FINAL REPORT FOR DISTRICT SCHOOL PERFORMANCE PREDICTION

MEMBERS:
Adhithiyaraj Sankaranarayan
Ashwin Ashokan
Krishna Sriram
Malliga Suresh Babu
Ramyaa Rathna Kumar

Two Paragraph Summary:

The stakeholder for this project is Educational Leadership Research Center (ELRC), Texas A&M. They wanted to develop an AI program to determine if a school is at risk of failure in the state testing and provide this as a service to different schools. Currently, it is manually determined whether a school is at risk of failure using data from individual student scores and different variables. This process is very time-consuming. Hence, the user would benefit from an AI application, where he/she will be able to upload the individual student data for prediction and get an automated prediction from a trained model that outputs the risk of failure for specific schools.

The work done is an extension of the model developed by a previous team. For this semester, the customer wanted us to focus on improving their existing machine learning (ML) model to predict student test scores for four different grades (3,4,5, and 6) and two subjects (Math and Reading). We were provided with raw data from Hearne school that had individual-level student data on past test scores, and other features such as ethnicity, early childhood development, English proficiency, socioeconomic status, and past two years' score for prediction. Multiple models in Linear Regression, Polynomial Regression, and Random Forest were developed to consider high test accuracy and interpretability. Linear regression models provided maximum accuracy and met the customer expectations. Finally, the end result was to collaborate with the front-end team to generate a working website with the required inputs and provide the end result in a CSV format.

Description of user stories:

- **Check if the new data is compatible with the existing ML model:**
  Points Assigned: 2
  Status: Completed
  Description: This user story was to check if the pre-existing model developed by the previous team were compatible with the new data provided by the client. The new data had extra features(x-variables) that needed to be compatible with their previous ML

model and we also planned to verify its accuracy based on the new training data, just to get a brief understanding of the project and the impact the new and pre-existing features had on the accuracy of the previous model. The second task of this User Story was to understand the basics of data cleaning code that the previous team implemented, and we used Jupyter notebook cells to understand their data flow for each set of independent routines. This user story was a brief introduction to the expectations of the project and trying to understand the previous team implementation.

- **Enhance the existing model to predict better than the human model:**

  Points Assigned: 3

  Status: Partially completed

  Description: The human model, which in turn are the personal predictions from the customer is still incomplete and we don't have those predictions yet. With the predictions from last year (although with different features), our current model predicts much better.

- **Discuss with the front-end team to agree on API and services for integration and presentation of the predicted results:**

  Points Assigned: 3

  Status: Completed

  Description: The front-end team wanted our trained model as a callback function for their Django app whenever a client uploaded a CSV file, their app would call our high-level routine, which would initially check if the corresponding schools' data and the subject has a pre-trained model, if so our routine would fetch the pickle file and predict the data, otherwise if we didn't have a pre-trained model for the given school and subject, we would train a model and save the pickle object in our folder. When returning the predicted results, we return it as a CSV file saved in a temporary folder. The front-end team would use that file for presentation on their web UI. Our routine also does pre-processing of the uploaded file to fill in NaN values and uses that for training and prediction.

- **Extract Dominant Features that are Responsible for the Students Predicted Scores:**

  Points Assigned: 2

  Status: Partially Completed

  Description: Our goal was to extract features that had the highest impact on students scores, but due to limited data set approximately 40-50 student scores/school/grade/subject, and the disproportionate number of datasets that represent a specific socioeconomic status, we weren't able to concretely say that the dominant features that we extracted were the actual dominant and deterministic features. Based on

the current dataset we believe it is highly difficult to extract the relevant features using any ML model.

Roles held by each team member:

- Adhithiyaraj Sankaranarayanan - Developer
- Ashwin Ashokan - Developer
- Krishna Sriram - Product Owner
- Malliga Suresh Babu - Developer
- Ramyaa Rathna Kumar- Scrum Master

There were no changes in the assigned roles during the implementation of the project.

Points completed for each iteration:

- **Iteration 0:** For this iteration, we met with Dr. Beverly Irby (head of this project) and Sukanya Sravasti (prior project owner) over a brief phone call and got an overview of the project. We created customer user stories and decided on the deliverables. A storyboard was then created keeping in mind the flow. (0 Points)

- **Iteration 1:** In this iteration, the user story implemented was to check if the new model was compatible with the existing model. We were able to extract features with high correlation and were unable to verify their validity due to the limited availability of the data set at that point in time. Several models were being implemented during the phase. Preprocessing like One hot encoding, filling the missing values, etc were completed and with this preprocessed data, a linear regression model was fit for both the subjects and, the 2019 STAAR scores were predicted for the students. (2 Points - user story 1)

- **Iteration 2:** In this iteration, we had completed the implementation of three different models: linear regression, random forest, and polynomial regression. We observed that Linear regression was the best model (figure 1) and hence extended that for the current 5^{th}, $6^{th,}$ and 7^{th} grades. To extract dominant features that are responsible for the students' predicted performance, we used the coefficients of our model as a measure of the feature's importance. We had also met with the front-end team Spongebob and discussed the required format of data for the web app. We decided on the interface for calling our saved models, storing trained models, and cleaning the standardized input CSV files. (2 + 1 Points - user stories 2 and 4)

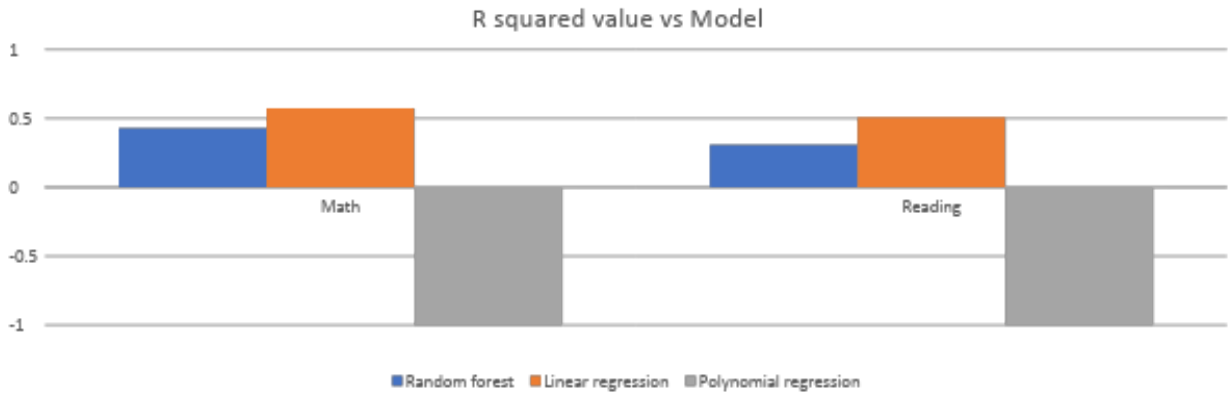

Figure 1: Model comparison chart for grade 4 Math and Reading

- **Iteration 3:** For this final iteration, we have completed our fourth user story: the integration of our ML model with the web app, by coordinating with the front-end team. We verified the correct working of the website to train the model or predict scores using the data provided by users. We have also received positive feedback from our customers for the successful implementation of our deliverables for the project. (3 Points - user story 3)

List each customer meeting dates and what you did:

- April 09 2021 at 8:00 PM: We met with Dr. Beverly Irby (head of this project) and Sukanya Sravasti over a brief phone call and got an overview of the project.

- April 10 2021 at 4:45 PM: We met with the customers over zoom to learn about their requirements and decide the project goals. The customers were also intimated about the shorter timeline this semester.

- April 13 2021 at 6:30 PM: In this meeting, we discussed the modification of a new set of raw data obtained from a district school user. We finalized the required features and output variables for each set of data and the actions to be performed on missing values were sorted. We confirmed the data format required for preprocessing and modeling with the previous product owner and the final format was modified according to code requirements. The customer wanted to implement the new data on the existing model to verify the accuracy of the model over a new dataset.

- April 20 2021 at 6:30 PM: In this meeting, we discussed the kind of metrics required for evaluating the model. We informed them regarding our meeting with the front-end team Spongebob on April 23 and discussed the planned agenda for the meeting. Following our discussion on the limited data size, we concluded to work with the training data even

though its size is comparatively small. The customers also proposed to pool in data of both the schools for increasing data points.

- April 28 2021 at 6:30 PM: In this meeting, the customer provided us with a new dataset for another school, Carmichael for grades 3,4, and 5. We were informed about fewer test scores available for the new school(Carmichael) and we discussed the resulting limitations in combining the dataset of two schools. We decided to compare the performance of our model for the combined dataset of the two schools with fewer STAAR scores, and one school with more STAAR scores. The dataset with better performance is to be finalized. We also discussed our ongoing work with the front-end team to integrate our model with the web app. The customers expressed the necessity of verifying the working of the website created by the front-end team using our model.

- April 28 2021 at 8:00 PM: We intimated the customer about the better performance of our model with the dataset of just one school over the appended dataset with lesser STAAR scores. We also discussed the time-crunch expressed by the front-end team to integrate with our model and the possibility of delay in integration. The customer verified the working of our model and expressed their satisfaction with our work.

Explain your BDD/TDD process:

We followed a BDD process where we incorporated weekly feedback from the client into our project goals. The client emphasized that they wanted an advancement of the existing ML model that would enhance the current predictions to serve as a working prototype when integrated with the frontend team. They wanted a complete web application that would cater to individual school's student performance for the STAAR score and they also wanted relevant features that had the highest impact on students' scores. We experimented with different models with an updated dataset and obtained a linear regression model that performed slightly better than the existing testing data that was evaluated by the customer.

Standardization:
We worked with the customer and the front-end team, emerging on a conclusive decision to develop a standardized interface for representing student scores on the input files. This is done to make it easier for parsing the data from our side and also for an interactive presentation by the front-end team. We also decided upon a simple interface that would be used by the front-end team to utilize our model services for inference and training.
The website takes in user-specified input format for the factors considered to predict test scores, which is again generated in the required CSV format, meeting the client's expectations.

Links to your Pivotal Tracker and GitHub repo:

GitHub:
The Github link for our team's project is:
https://github.com/Krishna2201Sriram/CSCE-606-thinkSmart

Pivotal Tracker:
https://www.pivotaltracker.com/n/projects/2495261

Links to demo and poster presentation:

Video:
https://www.youtube.com/watch?v=FwMyS-K1cew&t=2s&ab_channel=adhithiyarajsankaranarayanan

Poster:
https://github.com/Krishna2201Sriram/CSCE-606-thinkSmart/blob/master/Documentation/SPRING_2021/Project_poster.pptx