Online Shoppers Purchasing Intention

Kasyap Rayalacheruvu
Dept. of Mathematics
Stevens Institute of Technology
Hoboken, NJ
krayalac@stevens.edu

***Abstract*—This project is entailed to observe the activity of the shoppers using their click-stream data to predict whether a customer's visit to the online website leads to a purchase or not depending whether the visit lead to a revenue generation.**

Keywords - Shoppers, Revenue, EDA, Machine Learning, Online Shopping

I. Introduction

A. Purpose of this task

Consumer consumption habits have changed dramatically as a result of rapid advancements in computer technology and e-commerce. This not only includes business to consumer shopping but also business to business shopping and besides using a virtual online website on the internet you also have m-commerce with the increase in app development in the present days due to the ever increasing number of smart devices connected over the internet. This has further increased the percentage of purchase done from foreign websites too. Shopping online has opened up a new realm in the world of business which has been expanding at a constant rate. People nowadays are more prone to opening up the various websites or apps to look up the new inventory or the required essentials they might be in need of rather than going up to the stores. The popularity of online purchasing is growing. An accurate analysis system of online purchase patterns allows online shopping platforms to gain a better knowledge of customer psychology and develop better business tactics to enhance sales.[1] At this time, it has become a great ordeal for the various sellers and the online platforms to delve deeper into the shopping habits and behavior of their customers to assess their needs. This pattern in the customers purchasing intention can be easily predicted by analyzing the history of the customers. By getting valuable insights from the shoppers behavior the businesses can be benefited as they can understand which factors were more detrimental in resulting in a purchase adding onto their overall revenue. Based on the Online Shoppers Purchasing Intention Data Set, which comprises of aggregated page view data throughout the visit session as well as other user information, this report focuses on factors that may impact visitors' purchase intention.

B. About Data set

We are using the data set "Online Shoppers Purchasing Intention"[2] from the UCI repository a collection of databases that are used by the machine learning community for the empirical analysis of the multitude of machine learning algorithms[3]. Feature vectors from 12,330 separate sessions are included in the data set we're utilizing. To eliminate bias in the data set due to a proclivity for a single campaign or event, limits were placed on the number of users from a certain special day, user profile, or time. It was created such that each session belonged to a different user throughout the year with regard to these limitations. 85 percent (10,422) of the 12,330 sessions in the data set were negative class samples, which did not complete their shopping and generated no income, while the rest (1908) were positive class samples, which completed their shopping and generated money.The data collection has ten numerical and eight categorical features. The revenue feature is the class label, which is either true or false depending on whether the consumer purchased something, resulting in revenue generation for the vendor.
The terms "Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration" refer to the variety of different types of pages that can be accessed by the viewer during that session, as well as the total spending time on each of these page sections. These characteristics' values are obtained from the URL information of the user's visited pages and updated in real time when the user performs an action, such as switching from one page to another. The statistics measured by "Google Analytics" for each page on the e-commerce site are represented by the "Bounce Rate," "Exit Rate," and "Page Value" features. The proportion of visitors that reach the site through that page and subsequently exit ("bounce") without making any more requests to the analytics server during that session is the value of the "Bounce Rate" feature for that page. For a certain web page, the value of the "Exit Rate" feature is determined as the percentage of all page views to the page that were the last in the session. The "Page Value" function displays the average value of a web page seen by a user prior to making an e-commerce purchase. The "Special Day" feature highlights how near the site visiting time is to a certain special day (for example, Mother's Day or Valentine's Day), when sessions are more likely to be completed with a purchase. The value of this feature is established by taking into account e-commerce factors such as the time taken between an order and its delivery. For example, lets consider the event of valentines day, between February 2 and February 12 this value takes a nonzero value, zero before and after this date unless it is near to another special day and a maximum value of 1 on February 8. The data set also includes information about the operating system, browser, area, traffic type, visitor type (return or new), a Boolean value indicating if the visit was on a weekend, and the month of the year.

C. Machine Learning Background

The problem is characterized as a Machine Learning problem, as stated in the introduction. However, in this article, we will explain what Machine Learning is and why it's a good strategy for predicting client purchase intent. Machine learning is a sort of computer program that improves itself over time by interacting with data and accumulating new information observed from the various patterns found in it. Machine learning algorithms create a model using a sample dataset, commonly known as training data. It then uses the sample dataset to develop a model that can make predictions or judgments without having to be explicitly programmed to do so. Traditional programming can be practically hard to design a suitable solution since there are so many rules and edge circumstances. Creating a program to determine if a picture is of a cat, for example, is difficult because to multiple obstacles in establishing the criteria, dealing with edge situations, and other limitations. Rather of having human programmers define every required step, it is more beneficial to let the computer design its own algorithm. This challenge can be solved using machine learning, since the algorithm creates models that employ statistical and mathematical reasoning to approximate the image it is looking at to the desired image. The more numerous photographs of a cat it has seen, the better it will be at recognizing a cat image in the future. This is comparable to how the human brain works; individuals utilize their memories or facts from previous experiences to make sense of current information. In this report, we'll aim to forecast consumer purchase intent, which is difficult to do with standard programming owing to the limits mentioned earlier. As a result, we'll employ Machine Learning methods to create a model or algorithm that can help forecast the outcome of a customer's purchase intent with as much precision as feasible. There are two major classes of Machine Learning algorithms to consider. These are supervised and unsupervised. In this section, we'll aim to find the best Machine Learning model for the problem we're trying to address in this report.

1) Supervised Learning: The machine learning job of learning a function that translates an input to an output based on sample input-output pairs is referred to as supervised learning. It infers a function from a set of training examples and tagged training data. The training data for a supervised learning system will consist of inputs that are coupled with the proper outputs. The algorithm will look for patterns in the data that correspond with the intended outputs during training. Following training, a supervised learning algorithm will take in fresh unknown inputs and, using earlier training data, determine which label the new inputs will be classed as. A supervised learning model's goal is to anticipate the proper label for freshly provided data. During training, the machine learning model creates the function that connects input characteristics to a predicted output. There are two subcategories of supervised learning.

- classification - Data points with a specified category will be provided to a classification algorithm during training. A classification algorithm's goal is to take an input value and, depending on the training data supplied, assign it to a class, or category, where it belongs. To generate the mapping function indicated earlier, the model will look for association between characteristics within the data and class.
- Regression - Regression is a statistical method that aims to uncover a significant link between dependent and independent variables. A regression algorithm's purpose is to forecast by finding the elements that influence the prediction's result. We have dependent variables and independent variables in a regression model. The primary elements that influence the result are known as dependent variables.

2) Unsupervised Learning: Unsupervised learning, also known as unsupervised machine learning, analyzes and clusters unlabeled information using machine learning techniques. Without the need for human interaction, these algorithms uncover hidden patterns or data groupings. Because of its capacity to find similarities and contrasts in data, it's an excellent choice for exploratory data analysis, cross-selling techniques, consumer segmentation, and picture identification. Principal component and cluster analysis are two of the most common unsupervised learning approaches. In unsupervised learning, cluster analysis is used to organize, or segment, datasets with common properties in order to deduce algorithmic links. To put it another way, clustering is concerned with determining a structure in a set of unlabeled data by identifying unique groups within the dataset. We need to create a closeness metric for two data points in order to cluster them. The term "proximity" refers to how similar or unlike the samples are to one another. A "decent proximity measure" is very reliant on the application.Clustering algorithm can be classified as the following

- Exclusive clustering - it involves grouping data in such a way that each data item belongs to just one distinct cluster. One of the exclusive clustering techniques is K-means clustering.
- Overlapping clustering - Overlapping clustering clusters data using fuzzy sets, allowing each point to belong to two or more clusters with varying degrees of membership.
- Hierarchical Clustering - There are two types of hierarchical clustering algorithms: agglomerative clustering and divisive clustering. The union of the two closest clusters is used in agglomerative clustering. Setting each data item as a cluster achieves the starting condition. It achieves the desired clusters after a few repetitions. Essentially, this is a bottom-up approach. Divisive clustering begins with a single cluster that contains all of the data components. Clusters are gradually broken into smaller clusters based on some dissimilarity at each stage. This is essentially a top-down variation.
- Probabilistic clustering - Example of probabilistic clustering A totally probabilistic strategy is used to create a Gaussian mixture. K-means, Fuzzy K-means, and Mixture of Gaussians are some of the most prevalent clustering techniques.

3) Choosing Between Supervised and Unsupervised Learning: As can be observed, the challenge is based on predicting the "Revenue" class label using labelled characteristics, and hence the task is classified as a Supervised Learning problem.

This explains why implementing a Supervised Learning ML algorithm rather than an Unsupervised Learning ML method is preferable and acceptable since in an unsupervised learning situation, the data is unlabeled and the goal is to uncover structure in the data. As previously said, supervised learning may be divided into two kinds. These are Regression and Classification. Depending on whether we want to express the revenue feature with a Boolean value or integer values, we may use either of them.

D. Using SMOTE Techniques to Overcome Class Imbalance

Class imbalance occurs when there is an unbalanced distribution of class in a dataset, i.e. the negative class (majority class) has a greater number of data points relative to the positive class (minority class).[4] In general, we are more interested in the minority/positive class, and we strive to get the greatest outcomes in this class. The performance of the classifier model will be harmed if the unbalanced data is not addressed beforehand. The bulk of the predictions will be for the majority class, and the minority class characteristics will be treated as noise in the data and ignored. As a result, the model will have a significant bias.

Clearly, the accuracy metric is biased and not ideal in such circumstances where class distribution is skewed. As a result, we must rely on superior performance metrics, such as the F1-score, which is the harmonic mean of accuracy and recall, because the recall in our situation will undoubtedly be poor owing to the skewed dataset, which we must aim to improve.

One of the most typical techniques of dealing with an unbalanced dataset is to resample the data. Undersampling and oversampling are the two most common ways for this. Oversampling techniques are preferable over undersampling approaches in most circumstances. The reason for this is because when we undersample data, we tend to exclude occurrences that may contain crucial information. In this study, we'll go over several unique data augmentation oversampling approaches, such as SMOTE and its associated variants.

1) SMOTE: Synthetic Minority Oversampling Technique: SMOTE is an oversampling approach in which synthetic samples for the minority class are created. This approach aids in overcoming the problem of overfitting caused by random oversampling. It concentrates on the feature space in order to produce new examples by interpolating between positive instances that are close together.

The total number of oversampling observations, N, is put up initially. It is usually chosen such that the binary class distribution is 1:1. However, depending on the situation, this might be reduced. The iteration then begins with a random selection of a positive class instance. The KNNs for that instance are then retrieved. Finally, N of these K instances are chosen as the basis for creating new synthetic instances. To do so, the difference in distance between the feature vector and its neighbors is determined using any distance metric. This difference is now multiplied by any random number in the range (0,1) and added to the preceding feature vector[4].

Though this method is quite useful, it does have a few flaws.

- The synthetic instances are formed in the same direction, i.e. their diagonal instances are connected by an artificial line. As a result, the decision surface created by a few classifier algorithms becomes more complicated.
- In feature space, SMOTE tends to generate a significant number of noisy data points.

2) ADASYN: Adaptive Synthetic Sampling Approach: The ADASYN algorithm is a more generalized version of the SMOTE algorithm. By producing synthetic examples for the minority class, this technique also seeks to oversample it. However, it takes into account the density distribution, r_i, which determines the number of synthetic instances created for difficult-to-learn data. As a result, it aids in adaptively adjusting judgment limits based on difficult-to-learn data. This is the most significant distinction from SMOTE.

From the dataset, the total no. of majority N^- and minority N^+ are captured respectively. Then we preset the threshold value, d^th for the maximum degree of class imbalance. Total no. of synthetic samples to be generated, G = (N^- – N^+) x β. Here, β = (N^+/ N^-). For every minority sample x_i, KNN's are obtained using Euclidean distance, and ratio r_i is calculated as Δi/k and further normalized as r_x ¡= r_i / r . Thereafter, the total synthetic samples for each x_i will be, g_i = r_x x G. Now we iterate from 1 to g_i to generate samples the same way as we did in SMOTE[4].

3) SMOTE + Tomek Links: Combining undersampling and oversampling approaches is required. This is done in order to improve the performance of classifier models for samples generated using these strategies. SMOTE+TOMEK is a hybrid approach that tries to clear overlapping data points in sample space for each of the classes. The class clusters may be invading each other's space after SMOTE's oversampling. As a result, the model of the classifier will be overfit. Now, Tomek linkages are paired samples of the opposite class that are the nearest neighbors to each other. As a result, the bulk of class observations from these linkages have been eliminated since it is thought that this will enhance class separation around the decision borders. Tomek linkages are now applied to oversampled minority class samples created by SMOTE in order to obtain better class clusters.

4) SMOTE + ENN Links: Another hybrid strategy is SMOTE + ENN, which removes a larger number of observations from the sample space. ENN is yet another undersampling strategy in which the majority class's nearest neighbors are approximated. If the nearest neighbors incorrectly label that specific instance of the majority class, it is eliminated. Integrating this approach with SMOTE's oversampled data aids in significant data cleaning. Samples from both groups are excluded due to NN's misclassification. As a consequence, the class distinction is more apparent and straightforward.

II. Related Work

A data-level approach and feature selection approaches are proposed as a solution for the classifying of unbalanced data in [5]. One of the basic challenges in artificial intelligence, notably for classification in machine learning, is imbalance class categorization. Imbalanced data has been shown to degrade the performance of machine learning algorithms, where

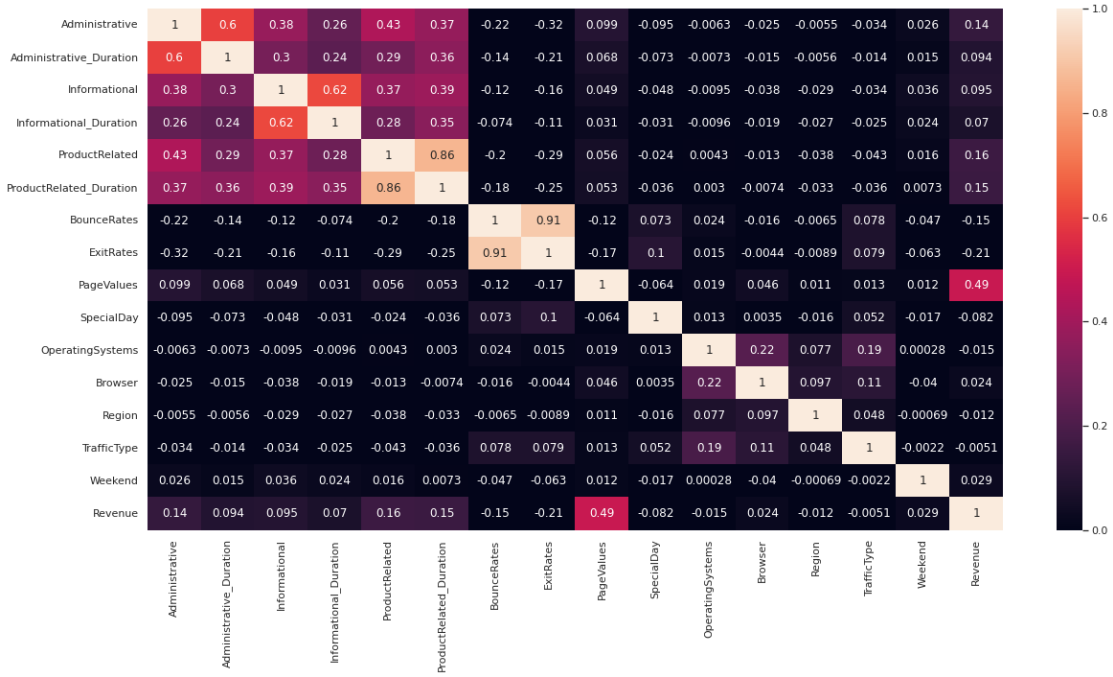

Fig. 1. Correlation Heatmap

imbalance data refers to the fact that the total data from each class differs substantially. The suggested approach is tested using a dataset from the UCI repository and the major testing metric is the area under the curve (AUC). In many cases, especially when the target group is a minority, the false positive rate (FPR) is set at an acceptable level, and the goal is to reduce the false negative rate (FNR) while keeping the FPR below a certain level. Various techniques have been offered to overcome this problem, in addition to tampering with the performance statistic. Algorithmic level techniques, data level techniques, ensemble classifications, and cost-sensitive techniques are the four categories of methods (a mixture of algorithmic and data level techniques).In order to overcome the problem of skewed class distribution in the learning phase, data preparation techniques adjust the data distribution. The algorithmic modification techniques change current algorithms to make minorities more significant. To deliver varying misclassification costs for each class in the learning process, cost-sensitive techniques use both algorithm and data modification approaches. Finally, to address the imbalance problem, ensemble of classifier sampling methods modify the ensemble learning algorithm, but they usually do not change the base classifier [6]. [5] claims that combining data preparation approaches with an ensemble classifier outperforms other methods. Before the model training step, the preprocessing data approach involves resampling uneven training data sets. Although no one strategy has been shown to perform effectively for all unbalanced data set situations, sampling methods have shown tremendous promise in that they aim to enhance the data set rather than the classifier. By either oversampling the minority samples or under sampling the majority samples, sampling procedures alter the distribution of each class observation. In the case of oversampling, sampling procedures create additional minority instances to balance the data set, but in the case of under sampling, sampling methods eliminate some majority instances. Because the removal of majority instances may remove essential information from the data set, under sampling approaches have been found to be less efficient than oversampling methods, especially when the data set is small. Random sampling is the most basic way of oversampling. It chooses a minority instance at random and replicates it until the minority class has grown to the appropriate size. Over-fitting occurs when random oversampling creates new instances that are highly close to the original instances.To solve this problem, the Synthetic Minority Oversampling Technique (SMOTE) is utilized, in which new synthetic instances are created by generating new synthetic instances from randomly picked minority instances and their N-nearest neighbors, where N is a user-defined variable. However, because new instances are created without taking into account the majority instances, this may result in over-generalization, increasing the overlap between minority and majority classes.When the data set has a higher imbalance ratio, over generalization can be magnified because the minority class instances are very limited and can become contained within the majority class after oversampling. This can make future classification performance much worse.

III. Our Solution

A. Descriptive Analysis of Data set

The data set is highly skewed with approximately 85 percent negatively classified samples.There are no missing values or NAN values in the data set.In this data set we have 10 numerical variables and 8 categorical variables along with 2 of them having Boolean data type.The target variable 'Revenue' is also Boolean.In pre-processing, One hot encoding is performed on the categorical variables 'Month' and 'Visitor Type'.Most of the features don't have a lot of distinct values.Using the count-plot function we got to know the

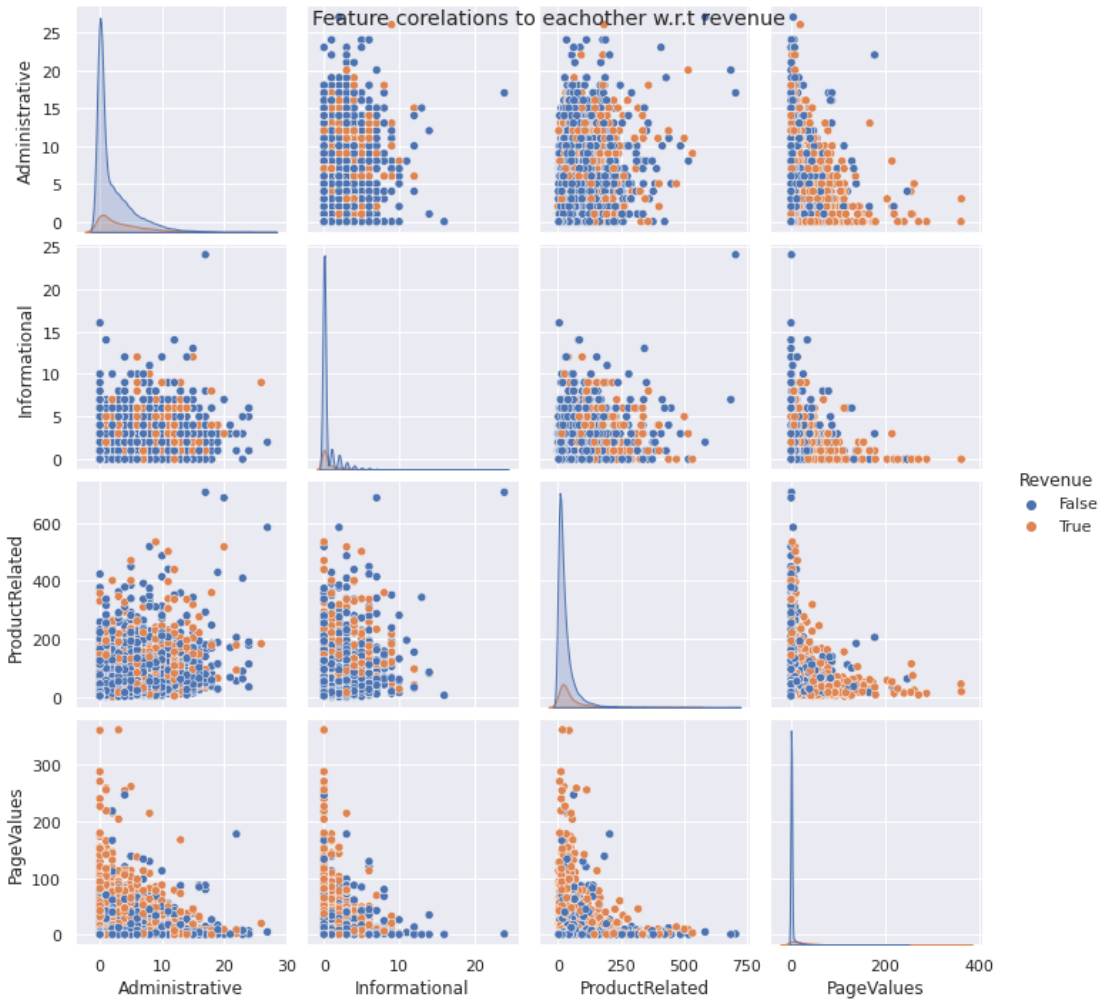

Fig. 2. Relation between the highly correlated feature w.r.t revenue

distribution of values of each features.In Fig.1 the correlation heat map is visualized to check the correlation among the features and the highly correlated features were then used in the pair plots to check the spread among the features in Fig.2 .The figures Fig 3,Fig 4 and Fig 5 visualize the counts of the features Month and Special day and the target variable Revenue respectively.We split the whole data set into training set and test set with a test size of 20 percent.

From the above heatmap as shown in fig 1, we observe the following that there is a limited number of features that correlate to eachother. The very few pair of features that have a high correlation i.e correlation greater or equal to 0.75 are : BounceRates and ExitRates (0.91), ProductRelated and ProductRelated Duration (0.86). features that have a mild correlation i.e between 0.35 and 0.75 are : Administrative and Administrative Duration (0.6), Administrative and Informational (0.38), Administrative and ProductRelated (0.43), Administrative and ProductRelated Duration (0.37), Administrative Duration and ProductRelated Duration (0.36), Informational and Informational Duration (0.62), Informational and ProductRelated (0.37), Informational and ProductRelated Duration (0.39), Informational Duration and ProductRelated Duration (0.35), PageValues and Revenue (0.49). From the plots in fig 2 we can see that a linear decision boundary will not be possible to segregate the data based on revenue using any two pair of the highly and moderately correlated pairs we had found earlier, hence this rules out the possibility of linear regression model form performing well. from fig 3 we can see January and April has no visitors at all. Maybe keeping some sort of sale during these months might increase visitors and hopefully the revenue generation. March, May and December has a lot of shoppers with may having the maximum number of visits amongst all the months but a very few of these visitors are contributing to revenue generation. Most revenue generation occur during November. From fig 4 The closer the visit date is to a special day (like valentines day, new year's eve etc) visitor do not purchase anything but they do on the special day itself. Fig 5 shows that the data set is highly unbalanced which will lead to issue regarding the generalisation and needs to be dealt with some data augmentation or sampling techniques.

B. Machine Learning Algorithms

Based on the data analysis that was done previously we can implement any classification algorithm to classify the data set.The algorithms which might be suitable for this

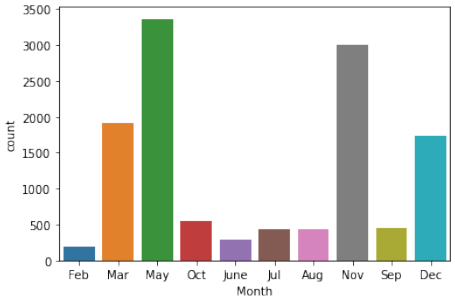

Fig. 3. Month feature

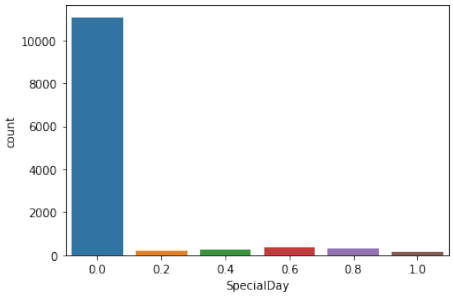

Fig. 4. Special Day feature

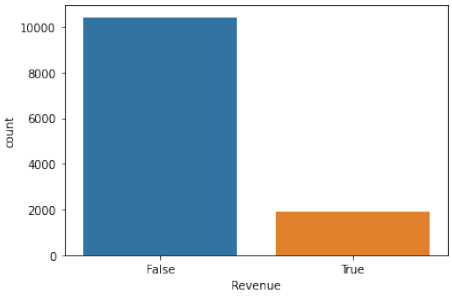

Fig. 5. Revenue

type of data set are Logistic Regression,Random Forest Classifier,Multilayer Perceptron,k-NN classifer,Adaptive boosting and Support Vector Classifier.We impemented these algorithms initially to check how they perform on the dataset and after that we applied oversampling using SMOTE and again trained the model on the oversampled data.

C. Implementation Details

In this paper, we are using a real- time shoppers based intention prediction system which basically predicts the intention of shoppers as soon as they visit a particular website.To do this, we are implementing 6 Machine Learning models namely Logistic Regression, K nearest neighbors, Random Forest, Support Vector Classification, Multilayer Perceptron and Adaptive Boosting. We can see from our results that Random Forest produces better accuracy and F-1 score compared to the other models.

Classification Algorithms -

1) Logistic Regression : This method is used to predict a dependent variable, given a set of independent variables, such that the dependent variable is categorical.This is achieved through a Logistic function,which has the shape of a Sigmoid curve. The dependent variable can hold values like 0 or 1,Yes or No, A,B or C. This method is very fast regarding prediction and training times, thus it is one of the most popular machine learning algorithms for binary classification.

2) Random Forest Classifier : A random forest classifier is basically a classification algorithm made up of several decision trees. One biggest advantage of Random forest classifier is it can be used for both classification and regression problems. This algorithm is more reliable than decision tree as each tree neutralizes the error of other trees[7].

3) Support Vector Classification : SVM is another machine learning technique which can be used for both classification and regression problems.The main function of SVM is to check for the hyperplane that best seperates the data into two different classes[7]. The best hyperplane is the one that has the maximum distance between the two classes.It is usually preferred over other algorithms because of less computational time and good accuracy.

4) K-Nearest Neighbor : KNN classifier is generally used for classification problems in Machine learning.It classifies a new data point based on similarity between the new case and the available case. It stores all the available data cases and classifies new data case based on similarity.Thus when new data appears, it can be easily classified into a suitable category by knn algorithm.

5) Multilayer Perceptron : Multilayer perceptron is a field of artificial neural network with multiple layers(3 or more).It consists of an input, an output and a hidden layer. Each neuron has it own activation function.This field investigates how simple models of brain can be used to solve predictive modeling tasks.Its predictive capability comes from hierarchical structure of the networks.

6) Adaptive Boosting : AdaBoost,best used with weak learners, can be used to boost the performance of any machine learning algorithm.Decision trees are the most common algorithms used with adaboost. They are often called decision stumps because they are very short and contain only one decision for classification.

IV. Comparison

So far based on the unbalanced data set(as the model have more negative samples), all the 6 models we have implemented are having less recall values as it wasn't able to learn the necessary relationships to avoid misclassifying revenue generating instances. The recall values for the models were well below 0.7. The Random forest model has given us the best precision of 0.77 so far as its a ensemble model making it easier to deal with the under-generalisation issue caused in the other models as a result of the skewed data set. The logistic regression model gives a decent precision of 0.73 but a lesser recall value of 0.54 as compared to Random Forest.In Multi layer perceptron the precision is high when compared to both previous models but the recall is too low at 0.31. We focus on the precision for now as we know the accuracy of our model will be less due to bad recall rates caused by not a proper balance between the revenue classes especially the the data set having more instances producing no revenue.

So, after oversampling the training data set we trained the same models on that. There was a significant increase in the recall.Even here we can see that Random forest classifier is the best model with a recall of 0.7 and F1-score 0.69 which there is not much improvement but the recall is higher which give the idea of representatives of True Revenue.In logistic regression though the recall is 0.74 and the F1-score is 0.64.There is a significant improvement in both recall and F1-score but falls behind Random Forest classifier.Even in Multi layer perceptron

there is some significant improvement but not enough focus to select it as the best model.The recall and F1-score of MLP are 0.44 and 0.54 respectively.

Comparison of Different Models before oversampling

Model	Precision	Recall	F1-Score
Random Forest	**0.75**	**0.52**	**0.61**
Logistic Regression	0.71	0.35	0.47
Multi Layer Perceptron	0.50	0.76	0.60
K-NN Classifier	0.78	0.19	0.31
Adaptive Boost	0.69	0.54	0.61
Support vector Classifier	0.83	0.01	0.02

Comparison of Different Models after oversampling using SMOTE

Model	Precision	Recall	F1-Score
Random Forest	**0.69**	**0.67**	**0.68**
Logistic Regression	0.62	0.72	0.67
Multi Layer Perceptron	0.66	0.60	0.63
K-NN Classifier	0.40	0.58	0.47
Adaptive Boost	0.62	0.65	0.64
Support vector Classifier	0.35	0.74	0.48

Comparison of Different Models after oversampling using ADASYN

Model	Precision	Recall	F1-Score
Random Forest	**0.68**	**0.66**	**0.67**
Logistic Regression	0.60	0.72	0.65
Multi Layer Perceptron	0.76	0.32	0.45
K-NN Classifier	0.37	0.61	0.46
Adaptive Boost	0.62	0.67	0.64
Support vector Classifier	0.32	0.75	0.45

Comparison of Different Models after oversampling using SMOTE + Tomek

Model	Precision	Recall	F1-Score
Random Forest	**0.70**	**0.66**	**0.68**
Logistic Regression	0.60	0.71	0.65
Multi Layer Perceptron	0.42	0.78	0.54
K-NN Classifier	0.39	0.55	0.46
Adaptive Boost	0.60	0.72	0.65
Support vector Classifier	0.36	0.73	0.48

Comparison of Different Models after oversampling using SMOTE + ENN

Model	Precision	Recall	F1-Score
Random Forest	**0.59**	**0.80**	**0.68**
Logistic Regression	0.56	0.80	0.66
Multi Layer Perceptron	0.49	0.79	0.61
K-NN Classifier	0.35	0.69	0.46
Adaptive Boost	0.59	0.76	0.67
Support vector Classifier	0.30	0.82	0.44

V. Conclusion

The main idea of the proposed work is related to the need to create a model that predicts the purchase intent of the visitor after visiting an e-commerce website. In this paper,the performance of the different supervised learning algorithms is studied based on the data of certain online shoppers.To identify this, 6 classification techniques have been implemented namely Logistic Regression,Random Forest, Support Vector Classifier,K-Nearest Neighbor,Multilayer Perceptron and Adaptive Boosting.Moreover, we have implemented SMOTE (Synthetic Minority Oversampling)technique to improve the performance and scalability of each classifier. Based on our evaluation, we have found out Random Forest as the most appropriate classifier in terms of different evaluation metrics.

We can implement Grid Search and Randomized Grid Search for getting the best hyperparameters for this data to increase the accuarcy and F1- score.There are many parameters in the model that we implemented which may have a significant role in finding the best model.It requires heavy computational time and power.There are also models such as XGBoost which gives us a better accuracy and F1-score. Apart from this,we can always use dimensionality reduction such as Principal Component Analysis where we try to reduce the number of features based on the importance.Without compromising on the variance in the dataset, we can combine different features into one to reduce the computational time.

References

[1] M. D. Houda Zarrad, "Online purchasing intention: Factors and effects," vol. 4, no. 1, pp. 37–47, January 2012. [Online]. Available: https://core.ac.uk/reader/236300560

[2] M. K. . Y. K. C. Okan Sakar, S. Olcay Polat, "Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and lstm recurrent neural networks." *Neural Computing and Applications*.

[3] D. Dua and C. Graff, "UCI machine learning repository."

[4] S. Satpathy, "Overcoming class imbalance using smote techniques," *Neural Computing and Applications*, October 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/#h2_3

[5] I. Kurniawan, A. Abdussomad, M. Akbar, D. Saepudin, M. Azis, and M. Tabrani, "Improving the effectiveness of classification using the data level approach and feature selection techniques in online shoppers purchasing intention prediction," *Journal of Physics: Conference Series*, vol. 1641, p. 012083, 11 2020.

[6] I. Nekooeimehr, "Oversampling methods for imbalanced dataset classification and their application to gynecological disorder diagnosis," 2016.

[7] R. A. Md Rahyan Kabir, Faisal Ashraf, "Analysis of different predicting model for online shoppers' purchase intention from empirical data," December 2019. [Online]. Available: https://www.researchgate.net/publication/340058413_Analysis_of_Different_Predicting_Model_for_Online_Shoppers'_Purchase_Intention_from_Empirical_Data