Pokémon Combat Power Prediction


JIARUI SONG


RAGINI SRINIVASAN


YIYANG SUN

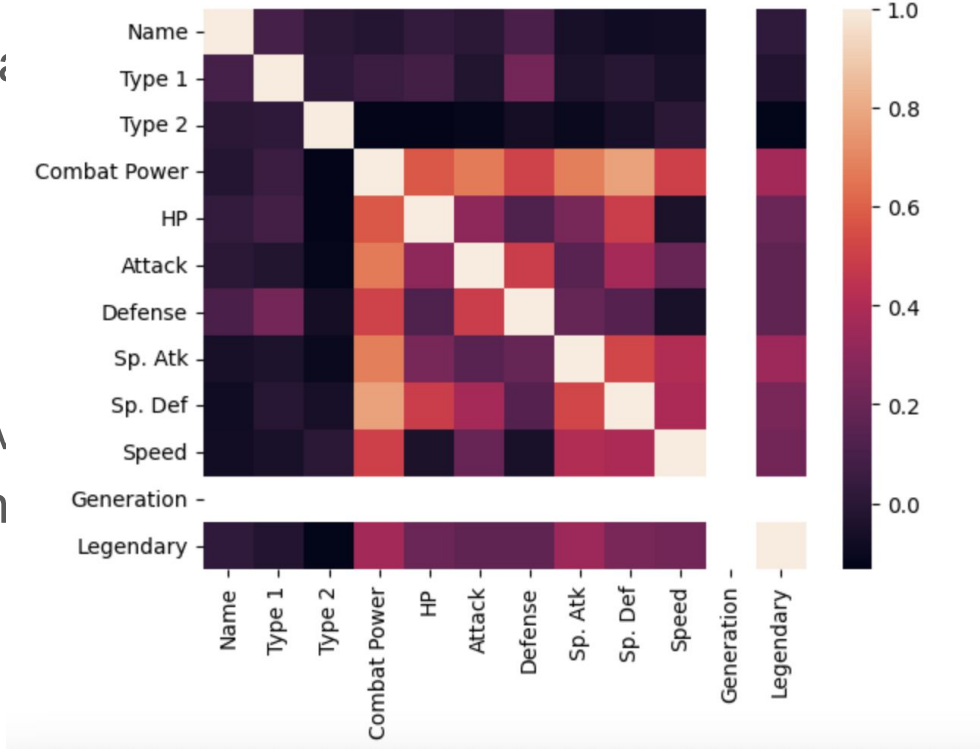

SIYU XIONG

About the Dataset

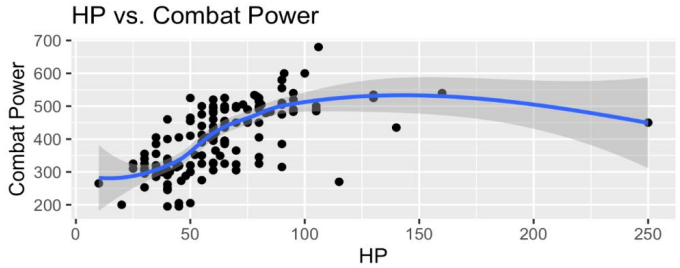
- Response variable: the total combat power of a Pokémon
- 10 predictor variables describing characteristics of a Pokémon that might affect its combat power
  - Attack power, defense power, speed, etc.
- N = 151 observations

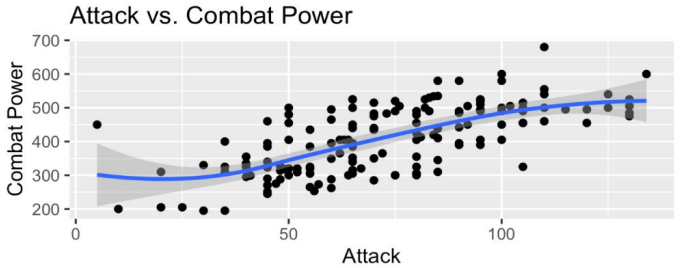
Data Preprocessing

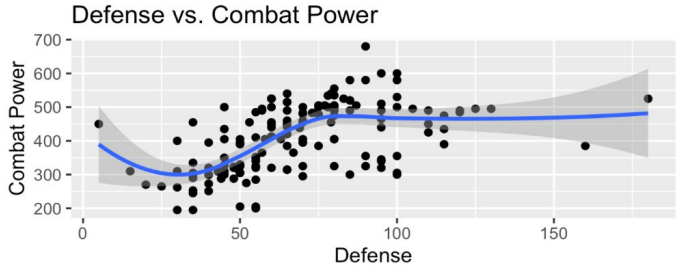
- Features with mostly NA values a removed
- All Variables are centered and scaled
- Based on the correlation matrix, categorical variables were remov as they have low correlation with our response variable Combat Power

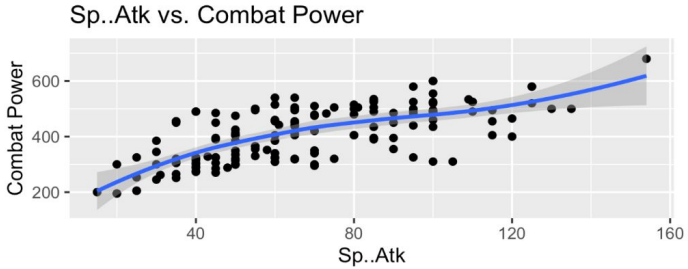


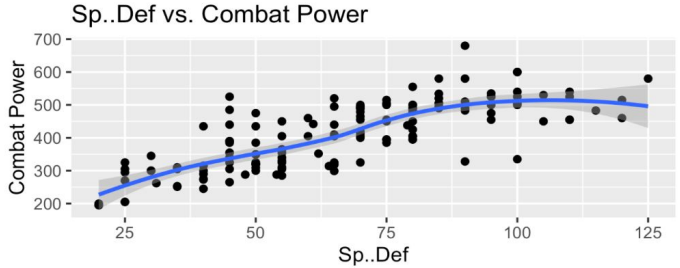
Predictors vs. Y-Variable Scatter Plots (Smooth Line)

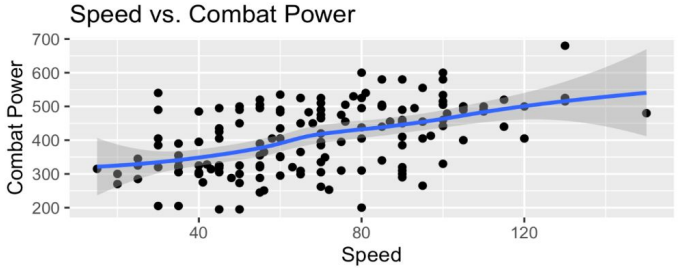




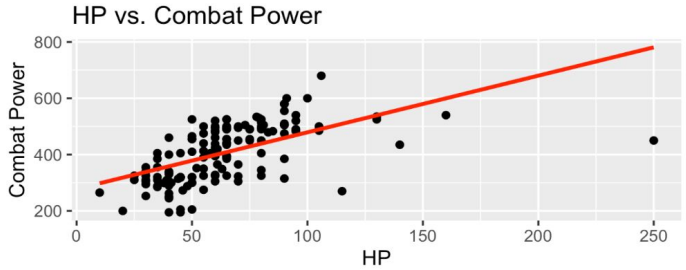


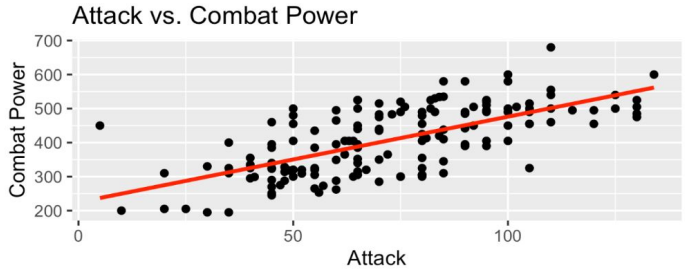


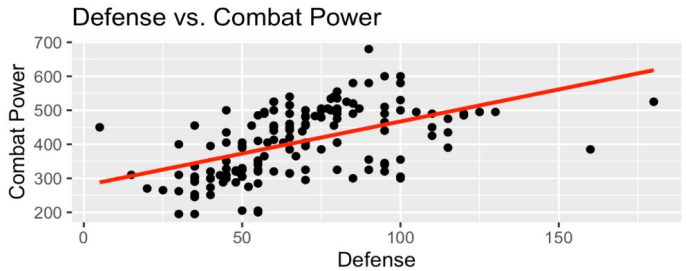


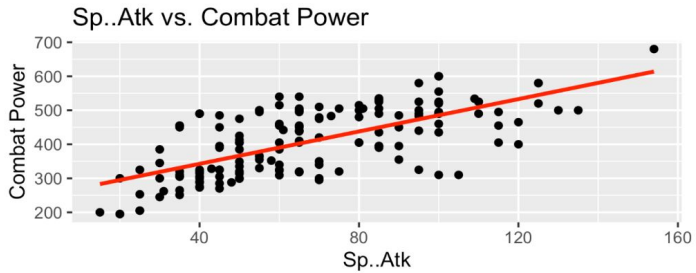


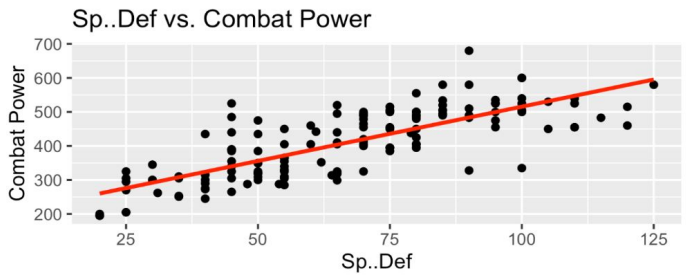
Predictors vs. Y-Variable Scatter Plots (Linear)

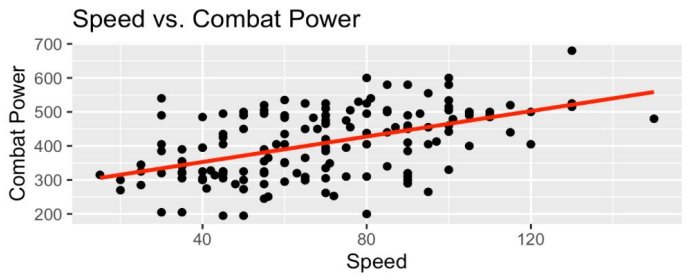




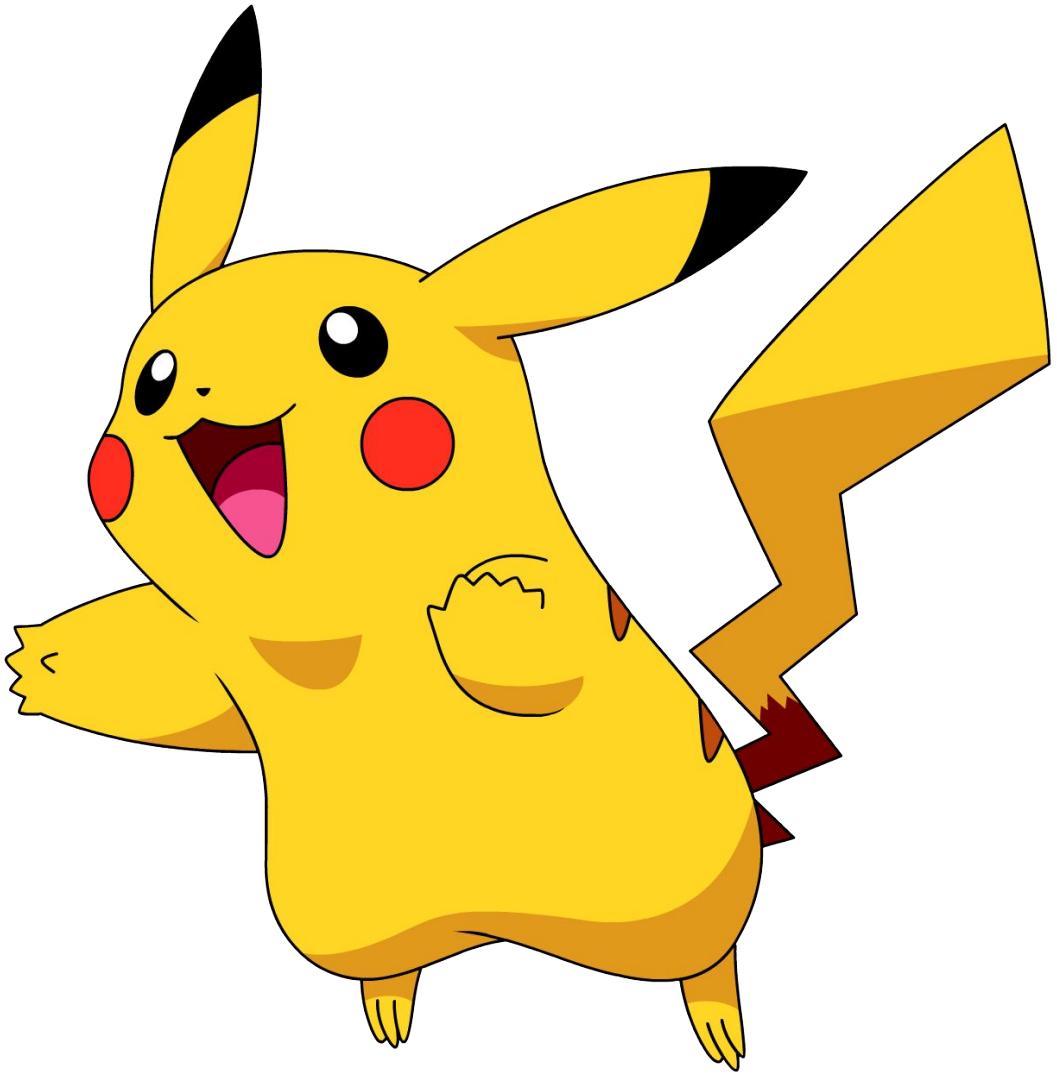




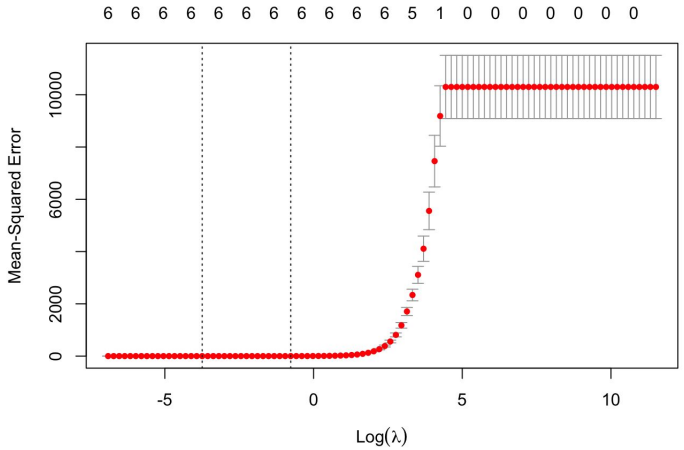




Linear Models



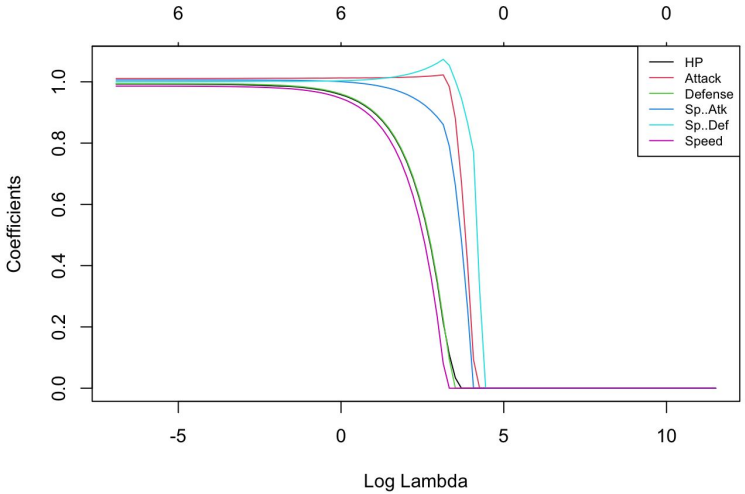
LASSO Model



Optimal λ: 0.001

Test RMSE: 0.1482682

Test R^2: 0.9999891

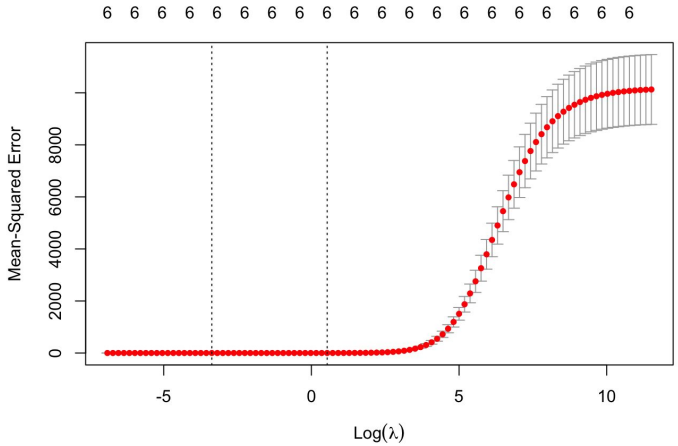


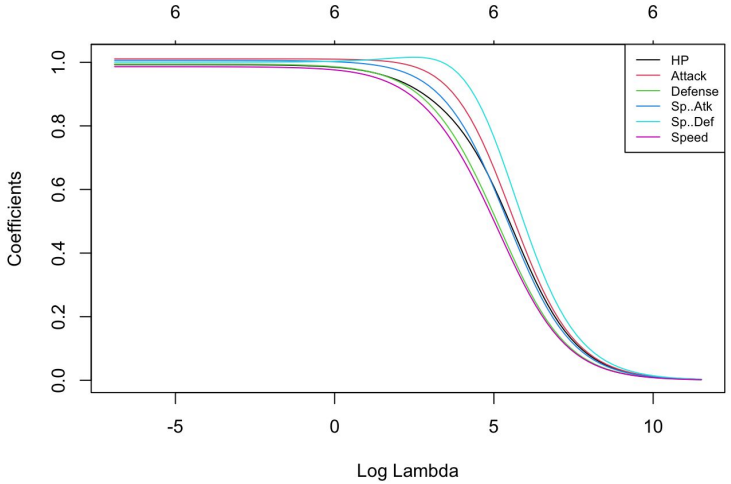
Optimal λ: 0.9

10-fold CV RMSE: 9.350761

10-fold CV R^2: 0.9987542

RIDGE Model





Optimal λ: 0.07220809

Test RMSE: 0.1750829

Test R^2: 0.9999881

Optimal CV λ: 0.0001

10-fold CV RMSE: 0.5156949

10-fold CV R^2: 0.9999525

Elastic Net Model

Optimal α: 0.10

Optimal λ: 1.5396741

Test RMSE: 2.979197

Test R^2: 0.9999459

CV RMSE: 3.063694

CV R^2: 0.9998308

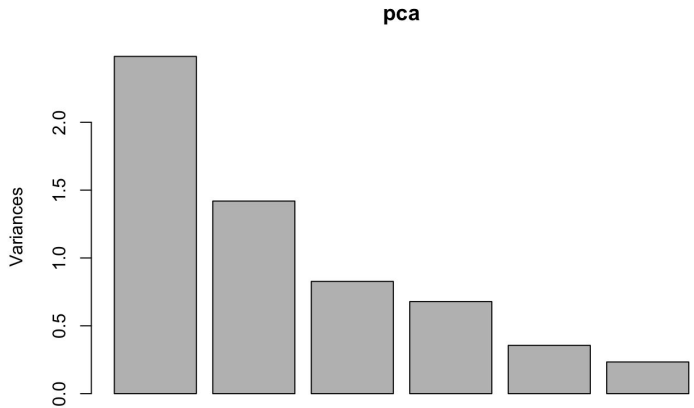
alpha	lambda	RMSE	Rsquared	MAE
0.10	0.1539674	3.063694	0.9998308	2.669048
0.10	1.5396741	3.063694	0.9998308	2.669048
0.10	15.3967405	8.482446	0.9992171	7.361019
0.55	0.1539674	3.080863	0.9998334	2.693164
0.55	1.5396741	3.080863	0.9998334	2.693164
0.55	15.3967405	17.336177	0.9960507	15.145405
1.00	0.1539674	3.098440	0.9998316	2.709611
1.00	1.5396741	3.099146	0.9998316	2.710212
1.00	15.3967405	27.370689	0.9841780	23.938302

PCA Model

	PC1	PC2	PC3	PC4	PC5	PC6
HP	0.4437172	0.2902156	0.5824899	0.1186440	0.3560326	0.48863057
Attack	0.4427849	0.3656014	-0.2945284	0.4961522	0.2099356	-0.54156267
Defense	0.2860453	0.4924948	-0.5227099	-0.5406416	-0.1372336	0.30211743
Sp..Atk	0.4209884	-0.3898564	0.1523431	-0.6001147	0.3269740	-0.42487946
Sp..Def	0.5223266	-0.2030213	0.1784992	0.1156187	-0.8003733	-0.01141514
Speed	0.2741866	-0.5885183	-0.4956435	0.2719519	0.2506499	0.44274397

Importance of components:

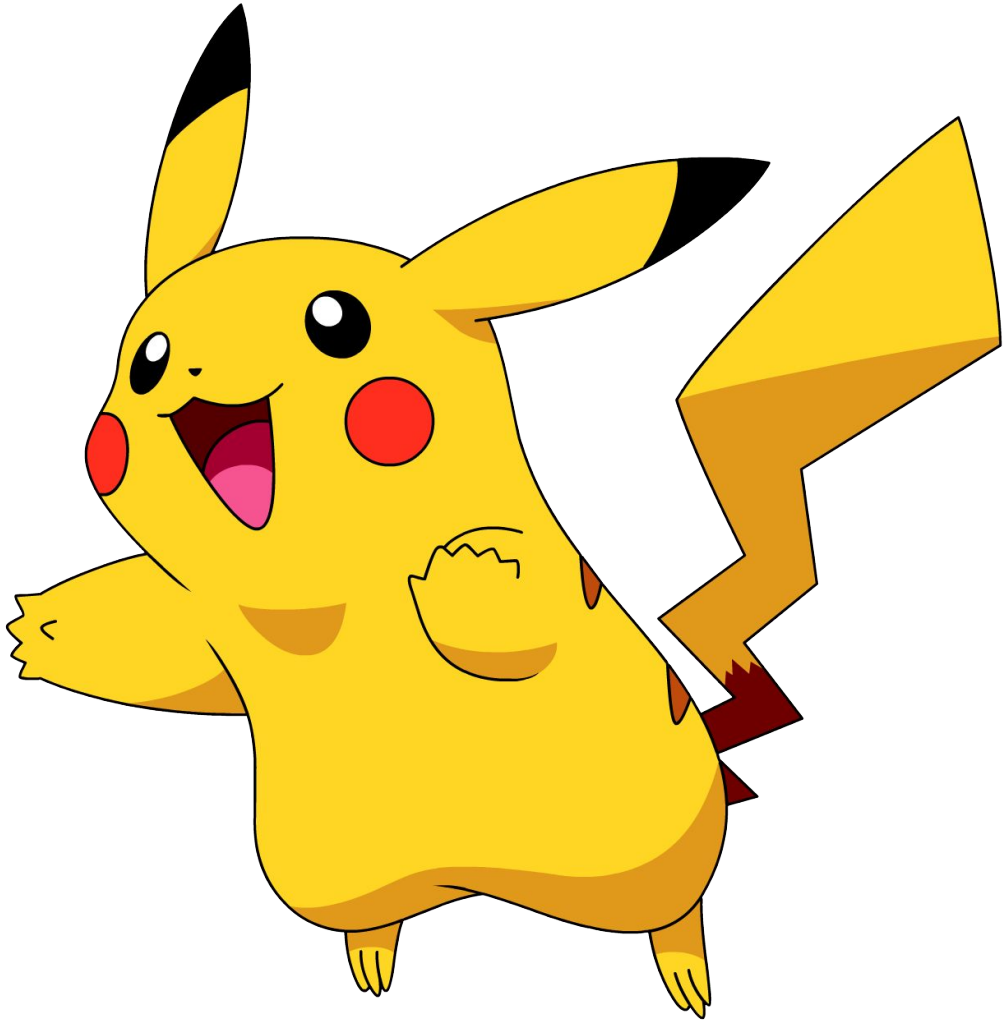
	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	1.5765	1.1914	0.9095	0.8237	0.59652	0.48334
Proportion of Variance	0.4142	0.2366	0.1379	0.1131	0.05931	0.03894
Cumulative Proportion	0.4142	0.6508	0.7887	0.9018	0.96106	1.00000



CV RMSE: 5.672624

CV R^2: 0.9969829

Non-Linear Models



Piecewise Orthogonal Polynomial Model

ANOVA Table for ploy = 1 ~ 5

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)	
1	144	95625					
2	138	85758	6	9866.9	2.6982	0.01671	*
3	132	80450	6	5308.2	1.4516	0.19984	
4	132	80450	0	0.0			
5	132	80450	0	0.0			

Optimal Degree: 2
Predictors Cut: 4
10-fold CV MSE: 10967.69

Comparison: Orthogonal Polynomial Model

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)	
1	144	89.856					
2	138	64.018	6	25.838	176.090	< 2.2e-16	***
3	132	31.660	6	32.358	220.527	< 2.2e-16	***
4	126	11.397	6	20.263	138.092	< 2.2e-16	***
5	120	2.935	6	8.462	57.673	< 2.2e-16	***

10-fold CV MSE Matrix

	Degree	CV_MSE
[1,]	1	0.7125979
[2,]	2	0.8271812
[3,]	3	0.8501690
[4,]	4	2.2138280
[5,]	5	2.4795033

Spline Model

B-Splines (bs()): using quantiles to split the predictors

RMSE	R-Squared	MAE
0.5989	0.9999	0.2473

Natural Cubic Splines (ns()):

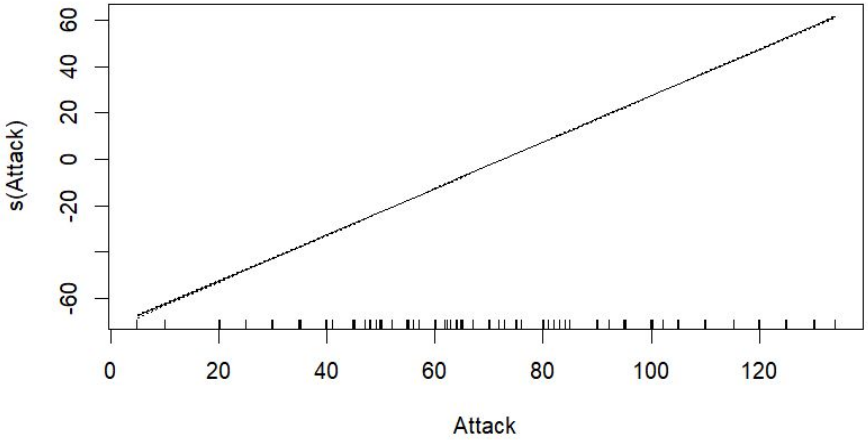
RMSE	R-Squared	MAE
0.7005	0.9999	0.3742

Smooth Splines (bs()):

The optimal smoothing parameter (lambda) is 132.5178.
The 10-fold Cross Validation RMSE is 0.8184

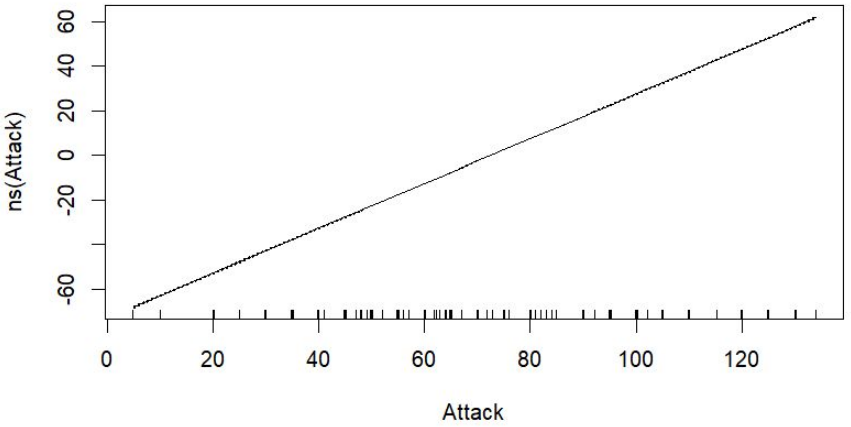
Generalized Additive Model

Smoothing Spline GAM

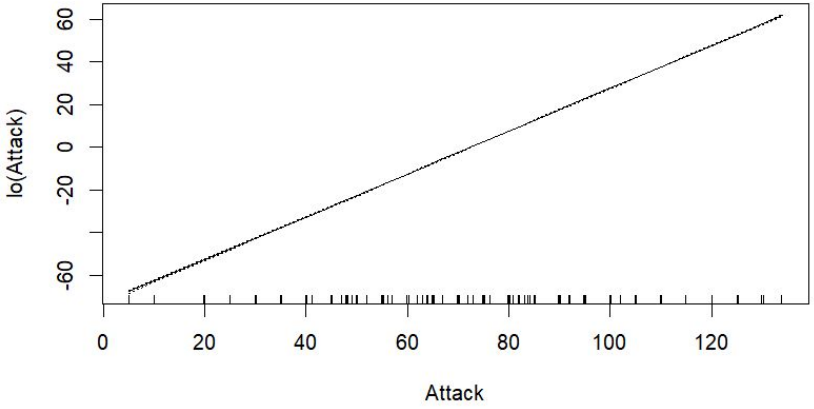


GAM Method	10-fold CV MSE
Smoothing Spline	0.7164
Natural Spline	0.7404
Local Regression	0.7404

Natural Spline GAM



Local Regression GAM



Model Comparison

	LASSO	RIDGE	Elastic Net	PCA	Piecewise Polynomial	Spline (B-Spline)	GAM
CV RMSE	9.350761	0.5156949	3.063694	5.672624	104.7267	0.5989	0.8464

- RIDGE seems to be the most accurate
- Low λ values for the linear models and high λ values for the nonlinear models
- Low α value for elastic net
- In general, a linear fit seems most appropriate


THANK
YOU!