The Factors that Influence a Pokemon's Battle Performance

Authored By:

Alessandro Montenegro

December 8th, 2024

Introduction

The world of Pokemon has introduced a brand new exciting and competitive stage where the smallest margins make the biggest difference. Consequently, competitors are led to dive into ways in which they can optimize their Pokemon's Battle Performance. Battle Performance is composed of several factors that ultimately make a Pokemon stronger on the battlefield. For this research analysis, we will be attempting to identify what these factors are most impacted by to gain a competitive edge over the opposition.

The dataset used to complete this research project is the Pokemon Base Stats Dataset obtained from Kaggle. The dataset consists of 1,014 observations, where each one belongs to a unique Pokemon. Each observation includes 17 variables, all of which detail various attributes of each Pokemon. These attributes include the inputs of this project such as Height (`Height`), Weight (`Weight`), Catch Rate (`Catch.rate`), Base Experience (`Base.Exp`), Growth Rate (`Growth.Rate`), Special Attack (`Sp..Atk`), Special Defense (`Sp..Def`).

The dataset also includes the output or response variables that we will be attempting to predict in this project. In this case, we are attempting to predict a Pokemon's Battle Performance. Although a designated variable for Battle Performance does not exist in this dataset, it does include the four variables required to measure Battle Performance, which are Speed (`Speed`), Attack (`Attack`), Defense (`Defense`), and Hit Points (`HP`). For this project, we aim to answer *Which factors have the biggest impact on a Pokemon's Battle Performance?* To complete this analysis, we will use two different methods: a decision tree model and a random forest model.

Methods

Decision Tree Model

Decision trees are ideal for this project because of their interpretability, visual intuitiveness, and their capability to handle both quantitative and categorical variables. They are particularly useful for handling any potential nonlinear relationships without too much data processing. Like all models, decision trees have their disadvantages. Some of these disadvantages include their sensitivity to small variations in data, and their proneness to overfitting data, especially with complex trees.

The following are the model formulas that correspond to decision trees and the four chosen response variables:

- Speed ~ *Height* + *Weight* + *Catch.rate* + *Base.Exp* + *Growth.Rate* + *Sp..Atk* + *Sp..Def*
- Attack ~ *Height* + *Weight* + *Catch.rate* + *Base.Exp* + *Growth.Rate* + *Sp..Atk* + *Sp..Def*
- Defense ~ *Height* + *Weight* + *Catch.rate* + *Base.Exp* + *Growth.Rate* + *Sp..Atk* + *Sp..Def*
- HP ~ *Height* + *Weight* + *Catch.rate* + *Base.Exp* + *Growth.Rate* + *Sp..Atk* + *Sp..Def*

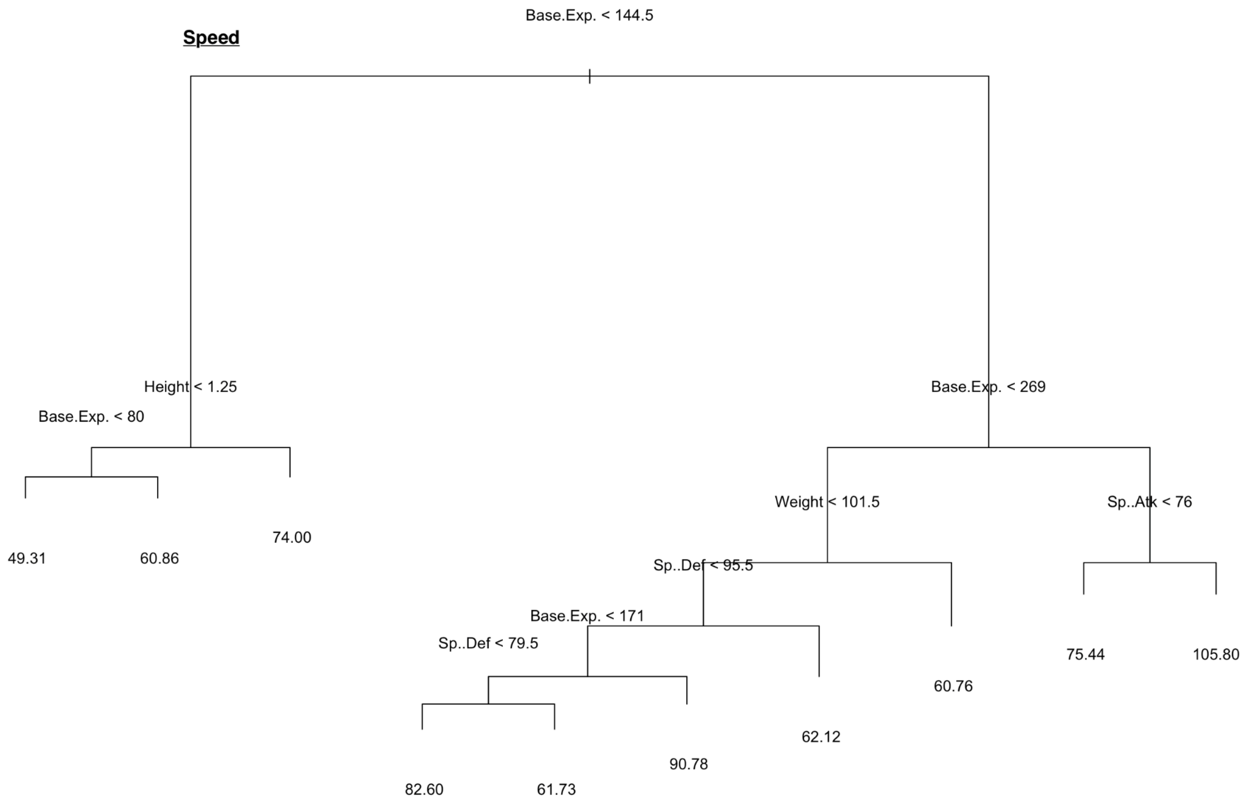
Throughout the process of fitting this model, two things were considered: the need for pruning, and excluding predictors. A decision tree can grow excessively complex especially if all splits are implemented, potentially leading to overfitting the training data. Pruning simplifies the tree by removing splits that do not significantly contribute to the model's accuracy, thus improving the model's ability to generalize to new data and avoid learning noise from the training set.

Subsequently, not all variables were able to be used as predictors when implementing a decision tree to any of the four response variables. Excluding the four response variables, there were a total 13 remaining variables in contention of being used. Predictors such as Type, Species, Abilities, and Gender were excluded due to excessive levels. Unfortunately, decision trees in R cannot handle any factor variable that contains over 32 levels efficiently. Additionally, the variables Name and Base Friendship were excluded because of the irrelevant information both variables contained. Lastly, any observations with missing fields were also excluded.

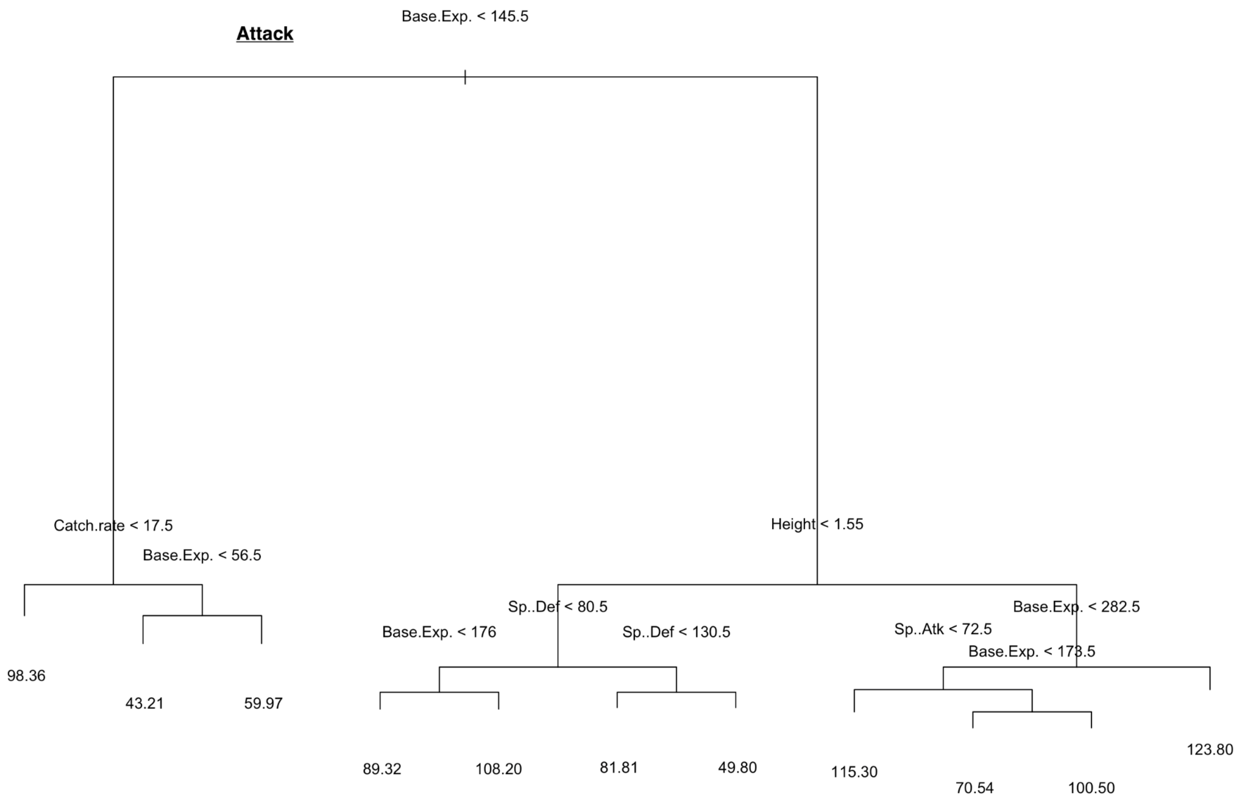
To perform this task, the data was first randomly subdivided 80% for training and 20% for testing. The models were executed using the training data for a total of ten iterations to account for randomness of the 80-20 splits. The mean of all ten root mean squared errors (RMSE), and the average R-squared values were the results considered when interpreting the outcome of the decision tree models. All of the decision tree models were performed under the set.seed(1) for reproducibility.

Results

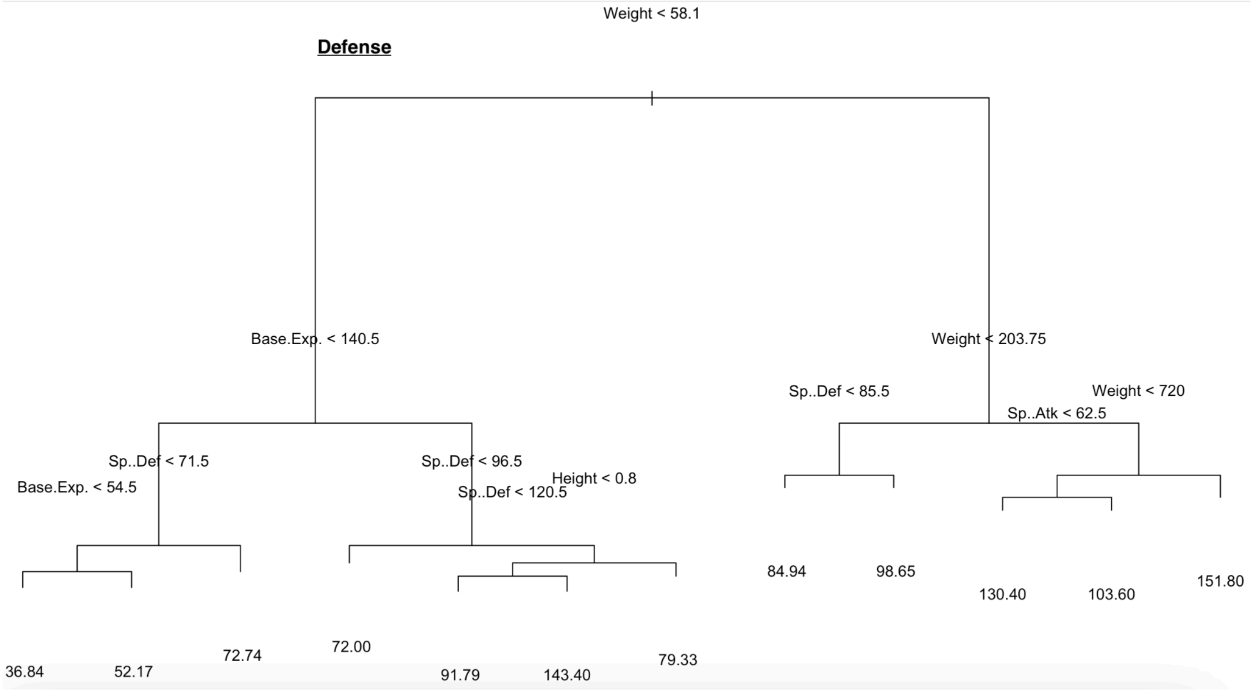
In this section, we present the results from our analysis using decision tree models to predict our four response variables (Speed, Attack, Defense, HP) based on the selected predictors. Key metrics such as tree visualizations, root mean squared errors, R-squared values, and overall interpretations are used to answer the primary research question. Below are four examples of pruned trees that were created from one of the ten iterations of each response variable.



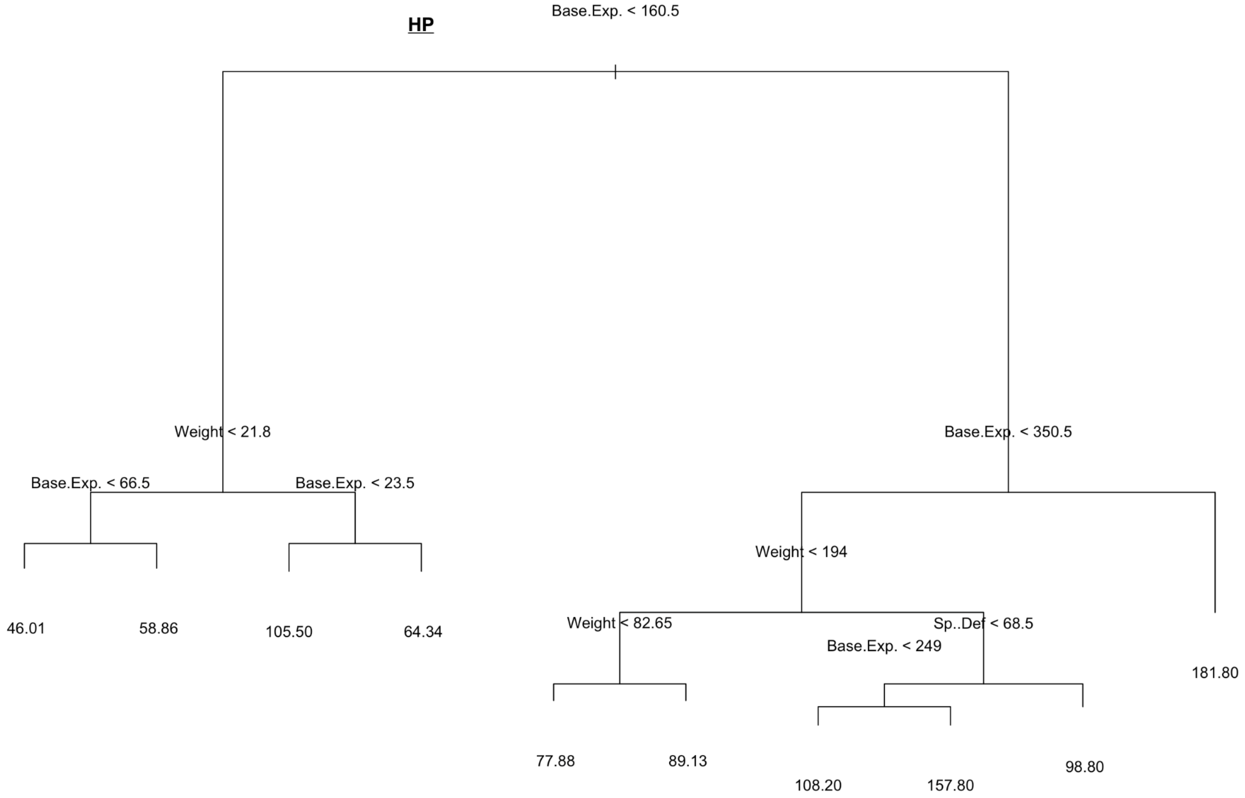
The pruned tree for Speed consists of ten terminal nodes. The most significant predictor was Base Experience followed by Height, Weight, and Special Attack.



The pruned tree for Attack consists of 11 terminal nodes. The most significant predictor was Based Experience followed by Catch Rate and Height.



The pruned tree for Defense consists of 12 terminal nodes. The most significant predictor was Weight followed by Base Experience.

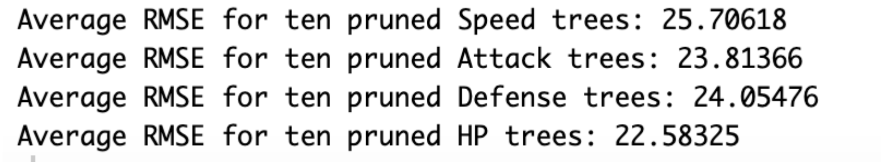


The pruned tree for HP consists of ten terminal nodes. The most significant predictor was Base Experience followed by Weight.

For the pruned Speed tree, the first big split occurs whether the Base Experience of the Pokemon is < 144.5 or not. If so, the next big split occurs at a Height of 1.25 (in meters), if the height is less than, there is an additional split of Base Experience, if not, a terminal node appears. Back to the original split, if Base Experience is greater than 160.5, then another Base Experience split occurs. If greater than 350.5, then a terminal node appears, if less, an additional four smaller splits occur yielding a total of five terminal nodes. A similar process occurs in the following three models where Base Experience is the first split, except in the Defense tree where Weight was first followed by Base Experience.

Throughout the four trees depicted above, the most significant variable used for predicting each of the four response variables was generally Base Experience followed by Weight then Height. Despite this depiction, each of the four pruned trees contained a large amount of terminal nodes, not only in the images above, but also on average throughout the ten iterations. A large number of terminal nodes indicates that the removal of any more nodes would significantly increase RMSE. Additionally, perhaps with the size of the dataset, or more variability with the predictors, the decision tree algorithm might require more splits to model the data effectively even after pruning. Despite pruning, the tree may still capture a significant level of detail in the data. This can be happening because the dataset does not have many nuanced patterns, or because the predictors are interacting in really complex ways

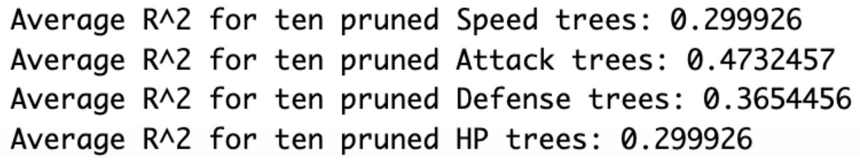
Next, we will take a look at how accurate these models are in predicting our four response variables. After running each of the four models for ten iterations, below are the mean RMSEs of all four models:



```
Average RMSE for ten pruned Speed trees: 25.70618
Average RMSE for ten pruned Attack trees: 23.81366
Average RMSE for ten pruned Defense trees: 24.05476
Average RMSE for ten pruned HP trees: 22.58325
```

Although the pruned decision trees were able to depict some complex trees across all four response variables, the average RMSE of all four models remained pretty high ranging from 22.58 to 25.71. A predictor error such as a high RMSE within this range indicates that all four models are not very accurate in predicting their respective response variables even after pruning. The reason being in part due to the complexity of the pruned trees. Even after pruning, it seems that the algorithm is having a difficult time predicting the testing data leaving doubt of perhaps still overfitting the training data.

Finally, we also calculated the R squared value of each of the four models to determine exactly how much the predictor variables were responsible for predicting each of the response variables. Below are the four R squared values followed by their interpretations:



```
Average R^2 for ten pruned Speed trees: 0.299926
Average R^2 for ten pruned Attack trees: 0.4732457
Average R^2 for ten pruned Defense trees: 0.3654456
Average R^2 for ten pruned HP trees: 0.299926
```

With the image above, all four models had somewhat similar R squared value averages throughout their respective ten iterations. The range of the four averages are between 29.99% and 47.32%. The lowest R squared average belongs to both the ten pruned Speed and HP trees while the highest belongs to the ten pruned Attack trees. The advantage of calculating these R squared averages is that they give us insight on what proportion of the dependent variables can be explained by our chosen independent variables. While ideally we'd like to have higher R squared values to better predict a Pokemon's battle performance, given that the chosen models were the decision trees, it's understandable why the R squared values are this low.

To conclude the debrief of the decision tree models, while the chosen predictor variables accounted for a significant amount of the variability of all four response variables, the decision tree model itself had difficulty in properly predicting the testing data with the available predictors. The analysis revealed that generally, Base Experience, Weight, and Height were the most impactful predictors across the four trees.

Random Forest Model

Random forests are well-suited for this project due to their robustness and ability to deliver highly accurate predictions. Random forests also excel at handling datasets with a mix of numerical and categorical variables and can capture complex interactions and nonlinear relationships in the data. However, random forests have their limitations, including their computational intensity, which can be a challenge for larger datasets especially if a large number of trees are used for training. They also have reduced interpretability compared to single decision trees, as it becomes difficult to visualize and understand the contribution of individual trees within the ensemble.

The following are the model formulas that correspond to the random forests and the four chosen response variables:

- Speed ~ *Height* + *Weight* + *Catch.rate* + *Base.Exp* + *Growth.Rate* + *Sp..Atk* + *Sp..Def*
- Attack ~ *Height* + *Weight* + *Catch.rate* + *Base.Exp* + *Growth.Rate* + *Sp..Atk* + *Sp..Def*
- Defense ~ *Height* + *Weight* + *Catch.rate* + *Base.Exp* + *Growth.Rate* + *Sp..Atk* + *Sp..Def*
- HP ~ *Height* + *Weight* + *Catch.rate* + *Base.Exp* + *Growth.Rate* + *Sp..Atk* + *Sp..Def*

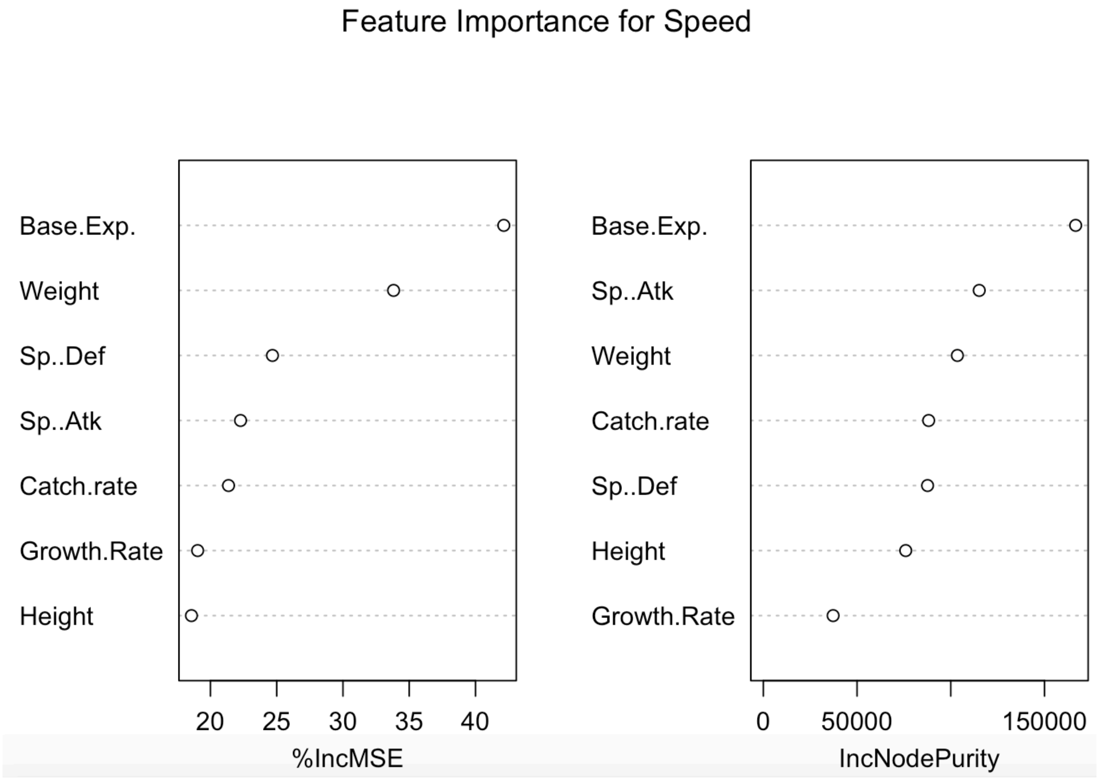
Throughout the process of fitting the random forest models, two things were considered: the amount of trees to be used in each random forest, and the essential predictors needed. The number of trees in each random forest was chosen to be 500 to account for both computational efficiency and model accuracy. Subsequently, the same predictors used for the decision tree models were used for the random forests due to the same problems faced when trying to include any of the previously excluded predictors.

To implement the random forests, the data was first randomly subdivided 80% for training and 20% for testing. The models were executed using the training data for a total of ten iterations to account for randomness of the 80-20 splits. The mean of all ten root mean squared errors (RMSE), and the average R-squared values were the results considered when interpreting the outcome of the random forests. All of the random forest models were performed under the set.seed(1) for reproducibility.

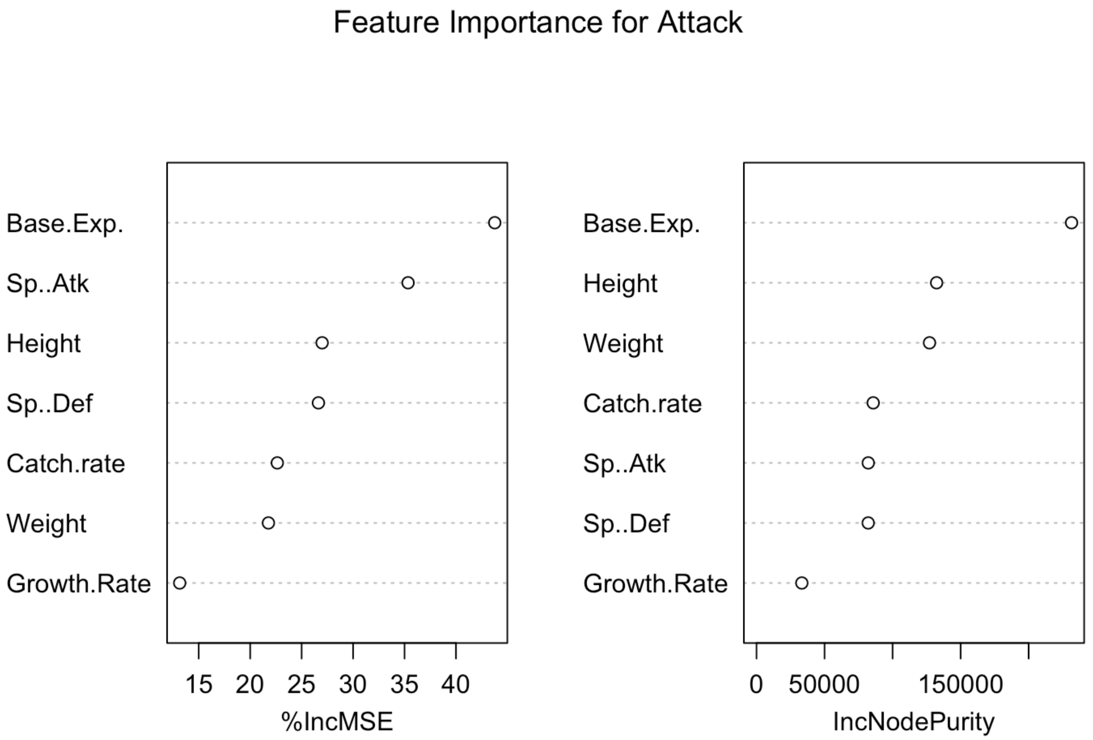
Results

In this section, the results from the random forest analysis will be presented to see how well they predicted the four response variables: Speed, Attack, Defense, and HP. Key metrics such as Feature Importance plots, root mean squared error values, R-squared values, and overall interpretations are used to answer the primary research question.

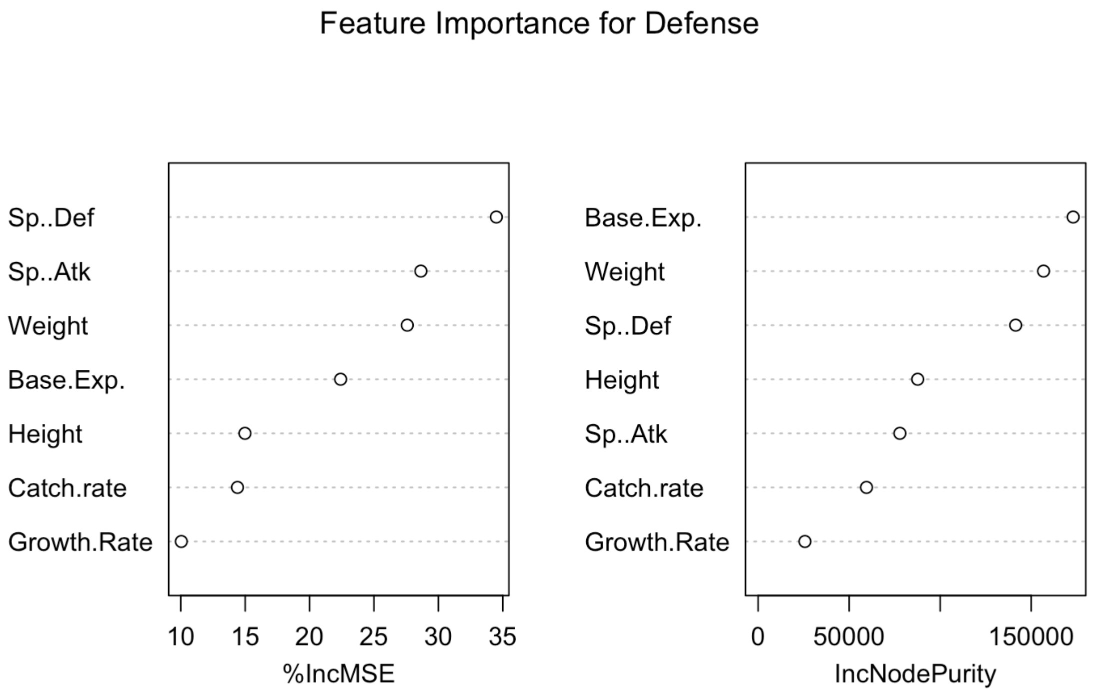
The first part of the results involves analyzing the respective Feature Importance plot of all four response variables. This plot displays two graphs, the first determines which variables are of most importance to the random forests. Since the random forest was implemented on each response variable a total of ten iterations, the four Feature Importance plots below are simply four examples from the forty that were generated.



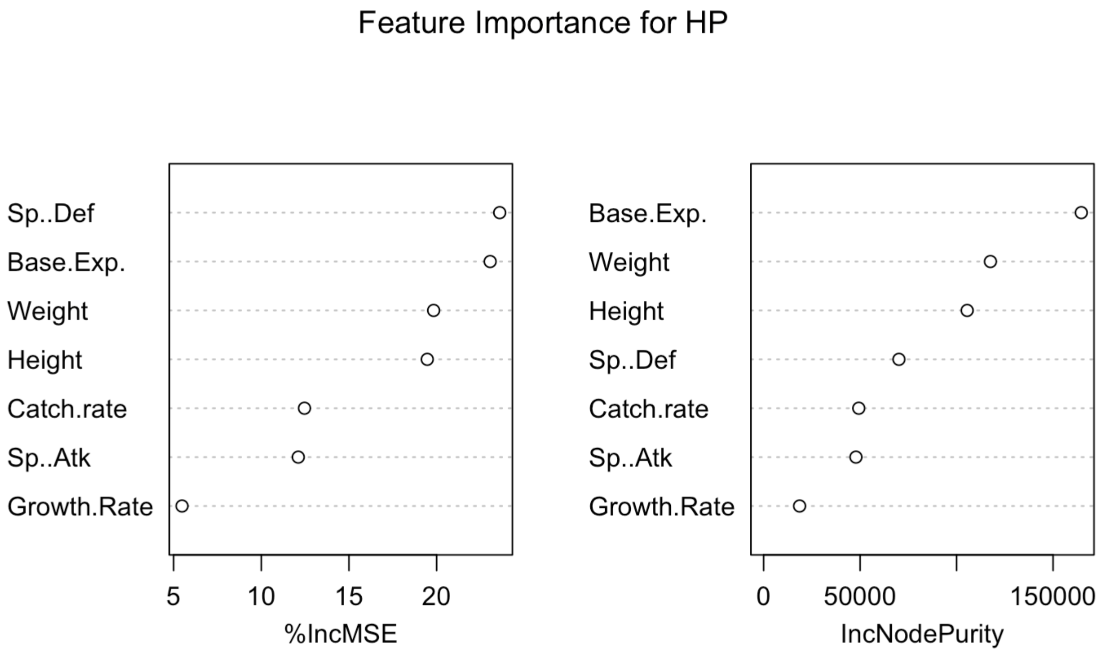
The Feature Importance plot of the Speed random forest depicts that Base Experience and Weight are by far the most important predictors. The random forest aligns with the decision tree example previously described in which Base Experience was also the most significant predictor, and Weight was the tree's third most important predictor.



The plot corresponding to the Attack random forest determined that Base Experience and Special Attack were the predictors with the most significance. The corresponding decision tree also determined Base Experience as its most important predictor, however, Special Attack was not among the most significant predictors of the tree.



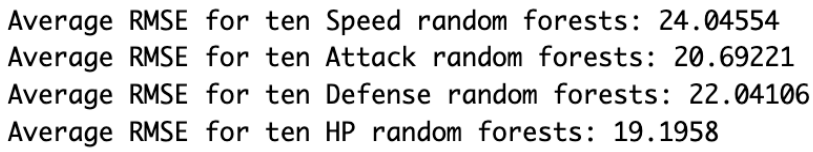
The Defense Feature Importance plot shows that Special Defense, Special Attack, and Weight are the three most significant predictors of the random forest respectively. Contrary to the respective decision tree, Special Defense and Special Attack were not part of its most important predictors. While Weight was the most important predictor of the Defense tree, it was only third in the corresponding random forest.



Finally, the HP plot determined that Special Defense and Base Experience were by far the most important predictors of the random forest. Similarly, Base experience was the decision tree's most important predictor, but Special Defense was not among the most important predictors.

Of the four examples of random forests used above, overall, Base Experience, Special Defense, and Weight were among the most important predictors in determining battle performance. This list of predictors is similar to the overall most important predictors of the decision trees: Base Experience, Weight, Height. The only difference of the two models is Special Defense from the random forests, and Height from the decision trees.

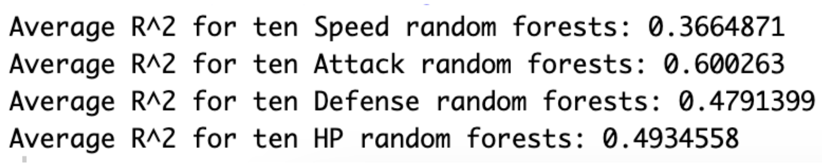
Next, we will take a look at how accurate these random forests are in predicting our four response variables. After running each of the four random forests for ten iterations, below are the mean RMSEs of all four forests:



```
Average RMSE for ten Speed random forests: 24.04554
Average RMSE for ten Attack random forests: 20.69221
Average RMSE for ten Defense random forests: 22.04106
Average RMSE for ten HP random forests: 19.1958
```

Across all four averages, there were improvements from the four averages calculated by the decision trees. Despite this improvement, the RMSEs of all four random forest models still remain pretty high. With all four averages ranging from 19.196 to 24.046, the reason for such high RMSEs could be a result from the data simply being too spontaneous given that each random forest uses 500 trees as samples.

Finally, we also calculated the R squared value of each of the four models to determine exactly how much the predictor variables were responsible for predicting each of the response variables. Below are the four R squared values followed by their interpretations:



```
Average R^2 for ten Speed random forests: 0.3664871
Average R^2 for ten Attack random forests: 0.600263
Average R^2 for ten Defense random forests: 0.4791399
Average R^2 for ten HP random forests: 0.4934558
```

As for the R squared averages, all four values had big improvements over their decision tree counterparts. The range for the random forest averages were between 36.65% and 60.02%. Similar to before, the lowest average once again belongs to Speed, while the highest also belongs to Attack. Although the R squared values grew significantly, they are still relatively low indicating that there are other factors, perhaps outside of the dataset, that greatly contribute to a Pokemon's battle performance. Despite the lower-than-wanted values, they provide us with some decent insight into the existing factors that contribute to a Pokemon's battle performance.

Conclusion

This project aimed to predict key attributes to a Pokemon's battle performance–Speed, Attack, Defense, and HP–using two machine learning models: decision trees and random forests. Through a detailed analysis process, we evaluated each model's predictive performance and interpretability. After comparing the results, it can be concluded that random forests were the better models for predicting a Pokemon's battle performance. The random forest models demonstrated: greater robustness to noise and overfitting, and improved generalization to testing data as evidenced by its respective lower RMSEs.

While decision trees are straightforward and interpretable, they exhibit higher prediction errors across all response variables. Even with pruning helping slightly to improve overfitting, the models still struggled to generalize well to the testing data. As a result, random forests consistently outperformed the decision trees in terms of predictive accuracy.

To conclude, despite decision trees providing an interpretable foundation for analyzing the contents that make up a Pokemon's battle performance, random forests are a better choice for predictive tasks on this dataset. Their ability to handle high-dimensional data and complex interactions between predictors makes them a more reliable and effective model for predicting our beloved Pokemon's battle performance attributes.

References

Choudhari, Piyush Shailesh - Pokemon Base Stats Dataset. Available at https://www.kaggle.com/datasets/crinklybrain2003/pokmon-base-stats-dataset

RStudio Libraries: tree, randomForest, corrplot, ggplot2