Online Shoppers Intention

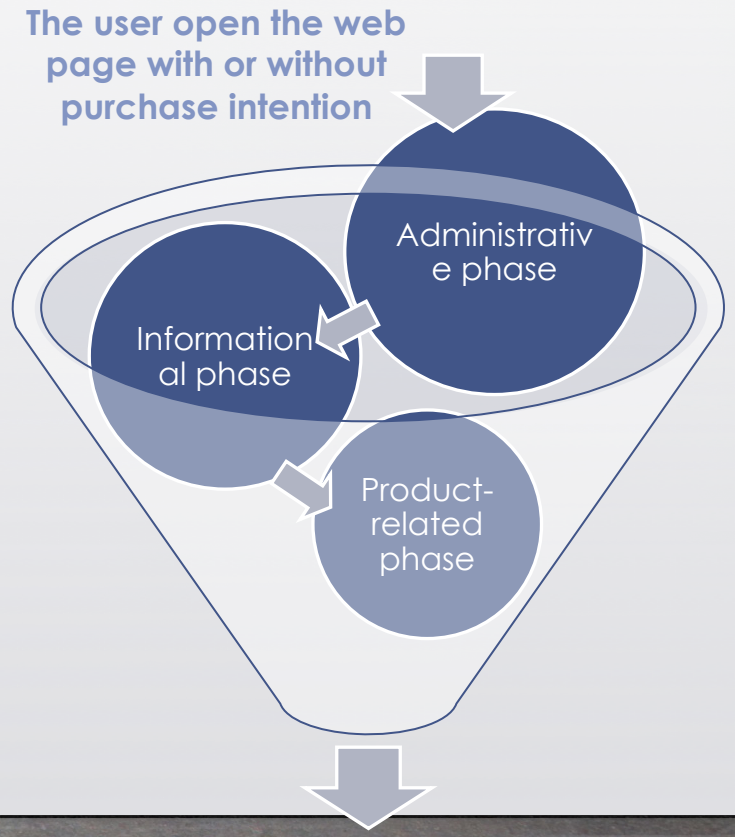
1 Think starters

The data set was obtained from the Kaggle website (Link). The data was extracted by C. Okan Sakar and Yomi Kastro in order to predict the visitor's shopping intent and Web site abandonment likelihood. Before purchasing, customers usually have many steps to do.

For example, a typical online purchasing path could be shown from the figure on the right. It started from the customer open the website:

1. **Administrative phase:** Customers log into their account or sign up for an account;
2. **Informational phase:** They keep searching for certain products if they have clear purchase intention or they have broad but vague categories and then have a purchase intention after browsing.
3. **Finally, product-related pages:** At this step, the customer comes to a product page, it might determine the final purchasing decision.

Hence, an intention will foster purchasing behaviors, in relation to different purchasing phase, but it does not necessarily lead to the final purchase.



2 Data Summary

Variable	Variable explanation		Variable	Variable description	
Administrative	Number of pages visited by the visitor about account management	numeric	Operating system	Operating system of the visitor	categorical
Administrative Duration	Total amount of time (in seconds) spent by the visitor on account management related pages	numeric	Browser	Browser of the visitor	categorical
Informational	Number of pages visited by the visitor about Web site, communication and address information of the shopping site	numeric	Region	Geographic region from which the session has been started by the visitor	categorical
Informational Duration	Total amount of time (in seconds) spent by the visitor on informational pages	numeric	TrafficType	Traffic source by which the visitor has arrived at the Web site (e.g., banner, SMS, direct)	categorical
Product related	Number of pages visited by visitor about product related pages	numeric	VisitorType	Visitor type as ‘‘New Visitor,’’ ‘‘Returning Visitor,’’ and ‘‘Other’’	categorical
Product related Duration	Total amount of time (in seconds) spent by the visitor on product related pages	numeric	Weekend	indicating whether the date of the visit is weekend	categorical
Bounce rate	Average bounce rate value of the pages visited by the visitor	numeric	Month	Month value of the visit date	categorical
Exit rate	Average exit rate value of the pages visited by the visitor	numeric	Revenue	Class label indicating whether the visit has been finalized with a transaction	Binary (Outcome)
Page value	Average page value of the pages visited by the visitor	numeric	Special day	Closeness of the site visiting time to a special day	categorical

3 Research Questions

Q1 Which factors contributed to user in purchasing the product (Revenue=True)?

Use different classification methods to better predict purchasing result

Q2 Can users types/intention be clustered based on different purchasing behaviors?

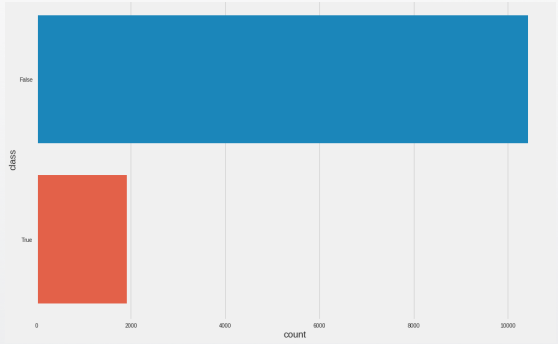
only on the user online behavior variables ("Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration") to cluster the customers and observe purchase-related patterns.

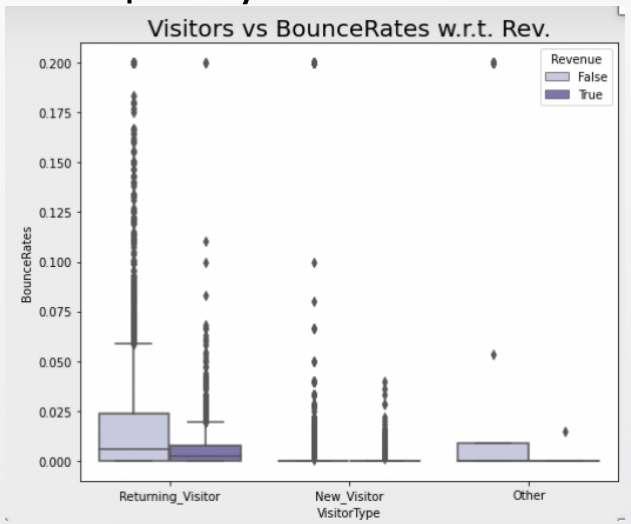
Q3 Whether downsamping or upweighting in this moderately imbalanced data set increase the model performance and prediction accuracy?

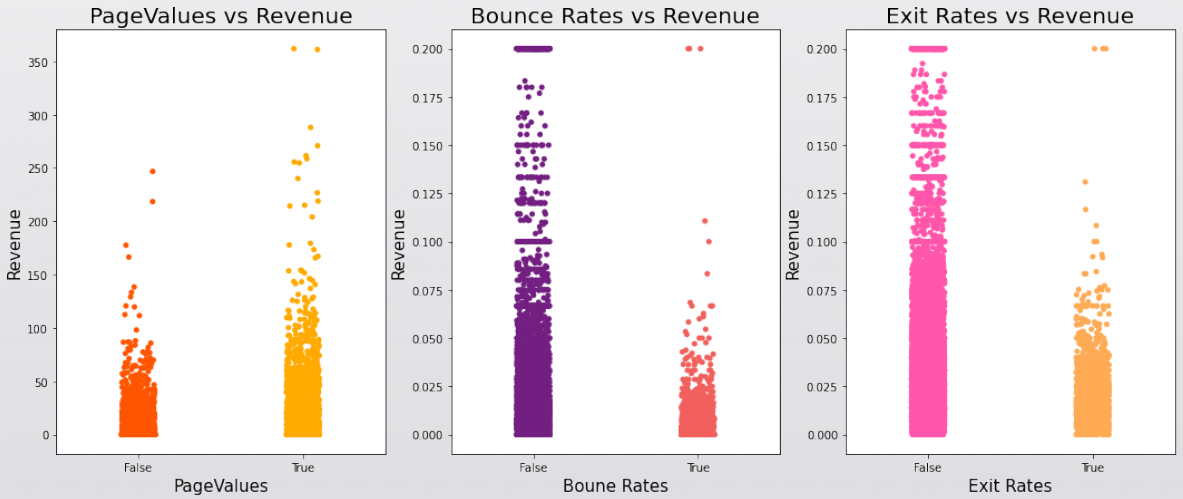
The data set is moderately imbalanced, which will leads to lower accuracy in minority score hence I attempted to use f_score as the metrics to compare model accuracy. Downsampling might effect data variability, Hence, Upweighting might be a way to increase accuracy for imbalanced data.

4 Several Figures (more in the report)

Revenue =True(1908) vs False(10408)



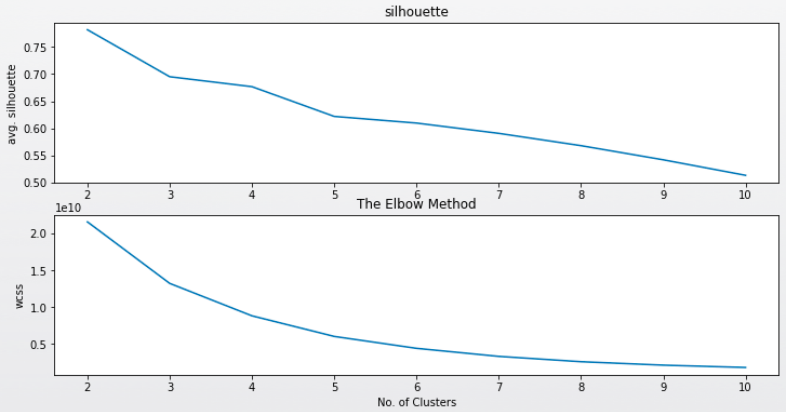




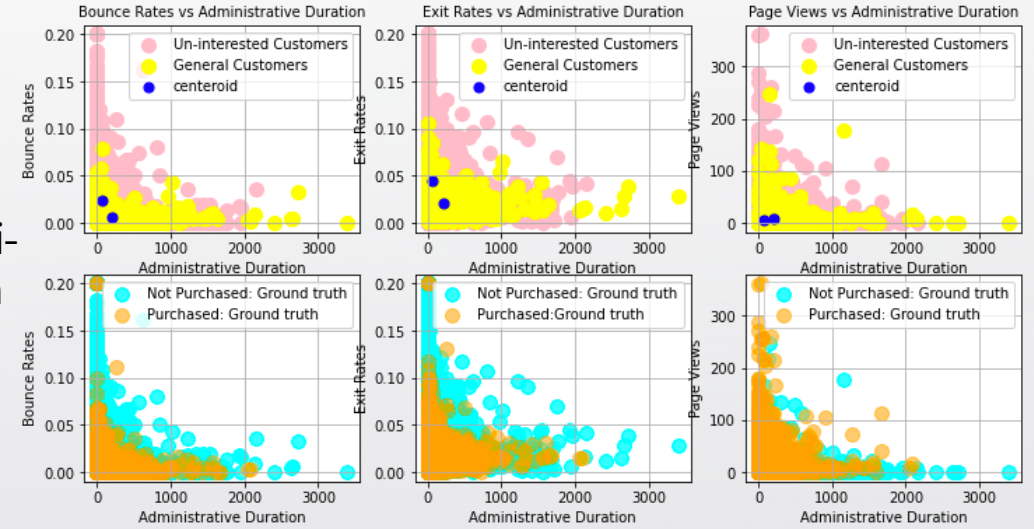
5 Result 1: Kmeans Clustering

I applied Kmeans only on the user online behavior variables ("Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration") to cluster the customers and observe purchase-related patterns.
Below, I ploted *Bounce Rates, Exit Rate* and *Page Views* vs different page categories (administrative,informational and product-related Duration, within the two clusters). I ALSO compared them to the ground truth (aka, the user purchased or not purchased, indicated by variable "Revenue"). Although a purchase intention will convert into these browsing behaviors, it does not definitely convert into purchase results (Revenue=True), we can still observed from the following plot that the clustered groups from Kmeans is highly relevant from the ground

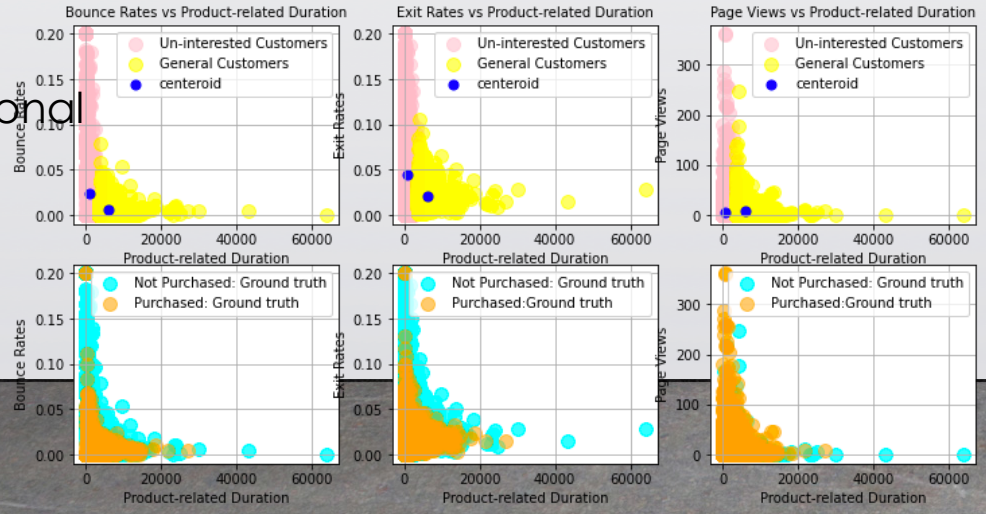
I used Elbow methods and Silhouette as the criterion to calculated best clusters (cluster=2)



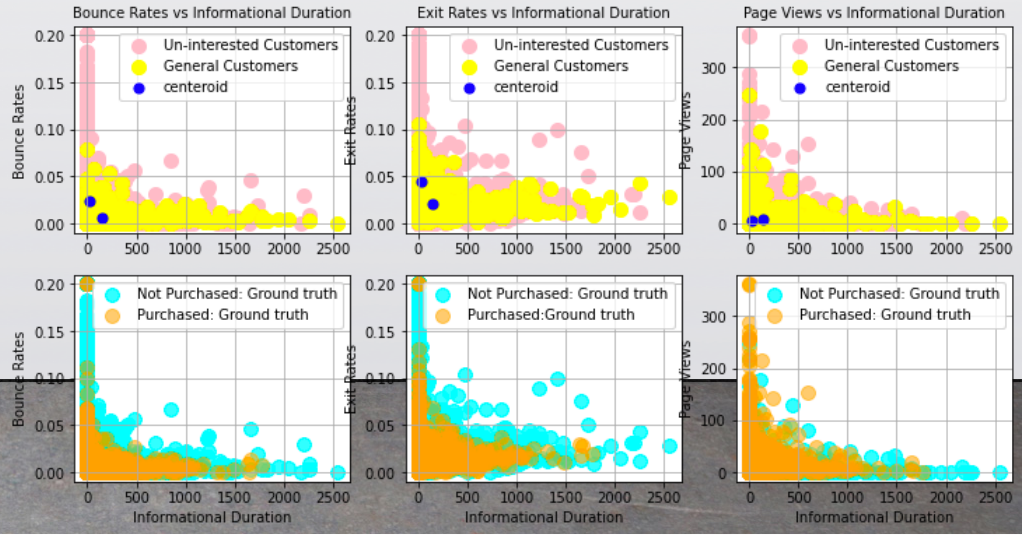
1 Administrative Duration



2 Informational Duration



3 Product-related Duration



6 Result 2:

Classification

Logistic regression & Linear Discriminant (LDA) & Random Forest

Aside from the intention clustering, I also predicted the target outcome (Revenue=True or False) with three classification methods.

The f_1 score for LDA is 0.1204937. The result from logistic regression is more optimal than LDA. A possible reason is that the linear discrimination rule is based on normality assumption. When the normality assumption is not valid, the linear discrimination rule may not be optimal. As shown in Figure 6, it indicates the histogram of several predictor variables, their distributions are highly skewed with many "0"s. Consequently, the normality assumption is not met here, which leads to the relatively poor performance of linear discrimination rule.

Random forest is based on constructing a forest, e.g., a set of diverse and accurate classification trees, using bagging resampling technique and combining the predictions of the individual trees using a voting strategy. The advantage of this method is that it introduce randomization and also returns the importance of each variables. The variable importance are shown on the right. However, random forest have many hyperparameters to set, hence the best combination can be chose by using cross validation and find the model with least error. Tunning hyperparameters can be tedious and time-consuming. Here, I tuned the hyperparameters using 3-fold cross validation. I tried the parameter n_estimators = [100, 300, 500, 800,1200], max_depth = [5, 8, 15, 25, 30], min_samples_split = [2, 5, 10, 15, 100], min_samples_leaf = [1, 2, 5, 10], the best model comes at (criterion='gini', max_depth=25, min_samples_leaf=5, min_samples_split=100, n_estimators=800).

The comparison of f_score with and without oversampling can seen on the right(Table 4 and Table 5). The f_score of Minority class (Revenue=True) has increased.

	precision	recall	f1-score
0	0.93	0.80	0.86
1	0.37	0.64	0.47
macro avg	0.65	0.72	0.66
weighted avg	0.84	0.78	0.80

Table 2: Accuracy of Linear Discriminant

	precision	recall	f1-score
0	0.89	0.98	0.94
1	0.77	0.35	0.48
macro avg	0.83	0.66	0.71
weighted avg	0.87	0.88	0.87

Table 3: Accuracy of Logistic Regression

	precision	recall	f1-score
0	0.92	0.95	0.94
1	0.68	0.55	0.60
macro avg	0.80	0.75	0.77
weighted avg	0.88	0.89	0.89

Table 4: Accuracy of Random Forest

	precision	recall	f1-score
0	0.91	0.97	0.94
1	0.77	0.51	0.62
micro avg	0.89	0.89	0.89
macro avg	0.84	0.74	0.78
weighted avg	0.89	0.89	0.88

Table 5: Accuracy of Random Forest (Over sampling)

Weight	Feature
0.1136 ± 0.0128	PageValues
0.0069 ± 0.0056	ExitRates
0.0058 ± 0.0051	BounceRates
0.0056 ± 0.0041	Month_Nov
0.0019 ± 0.0014	VisitorType_New_Visitor
0.0012 ± 0.0025	ProductRelated_Duration
0.0010 ± 0.0019	ProductRelated
0.0006 ± 0.0010	SpecialDay
0.0002 ± 0.0037	Region
0.0002 ± 0.0010	Month_Mar
0.0002 ± 0.0022	VisitorType_Returning_Visitor
0.0002 ± 0.0004	Month_June
0.0001 ± 0.0005	Month_Sep
0 ± 0.0000	Month_Feb
0 ± 0.0000	VisitorType_Other
-0.0002 ± 0.0006	Month_Aug
-0.0002 ± 0.0064	Administrative
-0.0003 ± 0.0007	Month_Oct
-0.0005 ± 0.0006	Month_Jul
-0.0006 ± 0.0004	Month_Dec