Effects of feature selection on the prediction of

Online Shoppers Purchasing Intention

Toronto Metropolitan University

CIND820: Big Data Analytics Project

Chang School of Continuing Education

Maryam Paknejad

Student Number: 500868652

Supervisor: Ceni Babaoglu

Submission Date: July 24th, 2022

Github link: https://github.com/Mary-PN/Chang-School_Big-data-certificate

Table of contents

Abstract

> "E-commerce refers to the purchase and sale of goods and/or services via electronic channels such as the internet. E-commerce was first introduced in the 1960s via an electronic data interchange (EDI) on value-added networks (VANs). The medium grew with the increased availability of internet access and the advent of popular online sellers in the 1990s and early 2000s." (Rivera, 2021)

Online shopping's accessibility and comfort paired with the option of comparing products from the convenience of our homes had already flooded our lives for many years. However, the lockdowns derived from the COVID-19 pandemic made the competition fiercer for e-commerce companies around the world.

To be a part of this upward trend or stay in online business competition, the companies are relying more and more on studying the customers' behaviors online, recognizing the strength and weaknesses of their portal, and adopting their marketing strategies accordingly.

The objective of this study is to recognize the patterns in customer trends and behavioral insights by using click stream analysis to propose business insights, and to improve the prediction algorithms results of the customer browsing routes using feature selection on classification models.

This project uses the "Online Shoppers Purchasing Intention Dataset" from the UC-Irvine Machine Learning Repository.

> "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. The dataset consists of 10 numerical and 8 categorical attributes."

Throughout this project, exploratory data analysis is used to determine the effects of attributes on the dependent variable using visualization and other techniques. Clustering techniques also are used to derive business insights. Then, the data is prepared for modeling. In predictive modeling, the focus of the study is on improving the performance by examining three feature selection criteria. Five machine learning models are trained, evaluation measures are calculated, and model performances are compared.

The final report includes insights to improve online e-commerce and marketing strategies, and suggestions on feature selections methods to better predict customer behavior.

Literature Review

> "E-commerce refers to the purchase and sale of goods and/or services via electronic channels such as the internet. E-commerce was first introduced in the 1960s via an electronic data interchange (EDI) on value-added networks (VANs). The medium grew with the increased availability of internet access and the advent of popular online sellers in the 1990s and early 2000s. Amazon began operating as a book-shipping business in Jeff Bezos' garage in 1995. EBay, which enables consumers to sell to each other online, introduced online auctions in 1995 and exploded with the 1997 Beanie Babies frenzy. Like any digital technology or consumer-based purchasing market, e-commerce has evolved over the years." (Rivera, 2021)

In recent years, the competitive marketing nature of e-commerce hasn't been limited to the website design, association rules, or offering similar items based on the customer interest anymore. Thanks to the recent advantages provided by neural networks and deep learning, the competition has been elevated to decreasing the risk of customers leaving the webpage as early as the moment they enter the website. There has been a lot of research and proposals made in this area to improve predicting the online customer behavior in real time. While Neural Networks have brought various options to satisfy the ongoing improvement of the designed models for categorizing the customers, the scientists have been constantly trying to improve the machine learning models' accuracy by experimenting new methodologies on the classification algorithms. There have been many experiments conducted in the area of customer intention in the past with some trying to improve the existing historical achievements.

In 2015, Suchacka et al. examined predicting buying sessions in a Web store using classification and two label classes: browsing and buying sessions. They used the k-Nearest Neighbors (k-NN)

on data from an online bookstore. They concluded that an 11-NN classifier is most effective both in predicting buying sessions, with high sensitivity of 87.5% and accuracy of 99.85%. Unfortunately the high accuracy of the kNN classifier does not compensate for the slow process of this lazy-learning algorithm. Hence this algorithm won't be included in this experiment.

In 2018, Sakar et al. proposed a system consisting of two modules to simultaneously predict customer's intention and the probability of leaving the website using data collected from online user sessions for a Columbia Sportswear company available on UCI Machine Learning Repository. The first module was a long short-term memory-based recurrent neural network, which would assign a score to the probability of a customer leaving the website without purchase. If a certain threshold was met, the second module would be activated, which would use classification algorithms to predict the purchasing intention, and only offer content to those visitors with high probability of purchasing intention. The same dataset, and the second model (classification) is the focus of this study. For the second module, Sakar et al. chose to evaluate random forest (RF), support vector machines (SVMs), and multilayer perceptron (MLP) classifiers. The training and test sets were selected randomly using oversampling for the 100 time modeling experiments, and the results were evaluated using accuracy and F1 score. To determine the significance in the difference between the achieved accuracies, a t-test was also conducted. The MLP produced significantly higher accuracy and F1 Score than RF and SVM.

To complement the Sakar et. al. model, in 2020, Baati et al. suggested a real-time online shopper behavior prediction system to predict the customer's shopping intention as soon as they visit the website using session and visitor information using naïve Bayes classifier, C4.5 decision tree and random forest. They used the same dataset but claimed to design a first type of marketing offer to all potential visitors as soon as they enter the website, and a second type of more generous offer

(similar to the second module proposed by Sakar et al.) to the customers with high purchasing intention. For modeling, a 70-30 percent split was used for training and testing sets respectively, and similar to Sakar et al. the experiment was repeated 100 times with randomly splitting training/test sets. The results ranked random forest at a significantly higher accuracy and F1 Score. A year later, Baati published an extension to the experiment with the goal of consolidating the previously suggested system by using hybridization of adaboost with random forest. The results showed that the novel system outperforms the previous one in terms of accuracy and F1 score.

In another study, in 2019, Kabir et al. explored the performance of various algorithms on the same dataset with the goal of identifying a more accurate predictive model. They used Random Forest, Decision Tree, Naive Bayes, and SVMK classification algorithms as well as ensemble methods such as Stacking, Voting, Bagging (Random Forest), Bagging (Extra Tree), Adaboosting, and Gradient Boosting . The result showed that Random Forest is more suitable than others amongst the algorithms but compared to ensemble methods, Gradient Boosting increased the accuracy even more.

In 2020, Kiki et al. performed a similar analysis on the same dataset. They studied the correlation between the attributes and inferred new information. Then, they fed the data to gradient boosting, artificial neural networks, and few other algorithms to predict the purchase intention, and used precision and F1- Score to evaluate the model performances. They demonstrated that the gradient boosting model performed better due to the new features used. Beside high interpretability, gradient boosting was promoted as a competitive algorithm to neural networks.

In the same year, Shagirbasha explored the same dataset based on the functional and visible design factors of websites such as the visual, informational and navigational design

characteristics of online portals and how they affect the perception of online consumers. Exploratory Data analysis was used to extract business insights. For the predictive analysis Artificial Neural Network, XGBoost, and Logistic Regression were designed, and ranked for performance in the mentioned order using accuracy as the evaluation measure.

Later in 2020, Kurniawan et al published an experiment to Improve

> "The Effectiveness of Classification UsingThe Data Level Approach and Feature SelectionTechniques in Online Shoppers Purchasing Intention Prediction". (Kurniawan et al. 2020)

The experiment was conducted in three sub-categories: "*Single Classifier*", "*SMOTE + AdaBoost Performance Results + Classification*", " *PSO+ SMOTE + AdaBoost Performance Results + Classification*". The article concluded that with the use of AUC as the evaluation measure, the second method outperforms the other two methods.

In 2021, Noviantoro et al. applied eight different classification algorithms such as Decision Tree, kNN, Random Forest, Naive Bayes, Neural Network, Logistics Regression, Deep Learning, and Rule Induction on the same dataset, and used accuracy, F1 Score and ROC (receiver operating characteristic) graphs to evaluate the models' performances. The results showed that the Neural Network has the best accuracy and F1 score, while Random Forest was the fittest in ROC curve. For this experiment, 10 fold stratified cross validation was used to reduce the bias on the majority class.

In the beginning of 2021, Mootha et al. published an experiment titled "*the stacking ensemble of Multi Layer Perceptrons to Predict Online Shoppers' Purchasing Intention*" using UCI dataset. They compared their experiment to more than 15 classification models including previous experiments that had used the same UCI dataset. They concluded that with the achieved accuracy

of 94% the proposed stacking classifier outperformed all previous ones. While this experiment has shown considerable improvement in classifying the customer purchasing intention using the UCI dataset, it solely relies on comparing the accuracy with previous studies . It should be noted that the high accuracy in this dataset is easily affected by the high number of True Negative predictions due to the imbalanced data, and it can only be a measure in cases where the target customers are **not** the ones with high purchase intention. This is contrary to the goal of the experiment conducted by Sakar et. al. where the purpose was only to offer content to those visitors with a high probability of purchasing intention, and therefore they have used F1-Score alongside Accuracy as the evaluation measure.

Furthermore, the majority of studies have focused on improving the accuracy by applying different algorithms and hyperparameter tuning the machine learning models. It appears that feature selection has not been the focus in the majority of previous studies. This approach was also proposed for future research by Mootha et al. in 2021.

To contribute to the process of improving the classification algorithms, the focus of this paper is on studying the feature selection effects on prediction results.

As mentioned, the goal of classifying customers can be aimed to offer incentives to the ones with high purchasing intention, or vice versa. For that reason this paper includes five evaluation measures in comparing the models to provide insight based on the business needs. Accuracy, Precision, Recall, F1- Score, and AUC-ROC are calculated and compared for each model and each set of features.

Dataset Description

This project uses the "Online Shoppers Purchasing Intention Dataset" from the UC-Irvine Machine Learning Repository.

“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.

The dataset consists of 18 attributes: “"Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration" represent the number of different types of pages 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 each page in the e-commerce site. The value of "Bounce Rate" feature for a web page refers to 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. The value of "Exit Rate" feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session. The "Page Value" feature 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 Valentine'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." (UCI Machine Learning Repository, 2018)

The Revenue attribute is selected as the classification attribute with 1908 sessions as positive class which resulted in purchase, and 10422 sessions of negative class which resulted in abandonment of the webpage without purchase. The proportion of positive and negative classes are 15.47% and 84.53% respectively, which demonstrates an imbalanced dataset.

Table 1 shows the numerical attributes, a short definition, and their descriptive statistics. Table 2 shows the categorical attributes, a short definition, and number of categories. Four of the categorical attributes already have numbers as their category values in the original dataset.

Table 1: Numerical attributes

Attribute	Description	mean	std	min	0.25	0.50	0.75	max
Administrative	Number of pages visited related to account management	2.32	3.32	0.00	0.00	1.00	4.00	27.00
Administrative_Duration	Total time spent on the account management pages (seconds)	80.82	176.78	0.00	0.00	7.50	93.26	3398.75
Informational	Number of pages visited related to website information	0.50	1.27	0.00	0.00	0.00	0.00	24.00
Informational_Duration	Total time spent on the website information pages (seconds)	34.47	140.75	0.00	0.00	0.00	0.00	2549.38
ProductRelated	Number of pages visited related to products	31.73	44.48	0.00	7.00	18.00	38.00	705.00
ProductRelated_Duration	Total time spent on the products pages (seconds)	1194.75	1913.67	0.00	184.14	598.94	1464.16	63973.52
BounceRates	Average percentage of site traffic that resulted in a single page visit	0.02	0.05	0.00	0.00	0.00	0.02	0.20
ExitRates	Average percentage exit rate value of the pages visited	0.04	0.05	0.00	0.01	0.03	0.05	0.20
PageValues[1]	Average page value of the pages visited	5.89	18.57	0.00	0.00	0.00	0.00	361.76
SpecialDay	a measure for closeness of the visiting time to a special day	0.06	0.20	0.00	0.00	0.00	0.00	1.00

[1] More information regarding how the Page Value is calculated: https://support.google.com/analytics/answer/2695658?hl=en

Table 2: categorical attributes

Attribute	Description	Category Count
Month	The month the visit occurred	10
OperatingSystems	Operating system of the user	8
Browser	Type of the browser of the user	13
Region	The location area of the user when entered the page	9
TrafficType	The source where the user was directed to the page from	20
VisitorType	Visitor types: New Visitor, Returning Visitor, and Other	3
Weekend	demonstrating if the day of visit was on a weekend	2
Revenue	Class label indicating if a purchase occurred or not	2

Methodology

In this case study, the following steps are taken to complete the project:

1. Exploratory Data Analysis: Univariate, bivariate and multivariate analysis is conducted on the original dataset to derive information and business insights. The results were also used for feature selection purposes.
2. Feature Selection: There are 3 experiments conducted to compare the results based on various combination of independent variables used in training the models:
    - ➢ Including all 17 independent variables in predictive analysis
    - ➢ Conducting Feature Selection based on correlation, and Chi-squared test (for discrete/categorical variables), which resulted in including 14 independent variables in predictive analysis
    - ➢ Conducting Feature Selection based on Mutual Information (threshold > 0.01) and Chi-squared test for discrete/categorical variables, as well as correlation, which resulted in including 9 independent variables in predictive analysis

In the first experiment, all 17 independent variables were used to predict the class label "Revenue". No feature selection technique was applied at this step so the results could be used as

a base for further comparison. During Exploratory Data Analysis, it was observed that there is a high positive linear correlation between two sets of the independent variables. As depicted in figure 1-1, “BounceRates” and “ExitRates” are highly correlated (Pearson: 0.91), regardless of the value for “Revenue”. In figure 1-2, the high positive linear correlation (Pearson: 0.86) between the “ProductRelated” and “ProductRelated_Duration” has been demonstrated which is also regardless of the value for “Revenue” . These were two of the applied feature selection results, and were included in all sets of experiments.

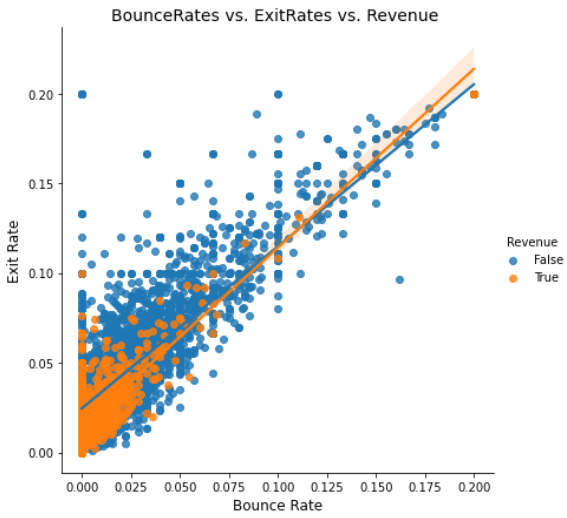


Figure 1-1

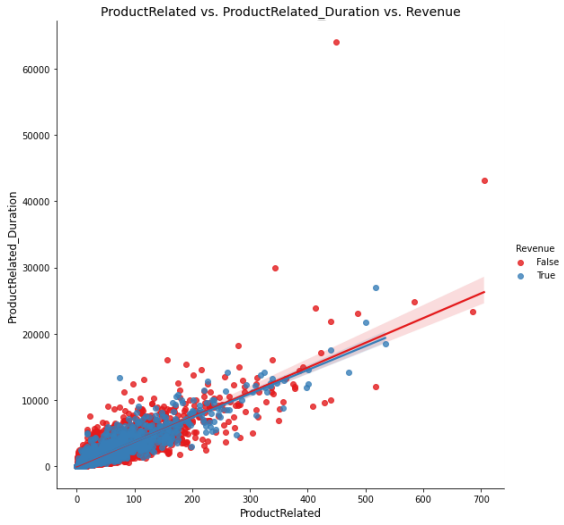


Figure 1-2

The selection of “ExitRates” and “ProductRelated” versus their perspective correlated variable was decided based on higher variance observed during bivariate analysis in relation to the dependent variable “Revenue” (figures 2-1 and 2-2)

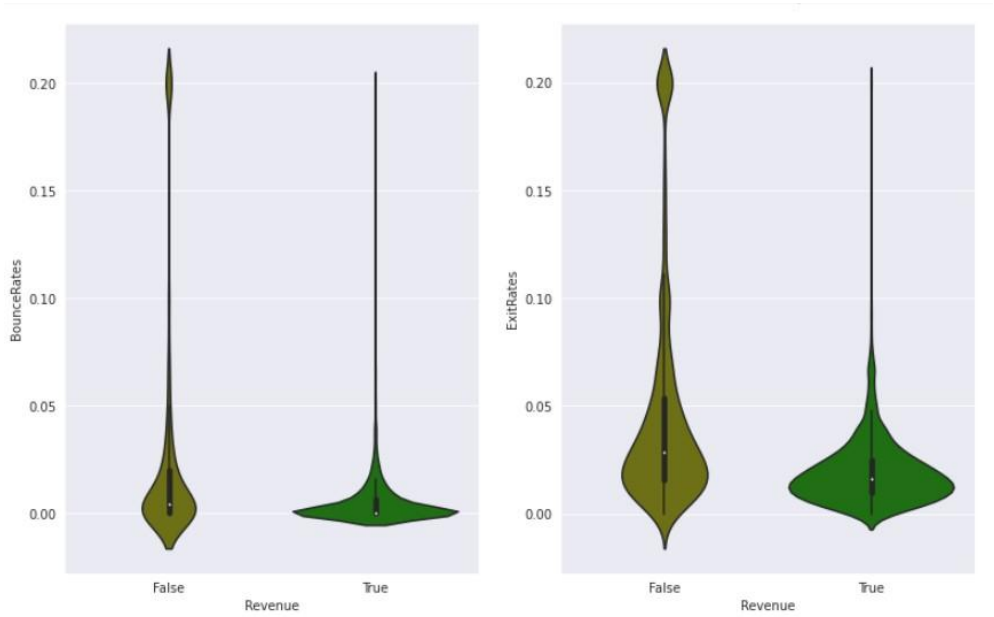


Figure 2-1

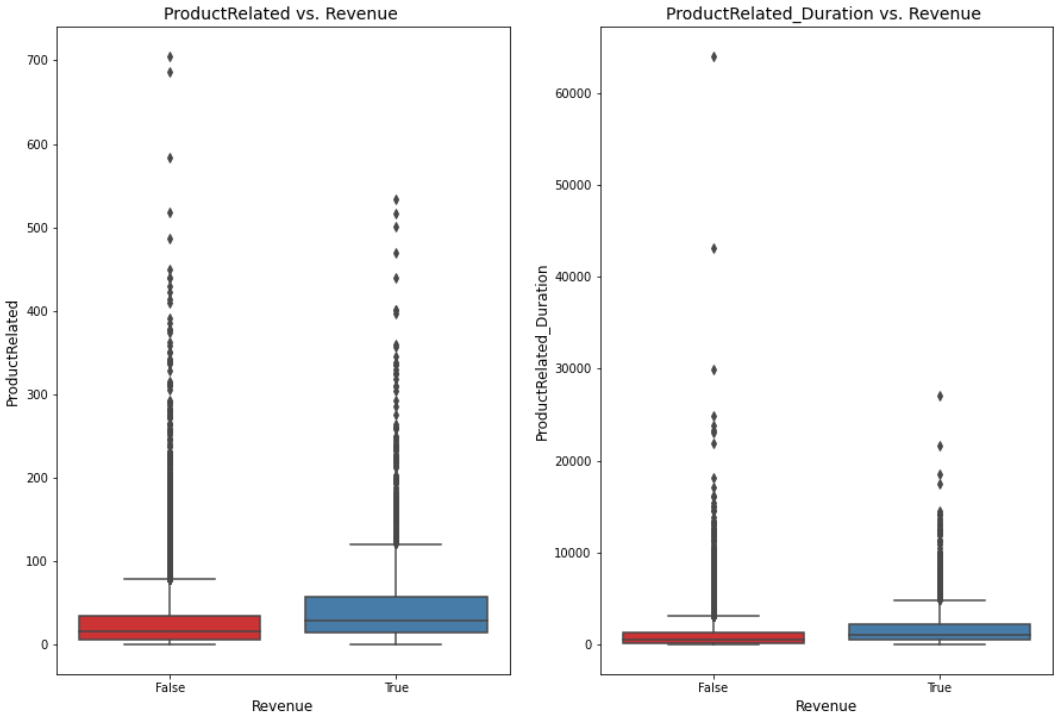


Figure 2-2

The next applied feature selection method was using the result of Chi-Squared Test of Independence on categorical and discrete variables to see which attributes are associated with predicting the dependent variable "Revenue". The result of the test showed that out of seven

categorical and 3 discrete independent variables (continuous variables were excluded), "Revenue" is independent of "Region" (CI: 95%, P_Value: 0.3214). This finding was applied to the second and third experiments.

The last feature selection method was Information Gain by applying the Mutual Information Classification algorithm from *scikit learn* library. According to the *scikit learn* documentation:

> "The Mutual information (MI) between two random variables is a non-negative value, which measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency. The function relies on nonparametric methods based on entropy estimation from k-nearest neighbors distances…"(*scikit learn library*)

As the documentation points out this method can only be applied to discrete variables, therefore the continuous variables were excluded from the test but they remained in the modeling stage. The categorical non-discrete variables were encoded by using discrete values. The result of the mutual information is demonstrated in figure 3. The line at MI=0.010 identifies the threshold.

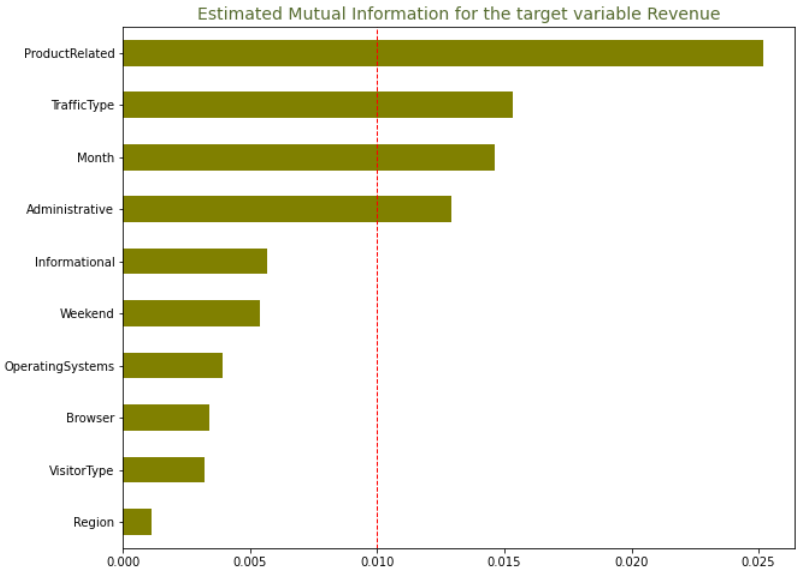


Figure 3

Therefore a total of eight (out of 17) independent variables were removed from the dataset. This included 6 variables from the information gain and 2 from correlation. It should be noted that the variable “Region” was considered independent of the “Revenue” by Chi-Squared test as well.

3. Preprocessing the dataset: After removing the appropriate features in each experiment, the data was preprocessed and prepared for modeling. First data normalization was applied to all numeric variables. Then the string variables were transformed into dummy variables with the exception of the dependent variable “Revenue”.
4. Machine Learning Parameters: In all stages of the experiment a single random State was used to ensure the model performances are comparable. The dependent and independent variables were separated, and they were split to 80% training set and 20% test set. The evaluation measures and cross validation measures were also pre-defined.
5. Machine Learning pipeline : A pipeline was set up using the following functions;
    - ➢ SMOTE function have been defined to be applied to the training set, to deal with the imbalance between two classes of dependent variables. SMOTE (Synthetic Minority Oversampling TEchnique) works by oversampling the minority class by creating synthesized samples.

        > “… SMOTE first selects a minority class instance at random and finds its k nearest minority class neighbors. The synthetic instance is then created by choosing one of the k nearest neighbors b at random and connecting a and b to form a line segment in the feature space. The synthetic instances are generated as a convex combination of the two chosen instances a and b.” (He H., Ma Y., 2013)

    - ➢ Then each classifier has been defined with their required parameters.

A 10 fold Cross validation has been set up and applied to the dataset using the pipeline above to train the model. The cross validation of folds is repeated 3 times, and has been set up with random state to ensure the consistency of folds between models.

The following Classification algorithms are applied during each experiment:

- Decision Tree:

  "A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions."

- XGB Classifier

  "XGBoost is an implementation of Gradient Boosted decision trees. In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model." (GeeksforGeeks, 2022)

In this study “logloss” was used due to binomial deviance of class label, and the maximum depth was set to 27 for all XGB models in all three experiments of feature selections.

- Random Forest:

  “Random Forest has multiple decision trees as base learning models. Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data and hence the output doesn’t depend on one decision tree but multiple decision trees. In the case of a classification problem, the final output is taken by using the majority voting classifier… ..The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.” (GeeksforGeeks, 2022)

- Support Vector Machine:

  “SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, then the data are transformed in such a way that the separator could be drawn as a hyperplane. Following this, characteristics of new data can be used to predict the group to which a new record should belong.” (IBM, 2021)

The Support Vector Classification was used with a linear kernel for all models in all three experiments of feature selection.

- Logistic Regression:

> "Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. In simple words, the dependent variable is binary…" (Tutorials Point)

The number of irritations for Logistic Regression was set to 2000 to prevent non-convergence of the algorithm with the low default number of irritations.

6. Assessing the models: If the algorithm performs well on the training/validation folds, then model assessment is performed on the unseen test set after fitting each model according to the same parameter used in the model training stage, and by predicting the class label.
7. Model Performance Evaluation: The model performance is measured by using the predefined scores:

   ➢ Accuracy, Precision, Recall, F1-Score: Please see the figure 4 for definitions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\text{-}score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Figure 4

   ➢ AUC_ROC:

   " The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the True Positive Rate against False Positive Rate at various threshold values and essentially

separates the ‘signal’ from the ‘noise’. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.” (Bhandari, A. 2022)

8. Results and Conclusion: The Business insights from EDA are presented. The performances of models during each experiment is discussed. The performance of models are also compared in regards to various feature selection methods. Potential future research questions are discussed.

In figure 5 the process has been described in a diagram.

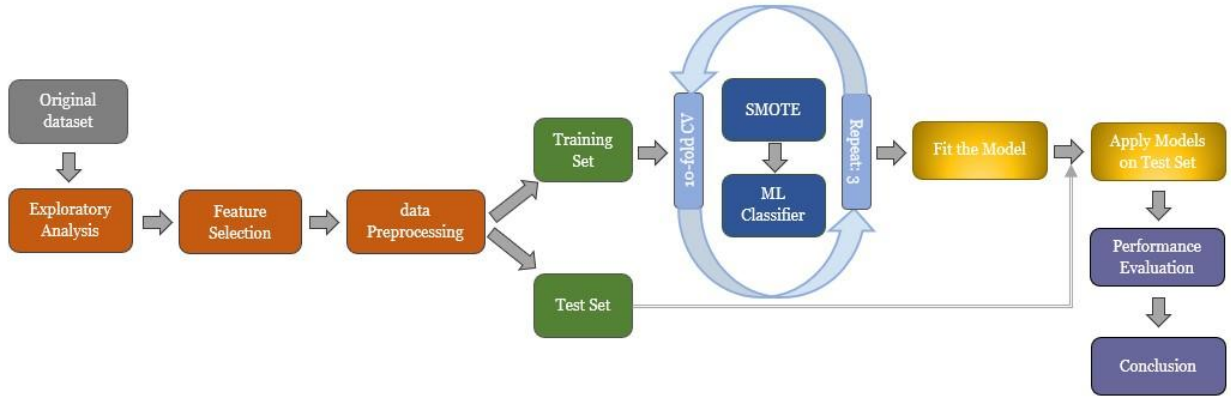


Figure 5

The results of the experiment are presented in two sections. First the insights derived from exploratory data analysis are presented. Next the predictive analysis models and their performance on unseen test data is compared.

Results: Exploratory Data Analysis

The following insights were derived from the original data, and can be used for marketing strategies and future decision making:

- The average number of visits on a weekday is higher than a weekend day. However the conversion rate on weekends (17.2%) is higher than weekdays (14.8%). While this may happen due to various reasons, marketing strategy could be directed at having specific sales during the weekdays to encourage the finalization of the purchase.
- The majority of the visitors (85.6%) are a "Returning_Visitor". This is a positive note for the company, however it also relays the need for focus on advertisement to attract new customers. On the other hand, the new visitors have a higher conversion rate than returning ones. Many marketing strategies can help increase the conversion rate for returning visitors; such as: offering discount on second purchase after the first purchase is made, creating a points system for every dollar spent, offering a certain percentage discount after a certain amount spent in a time frame.
- Region 1 has significantly higher visitors than others. In general the “Region” - ”Revenue” frequency chart (figure 6) provides info to the marketing team to focus on advertisement in specific areas to increase the number of visits by considering marketing strategies specific to region. For example, in regions with low number of visits like 5, finding which competitors are in upward conversion trends and why! It's worth taking a survey from people visiting from region 1 to determine the main reason for high number of visits and low conversion rate. The survey can help the marketing team to adapt strategies for advertisement in regions 2 to 9 to increase the number of visits. For example, sometimes fans are drawn to a specific sport e-commerce website, because the aesthetics of the website includes the color of their states’/city’s sport team. Therefore, making customized colors for each region can have an impact on the number of visits.

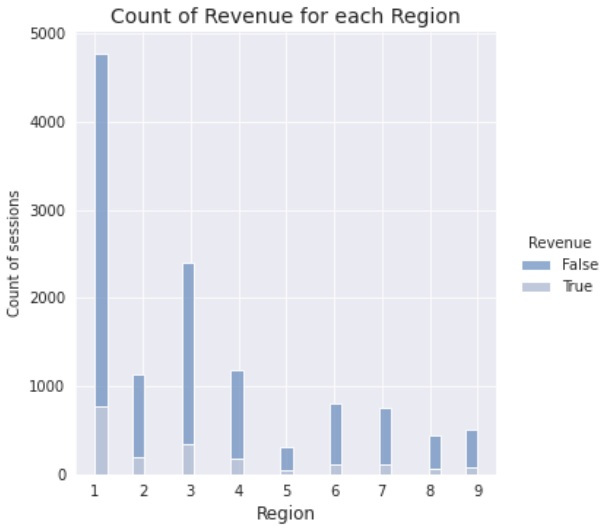


Figure 6

- There are no visits from any customers in "January" and "April". The month of May has the highest number of visits probably due to closeness to the summer vacations period, followed by November due to closeness to the Christmas time and Holidays. Seasonal sale periods can be drawn from this information.
- The majority of visits do not happen close to a special day. This seems to be the result of considerations regarding the delivery period for the purchase.
- Browsers 12 & 13 have higher conversion rates than type 1 & 2, while 1 & 2 have the highest number of visits. Aside from popularity of some browsers, the marketing team must compare the differences between the browser types and find out what advantages certain browsers are providing that may have a positive effect on the company's revenue. For example: some browsers such as "Edge" from "Microsoft" have an accessibility option ("Read aloud this page") enabled in the browser by default, while for "Google" such option needs to be found from the setting, and activated. In such a case, adding an accessibility option to the e-commerce website can be beneficial.
- TrafficType 1, 2, and 3 together count for the majority of visits as well as high conversion rate overall. This can be explored further to discover the reason for this result. For

example, if these TrafficTypes are being directed from highly trusted websites, the focus can be placed in expanding redirection from such trusted webpages

- The majority of visitors do not access "Administrative" pages during their visit, and they only spend 0-100 seconds on the "Administrative" Pages. The data shows the same trend for "informational" pages and its duration. However, the majority of visitors access between 0-20 “ProductRelated” pages and spend 0-1000 seconds on them, which is significantly higher than other types of pages as expected. The higher the duration spent on any type of pages, the lower the Exit Rates, and higher conversion rate (figure 7). Making these types of pages more engaging and interesting can result in more time spent on the pages and consequently higher conversion rate.

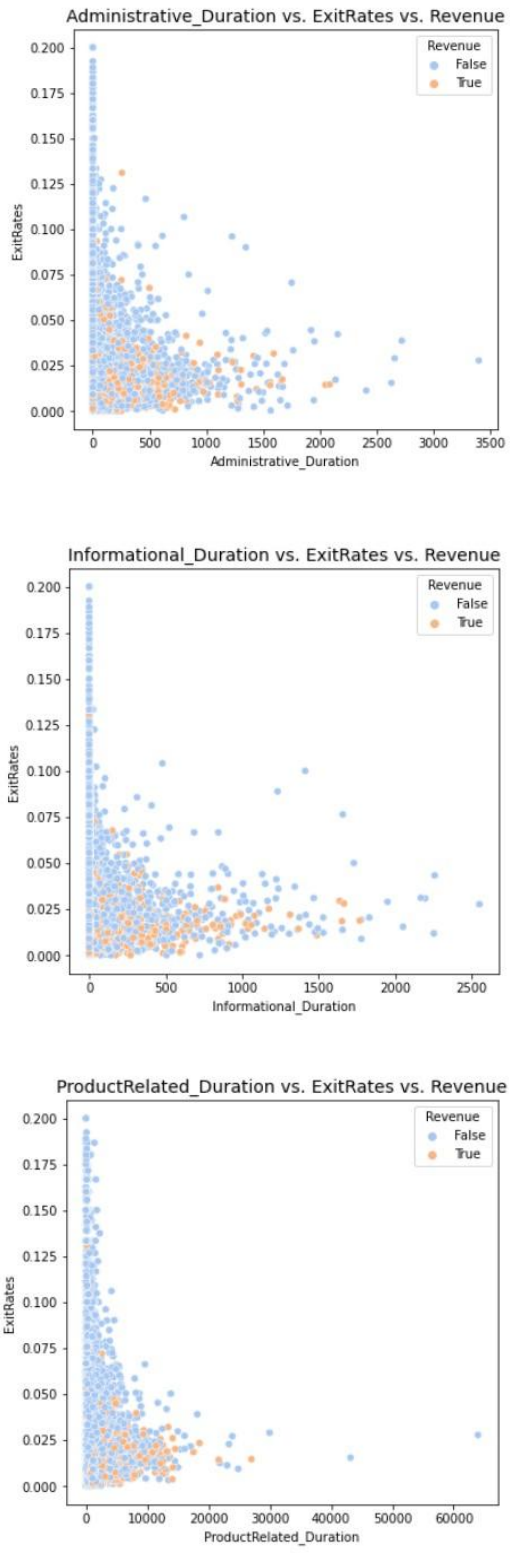
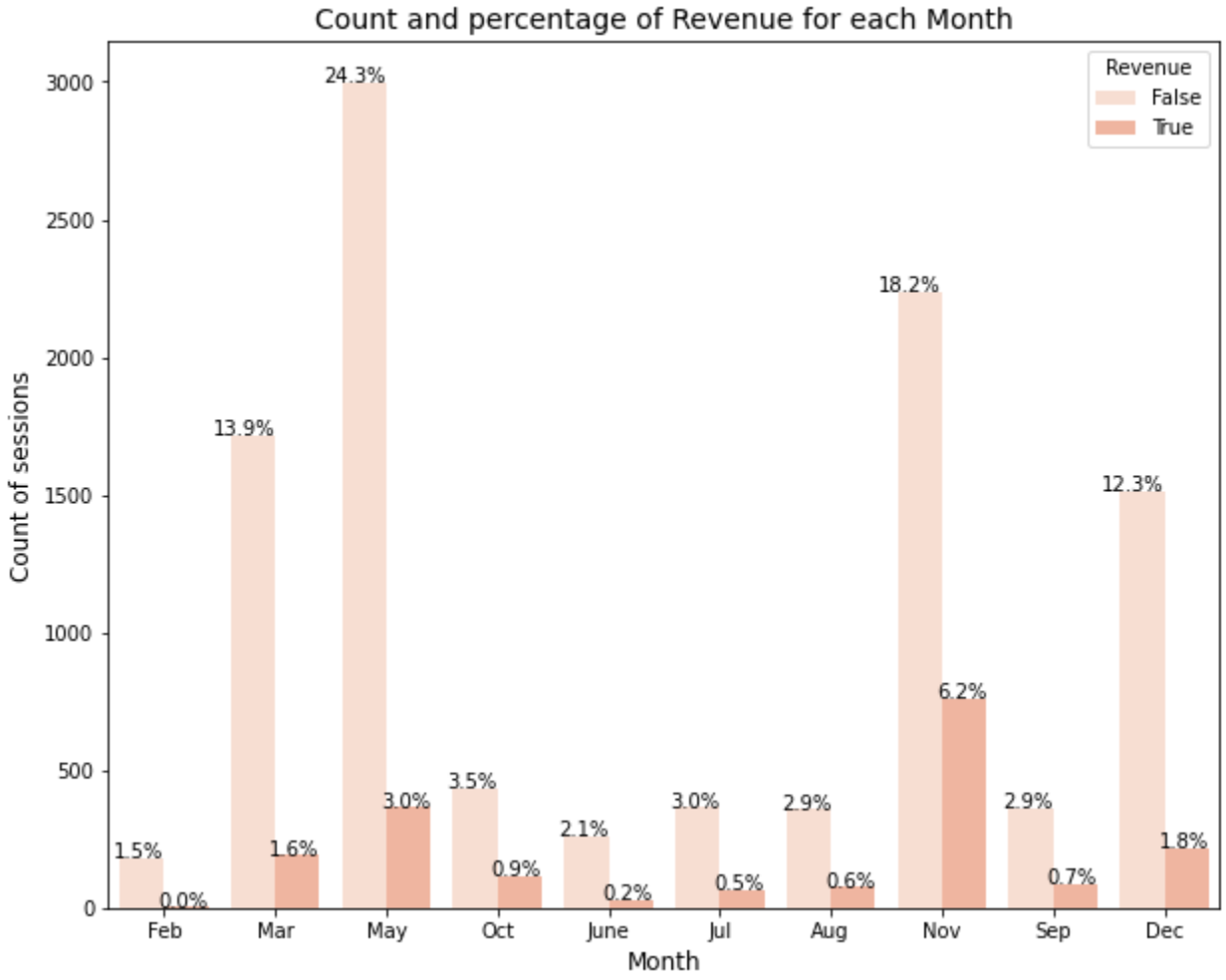


Figure 7

- Despite a high number of visits in May (24.3% of annual visits), the conversion rate is half of November. Clearly Christmas Holidays have a big effect on the high purchase rate in November. (figure 8). The company needs to develop new marketing strategies (i.e. offering special discounts for summer gear) specific to May and March to increase the purchase rate as these two spring months have high number of visits and low conversion rate. They should also look into fast delivery options for December as this month also has a high number of visits and a very low conversion rate.



Results: Predictive Analysis

As mentioned in the methodology section, five machine learning base models were selected with no hypertuning, to compare the effects of feature selection on the model performance. Three different sets of independent variables were fed to all the models and the results are presented in Table 3 to 5. The following independent variables were removed from each experiment:

1. No feature selection was applied
2. Excluded “Productrelated_Duration”, “BounceRates”, “Region”
3. Excluded “Productrelated_Duration”, “BounceRates”, “Region”, “VisitorType", "Browser", "Weekend","OperatingSystems", "Informational"

Classifier	Accuracy	Recall (False)	Recall (True)	Precision (False)	Precision (True)	F1-Score (False)	F1-Score (True)	ROC_AUC
XGB	0.895	0.960	0.581	0.917	0.752	0.938	0.655	0.770
Random Forest	0.890	0.964	0.531	0.909	0.752	0.935	0.622	0.747
Decision Tree	0.857	0.922	0.543	0.907	0.589	0.914	0.565	0.732
SVC (Linear)	0.851	0.993	0.164	0.852	0.821	0.917	0.273	0.578
Logit	0.845	0.996	0.118	0.845	0.847	0.914	0.208	0.557

Table 3: Result for modeling with all 17 Independent Variable

Classifier	Accuracy	Recall (False)	Recall (True)	Precision (False)	Precision (True)	F1-Score (False)	F1-Score (True)	ROC_AUC
XGB	0.883	0.951	0.557	0.912	0.699	0.931	0.620	0.754
Random Forest	0.885	0.961	0.517	0.976	0.732	0.933	0.621	0.739
Decision Tree	0.853	0.914	0.557	0.909	0.573	0.912	0.565	0.736
SVC (Linear)	0.869	0.978	0.344	0.878	0.763	0.925	0.474	0.661
Logit	0.872	0.977	0.365	0.882	0.762	0.927	0.494	0.671

Table 4: Result for modeling with 14 Independent Variable

Classifier	Accuracy	Recall (False)	Recall (True)	Precision (False)	Precision (True)	F1-Score (False)	F1-Score (True)	ROC_AUC
XGB	0.876	0.948	0.526	0.906	0.675	0.927	0.591	0.737
Random Forest	0.881	0.954	0.531	0.908	0.702	0.930	0.605	0.742
Decision Tree	0.853	0.926	0.502	0.900	0.582	0.913	0.539	0.714
SVC (Linear)	0.870	0.978	0.344	0.878	0.767	0.926	0.475	0.661
Logit	0.871	0.976	0.363	0.881	0.754	0.926	0.490	0.669

Table 5: Result for modeling with 9 Independent Variable

In comparing the results above, it becomes apparent that the selection of the model for predictive purposes will mainly depend on the marketing strategy. If the aim is to only offer incentives to customers that have lower intention of purchase to encourage them to make a purchase, then the Recall for class “False” becomes more important which means the need to minimize the False Negative (customers that had lower intention but were wrongly missed by model due to being classified as high intention. purchase). Lower False Negative leads to a higher Recall score for class “False”.

On the other hand, if the plan is to offer incentives to customers that have higher intention of purchase, so they don't leave the website without finalizing their purchase, then the Recall for class “True” is more important. Specifically due to the imbalanced dataset this is one the most important considerations in comparing the performance of the models based on feature selection, and consequently in classifying the customer’s purchase intention. The next measure to consider for the class “True” is the ROC_AUC score which takes into account the True Positive Rate/False Positive Rate. The higher the ROC_AUC, the better a model is distinguishing the positive outcome.

If the purpose is to target both types of customers with a different marketing strategy for each then Accuracy and F1-Score are the measures to consider.

Focusing on the main purpose of the study, both feature selection methods show considerable improvement in the performance of all models based on the selected evaluation measures compared to the base model without any feature selection. Mainly, the improvement is noticeable on the class “True” in “Recall” and “F1-Score”. Furthermore, the 14 features selected on the second experiment perform in a more balanced way for both class “True” and “False” across all the models. This also includes better results on ROC Area under the curve results.

It’s worth noting that the results from tree based algorithms are less impacted by feature selection compared to the algorithms such as Logistic Regression which is more sensitive to high correlation amongst the independent variables. Overall picture also identifies that tree base models with combined decision trees perform better in these types of classifications.

Conclusion

This experiment focused on the effects of feature selection in improvement of model performance in predicting the online shoppers purchasing intention. Due to the time constraints in delivering the result of the experiment only 5 algorithms were selected to be examined. The results confirmed the improvement across all evaluation measures by effective feature selection on all 5 algorithms, and included balancing the performance of the predictive modeling in both classes of this imbalanced dataset. The effective feature selection in this experiment was the use of correlation and the Chi-squared test of independence however the information gain method using mutual information classification did not significantly increase the achieved improvement compared to the primary feature selection method. The added benefit of reducing the number of features is to reduce the runtime of the classification task, specifically when such models can be

designed alongside a neural network with the purpose of real time prediction of the customer behaviour.

In the future, the feature selection methods and hyperparameter tuning of the models can be applied at the same time to further investigate the performance of predictive modeling. Furthermore, other Ensemble Methods and Neural Networks models can be also added to the experiments.

References

1. Baati, K. (n.d.). *Hybridization of adaboost with random forest for real-time prediction ...* Retrieved May 29, 2022, from https://www.researchgate.net/publication/350937568_Hybridization_of_Adaboost_with_Random_Forest_for_Real-Time_Prediction_of_Online_Shoppers'_Purchasing_Intentio
2. Baati, K., & Mohsil, M. (1970, January 1). *Real-time prediction of online shoppers' purchasing intention using Random Forest*. SpringerLink. Retrieved May 29, 2022, from https://link.springer.com/chapter/10.1007/978-3-030-49161-1_4
3. Bhandari, Aniruddha. AUC-ROC Curve in Machine Learning Clearly Explained. Retrieved on July 21, 2022 from https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/#:~:text=The%20Area%20Under%20the%20Curve,the%20positive%20and%20negative%20classes
4. GeeksforGeeks, XGB. Retrieved on July 16th from https://www.geeksforgeeks.org/xgboost/
5. Haibo He, Yunqian Ma. Imbalanced Learning: Foundations, Algorithms, and Applications 1st Edition. Retrieved on Jul 16, 2022 from: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/
6. IBM, How SVM works? Retrieved on July 16, 2022 from https://www.ibm.com/docs/it/spss-modeler/SaaS?topic=models-how-svm-works3
7. Kabir, M. R., Ashraf, F. B., & Ajwad, R. (n.d.). *Analysis of different predicting model for online shoppers' purchase ...* Retrieved May 29, 2022, from

https://www.researchgate.net/publication/340058413_Analysis_of_Different_Predicting_Model_for_Online_Shoppers'_Purchase_Intention_from_Empirical_Data

8. Kiki, Y., & Houndji, V. R. (2020, October 30). *Prediction of the purchase intention of users on ECommerce platforms using gradient boosting*. ResearchGate. Retrieved May 29, 2022, from https://www.researchgate.net/publication/354492965_Prediction_of_the_Purchase_Intention_of_Users_on_ECommerce_Platforms_using_Gradient_Boosting
9. Kurniawan, I., Abdussomad, A., Akbar, M. F., Saepudin, D. F., Azis, M. S., & Tabrani, M. (n.d.). *(PDF) improving the effectiveness of classification using the data ...* Retrieved May 29, 2022, from https://www.researchgate.net/publication/347138247_Improving_The_Effectiveness_of_Classification_Using_The_Data_Level_Approach_and_Feature_Selection_Techniques_in_Online_Shoppers_Purchasing_Intention_Prediction
10. Mootha, S., Sridhar, S., & Devi M S, K. (n.d.). *A stacking ensemble of multi layer perceptrons to predict online ...* Retrieved May 29, 2022, from https://www.researchgate.net/publication/348447252_A_Stacking_Ensemble_of_Multi_Layer_Perceptrons_to_Predict_Online_Shoppers'_Purchasing_Intention
11. Noviantoro, T., & Huang, J.-P. (n.d.). *Applying data mining techniques to investigate online shopper purchase ...* Retrieved May 29, 2022, from https://www.researchgate.net/publication/352794957_APPLYING_DATA_MINING_TECHNIQUES_TO_INVESTIGATE_ONLINE_SHOPPER_PURCHASE_INTENTION_BASED_ON_CLICKSTREAM_DATA

12. Paul, S. (2018, October 10). *Diving deep with imbalanced data*. DataCamp. Retrieved May 29, 2022, from https://www.datacamp.com/tutorial/diving-deep-imbalanced-data#!
13. Rivera, A. (2021, December 21). *What is e-commerce? learn the basics: Business news daily*. Business News Daily. Retrieved May 29, 2022, from https://www.businessnewsdaily.com/4872-what-is-e-commerce.html
14. Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2018, May 9). *Real-time prediction of online shoppers' purchasing intention using Multilayer Perceptron and LSTM recurrent neural networks - neural computing and applications*. SpringerLink. Retrieved May 29, 2022, from https://link.springer.com/article/10.1007/s00521-018-3523-0
15. Scikit-learn. *Sklearn.feature_selection.mutual_info_classif.* Retrieved July 16, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_classif.html
16. Shagirbasha, S. (n.d.). *Online shopping behaviour: Predicting consumer's purchase intention*. Retrieved May 29, 2022, from https://www.researchgate.net/publication/357579024_Online_shopping_behaviour_Predicting_consumer%27s_purchase_intention
17. Suchacka, G., Skolimowska-Kulig, M., & Potempa, A. (n.d.). *A K-nearest neighbors method for classifying user sessions in e-commerce scenario*. Journal of Telecommunications and Information Technology. Retrieved May 29, 2022, from https://yadda.icm.edu.pl/baztech/element/bwmeta1.element.baztech-40e29335-8f5f-4d8c-aa93-8c13a90d1b2d

18. Tutorials Point. Machine Learning - Logistic Regression. Retrieved July 16th, 2022 from https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_logistic_regression.htm

19. UCI Machine Learning Repository: Online Shoppers Purchasing Intention Dataset Data set. (2018). Retrieved May 29, 2022, from http://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset#