Project Proposal

Describe the problem:

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 both regression and classification in my project. Since there are multiple features, some of them might not be as important and influential as others. As a result, the regression analysis and the model selection process could help me to identify the most crucial features, preparing for the classification. 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 figure out the relationship or any pattern between the time we spend on website and the final decision.

Describe the dataset:

The dataset consists of 10 numerical and 8 categorical attributes.

- The number of different types of pages: "Administrative", "Informational", "Product Related"
- The total time spent: "Administrative Duration ", "Informational Duration", "Product Related Duration"
- "Bounce Rate" : The percentage of visitors who enter the site but did not take any other requests.
- "Exit Rate" : The percentage of the last views in the session in all pageviews
- "Page Value" : average value for a web page before visitor making transactions
- "Special Day": Whether it is a holiday
- “Operating system”, “browser”, “region”, “traffic type”, “visitor type” are variables that could be easily understood from the name.
- “Revenue”: the target variable, indicating whether the shopping sites make revenue

There are several projects on Kaggle applied this dataset. One of them used the clustering method to study for users characteristics on shopping website in terms of the time spent on the specified website. The other one did the logistic regression by treating the target variable as the response variable, marking as 0s or 1s.

Preprocess the dataset:

Since the number of types of pages that customers visited has the limitation, I apply MinMaxEncoder to those variables, including "Administrative", "Informational", and "Product Related". Meanwhile, the total time spent on those websites is unlikely to have the upper limit. So, I apply StandardScaler to "Administrative Duration", "Informational Duration" and "Product Related Duration". "Bounce Rate" and "Exit Rate" refer to specific percentages, which should be dealt by MinMaxEncoder. Then, I cleaned the "Page Value" with StandardScaler. Besides, there are several categorical variables with no order. So, I applied OneHotEncoder to the rest of the variables except the target variable. I used LabelEncoder to the target variable to prepare for the classification. At the end, I got 81 features after preprocessing the data.

Reference:

https://www.kaggle.com/roshansharma/online-shoppers-intention

Github Link:

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