Project 4
Group 9
Tesla Superchargers
October 3, 2024

Robert Williard
Amber Venes
Zachary Dallas
Earl Lewis

Pokemon Writeup

Pokémon is a well-known and beloved game that has been around for decades. Initially released in 1996, the first generation introduced 151 unique Pokémon, which has now grown to an impressive 1,025 across nine generations. The main objective of the game is to catch every Pokémon and improve as a Pokémon Trainer. Each Pokémon has its own unique stats, offering players a wide range of possibilities for progressing in various ways with their Pokémon. This raises several questions: Which Pokémon are the easiest and most difficult to catch? Which generation of Pokémon is the best, based on average stats like Attack, Defense, Speed, Special Attack, and Special Defense? Finally, which type of Pokémon is the most common, and which is the least?

What is Our Dataset

The dataset selected for this project contains the information necessary to address these questions. It includes details such as the Pokémon's name, type, species, height, weight, abilities, catch rate, base friendship, base experience, growth rate, gender, HP, attack, defense, special attack, special defense, and speed. Using this data, we created a website featuring two dashboards with visualizations and developed a machine learning model trained to predict whether a Pokémon can be caught based on its catch rate. The goal of this project is to utilize the given data to train a machine learning model that can predict outcomes based on the provided information while ensuring the platform is user-friendly for any user.

Data Cleaning

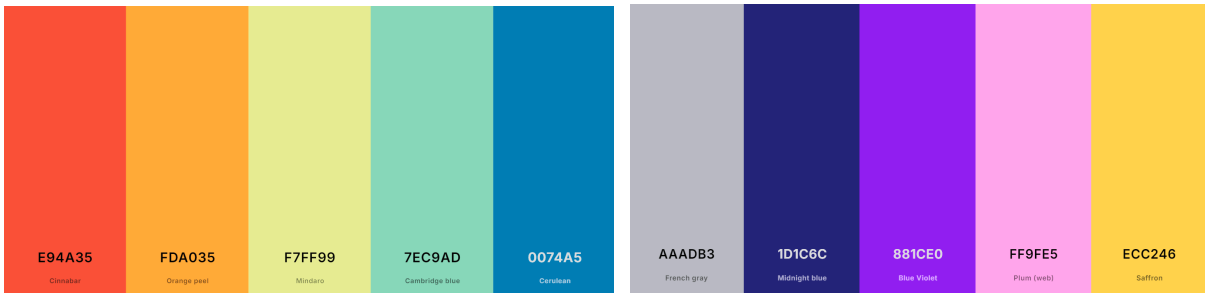
The first step of our project was to clean the dataset. Fortunately, this dataset required minimal cleaning. There were no null values, and only a small amount of data manipulation was needed. One of the main adjustments involved the "type" column. Instead of allowing Pokémon to have two types, we simplified the data to include only one type per Pokémon, dropping the second type. This was necessary because the dashboard visuals and the model had difficulty properly recognizing both types for the same Pokémon.

For the machine learning model, our primary focus was on the "catch rate" column, while for the dashboards, we mainly concentrated on attributes such as height, weight, type, name of the Pokémon, and stats like attack, defense, and others. Another

issue we encountered was that the dataset did not include the generation of each Pokémon. To address this, we had to import a second dataset to obtain the generations.

Color Design Considerations

The next step was to decide on a design concept for the visuals and website, as well as a color theme.

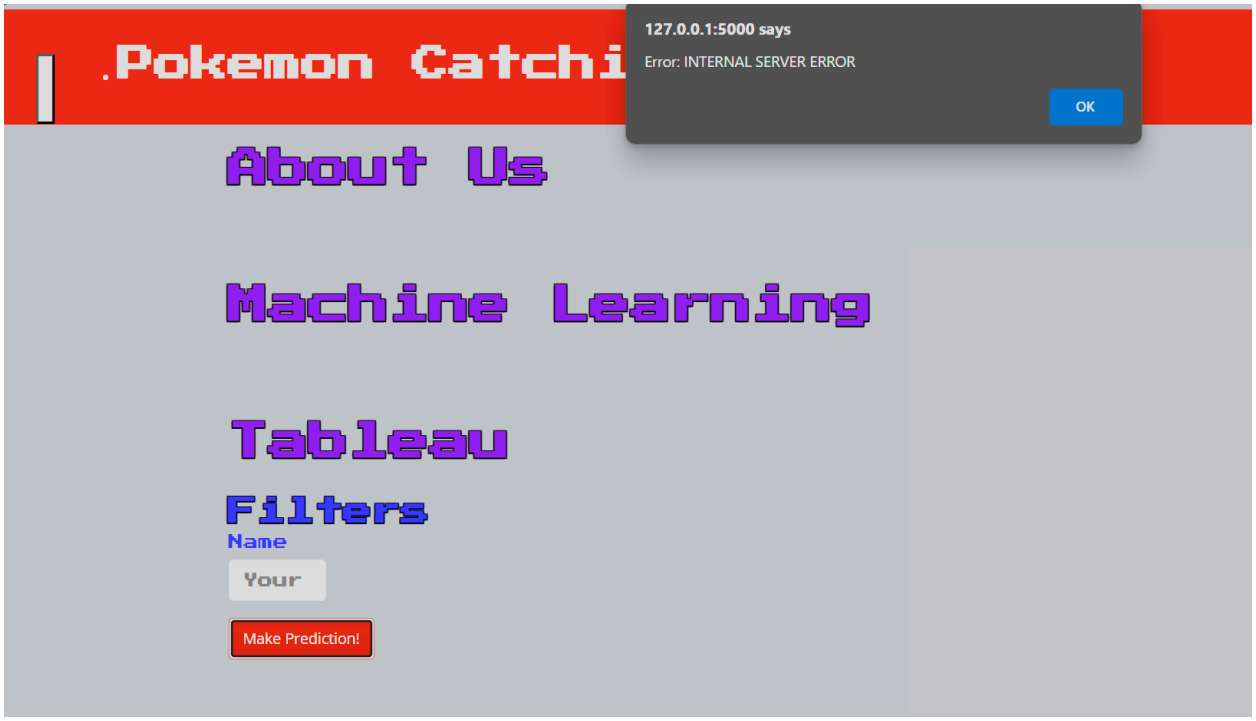


We decided to use a wide range of bright colors, as we felt this would best represent the vibrant world of Pokémon. Additionally, we aimed to keep the website design simple and clean, using a Nintendo-inspired font to capture the retro feel of the early Pokémon games. This combination of colors and design elements helped create a nostalgic yet visually appealing experience for users.

Machine Learning

For our machine learning model, we chose the Random Forest Regressor because it excels at handling complex datasets and delivering high accuracy through the use of multiple decision trees. This algorithm reduces overfitting by averaging predictions, resulting in reliable outcomes even in varied datasets. It is capable of

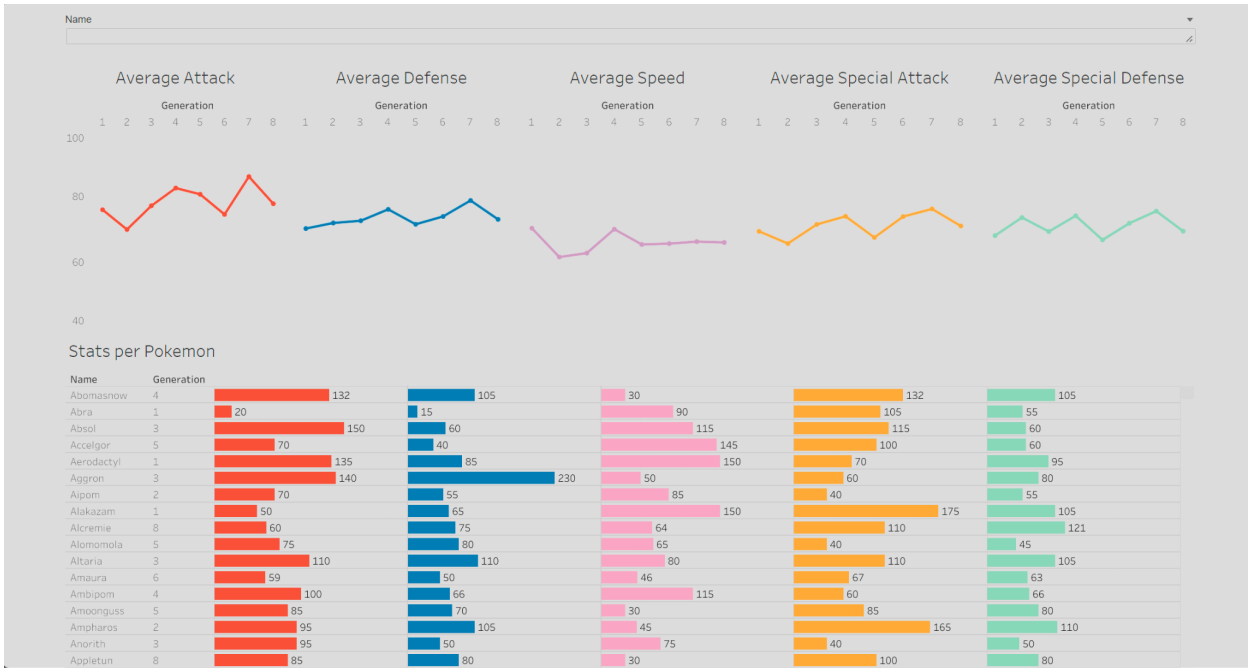
managing both numerical and categorical data while providing valuable insights into feature importance. Our evaluation revealed a Root Mean Square Error (RMSE) of 66%, demonstrating strong performance compared to other models. This indicates that the Random Forest Regressor effectively balances bias and variance, producing predictions that closely align with actual values.



However, we encountered some challenges during deployment, including version incompatibility and data type mismatches, which underscored the importance of ensuring all dependencies are aligned before going live.

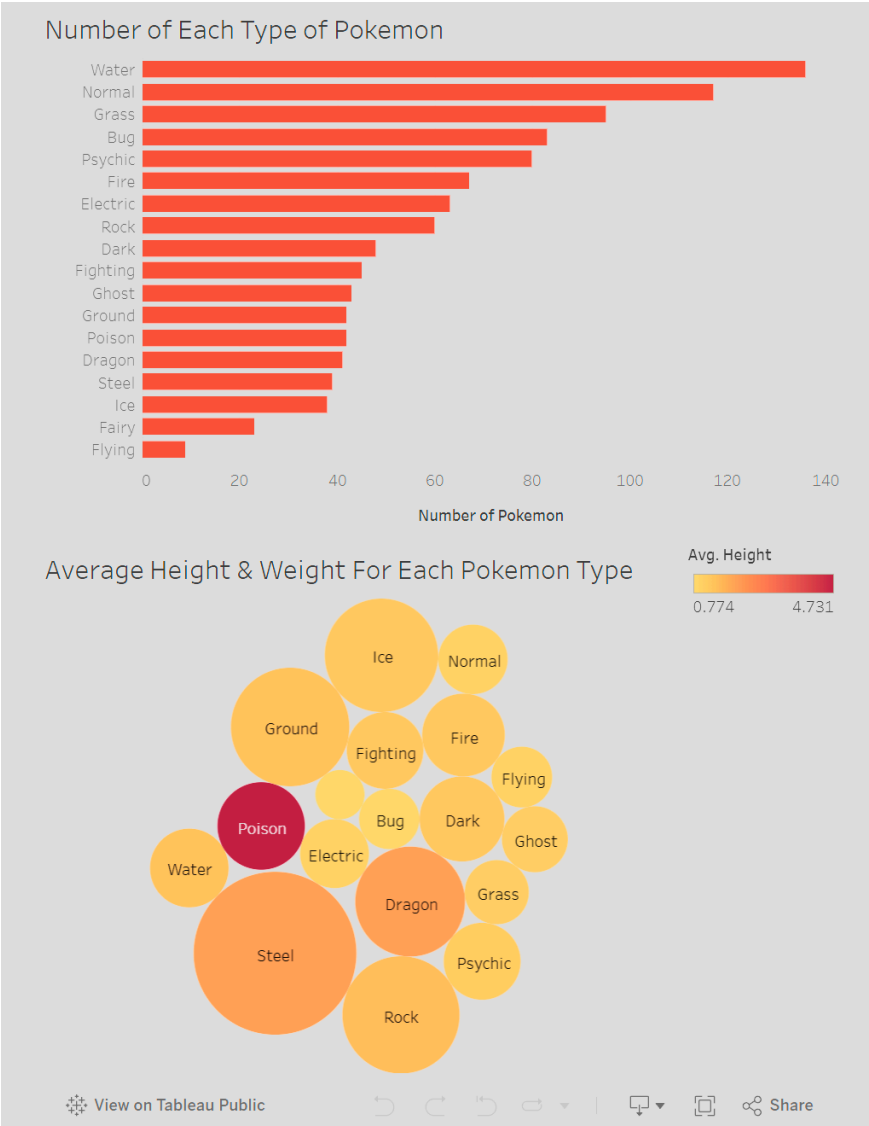
Dashboard Design Concepts and Answers To Our Questions

For both of our dashboards, we chose to keep the background off-white while using bright colors for the visuals to make them stand out and capture the attention of viewers. In the first dashboard, users are greeted with line charts that visualize the overall average stats of eight generations. This allows users to clearly see and compare the average attack, defense, speed, special attack, and special defense for each generation. The second visual is a horizontal bar chart that displays the stats of each individual Pokémon. This chart enables users to compare the statistics of each Pokémon rather than focusing solely on generational averages.



Our second question can be addressed by examining the very first visual. It shows that the top generation for average attack, defense, special attack, and special defense is the seventh generation, while the highest average speed belongs to the fourth generation. By visualizing this information based on generations, users can strategically target Pokémon according to the average statistics of each generation. From this analysis, it is reasonable to conclude that the seventh generation of Pokémon would be the most effective in battles, as they possess the highest average stats compared to other generations.

In our second dashboard, viewers are first presented with a horizontal bar chart that displays the number of Pokémon for each type. This chart includes data for types such as water, normal, grass, bug, psychic, fire, electric, rock, dark, fighting, ghost, ground, poison, dragon, steel, ice, fairy, and flying. The second visual available is a bubble chart that illustrates the average height and weight for each Pokémon type, providing users with additional insights into the characteristics of different types.



In the second dashboard, we answer our final question: which type of Pokémon is the most common compared to the least common. Upon analyzing the first visual, we found that water-type Pokémon are the most abundant, while flying-type Pokémon are the least represented. This insight helps users understand the distribution of different Pokémon types within the dataset.

Bias and Limitations

The first limitation we encountered in our original dataset was the absence of a generation column. As a result, we had to import a secondary dataset solely to provide this information. Another limitation pertains to the "catch rate" column. At first glance, it appears that each Pokémon has a straightforward, predetermined catch rate. However, the catch rate is actually determined by a formula, rather than a single static value.

External factors within the game, such as weather conditions, the type of Poké Ball used, and other variables, all influence the catch rate. Consequently, the catch rate provided in the dataset is unrealistic and unreliable.

Conclusion and Reflection

Our analysis using the dashboards successfully revealed important insights into Pokémon statistics and the distribution of types. The visuals enabled users to easily compare average stats across different generations, clearly identifying that the seventh generation excels in attack, defense, special attack, and special defense, while the fourth generation has the highest average speed. Additionally, the dashboards highlighted the prevalence of various types, showing that water-type Pokémon are the most abundant while flying-type Pokémon are the least represented. However, we encountered notable challenges with the catch rate data. At first glance, it appeared that each Pokémon had a clear and predetermined catch rate; however, we soon learned that these values are derived from a complex formula influenced by multiple external factors. This complexity rendered the catch rate information in our dataset unrealistic and unreliable, which ultimately impacted our predictive capabilities. This experience highlights the critical need for accurate data and thorough verification processes in future projects, as it is essential to ensure that our analyses are based on trustworthy information to drive meaningful conclusions.

In conclusion, our project aimed to develop a machine learning model for predicting Pokémon data, and we chose the Random Forest Regressor for its effectiveness with complex datasets. However, we faced challenges during deployment, including version incompatibility and data type mismatches, which underscored the need to align all dependencies before going live. Despite these issues, our model demonstrated reasonable accuracy with an RMSE of 66%, validating our choice of algorithm.

Reflecting on this experience, it's clear that careful testing, validation, and management of library versions are essential for success. The challenges we encountered highlighted the importance of diligence in both development and deployment phases. Going forward, we will focus on maintaining a consistent environment and checking data compatibility to avoid similar problems. Overall, this project deepened our understanding of the complexities involved in moving machine learning models from development to deployment.