Pokemon Fight Winner Predictor

Name:	**V.K. Susheel**
Registration No./Roll No.:	20297
Institute/University Name:	IISER Bhopal
Program/Stream:	EECS
Problem Release date:	January 29, 2023
Date of Submission:	April 16, 2023

1 Introduction

The project's objective is to predict the winner of a battle between two Pokemon based on their characteristics. The Pokemon data gives each of the 800 Pokemon a numeric id, their First and Second Types, and their HP, Attack, Defense, Special Attack, Special Defense and Speed stats. The data also indicates the Generation to which the Pokemon belongs and if the Pokemon is Legendary. The combat data consists of three columns. The first two columns contain the id of the combatants, and the third gives the id of the winner. The Pokemon in the first column of the data set attacks first. There are 50000 battles in the training data and 10000 in the test data.

For each battle, new features are created, namely, HP Diff, Attack Diff, Defense Diff, SpAtk Diff, SpDef Diff and Speed Diff, which give the difference in stats between each pair of combatants. In order to deal with cases where these differences are negligible or even non-existent, features that determine the type advantages and disadvantages of the combatants are also created. Under the assumption that Pokemon will always choose the strongest move, i.e., the most effective one, the maximum of the advantages or minimum of the disadvantages of both types are considered for each Pokemon. Further, apart from the difference between each stat of each combatant, we also consider the maximum attack ratio of each Pokemon, i.e., the maximum between the ratio of Attack and Special attack of one Pokemon to the Defense and Special Defense, respectively, of the other. After this feature engineering is completed, each data point consists of 9 features, six individual stat differences, the maximum attack ratio of each Pokemon and the maximum type advantage of each Pokemon against its opponent's types. The class labels are 0 if the first Pokemon wins and 1 if the second Pokemon wins.

2 Methods

In order to most accurately determine the winner of a particular fight, the following models are used: Random Forest, Support Vector Machine(SVM), K-Nearest Neighbours(KNN), Logistic Regression, Decision Tree and Gaussian Naive Bayes. Since we are using existing models, we perform an exhaustive Grid Search to determine the most appropriate parameters for each model. We consider 10-fold cross-validation with f1 score as the scoring parameter for the Grid Search. In order to use the Logistic Regression model, we scale the data using the inbuilt Standard Scaler function, which sets the variance of each data point to unity. The parameters and accuracy obtained after implementing multiple Grid Searches with various more finely tuned parameters are:

Table 1: Performance Of Different Classifiers With Corresponding Parameters

Classifier	F1 Score	Parameters
LR	0.909	C = 30, multi_class = ovr, penalty = l2
GNB	0.862	var_smoothing = 0.000000000001
RF	0.95	criterion = gini, max_depth = 10, max_features = log2, n_estimators = 500
DT	0.945	criterion = gini, max_depth = 8, max_features = None
SVM	0.566	C = 30, kernel = linear, max_iter = 500
KNN	0.913	metric = manhattan, n_neighbors = 15

3 Evaluation Criteria

The data was split into train and test sets in the ratio of 4:1, creating a train set with 40,000 data points and a test set with 10,000. Here stratification is not necessary since the classes are balanced. As mentioned previously, we use Grid Search Cross Validation bypassing the train set along with various model parameters. The f1 score, train and test accuracy were the primary factors used to determine the best model. The best model was run on the test data, and a CSV file containing the winner of each battle was created. From the first table, we observe that the Random Forest and Decision Tree have the highest accuracy on the test data of 0.95 and 0.94, respectively. Based on these models, we analyze the most important features and interpret their results in the following sections.

Table 2: Performance Of Different Classifiers Using All Terms

Classifier	Precision	Recall	F-measure
LR	0.89	0.89	0.89
GNB	0.80	0.80	0.80
RF	0.95	0.95	0.95
DT	0.94	0.94	0.94
SVM	0.57	0.56	0.50

Table 3: Confusion Matrices of RF and DT Respectively Using All Terms

	Predicted Class	
	First Pokemon Wins	Second Pokemon Wins
First Pokemon Wins	4553	200
Second Pokemon Wins	299	4948

RF

	Predicted Class	
	First Pokemon Wins	Second Pokemon Wins
First Pokemon Wins	4534	219
Second Pokemon Wins	354	4893

DT

4 Analysis of Results

From the precision and recall given in Table 3, we observe that the number of data points wrongly classified as 1 or the second Pokemon winning the battle is relatively higher than the number misclassified as 0, relative to the number of correctly classified points. This could be attributed 2to the misinterpretation of the information given in the problem statement that the first Pokemon attacks first. Since the determinant of which Pokemon attacks first is the difference in their speeds, which we will later see, is the most significant determinant of the winner of the battle, it would be unlikely that the first Pokemon should have the advantage in all battles. However, the difference in the False Zero to True Zero and False One to True One is so negligible that it may simply be within the variance expected by the model's performance. In order to predict the winners of the battles in the test set, we use the Random Forest model due to its slightly improved accuracy over the Decision Tree model. We also determine the most important features for both the Decision Tree model in Figure 1 and the Random Forest model in Figure 2, which indicate the importance of Speed Difference in determining the winner of each battle.

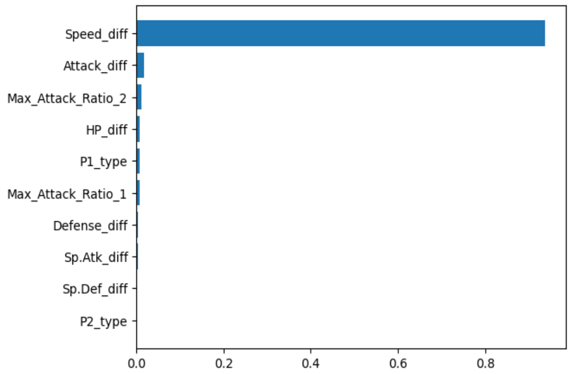

Figure 1: Importance of Each Feature in Decision Tree Model

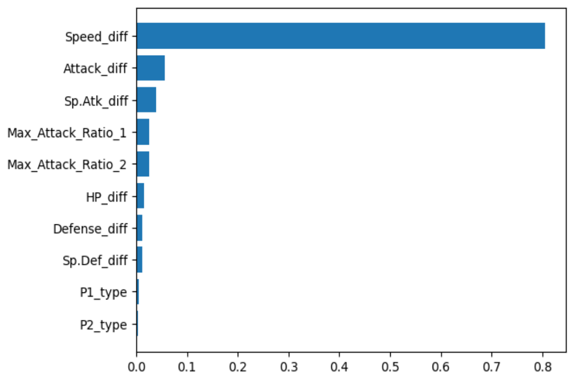

Figure 2: Importance of Each Feature in Random Forest Model

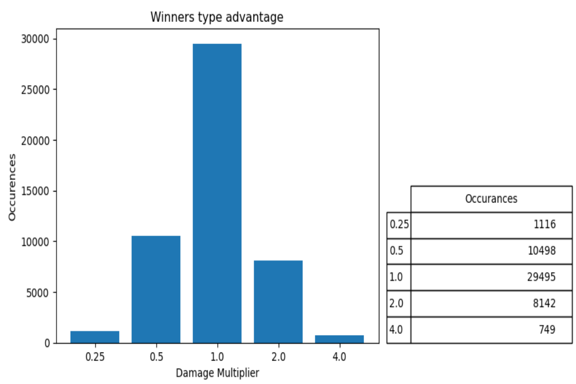

Figure 3: Type Advantage of Winner of Each Battle

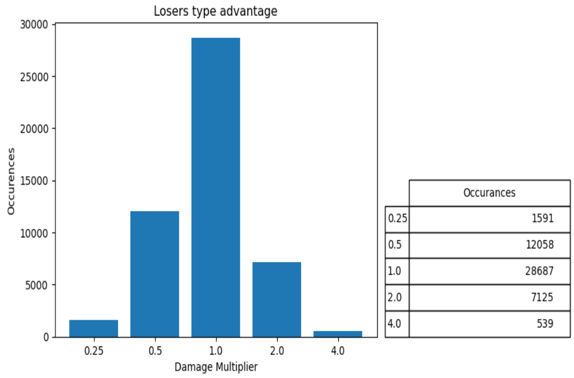

Figure 4: Type Advantage of Loser of Each Battle

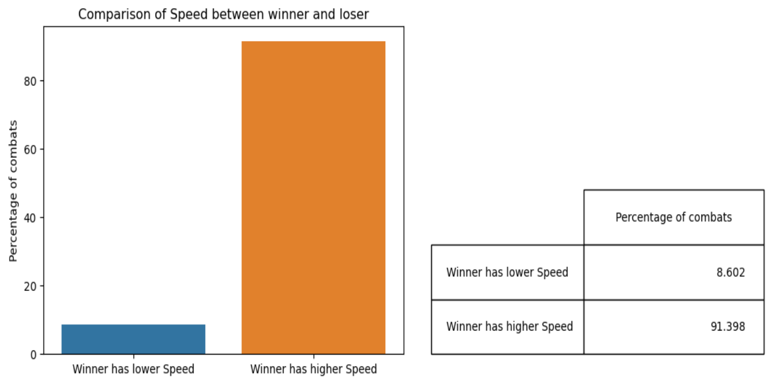

Figure 5: Percentage of Battles Where Faster Pokemon Wins

5 Discussions and Conclusion

From the results, it can be inferred that the difference in Speed between the combatants is the most important feature in determining the winner of each fight. This is likely because the battles were conducted with each combatant choosing the most damage-dealing move at each turn, favouring the faster Pokemon in all cases except where the slower Pokemon is capable of taking and dealing more damage than its opponent. It is also apparent that the type effectiveness has a minimal effect over the large sample of battles. This is because most battles took place between types that have no additional effectiveness against one another, giving more impact to the statistical difference, as evidenced by Figure 3 and Figure 4 above. However, the feature is retained to handle the cases where statistical differences are minimal, but Type Advantages are significant. In particular, keeping the feature allows us to use the model effectively in samples containing Pokemon with specific types that have some relation with one another. The effectiveness of the difference of each stat also correlates with the graphs obtained from the labelled data given below. In particular, the difference in Speed is by far the most consistent indicator of victory, with 91 per cent of the winners having greater Speed than the losers, as shown in Figure 5.

This project can form a basis for a more complicated battle predictor with more information about each combatant, such as the levels, abilities, held items, etc. Further, it can be generalized to predict the winner of a battle in any strategy game where combatants have individual stats and types. Lastly, given the characteristics of its team and opponent, it can motivate the construction of a Pokemon AI that aims to win Pokemon battles.