**Introduction**

The objective of this project is to review the predictive web analytics case study performed by Suchismita Sahu. A few technical terms must first be understood prior to reviewing the case study. Ecommerce and e-business are common nowadays due to the mass quantity of internet users worldwide. The article written by ecommerce guide explains that ecommerce, “covers particularly the sales and purchases made on the internet-while e-business involves any online business activity, including sales calls, procurement of materials, signing contracts, and so on” (2022). Essentially, ecommerce falls under e-business as an important method of making sale transactions online. Some businesses may solely rely on ecommerce as their sole sales strategy where users only make purchases through an organization’s website. Other businesses may utilize ecommerce as an added sales tactic to improve revenue by making consumer purchases more widely available. A fascinating statistic from ecommerce guide states, “In 2021, retail ecommerce sales amounted to 4.9 Trillion U.S. dollars worldwide. And it’s forecast to grow over 50% within the next four years. People are buying products online, and they’ll only buy more” (2022). Enter web analytics. Web analytics aims to capture and analyze website data based on various metrics from users visiting the site. Some common metrics a business may capture are quantity of users visiting the site, commonly clicked links, time spent on different pages, and where the website traffic stems.

Now that the general terms of ecommerce, e-business, and web analytics are outlined, it is time to dive into the case study. The high-level objective for this study was to construct a Predictive Model of whether a customer will make a purchase based on web analytics data from an organization’s site. The project aimed to understand insights through various visualizations, understand customer behavior, and construct a predictive model. This is an important problem to solve due to the major reliance from businesses on ecommerce. Understanding how a customer will behave while browsing on a site and whether they will likely make a purchase helps an organization recognize actionable insights for marketing and advertising. The organization could recognize great monetary gains based on the trajected growth of ecommerce in the short term. The data was acquired from UCI Machine Learning Repository through the link below (online\_shoppers\_intention.csv):

[UCI Machine Learning Repository: Online Shoppers Purchasing Intention Dataset Data Set](https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset)

The dataset consists of 18 features (18 categorical and 10 numerical) with 12,330 records. The following features focus on the type of page visited along with the time spent on that page: “Administrative”, “Administrative Duration”, “Informational”, “Informational Duration”, “Product Related”, “Product Related Duration”. The next features focus on the movement between pages of the site: “Bounce Rate” (enter site and leave right away), “Exit Rate”, and “Page Value” (time on web page prior to ecommerce transaction). The attribute “Special Day” measures the time on the site around a Holiday. Other features within the dataset include browser, visitor type, operating system, and region (UCI Machine Learning Repository, 2018).

**Methods and Results**

Python was utilized to perform the analysis. The dataset was relatively clean straight from the UCI Machine Learning Repository. The data was loaded into a data frame via CSV file. The data frame was checked for any missing or null values in which none were found. Next, exploratory data analysis was used to evaluate the features individually and together (bivariate). Sahu utilized countplots to further understand Revenue, Visitor Type, Visitor Type by Weekend, Traffic Type by Revenue, Traffic Type Distribution, Region Distribution, Operating System Distribution, Month Distribution, Traffic Type vs. Revenue, and Region by Revenue. Additional bivariate analysis was performed with stripplots (scatterplots with one categorical variable) to understand the relationship between Revenue vs. Page Value, Revenue vs. Bounce Rate, and Administrative vs. Informational Pages (Linear Regression Plot). The last plots exploring the features were boxplots with respect to revenue. The box plots included Month vs. Page Values, Visitor Type vs. Bounce Rates, and Month vs. Exit Rates. There were no notes of removing any of the data throughout the study. The EDA plots helped better understand the features included within the dataset.

The problem was addressed through several steps. To learn about the customer behavior, Sahu used k-Means clustering to segment the users into common clusters. Administrative Duration vs. Bounce Rate was the first focus with a clustering model. The number of clusters was refined to three with an elbow plot. The elbow plot showed the Within Cluster Sum of Squares (WCSS) vs. the number of clusters for the model. The three clusters helped identify un-interested customers, general customers, and target customers based on how long the user stayed on an administrative page or immediately left. The same clustering approach was used for Informational Duration vs. Bounce Rate, but the elbow plot showed two main clusters. The two clusters signified un-interested customers and target customers based on the same strategy for time spent on informational pages. The last clustering model focused on Region vs. Traffic. This cluster model had two main clusters which were also un-interested customers and target customers. The predictive model phase was next. The categorical variables were prepared with one hot encoding using Python’s get\_dummies() method. The target variable, Revenue, was transformed into a categorical variable to predict whether a customer would make a purchase. The data was then split into a 70% Training and 30% Testing subset. A Random Forest Classifier model was constructed, trained, and fit to the training data. The model was used to make predictions based on the test data. The analyst used a confusion matrix, model accuracy, precision, recall and F1-scores to evaluate the model performance. A Receiver Operating Characteristic (ROC) curve was also utilized as an evaluation metric. The model accuracy for the Random Forest Classifier ended up being 89% and the AUC was 0.92. A Logistic Regression model was also used to compare against the Random Forest Classifier. The same approach was utilized with the training and test data to construct and make predictions with the Logistic Regression model. The same evaluation metrics were utilized for this model which resulted with an accuracy of 86% and an AUC of 0.89. Based on these evaluation metrics, the Random Forest Classifier turned out to be the favorable model.

The modeling techniques used in this case study were k-Means Clustering, Random Forest Classifier, and Logistic Regression. The intent for the k-Means Clustering models were to understand customer segmentation for site users based on time of webpages visited versus bounce rate and region of user against site traffic. The team chose this method because a clear target customer base was identified for each intended model. The Random Forest Classifier and Logistic Regression models were both constructed to predict whether a customer will make a purchase. The target variable was converted into a binary feature which most likely led Sahu to use these two models. The confusion matrix, ROC curve, accuracy, precision, recall, and F1-score helped compare the two models as evaluation metrics. The Random Forrest Classifier appeared to be the more favorable model.

**Conclusion**

Both sets of predictions from the Random Forest Classifier and Logistic Regression models were exported to a CSV file to provide to the decision makers within the organization. In conjunction with the results from the clustering models, the business now has insights into which customers will generate revenue. The website used in the case study appeared to be anonymous without any specific mention from the article. As a result, there was no additional information for quantitative monetary improvements resulting from this case study. However, the business decision makers were provided with useful insights to act upon. Some of these actionable insights could be marketing campaigns, updated online advertisements, and website restructuring to improve site visit duration. During this case study, the main learnings centered around web analytic metrics used to predict website user purchases. Target customers were also identified based on these key web analytic metrics. In the future, another analyst might approach the problem with a similar strategy. However, there are a few recommendations for improvement. First, a few additional models could be constructed and compared for performance based on similar evaluation metrics. Some examples of other binary classification models are Support Vector Machines, Naïve Bayes, or K-Nearest Neighbor. If the model is deployed, then these additional models would be worth investigating to determine the best model is selected. Second, it would be wise to tune the models once built, trained, and tested. Lastly, it might be beneficial to attempt an additional approach to target how much a customer will spend rather than just predict that they will spend. This would be a good next step to take in the analysis.

**References**

Ecommerce Guide. (2022). What is Ecommerce? [What is Ecommerce? Definitions, Examples, And The Origin of Online Shopping (ecommerceguide.com)](https://ecommerceguide.com/guides/what-is-ecommerce/)

GobiemoUSA. (2022). Web Analytics Basics. *usability.gov*. [Web Analytics Basics | Usability.gov](https://www.usability.gov/what-and-why/web-analytics.html)

Sahu, Suchismita. (2021, September 6). PREDICTIVE WEB ANALYTICS: A CASE STUDY. *Analytics Vidhya*. [PREDICTIVE WEB ANALYTICS: A CASE STUDY | by Suchismita Sahu | Analytics Vidhya | Medium](https://medium.com/analytics-vidhya/predictive-web-analytics-a-case-study-f30feda45002)

Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018). Online Shoppers Purchasing Intention Dataset. *UCI Machine Learning Repository*. [UCI Machine Learning Repository: Online Shoppers Purchasing Intention Dataset Data Set](https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset)

**Assignment 5.2 Requirements:**

Find an example of a predictive analytics case study either in the textbooks or online. From this case study, write a four-page minimum write-up that includes the following.

Introduction

* What was the problem being solved?
* Why was this problem important to solve?
* How was the data acquired?

Methods and Results

* What steps were taken to prepare the data?
* How was this problem solved?
* What modeling techniques were used?
* Why did the team choose the methods/models they did?
* What metrics were used to evaluate the results? Why was this metric chosen?

Conclusion

* How were the results or model implemented?
* What were the actionable consequences of the case study?
* What did the team learn from the case study?
* How should or would the team approach the problem differently in the future?

You do not need to do any of the technical work for this project. This is more of a high-level assignment where you are explaining the process that was used in a project.