The (Un?) Reasonable Effectiveness of Data: Report

Team: 22

Members:

Joseph Fitzpatrick – 14312993
Callum Duffy - 14315135
Kevin Morris – 14315027

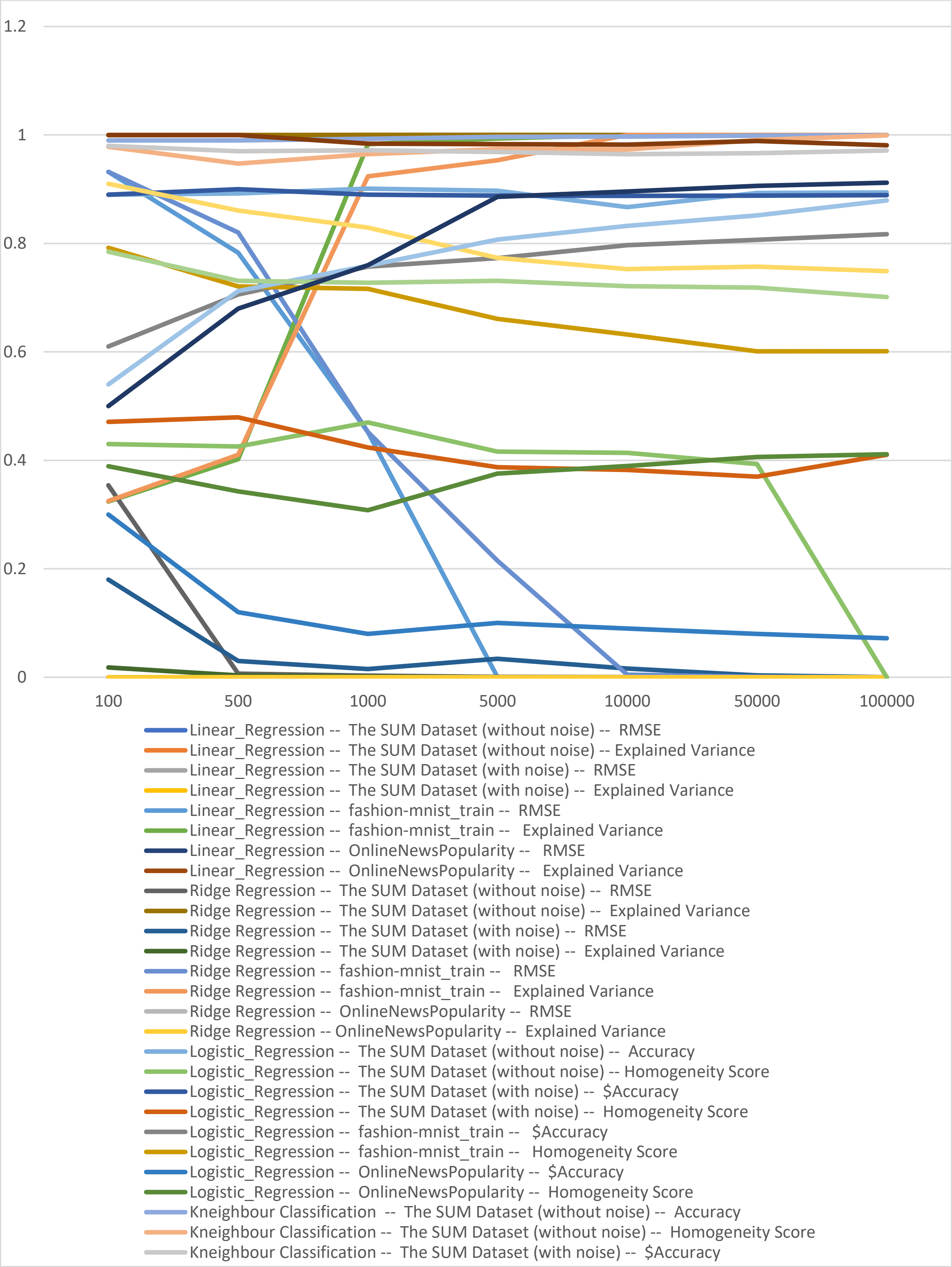
Time Required: 12 Hours

Results:

	100	500	1000	5000	10000	50000	100000
Linear_Regression -- The SUM Dataset (without noise) -- RMSE	0	0	0	0	0	0	0
Linear_Regression -- The SUM Dataset (without noise) -- Explained Variance	1	1	1	1	1	1	1
Linear_Regression -- The SUM Dataset (with noise) -- RMSE	0	0	0	0	0	0	0
Linear_Regression -- The SUM Dataset (with noise) -- Explained Variance	1	1	1	1	1	1	1
Linear_Regression -- fashion-mnist_train -- RMSE	0.932	0.783	0.452	0.0002	0	0	0
Linear_Regression -- fashion-mnist_train -- Explained Variance	0.324	0.4022	0.9831	0.9936	1	1	1
Linear_Regression -- OnlineNewsPopularity -- RMSE	0	0	0	0	0	0	0
Linear_Regression -- OnlineNewsPopularity -- Explained Variance	1	1	1	1	1	1	1
Ridge Regression -- The SUM Dataset (without noise) -- RMSE	0.3539	0.006623	0.00306	0.000632	0.00032	0.000006	0.000003
Ridge Regression -- The SUM Dataset (without noise) -- Explained Variance	0.999926	0.999902	0.999996	0.999998	1	1	1
Ridge Regression -- The SUM Dataset (with noise) -- RMSE	0.18	0.03	0.0152	0.034	0.0159	0.0032	0.0002
Ridge Regression -- The SUM Dataset (with noise) -- Explained Variance	0.018	0.003521	0.00152	0.000159	0.00032	0.00015	0.00002
Ridge Regression -- fashion-mnist_train -- RMSE	0.932	0.82	0.452	0.215	0.0049	0.0007	0.00003
Ridge Regression -- fashion-mnist_train -- Explained Variance	0.325	0.4102	0.9236	0.9534	1	1	1
Ridge Regression -- OnlineNewsPopularity -- RMSE	0	0	0	0	0	0	0
Ridge Regression -- OnlineNewsPopularity -- Explained Variance	1	1	1	1	1	1	1
Logistic_Regression -- The SUM Dataset (without noise) -- Accuracy	0.89	0.892	0.901	0.897	0.867	0.893	0.894113
Logistic_Regression -- The SUM Dataset (without noise) -- Homogeneity Score	0.429854	0.425439	0.469904	0.41606	0.413808	0.393135	N/a
Logistic_Regression -- The SUM Dataset (with noise) -- $Accuracy	0.89	0.9	0.89	0.8882	0.8877	0.8882	0.8889
Logistic_Regression -- The SUM Dataset (with noise) -- Homogeneity Score	0.471	0.47932	0.42363	0.38737	0.38226	0.369892	0.4102
Logistic_Regression -- fashion-mnist_train -- $Accuracy	0.61	0.706	0.757	0.773	0.797	0.8069	0.8171
Logistic_Regression -- fashion-mnist_train -- Homogeneity Score	0.792	0.720843	0.71623	0.660993	0.63214	0.60122	0.6012
Logistic_Regression -- OnlineNewsPopularity -- $Accuracy	0.3	0.12	0.08	0.1	0.09	0.08	0.0719
Logistic_Regression -- OnlineNewsPopularity -- Homogeneity Score	0.3889	0.342855	0.30804	0.37579	0.389691	0.40633	0.41122
Kneighbour Classification -- The SUM Dataset (without noise) -- Accuracy	0.99	0.99	0.993	0.9964	0.997	0.99908	0.999784
Kneighbour Classification -- The SUM Dataset (without noise) -- Homogeneity Score	0.9783	0.947305	0.964686	0.97424	0.973324	0.990619	0.99927
Kneighbour Classification -- The SUM Dataset (with noise) -- $Accuracy	0.98	0.97	0.972	0.9686	0.9646	0.96624	0.9712

Kneighbour Classification -- The SUM Dataset (with noise) -- Homogeneity Score	0.91	0.8609	0.82929	0.7734	0.7527	0.7571	0.7489
Kneighbour Classification -- fashion-mnist_train -- $Accuracy	0.54	0.712	0.759	0.807	0.8324	0.8516	0.8793
Kneighbour Classification -- fashion-mnist_train -- Homogeneity Score	0.7847	0.7309	0.727566	0.731185	0.7211	0.71872	0.7011
Kneighbour Classification -- OnlineNewsPopularity -- $Accuracy	0.5	0.68	0.76	0.886	0.896	0.906	0.912
Kneighbour Classification -- OnlineNewsPopularity -- Homogeneity Score	1	1	0.984	0.9831	0.9821	0.9891	0.981

Graph of Results:



Finding/Answer

Question 1: To what extent does the effectiveness of machine-learning algorithms depend on the size and complexity of the data?

From our tests, we have noticed that there is an increase in the effectiveness of machine-Learning algorithms when the number of instances of data increases. As we increase the number of instances of the data for our regression tests the accuracy generally tended toward the optimal depending on the scoring metric that was used. When carrying out Classification testing we noticed that while the accuracy generally increased when we added more data there was a tendency for some of the training data to be over-fit resulting in a slight decrease in the correctness of the classification.

The number of features allows us to more accurately predict the value we were trying to find. This suggested that the more data related to the target the better we could predict the target. However, when the variables we chose closely related to the target value we could didn't notice a significant reduction in accuracy. When the data had many features, it increased the runtime complexity significantly. This meant that we would have to wait a very long time for the values to be produced. This would need to be factored into the decision when deciding how many features to use.

Question 2: Looking only at the performance of your best performing machine-learning algorithm on "The SUM dataset (without noise)": how well was machine-learning suitable to solve the task of predicting a) the target value and b) the target class? Consider in your assessment, how well a simple rule-based algorithm would have performed.

From the test, I ran I found that Linear Regression performed the best on predicting the target class and value. This algorithm correctly predicted both the target value correctly. We think a rile based algorithm would not have performed as well as our model. This is because the data wasn't perfect and thus wouldn't have allowed for a rigid definition of the model for prediction to be as accurate as our model.

Datasets and Algorithms

Algorithm 1	Linear Regression
Algorithm 2	Ridge Regression
Algorithm 3	Logistic Regression
Algorithm 4	Kneighbour Classification
Dataset 1	Sum Dataset, with noise
Dataset 2	Sum Dataset, without noise
Dataset 3	Fashion-mnist_train
Dataset 4	OnlineNewsPopularity
Metric 1	RMSE
Metric 2	Explained Variance
Metric 3	Accuracy
Metric 4	Homogeneity Score

Contributions

For this task1 we all worked together on decided which datasets and algorithms. Kevin Morris and Callum decided on the scoring metrics to be used in the task. Joseph Fitzpatrick worked on the data regularization to allow for the data to be used easily by the algorithms and wrote the code to implement the different algorithms with help from Kevin and Callum.

Additional Information

For Explained Variance and Accuracy the best score is 1. For RMSE and Homogeneity Score the best value is 0. For this reason, the image isn't the nicest to look at, as there are trends to look at which seem to be going in the opposite direction but they are tending toward the optimal score for that scoring metric.