Capstone Project-2

Ted Talk Views Prediction

By-Prabir Debnath

Points of Discussion

1. The Problem Statement
2. Concept of ML Model
3. Summary of the Experience Set
4. Exploration and Pre-processing of Data
5. Building and Evaluation of Model-1
6. Building and Evaluation of Model-2
7. Building and Evaluation of Model-3
8. Final Conclusion

The Problem Statement

We are provided with a labeled dataset on the details of Ted Talk Videos. The task is to explore and analyze the data and to build a regression model for Ted Talk Videos Views Prediction.

Concept of ML Model

The performance of a machine learning model depends on three factors:

i. Quality of Data

(cleaner experiences for better learning)

ii. Quantity of Data

(more experiences for better learning)

iii. Quality of Model

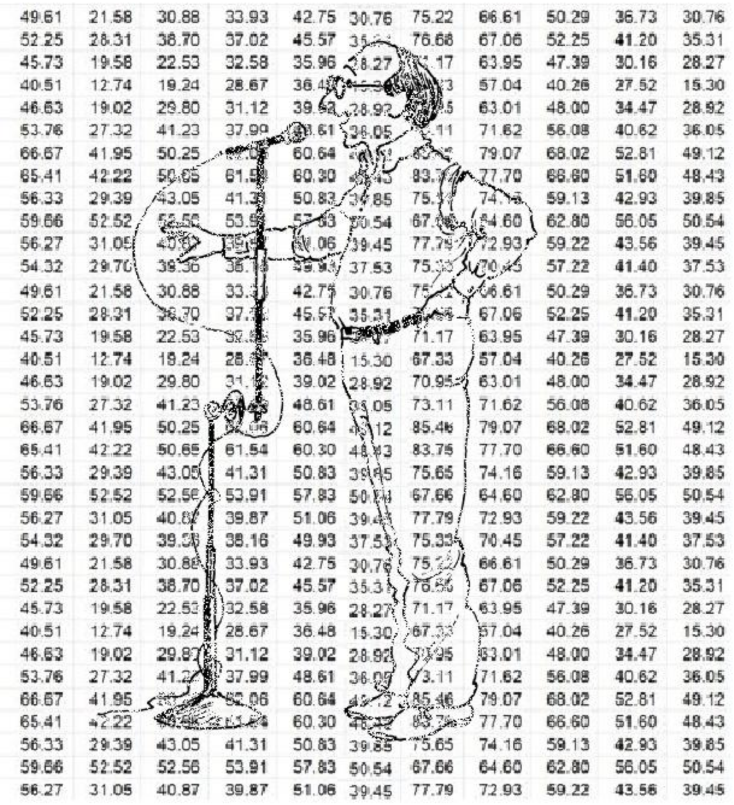
(right model and right hyperparameters for better learning)

Summary of the Experience Set

Here, the dataset has 4,005 rows, which means 4,005 experiences about Ted Talk Videos and

It has 19 columns, which means each experience is observed along 19 features or dimensions.

Let’s Decode the Experiences !



49.61	21.58	30.88	33.93	42.75	30.76	75.22	66.61	50.29	36.73	30.76
52.25	28.31	38.70	37.02	45.57	[illegible]	76.66	67.06	52.25	41.20	35.31
45.73	19.58	22.53	32.58	35.96	28.27	[illegible]	63.95	47.39	30.16	28.27
40.51	12.74	19.24	28.67	[illegible]	[illegible]	[illegible]	57.04	40.26	27.52	15.30
46.63	19.02	29.80	31.12	[illegible]	28.92	[illegible]	63.01	48.00	34.47	28.92
53.76	27.32	41.23	37.99	[illegible]	36.05	[illegible]	71.62	56.08	40.62	36.05
66.67	41.95	50.25	[illegible]	60.64	[illegible]	[illegible]	79.07	68.02	52.81	49.12
65.41	42.22	50.65	[illegible]	60.30	[illegible]	[illegible]	77.70	66.60	51.60	48.43
56.33	29.39	43.05	[illegible]	50.83	39.85	[illegible]	[illegible]	59.13	42.93	39.85
59.66	52.52	[illegible]	[illegible]	[illegible]	50.54	[illegible]	[illegible]	62.80	56.05	50.54
56.27	31.05	[illegible]	[illegible]	[illegible]	39.45	[illegible]	72.93	59.22	43.56	39.45
54.32	29.70	39.36	[illegible]	[illegible]	37.53	[illegible]	[illegible]	57.22	41.40	37.53
49.61	21.58	30.88	[illegible]	42.75	30.76	[illegible]	[illegible]	50.29	36.73	30.76
52.25	28.31	[illegible]	[illegible]	[illegible]	35.31	[illegible]	67.06	52.25	41.20	35.31
45.73	19.58	22.53	[illegible]	35.96	[illegible]	71.17	63.95	47.39	30.16	28.27
40.51	12.74	19.24	[illegible]	36.48	15.30	67.33	57.04	40.26	27.52	15.30
46.63	19.02	29.80	[illegible]	39.02	28.92	70.95	63.01	48.00	34.47	28.92
53.76	27.32	41.23	[illegible]	48.61	[illegible]	73.11	71.62	56.08	40.62	36.05
66.67	41.95	50.25	[illegible]	60.64	[illegible]	85.46	79.07	68.02	52.81	49.12
65.41	42.22	50.65	61.54	60.30	[illegible]	83.75	77.70	66.60	51.60	48.43
56.33	29.39	43.05	41.31	50.83	[illegible]	75.65	74.16	59.13	42.93	39.85
59.66	52.52	[illegible]	53.91	57.83	[illegible]	67.66	64.60	62.80	56.05	50.54
56.27	31.05	[illegible]	39.87	51.06	[illegible]	77.79	72.93	59.22	43.56	39.45
54.32	29.70	[illegible]	38.16	49.93	[illegible]	75.33	70.45	57.22	41.40	37.53
49.61	21.58	[illegible]	33.93	42.75	[illegible]	[illegible]	66.61	50.29	36.73	30.76
52.25	28.31	38.70	37.02	45.57	[illegible]	[illegible]	67.06	52.25	41.20	35.31
45.73	19.58	[illegible]	32.58	35.96	28.27	71.17	63.95	47.39	30.16	28.27
40.51	12.74	19.24	28.67	36.48	15.30	[illegible]	57.04	40.26	27.52	15.30
46.63	19.02	[illegible]	31.12	39.02	[illegible]	[illegible]	[illegible]	48.00	34.47	28.92
53.76	27.32	[illegible]	37.99	48.61	[illegible]	[illegible]	71.62	56.08	40.62	36.05
66.67	41.95	[illegible]	[illegible]	60.64	[illegible]	[illegible]	79.07	68.02	52.81	49.12
65.41	42.22	[illegible]	[illegible]	60.30	[illegible]	[illegible]	77.70	66.60	51.60	48.43
56.33	29.39	43.05	41.31	50.83	39.85	75.65	74.16	59.13	42.93	39.85
59.66	52.52	52.56	53.91	57.83	50.54	67.66	64.60	62.80	56.05	50.54
56.27	31.05	40.87	39.87	51.06	39.45	77.79	72.93	59.22	43.56	39.45

Exploration and Pre-processing of Data

We have done the exploration and pre-processing in seven steps to transform raw data into quality data for our ml model.

1. Connection with the Data
2. First Feelings of the Data
3. Deeper Understanding of the Data
4. Cleaning the Data
5. Treating Anomalies in the Data
6. Final Feature Selection from the Data
7. Preparation of Input and Output Data

Cleaning the Data

> We have handled 522 null values in ‘occupations’ column with imputation, thus there is no loss of data

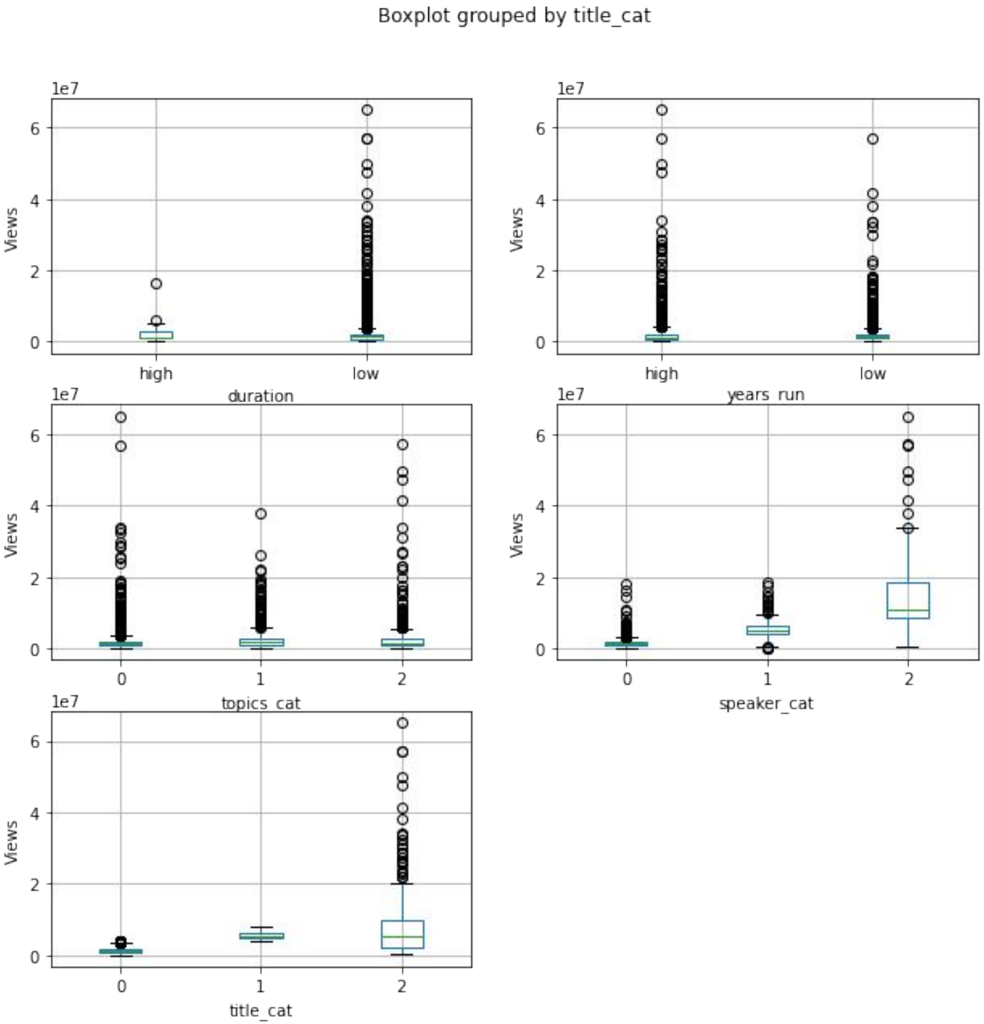
> We have done feature engineering on ‘published_date’ to extract ‘years_run’ feature

> There are very few experiences (only 1-27) for most of the 'native_lang' as compared to 'en'. We have removed those experiences and lost 48 experiences. Thus our model will predict the ted talk views only for english language

> We have encoded ‘topics’ column with three categories: Highly Favourite:2, Medium Favourite:1, Least Favourite:0
> We have encoded ‘speaker_1’ column with three categories: Highly Famous:2, Medium

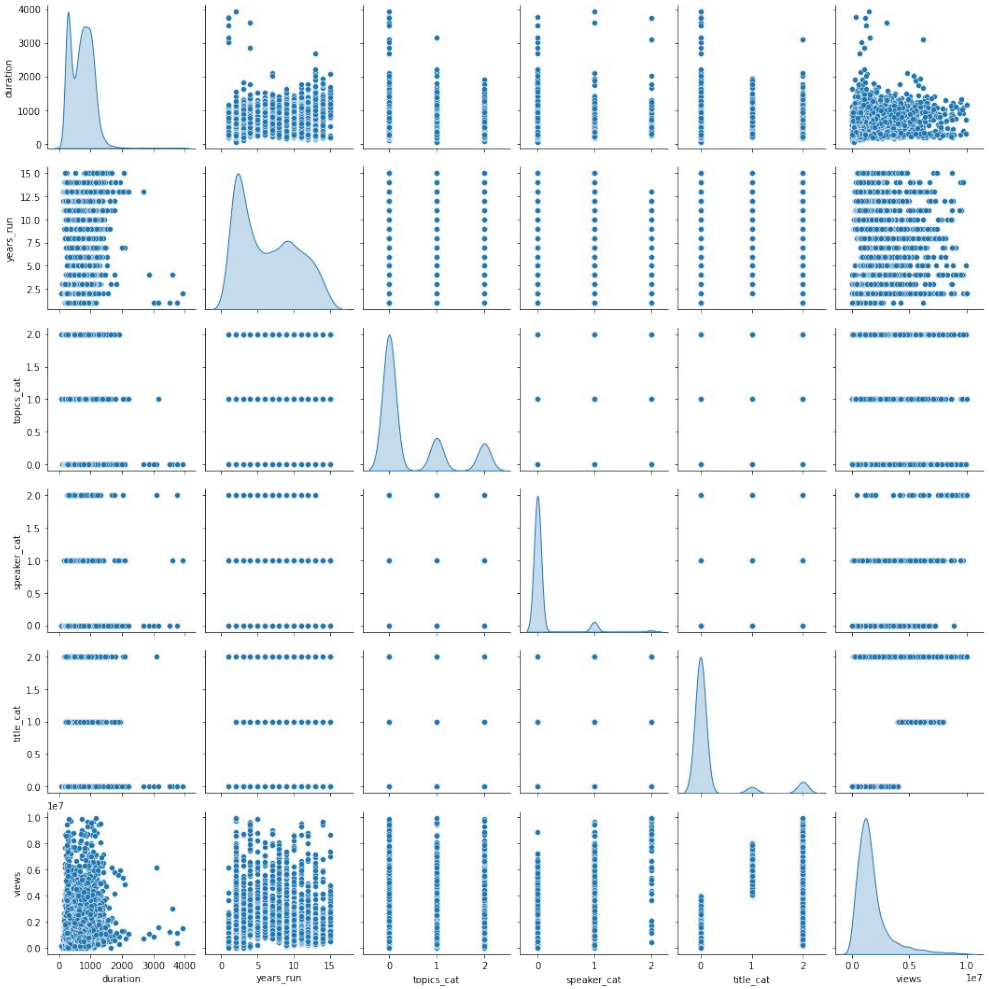
Famous:1, Least Famous:0

> We have encoded ‘title’ column with three categories: Highly Attractive:2, Medium Attractive:1, Least Attractive:0



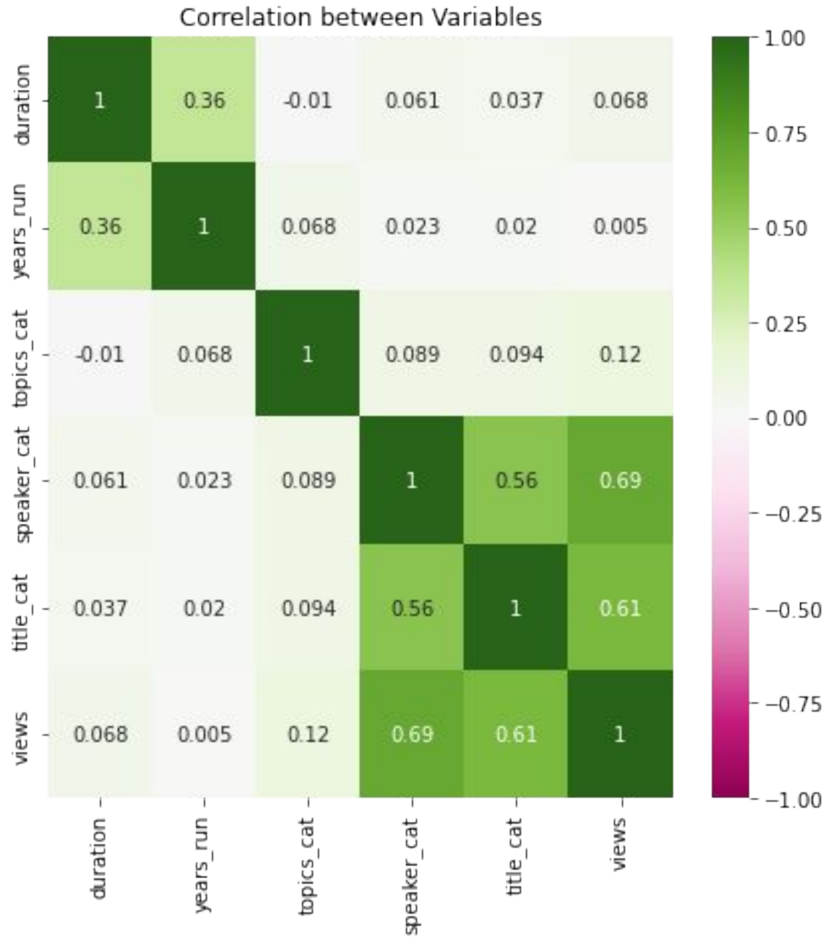
Treating Anomalies in the Data

We have removed 89 experiences with more than 10 million views as they are exceptional experiences for all of the features.



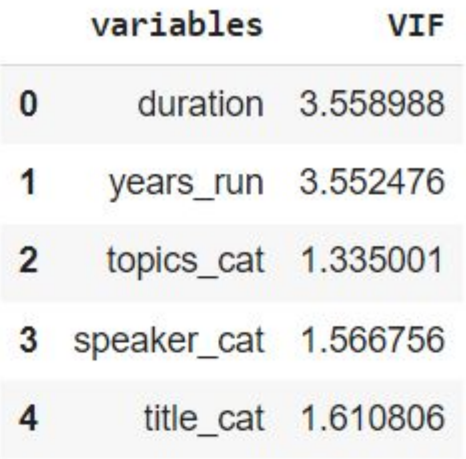
Overall Feature Understanding

Here, the distribution of ‘views’, and 'duration' are positively skewed. Thus we have done log transformation on these features to normalize their distribution.



Looking for Truly Independent Features

Here, all our input variables are truly independent as all the VIF values are below 10



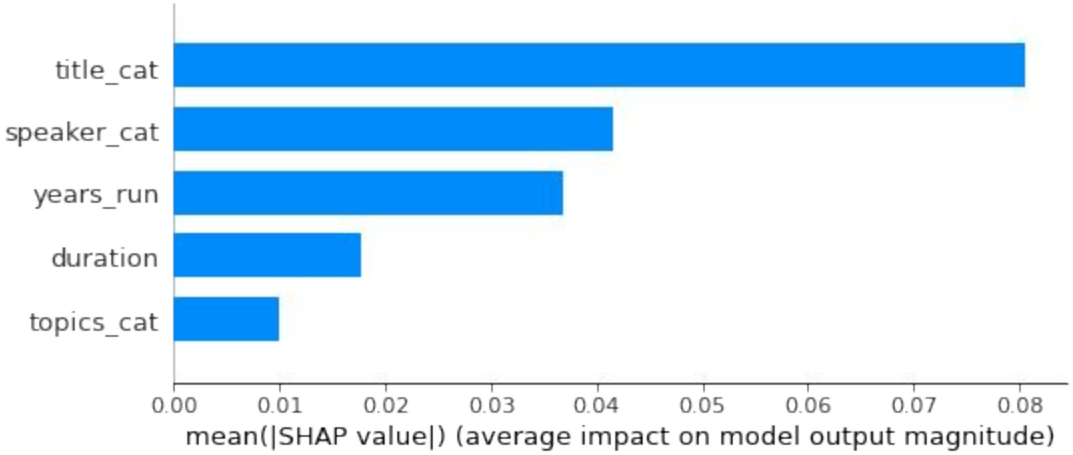
	variables	VIF
0	duration	3.558988
1	years_run	3.552476
2	topics_cat	1.335001
3	speaker_cat	1.566756
4	title_cat	1.610806

Let’s Train the Models !

Building and Evaluation of Model-1

Final Random Forest Model:

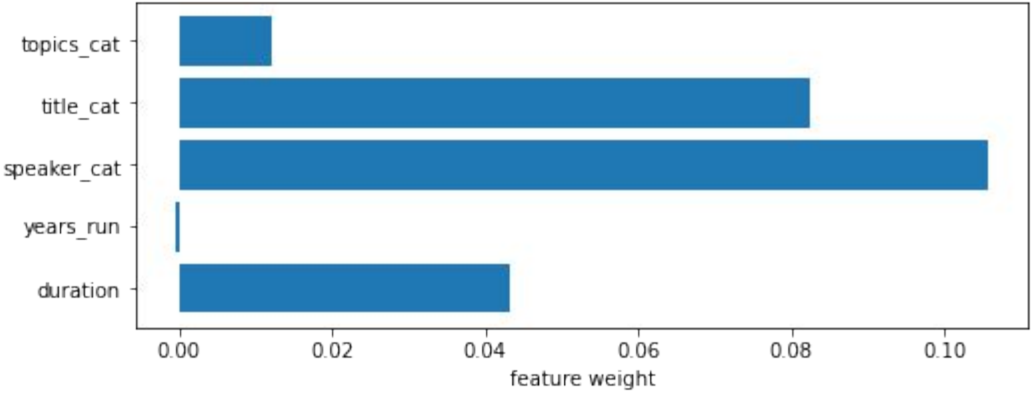
After cross validation and hyperparameter tuning, the best parameters are {'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 10, 'min_samples_split': 5, 'n_estimators': 200} (test r2_score is 63% and variance in prediction is 3%)



Building and Evaluation of Model-2

Final Ridge Model:

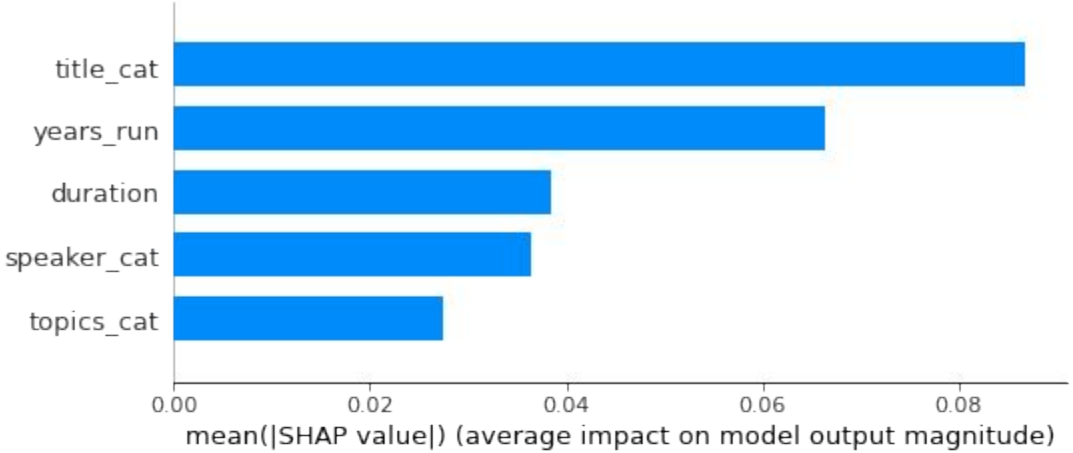
After cross validation and hyperparameter tuning, the best parameters are {'alpha': 100, 'fit_intercept': True, 'max_iter': 150, 'tol': 0.2} (test r2_score is 50% and variance in prediction is 4%)

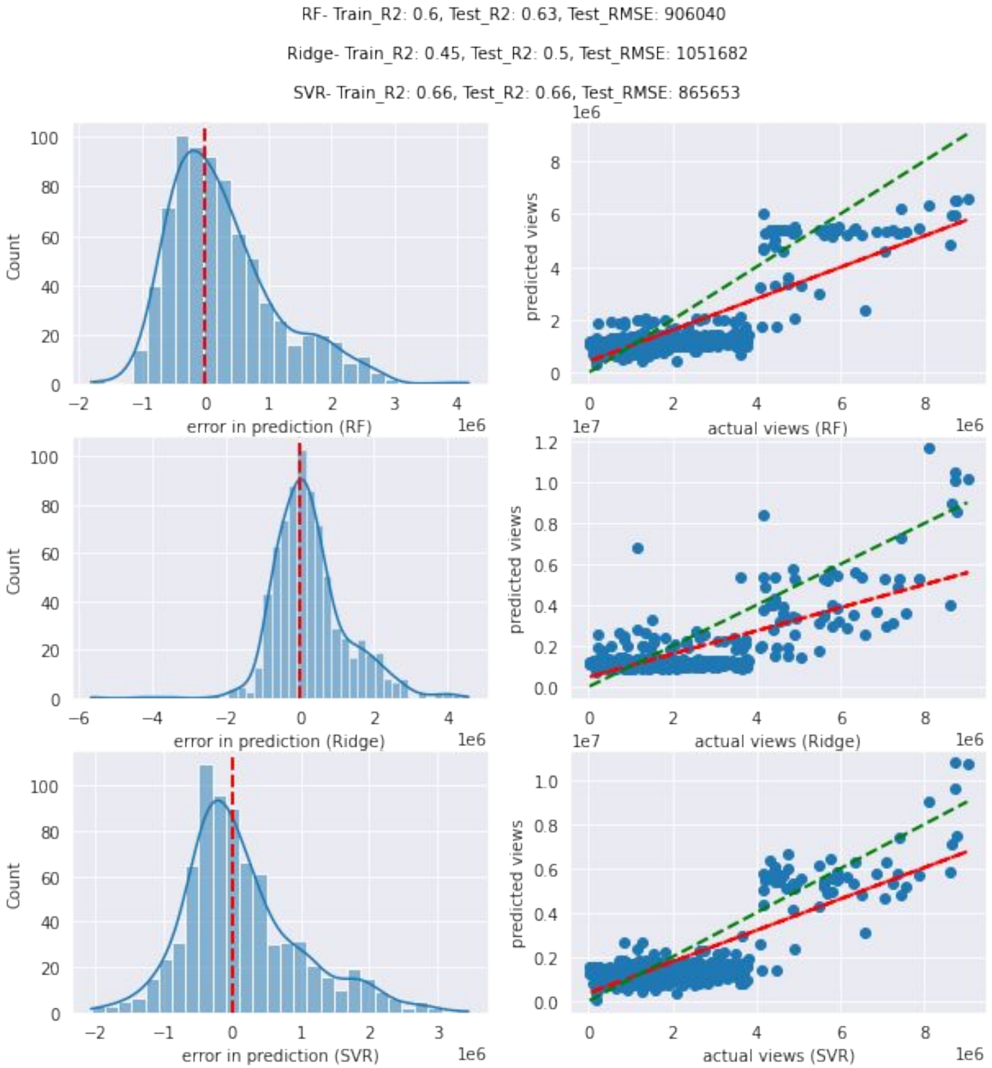


Building and Evaluation of Model-3

Final SVR Model:

After cross validation and hyperparameter tuning, the best parameters are {'C': 6, 'gamma': 0.1} (test r2_score is 66% and variance in prediction is 0%)





RF- Train_R2: 0.6, Test_R2: 0.63, Test_RMSE: 906040

Ridge- Train_R2: 0.45, Test_R2: 0.5, Test_RMSE: 1051682

SVR- Train_R2: 0.66, Test_R2: 0.66, Test_RMSE: 865653

Final Conclusion

On the basis of the performance study of our three models, we are selecting **Support Vector Regressor** (*the best warrior*) for predicting ted_talk views, as it has the lowest variance in prediction and highest r2_score among all three models

Thank you !