Data 1030 Project – Customers' Intention

Qinyun Wu
Data Science Initiative, Brown University

https://github.com/Qinyun718/Data1030-OnlineShopping.git

12/03/2019

1 Introduction

The dataset is collected from Gözalan Group (http://www.gozalangroup.com.tr). It contains several related features to describe customers intention when shopping online. The target variable is the "Revenue" in the online shopping dataset. My goal is to apply classification in my project. During classification, the 'Revenue' attribute can be used as the class label, indicating whether the shopping site makes revenue in each day. In our daily life, we use online shopping even every day. For most of the time, we just look through the website and check for goods. Sometimes, we stay in one website for a long time but buy nothing. As a result, it is interesting if we could predict customers' final decision from their visiting information.

The dataset consists of 10 numerical and 8 categorical attributes:

1. The number of different types of pages: "Administrative", "Informational", "Product Related"
2. The total time spent: "Administrative Duration ", "Informational Duration", "Product Related Duration"
3. "Bounce Rate" : The percentage of visitors who enter the site but did not take any other requests.
4. "Exit Rate" : The percentage of the last views in the session in all pageviews
5. "Page Value" : average value for a web page before visitor making transactions
6. "Special Day": Whether it is a holiday
7. "Operating system", "browser", "region", "traffic type", "visitor type" are variables that could be easily understood from the name.

8. “Revenue”: the target variable, indicating whether the shopping sites make revenue

2 Exploratory Data Analysis (EDA)

Before applying any methods on the data set, the Exploratory Data Analysis helps us understand the features and the target variable, preparing for the future modelling.

In order to evaluate the accuracy of classification model, I calculated the baseline accuracy for the data set which is 0.845 and drew the bar plot of ”Revenue” (Figure 1).

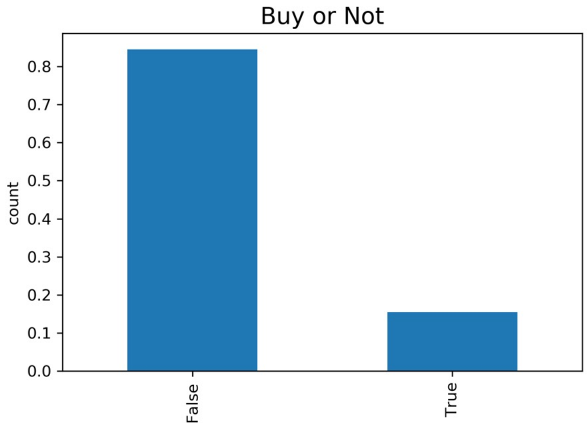

Figure 1: Revenue - histogram

In terms of numerical features, their distributions are important to be analyzed. Figure 2-4 shows the distribution for all of the numerical variables. Obviously, they are all right skewed distributions and concentrated on one extremely small values.

Other than numerical variables, categorical features are crucial for building the classification models. I plot the histograms for each of the categorical features (Figure 5-7). From the figures, it is obvious that revenue is not earned for most of the time, which is consistent with the balance of the whole data set. When taking a closer look to the first graph in Figure 5, we could tell the importance of the special day relative to customers’ intention of buying products. ”Special Day” means that how close the day customer entered in the website to one of the holidays. When ”Special Day” equals 0, it means that the day is the holiday. Then, we figured out that when the day is holiday, customers are

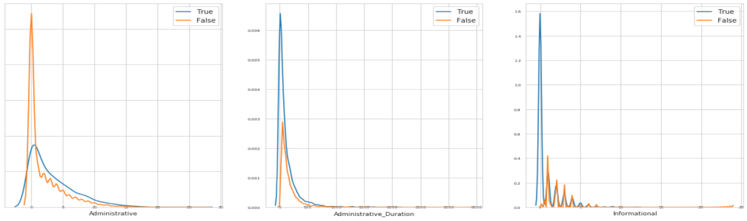

Figure 2: Numerical features - Distribution

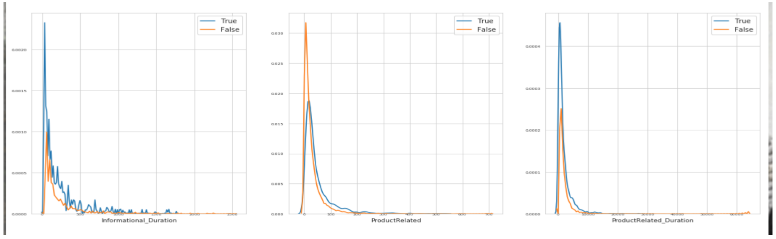

Figure 3: Numerical features - Distribution

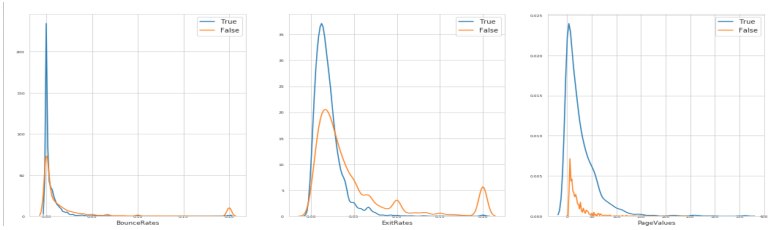

Figure 4: Numerical features - Distribution

much more likely to make the buying decision comparing to the other days. The result is interesting to be figured out and is also consistent with our expectations.

3 Methods

3.1 Preprocessing

Regarding data prepossessing, I applied MinMaxScaler on ”Administrative”, ”Informational”, ”Page Value” and ”Product related”, since their ranges are within some boundaries. In terms of those numerical features without specific boundaries, I applied StandardScaler, including ”Administrative duration”, ”Informational duration”, and ”Product related duration”. Besides, ”Bounce rate” and ”Exit rate” have already in the range of 0 to 1, so there is no need to apply any scalers on them.

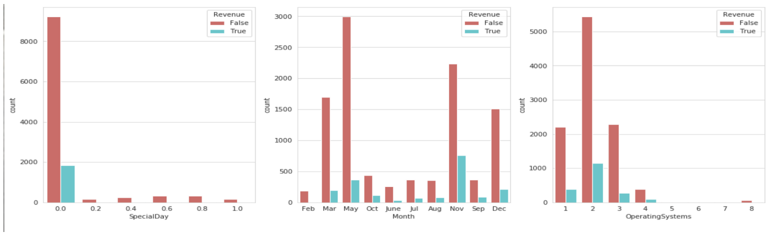

Figure 5: Categorical features - Histogram

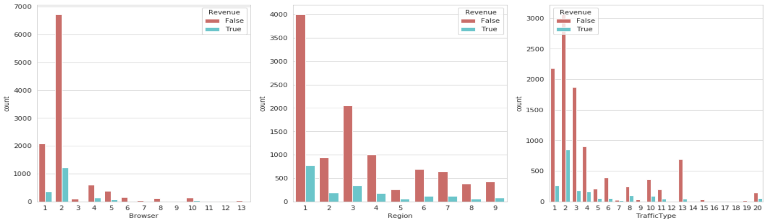

Figure 6: Categorical features - Histogram

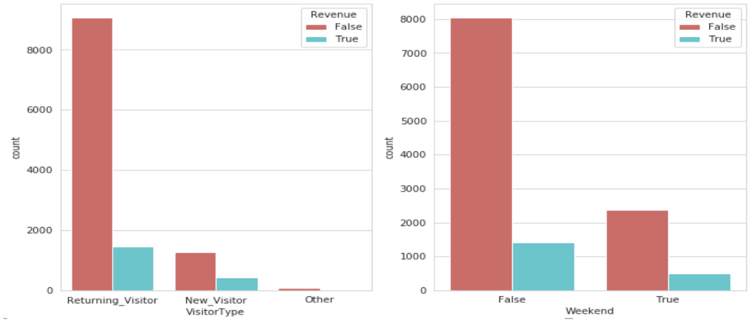

Figure 7: Categorical features - Histogram

In terms of categorical features (except taget variable), I applied OneHotEncoder on all of them, since they do not have priority over their labels. Finally, I applied LabelEncoder on "Revenue", using 0 for representing "False" and 1 for "True".

Then, the data set does contain missing values. After calculating the percentage of the observations with missing values, I got 0.001135 which is extremely small. As a result, I made the decision to drop those variables.

3.2 Cross Validation

The data set is split into two groups, including testing and training. Specifically, 20% of data use for testing and 80% for training. In terms of cross validation, I set the number of folds as five, meaning that among the training data, there are 20% of them are use as cross validation and rest of them for training.

3.3 ML methods

I developed three different methods to solve the classification problem. The first one is the logistic regression with l1 norm. When applying this classification method, I tuned alpha as the target parameter in order to find the one generating the highest score. In this example, I used the accuracy score to evaluate the effectiveness of the classification model, which will be compared with the baseline score in the result section. Specifically, I tried 20 values of alpha in the range of 10^{-5} to 10^{2}. The range guarantee that the best alpha is in the range but not on the limit.

Second, I used random forest to train the best classifier. Using random forest, I tuned both the depth of the tree and the minimum number of splits concurrently. Since random forest is insensitive with the preprocessing, I used the data before preprocessing in order to improve the efficiency of the model. In order to guarantee the best parameters are not on the limit, I set the depth in the range of 1 to 10 and the other one in the range of 3 to 15.

Then, I trained the classifier with Support Vector Machine, specifically, the SVC method. Under this method, I tuned both the values of gamma and C. C is set to be in the range of 10^{-3} to 10^{5} and gamma is set to be in the range of 10^{-10} to 10^{3}. Lastly, I used K-Nearest-Neighbors to train a classifier. I tuned the number of neighbors from 5 to 50 with a step of 5.

The reason why I choose accuracy as my evaluation metric is that it can tell us the effectiveness of the classifier most directly. Besides, I calculated the baseline accuracy score at the very beginning, so it is reasonable to compare the actual accuracy with the initial value.

In general, the measure uncertainties are most generated by the selected value of random state. With the consideration of the uncertainties, I trained all classifier with 5 different random states for each method.

4 Results

The baseline accuracy for this data set is around 0.8450. I used this baseline score to compare with the results from three different methods, drawing the bar plot to present their differences (Figure 8). The logistic regression with l1 norm generated an accuracy with the mean of 0.8820, while the random forest performed better with the accuracy of 0.8940. SVC generated a similar accuracy as random forest but a little bit higher, which is 0.8890. The accuracy of KNN classifier is 0.844, which is a little bit lower than the baseline. The result from KNN indicated that KNN is not a suitable method for the online shopping data set.

Since random forest generated the highest accuracy, I applied the global feature importance function to find the top ten most important features to interpret the classification models. Figure 9 provides the details about the important features when training a random forest classifier. I generated the bar plot several times with different parameter that I tuned from the method section. We could obtain the same selection of features as the most important features, although

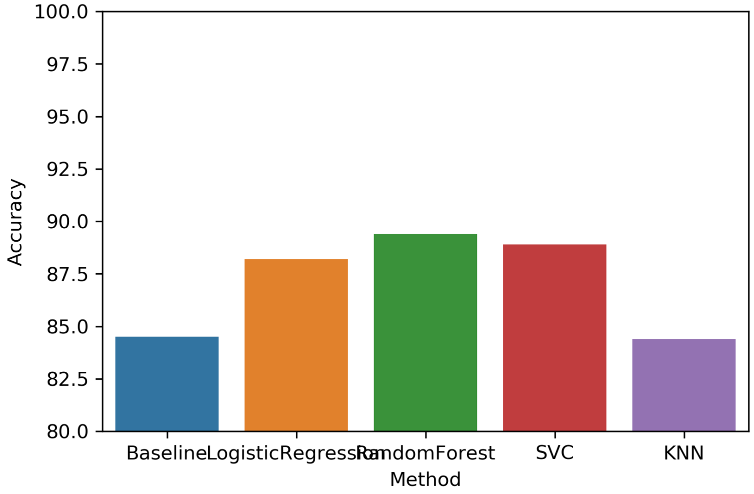

Figure 8: Accuracy Comparison

their values varies. From figure 9, we could conclude that ”Month_AUG” has the exceptional influence on the random forest classification model when comparing to any other features. In our real lives, summer vacation happens in August, indicating a potential peek season for shopping. As a result, the conclusion obtained from global feature importance is consistent with our real life experience. Other than that, ”ProductRelated” is a crucial feature as well as several one hot features related to ”Special day”. We could interpret the result that the number of pages related to products themselves that customers visited is highly correlated to their decision of buying or not. In terms of ”Special Day”, we expected that the value of 0.0 would be more important to the customers' decision in EDA section. Oppositely, the global feature importance shows that when the values of 0.4 and 1.0 performed more importantly than the value of 0.0. Additionally, ”informational”, ”Operating System_0.5” and ”informational Duration” are in the top ten important features.

Generally, all three of classification models improved to some extent from the baseline model. Therefore, it is reasonable to apply those classifier to the real life in order to predict and understand customers' intention if the important

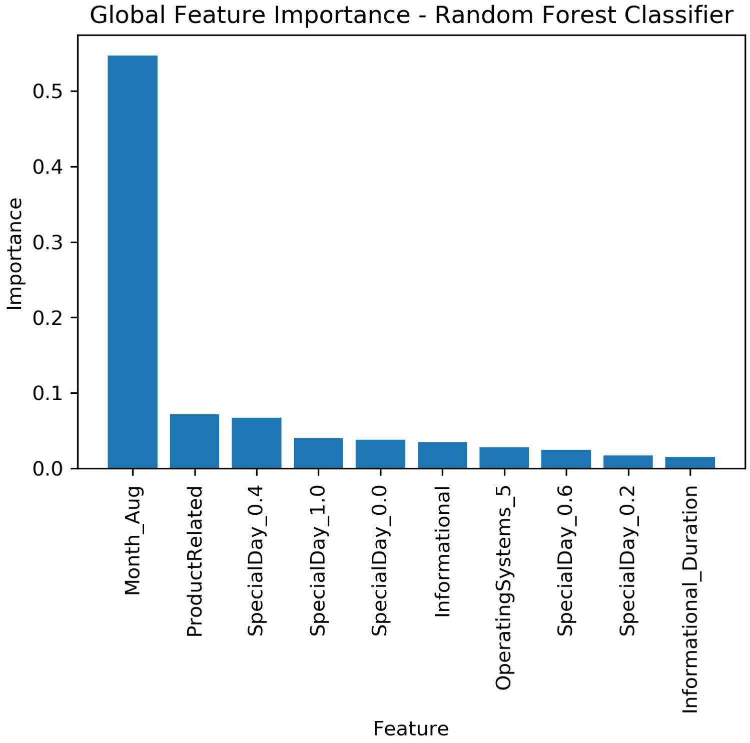

Figure 9: Global feature importance

features are given.

5 Outlook

Due to the time limitation, I only tried 5 random state for each model, indicating a potential limitation toward the final result. Besides, the data set is imbalanced. For future works, if the imbalanced problem could be solved, the results could be improved and become more generalized.

In addition to three methods I used above, XGB would be another option to train a classifier, which could potentially improve the accuracy result.

Then, the additional data could be collected to improve the classification model. For example, customers' intention might be influenced by the time period that they enter each website. Specifically, people tend to make buying decision in the evening rather than the early morning. Similarly, lots of features could be connected to help us understand customers' intention.

6 Reference

Sharma, R. (2019, May 23). Online Shopper's Intention. Retrieved from https://www.kaggle.com/roshansharma/online-shoppers-intention/kernels.

Roshansharma. (2019, October 30). Online Shopper's Intention. Retrieved from https://www.kaggle.com/roshansharma/online-shopper-s-intention.

Kageyama. (2019, June 25). [LGBM] Online Shopper's EDA and Classification. Retrieved from https://www.kaggle.com/kageyama/lgbm-online-shopper-s-eda-and-classification.

Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2019) 31: 6893. https://doi.org/10.1007/s00521-018-3523-0