- **Encoding -** Label encoding was applied since many fields of the data were categorical (Yes/No).
- **Discarding irrelevant features -** The correlation heatmap and KD scatterplots were referred to before discarding the irrelevant features. Fields such as Mother's and Father's job were discarded since they do not affect the outcome in the absence of data related to their financial situation.

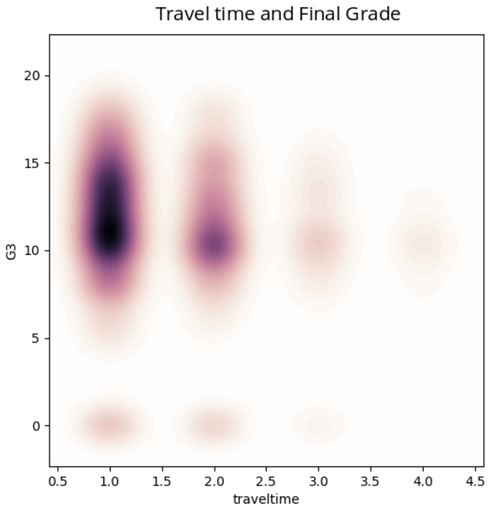

Fig. 4. Travel time to school and score

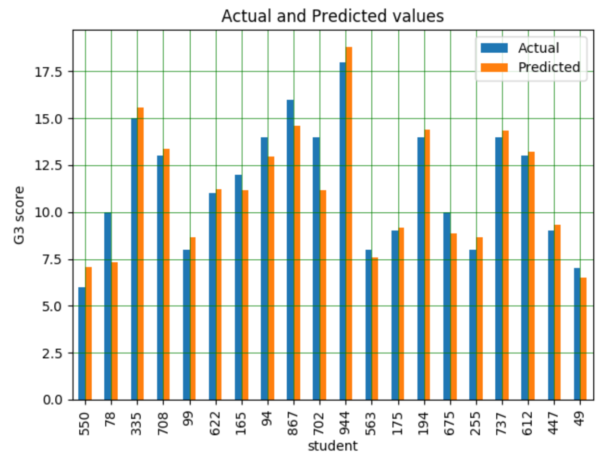

Fig. 5. Predicted and actual scores of students

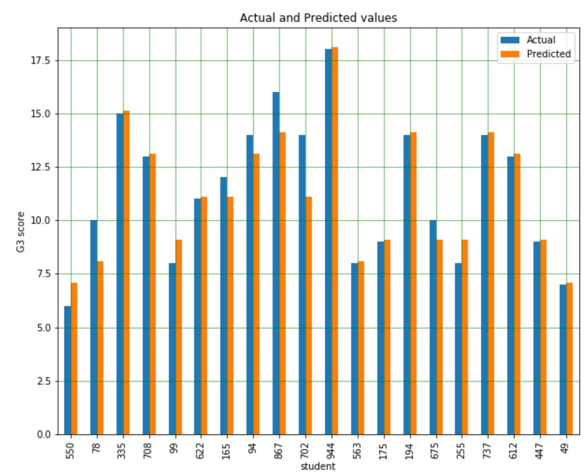

Fig. 6. Predicted and actual scores from the SVR

- **Data Normalization -** The fields of the data are recorded in different scales. For example : Alcohol consumption was recorded on a scale of 1 - 5 and previous grades are recorded out of 20. Moreover, many binary fields exist. Thus, normalization was important.

V. Models and Training

A. Regression

The problem of predicting a student's performance by grade is a regression problem. We have used Linear Regression.

1) Linear Regression: The features were first scaled appropriately and a Linear Regression Model was trained.

2) Non-Linear Regression: The previous scores of students was excluded from the dataset. Predicting the performance of students without considering their previous exam results is challenging as well as useful. After discarding the said two features, the remaining features do not share a linear relationship with the outcome. Thus, this is a non-linear regression problem.

VI. Results

A. Linear Regression

The dataset was split, and a linear regression model was trained and tested. The actual and predicted scores of 20 students is shown in Figure 5. The performance of a linear regression model is evaluated based on the Root Mean Squared Error : $RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{d_i - f_i}{\sigma_i}\right)^2} = 1.38$

B. Linear Support Vector Regression

A Support Vector Regression(SVR) model with Linear Kernel was trained. The actual and predicted values of scores of 20 students is shown in Figure 6.

$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{d_i - f_i}{\sigma_i}\right)^2} = 1.36$

C. Artificial Neural Networks

A thorough data-visualization showed features having a strong, but non-linear influence on the target field. One such feature has been represented in Fig. 7. which shows dependence of the student's grades to their frequency to go out with friends (at a scale of 1 to 5). Linear models such as linear-regression tend to give sub-optimal performance when attempting to model non-linear relations. Hence, our next step was to turn to Artificial Neural Networks(ANNs). ANNs have given state-of-the-art results in modeling complex non-linear relations for problems over various domains. Several architectures and hyperparameter values were tried in order to get to the best fit for the data. A few of the values which provided better results are presented in Figure 5. On the basis of these analysis, Model 6 was chosen as the final model. The RMSE was calculated over 10 random partitioning of the dataset and the final RMSE to compare within the model

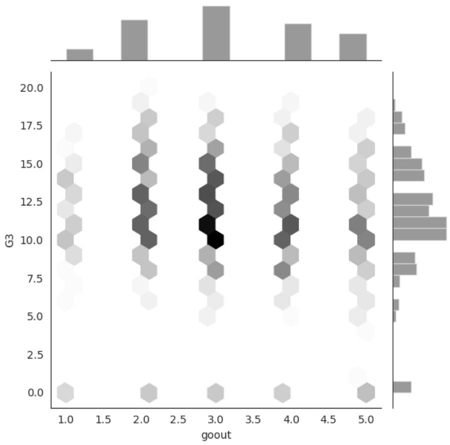

Fig. 7. Frequency of going out and grades

was taken as the average obtained through a 90-10, 80-20 and 70-30 Train/Validation split. Random Normal Distribution was used for weight initialization for all the layers and the activation function used was Rectified Linear-Unit (ReLU). Adam Optimizer was used over Root Mean-squared error to train the model over.

Model Name	Train/Test Split	Architecture	Epochs	Root-Mean-Square-Error
Model 1	90-10	2 Layers : 20 – 1	10	3.2
Model 2	90-10	2 Layers : 25 – 1	10	3.2
Model 3	90-10	2 Layers : 35 – 1	10	3.18
Model 4	**90-10**	**2 Layers : 50 – 1**	**10**	**3.16**
Model 5	90-10	2 Layers : 100 – 1	10	3.18
Model 6	90-10	3 Layers : 75 – 25 – 1	10	3.21
Model 7	90-10	3 Layers : 32 – 8 – 1	10	3.17
Model 8	90-10	4 Layers : 50 – 25 – 10 – 1	10	3.22
Model 9	90-19	5 Layers : 100-50-20-10-1	10	3.21

Fig. 8. Artificial Neural Network Architectures

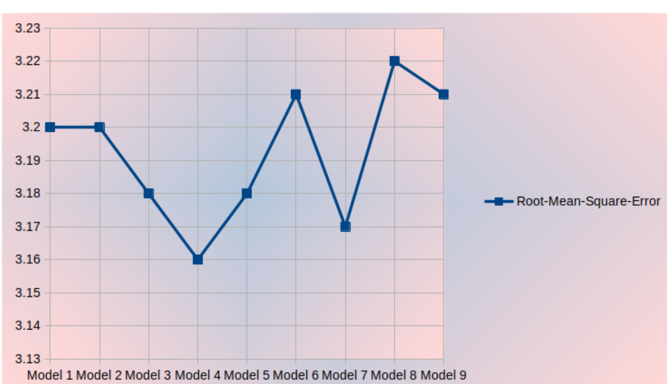

Fig. 9. Artificial Neural Network Architectures 2

However, due to lack of time and resources trying to find an adequate architecture became a massive challenge. Due to the enormous number of hyperparameters and their tendency to overfit, Artificial Neural Network Models often deliver inadequate performance for small datasets (the case with our present project).

D. Non Linear Support Vector Regression

Contrary to Artificial Neural Networks, the development of SVMs involved sound theory first, then implementation and experiments. Where ANN can suffer from multiple local minimas, the solution to an SVM is global and unique. The reason that SVMs often outperform ANNs in practice is that they deal with the biggest problem with ANNs, SVMs are less prone to overfitting[3].Hence, our next logical step was to use Support Vector Regression which could account non-linear relationships in the data.

A Support Vector Regression (SVR) model with RBF kernel was used. The epsilon parameter for the SVR model was reduced to 0.0005. The value of epsilon defines a margin of tolerance where no penalty is given to errors. The smaller the value, the lesser errors are admitted into the solutions.

VII. Summary of Results

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{d_i - f_i}{\sigma_i}\right)^2}$$

A. Predicting grades on the UCI Machine Learning Repository

- Considering the past academic performance, represented by the linear features G1 and G2. (17 features)
  - Using Linear Regression

Model	Linear Regression
Mean Absolute Error	0.840323357806673
Mean Squared Error	1.9122791160804276
Root Mean Squared Error	1.3828518055382606

  - Using SVR with a Linear Kernel

Model	SVR (Linear)
Mean Absolute Error	0.7474999842033714
Mean Squared Error	1.8723460779853631
Root Mean Squared Error	1.3683369753044616

- Considering only the social factors and dropping the past academic performance, thus leaving only non-linear features. (15 features)
  - Using SVR with RBF Kernel

Model	SVR (RBF)
Mean Absolute Error	1.8156440017400712
Mean Squared Error	5.659504043024194
Root Mean Squared Error	2.378971215257594

– Using ANN

Model	ANN
Root Mean Squared Error	2.43520118

B. Predicting grades on the data collected from various Indian Universities.

- Considering only the social factors and dropping the past academic performance, thus leaving only non-linear features. (11 features)

  – Using SVR with Linear Kernel

Model	SVR (Linear)
Mean Absolute Error	2.37916119972691
Mean Squared Error	9.250382792841222
Root Mean Squared Error	3.041444195253502

  – Using SVR with RBF Kernel

Model	SVR (RBF)
Mean Absolute Error	5.324963961446334
Mean Squared Error	32.507283700156634
Root Mean Squared Error	5.701515912470703

VIII. Conclusion

Our problem statement essentially comprises of 3 sub problem statements. Objective A has been subdivided into 2 parts while Objective B remains an individual sub problem statement. SVR with a linear kernel gave us the best results in the first part of Objective A, which was to predict the grades on the UCI Machine Learning Repository taking into account the past academic performance. SVR with RBF kernel outperformed ANN in the second part of Objective A, which was t predict the grades solely on social and demographic factors in a student's daily life. SVR with Linear Kernel gave us the best results on the dataset that we collected from various Indian Universities.

IX. References

1 http://www3.dsi.uminho.pt/pcortez/student.pdf

2 https://archive.ics.uci.edu/ml/datasets/student+performance

3 https://www.svms.org/anns.html