Online Shopper Intention Analysis using Clustering and Classification techniques

Junyi Li

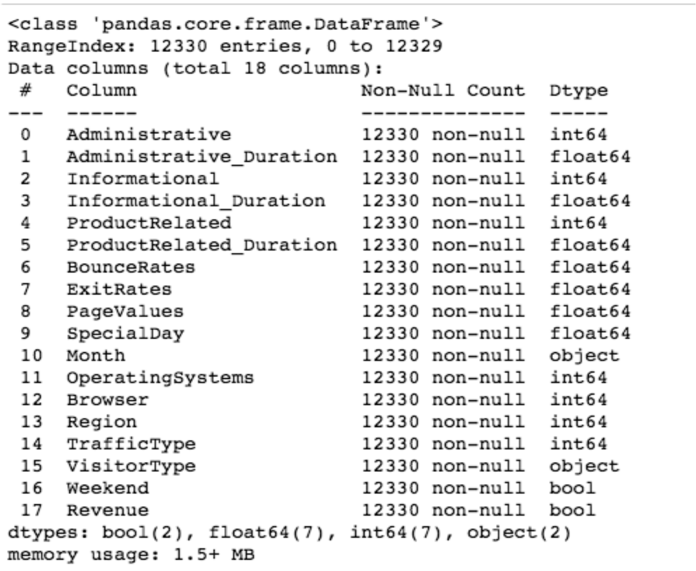
Zhou, Chuxiang

Abstract

The dataset is about the consumers shopping action. We need to determine whether a consumer will make purchase or not. In order to make machine learning report, we first clear data, including removing outliers and checking duplicate data and missing values. Then, we use clustering and classification to analysis data and then compare their accuracies, made confusion matrixes and compare AUROC values. Finally, we find that Tuning Hyperparameter has the highest accuracy and AUROC value.

1. Introduction

Today, more and more people use internet to purchase products. The factors that have high correlation with revenue is important for marketing strategies. This shopper dataset was from UCI Machine Learning Repository and it consists of 10 numerical and 8 categorical attributes and 8 categorical attributes. The 'Revenue' attribute can be used as the class label y.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12330 entries, 0 to 12329
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Administrative           12330 non-null  int64
 1   Administrative_Duration  12330 non-null  float64
 2   Informational            12330 non-null  int64
 3   Informational_Duration   12330 non-null  float64
 4   ProductRelated           12330 non-null  int64
 5   ProductRelated_Duration  12330 non-null  float64
 6   BounceRates              12330 non-null  float64
 7   ExitRates                12330 non-null  float64
 8   PageValues               12330 non-null  float64
 9   SpecialDay               12330 non-null  float64
 10  Month                    12330 non-null  object
 11  OperatingSystems         12330 non-null  int64
 12  Browser                  12330 non-null  int64
 13  Region                   12330 non-null  int64
 14  TrafficType              12330 non-null  int64
 15  VisitorType              12330 non-null  object
 16  Weekend                  12330 non-null  bool
 17  Revenue                  12330 non-null  bool
dtypes: bool(2), float64(7), int64(7), object(2)
memory usage: 1.5+ MB
```

Before we start analysis, we need to take a look at our dataset and their meanings, which are explained very clearly in later content. "Administrative" means page visited. "Administrative Duration" means time was spent. "Informational" means page visited. "Informational Duration" means time spent. "Product Related": means pages visited. "Product Related Duration" means time spent. "Bounce Rate" feature for a web page refers to the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session.

The value of "Exit Rate" feature for a specific web page is calculated as for all page views to the page, the percentage that were the last in the session. The "Page Value" feature represents the average value for a web page that a user visited before completing an e-commerce transaction.

The "Special Day" feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother's Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction. The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. For example, for Valentina's day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8.

The dataset also includes operating system, browser, region, traffic type, visitor type as returning or new visitor, a Boolean value indicating whether the date of the visit is weekend, and month of the year.

2. Data Preparation, Exploratory Data Analysis and Visualization

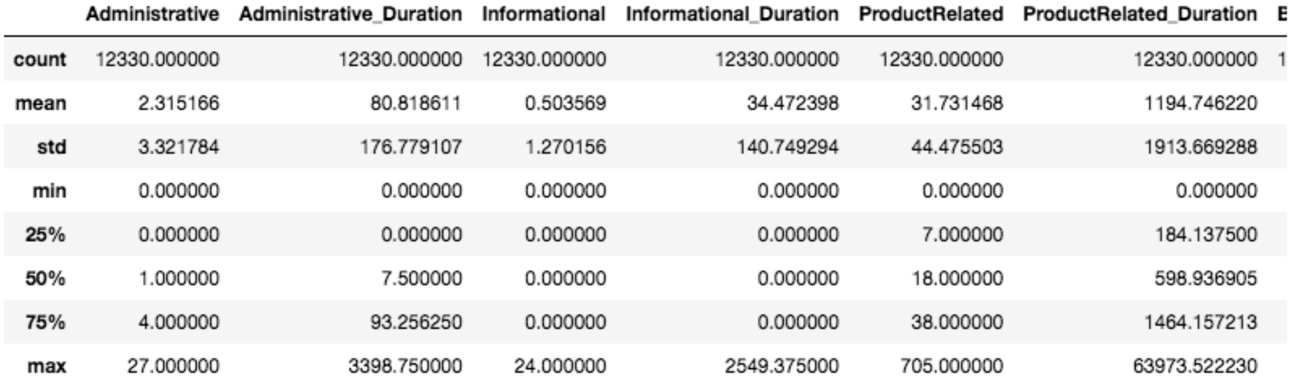
2.1 Missing value and Duplicate Value

First, we find that dataset has no missing values, since it was from UCI website and it may have been cleared by the publisher.

Second, we find that there are 125 duplicate rows. However, we decide not to delete them, become total dataset is 12,330 rows. The number of 125 duplicate rows is only small fraction of total dataset, and they may be accidently same as other rows, so there is no need to delete them.

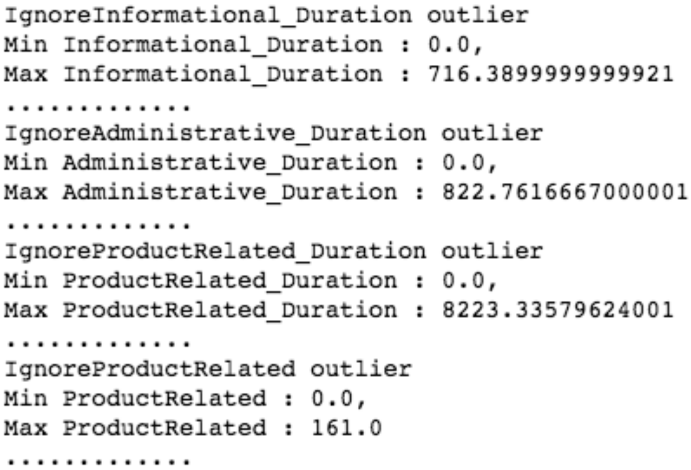
2.2 Outliers Removal

We check all the standard deviation of value of columns, we find the stds of columns related to duration time spent on website are very high.



	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	E
count	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	12330.000000	1
mean	2.315166	80.818611	0.503569	34.472398	31.731468	1194.746220	
std	3.321784	176.779107	1.270156	140.749294	44.475503	1913.669288	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	0.000000	7.000000	184.137500	
50%	1.000000	7.500000	0.000000	0.000000	18.000000	598.936905	
75%	4.000000	93.256250	0.000000	0.000000	38.000000	1464.157213	
max	27.000000	3398.750000	24.000000	2549.375000	705.000000	63973.522230	

For example, we can see that "Administrative_Duration" has standard deviation of 176.8 and that "Informational_Duration" has standard deviation of 140.749, and their max value is much higher than the value at 75%. To solve this outlier problem, we remove out top 1% value for columns with high "std". After this processing, these columns max value decrease a lot.



```
IgnoreInformational_Duration outlier
Min Informational_Duration : 0.0,
Max Informational_Duration : 716.3899999999921
.............
IgnoreAdministrative_Duration outlier
Min Administrative_Duration : 0.0,
Max Administrative_Duration : 822.7616667000001
.............
IgnoreProductRelated_Duration outlier
Min ProductRelated_Duration : 0.0,
Max ProductRelated_Duration : 8223.33579624001
.............
IgnoreProductRelated outlier
Min ProductRelated : 0.0,
Max ProductRelated : 161.0
.............
```

2.3 Boolean

For better analysis, We change the value of weekend and revenue from boolean to 1 or 0. (1-weekends and 1- means not weekend.)

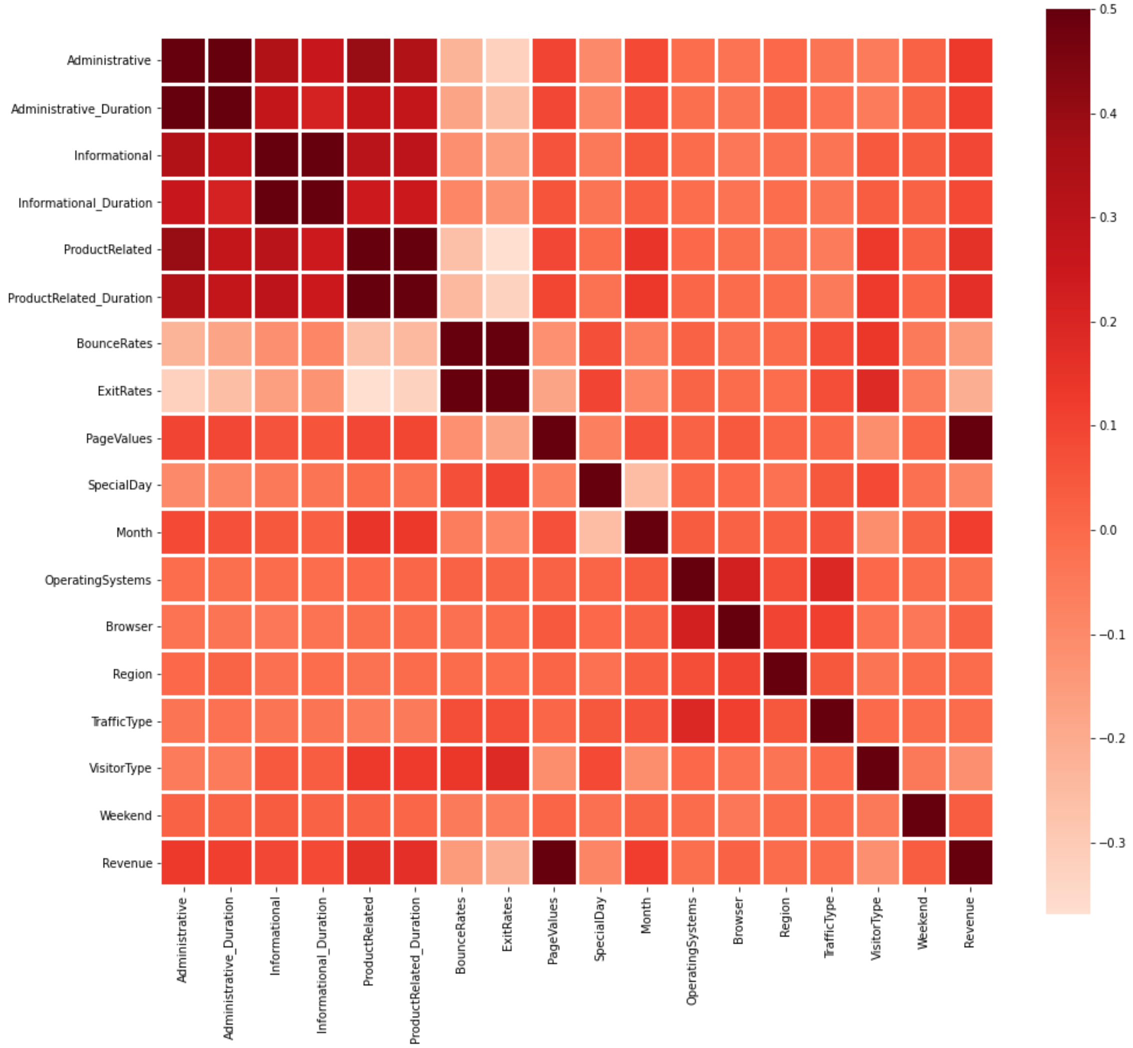
Besides, Visitor Type, which includes 'Returning_Visitor' and 'New_Visitor', have Boolean type, we change them to 1 and 0 respectively. (1- returning visitor and 0- new visitor.)

Last but not least, we change the month values from strings (Jan –Dec) to integer (1-12).

Those changes are better for later analysis.

2.4 Check correlation and delete low correlation columns

Revenue has high correlation with PageValues, which has dark red on their crossing point. Then it has moderate correlations with webpage visited and duration spent on the information ,administration, product related pages. For more precise detail analysis, we use corr() function.



From corr(), we find that the lowest correlations are TrafficType, Region, Operating Systems, Brower, Weekend. We decide to delete them.

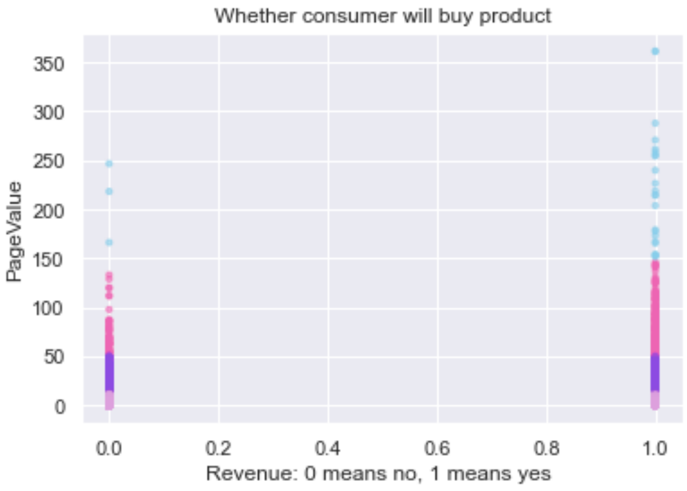
2.5 Dataset Standardized

Then, in order to better use clustering and classification, we use StandardScaler() to standardize dataset.

3. Clustering

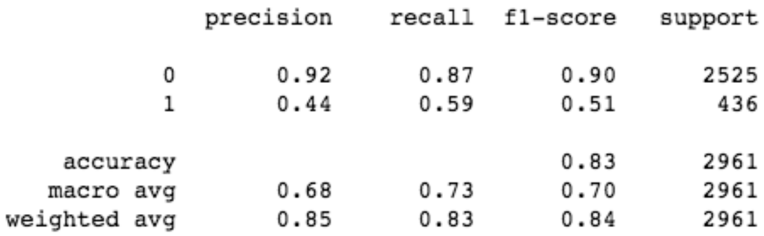
3.1 Kmean method(The best k)

We use the elbow method to identify optimal k, which is 4. Then we use this 4 to visualize Revenue and PageValues clusters. We can see that blue color cluster shows higher the Pagvalues, the more dots close to 1.



3.2 GaussianNB method

We split dataset 75% as training data, and 25% as testing data. The accuracy is 83%, but it has lower score for recall and precision.



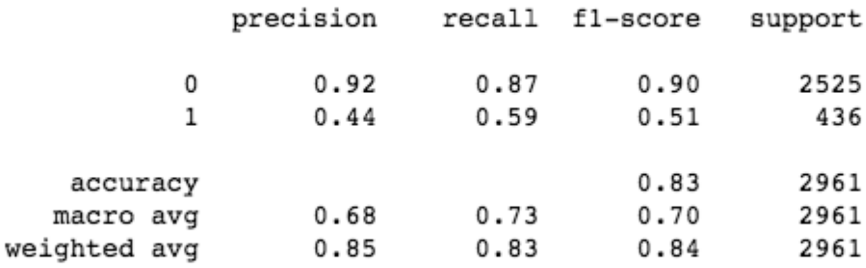
```
              precision    recall  f1-score   support

           0       0.92      0.87      0.90      2525
           1       0.44      0.59      0.51       436

    accuracy                           0.83      2961
   macro avg       0.68      0.73      0.70      2961
weighted avg       0.85      0.83      0.84      2961
```

4 Decision Tree and Tuning Hyperparameters

We use DecisionTreeClassifier() and GridSearchCV(). The accuracy is 90%.



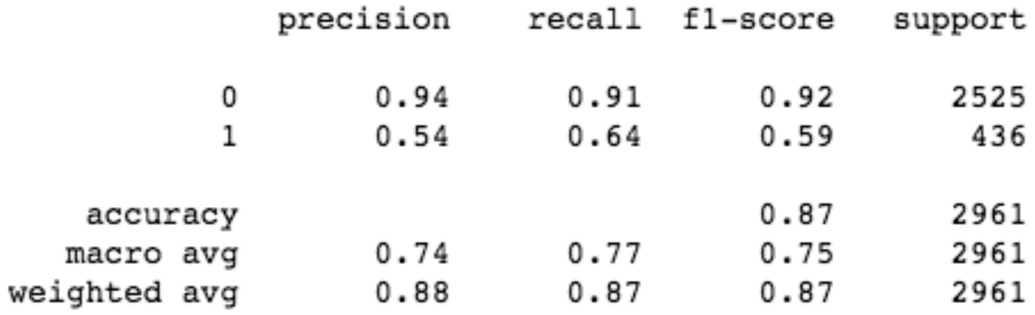
```
              precision    recall  f1-score   support

           0       0.92      0.87      0.90      2525
           1       0.44      0.59      0.51       436

    accuracy                           0.83      2961
   macro avg       0.68      0.73      0.70      2961
weighted avg       0.85      0.83      0.84      2961
```

5. Naive Bayes

We use BernoulliNB() to do anther classification. Its accuracy is 87% . According to its confusion matrix, it has low precision for predicting True for Revenue.



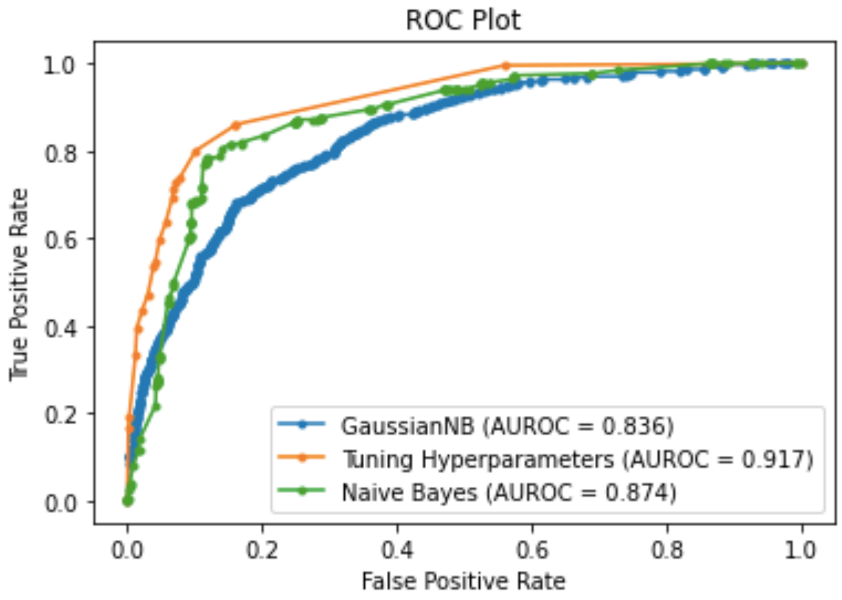
```
              precision    recall  f1-score   support

           0       0.94      0.91      0.92      2525
           1       0.54      0.64      0.59       436

    accuracy                           0.87      2961
   macro avg       0.74      0.77      0.75      2961
weighted avg       0.88      0.87      0.87      2961
```

6 ROC and AUC

As we know that The ROC curve summarizes the prediction performance of a classification model at all classification thresholds. Particularly, the ROC curve plots the False Positive Rate (FPR) on the X-axis and the True Positive Rate (TPR) on the Y-axis. From below picture, Tuning Hyperparameters has higher AUROC scores and it has higher accuracy among those models. It turns out to do the best.



Reference:

Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018). [Web Link]