Machine Learning

We chose pokemon battles for our project due to the high number of numeric fields such as attack, defense, speed, etc., additionally the effectiveness of the types of pokemon against the others was listed in a decimal. This lent itself to creating a good machine learning model, however it still took some development. When we initially sourced our data there were also the categorical fields for types of pokemon included which seemed at the start to be a contributing factor to who would win in a battle, so we set out to dummify that data which led to 36 new columns 18 for each of the 2 types a pokemon can possibly have. However upon testing the various models we found a few things that seemed like noise by checking the feature importance on our models, so we removed columns such as experience scale, steps to hatch egg of the pokemon, the pokemon's generation it came from, if the pokemon was legendary, and all 36 of our dummified type columns. I believe the type's importance was low because we already had the effectiveness conveyed by decimal. Once our data was cleaned to its final state we tried a linear regression, which failed abysmally due to the fact there is no 1-1 correlation, as well as a random forest, light gradient boost machine, and XGBoost model. With the returned scores we landed on the XGBoost model as it had the least over-fitting and best ROC curve with a value over .9. However with our model there are some limitations that have to do with the intangibles of pokemon battles, some pokemon have move-sets that take advantage of being high speed to limit their opponents options and these kinds of pokemon are not represented well with our data and will lose to most pokemon even if they obviously shouldn't. Additionally we have an upper limit for our input of pokemon IDs in the battle page of 801, if you input anything beyond that it breaks the model, however even with a max tag on the input element in HTML it does not stop users from typing in something like 999. This is unable to be blocked without some further work in jQuery.