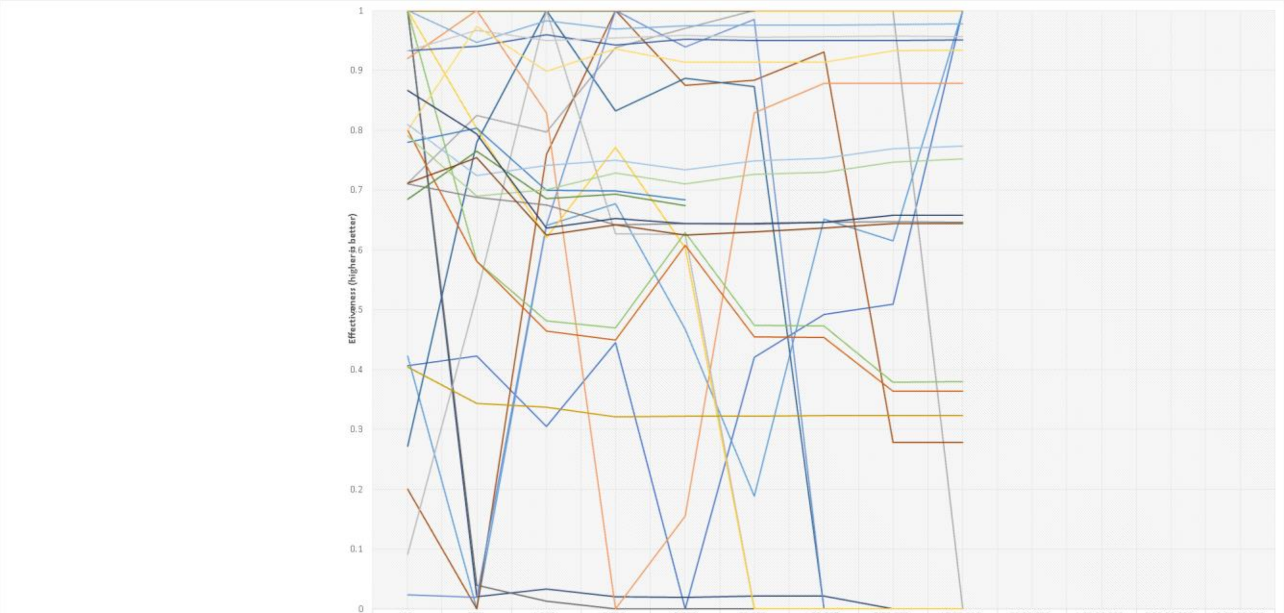
The (Un?) Reasonable Effectiveness of Data: Report

Team: team_26
Student IDs: 17306092, 17317559, 17302431
Total Time Required (in hours): 45 hours



	100	500	1,000	5,000	10,000	50,000	1,00,000	5,00,000	10,00,000	50,00,000	1,00,00,000	5,00,00,000	10,00,00,000
Linear Regression -- The SUM Dataset (without noise) -- RMSE(Normalised)	0.406660921	0.422984491	0.304638613	0.445429398	0	0.420331911	0.492461683	0.509511651	1				
Linear Regression -- The SUM Dataset (without noise) -- R2 Score (Normalised)	1	1	1	1	1	1	1	1	1				
Linear Regression -- The SUM Dataset (with noise) -- RMSE(Normalised)	0.711401636	0.825362305	0.797414409	0.938506703	0.970457143	1	0.999043035	0.999140644	0				
Linear Regression -- The SUM Dataset (with noise) -- R2 Score (Normalised)	0.999519123	0.9997003	0.999705288	0.999669465	0.999662138	0.999655868	0.999655271	0.999654472	0.999654931				
Linear Regression -- New York Taxi Dataset -- RMSE(Normalised)	0.423238057	0	0.640576519	0.677752835	0.468135608	0.188226547	0.652015385	0.615153638	1				
Linear Regression -- New York Taxi Dataset -- R2 Score (Normalised)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00				
Linear Regression -- Year Prediction MSD -- RMSE(Normalised)	1	0.020616333	0.033211074	0.020470007	0.019706167	0.021614366	0.022024584	0	0				
Linear Regression -- Year Prediction MSD -- R2 Score (Normalised)	0.20048198	0	0.759529653	1	0.874734671	0.884074422	0.930882513	0.277919369	0.277919369				
Random Forest -- The SUM Dataset (without noise) -- RMSE(Normalised)	1	0.040011797	0.013172542	0.000432775	0.000109427	0	0	0	0				
Random Forest -- The SUM Dataset (without noise) -- R2 Score (Normalised)	1.00	1.00	1.00	1.00	1.00								
Random Forest -- The SUM Dataset (with noise) -- RMSE(Normalised)	0.27251829	0.779655598	1	0.83253477	0.887207497	0.873200692	0	0	0				
Random Forest -- The SUM Dataset (with noise) -- R2 Score (Normalised)	1.00	1.00	1.00	1.00	1.00	1.00							
Random Forest -- New York Taxi Dataset -- RMSE(Normalised)	0.024181503	0.01961832	0.642037437	1	0.939484817	0.985116401	0	0	0				
Random Forest -- New York Taxi Dataset -- R2 Score (Normalised)	0.920176787	1	0.829346218	0	0.155072007	0.828621385	0.878571971	0.878571971	0.878571971				
Random Forest -- Year Prediction MSD -- RMSE(Normalised)	0.091388429	0.524715868	1	0.627291141	0.626083881	0	0	0	0				
Random Forest -- Year Prediction MSD -- R2 Score (Normalised)	1	0.803917823	0.620469766	0.771427695	0.602974771	0	0	0	0				
Logistic Regression -- The SUM Dataset (without noise) -- Accuracy	1.00	0.95	0.98	0.97	0.97	0.98	0.98	0.98	0.98				
Logistic Regression -- The SUM Dataset (without noise) -- Precision Score	1.00	0.58	0.48	0.47	0.63	0.47	0.47	0.38	0.38				
Logistic Regression -- The SUM Dataset (with noise) -- Accuracy	0.93	0.94	0.96	0.94	0.95	0.95	0.95	0.95	0.95				
Logistic Regression -- The SUM Dataset (with noise) -- Precision Score	0.80	0.58	0.46	0.45	0.61	0.45	0.45	0.36	0.36				
Logistic Regression -- New York Taxi Dataset -- Accuracy	0.71	0.69	0.68	0.64	0.64	0.64	0.65	0.65	0.65				
Logistic Regression -- New York Taxi Dataset -- Precision Score	0.41	0.34	0.34	0.32	0.32	0.32	0.32	0.32	0.32				
Logistic Regression -- Year Prediction MSD -- Accuracy	0.78	0.804	0.7	0.6984	0.6841								
Logistic Regression -- Year Prediction MSD -- Precision Score	0.684325397	0.765445604	0.685409491	0.692984175	0.673681045								
Decision Tree Classification -- The SUM Dataset (without noise) -- Accuracy	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00				
Decision Tree Classification -- The SUM Dataset (without noise) -- Precision Score	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00				
Decision Tree Classification -- The SUM Dataset (with noise) -- Accuracy	0.93	0.97	0.95	0.95	0.96	0.96	0.96	0.96	0.96				
Decision Tree Classification -- The SUM Dataset (with noise) -- Precision Score	0.80	0.97	0.90	0.94	0.91	0.91	0.91	0.93	0.93				
Decision Tree Classification -- New York Taxi Dataset -- Accuracy	0.81	0.72	0.74	0.75	0.73	0.75	0.75	0.77	0.77				
Decision Tree Classification -- New York Taxi Dataset -- Precision Score	0.79	0.69	0.70	0.73	0.71	0.73	0.73	0.75	0.75				
Decision Tree Classification -- Year Prediction MSD -- Accuracy	0.866666667	0.793333333	0.636666667	0.652666667	0.644666667	0.643666667	0.646066667	0.657753333	0.657984269				
Decision Tree Classification -- Year Prediction MSD -- Precision Score	0.711538462	0.753971572	0.625113122	0.642338848	0.625355135	0.630677618	0.636644653	0.644455957	0.644483285				

Dataset size

Findings/Answer

Question 1: To what extent does the effectiveness of machine-learning algorithms depend on the size and complexity of the data? [200-300 words]

The effectiveness of machine-learning algorithms depends to a vast extent on the size of the data. From the chart given above, it is observed that the prediction metrics vary with the number of data-points taken into consideration. Although, it cannot be conclusively said that the higher the size of the data, the better the effectiveness of machine-learning algorithms. The chart clearly depicts instances where the size of the dataset has noticeable effect on the effectiveness but there are instances where a larger dataset has resulted in the lowering of effectiveness of the algorithm. The effectiveness of the algorithm depends on the complexity of the problem i.e. how a dependent variable is related to the independent variable. There needs to be enough data to capture this relationship that may or may not exist between input and output variables.

From the above chart, it was also observed that the effectiveness of the algorithm also depends on the complexity of the data. It was observed that for data without noise, the predictions were accurate or close to accurate thus increasing effectiveness. From our observation, if a dataset contains many number of features that are irrelevant to the data that is to be predicted, the prediction quality of the model decreases. Hence, simply adding features which are not relevant to the output, will reduce the effectiveness of the algorithm.

Question 2: Looking only at the performance of your best performing 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 could have performed. [100 words max]

a) Considering the best performing algorithm, in our case Decision Tree Classification, we could predict accurately, all the data in the test sets with 100% accuracy and precision. Hence, we can safely say that machine-learning was suitable approach to predict the data in test set from the given data in the training set.

b) A rule-based algorithm would have performed poorly as compared to Decision Tree Classification. One of the reasons is that the rules may miss scenarios or perform incorrectly due to missing/incorrect values. Machine Learning works on creating a model and few false cases would not affect predictions.

Conclusion: We concluded that the effectiveness does depend on the size and complexity of the data in results but are susceptible various factors like choice of dataset & choice of algorithm. Both underfitting and overfitting will decrease the performance of the machine learning algorithms and a right balance should be maintained on deciding the size of the data. Also, the proper selection of dependent variables to determine the target variable according to the algorithm used, plays an important role in determining the results.

Limitation:

1) Fewer data points in a dataset will generally give inappropriate results about the algorithm.
2) Business scenarios will play a crucial role in deciding the algorithm to be used.
3) Data pre-processing and cleaning is necessary for accurate result prediction.

Data, Algorithms, etc.

Algorithm 1	Linear Regression
Algorithm 2	Random Forest
Algorithm 3	Logistic Regression
Algorithm 4	Decision Tree Classification
Dataset 1	The SUM Dataset (without noise)
Dataset 2	The SUM Dataset (with noise)
Dataset 3	New York Taxi Data
Dataset 4	Year Prediction MSD
Metric 1	RMSE(Normalised)
Metric 2	R2 Score(Normalised)
Metric 3	Accuracy
Metric 4	Precision Score

Contributions (max. 200 words)

17306092 did research and findings to decide which classification algorithm to use. He also selected the metrics to use for classification algorithms after discussion with everyone in group.
17302431 did research and findings to decide which regression algorithm to use. He also selected the metric to use in regression algorithms after discussion with everyone in group.
1731755 did research on selecting the best possible dataset to maximize the tasks that we can complete for both regression and classification. He also captured data and helped prepare the report and chart.
We all divided the algorithms and datasets into individual tasks and each person coded, noted and shared their findings(metrics) with the team. We divided the work such that no one person was just working on one particular dataset or just one particular algorithm to help each of us to gain better understanding of the algorithms and get familiar with python and scikit. After completing all coding and after capturing all the necessary metrics, we all made the excel sheet, the chart and the report using the template provided and collectively answered the questions.

Additional Information

If you feel that any additional information is needed to understand your work, please provide it here.

The datasets and the metrics used have been provided in the table above and the excel sheet. Points to note are: -
1. The Sum (without noise) was used in Logistic Regression, but since the prediction was taking too long, we did not use KFold and used a 70:30 split.
2. The Sum (with noise) was used in Logistic Regression, but since the prediction was taking too long, we did not use KFold and used a 70:30 split.
3. The Sum (without noise) was used in Decision Tree Classification, but since the prediction was taking too long, we did not use KFold and used a 70:30 split.
4. The Sum (with noise) was used in Decision Tree Classification, but since the prediction was taking too long, we did not use KFold and used a 70:30 split.
5. New York Taxi Data was used in Logistic Regression and a column was added “long_or_short” which has a value of 1 when “trip_duration” is greater than 500, otherwise it is 0.
6. New York Taxi Data was used in Decision Tree Classification and a column was added “long_or_short” which has a value of 1 when “trip_duration” is greater than 500, otherwise it is 0.
7. In the New York Taxi Data, the variables “pickup_datetime” and “dropoff_datetime” were converted Unix timestamps. Also, a new variable ‘distance in kms’ was calculated on the basis of pickup and drop-off latitude and longitude.