Machine Learning Project
Student grade prediction

Padraigh Jarvis

Table of contents

1. Abstract

The prediction student grades is a potential application of machine learning that can be use to make colleges or schools more aware of a student's capabilities and provide support to them so that they can reach their fullest potential as well as providing a increased retention rate in colleges. Projects have been undertaken in the past to predict a student's final grade based on ether their previous grades or their opinion of the subject, but none have taken into consideration the socio-economic factors that the student might face such as their alcohol consumption or their family status.

Using a dataset provided by Kaggle.com that provided socio-economic features of students as well as their grades. While this dataset had a limited amount of data entries a model was created to predict the final results of a student by putting them in to 1 of 4 categories. The first category being results of 0-25%, second being 26-50%, third being 51-75% and fourth being 76-100%.Several algorithms were tested as well as hyper-parameter optimization for each of them and it was found that the SVC algorithm best suited for this model, providing results of 83% accuracy.

2. Introduction

The are of grade prediction is the prediction of a student's future grades based off of their previous grades and a number of social factors such as family conditions etc.

2.1 Motivation

The Higher Education Authority have found that between the years 2014 and 2015 there was an average dropout rate of almost ¼ in Cork Institute of Technology [1]. This number can be reinforced by the number of students that have even dropped out of the software development 2014-2018 course numbering somewhere between 10 and 20 with a starting number of 40 students in 2014 with second year having by far the highest failure rate.

Machine learning allows for the creation of models that can be used to predict outcomes based off of a number of feature vectors, a good example of this is the breast cancer classification model that uses features to predict if a tumor is either benign or malignant.

Using a data set that contains a number of student features such as their previous grades and their socio-economic situation the goal for this project is to create a machine learning model that will be able to predict the final grade for a student. The hope is in a real life situation where a student’s grade are predicted to be a fail(<40%) then extra learning assets could be assigned to them in the hopes of improving the chances of them obtaining a higher final grade.

The potential increase in final grades would save money on paying staff to create and oversee repeat examinations or allocate resources to repeating students, the reduction in dropouts would also reflect well on their student enrolment to graduate ratio.

3. Related Research

Many papers in the past have focused on student grade prediction during previous grading data gathered from students, there are a lack of papers that focused on the use of social/economic features that could impact the final grade of a student.

A paper released in 2015 focused on the prediction of student grades for the next semester based off of their grades from the previous semesters.The potential grades for students were split into 6 classifications (A-F, A being best F being worst) and was decided that for this ‘rating prediction’ problem 3 different types of algorithms would be tested. The first being SVD, second being SVD but pre-processed with KNN and lastly being a factorization machine model. Using a dataset of 310,557 students it was found that using the Factorization machine model provided the lowest root-mean-square error with the SVD with KNN doing second best and pure SVD doing the worst.While the model was able to predict passing grades accurately enough , it had a problem prediction failing grades.[2]

An observation made during this study was that with an increased amount of data the more accurate the model became, this means that the model they used showed a low bias in the model learning curve.

Another research paper that took place in 2014 focused on student comments which reflect the students attitude to a subject , if they are having learning difficulties and other factors relating to an academic subject that the student is taking. The paper focused on using an artificial neural network to develop a model for predicting their grades , using this they managed to achieve results up to 82.6% accuracy. They mapped the comments made by a student to their final grade, a similarity measuring method was then used to calculate the similarity of a new comment and a comment in the nearest cluster. A three layered perceptron was used to create a model for each lesson for the students. The first layer was the input layer that held the latent semantic analysis results of the similarity between words. The second was a hidden layer with 30 hidden neurons and the 3rd was the output layer that gave back a 5 neurons indicating the students resulting grade. The accuracy of the model developed was evaluated using k-fold cross fold validation with a k value of 10.[3]

4. Algorithm/Model Detail

4.1 Dataset used

The dataset used for this project was the student alcohol consumption dataset found on kaggle.The dataset contains 395 rows and 32 features, the dataset used was taken from students at Mousinho da Silveira which is a mathematics school in portugal. A list of the features and their description can be found in appendix 1.

This data provided a wide number of socio-economic features about a student that could be used during the development of the model as well as academic features like previous grades and absences from school. A problem with the data set is the low number of data entries in it, only having 395 row means it is not feasible to split the data set into a training and a test set.As a result other method of testing such a k-fold cross validation must be used to test the model created on unseen data.

The dataset will go through some cleaning changing the string values(such as yes or no) into numeric representation such as 1 and 0. The school feature will be dropped from the dataset due to the fact that all of the data entries in this dataset are from Mousinho da Silveira and will provide no useful data to the model.Due to the absence of null or missing values the dropping of data entries or imputation will not have to take place.

The G1,G2 and G3 features will be converted into number from 0-3. 0 representing a score of 0-25% , 1 representing a score of 26-50%, 2 representing a score of 51-75% and 3 representing 76-100%. In this way we have a multiclass classification problem of what potential category a student's score will fall under.

4.2 Algorithms used

A range of algorithms will be used to created models for this dataset, these models will then be tested via k-fold validation to predict the accuracy for that particular algorithm. The algorithms that will be tested for preliminary accuracy will be decision tree classifier, naive bayes, nearest neighbor, random forest, logistical regression and SVC

After all of the algorithms have been run hyper parameter optimization will then be run on the algorithms(baring naive bayes). This is to allow us to see if there are any jumps in accuracy due to more optimized constructor arguments being passed into the algorithm that might suit our model better. When the hyper parameter results have been obtained a confusion matrix will be generated for the model that provided the best results during hyper parameter optimization to find out if the accuracy obtained are the results of miss classifications or not.

5. Empirical Evaluation

After the preprocessing took but before any model processing took place a heat map of the features in the dataset took place to give us a better idea on how each feature correlates to one another.See figure 5.1 for the heat map

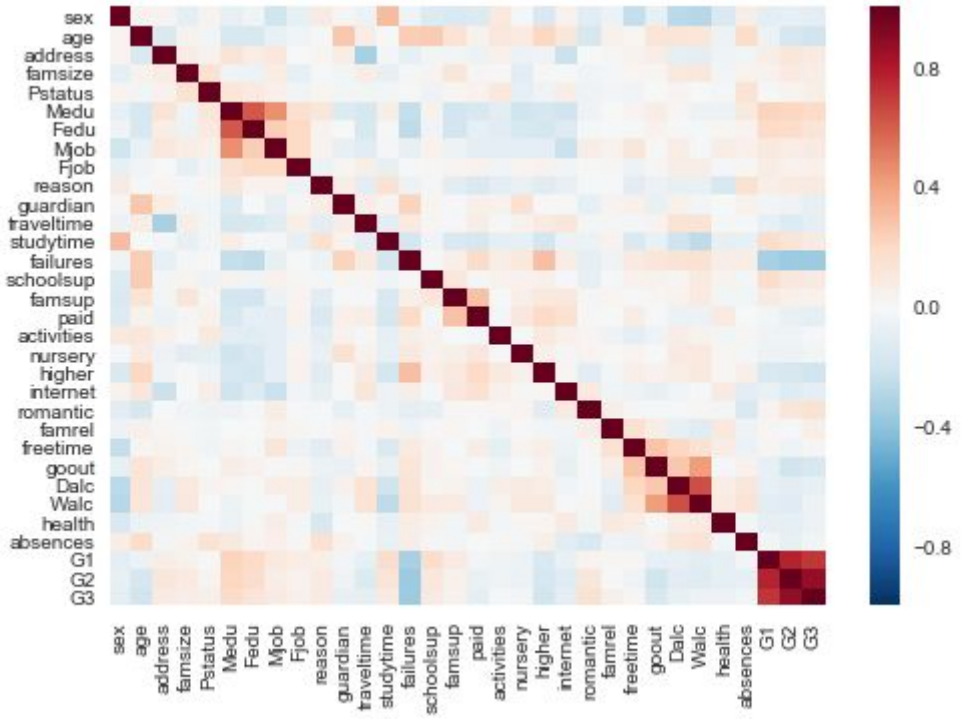

Figure 5.1 heat map

From this heat map we can see that very few features have a high impact on G3. G2 and G1 have by far the biggest impact with failures having the next highest impact on G3 after them.This will be reflected more in the feature removal section.

5.1 Initial model algorithm results

During the initial testing on the basic algorithms with no hyper parameter optimization results found that a random forest algorithm provided the best accuracy in 10 fold cross fold validation.A table of the different algorithms and their results can be found in table 5.1

Algorithm name	Initial accuracy
Decision Tree	73.60%
Naive Bayes	58.62%
Nearest Neighbor	54.37%
Logistical Regression	73.90%
SVC	65.12%

Random Forest	77.97%

Table 5.1

One the initial results were compiled hyper parameter optimization was then run, after this is was found that SVC provided the best accuracy with the hyper parameter optimization providing a massive jump in the initial accuracy. A full list of the best parameters used for hyper optimization and the results for the individual algorithms can be found in table 5.2

Algorithm	Best parameters	Accuarcy
Decision Tree	'criterion': ‘entropy', 'random_state': 1	75.69%
Nearest Neighbor	'n_neighbors': 10, 'p': 1, 'weights': 'uniform'	65.82%
Logistical Regression	'C': 1, 'random_state': 0	73.92%
SVC	'C': 10000, 'gamma': 1e-05	83.79%
Random Forest	'criterion': 'gini', 'max_features': 'sqrt', 'n_estimators': 80	83.29%

Table 5.2

After reviewing the results of the hyper parameter optimization SVC was used as the selected algorithm for the development of this model.

5.2 Confusion matrix results

Using a model, generated from a SVC algorithm with the best parameters produced from the hyper parameter optimization, a 4x4 confusion matrix was generated to reflect the true results of the model's predictive power. The resulting confusion matrix showed that the majority of results where being returned as a true positive ,thus reaffirming the accuracy of this model. See figure 5.2.

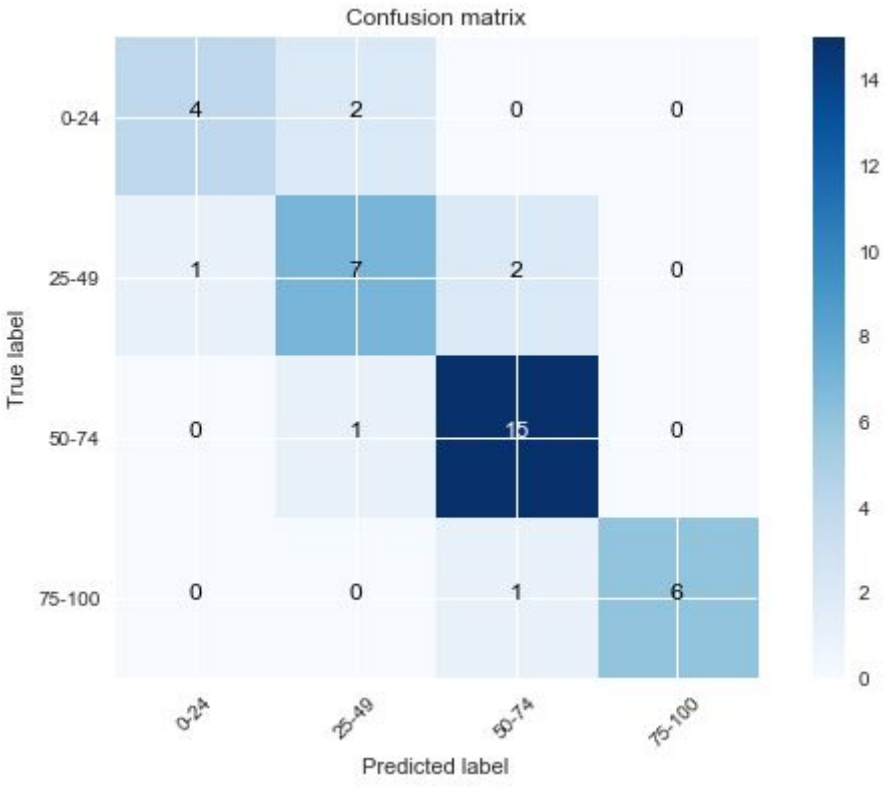

Figure 5.2 Confusion Matrix

5.3 Feature removal results

After reviewal of the heat map and seeing how a majority of features in the dataset (most notably the social-economic features) it was decided to run feature removal to see if there was a decrease or increase in accuracy. It was found that after an initial dip in accuracy after the first few features are removed there is an increase in the overall accuracy of the model. The increase in accuracy was minimal however with a total increase of only ~2%. See figure 5.2 (Note: Y axis represents accuracy out of 1 and X axis represents the number of features removed)

Figure 5.3 Feature removal graph

6. Conclusion and Future Work

At the conclusion of this project it was found that the socio-economic features that were found in the dataset had little to no impact on the final grade that the student obtains. By far the most impactful socio-economic feature is the student's mother's education.As a result a large amount of the features available in this a dataset could be dropped to provide a small increase in accuracy for the model, this could be because of the low number of data entries in the dataset.

Future work for this project would include a increased volume of data entries, this increase of data could show more potential correlations between the socio-economic features and the final grade of the student. The features including in the dataset could also be expanded to include things like amount of money the family earns or even personality features such as if the student is an introvert or extrovert to see if there are any correlations between those features and their final grade. An increased amount of data entries would also allow for the increase in the amount of classes present in this multiclass classification problem going from quartiles to something that pertains more to the domain like A-F grading of schools or class honours of college systems.

Bibliography

[1] *A STUDY OF PROGRESSION IN IRISH HIGHER EDUCATION,* D. Frawley, V. Pigott and D. Carroll
[2] *Next-term student grade prediction.*M. Sweeney , J. Lester ,H. Rangwala
[3] *Predicting students' grades based on free style comments data by artificial neural network.* S. E. Sorour, T. Mine, K. Goda, S. Hirokawa

Apendex 1

Feature	Description
School	If the school is Gavriel Perira (GP) or Mousinho da Silveira(MS)
Sex	Female(F) or Male(M)
Age	Age of the student between 15 and 22
Address	If the home address of the student is in an urban area(U) or a rural area(R)
Famsize	The size of the family(less or equal to 3(LE3) or greater than 3(GT3))
Pstatus	If parents are living together or not (A for appart , T for together)
Medu	Education level of the student's mother(none(0),primary education(1),5th to 9th grade(2),secondary education(3),higher education(4)
Fedu	Education level of the student's father (none(0),primary education(1),5th to 9th grade(2),secondary education(3),higher education(4)
Mjob	Mother's job(teacher,health,civil service ,at_home or other)
Fjob	Father's job(teacher,health,civil service,at_home or other)
Reason	Reason for choosing current school(close to home, school's reputation , course preference or other)
Guardian	The student's guardian(mother,father or other)
traveltime	Time to travel from school to home(less than 15 minutes(1),15-30 minutes(2),30 minutes to an hour (3) or more then an hour(4))
studyTime	Weekly study time(less than 2 hours(1),2-5 hours(2),5-10 hours(3) or more then 10 hours(4))
Failures	Number of past failures(1 if 1 fail, 2 if 2 fail, 3 if 3 fail , 4 if more than 3 fails)
schoolsup	Extra educational support(yes or no)

Famsup	Family educational support(yes or no)
Paid	Extra paid classes in course subject(yes or no)
activities	Extra-curricular activities(yes or no)
Nursery	Attended a nursery school(yes or no)
Higher	Interested in higher education(yes or no)
Internet	Has access to internet at home(yes or no)
Romantic	Has a romantic relationship(yes or no)
famrel	Quality of family relationships(1 being very bad, 5 being very good)
freetime	Free time after school(1 for very low, 5 for very high)
goout	Going out with friends(1 for very low , 5 for very high)
Dalc	Work day alcohol consumption(1 for very low , 5 for very high)
Walc	Weekend alcohol consumption(1 for very low, 5 for very high)
health	Current health status(1 for very bad, 5 for very good)
absences	Number of school absences(0-93)
G1	First period grade(0-20)
G2	Second period grade(0-20)
G3	Final grade(0-20)