Science and Technology and Social Development Proceedings Series

Article

2024 International Conference on Education, Economics, Management, and Social Sciences (EMSS 2024)

Financial Aid and Educational Outcomes: A Study of Grant Effectiveness in Rural China

Siran Jin [1,*]

[1] Shanghai High School International Division, 1399 Dingxiang Road, Shanghai, China
* Correspondence: Siran Jin, Shanghai High School International Division, 1399 Dingxiang Road, Shanghai, China

Abstract: This paper investigates the effectiveness of financial grants in rural China. The financial grant system's selection criteria are difficult to quantify and subjective because they vary from school to school. Statistical methods such as ANOVA are utilized to explores correlations between student academic performances with their socio-economic backgrounds. From the experiments conducted, it can be shown that certain factors such as attitude towards study and grade contribute more than other factors to the overall academic performances. Furthermore, machine learning methods are incorporated into the study to try to determine non-linear relations between academic performances with socio-economic factors. According to the results, it can be concluded that the current allocation model of financial grants in rural China is insufficient. Thus, in this article the author proposes a more scientific and statistical approach of such allocation.

Keywords: financial grants; education; China; machine learning; statistical methods

Published: 02 October 2024

1. Introduction

Financial grants are free and accessible monetary aids that can cover most of the tuition fees for students who are unable to pay for their education. In the United States, the average scholarship and grant aid award for students attending four-year public schools in 2020-21 was $7,813 [1], with 610,000 new students will benefit from the Pell program, which is a financial grant program aiming to fund students for accessible education [2]. United States has established a well-developed and influential financial granting system.

On the other hand, in China, financial grants for students are inefficient and comparatively unorganized. First and foremost, minimal financial grants are provided by the government, ranging from 750 CNY to 1500CNY per year for middle school students and about 2000CNY per year for high school students. Thus, private enterprises and schools are responsible for providing a certain portion of the financial grants, contributing to the differentiation of financial grants in both distribution channel and amount of aid provided. Moreover, the six main requirements for applying for federal financial grants: supporting the Chinese government, aligning with Chinese laws and regulations, possessing outstanding morals and ethics, consistently working hard, and suffering from financial dilemmas [3], are unable to quantify as specific qualitative requirements, including the average annual income to be determined as "poor," are not provided.

There are minimal papers addressing the inefficiency and differentiation of financial grants in China, while the few existing papers base their suggestions more upon experience rather than scientific analysis. Some papers suggest rating ideology and morality, attitude towards study, daily mentality, and economic hardship with weights of 0.1, 0.1, 0.2, and 0.6, respectively [4]. While other papers rate the student's family wellbeing at 60%, ideology at 15%, academic performance at 15%, and miscellaneous, consisting of community contribution, part-time work experiences, and club attendance, at 10%. The weights of the financial grants' measurements lack scientific reasoning, indicating that these measurements are solely based upon experience.

In this paper, the author proposes that financial aids have a varying degree of impact on students' performance. Geographical and other social-economic factors also contribute to predicting the overall performances of students, with some of these factors more important than others. To test the hypothesis, this paper will compare the academic performance of students who received grants with students who did not in both rural and urban regions through ANOVA and determine the difference in the impact of grants in rural and urban areas. Moreover, the author utilizes other more advanced modeling methods such as machine learning to construct a mapping from a range of social-economic factors to the overall academic performance of students.

2. Existing Studies

Over the past decades, multiple countries had tried to make education affordable for students with financial difficulties. The Pell Program, or financial granting program in the United States, is acknowledged to have positive effects on students' academic performance. Significant increases in full-time enrollment (between 4 to 7%) and suggestive evidence of increases in GPAs (between 0.06 and 0.08 points) are demonstrated by students receiving Pell Grants [5]. Yet, there are still some potential flaws in the grant system. For instance, low-income students do not receive the full benefit of these subsidies[6]. Other studies propose reforms of the Pell Program by adding support services for freshmen, lowering the eligibility and application difficulty for all students, and creating completion incentives that urge students to devote time to their studies [7].

Similarly, in China, flaws in the granting system are present. Student loans and need-based grants that are irrespective of the applicants' academic performance are the main funding sources for financially disadvantaged students. However, these financial grants that disregard students' previous academic performance show no significant positive effects. A direct correlation between student loans or need-based grants and students' post-graduation plans is absent [8]. Due to the limited research devoted to this field in China, financial grants, and especially the effective distribution of them, are largely unexplored territory.

2.1. Statistical Methods

Hypothesis testing, formalized in the twentieth century by R.A. Fisher and Jerzy Neyman with Egon Pearson, is a scientific process to either reject the null hypothesis, which is the basic hypothesis, in favor of the alternative hypothesis, the hypothesis that completely opposes the null hypothesis, or to fail to reject the null hypothesis [11]. It consists of five crucial steps: state the null and alternative hypotheses, collect data, perform a statistical test, reject or fail to reject the null hypothesis, and interpret the results [12]. Hypothesis testing attempts to seek statistically significant proof that builds from either the null or alternative hypothesis. As one hypothesis is favored, the other is rejected because of their mathematical asymmetry [13].

One of the statistical tests performed in hypothesis testing is ANOVA (analysis of variance), which compares means of a continuous variable in two or more independent comparison groups [14]. The F test statistic and p-value are the two most important outputs in a one-way ANOVA test. The F test statistic represents the ratio of mean square variances between groups (MSB) to mean square variances within groups (MSW) [15], where the p-value indicates the probability of the result occurring under the null hypothesis[16]. The F-value is calculated by the equation below [17]:

$$F_{score} = \frac{\frac{SSB}{DF_B}}{\frac{SSW}{DF_W}} \qquad (1)$$

Where SSB is the between sum of square, SSW is within sum of square, DF_B is the degree of freedom between groups, and DF_W is the degree of freedom within groups [18]. Only when the p-value is lower than a certain value (mostly 0.05) can the null hypothesis be rejected.

2.2. Machine Learning

Artificial intelligence has been thriving in recent decades, and several novel statistical methods have been introduced. The machine learning algorithm consists of three parts: decision processing, where the AI, based on some input data, produces an estimate of a pattern in the data; error functioning, where known examples of errors help the AI evaluate its prediction; and model optimization, where weights are adjusted to reduce the discrepancy between the known example and the model estimate until a threshold of accuracy has been met [19]. There are four types of machine learning, where supervised learning involves regression and classification. [20].

A typical type of supervised machine learning is the decision tree. A decision tree is a nonparametric machine learning algorithm that contains a hierarchical tree structure that comprises roots, branches, internal nodes, and leaf nodes [21]. The tree selects the best attribute using Gini impurity, entropy, and information gain; splits the dataset based on the selected attribute; and repeats the process until a stopping criterion is met or all data perfectly align with the tree [22]. Entropy of a set of data with n different values is calculated by the equation 2 below [23]:

$$Entropy = -\sum_{i=1}^{n} P_i \times \log_2 P_i \qquad (2)$$

Where P_i is the probability of obtaining the i^{th} value when randomly selecting one data from the set. Information gain (I) is found by the equation 3 below [24]:

$$I = E_i - \overline{E_j} \qquad (3)$$

Where E_i is the initial data's entropy and $\overline{E_j}$ is the weighted average entropy of subsets of initial data. However, the tree exhibits a vital flaw: overfitting. The decision tree fits the tree to ensure it perfectly aligns with the training data while overlooking potential differences between the training and real data [25]. Stopping the tree from growing when the data split is not statistically significant, acquire more training data, or remove irrelevant nodes eliminate overfitting. One method to prevent overfitting is to normalize the data using the z-score, scale normalization, and min-max normalization.

3. Methodology and Results

3.1. Data Collection

The data discussed in this paper were collected by the author through several sources, including face-to-face interviews with the director of the Office of Academic Affairs at the three schools, on-campus investigations, and online platforms. Data from two schools in rural China were used in this research. The first school, School A, is located in southeastern China near the East China Sea, and 87 fourth grade to eighth-grade students participated in this experiment. The author arranges an on-campus visit in July 2024. The second school, School B, is situated along the coast of Eastern China, and 143 students participated in the experiment.

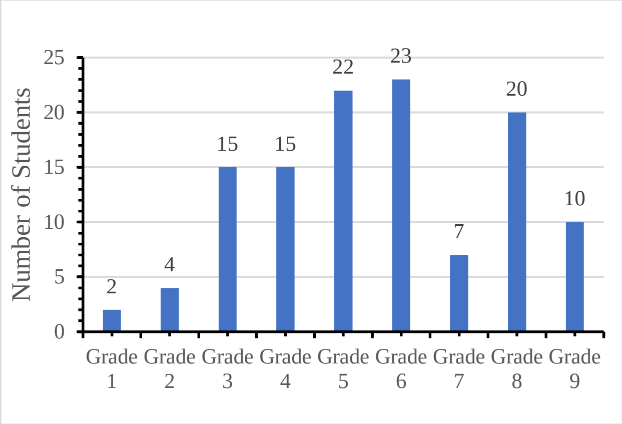
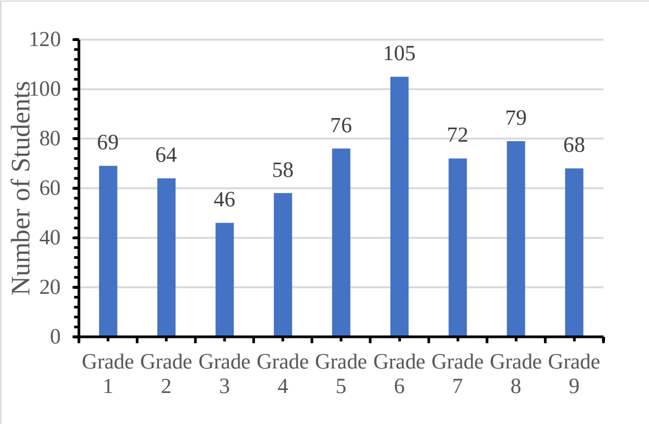


Figure 1. Distribution of Students by Grade in School A



.

Figure 2: Distribution of Students by Grade in School B.

Basic information about these schools and their students was collected through interviews with the head teachers at the corresponding schools as shown in Figure 1 and Figure 2. In the interviews, the status quo of the region where the schools are located, the basic structure of the school, and questions relating to financial grants are investigated. Additionally, data corresponding to students' academic performance and family status are collected in different ways depending on the school size. For schools with low capacity like School A, the author organized the family status of students by examining previous applications for financial grants and aligning them with these students' scores in 2023. For larger schools like School B, the author provided sample data in Excel and asked the head teachers to complete the questionnaire that consists of the required data. These data consist of students' Math, Chinese, English, and total scores, as well as students' family status, annual income, attitude towards learning, and any kinds of financial or material grants that the student received as shown in Table 1.

Table 1. Sample Student Information Collected from School A.

Grade	Gender	Chinese	Math	Science	English	…	Financial Grants	Injuries/ sicknesses	Debt	Attitude towards study	Free lunch
9	Female	82.5	96	120	109	…	0	0	0	above average	0
9	Male	45.5	57	32	28.5	…	1	0	0	average	1
9	Male	54.5	18	4	30	…	1	0	0	below average	0

8	Female	62.5	55	63	39.5	…	1	0	0	above average	0
8	Female	71	83	69	60	…	1	0	0	average	0
8	Female	42.5	77	52	30.5	…	1	0	0	above average	0
8	Female	60.5	43	47	47	…	0	0	0	average	0

In the following essay, the author is going to discuss which of the different factors collected display a consistent impact on the student's academic performance. These factors are later used to establish a connection between students' academic performance and their backgrounds. In Figure 3, a comparison between students' grades from School A and School B is established, signifying a lower average score for students in School A compared to students in School B. Furthermore, the author converted the collected data from categorical variables to numerical variables by feature engineering, which facilitated comparison between multiple students. For example, students with poor, average, or positive attitudes toward study are weighted with numerical values---0, 0.5, and 1 respectively---that positively correlate with their behavior.

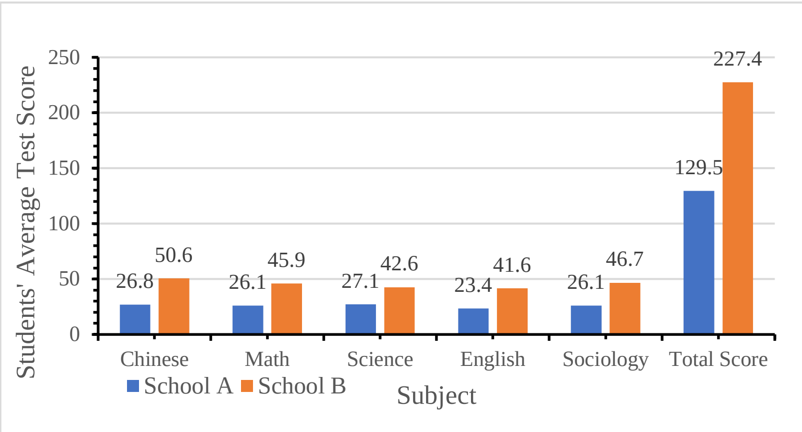


Figure 3. 8th Grade Students' Average Test Scores in School A and School B.

The data collected are rather subjective and might vary from region to region. For instance, in School A, only seventh and eighth-grade students have the score for sociology. Moreover, some students in School A received free pillows and subsidies for daily expenses while students in School B only received grants due to the difference in social-economic factors. To resolve these problems, the author modified these data by overlooking factors that are unique and ambiguous.

3.2. Data Processing

As with data collected from any real-world source, they come with certain missing values and imperfections that need to be fixed. Some data are either blank or filled with improper figures, for instance, Chinese characters were inserted in the students' Math Score. One example of such error occurs when School A uses the Chinese character '是' to represents an affirmative attitude while School B highlights students' names. Thus, the missing values were proposed before analyzing the significance of these data. Basic feature engineering methods were used to convert categorical features such as "Injuries/Sicknesses" or "Free Lunch". The author replaced 'affirmative' with 1 and 'denial' with 0. Moreover, for numerical features such as "Math" or "English" scores, the author replaced null values by the average score of this subject in this school, as these newly added values do not affect the mean, medium, or extreme values. To make ease for the machine learning and statistical methods later in this essay, all Chinese characters were translated to English characters.

3.2.1. ANOVA Pre-requisites

There are three basic prerequisites before performing a one-way ANOVA [27]:

1) Normality of data, where the statistical error of the data collected is a normal random variable with zero or minimal mean.
2) Homoscedasticity, where the residual variance is the same in data sets and the difference of the largest and the smallest variance within groups should not differ by more than one order of magnitude, and
3) Independence of probability, where any one data does not depend on any other data.

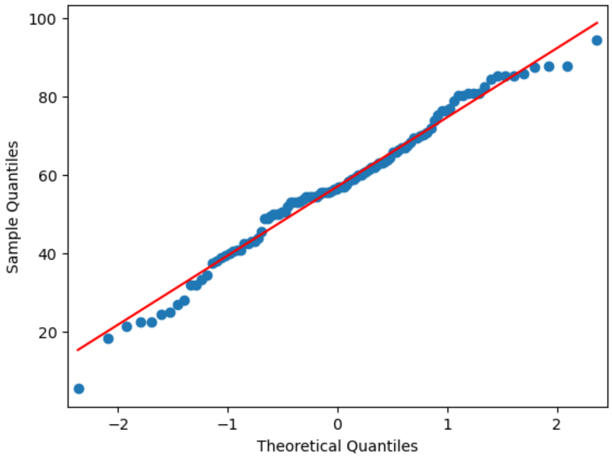


Figure 4. QQ plot of Math Score in School B.

To test the normality of the collected data. the author performs the Normal Quantile-Quantile (Q-Q) plot test, where quantiles of data distribution versus quantiles of the normal distribution are plotted [28]. On the x-axis, theoretical quantiles are plotted, which assumes that data has a mean of 0 and a standard deviation of 1. On the y-axis, the actual value of data is depicted. If the data collected are statistically normalized, the result would be a straight diagonal line y=x [29]. After several trials, it can be concluded that most of the scores collected follow the theoretical normalized data as shown in Figure 4

The Levene test, on the other hand, is commonly used to test homoscedasticity of data. The Levene test starts with a null hypothesis that all variances are equal across all groups or samples [30]. If the p-value results of the null hypothesis are smaller than a nominal percentage of 5%, it rejects the homoscedasticity of data [31]. Resolving the absence of homoscedasticity of data requires data transformation, including reciprocal, log, and square-root transformation [32]. These transformations remove correlations between variation, mean, and standard deviation, consequently creating homoscedasticity [33]. The author used log transformation to reduce p-value in Levene test.

3.3. ANOVA

After confirming that all data follow the three prerequisites, an ANOVA test can be performed. For every factor, separate the data by different subjects and then by different values. For example, “Financial Grants” is rearrange data into 5 categories: “Math”, “Chinese”, “Science”, “English”, and “Total Score”. Then separate the data by data that labels “Financial Grants” = “1” and data that labels “Financial Grants” = “0”. Repeat this process for every factor in the dataset.

3.3.1. ANOVA Results

The ANOVA results vary in School A and School B. In School A, there are minimal factors that show a significant effect on students' academic performance, which is represented by a p-value less than 0.05. Only grade and attitude towards study demonstrate some influence on students' scores, with grade having a statistically significant influence on Math, Chinese, Science, and Total scores, while attitude towards study has a statistically significant influence on Math and Total Score. These data are shown in Table 2 with a darker color indicating a lower p-value.

Table 2. P-value of ANOVA Factors with Academic Performance in School A.

	Financial Grants	Income	Gender	Work Force	Parent Status	Parental Care	Attitude Towards Study	Debt	Injuries/Sicknesses	Grade
Math	0.0779	0.2907	0.6804	0.2109	0.9651	0.1311	0.0245	0.4079	0.0493	1.89E-06
English	0.1753	0.091	0.2201	0.1843	0.3495	0.2568	0.067	0.1143	0.1436	0.0115
Chinese	0.2491	0.623	0.1658	0.5646	0.627	0.184	0.1225	0.0867	0.1245	3.55E-04
Science	0.2524	0.5954	0.1426	0.5739	0.9369	0.212	0.2234	0.0912	0.0921	3.52E-04
Total Score	0.2714	0.5497	0.1021	0.3699	0.746	0.1225	0.0306	0.1304	0.0369	2.52E-05

In School B, multiple factors demonstrate statistical significance in influence, including income, work force, attitude toward study, and grade. These data are shown in Table 3 with a darker color indicating a lower p-value.

Table 3. P-value of ANOVA Factors with Academic Performance in School B.

	Financial Grants	Income	Gender	Work Force	Parent Status	Parental Care	Attitude Towards Study	Debt	Injuries/Sicknesses	Grade
Math	0.0238	0.0052	0.2138	0.0351	0.1378	0.1959	1.47E-18	0.6994	0.2517	0.0023
English	0.7709	0.1406	0.3165	0.4412	0.9057	0.7632	1.70E-11	0.9113	0.9713	2.40E-04
Chinese	0.4574	6.44E-04	0.0173	0.0362	0.1308	0.7039	1.94E-14	0.3929	0.6671	0.0117
Science	0.0465	0.0277	0.8216	0.2286	0.135	0.6434	4.39E-14	0.2756	0.316	0.2537
Sociology	0.2894	0.028	0.0956	0.0151	0.0818	0.3079	7.75E-16	0.7219	0.5094	0.8303
Total Score	0.0715	0.0055	0.1545	0.0176	0.1382	0.6529	7.22E-23	0.6056	0.4343	0.0059

The heterogeneity of ANOVA p-values in several factors, in which factors in school A show less significant influence over academic achievement than school B, might be caused by geological influences. Other subtle factors, including extracurricular activities or experience, are expected to demonstrate a profound impact on students' academic performance.

3.3.2. Correlation Matrix

The ANOVA test can only indicate if there is significance between different magnitudes of factors exists, but it cannot tell how the factors' magnitude correlates with academic performance. Thus, the author utilizes a correlation matrix to identify the correlation between significant factors tested by ANOVA and test scores.

The correlation matrix is a set of Pearson correlation coefficients that indicates the strength of relationships between variables in a dataset. The Pearson correlation coefficient [34], also known as the product-moment correlation coefficient, is denoted as r and ranges from −1 through 0 to +1. If the sign of the correlation coefficient was positive, then a positive correlation would have existed, and vice versa with a positive coefficient. A correlation matrix is produced by inserting a dataset in a NumPy correlation matrix function.

In Figure 5, only attitude towards study and grade exhibits a certain extent of correlation with academic performance, in which attitude towards study poses a positive correlation while grade poses a negative correlation. This shows that students with positive attitudes towards study in School A are more likely to perform academic sufficiency, while students in lower grades in School A are expected to receive better scores than students with higher grades. The results of the correlation matrix are similar to the ANOVA results, where attitude towards study and grade heavily impact students' academic performance.

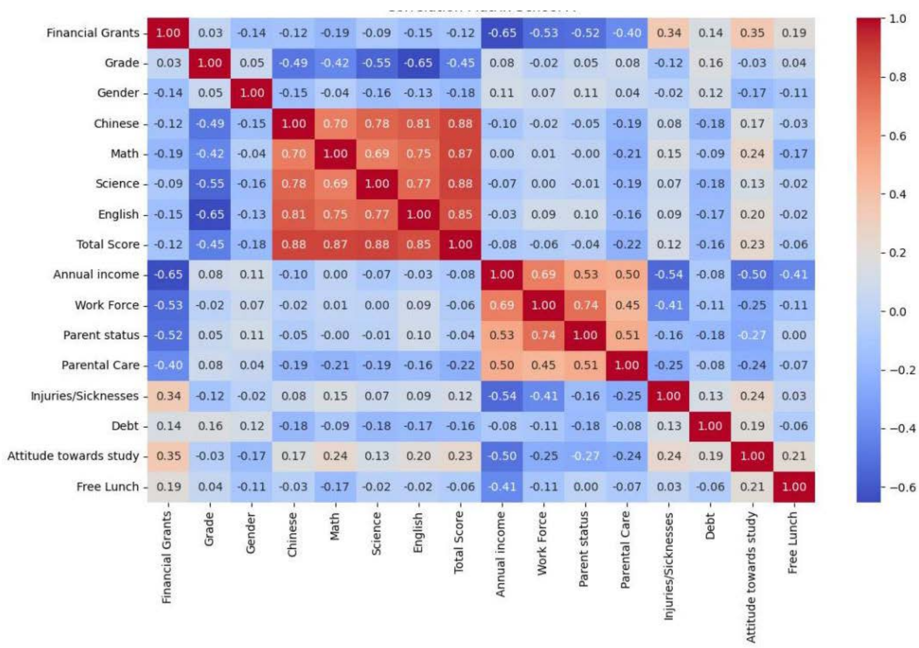


Figure 5. Correlation Matrix of School A.

In Figure 6, attitude toward study has the highest Pearson correlation coefficient among all the other factors, suggesting that attitude toward study is significantly and positively associated with students' academic performance in School B. Other factors, including annual income and workforce, have some extent of correlation with academic performance. However, grades do not exhibit any significant correlation with students' scores, which contradicts the ANOVA results.

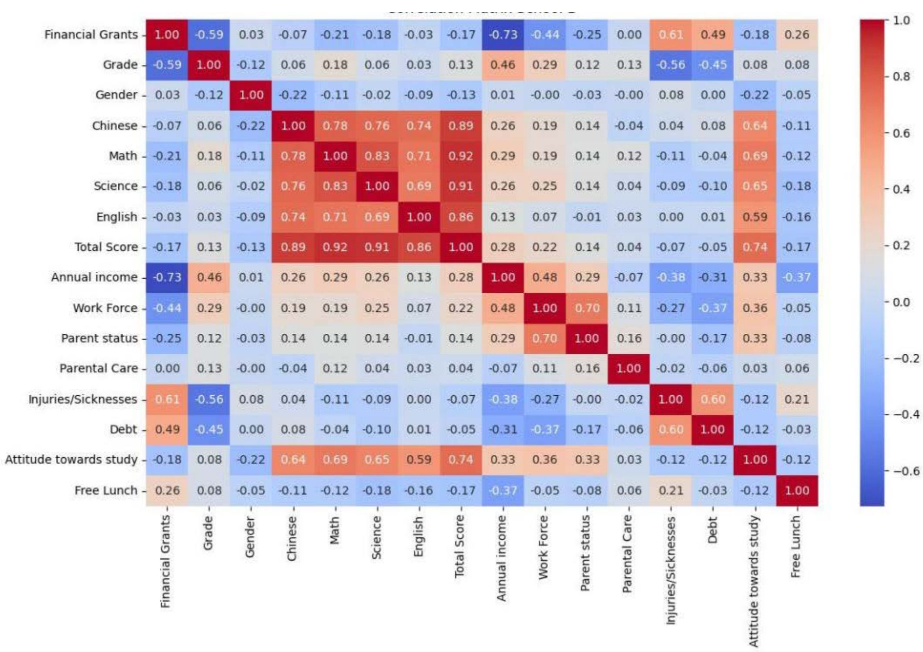


Figure 6. Correlation Matrix of School B.

The correlation matrix only identifies linear correlation between two variables (i.e. Annual income and Math score) while it does not identify any non-linear relation or relations among multiple variables. Conventional statistical models are unable to depict such a relationship. Novel statistical models in machine learning are required to investigate the underlying correlations.

3.4. Decision Tree

The decision tree is modeled by a group of training data and a group of testing data. The tree learns from the training data and identifies a possible trend between factors and scores. The tree then tests its hypothesis on the testing data to determine its accuracy. Mean-squared error (MSE) is a commonly used index to determine the accuracy of the tree model, as it is negatively correlated with accuracy. MSE is calculated by equation 4 below [35]:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i^2 - Y^2) \qquad (4)$$

Where n is the number of predicted data, Y_i is the actual value in the testing data, and Y is the predicted value. The author split 80 percent of the data as training data, while the 20 percent left as testing data.

The author normalized the data and combined the data from both schools to create a new dataset and train a decision tree model based on that. In this dataset, scores range from 0 to 1 with over 200 data. Factors that are used in this data include Financial Grants Grade, Gender, Annual income, Work Force, Parent status, Parental Care, Injuries/Sicknesses, Debt, Attitude towards study, and Free Lunch. Among those factors, attitude towards study and Grade are the two most influential factors in the decision tree model as shown in Figure 7.

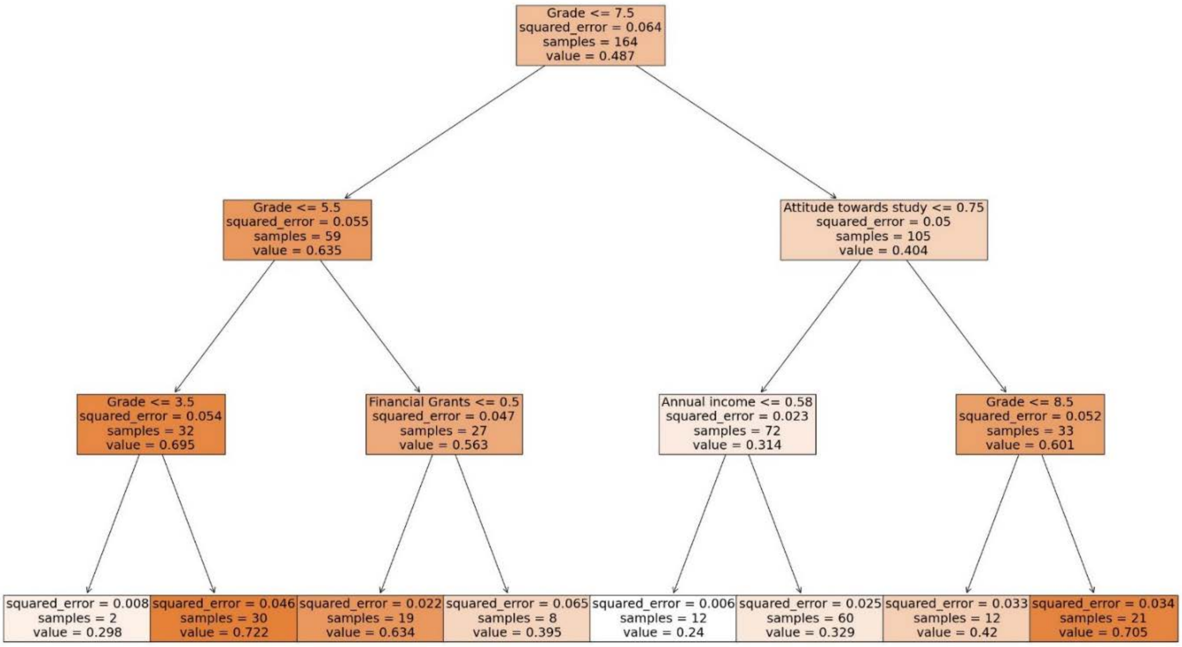


Figure 7. Visualization of Decision Tree Model of Total Score.

To evaluate the effectiveness of the decision tree model, the author compares the decision tree's prediction with the mean prediction, which is an array of the mean of the testing data, in Table 4. The mean prediction model predicts students' scores by the mean testing score. In this scenario, it is used as a reference model to test the decision tree's accuracy by comparing its MSE with the tree's MSE. Since the mean prediction model simply based its prediction on mean values, it should be rather inaccurate. The decision tree models are proven more accurate than mean prediction since its MSE values are less than that of mean prediction.

Table 4. Percentage Difference in MSE Between Mean Prediction and Decision Tree.

	Chinese	English	Math	Science	Total
Mean Prediction	0.05358	0.05983	0.07056	0.0709	0.05492
Decision Tree Model	0.04783	0.03589	0.06271	0.0293	0.04076
Percentage Difference	10.73%	40.01%	11.13%	58.67%	25.78%

4. Discussion

The aim of ANOVA tests and the decision tree model is to determine the relationship between financial grants and academic performance, as well as to investigate the factors that significantly influence students' test scores. After discovering the correlation, the effectiveness of financial grants, which is the distribution of monetary resources, on academic performance can be indicated.

In both School A and School B, attitude towards study and grade exhibit statistical significance on students' academic performance because these factors have a p-value lower than 0.05 for most of the subjects. This indicates that a statistical correlation can be established between these two factors with academic performance. However, for School B, more factors, consisting of annual income and force, are statistically significant, highlighting the variation between School A and School B. Comparing School A and B, School B's ANOVA results tend closer to the author's hypothesis: students with better economic status, which can be measured by annual income and work force, show better academic performance. Nonetheless, for students in School A, annual income and workforce have minimal impact, which departs from the hypothesis.

One possible explanation for the differentiation is due to the varying economic conditions in School A and School B. School A is located in a comparatively undeveloped region in respect of School B. The average monthly disposable income is 32215 CNY in the region School A situates and 42247 CNY in the region School A situates in 2022 [38], not to mention students in School A are mostly from villages, while students in School B came from various backgrounds. As a result, most students, even students that are classified as 'poor' for annual income in the dataset, in School B have better economic backgrounds than students with 'average' economic backgrounds in School A. Moreover, students in School A are expected to have similar economic conditions because School A is exclusively meant for local students. On the contrary, students are required to submit a formal Senior high school entrance examination report to be accepted by School B. Thus, students in School A have similar economic backgrounds, while students in School B have various economic backgrounds, causing the actual income difference is not as significant, regarding '0' and '1's, in the 'annual income' factor for School A.

The correlation matrix results reflect the insufficiency of ANOVA tests. In both Schools A and B, grade demonstrates statistical significance in academic performance. However, in the correlation matrix, the grade is poorly correlated with scores. This indicates that even though students in different grades receive different scores, there is no correlation between the magnitude of grades and students' scores. This phenomenon reveals the potential drawbacks of ANOVA tests: ANOVA tests can only show if there is a correlation, but cannot indicate how it is correlated, and this drawback is more significant when the input is distinguished by several groups. An example is students' grades, which are grouped from 4 to 8 for School A. Thus, relying solely on ANOVA tests cannot precisely describe the correlation between factors and students' scores, other statistical methods like machine learning can depict a better picture of the correlation.

The decision tree model exhibits a good overall prediction for students' performances by utilizing the training data, with an expected error of about 5 percent of the average score. This MSE value is later deducted by reducing the level of the decision tree model to prevent overfitting. Moreover, this method is statistically significant as its MSE errors are lower than that of the mean prediction method, with certain subjects, including English and Science, remarkably lower than the mean prediction. However, for other subjects like Chinese and Math, the difference between the decision tree model and mean prediction is subtle. This indicates that even though the decision tree model is based on machine learning algorithms, the model is only a basic machine learning model that does not notably differentiate from conventional statistical methods. More advanced models like random forest, K-means clustering, and logistic regression are expected to produce an accurate prediction. In addition, the decision tree model be improved by providing more training data, causing the small data size, around 250 students, to become a drawback of the decision tree model.

5. Conclusion

Financial grants are monetary resources for students who cannot afford tuition fees and daily expenses. However, after investigating its relationship with students' academic performance with the help of statistical tools, including ANOVA, correlation matrix, and decision tree under the current distribution scheme, financial grants do not directly translate into an improvement in students' performances. In ANOVA tests, attitude towards study and grade are the two most influential factors in both School A and School B, while annual income and workforce are significant factors only in School B. In the correlation matrix, both schools exhibit a positive correlation between attitude toward study and students' scores, but the positive correlation between grades and academic performance cannot be spotted. In the decision tree models, attitude towards study and grade are at the initial nodes, proving their significance when predicting students' academic performance.

Nevertheless, attitude towards study demonstrates a clear positive correlation with academic performance in the decision tree model, while grade influences scores differently. For example, in one model students who are eight-grade or above obtain better scores than students who are below, while in another model, students who are in seventh-grade or above obtain lower scores than students who are below. Factors that reflect economic status, annual income, and work force, have weak positive correlations with scores in tree models.

There are still many insufficiencies and areas to investigate. First, the number of data is insufficient, as only about 250 students participated in the research. More data is required to create a more accurate decision tree model as it is based on machine learning. The shortage of data undermines the reliability of the results in the essay, making it incapable of representing all schools in China. Second, the decision tree model is only one basic model in machine learning, other advanced machine learning models like random forest are expected to provide more precise predictions than the decision tree model. Third, more factors can be taken into consideration when discussing the relationship between factors and academic performance. Such factors include extracurricular interests, the presence of leadership, volunteering activities, and quality of education. Lastly, the quantification of factors can be improved. For example, if these data are available, providing the exact value of annual income instead of magnitudes from 0 to 1 can help enhance the accuracy of prediction.

The current method of distributing financial grants has been proven to be inefficient and the resources are not utilized as best as they can due to the ambiguity of criteria when selecting students who receive financial grants. To resolve inefficiency in financial grant system, the author proposes an improved criterion that utilizes statistical methods to identify factors that have the most impact on students' overall performance. By collecting data from different schools in rural China, it can be shown that by employing a selection criterion that identifies the influential factors using statistical methods and machine learning, the overall impacts of these grants can be further improved. Thus, based on the results of this experiment, the author proposes that weighing attitude towards study more than other factors can improve efficiency of financial grants. It is true that under the current experiment setup, there are areas to be improved and the criterions might change from school to school, but nevertheless this provided us with a new and improved way of looking at grant allocation.

References

1. H. Bareham, "FAFSA statistics 2023: What is the average financial aid for college?," Bankrate, Jun. 11, 2024. https://www.bankrate.com/loans/student-loans/fafsa-statistics/
2. "U.S. Department of Education." https://www.ed.gov/news/press-releases/us-department-education-releases-new-data-highlighting-how-simplified-streamlined-and-redesigned-better-fafsa%C2%AE-form-will-
3. China Ministry of Education http://www.moe.gov.cn/jyb_xxgk/moe_1777/moe_1779/201410/t20141021_178433.html
4. Y. Jia, 2018, doi: 10.3969/j.issn.1003-0166.2018.05.005.
5. R. S. E. Park and J. Scott-Clayton, "The impact of Pell grant eligibility on community college students' financial aid packages, labor supply, and academic outcomes," Educational Evaluation and Policy Analysis, vol. 40, no. 4, pp. 557–585, Jul. 2018, doi: 10.3102/0162373718783868.
6. L. Turner, "The Incidence of Student Financial Aid: Evidence from the Pell Grant Program," Department of Economics at University of Toronto, Jan. 2012.
7. S. Baum and J. E. Scott-Clayton, "Redesigning the Pell Grant program for the Twenty-First Century," The Hamilton Project, Oct. 2013, doi: 10.7916/d8vd6wfs.
8. W. Huang, F. Li, X. Liao, and P. Hu, "More money, better performance? The effects of student loans and need-based grants in China's higher education," China Economic Review, vol. 51, pp. 208–227, Oct. 2018, doi: 10.1016/j.chieco.2017.09.005.
9. S. Valcheva, "Statistical Methods for Data Analysis: a Comprehensive Guide," Blog for Data-Driven Business, Mar. 18, 2024. *https://www.intellspot.com/statistical-methods/#:~:text=Statistical%20methods%20for%20data%20analysis%20are%20the%20tools,foundational%20to%20making%20informed%20decisions%20based%20on%20data*.

10. O. Michelle Libretexts, "7.2.2: Descriptive versus Inferential Statistics," Statistics LibreTexts, Jun. 13, 2024. https://stats.libretexts.org/Courses/Taft_College/PSYC_2200%3A_Elementary_Statistics_for_Behavioral_and_Social_Sciences_(Oja)/02%3A_Mean_Differences/2.01%3A_Inferential_Statistics_and_Hypothesis_Testing/2.1.02%3A_Samples_and_Populations_Refresher/2.1.2.02%3A_Descriptive_versus_Inferential_Statistics
11. Z. Bobbitt, "Introduction to hypothesis testing," Statology, Jun. 08, 2022. https://www.statology.org/hypothesis-testing/
12. R. Bevans, "Hypothesis Testing | A Step-by-Step guide with easy examples," Hypothesis Testing, Nov. 08, 2019. https://www.scribbr.com/statistics/hypothesis-testing/ (accessed Aug. 14, 2024).
13. S. Kyla, Libretexts, "9.1: Introduction to hypothesis testing," Statistics LibreTexts, Apr. 24, 2022. https://stats.libretexts.org/Bookshelves/Probability_Theory/Probability_Mathematical_Statistics_and_Stochastic_Processes_(Siegrist)/09%3A_Hypothesis_Testing/9.01%3A_Introduction_to_Hypothesis_Testing
14. S. Lisa, "Hypothesis Testing - Analysis of Variance (ANOVA)." https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_HypothesisTesting-ANOVA/BS704_HypothesisTesting-Anova_print.html
15. W. Rachel, Libretexts, "11.1: One-Way ANOVA," Statistics LibreTexts, Mar. 12, 2023. *https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Mostly_Harmless_Statistics_(Webb)/11%3A_Analysis_of_Variance/11.01%3A_One-Way_ANOVA*
16. B. Rebecca, "Understanding P-values | Definition and Examples," Jul. 16, 2020. https://www.scribbr.com/statistics/p-value/#:~:text=The%20p%20value%2C%20or%20probability%20value%2C%20tells%20you,calculated%20by%20a%20statistical%20test%20using%20your%20data. (accessed Aug. 12, 2024).
17. M. Hassan, "ANOVA (Analysis of variance) - Formulas, Types, and Examples," Research Method, Aug. 14, 2023. *https://researchmethod.net/anova/*
18. "What Are Degrees of Freedom? on JSTOR," www.jstor.org, [Online]. Available: *https://www.jstor.org/stable/42659677*
19. "What is a machine learning algorithm? | IBM." https://www.ibm.com/topics/machine-learning-algorithms#:~:text=What%20Is%20a%20Machine%20Learning%20Algorithm%3F&text=IBM,given%20set%20of%20input%20variables.
20. E. Eaton, "Introduction to machine learning," 2017 fall. https://www.seas.upenn.edu/~cis5190/fall2017/lectures/01_introduction.pdf
21. "What is a Decision Tree? | IBM." https://www.ibm.com/topics/decision-trees
22. GeeksforGeeks, "Decision tree," GeeksforGeeks, May 17, 2024. *https://www.geeksforgeeks.org/decision-tree/*
23. F. Aiolli, "Dipartimento di Matematica 'Tullio Levi-Civita,'" 2008. https://www.math.unipd.it/ (accessed Aug. 30, 2024).
24. University of Washington Computer Science & Engineering, "Information gain." https://homes.cs.washington.edu/
25. GeeksforGeeks, "Overfitting in decision tree models," GeeksforGeeks, May 02, 2024. *https://www.geeksforgeeks.org/overfitting-in-decision-tree-models/*
26. L. A. Shalabi, Z. Shaaban, and B. Kasasbeh, "Data mining: a preprocessing engine," Journal of Computer Science, vol. 2, no. 9, pp. 735–739, Sep. 2006, doi: 10.3844/jcssp.2006.735.739.
27. Keya Rani Das, A. H. M. Rahmatullah Imon. A Brief Review of Tests for Normality. American Journal of Theoretical and Applied Statistics. Vol. 5, No. 1, 2016, pp. 5-12. doi: 10.11648/j.ajtas.20160501.12
28. C. Ford, "Understanding QQ plots | UVA Library," Aug. 25, 2015. https://www.library.virginia.edu/data/articles/understanding-q-q-plots#:~:text=QQ%20plots%20take%20your%20sample%20data%2C%20sort%20it,to%20match%20the%20size%20of%20your%20sample%20data. (accessed Aug. 12, 2024).
29. NIST, "1.3.5.10. Levene Test for Equality of Variances." https://itl.nist.gov/div898/handbook/eda/section3/eda35a.htm
30. R. J. Carroll and H. Schneider, "A note on levene's tests for equality of variances," Statistics & Probability Letters, vol. 3, no. 4, pp. 191–194, Jul. 1985, doi: 10.1016/0167-7152(85)90016-1.
31. Nenamalikah, "Python: Homogeneity of Variance Check," Kaggle, Mar. 23, 2023. https://www.kaggle.com/code/nenamalikah/python-homogeneity-of-variance-check
32. M. S. Bartlett, "The Use of Transformations," Biometrics, vol. 3, no. 1, p. 39, doi: 10.2307/3001536.
33. GeeksforGeeks, "Create a correlation Matrix using Python," GeeksforGeeks, Jun. 28, 2024. https://www.geeksforgeeks.org/create-a-correlation-matrix-using-python/
34. P. Sedgwick, "Pearson's correlation coefficient," BMJ, vol. 345, no. jul04 1, p. e4483, Jul. 2012, doi: 10.1136/bmj.e4483.
35. N. Piepenbreier, "How to calculate mean squared error in Python," Datagy, Nov. 21, 2023. https://datagy.io/mean-squared-error-python/ (accessed Aug. 26, 2024).
36. Pingyang Municipal Government Feb. 22, 2023. https://www.zjpy.gov.cn/art/2023/2/22/art_1229549064_4151711.html (accessed Aug. 22, 2024).