Angela Scarano
04/15/2020
CMSC 471

Basic Random Forest Classifier Report

NOTE: I coded the homework before I realized there was attribute information given on the website where the data was retrieved. This chart notates which feature belongs to which attribute.

FEATURE	ATTRIBUTE
Feat 1	Age
Feat 2	Workclass
Feat 3	Fnlwgt
Feat 4	Education
Feat 5	Education-num
Feat 6	Marital-status
Feat 7	Occupation
Feat 8	Relationship
Feat 9	Race
Feat 10	Sex
Feat 11	Capital-gain
Feat 12	Capital-loss
Feat 13	Hours-per-week
Feat 14	Native-country

Angela Scarano
04/15/2020
CMSC 471

1.) Tweak RFC parameters and record the change and accuracy change in a table

Default	Accuracy = 0.8461 =0.8460782507217002
Max_Depth = 3	Accuracy = 0.8139 = 0.8138934954855352
Max_Depth = 40	Accuracy = 0.8416 = 0.8415944966525398
Max_Depth = 20	Accuracy = 0.8518 = 0.8517904305632332
N_estimators = 200	Accuracy = 0.8503 = 0.8502548983477674
N_estimators = 400	Accuracy = 0.8506 = 0.8505620047908605
N_estimators = 500	Accuracy = 0.8495 = 0.8495178428843437
N_estimators = 450	Accuracy = 0.8495 = 0.8494564215957251
Max_depth = 25	Accuracy = 0.8467 = 0.8466924636078865
Max_depth = 20, n_estimators = 400	Accuracy = 0.8582 = 0.8582396658681899
Max_depth = 17, n_estimators = 300	Accuracy = 0.8579 = 0.8579325594250967

2.) Check feature importances for the most accurate model, what does this tell you?

Feature Importances Printed (Features 1-14, left to right, top to bottom):

0.1201361	0.03483692	0.10425134	0.03831453	0.11347783
0.09599234	0.05943447	0.1250786	0.01312358	0.01573626
0.14147451	0.0438672	0.07772007	0.01655625	

This information suggests there is not one feature from the given list of features that is extremely important in determining income level, and there are a few that hardly influence that decision for the Random Forest Classifier (RFC) (Especially features 9, 10, and 14). There are a few that are each about 10-15% influential (Features 1, 3, 5, 6, 8, and 11) that appear significantly influential in the decision-making process.

Angela Scarano
04/15/2020
CMSC 471

3.) For each feature (there are about 14), remove them one at a time. Retrain the model with that feature missing. How much does accuracy dip for each one? What does this tell you about each feature? Make some conclusions about what features are important.

NOTE: The most accurate model (parameters Max_depth = 20, n_estimators = 400) was used for this chart

FEATURE REMOVED	ACCURACY	CHANGE FROM DEFAULT	% CHANGE
N/A	0.8568269762299613	0	0
Feat 1 (age)	0.8522203795835637	- 0.0046065966463976205	- 0.46%
Feat 2 (workclass)	0.8575640316933849	+ 0.0007370554634236193	+0.07%
Feat 3 (fnlwgt)	0.8544315459738345	- 0.0023954302561267626	-0.24%
Feat 4 (education)	0.8586696148885203	+ 0.0018426386585590482	+0.18%
Feat 5 (education-num)	0.8581168232909526	+ 0.0012898470609913337	+0.13%
Feat 6 (marital-status)	0.8559670781893004	- 0.0008598980406608892	-0.09%
Feat 7 (occupation)	0.8540015969535041	- 0.002825379276457207	-0.28%
Feat 8 (relationship)	0.8573183465389104	+ 0.0004913703089490795	+0.05%
Feat 9 (race)	0.8565198697868681	- 0.0003071064430931747	-0.03%
Feat 10 (sex)	0.8578711381364781	+ 0.001044161906516794	+0.10%
Feat 11 (capital-gain)	0.8394447515508875	- 0.0173822246790738	-1.74%
Feat 12 (capital-loss)	0.8511762176770469	- 0.005650758552914414	-0.57%
Feat 13 (hours-per-week)	0.8555985504575886	- 0.0012284257723726988	-0.12%
Feat 14 (native-country)	0.856704133652724	- 0.00012284257723726988	-0.12%

The % change in accuracy is somewhat parallel to the feature importance values above – when some features valued more important to the selection process were removed overall accuracy was decreased, while when some features valued less important to the selection process were removed overall accuracy increased. There are some exceptions to this: Features 5 (education-num) and 8 (relationship) were thought to be influential but removing them improved accuracy. The change for relationship could be attributed to random chance as it is so close to zero. Features 9 (race) and 14 (native-country) were expected by the RFC to not have much influence on determining income level, but removing them as features caused overall accuracy to decrease. As

with the removal of relationship, race had such a low change in accuracy it could be attributed to chance.

The largest impacts on accuracy based on feature removal were all negative changes. Removing age, capital-gain, or capital-loss resulted in the largest changes, with the removal of capital-gain resulting in a whopping 1.74% decrease in accuracy. Other notable changes to accuracy occurred when removing fnlwgt and occupation, also decreasing accuracy when removed. This leads me to suspect age, capital-gain, and capital-loss are important when predicting income levels, with occupation and fnlwgt having some (but less) importance in the prediction as well. Meanwhile race and relationship likely have little to no importance to predicting income given the comparatively small change in accuracy reported when those features were missing.

Angela Scarano
04/15/2020
CMSC 471

Extra Credit Report: Compare Random Forest to Other Classifier Algorithms

CLASSIFIER ALGORITHM	BEST CHANGES TO DEFAULT PARAMETERS	ACCURACY	ACCURACY %
Random Forest	Max_depth = 20 n_estimators = 400	0.8582396658681899	85.82%
K-Nearest Neighbors	N_neighbors = 20 P = 1	0.801486395184571	80.15%
Gaussian Naïve Bayes	No changes to default (no parameters able to be changed)	0.821018364965297	82.10%
Linear Discriminant Analysis	No changes to default (the changes did not change the accuracy)	0.8206498372335852	82.06%

Random Forest vs. K-Nearest Neighbors

Random Forest had an overall better accuracy than K Nearest Neighbors. Of all the algorithms tested, K Nearest Neighbors performed the worst with an accuracy just over 80%. K Nearest Neighbors assumes that the data points that are most like one another will share a label. This isn't necessarily true for the data used here – the feature importance check from question two showed there weren't many, if any, features that were highly important when selecting a label. By extension using similarity of these features between data points would not lead to accurate labeling, which explains why the K Nearest Neighbors algorithm performed the worst of the tested algorithms.

Random Forest vs. Gaussian Naïve Bayes

Random Forest had an overall better accuracy than Gaussian Naïve Bayes. Gaussian Naïve Bayes algorithms are more accurate than others and perform better when there is less data to work with, so with the large training data set used it's predictable that Random Forest would end up more accurate. Gaussian Naïve Bayes is still fairly accurate though – this could be attributed to the fact that many of the features listed are likely mutually independent of each other, a bonus for a Naïve Bayes algorithm.

Angela Scarano
04/15/2020
CMSC 471

Random Forest vs. Linear Discriminant Analysis

Random Forest had an overall better accuracy than Linear Discriminant Analysis (LDA). LDA algorithms make assumptions about the data used that are not true of our dataset. LDA assumes each feature in the data to have the same variance, and for a Gaussian distribution of those features. Also, any outliers in the data will skew the statistics used to separate the labels. The data used for this homework does not neatly meet those assumptions and explain why LDA performed worse than Random Forest.