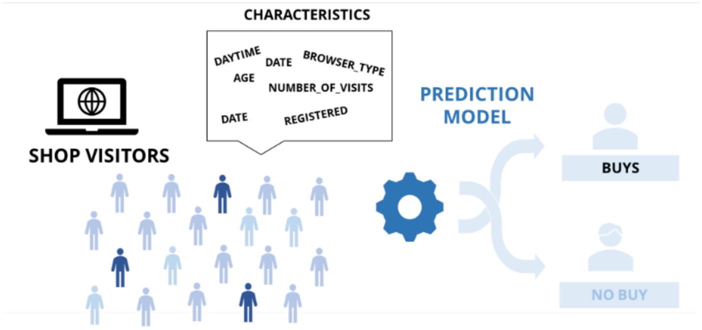
Machine Learning Module



Online Shopper's Intention Prediction

A Machine Learning Report

By

Salisu Abdullahi

C2224671

Table of Contents

Abstract

The project focused on applying machine learning classification models to e-commerce website data to outline which of the models is the most effective in predicting whether a shopper will make a purchase and which features are important in the purchase decision. Exploratory analysis and some data pre-processing techniques were applied to the data before splitting into a train, validation, and test set with oversampling techniques applied to the training data to prevent bias and improve training. These data sets were run through seven (7) classification model algorithms and the model performance assessed using accuracy and recall score. The three most effective models were outlined and hyperparameter tuning was applied to them to determine which is better and reduce assumptions of overfitting. It was determined that Extreme Gradient Boost was the best model suited to my dataset and research question. I further investigated to find out which of the features of my dataset were more important in my model towards the purchase outcome and used SHAP practice to determine this. This highlighted Page values, Exit Rate, Bounce Rate, and Month as top features important in purchase prediction. Finally, I concluded that using machine learning models to predict purchase conversion for an e-commerce website can be useful in helping businesses make feature or business decisions.

Introduction

With the rapid adoption of technology and the internet of things, e-commerce has become one of the prominent ways of shopping with the number of online shopping customers increasing continuously over the years (Ozen and Engizek, 2014). This is largely because it can be done from anywhere and at any time; and in some cases, might be cost effective. While these advantages have created an ecosystem, e-commerce websites are faced with the challenge of how and what influences customer intention to purchase? Studies which use Technology Acceptance Model (Davis *et al*, 1989) have highlighted ‘ease of use’ as one of the influences of customer shopping intention (Gefen et al., 2003a; Gefen et al., 2003b; Ha, 2020; Ha et al., 2019). Thus, to improve online customer chances of making a purchase, e-commerce websites and companies need to outline and understand how easy it is to use their platform and further identify the features which are important towards influencing customer’s purchase intention. In recent times, e-commerce websites have integrated analytical systems which helps businesses understand how their users interact with their platforms. Machine Learning as a form of artificial intelligence is quickly emerging as an attractive alternative to statistical methods in various industries with possible applications in marketing and e-commerce (Adrian *et al*, 2019).

Research Questions

I am attempting to answer the predictive research questions stated below:

- With provided session data, determine the best machine learning algorithm model to use in predicting whether shoppers will make a purchase.
- Determine the order of importance of the website features in purchase prediction.

Data Description, Exploration and Preparation

Data Collection

Data was created by Sakar *et al* (2018), gotten from UCI Machine Learning Repository (Online Shoppers Purchasing Intention Dataset) and can be downloaded here. The data was formed over a one-year period with each session belonging to a different user. The data is in CSV format.

Data Loading and Cleaning

Data cleaning is a very crucial part of any machine learning analysis as it allows identification of missing values or duplicated values in a dataset which might affect the machine learning algorithm outcome. Although Pandas and NumPy libraries are useful in handling such cases. However, in this dataset, there was little or no missing values. Data was loaded as a pandas dataframe.

Data Description

From observation, the data contains 18 columns and 12,330 rows with 17 of the columns being feature variable and one column (Revenue) being a target variable/class. The rows represent each session totalling 12,330 sessions collated over the one-year period. The columns attributes have been described below:

- **Administrative**: represents the number of admin/login pages visited in a session.
- **AdministrativeDuration**: represents the total time spent on admin/login pages in a session.
- **Informational**: represents the number of informational pages visited in a session.
- **InformationalDuration**: represents the total time spent on informational pages in a session.
- **ProductRelated**: represents the number of product-related pages visited in a session.
- **ProductRelatedDuration**: represents the total time spent on product-related pages in a session.
- **BounceRates**: represents the percentage of shoppers who enter the website through that page and exit without triggering any additional tasks during that session.
- **ExitRates**: represents the percentage of pageviews on the website that end at that specific page.
- **PageValues**: represents the average value for a web page that a user visited before completing an e-commerce transaction.
- **SpecialDay**: indicates the closeness of the site visiting time to a specific special day (e.g., Mother's Day, Christmas) in which the sessions more likely end in purchase.
- **Month**: Contains the month the session occurred.
- **OperatingSystems**: represents the operating system that the shopper was on during the session.
- **Browser**: represents the browser that the shopper was using during the session.
- **Region**: represents the region the user is in during the session.
- **TrafficType**: represent how the shopper accessed the website (direct URL input, referral link or through search engine).
- **VisitorType**: represents whether a visitor is New Visitor, Returning Visitor, or Other.
- **Weekend**: represents whether the session is on a weekend.
- **Revenue**: represents whether the user completed the purchase or not.

Data Exploration – Univariate Analysis

Univariate analysis explores variables in a dataset separately analysing them each. Here, I explored a few variables in my dataset, observing them and understanding the trend of distribution in each variable. The variables analysed in my dataset were Revenue, VisitorType, Weekend and Month with visualisation shown below.

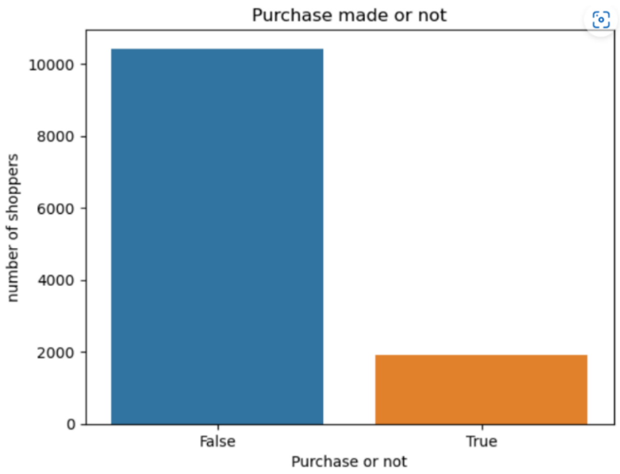


figure 1: Distribution of Purchase decision

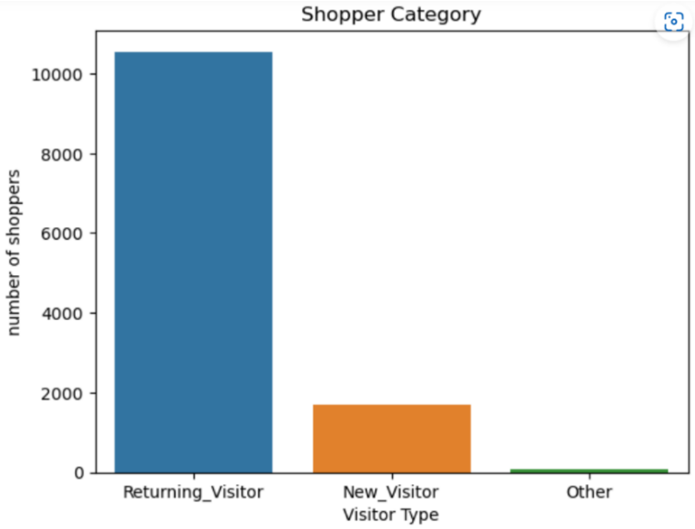


figure 2: Distribution of Visiting Shopper Type

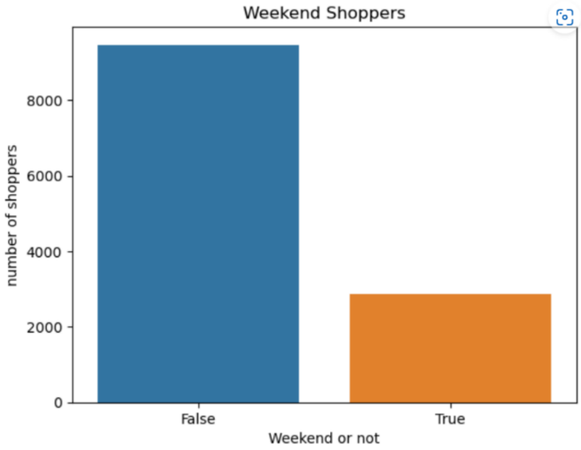


figure 3: Distribution of Shoppers on a weekend

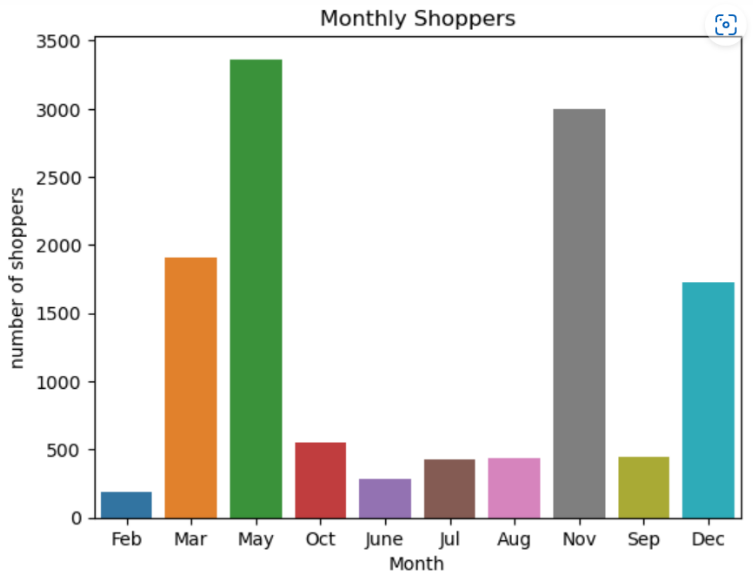


figure 4: Distribution of Shopper in a Month

From above visualizations, I noticed that most of the feature variables have an imbalanced decision. In summary, univariate analysis can also provide insights for business and marketing decisions specific to the features I outlined.

Data Exploration – Bi-Variate Analysis

Bi-Variate analysis involves analysis of two variables for the purpose of determining the relationship or association between them. For my dataset, I analysed ProductRelated Duration, bounce rate, administrative duration, visitor type, Page Values each against revenue. Analysis visualization can be seen below.

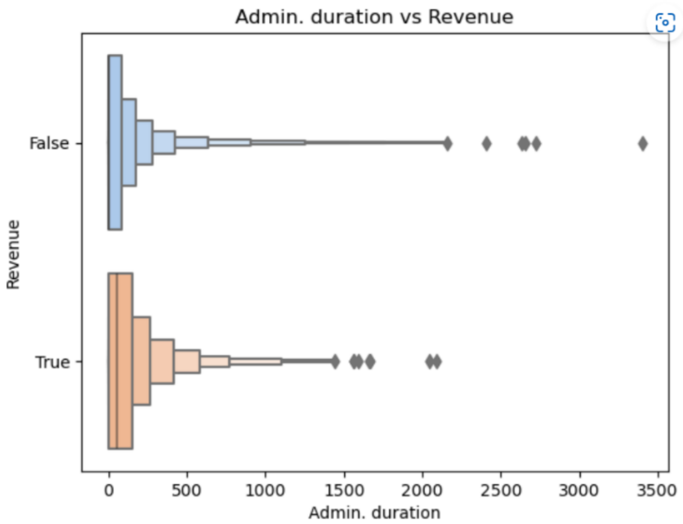


figure 5: Distribution of Revenue by Administrative duration

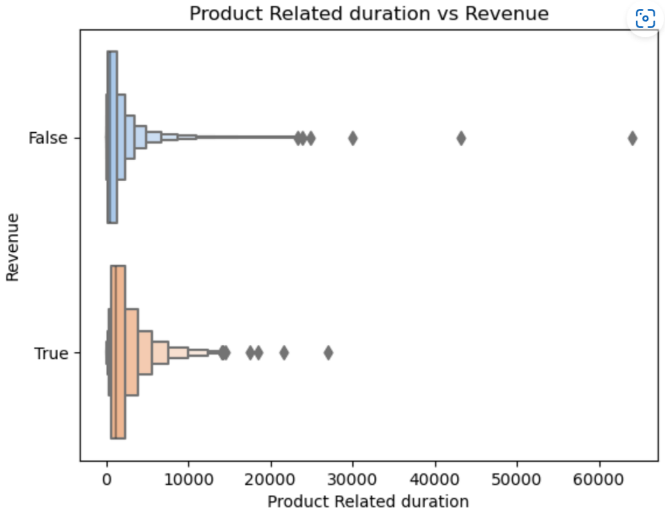


figure 6: Distribution of Revenue by Product Related duration

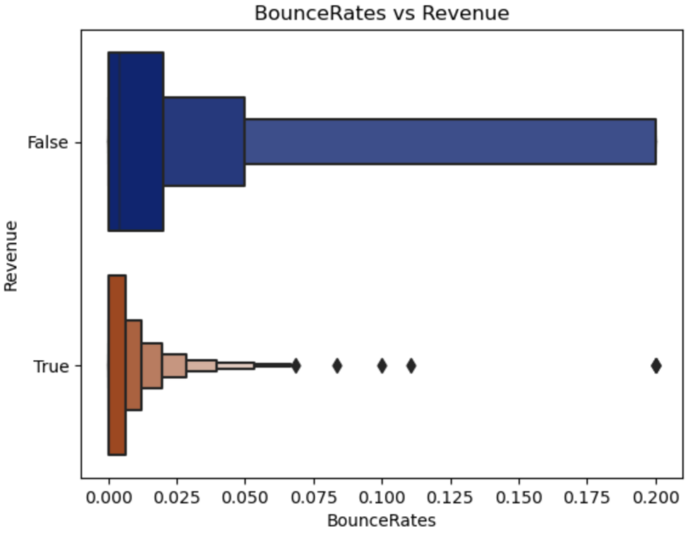


figure 7: Distribution of Revenue by Bounce rates

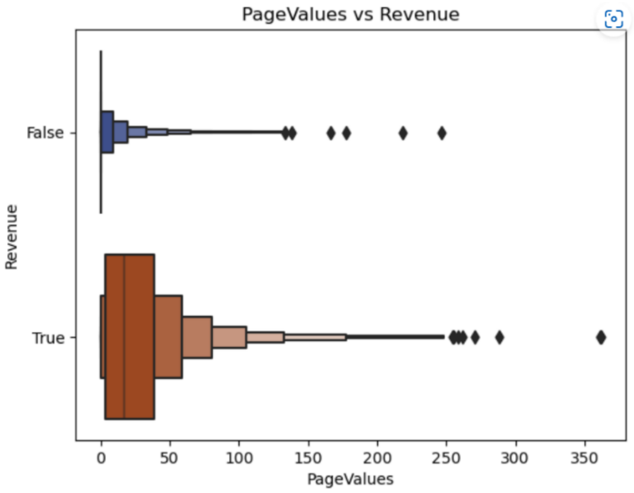


figure 8: Distribution of Revenue by PageValues

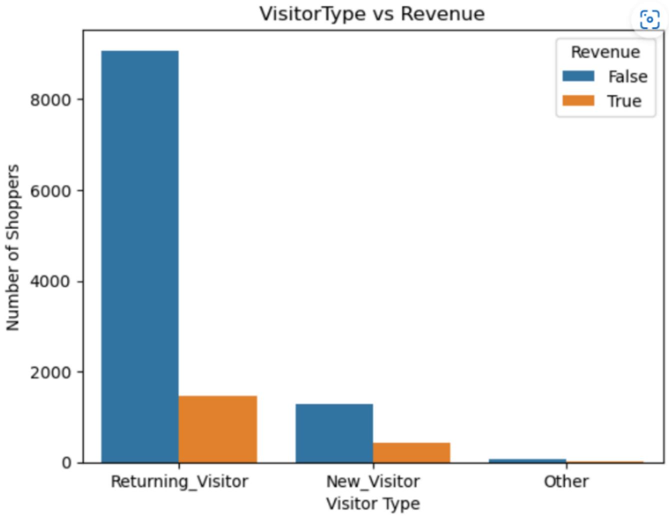


figure 9: Distribution of Revenue by VisitorType

From visualization, I noticed that there were some outliers present in some features. In summary, Bi-Variate analysis can be useful in decision making and provide insight into customer behaviour and influences to improve sales.

Feature Engineering

Label encoding was done to convert categorical data into binary data to allow for easy reading and analysis by the machine learning model. Label encoding was selected because it does not affect the dimensionality of the dataset. Columns converted include Revenue, VisitorType, Month and Weekend.

Upon conversion into binary data, the independency of features was determined. This is to eliminate repetition and assumption of bias. To do this, I performed a correlation matrix using Pearson Correlation Matrix with correlation threshold set at 0.60 (60%).

From the correlation analysis, there was a high correlation between administrative data (both duration and actual), Informational, Product Related, and Rates (Exit and Bounce). To avoid multi-collinearity that exists between the features, we eliminate the feature with less correlation to revenue (administrative duration, informational duration, product related duration and exit rates). Also, PageValues was the most correlated feature to revenue while traffic type was the least correlated feature.

Data Preparation – Splitting, Scaling and Oversampling.

The data was split into 3 sets (training, validation, and test sets) using the splitting formula 60:20:20. From earlier visualizations, some feature outcomes were imbalanced and contained outliers. Thus, to avoid bias and inaccuracies, I decided to scale the dataset and oversample the train data as well to account for the imbalance preventing bias during training. The scaling technique used was the Standard Scaler and the oversampling technique used on the training dataset was the Random Oversampling technique. Standard scaler was used as not all features contained data with outliers. Also, features such as PageValues have information that are useful as outliers.

Finally, after data cleaning, pre-processing and preparation, the data translated into three sets of data with the train data containing 12,516 sessions (after oversampling), the validation dataset containing 2,466 sessions and the test dataset containing 2, 466 sessions.

METHOD

Model Selection

The research question required selection of a machine learning algorithm that best evaluates the dataset using all its parameters. Thus, Machine learning algorithms models selected were those that are best for binary classification problems as they predict one of two possible outcomes (purchase or not). Thus, the algorithms selected are K-Nearest neighbour, naïve bayes, support vector machine, logistic regression, decision trees, Random Forest, and extreme gradient boosting.

Model Performance Evaluation

While accuracy may seem like an effective way to determine how well an algorithm is learning and predicting, it may not be a true reflection of the degree to which the algorithm prediction is correct. This is because the accuracy might be because of assumption in data. Since the aim of this project is to identify the best algorithm that predicts the online shopper intention to purchase and is based on features of the website, I evaluated the models using precision, f1 score, accuracy, and recall rate. However, to select the top 3 best performing algorithms, accuracy and recall rate were the metrics used. This is in line with the project goal as recall rate determines how well a model predicts all the labels (outcomes) correctly (Khalid, 2020).

Hyperparameter Tuning

For all models to be used, the parameters used in each model are the default parameters. After running all models, the best three models are selected. These three undergo hyperparameter tuning. There is no correct approach to choosing hyperparameters (Wade and Glynn, 2020). The aim of this is to try and reduce bias in training data as well as improve the accuracy and recall rate of the models by finding the optimal parameters in which each model will perform best. RandomizedSearchCV and GridSearchCV were the hyperparameter tuners used in this procedure to find the optimum parameters.

Feature Importance

This is useful as it helps with the understanding of how the model makes its predictions giving insights into the features that make the model work well. To get this, I used Shapley Additive Explanation technique (SHAP). ShAP is considered better than other traditional methods of determining feature importance because they can be inconsistent, which means that the most important features might not be given the highest importance score (Rathi, 2020).

RESULTS

Table 1: Model Performance Results

MODEL	ACCURACY (%)	RECALL SCORE (%)	TRUE POSITIVE	TRUE NEGATIVE
K-Nearest Neighbour	84.18	40.62	0.91	0.41
Naïve Bayes	82.31	64.20	0.85	0.64
Logistic Regression	86.86	72.44	0.89	0.72
Support Vector Machine	67.88	76.13	0.67	0.76
Decision Trees	84.54	51.42	0.90	0.51
Extreme Gradient Boosting	87.18	69.60	0.90	0.70
Random Forest	89.45	67.61	0.93	0.68

DISCUSSION

From the results above, the three best algorithms were random forest, extreme gradient boosting, and logistic regression. Random forest classifier is a very efficient algorithm because it can handle large date producing predictions that can be understood easily and providing the highest accuracy amongst available classification methods. It is also not susceptible to overfitting and trains faster than decision tree classifier. Extreme gradient boosting algorithm is a flexible model consisting of multiple decision trees built in parallel and are very efficient as classifiers as they minimize bias and underfitting in a dataset. Logistic regression is a simpler model to the other two and is mostly used in binary or linear classification problems.

Upon hyperparameter tuning, extreme gradient boosting was the best model slightly edging logistic regression and random forest. There were some limitations when running through random forest and extreme gradient boosting algorithms which may have an impact on the accuracy and recall value of the models. Important features as determined by ShAP are Page Values, Month, Product related duration, Bounce rate. This agrees with the exploratory analysis done at the beginning of the project.

CONCLUSION

From our procedure and results, the best algorithm for determining online shoppers' intention to make purchase is the Extreme Gradient Boosting. A few factors could have affected my result and its accuracy. For example, the limitation on number of estimators may be because of my computer's processing capacity and this limitation may be solved when algorithm is run on a higher capacity system as the more trees (estimators) in the random forest, the better the performance (Plonski, 2020). Furthermore, the difference in accuracy in extreme gradient boosting and random forest before and after hyperparameter tuning suggests that both models can be used interchangeably and running the model again might provide a result in which random forest is the best algorithm. There is also need for more data in which there is actual purchase by shoppers in a session as that class was a minority class. More features and data should be provided to further test algorithm performance. Finally, inferences from exploratory data analysis can be implemented to further improve outcome further improving performance of the model.

REFERENCES

1. Adrian, M. *et al* (2019). ‘Leveraging E-Commerce Performance through Machine Learning Algorithms’, *Annals of “Dunarea de Jos” University of Galati Fascicle I. Economics and Applied Informatics,* 25(2), 162-171. doi: https://doi.org/10.35219/eai1584040947
2. Davis, F. D., *et al* (1989). ‘User acceptance of computer technology: a comparison of two theoretical models’, *Management Science*, 35(8), 982-1003. doi: https://doi.org/10.1287/mnsc.35.8.982
3. Gefen, D., *et al* (2003a). ‘Trust and TAM in online shopping: an integrated model’, *MIS Quarterly*, 27(1), 51-90. doi: https://doi.org/10.2307/30036519
4. Gefen, D., *et al* (2003b). ‘Inexperience and experience with online stores: the importance of TAM and trust’, *IEEE Transactions on Engineering Management*, 50(3), 307-321. doi: https://doi.org/10.1109/tem.2003.817277
5. Ha, N. T., *et al* (2019). ‘The effect of trust on consumers' online purchase intention: An integration of TAM and TPB’. *Management Science Letters*, 9(9), 1451-1460. doi: https://doi.org/10.5267/j.msl.2019.5.006
6. Ha, N. T. (2020). ‘The impact of perceived risk on consumers' online shopping intention: An integration of TAM and TPB’, *Management Science Letters*, 10(9), 2029-2036. doi: https://doi.org/10.5267/j.msl.2020.2.009
7. Khalid I. (2020). ‘Greater accuracy does not mean greater machine learning model performance’, *Towards Data Science,* 13th May. Available at: https://towardsdatascience.com/greater-accuracy-does-not-mean-greater-machine-learning-model-performance-771222345e61 (Accessed: 6 May 2023)
8. Ozen, H., and Engizek, N. (2014). ‘Shopping online without thinking: being emotional or rational?’, *Asia Pacific Journal of Marketing and Logistics*, 26(1), 78-93. Available at: https://www.researchgate.net/publication/280193248_Shopping_online_without_thinking_Being_emotional_or_rational (Accessed: 6 May 2023)
9. Plonski P. (2020). ‘How to reduce memory used by Random Forest from Scikit-Learn in Python?’, *Mljar*, 24th June. Available at: https://mljar.com/blog/random-forest-memory/ (Accessed: 6 May 2023)

10. Rathi, P. (2020). ‘A novel approach to feature importance – Shapley Additive Explanations’. *Towards data Science.* 2nd July. Available at: https://towardsdatascience.com/a-novel-approach-to-feature-importance-shapley-additive-explanations-d18af30fc21b (Accessed: 6 May 2023)

11. Sakar, C. *et al* (2018). ‘Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks’, *Neural Computing and Applications*, 31, 6893–6908. Available at: https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset (Accessed: 6 May 2023)

12. Wade C., Glynn K. (2020). *Hands-On Gradient Boosting with XGBoost and scikit-learn,* Page 57. Available at: https://books.google.co.uk/books?id=2tcDEAAAQBAJ&lpg=PP1&ots=s4yMHjmfiK&dq=How%20to%20reduce%20memory%20used%20by%20Random%20Forest%20from%20Scikit-Learn%20in%20Python%3F&lr&pg=PP1#v=onepage&q&f=false (Accessed: 6 May 2023)