Project Proposal (Sep 30, 2019)

Shiqi Lei

Dataset: Online Shopper's Intention https://www.kaggle.com/roshansharma/online-shoppers-intention

Github Repository: https://github.com/beckyleii/data1030-f19-project.git

---

Problem to solve.

- **Describe the problem you want to solve**: What is the target variable? Is the problem regression or classification? Why is this interesting/important?

The target varibale of the dataset is the categorical featuer "revenue". The dataset consists feature vectors belongs to 12,330 sessions, and it was formed so that each session would belong to a different users in a 1-year period to avoid any tendency to a specific campaign, special day, user profile or period. Thus, I believe this dataset could provide reasonable conclusion based on my model.

One of the reason why I chose this dataset is because the e-commerce becomes more and more popular nowadays, and it becomes our life necessity especially for college students and people who are too busy to go shopping in stores. Thus, the dataset has great potential in analyzing the trend of decision making by visitors in making purchases, and the result might suggest what improvement the e-commerce site might do to stimulate the revenue.

In this project, I want to predict whether the revenue will be generated based on the duration and special period of the time, and I will also analyze the impact of region, browser and different e-commerce categories on the target variable 'Revenue'.

Describe the dataset

- **Number of data points and number of features**:

The dataset consist of 10 numerical and 8 categorical attributes. In this dataset, we are setting **'Revenue'** attributes as the target variable, which returns boolean value indicating whether the visitor generates revenue or not.

- **It feature is categorical, describe each category; if feature is numerical, include the unit of the quantity and what it measures**.

First we have three features: **"Administrative", "Informational","Product Related"**, which are three different type of pages that were visited by the visitors in that session, and respectively, there are three numerical data: **"Administrative Duration", "Informational Duration","Product Related Duration"**, which are the total time visitors spent in each of these page categories. The values are derived from the URL information and the information are updated when a user moving from one page to another.

Then there are three cetegorical data:

'Visitor Type', has sub-categories of 'returnning_visitor', 'new_visitor' or 'others',

and **'Weekend'**, returns boolean value indicating whether the date of the visit is weekend.

'Month', keeps track of the month of the year that the session occurs.

'Traffic Type',**'Region'**, **'Operating Systems'**, uses numerical values to represent different categories.

'Browser', indicate the type of browser that the visitor used; 'Operating System', 'Region' and 'Traffic Type' of the session.

Then we have **numercial features**:

'Bounce Rate', which refers to the percentage of visitors who enter the site from that categorical page and then leave ("bounce") without triggering another request to the analytics server during that session;

'Exit Rate', refers to the percentage that the specific web page were last in the session;

'Page Value', represent the average value for a web page that a user visited before completing an e-commerce transaction. (Note, three numerical metrics are measured by 'Google Analytics' for each page in the e-commerce site.)

'Special Day', measures 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.

- **If dataset is from Kaggle, write a short description about 2-3 public projects where the data has been used, and how the features were used**.

From the Kaggle public project, one of the project utilized the clustering analysis to learn the user characteristics in terms of time spent on the websites. And another public project uses the logistic regressing classification, however, the project did not have any information about how the data were sued and what result was derived from the data.

Proprocess the dataset

- Apply MinMaxEncoder or StandardScalar on the continuous features.
- Apply OneHotEncoder or OrdinalEncoder on categorical features.
- Apply the LabelEncoder on the target variable if necessary.
- Describe why you chose the preprocessor you used for each feature.
- How many features do you have in the preprocessed data? **There are 18 features in teh preprocessed data.**

As I go through the dataset, it seems that three durations are tailed distributed, and thus, we apply StandardScalar on the continuous features:"Administrative Duration", "Informational Duration", "Product Related Duration".

"Administrative", "Informational", "Product Related" are also continuous features containing the number of different types of pages visited by the visitor. Thus we can use MinMaxEncoder to fit the these three features.

In addition, we can fit "Bounce Rate", "Exit Rate", "Page Value" and "Special Day" using MinMax Encoder since they are percentage and average values that are reasonably bounded within certain interval.

"Browser", "Operating System", "Traffic type","vistor type" and "Region" are unordered categorical data, with "Weekend" as binary categorical data, all of these can use One Hot Encoder.

"Month" is a order categorical data, and thus we can use OrdinalEncoder to fit.

Finally, since the target variable "Revenue" is a binary categorical data that returns True and False, we can use LabelEncoder to fit it.