Prediction of Purchasing Intention of Online Shoppers on a Website

Satwik Murarka
Dept. of Chemical Engineering
Indian Institute of Technology Bombay
Mumbai, Maharashtra
190020101@iitb.ac.in

Tanmay Choudhary
Dept. of Chemical Engineering
Indian Institute of Technology Bombay
Mumbai, Maharashtra
190020121@iitb.ac.in

Rhythm Shah
Dept. of Metallurgical Engg. & Materials Sci.
Indian Institute of Technology Bombay
Mumbai, Maharashtra
19011074@iitb.ac.in

Abhi Manjunath
Dept. of Chemical Engineering
Indian Institute of Technology Bombay
Mumbai, Maharashtra
190020036@iitb.ac.in

***Abstract*—In the current age, e-commerce websites have become a major source of sales for businesses and industries. They face several challenges due to intense competition, non-fulfillment of expectation of customers, logistical issues. To increase these sales, it is important for a business to understand the customer behaviour and factors which affect the decision of purchase. Theses would help them to improve their websites and create personalized plans for customers. In the following paper, we have analyzed the user activity of a website over a period of one year to gain insights over purchase patterns and compared various Machine Learning Models namely Random Forests, Support Vector Machines, Logistic Regression and Neural Networks, to predict the outcome of a session.**

I. Introduction

As the name suggests, E-commerce or electronic commerce is the buying and selling of goods,services as well as transfer of funds via an electronic medium, mostly the internet. It has made the life of customers much easier as they can purchase goods and services from the comfort of their homes in a matter of few clicks. With the coming digital age, e-commerce is taking a major role and has already become a major source of sales for many businesses and industries. In India, the market size of the e-commerce industry has grown exponentially from 14 billion USD in 2014 to 84 billion currently and it expected to become 200 billion by 2027. E-commerce provides people with an online store where they can browse products and place orders on their device itself. It works by communication between the customers web browser and the server hoisting the online website. Data regarding the order is then transferred to a central system where it checks with databases to validate the order. Once validated the order is successfully confirmed and the customer receives a confirmation on their browser.The data of order is then sent to the warehouse from where the product gets dispatched.

Every method has its own advantages, disadvantages and challenges associated with it and so is the case with E-commerce.The advantages of E-commerce are 24×7 availability as customers can shop without any time restrictions, easy availability and accessibility, access to international products and a personalized experience with product recommendations. Although it has numerous advantages, it also certainly has its downsides. No physical touch of the product, limited customer service and security concerns are some of the disadvantages of E-commerce websites. These also face certain challenges to remain profitable and competitive. These are capitalization of data which is the ability to use data generated to facilitate business decisions for higher profits , high customer expectations, personalization and many more. Capitalization of data is a challenge due to large amounts of data being generated by these sites. One has to choose the right metrics and methods to gain benefits out of the it. This is critical since a bad analysis can severely affect the revenue. In our paper we tackle the first challenge and use the data generated from user sessions over the period of one year to answer some interesting questions like the factors on which a purchase depends upon and then compare machine learning models to predict the outcome of session. We do this by doing exploratory analysis, hypothesis testing and model development.

II. Background & Prior Work

The challenges associated with the predicting the given problem involve large amount of data which needs to be modelled using right metrics for good performance. Certain papers have been published discussing various algorithms and techniques for these. Some of these are:

- C. Okan Sakar, S.Olcay Polat, Mete Katircioglu and Yomi Kastro first tackled the problem by publishing a paper to propose a prediction online shopper behaviour analysis system which purchasing intention and abandonment chances of a customer. They used Random Forests, Support Vector Machines and MultiLayer Perceptrons for the task. They also used feature selection to improve the scalability and performance of the model. They did

not applied exploratory data analysis techniques to the dataset. The final/best accuracy and F1 score achieved were **0.87** and **87.94%** respectively.
- Karim Baati and Mouad Mohsil suggested a real-time prediction system using Naive Bayes Classifier, C4.5 Decision Tree and Random Forests. Further they used oversampling to improve performance and scalability and used F1 score and accuracy to compare performance.

III. Dataset & Methodology

The dataset consists of **12330** sessions with each session belonging to a different user in a 1 year period to avoid any bias which could be caused by specific special days, campaigns or events. Out of the 12330 sessions, 10422 (84.5%) were negative class samples which did not end up in shopping or no revenue generated while the rest 1908 (15.5%) were positive class samples which ended up in shopping. The dataset consists of 18 features or variables out of which "Revenue" is used as class label. In this, "True" means a successful purchase whereas "False" means no purchase. There are 17 attributes or features of which 10 are numerical and 7 are categorical. The features consists of Administrative,Informational and Product Related which tells the number of pages visited by the user of the particular category. Further, there are features which give the duration spent by the users on each of the sites. Another category of features includes the Google Analytics metrics namely the Bounce Rate, Exit Rate and Page Values. The Bounce rate refers to the percentage of visitors who enter the site and leave without triggering any other action. Page Value represents the average number of pages the user visited before completing the transaction. Exit Rate for a webpage is the percentage that were last in the session.

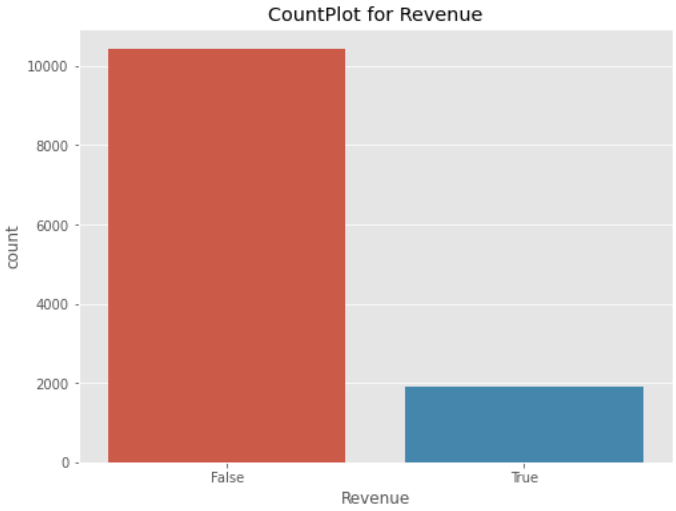

Fig. 1. Frequency of categories for Revenue class label

A through exploratory data analysis was done followed by hypothesis testing and Principal Component Analysis. Due to this the number of features were successfully reduced from 29 to 14. The dataset did not contain any missing values and hence no data imputation was required. As the original dataset is imbalanced, as seen in Fig 1, the models were trained twice, once for the original sample (D1,D2) and once on a balanced dataset (D3,D4) with equal number of class labels for both the categories. In D1 we do not use PCA while in D2 we use PCA. A similar approach is followed for datasets D3 and D4.

IV. Experiments & Models

A. Data Analysis and Interpretation

In data analysis we do: Exploratory Data Analysis, Hypothesis Testing and Principal Component Analysis.

Exploratory Data Analysis: We provide an overview of our data, take care of missing variables ad separate the features in numerical and categorical. We also look at the distributions of each variable and find the relative composition of each categorical variable. This was further divides into:

- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis

Hypothesis Testing: We use hypothesis testing to find the correlation between the target variable and the categorical columns. This is done using a test known as chi-squared contingency test.

Principal Component analysis: Our data originally had 18 dimensions, which were increased to 29 dimensions after one-hot encoding the categorical variables. Due to such a high dimensionality, observations become harder to cluster and more difficult to process. PCA is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. By utilising PCA, we were able to bring down the number of features from 29 to 14 (as mentioned earlier).

B. Computational Environment

In this project we have used four binary classification models as implemented in the `scikit-learn` library in `Python`:

- Logistic Classifier
- Neural Network Classifier
- Random Forests Classifier
- Support Vector Machine (SVM) Classifier

The entire dataset was normalised before proceeding with these models and the categorical variables were one-hot encoded to from numeric variables.

C. Machine Learning Models

The results obtained from the **Logistic Classifier** have been used as a benchmark for more complex models, since Logistic Regression is probably the most commonly used supervised learning classification method. The *saga* solver,is an extension of the stochastic average gradient descent approach, has been used here, which takes a random sample of prior gradient values and allows for L1 regularisation. To increase the projected performance of the machine learning model and

lower the standard error associated with the results, a 4-fold cross-validation approach was adopted.

We used a multilayer perceptron architecture with a single hidden layer (H) to create the **Neural Network Classifier**. This H is a hyperparameter that determines the architecture's complexity. We employed the *Adam* optimising method, which combines the benefits of RMSPRop and AdaGrad, two modifications of the standard stochastic gradient descent approach. The most significant advantage of Adam over other conventional optimization techniques is that it converges faster by avoiding the oscillation problem that other optimization techniques face during gradient descent, and it does so by using adaptive learning rate and "momentum", a weighted average of previous weight updates, while updating during backpropagation.

The **Random Forest Classifier** is made up of a group of decision trees (DT), each of which is a branching structure that represents a set of rules for categorising information in a hierarchical manner. The hyperparameters that were chosen after executing a grid search on the training set for each ensemble were the depth of each DT and the number of such DTs to be employed in each ensemble.

The **Support Vector Machine** Classifier, which employs a Kernel technique, was also employed in the hopes of further improving the F1 Score. The SVM uses the kernel trick to transform the input into a higher-dimensional vector space and tries to find the maximum margin decision boundary that separates the positive and negative examples. The Regularization Parameter (C) and the nature of the Kernel Function were the hyperparameters tuned for the SVM. Due to the enormous dimension of the parameters following the Kernel transform, SVMs are prone to overfitting and do not directly provide probability estimates; these are determined using a costly five-fold cross-validation procedure (in the `scikit-learn` library).

D. Performance Metrics

Precision and **recall** are two important performance metrics which would be used to evaluate our model. The precision of the model will give the people who would buy a product from the website, divided by the number of people the model has considered as positive. Recall is given by the ratio of true positives by the sum of true positives and false negatives.

$$Precision\,(P) = \frac{TP}{TP + FP} \tag{1}$$

$$Recall\,(R) = \frac{TP}{TP + FN} \tag{2}$$

$$F1\ Score = \frac{2PR}{P + R} \tag{3}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Where TP is the number of True Positives, TN is the number of True Negatives, FN is number of false negatives and FP is the number of false positives. To account for both the metrics as a single metric, we use F1 score which is the harmonic mean of precision and recall.

Our model has been trained once for a unbalanced and a balanced dataset. For the unbalanced dataset we cannot use accuracy as a metric due to the severe class imbalance and hence F1 score is used as an evaluation metric. Accuracy is not suitable for these cases. Suppose we have a dataset with 90% majority class and 10% minority class. To get an accuracy of 90% we can simply label all the examples to the majority class. This is not good as it is not able to classify the minority class correctly. This is a goal of our model as well and hence we do not use accuracy for the imbalanced case.

The balanced dataset uses accuracy and F1 Score as the model evaluation metrics.

Hence, the goal of the models is to maximize the F1 score for both the cases and accuracy for the balanced case.

V. Evaluation & Results

A. Exploratory Data Analysis

The aim of our exploratory data analysis was to find patterns in data, distributions of each variables and the relative composition of it. It was found that the dataset did not contain any missing values and hence no data imputation was required. We also derived certain relations between the features. It was found that around **85%** of the visitors were returning which meant that the site has a high retention rate. Also the percentage of purchase was slightly higher for the new visitors as compared to the returning visitors. So it would be beneficial to target new customers through promotional events and targeted campaigns. We also find that the bounce rate, exit rate are negatively correlated while the page value is positively correlated with the revenue. It was also found that the dataset consisted of a significant number of outliers and hence they could not be neglected for better performance. Some other interesting facts derived from EDA were:

- Traffic Type 2 is the most common from of traffic
- Majority of the visitors belong to region 1 and 3
- 80% of the users do not visit any informational sites
- The site witnesses the most number of visitors in the month of May
- 95% of the users use Operating Systems 1,2 or 4

B. Hypothesis Testing and Principal Component Analysis:

We used hypothesis testing to statistically verify the categorical features which affected the revenue. It was found that for the p-value for *Region* was **0.32** and hence the *Revenue* does not depend upon it. Upon applying PCA, it was found that the first 18 principal components explain 90% of the variance in the data.

C. Machine Learning Model Implementation

The hyperparameters obtained for the 4 cases of the 4 models are given as follows:

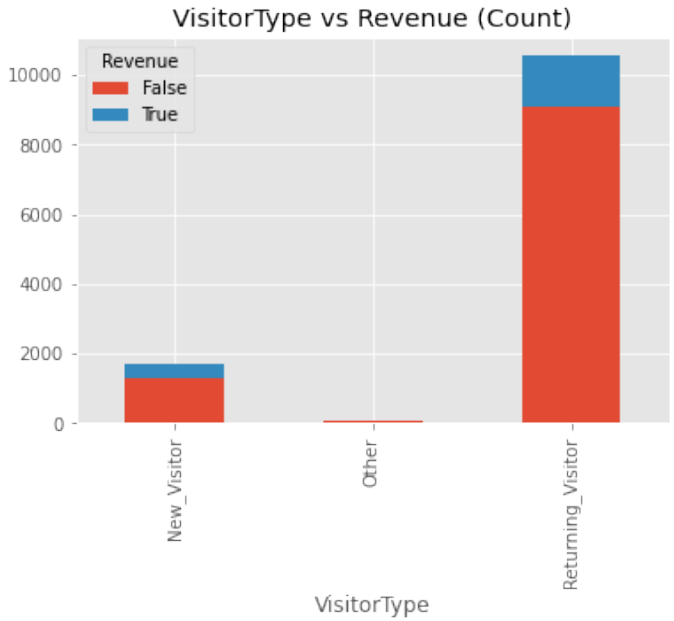

Fig. 2. Frequency plot based on type of visitor

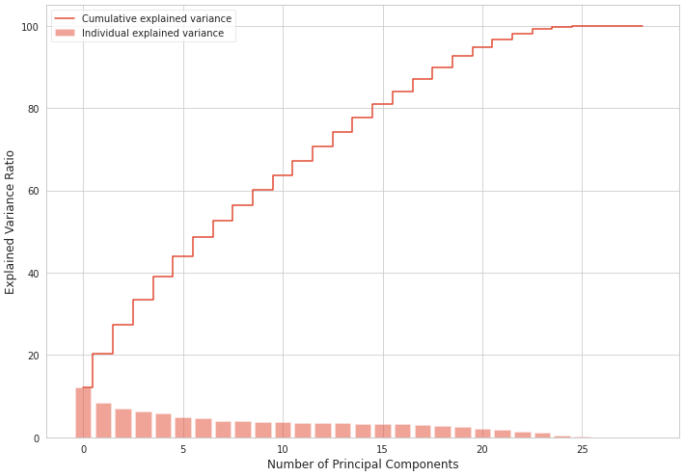

Fig. 3. Explained Variance Ratio vs Number of Principal Components

1) Hyperparameter Tuning: For **original dataset D1 with no PCA**:

- Logistic Regression: C = 5
- Neural Network: Hidden Layer Size = (75, 75) & Learning Rate (α) = 0.5
- Random Forest Classifier: Number of Trees = 100 , Maximum Depth of Each Tree = 23
- Linear SVM : C = 10

For **original dataset D2 with PCA**:

- Logistic Regression: C = 1
- Neural Network: Hidden Layer Size = (100,) & Learning Rate (α) = 0.1
- Random Forest Classifier: Number of Trees = 125 , Maximum Depth of Each Tree = 12
- Linear SVM : C = 10

For **oversampled dataset D3 with no PCA**:

- Logistic Regression: C = 10
- Neural Network: Hidden Layer Size = (75, 75) & Learning Rate (α) = 0.05
- Random Forest Classifier: Number of Trees = 50 , Maximum Depth of Each Tree = 25
- Linear SVM : C = 5

For **oversampled dataset D4 with PCA**:

- Logistic Regression: C = 10
- Neural Network: Hidden Layer Size = (75, 75) & Learning Rate (α) = 0.05
- Random Forest Classifier: Number of Trees = 50 , Maximum Depth of Each Tree = 25
- Linear SVM : C = 5

2) Test Data Results: The classifier models underwent Hyperparameter Tuning and were then evaluated on the test dataset. The results have been summarized as follows:

TABLE I
UNBALANCED DATASET WITH NO PCA

Model	Evaluation Metrics			
	F1 Score	Accuracy	Recall	Precision
Logistic Regression	0.51	89.21%	0.39	0.72
Neural Network	0.62	90.15%	0.56	0.69
Random Forest	0.65	91.08%	0.58	0.74
Support Vector Machine	0.01	83.29%	0.06	0.03

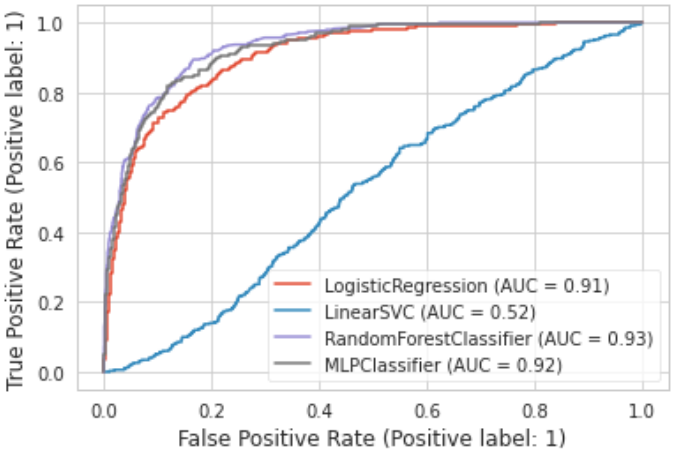

Fig. 4. AUC-ROC curve for D1

The Random Forests Classifier has the best score in the terms of all aspects for dataset D1. The SVM classifier had a very poor performance.

Now we use the features obtained using PCA and make dataset D2 whose results are as follows:

After this we use oversampling to make a balanced dataset D3 and use all the feature. The results obtained are as follows:

TABLE II
UNBALANCED DATASET WITH PCA

Model	Evaluation Metrics			
	F1 Score	Accuracy	Recall	Precision
Logistic Regression	0.49	88.80%	0.38	0.70
Neural Network	0.62	89.94%	0.58	0.67
Random Forest	0.62	89.78%	0.58	0.66
Support Vector Machine	0.07	40.71%	0.17	0.05

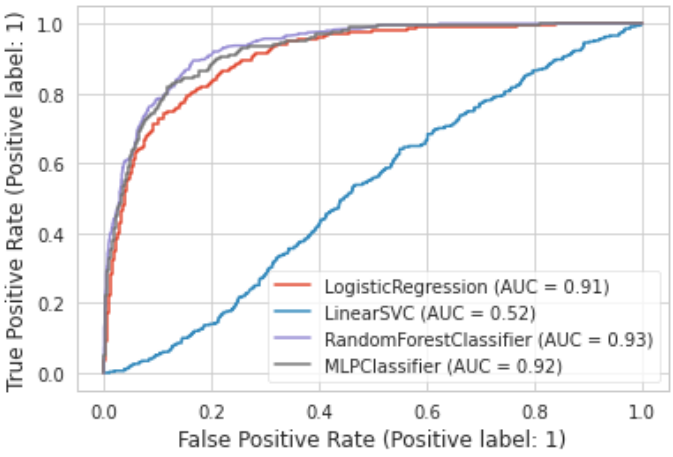

Fig. 5. AUC-ROC curve for D2

TABLE III
BALANCED DATASET WITH NO PCA

Model	Evaluation Metrics			
	F1 Score	Accuracy	Recall	Precision
Logistic Regression	0.80	82.44%	0.75	0.86
Neural Network	0.94	94.46%	0.98	0.91
Random Forest	0.96	96.26%	1.00	0.92
Support Vector Machine	0.56	51.81%	0.65	0.50

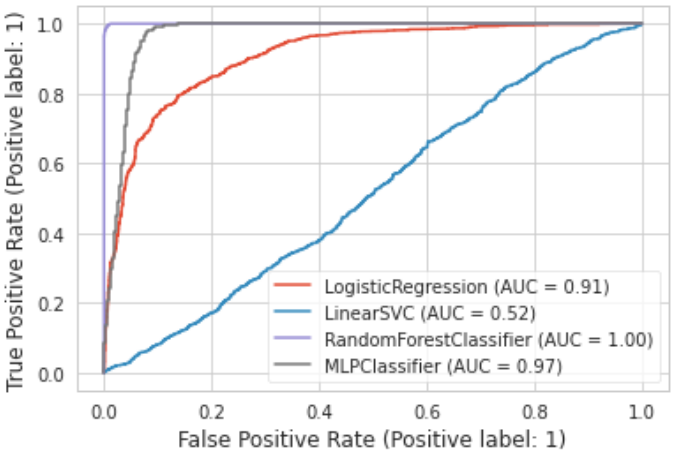

Fig. 6. AUC-ROC curve for D3

Now using balanced dataset D4 which also uses PCA, we get the following results:

TABLE IV
BALANCED DATASET WITH PCA

Model	Evaluation Metrics			
	F1 Score	Accuracy	Recall	Precision
Logistic Regression	0.78	81.26%	0.72	0.86
Neural Network	0.88	88.13%	0.91	0.85
Random Forest	0.95	95.44%	0.99	0.91
Support Vector Machine	0.64	63.18%	0.70	0.59

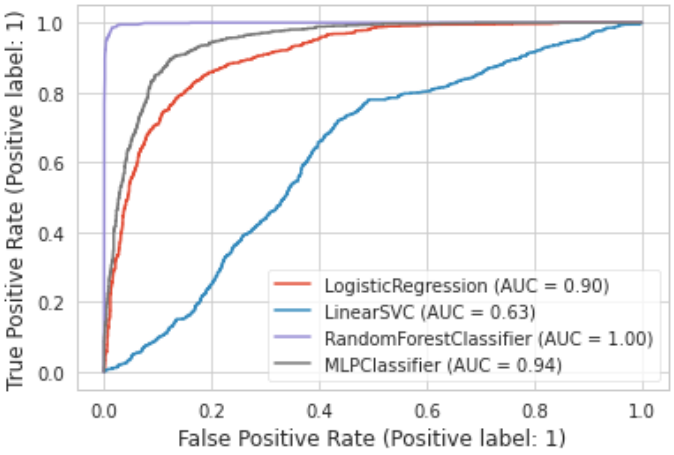

Fig. 7. AUC-ROC curve for D4

VI. CONCLUSIONS / DISCUSSIONS

With the help of the analysis made earlier, we can make the following conclusions and recommendations:

- The website has a **good retention rate** as most of the users are returning users. We also see a slightly higher conversion rate for the new users. So it is recommended to invest in advertisements, promotional campaigns, discounts to attract new visitors.
- Majority of the users belong to Region 1 and 3. Emphasise should be given for inter-regional promotions to expand the website to audience from other regions also.
- Bounce Rates and Exit Rates are negatively correlated with the Revenue whereas Page Values is positively correlated. This can be improved by optimizing the page to make its UI friendlier, popup discounts and introducing short descriptions.
- Majority of the users use Operating System 1,2 and 3 and Browsers 1,2. The website can be optimized for these browsers and operating systems to give the user the best experience.
- With the help of Principal Component Analysis we were able to reduce the number of feature from 29 to 14,

without significant loss of information. This also helped in improving the performance of the models.
- Random Forests Classifier and Neural Networks were the best performing models across all datasets with Random Forests performing slightly better than Neural Networks
- Support Vector Machine (Classifier) was the worst performing classifier among all the classifiers for all the datasets.
- The best result was obtained on the dataset D3 (balanced and without PCA) with the **Random Forest Classifier**. The F1 score for this case was **0.96** and an accuracy of **96.26%**.

VII. Learning Outcome & Future Work

The project helped us to learn on how to perform a data analysis and model development in a systemic manner. Also many aspects of EDA, feature selection and model development was learned in the process. For the future, we would like to implement some more analysis methods like clustering analysis and carry out a more study of each of the parameters affecting the target variable. Additionally, some better *Deep Learning* models like LSTMs can be used for real time prediction, regarding which literature can be found on the internet. Apart from that, some other sophisticated methods to deal with the class imbalance present in the data can be tried out.

Contribution

- Satwik Murarka : EDA on the dataset | 60% of the report
- Rhythm Shah : PCA, hypothesis testing on the dataset | Collaborated with Abhi on the Video submission
- Tanmay Choudhary : Worked on Machine Learning models, and performance metrics | 40% of the report
- Abhi Manjunath : Worked on Machine Learning Models, and performance metrics | Worked on the video submission

References

[1] Esmeli, R., Bader-El-Den, M. Abdullahi, H. Towards early purchase intention prediction in online session based retailing systems. Electron Markets 31, 697–715 (2021). [https://doi.org/10.1007/s12525-020-00448-x]

[2] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, Vol. 16 (2002), pp. 321 – 357.

[3] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[4] Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. Neural Comput Applic 31, 6893–6908 (2019). https://doi.org/10.1007/s00521-018-3523-0

[5] L. A. Jeni, J. F. Cohn and F. De La Torre, "Facing Imbalanced Data–Recommendations for the Use of Performance Metrics," 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, 2013, pp. 245-251, doi: 10.1109/ACII.2013.47.

[6] Paula Branco, Luis Torgo, Rita Ribeiro. A Survey of Predictive Modelling under Imbalanced Distributions. arXiv, 1505.01658, 2015.

[7] Hand, D.J. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach Learn 77, 103–123 (2009). https://doi.org/10.1007/s10994-009-5119-5