DECISION TREES AND MAP REDUCE

Getting shaded under a decision tree

Qiushi Yan | KC Barrett

Agenda

1. DATA
2. MODELS
3. CODE
4. RESULTS

SECTION 1

Data

Brief Overview of Data:

Summary Statistics for Numerical Variables

Variable	Mean	Std. Dev.	Min	25%	50%	75%	Max
Age	38.58	13.64	17	28	37	48	90
Final Weight	189,778	105,550	12,285	117,827	178,356	237,051	1,484,705
Education Num.	10.08	2.57	1	9	10	12	16
Capital Gain	1,077.65	7,385.29	0	0	0	0	99,999
Capital Loss	87.30	402.96	0	0	0	0	4,356
Hours per Week	40.44	12.35	1	40	40	45	99

Brief Overview of Data:

Occupation, Marital Status, and Relationship

Occupation	Count
Prof-specialty	4,140
Craft-repair	4,099
Exec-managerial	4,066
Adm-clerical	3,770
Sales	3,650
Other-service	3,295
Machine-op-inspct	2,002
? (Unknown)	1,843
Transport-moving	1,597
Handlers-cleaners	1,370
Farming-fishing	994
Tech-support	928
Protective-serv	649
Priv-house-serv	149
Armed-Forces	9

Marital Status	Count
Married-civ-spouse	14,976
Never-married	10,683
Divorced	4,443
Separated	1,025
Widowed	993
Married-spouse-absent	418
Married-AF-spouse	23

Relationship	Count
Husband	13,193
Not-in-family	8,305
Own-child	5,068
Unmarried	3,446
Wife	1,568
Other-relative	981

Brief Overview of Data:

Race, Sex, Working Class, and Education

Race	Count
White	27,816
Black	3,124
Asian-Pac-Islander	1,039
Amer-Indian-Eskimo	311
Other	271

Sex	Count
Male	21,790
Female	10,771

Workclass	Count
Private	22,696
Self-emp-not-inc	2,541
Local-gov	2,093
Unknown	1,836
State-gov	1,298
Self-emp-inc	1,116
Federal-gov	960
Without-pay	14

Education	Count
HS-grad	10,501
Some-college	7,291
Bachelors	5,355
Masters	1,723
Assoc-voc	1,382
11th	1,175
Assoc-acdm	1,067
10th	933
7th-8th	646
Prof-school	576
9th	514
12th	433
Doctorate	413
5th-6th	333
1st-4th	168
Preschool	51

Brief Overview of Data:

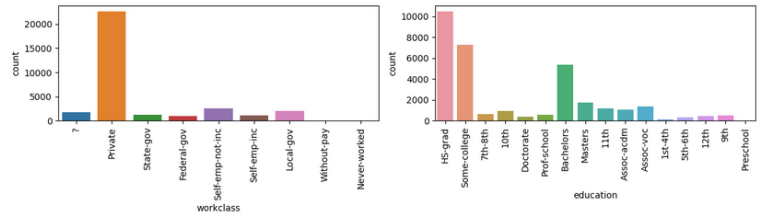
Native Countries

Native Country	Count
United-States	29,170
Mexico	643
Unknown	583
Philippines	198
Germany	137
Canada	121
Puerto-Rico	114
El-Salvador	106
Vietnam	67
Japan	62
Columbia	59

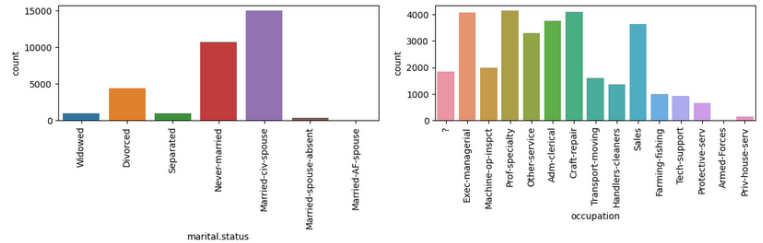
Native Country	Count
Haiti	44
Portugal	37
Peru	31
France	29
Ireland	24
Cambodia	19
Laos	18
Yugoslavia	16
Hungary	13
Scotland	12
India	100

Native Country	Count
Cuba	95
England	90
Jamaica	81
South	80
China	75
Italy	73
Dominican-Republic	70
Guatemala	64
Poland	60
Taiwan	51
Iran	43

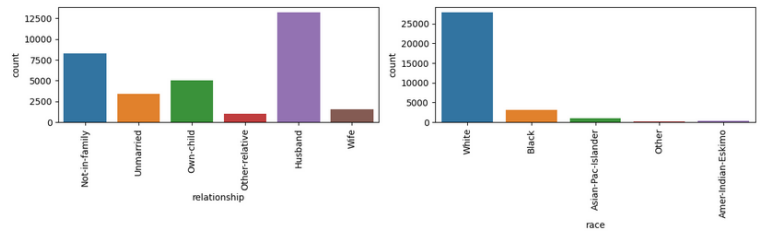
Native Country	Count
Nicaragua	34
Greece	29
Ecuador	28
Hong	20
Trinadad&Tobago	19
Thailand	18
Outlying-US	14
Honduras	13
Holand-Netherlands	1


count
workclass
?
Private
State-gov
Federal-gov
Self-emp-not-inc
Self-emp-inc
Local-gov
Without-pay
Never-worked

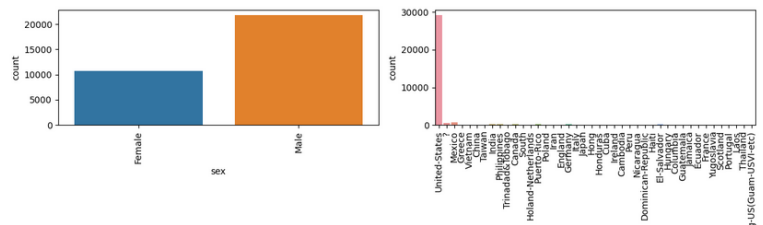
count
education
HS-grad
Some-college
7th-8th
10th
Doctorate
Prof-school
Bachelors
Masters
11th
Assoc-acdm
Assoc-voc
1st-4th
5th-6th
12th
9th
Preschool


count
marital.status
Widowed
Divorced
Separated
Never-married
Married-civ-spouse
Married-spouse-absent
Married-AF-spouse

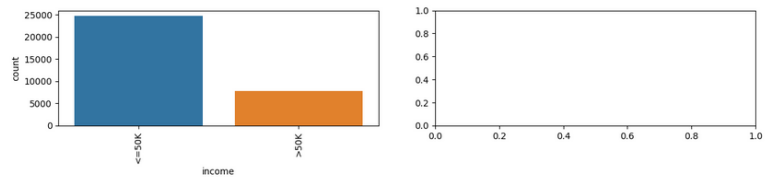
count
occupation
?
Exec-managerial
Machine-op-inspct
Prof-specialty
Other-service
Adm-clerical
Craft-repair
Transport-moving
Handlers-cleaners
Sales
Farming-fishing
Tech-support
Protective-serv
Armed-Forces
Priv-house-serv

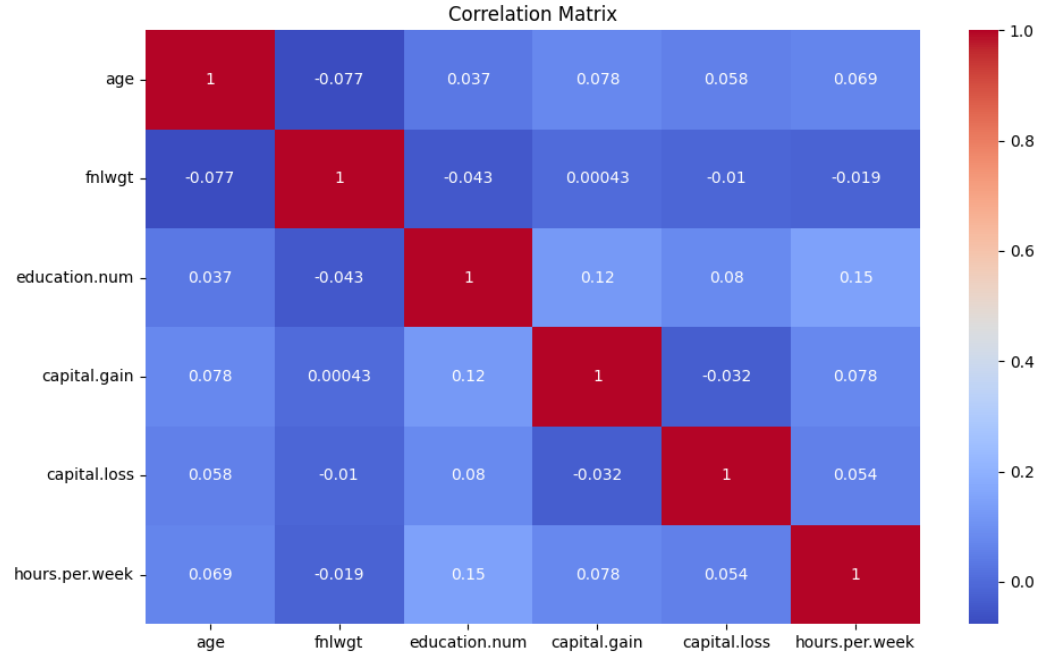

count
relationship
Not-in-family
Unmarried
Own-child
Other-relative
Husband
Wife

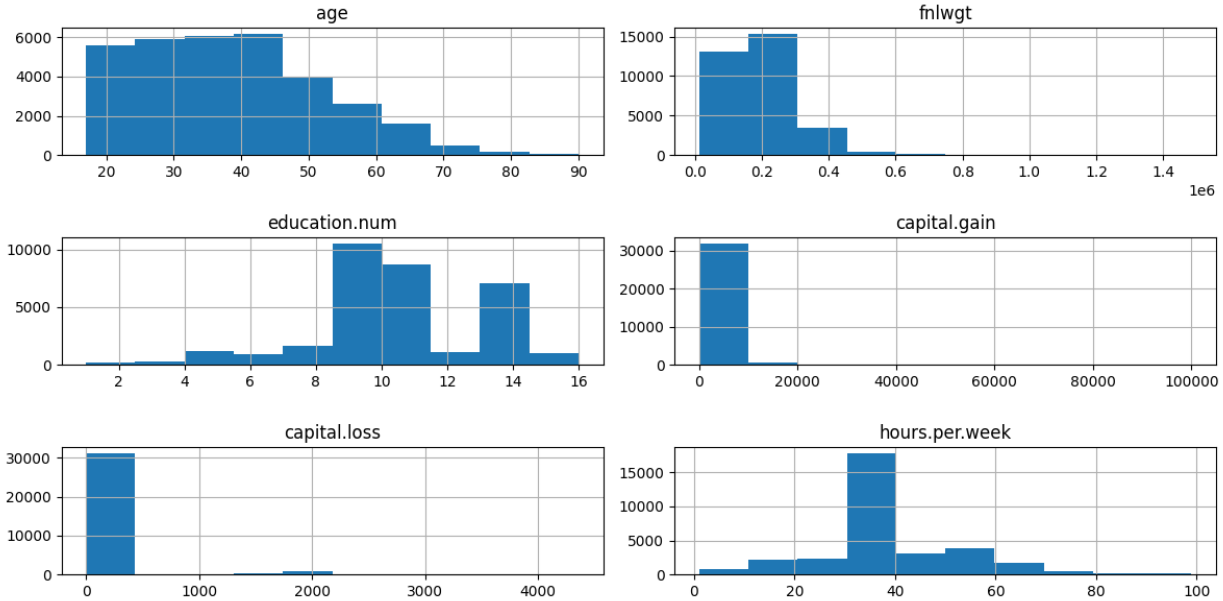
count
race
White
Black
Asian-Pac-Islander
Other
Amer-Indian-Eskimo


count
sex
Female
Male

count
native.country
United-States
Mexico
Greece
Vietnam
China
Taiwan
India
Philippines
Trinadad&Tobago
Canada
Portugal
Holand-Netherlands
Puerto-Rico
Poland
Iran
England
Germany
Italy
Japan
Hong
Honduras
Cuba
Ireland
Cambodia
Peru
Nicaragua
Dominican-Republic
Haiti
El-Salvador
Hungary
Columbia
Guatemala
Jamaica
Ecuador
France
Yugoslavia
Scotland
Portugal
Laos
Thailand
Outlying-US(Guam-USVI-etc)


count
income
<=50K
>50K


Correlation Matrix
age
fnlwgt
education.num
capital.gain
capital.loss
hours.per.week
1
-0.077
0.037
0.078
0.058
0.069
-0.077
1
-0.043
0.00043
-0.01
-0.019
0.037
-0.043
1
0.12
0.08
0.15
0.078
0.00043
0.12
1
-0.032
0.078
0.058
-0.01
0.08
-0.032
1
0.054
0.069
-0.019
0.15
0.078
0.054
1
1.0
0.8
0.6
0.4
0.2
0.0


age
fnlwgt
1e6
education.num
capital.gain
capital.loss
hours.per.week

SECTION 2

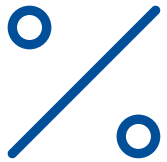
Models

Overview of Decision Tree Classifier

The Decision Tree Classifier is a powerful technique for classifying data, leveraging Gini Impurity and Entropy metrics to find the most effective split for the Decision Tree, and utilizing MapReduce for parallel processing of large datasets.

DECISION TREE:

A tree-like model used for classification tasks. Each internal node represents a decision based on a feature value, and each leaf node represents a class label.



GINI IMPURITY & ENTROPY:

Metrics used to measure the impurity of a set of labels. They are used to determine the best split for a decision tree.

MAPREDUCE:

A programming model for processing large datasets in parallel across a distributed cluster of computers.

DecisionTreeClassifier

Class Attributes

The `DecisionTreeClassifier` class implements a decision tree classifier for classification tasks and supports two training methods: standard single-threaded implementation and multi-threaded implementation with the MapReduce programming model.

- **MAX_DEPTH:**
  The maximum depth of the tree. If None, there is no limit.
- **METRIC:**
  The splitting criterion, either 'gini' for the Gini impurity or 'entropy' for the information gain.
- **SPLIT_METHOD:**
  The splitting method for continuous features. If 'chimerge', the ChiMerge algorithm is used.
- **CHIMERGE_THRESHOLD:**
  The threshold for the ChiMerge algorithm.
- **CHIMERGE_MAX_INTERVALS:**
  The maximum number of intervals for the ChiMerge algorithm.
- **N_WORKERS:**
  The number of worker processes used for the MapReduce implementation.

DecisionTreeClassifier

Class Methods

- **__INIT__:**
  Initializes a new instance of the DecisionTreeClassifier.
- **FIT(SELF, X, Y):**
  Fits the decision tree classifier on the given dataset.
- **FIT_MAPREDUCE(SELF, X, Y):**
  Fits the decision tree classifier on the given dataset using the MapReduce implementation.
- **PREDICT(SELF, X):**
  Predicts the class labels for the given dataset.
- **DRAW(SELF):**
  Draws the decision tree.
- **_PREPARE_FIT(SELF, X, Y):**
  Prepares the fitting process by checking the input data and setting some attributes.
- **_GROW_TREE(SELF, X, Y, DEPTH):**
  Grows the decision tree recursively.
- **_SHOULD_STOP(SELF, X, Y, DEPTH):**
  Determines whether to stop growing the tree at the current node.
- **_GROW_TREE_MAPREDUCE(SELF, X, Y, DEPTH, MASK):**
  Grows the decision tree recursively using the MapReduce implementation.
- **_BEST_SPLIT(SELF, X, Y):**
  Finds the best feature and threshold for the current split.

DecisionTreeClassifier

Class Methods continued

- **_BEST_SPLIT_MAPREDUCE(SELF, PARTITIONS):**
  Finds the best feature and threshold for the current split using the MapReduce implementation.
- **_BEST_SPLIT_MAPPER(SELF, PARTITION):**
  Maps the partition data to the best split scores for the partition.
- **_BEST_SPLIT_REDUCER(SELF, FEATURE_THRESHOLD_SCORES):**
  Reduces the partition scores to a single best split.
- **_SPLIT_SCORE(SELF, LEFT, RIGHT):**
  Calculates the split score for a given left and right dataset.
- **_SPLIT_SCORE_MAPREDUCE(SELF, LEFT, RIGHT):**
  Calculates the split score for a given left and right dataset for the MapReduce implementation.
- **_GET_THRESHOLDS(SELF, X, Y):**
  Gets the splitting thresholds for each feature.
- **_MOST_COMMON_LABEL(SELF, Y):**
  Finds the most common label in the target variable `y`.
- **_PREDICT(SELF, X):**
  Predicts the class label for a single instance.
- **_PARTITION_DATA(SELF, X, Y):**
  Partitions the dataset into smaller chunks for the MapReduce implementation.
  Initializes a new instance of the DecisionTreeClassifier.

RandomForestClassifier:

Overview of RandomForestClassifier

RandomForestClassifier is a powerful ensemble learning method for classification tasks that can be used to make accurate predictions.

1 **ENSEMBLE LEARNING METHOD FOR CLASSIFICATION TASKS**

RandomForestClassifier is an ensemble learning method that combines multiple decision trees to make predictions for classification tasks.

2 **CONSTRUCTS MULTIPLE DECISION TREES DURING TRAINING**

RandomForestClassifier builds multiple decision trees during the training process.

3 **OUTPUTS CLASS WITH MAJORITY VOTE FOR CLASSIFICATION**

RandomForestClassifier outputs the class with the majority vote from the decision trees for classification.

RandomForestClassifier:

Class Attributes

- **N_ESTIMATORS:**
  Number of decision trees in the ensemble. Icon keyword: tree
- **MAX_DEPTH:**
  Maximum depth of the trees. Icon keyword: depth
- **METRIC:**
  Splitting criterion (either 'gini' or 'entropy'). Icon keyword: split
- **N_WORKERS:**
  Number of worker processes for parallel training. Icon keyword: workers
- **SPLIT_METHOD:**
  Splitting method for continuous features. Icon keyword: method
- **CHIMERGE_THRESHOLD:**
  Threshold for the ChiMerge algorithm. Icon keyword: threshold
- **CHIMERGE_MAX_INTERVALS:**
  Maximum number of intervals for the ChiMerge algorithm. Icon keyword: intervals
- **MAX_FEATURES:**
  Number of features to consider when looking for the best split. Icon keyword: features
- **BOOTSTRAP:**
  Whether to use bootstrap samples for training individual trees. Icon keyword: bootstrap
- **RANDOM_STATE:**
  Random seed for reproducibility. Icon keyword: seed

RandomForestClassifier:

Class Methods

- **__INIT__():**
  Initializes a new instance of the RandomForestClassifier.
- **FIT(SELF, X, Y):**
  Fits the random forest classifier on the given dataset.
- **PREDICT(SELF, X):**
  Predicts the class labels for the given dataset.
- **PREDICT_PROBA(SELF, X):**
  Predicts the class probabilities for the given dataset.
- **_SAMPLE_FEATURES(SELF, X):**
  Samples the specified number of features to be used in each decision tree.
- **_BOOTSTRAP_SAMPLE(SELF, X, Y):**
  Creates a bootstrap sample of the input dataset by randomly sampling with replacement.
- **_FIT_SINGLE_TREE(SELF, X, Y):**
  A helper method that creates and fits a single decision tree with the given parameters.

SECTION 3

Code

Code:

The rpart package

Install !pip install rpart

Package directory contains subdirectories for data, examples, rpart models, and tests.

```
├── README.rst
├── data
│   └── adult.csv
├── examples
│   ├── fit_worker.ipynb
│   ├── mp.ipynb
│   ├── no-mp.ipynb
│   └── rf.ipynb
├── pyproject.toml
├── rpart
│   ├── DecisionTreeClassifier.py
│   ├── Node.py
│   ├── RandomForestClassifier.py
│   ├── __init__.py
│   ├── dataset.py
│   ├── entropy.py
│   └── utils.py
└── tests
    ├── __init__.py
    └── test_rpart.py
```

SECTION 4

Results

DecisionTreeClassifier had an accuracy of 83.8%

- **OVERALL ACCURACY OF 83.9%:**

The Decision Tree classifier correctly predicted 83.9% of the instances in the test set as either "<=50K" or ">50K" income categories.

- **HIGH PRECISION AND RECALL FOR <=50K:**

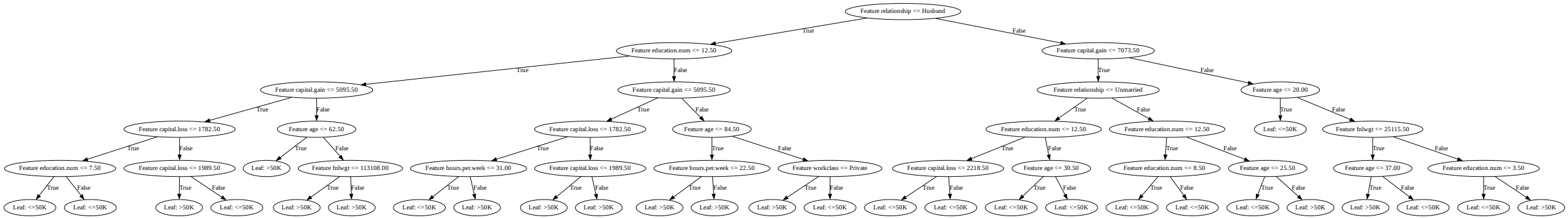
The model demonstrated high precision (0.85) and recall (0.95) for the "<=50K" class, indicating that it performed well in identifying and classifying true positives within this income category.

- **LOWER PERFORMANCE FOR >50K:**

The model showed lower precision (0.77) and recall (0.47) for the ">50K" class, suggesting that it struggled to accurately identify and classify true positives within this income category, and thus had a lower F1-score (0.58) compared to the "<=50K" class.

	precision	recall	f1-score	support
<=50K	0.85	0.95	0.9	7,429
>50K	0.77	0.47	**0.58**	2,340
accuracy			0.84	9,769
macro avg	0.81	0.71	0.74	9,769
weighted avg	0.83	0.84	0.82	9,769

Graph of DecisionTreeClassifer



RandomForestClassifier had an accuracy of 76.5%

- **OVERALL ACCURACY OF 76.6%:**

The Random Forest Classifier correctly predicted 76.6% of the instances in the test set as either "<=50K" or ">50K" income categories.

- **HIGH PRECISION BUT LOW RECALL FOR >50K:**

The model demonstrated perfect precision (1.00) for the ">50K" class, indicating that all instances it predicted as ">50K" were correct. However, the recall was extremely low (0.02), meaning it identified and classified only a small fraction of true positives within this income category, resulting in a very low F1-score (0.04).

- **HIGH RECALL BUT LOWER PRECISION FOR <=50K:**

The model showed high recall (1.00) for the "<=50K" class, capturing all true positives in this income category. The precision was slightly lower (0.76), which indicates that some instances were falsely classified as "<=50K". The overall F1-score for this class was relatively high (0.87) due to the high recall.

	precision	recall	f1-score	support
<=50K	0.76	1	0.87	7,429
>50K	1	0.02	0.04	2,340
accuracy			0.77	9,769
macro avg	0.88	0.51	0.45	9,769
weighted avg	0.82	0.77	0.67	9,769