Predicting student grade

By Nawat Sunthornyanakit, Karn Phanawadee

Table of Contents

Points to discuss:

- Raw data analysis
- Excel data processing (Input & Output)
- What features should we use
- Type of machine learning use
- Final results

Goal

Our goal in this project is to predict student grade with the highest accuracy

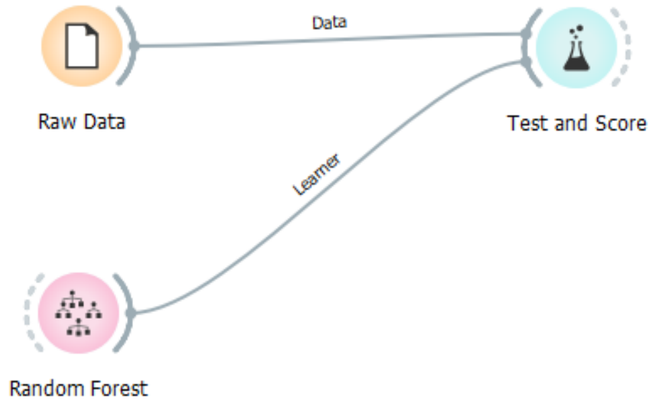
Raw data

- There are 8 columns and 50,000 rows of data including
  - FakeID
  - Year
  - Sem
  - System
  - Course
  - Credit
  - Grade

Problems

- Most of the data are catergoical but was presented as numerical
- Some data have too many catergorical (Course and FakeID)
- Missing values for systems
- Some of the data in the Unseen data is not in the seen data

Prediction using the raw data



Trying to predict without any data processing

Without any data processing the accuracy is

Test and Score

Settings

Sampling type: Stratified 5-fold Cross validation

Scores

Model	MSE	RMSE	MAE	R2
Random Forest	0.49716219886295926	0.7050972974440898	0.5137528110139861	0.15864522656990387

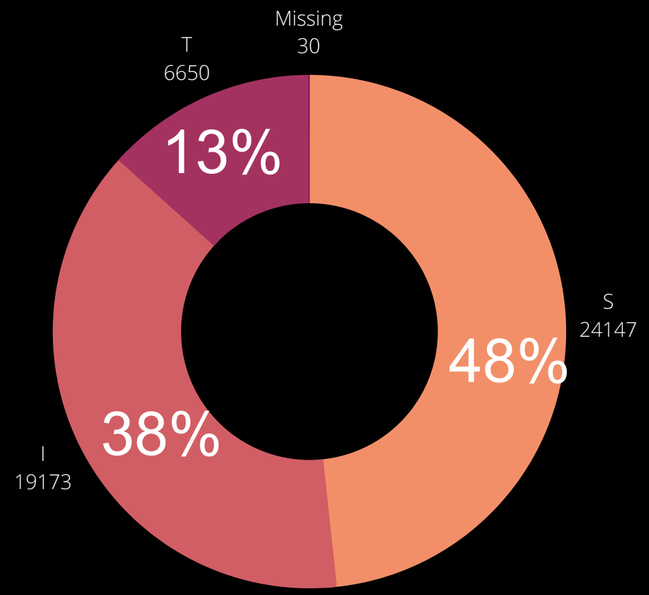
MSE = 0.497

Excel data processing

Missing values

Missing System

Using the mode we replace any missing value



1006	4716	2017	2	T	2	2949656	3	3.5
1007	1679	2012	2	I	1	2952303	4	1.5
1008	2977	2013	2	I	2	2601120	3	4
1009	506	2010	1		0	2952474	3	3.5
1010	691	2012	2	S	10	2605311	3	3.5
1011	1019	2012	1	S	1	2900305	2	4
1012	5073	2019	1	I	1	2952321	3	4
1013	679	2012	2	S	1	201122	3	4

30 Missing values

There are total of 30 missing values, so using the REPLACE function we replace the missing data with "S"

=IFS(X="",S)

Too much catergorical value (FakeID and Course)

How to deal with too many categorical values?

Using average we are able to replace FakeID and Course with the probability and mean of each value

Why not one hot encoding?

There are too many variables so if we decided to use 1 hot the list of features would be too long.

New features

2.Course_0-4

There are a total of 731 courses. For each course, we create 8 new features to get a probability of getting each scores.

FakeID	Fake ID fix	Year	Sem	System	Section	Course	Course_0	Course_1	Course_1.	Course_2	Course_2.	Course_3	Course_3.	Course_4	Credit	Grade
569	3.8	2012	1	S	2	201131	0	0	0	0	0	0	0.333333	0.666667	3	4
1770	3.823529	2011	3	I	1	2952216	0	0	0	0	0.052632	0.263158	0.473684	0.210526	3	4
570	3.15	2010	2	I	0	2940309	0.013187	0.030769	0.081319	0.186813	0.186813	0.173626	0.158242	0.169231	4	2.5
2067	3.416667	2011	2	T	1	2942612	0	0	0	0	0.036585	0.304878	0.402439	0.256098	3	3.5

1.Fake ID fix

Total of 5791 individuals. Each individual will get their own average score as a new feature.

What formulas did we use?

1. =SUMIF($F$2:$F$50001,F2,$H$2:$H$50001)/COUNTIF($F$2:$F$50001,F2)
2. =COUNTIFS($H$2:$H$50001,0,$F$2:$F$50001,F2)/COUNTIF($F$2:$F$50001,F2)

How to prepare the unseen data

Three main problems

1. Missing features
   - Import the feature from the seen data using match and index
2. Missing data
   - Replacing the missing data with the mode of the seen data
3. New data
   - Using the average we replace all of the new data with the average of all the seen data

=IFNA(MATCH(F2,Data!$G$2:$G$50001,0),50001)
=INDEX(Data!$A$2:$P$50002,Predict!$G3,Predict!H$1)

=IFS(X="",S)

=AVERAGE(B2:B20001)

What features should we use

Correlations

Pearson correlation

1	+0.601	Fake ID fix	Grade
2	+0.405	Course_4	Grade
3	-0.403	Course_2.5	Grade
4	-0.397	Course_2	Grade
5	-0.370	Course_1.5	Grade
6	-0.325	Course_1	Grade
7	-0.233	Course_3	Grade
8	-0.224	Course_0	Grade
9	-0.211	Credit	Grade
10	+0.071	Grade	Sem
11	+0.069	Course_3.5	Grade
12	+0.069	Grade	Year
13	-0.057	Grade	Section
14	+0.034	Grade	System (2)

Trial 1

-Fake ID

-Course

-Credit

-Year

Trial 2

All of the features

Trial 1

Test and Score

Settings

Sampling type: Stratified 5-fold Cross validation

Scores

Model	MSE	RMSE	MAE	R2
Random Forest	0.2516245699007402	0.5016219392139265	0.37037831249893594	0.5580207265648756
Gradient Boosting	0.25194391439766456	0.5019401502148085	0.36878402843335195	0.557459797046816 9

Trial 2

Test and Score

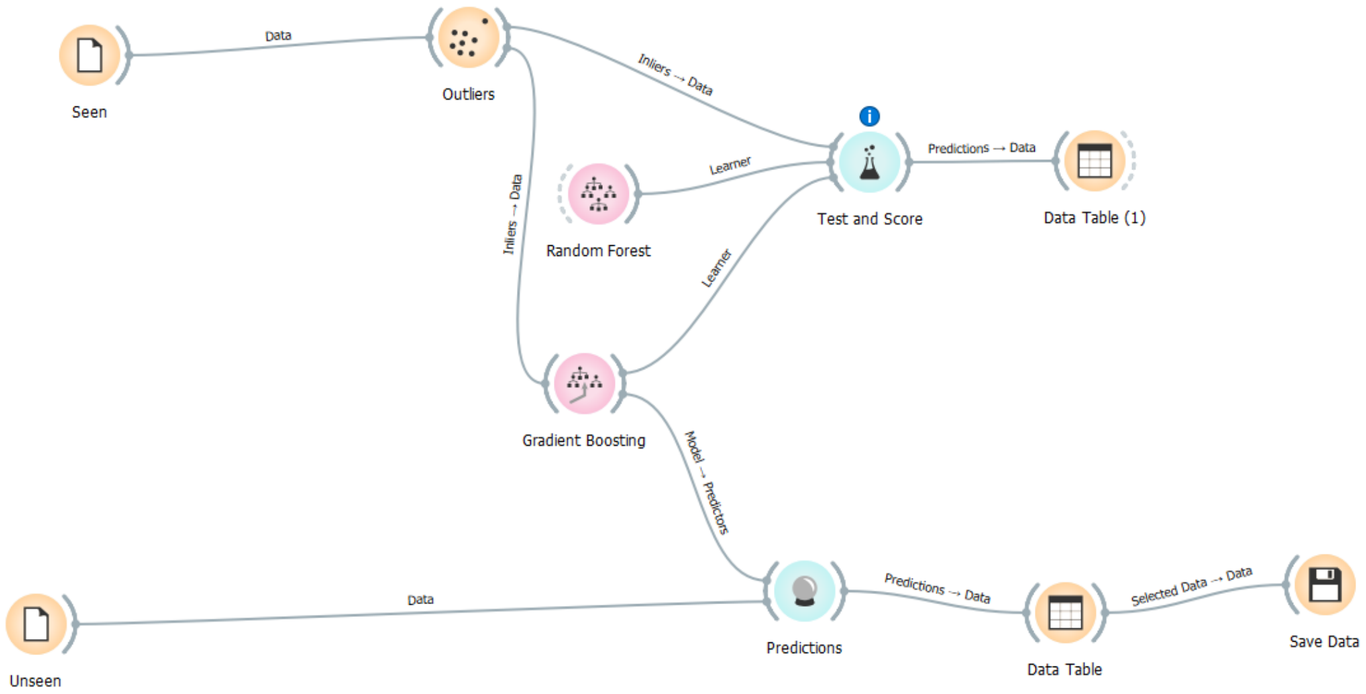
Settings

Sampling type: Stratified 5-fold Cross validation

Scores

Model	MSE	RMSE	MAE	R2
Random Forest	0.268489483642968	0.5181597086256012	0.3894679053945378	0.49131361923812955
Gradient Boosting	0.24439601681241532	0.4943642551928844	0.3646948811875259	0.5369616583186391

Types of tools used



01 Outliers

To get a more accurate model our team have decided to remove an outline before creating the model

02 Gradient boosting

Pros: Accurate, and work well with categorical data

Cons: Prone to overfitting, difficult to understand difficult tuning.

03 Random Forest

Pros: Works well on categorical data and is easy to understand

Cons: Might be prone to overfitting, high variance, not very accurate, and we can't guarantee an optimal trees

Result

50% improvement!!

Gradient Boosting vs Random Forest

We can see that Gradient boosting is slightly more accurate than the random forest method, with the MSE of 0.244 and 0.268 from Gradient boosting and random forest respectively.

Test and Score

Settings

Sampling type: Stratified 5-fold Cross validation

Scores

Model	MSE	RMSE	MAE	R2
Random Forest	0.268489483642968	0.5181597086256012	0.3894679053945378	0.49131361923812955
Gradient Boosting	0.24439601681241532	0.4943642551928844	0.3646948811875259	0.5369616583186391

THANKYOU