Menu

Go to

Regressor

In this section, we aimed to identify the best regression model for our dataset by utilizing a combination of **AutoML** and **GridSearch** techniques.

- We used a dataset containing **12,000** samples for training and **3,000** for testing.
- To extract meaningful representations from the text data, we employed **BERT-based sentence embeddings**, which effectively capture the semantic relationships between words.

To find the most effective regression model, we explored two approaches:

1. **AutoML using TPOT** – An automated machine learning pipeline optimization tool that systematically searches for the best model and hyperparameters.
2. **GridSearch** – A traditional hyperparameter tuning method that evaluates predefined parameter combinations to optimize model performance.

Best Performing Model

Both approaches independently determined that **Random Forest Regression** was the most effective model for our dataset.

This indicates that an ensemble-based method, which leverages multiple decision trees to improve accuracy and reduce overfitting, provided the best generalization to unseen data.

Result:

Random Forest Regression Results

	Paper ID	Actual Citations	Predicted Citations
0	764015	1.0000	48.4293
1	822808	20.0000	47.6628
2	836090	50.0000	34.7556
3	60912	122.0000	40.9289
4	978063	50.0000	63.0062
5	576953	1.0000	38.0120
6	585731	50.0000	42.9059
7	615581	0.0000	33.2350
8	475637	1.0000	36.6131
9	839929	0.0000	30.9720
10	913465	2.0000	32.5011
11	578924	2.0000	52.3052

	Metric	Value	Interpretation
0	RMSE (Root Mean Squared Error)	440.1746	High RMSE indicates **large deviations** between predicted and actual citation counts.

	Metric	Value	Interpretation
1	MAE (Mean Absolute Error)	93.9885	On average, predictions deviate by **94 citations** from actual values.
2	R^2 Score	-0.0224	A negative R^2 suggests the model **performs worse than a simple mean predictor**, meaning **very weak predictive power**.

XGBoost Regression Results

	Paper ID	Actual Citations	Predicted Citations
0	764015	1.0000	2.0623
1	822808	20.0000	7.7063
2	836090	50.0000	10.9793
3	60912	122.0000	20.4241
4	978063	50.0000	12.1711
5	576953	1.0000	5.3041
6	585731	50.0000	8.9392
7	615581	0.0000	6.7033
8	475637	1.0000	11.0652
9	839929	0.0000	2.0623
10	913465	2.0000	5.0930
11	578924	2.0000	28.7883
12	443603	2.0000	12.3324
13	980326	0.0000	2.0623

XGboost Regression Evaluation Metrics

	Metric	Value	Interpretation
0	RMSE (Root Mean Squared Error)	123.1483	The model's **prediction error** is significantly lower than previous models.
1	MAE (Mean Absolute Error)	32.8007	On average, predictions deviate by **~33 citations** from actual values.

	Metric	Value	Interpretation
2	R^2 Score	-0.0304	**Slightly negative R^2** suggests **poor model performance**, though better than before.

LightGBM Regression Results

	Paper ID	Actual Citations	Predicted Citations
0	764015	1.0000	2.0845
1	822808	20.0000	12.9575
2	836090	50.0000	8.4029
3	60912	122.0000	12.5077
4	978063	50.0000	12.5412
5	576953	1.0000	2.8591
6	585731	50.0000	9.2386
7	615581	0.0000	5.0948
8	475637	1.0000	9.9395
9	839929	0.0000	2.0845
10	913465	2.0000	13.6013
11	578924	2.0000	29.1868
12	443603	2.0000	11.8679
13	980326	0.0000	2.0845
14	766729	2.0000	12.9251
15	492331	2.0000	8.5311

LightGBM Regression Evaluation Metrics

	Metric	Value	Interpretation
0	RMSE (Root Mean Squared Error)	123.4998	The prediction error remains comparable to XGBoost, with a **slightly higher RMSE** than XGBoost (123.15).

	Metric	Value	Interpretation
1	MAE (Mean Absolute Error)	32.8441	Mean absolute error is **almost identical** to XGBoost (32.8007), suggesting similar average prediction deviation.
2	R^2 Score	-0.0363	Slightly lower than XGBoost (-0.0304), indicating a **marginal decrease** in the model's ability to explain variance.

Challenges:

Older papers tend to accumulate more citations compared to newer papers simply due to having more time to be cited.Test sets containing newer papers may have lower citation counts compared to what the model expects based on historical data.Certain topics may become more popular over time, while others may decline in interest. The model trained on historical data may struggle to predict citations for topics that were not popular during the training period.

Test data may include new research topics or keywords that were not present during the training phase.The model cannot infer the importance or future citation potential of these novel concepts since they were not seen in the training set.Test data may include new researchers with no prior publication history in the training set.Author-related features may not generalize to new authors, leading to poor predictions.