1. Letter of transmittal

Dear Mr. Smith,

It is a well-known fact that most school systems have been under funded for years and trying where best to allocate funds can be a real challenge. We here at Acme feel like we have a solution that will allow you to best use what limited resources are available. We will create a product that with a limited amount of data will be able to help determine which programs and students will benefit most your current test preparation course.

This product will be a browser-based application using data currently being pulled from existing test scores and student data. Using this data, it will then create several visual displays comparing various metrics as well as predicting increases or decreases based on the use of your test preparation course.

The objective of this will be to allow the school board to better target which students are more likely to benefit from the use of a test preparation course since time and money are always in short supply. On top of that you will be able to better show to your stakeholders the added benefit that the test preparation course is having and through that possibly secure more funding for your school.

We will be using an Agile methodology for this project, so you can expect daily updates and a small number of people working this project to keep costs to a minimum. With the small team we project it will cost $30,000 to develop this application. Tech support for after the project is finished will continue for a year and if satisfied you can retain that support for a fee to be determined at a later date.

The impact of this solution on your stakeholders will be evident rather quickly once you see the correlation between several factors and test scores. Knowing where best to allocate funds when on a limited budget is crucial if you wish to maintain the high standards your school has. We here at Acme hope to hear back from you and look forward to working with you.

Sincerely,

Sean Tate

2 Problem Statement

The main issue hounding every school district is a lack of funding. With unlimited funding all projects and avenues could be explored, but in the real world there is only so much money to go around, so knowing where best to spend that money is key. Taxpayers are not always willing to pay more to fund projects they feel will not give an adequate return on their investment. With our application you will be able to easily visualize data points in relation to your existing test preparation program. By using historical data, this browser-based application can project changes to test scores based on the use of an existing test preparation program and the demographics of the student. As long as the data format stays the same more data can be added thus increasing the accuracy of the program as time goes on. As it currently stands all data regarding the current test preparation program is calculated by hand and is a time-consuming process, but with this application that time can be cut down significantly.

3 Data product benefits

This application will automate many of the data analysis tasks that would normally be performed by someone regarding your existing test preparation program. This will cut down on man hours and save money in the long run. On top of that it will generate several different graphs to help illustrate improvement or the lack there of. Lastly the user will be allowed to select the education level of the student’s parents to see the projected changes to their test scores if a test preparation course was offered.

4 Outline

The application will be created using Python and developed in a Jupyter Notebook which is a locally hosted server that acts as a virtual machine for the purpose of running the code. The application will take student data, test scores, and several other factors and attempt to find a correlation between them.

5 Data

Data will be stored on a .csv file which will be loaded automatically into the application. Any new data will need to be in this format and conform to the example below.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| gender | race/ethnicity | parental level of education | Lunch | test preparation course | math score | reading score | writing score |
| female | group B | bachelor's degree | standard | none | 72 | 72 | 74 |
| female | group C | some college | standard | completed | 69 | 90 | 88 |
| female | group B | master's degree | standard | none | 90 | 95 | 93 |
| male | group A | associate's degree | free/reduced | none | 47 | 57 | 44 |
| male | group C | some college | standard | none | 76 | 78 | 75 |
| female | group B | associate's degree | standard | none | 71 | 83 | 78 |
| female | group B | some college | standard | completed | 88 | 95 | 92 |
| male | group B | some college | free/reduced | none | 40 | 43 | 39 |
| male | group D | high school | free/reduced | completed | 64 | 64 | 67 |

As you can see race/ethnicity has been changed simply to groups to prevent any racial bias to be perceived. The projections used in this application will only use students whose parents have similar educational levels such as everyone with some college or a master’s degree. No names will be used so there is no fear of any personal information being leaked or used for malicious purposes.

6 Objectives and Hypothesis

The objective with this application is to allow a person to quickly and easily analyze the large amount of test data in the database and then to present this data in a user-friendly way that can be presented to a school board. Our hypothesis is that parents with a higher level of education will have a child who will benefit less from a test preparation class versus a child whose parents have a lower education level. This program will essentially try to show who will get the most benefit from the test preparation class, so that the school can target those who will benefit the most since there is not always enough room for everyone.

7 Methodology

We will be using an Agile methodology for this project, so you can expect daily updates and a small number of people working this project to keep costs to a minimum. The variation of Agile we will use is one based around Scrum. Feedback from the customer will be one of the major keys to success as we iterate through the project. Requirements will not be rigidly defined as the focus will be more on iterable products rather than paperwork and strictly defined documents. There will be a series of Sprints at which point the development team will work towards that current sprint goal. After each sprint has finished the product will go through brief testing and then sent to the customer for review and feedback. The feedback will then be used to help drive focus for the next sprint. Once we receive acceptance from the client then we will turn over the application and provide access to the environment that the application runs on.

8 Funding

Development team – 5-person team – 210 hours - $21000

Testers – 50 hours - $5000

Maintenance and licensing - $1000 renewed yearly

10% buffer - $3000

9 Stakeholder Impact

This web application will have an impact on a couple of different stakeholders. The first and foremost of these stakeholders is the school district. By being able to better target their test preparation classes they will see a better return on the money and time spent on that program. That in turn will save the taxpayer money in the long run. Secondly by being able to better display the results from their test preparation program it will show school board members that progress is being made and to continue if not boost funding. Lastly students will benefit by being placed into programs that are more beneficial to them instead of wasting valuable time and space.

10 Data Precautions

As stated earlier all names and race/ethnic types have been scrubbed from the data set, so no personal information could potentially be lost or stolen. The school district will oversee curating any future data that they wish to include in the existing data set. As the data set grows the application can be fitted with a username and password function, but that will not be included in the first version. Any and all communications that do include any personal information will be sent via an encrypted email and or zipped with the password for those attachments to be sent in a separate email.

11 Expertise

I have been working in the IT field for over 10+ years and my wife is an educator, so I am acutely aware of how technology and the education sector go hand in hand. My team members will consist of a number of programmers with 5+ years of experience along with a team of seasoned testers that can shake out any bugs before the application goes live.

Section B

1 Decision-support problem

The main issue hounding every school district is a lack of funding. With unlimited funding all projects and avenues could be explored, but in the real world there is only so much money to go around, so knowing where best to spend that money is key. Taxpayers are not always willing to pay more to fund projects they feel will not give an adequate return on their investment. With our application you will be able to easily visualize data points in relation to an existing test preparation program. By using historical data, this browser-based application can project changes to test scores based on the use of an existing test preparation program and the demographics of the student. Using Python as its underlying base and taking advantage of the Pandas, Numpy, Pyplot, and sklearn libraries we will be able to use data analysis standards and methods to accurately derive useful information from the data set. As long as the data format stays the same more data can be added thus increasing the accuracy of the program as time goes on. As it currently stands all data regarding the current test preparation program is calculated by hand and is a time consuming process, but with this application that time can be cut down significantly.

2 Customer Description

This application is designed mainly for the use of school board members or school staff. The need for this product comes from the constant need to spend funds wisely since budgets are always tight when it comes to schools. Being able to apply funds where they are needed most is crucial to the success of any school and its students. Being able to accurately predict which students will benefit the most from a program will help alleviate overcrowding of classes and ideally increase a student’s engagement with the class since they will not be bored. If a student feels like they are not benefiting from a class, they are less likely to pay attention and more likely to cause disruptions in class.

3 Existing System

The current system they are using is a manual one that is very time consuming to correlate all the data into anything meaningful.

4 Data Set

The data set we are using is one from the school system that they have already established. The data is in a .csv file and includes all the data that is needed to predict score changes. As of now it contains gender, race/ethnicity (labeled as groups so as to avoid any perceived bias), scores in math, reading, writing, lunch type be it free or standard, parental level of education, and whether they have completed a test preparation course or not. This data was collected from: <https://www.kaggle.com/spscientist/students-performance-in-exams> The school will add to this data set as time goes on, but it must remain in its current format.

1. Methodology

We will be using an Agile development methodology. Continuous testing and feedback from the customer will ensure their needs are being met by the application. The iterative process we will use will allow for changes to be incorporated during development in a relatively short timeframe.

Requirement phase --- Working with the customer and the team lead a clear outline of the application requirements will be drawn up. This will give us a good idea of what items to include in the final product, so time is not wasted on features that are not wanted or needed.

Development phase --- The requirements we gathered will then be used to begin the development of the application being sure to use an iterative approach. At defined intervals we will present the current version of the product for the client to review. We will then run testing based on customer feedback and iterate further. Black box testing will be used to ensure proper data projections using defined inputs, as well as full systems testing.

Delivery --- Once development has concluded the application will be delivered to the client for final review and acceptance. Instructions for running the program will be included along with training and support.

1. Deliverables

The deliverables for this project will consist of a breakdown of costs associated with the project, a sample data set for testing and demonstration purposes, and the application itself. The application will be a Jupyter notebook that contains all projection models as well as data visualizations.

1. Implementation

The initial implementation of the application will be a rough framework that will consist of the basic functionality of the application. The sample data will be used to generate multiple graphs along with running an initial prediction model. Once these predictions and models have been created, they will be presented to the customer and subsequent changes will be made. All changes and upgrades will be documented, so cost overruns can be monitored. Once the application receives acceptance from the client, we will incorporate their actual data for final testing. As this program is browser-based there is no need for any hardware rollouts. For the first year we will aid in maintaining the database to ensure it does not become corrupt, but we will also train client staff on how to properly enter new data. Updates will be performed on an as needed basis so long as their support contract dictates.

1. Validation

There will be a small team of testers who will test and validate the sample data. They will take previous reports that were done by hand by the client and compare the results that the application provided to ensure accuracy. Multiple iterations of the application will then be presented to the client for feedback to ensure acceptance goals are being met. The projection model will be verified by our team using known data points and adjusted if any discrepancies are encountered. Once the client has given sign off on the performance of the sample data we will then progress to the larger data set of the client. Graphical output will then be compared to what was previously done by hand to ensure accuracy of the program.

1. Costs

The programming environment being used is an open-source platform that is free to so there is no cost associated with that as such. Also, as this is a browser based application there is no special hardware needed to run the application by the client other than their current PC setup. The human resources portion are as follows: Development team – 5-person team – 210 hours - $21000, Testers – 50 hours - $5000, and a 10% buffer - $3000 for any over runs.

1. Timeline

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Milestone | | Start Date | End Date | Acceptance Criteria |
| Requirement clarification | 3/1/2023 | | 3/5/2023 | When all requirements are agreed upon |
| First application iteration | 3/5/2023 | | 3/12/2023 | Present application to testers with demo data for review |
| Testing | 3/12/2023 | | 3/14/2023 | Testers verify that data is displaying correctly |
| Customer data/visualization data added | 3/14/2023 | | 3/24/2023 | Customer database is used to replace demo data and graphical representations are added |
| Testing | 3/24/2023 | | 3/26/2023 | Testers verify results for consistency and graphs for accuracy |
| Final iteration | 3/26/2023 | | 4/1/2023 | Acceptance by the customer of the final product |
| Maintenance | 4/1/2023 | | 4/1/2024 | 1 year of updates on an as needed basis along with training. |

Section D

1 Business Vision

Acme originally created this program because it saw a need in the education sector for cost cutting. Funding was an ever-present issue with many school districts, so many needed programs were being cut due to poor performance. It was this need Acme endeavored to fill with a relatively simple program. Acme strove to create a modular program that with certain data sets will predict which students will benefit from certain classes thus saving the school money by getting the right students into the right classes. With this program Acme will be able to customize it for many different school districts and their needs giving Acme a larger market share of the educational IT budget. By selling licensing and support to the program it will generate renewable income which in turn will fund innovation within the company.

2 Dataset

The data that was given to Acme had already been cleaned and formatted properly. We did add a check to ensure there were no null values with that code being: score\_df.isnull().sum() which in this case returned all 0s since there were none. The data set came as a .csv file and was simply adapted to the program. Below is a sample of the data set as it was given to Acme.

"gender","race/ethnicity","parental level of education","lunch","test preparation course","math score","reading score","writing score"

"female","group B","bachelor's degree","standard","none","72","72","74"

"female","group C","some college","standard","completed","69","90","88"

"female","group B","master's degree","standard","none","90","95","93"

"male","group A","associate's degree","free/reduced","none","47","57","44"

"male","group C","some college","standard","none","76","78","75"

For ease of coding no further cleaning was done of the data and the set was passed straight into the program. Once into the program it was formatted thusly:

| **gender** | **race/ethnicity** | **parental level of education** | **lunch** | **test preparation course** | **math score** | **reading score** | **writing score** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | female | group B | bachelor's degree | standard | none | 72 | 72 | 74 |
| **1** | female | group C | some college | standard | completed | 69 | 90 | 88 |
| **2** | female | group B | master's degree | standard | none | 90 | 95 | 93 |
| **3** | male | group A | associate's degree | free/reduced | none | 47 | 57 | 44 |

3 Analysis Code

The projection model was created using historical student data with each student’s testing data being grouped with other students who have similar demographics. Furthermore, they were then grouped into students who had taken the test preparation course and those who had not. This is formed the basis for the model. Then the averages of the test scores would be calculated for both the students who took the course and those who did not. Those numbers would be compared grouping them by similar demographics and this would allow the user to see the projected test score changes.

Once that was finished to add value to the program, we decided to run the data through another prediction model. This model was a random forest regressor that used all the data available to attempt to predict future cumulative test scores. We accomplished this by first turning all data that wasn’t a number into a numeric form. We then did a training and test split which means that the training set contains known output from which the model learns off of. The test set then tests the model’s predictions based on what it learned from the training set. We set the number of estimators at 100 currently since that is the default for Scikit Learn, but it can be increased or decreased as needed. The higher the number of estimators the longer the code will take to execute though. Once that was finished, we then plotted the resulting data on a scatterplot to help illustrate the accuracy of the predictions versus actual cumulative testing data. We then calculated the R^2 score which tells us how well our model is fitted to the data by comparing it to the average line of the dependent variable. The closer this number is to one indicates how well our model performs. When we ran it our R^2 number came back at .9998 , so we had an accuracy score of 99%. The mean absolute error, mean squared error, and the root mean squared error were then calculated to further evaluate the performance of the model. The MAE came back at .18426, the MSE was at .2324, and the RMSE was at .4820 so overall the performance of the model was quite good. These results can be shown to the customer to help encourage future versions of this program that will expand its prediction capabilities and to help display the accuracy of the predictions. The reason we decided to add the random forest regressor was twofold. First and foremost was that it can help predict future scores and that is exactly what the customer needed. Secondly thanks to its ‘wisdom of the crowds’ approach, random forest regression can achieve extremely high accuracies. While the random forest regressor in its current state will not predict test score changes based on the test preparation class it does give a general idea of where it thinks the overall test scores will be in the future.

Included in the program were a number of different graphs that help represent the student distribution. The first of the graphs breaks out each of the test scores by sex and whether they took the test preparation course or not. The second gives an overall distribution of the education level of the student’s parents and then a breakdown of cumulative scores vs parental level of education. Next, we show a histogram for each the three different tests along with a cumulative score. We then go one step further and show a boxplot of the test scores versus having taken a test prep or not. Lastly a visualization of the data was created to show projected changes to test scores with a drop down for the use to select the particular parental education level they wanted to see.

4 Hypothesis

The original hypothesis was that a student with parents who had a higher education level would benefit less from a test preparation class versus a student whose parents had less education. After running the program with the data on hand it seems our hypothesis was correct. Students whose parents had the lowest education level showed the greatest increase in test scores where conversely the higher education level parents had students who only got a slight bump in test scores. By adding more students to the data set we would hopefully see a greater increase in those scores.

5 Visualizations

The first graph was created by breaking down the data first by gender, then by whether a test prep was completed or not, and lastly by subject such as math, writing, and reading. These graphs did not show any major gaps between the sexes as far as test score. It did show that both genders faired better on all tests when they took a test preparation course. Due to these findings, it was determined to disregard the gender aspect of the model and focus more on the parental level of education.

The second visualization was a pie chart with a breakdown of the education levels of all the students. In this data set only 17.70% of the students had parents that did not have some kind of college education with the largest portion of the parents having some kind of bachelor’s degree. This kind of a chart will be helpful to the client because it is very easy to see the breakdown of their school and will allow them to quickly see what percentage of the population will be affected by whatever decision they come to.

The third is a box graph that shows the cumulative score versus the parental level of education which helps to further support our hypothesis. The boxes represent the scores that make up the 25th to 75th percentile of each level. The median of all scores for that demographic is denoted by the line inside the box. The dots represent outliers that could not be grouped. Each line extends to include all the scores minus the few outliers. The data shows that as the education level seems to decrease the test score spread seems to increase as well with the largest spread coming from the some high school group.

The fourth display is a histogram of each subject along with a cumulative score. These graphs include all scores so the school board can get an overall feel for totals for each subject. They can see at a glance where the bulk of the students fall when it comes to each of the three subjects.

Next, we have another boxplot, but this time it is of the range of scores in each subject versus having had a test preparation class or not. With these graphs the school board can clearly see the benefit from taking the test preparation class. The range of scores shrinks considerably when you look at having taken a test preparation class versus having not.

The second to last display is an interactive graph that shows the projected score change by parental education level when a test preparation class is taken versus not. This graph is what ties it all together in so much as that the school board can plainly see the increase in scores. The rate of rise is not as steep the higher the education level gets which indicates less of an impact the test preparation class having. The data in the some high school group show an almost 10 point increase in the writing score versus only a 2 point increase in math score for a student with a parent who has a master’s degree.

Lastly, we added a scatter plot of the data from when the random forest regressor was ran. This gives a visual representation of how the programs predictions of cumulative scores faired versus the actual scores. By looking at the mean absolute error, mean squared error and the root mean squared error you can see quite clearly that the predictions were not very far off from the actual scores.

6 Accuracy

The projection model only focused on the differences in scores based on the level of education on the parent mainly due to the fact the application would become skewed as more criteria were added. With the current data set we sat a 10-point boost in writing by taking the test prep class if the parent had some high school versus only 6 points if the parent had a master’s degree. These gaps would be likely to increase as the dataset grew. We then used a second prediction model to attempt to predict future cumulative scores in general this model was a random forest regressor. We accomplished this by first turning all data that wasn’t a number into a numeric form. We then did a training and test split which means that the training set contains known output from which the model learns off of. The test set then tests the model’s predictions based on what it learned from the training set. We set the number of estimators at 100 currently since that is the default for Scikit Learn, but it can be increased or decreased as needed. The higher the number of estimators the longer the code will take to execute though. Once that was finished, we then plotted the resulting data on a scatterplot to help illustrate the accuracy of the predictions versus actual cumulative testing data. We then calculated the R^2 score which tells us how well our model is fitted to the data by comparing it to the average line of the dependent variable. The closer this number is to one indicates how well our model performs. When we ran it our R^2 number came back at .9998 , so we had an accuracy score of 99%. The mean absolute error, mean squared error, and the root mean squared error were then calculated to further evaluate the performance of the model. The MAE came back at .18426, the MSE was at .2324, and the RMSE was at .4820 so overall the performance of the model was quite good. While the random forest regressor in its current state will not predict test score changes based on the test preparation class it does give a general idea of where it thinks the overall test scores will be in the future. As the dataset is expanded, we hope that will help train the model, so it can make increasingly more accurate predictions. The charts and graphs help to illustrate this data and the score variances. We feel that the application in its present state is complete in its accuracy to project change in test scores, but further expanding the dataset will only help improve accuracy in the future.

7 Testing

A small testing team was used to verify the data after each iteration of the program. They would review previously calculated data from the customer to ensure that what was being returned by the current version was accurate. The projection model has limited inputs currently in the drop down as it would crash if unknown variables were passed to it. To ensure its operation it was given a minimal set of inputs but can be expanded in future versions.

Each iteration of the program after testing was passed on to the client for acceptance and to get their feedback on features they would like to see or things they wanted removed. This allowed us to further customize the program to meet their specific needs and goals for this program.

8 Files

All files have been submitted along with this document.

9 Quick Start

1. Ensure whichever browser you plan on using has all the latest patches installed first.
2. Navigate to the website <https://github.com/manson627/c964_test>
3. Towards the bottom of the page there is a button that says, “Launch binder” This will launch the environment with the Jupyter Notebook that contains the application. Click the button
4. Once the application has loaded all dependencies and launched double click the Task2.jpynb icon on the left of the screen. This will bring up the application.
5. Click the 2 triangles to restart the kernel and run all cells. A pop up will appear confirming this is what you wish to do. Click the red restart button.
6. Tables and charts will then start to populate to show that everything is running as it should.
7. Any code changes made here are separate from the GitHub repository so they will not be permanent.

Reference

Seshapanpu, J (2018) Students Performance in Exams, Marks secured by the students in various subjects. Kaggle.

https://www.kaggle.com/spscientist/students-performance-in-exams