AV-JOB-A-THON-AUGUST-2022

Build a Machine Learning model to predict the CTR(click through rate) of an email campaign based on the email campaigning information.

- Basic exploratory data analysis using pandas, matplotlib, seaborn packages.
- Data pre-processing
  - Numerical feature engineering
    - Groupby numerical summary(min,mean, median, max) of numerical columns.
  - Apply the label encoder to times_of_day column
  - Created number of users by grouping the sender, category, and product columns.
  - Created a body_len_grt_para_len column(whether the body length is greater than the average paragraph length).
- The final features for the model
  - 0_sender
  - 1_subject_len
  - 2_body_len
  - 3_mean_paragraph_len

- 4_day_of_week
- 5_is_weekend
- 6_times_of_day
- 7_category
- 8_product
- 9_no_of_CTA
- 10_mean_CTA_len
- 11_is_image
- 12_is_personalised
- 13_is_quote
- 14_is_emoticons
- 15_is_discount
- 16_is_price
- 17_is_urgency
- 18_target_audience
- 19_user_count
- 20_grp_subject_len_mean
- 21_grp_subject_len_max
- 22_grp_subject_len_median
- 23_grp_subject_len_min
- 24_grp_body_len_mean
- 25_grp_body_len_max
- 26_grp_body_len_median
- 27_grp_body_len_min
- 28_grp_mean_paragraph_len_mean

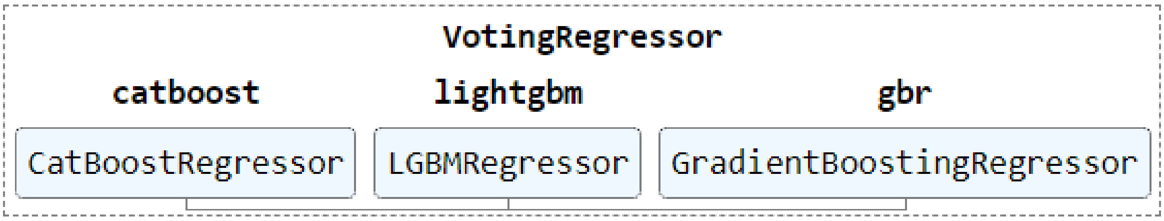
- 29_grp_mean_paragraph_len_max
- 30_grp_mean_paragraph_len_median
- 31_grp_mean_paragraph_len_min
- 32_grp_no_of_CTA_mean
- 33_grp_no_of_CTA_max
- 34_grp_no_of_CTA_median
- 35_grp_no_of_CTA_min
- 36_grp_mean_CTA_len_mean
- 37_grp_mean_CTA_len_max
- 38_grp_mean_CTA_len_median
- 39_grp_mean_CTA_len_min
- 40_body_len_grt_para_len
- 41_click_rate

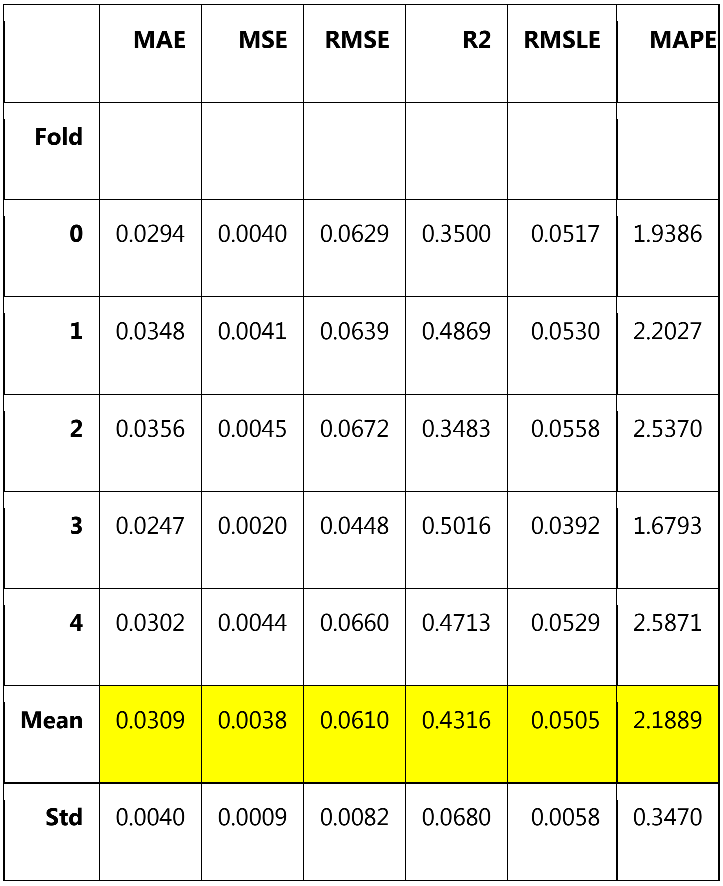
- By using pycaret regressor compared more than one regressor model with 5-fold cross-validation and evaluated by the r2 score.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
catboost	CatBoost Regressor	0.0307	0.0039	0.0614	0.4253	0.0509	2.1820	3.6740
lightgbm	Light Gradient Boosting Machine	0.0333	0.0040	0.0628	0.3976	0.0520	2.3685	0.4340
gbr	Gradient Boosting Regressor	0.0326	0.0041	0.0634	0.3800	0.0527	2.3611	0.6000
rf	Random Forest Regressor	0.0332	0.0043	0.0648	0.3612	0.0539	2.6557	1.4860
et	Extra Trees Regressor	0.0315	0.0043	0.0648	0.3590	0.0539	2.3666	2.1240
xgboost	Extreme Gradient Boosting	0.0335	0.0043	0.0651	0.3515	0.0544	2.6346	0.5000
omp	Orthogonal Matching Pursuit	0.0433	0.0059	0.0767	0.0994	0.0638	3.9964	0.0160
br	Bayesian Ridge	0.0452	0.0063	0.0787	0.0528	0.0658	4.5113	0.0360
knn	K Neighbors Regressor	0.0423	0.0062	0.0785	0.0454	0.0667	3.8856	0.2140
lasso	Lasso Regression	0.0461	0.0063	0.0789	0.0450	0.0662	4.7002	0.1900

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
en	Elastic Net	0.0462	0.0063	0.0790	0.0439	0.0662	4.7024	0.1940
llar	Lasso Least Angle Regression	0.0464	0.0067	0.0811	-0.0049	0.0684	4.6544	0.0160
dummy	Dummy Regressor	0.0464	0.0067	0.0811	-0.0049	0.0684	4.6544	0.0040
ada	AdaBoost Regressor	0.0689	0.0073	0.0849	-0.1114	0.0753	9.0867	0.2560
dt	Decision Tree Regressor	0.0400	0.0080	0.0873	-0.1650	0.0723	2.5887	0.0400
par	Passive Aggressive Regressor	0.0960	0.0143	0.1193	-1.2398	0.1025	12.4774	0.0240
ridge	Ridge Regression	0.0958	0.0190	0.1337	-1.9614	0.1069	11.8459	0.0120
lr	Linear Regression	2.9415	74.0284	5.3367	-17421.1808	0.7114	379.1522	0.0140
lar	Least Angle Regression	479.897	140.00	544.51	-185.00	11.459	129.968	0.0620

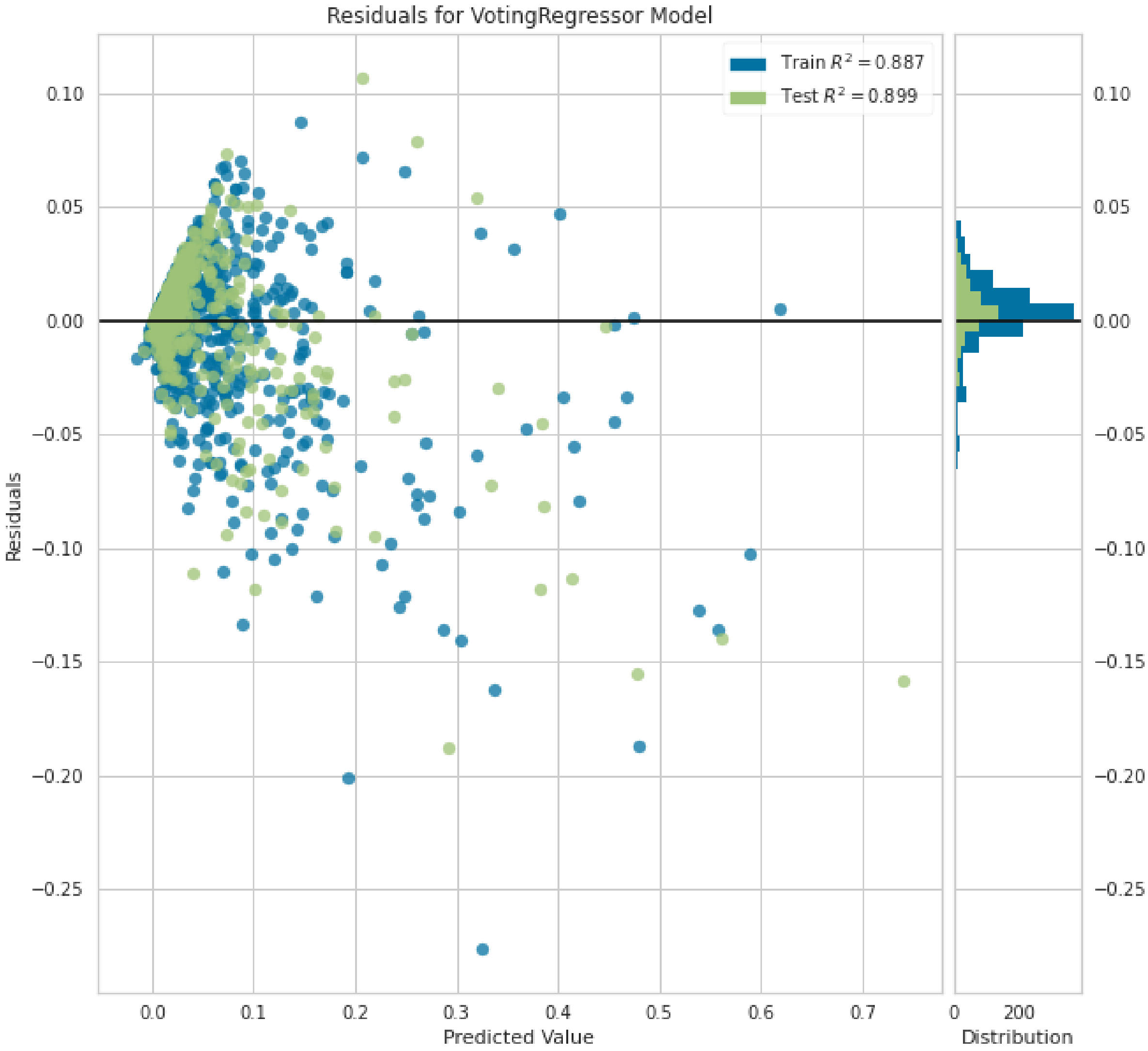
- Blended the top 3 model



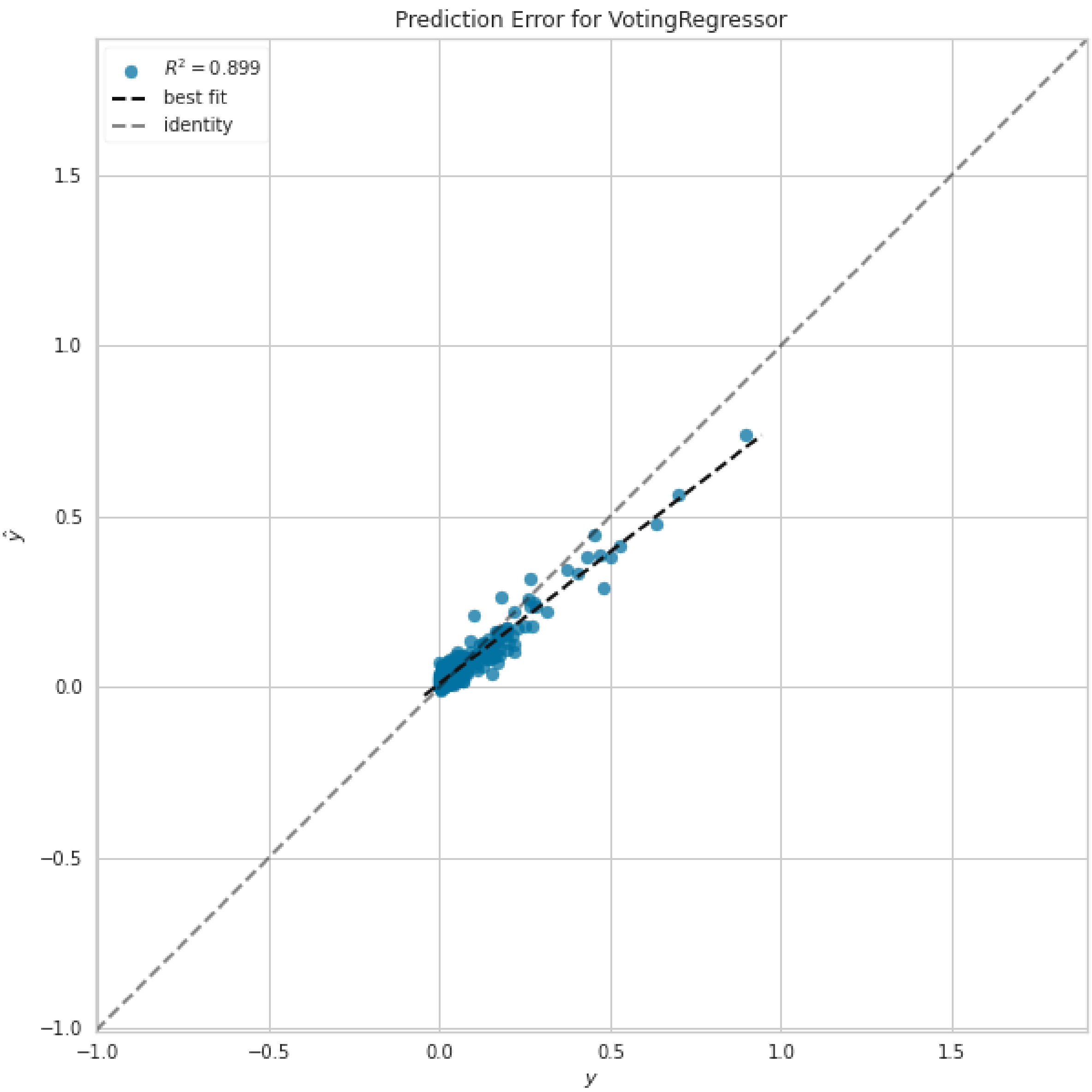


	MAE	MSE	RMSE	R2	RMSLE	MAPE
Fold						
0	0.0294	0.0040	0.0629	0.3500	0.0517	1.9386
1	0.0348	0.0041	0.0639	0.4869	0.0530	2.2027
2	0.0356	0.0045	0.0672	0.3483	0.0558	2.5370
3	0.0247	0.0020	0.0448	0.5016	0.0392	1.6793
4	0.0302	0.0044	0.0660	0.4713	0.0529	2.5871
Mean	0.0309	0.0038	0.0610	0.4316	0.0505	2.1889
Std	0.0040	0.0009	0.0082	0.0680	0.0058	0.3470

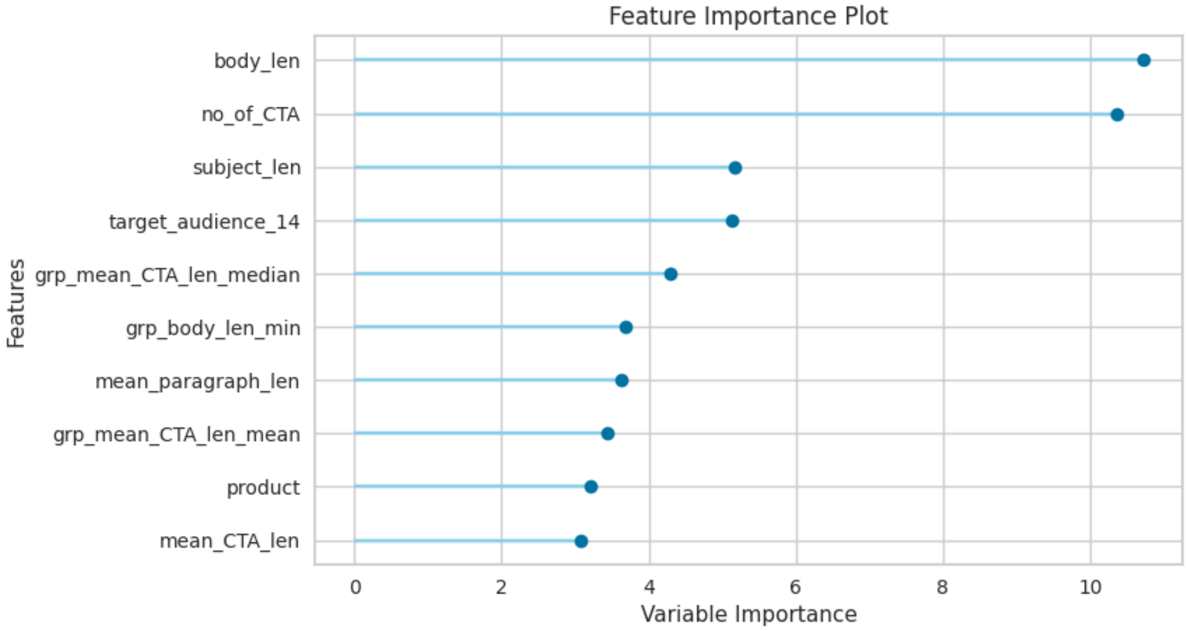
- Voting Regressor Residual Plot



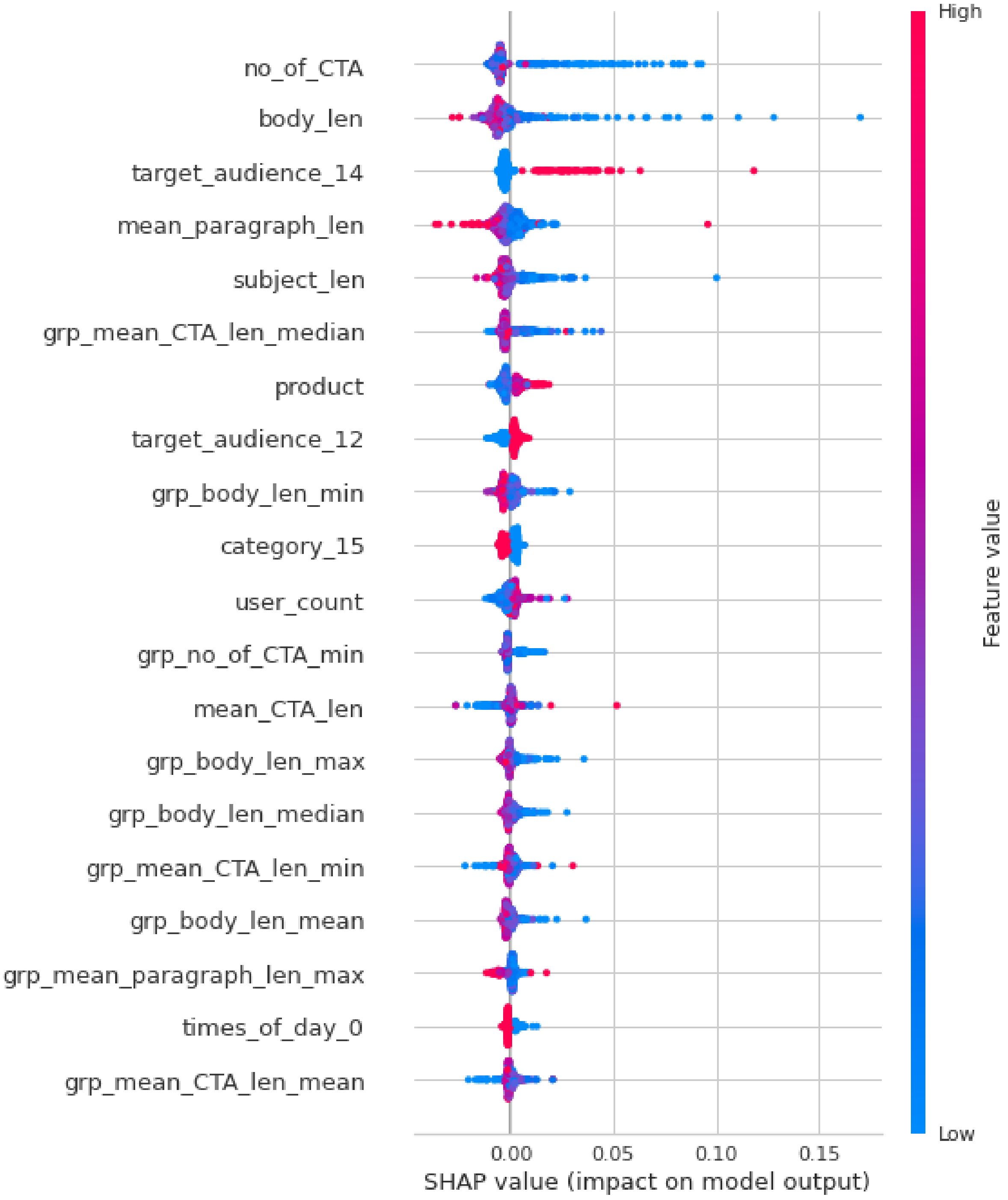
- Voting Regressor Prediction Error Plot



- Catboost Model Feature Importance Plot



- SHAP - Catboost Model Feature Importance Plot


no_of_CTA
body_len
target_audience_14
mean_paragraph_len
subject_len
grp_mean_CTA_len_median
product
target_audience_12
grp_body_len_min
category_15
user_count
grp_no_of_CTA_min
mean_CTA_len
grp_body_len_max
grp_body_len_median
grp_mean_CTA_len_min
grp_body_len_mean
grp_mean_paragraph_len_max
times_of_day_0
grp_mean_CTA_len_mean
SHAP value (impact on model output)
High
Low
Feature value