Findings

KNN Model:

The KNN model was trained on the dataset with missing values replaced and scaled features. It achieved an accuracy score of approximately 46.67% on the training data. The predicted skill levels were categorized into six classes: 1, 2, 3, 4, 5, and 6. The distribution of predicted skill levels is as follows: [1, 2, 3, 4, 5, 6].

Random Forest Model:

The Random Forest model was trained on the dataset with missing values replaced, scaled features, and using the same training-test split ratio. It achieved an accuracy score of approximately 37% on the test data. The predicted skill levels were categorized into seven classes: 1, 2, 3, 4, 5, 6, and 7. The model's performance varied across different skill levels, with f1-scores ranging from 0.21 to 0.50.

Comparison:

The KNN model achieved a higher accuracy score (46.67%) on the training data compared to the Random Forest model (37%) on the test data. However, the Random Forest model provides a more detailed classification with seven skill levels compared to six in the KNN model. The Random Forest model shows varying performance across different skill levels, with higher scores for skill levels 4, 5, and 6.

Please note that these findings are based on the specific dataset and model configurations used. It's important to consider other factors and conduct further evaluation to make more robust conclusions.

Based on the EDA and model results, here's my advice to the stakeholders regarding collecting more data:

KNN Model Performance: The KNN model achieved an accuracy score of approximately 46.67% on the training data. While this accuracy score is better than random guessing, it indicates that the model's performance is not very high. The model's ability to predict the skill levels of players could potentially be improved with more data.

Random Forest Model Performance: The Random Forest model achieved an accuracy score of approximately 37% on the test data. This score suggests that the model's performance is also not very high. However, it's important to note that the Random Forest model provides a more detailed classification with an additional skill level.

Based on these findings, my advice to the stakeholders would be as follows:

Collecting More Data: Collecting more data is likely to be beneficial for improving the performance of both models. A larger and more diverse dataset can provide the models with more examples and patterns to learn from, potentially leading to better predictions.

Data Collection Strategy: When collecting more data, it is crucial to ensure that it is representative of the target population. The data should cover a wide range of skill levels, playing styles, Unique Units Made, Total Map explored and other relevant characteristics to capture the true variation in the dataset.

Feature Selection and Engineering: During the data collection process, it would be beneficial to consider additional features that may have an impact on player skill levels. Gathering more comprehensive information about player behavior, strategies, or game-related metrics could potentially enhance the models' predictive power.

Model Evaluation and Comparison: Once the new data is collected, it's important to evaluate the models again using the expanded dataset. Compare the performance of the KNN model and the Random Forest model using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score to determine which model performs better on the updated dataset.

Overall, collecting more data and re-evaluating the models on the expanded dataset is a valuable step to potentially improve the accuracy and predictive power of the models. It's important to emphasize the need for quality data, proper feature selection, and continuous model refinement to achieve the best possible results.