STAT 6130 Multivariate Statistics Final Project: Online Shopper Intention Prediction

Erzhen Hu (eh2qs)

May 2020

1 Background Information

The data set was obtained from the Kaggle website (https://www.kaggle.com/roshansharma/online-shoppers-intention). 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. Purchase intention is always hard to predict or determine by online behavioral data. Before purchasing, customers usually have many steps to do. For example, there are administrative phase that customers log into their account or even sign up for an account; Informational phase that they keep searching for certain products or broad and vague categories without knowing what to purchase. And finally, a product-related page might determine the final purchasing decision. Hence, an purchasing intention will foster, directly or indirectly lead to a purchasing behavior, in relation to different purchasing phase, but it does not necessarily lead to the final purchase. The data are used together to determine the visitors who have purchasing intention but are likely to leave the site in the prediction horizon, like browsing the informational data but not purchase, browsing the product-related pages but not purchase. Learning those phases can help business take actions accordingly to improve the Web site abandonment and purchase conversion rates.

The data also have a target variable (Revenue=True or False) to indicate whether the customer finally purchase the product. Additionally, the data are used together to determine the visitors who have purchasing intention but are likely to leave the site in the prediction horizon, like browsing the informational data but not purchase, browsing the product-related pages but not purchase. Learning those phases can help business take actions accordingly to improve the Web site abandonment and purchase conversion rates.

The dataset consists of feature vectors belonging to 12,330 sessions. The dataset was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period.

2 Summary of Data

The data consists of 12330 observations collected in a 1-year period. It contains 18 columns, the response variable is **"Revenue"**, which indicates whether revenue will be generated or not. The distribution of this target is not balanced, with 10422 False and 1908 True, as shown in Figure 1.

The first six columns ("Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration") represent the number of different types of pages (administrative, informational or Product-related) visited by the visitor in that session and total time spent in each of these page categories. The values of these features are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, e.g. moving from one page to another.

The "Bounce Rate", "Exit Rate" and "Page Value" features represent the metrics measured by "Google Analytics" for a specific web page in the e-commerce site.

- **Bounce Rate:** Numeric, the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session.

- **Exit Rate:** Numeric, calculated as for all page views to the page, the percentage that were the last in the session.
- **Page Value:** Numeric, represents the average value for a web page that a user visited before completing an e-commerce transaction.

The "Special Day" feature indicates 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. The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. For example, for Valentina's day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8.

The dataset also includes operating system, browser, region, traffic type, visitor type as returning or new visitor, a Boolean value indicating whether the date of the visit is weekend, and month of the year.

3 Research Problem

3.1 Exploratory Data Analysis

The data is highly imbalanced, with the minority class (revenue=true) in 15.47 percent of the data set. There are 14 missing data in the data set, I directly delete them. The EDA can be viewed from Figure 2 to Figure 4.

3.2 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 Model and Method

I tried logistic regression, linear Quadratic Discriminant, random forest, support vector machines, classifiers and clustering methods using 70 % of data set as training and the rest 30 % for validation.

The classifiers tend to minimize their errors on majority class samples(Revenue=FALSE), which leads to an imbalance between the accuracy rates of the positive and negative classes, as shown in every methods, the no-revenue class gives higher accuracy than revenue. This is directly related to the objective of this project. Mostly, in a real-time customer analysis model, correctly identifying directed buying visits, which are represented with minority class (revenue=TRUE) in our dataset, is as important as identifying (revenue=FALSE).

Here, because the data is highly imbalanced, I used f-score as the metric to compare models.

To further deal with class imbalance problem, I used a upweighting method, in which a uniform distribution over the classes is aimed to be achieved by adding more of the minority (revenue=True) class instances. I did not use downsampling because it will decrease the training data set very small like the minority class. Since this data set is created by selecting multiple instances of the minority class more than once. However, if I first upweighting the data set and then dividing it into training and test sets, this may lead to biased

results and over training, because the same minority class (revenue=True) may be used both for training and test. For this reason, 30% of the data set consisting of 12,330 original data is first left out for testing and the upweighting method is applied to the remaining 70% of the samples.

4.1 Logistic Regression

Logistic regression gives, and this can tell us some intuitive relations between predictors and outcomes just like the linear regression. For example, the results shows one unit increase in Product Related Duration with got a 6.068e-05 unit increase in the probability of the user in buying the product (revenue=True, p=0.024891*); A unit decrease in Exit Rate will cause the probability (1.555e+01 unit decrease) of the user in not buying the product (revenue=False, p=9.32e-11***). These patterns can also be observed through Exploratory Data Analysis, i.e.,Figure **??**, though. The results from logistic regression is shown in Table 3

4.2 Linear Discriminant

Linear Quadratic Discriminant can be The classification error is 0.1204937 for Linear Quadratic Discriminant (LDA), the result from logistic regression is more optimal than LDA. The coefficients has been shown in **Table 1**. 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.

4.3 Random Forest

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). As shown in these two plots, The f_score of Minority class (Revenue=True) has increased.

4.4 KMeans

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. I used both elbow method and silhouette as criterion to choose the optimal cluster number. **The elbow method** in clustering is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use. As for **silhouette**, the functions evaluates how well one observation fits in its own cluster comparing to how well it fits in the next closest cluster. The Elbow Method is more of a decision rule, while the Silhouette is a metric used for validation while clustering. Thus, it can be used in combination with the Elbow Method. Also, Silhouette coefficient exhibits a peak characteristic as compared to the gentle bend in the elbow method. This is easier to visualize. I ploted 2 to 11 clusters and as shown in Figure 7, the optimal cluster number is 2.

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 in each

figure(aka, the user purchased or not purchased, indicated by variable "Revenue"). The Bounce Rates,Exit Rate and Page Views vs administrative duration, compared with Ground Truth can be viewed on Figure 8. It can be observed that every plot looks similar to the ground truth, except for a misclassification in longer administrative duration (range 2000 to 3000)at the tail, which might be impacted due to those potential outliers.

The Bounce Rates,Exit Rate and Page Views vs Informational duration, compared with Ground Truth can be viewed on Figure 9. The Bounce Rates,Exit Rate and Page Views vs Product-related duration, compared with Ground Truth can be viewed on Figure 10. It did not that good in clustering Product-related duration and the three predictors (Bounce rate,exit rates, page views), this indicates that clicked into the product-related duration might not definitely leads to a purchase result, smaller product-related duration (aka quick product-related duration) also leads to a purchase results, which were misclassified as un-interested customers.

In conclusion, 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.

4.5 Model Based Clustering: Mixture Multivariate Normal

The K-mean clustering method do not impose distributional assumption on the data. In order to take into account of the fact that observations from different clusters should follow different multivariate normal distributions, I also tried mixture distribution model.
There are three clusters based on BIC criterion; Assumption is EEV (ellipsoidal, equal volume and shape) model with 3 components, more detailed information is shown in Table 6.

5 Analysis, Conclusion and Discussion

Purchase intention is hard to predict or determine.Before purchasing, customers usually have many steps to do. Learning the conversion rate in each step/purchasing phase and its relation to purchase intention and final purchase decision can help business. Hence I attempted the Kmeans method to cluster customer types based on the purchasing behavior variables ("Administrative", "Administrative Duration","Informational", "Informational Duration", "Product Related" and "Product Related Duration"). Although a purchase intention will convert into these browsing behaviors, the result shows that 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. Quicker product-related page views might indicate a purchase result, yet are misclassified as not interested customers in Kmeans.

I also tried different classification methods on the whole data set, after removing the 14 missing data. Random forest performs best, compared to logistic regression and linear Discriminant methods. The linear Discriminant have a higher f_score on minority class(Revenue=True) than linear regression, but have a bad performance overall accuracy, which might be attributed to its normal assumption.

Additionally, due to the fact that the data is highly imbalanced, with the minority class (revenue=true) containing 15.47 percent of the whole data set. I used f_score as the metric to compare different models. I also attempted over-sampling and applied randon forest model on it, which increases the f_score in minority class by 0.02.

6 Tables and figures

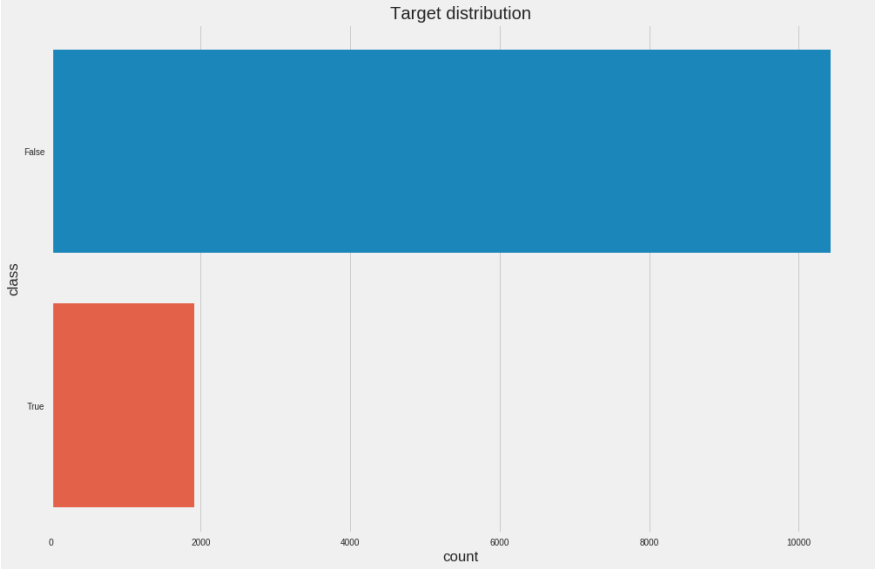

Figure 1: Distribution of target

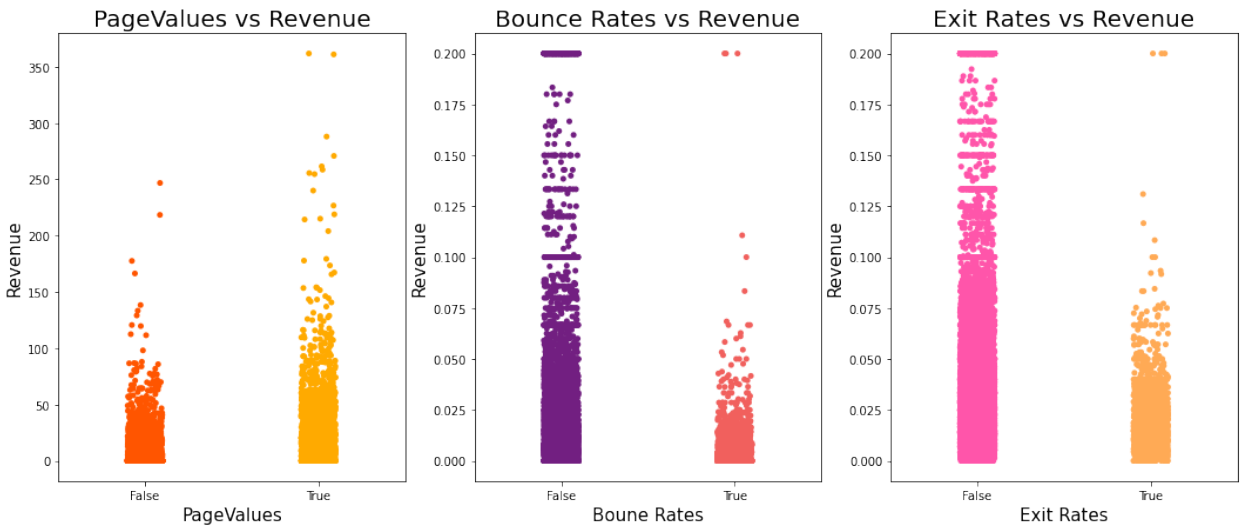

Figure 2: Different types of page vs Revenue

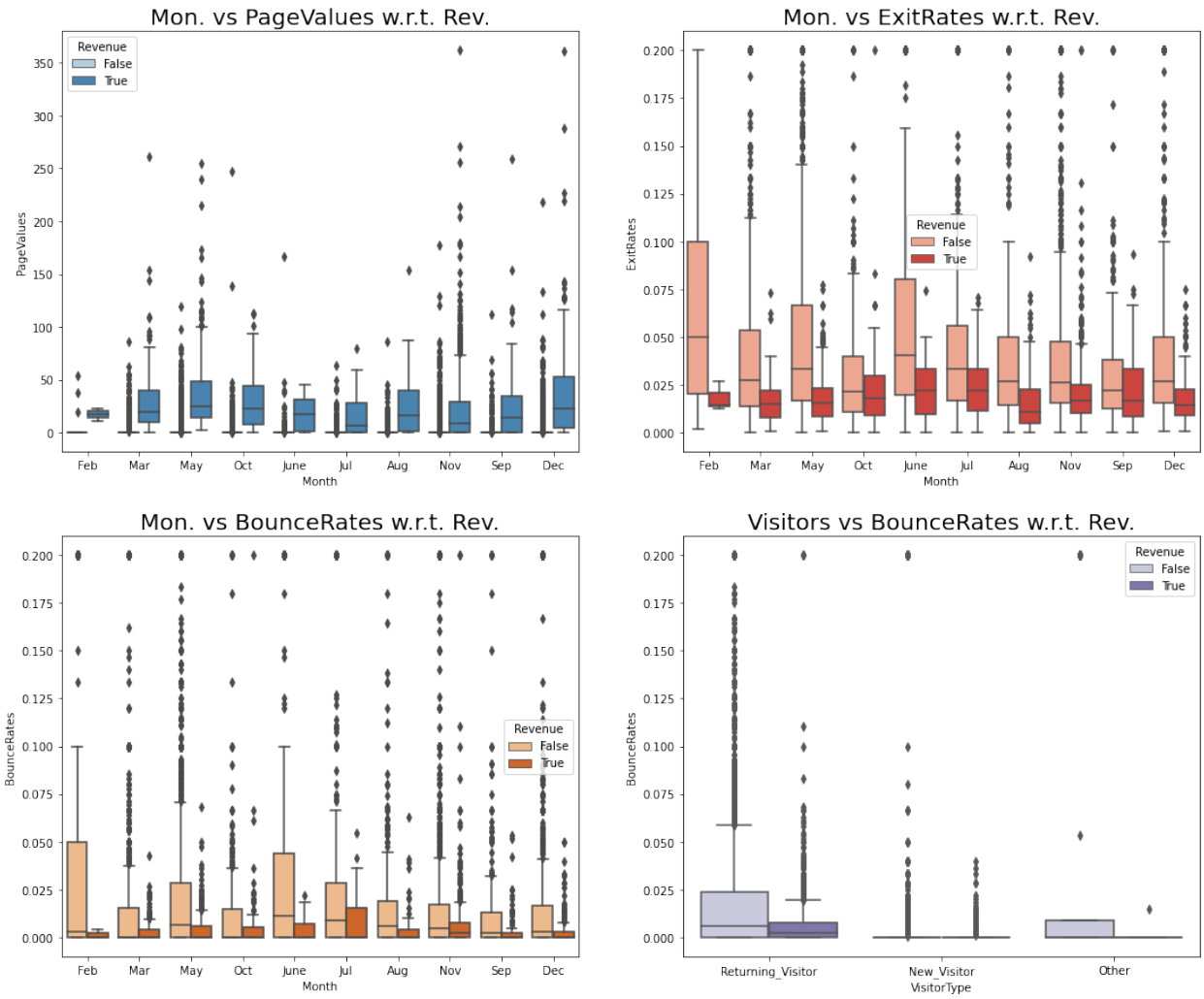

Figure 3: EDA: Multivariate Analysis

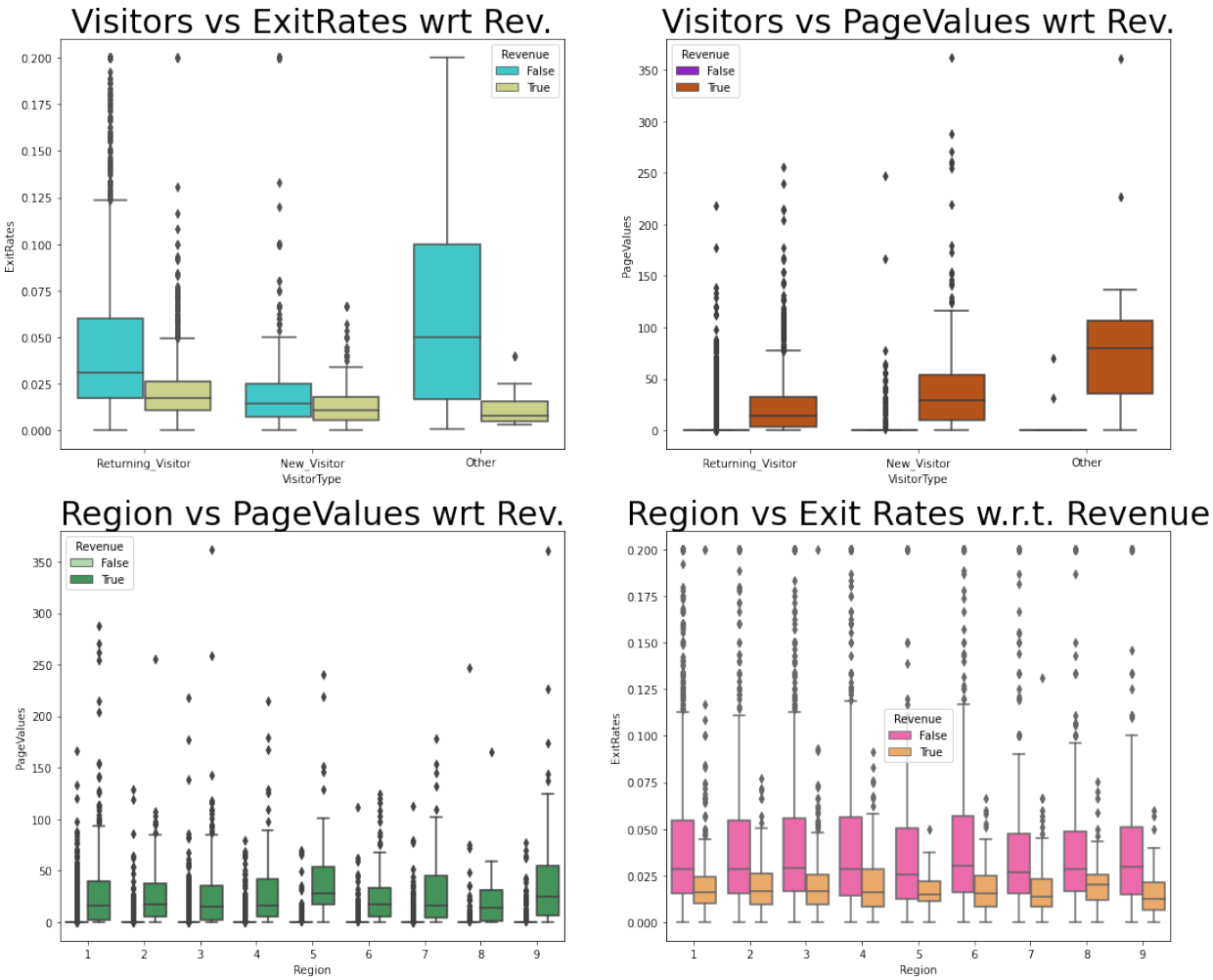

Figure 4: EDA: Multivariate Analysis

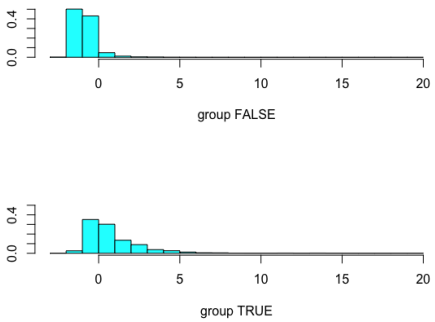

Figure 5: Coefficient plots for both groups

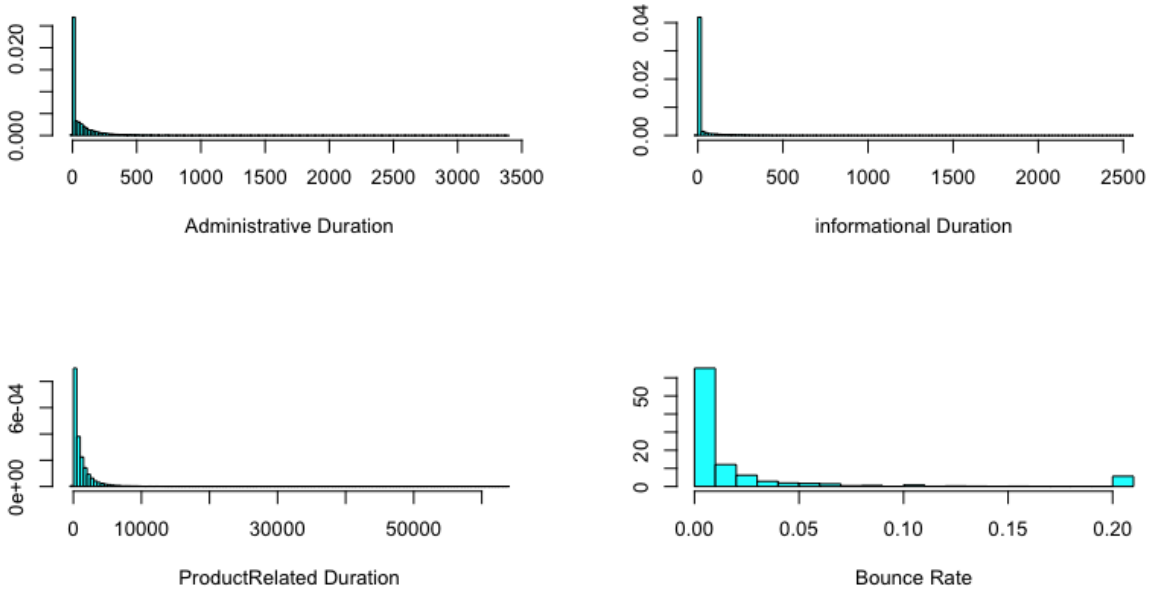

Figure 6: Highly Skewed Histogram of Predictors

Coefficients of linear discriminants:	LD1
Administrative	1.133453e-02
Administrative Duration	-9.808800e-05
Informational	2.061340e-02
Informational Duration	9.284148e-05
ProductRelated	1.409665e-03
ProductRelated Duration	6.285704e-05
BounceRates	2.933386e+00
ExitRates	-6.233819e+00
PageValues	5.457847e-02
SpecialDay	-1.003143e-01
MonthDec	-3.048819e-01
MonthFeb	-3.681651e-01
MonthJul	1.509911e-02
MonthJune	-2.005265e-01
MonthMar	-2.362412e-01
MonthMay	-2.448489e-01
MonthNov	3.810352e-01
MonthOct	1.974154e-02
MonthSep	-3.090028e-02
OperatingSystems	-4.932098e-02
Browser	2.369347e-02
Region	-1.124459e-02
Traffic Type	-8.920776e-04
VisitorTypeOther	-4.174423e-01
VisitorType (Returning Visitor)	-2.932398e-01
WeekendTRUE	4.758638e-02

Table 1: Coefficients of linear discriminants

	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)

log-likelihood	n	df	BIC	ICL
-437558.1	12316	478	-879618.4	-879799

Table 6: Result from Model Based Clustering

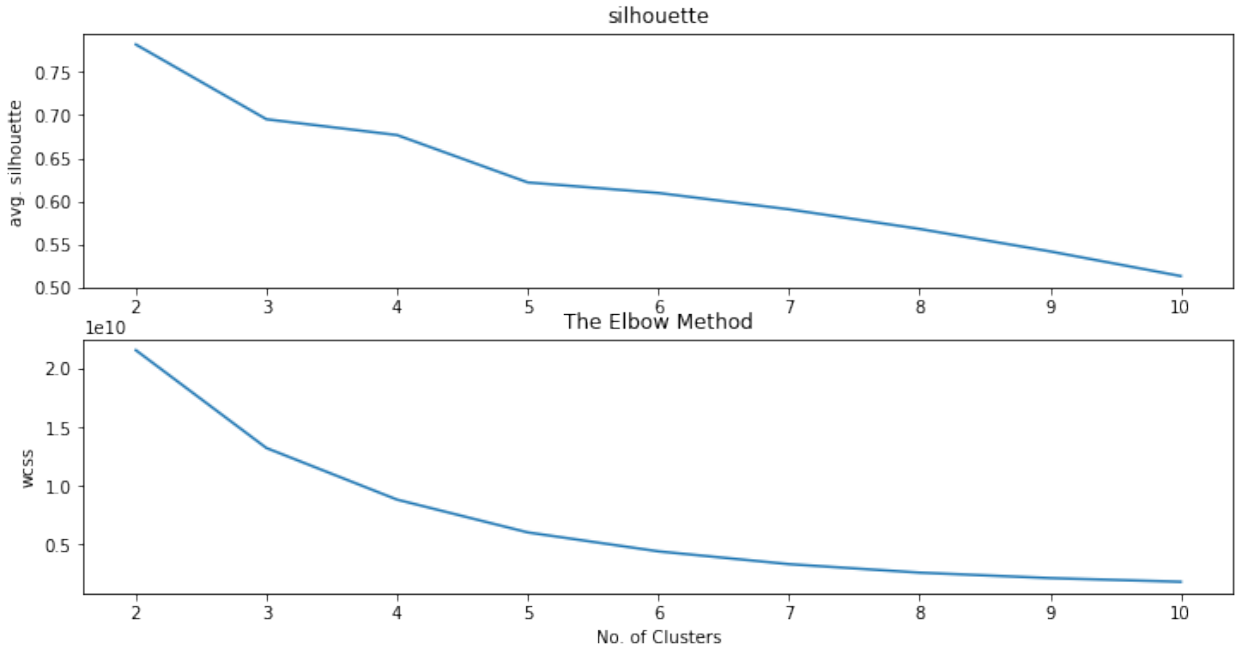

Figure 7: Elbow and Silhouette method

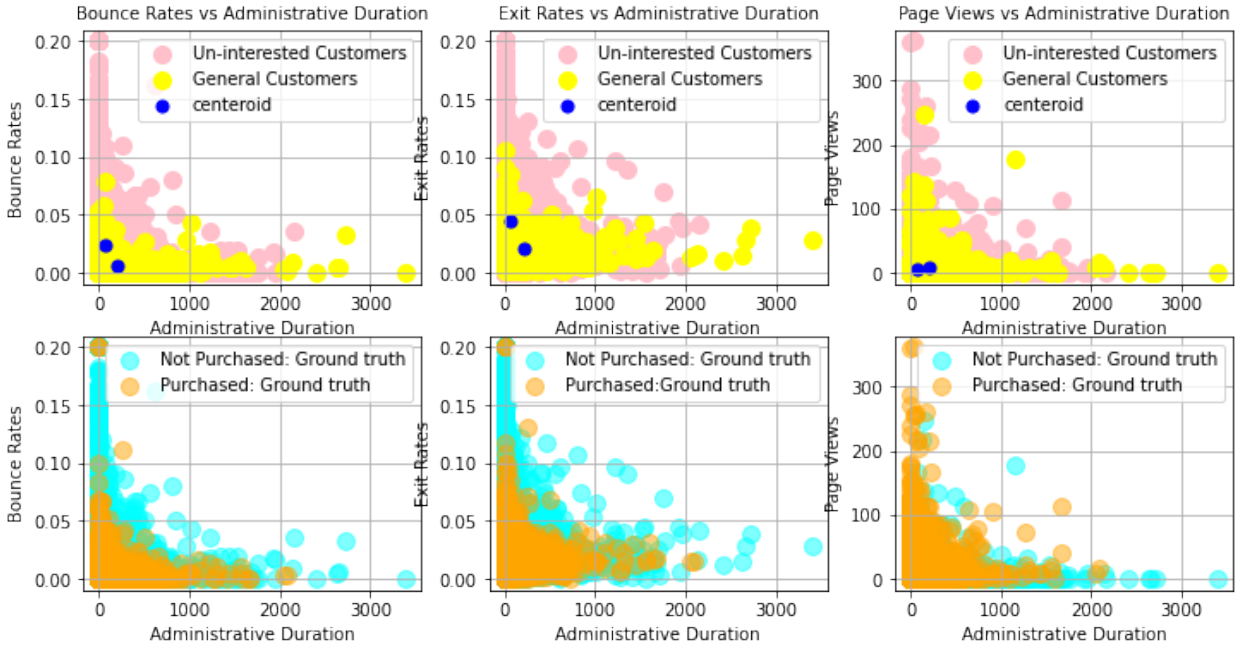

Figure 8: Administrative Duration vs Bounce Rates/Exit Rates/Page Views, compared to ground truth

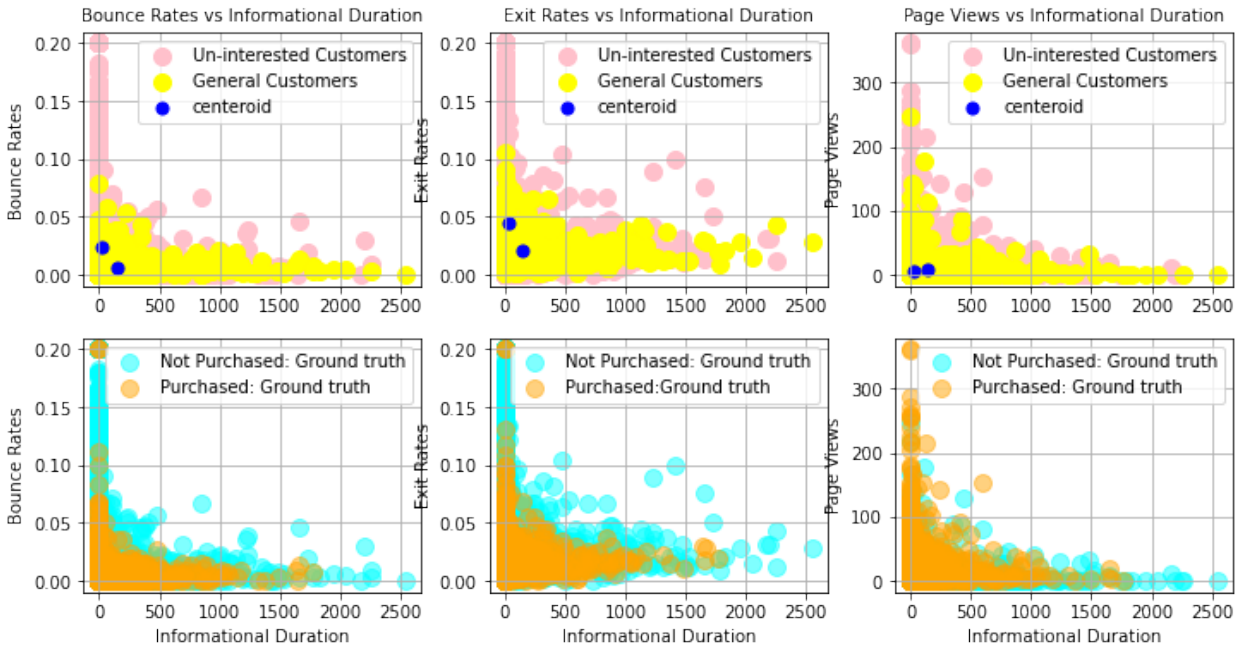

Figure 9: Informational Duration vs Bounce Rates/Exit Rates/Page Views, compared to ground truth

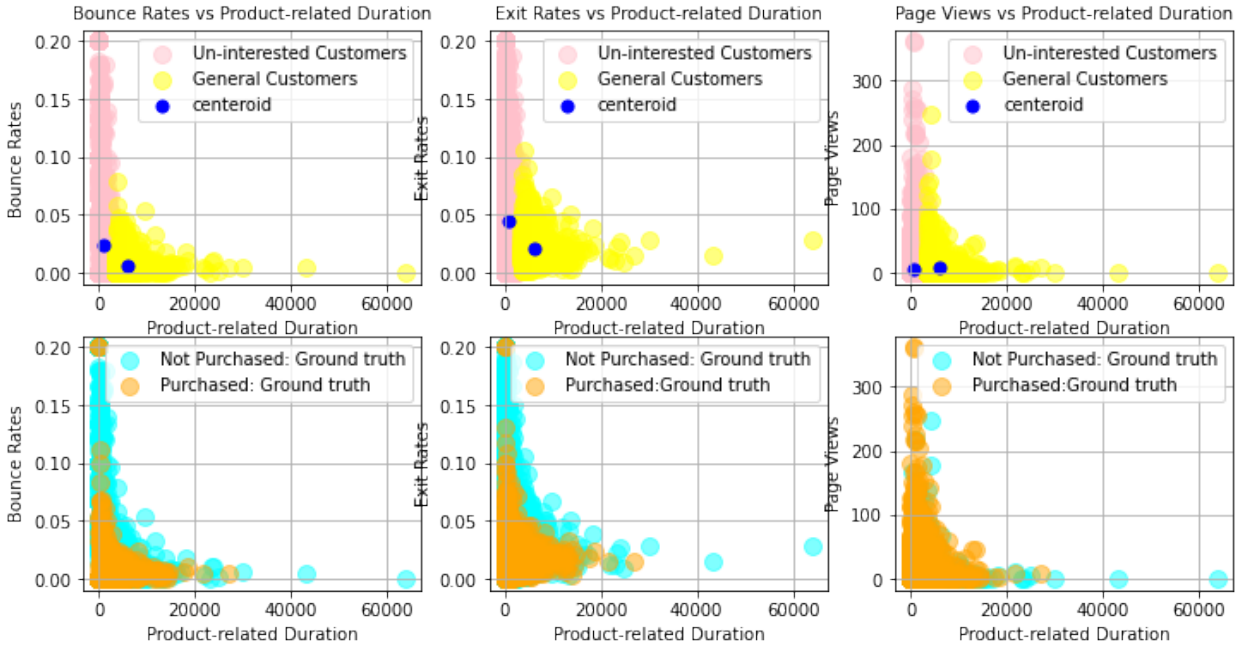

Figure 10: Informational Duration vs Bounce Rates/Exit Rates/Page Views, compared to ground truth

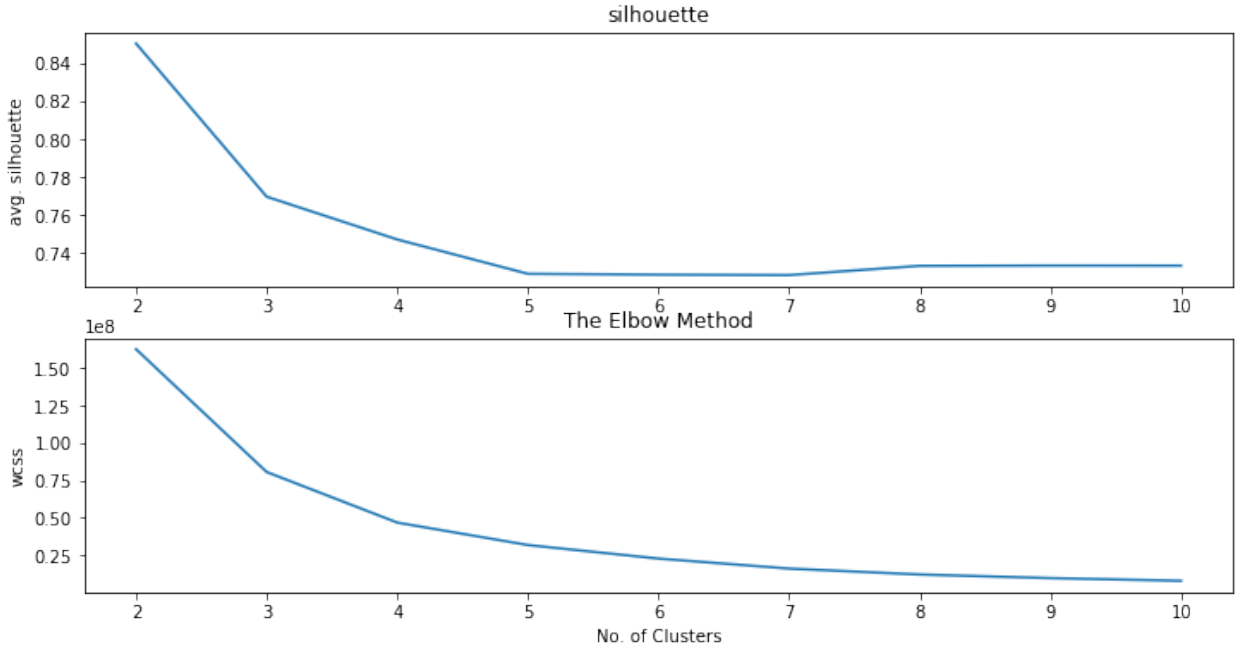

Figure 11: Administrative Duration vs Bounce Rates: Elbow and Silhouette methods

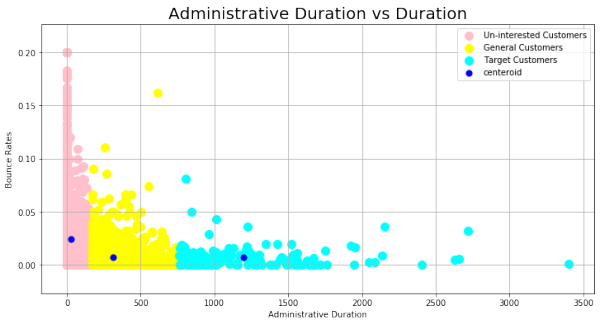

Figure 12: Administrative Duration vs Bounce Rates: Three clusters

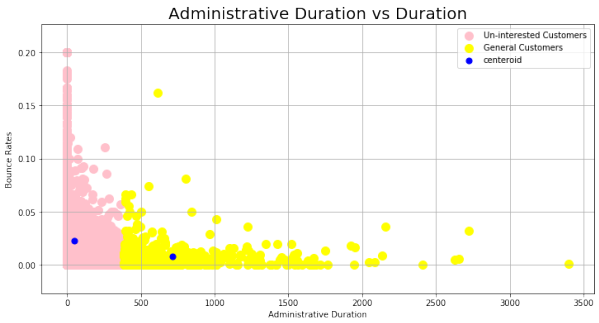

Figure 13: Administrative Duration vs Bounce Rates: Two clusters

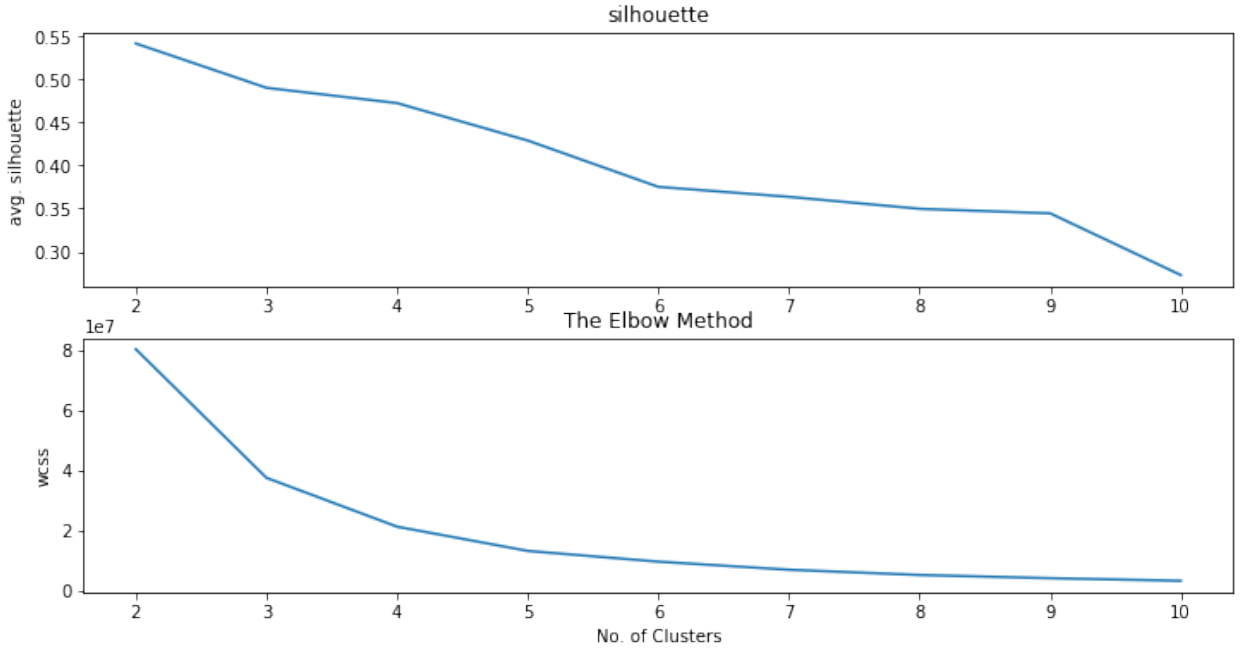

Figure 14: Informational Duration vs Bounce Rates: Elbow and Silhouette methods

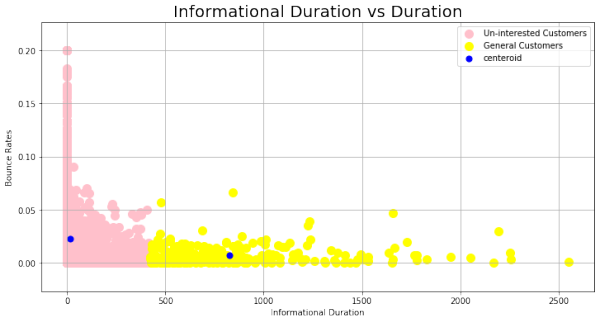

Figure 15: Informational Duration vs Bounce Rates: Two clusters

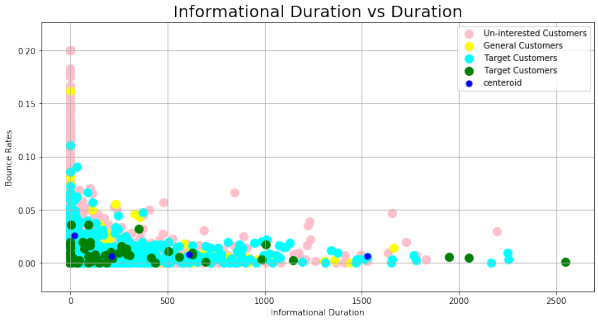

Figure 16: Informational Duration vs Bounce Rates: Four clusters

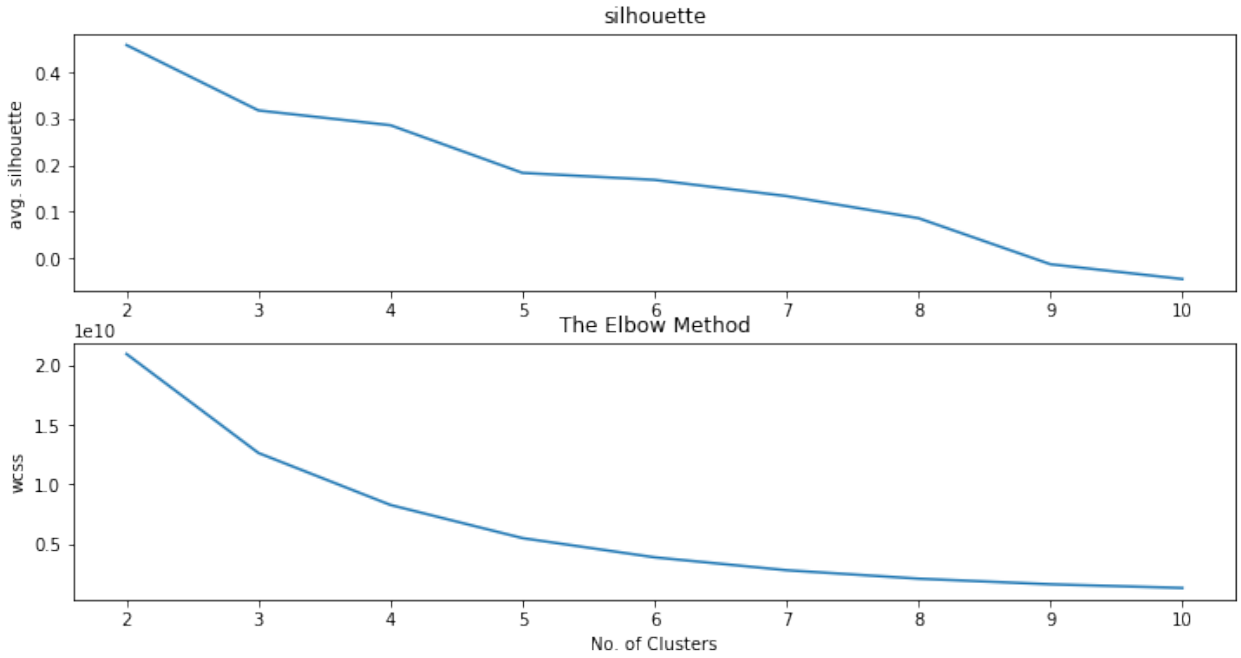

Figure 17: Product-Related Duration vs Bounce Rates: Elbow and Silhouette methods

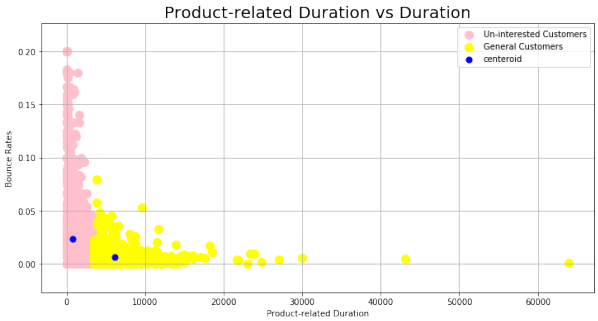

Figure 18: Product-Related Duration vs Bounce Rates: Two clusters

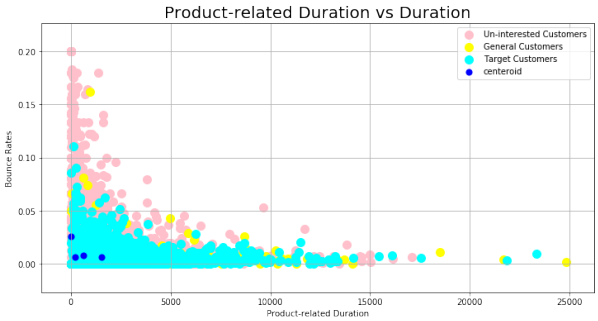

Figure 19: Product-Related Duration vs Bounce Rates: Four clusters

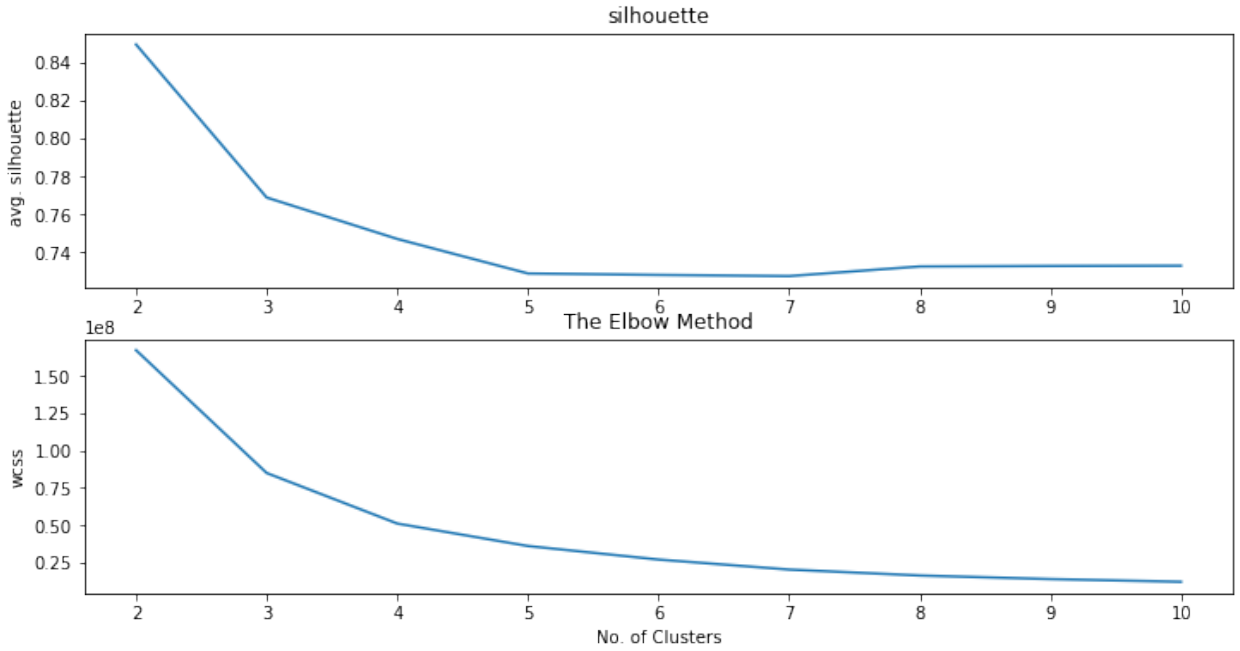

Figure 20: Product-Related Duration vs Page Views: Elbow and Silhouette methods

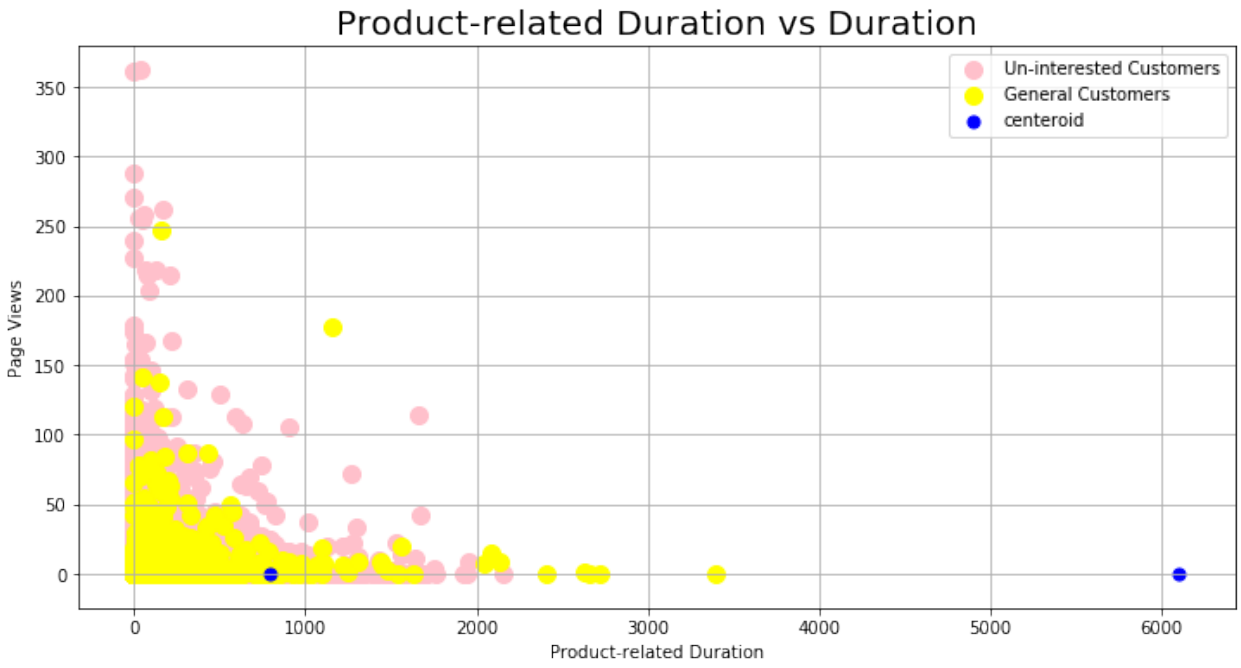

Figure 21: Product-Related Duration vs Page Views: 2 clusters