STAR TREK

THE NEXT GENERATION

PREDICTING IMDB RATINGS USING NLP & MACHINE LEARNING

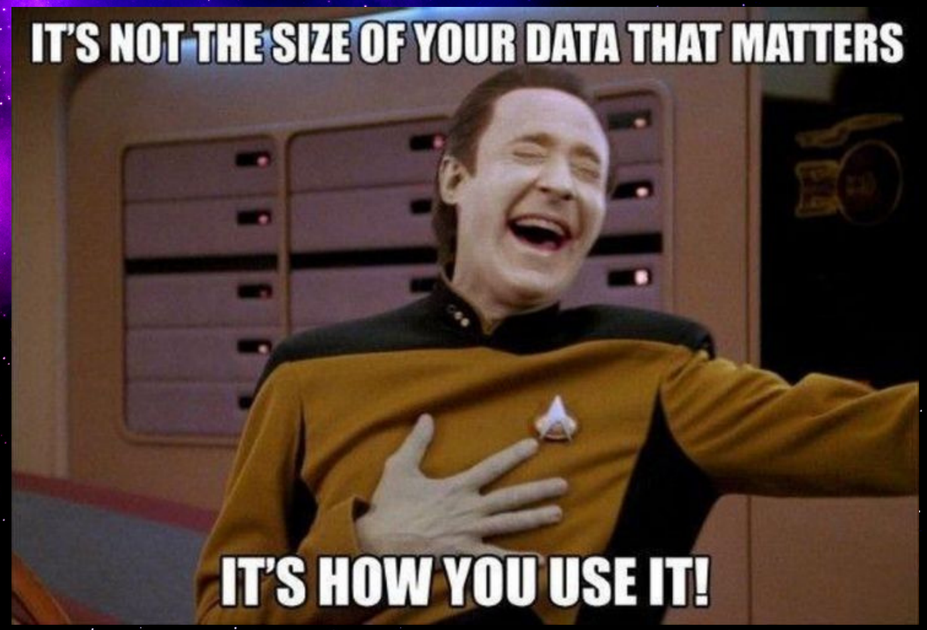
CAPSTONE PROJECT BY KATYA KOGAN

THE MISSION

PREDICT THE IMDB RATINGS FROM THE SCRIPT BY EACH CHARACTER'S LINES

COST AN AVERAGE OF $1.3 MILLION DOLLARS TO PRODUCE AN EPISODE

USING PAST EPISODES' FEATURES TO DETERMINE IF THE NEW EPISODE WRITTEN WILL RECEIVE HIGH RATINGS OR NOT



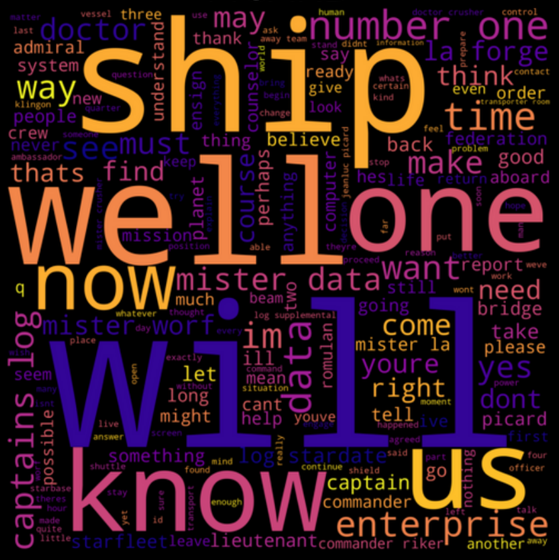
DATA CLEANING & FEATURE ENGINEERING

	Unnamed: 0	episode	productionnumber	setnames	characters	act	scenenumber	scenedetails	partnumber	type	who	text	speechdescript	Released	Episode	imdbRating	imdbID	Season
62465	62466	disaster	#40275-205	USS ENTEF	PICARD,LIE	FOUR	49B		461	description		Patterson closes his	False	1991-10-	5.0	7.8	tt070870	5.0
98609	98610	force of nature	#40277-261	USS ENTEF	PICARD,PR/	THREE	25		323	speech	RABAL	Captain, according t	False	1993-11-	9.0	6.4	tt070871	7.0
10245	10246	when the bough bre	#40271-118	USS ENTEF	PICARD,ALI	FOUR	58		488	speech	BEVERLY	No. That doesn't fit th	False	1988-02-	16.0	6.3	tt070884	1.0
85461	85462	tapestry	#40276-241	USS ENTEF	PICARD,Q,F	TWO	15		123	speech	COREY	Of course it's me... I	True	1993-02-	15.0	8.8	tt070878	6.0
60136	60137	redemption ii	#40275-201	USS ENTEF	PICARD,ADI	FIVE	76		570	description		Data EXITS. Hold or	False	1991-09-	1.0	8.3	tt017050	5.0
3261	3262	where no one has g	#40271-106			FOUR	92		347	description		at the unknown part	False	1987-10-	5.0	7.6	tt070884	1.0
12670	12671	skin of evil	#40271-122	USS ENTEF	PICARD,ARI	THREE	51		391	description		ARMUS pulls the tric	False	1988-04-	22.0	6.8	tt070877	1.0
69560	69561	ethics	#40275-216	USS ENTEF	PICARD,DO	TWO	26		225	speech	WORF	Please proceed.	False	1992-02-	16.0	7.2	tt070870	5.0
75382	75383	the inner light	#40275-225	USS ENTEF	PICARD,ELI	TWO	29		204	speech	ELINE	I think you're still tryi	False	1992-05-	25.0	9.3	tt070880	5.0
62585	62586	disaster	#40275-205	USS ENTEF	PICARD,LIE	FIVE	77		581	speech	KEIKO	Well, I'm sorry...	False	1991-10-	5.0	7.8	tt070870	5.0

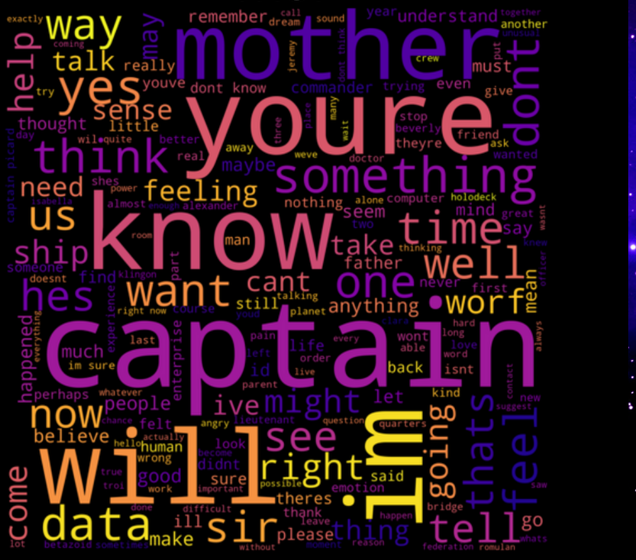
PICARD_PCT	RIKER_PCT	DATA_PCT	WORF_PCT	TROI_PCT	BEVERLY_PCT	GEORDI_PCT	Q_PCT	LORE_PCT	WESLEY_PCT	GUINAN_PCT	TASHA_PCT	PULASKI_PCT	RATING
33.46	31.58	10.44	1.69	0.0	4.26	5.21	6.05	0.0	4.52	0.0	2.79	0.0	8.0
11.47	3.01	26.08	25.44	14.94	4.23	14.82	0.0	0.0	0.0	0.0	0.0	0.0	8.0
27.32	42.75	7.88	8.6	0.0	0.0	0.0	0.0	0.0	10.2	0.0	0.0	3.25	8.0
39.13	24.34	11.6	1.94	3.09	0.92	16.36	0.0	0.0	2.63	0.0	0.0	0.0	7.0
41.64	9.67	18.77	5.32	2.31	5.44	16.85	0.0	0.0	0.0	0.0	0.0	0.0	7.0
39.27	6.1	14.12	4.62	4.9	11.55	7.46	9.44	0.0	0.0	0.0	2.54	0.0	8.0
42.47	18.29	8.61	2.45	3.18	9.29	2.8	10.24	0.0	2.67	0.0	0.0	0.0	7.0
12.8	26.11	18.57	2.77	10.22	13.69	6.21	0.0	0.0	2.23	0.0	7.39	0.0	6.0
8.72	15.11	0.94	4.65	0.03	12.74	25.84	31.97	0.0	0.0	0.0	0.0	0.0	6.0
30.28	26.48	3.75	3.8	1.84	33.85	0.0	0.0	0.0	0.0	0.0	0.0	0.0	7.0
8.81	1.26	41.32	29.01	3.88	2.45	13.28	0.0	0.0	0.0	0.0	0.0	0.0	8.0
9.55	1.78	1.53	67.03	1.04	0.59	1.53	16.93	0.0	0.0	0.0	0.0	0.0	7.0
53.02	4.1	15.48	3.45	3.0	10.6	10.35	0.0	0.0	0.0	0.0	0.0	0.0	6.0
24.63	13.12	12.72	3.24	0.4	1.74	39.52	0.0	0.0	2.28	2.36	0.0	0.0	7.0
15.48	18.46	22.96	4.92	0.9	8.86	10.66	0.0	15.06	2.7	0.0	0.0	0.0	8.0

EDA & VISUALISATIONS

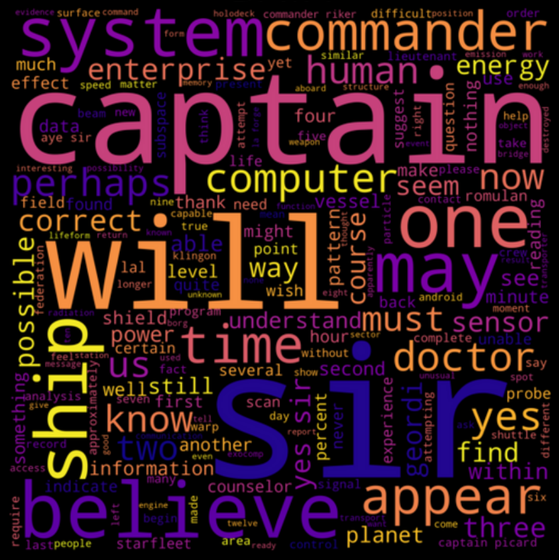
PICARD



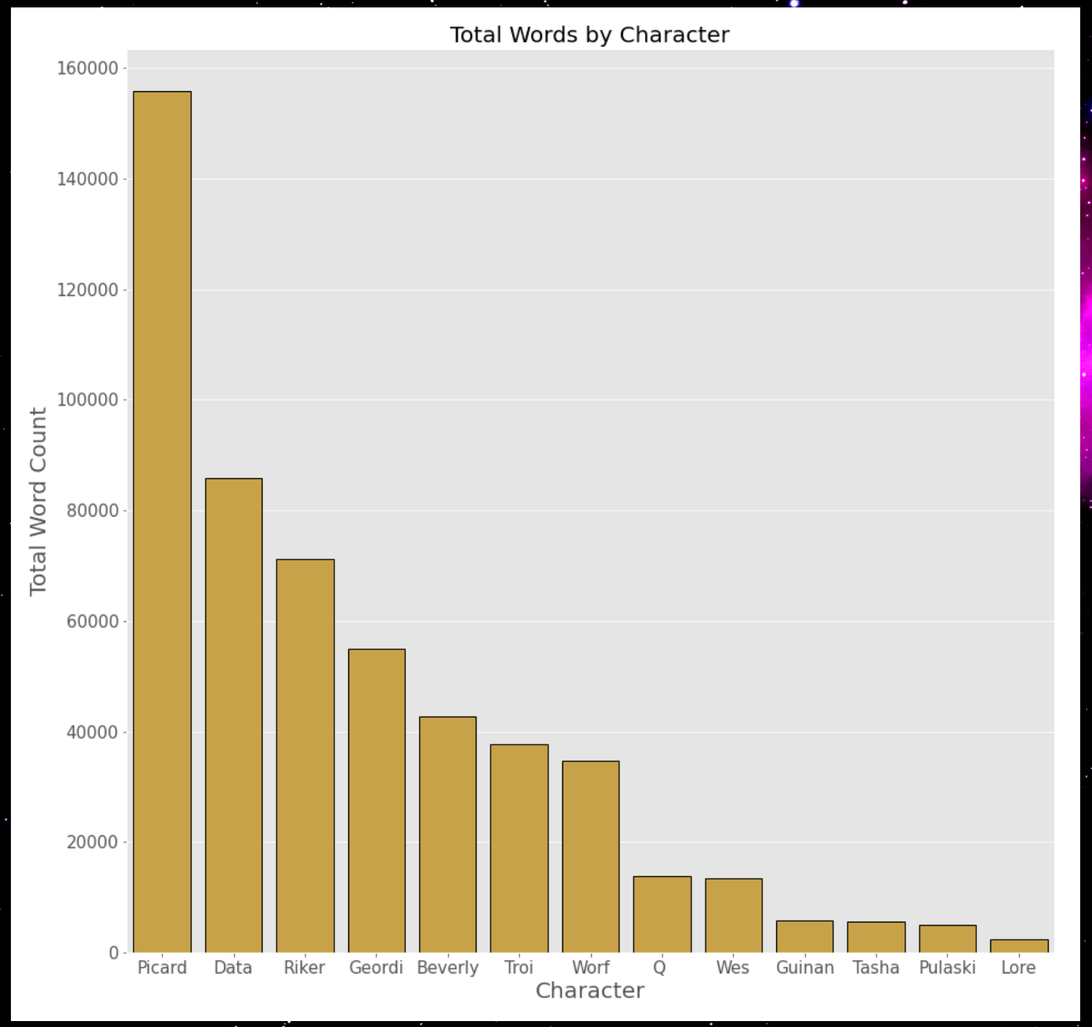
TROI

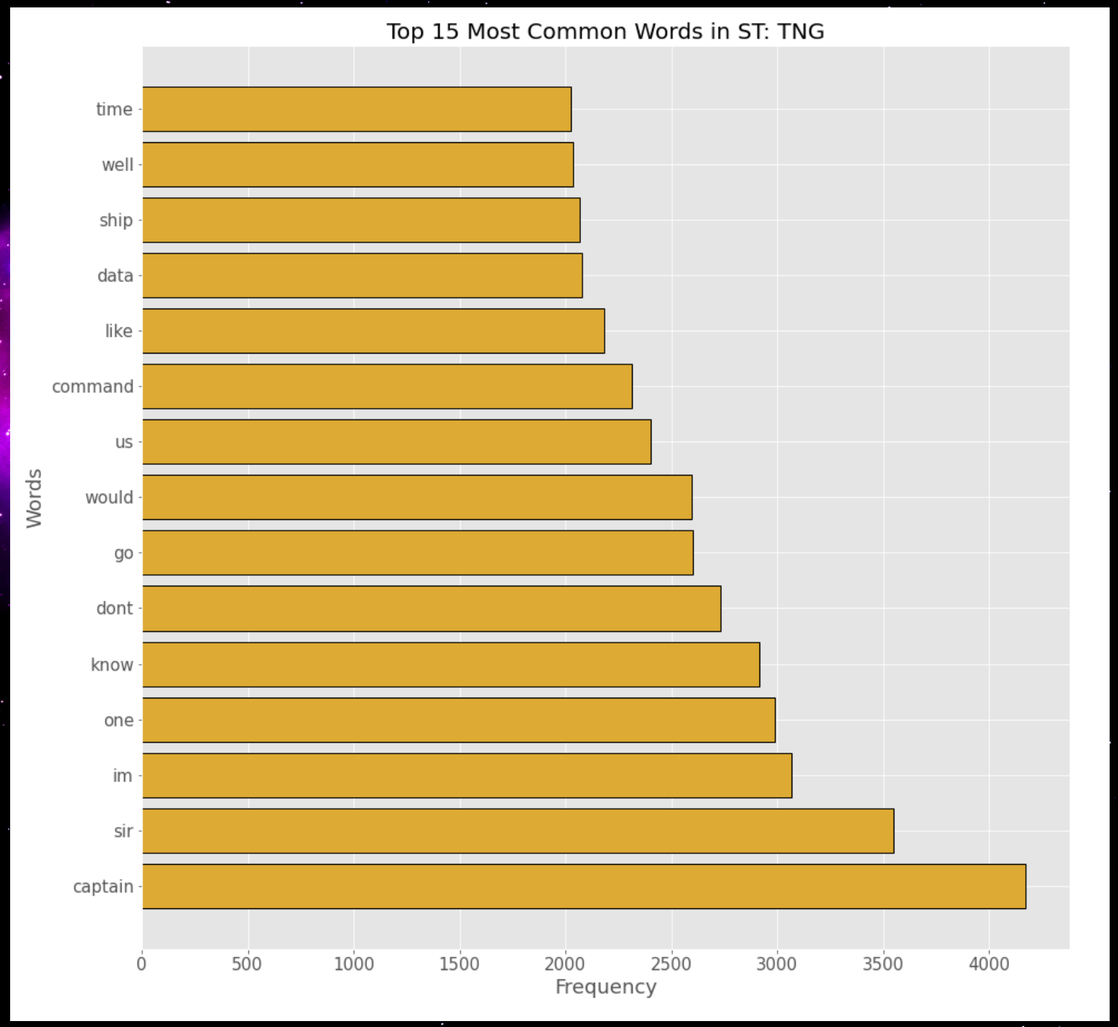


DATA

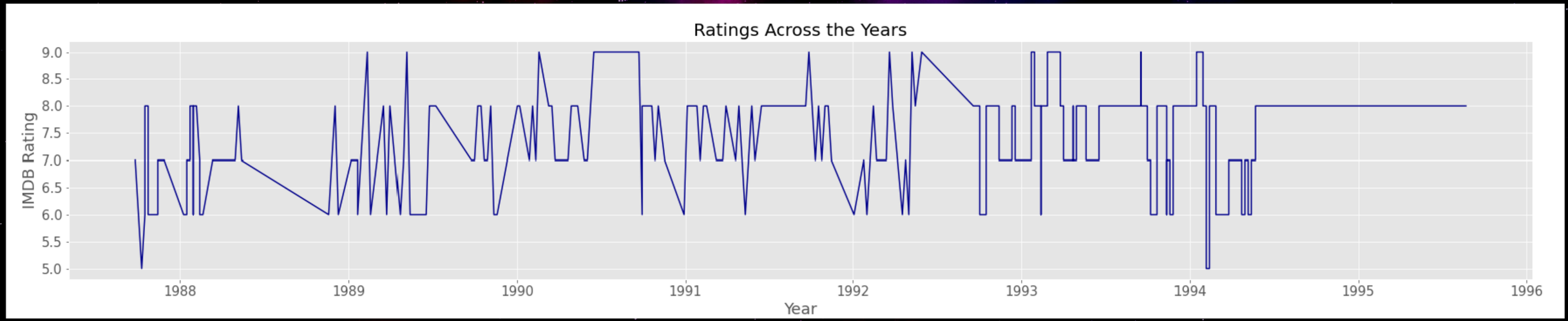


EDA & VISUALISATIONS

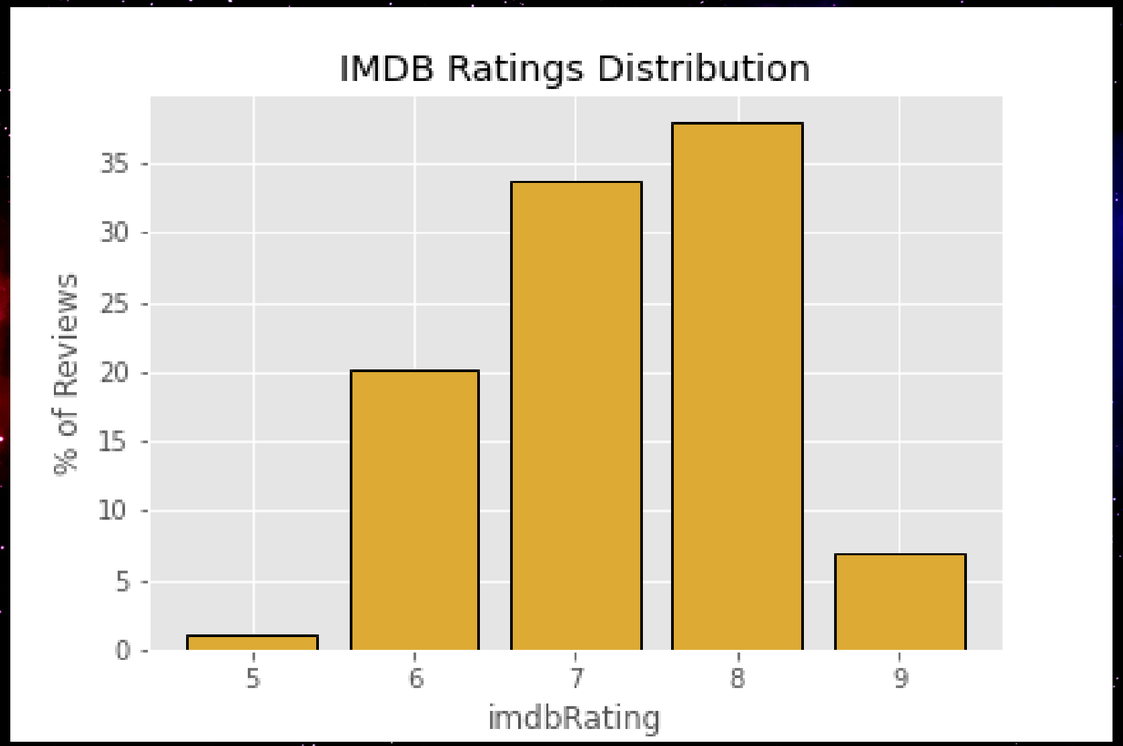




EDA & VISUALISATIONS



EDA & VISUALISATIONS

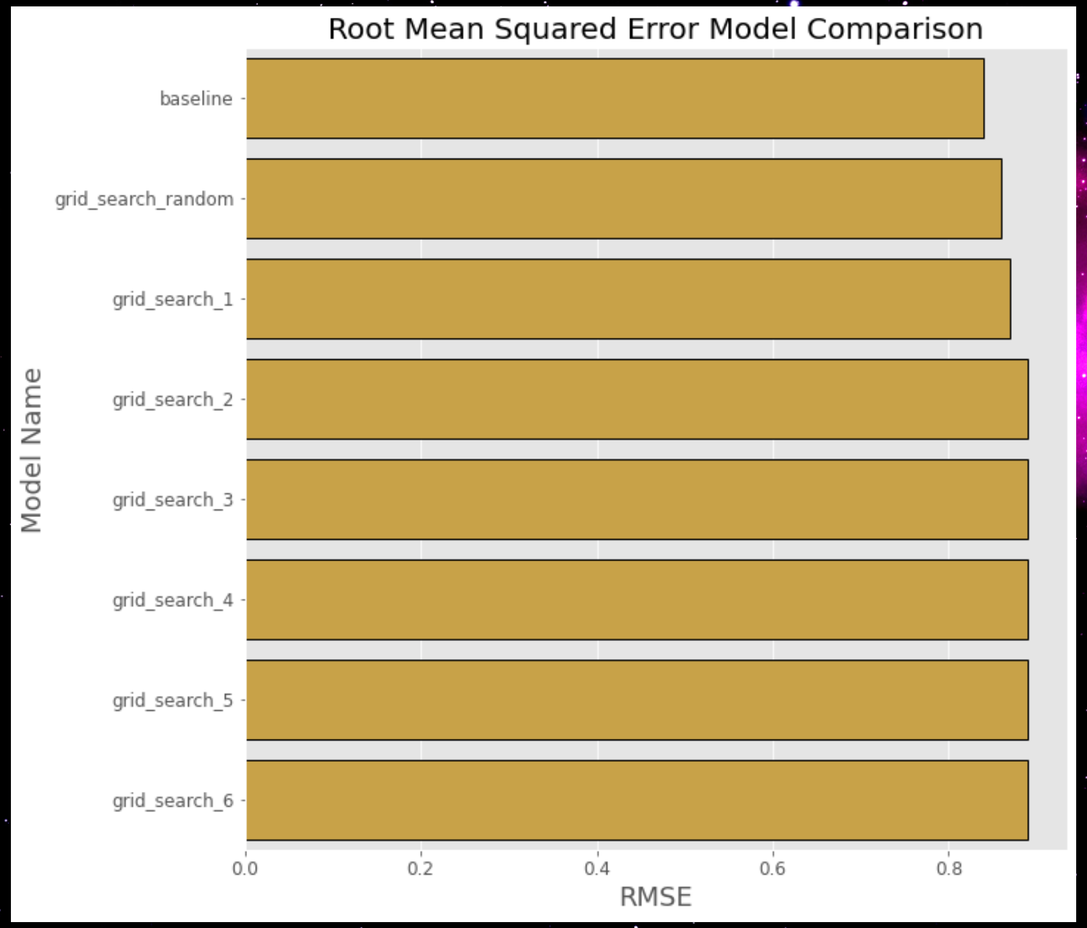


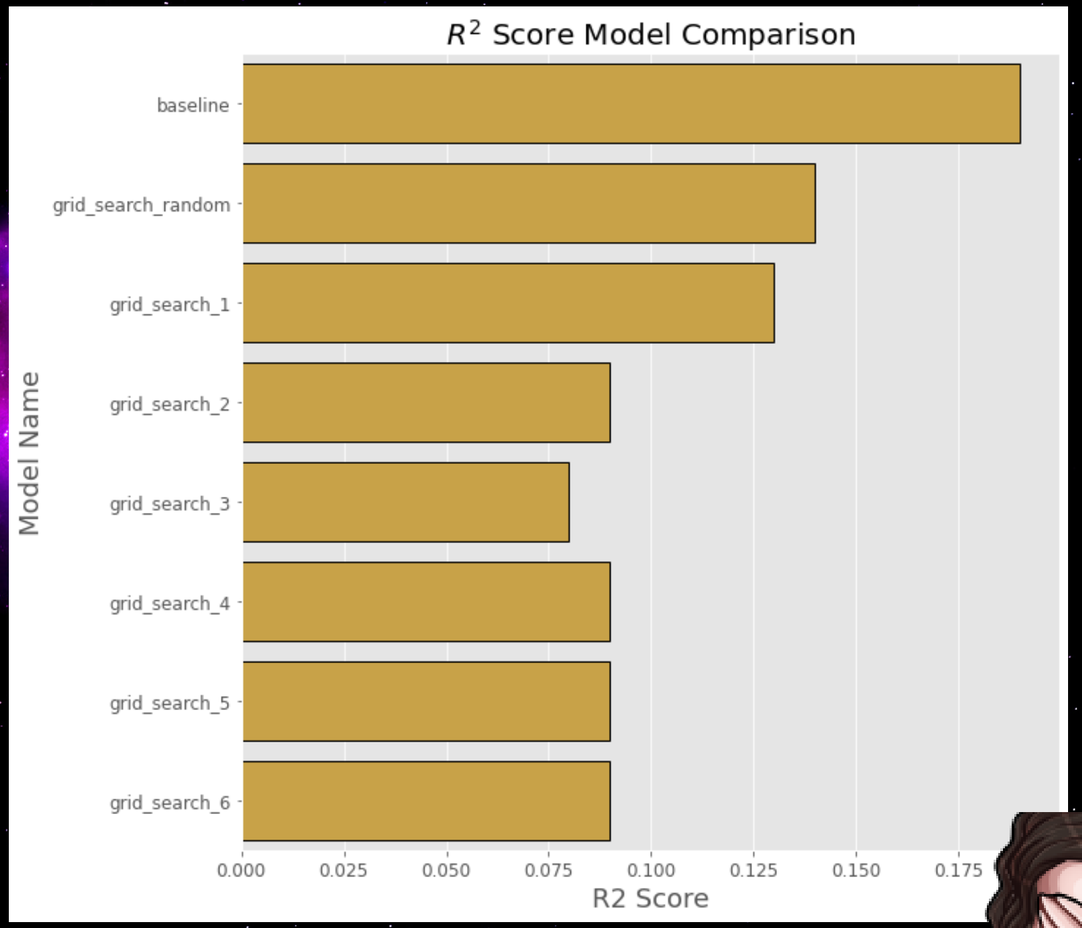
MODELLING

USING RMSE (ROOT MEAN SQUARED ERROR) AS A METRIC

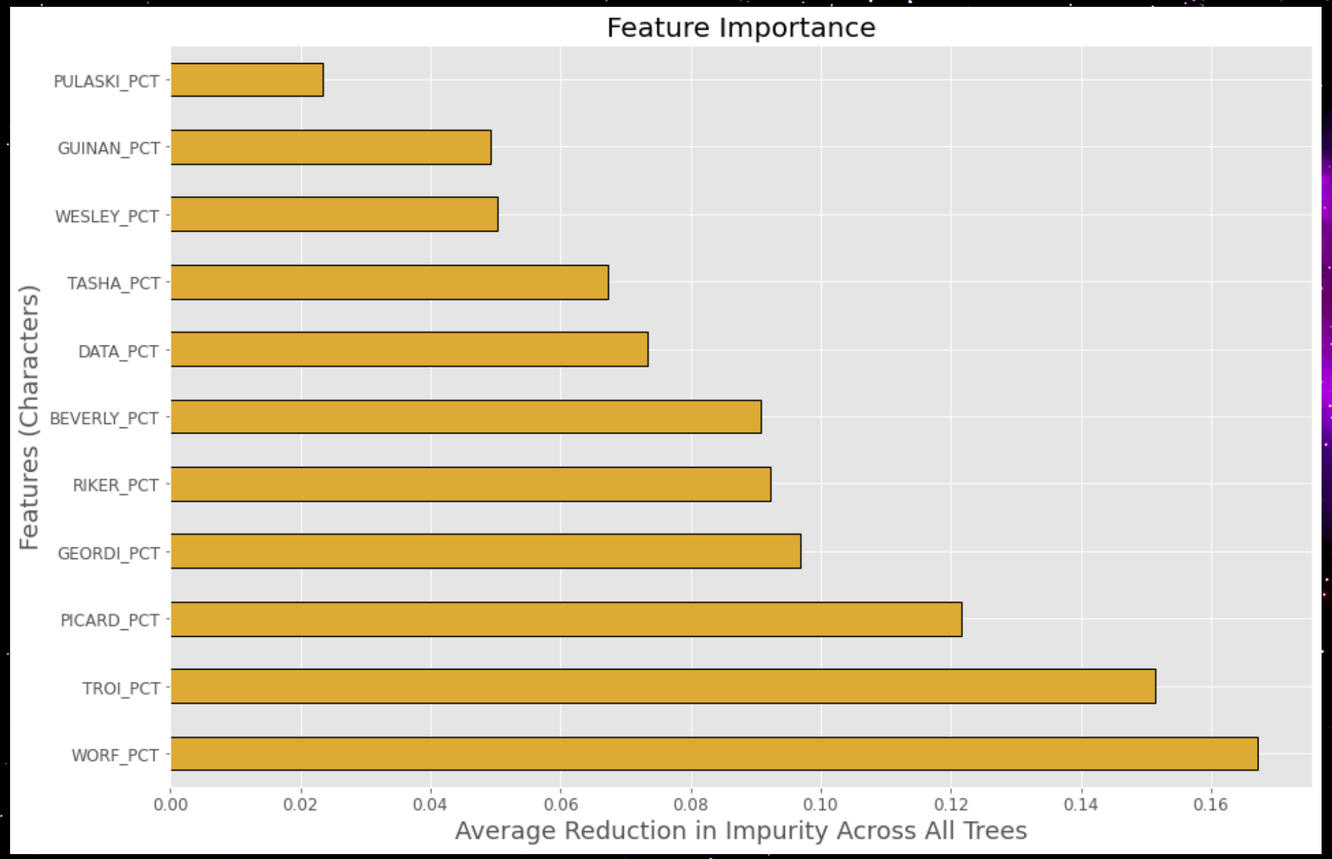
	Linear Regression	Decision Tree Regressor	Random Forest Regressor	SVM Regressor	OrdinalRidge (MORD)	LAD (MORD)	GradientBoost Regressor
Training RMSE	0.81	0.00	0.36	0.62	0.85	0.88	0.30
Testing RMSE	0.87	1.34	0.87	0.94	0.93	0.94	0.93
Mean Cross Validation Score	0.98	1.13	0.97	0.93	1.01	1.01	1.02

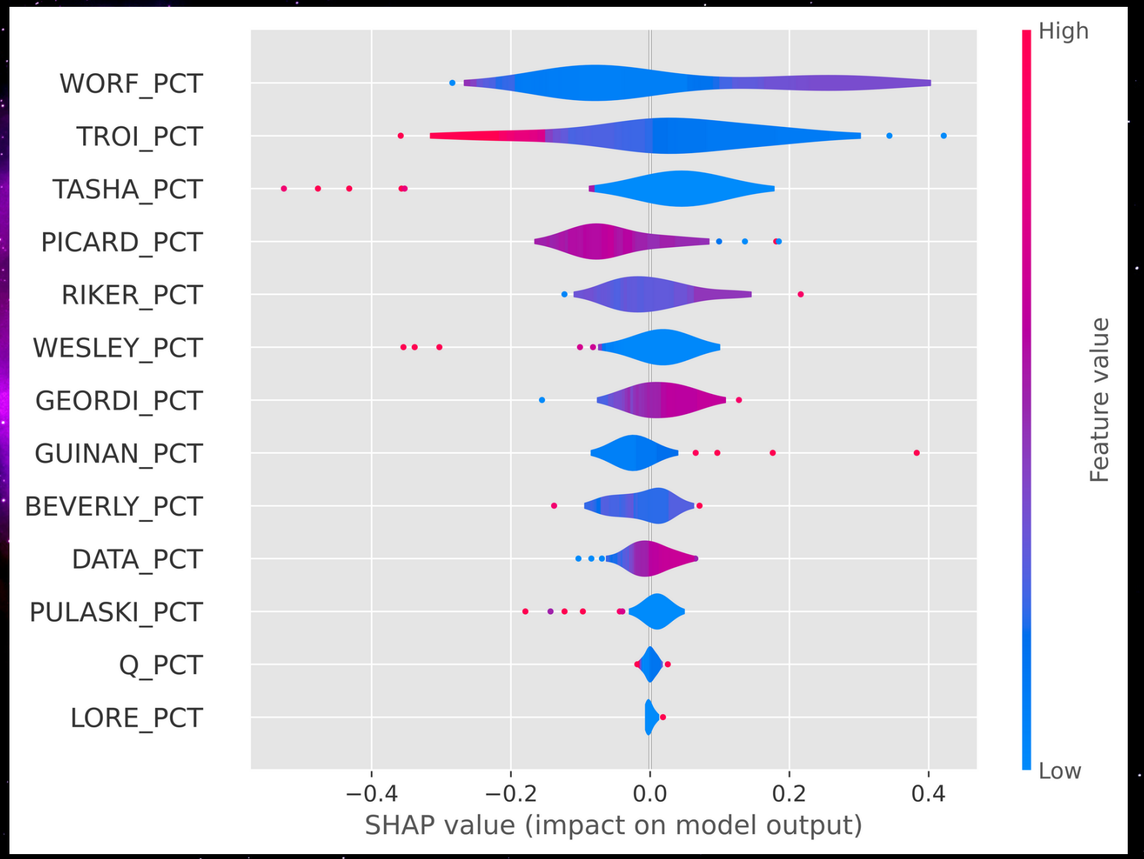
MODELLING


Root Mean Squared Error Model Comparison
baseline
grid_search_random
grid_search_1
grid_search_2
grid_search_3
grid_search_4
grid_search_5
grid_search_6
Model Name
0.0
0.2
0.4
0.6
0.8
RMSE


R^2 Score Model Comparison
baseline
grid_search_random
grid_search_1
grid_search_2
grid_search_3
grid_search_4
grid_search_5
grid_search_6
Model Name
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
R2 Score

MODELLING


Feature Importance
PULASKI_PCT
GUINAN_PCT
WESLEY_PCT
TASHA_PCT
DATA_PCT
BEVERLY_PCT
RIKER_PCT
GEORDI_PCT
PICARD_PCT
TROI_PCT
WORF_PCT
Features (Characters)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
Average Reduction in Impurity Across All Trees


WORF_PCT
TROI_PCT
TASHA_PCT
PICARD_PCT
RIKER_PCT
WESLEY_PCT
GEORDI_PCT
GUINAN_PCT
BEVERLY_PCT
DATA_PCT
PULASKI_PCT
Q_PCT
LORE_PCT
−0.4
−0.2
0.0
0.2
0.4
SHAP value (impact on model output)
High
Low
Feature value

NEXT STEPS

CLEANUP MY NOTEBOOKS, AND FINALIZE THE REPORT

FIND A METHOD OF DEMONSTRATION, EITHER THROUGH STREAMLIT OR FLASK

AFTER BOOTCAMP

TEST WITH A NEURAL NETWORK, AND MORE NLP FOCUSED ANALYSIS


python
matplotlib
SHAP
seaborn
NumPy
THANK YOU
scikit learn
TRACK MY PROGRESS
jupyter
GITHUB.COM/KATYAZEROSS
pandas
GitHub