Coding Challenge Write-up Vicki Wu

Background / Dataset Introduction:

MathE is a pioneering collaborative e-learning platform that enhances the users' mathematical learning processes in higher education. Its core objective is to cultivate virtual learning and foster knowledge exchange (Azevedo et al., 2024). The dataset has 9546 responses from students in 8 countries to mathematical questions in 14 topics taught in higher education. The file has eight features, named: Student ID, Student Country, Question ID, Type of answer (correct or incorrect), Question level (basic or advanced), Math Topic, Math Subtopic, and Question Keywords.

Focus Question:

Given the information provided by the dataset, the question raised is: How can we use students' country, question level, math topic, and subtopic to predict the correctness of the response? My main aim is to determine whether the response is correct or incorrect, making this a classification problem.

Model Selection:

I selected a random forest model to address this specific question because of the high number of categorical variables and the complexity of the dataset. The random forest algorithm is an ensemble learning method that combines the predictions of multiple decision trees to improve accuracy and generalization. It is particularly effective in handling large datasets with nonlinear relationships and categorical features.

Initially, my objective was to calculate the average correctness of student responses using a simple linear regression model. However, I soon realized that this calculation could be done directly without the need for modeling, so I shifted my focus to a classification problem. I chose to use logistic regression to predict whether a student's response would be correct or incorrect. Logistic regression seemed promising because of its simplicity and the linear relationships it assumes. However, I encountered challenges, particularly due to the large number of features and the issue of multicollinearity, where many features were highly correlated, making logistic regression unsuitable for this task.

To overcome these challenges, I decided to switch to a random forest model. Random forest can effectively handle large numbers of features, deal with multicollinearity, and capture complex, non-linear relationships. Additionally, a random forest's ability to average multiple decision trees reduces the risk of overfitting, providing more robust predictions. It also ranks feature importance, which helped me identify the most significant predictors, such as student country and question level, in determining the correctness of responses. This made random forest the most suitable choice for my dataset, especially considering its multidimensional and categorical nature.

Result and Discussion:

The accuracy of the model was 0.61, meaning it correctly classified about 61% of the test samples. This is relatively low and suggests room for improvement. The precision for the incorrect answers (0) is 0.59, and 0.62 for the correct answers (1). While the recall is 0.71 for class 0 and 0.55 for 1. The F1-score reflects the balance between precision and recall, and it

suggests that the model struggled more with identifying the “correct” answers, which is 0.55 compared to 0.65 for the “incorrect” answers.

The feature importance analysis shows that the model heavily relies on the student country, particularly Slovenia (0.205), as the most influential factor in predicting correctness, followed by question level (0.129). Other countries such as Italy (0.074), Portugal (0.072), and Lithuania (0.069) also play significant roles. While math subtopics like Elementary Geometry and Linear Systems contribute to the predictions, their importance is considerably lower. This suggests that the model is more focused on geographic and contextual factors rather than specific mathematical topics.

Since the support for 0 and 1 are 871 and 838 respectively, the bias should not be caused by an imbalanced class. The model's relatively low accuracy may be due to its over-reliance on student country, leading to potential bias and insufficient use of math-related features. Additionally, although I have selected the most significant topics, the low importance of math subtopics indicates that these features might need better preprocessing or aggregation to enhance their predictive power.

Links:

Dataset:

Azevedo, B. F., Pacheco, M. F., Fernandes, F. P., & Pereira, A. I. (2024). Dataset of mathematics learning and assessment of higher education students using the MathE platform. *Data in Brief*, *53*, 110236. https://doi.org/10.1016/j.dib.2024.110236

Data Cleaning:

Ang Li-Lian. (2022, July). *So You've Got a Really Big Dataset. Here's How You Clean It.* Medium; Towards Data Science. https://towardsdatascience.com/so-youve-got-a-dataset-here-s-how-you-clean-it-5d0b04a2ed86

Encoding Categorical Variables:

Moffitt, C. (2017). *Guide to Encoding Categorical Values in Python - Practical Business Python*. Pbpython.com. https://pbpython.com/categorical-encoding.html

Model Selection:

Ph.D, J. M., & Kavlakoglu, E. (2024, August 29). *What are classification models?* Ibm.com. https://www.ibm.com/topics/classification-models

GeeksforGeeks. (2024, February 23). *Difference Between Random Forest and Decision Tree*. GeeksforGeeks. https://www.geeksforgeeks.org/difference-between-random-forest-and-decision-tree/

Evaluation Metrics:

Kumar, S. (2021, July 20). *Metrics to Evaluate your Classification Model to take the Right Decisions*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/