20/09/2024

Pokemon Battle Analysis

DATA1030 Final Presentation

Taemin Huh

Brown University DSI

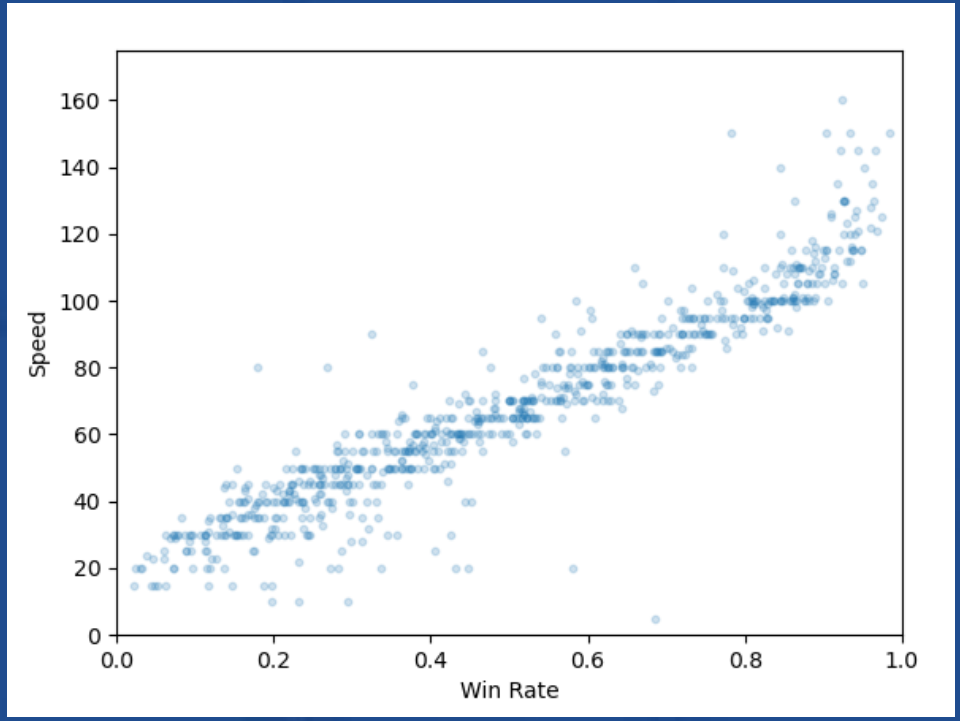
https://github.com/taemin-huh/data1030-project/

Introduction

- Target variable: Win Rate (regression problem)
- 12 features: pkmn dataset (800 Pokemon info datapoints – Kaggle)
  - ID: Pokedex Number, Pokemon Name
  - Type: Type 1, Type 2
  - 6 stats: HP, Attack, Defense, Sp. Atk, Sp. Def, Speed
  - Class: Generation, Legendary
- 3 features: battle dataset (50,000 Pokemon battle datapoints – Kaggle)
  - First Pokemon, Second Pokemon, Winner

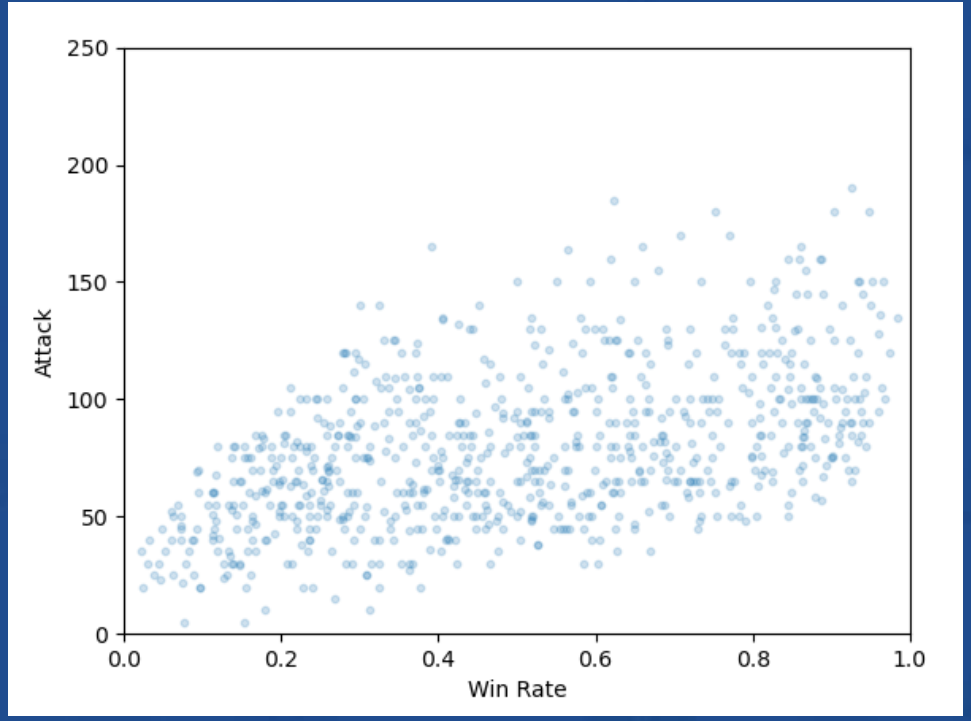
Exploratory Data Analysis

Scatter Plot: Speed vs. Win Rate



Strong correlation (~0.94)

Scatter Plot: Attack vs. Win Rate



Some correlation (~0.50)

Pre-Processing

- Basic split (IID, large # of datapoints)
  - 60%/20%/20% for train/test/split
- Pre-processors
  - OneHotEncoder: Type1 (18), Type2 (19), Generation (6), Legendary(2)
  - MinMaxScaler: HP, Attack, Defense, Sp. Atk, Sp. Def, Speed (0-255 each)
- 51 features after pre-processing (15-5+17+18+5+1)

```
X_train shape: (469, 15)
X_train_prep shape: (469, 51)
```

Cross-Validation

- Attempted supervised ML algorithms
  - Multiple linear regression, SVM, decision tree, random forest, XGBoost
- GridSearchCV w/ 5-fold CV to tune hyperparameters
- Tuned hyperparameters
  - Linear regression: fit_intercept
  - SVM: kernel, C, gamma
  - XGBoost: n_estimators, learning_rate, max_depth, subsample, colsample_bytree
  - Decision tree: max_depth, min_samples_split
  - Random forest: n_estimators, max_depth
- Repeated 5-fold CV w/ 10 repeats for uncertainty estimation

Results

Table: Performance & Uncertainty Evaluation

	Model	Test MAE	Std Dev from Baseline	Splitting Uncertainty	Non-Deterministic Uncertainty
0	LinearRegression	0.048137	41.005280	0.004128	NaN
1	SVM	0.052230	33.929851	0.004868	NaN
2	DecisionTree	0.046020	39.797189	0.004306	NaN
3	RandomForest	0.040605	43.177033	0.004095	0.000253
4	XGBoost	0.038758	44.898587	0.003979	0.000278

Primary Evaluation Metric

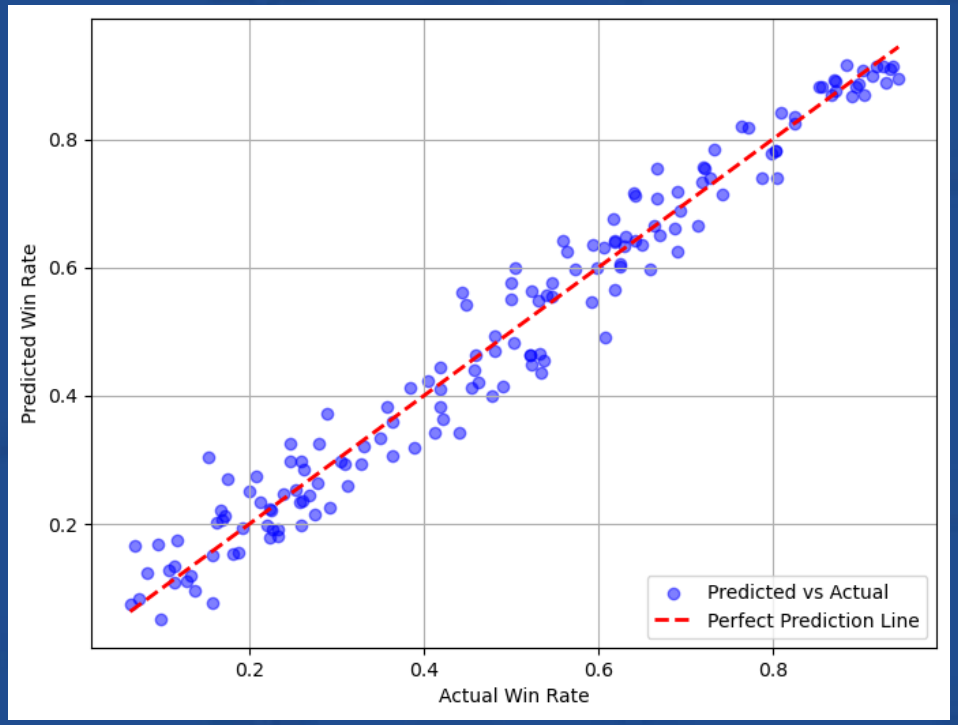
Best Performing Model

Baseline MAE

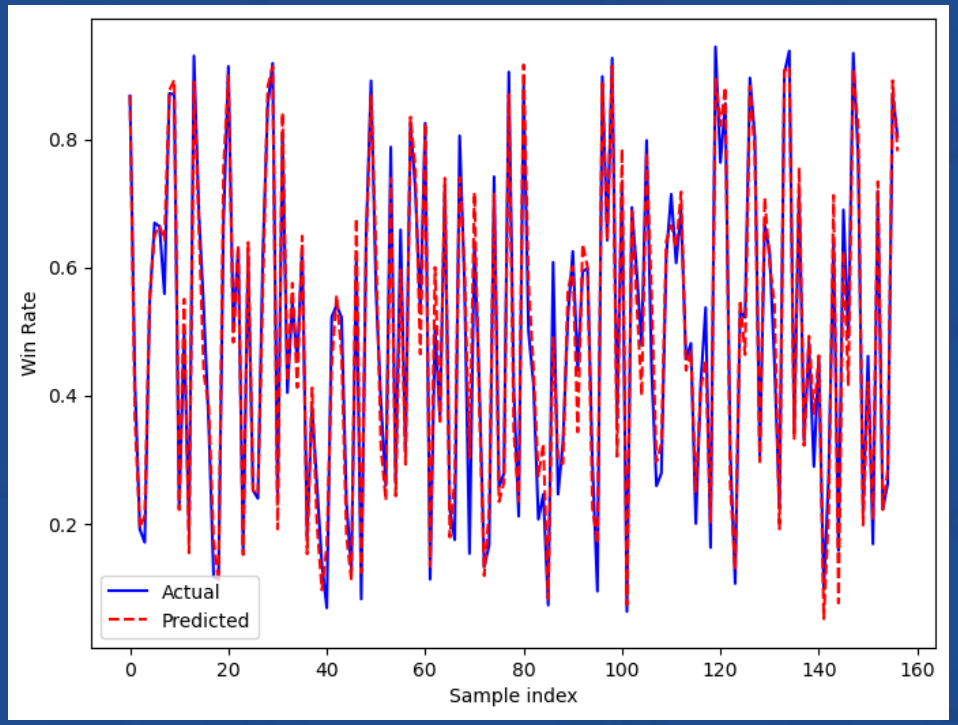
```
Baseline prediction (mean of y_train): 0.4994157900730197
R^2 Score: -0.001114365822402652
MAE: 0.2173976676130286
```

Results (Cont'd)

Scatter Plot: Actual vs. XGBoost Predicted Win Rate



Line Plot: Actual vs. XGBoost Predicted Win Rate



Both plots show meaningfully close predictions vs. actual

Results (Cont'd)

Global Feature Importances

Permutation Importance (Linear Models)

LinearRegression Global Feature Importance:

	Feature	Importance
50	Feature 50	2.184277e-01
46	Feature 46	2.034551e-02
43	Feature 43	7.043117e-03

SVM Global Feature Importance:

	Feature	Importance
50	Feature 50	2.088433e-01
46	Feature 46	2.335917e-02
10	Feature 10	2.566417e-03

Gini Importance (Tree-Based Models)

DecisionTree Global Feature Importance:

	Feature	Importance
50	Feature 50	0.955395
46	Feature 46	0.042320
26	Feature 26	0.001248

RandomForest Global Feature Importance:

	Feature	Importance
50	Feature 50	0.922071
46	Feature 46	0.037326
45	Feature 45	0.010574

XGBoost Global Feature Importance:

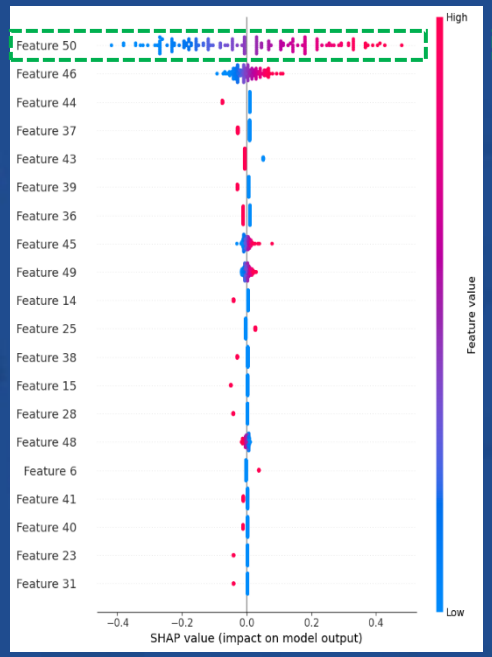
	Feature	Importance
50	Feature 50	0.424124
43	Feature 43	0.111956
25	Feature 25	0.073364

Feature 50 (Speed) has the highest importance score across all regressions

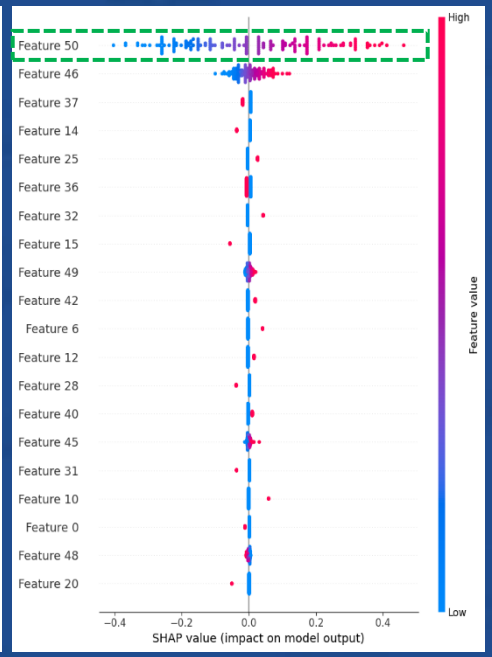
Results (Cont'd)

Global Feature Importances: SHAP Summary Plots

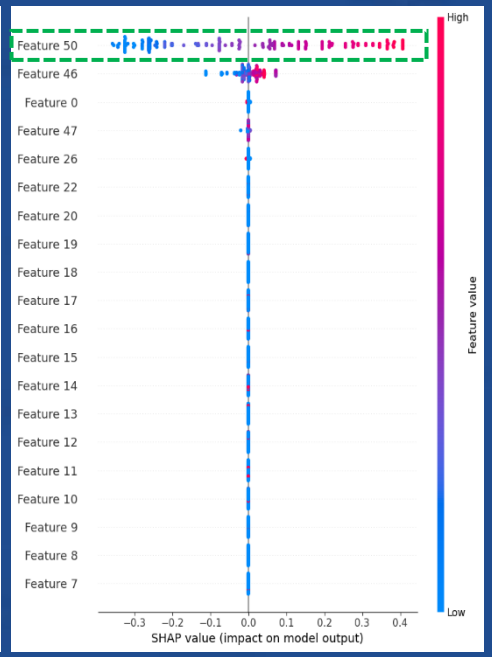
Linear Regression



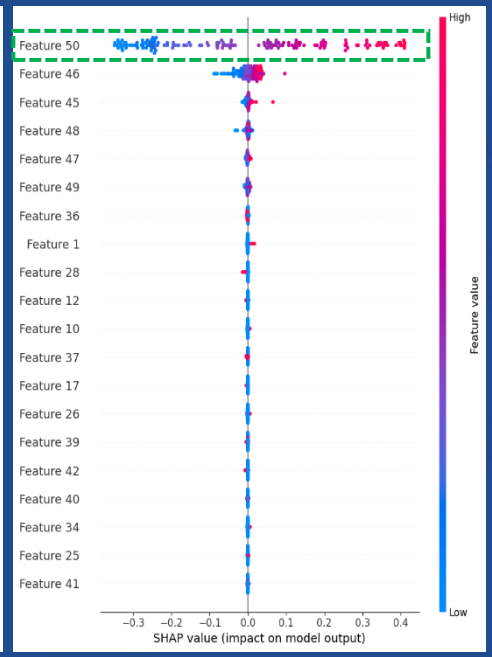
SVM



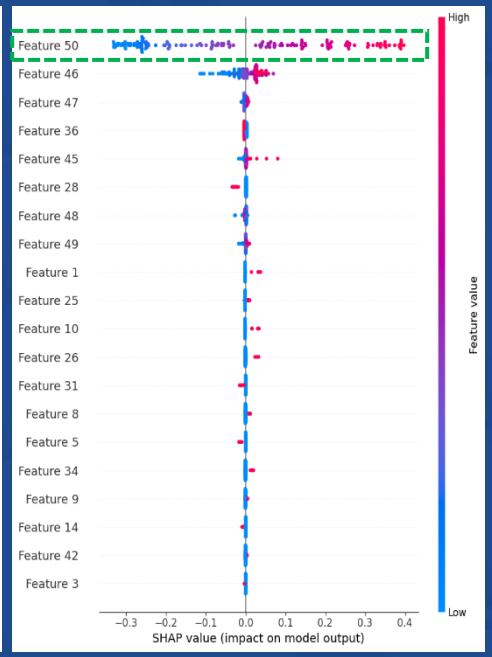
Decision Tree



Random Forest



XGBoost

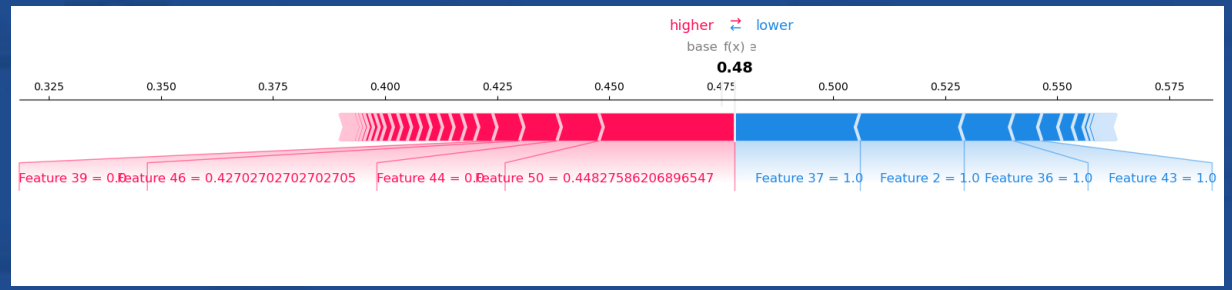


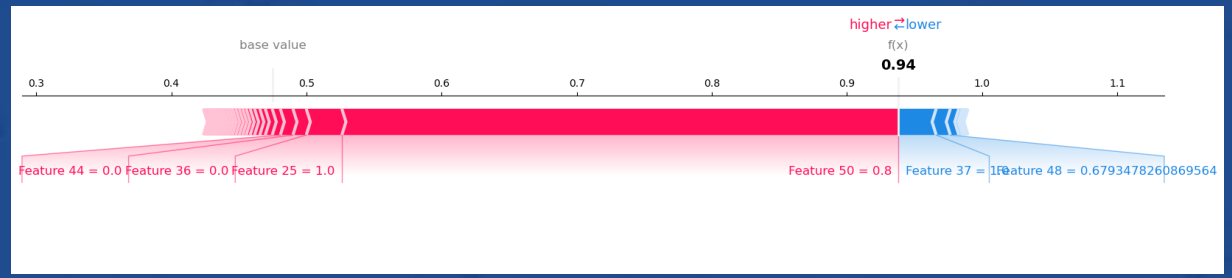
Feature 50 (Speed) has the highest absolute SHAP values across all regressions

Results (Cont'd)

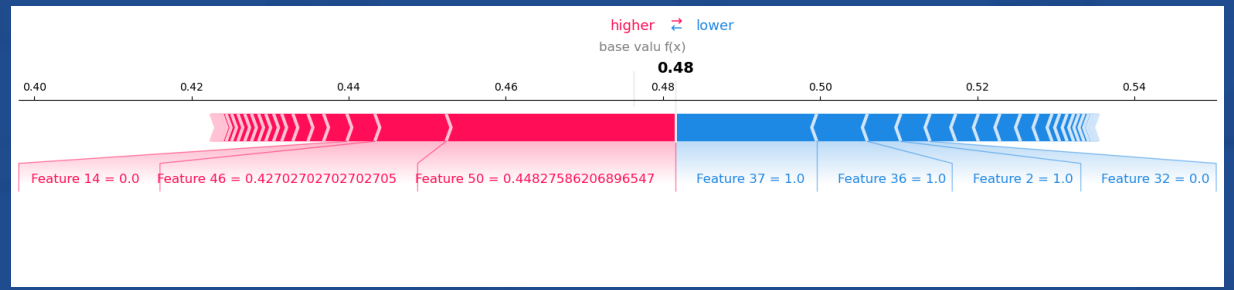
Local Feature Importances: SHAP Force Plots

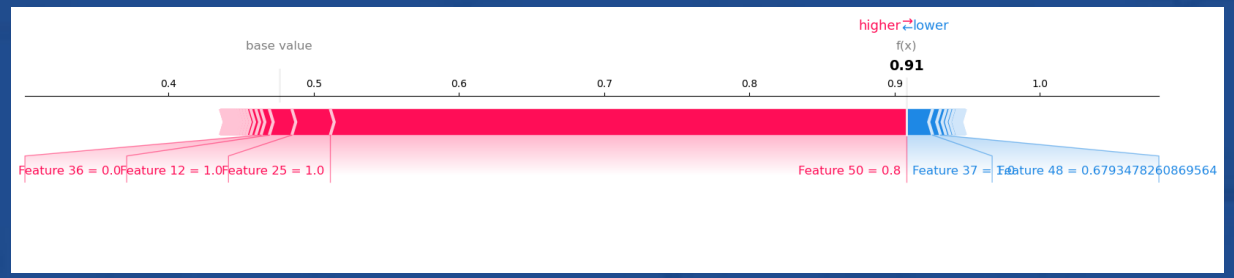
Linear Regression: Typical Case (Above) & Edge Case (Below)





SVM: Typical Case (Above) & Edge Case (Below)

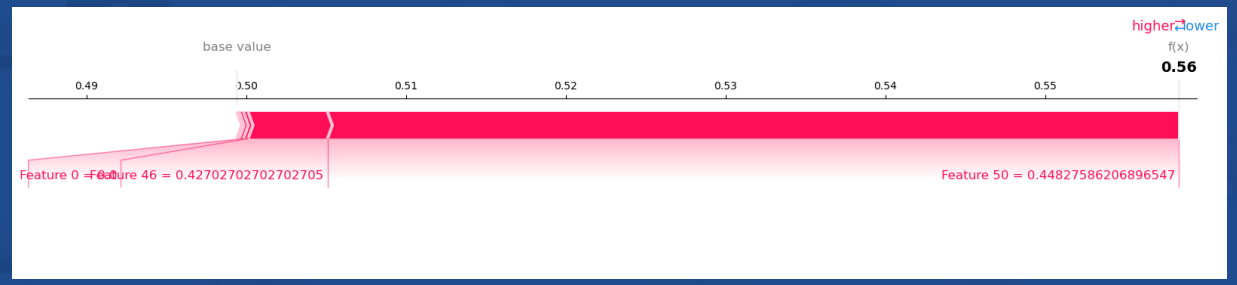


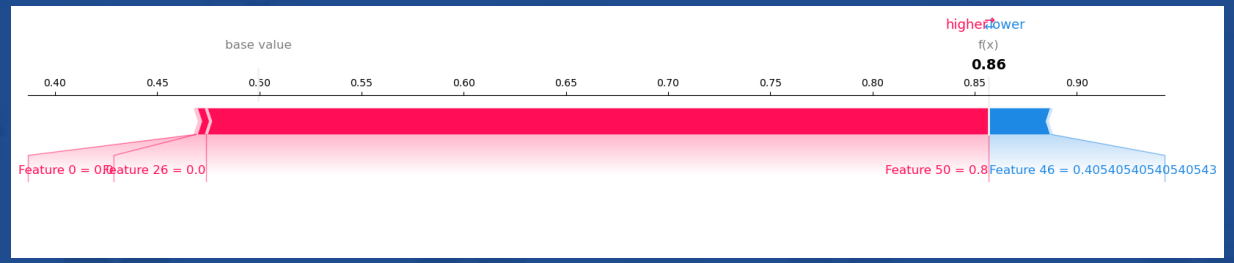


Results (Cont'd)

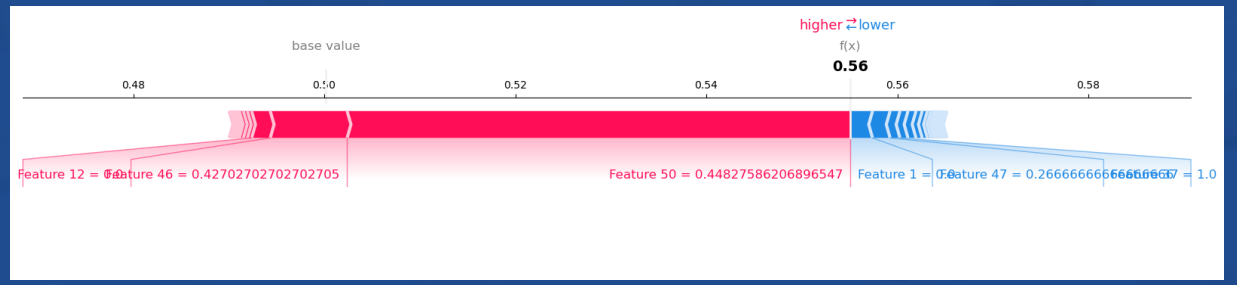
Local Feature Importances: SHAP Force Plots

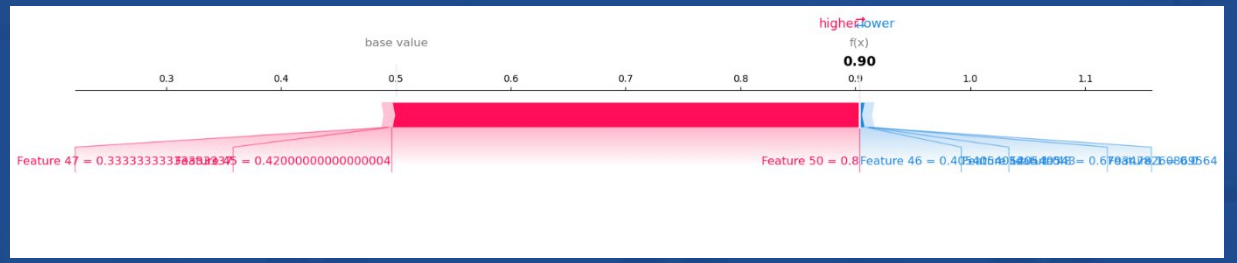
Decision Tree: Typical Case (Above) & Edge Case (Below)





Random Forest: Typical Case (Above) & Edge Case (Below)

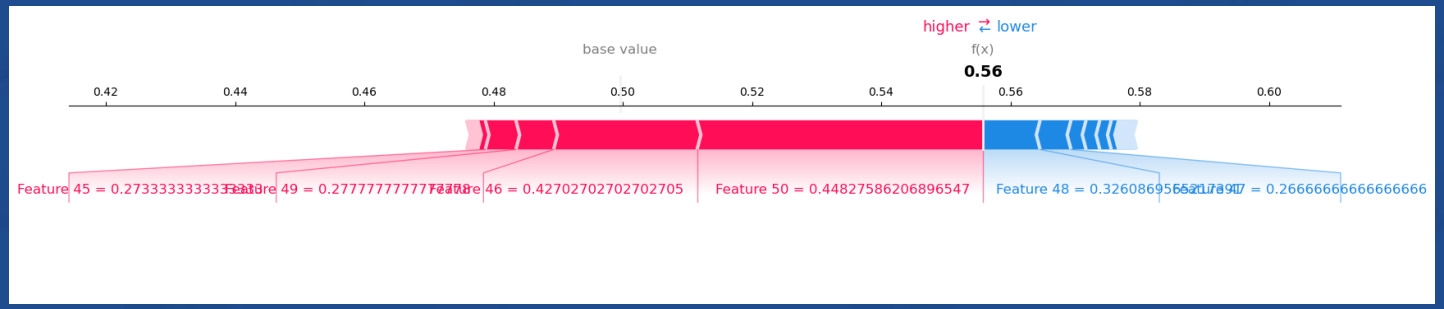


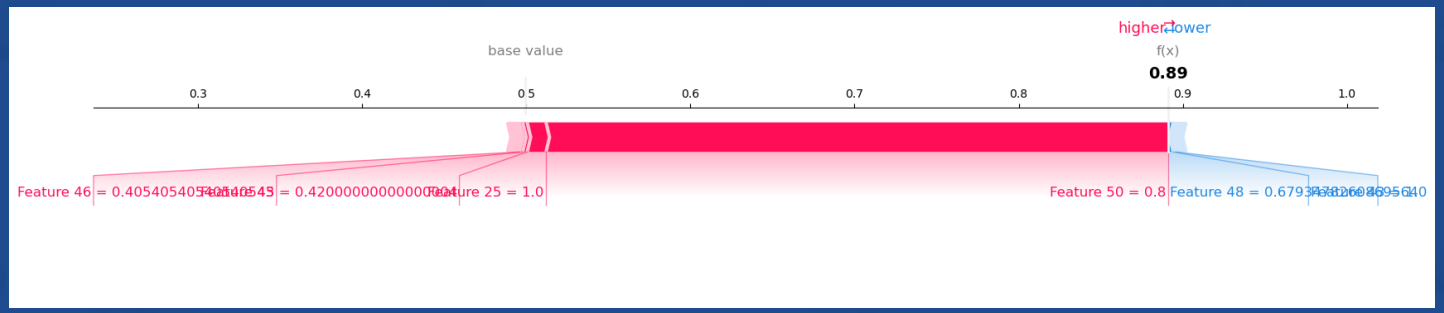


Results (Cont'd)

Local Feature Importances: SHAP Force Plots

XGBoost: Typical Case (Above) & Edge Case (Below)





Outlook

- Additional data
  - 25 Pokemon natures
  - Level
  - Held items & abilities
  - Movesets
  - Competitive play pick rate
- Combining multiple models through stacking
- Creating partial dependence plots (PDP's)
- Implementing neural networks given enough data

Thank You