Predicting Student's Portuguese Final Grade based on Math Performance using CV Lasso

Wing Cheong Lo
ECN 190
University of California, Davis

Keywords

Machine Learning (ML), Data mining (DM), CV Lasso, Nonparametric Bootstrap ACT/SAT, higher education, prediciton

Abstract

Discovering youngsters with the highest potential to receive higher education is always a great challenge to the admission office of every university. Between 1967 and 2012, the share of workers with minimum a bachelor's degree has changed 19% positively (Carnevale 13). We believe having a higher portion of postsecondary holders is definitely beneficial to the U.S economy, however seats available for university are still scarce and ACT/SAT could be biased since some students are bad exam takers. The admission office may be interested in our concern in this paper, that is whether students' ACT/SAT scores can be estimated solely using their background and normal school performance rather than taking ACT/SAT under great pressure. Although we understand that this method may have much more problems than actual examination, the results can be used as a reference by the admission office. Therefore, we are interested in methodology to predict exam scores in secondary school, so that students who don't do their best in ACT/SAT may also be reconsidered. In this paper, although we use datasets unrelated to ACT/SAT, we are trying to examine whether students' final scores, G3 can be predicted precisely with DM method.

Introduction

The arise of ML and data science have changed the potential future development of

econometrics. Data is frequently described as the future petroleum. In 2008, Professor Paulo Cortez of the University of Minho has already proved students' grades can be predicted effectively under the application of 4 DM algorithms (D-Tree, Random Forest, ANN and Support Vector Machine) with the datasets we are going to use (Cortez 2). However, when Professor Paulo attempts to seek out the best DM model using two datasets independently, we try to use the datasets dependently. In this paper, we will try to use a different DM method, CV Lasso to select the best regression model. And then use Nonparametric Bootstrap to generate a confidence interval for our prediction. The datasets have same numbers of regressors and different numbers of observations. In particular, student-mat.csv (Math performance) is our training dataset and student-por.csv (Portuguese performance) is the test dataset. We believe this usage is beneficial to prediction. The datasets are real world statistics and were obtained from two schools in Portugue. One concern is, using math performance to predict portuguese may be ineffective, as portuguese students may have better performance on portuguese than mathematics. Therefore, in this paper, our major goal would be if it is useful to predict one's potential solely based on his background and concentration to the school, rather than his talent on the subject. G3, the final score is our parameter of interest.

Methodology

Dataset Description

In our training dataset (see next page for description of attributes), student-mat.csv, we have totally 395 observations, while that of student-por.csv have totally 649 observations. Both datasets have 33 same attributes. Some observations in both datasets are the same person and we will discuss more how we modify our dataset in

Table A-1 : Description of the dataset

```
# Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese
language course) datasets:
1 school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)
2 sex - student's sex (binary: "F" - female or "M" - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: "U" - urban or "R" - rural)
5 famsize - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
6 Pstatus - parent's cohabitation status (binary: "T" - living together or "A" - apart)
7 Medu - mother's education (numeric: 0 - none,  1 - primary education (4th grade), 2 – 5th
to 9th grade, 3 – secondary education or 4 – higher education)
8 Fedu - father's education (numeric: 0 - none,  1 - primary education (4th grade), 2 – 5th
to 9th grade, 3 – secondary education or 4 – higher education)
9 Mjob - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g.
administrative or police), "at_home" or "other")
10 Fjob - father's job (nominal: "teacher", "health" care related, civil "services" (e.g.
administrative or police), "at_home" or "other")
11 reason - reason to choose this school (nominal: close to "home", school "reputation",
"course" preference or "other")
12 guardian - student's guardian (nominal: "mother", "father" or "other")
13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 -
30 min. to 1 hour, or 4 - >1 hour)
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10
hours, or 4 - >10 hours)
15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or
no)
19 activities - extra-curricular activities (binary: yes or no)
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)
22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29 health - current health status (numeric: from 1 - very bad to 5 - very good)
30 absences - number of school absences (numeric: from 0 to 93)

# these grades are related with the course subject, Math or Portuguese:
31 G1 - first period grade (numeric: from 0 to 20)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)

Additional note: there are several (382) students that belong to both datasets .
These students can be identified by searching for identical attributes
that characterize each student, as shown in the annexed R file.
```

the next section. We are now going to describe what attributes in these two datasets. Provided by the dataset, there are 30 attributes related to one's background and overall school performance while G1, G2, G3 are related to course subject. Among these attributes, some of them are categorical variables. What we care the most

in the dataset is correlation between G3 and students' background, his school performance. A very interesting point is one of the variables, health, has a negative interpretation on course grade. We will talk about this later. We should also notice that G1, G2 have a high correlation with G3, since G3 is the final grade of the course. But we will exclude G1, G2 from our regression model since we believe this enhances predictability of the model.

Dataset Modification

There are some important modifications that we will make to the dataset.

First of all, the same person between both datasets will be excluded from student-por.csv to avoid overfitting. We are making this decision because student-por.csv has more observations than student-mat.csv and we want there to be more observations in our training dataset than test dataset. There are totally 382 observations overlapping. A simple linear search will be used to sort these samples, based on suggestions by the dataset editor in student-merge. They are: *{school, sex, age, address, famsize, Pstatus, Medu, Fedu, Mjob, Fjob, reason, nursery, internet}*

Secondly, we will modify some categorical variables to be dummy variables. Notice we decide not to modify all numeric categorical variables (i.e Fedu) as they are ratings. Now we are having 59 attributes. The modification of those variables will be presented below as a table in the next page. Meanewhile, in order to avoid perfect multicollinearity from dummy variables trap, we are going to drop these variables from our regression model : *{school.MS, sex.M, address.U, famsize.LE3, Pstatus.T, Mjob.other, Fjob.other, reason.other, guardian.other, schoolsup.no, famsup.no, paid.no, activities.no, nursery.no, higher.no ,internet.no, romantic.no}*

which means that we are reducing the number of attributes from 59 to 42.

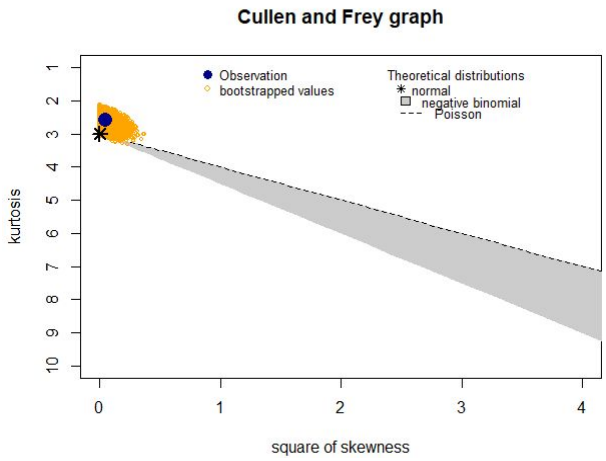
Table A-2 : Modification of Dataset

Original Variable	*Converted Variables*
School	*School.GP, ~~School.MS~~*
Sex	*Sex.F , ~~Sex.M~~*
Address	*Address.R, ~~Address.U~~*
Famsize	*~~Famsize.LTE3~~, Famsize.GTE3*
Pstatus	*Pstatus.A , ~~Pstatus.T~~*
Mjob	*Mjob.at_home, Mjob.health, Mjob.teacher, Mjob.services, ~~Mjob.other~~*
Fjob	*Fjob.at_home, Fjob.health, Fjob.teacher, Fjob.services, ~~Fjob.other~~*
Reason	*Reason.course, Reason.home, Reason.reputation, ~~Reason.other~~*
Guardian	*Guardian.father, Guardian.mother, ~~Guardian.other~~*
Schoolsup	*Schoolsup.yes, ~~Schoolsup.no~~*
Famsup	*Famsup .yes, ~~Famsup.no~~*
Paid	*Paid.yes , ~~Paid.no~~*
Nursery	*Nursery.yes, ~~Nursery.no~~*
Higher	*Higher.yes, ~~Higher.no~~*
Internet	*Internet.yes, ~~Internet.no~~*
Romantic	*Romantic.yes, ~~Romantic.no~~*

****Variables not included (i.e. G1,G2,G3) above have not any modification***

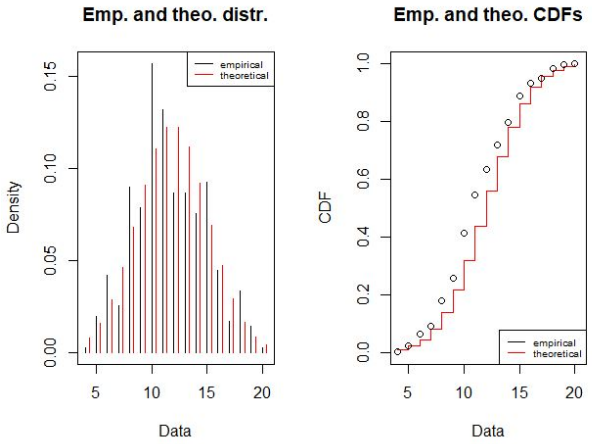
Last but perhaps the most important one, we decide to drop observations who get 0 in their final from both test and train datasets. Professor Paulo doesn't make such a change in his five-level classification. Here, we are trying to predict an actual score of a student and we think it would be better to treat '0' grade students as outliers. Our range of grade would now be (1,20) in the modified training dataset.

Table A-3 : modified Math G3 distribution



**** Yellow area is the result of 100000 Nonparametric Bootstrap, as an estimation of sample Skewness and Kurtosis of G3 in the modified train Dataset, trainDataSet_reduced_dummy***

Table A-4 : modified G3 fit with Bell Curve



**** fitting Method is Maximum likelihood Estimation***

Data Distribution -- Modified Math G3

We are particularly concerned about the distribution of our prediction target, G3/Final Grade in our training dataset. Table A-5 and A-3 have indicated the kurtosis and skewness are closed to 3 and 0 respectively, hinting that a normal distribution may fit the data. In Table A-3, we use Nonparametric Bootstrap to estimate sample skewness and kurtosis since we care whether the distribution of G3 fits a normal distribution or not. We also run a Jarque Bera Test (See eq. (1.2) (Thadewald 5) and the p-value is 0.07723, meaning null hypothesis as S = 0 and K = 3 cannot be rejected when P-value > 0.05.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (1.1)$$

$$\mathrm{JB} = \frac{n}{6} \cdot \left(S^2 + \frac{(K-3)^2}{4}\right) \quad (1.2)$$

Technically, it is wrong to describe it as a normal distribution as $\boldsymbol{\mu}$ in eq. (1.1) can't theoretically be zero. However, it is fine that $\boldsymbol{\mu} \pm 2\sigma$ includes most examples, as examples like SAT score can be distributed by a bell curve. The P.D.F should be discrete as the values range from 0 to 20 and is integer.

Table A-5 : Summary of modified G3

```
MathG3_distribution
summary statistics
------
min:  4   max:  20
median:  11
mean:  11.52381
estimated sd:  3.227797
estimated skewness:  0.2092559
estimated kurtosis:  2.598051

        Jarque Bera Test

data:  trainDataSet_reduced_dummy$G3
X-squared = 5.122, df = 2, p-value = 0.07723
```

Extra Information about the datasets

The datasets were collected in two secondary schools via questionnaires on

paper sheets from students in 2005 - 2006 school years. The questions were first sent to school professionals to get feedback and edited then distributed. (Cortez 2)

Model

In this paper, we are going to use CV-lasso to select the turning parameter λt.

As we have discussed earlier, we find data potentially fit normal distribution, therefore in our glm model, we will use "gaussian":

glm(G3 ~.-G1-G2, family=gaussian)

The in-sample R^2 of the model is 0.337484.

After that, we are going to use Cross Validated along with Lasso regularization to shrink variables which are not good for prediction.

L1 Regularization Formula:

$$\hat{\beta}_{\lambda} = \arg\min_{\beta} \left\{ \frac{1}{n} dev(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \quad (1.3)$$

Penalty Weight λ in eq. (1.3) is a sign-to-noise filter. When λ is higher, more information is shrinked. Notice that lasso alone is unable to do any model selection (Matt 79). To select the best model for prediction, we will use CV-lasso for best OOS predictive performance.

In our model, we decide to use CV-min rule as we are focused on OOS predictive performance, while our K in eq. (1.4) is 5.

K-Fold CV Lasso Formula:

$$\hat{\beta}_t^k = \arg\min_{\beta} \left\{ \frac{1}{n \cdot (K-1)/K} dev_{-k}(\beta) + \lambda_t \sum_{k} |\beta_j| \right\} \quad (1.4)$$

Table A-6 : CV-lasso

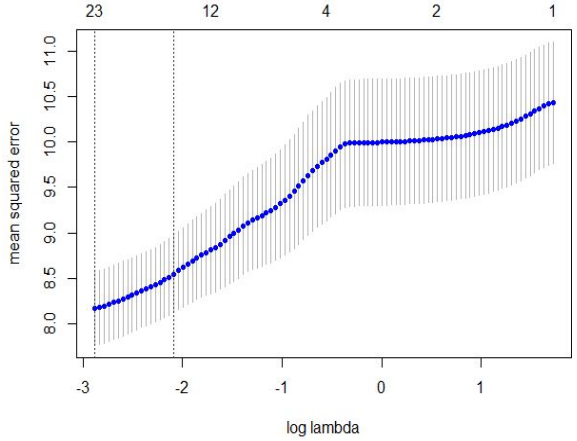


Table A-7 : Lasso Path Plot

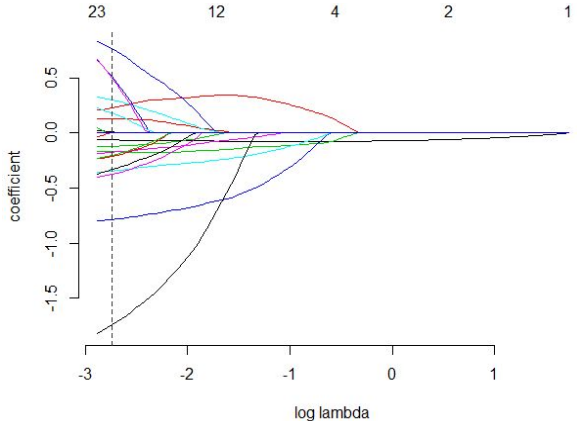


Table A-7 shows λ selection when we choose CV-min rule. Almost one-third of the attributes are shrinked from our glm model. Now, for the next step, we are going to predict the test data set and see the result.

A smart conclusion on the result

Table A-8 shows coefficients of our penalized glm model under CV-min rule.

Table A-9 : Predict G3 of modified Test DS

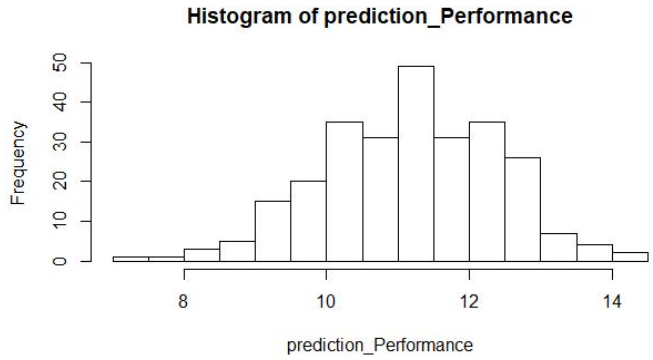


Table A-10 : Actual G3 of modified Test DS

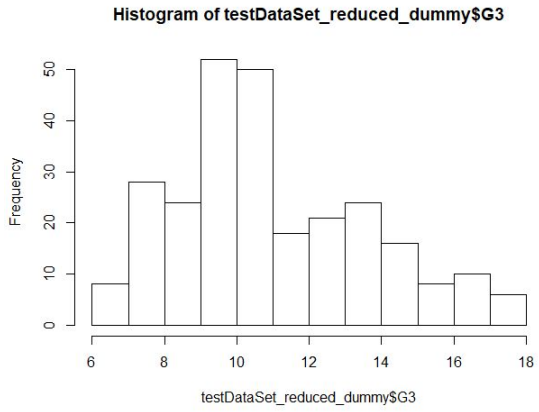


Table A-8 : Real Result in Data Set

```
> betamin
40 x 1 sparse Matrix of class "dgCMatrix"
seg100
intercept          14.959076586
school.GP           0.048690068
sex.F              -0.398690857
age                -0.124724527
address.R          -0.229722336
famsize.GT3        -0.033988131
Pstatus.A           .
Medu                0.206874184
Fedu                0.133356454
Mjob.at_home        .
Mjob.health         0.662964205
Mjob.services       0.831540208
Mjob.teacher        .
Fjob.at_home        .
Fjob.health         .
Fjob.services       .
Fjob.teacher        0.668839544
reason.course       .
reason.home         .
reason.reputation   .
guardian.father     .
guardian.mother     .
traveltime          .
studytime           0.329344691
failures           -0.798264428
schoolsup.yes      -1.819176976
famsup.yes         -0.369352431
paid.yes           -0.232358549
activities.yes      .
nursery.yes         .
higher.yes          .
internet.yes        0.225895026
romantic.yes        .
famrel              0.024980076
freetime            0.005866417
goout              -0.358618113
Dalc                .
Walc               -0.163289630
health             -0.183947061
absences           -0.061673218
```

Table A-9 describes the distribution of 265 predicted G3 using the model in Table A-7. Table A-10 describes the distribution of 265 actual G3 in our modified dataset (see p.5).

In brief, we feel delighted as the distributions are similar, however A-9 looks more left-skewed. The distance between actual value and predicted value is 1.95044 averagely (We calculate by adding the sum of distance and divides by 265). We are not yet satisfied by our result. From here, we will try to use another method, Nonparametric Bootstrap, to calculate a 95% confidence interval for the predicted value.

Confidence Interval for Prediction

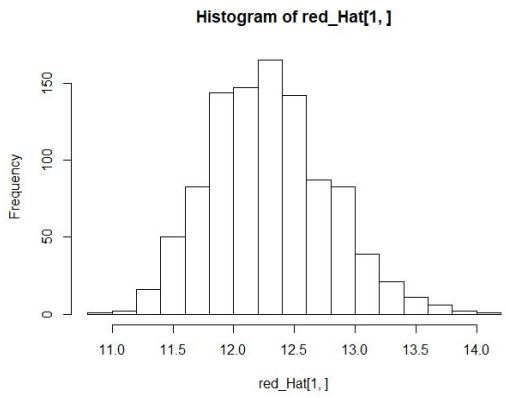
Lasso Regularization is more biased when λ increases, because more information is lost. Table A-8 shows over one-third regressors are penalized. Considering this fact, we decide to use Nonparametric Bootstrap to make Confidence Interval for G3. Although Nonparametric Bootstrap doesn't work with model selection, it should be fine if we use it to generate C.I. for predicted G3 (Taddy 99). Moreover, Amiri shows that performance of parametric and nonparametric bootstrap depends on sample kurtosis (Amiri 9). If sample kurtosis, K^*, in eq. (1.4) is greater than kurtosis, K, then Parametric bootstrap should be considered as it is more accurate for variance estimation in small sample groups.

$$K^* = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4/n}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2/n\right)^2} - 3. \qquad (1.4)$$

Notice that we have made 100000 Nonparametric bootstrap to estimate sample kurtosis and skewness of G3 in Table A-3. There are more yellow plots (bootstrap values) below the observation than that of above, which implies more samples having a higher kurtosis than the observation. However, we are using Nonparametric bootstrap because we want to give more bias on prediction when lasso is biased. Nonparametric bootstrap would although give us a less accurate, but wider confidence interval for the prediction value. Since we care about OOS predictive ability, using nonparametric methods should be favored.

Result

Table A-11 : 1st Observation prediction



Normal Interval of Bootstrap CI :

$$T_n \pm z_{\alpha/2}\,\widehat{\text{se}}_{\text{boot}} \qquad (1.5)$$

Eq. (1.5) has indicated the easiest Confidence Interval for Bootstrap (Wasserman 108). Here, we decide to use a 95% C.I. to see whether our prediction is closer or not. Our result is : (11.32434 13.26833), with S.D. = 0.2440211 The actual value of first Observation in the test dataset is 13. We have a result of 12.89015 for first observation using solely CV-lasso. Since Lasso is biased and this result may be out of luck, it would be good for us to construct a C.I. We now run a for loop to see how many G3 values fall inside the intervals. Table A-12 shows that 101 out of 265 observations fall into the interval. We believe the percentage is already high, near 40% of successful rate. Notice we are using two datasets for different purposes, one as test dataset and one as train dataset. If we have a super high successful rate, we may make a mistake in terms of overfitting. In this paper, what we really care about is model's OOS (out-of-sample) predictive ability, rather than their in-sample predictive ability.

Table A-12 : Actual inside 95% C.I.

```
> # Run a test to see if the predicted value falls into the cofidence interval
> pass = 0
> for (w in 1:nrow(red_Hat)) {
+   lower_limit = mean(red_Hat[w,] - 2*sd(red_Hat[w,]))
+   upper_limit = mean(red_Hat[w,] + 2*sd(red_Hat[w,]))
+   target      = testDataSet_reduced_dummy$G3[w]
+   if (target > lower_limit & target < upper_limit) {
+     pass = pass + 1
+   }
+ }
> pass
[1] 101
>
```

Notice that we use Parametric Bootstrap for estimation, implying that we have a more accurate, but shorter interval for Confidence Interval, as we have a small sample.

Conclusion

In this paper, we attempt to create a DM model for score prediction. Actual score prediction, however, is very hard to do

precisely as we cannot hold Conditional Ignorability to be true, that is, we believe that we control all elements affecting the result. However, when we are data miners, observators, instead of the dataset collectors, we are not familiar with our dataset. Moreover, observation error can exist from merely doing questionnaires, i.e. students can hardly remember how long they study. We also notice that multiple assumptions, i.e. the modified train dataset is normally distributed, '0' G3 observations should be omitted, bootstraps are used correctly, must be correct. Otherwises, our logic can be entirely wrong. However, this paper just tries to discover potential method for prediction, rather than proving something. Last but the foremost, we choose not to continue the work and follow the path of Professor Cortez, and select to use the datasets dependently. This is because we care about overfitting and OOS predictive ability on accurate grades the most in this paper. After the work, we believe there is a potential for the admission department to consider using ML methods, like ANN, to estimate a student's potential and performance on the test. Yet, our model may still have a lot of problems.

In p.4, we observe that health is negatively correlated with G3. Table A-8 shows that health is also included in the penalized model. However, we are unable to discover the reasoning behind, making this as a regret, when variable health is a rating of health status (from 1 to 5).

Acknowledge

This paper is finished in March 2020 under the supervision of Prof. Takuya Ura and Francis Graham. We (I tend to write I as We, it is just a habit) are glad for their assistance, and believe both of them are discovering the latest development of

implementation of DM/ML on econometrics.

This work is finished by an undergraduate student, potentially exists lots of mistakes and should not be treated seriously.

We also want to thank Matt Taddy, once professor of econometrics in the University of Chicago, as he leads us in a beautiful journey in machine learning.

Reference

Taddy, M, *Business Data Science,* New York, McGraw-Hill Education, 2019

Wasserman, L, *All of Statistics: A Concise Course in Statistical Inference*, Springer-Verlag New York Inc, 2004

Thadewald, Thorsten & Buning, Herbert. (2007). Jarque-Bera Test and its Competitors for Testing Normality - A Power Comparison. Journal of Applied Statistics. 34. 87-105. 10.1080/02664760600994539.

Carnevale, Anthony P.; Rose, Stephen J, "The Economy Goes to College: The Hidden Promise of Higher Education in the Post-Industrial Service Economy", Georgetown University Center on Education and the Workforce, 2015

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds, Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

Amiri, S and von Rosen D. and Zwanzig S., On the comparison of parametric and nonparametric bootstrap, U.U.D.M. Report 2008:15, 2008