Arun Agarwal

Principles of Data Science, CIS 3715

March 25th, 2021

Progress Report I

Deep Analysis and Prediction of the Effects of Various Factors on the Performance of Students on Examinations

Project Summary/Overview:

The purpose of an examination is to understand the ability and learning of a student; thus, through careful analysis of patterns between parental level of education, gender, race, reading scores, writing scores, math scores, and other factors from the Royce Kimmons Kaggle dataset named “StudentPerformance.csv, one can understand which aspects have the largest impact on test outcomes as well as what are the best ways to improve student scores for the future. Furthermore, the improvement of test scores would benefit the economy at a national level by creating equality, leading to major GDP growth, and building a modern society. Correlated with improved test scores is a stronger education system, which for the learner, creates more employment opportunities, secures a higher income, develops problem-solving skills, provides a prosperous and happy life, and educates them to give back to the community.

Prediction of student academic performance in mathematics, reading, and writing based on various demographic and socioeconomic statistics can be performed through creation of various data science models such as linear regression, logistic regression, and k-NN. While the dataset includes this information for 1000 students, not all useful pieces of data, such as hours spent studying, are provided, so I can only perform analysis on the effects these features have on performance. A model with good RMSE, MSE and MSAE scores signifies the model predicts student’s performance well, making the data useful in identifying methods to improve student performance for the future. For each model, I will use different hyperparameters and preprocessing techniques to optimize performance. Various types of graphs and plots will also be utilized for exploratory data analysis. The early detection of students who are vulnerable to suffering academic failure (through use of these models) can subsequently be used to design new teaching/mentoring strategies for an overall strengthening education system and society.

Progress Report (What has been done):

The beginning steps of the proposed timeline in the project proposal was as follows. For week 1, I was to write the project proposal, import the dataset, import required libraries, complete data preprocessing (including cleaning the data, making graphs to understand the data, label encoding, and splitting the dataset), and exploratory data analysis. All of this and more was completed by the end of Week 1 for the project, taking a total of approximately 12 hours, which is 5 hours longer than the proposed time. I spent a larger amount of time for this week because both the project proposal and progress report I had to be completed for week 1. Furthermore, I did more than what was proposed for this week, as will be explained in detail later, and was very detailed in the writing of my notebook, so this added a significant amount to the time. In the project proposal, 10 steps were outlined in the procedure to reach full completion of the project, and the first three steps (each with multiple parts) were completed for week 1, and steps four through ten were outlined out in the Jupyter Notebook.

The first thing I did for this week (after finishing the project proposal) was find and download the dataset. The dataset is provided by Royce Kimmons and can be found on the Kaggle website under the Student Performance on Examinations project. I did a preliminary exploration of the features and size of the dataset, and then moved on to setting up the coding environment. For consistency, I utilized the same program that is used in the class’s labs, Jupyter Notebook, accessed through the Anaconda Navigator. Since I created my own environment to work with, I needed to import multiple libraries to be able to perform all analysis that I plan to on this project at some point. To know which libraries I needed to import, I reviewed our class examples (Slides and Labs 3-8) and previous work with this dataset, available from the Kaggle dataset. I also revisited the similar projects listed in the Project Proposal document. I reviewed the code and setup of these notebooks not only to understand the necessary libraries to import but also to get an understanding of how I should go about working on the project and what I can include. Once I learned all the libraries necessary to complete my project, I started the import process in the Anaconda Navigator. I installed plotly, lightgbm, py-xgboost, and seaborn into the environment; I also needed to install bayes\_opt, but I could not see any option to install this library into my environment. Therefore, I may decide to not perform bayes optimization in my project. This might work out to my favor as we have never learned about the optimization project in class, so I am not sure if I would be able to write the necessary code for it. Other libraries that I would use for my project that are already imported include numpy, pandas, matplotlib, sklearn, and pylab.

With the environment set up, the Jupyter Notebook software could be opened, and creation of the notebook could begin. A folder was created for the project, called ‘Final Project 3715’, with the csv dataset uploaded into it. Also added into the folder was a new notebook, which when opened was named ‘Agarwal—001—Final Project Student Performance on Examinations’. Next, due to my lack of experience with the intricacies of Jupyter Notebook, numerous things had to be looked up. For example, I had to learn how to make different headers, how to write plain text, how to write italics, how to make bullet points, how to organize the code well, how to reorder cells, how to add numbered points, how to create links in the document (for a table of contents), and how to do numerous shortcuts (delete cells, add cells above, add cells below, move cells, etc.). I learned this information through simple searches and through the help page in the software itself. Next, I added a title to the beginning of the document, my name, and the class name. Next, I created a cell called ‘Table of Contents’. When examining other projects with the dataset, I found that using a Table of Contents in a Jupyter Notebook is a great way for one to organize their dataset. Thus, as the project goes on, I will continue to add links to various areas in my projects for ease of access and organization in the notebook. Currently, there is a link to the ‘Goals’ tab in the document. I will make more links later as I may decide to change the orderings of things at a later point, which will cause disordering in the Table of Contents. Next, I added sections called ‘Introduction’ and ‘Goal’ which, as their title states, introduces the project (explain what will be done, why it will be done, how it will be done, etc.) and describes the end-goals. Another section, named “Questions to Try to Answer,” lists all of the essential questions I will try to answer through the analysis I will perform on this dataset (8 questions are listed).

Next, the coding part of the notebook began with importing the libraries stated earlier. The imports are broken down in the notebook based on what they will be used for. After this, I made a section for reading in the csv file, similar to how we do it with every lab. I also called the head() function to print the first few rows to get a superficial understanding of the dataset and its features. Furthermore, I added a target column to the data called ‘average score’, which is simply the average of the math, reading, and writing scores. This will be useful for creating models in which having one overarching target column will be helpful. After this, I listed out the targets of the data (the scores) and the features (gender, race/ethnicity, parental level of education, lunch, and test preparation course). I next printed out the size of the dataset; seeing that the dataset has only 1000 points, I chose not to remove any outliers in hopes of preserving/holding on to as much of the data as possible. I also used the describe method to understand basic information about the numerical features such as the count, mean, std, min, and max and analyzed the meaning of these results. Moving on, I checked for missing values using isnull().sum(), which demonstrated that there are no missing values and no filling in would need to be completed. I also used the info() function to demonstrate which features were categorical and which numerical (listed earlier).

With preliminary data examination completed, I started the exploratory data analysis/data visualization section of the project. First, I visualized all numerical features with histogram plots to see their distributions and analyzed the results. I decided 50 bins was a decent amount to display the data well. I then created a pair plot as another way to visualize the numerical features. Next, I used the corr() function of Pandas and the heatmap plot to show the correlation between different numerical features. The high correlation between the reading and writing scores was analyzed in detail. I also visualized the numerical features with density plots. After data visualization of the numerical features, I examined the categorical features using a variety of graphs. By coding a method to create pie charts, I was able to make a pie chart for each categorical feature at the same time. These pie charts were able to demonstrate a lot of proportions as well as pictorially demonstrate the unique values in each feature. Further categorical feature analysis was performed by creating bar graphs to compare each feature’s unique values to the average score for those values. The information collected from these bar graphs was then analyzed. Next, I explicitly coded a pass/fail mark, where 50% or lower is fail and standard letter grades are coded for the rest of the percentages. Then, I was able to see, for each student, which student is passing, and which is failing for each examination score. These new features were added to a copy of the dataframe, named df\_extended because I will only need these columns a few times. Using the previously created pie chart method, I was then able to create pie charts for the newly added categorical features. Two more columns were added to the extended dataframe, including ‘Total Marks’ and ‘Percentage’ to later find the percentage of marks. With the created grading scale, I plotted the number of students who obtained each grade to get a general sense of how well student do on examinations. I also made a count plot for parental level of education broken down by student grade. Finally, away from the grading scale, I performed analysis based on gender by developing plot density graphs for the three scores based on gender. The results of the gender differences were then analyzed.

Finally, as done in labs, I converted the categorical features to numerical features using Label Encoding. This had to be done for a total of five columns, and the info method was called on the dataset to confirm that the whole dataset is now numerical values. Next, I split the dataset with 80% of the data being in the training set, and the final 20% in the test set to start the linear regression process for week 2. For organization purposes, I also outlined some of the other things I will do in the notebook including training the linear regression model, evaluating the linear regression model, using ridge regression model to do prediction, building a support vector machine (SVM), building a logistic regression model, building a linear least squares regression model, performing principal component analysis (PCA), and doing k-NN classification. Sections for the Conclusion and References are also included at the end of the notebook.

Progress Report (What has not been done):

As outlined in the previous paragraph, there are still multiple things to be done in the coming weeks. I need to first finish the linear regression and ridge regression models and their analysis. Model metrics will be written, and scores for MSE, MAE, RMSE, and R-squared will be given for both the training set and test set. Model performance will also be analyzed through creation of numerous graphs and plots. Next, I will build a logistic regression model using k-fold cross validation. I will need to tune the hyperparameter for the regularization term. Also, I will evaluate the accuracy, precision, recall, and F1 scores of the classifications for each preprocessing technique for an understanding of performance. Moving on, I will create another linear regression model, called linear least square regression to understand the strength of the relationship between various provided factors and exam scores. Residuals—the deviation of the fitted values from the actual values—will be created for this and used to determine if it is good to predict the reading scores from one’s writing scores. Furthermore, a coefficient of determination (R^2 value) and a plot of the two features graphed will need to be included. I will also build a support vector machine (SVM) if I deem it useful after doing these models. Next, I will conduct Principle Component Analysis (most likely the sklearn Neighborhood Component Analysis) to prepare for k-NN on the preprocessed dataset using the most effective preprocessing technique. Then, I will of course perform k-NN classification using sklearn.neighbors and plot the results. After all these models and plots have been made, I will compare their results and make general conclusions about which factors are interrelated to each other and to performance on examinations (written in the conclusion section of the notebook and in the final report). Throughout the process, I will also need to write a second progress report, complete a lightning talk, and submit a final report. Everything listed here should take approximately twenty-two to twenty-five hours.

Progress Report (What will be done in the following week):

As outlined in the project proposal, there are multiple things that need to be completed in the following week. First, I will start, write, and finish the linear regression, ridge regression, linear least squares regression, and logistic regression models as well as all corresponding components. These components are listed in the previous section. With completion of these models comes all corresponding pattern and model performance analysis. I will also need to write progress report II at the end of the following week. This should take approximately twelve to fifteen hours to complete.