**Problem Statement:**

Predict the number of shares in social networks using a dataset

**Dataset:**

Online News Popularity Data Set

This advanced level data set has 39644 rows and 61 columns.

This dataset summarises a heterogeneous set of features about articles published by Mashable in a period of two years.

**Data Preprocessing:**

1. Handling NULL values

There are no null values so there is no need to replace them with the mean or to delete any rows or columns.

1. Removing columns that are non predictive

Removed the ‘URL’ and the ‘Timedelta’ columns as they were already marked as non predictive. This also takes care of the only Categorical variable (‘URL’) so there is no need for Encoding Categorical Variables.

1. Feature scaling was done where ever it was needed

80% of the data was used for training the model and the remaining 20% was used for testing

**Models and Techniques Used:**

1. Decision Tree Regression

* First Run : r2 of -2.8, not a very good result
* Second Run: r2 of -2.6, not a very good result

1. Multiple Linear Regression

* First Run (): r2 of -0.59, still not a very good result but a better one than Decision Trees

1. Random Forest Regression

* First Run(n\_estimators = 10 , random\_state = 0) : r2 of -0.33, not a very good result but the best one so far
* Second Run(n\_estimators = 20) : r2 of -0.24, slight improvement
* Third Run(n\_estimators = 100) : r2 of -0.139, slight improvement
* Fourth Run(n\_estimators = 1000, random\_state = 50) : r2 of -0.08, slight improvement but still not great

1. Support Vector Regression

* First Run(kernel = ‘rbf’) : r2 of 0.008, best result until now

1. Using Bagging Regressor

* with estimator = random forest: -0.07
* with estimator = svr: 0.0137, new best result

1. Boosting Methods

* LightBGM: r2 of -0.03
* XGBoost :. r2 of -0.26.

1. CatBoost Regressor

* r2 of -0.05

1. Polynomial Regression using the column with highest correlation with shares

* 0.0225, **best result**

**Final Results:**

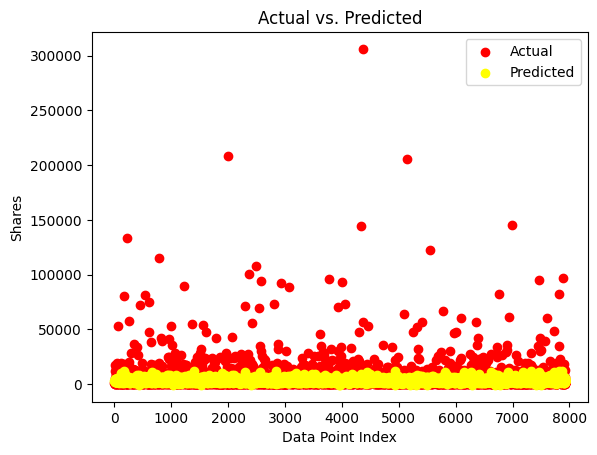
In the evaluation of different regression models and techniques for the given dataset, the Support Vector Regression (SVR) with the radial basis function (RBF) kernel and the Bagging Regressor using the SVR as the base estimator stood out as the second best-performing approaches. The R-squared value achieved by the SVR with RBF kernel was 0.008, which was the highest among all the models tested.

Further, the Bagging Regressor with SVR as the base estimator achieved an improved R-squared value of 0.0137.

The other models, including Decision Tree Regression, Multiple Linear Regression, Random Forest Regression, LightBGM, and XGBoost, showed lower and even negative R-squared values, indicating that they were not suitable for capturing the underlying patterns in the dataset.

However, the best performance was from using a simpler model which was the Polynomial Regression model with the variable that had the highest correlation with the shares(column to be predicted) column for which the r2 was 0.225.

Overall, the results suggest that the dataset might have a complex non-linear relationship.



Data Visualisation for the model with the best results