**Question 1) Describe your strategy for addressing challenges such as handling missing values and categorical features. Could you also elaborate on your feature selection metrics and explain the rationale behind them?**

Since we only had 34 rows with missing values and all these 34 rows had missing values in both the outcome column and the interferenceOnShooter column, considering the small proportion of 34 rows to the total data and the fact that our analysis was focused on the outcomes, we removed them. For the interferenceOnShooter column, we used ordinal encoding because it had values: low, medium and high, and for the other two columns playType and bodyPart, we used a one-hot encoding to give us the best output. We rounded minute and second values, and instead of using x and y, we used distance and angle.

**Question 2) Why didn’t we use regression to predict whether a shot would result in a goal?**

1. Non-linear relationships: The relationship between the predictors (e.g. shot distance, angle, player position) and the outcome (goal or no goal) may not be linear. Regression assumes a linear relationship between the predictors and the outcome, which may not be the case in this scenario.
2. Classification problem: Predicting whether a shot results in a goal is a binary classification problem (goal or no goal), rather than a continuous regression problem. Classification algorithms, such as logistic regression or decision trees, are better suited for this type of problem.
3. Imbalanced data: In soccer, the number of shots that result in a goal is typically much smaller than the number of shots that do not result in a goal. This can lead to imbalanced data, which can affect the performance of regression models. Classification algorithms are better equipped to handle imbalanced data.
4. Non-linear interactions: There may be complex interactions between the predictors that affect the likelihood of a shot resulting in a goal. Regression may not be able to capture these non-linear interactions effectively.

**Question 3) How would you go about verifying the accuracy of the given formula used to calculate the shot angle in the preprocessing section?**

There was a ready-made function that we used to avoid verifying and we made sure that the resulting angle falls within the first and fourth quadrants of the trigonometric circle.

**Question 4) Discuss the advantages and disadvantages of k-fold cross-validation. Can you also explain other types of cross-validation methods that could address the limitations and issues associated with k-fold cross-validation?**

Advantages of k-fold cross-validation:

1. It provides a more reliable estimate of model performance compared to a single train-test split.
2. It utilizes the entire dataset for both training and testing, leading to a more robust evaluation.
3. It helps in reducing variance in the performance estimate by averaging results from multiple folds.
4. It can be used to tune hyperparameters and evaluate model generalization.

Disadvantages of k-fold cross-validation:

1. It can be computationally expensive, especially for large datasets or complex models.
2. It may not be suitable for time-series data or when there is a specific temporal order in the data.
3. It may lead to overfitting if hyperparameters are tuned based on the validation set in each fold.

Other types of cross-validation methods that can address the limitations of k-fold cross-validation include:

1. Leave-One-Out Cross-Validation (LOOCV): In LOOCV, a single data point is held out as the validation set, and the model is trained on the remaining data points. This process is repeated for each data point. LOOCV is computationally expensive but provides a less biased estimate of model performance.
2. Stratified K-Fold Cross-Validation: This method ensures that each fold contains a proportional representation of each class in the target variable. It is useful for imbalanced datasets and can improve the generalization of the model.
3. Time Series Cross-Validation: For time-series data, where the order of data points is crucial, time series cross-validation methods like Time Series Split or Walk-Forward Validation can be used. These methods preserve the temporal order of data points during the cross-validation process.
4. Group K-Fold Cross-Validation: In cases where data points are grouped or clustered, Group K-Fold Cross-Validation ensures that data points from the same group are kept together in the same fold. This can prevent data leakage and provide a more realistic estimate of model performance.

Each cross-validation method has its strengths and limitations, and the choice of method should be based on the specific characteristics of the dataset and the modeling task at hand.

**Question 5): What metrics did you use to evaluate your manual implementations of multivariate regression and k-fold cross-validation, and why did you choose them?**