Raghad saleh alkhathran
Project proposal
21 October

Online Shoppers Intention

Abstract:

The goal of this project was to use classification models to predict the revnue of website in order to help improve the customers experience with the website, to gain more revenues. I worked with data provided by Kaggle, start with preparation the data then start using visualized to get deep understanding for the data, ending by build a classification model to predict my target column.

Introduction:

Now days data science become important aspect in different fields, as in Business field to understand the costumers. With so many potential sources of customer data, a foundational understanding of data science could help make sense of it, Making sense of data will reduce the horrors of uncertainty for organizations so, data science is important as a foundation for taking your businesses to the next level.

Design:

Since the customer is important, We need to study how they behave to make any future decision the origination will take.

In this project I analyzed ‘Online Shoppers Intention ‘data to try improve the customers experience with the website, to gain more revenues. By using the classification model to predict the revenue values. Also using the visualization of the data to answer the following questions:

- Is the weekend a factor to raise the revenue?
- the closeness of the site visiting time to a specific special day (e.g., Mother’s Day, Valentine's Day) effect the revenue?
- What features has the strong correlation with revenue?

Dataset overview:

The dataset is an open source from Kaggle represent user who visited a website. The dataset consists of 10 numerical and 8 categorical attributes and 12,330 rows. some features represent the page that user had been visit and the duration the suer spend on it. And other represent the metrics measured by "Google Analytics”. Also, the ‘Special Day’ feature one of the most important features when the user happens to visit the website in such a specific time like pre Mother’s Day. In addition to the target column ‘revenue’ where represents if the website gains a revenue from the visitor or not.

Data Link : https://www.kaggle.com/imakash3011/online-shoppers-purchasing-intention-dataset

Raghad saleh alkhathran
Project proposal
21 October

Algorithms:

Started work to clean the dataset to find any null value or inconsistent in the data and other cleaning process, then use the visualization Tanique to help understand the data before starting build the model.
Since my goal was to predict a categorical value so we need a classification. Support Vector Machine" (SVM) and logistic regression were used after splitting the data into train and test set.

Models :

"Support Vector Machine" (SVM): is a supervised machine learning algorithm that can be used for both classification or regression challenges.

Logistic regression: machine learning algorithms for binary classification. It is a simple algorithm that performs very well on a wide range of problems.

Model Evaluation :

The data was imbalance class so to handle this issue I used oversampling method to get more records.The entire training dataset after the oversampling was 15633 records. And before oversampling 8631. all scores reported below were calculated before and after the oversampling.

Before oversampling:

Logistic regression:

Accuracy for training set = 0.880
Accuracy for test set = 0.8834
train error = 0.119
test error = 0.116

Recall= 0.895
precision= 0.976
specifity= 0.732

Support Vector Machine (SVM):

Accuracy for training set = 0.880
Accuracy for test set = 0.883
train error = 0.153
test error = 0.150

Recall= 0.849
precision= 1.0

specifity= 1.0

after oversampling:

Logistic regression:

Accuracy for training set = 0.683
Accuracy for test set = 0.679
train error = 0.316
test error = 0.320

Recall= 0.6143996134331964
precision= 0.9720948012232415
specifity= 0.9319029850746269

Support Vector Machine" (SVM):

Accuracy for training set = 0.505
Accuracy for test set = 0.507
train error = 0.494
test error = 0.492

Recall= 0.504
precision= 1.0
specifity= 1.0

Tools:

Jupyter Notebook: Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations.

Python libraries:

- NumPy: the fundamental library for scientific computing in Python on which Pandas was built.
- Pandas: to do data manipulation and analysis.
- Matplotlib: for data visualization
- Seaborn: data visualization library for statistical graphics plotting in Python.
- imblearn.over_sampling: for SMOTE (is a type of data augmentation that synthesizes new samples from the existing ones).
- Sklearn: for import the model

Communication :

In addition to the slides and code, a Poster with summary of the result.

Raghad saleh alkhathran
Project proposal
21 October

Reference :

- https://online.hbs.edu/blog/post/what-is-data-science
- https://jupyter.org
- https://www.datacamp.com/community/blog/python-pandas-cheat-sheet?utm_source=adwords_ppc&utm_campaignid=12492439802&utm_adgroupid=122563403481&utm_device=c&utm_keyword=panda%20package%20python&utm_matchtype=b&utm_network=g&utm_adposition=&utm_creative=504158804617&utm_targetid=aud-1282405912307:kwd-614516587896&utm_loc_interest_ms=&utm_loc_physical_ms=9077037&gclid=CjwKCAjwqeWKBhBFEiwABo_XBpk2ARUn0ShWHZP8Ginyu4OJPPd5SD2O-ajSgdUwx_v5WUCPy6CTQRoCQrYQAvD_BwE
- https://www.mygreatlearning.com/blog/seaborn-tutorial/
- https://machinelearningmastery.com/logistic-regression-tutorial-for-machine-learning/
- https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/