Legendary Pokemon Prediction

Seungmin Lee
ELEC400M Final Report
2022-04-20

1 Introduction

The Pokemon franchise was created by Satoshi Tajiri in 1996, and initially introduced as video game series. The games were very successful and brought huge profit to the company. Pokemon was also introduced in other media such as animations, films, and trading-card-game. Currently, in the year 2022, the Pokemon brand is the world biggest media franchises with estimated all-time sales of $100 billions [1].

One of the fascinating part of Pokemon is existance of legendary Pokemons. They are very rare and generally stronger than other pokemons. Through this report, the data-set of 721 Pokemons will be used to develop several machine learning models to correctly classify legendary Pokemons. In details, for the classification methods: logistic regression, support vector machine (SVM), random forest, and K-nearest neighbors (KNN) are utilized. Also, to improve the performance of classification further, Principal Component Analaysis (PCA) are applied on the logistic regression and KNN algorithms. The data-set used in this report is derived from the Kaggle website created by Asier[2].

Note that Machine Learning(ML) models are trained on a Laptop with Intel Core i7-1165G7 CPU, Intel Iris Xe GPU and 16G memory. Also jupyter notebook's source code is publicly uploaded on GitHub https://github.com/seung7/Legendary-Pokemon-Prediction-ML

2 Data Analysis and Pre-Processing

2.1 Data Attributes

Pokemon's attributes are described in below:

- #(Number): ID for each Pokemon
- Name: Name of the Pokemon
- Type 1: Primary Type
- Type 2: Secondary Type
- Total: Sum of HP, Attack, Defense, Sp Atk, Sp Def, Speed
- HP: Health Point
- Attack: Attack point
- Defense: Defense point
- Sp_Atk: Special Attack point
- Sp_Def: Special Defense point
- Speed: Speed of the Pokemon
- Generation: Number of generation of the Pokemon when the Pokemon was introduced
- isLegendary: if true, the Pokemon is the legendary Pokemon
- hasGender: if True, the Pokemon has a gender

- Pr_male: if the Pokemon has a gender, the probability of being a male
- Egg_Group_1: if two Pokemons are belong to the same Egg Group, They are interbreedable
- Egg_Group_2: Secondary Egg Group
- hasMegaEvolution: if true, the Pokemon can evolve into its mega form
- Height_m: Height of the Pokemon in meters
- Weight_kg: Weight of the Pokemon in kilograms
- Catch_Rate: Catch Rate. higher rate indicates that the Pokemon is can be easily catched.
- Body_style: Body shape of the Pokemon

2.2 Number of Legendary Pokemons

We are interesting in classifying legendary Pokemons. In our sample, there are 46 legendary Pokemons, and 675 non-legendary Pokemons.

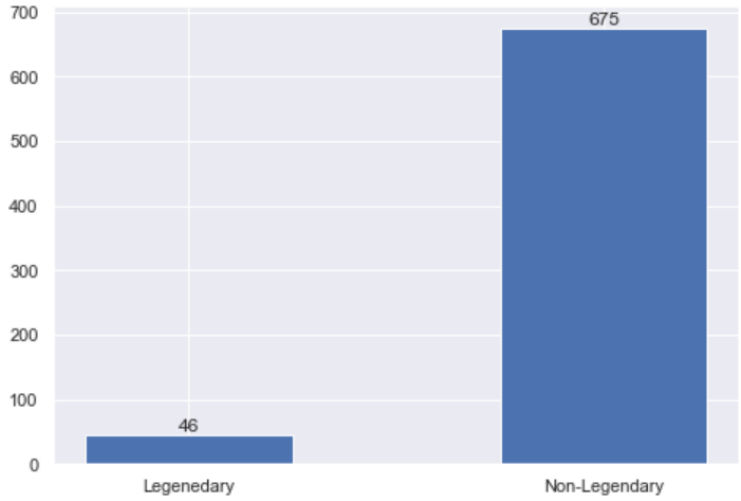

Figure 1: Legendary vs Non-Legendary Bar Chart

2.3 Data Pre-processing

If an attribute has any empty values, it has to be opted out before training the machine learning(ML) models. 371 Pokemons are missing 'Type_2', 77 Pokemons are missing 'Pr_male', and 530 Pokemons are missing 'Egg_Group_2'. These fields cannot be used for training. Therefore, they are dropped.

'Name' and 'Number' fields of the Pokemons are opted out as well because each Pokemons has their unique name and number, and they have no relationship for a Pokemon being a legendary.

Categorical values have to converted into numerical values as well. Firstly, 'True' value of any Boolean fields(isLegendary, hasGender, hasMegaEvolution) is converted into integer 1, and 'False' is converted to 0. Secondly, non-Boolean categorical values(Type_1, Color, Body_style, Egg_Group_1) are converted into numerical values by using *Pandas' get_dummies* method.

After pre-processing the data-set, out of 721 data, 70 percentage of them are randomly set as a training data-set, and remaining 30 percentage are set as a testing data-set.

2.4 Correlation

Before performing classification, correlation between 'isLegendary' field and all the other fields are analyzed by using *Pandas' corr* method. The pie chart in Figure 2 shows the top 10 attributes that have the most correlation to the 'isLegendary' field.

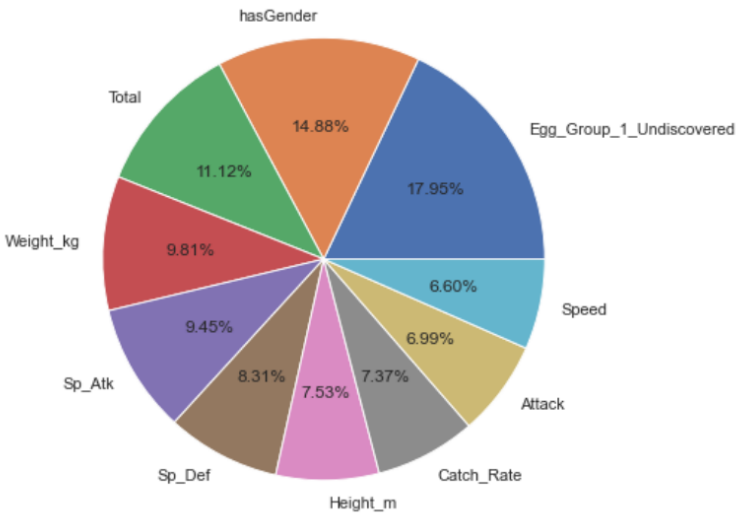

Figure 2: Correlation Pie chart

3 Classification

3.1 Evaluation Methods

To evaluate classifiers, two evaluation metrics are used throughout this report: accuracy score and precision score

$$\text{Accuracy Score} = \frac{TP + TN}{TP + FP + TN + FN}$$
$$\text{Precision Score} = \frac{TP}{TP + FP}$$
$$\text{where, } TP : number\ of\ True\ Positive$$
$$TN : number\ of\ True\ Negative$$
$$FP : number\ of\ False\ Positive$$
$$FN : number\ of\ False\ Negative$$

3.2 Logistic Regression, Linear SVM, Poly SVM, Random Forest, and KNN

At first, using logistic regression classifier, the ML model is trained.

Next, models are developed with linear SVM, poly SVM, random forest, and KNN. Before developing these models, cross-validation with 5 folds is used to tune their hyper parameters for each models. The Result is shown below:

- Logistic Regression -> Accuracy: 0.977, Precision: 0.818
- Linear SVM -> Accuracy: 0.982, Precision: 0.900
- Poly SVM -> Accuracy: 0.968, Precision: 0.727
- Random Forest -> Accuracy: 0.991, Precision: 0.857
- KNN -> Accuracy: 0.977, Precision: 0.818

The result is also represented as bar graphs, located at the end of report as Figure 4.

3.3 Discussion

Linear SVM's result of accuracy score of 0.982 which is higher than poly SVM's accuracy score of 0.968. Therefore, the 'isLegendary' field is more linearly related to the rest of data-set than being polynomially related.

Also the random forest classifier, the method of randomly reducing dimension of the data-set, yields the highest accuracy score of 0.991. High scores on the random forest means that there exist attributes that are not correlated for a Pokemon being a legendary. Therefore, getting rid of these attributes will yield better scores. One method that can effectively reduce the dimensions of the data-set is using PCA. To improve the performance of the classifiers, in the next section, the method of PCA will be discussed.

4 Improving Performance with PCA

4.1 PCA with 2 principal axes

Before training the PCA model, to confirm if PCA is effective, the training data-set are projecting onto 2 principal axes. The scatter plot is shown below, and it successfully separates legendary Pokemons from non-legendary Pokemons.

Figure 3: PCA transformation with 2 principal axes

4.2 PCA Result and Discussion

PCA is applied on the logistic regression and KNN models. These models are selected because they yielded relatively low accuracy and precision scores.

To find the appropriate number of principal axis, PCA models were initially trained with 2 axes and incremented up to 15 axes, and one that gives the highest accuracy and precision scores is selected. For the logistic regression, 6 principal axes gave the highest scores. For the KNN model, 8 principal axes gave the highest scores.

Using less number of principal axes then these selected number of axes would underfit the data-set, while using more than these number of axes would overfit the data-set.

As a result, for both cases, the accuracy and precision scores are improved. Logistic regression model is improved from accuracy, precision score from 0.977, 0.818 to 0.995 to 0.923, and KNN model is improved from accuracy, precision scores from 0.977, 0.818 to 0.991, 0.917. The bar graph of the result is shown in Figure 4.

As a conclusion, reducing the dimensions of data-set with PCA improves the classifications performance for our data-set.

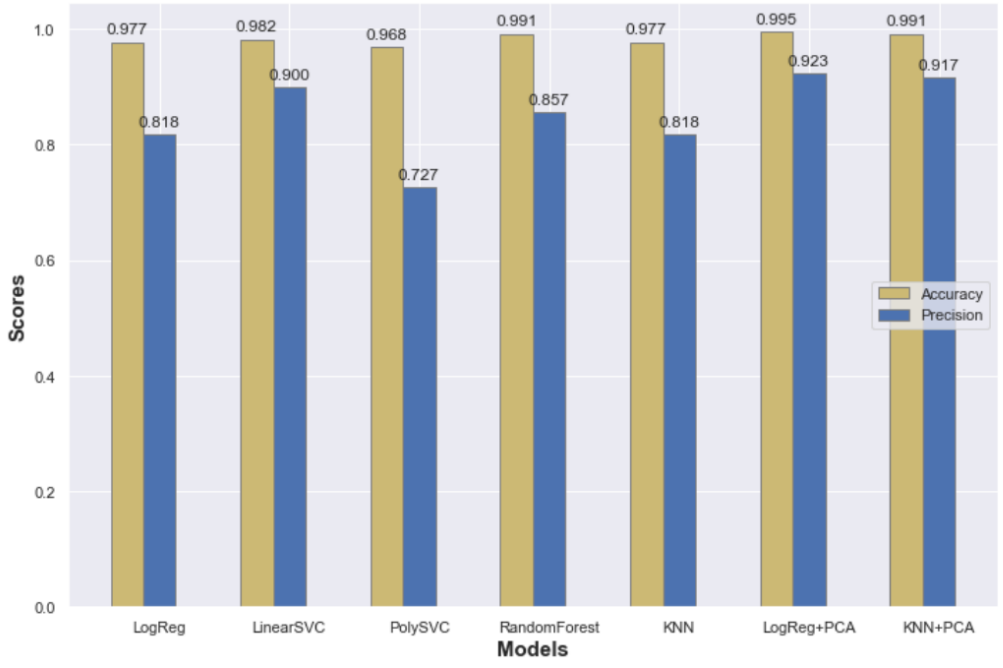

Figure 4: Accuracy and Precision Scores for all Classifiers

5 References

[1] K. Buchholz and F. Richter, “Infographic: The Pokémon Franchise Caught ’Em All,” Statista Infographics, 24-Feb-2021. [Online]. Available: https://www.statista.com/chart/24277/media-franchises-with-most-sales/. [Accessed: 17-Apr-2022].

[2] A. L. Zorrilla, “Pokémon for Data Mining and Machine Learning,” Kaggle, 05-Mar-2017. [Online]. Available: https://www.kaggle.com/datasets/alopez247/pokemon. [Accessed: 17-Apr-2022].