DETECTION of ONLINE SHOPPERS' PURCHASING INTENTION

Gargee Singh

July 2024

TORONTO METROPOLITAN UNIVERSITY

DECLARATION

I hereby declare that this project report is based on my original work expect for Citations and quotations which have been duly acknowledged.

Name : Gargee Singh

Date : 2024-07-28

ABSTRACT

Shoppers go online first in over 60% of shopping occasions (ThinkwithGoogle 2018, n.d.) and this constantly flowing stream of clicks holds immensely insightful data about how and at exactly what point during their web session, do customers decide to make the payment or exit the website or not complete the transaction. Naturally, this has led to companies wanting to harness the power of all this data, learn from it and make their offerings increasingly more lucrative. The combined power of e-commerce, logistics and social media has always intrigued me, and I chose Clickstream analytics for this project using the Online Shoppers Purchasing Intention Data Set (UC IrvineMachine Learning Repository, n.d.).

Using this sequential clickstream dataset, my project's primary focus will be to answer the question whether machine learning can be used to predict a customer's intention to make the purchase or not followed by the second question which is a customer's likelihood to leave the website without buying a product more commonly known as website abandonment rate. I will try to answer these questions and also try to establish whether or not there is a significant relationship between different pages a customer visits on the website. The combined information from the model should give us answers that can help predict a website's purchase conversion as well as abandonment ratio.

The data set I will be using has ten numerical and eight categorical attributes and although there are no null values, it is imbalanced. As initial pre-processing steps I will be working on methods (e.g., Over Sampling, Under Sampling or resampling with different ratios) to balance the data and check for correlation between numerical attributes which would need to be normalized before being used in the model. Once finalized I will proceed to feature selection methods to use only the most relevant features for the models. Once finalized with a balanced ratio between class labels and relevant features the data set will be fed to the different models using the standard train to test ratios.

I propose to start with a classification model using the attribute 'Revenue' as the class Label to establish a decision tree of customer's making the purchase or not. For comparison I also plan to

run a random forest model as it is a stronger modeling technique and will help me limit overfitting as well as errors due to any biases in the data which I might overlook in the decision tree. For modeling I plan to use aggregate page views data along with session and user information present in the data. I also plan to explore Support Vector Machine learning models as well as Logistic Regression on my data set so as to have a few good comparisons to make.

In the end I will be using a common matrix to measure the scalability and performance of each model before recommendations about the best model that predicts whether the shopper will make the purchase or is merely scrolling through the shop in their current session.

Link to the Data Set, I downloaded the CSV file from this link – https://archive.ics.uci.edu/dataset/468

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 General Information

Global ecommerce sales reached $5.82 trillion, marking a 10% increase from the previous year. This figure is projected to rise to $6.33 trillion in 2024, representing an 8.8% annual growth rate - the third-highest growth rate in the forecast period from 2021 to 2027 (Shopify 2024). As online sales continue to expand and capture a larger share of the retail market, the ecommerce landscape is becoming increasingly competitive. Companies are intensifying their efforts to attract consumer attention and prevent customers from shifting to competitors. To achieve this, businesses are allocating substantial marketing budgets and focusing on creating personalized online shopping experiences.

Machine learning plays a crucial role in these efforts, with companies utilizing streaming analytics and investing in predictive systems designed to enhance customer engagement. However, these predictive systems do not always achieve high conversion rates, highlighting areas for improvement. One significant challenge is the "rare class problem," which arises from imbalanced datasets and overlapping data issues. This problem, prevalent in machine learning, can reduce the effectiveness of predictive models. The rare class problem occurs when there is a disparity between the majority and minority classes within a dataset . The second problem occurs when the majority and the minority samples overlap in the data space.

1.2 Importance of the Project

This project aims to help online retailers utilize machine learning to boost their conversion rates. The objective is to create a robust machine learning model that continuously learns from shopper data and accurately predicts a customer's likelihood of making a purchase. By implementing this model, companies can enhance their targeted advertising and discount campaigns, thereby attracting more customers to their websites and increasing the likelihood of purchases.

This project also aims to make a significant contribution to the research field by enabling researchers to compare the effectiveness of various over-sampling and under-sampling techniques. By evaluating the performance of different models using these techniques, the project seeks to provide valuable insights into their impact on machine learning outcomes.

1.3 Problem Statement

Using this sequential clickstream dataset, my project's primary focus will be to answer the question whether machine learning can be used to predict a customer's intention to make the purchase or not followed by the second question which is a customer's likelihood to leave the website without buying a product more commonly known as website abandonment rate. I will try to answer these questions and also try to establish whether or not there is a significant relationship between different pages a customer visits on the website. The combined information from the model should give us answers that can help predict a website's purchase conversion as well as abandonment ratio. Also since the data set is imbalanced , the project aims to deploy various data balancing methods before prediction models.

1.4 Scope of the Project

This project tries to detect online users' purchasing intention while working with an imbalanced data set. The project focuses on the unbalanced and overlapped data set. In order to balance the data set before recommending a model , the project utilizes

1. RUS as the undersampling technique.
2. Synthetic Minority Oversampling Technique (SMOTE)
3. Borderline – SMOTE
4. SMOTE – ENN

The project also aims to compare the performances of the classifiers with under sampling, oversampling and without any sampling.

- Logistic Regression
- Random Forest
- Decision Tree
- Support Vector Classification
- ADA Boost Classification
- Gradient Boost Classifier
- XGB Classifier

Python is the programming language for this project and a range of libraries and tools, including pandas, NumPy, Matplotlib, Scikit-learn, Imbalanced-learn, smote variants and seaborn have been used for creating these models.

1.5 Proposed Solution

The project will start with a base line model , based off the imbalanced and unscaled data set. Using that as a starting point the project will first employ various other models on imbalanced data set. In order to handle the imbalance and overlap I will utilize random under sampling as well as SMOTE over sampling methods to compare various model performances. Finally the project aims to recommend a model to be deployed for accurately predicting a customers' online purchasing intention.

CHAPTER 2

EDA & Literature Review

2.1 Literature Review

Online shopping has permeated our societies so well that it is hard to think of a time when we were not shopping on our mobile phones and any shopping was a trip to store near or far and sometimes may be another city, depending on where one lived. For instance, in 2021 Global Online retail sales saw a 27.6% increase , totalling 4.3 trillion [6]. Now all this online shopping has not only created an opportunity for business to be making sales at all times of the day but has also brought along challenges like how to stay competitive, retain previous customers while attracting new customers who will make purchases. Considering the wide reach and potential for growth there have been numerous studies aimed at understanding various aspects of online visits on an e-commerce website, some of which I researched.

At first the study by Sakar [1] which proposed a real time behaviour analysis system that aimed to simultaneously predict a customers' shopping intent and in contrast their likelihood to leave the website without making a purchase. Their methodology used the aggregated page view data tracked during the visit along with some session and user information and employed two modules wherein the first module determined if the user should be offered content, and the second module came into action only if a customer seems likely to abandon the site. A potential concern addressed by the second [2] research was the risk of abandonment implied in their method. Their research used ML algorithms such as Random Forest, Support Vector Machines (SVM) and Multi Layer Perceptron Classifiers.

Another study along similar lines are the one by [2] Karim Baati and Mouad Mohsil where they intended to build on existing research and suggest a system that allows to detect users with high purchase intention once connected to the website so the website could attract them into making a purchase. Though their study was based off the same data as that of Sakar [1] , they only worked with those pertaining to session and user information. Their research uses similar dataset as the second module of the first [1] research and while the first one predicts likelihood of abandonment; this research aims to predict customers with high purchasing intention or conversely likelihood of not leaving the website and then call another system to make recommendations for a potential purchase conversion. Their research primarily used Random Forest for their model. Another interesting study in the same arena though a different research question is one specific to the time period after COVID-19 and consumers in Lebanon and Bahrain which aimed at exploring the intention of the customers to shop online in the post pandemic era. Their study extended the technology acceptance model [TAM] to prove two key Hypotheses ,first Perceived Usefulness (PU) is positively associated with Attitude Towards Online Shopping(ATU) for online shopping post COVID-19 and the second that PU is positively associated with Intention Towards Online Shopping (ITOS) post COVID-19.

Another similar research is the one by Shierly and Sihombing [3] which explored the hypothesis that the factors that affect consumers' purchase intention are perceived benefit , perceived risk , hedonic motivation, trust and attitude towards online shopping. Their research used Structural Equation Modelling (SEM) to validate these hypotheses and concluded that there was no indication of a relationship between either perceived benefits or perceived risk to online shopping. A very important aspect of online shopping is reviews by previous customers which is explored in the article by Jin Yang, Rathindra Sarathy and JinKyu Lee [4]. Reviews play an important role in reducingthe risk and uncertainty that online buyers would assume before purchasing a product. This study examines the effect of general feedback or opinion of previous shoppers in product reviews represented by review balance and the number of reviews for a product. Their study establishes that no significant causal effect between perceived uncertainty and purchase intention which would mean a customers' decision to leave a website may or may not have been directly caused by perceived uncertainty related to a product.

As expected with the ongoing interest and research in this domain there have been a lot of similar research and papers that have explored various interesting aspects of customer's online shopping behaviour. The research in this field has been diverse with some looking at data for specific geographical region , before or after certain time periods. Some of them have been more focused on predicting behaviour patterns like customers who are more or less interested than other customers and in turn helping business improve their website or processes to improve purchase conversion. Some other researches have been more focused on understanding behaviour patterns and not necessarily making predictions like the study [3] for the region of Lebanon and Bahrain which extend the Technology Acceptance model among the others. Among all this wonderful research my research question is more inclined towards making a prediction with analysing data collected during a user’s website session, I will hope to be able to predict acustomers’ likelihood to make a purchase or not using machine learning models like Logistic Regression, Random Forest , Decision Trees etc. Though there is a lot of wonderful research on this subject I wish to be able to explore the problem using different algorithms which have not been very common for e.g. Cross Fold Validation , KNN etc. I wish to compare and rank results produced by each algorithm.

2.2 EDA

2.2.1 Data Set Features

Table 1.1 Numerical Features of the Online Purchasing Intention Data Set by Sakar et al.(2019)

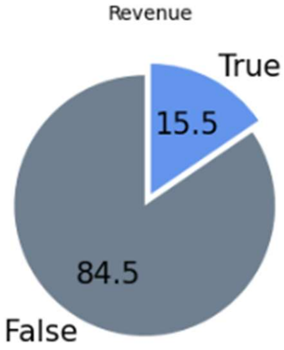
Administrative	The number of unique page categories visited by the visitor during the session.
Administrative_Duration	Total time duration spent (in sec) on administrative pages
Informational	The total number of Informational pages that the user visited
Informational_Duration	Total time spend (in sec) on Informational pages
ProductRelated	The total number of Product Related pages that the user visited
ProductRelated_Duration	Total time spend (in sec) on Product Related pages
BounceRates	The percentage of visitors who enter the website through that page and exit without triggering any additional tasks
ExitRates	The percentage of pageviews on the website that end at that specific page
PageValues	The average value of the page averaged over the value of the target page and/or the completion of an eCommerce transaction. More information about how this value is calculated
SpecialDay	The proximity of the site visit to a memorable day

Table 1.2 Categorical Features of the Online Purchasing Intention Data Set by Sakar et al.(2019)

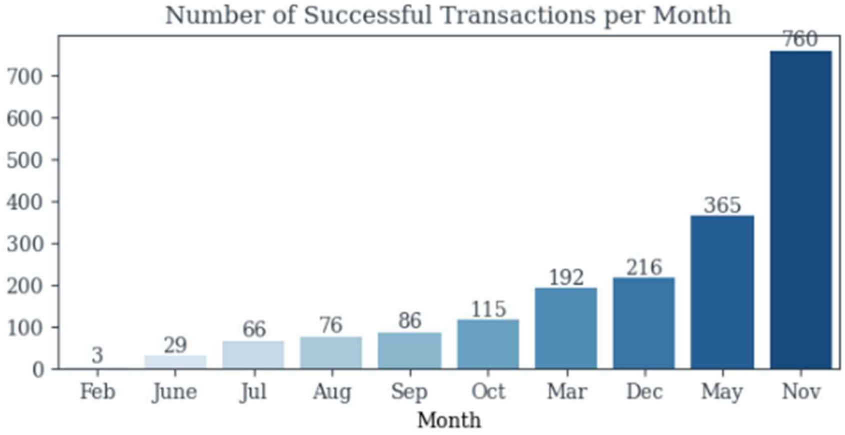
Month	Month value of the visit date
OperatingSystems	Operating system of the visitor
Browser	Browser used by the visitor
Region	Geographic region from which the visitor initiated the session
TrafficType	An integer value representing what type of traffic the user is categorized into. Read more about traffic types here.
VisitorType	Visitor type as ‘‘New Visitor,’’ ‘‘Returning Visitor,’’ and ‘‘Other’’
Weekend	Weekend value signifying whether the visit date is a weekend.
Revenue	Class label signifying whether a transaction was completed during the visit.

2.2.2 Visuals

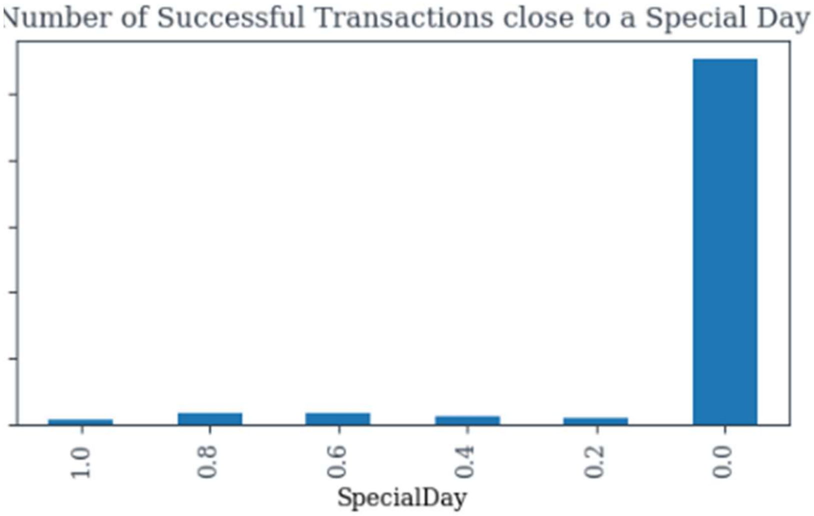
The data set is imbalanced with only 15.5% of the instances belonging to the minroity class.

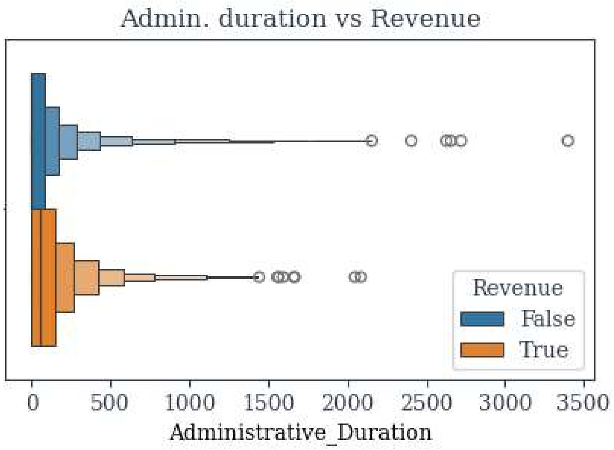


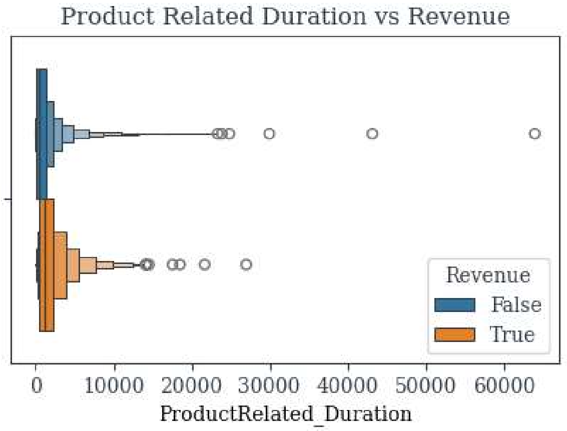
One significant analysis is that the number of successful transactions per month is the highest in November.

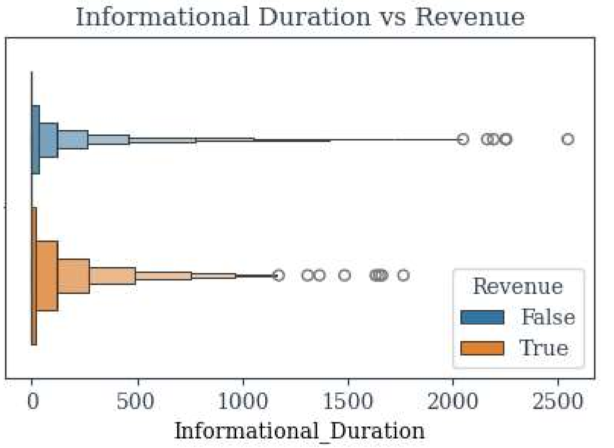


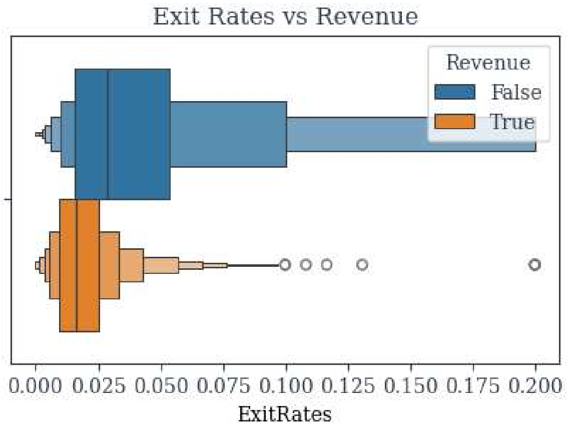
Another point worth noting is that the number of successful transactions surge dramatically on special days.




Admin. duration vs Revenue

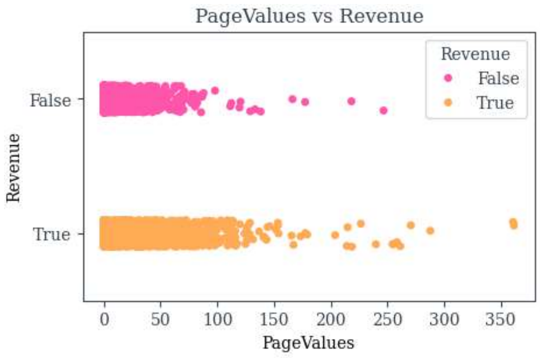

Product Related Duration vs Revenue

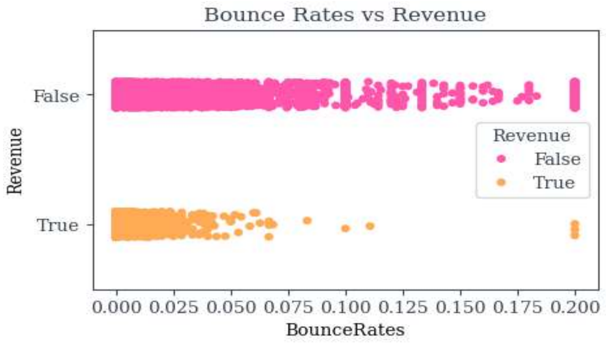

Informational Duration vs Revenue


Exit Rates vs Revenue

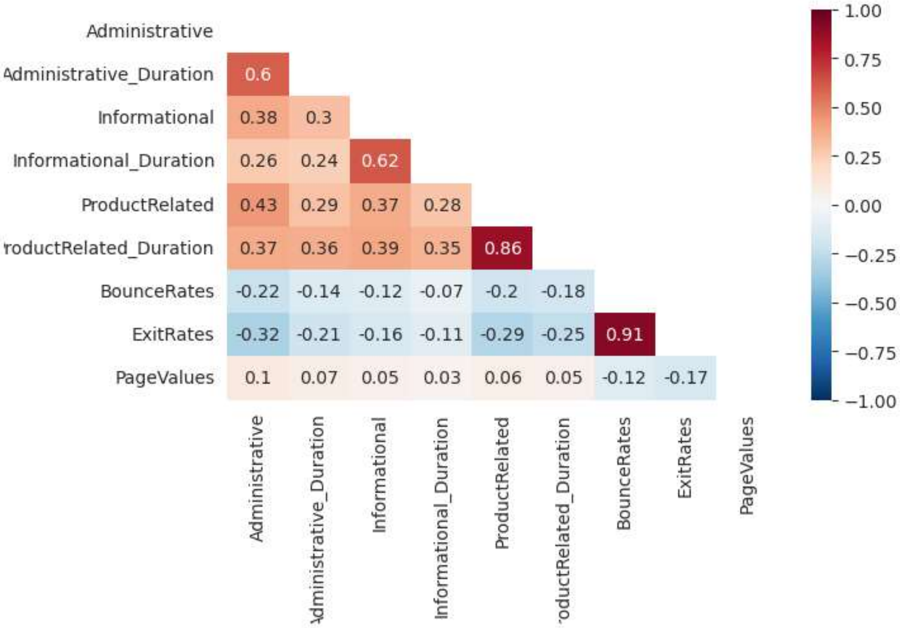
We see that Admin Duration, Product Related Duration and Informational Duration are exponentially distributed and have outliers for both TRUE and FALSE. Exit Rate is normally distributed but has outliers for records where purchase made was True.

Page Values is also exponentially distributed and has outliers , but is also influential on the Revenue Column and is a significant feature.

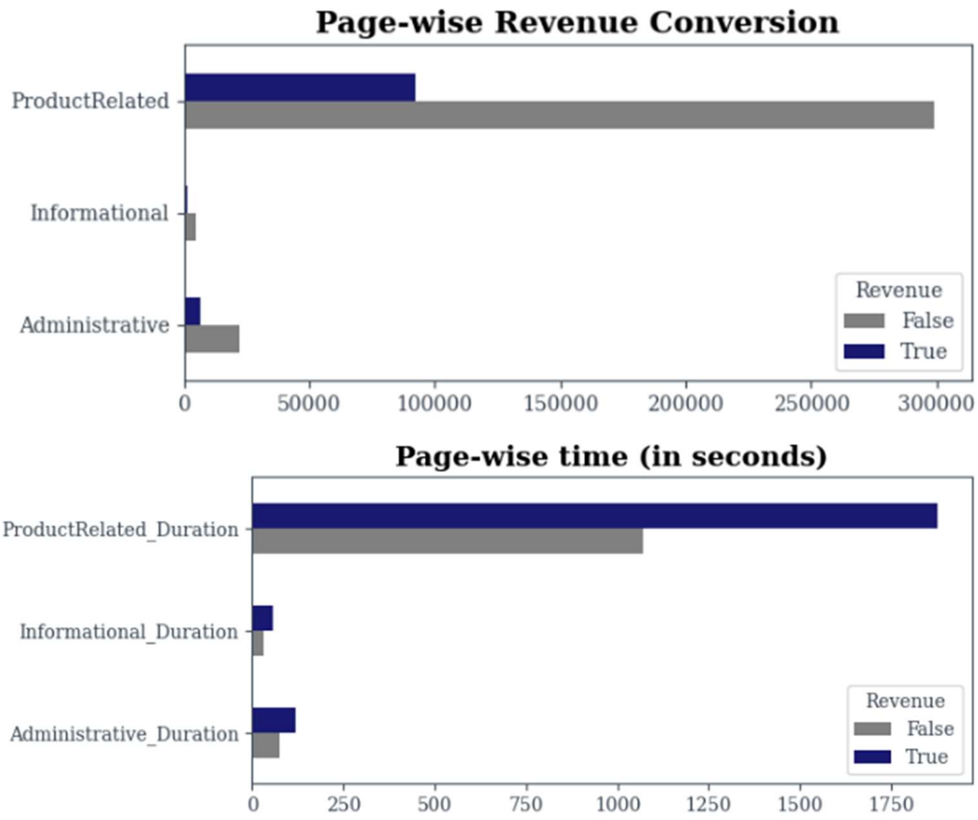

PageValues vs Revenue


Bounce Rates vs Revenue

Correlation within Numerical columns. Two of the feaures are highly correlated which will need to be normalized.

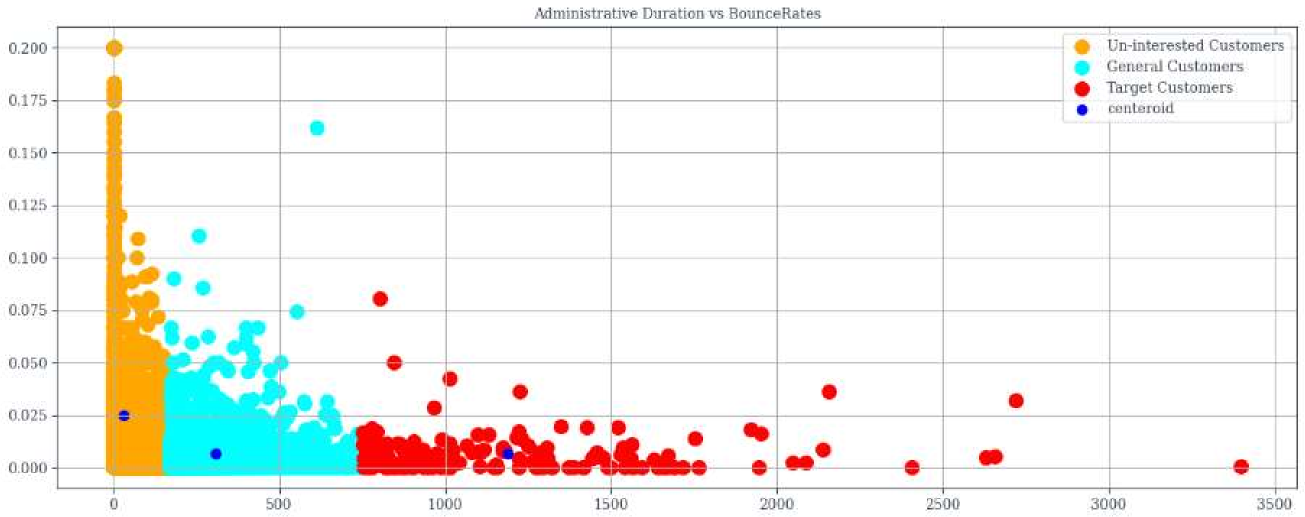


As can be expected amongst all the website pages and product related pages have the highest contribution to revenue generation.

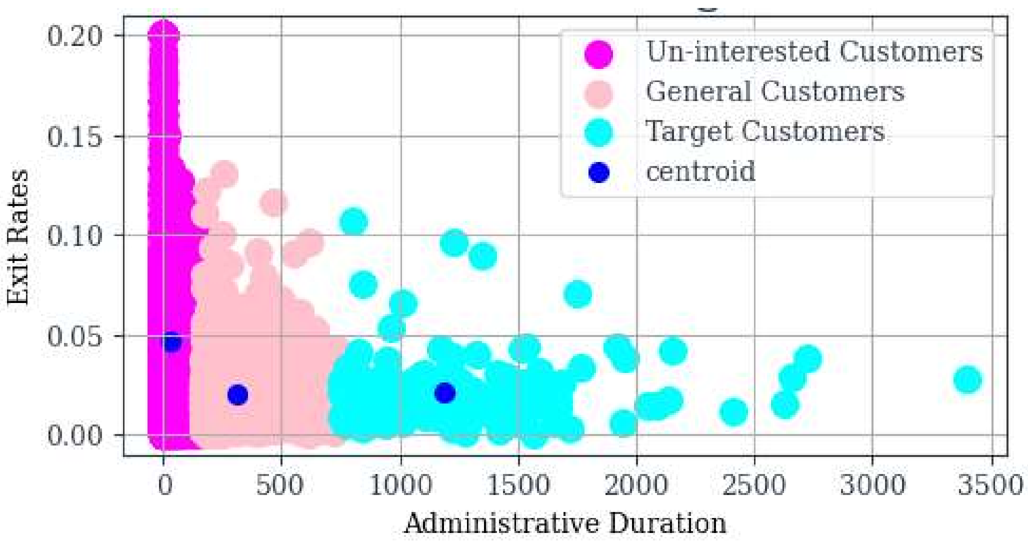


Understanding User Characteristic with Clustering Analysis

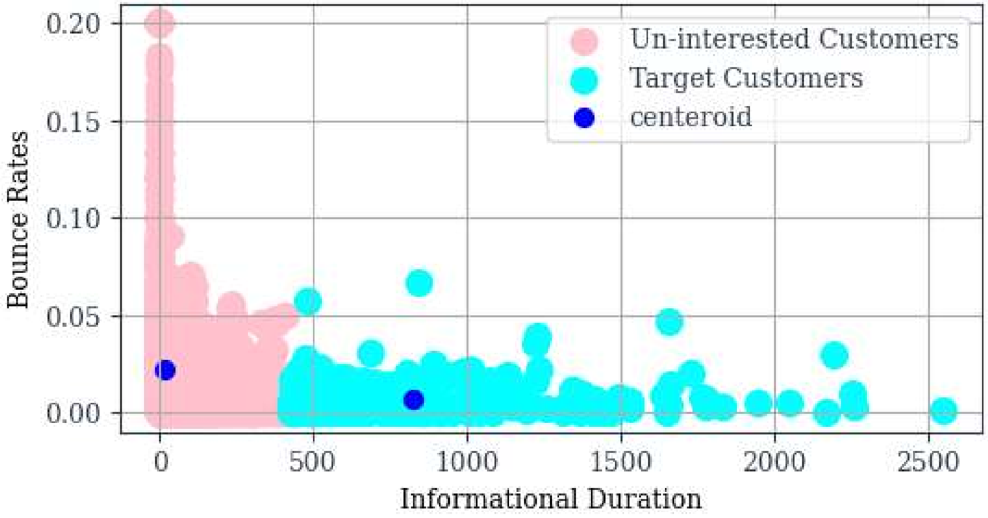
User's can be categorized into 3 groups based on time spent on administrative pages , we see that the bounce rates drop as the time spent on administrative pages increases and users who spent less than 500 seconds on administrative pages have the highest bounce rates.



Administrative Duration vs Exit Rates.

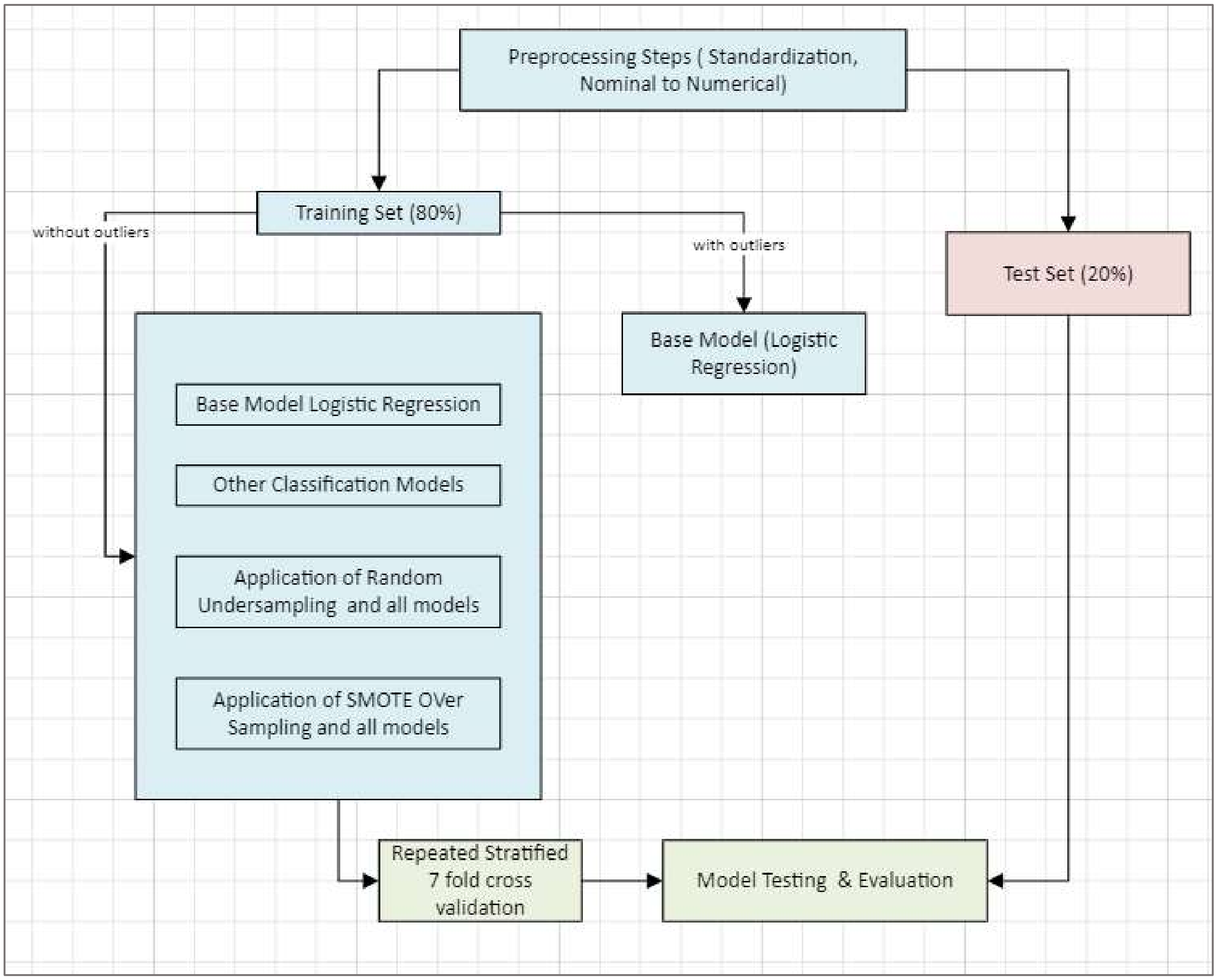


Similarly there are two groups of customers based on time spent on Informational Pages.



CHAPTER 3

Flow Chart of Workflow


Preprocessing Steps (Standardization, Nominal to Numerical)
Training Set (80%)
without outliers
with outliers
Test Set (20%)
Base Model (Logistic Regression)
Base Model Logistic Regression
Other Classification Models
Application of Random Undersampling and all models
Application of SMOTE OVer Sampling and all models
Repeated Stratified 7 fold cross validation
Model Testing & Evaluation

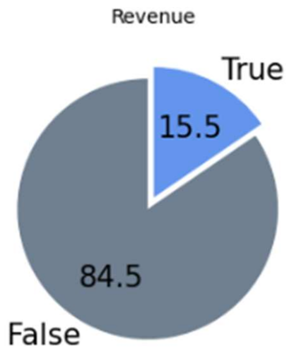
CHAPTER 4

ANALYSIS and WORKFLOW

3.2 Data Set Overview

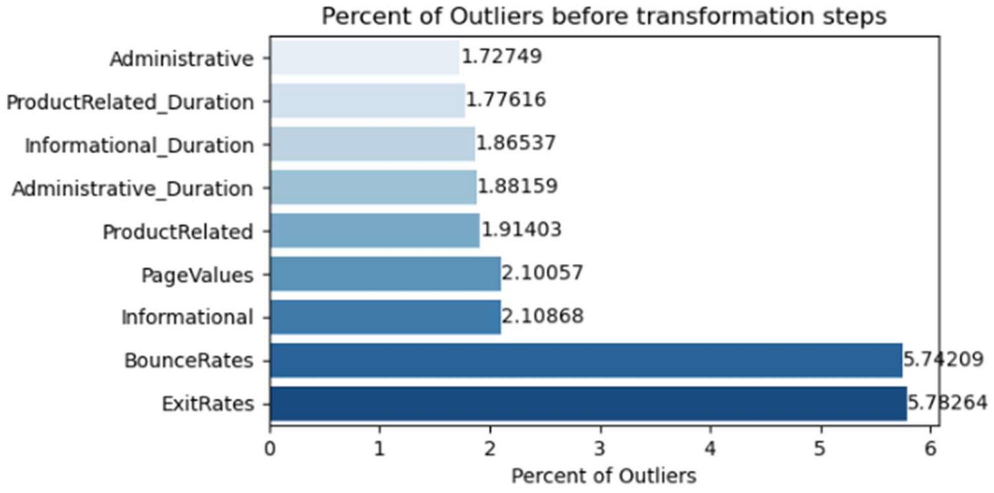
In this study, I have used the Online Shoppers Purchasing Intention Data Set from Sakar et al.(2018). The data set 12,330 instance and shows each session that belong to a different user in a 1-year period. The target column "Revenue" is a binary column where the values are True or False.

The data set is imbalanced , with only 15.5% (1908) of instances belonging to the minority class and 84.5%(10,422) belong to the majority class. Minority Class here is customer's who made a purchase while the majority instances are about customers who did not make a purchase.



The data set does not have any NULL values though there are few (125) duplicate records. .Due to the small volume and nature of the data as each line represent a user who accessed the website. It is possible that two users had a similar behaviors, we will not be removing duplicates and keep them in our data set.

The numerical columns have some outliers , but since most of these outliers belong to sessions where the clients made a purchase (minority class label) I will not be removing these outliers and try to minimize their impact with standardization.



3.3.1 Preprocessing Steps

Since the data does not contain any NULL values data cleaning is not required however there are a few issues to overcome.

1. First one being some outliers in numerical columns , these will be normalized with the help of BOXCOX transformation. The Box-Cox transformation is a family of power transformations designed to stabilize variance and make the data more closely approximate a normal distribution.
2. The second being the existence of nominal features where some of these have an integer data type. In order to avoid these features impacting the model as a numerical , the columns – (Month, Visitor Type Browser, Operating Systems and Traffic Type) been transformed into an indicator variable with the help of one hot encoding / label encoding. The two Boolean columns (Weekend and Revenue) were converted to integers.
3. To address the potential impact of feature scale on model performance, the Min-Max Scaler was employed. This technique transforms all features in the dataset to a uniform range, typically [0, 1]. By normalizing features within this specified range, the Min-Max Scaler helps to minimize bias that could arise from differences in feature scales, ensuring that no single feature disproportionately influences the model. This preprocessing step promotes a more balanced and effective model training process.
4. Feature Selection : chi2_contingency for categorical variables. For numerical variables I have used the below tests:
    1. Shapiro Wilk test
    2. Levene's Test
    3. Man Whitney U Test

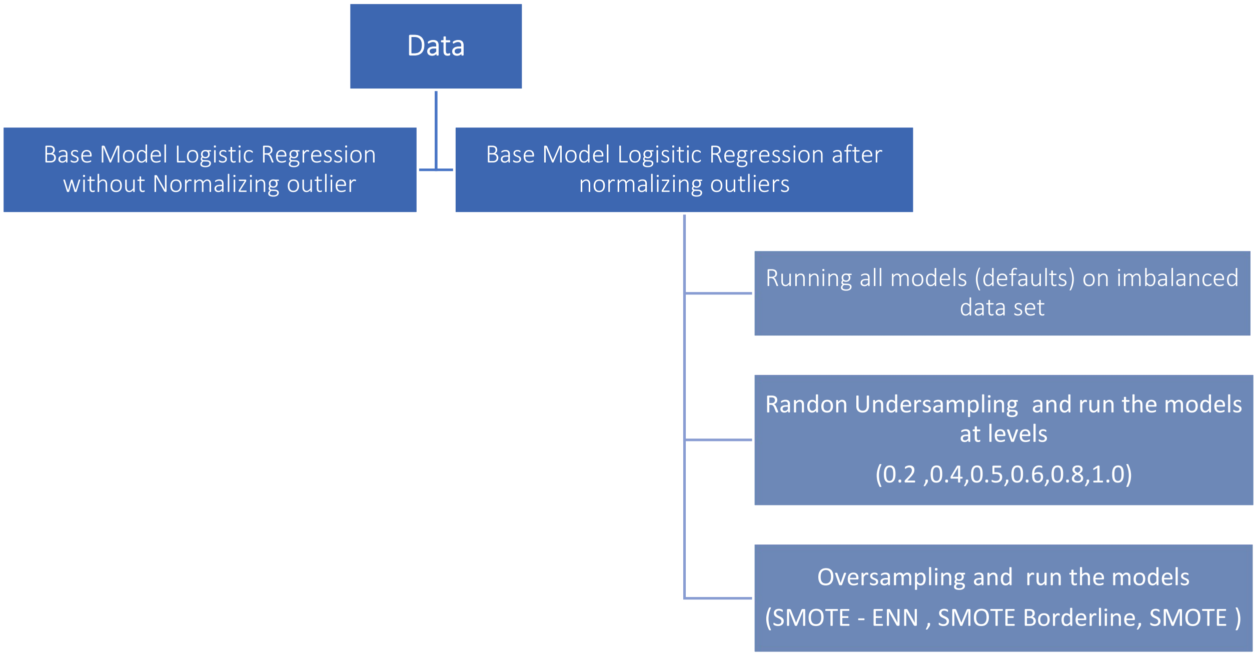
3.3.2 Train Test Split

The data set was split into training set and Test for various experiments to be run and compared. Splitting the data ensures that size of the majority and minority classes were same for every experiment. I have used a 80:20 train : test ratio , hence training the models on 80% of the data while testing it on 20% of the data set.

3.3.3 Sampling Methods

82 experiments were carried out on the data set. To establish a base performance I have a run a base model on the data without normalizing the outliers followed by running the model once again after normalizing the outliers. Observing some improvement in the model , I have continued to use the data with normalized outliers and further scaled it before feature selection and running the various other models on imbalanced data set.

Next I applied random undersampling at different level (0.2 to 1.0) and SMOTE oversampling (BorderLine and SMOTE ENN) . This helps establish a base line for comparing models with different methods and different levels of imbalance.


Data
Base Model Logistic Regression without Normalizing outlier
Base Model Logisitic Regression after normalizing outliers
Running all models (defaults) on imbalanced data set
Randon Undersampling and run the models at levels
(0.2 ,0.4,0.5,0.6,0.8,1.0)
Oversampling and run the models
(SMOTE - ENN , SMOTE Borderline, SMOTE)

3.3.4 Cross Validation and Evaluation

I have used Repeated Stratified kfold for cross validation , this involves simply repeating the cross-validation procedure multiple times (I have used split = 7 and repeated the process 3 times) . This methodsreports the mean result across all folds from all runs. This mean result is expected to be a more accurate estimate of the true unknown underlying mean performance of the model on the dataset, as calculated using the standard error.

Accuracy – Since the data set is imbalanced , I have not relied on accuracy alone in order to understand model performance , since most models will predict the majority labels quite often.

Precision &Recall : Since this is an online shopping business and the cost of missing a purchase prediction (minority label) is higher than the cost of a false positive , I have focused on recall since I am the model is able to predict most of the minority class ,i.e. scenario when a customer is likely to make a purchase. I think there will be higher tolerance for a false positive in this case i.e. based on a model prediction if the company offered a client to make a purchase but the client did not make a purchase.

F1 & F2 : For similar reason as above , since F2 places more weight on the recall value , I have used F2 score more model comparison.

CHAPTER 4

RESULTS AND DISCUSSION

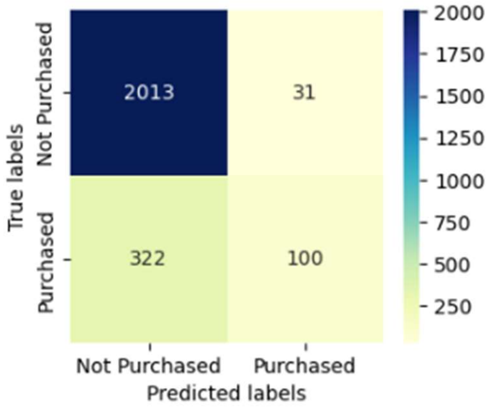
4.1 Base Model (Logistic Regression)

4.1.1 Logistic Regression without normalizing outliers

(Notes : Ro indicates Majority recall, R1 indicates Minority recall)

The first experiment is run on the data set without normalizing the outliers (one hot encoding and min max scaler were applied) as to keep that as a base line for all models.

This first base model runs with 0.86 accuracy which is decent start but we have to keep in mind that the data set is imbalanced so accuracy will not be an ideal measure. The imbalance is clearly reflected by the recall values, recall of the majority class (R_0)is at 0.98 while the minority class (R_1)is only at 0.24 which demonstrates that the classifier has a strong bias towards the Do Not Buy category .I have also looked at the F2 since for our model we would want to minimize false negatives but the F2 score is considerably lower than the F1 score which tell us that there is room to improve false negatives.



	Misclassifications	Accuracy	Precision	R_1	R_0	F1 Score	F2 Score	Type I errors	Type II errors
Logistic Regression	353	0.86	0.76	0.24	0.98	0.36	0.28	31	322

Next we compare the training and the test data set scores for this made model. The accuracy is on training is higher as would be expected but the lower precision on training score , indicates a possibility that we have a bad ratio of negative to positive in the test set , once again potentially caused due to the imbalanced data set. The difference in training and test accuracy is not high enough to suggest the model is overfitting.

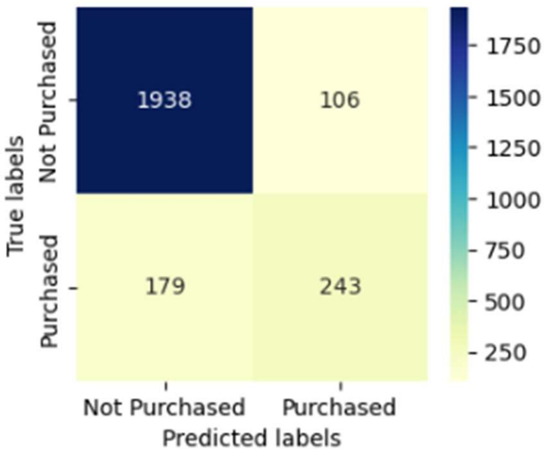
Data Set	Accuracy	Precision	Recall	F1 Score
Training Data	0.89	0.74	0.37	0.50
Test Data	0.87	0.75	0.36	0.49

As our performance measures have indicated the imbalance in the data set is affecting performance of the models so we will be deploying balancing techniques , I will be using SMOTE Oversampling (Standard and Border Line) and Random Under sampling to balance the data set.

4.1.2 Logistic Regression after normalizing outliers

As a next step we will perform the same experiment but this time after normalizing the outliers in order to see if it has any impact on model performance. The data set at this stage does not have the same ratio of outliers , the BOXCOX transformation has also reduced the correlation. One hot encoding and MinMax Scaler have also been applied as the base model.

We clearly see the impact of normalization on outliers ,accuracy of the model has slightly improved and Recall for the minority class and F2 score both have shown considerable improvement, which shows that despite the imbalance removing outliers has also improved model performance. On the flip side we see that though overall misclassification has gone down , false positive values have gone up ,we will monitor how this changes with other models however keep our focus on reducing the False Negatives.



	Misclassifications	Accuracy	Precision	R_1	R_0	F1 Score	F2 Score	Type I errors	Type II errors
Logistic Regression	285	0.88	0.7	0.58	0.95	0.63	0.6	106	179

Since we have seen positive impact of removing outliers on our base model, we will proceed with this data and apply other models on the data to compare performance.

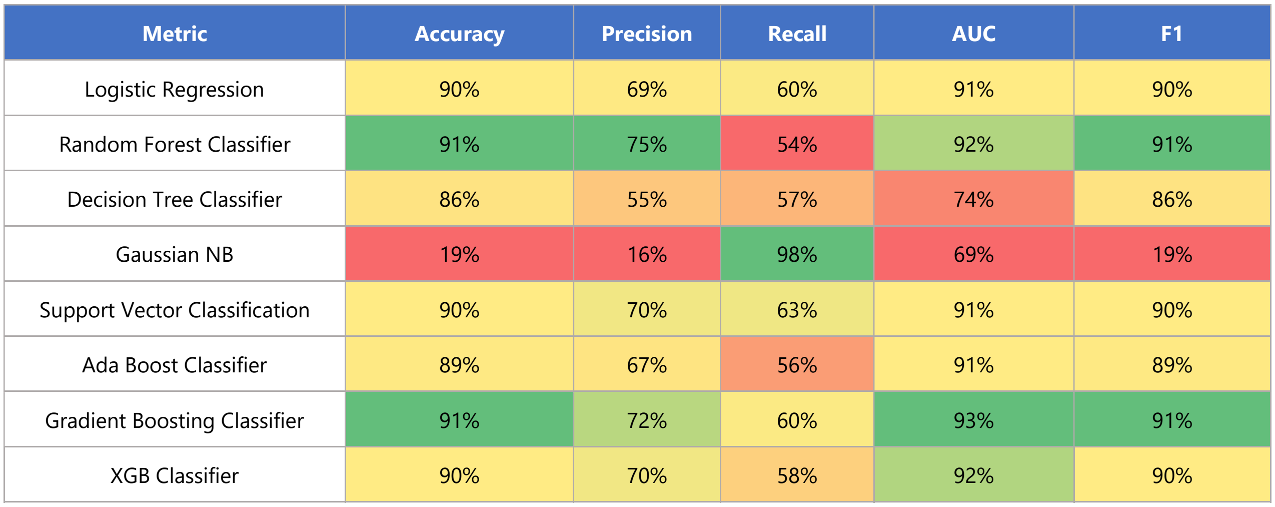
4.1.3 Feature Selection and running all other models with default parameter

Starting with 10 numerical and 8 categorical features , the data set ends up with 66 features (after one hot / label encoding). Before proceeding with modelling , feature selection was performed at this stage and the data set ends up with a count of **51 significant features**.

After feature selection , a total of 8 experiments were performed on the data set at this stage and their performance was compared with repeated stratified cross validation across 7 splits , repeating the process 3 times.

At this stage ,with imbalanced data set and having run these models at default , Random Forest Classification and Gradient Boost Classifier seem to return the best results followed by XGB Classifier. Gaussian Naiye Bayes seems to be the worst performing model with only 19% accuracy and despite the data being imbalanced towards the **“did not purchase”** class, the model is predicting 98% of instances as a **“purchased”** which is incorrect.

Cross Validation Scores :

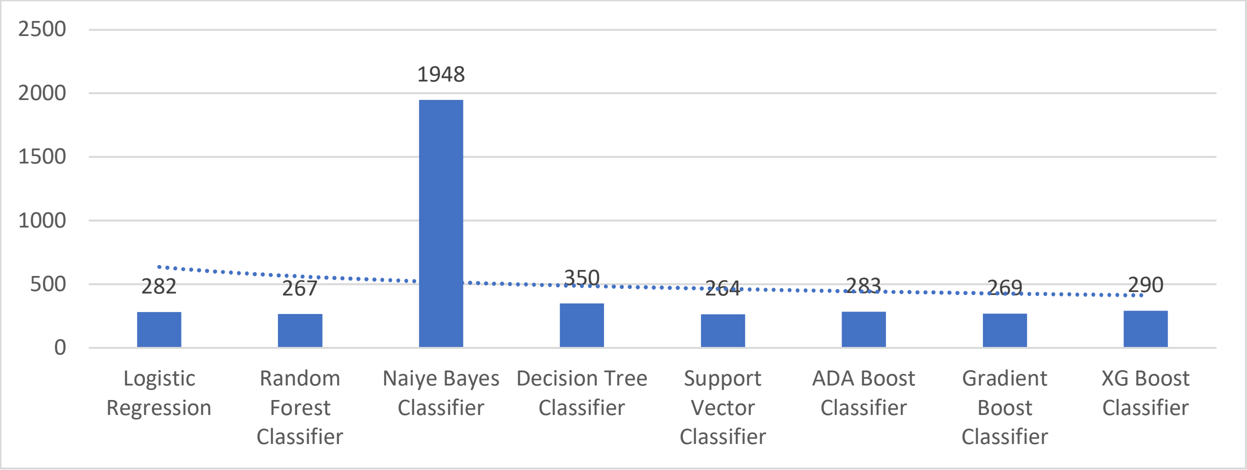


Metric	Accuracy	Precision	Recall	AUC	F1
Logistic Regression	90%	69%	60%	91%	90%
Random Forest Classifier	91%	75%	54%	92%	91%
Decision Tree Classifier	86%	55%	57%	74%	86%
Gaussian NB	19%	16%	98%	69%	19%
Support Vector Classification	90%	70%	63%	91%	90%
Ada Boost Classifier	89%	67%	56%	91%	89%
Gradient Boosting Classifier	91%	72%	60%	93%	91%
XGB Classifier	90%	70%	58%	92%	90%

Base Model Evaluation

1.1 Misclassifications

Naiye Bayes Classification has the highest number of misclassifications because unlike other models it is also misclassifying the majority class and predicting them as belonging to the majority class.



1.2 Majority and Minority Recall for each model

As expected due to the imbalance all models (except Naiye Bayes) show a very high majority recall and a weak minority recall.

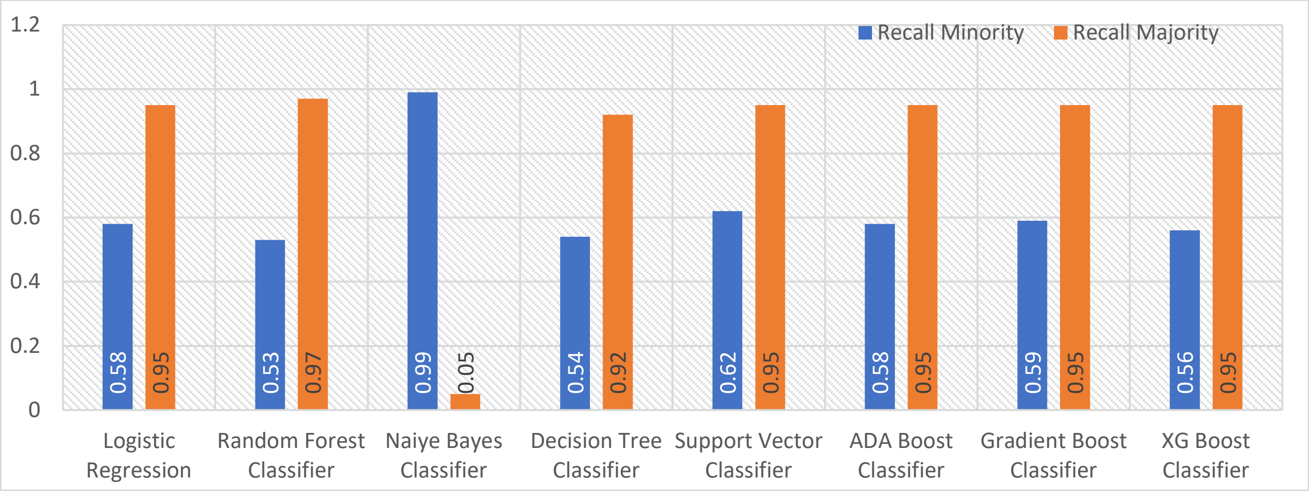


Table with model scores:

Description	Misclassifications	Accuracy	Precision	Recall Minority	Recall Majority	Type I errors	Type II errors	F1 Score	F2 Score
Logistic Regression	282	0.89	0.7	0.58	0.95	105	177	0.63	0.6
Random Forest Classifier	267	0.89	0.76	0.53	0.97	70	197	0.63	0.56
Naiye Bayes Classifier	1948	0.21	0.18	0.99	0.05	1945	3	0.3	0.52
Decision Tree Classifier	350	0.86	0.59	0.54	0.92	157	193	0.57	0.55
Support Vector Classifier	264	0.89	0.72	0.62	0.95	102	162	0.66	0.64
ADA Boost Classifier	283	0.89	0.7	0.58	0.95	106	177	0.63	0.6
Gradient Boost Classifier	269	0.89	0.72	0.59	0.95	95	174	0.65	0.61
XG Boost Classifier	290	0.88	0.69	0.56	0.95	105	185	0.62	0.58

4.1.4 Under Sampling Results

Comparing the best R_1 and F2 scores for each model after random under sampling the scores before sampling.

Chart 12.1 Best Recall after under sampling and before sampling

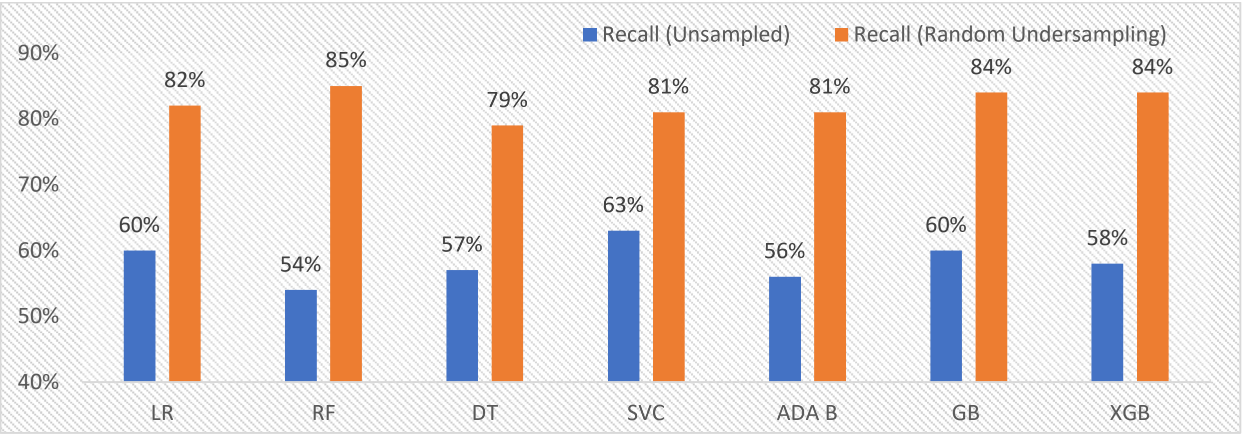
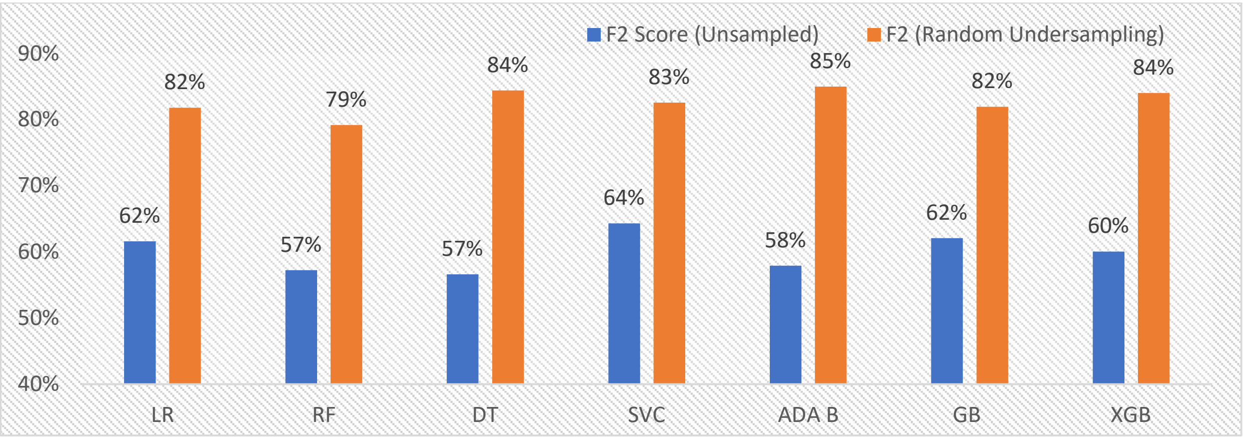


Chart 12.1 Best F2 after under sampling and before sampling



Most models show a significant improvement in the recall as well as the F2 score after random under sampling which shows the balancing the data has improved the model performance. Random Forest classifiers has shown the highest gain in the recall value followed by XGBoost Classifier.

Comparing each model’s cross validation score at different under sampling levels (0.2 to 1) .Since imbalanced models often lead to high accuracy as the model mis classifies most instances to the majority class it is understandable that accuracy begins to drop with the increase in under sampling ratio and as the data reaches an equal balance the accuracy is down by almost 6% for all models.

	Unsampled	Random Under sampling Levels						
	Accuracy	0.2	0.4	0.5	0.6	0.8	1	Difference
LogisticRegression	90%	89%	87%	86%	86%	85%	84%	6%
RandomForestClassifier	91%	90%	87%	87%	86%	86%	85%	6%
DecisionTreeClassifier	86%	86%	82%	81%	80%	80%	79%	7%
SVC	90%	90%	87%	86%	85%	85%	84%	6%
AdaBoostClassifier	89%	88%	85%	85%	84%	84%	83%	6%
GradientBoostingClassifier	91%	90%	87%	87%	87%	86%	85%	6%
XGBClassifier	90%	89%	86%	85%	85%	85%	84%	6%

Recall on the other hand shows considerable improvement with increase in under sampling ratio and the best recall being at under sample ratio of 1.

	Recall	0.2	0.4	0.5	0.6	0.8	1	Difference
LogisticRegression	60%	62%	76%	78%	79%	81%	82%	22%
RandomForestClassifier	54%	58%	74%	78%	80%	83%	85%	31%
DecisionTreeClassifier	57%	59%	68%	71%	74%	77%	79%	22%
SVC	63%	66%	79%	80%	80%	81%	81%	18%
AdaBoostClassifier	56%	58%	69%	73%	75%	79%	81%	25%
GradientBoostingClassifier	60%	63%	75%	78%	81%	83%	84%	24%
XGBClassifier	58%	61%	73%	76%	79%	83%	84%	26%

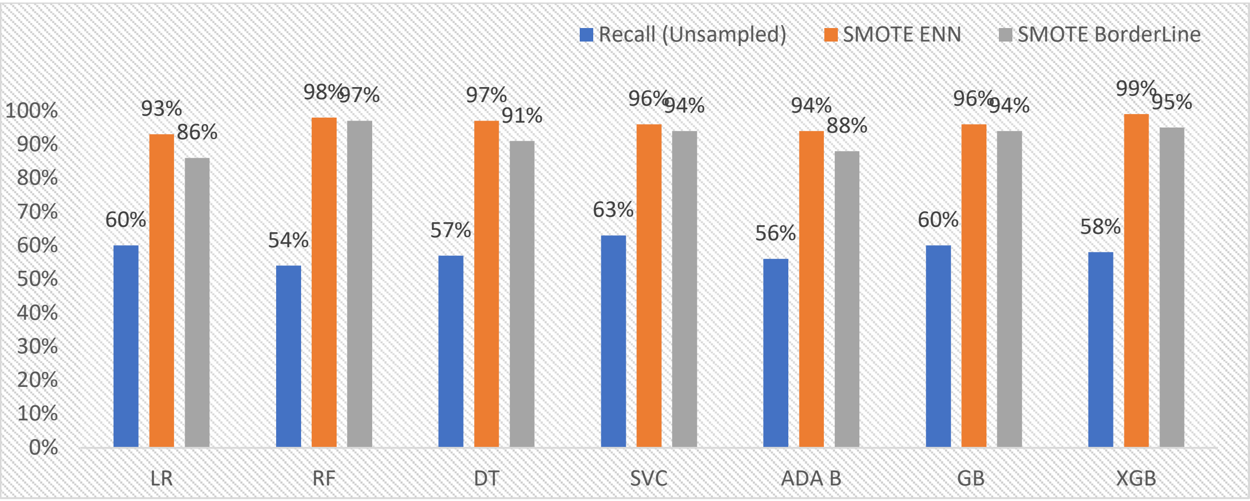
F2 scores exhibit a similar patter as the recall score .

	F2 Score	0.2	0.4	0.5	0.6	0.8	1	Difference
LogisticRegression	62%	63%	76%	78%	80%	82%	83%	21%
RandomForestClassifier	57%	61%	75%	79%	81%	83%	85%	28%
DecisionTreeClassifier	57%	59%	68%	71%	74%	77%	79%	23%
SVC	64%	67%	78%	80%	80%	82%	82%	18%
AdaBoostClassifier	58%	60%	71%	74%	76%	80%	82%	24%
GradientBoostingClassifier	62%	65%	76%	79%	81%	83%	84%	22%
XGBClassifier	60%	63%	74%	77%	79%	83%	84%	24%

Comparing the models above Random Forest Classifier is the best performing model which has shown considerable improvement in recall score while a similar drop in accuracy. Another approach possible would be try various under sampling levels which result in the best performance score while balancing the trade- off between accuracy and recall. An undersampling ratio between 0.7 to 0.8 might result in good recall while also balancing the accuracy of the classifier.

4.1.5 Over Sampling Results

Comparing Recall Values of the imbalanced data and to those of the two different SMOTE methods.



4.1.6 Conclusion & Hyperparameter Tuning

After conducting 85 experiments using hybrid sampling with nine undersampling and ten oversampling ratios, the project has achieved its objectives.

Overall, hybrid sampling has significantly increased the purchase intention detection rate. The best hybrid sampling technique is Random Undersampling (80%) and Standard SMOTE (80%) with Random Forest, yielding a Recall of 0.8521 for the majority class and 0.8564 for the minority class.

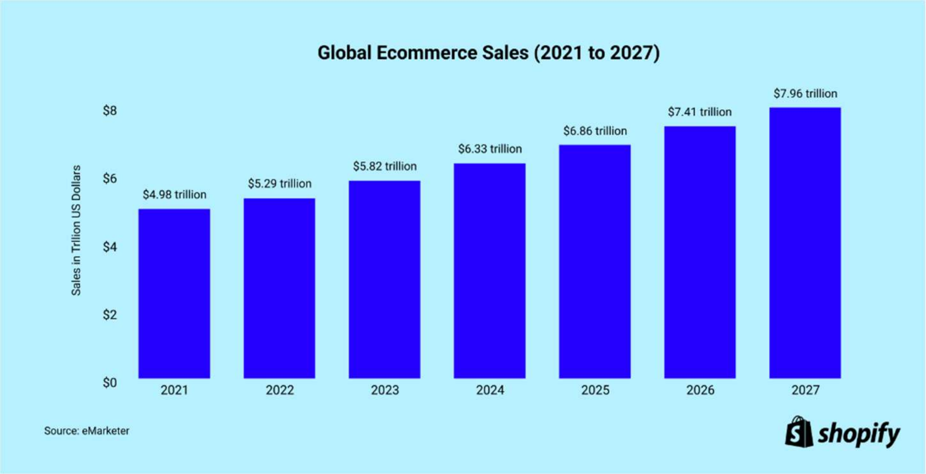
Random Forest functions well with all hybrid sampling techniques compared to the other classifiers,. Random Forest with the hybrid sampling technique Random Undersampling + Standard SMOTE produces the finest results.

5.2 Recommendations

Applying feature selection to future projects could increase accuracy and clarity while decreasing computational complexity. According to a study by Singh and Jain (2019), the true positive rate (TPR) can be substantially increased by employing feature selection techniques such as filter and wrapper. In the paper, except for the Random Forest classifier, applying a filter or wrapper enhances the TPR of J48, AdaBoost, Naive Bayes, and PART classifiers. Another suggestion would be to include algorithm fairness within the project's scope. Since this project demonstrates that certain classifiers are susceptible to bias towards a particular class, addressing algorithm fairness would aid in illuminating the factors influencing the detection rate. Hasanin and Khoshgoftaar (2018) stated that RUS often leads to losing important information as it randomly eliminates patterns of the majority class. Consequently, additional undersampling techniques can be incorporated into the experiments by exploring more undersampler options available in the research field. Koziarski (2021) proposed an undersampling technique, Synthetic Majority Undersampling Technique (SMUTE), which has proven a viable alternative to RUS.

Limitations

Citations



https://www.shopify.com/ca/blog/global-ecommerce-sales

[1] Sakar,C. and Kastro,Yomi. (2018). Online Shoppers Purchasing Intention Dataset. UCI Machine Learning Repository. https://doi.org/10.24432/C5F88Q.

[2] Baati K, Mohsil M. Real-Time Prediction of Online Shoppers' Purchasing Intention Using Random Forest. Artificial Intelligence Applications and Innovations.

2020 May 6;583:43–51. doi: 10.1007/978-3-030-49161-1_4. PMCID: PMC7256375.

[3] Shierly, L., & Sihombing, S. O. (2015, October). Predicting Online Purchase Intention: An Empirical Study. In Management Dynamics Conference (pp. 3-13).

[4] Jing Yang, Rathindra Sarathy, JinKyu Lee, The effect of product review balance and volume on online Shoppers' risk perception and purchase intention,

Decision Support Systems, Volume 89, 2016, Pages 66-76, ISSN 0167-9236,

https://doi.org/10.1016/j.dss.2016.06.009. (https://www.sciencedirect.com/science/article/pii/S0167923616301014)