Marcel Kapica Linear Regression project raport.

I have decided to choose student performance as a topic because it is related to my day-to-day academic life, and I wonder what features of students are the strongest factors in getting best grades. I chose this topic because I want the best grades possible and I want to know which factors have the most input in my grade so I can influence them or draw some interesting conclusions for my academic life.

Methodology

A) The objective is to get to know which factors have the greatest influence on grades

B) The idea is to influence these factors or draw conclusions

C) The program I used is Microsoft Machine Learning Studio Azure

D) The source of data was kaggle

Source of data: <https://www.kaggle.com/datasets/desalegngeb/students-exam-scores/data>

The dataset was uploaded on April 17 2023 and has not been updated since.

The dataset contains 30641 cases and 15 features from which one is ID, three are test scores and eleven are possible predictors.

Raw dataset contains 10 string features and 5 numeric features.

The algorithm I am creating is predicting exam score based on student’s features.

Variable’s characteristics

ID (Column 0) was in the range from 0 to 999. ID has no use for predicting math score.

Central tendency measures, dispersion measure and measures of the shape of distribution for this feature are :

M = 499**.**5566;SD = 288.7479; Md = 500; K = -1.200814; SKE = -0.000514

Even distribution, no missing values, no potential outliers.

Gender variable included 2 categories (male, female). There are slightly more females in the dataset (mode).

This variable had no missing values and no potential outliers.

Ethnic group (EthnicGroup) variable included 5 categories (group A, group B, group C, group D, group E). Group C is the most frequent category (mode).

Variable ethnic group had 1840 missing values (6%) and no potential outliers.

Parent education (ParentEduc) variable included 6 categories (some collage, high school, associate’s degree, some high school, bachelor’s degree, master’s degree). Some collage is the most frequent category (mode).

Variable parent education had 1845 missing values (6%) and no potential outliers.

Lunch type (LunchType) variable included 2 categories (standard, free/reduced). Standard is the most frequent category (mode).

This variable had no missing values and no potential outliers.

Test preparation (TestPrep) variable included 2 categories (none, completed). None is the most frequent category (mode).

Variable test preparation had 1830 missing values (6%) and no potential outliers.

Parent marital status (ParentMaritalStatus) variable included 4 categories (single, divorced, married, widowed). Married is the most frequent category (mode).

Variable parent marital status had 1190 missing values (4%). Category widowed was a potential outlier.

Practice sport (PracticeSport) variable included 3 categories (somtimes, regularly, never). Sometimes is the most frequent category (mode).

Variable practice sport had 631 missing values (2%) and no potential outliers.

Is first child (IsFirstChild) variable included 2 categories (yes, no). Yes is the most frequent category (mode).

Variable is first child had 904 missing values (4%) and no potential outliers.

Number of Siblings (NrSiblings) was in the range from 0 to 7

Central tendency measures , dispersion measure and measures of the shape of distribution for this feature are :

M = 2.1459; SD = 1.4582; Md = 2; K = 0.297531; SKE = 0.676266

Variable number of siblings had 1572 missing values (5%) and leptokurtic distribution. High numbers of siblings were potential outliers.

Transport Means (TransportMeans) variable included 2 categories (school\_bus, private). school\_bus is the most frequent category (mode).

Variable transport means had 3134 missing values (10%) and no potential outliers.

Weekly study hours (WklyStudyHours) variable included 3 categories (5 - 10, < 5, > 10). 5 - 10 is the most frequent category (mode).

Variable weekly study hours had 955 missing values (3%) and no potential outliers.

Math score (MathScore) was in the range from 0 to 100. Math score is the target statistic we are going to predict.

Central tendency measures, dispersion measure and measures of the shape of distribution for this feature are :

M = 66.5584; SD = 15.3616 and Md = 67 K = -0.239841; SKE = -0.162862

This variable had no missing values and normal distribution. Math score had potential outliers in the lower scores as it was left skewed.

Reading score (ReadingScore) was in the range from 0 to 100. Reading score has no use to us as we want to predict math test result based on features of a person not based on results of other tests which are similar to the one we predict.

Central tendency measures, dispersion measure and measures of the shape of distribution for this feature are :

M = 69.4775; SD = 14.759 and Md = 70 K = -0.277254; SKE = -0.181288

This variable had no missing values and normal distribution. Looking for outliers in this variable is pointless because this variable has no use to us.

Writing score (WritingScore) was in the range from 0 to 100. Writing score has no use to us as we want to predict math test result based on features of a person not based on results of other tests which are similar to the one we predict.

Central tendency measures, dispersion measure and measures of the shape of distribution for this feature are :

M = 68.4186; SD = 15.4435 and Md = 69 K = -0.300259; SKE = 0.15983

This variable had no missing values and normal distribution. Looking for outliers in this variable is pointless because this variable has no use to us.

A graph of blue bars

Description automatically generatedA graph of a number of numbers

Description automatically generated

A graph of a bar graph

Description automatically generated with medium confidenceA graph of writing score

Description automatically generated

The dataset contained outliers and missing values to be dealt with, variables with string type which should be made categorical and features with categories which should be grouped.

I was using linear regression because the target (math score) is a ratio feature.

Steps I have undertaken

1

(SELECT)

First, I have deleted reading and writing scores by “Select Columns In Dataset” because these variables have no use to us as I want to predict math test result based on features of a person not based on results of other tests which are similar to the one I predict.

2

(OUTLIERS 1) (math score)

Then I have used ZScore in “Normalize Data” to normalize the math scores so I can find outliers.

3

(OUTLIERS 1) (math score)

Then by “Clip Values” I used a upper threshold of 3.29 for and lower threshold of -3.29 to make peaks and subpeaks missing values.

4

(OUTLIERS 1) (math score)

Then in “Clean Missing Data” I chose the new column I created (MathScore(2)) and used “remove entire row” to delete 52 outliers I just found. The outliers were math scores smaller than 21 out of 100.

5

(SELECT)

Then I have deleted the new column by “Select Columns In Dataset” as I used it and I no longer need it.

6

(WRONG FEATURE TYPE)

Then by “Edit Metadata” I made categorical features categorical as the program have set them for string at the start (gender, ethnic group, parent education, lunch type, weekly study hours [because in bins], transport means, is first child, practice sport, parent marital status, test preparation).

7

(MISSING VALUES)

Then in “Clean Missing Data” for the number of siblings feature I used “replace with mode” which turned all missing cases into “1” which in my opinion is more appropriate than mean and median (2).

8

(MISSING VALUES)

Then in “Clean Missing Data” for features with more than 5% missing values I used “Custom substitution value” which turned missing values into a new category (unknown) because I did not want to delete large amounts of cases or neither delete useful columns. Replacing these features with mode did not feel right as well because there is a small number of categories with similar amounts of cases in each, replacing large amounts of cases with mode would disrupt the distribution and produce fake data (test prep, parent education, ethnic group, transport means).

9

(MISSING VALUES)

Then in “Clean Missing Data” for features with less than 5% missing values I used “delete rows” to delete small amounts of cases (3527) that would not greatly impoverish the data. I did not want to delete useful columns. Replacing these features with mode did not feel right as well because there is small number of categories with similar amounts of cases in each, replacing large amounts of cases with mode would disrupt the distribution and produce fake data (parent marital status, practice sport, is first child, weekly study hours).

10

(OUTLIERS 2) (number of siblings)

Then I have used ZScore in “Normalize Data” to normalize the number of siblings so I can find outliers.

11

(OUTLIERS 2) (number of siblings)

Then by “Clip Values” I used an upper threshold of 3.29 for and a lower threshold of -3.29 to make peaks and subpeaks missing values.

12

(OUTLIERS 2) (number of siblings)

Then in “Clean Missing Data” I chose the new column I created (NrSiblings(2)) and “remove entire row” to delete 264 outliers I just found. The outliers were values equal to 7.

13

(SELECT)

Then I have deleted the new column by “Select Columns In Dataset” as I have already used it and I will no longer need it.

14

(GROUPING CATEGORIES) (parent education)

Then by “Convert to dataset” I have turned “some high school” into “high school” as these are the same and I want to group them, so the model is cheaper and faster.

15

(GROUPING CATEGORIES) (parent education)

Then by “Convert to dataset” I have turned “associate’s degree” into “licentiate” as I want to make categories more general so there are less and the model is cheaper and faster.

16

(GROUPING CATEGORIES) (parent education)

Then by “Convert to dataset” I have turned “bachelor’s degree” into “licentiate” as I want to make categories more general so there are less, and the model is cheaper and faster.

17

(GROUPING CATEGORIES) (parent marital status)

Then by “Convert to dataset” I have turned “widowed” into “single” as these are similar, and I want to group them to have a better distribution and for the model to be cheaper and faster.

18

(GROUPING CATEGORIES) (parent marital status)

Then by “Convert to dataset” I have turned “divorced” into “single” as these are similar, and I want to group them to have a better distribution and for the model to be cheaper and faster.

19

(FINAL STEPS)

Then by “Filter Based Feature Selection” I used Spearman corelation to get to know which variables are most correlated with the math score. I chose Spearman corelation because the data is not normally distributed and there are ordinal features (eg. weekly study hours [bins]). At this point I noticed very low corelation between features and the score (highest is 0.362137 for lunch type). I have selected all the features to know which ones are best predictors. If the corelation was high is some of the variables I would delete the low correlation ones and leave the effectively predicting ones for the model to be fast and cheap.

20

(FINAL STEPS)

Then I used “Split Data” to split the data into training part (75%) and test part (25%). I have used randomized split to be sure, but I did not think the cases were arranged in specific order.

21

(FINAL STEPS)

Then I used “Train Model” connected it to the training data part and “Linear Regression” and set the target to math score to train the model to predict math score using linear regression based on our training data part.

21

(FINAL STEPS)

Then I used “Score Model” connected it to the test data part and “Train Model” to use the model to predict math scores in the test data part based on the trained model

22

(FINAL STEPS)

Then I have used “Evaluate Model” to compare predicted scores with real ones and to evaluate the model. Coefficient of Determination turned out to be very low (0.240512). Relative Absolute Error turned out to be very high (0.869914), and Relative Squared Error was also very high (0.759488). Mean Absolute Error was 10.774173 and Root Mean Squared Error was 13.294831.

Reduction of dimensionality report

After all modifications the dataset contained 26798 cases (30641 at the start).

I have deleted 3842 cases (12,5%).

Steps 1-6

A screenshot of a computer

Description automatically generated

Steps 7-12

A screenshot of a cell phone

Description automatically generated

Steps 13-18

A screenshot of a computer

Description automatically generated

Steps 18-22

A screenshot of a computer

Description automatically generated

Whole model

A screenshot of a computer

Description automatically generated

Conclusion

When I saw the low correlation between variables and the target, I figured something was not right. I investigated it and noticed that in the ID column there are 1000 unique values while there was more than 30000 cases. Every case should have a different ID so I thought someone might have copied the data a couple of times and I could cut the thousand-case piece out of it, but different cases with same ID numbers had different statistics so I think this data was generated without a greater structure. Then unfortunately the model has no use, and I do not get to know which features of students are the greatest factors. But still I have scored and evaluated the model and the accuracy of it was obviously low.

If the model had worked, I could know any student’s test grade before it takes place. Additionally, I could draw conclusions like:

Which is a greater factor doing test preparation or studying certain amount of time weekly?

Are our physical features like ethnic group and gender affecting our grades?

How are our lifestyle features like lunch type, transport means and practicing sport affecting our performance?

How is our family status affecting our performance?

Which has greater influence on grade efforts or characteristics?

Then I could alter my studying to preparing before test more intensively or more intensively studying on a week-to-week basis. I could create programs for spreading gender and ethnic equality knowledge if the results would show that they have low correlation with performance. If they would show that these are strong determinants of the performance I could hire only people of a certain social group in my future company, or I could create programs for these groups to make them equal. I could alter my lifestyle for getting the best grades if the results would show a strong correlation. If they would show a low correlation, I would be free of any thoughts that my lifestyle could affect my grades. I could create programs for people with bad family status if the results would show that it has a great influence, or I could avoid hiring people with bad family status in my future company. If the results had shown that characteristics have minimal influence on grade compared to efforts, I would have a strong motivation to put efforts in studying. To summarize I did not reach my goal and did not get insights in influencing student performance but still throughout this project I have learnt a whole lot!