Assignment – 2 Report

NEERAJA VUDHANTHI

The dataset used for this assignment is the "ozone_level" regression dataset from the OpenML website, which is a collection of atmospheric data of the ozone levels. This aim of this “ozone_level” dataset is to predict the ozone level (target) based on various environmental factors (features). The features in this dataset are of both nominal and numeric types. These nominal and numeric features of the dataset describe atmospheric conditions such as temperature, wind speed, humidity, and pressure. The “ozone_level” dataset has 2538 instances with 73 attributes/features.

Few of the features in the "ozone_level" dataset may include:

1. Temperature - numeric
2. Wind speed - numeric
3. Humidity - numeric
4. Atmospheric pressure - numeric
5. Solar radiation - numeric
6. Wind direction – numeric
7. Precipitation - numeric
8. Nitrogen oxides concentration - numeric
9. Carbon monoxide concentration - numeric
10. Particulate matter concentration - numeric
11. Location/Region - nominal
12. Weather Conditions – nominal

These features provide essential information about the atmospheric conditions that influence ozone levels. The aim of this dataset is to provide datapoints for predictive modeling to accurately forecast ozone levels based on these environmental variables.

Since the assignment asks us to use any metric for the learning curve, we will be using the RMSE values of trained regression models.

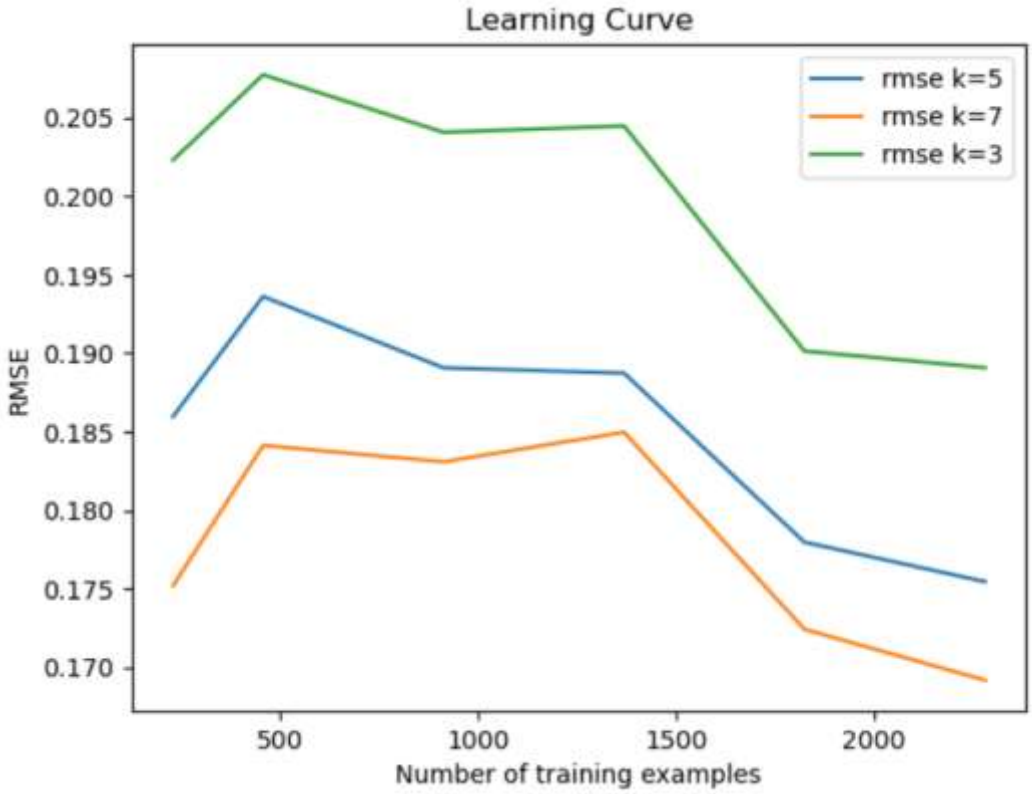

Fig 1: Learning curve for the different values of k with RMSE represented on the Y-axis and Number of training examples represented on the X-axis.

K	Values for cross validation = 10 (last point with maximum training data)									
7	0.209	0.178	0.210	0.275	0.142	0.145	0.109	0.165	0.115	0.140
5	0.214	0.183	0.214	0.282	0.144	0.153	0.113	0.171	0.130	0.147
3	0.230	0.189	0.229	0.292	0.164	0.176	0.117	0.182	0.146	0.162

Table 1: Recorded RMSE values for the Number of Neighbors with K=3,5,7.

Analysis:

k=3

1. Model Complexity: As the size of the training data increases, the model's performance tends to improve. This suggests that the KNN regressor with k = 3 benefits from more data for learning and generalizes better to unseen examples.

2. Generalization: The improvement in test scores with more data indicates that the model generalizes well. There seems to be a consistent trend of better performance with increasing training data size, suggesting good generalization ability.

k = 5:

1. Model Complexity: Similar to k = 3, as the training data size increases, the model's performance tends to improve, indicating better generalization with more data.

2. Generalization: The improvement in test scores with more data suggests good generalization ability. The trend of increasing test scores indicates that the model is learning effectively from the training data.

3. Model Selection: As with k = 3, the gap between training and test scores helps in determining that the model is not overfitting..

k = 7:

1. Model Complexity: The model's performance improves with increasing training data size, suggesting better generalization.

2. Generalization: The consistent improvement in test scores indicates that the model generalizes well to unseen data. This suggests that the KNN regressor with k = 7 is learning effectively from the training data performs when compared to the regressors with k=3 and k=5.

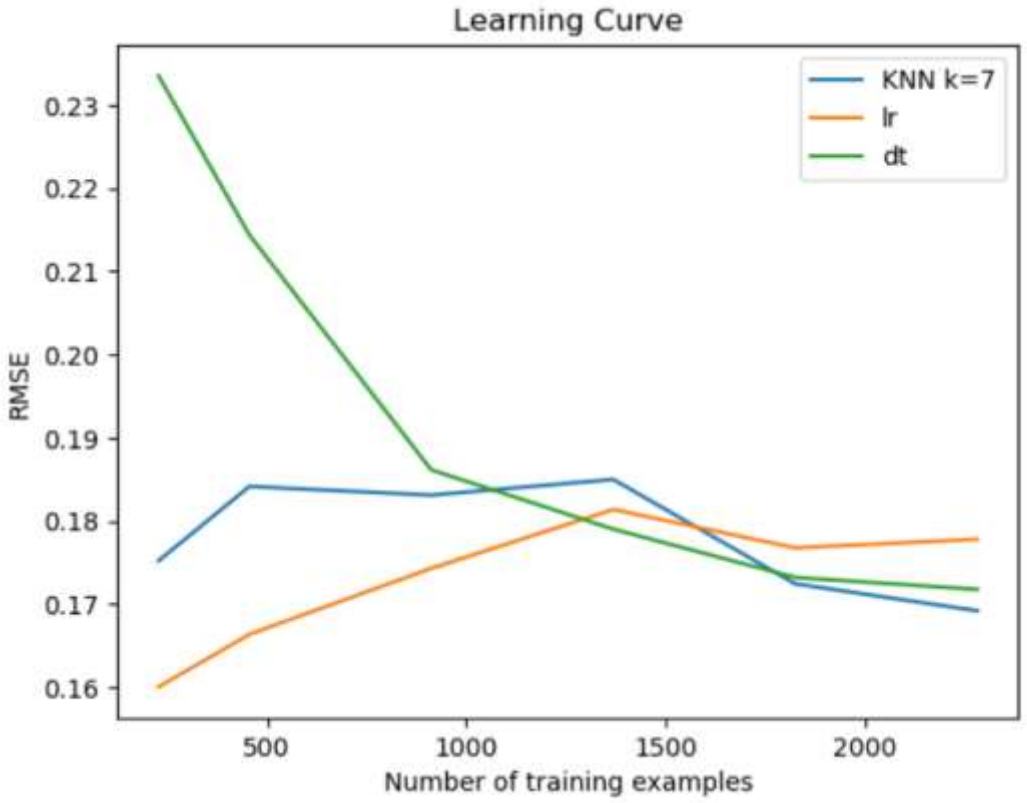

Fig 2: Learning curve for the different models trained with different methods. Here, RMSE represented on the Y-axis and Number of training examples represented on the X-axis.

Method	Values for cross validation = 10 (last point with maximum training data)									
KNN	0.209	0.178	0.210	0.275	0.142	0.145	0.109	0.165	0.115	0.140
DT	0.271	0.163	0.205	0.289	0.166	0.158	0.089	0.152	0.093	0.125
LR	0.214	0.185	0.215	0.273	0.162	0.160	0.114	0.177	0.123	0.150

Table 2: Recorded RMSE values for the KNN, Decision Tree(DT) and Linear Regression(LR) models.

Analysis:

KNN (K-Nearest Neighbors):

- **RMSE Values:** The RMSE values for KNN range from approximately 0.109 to 0.275.

- **Analysis:** KNN tends to have varying performance across different datasets or folds, as indicated by the range of RMSE values. It can be sensitive to the choice of k and the distance metric used. In this case, the performance seems relatively consistent, with some fluctuations.

Decision Tree (DT):

- RMSE Values: The RMSE values for Decision Tree range from approximately 0.089 to 0.289.

- Analysis: Decision Trees can capture complex nonlinear relationships in the data and are capable of fitting to the training data well. However, they might suffer from overfitting if not properly pruned or regularized, as indicated by the relatively wide range of RMSE values.

Linear Regression (LR):

- RMSE Values: The RMSE values for Linear Regression range from approximately 0.114 to 0.273.

- Analysis: The RMSE values suggest relatively stable performance across different folds.

Since, there is a huge gap between the linear regression and the other methods, Linear regression can achieve higher accuracy for this “ozone_level” dataset.