Pokémon Classification Project

Charles Christopher Hyland

450411920

Table of Content

Business Understanding

Business Objectives

Background

With the discovery of new regions in the Pokémon world, strange and unidentified Pokémon have been appearing in the wilderness. With such unknowns, this can lead to a myriad of dangers from novice Pokémon trainers facing such dangerous unknown Pokémon to entire cities being threatened with the potential destructive capabilities of Pokémon. Therefore, it is imperative that we are able to classify Pokémon and record as much information about them. Professor Oak, renowned expert in the field of Pokémon has already begun the initiative of classifying Pokémon through the invention of the Pokédex, a portmanteau of the words "Pokémon" and "Index", a device which acts as a digital encyclopedia of currently identified Pokémon. As the sole inventor of the Pokédex, he is the key individual in the organization. Additionally, Devon Corporations, a large trading company, has agreed to finance the project for its entirety. As this project is based within a small scale organization, there is a lack of hierarchy and participants in the project. However, research assistants to professor Oak and other aides will be impacted as we will require their input into the project. Currently, no steering committee has been or will be created.

Currently, the Pokédex is a database containing information about the different types of Pokémon. Currently, a Pokémon's is added to the Pokédex's database when a trainer encounters a new Pokémon. However, any information regarding the Pokémon is only collected once a Pokémon is caught by the Pokémon trainer. From this, it is easy to see discrepancies created in the Pokédex's database whereby if a trainer encounters a new Pokémon but fails to capture it can lead to the database being filled with null information. Therefore, a classification system whereby upon encountering and "battling" a Pokémon, the Pokédex will be able to classify what type of Pokémon the trainer has encountered. More specifically, if given the attributes of a Pokémon observed from battle (e.g. it's attack, defence, speed, and other attributes), can we then classify what type of Pokémon it is (e.g. whether it is a fire, water, or any other type of Pokémon)

From this, there is a lack of solutions and alternatives to this issue and therefore this emphasises the need for such a project to be in place. There are a plethora of advantages to the project such as being able to fix such discrepancies and prevent medical injuries to trainers. Furthermore, discussion with key stakeholders such as professor Oak and others such as Devon Corporations has seen that this solution been highly accepted within the organisation and also called for.

Business Objectives

The project is commissioned with the following objectives:

- Given an unknown Pokémon, the program is able to classify based on their characteristics on what type of Pokémon
- Classify whether the Pokémon has a secondary type and what that type is
- Identify potential “legendary” Pokémon
- With all this information, we then like to compile information about new unseen Pokémon to remove any discrepancies in the Pokédex and identify new Pokémon

Business Success Criteria

There has been previous projects attempting to classify Pokémon and therefore, the results from those studies will be used as a benchmark for this project[1].

Tentatively, the project will be a success if:

- An error prediction rate of less than 10%
- Ability to classify all 18 types of Pokémon
- Works for observations with a quarter of variables having incomplete information[2]

Assess Situation

Inventory of Resources

The hardware needed to be supported is the Pokédex whilst the current system that is also needed to be supported is the in-house database constructed by Professor Oak, which is also populated from other Pokémon trainers who has access to the Pokédex. This Data Based Management System takes the form of an operational database, whereby it is highly suited to deal with Online Transaction Processing (OLTP). Furthermore, a data warehouse is not yet implemented, therefore the data is not structured in such a way to allow for

[1] https://github.com/dimart/pokemon_recognition

[2] https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/

efficient analytics. Thus, there is a need to improve the classification algorithms on the Pokédex to identify the Pokémon before storing it into the database.

Figure 1. Pokédex

Data

The current dataset we possess contains information on Pokémon across 6 different regions: Kanto, Johto, Hoenn, Sinnoh, Unova, and Kalos. In total, there are 721 identified Pokémon with many more yet to be discovered in new regions. All information regarding the Pokémon has been provided to us upon request.

Personal

Currently, only Professor Oak is available upon contact and no other individuals who are tasked with the Pokédex. There appears to be little expertise with regards to database management, and therefore a database specialist will be consulted.

Risk

There are no particular risks at this stage of the project. However, a full detailed outline of all possible risks and their impact is explored later in this report.

Any algorithms developed for the Pokédex, will be able to be rolled out to all other devices.

Requirements, Assumptions, and Constraints

Requirements

There are no legal and security restrictions on the project's result as the project is intended to be deployed for the public good in combatting against dangerous Pokémon. All stakeholders involved in the project have been notified and has given consent to the scheduling on the project. Any results from this project will be swiftly implemented across all Pokédexes and Pokémon databases in order to enhance the classification algorithms of current software.

Assumptions

Economic factors could affect project

There are no obvious economic factors affecting this project as the dataset has been collected already. Furthermore, being the only product on the market, the Pokédex enjoys a unique position of being a monopoly for devices that can be used to identify Pokémon. Therefore, no competition is faced against other products.

Data quality assumptions

There is the possibility of measurement error whereby the attributes of each Pokémon are recorded incorrectly. This then lead to a measurement error in the independent/predictor variable. This may be an issue as measurement errors in the independent variable will lead to inferences made regarding the dataset to be incorrect and we may even face the **Classical Errors-in-Variables (CEV)** problem. However, the assumption has been made that all research associates on the Pokédex has correctly recorded the information. Likewise, the possibility of a ***systematic error,*** whereby increasing the sample size does not reduce the error, is assumed to not be present in the dataset and that those who have collected the data has undertaken the necessary steps to obtain valid results. Therefore, with regards to the veracity of the data, we assume the data's integrity to not have been compromised at any stage prior to the project.

With regards to the final results, stakeholders who would want to utilise the results of the project will primarily be interested in the final results of the project. Therefore, upon deployment, all the underlying mechanics in the project will be part of a black-box whereby the results of the processes will be presented to the user in a friendly Graphical User Interface model.

Constraints

As mentioned earlier, there has been no accessibility and legal issues regarding the data, as the data is part of the public domain to those who possesses a Pokédex. All funding for the project is covered by Devon Corporations.

Risks and Contingencies
These are the possible risks to the project and evaluation of the risks.

Risk	**Impact**
Scheduling	There is no issue with regards to scheduling risks as the project's length does not impact any stakeholder's greatly if it is delayed.
Financing	Budgetary problems should not be faced as the costs to the project is minimal. Furthermore, with generous funding from Devon Corporations, this risk is further mitigated.
Data	The possibility for the data to be incorrect is highly likely. Therefore, numerous Econometrics and Statistical measures should be in place to control for such errors.
Results	Even if the initial results are less dramatic than expected, it does not make too much of a difference to the project whereby the results of the project can still be used for research purposes and improved upon for further iterations and improve the classification of the algorithms.

No immediate risks except for the fact that as this project's Pokémon classification is displayed in dangerous scenarios, such as a Pokémon trainer battling a new unknown dangerous Pokémon, where failure of the classification algorithm can lead to terrible consequences.

With all this in mind, there are no pressing risks to the projects and therefore the contingency plan is minimal. The contingency plan consists of halting the project and requesting the researchers resample Pokémon, analyse them, and record them to obtain more representative data.

Terminology

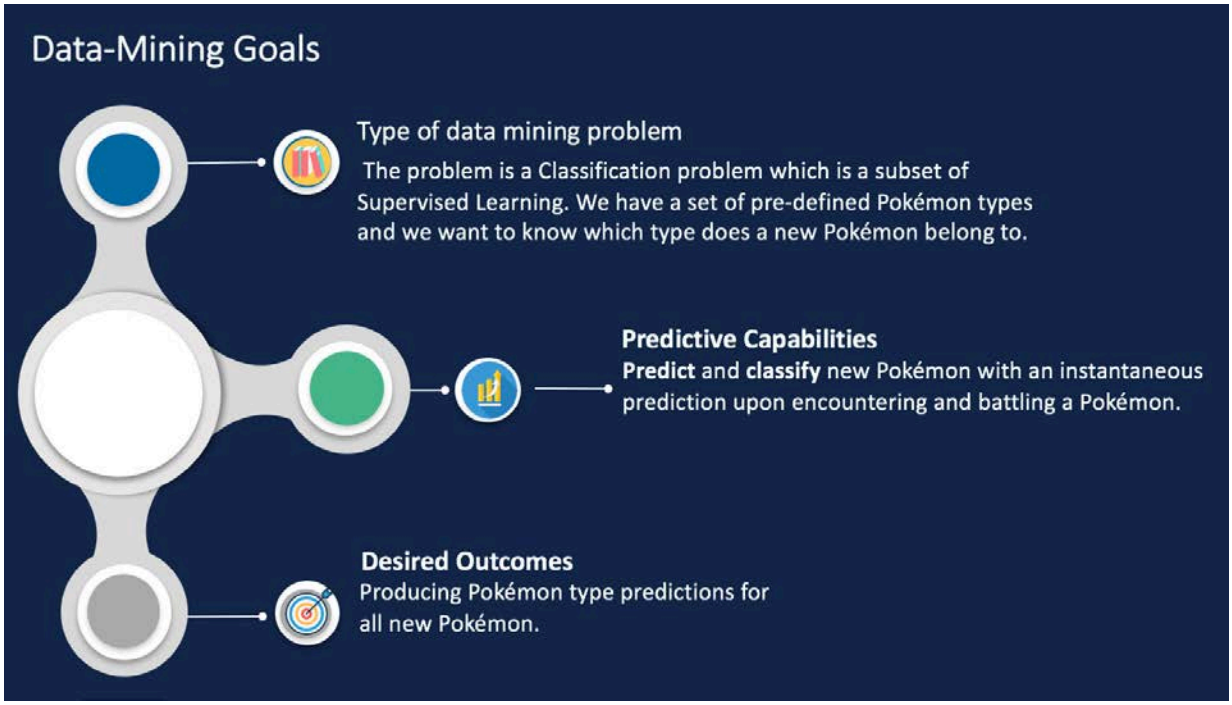
Type – There are currently 18 different types of Pokémon ranging from fire, water, grass, and many more.
Pokémon Trainers – Person who catches, trains, cares for, and battles with Pokémon.
Pokéball – Devices that are used to capture Pokémon. The Pokémon is captured and stored in the ball.

Figure 2. Pokéball

Costs and Benefits Analysis

Benefits	Cost
The classification algorithm is able to save money spent on medical costs for injured Pokémon trainers, as upon encountering a dangerous Pokémon, they will now know it is wise to not engage in combat.	There are no costs in collecting the data as it has been collected already and no external data is being utilised.
Reparation fees of cities and other public properties are saved as the algorithm is able to identify dangerous Pokémon and take preventive measures and contingency plans to protect the aforementioned properties.	There are minimal costs with the deployment of the results as the software simply needs to be installed across all Pokédexes.
The advancement of knowledge regarding Pokémon is also highly valuable whereby the understanding of Pokémon will greatly be increased.	There are costs associated with operation of the software whereby upon a Pokémon trainer discovering a new Pokémon, the Pokédex will utilise the cloud in order to carry out the classification of the Pokémon.

Determine Data Mining Goals



Data Mining Success Criteria

The methods utilised for evaluation of the model will include:

- Confusion Matrices
- Root Mean Squared Error
- R-squared
- % of inaccurate classifications

Benchmarks used for the project are:

- 90% accuracy in identifying and classifying Pokémon[3]
- Confusion Matrix whereby classification accuracy for each Pokémon type is over 80%

The successful deployment of the project is also tantamount to the success of the project

Produce Project Plan

Project Plan

A project plan was constructed:

Phase	Time	Resources	Risks
Business Understanding.	2 weeks.	All analyst.	Data problems.

[3] https://github.com/dimart/pokemon_recognition

Data Exploration.	4 weeks.	Database analyst.	Data problems. Technology problems.
Model Construction.	3 weeks.	Data mining consultant, business analyst.	Inability to find adequate model. Data problems.
Review and Presentation of results.	2 weeks.	Business analyst.	Inability to implement results.

Initial Assessment of Tools and Techniques
A possible way of structuring this problem is a nested multinomial classification problem. Therefore, the potential tools to use are:

- KNN algorithm
- Nested Logit Model
- Random Forest
- Boosting
- Mixed discriminative Analysis
- Multinomial Logit Model

Data Understanding

Collecting Initial Data
The data comes from existing data scraped over numerous websites via APIs. No other data is required for the project at this stage.

Currently, there are 800 observations from a cross-sectional dataset. From this it appears that the data allows us to draw generalizable conclusions and also assists us in making accurate predictions.

However, a possible issue is that there are not many attributes. This may be a problem as our predictions may suffer from omitted variable bias and that important predictors we could be using are not present in the data.

The data has already been merged and pre-processed. This brings into question the veracity of the data as the data has been processed by other parties and it is uncertain whether strict data processing methodology was adhered to.

Any missing value issues in the data was rectified as follows:

1) Find out the type of Pokémon it is.
2) Calculate the median of the attribute that the missing value belongs to from the rest of the Pokémon of the same type.
3) Use the median to replace the missing value.

It is worth acknowledging the flaws in this methodology, yet it is sufficient for the task at hand.

Describe Data

The data comes in as a tabular format, which contains 12 attributes.

#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	False
2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60	1	False
3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	False
3	VenusaurMega Venusaur	Grass	Poison	625	80	100	123	122	120	80	1	False
4	Charmander	Fire	NaN	309	39	52	43	60	50	65	1	False

Fig. 3 Snapshot of sample of data.

Here, in a snapshot of the data, we can see it is organized in a table format. Furthermore, if a Pokémon does not have a second type, then NaN is used in place. The “Legendary” attribute return False if the Pokémon is not a legendary type Pokémon and returns True if it does.

The variables present in the data and its respective type is as follow:

Variable	Type
Type 1	String
Type 2	String
Total	Numeric
HP	Numeric
Attack	Numeric
Defence	Numeric
Special Attack	Numeric
Special Defence	Numeric
Speed	Numeric
Generation	Numeric
Legendary	Boolean

Fig 4. List of variables

The binary variable of Legendary will have to be converted into a numerical binary format (0 for False and 1 for True) to make data manipulation easier. Generation is coded as a numeric range from 1-6, where each number indicates a different region.

The attributes that seem most promising from the data in terms of using to classify Pokémon are:

- Attack

- Defence
- Special Attack
- Special Defence
- Speed

The generation (the region in which the Pokémon comes from) of the Pokémon currently seems irrelevant. However, it may be the case that Pokémon of a certain type from a particular region, has a combination of features which differ to Pokémon of the same type in other regions.

Amount of data

Having a large data set ensures a more accurate model and gives us more information. However, the drawback to this is that the processing time with Python will be longer due to the polynomial complexity of many algorithms to be implemented in the model.

	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation
count	800.00	800.00	800.00	800.00	800.00	800.00	800.00	800.00
mean	435.10	69.26	79.00	73.84	72.82	71.90	68.28	3.32
std	119.96	25.53	32.46	31.18	32.72	27.83	29.06	1.66
min	180.00	1.00	5.00	5.00	10.00	20.00	5.00	1.00
25%	330.00	50.00	55.00	50.00	49.75	50.00	45.00	2.00
50%	450.00	65.00	75.00	70.00	65.00	70.00	65.00	3.00
75%	515.00	80.00	100.00	90.00	95.00	90.00	90.00	5.00
max	780.00	255.00	190.00	230.00	194.00	230.00	180.00	6.00

Fig 5. Summary statistics of the data

We can also noticed that the mean is greater than the median for the attributes of the Pokémon, thereby suggesting that the distribution of such variables are positively or right skewed.

Currently, the summary statistics of the overall data does not provide us any insights into the business problem. We would be able to derive more information by breaking down the dataset by Pokémon type.

When splitting the data up by the type of Pokémon and summarizing the data based on certain variables, a few noticeable results appeared:

Variable	Maximum	Type	Minimum	Type
Attack	112.13	Dragon	70.97	Bug
Defence	126.37	Steel	65.71	Fairy
Special Attack	96.84	Dragon	53.87	Bug
Special Defence	88.84	Dragon	62.75	Ground

Speed	102.5	Flying	55.26	Steel

From this, it appears that certain Pokémon types are associated with certain characteristics and due to the discrepancies between the maximum and minimum value for each type, it is worthwhile to assume that different types of Pokémon have unique attributes associated with them. From this, we are able to proceed with the business question and makes it worthwhile to further examine this trend.