Competitive Pokemon Usage Tier Classification

Devin Navas
Fordham University
dnavas1@fordham.edu

Dylan Donohue
Fordham University
ddonohue7@fordham.edu

***Abstract*—This paper investigates competitive Pokémon usage tier classification given a Pokémon's stats and typing. Pokémon were classified into the usage tiers defined by the competitive battling website *Pokémon Showdown* based on their individual base stats, the sum of all their base stats (BST), and their number of type weaknesses and type resists. Classifications were done using Weka's J48 Decision tree, Weka's Lazy IBk nearest neighbor, and Weka's Logistic Regression algorithm. There were four different sets of tests run on these algorithms: one using cross validation without undersampling, one using cross validation with undersampling, one using a test set without undersampling, and one using a test set with undersampling. The algorithms were evaluated by the metrics of accuracy and precision. Lazy IBk had the highest accuracy and precision out of all the algorithms, and the highest individual accuracies for each algorithm were in the runs using the test set with no undersampling, though we suspect this may have been affected by overfitting.**

I. Introduction

The gameplay of the Pokémon video game series focuses on the raising and battling of fictional creatures called Pokémon. Pokémon battles follow a simple format—two teams of Pokémon take turns attacking each other until all of one team's Pokémon have been knocked out. All Pokémon have a typing that determines what Pokémon types they are weak to and strong against, as well as unique numerical stats that help determine the level of damage they can give to and take from other Pokémon. *Pokémon Showdown* is a website used by competitive Pokémon players to build and battle with teams of Pokémon. Since it is much easier to build competitive Pokémon on *Pokémon Showdown* than it is in the Pokémon video games, many competitive Pokémon players use *Pokémon Showdown* to build and test teams that they plan to use in official Pokémon battle tournaments, whereas others use it as a method for competitive battling on its own. Pokémon on *Pokémon Showdown* are ranked into usage tiers based on how commonly they are used on players' teams. A Pokémon's competitive viability takes into account many factors, most important being their stats, and their typing. It is typically the case that the Pokémon that are most commonly used in competitive battles have either high stats, good typing, or both.

Every competitive season for *Pokémon Showdown* determines the Pokémon available for use on teams based on what Pokémon are available in the most recently released Pokémon game. For our project, we used Weka to classify all the Pokémon that are currently available for use in competitive battle into the usage tiers from *Pokémon Showdown*, listed here from most to least popular: Ubers, Overused (OU), Underused Borderline (UUBL), Underused (UU), Rarely Used Borderline (RUBL), Rarely Used (RU), Never Used Borderline (NUBL), Never Used (NU), and Partially Used (PU). The features evaluated for each Pokémon were each of their individual base stats, the sum of all their base stats— also known as their base stat total (BST)— how many type weaknesses the Pokémon had, and how many type resistances the Pokémon had.

In our search for related work, we were unable to find any papers covering anything about our specific topic, but we did find an article summarizing an experiment where random forest was used to try and identify whether a Pokémon was a legendary based on its stats and type (Cardorelle, 2019). Legendary Pokémon are not categorized by any metrics, rather a Pokémon is just stated to be legendary according to the storyline of the game they appear in. However, many legendary Pokémon have very high stats and particular typings, so you can try to identify if a Pokémon is legendary by examining these features. Cardorelle's experiment is related to our project in this way, since he also used stats and typing as features in his data set. Interestingly, our results tended to classify legendary Pokémon into the highest usage tier— Ubers— which is likely a result of many of them having high stats.

II. Background

A. Pokémon Typing

The most basic component to Pokémon battle strategies have to do with what "type" the Pokémon is, and consequently, that Pokémon's type weaknesses and type resists. A Pokémon can have up to two types, and can have any combination of types. A Pokémon's attacks also have a typing that corresponds to a single type— the move Water Gun is a water type move, Flamethrower is a fire type move, and so on. Type weaknesses are when a Pokémon type takes high damage from the moves of another type. For example, grass type Pokémon are weak to fire type moves, fire type Pokémon are weak to water type moves, and water type Pokémon are weak to grass type moves. The same goes for a Pokémon's type resists, which are the moves of a certain type that are not very effective against that Pokémon's type. One example of this is that a water type move would do very little damage against a grass type Pokémon. Therefore, continuing with this example, if a water type Pokémon and a grass type Pokémon faced each other in battle, the grass type Pokémon would have a natural advantage, since it has grass type moves that are strong against water Pokémon, and is resistant to water type moves. This is an abridged explanation, since there are many other factors that would come into play regarding what types of moves a Pokémon could learn, but for simplicity's sake, we considered the typing of a Pokémon to encapsulate all of this information. Overall, if a Pokémon has many type resists, and few type weaknesses, this would be considered an ideal scenario.

B. Base Stats

The next most basic component to interpreting a Pokémon's viability in competitive play is the Pokémon's base stats. Base stats are numerical values that range from 1 to 255. Every Pokémon has six stats: HP, Attack, Defense, Special Attack, Special Defense, and Speed. The actual stats of a Pokémon in the games would be much higher than their base stats—oversimplifying the definition a bit, base stats are the essence of that Pokémon's potential for a specific stat. For example, since the Pokémon Diglett has a base HP stat of 10, it can be inferred that that Pokémon's in-game HP stat would be extremely low. Conversely, if a Pokémon's base Defense stat was rather high, something like 150, that would indicate that their in-game Defense would be high. Each stat contributes something different to a Pokémon's power, and while it is generally best to have the highest stats possible, any single base stat having a value greater than 100 is considered to be good in the competitive sphere. A Pokémon's base stat total (BST) is the sum of each of its base stats, and as is the case with base stats, generally the higher a Pokémon's BST, the better it is in competitive play.

III. Experiment Methodology

A. Data Sets

Since the number of type resists and type advantages, and the value of base stats are the most easily quantifiable values relating to a Pokémon's competitive viability and therefore their usage on *Pokémon Showdown*, these are the features we decided to use when compiling our data set. We were unable to find any readily available sets online with these specific features for the most up-to-date competitive Pokémon, so we compiled the data ourselves. Our data set consists of 297 Pokémon, each of their base stats, their BST, their number of type resists, and their number of type weaknesses. We gathered this data using information from *Smogon*, an online forum that helps run *Pokémon Showdown*, and has statistics for every Pokémon currently usable on *Pokémon Showdown*. The data was compiled in a spreadsheet form on Google Sheets, and was then converted into a .arff file using the Weka Arff Viewer. In total, there were 17 Pokémon in Ubers, 39 Pokémon in OU, 9 Pokémon in UUBL, 59 Pokémon in UU, 5 Pokémon in RUBL, 44 Pokémon in RU, 4 Pokémon in NUBL, 43 Pokémon in NU, and 77 Pokémon in PU, making PU the majority class for this data set.

B. Algorithms

The three algorithms we used in this experiment are Weka's J48 Decision tree, Weka's Lazy IBk nearest neighbor, and Weka's Logistic Regression algorithm. At first we were only going to run tests using decision trees and nearest neighbor, but since these algorithms can suffer from data fragmentation, we decided to use logistic regression as well. Logistic regression was also advantageous to use because it provided us with numerical weights for our features and thus indicated which features were considered the most important when classifying Pokémon into the usage tiers. Seeing which features had the highest weights would also allow us to try and determine what was the main cause of misclassification errors for this algorithm.

C. Setup Methodology

We used several combinations of methods to prepare our data to be run through each of the algorithms. Overall, we did four different sets of runs with each of the algorithms, for a total of 12 runs overall. For two sets of runs, we used 10-fold cross validation. The other two sets were run using a test set we created ourselves by taking 40% of Pokémon from each tier, and thus 40% of the entire data set, to be used in a test set. One set of runs from each of the cross validation and test set groups were, additionally, put through some preprocessing. Since the majority class, the PU class, had a far greater amount of Pokémon in it than many of the other classes, we decided to use a Weka preprocessing filter called "Resample" that produced a random subsample of the data set using sampling without replacement. This reduced the size of some of the classes to make them all closer in size; any class with a size over 33 went down to 33 and every other class stayed the same. This allowed us to undersample the majority class and gain completely different results in two of our sets of runs.

D. Evaluation Metrics

After running each of the different algorithms on the data sets, we looked at the confusion matrix, and other measures such as accuracy and precision for each of the runs to evaluate our results. The confusion matrix allowed us to see where misclassification errors occurred, and what the nature of these outliers were. Though we did aim to get the highest precision and accuracy possible, we were also interested in seeing what Pokémon were commonly misclassified into the wrong usage tier. If a Pokémon was often classified as a lower tier than its actual tier, it could imply that Pokémon has a value on competitive teams that is not discernible by looking only at its stats and typing. Conversely, if a Pokémon was usually put into a higher tier than its actual tier, it could imply that it might be used less for very specific reasons. An example of this would be if a Pokémon had good typing and stats, but just happened to be weak to many of the Pokémon in the higher usage tiers. Analyzing these outliers and trying to interpret why a Pokémon may have been classified in a certain way was one of the most engaging parts of our evaluation process.

IV. Results

The runs using the test set gave the overall highest accuracy results (Table I). The runs using the test set with no undersampling had the highest accuracies for J48 and Lazy IBk, though we believe this may have been the result of overfitting, as evidenced by the Lazy IBk run for this set having an accuracy of 100.0%. Lazy IBk had the highest accuracy for the runs using the test set, whereas Logistic Regression had the highest accuracy in the runs using cross validation. Both J48 and Lazy IBk improved greatly with the use of the test set, whereas Logistic Regression only had a slight improvement. Logistic Regression stayed within the accuracy range of 29.50%-47.42% in each of its runs. We looked at what feature the Logistic Regression algorithm was giving the highest weight to try and figure out what might have been causing these numbers, and learned that it was weighing the type resist feature the most heavily with a weight .3, whereas the second highest weight was BST with a weight of .1. Having a high number of type resists

is good for a Pokémon to have, but having high stats is certainly more important in the consideration of whether a Pokémon is competitively viable or not, so having the classification consider the wrong feature the most important is likely what caused the accuracy to remain low for Logistic Regression even with the use of the test set.

TABLE I. ACCURACY OF CLASSIFICATIONS(%)

Methods	**Algorithm**		
	J48	***Lazy IBk***	***Logistic Regression***
Cross Validation without Undersampling	32.32	29.60	32.65
Cross Validation with Undersampling	26.00	27.00	29.50
Test Set without Undersampling	78.85	100.0	47.40
Test Set with Undersampling	64.00	78.28	47.42

For the runs using cross validation without undersampling, the accuracy for each of the algorithms was a bit low, with the highest accuracy being Logistic Regression with an accuracy of 32.65% (Table I). For each of the algorithms, the most precise predictions made were in Ubers and PU (Table II). This makes sense, because Pokémon in Ubers are generally Pokémon with the highest BSTs and would be easy to identify as strong, although there were some cases where Ubers Pokémon were placed in lower tiers despite their high stats. These types of misclassifications were the most notable since several of the higher tier Pokémon from Ubers and OU were classified into the lower tiers of PU or NU. One reason PU may have had such a high precision is that PU Pokémon often have very low stats, so any Pokémon with lower stats could have been correctly assumed to be in PU. In addition, since PU is the majority class it is expected that many Pokémon might be misclassified into it.

TABLE II. PRECISION OF CLASSIFICATIONS- CROSS VALIDATION WITHOUT UNDERSAMPLING

Usage Tier	**Algorithm**		
	J48	***Lazy IBk***	***Logistic Regression***
Ubers	.737	.867	.750
OU	.333	.205	.237
UUBL	0.00	0.00	0.00
UU	.196	.288	.304
RUBL	0.00	0.00	0.00
RU	.250	.136	.188
NUBL	0.00	0.00	0.00
NU	.116	.116	.148
PU	.506	.506	.378

Some specific Pokémon were misclassified in the same way by all of the algorithms in this set of runs. The Pokémon Ditto and Mimikyu were classified as PU and NU, respectively, by all of the algorithms, but both of these Pokémon belong to the OU tier. In addition, the Ubers tier Pokémon Darmanitan-Galar was put into the PU tier by each algorithm, a drastic discrepancy in classification. The borderline tiers, UUBL, RUBL, and NUBL, were never predicted correctly for this set, but there are so few of these Pokémon, only 18 in total, that we expected these tiers to have low precision for most of the runs.

The runs using cross validation with undersampling had lower accuracies for each of the algorithms than the runs without undersampling, with the highest accuracy being Logistic Regression again, with an accuracy of 29.50% (Table I). Although the overall accuracy was lower than the runs without undersampling, certain classes were predicted with higher precision. Ubers and PU were classified with the highest precision again, but some of the borderline tiers were actually predicted correctly this time, with J48 predicting NUBL with a precision of .250 and Nearest Neighbor predicting UUBL with a precision of .429 (Table III).

TABLE III. PRECISION OF CLASSIFICATIONS- CROSS VALIDATION WITH UNDERSAMPLING

Usage Tier	**Algorithm**		
	J48	***Lazy IBk***	***Logistic Regression***
Ubers	.743	.867	.824
OU	.333	.300	.200
UUBL	0.00	.429	0.00
UU	.182	.237	.281
RUBL	0.00	0.00	0.00
RU	.115	.103	.233
NUBL	.250	0.00	0.00
NU	.194	.194	.233
PU	.289	.323	.300

Many of the outliers in this run were similar to the runs without undersampling, though Darmanitan-Galar was placed into the slightly higher NU tier rather than the PU tier. PU was actually predicted with the lowest precision than it had in any of the other sets of runs, likely due to the undersampling. This is also likely what allowed for other classes, like Ubers and the borderline tiers, to be predicted correctly more often.

In the runs using the test set without undersampling, the accuracy for each of the algorithms was significantly higher than the cross validation tests. Lazy IBk had an accuracy of 100.0%

for this run, which we assumed was a result of overfitting the training set, so we decided to consider this result erroneous (Table I). Taking this into consideration, the algorithm with the highest valid accuracy for this set was then J48 with an accuracy of 78.85%. Even without undersampling, the use of the test set allowed for a more accurate classification of borderline tiers, with UUBL and NUBL both having a precision of .500 with J48, and an overall higher precision for all of the classes than the runs that used cross validation (Table IV).

TABLE IV. PRECISION OF CLASSIFICATIONS- TEST SET WITHOUT UNDERSAMPLING

Usage Tier	Algorithm		
	J48	***Lazy IBk***	***Logistic Regression***
Ubers	.909	1.00	.900
OU	.731	1.00	.478
UUBL	.500	1.00	0.00
UU	.763	1.00	.387
RUBL	0.00	1.00	0.00
RU	.941	1.00	.438
NUBL	.500	1.00	0.00
NU	.840	1.00	.375
PU	.796	1.00	.475

Obviously, due to the higher accuracy of these runs, there were less misclassification errors, but some of the same Pokémon that were misclassified in the runs that used cross validation were misclassified in these run as well, such as Ditto, Mimikyu, and Darmanitan-Galar. Interestingly, certain Pokémon that are considered competitively viable thanks to having one or two particularly high stats, but have comparatively lower values for all their other stats, were classified correctly in these runs, whereas they were misclassified in the runs using cross validation. One example of this is the Pokémon Gengar, who was put into the PU tier by all of the cross validation runs, but was correctly classified as OU in all of these runs.

The runs using the test set with undersampling still had significantly higher accuracy for each of the algorithms than the cross validation tests, but generally not as high as the tests without undersampling. The highest accuracy was Lazy IBk with an accuracy of 78.28%. Logistic Regression did have a 0.02% higher accuracy than the run without undersampling, but this difference is a very minute improvement (Table I). The Lazy IBk run in this set was the only run out of all the ones we did (excluding the Lazy IBk run with 100.0% accuracy) that had a precision higher than 0.00 for all of the borderline classes (Table V). Like the 100.0% accuracy run though, this could also be a result of overfitting.

TABLE V. PRECISION OF CLASSIFICATIONS- TEST SET WITH UNDERSAMPLING

Usage Tier	Algorithm		
	J48	***Lazy IBk***	***Logistic Regression***
Ubers	.909	1.00	.909
OU	.514	.778	.419
UUBL	.667	.833	0.00
UU	.647	.758	.348
RUBL	0.00	.600	0.00
RU	.606	.690	.417
NUBL	0.00	.667	0.00
NU	.533	.792	.367
PU	.875	.842	.585

Similar Pokémon were misclassified in these runs to the test set runs without undersampling, most notable again being Ditto, Mimikyu, and Darmanitan-Galar.

A. Interpreting Misclassifications

Since certain Ubers and OU Pokémon were misclassified into the lower tiers so consistently, we analyzed these cases to try and figure out if there were any common characteristics between these misclassified higher-tier Pokémon. The cases we found the most interesting were the Pokémon Ditto, Mimikyu, and Darmanitan-Galar, all of whom are high-tier Pokémon that are known to be very popular on competitive teams, but were never classified correctly throughout all of the sets of tests. Darmanitan-Galar is in Ubers, while Ditto and Mimikyu are in OU, but it makes sense for all of them to have been frequently misclassified because the strength of all of these Pokémon cannot be captured purely by their stats or typing. Each of these Pokémon have a unique skill called their "ability", that gives these Pokémon certain tactical advantages once in battle. Every Pokémon has an ability, though each Pokémon only has a certain pool of abilities that they can choose from. Most Pokémon share the same abilities amongst each other, but certain Pokémon have abilities unique to only that Pokémon. This is the case for Ditto, Mimikyu, and Darmanitan-Galar, all of whom have an ability that no other Pokémon has, with all of these abilities also being extremely useful in battle.

a) Ditto: In Ditto's case, all of its stats have a value of 48, giving it a very low BST of 288, and while it only has one type weakness, it also has only one type resist. As such, it would make sense for Ditto to be classified into a lower tier based purely upon its stats and typing. However, Ditto's unique ability Imposter allows it to transform itself into the opponent's Pokémon once it enters a battle. This then grants Ditto the stats and typing of the opposing Pokémon, essentially giving Ditto the potential to match the strength level of any opponent. This versatility allows Ditto to be extremely flexible on any competitive team, and is thus why Ditto is in the OU tier.

b) Mimikyu: Mimikyu's stats are generally average, with three of it's stats being near or just above 100 (Attack: 90, Sp.Def: 105, Speed:96), and a BST of 476. It has 4 type resists and only 2 type weaknesses, which is a pretty good ratio, but overall Mimikyu's typing and stats are pretty average in comparison to most competitively viable Pokémon. What sets Mimikyu apart is its ability, Disguise, which protects it from taking damage from the first attack it receives from an enemy Pokémon. Getting one turn to take a free hit without any damage gives Mimikyu a huge advantage, since it can utilize this damage-free turn to use moves that increase its stats, or some other set-up based move that will give it the advantage over its opponent.

c) Darmanitan-Galar: Darmanitan-Galar's stats are generally better than those of both Mimikyu and Ditto, but its BST of 480 is not much higher than Mimikyu's, and though its Attack of 140 is impressive, its ratio of 3 type weaknesses to 1 type resist make it understandable that it might have been classified into a lower tier. However, Darmanitan-Galar's ability Gorilla Tactics helps give it a major advantage in battle by multiplying its Attack by 1.5, which is a huge boost to this already high stat. Darmanitan-Galar is often given certain stat-boosting items in battle that increase its stats even higher in addition to the boost it gets from its ability. Pokémon can each hold one item while in battle, and these items are generally used to give Pokémon certain stat boosts or status effects. The items commonly given to Darmanitan-Galar in battle, the Choice Scarf and Choice Band, are both items that increase stats; Choice Scarf multiplies its holder's Speed by 1.5 while Choice Band multiplies its holder's Attack by 1.5. As a result, a Darmanitan-Galar could enter a battle with its stats already ridiculously boosted, having either its Attack and Speed both multiplied by 1.5 while holding the Choice Scarf, or, outrageously, its Attack multiplied by 3 while holding the Choice Band. These factors are what cause Darmanitan-Galar to be in the Ubers tier, since it can, given the right circumstances, sweep the entire enemy team by itself with the help of these stat boosts.

Other notable classification errors were those of Pokémon that do have objectively high stats and good typing, but do not fit into the current competitive metagame of its respective tier. An example of this is the Pokémon Escavelier, which has 9 type resists and only one type weakness, and three stats over 100. It would make sense that a Pokémon like this might belong to the OU tier, where Escavelier was commonly misclassified into. However, Escavelier's one type weakness is that it is 4x weak to fire type attacks, meaning any fire type move used against it is multiplied by 4. This makes Excavalier much less viable despite its good stats and high number of type resists, since many commonly used Pokémon in the current metagame can easily defeat it as a result of this weakness, which is why it is currently placed in the RU tier.

V. Conclusion

Overall, we received our best results in the runs using the test sets rather than cross validation. The highest accuracies for J48 and Lazy IBk were achieved with the runs using the test set with no undersampling, though we suspect some overfitting occurred in these runs, as evidenced by the 100.0% accuracy in the Lazy IBk run. Logistic Regression performed the best in the run using the test set with undersampling. The Ubers and PU classes had the highest precision throughout each of the sets of runs, though for PU this was likely the case because it was the majority class. The borderline tiers of UUBL, RUBL, and NUBL consistently had the lowest precision of any of the other tiers, probably due to how small each of these tiers is, but it did seem that undersampling and using the test set helped to increase the precision for these tiers.

One of the more interesting aspects of analyzing the misclassification results was to see what kinds of Pokémon were misclassified most frequently. There was certainly a noticeable trend of Pokémon who are competitively viable thanks to the effects of their ability rather than just their stats or typing being misclassified into much lower tiers than their actual tier, as we saw with Ditto, Mimikyu, and Darmanitan-Galar. As such, it is clear that not incorporating abilities as a feature is one of the shortcomings of this project. Initially, we had decided not to use abilities as one of our features because they are not easily quantifiable, and we feared it might create excess noise as a result of us not being able to quantify the usefulness of a Pokémon's ability properly. However, after seeing how big of an impact not considering abilities had on the misclassification of many of the higher-tier Pokémon, if we were to redo this project in the future, it is clear that abilities should be taken into account as one of the features we evaluate on. This could possibly be done by trying to "rank" every ability by assigning it a numerical value on a small scale, like 1-5, that would try to assess how useful a Pokémon's ability is in battle. Given this example, Pokémon like Ditto, Mimikyu, and Darmanitan-Galar would all have a 5 on this scale, and perhaps this could help to increase the precision of the higher-tier classifications. Other qualitative aspects to a Pokémon's competitive viability, like the strength of what moves they can use, could also be taken into account using a scale like this if we were to expand upon this project in the future.

References

[1] Cardorelle, Sylvester. "Identifying Legendary Pokémon Using The Random Forest Algorithm." Medium, Towards Data Science, 6 June 2019, towardsdatascience.com/identifying-legendary-pok%C3%A9mon-using-the-random-forest-algorithm-ed0904d07d64. Accessed 5 Apr. 2020.