 FINAL PROJECT

INITIAL ANALYSIS REPORT

Student Performance Analysis in Portugal

Lakshmi Sravya Vedantham

ID: 002989684

Northeastern University -Toronto

Feb 16, 2021

Instructor- Oleksandr Maizlish

ALY6015 22280 Intermediate Analytics

|  |
| --- |
| 1. **Introduction**   The educational system is a critical part of today's environment. I'm always interested in learning more about the educational system. Being a coding instructor to a diversity of students has always forced me to learn new things while educating them. But there were many moments when I was astounded to see kids' performance dramatically improve or decline. As a result, I decided to conduct research on student data in Portugal.  Though the quality of education in Portugal has improved, there are still several challenges in the educational system as well as the surrounding environment that contribute to poor student performance and comprehension, particularly in mathematics and Portuguese courses. The purpose of this study is to investigate the elements that affect student learning performance. I'm curious if the education of mothers and fathers has an impact on student performance. Aside from that, I wanted to see if student failures, where they live, whether they have access to the internet, and how many hours they study per day have an impact on their grade.  There are a variety of machine models that can categorize data and forecast categorical classifications. I also created a prediction model that classifies the student's grades using classification and regression approaches. Linear regression, logistic regression, Naive Bayes, and Decision Trees are the models that I have taken into consideration. In the case of underfitting or overfitting, there is always a need to improve linear models. As a result, to avoid overfitting, I employed Lasso and ridge regression.  The information from Portugal's dataset will be presented in the following section. |

|  |
| --- |
| **2.1 Data Analysis**  **Data set Description**  The dataset is taken from two Portuguese secondary schools, and it applies to secondary school student accomplishment. Data on student grades, demographic, socioeconomic, and school-related variables were gathered through school reports and surveys. For performance in two separate courses, there are two datasets available: mathematics (mat) and Portuguese language (por). [Cortez and Silva, 2008] used binary/five-level classification and regression tasks to model the two datasets. I have combined these two datasets for further analysis.    There are 1044 items in the collection, each with 33 characteristics. That dataset is intriguing because it has both continuous and categorical variables. 17 of the 33 variables are categorical, while the remaining 16 are continuous variables. Another intriguing feature of the dataset is that it is devoid of null values. Before moving on to the descriptive analysis, I've added a new feature called "final status," which shows each student's performance based on the final score. If the score is between 15 and 20, the status is considered good; if it is between 8 and 14, the status is considered fair; and if the score is less than 7, the status is considered poor. There is no specific reason behind this division. It is just to determine the student performance in the form of categorical variable.  Dataset is now ready for Descriptive Analysis. |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **2.2 Descriptive Analysis**  To gain a better understanding of the data. Let's get started with ggplots and descriptive analysis. I'm particularly interested in the impact of categorical variables on the outcome. The graphs below supply a clearer picture.   |  |  | | --- | --- | |  | Based on the dataset, this graph gives a clear picture of the number of students in each status. It was discovered that about 750 students have a final status “fair” (those who scored between 8 and 14 on the final exam), 210 students are toppers, and the remaining 100 students are below average. | |  | This graph depicts the impact of a mother's job on a student's performance. Surprisingly, mothers of kids who work outside the home have a better impact on student performance than mothers who stay at home, according to this dataset. Mothers of students working in the health field had the least impact on their children's academic performance. I used a chi-square test to conclude that a mother's job has an impact on a student's performance. (Appendix -1) | |  | Like the role of mothers' jobs, fathers' jobs have a minor impact on student educational attainment. Students whose fathers work in different fields or supply services have a stronger impact on their schooling. When comparing the mothers' and fathers' jobs of students, mothers' jobs appear to have a greater impact on child achievement than fathers' jobs. I used a chi-square test to examine if the student's grade performance was dependent on the father's job, and it appeared that it was. (Appendix 2) | |  | This graph depicts the impact of romantic relationships on student performance. Students who were not in a romantic connection did better than students who were in a romantic relationship. This graph suggests that students' performance can be improved if they focus on their studies rather than their relationships. I used a chi-square test to determine the impact of romantic relationships on student grades, and it appears that there is insufficient data to show that romantic relationships have an impact on students' grades. (Appendix-3) | |  | The internet is a fantastic resource for students. It has advantages and drawbacks in terms of student performance. This graph shows that having better internet has a significant impact on student's performance compared to those who do not. I also used a chi-square test to see if the internet influenced student grade performance, and it seemed that it did. (Appendix –4) | |  | This graph depicts how the student's location has an impact on their education. We can see that students who live in urban areas, or cities, do better than students who live in rural areas. There could be a variety of reasons why pupils in rural areas struggle. However, if you want to improve student performance, you need to move to the city. The null hypothesis that there is no association between the address and student score is rejected by chi-square tests. (Appendix - 5)  I looked at other categorical variables to see whether they were dependent on the final score, but they don't appear to have much of an impact on the correlations between them. Hence, I ran a correlation test on numerical variables. | |  | I created a Corr plot of all numerical continuous data that influence student performance. Student failures, study time, and mother and father education are all critical factors affecting student achievement, as shown in this graph. | |  | The failure history of students is depicted in this graph. There are over 850 kids that did not fail and performed admirably. At least once, 125 students have failed math or Portuguese. Thirty to forty children failed mathematics and Portuguese multiple times. I can infer from this that about 200 children out of a total of 1040 are performing poorly, and there appear to be some factors contributing to this. This is something that should never be overlooked. | |  |  | |  |  |  |  |  | | --- | --- | |  | This boxplot proves how student performance changes as study time increases. When compared to students who study less than 2 hours per day, students who study 2 to 3 hours per day have a higher score (14 to 20). Increased student study time is recommended to improve student performance. |   This section supplies a detailed description of the descriptive analysis. The machine learning modeling is discussed in the following section.  **2.2 Modeling**  Modeling entails training a machine-learning algorithm to predict labels (final status) from features, fine-tuning it for business purposes, and testing it with holdout data. The result of modeling is a trained model that can be used to infer new data points and make predictions.  Now it's time to partition the dataset. The original dataset was split into two parts: 70 percent training and 30 percent testing. The training dataset is used to train the model, and the testing dataset is used to test it once it has been trained.  **Linear Model: (Model 1)**  The term "linear model" refers to a model that is defined as a linear collection of features. This model computes one weight for each feature based on training data and predicts the target value. For the given data, I ran linear regression on the final score and other characteristics. When running linear regression, it appears that a blunder is made every time because the final status and a final score are the same columns, resulting in a strong correlation. As a result, before sending it to the model training set, I tried to remove the final status column.  Hence, the equation for the linear model is as follows  **Final score ~ school + sex + age + address + family size + parents' status + reason + mother education + father education + mother job + father job + reason + guardian + commute time + activity time + study time + school support + family support + paid classes + romantic + family quality + free time + go out + health + absences + weekend alcohol + weekday alcohol.**  The model's predictions are in terms of the final score, which I converted to final status and compared to the training data's final status, yielding the following results.    As can be seen in the diagram above, the left figure describes the training dataset results, while the right figure depicts the testing dataset results. The model works well without overfitting because the training (69.1%) and testing (69.45%) accuracies are quite close. But, because I was curious, I ran lasso and ridge regression on the above model to see what the outcomes were.  **Lasso Regression for the linear model**  Lasso regression is a technique for improving the accuracy of regression methods by regularising them. In this model, shrinkage is used. Data values are reduced to a bare minimum. The lasso technique encourages simple, sparse models with fewer parameters. This sort of regression is best for models with a lot of multicollinearities or when you want to automate model selection steps like variable selection and parameter removal.    The results of the lasso model on the training set are shown on the left, while the results of the testing set are shown on the right. We can deduct from this that this model is not as good as the linear model because most of the features have been removed. Training accuracy appears to be 64 percent, while testing accuracy appears to be 65 percent. This model seemed to have comparable results as the linear model. As a result, I continued to run a linear model with ridge regression.  **Ridge Regression for the linear model**  Ridge regression is a model tuning technique that is applied to multicollinear data analysis. This approach is used to produce L2 regularisation. When there is a problem with multicollinearity, least-squares are unbiased, and variances are significant, leading to predicted values that are far from the actual values.    The above numbers, on the left and right, show the outcomes of the training and testing tests. This model produces comparable outcomes as the linear model, although it is not as good as the linear model, which has training and testing accuracies of 64% and 65%, respectively. From these lasso and ridge models, we may deduce that the linear model is better suited without overfitting and it's not necessary to perform lasso and ridge tests.  **Classification Models**  A classification model tries to deduce some inferences from the training values provided. It will expect the new data's class names (final status) and categories.  **Multinomial Logistic regression (Model 2)**  Logistic regression is a model in statistics that uses a logistic function to be a binary dependent variable in its most basic form. (“by the use of different Data Mining Techniques”) It is a technique for estimating the parameters of a logistic model in regression analysis. I used multiclass classification here since the final status category has three levels (bad, fair, and good). To perform multinomial logistic classification, I used the “multinom” function ().  The equation for this model is as follows:  **Final status ~ school + sex + age + address + family size + parents' status + reason + mother education + father education + mother job + father job + reason + guardian + commute time + activity time + study time + school support + family support + paid classes + romantic + family quality + free time + go out + health + absences + weekend alcohol + weekday alcohol.**  The final status prediction is compared to the training and testing final statuses, yielding the following findings.    The logistic model uses categories to classify the training and testing dataset input (poor, fair, good grades). The training set results in the above-left figure show that the model supplies 77% accuracy, whilst the testing set supplies 66% accuracy (right figure). This shows the overfitting of the multinomial model. As a result, I choose to use lasso and ridge multinomial regression on this logistic multinomial regression.  **Multinomial Lasso Regression for the logistic model**    On the left, the lasso model's results on the training set are shown, while the testing set's results are shown on the right. Because most of the features have been dropped, we can conclude that this model is not as good as the multinomial logistic model. The accuracy of training and testing is roughly 70%. Let's have a look at how ridge regression works.  **Multinomial Ridge Regression for the logistic model**    The results of the training and testing tests are shown to the left and right of the figures above. This model outperforms the logistic model, which has 72 percent training accuracy and 72 percent testing accuracy, respectively. As a result, we may conclude that, when compared to lasso and logistic models, ridge multinomial regression is a better fit.  **Classification Model: Naïve Bayes**  The naive Bayes classifier is another classification classifier that I could use to classify this data set. It's a Bayes' theorem-based probabilistic classifier with strong independence assumptions between the features.  The equation for this model is as follows:  **Final status ~ school + sex + age + address + family size + parents' status + reason + mother education + father education + mother job + father job + reason + guardian + commute time + activity time + study time + school support + family support + paid classes + romantic + family quality + free time + go out + health + absences + weekend alcohol + weekday alcohol.**  The training and testing final statuses are compared to the final status prediction, supplying the following results.    As seen in the above picture, this classifier is well-known for classification and supplies 66 percent accuracy for both training and 59 percent testing accuracy. It performs better when the training dataset is larger and the hyperparameters are fine-tuned.  I discovered a model that is even better than naive Bayes: decision trees.  **Classification Model: Decision Trees**  For classification and regression, Decision Trees (DTs) are better models that will learn simple decision rules from data attributes to develop a model that predicts the value of a target variable. (“Decision tree for healthcare analysis | Detect breast cancer”) With a set of if-then-else decision rules, decision trees learn input data and make decisions. As the if-else tree increases, the model becomes more complex and more accurate.  The equation for this model is as follows:  **Final status ~ school + sex + age + address + family size + parents' status + reason + mother education + father education + mother job + father job + reason + guardian + commute time + activity time + study time + school support + family support + paid classes + romantic + family quality + free time + go out + health + absences + weekend alcohol + weekday alcohol.**  The final statuses of training and testing are compared to the final status prediction, yielding the following findings.    By using if-else logic, decision trees deal with the construction of trees. As illustrated in the above figures, this performs quite well for this dataset, supplying 75 percent accuracy for training tests and 69 percent testing datasets.  It's time to decide which model is the best out of all the options.   |  |  | | --- | --- | |  | The multinomial logistic model appears to be overfitted in this graph. As a result, ridge regression appears to be a better model in comparison to lasso multinomial regression. As a result, we can conclude that linear, lasso multinomial, ridge multinomial models are better models. However, if I had to choose, I'd go with ridge multinomial regression because it is a perfect fit for this dataset. | |

|  |
| --- |
| **3. Summary and Conclusion**  The following conclusions are made as per the preceding analysis   * Students who have parents who work in other fields and supply services perform better. * Students who were not in a relationship did better on tests. * The internet has a positive effect on student achievement. * Students in metropolitan areas performed better. * Students that study for more than 2 hours every day score higher. * Educated Parents of students play a vital role in their performance.   To get good grades, a student should have educated students who work in various fields and supply services, and the student should not be in a romantic relationship and study 2 to 3 hours every day. He or she should live in cities where there is access to the internet and should not fail any topics. However, these are only inferences based on the preceding evidence; it would be ideal if the student remained happy and studied according to his interests without being pressured.  The strongest models for grade categorization were linear regression, multinomial lasso regression, and multinomial ridge regression. However, ridge multinomial regression with 72 percent performance in both training and test sets appears to be best for this student dataset based on the accuracies. All the models appear to be less accurate. In the future, I'd like to use gradient boosting techniques or advanced machine learning models to improve the accuracy of these models for this dataset.  The upcoming section (Section 4) gives all citations followed by the Appendix section |

|  |
| --- |
| **4. BIBLIOGRAPHY**   * 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. * *UCI Machine Learning Repository: Student Performance Data Set*. (2008). Archive. <https://archive.ics.uci.edu/ml/datasets/Student+Performance> * *“Chi-Square Test of Independence in R - Easy Guides - Wiki - STHDA*.” (“Chi-Squared test of independence - General - RStudio Community”) (2018). STHDA. <http://www.sthda.com/english/wiki/chi-square-test-of-independence-in-r> * *Data Science and Analysis Tutorials | Data Camp*. (2018). Data Camp Community. <https://www.datacamp.com/community/tutorials> * *Google Collaboratory*. (2021). Collab. <https://colab.research.google.com/github/meizmyang/Student-Performance-Classification-Analysis/blob/master/Student%20Performance%20Analysis%20and%20Classification.ipynb#scrollTo=hJWwOk8XzjRZ> * *“Penalized Logistic Regression Essentials in R: Ridge, Lasso and Elastic Net*.” (“Penalized Logistic Regression Essentials in R: Ridge ...”) (“Articles - Classification Methods Essentials - STHDA”) (2018, March 11). Articles - STHDA. <http://www.sthda.com/english/articles/36-classification-methods-essentials/149-penalized-logistic-regression-essentials-in-r-ridge-lasso-and-elastic-net/> * *RPubs - Student Analytics - Prediction of Academic Success for Highschool Students*. (2017, October 8). Rpubs. <https://rpubs.com/Vasanth_Kailasam/StudentAnalytics> * GeeksforGeeks. (2021, September 24). *Multiple barplots in R*. <https://www.geeksforgeeks.org/multiple-barplots-in-r/> * Li, S. (2019, February 27). *Building A Logistic Regression in Python, Step by Step*. Medium. <https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8#:%7E:text=Logistic%20Regression%20is%20a%20Machine,%2C%20failure%2C%20etc>.). * *Regression Model Accuracy Metrics: R-square, AIC, BIC, Cp and more*. (2018, March 11). Articles - STHDA. <http://www.sthda.com/english/articles/38-regression-model-validation/158-regression-model-accuracy-metrics-r-square-aic-bic-cp-and-more/> * *RPubs - Adaptive LASSO Examples*. (2017, October 31). Rpubs. <https://rpubs.com/kaz_yos/alasso> * *RPubs - Student Analytics - Prediction of Academic Success for Highschool Students*. (2017, October 8). Rpubs. <https://rpubs.com/Vasanth_Kailasam/StudentAnalytics> |

|  |
| --- |
| **5. Appendix (code)**  **Chi-square tests Conclusion**   1. **Mothers job vs student performance**   H0: There is no dependency of Mothers job on the grade of the student  H1: There is a dependency of Mothers job on the grade of the student     1. **Fathers Job vs Student performance**   H0: There is no dependency of fathers job on the grade of the student  H1: There is a dependency of father's job on the grade of the student     1. **Romantic relationship vs Student performance**   H0: There is no dependency of romantic relationship on the grade of the student  H1: There is a dependency of romantic relationship on the grade of the student     1. **Internet vs Student performance**   H0: There is no dependency of internet on the grade of the student  H1: There is a dependency of internet on the grade of the student     1. **Address vs Student performance**   H0: There is no dependency of location on the grade of the student  H1: There is a dependency of location on the grade of the student |