Case study on School test scores

DISCLAIMER:

This report describes a case study set up according to CRISP-DM in which regression is applied. Only the highlights of the study are presented!

The report should not be taken as a guideline for the elaboration of the final assignment! For that, the instructions in the final assignment and the rubrics for assessing the assignment should be followed.

Advanced topics are preceded by an asterisk (*): these topics are good to know but not mandatory.

Author: Ronald Fokkink
Date: 09/25/2022

Table of Contents

1. Business Understanding

We present two variations of the case study:

1. Prediction of school test scores
2. Explanation of school test scores

Prediction of School test scores

DM Goal

Predict the expected test scores for California schools given the school characteristics and the demographics of their students (regression task).

Business goal (fictional)

As the state of California, we want to compare test scores of schools using the SAT-9 test system to the test scores of schools using alternative test systems.

*Explanation of School test scores

DM Goal

Determine:

- which characteristics of California schools and the demographics of their students are related to school test scores and
- how much these factors 'contribute' to the test scores
- by developing a linear regression model

Business Goal (fictional)

As the state of California, we want to know what school characteristics can be used to improve school test scores given the demographics of the students. **This will only be possible if these characteristics are causal to school test scores, which cannot be conclude from the data analysis!**

NOTE: During model building (ch. 4), a linear regression model will be included. This model will be used as a predictive and an explanatory model, i.e. for both variations.

2. Data Understanding

Data collection

Dataset on test performance of schools in California and school characteristics and student demographic backgrounds.

Data description

Description of variables in the dataset (*school.xlsx*).

School and test performance:

- *ID* Unique identifier of the school
- *school* Name of the school
- *county* County where the school is located
- *test_score* Average test score on SAT-9 (Stanford 9 standardized test) administered to 5th grade students

School characteristics:

- *students* Total numbers of students enrolled in the school
- *teachers* Total number of teachers (in FTE) at the school

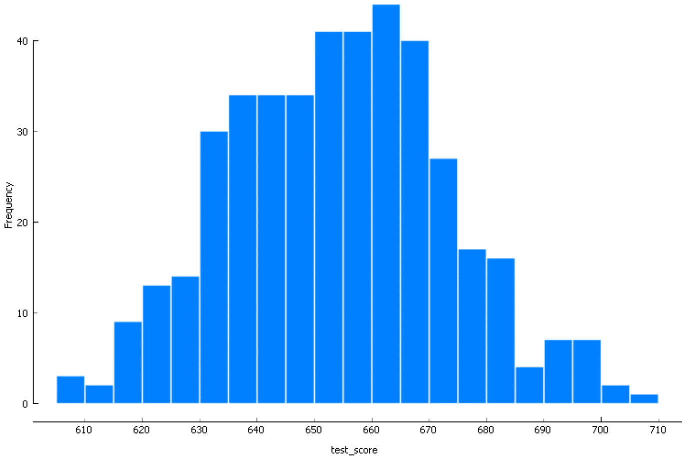
Student demographic backgrounds:

- *english_pct* Percentage of students that are English learners (that is, students for whom English is a second language) in the district where the school is located
- *income* Average income (in USD 1,000) in the district where the school is located

Data exploration

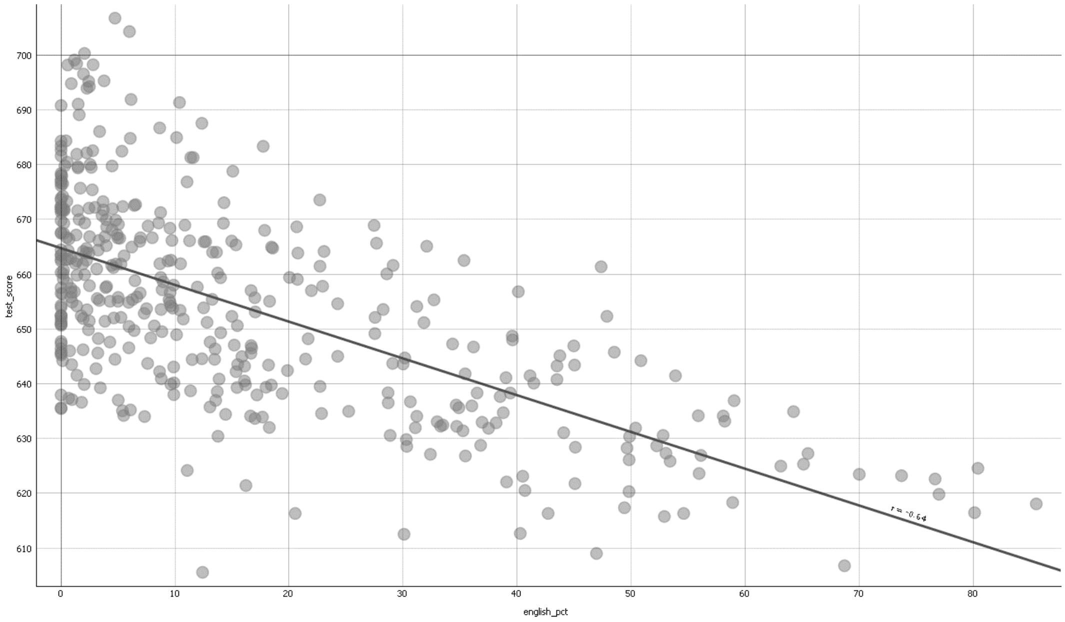
Explore the distribution of the variables in the dataset (histograms):

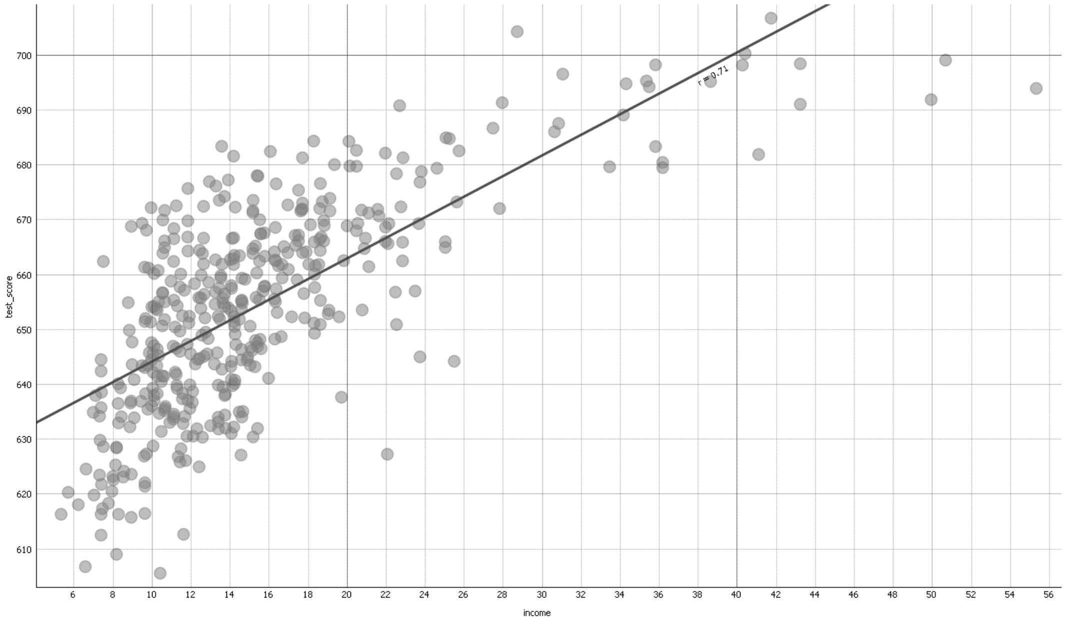
- Check for missing values and duplicates (there are none)
- Investigate outliers (f.i. high income districts) and missing values (there are none)
- The distribution of the *test_score* variable is quite symmetric. Thus, there is no need to consider (logarithmic) transformation of the dependent variable to improve the predictive performance of models (such as random forest).



Explore the joint distributions between the dependent variable *test_score* and each independent variable to identify and examine relationships (scatterplots and Pearsons's r):

- There exist strong relationships with *english_pct* (r = -0.64) and *income* (r = 0.71)
- There exist very weak relationships with *students* (r = -0.15) and *teachers* (r = -0.14)
- The ratio of the number of students and teachers might be (negatively) related with the dependent variable. Therefore, the derived variable *student_teacher_ratio* = *students* / *teachers* was created. However, there is only a weak relationship with this variable (r = -0.23).





3. Data preparation

Variable selection
All variables for school characteristics and student demographic backgrounds are included in the dataset for model building. Since there are enough observations (420), there is no need to reduce the number of variables to improve the performance of the prediction models, in particular the linear regression model.

Variable construction
The variable *student_teacher_ratio* was created as explained earlier.

Cleaning, integrating and formatting
Not applicable.

4. Modeling

Model building

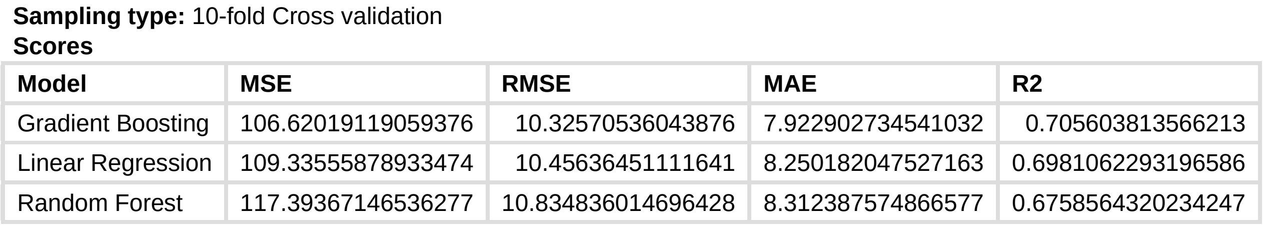
The following prediction models were build using Orange 3.31:

- Linear regression
- Random Forest
- Gradient Boost.

The first model is a linear model that is also used for the explanatory variation of the study. The other models are nonlinear ensemble models that usually have high predictive performance.

Model assessment - Prediction of School test scores

The predictive performance of the models is presented below using various measures (report of Orange's *Test and Score* widget).



Sampling type: 10-fold Cross validation

Scores

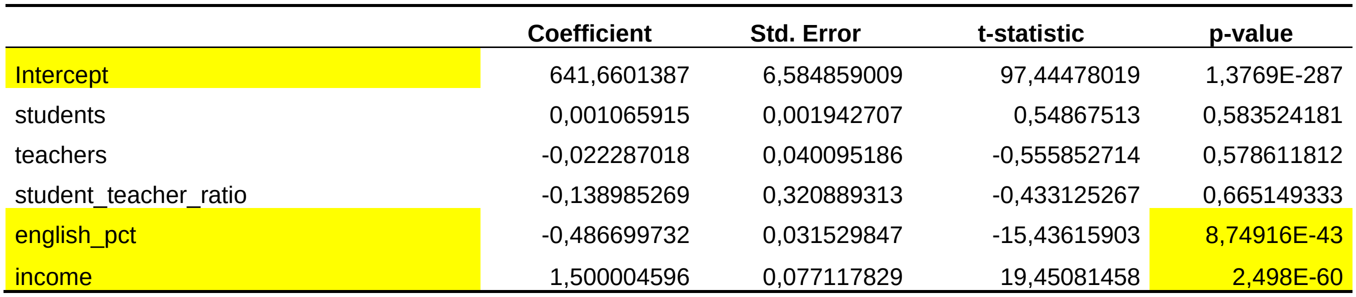
Model	MSE	RMSE	MAE	R2
Gradient Boosting	106.62019119059376	10.32570536043876	7.922902734541032	0.705603813566213
Linear Regression	109.33555878933474	10.45636451111641	8.250182047527163	0.6981062293196586
Random Forest	117.39367146536277	10.834836014696428	8.312387574866577	0.6758564320234247

The Gradient Boosting model has the highest accuracy based on the MAE (Mean Absolute Error). However, the differences between the MAE of the models are small relative to the mean test score of 654.

***Model assessment - Explanation of School test scores**

The coefficients of the Linear Regression model are available as output of Orange's *Linear Regression* widget. The small coefficient values (almost zero) of the independent variables *students* and *teachers* are an indication that these variables are not significant.

Using Excel's *Analysis Toolpak*, we get additional information about the significance of the coefficients. See the report below.



	Coefficient	Std. Error	t-statistic	p-value
Intercept	641,6601387	6,584859009	97,44478019	1,3769E-287
students	0,001065915	0,001942707	0,54867513	0,583524181
teachers	-0,022287018	0,040095186	-0,555852714	0,578611812
student_teacher_ratio	-0,138985269	0,320889313	-0,433125267	0,665149333
english_pct	-0,486699732	0,031529847	-15,43615903	8,74916E-43
income	1,500004596	0,077117829	19,45081458	2,498E-60

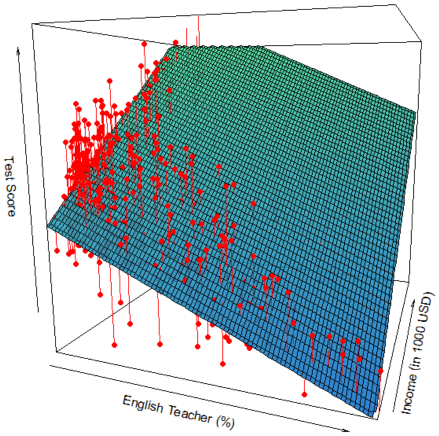
From this we can conclude that the variables *students* and *teachers* are indeed not significant, as the p-values are (much) greater than 0.05. **Note that the information presented is valid only if certain conditions, such as linearity, normality and homoscedasticity, are met!!**

Omitting the non-significant variables yields the following linear regression formula:

$$test\_score = 638.93 - 0.49 \times english\_pct + 1.50 \times income$$

Interpretation:

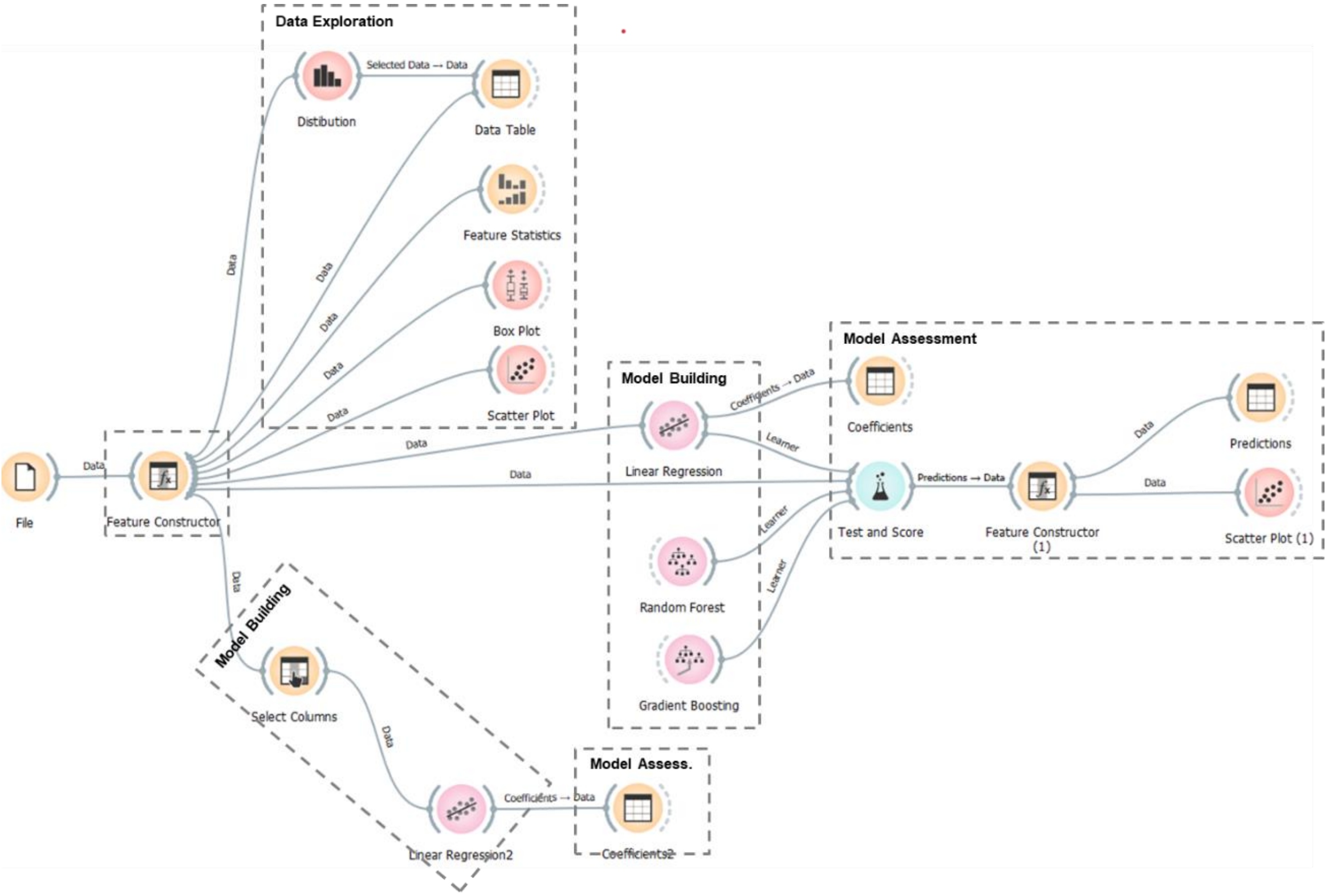
- 1 unit increase in pct. of English learners is associated with a decrease in average test score by .49 point, all other variables held constant
- 1,000 USD increase in average income is associated with an increase in average test score by 1.50 point, all other variables held constant.



We only have variables on student demographic backgrounds, none on school characteristics.

As for answering the business objective, we found no school characteristics related to school test scores. Thus, the study found no school characteristics that the state of California could use to improve school test scores, given the demographic characteristics of the students.

Orange workflow



Appendices

• *school performance.ows*	Orange workflow
• *school.xlsx*	dataset