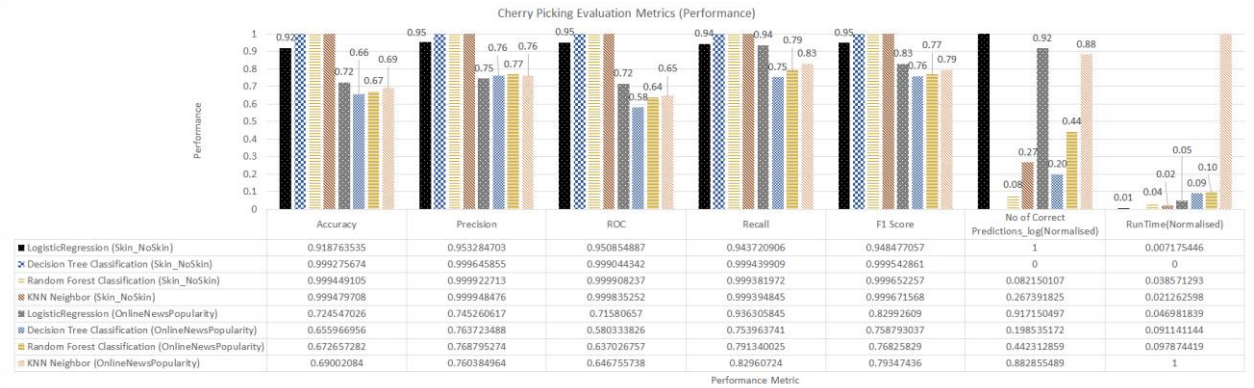
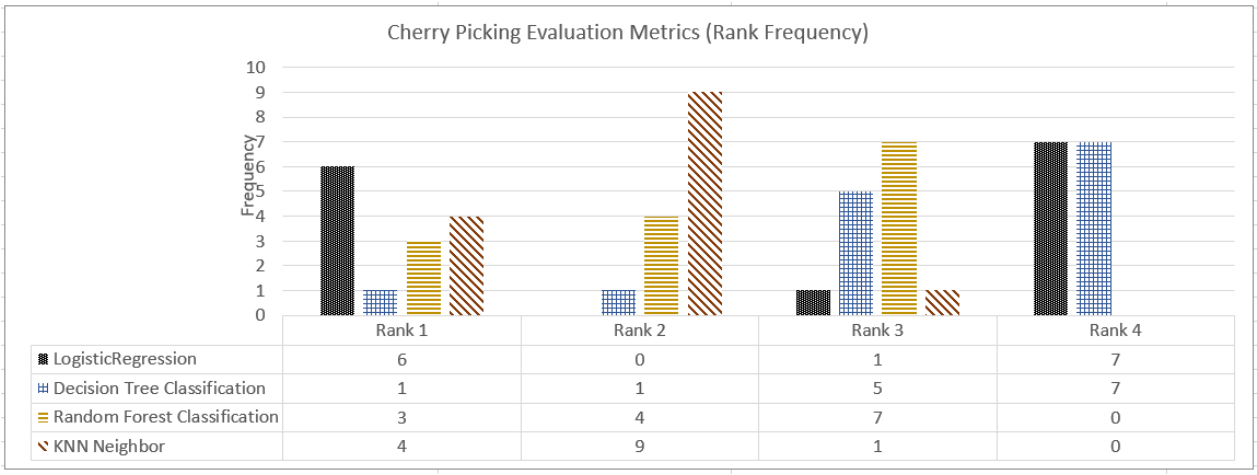
Cherry-Picking Evaluation Metrics: Report

Team: team_26
Student IDs: 17306092, 17302431, 17317559

Total Time Required (in hours): 30 hours



	Accuracy	Precision	ROC	Recall	F1 Score	No of Correct Predictions_log(Normalised)	RunTime(Normalised)
LogisticRegression (Skin_NoSkin)	0.918763535	0.953284703	0.950854887	0.943720906	0.948477057	1	0.007175446
Decision Tree Classification (Skin_NoSkin)	0.999275674	0.999645855	0.999044342	0.999439909	0.999542861	0	0
Random Forest Classification (Skin_NoSkin)	0.999449105	0.999922713	0.999908237	0.999381972	0.999652257	0.082150107	0.038571293
KNN Neighbor (Skin_NoSkin)	0.999479708	0.999948476	0.999835252	0.999394845	0.999671568	0.267391825	0.021262598
LogisticRegression (OnlineNewsPopularity)	0.724547026	0.745260617	0.71580657	0.936305845	0.82992609	0.917150497	0.046981839
Decision Tree Classification (OnlineNewsPopularity)	0.655966956	0.763723488	0.580333826	0.753963741	0.758793037	0.198535172	0.091141144
Random Forest Classification (OnlineNewsPopularity)	0.672657282	0.768795274	0.637026757	0.791340025	0.76825829	0.442312859	0.097874419
KNN Neighbor (OnlineNewsPopularity)	0.69002084	0.760384964	0.646755738	0.82960724	0.79347436	0.882855489	1



	Rank 1	Rank 2	Rank 3	Rank 4
LogisticRegression	6	0	1	7
Decision Tree Classification	1	1	5	7
Random Forest Classification	3	4	7	0
KNN Neighbor	4	9	1	0

Findings/Answer (200-300 words)

1. How meaningful are evaluation metrics (“meaningful” in terms of how consistent are they in assessing the performance of machine-learning algorithms)? [200-300 words]

According to our findings on the basis of the above task, we obtained quite conclusive results that evaluation metrics are meaningful indicators to assess the performance of machine-learning algorithms. From the above given charts and data, we can see Decision Tree Classification is being indicated as the worst algorithm by nearly all of the metrics, which shows consistent behavior throughout. Also, K-Nearest Neighbor algorithm has been consistently ranked high by most of the metrics. Both the above examples show that evaluation metrics are indeed quite consistent in ranking the performance of an algorithm. However, there are exceptions when a particular metric might give inaccurate results. In our task, we found this inconsistency in logistic regression which we had randomly selected as our novel

algorithm. For example, accuracy for linear regression was best for one of the dataset, but was the lowest for the other dataset. This gives contradictory information about the algorithm whether it performs well or not and how it can be improved. Hence, dataset plays an important role in evaluation of metrics.
To substantiate the above argument, we created two metrics. The first metric (Number of correct prediction) shows that linear regression is best performing algorithm for both the datasets and that Decision Tree Classification is the worst performing algorithm for both the datasets. This shows that the new metric is giving quite consistent results. However, the second metric(Runtime), ranks linear regression as one of the worst performing algorithms.
Conclusion: We concluded that evaluation metrics are generally consistent in results but susceptible to wide variations due to factors like choice of dataset & choice of algorithm. Hence, consideration should be given to these factors by using multiple metrics to judge an algorithm to get a proper result.
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.

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. OnlineNewsPopularity dataset was used in Logistic Regression, Decision Tree Classification, Random Forest Classification and KNN Neighbors. The target variable ' shares' was replaced by 'High and Low Shares' which has a value of '1' when 'Number of shares' is greater than or equal to 1000, otherwise it is '0'.
2. Two new evaluation metrics were created: a) Number of correct predictions b) RunTime

Data, Algorithms, etc.

Novel Algorithm	LogisticRegression
Baseline Algorithm 1	Decision Tree Classification
Baseline Algorithm 2	Random Forest Classification
Baseline Algorithm 3	KNN Neighbor
Dataset 1	Skin_NoSkin
Dataset 2	OnlineNewsPopularity
Common Metric 1	Accuracy
Common Metric 2	Precision
Common Metric 3	ROC
Common Metric 4	Recall
Common Metric 5	F1Score
Made Up Metric 1	No of Correct Predictions_log(Normalised)
Made Up Metric 2	RunTime(Normalised)

Contributions (max. 200 words)

17306092 developed wrapper to execute all algorithms. Played part in implementation of custom metrics. Helped in preparing the reports and charts.
17302431 did research and findings to decide which custom metrics and algorithm to use. He also and helped prepare the report and chart. Played part in implementation of custom metrics.
1731755 did research on selecting the best possible dataset to maximize the tasks that we can complete for classification. He also captured data and helped prepare the report and chart. Played part in implementation of custom metrics.

We all divided the algorithms and datasets into individual tasks and each person coded, noted and shared their findings(metrics) with the team. The task of creating a new metric was not done individually by just one person, it was discussed collectively and pseudo implementation was done as a group. 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.