Summary for task Student Grade Estimation

Israel Cuautle Muñoz

File DT.R has code used to this topic, you can use it to train, and/or use test sets for getting results.

All tasks regarding, creating new columns, renaming columns, deleting non-useful columns, etc., were performed using Excel.

- Records total = 127924
- 60% = 76754 records for training.
- 40% = 51169 records for testing splitted in two, for getting TEST1 and TEST2
- Rows deleted due lack of information in most of provided (original) columns = 67

Sorted rows using excel through rand() function as temporal added column and sorted several times in orderto get as much randomness as possible.

There are three sets: Train (60% of random records), Test1 (20% of random records), and Test2 (20% ofrandom records).

I tried with 6 different algorithms:

1. Recursive Partitioning.
2. Decision Trees.
3. Random Forest.
4. Naive Bayes.
5. Support Vector Machines.
6. Evolutionary Learning. (Just tried, but no results)

Where only Naive Bayes provided “tangible” results, the worst was SVM with a lot of time consumption and processing without any results.

Basically, formula used on Naive Bayes algorithm worked, which is:

Mark ~ Attempt + Semester_en + EarnedECP + FY + SY + TY + EX + EXT.

Columns {Pass, Fail, Semester(0=Spring; 1=Autumn), Bachelor, Master, StateOfStudies_en(was changed as [1stYear=1, ..., ExchangeStudent=4, External=5, Unknown=0]), used to created {FY, SY, TY, EX, EXT} correspondingly }, were created as representation of "knowledge" from data, simplifying meaning, and to be more useful for algorithms, they were added with intention to be discrete variables. Only column EarnedECP is a continuous variable.

The rest of mentioned algorithm provided only probabilities of getting a non-specified either differentiated or non-differentiated mark. Files used for getting results are included on data folder, and an R image as well (wsTopic.RData).

As a result on example_result folder, you can see a file named nB_solution.csv, which basically is concrete result set since other algorithms (listed previously) did not provided good results, this file consists of three columns:

1. Column without specific name: sequential number of row.
2. StudentID: Student identification number.
3. Grade: The estimated possible result of a specific student (StudentID) based on formula described prebiously. Grade could be (Pass/Fail) for non-differentiated grades and A,B,…,F for differentiated.

Estimated possible grade is based on courses previously passed or failed, the number of attempts done for passing courses, the semester where student "is", the number of ECP that has “so far”, and if the student is from 1st, 2nd, 3rd year, if he or she is either external of exchange student.

This very-basic model answers following question:

"Based on current student status, what's the possible grade for a future course? Considering Pass/Fail for non-differentiated grades and A, B ... or F for differentiated grades".

In folder imas you can see a couple of results for prediction tryouts regarding Recursive Partitioning (RP) and Decision Trees (DT) algorithms which are not useful at all for estimating a grade. RP results is not enough structure and DT is too much for “estimating” and A for every single possible course, conclusion regarding DT and RP is they are not a good option for estimating a possible realistic grade.

A way to see how useful this model is, is to use a confusion matrix, in this case we have one for test1 and another for test2.

Test1:

25583	A	B	C	D	E	F	Fail	Pass
A	648	640	625	453	399	801	158	1558
B	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0	0
E	32	58	121	256	616	643	10	109
F	0	0	0	0	0	0	0	0
Fail	1851	1550	1732	1595	1389	3864	718	5693
Pass	19	5	7	4	1	20	1	7

In this table we can see model is useful for predicting grades such as: A, E, Fail, and Pass. Using confusionMatrix() function from library caret, we get result that in general, accuracy of this model is about 0.0777.

Test2:

25586	A	B	C	D	E	F	Fail	Pass
A	724	625	629	489	382	816	148	1514
B	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0	0
E	26	77	120	271	549	689	16	112
F	0	0	0	0	0	0	0	0
Fail	1815	1624	1569	1598	1420	3861	710	5736
Pass	20	8	10	4	2	11	0	11

Using same function confusionMatrix() from library caret, we get result in general, that accuracy of this model is about 0.0779.

Since results is considerable small to consider this model as relevant, it seems that the knowledge extracted from data is not enough for creating a model to be confident enough on it and predict a student grade accurately.