Online Shoppers Intention : Project Overview

* Build a model to help Business team to know about the intention online shoppers about their website
* Dataset name obtained from Kaggle with name of dataset Online Shoppers Intention
* This Dataset have 12330 data
* From the start, we pick numeric and categoric column for look the big picture from dataset
* There’s no missing value and have 125 duplicated data
* In this model, replace other to Returning visitor and drop the new visitor. This model, focused on Returning Visitors
* Split data for pick the best model 70:30 and handle class imbalance
* Try with 5 modeling and done the hyperparameters for each modelling and pick the best model for this dataset.

Problem

* On an online shop website, have 12330 visitors, that include Returning and New visitor.
* On an online shop website, the Returning Visitor have lower conversion rate than the New Visitor.
* Based on data, There are many Returning Visitor visit that online shop website.

Objective:

* Make a modelling that can predict user who come to the online shop website and check out their shopping cart (purchase).

Purpose:

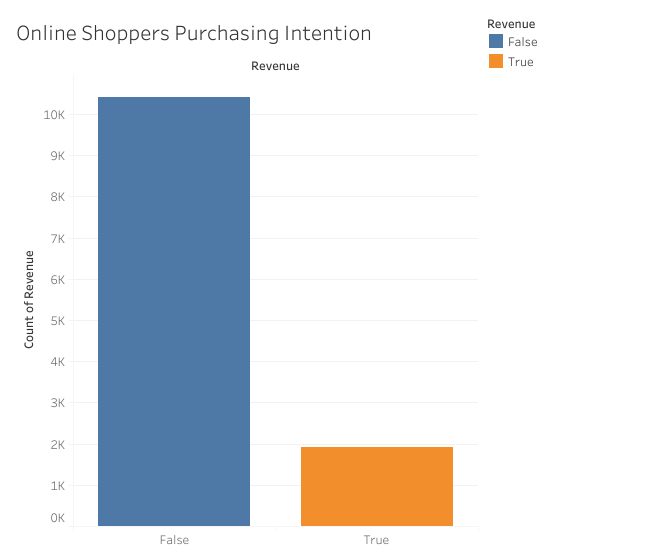
* Increase Revenue (Conversion rate) in particular for Returning Visitor by following up user with the prediction to check out their shopping cart (purchase).
* Call to action and giving the promo for the target.

Data:

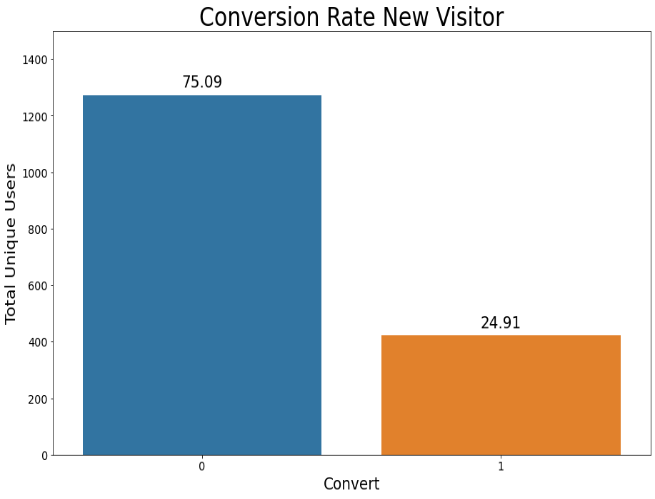
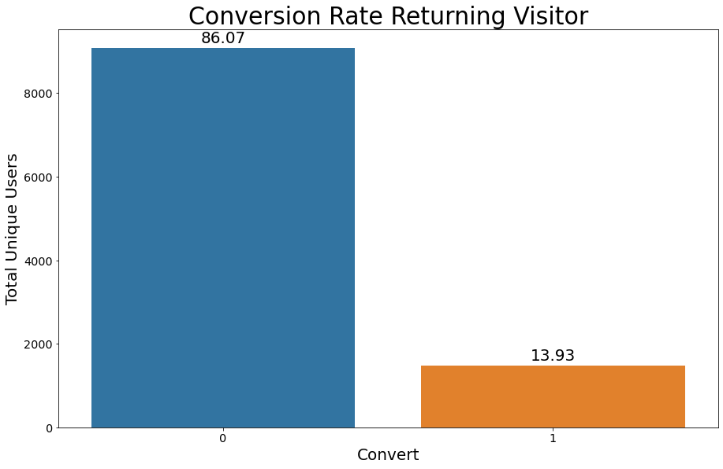
* 12330 rows, O null missing values, and 125 duplicates
* In this dataset, have 18 columns with details as follows:
  + 10 Numericals
  + 8 Categoricals
* For modelling, we choose 8 features with 1 target.

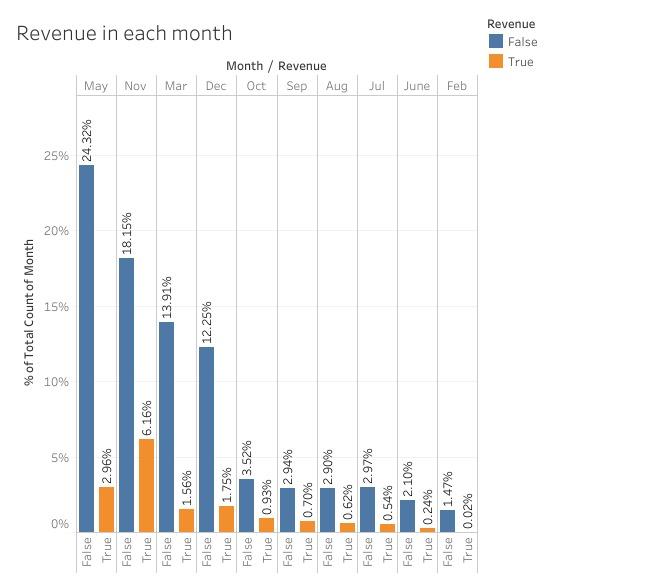
Exploratory Data Analysis (EDA):

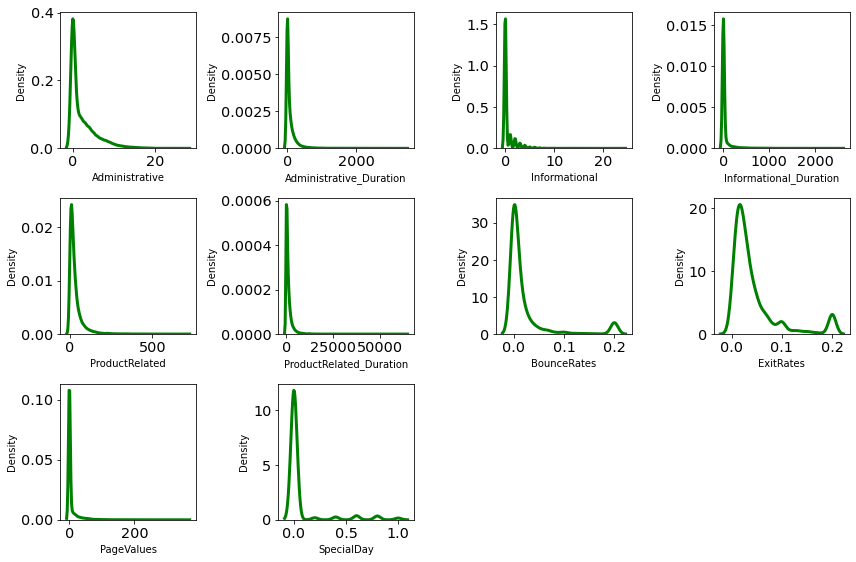
1. There are many Returning Visitor rather than New Visitor
2. Revenue have a many False data than a true, that’s why we need to increase a conversion rate

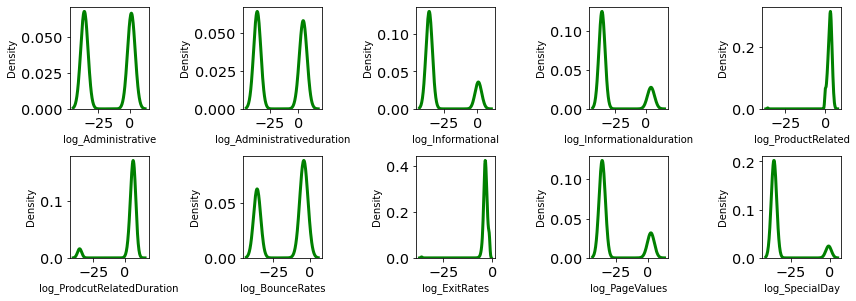


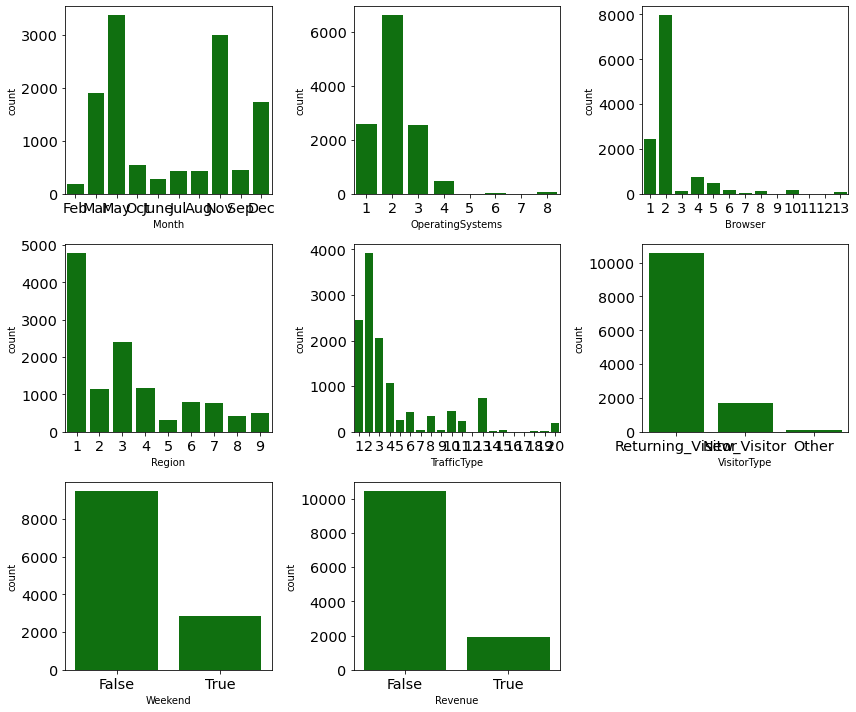
1. If we combine Visitor type and Revenue, New visitor have a many purchase rather than Returning visitor. So, in this dataset, drop a new visitor and focused for returning visitor.



1. For month, there are 3 top visitor type visit the online shop website (March, May, and November) and there are 3 top conversion rate (November, May, and December).
2. Univariate analysis for numeric



The result of KDE plot, the features have a positively skew, this features conducted log transformation. With the result:

1. Univariate analysis for categoric
2. In this dataset, have 125 duplicates, but for this analyst we don’t drop this duplicated. This duplicated assumed that evert row in this dataset is a user visit session which happen to have the same patterns with another.

Handling visitor type, we drop all the row that have New visitor type and replace other visitor to Returning Visitor. After the drop, we have 10636 rows.

Handling outlier, Using z-score. After use z-score, we have 9883 rows.

1. For the feature, we need feature transformation, like log transformation. Another feature is Encoding. Label Encoding for Weekend and Revenue and One Hot Encoder for Month, Operation Systems. Browser, Region, Traffic Type, and Visitor Type.
2. After running correlation for the dataset, with a>0.05 and a<-0.05, we choose 8 features. With the details:

* Log\_Administrative
* Log\_Informational
* Log\_ProductRelated
* Log\_ExitRates
* Log\_PageValues
* Agg\_Month\_nov
* Agg\_OperatingSystems\_2
* Agg\_trafficType\_2

1. From the feature selection, we split 70% for train and 30% for test. Handling class Imbalance with SMOTE data train (1:1).
2. For Model, we try 5 model,

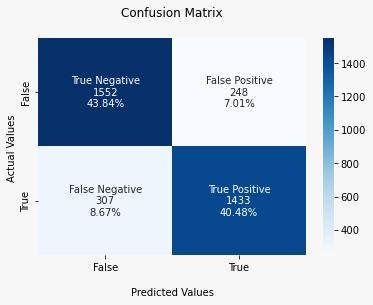
* Logistic Regression
* Decision Tree
* Random Forest
* AdaBoost
* X
* GBoost

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy(%) | Precision(%) | Recall(%) | AUC(%) | F1(%) |
| Logistic Regression | 86 | 51 | 80 | 83 | 62 |
| Decision Tree | 84 | 47 | 62 | 74 | 53 |
| Random Forest | 86 | 53 | 70 | 79 | 60 |
| AdaBoost | 87 | 53 | 79 | 83 | 64 |
| XGBoost | 87 | 56 | 63 | 77 | 59 |

1. Tuning Hyperparameters

|  |  |
| --- | --- |
| Solver | Newton-cg, lbfgs, liblinear |
| Penalty | L2, l1, elasticnet, none |
| C | 100, 10, 1.0, 0.1, 0.01, 0.001, 0.0001 |

1. Best Model Logistic Regression. With the top features, Log\_pageValues; agg\_Month\_Nov; Log\_ProductRelated; agg\_trafficType\_2; Log\_Informational; and Log\_Administrative.
2. From the Feuture Selection Logistic regression, we make a confussion matrix.



Conclusion:

* After Exploratory Data Analyst and Data processing, we choose 8 features with 1 target.
* This features selection, we try 5 models. Logistic Regression, Random Forest, Decision Tree, AdaBoost, and XGBoost. After that, we also try this model with hyperparameters.
* For this analyst, we choose logistic Regression for the model. The best Feature from this model is Page Values.
* From the matrix, we can assumed profit this e-commerce have an increase 75%.
* Form the Page values, we can give the recommendation product or coupon discount so the visitor don’t need to take a lot of time to check out their shopping cart.

Sugeestion:

* For this online shopper website, can give a special treatment for visitor who only want to visit not buy with pop up message or in their social media.
* Increase time to checkout before the visitors experiences search fatigue with coupon discount or recommendation product.
* Try to use A/B Test to know :
  + Page Value with marketing methods in months, especially in November and May.
  + Visitor Traffic type that’s not productive at all and why that’s happen.
  + Payment methods from the website.
  + History product from the visitors.

The Team Behind this Project:

* This Project have 8 members: Arif B, Desy N, Maya M, Inggriani, Danang Rizki, and M. Donny.
* This project is a final project for Rakamin Bootcamp for Data Science.
* If you want to know about the detail, you can visit gdrive with this link : https://drive.google.com/drive/folders/1cK\_bYaqN853SYLVfwWOv6ZD72f0UAIIx?usp=sharing