Data 311: Pokémon The Myths and The Legends

Can you catch them all?

Introduction

The Pokémon franchise began in 1996 with the release of Pokémon Red and Blue and has since spawned many video game sequels across many platforms, an anime series, several anime films, a wildly popular card game, and in 2019, a live-action film with Ryan Reynolds providing voice-work[1]. The wild popularity of Pokémon has resulted in seven generations of Pokémon creatures, numbering 809 in total with an eighth generation being released in summer 2019 with the Sword and Shield editions on the Nintendo Switch[2]. The number of unique pokemon in existence is increasing at a rapid rate and players have to adopt new training strategies as the world of gaming changes. For this reason, this study examines a dataset comprising the first six generations of Pokemon with 721 observations. Machine learning techniques were employed to determine useful trends for trainers in their journey to catch them all.

Data

The dataset examined, accessed from Kaggle, is Pokémon for Data Mining and Machine Learning[3], a set of Pokémon with related statistics and descriptors. The set contains 721 observations and 23 variables, several of which are categorical. Of the numerical predictors, the majority are the Pokémon statistics, such as Total, Attack, Defense, HP, while others relate to body descriptors such as Weight_kg and Height_m. The other numerical predictor, Pr_Male is indicative of the probability that when a Pokémon is caught it will be male. Finally, Catch_Rate is a number bounded by 3 and 255 that refers to the ease of catching a Pokémon, where the higher the catch rate, the easier it is to catch the Pokémon[4]. The statistics of Pokémon were analyzed through a variety of models attempting to predict whether or not a Pokémon is of the legendary type. In order to predict legendary status, the isLegendary and Total variables were tested against a range of predictors within the Pokémon dataset. Model accuracy varied depending on the method used. The models that were particularly useful in predicting legendary status were k-means clustering, neural nets, KNN, and random forests, though testing was not limited to these models. Gender analyses of Pokémon were another significant aspect of the investigation. Since gender appeared to be linked to the legendary status, a large focus was on whether or not there is a correlation between Pokémon statistics and gender. These analyses were run under the assumption that the higher or rather "better" the Pokémon's statistics, the greater likelihood of legendary status. The investigation analyzed Pr_Male as it relates to Pokémon statistics, with the goal of determining if a correlation exists between statistics such as Defense and Attack and whether a Pokémon is typed as male or female.

---

[1] "About the Pokémon Company International." The Pokémon Company International, accessed April 4, 2019. https://api.pokemon.com/us/about-pokemon/

[2] "Generation." Bulbapedia, accessed April 4, 2019. https://bulbapedia.bulbagarden.net/wiki/Generation

[3] Alopez247. "Pokémon for Data Mining and Machine Learning." March 05, 2017. Accessed April 4, 2019. https://www.kaggle.com/alopez247/pokemon.

[4] "Catch Rate." Bulbapedia, accessed April 4, 2019. https://bulbapedia.bulbagarden.net/wiki/Catch_rate

After loading in the dataset, preliminary analyses and plots were run to visualize the distribution of Pokémon in the dataset based on their Type_1(Figure 1). Type is a fairly important aspect of a Pokémon as Pokémon types determine special attributes that determine the strengths and weaknesses of different Pokémon species. From the visualization, it appears that water is a very popular type of Pokémon. [5]

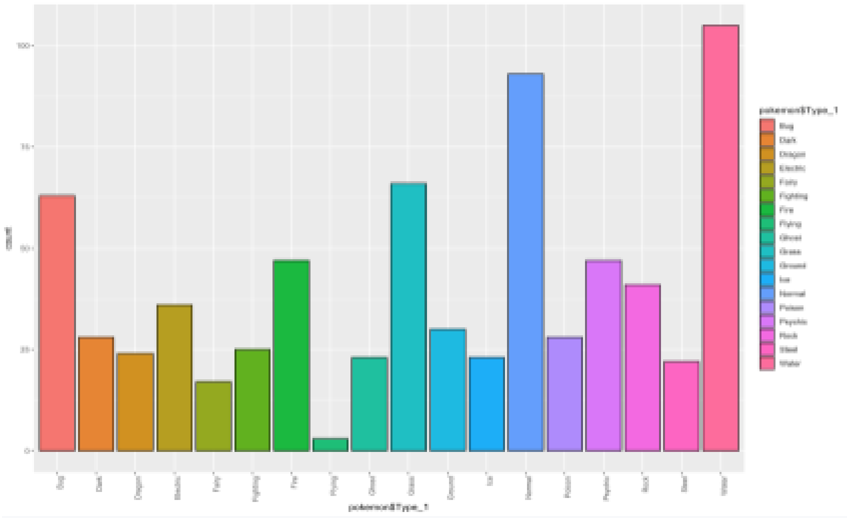


Figure 1

Initially, clustering was run in order to determine the approximate grouping of the data. Both Euclidean and Manhattan distance and complete, single, and average linkage was used. Single Linkage featured a large amount of chaining, though Manhattan distance did a slightly better job than Euclidean; on the other hand, average linkage featured less chaining but the distribution of the groups was skewed. There was one extremely large group containing the majority of the dataset and one small one containing the minority. Complete linkage featured the most reasonable grouping with Manhattan distance slightly outperforming Euclidean (See Appendix A for Graphs).

After clustering, Mclust was applied to the dataset using the predictors HP, Attack, Defense, Sp_Atk, Sp_Def, Weight_kg, and Height_m. Mclust was also run with varying numerical predictors that are not listed, as they did not prove beneficial for discussion. Model-based clustering describes the modeling technique that uses the EM algorithm in order to perform clustering using Gaussian mixture models[6]. The Bayesian Information Criterion (BIC) suggests the optimal model and number of groups given the dataset Mclust is performed on. The

Figure 2

[5] "Pokemon Go Type Chart: Every Type Strength and Weakness for Battle Counters." VG247. January 29, 2019. Accessed April 05, 2019. https://www.vg247.com/2019/01/29/pokemon-go-type-chart-strength-weakness-effectiveness-counters/.

[6] “Mclust: Model-Based Clustering.” R Package Documentation, accessed April 4, 2019. https://rdrr.io/cran/mclust/man/Mclust.html.

marginal and uncertainty plot generated projected four groups, with what appears a significant amount of uncertainty of group membership (See Figure 2).

The BIC indicated the primary model VVV and 5 or 4 groups. When the data was scaled the BIC still suggested the model VVV, though suggested 5 groups exclusively. It is reasonable to suggest that using Mclust on the Pokémon dataset may not be an effective model for the data, as the dataset contains a large proportion of variables. At least it may not be useful in such a way where group membership can be determined by Pokémon statistics such as HP, Attack, Defense, Weight_kg, or Height_m. A variety of numerical predictors were tested against Mclust and Mclust plotted on classification, however, the results did not provide interpretable relationships using these methods. Regardless of predictors chosen, adjustedRandIndex values fell above 0, but significantly less than 1. Overall, it appears that Mclust and adjustedRandIndex are not particularly useful in better understanding the Pokémon dataset, and may paint an inaccurate representation of groups that cannot be confirmed nor denied through use of the Mclust and adjustedRandIndex functions.

However, the best grouping was performed by K-Means, a clustering method that iteratively computes the clusters based upon the mean and the given K-value. Six clusters gave the best representation of the data and a reasonably low WSS (Figure 3). For instance, with six clusters, Cluster 3 contained the shortest and most lightweight Pokémon, whereas, Cluster 5 contained the lowest stats/easiest to catch Pokémon with the least likelihood of being Male. Cluster 6 contained every legendary Pokémon, had the highest statistics and the lowest catch rate. They also happen to be the heaviest and tallest on average and were the most likely to be Male. This correlation between legendary status, high statistics, and high probability of being male informed later analyses on Pr_Male to search for male bias in Pokémon.

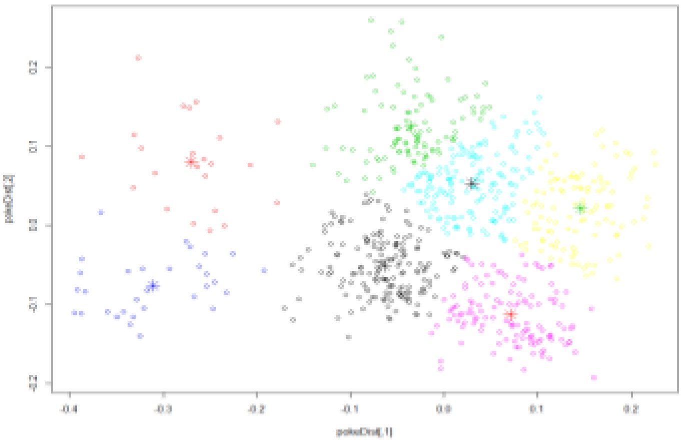


Figure 3

Methodology

isLegendary Analyses

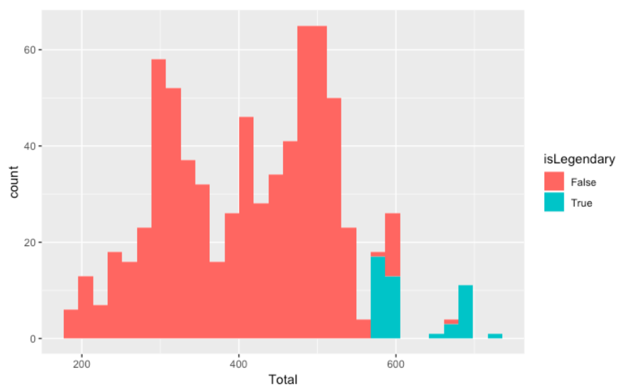


Figure 4

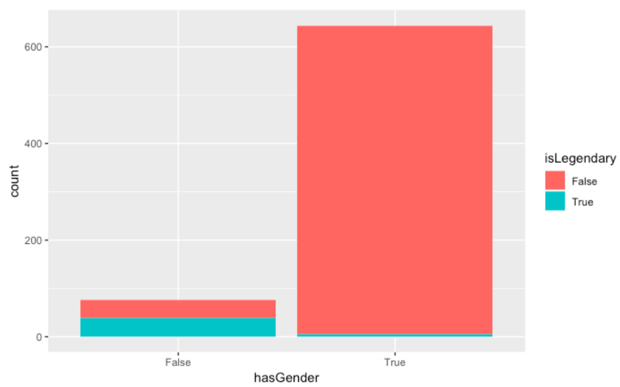


Figure 5

The predictor isLegendary was visualised against Total and hasGender, as isLegendary was one of the primary interests of analyses. The visualization suggests that the higher the total, the more likely a Pokémon is to be legendary (Figure 4)[7]. In fact, most Pokémon with totals between 550-625 are legendary and all Pokémon with totals above 650 are legendary. It also appears from the data set that most legendary Pokémon do not have a gender except for a couple outliers. From the visualization of legendary type Pokémon organized by gender, it is clear that the majority of legendary Pokémon (isLegendary = TRUE), do not have a gender (hasGender = FALSE). Thus, it is evident that Total and hasGender may be indicative of the legendary status of a Pokémon.

A linear model was fit to the Pokémon dataset in order to determine if there are any linear relationships present between the data. Name and Number were excluded from examination, as the two variables would not be useful predictors. A linear model was fit to Total and statistics HP, Defense, and Attack. The following is known as multiple linear regression. Multiple linear regression is a model designed to fit the relationship between a response variable, in this case, Total, and multiple explanatory/predictor variables, HP, Attack, and Defense. The result of multiple linear regression suggests a fairly linear model (See Figure 6) .

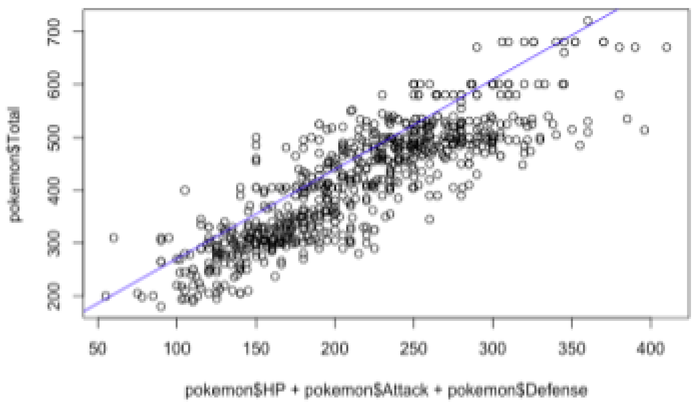


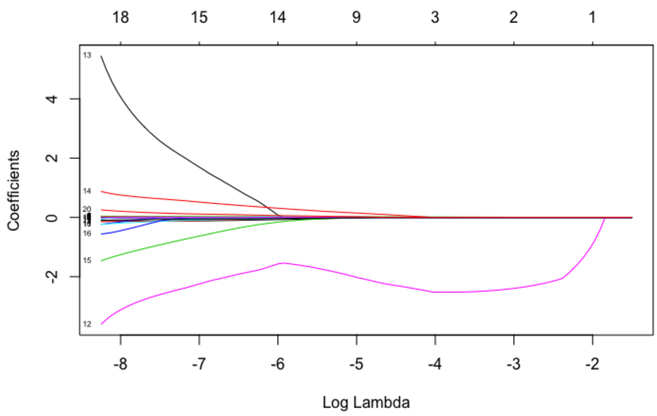
Figure 6

T-test and significance values reveal that the previously mentioned predictors are significant, thus there appears to in fact be a linear relationship between Total and these fundamental Pokémon statistics. This makes sense under assumptions of Total. As fundamental Pokémon statistics increase then the resultant total should too.

The relationship between Total and base statistics suggests the existence of useless predictors in a linear model. The relationship between isLegendary and Total and the relationship between Total and HP, Attack, and Defense led to overparameterization in a linear regression model for isLegendary. Since linear regression good for interpretation, it is still a useful model but can be victim to overparameterization. Student's T test can use standard statistical theory and hypothesis testing for variable selection but because of Type

[7] "Predicting Legendary Pokémon." Accessed April 4, 2019. https://www.kaggle.com/excaliburzero/predicting-legendary-pokemon.

1 error, there is a chance that a useless predictor will be selected as important by chance alone. The least absolute shrinkage and selection operator (lasso) model was fit to investigate variable importance, acting as a shrinkage method via the addition of a penalization term in order to force the coefficients of some variables to zero. Cross validation was used to determine the best value for the penalty term so predictors were removed without adding much error, as more predictors is always more flexible.



Lasso was performed on isLegendary using all predictors. The cross-validated true minimum had 14 useful predictors with an MSE of 0.08835933 and one standard deviation of the minimum gave 12 useful predictors. Both lasso models removed base statistics in favour of Total and the most important predictor for determining isLegendary is hasGender. This matches the rest of the analyses. HasGender increases in importance as Pr_Male is removed, which is NA whenever hasGender is false. In this way, the removal of Pr_Male is supported as the removal of a useless predictor that has information repeated in another predictor.

The removal of base statistics is supported by performing the lasso method on Total, which found that main effects are HP, Attack, Defense, Sp_Atk, Sp_Def, Speed, and Catch_Rate with no interactions. This shows a strong relationship between Total and these predictors and any model using Total should avoid using these as predictors to avoid overparameterization. Lasso with interactions on isLegendary found that the main effects are Total, Sp_Atk, Sp_Def, and Weight. Significant Interactions were (Total, Sp_Atk), (Total, Sp_Def), (Total, Catch_Rate), and (Sp_Atk, Type_1). The base statistics were removed from the model, keeping Total in place of its predictors Attack, Defense, and Speed. Adding interactions increased the cross validated MSE of the model to 0.09163895, which is less flexible but provides insight on how predictors interact when determining legendary status.

It is likely that pokemon with similar statistics are more likely to be like one another, especially in terms of legendary status. To test this connection, the k-nearest neighbours (KNN) algorithm was used to classify the observations via the probability of a pokemon being Legendary based on its k-nearest neighbours. The numerical statistics HP, Attack, Defense, Sp_Atk, Sp_Def, Speed, and Catch_Rate were used to classify the binary response for isLegendary, all numerical predictors. Only 7 of the 22 possible predictors were used in the analysis firstly because some of the predictors have missing values for a large amount of the observations (e.g. Pr_Male is only applicable to observations with a gender). The predictors that were chosen are more indicative of the performance of the Pokémon, likely serving as useful for predicting Legendary status. Also, Height_m and Weight_kg were initially used as predictors but they proved to be detrimental to the model. A randomized training set containing 432 observations and testing data set containing the other 289 observations were generated. The KNN algorithm was applied on the training and testing data sets, resulting in a classification performance of 98% accuracy on the testing data set and an F1 score of 0.989. Cross-validation on the entire dataset using the KNN model resulted in a classification performance of 98% and an F1 score of 0.993. This model shows a clear relationship between the numerical statistics and legendary status in Pokémon. A classification performance this high seems unlikely without the model overfitting; however, the cross validated values and testing set are indicative of the true population. This shows that the KNN is not overfit nor as variable. This is reflected in the predictors used, as Legendary Pokémon are the most powerful and difficult to find Pokémon, so a model that classifies the Pokémon with the best stats and lowest catch rate as legendary would seem to be doing its job, with the low misclassification rates backing up that claim. Therefore, KNN solidifies the relationship between the numerical statistics of Pokémon and Legendary status.

Figure 8.1

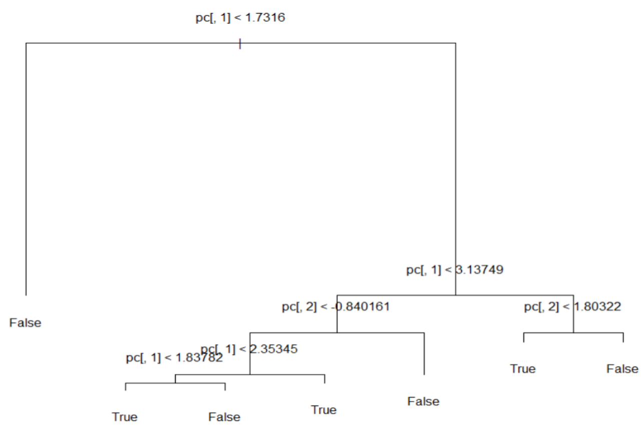
KNN Classification on Testing Set		
	False	True
False	269	3
True	3	14

Figure 8.2

Cross-validated KNN Classification		
	False	True
False	669	3
True	6	43

A principal components analysis was performed to attempt to simplify HP, Attack, Defense, Sp_Atk, Sp_Def, Speed, Height_m, Weight_kg, Catch_Rate. Pr_Male was not used since it would exclude the Pokémon with no gender. When performing an analysis on a dataset with many, potentially correlated, predictors, PCA is a method that is used to combine these variables into a smaller set of variables that still explain a fair amount of the variation in the dataset. The first two principal components satisfied the Kaiser criterion, and although, cumulatively, they only explain 59% of the variation in the data, they were the two components kept. The Kaiser criterion indicates that the principal components from scaled data that have a standard deviation greater than 1 should be used for analysis.

The first principal component (PC1) had high positive loadings for most numeric stats other than speed, and a relatively large negative loading for Catch_Rate, suggesting that it may be a good predictor for Legendary status. A linear discriminant analysis was performed using PC1 as a predictor and isLegendary as the response. Linear discriminant analysis (LDA) is a classification model that uses the prior and posterior probabilities of observations falling into one or the other classes. In turn, these probabilities approximate the Bayes classifier and indicate a decision boundary for the given class. Although this model had a classification performance of 95% (i.e. predicted isLegendary correctly 95% of the time), its F1 and Recall scores were 0.51 and 0.39 respectively. Since the model predicted False for Legendary Pokémon 28 out of 46 times, this model is biased to the majority class of not legendary, and so is not very useful. Note that Recall and F1 were used



as classification performance indicators since the models created using PCA all suffered from a bias towards the majority class.

To interpret the resulting principle components, a classification tree was created using PC1 and PC2 as predictors and isLegendary as the response. Classification trees are a method of observing the multidimensional space occupied by several predictors, then splitting this space progressively such that most observations on either side of the split match. This results in a tree-shaped model which can be very useful for inference. The simple tree had a F1 and Recall of 0.85 and 0.78 respectively, indicating a much more useful classifier than LDA. The resultant tree, however, had 13 terminal nodes, which is overly complicated for a model with two predictors. Cross-validation yielded a pruned tree with 7 terminal nodes, and obtained F1 and Recall scores of 0.82 and 0.80 respectively, so far the best model for predicting isLegendary with the PCA analysis. Using bagging, the confusion matrix gave worse values for F1 and Recall, 0.68 and 0.80 respectively, although these numbers give a better idea for long-run performance because of the cross-validation used in bagging. Similarly, using random forests, F1 and Recall for the resulting confusion matrices were 0.63 and 0.76 respectively. Based on the above, PC1 and PC2 can result in a decent model for predicting isLegendary using a bagged tree model; however, the resultant model suffers significantly in interpretability. Despite being a decent classifier for isLegendary, the model was rejected for lack of interpretability and subpar performance compared to other classifiers for legendary status.

A more predictively capable model is the neural network. Neural networks process information in such a way that is often compared to the human nervous system[8]. Likewise, neural nets use language that references neurological terminology. When referencing the number of nodes a neural net contains, these can be considered analogous to neurons. While neural networks are considered a useful tool for predictive power, its limitation is a lack of interpretability. Neural networks use training data in order to assess linkages between predictors and then are ran using testing data in order to assess for model accuracy. For this analysis, the predictors HP, Attack, Defense, Sp_Atk, Sp_Def, and Speed were used to generate an isLegendary neural network. Through repetitive testing, experiments determined that 5 nodes within one hidden layer appeared to give the optimal misclassification on the testing set (See Figure 10.1 and 10.2).

Figure 10.1

Training misclassification

	1	2
False	461	5
True	0	39

Figure 10.2

Testing misclassification

	1	2
1	203	6
2	0	7

The predictive capability of the neural net on isLegendary using HP, Attack, Defense, Sp_Atk, Sp_Def, and Speed supports the assumption that the “better” the statistics, the greater likelihood the Pokémon is legendary. The fact that these predictors were valuable in terms of predicting legendary status indicated that they would likely be useful predictors for determining the relationship between Pokémon statistics and gender. Using the context from analyses on isLegendary, analyses on Pr_Male followed similar assumptions.

[8] “Creating & Visualizing Neural Network in R.” Accessed April 4, 2019. https://www.analyticsvidhya.com/blog/2017/09/creating-visualizing-neural-network-in-r/

Pr_Male analyses

The presence of gender is a very important predictor for legendary status but the overlap of having gender and the probability of being male made Pr_Male a useless predictor in most models. If a pokemon has no gender, then Pr_Male is null as it is meaningless to say what the probability of catching a male would be. The effect of gender on other statistics and how gender is expressed within a pokemon is an important topic when discussing gamer demographics and how these have changed over time.

Shown below is the distribution of the Pr_Male variable. 139 Pokémon had a higher probability of being male, 458 had an equal probability of being either male or female, and 47 Pokémon had a higher probability of being female. Generations 1 and 5 had the highest number of observations across all probability distributions, likely a result of these generations introducing the largest amount of new Pokémon (150 and 153 respectively).

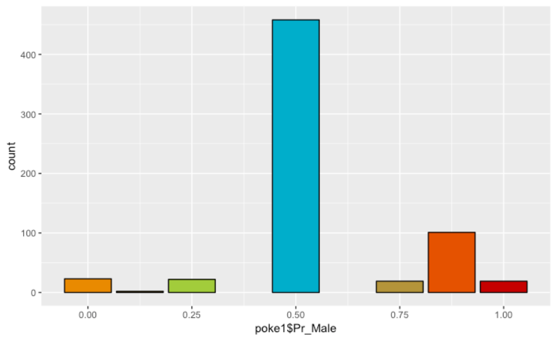


Figure 11

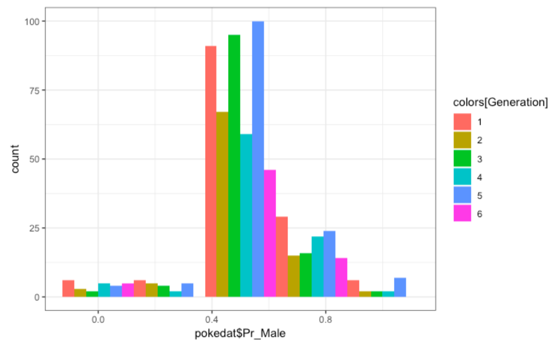


Figure 12

To test the hypothesis that the numeric statistics could have an influence on the probability that a given Pokémon is male (Pr_Male>0.5), the predictors HP, Attack, Defense, Sp_Atk, Sp_Def, Speed, Height_m, Weight_kg, and Catch_Rate were used to construct regression trees on Pr_Male. First, a simple regression tree provided a MSE of 0.027, meaning that on average, its predictions would be off by about 16% for Pr_Male. Performing cross-validation 100 times to obtain the number of terminal nodes that would result in the lowest deviance pruned the tree down to two terminal nodes, increasing the MSE to 0.034. This tree only used Catch_Rate as a predictor, where a Catch_Rate lower than 47.5 would result in a higher average value for Pr_Male (See Figure 13). Such a simplistic model is not inherently useful and highly biased; pruning the original tree down to four terminal nodes instead resulted in an MSE of 0.032 and kept Catch_Rate, Defense, and Height_m as predictors (Pokémon with

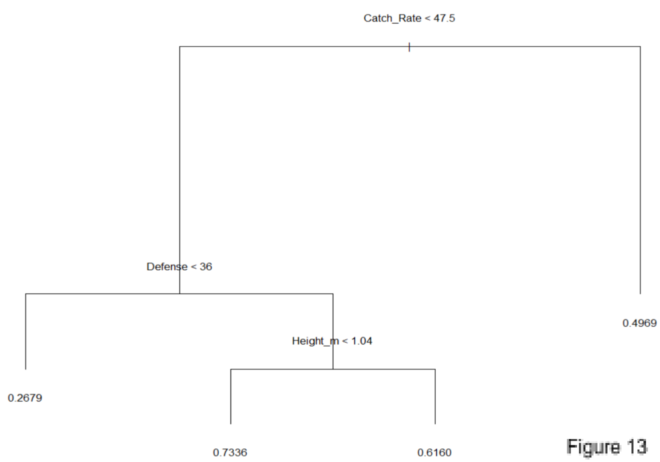


Figure 13

a Catch_Rate less than 47.5 and Defense less than 36 were least likely to be male at 26.8% and Pokémon with a Catch_Rate less than 47.5, Defense greater than 36, and Height_m less than 1.04 m were most likely to be male at 73.4% based on this model).

Performing bagging on the original tree resulted in a model with an MSE of 0.0305, explaining 23.6% of the variance in Pr_Male. The variable importance plot showed Catch_Rate is the most important predictor for Pr_Male, with Defense being second most important (corroborated by the cross-validated tree earlier). A random forest model using stubs (m=1) resulted in an MSE of 0.0305 and explained 23.7% of the variance in the model. Once again, Catch_Rate was the most important variable in this model, although Attack was rated second most important. None of these models proved to have great predictive power for Pr_Male; however, it did establish that Catch_Rate, Defense, and Attack were the most important variables for predicting Pr_Male. Creating a new regression tree which cuts out the least important variables, then performing bagging and random forests, did not result in better models (meaning that the less important variables in the original models still had some explanatory power for Pr_Male). Despite the tree model not having great predictive power, it is easy to interpret, and is an easy way to visualize the data. The cross-validated, pruned tree is therefore not great for predictions, but is a good starting point to visualizing the relationship between Pr_Male and the numeric stats.

Although performing PCA on the full data set did not result in overly useful models, PCA using the same predictors might result in principal components that are useful for predicting Pr_Male. Since the observations where hasGender is False have null values for Pr_Male, the dataset was split into a subset containing only observations where hasGender is true. PCA was then performed on the resulting dataset, and similar to before, the first two components met the Kaiser criterion. The next three principal components were not far below satisfying the Kaiser criterion (e.g. standard deviation of 0.84 for PC5) and cumulatively, the first five components explained 84% of the variation in the data, those five components were kept. A linear model was fit on the first five components, and using backward selection with a cutoff p-value of 0.05, reduced the model down to looking at PC1 and PC4, with p-values of 0.0030 and 6.28e-07 respectively in the linear model. The final linear model had an intercept of 0.56, and coefficients of 0.014 and -0.046 respectively for PC1 and PC4. Looking at the loadings for PC1, Attack, Sp_Atk, Sp_Def, Height_m, and Weight_kg had high positive loadings, while Catch_Rate had a high negative loading. Based on previous analyses showing that Legendary Pokémon had a low catch rate, high attack, and large heights, PC1 can loosely be interpreted as the Legendary Pokémon. PC4 was a little more tricky to interpret; there was a high loading for HP, and high negative loadings for Attack and Defense. So loosely, PC4 refers to Pokémon with a high amount of health but low Attack and Defense. Taking this back to the linear model, the taller, heavier, and more powerful a Pokémon is, the more likely it is to be male, while higher HP but lower attack/defense make

	PC1	PC4
HP	0.33	0.46
Attack	0.35	-0.56
Defense	0.31	-0.47
Sp_Atk	0.33	0.24
Sp_Def	0.34	0.22
Speed	0.22	-0.18
Height_m	0.35	0.29
Weight_kg	0.36	0.11
Catch_Rate	-0.40	0.16

Figure 14

a Pokémon more likely to be female. This analysis lends support to the idea that the Pokémon developers have a gender-normative bias in developing Pokémon. In other words, the developers reinforce the stereotypes that males tend to be more powerful than females in the Pokémon universe.

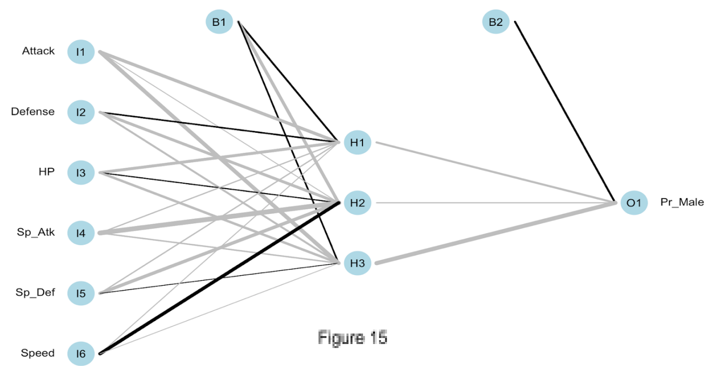


Figure 15

A neural net predicting for Pr_Male was generated using the predictors Attack, Defense, HP, Sp_Atk, Sp_Def, and Speed (See Figure 15). To predict for Pr_Male, the training and testing set were cleaned to omit for N/A values in the Pr_Male column. This was achieved by applying “which(hasGender=='True'),]” to the testing and training set. From testing the node amount on one and two hidden layers of the Pr_Male neural net, it appears that three nodes to provide the lowest MSE, with a value of 0.03845484. A linear model of Pr_Male using the same predictors as the neural net appears to provide a close, though slightly better MSE value of 0.03764972. The neural net was also generated with the addition of Weight_kg, Height_m, and Catch_Rate predictors, however, the MSE did not appear to benefit regardless of the number of nodes implemented with the addition of these predictors. In terms of understanding the correlation of Pokémon stats and Pr_Male, K-Means clustering appears to provide more support via interpretability.

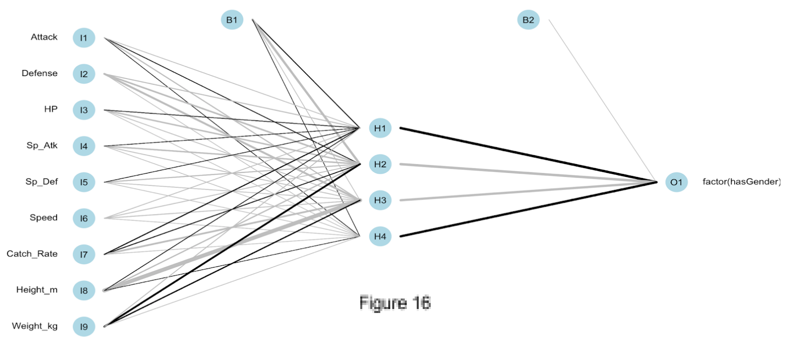


Figure 16

The predictor hasGender was tested for predictability modeled by a neural network using the predictors HP, Attack, Defense, Sp_Atk, Sp_Def, Speed, Weight_Kg, and Height_m (See Figure 16). It appears that the neural net model is effective in predicting hasGender. The neural net fit so effectively to hasGender that 9 nodes within 1 hidden layer used to generate the neural net resulted in 0 misclassifications. This is clearly an example of overfitting the model. In order to avoid overfitting that would contribute to long-run misclassification, a variety of node sizes were used on the neural net to provide the best misclassification on the testing set (See Figure 17.1 and 17.2).

The optimal node size for the best testing misclassification appears to be 4 nodes within 1 hidden layer. Based on the results of previous analyses in this investigation, it should not come as a surprise that a neural net was effective in predicting hasGender. Given that "better" Pokémon statistics increase the likelihood of being of the legendary type, than so too should the likelihood that the Pokémon is genderless.

Training Misclassification

	1	2
False	49	8
True	11	437

Figure 17.1

Testing misclassification

	1	2
1	15	5
2	11	185

Figure 17.2

Conclusion

It was determined through analyses of the Pokémon dataset that Pokémon can be surprisingly progressive, however, there remain residual aspects of sexism specific to gendering and fundamental statistics[9]. Through analyses, the investigation suggested that the statistics of Pokémon can help in predicting a Pokémon's gender, and as well its likelihood of being legendary. From the use of regression trees, it was found that Pokémon with the highest probability of being female, according to Pr_Male, had the lowest Attack and Defense values. Similarly, it appears that there is a correlation between catchability and gender, such that Pokémon with lower statistics like Attack or Defense are easier to catch and more likely to be female. It could be said that this bias towards males could be a product of Pokémon's original target audience. Pokémon was released in 1996[10], during a time when video games were primarily targeted at males. The distinction of "stronger" traits ascribed to male Pokémon may also be a product of the game developers own internalized biases. Where Pokémon may be considered more progressive is in its depiction of the legendary Pokémon type, which are the "strongest" types of the Pokémon universe. While there is a functional reason for the majority of legendary Pokémon being genderless since it disallows for the ability to breed[11], the representation of non-binary gender identity is progressive and incredibly important as game designers could have simply have made all legendary Pokémon one gender. While the distribution of genders in the dataset is fairly even, the statistics that determine a Pokémon's strength are biased towards male gendered Pokémon. Simply put, Pokémon gendering and statistics are inherently sexist, regardless if the developers were mindful of these distinctions or not. Altogether, Pokémon is an example of how culture and internalized biases can be reflected in game design.

---

[9] Ballou, Elizabeth. "Women Play Video Games Too." Bustle. December 17, 2018. Accessed April 05, 2019. https://www.bustle.com/articles/32730-sexism-in-gaming-culture-women-can-be-gamers-too-and-dont-you-forget-it.

[10] "History of Pokémon." History of Pokémon - Bulbapedia, the Community-driven Pokémon Encyclopedia. Accessed April 05, 2019. https://bulbapedia.bulbagarden.net/wiki/History_of_Pokémon.

[11] Xoxo. "Pokeffects: Exploring Gender Roles in the Pokemon Franchise." Pokeffects: Exploring Gender Roles in the Pokemon Franchise. January 01, 1970. Accessed April 05, 2019. http://barbaricpoetries.blogspot.com/2012/05/pokeffects-exploring-gender-roles-in.html.

Bibliography

"About the Pokémon Company International." The Pokémon Company International, accessed April 4, 2019. https://api.pokemon.com/us/about-pokemon/

Alopez247. "Pokémon for Data Mining and Machine Learning." Kaggle. March 05, 2017. Accessed April 05, 2019. https://www.kaggle.com/alopez247/pokemon.

Ballou, Elizabeth. "Women Play Video Games Too." Bustle. December 17, 2018. Accessed April 05, 2019. https://www.bustle.com/articles/32730-sexism-in-gaming-culture-women-can-be-gamers-too-and-dont-you-forget-it.

"Catch Rate." Bulbapedia, accessed April 4, 2019. https://bulbapedia.bulbagarden.net/wiki/Catch_rate.

"Creating & Visualizing Neural Network in R." Accessed April 4, 2019. https://www.analyticsvidhya.com/blog/2017/09/creating-visualizing-neural-network-in-r/.

"Generation." Bulbapedia, accessed April 4, 2019. https://bulbapedia.bulbagarden.net/wiki/Generation

"History of Pokémon." History of Pokémon - Bulbapedia, the Community-driven Pokémon Encyclopedia. Accessed April 05, 2019. https://bulbapedia.bulbagarden.net/wiki/History_of_Pokémon.

"Mclust: Model-Based Clustering." R Package Documentation, accessed April 4, 2019. https://rdrr.io/cran/mclust/man/Mclust.html.

"Predicting Legendary Pokémon." Accessed April 4, 2019. https://www.kaggle.com/excaliburzero/predicting-legendary-pokemon

"Pokemon Go Type Chart: Every Type Strength and Weakness for Battle Counters." VG247. January 29, 2019. Accessed April 05, 2019. https://www.vg247.com/2019/01/29/pokemon-go-type-chart-strength-weakness-effectiveness-counters/.

Xoxo. "Pokeffects: Exploring Gender Roles in the Pokemon Franchise." Pokeffects: Exploring Gender Roles in the Pokemon Franchise. January 01, 1970. Accessed April 05, 2019. http://barbaricpoetries.blogspot.com/2012/05/pokeffects-exploring-gender-roles-in.html.

Appendix A

Clustering dendrograms were produced of the Single, Average, and Complete Linkage in both Euclidean and Manhattan distance. These dendrograms are included here.