PREDICTING THE ONLINE CUSTOMERS' PURCHASE INTENTION COMPARING MACHINE AND DEEP LEARNING MODELS

ANNABELLE SONNEVELDT

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

462078

COMMITTEE

dr. Nápoles
dr. Postma

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

February 4, 2022

Word count: 8767

Preface

Dear reader,

Thank you for taking the time to read this thesis on the prediction of the customers’ purchase intention. I would like to thank my supervisor, Dr. Nápoles for his feedback throughout the thesis period.

Best regards,

Annabelle Sonneveldt

CONTENTS

PREDICTING THE ONLINE CUSTOMERS' PURCHASE INTENTION COMPARING MACHINE AND DEEP LEARNING MODELS

ANNABELLE SONNEVELDT

Abstract

In recent years, the number of online customers using e-commerce websites has increased significantly, although conversion rates have remained unchanged. Given the unchanging conversion rates, understanding the purchase intentions of online customers is significantly important for online decision-makers. This gives decision-makers a better understanding of the customers' intention to purchase a product, and as a result, this might improve the customer experience and thereby increase conversion rates. Existing research has previously attempted to predict the customers' purchasing intent. This study will focus on finding the best model and features for predicting the customers' purchase intentions. The Online Shoppers Purchasing Intention Dataset used has an unbalanced class label due to the few positive transactions relative to the negative transactions (Sakar, Polat, Katircioglu, & Kastro, 2018). The results of previous studies have shown that the Bagging and Boosting ensemble learning models handle the issue of imbalance (Algawiaz, Dobbie, & Alam, 2019). However, Kabir, Ashraf, and Ajwad (2019) state that the accuracy can be improved with the use of a deep learning model. Therefore, the additional value of the deep learning model Multilayer Perceptron is being studied. Furthermore, another research gap is the identification of the most important features when comparing ensemble learning models and a deep learning model. The findings show that the Multilayer Perceptron performs better than the baseline model and the Bagged Decision Trees. However, Gradient Boosting outperforms the other models with an F1-score of 69.2%. Finally, page value and exit rate are the features that have a significant impact on the customers' purchase intentions.

DATA SOURCE/ CODE/ ETHICS STATEMENT

Work on this thesis did not involve collecting data from human participants or animals. The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis. The author of this thesis acknowledges that they do not have any legal claim to this data or code. The code used in this thesis is publicly available on `https://github.com/AFL1997/Master_thesis.git`.

1 INTRODUCTION

In recent years, online buying has become increasingly popular. On the other hand, the conversion rates of these e-commerce websites have remained unchanged, which means that the total number of sales has stayed constant (Behera, Gunasekaran, Gupta, Kamboj, & Bala, 2020; Liu, Lee, & Srinivasan, 2019; Zhou, Mishra, Gligorijevic, Bhatia, & Bhamidipati, 2019). Due to the unchanged conversion rates, online customers' purchase intents are significantly important to online decision-makers. This will provide an online decision-maker with a better understanding of the customers' intentions and, if necessary, allow them to adjust to the factors influencing their purchase intention. This might improve the customer experience and thereby, increase conversion rates.

One of the most commonly observed challenges when working with consumer data for online purchases is that only a small percentage of the visitors complete a transaction (Linoff & Berry, 2004). The Online Shoppers Purchasing Intention Dataset used has an unbalanced class label due to the few positive transactions relative to the negative transactions, which indicates that the dataset is imbalanced (Sakar et al., 2018). Previous research has shown that Bagging and Boosting ensemble learning models solve imbalance difficulties, which is discussed in Section 2.2.

The goal of this study is to find the best model and features for predicting the customers' purchase intentions. In previous studies, the Bagging and Boosting ensemble learning algorithms have shown promising results (Baati & Mohsil, 2020; Kabir et al., 2019; Martínez, Schmuck, Pereverzyev, Pirker, & Haltmeier, 2020). However, Kabir et al. (2019) state that the accuracy can be improved with the use of a deep learning model. From a scientific point of view, this study aims to contribute to the field of research by examining the added value of a deep learning model in comparison with ensemble methods. Furthermore, it is essential to understand the impact of the models' features in order to understand the predictions made by this study (Bugaj, Wrobel, & Iwaniec, 2021). In this study, the impact of the features is measured using the feature importance scores of Shapley

Additive Explanation (SHAP). This method was chosen because it works effectively with complex models like Gradient Boosting and Multilayer Perceptron (Lundberg, Erion, & Lee, 2018; Lundberg & Lee, 2017). Therefore, this study fills a gap in the literature by determining the important features of ensemble learning models and a deep learning model when predicting the customers' purchase intention.

As a result, this thesis will address the following research question:

> *To what extent can the customers' purchase intention be predicted using machine and deep learning models?*

This main research question is divided into two sub-questions.

> The first sub-question is: *"Which model performs significantly better when predicting the customers' purchase intention?"*

In order to answer the first sub-question, the Bagging and Boosting ensemble learning algorithms and a Multilayer Perceptron are trained and compared with each other. The classification evaluation metrics will determine the models' performance, with the F1-score playing a significant role due to the imbalanced dataset.

> The second sub-question is: *"Which features have the largest impact on the prediction of the customers' purchase intention?"*

In order to answer the second sub-question, the impact of the features is determined using the feature importance scores of Shapley Additive Explanation (SHAP). This method works effectively with complex models like Gradient Boosting and Multilayer Perceptron. The results of this method will be compared with the baseline model and the findings from previous research.

The findings of this study show that the ensemble learning model Gradient Boosting remains the best model for predicting the customers' purchase intent. The use of a deep learning model adds limited value, as the Multilayer Perceptron performs better than the baseline model and Bagged Decision Trees. However, Gradient Boosting outperforms all the other models with an F1-score of 69.2%. Furthermore, in both the different types of models and the previous research, the features page value and exit rate are referred to as important features that have a significant impact on the customers' purchase intention (Shi, 2021). This indicates that the time a visitor spends on a web page influences whether they make a purchase, and that a lower exit rate has a positive impact on whether the visitor makes a purchase.

This paper is organised in the following way. In Section 2, the related work part focusing on the purchase intention is given. Section 3 describes the methods used to answer the research questions. Section 4 explains the procedures for completing the research. Section 5 summarizes the results of the study concerning the research question. In Section 6 and Section 7, the results of this study are discussed.

2 RELATED WORK

Online purchasing has become increasingly popular. However, the significant increase in the use of e-commerce websites has not resulted in higher conversion rates (Behera et al., 2020; Liu et al., 2019; Zhou et al., 2019). Therefore, the understanding of the customers' purchase intentions is significantly important to identify. This gives online decision-makers greater clarity, allowing them to improve the customer experience and, as a result, increase conversion rates.

Several studies have already attempted to develop a model that can predict online customers' purchase intentions using the Online Shoppers Purchasing Intention Dataset (Baati & Mohsil, 2020; Esmeli, Bader-El-Den, & Abdullahi, 2020; Kabir et al., 2019; Sakar et al., 2018). These research findings are presented in Section 2.1. One of these findings shows that ensemble learning models outperform other classification models, which will be discussed in greater detail in Section 2.2. Apart from identifying the model that produces the most accurate results, it is significantly important to determine which features have the largest impact on predicting the customers' purchase intentions. This will be discussed in Section 2.3. Finally, the research questions and goals are formulated in relation to the literature review.

2.1 *Previous research*

Previous related research identified the most suitable algorithm for predicting purchase intention. This study used the classification algorithms on the Online Shoppers Purchasing Intention Dataset (Kabir et al., 2019). This dataset has an unequal class label due to the few positive transactions relative to the negative transactions, which indicates that the dataset is imbalanced (Sakar et al., 2018). In terms of accuracy, Random Forest outperformed the other algorithms because this is a weighted model that works well with imbalanced datasets (Kabir et al., 2019). The prediction of unweighted models tends to predict the majority class, while the weighted models are more balanced. These models take the minority class into consideration, which in this case is the positive transaction (Shi, 2021). Therefore, Random Forest is more accurate when predicting the customers who have made a positive transaction.

The performance of the Random Forest is improved with the use of other ensemble methods. The ensemble method Gradient Boosting performs best in this situation due to the best number of splits and trees (Kabir et al., 2019). As a result, Kabir et al. (2019) suggest that ensemble models outperform all other classification models. These ensemble models will be discussed in greater depth in Section 2.2.

This area of research is also being investigated in terms of predicting real-time shopper behavior rather than predicting the customers' purchase intentions afterwards. The real-time shopper behavior framework is composed of two components. These two components continually forecast purchase intents and the churn rate of the site throughout the prediction horizon, which allows these components to take appropriate steps to enhance website dropout and conversion rates (Sakar et al., 2018).

The customers' purchase intention is predicted in the first part based on information about the session and the user. According to the findings of this study, the Multilayer Perceptron outperforms the other algorithms. The churn rate is predicted in the second part based on sequential clickstream data using a Long-Short-Term Memory Recurrent Neural Network. This information enables businesses to provide material only to visitors who are expected to leave the website within the specified time period. Furthermore, the clickstream data features along with the session information-based features provide significant information for predicting the customers' purchase intentions (Sakar et al., 2018).

One disadvantage of the previously mentioned real-time prediction framework is that it predicts purchases after customers have browsed two or more products (Sakar et al., 2018). In contrast to this studies, a real-time system is built for analyzing online customer behavior. The moment a visitor connects to a website, the system detects visitors with a high purchase intention. This enables businesses to make promotional offers to visitors who visit the website with a high purchase intention. The algorithms used for this study are Naïve Bayes, C4.5, Decision Tree and Random Forest. According to the findings of this study, Random Forest yields a higher accuracy and F1-score than the other algorithms. Due to the imbalanced dataset, this study shows the performance of the Random Forest as a weighted model once more (Baati & Mohsil, 2020).

Unlike predicting the purchase intentions in real-time, the purchase intentions could also be predicted within a predefined time frame in the future. A new collection of customer-relevant features was employed to forecast this, which were generated from the times and values of past purchases. The algorithms used for this study are Logistic Lasso Regression, Extreme Learning Machine and Gradient Tree Boosting. According to the findings of this study, the Gradient Boosting turns out to be the best performing method (Martínez et al., 2020). The ensemble learning model Gradient Boosting is discussed in further detail in the following section of this chapter.

2.2 *Ensemble learning methods*

Previous studies have demonstrated that Random Forest and Gradient Boosting outperform the other classification models when predicting the customers' purchase intentions using the Online Shoppers Purchasing Intention Dataset (Baati & Mohsil, 2020; Kabir et al., 2019; Martínez et al., 2020). These models are ensemble learning models, a type of machine learning methodology and this can be used to enhance an algorithm's performance (Sridhar, Mootha, & Kolagati, 2020). Ensemble learning is a method that combines multiple learning models to provide better forecasts and better results in terms of accuracy and generalization (Dong, Yu, Cao, Shi, & Ma, 2019).

Bagging and Boosting are some examples of ensemble learning. Bagging takes several bootstrap samples, fits a weak learner to each of them, and then aggregates the results so that an average can be calculated. An example of a Bagging ensemble algorithm is Random Forest (Breiman, 1996; Syarif, Zaluska, Prugel-Bennett, & Wills, 2012). Boosting generates a number of base learners, each of which is re-weighted according to its performance in a sequential manner, making the base learners stronger (Syarif et al., 2012). It has been shown that Bagging and Boosting learning models solve imbalance difficulties since these models rely on weak classifiers that raise the weight of incorrectly classified examples with each iteration (Algawiaz et al., 2019).

The Boosting ensemble learning model AdaBoost has been implemented into a recommender system designed to predict customers' purchasing intentions. Predicting this can be viewed as critical information in a recommendation system because it enables a better understanding of the user and, as a result, more accurate recommendations. The suggested model classifies eight features that can affect a customer' purchasing intention. These features are put into the AdaBoost algorithm, which uses them to predict a customers' purchase intent. The other models used for this study are Multilayer Perceptron and Recurrent Neural Network. According to the results of this study, AdaBoost outperforms the other algorithms with an accuracy of 91%. The reason is that AdaBoost is good at predicting imbalanced datasets (Algawiaz et al., 2019).

Gradient Boosting is another well-known Boosting ensemble model. As previously stated, this ensemble learning model has been utilized to predict the customers' purchase intentions. These findings demonstrate that Gradient Boosting is a model that outperforms other classification models due to its ability to deal with imbalanced difficulties (Algawiaz et al., 2019; Martínez et al., 2020). This is also the case when researchers utilize statistical analysis to develop new features from existing ones. These features are stored in the Gradient Boosting ensemble method, Decision Tree, Support Vector Machine, and Multilayer Perceptron. In this study, Gradient Boosting outperforms the other models due to the new features. As a conclusion, feature engineering demonstrates that it is still advantageous to employ traditional machine learning models for classification rather than rely on deep learning models (Kiki & Houndji, 2020).

2.3 *Features*

Previous studies discussed the best models for predicting customer purchase intentions. In order to increase conversion rates, it is critical to understand customers' purchasing patterns and the factors influencing purchasing intention (Shi, 2021). The features that could impact the customers' purchasing intentions are analyzed by descriptive statistical analysis using the Online Shoppers Purchasing Intention Dataset. The results of this study show that features like time spent on the website and page value are positively correlated with the customers' purchase intent. However, features including bounce rates and exit rates are negatively correlated with the customers' purchase intent. The algorithms used in this study are Logistic Regression, Decision Tree and Random Forest. In this study, Random Forest outperforms the other algorithms with an accuracy of 87.5%. As previously stated in the previous research section, the same appears to be true for this study. Random Forest is a weighted model with a more balanced accuracy rate, whereas an unweighted model prefers to predict the majority class of the dataset (Shi, 2021).

2.4 *The current study*

Unlike previous studies, during this study the focus will be on finding the best model and features for predicting the customers' purchase intention when comparing machine and deep learning models. Machine learning will use different Bagging and Boosting ensemble learning models. These models solve imbalance difficulties by relying on weak classifiers that raise the weight of incorrectly classified examples with each iteration (Algawiaz et al., 2019). According to Kabir et al. (2019), the accuracy might be improved by the use of a deep learning model. For this reason, the ensemble learning algorithms are compared to the deep learning model Multilayer Perceptron.

The identification of the most important features using ensemble learning models and a deep learning model is another research gap that will be investigated. This means that the most relevant features for each model are identified, and the outcomes of each model are compared. To my knowledge, the above will expand the research's contribution to customer purchase intent. As a result, the following main research question and sub-question will be addressed: *"To what extent can the customers' purchase intention be predicted using machine and deep learning models?"* This main research question is divided into two sub-questions.

> The first sub-question is: *"Which model performs significantly better when predicting the customers' purchase intention?"*

The different machine learning and deep learning models will be compared to answer this sub-question. In this study, machine learning will use Bagging and Boosting ensemble learning algorithms instead of single machine learning models to predict the customers' purchase intentions. This is based on the finding that ensemble learning algorithms provide better forecasts and results by combining learning models in terms of accuracy and generalization (Dong et al., 2019). Bagging and Boosting ensemble learning models also address the issue of imbalance by relying on weak classifiers that raise the weight of incorrectly classified examples (Algawiaz et al., 2019). A few studies have already used Random Forest and Gradient Boosting (Baati & Mohsil, 2020; Kabir et al., 2019; Martínez et al., 2020). However, Kabir et al. (2019) state that the accuracy can be improved with the use of a deep learning model. That is the reason that Multilayer Perceptron is included in this study, which has been successfully applied to the same dataset (Sakar et al., 2018).

This study aims to contribute to this field of research by examining the added value of a deep learning model in comparison with ensemble learning models when predicting the customers' purchase intention. As a result, the Bagging and Boosting ensemble learning algorithms and the Multilayer Perceptron will be trained and compared with each other to predict the customers' purchase intentions. The classification evaluation metrics will determine the models' performance, with the F1-score playing a significant role due to the imbalanced dataset.

> The second sub-question is: *"Which features have the largest impact on the prediction of the customers' purchase intention?"*

The features that have the largest impact on predicting the customers' purchase intent are searched for to answer this sub-question. It is essential to understand the impact of the models' features in order to comprehend the predictions for this study (Bugaj et al., 2021). There are different ways to measure the largest impact of the features in the dataset. Shi (2021) investigated the features that impact the customers' purchasing intentions through descriptive statistical analysis.

For this study, the impact of the features is measured using the feature importance scores of Shapley Additive Explanation (SHAP). This method was chosen because it works effectively with complex models like Gradient Boosting and Multilayer Perceptron (Lundberg et al., 2018; Lundberg & Lee, 2017). Therefore, this study fills a gap in the literature by determining the important features of ensemble learning models and a deep learning model when predicting the customers' purchase intention. The results of these models will be compared with the baseline model and the findings from the previous research. This provides a greater explainability of the complicated models predicting customers' purchase intent than in previous research (Bugaj et al., 2021).

The dataset for this study is the Online Shoppers Purchasing Intention Dataset. Each instance in the dataset represents a unique user's intent to complete the transaction (Sakar et al., 2018). The dataset consists of 12,330 instances and 18 features, from which 10 numerical features and 8 categorical features. These categorical features will be numerically transformed so that these features can be utilized to build and train the models. Furthermore, the dataset has an unequal class label due to the 1908 positive transactions versus the 10,422 negative transactions. To address this imbalanced problem, the oversampling method SMOTE is used. The dataset and oversampling method SMOTE are further explained in Section 4 of this thesis.

3 METHOD

This chapter describes the different machine learning and deep learning models that were compared to determine the best model for predicting the customers' purchase intention. The different machine learning models are Boosting and Bagging ensemble learning algorithms. These algorithms combine multiple learning models to provide better forecasts and better results in terms of accuracy and generalization (Dong et al., 2019). The Bagging ensemble algorithms used are Bagged Decision Tree and Random Forest. The Boosting ensemble algorithms used are AdaBoost and Gradient Boosting. These ensemble learning algorithms are compared to the deep learning model Multilayer Perceptron. The Multilayer Perceptron was chosen because the accuracy might be improved by the use of a deep learning model (Kabir et al., 2019).

Furthermore, the feature importance score has been used to find the features having the largest impact on predicting the customers' purchase intention. The feature importance scores were calculated using the SHAP values. SHAP is chosen because it works effectively with complex models like Gradient Boosting and Multilayer Perceptron and it can be used with both machine and deep learning models (Lundberg et al., 2018; Lundberg & Lee, 2017). This enables the comparison of the feature importance scores of different models. The models and the SHAP values are described in further detail in this chapter.

3.1 *Bagging ensemble algorithms*

Bagging also known as bootstrap aggregation is one of the most often used ensemble learning algorithms (Breiman, 1996). Bagging takes several bootstrap samples, fits a weak learner to each of them, and then aggregates the results so that an average can be calculated (Breiman, 1996; Syarif et al., 2012). The bagging ensemble algorithms used for this study are Bagged Decision Tree and Random Forest.

3.1.1 *Bagged Decision Trees*

The Bagged Decision Tree combines multiple Decision Trees to construct a powerful prediction model, which decides on the outputs of the different Decision Trees by a majority vote. This model builds and trains N decision trees on N training sets at random with replacement (Breiman, 1996). Equation 1 represents the bagged prediction, where X is the record for which the forecast is created and $f_b\hat{(X)}$ is the prediction of each individual base learner (Boehmke & Greenwell, 2019).

$$\hat{f_{bag}} = \hat{f}_1(X) + \hat{f}_2(X) + ... + \hat{f}_b(X) \tag{1}$$

3.1.2 *Random Forest*

Random Forest is another Bagging ensemble method that is used during the study. Random Forest, like Bagged Decision Trees, generates a significant number of associated decision trees. The difference between Random Forest and Bagged Decision Trees is that Random Forest randomly chooses the best split of features from a subset to divide a decision tree, whereas Bagged Decision Trees consider all features. When it comes to regression, the forecasts are averaged from a significant number of individual decision trees, and classification makes use of majority voting, in which the class label with the highest number of votes is classified (Breiman, 2001). By considering a subset and building a significant number of individual trees, the generalization accuracy will be increased (Schonlau & Zou, 2020). Furthermore, no pruning is utilized in Random Forest, which allows each decision tree to grow to its maximum significant potential (Breiman, 2001).

3.2 *Boosting ensemble algorithms*

An ensemble learning algorithm is a method that combines multiple learning models. Another commonly used ensemble learning method, Boosting, was presented by Schapire, Freund, Bartlett, and Lee (1998). Boosting generates a number of base learners, each of which is re-weighted according to its performance in a sequential manner, making the base learners stronger (Syarif et al., 2012). The Boosting ensemble algorithms used for this study are AdaBoost and Gradient Boosting.

3.2.1 *AdaBoost*

AdaBoost is the first boosting ensemble method described. The AdaBoost algorithm, also known as Adaptive Boosting, was created to build stronger classifiers from weak classifiers by training these multiple classifiers with a set of weights. Weak classifiers can be considered as guessing at random, whereas strong classifiers can be compared to accurate classification (Freund & Schapire, 1997). The advantage of AdaBoost is that the model is quick, accessible, and straightforward to apply. It is not necessary to be familiar with the weak classifiers beforehand. The AdaBoost algorithm can be expressed in Algorithm 1, adapted from Hastie, Tibshirani, and Friedman (2009).

Algorithm 1 AdaBoost

```
Initialize observation weights $w^i = 1/n, i = 1, ..., n$.
For $m$ = 1 to $M$:
        a. Fit a classifier $C_m(x)$ to the training data by using weights $w_i$.
        b. Compute $error_m \sum_{i=1}^{n} w_i I(y_i \neq C_m(x_i)) / \sum_{i=1}^{n} w_i$, where $I(A)$ is
    the indicator function, which is equal to 1 when $A$ materializes and 0
    otherwise.
        c. Compute $a_m = log((1 - error_m)/error_m)$.
        d. Set $w_i$ equal to $w_i exp(a_m I(y_i \neq C_m(x_i))), i = 1, ..., n$.
Predict $C(x) = sign(\sum_{m=1}^{M} a_m \cdot c_m(x))$ (i.e. by majority voting), where
    sign denotes the sign function.
```

Where the weights are represented as w_i where i denotes the training examples on a round m. Wang (2012) explains that with each round, the weights of incorrectly classified examples grow, encouraging the weak learners to focus on the challenging instances. Freund and Schapire (1997) and Hastie et al. (2009) indicate that the weights of incorrectly classified examples change because the weak hypothesis is identified by the weak learner and its importance is taken into account.

3.2.2 *Gradient Boosting*

Gradient Boosting is another Boosting ensemble method that is used during the study. Gradient Boosting, in contrast to AdaBoost, aims to reduce a loss function by boosting in the opposite direction of the gradient. The loss function indicates how well a model performs when it comes to predicting a problem. The difference between AdaBoost and Gradient Boosting is that AdaBoost uses strong weight samples to identify weaknesses, while Gradient Boosting uses gradients to identify weaknesses (Bahad & Saxena, 2020). Gradient Boosting is a type of gradient descent that is used to reduce the size of complex loss functions that cannot be reduced immediately.

Gradient Boosting can be expressed in Algorithm 2, adapted from Friedman (2001). Where M is the number of iterations. g is the gradient of the loss function, which is denoted as L with respect to the prediction value $f_1(x)$. h corresponds to the base learner for the gradient components. Update the prediction value $f(x)$ by computing the step magnitude magnifier. The process is repeated until the final predictive function is established (Friedman, 2001).

Algorithm 2 Gradient Boosting

```
Initialize $f_0$ with a constant.
For $m$ = 1 to $M$:
        a. Compute the negative gradient $g_m(x_i)$ of the loss function $L$ at $f_{m-1}(x_1), i = 1, ..., n$.
        b. Fit a new base learner function $h_m(x)$ to $(x_i, g_m(x_i)), i = 1, ..., n$.
        c. Update the function estimate $f_m(x) \leftarrow f_{m-1}(x) + ph_m(x)$.
Predict $f_M(x)$.
```

3.3 *Multilayer Perceptron*

Multilayer Perceptron (MLP) is a feedforward artificial neural network with numerous layers of nodes, each of which is fully connected to the next layer. The MLP is made up of three layers: an input layer, an output layer, and a single or multiple hidden layers (Ramchoun, Amine, Idrissi, Ghanou, & Ettaouil, 2016). MLP enables the prediction of the output data from given input data by using a nonlinear activation function in the hidden

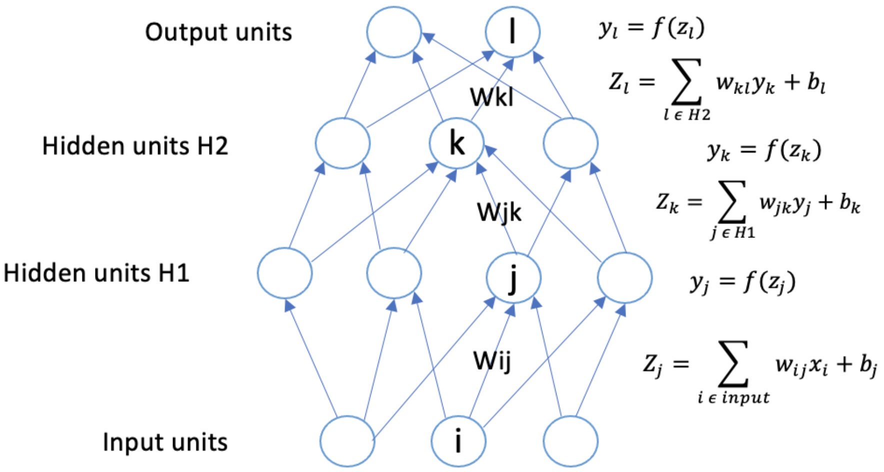
Output units
Hidden units H2
Hidden units H1
Input units
l
k
j
i
Wkl
Wjk
Wij
$y_l = f(z_l)$
$Z_l = \sum_{l \in H2} w_{kl} y_k + b_l$
$y_k = f(z_k)$
$Z_k = \sum_{j \in H1} w_{jk} y_j + b_k$
$y_j = f(z_j)$
$Z_j = \sum_{i \in input} w_{ij} x_i + b_j$

Figure 1: Multilayer Perceptron Algorithm. Source: (Shrestha & Mahmood, 2019)

layer (Taud & Mas, 2018). These activation functions assist the network in retraining important knowledge and discarding irrelevant data. An activation function determines which neurons should be stimulated and which should be left inactive (Ding, Qian, & Zhou, 2018). Furthermore, the loss function is calculated to determine the performance of the network. A low loss means that the network is working well, while a large loss means that the network is not working well (Janocha & Czarnecki, 2017).

The Multilayer Perceptron algorithm is expressed in Figure 1, adapted from Shrestha and Mahmood (2019). The weighted total of the inputs is denoted as Z. X corresponds to the input features of the dataset that are multiplied by the weight matrices of the node in the hidden layer w. In the hidden nodes, b is the bias value threshold. At each layer, Y represents the non-linear activation function f of Z (Shrestha & Mahmood, 2019).

3.4 *SHAP values*

As described at the beginning of this chapter, SHAP values are used to calculate the feature importance scores. SHAP values are chosen because they can be used for both machine and deep learning models. SHAP includes a variety of functions that can be applied to various models. The Bagging and Boosting ensemble learning models are both tree-based models that make use of the Tree SHAP (Lundberg et al., 2018). The Multilayer Perceptron model is a deep learning model that makes use of the Deep SHAP (Lundberg & Lee, 2017). This ensures that the results of the models' output may be compared to one another. Lundberg and Lee (2017) proposed SHAP, also known as Shapley Additive Explanations. SHAP assigns each feature an importance score for a prediction. The SHAP model does this by analyzing the feature values and substituting them with random feature values in a black-box model. SHAP analyzes how these feature values have evolved throughout time. The model then adds absolute values to calculate feature importance scores. This is expressed in the following equation, adapted from Lundberg and Lee (2017):

$$g(z') = \varnothing o + \sum_{j=1}^{M} \varnothing j z' j \tag{2}$$

where $z^{'} \in 0,1^{M}$relates to the presence of a feature value, with 1 indicating that the feature is present and 0 indicating that it is not, $\varnothing o$ is the predicted value, $\varnothing j$ is the feature identification of feature j, and M is the maximum vector size (Lundberg & Lee, 2017).

4 EXPERIMENTAL SETUP

This chapter describes the dataset and procedures used in the experiments during this study. The findings of the exploratory data analysis are presented, as well as the preprocessing steps utilized. In addition, the experimental approach is well defined, including a description of the task being researched. Finally, the programming language and packages used in this work, as well as the techniques of evaluation used, are discussed.

4.1 *Dataset*

The dataset for this study is the Online Shopper Purchasing Intention Dataset created by (Sakar et al., 2018). Each instance in the dataset represents a unique user's intent to finish the transaction. The dataset was structured such that each session belonged to a different user throughout the year, minimizing any bias towards a particular campaign, special day, user profile, or timeframe (Sakar et al., 2018). The dataset consists of 12,330 instances and 18 features. There are ten numerical features such as administrative duration, bounce rate, exit rate, and page value. The remaining eight are categorical features such as visitor type, weekend, month, and revenue. The features of this dataset are explained in detail in Appendix A. 10,422 of the total number of instances in the dataset are negative, indicating that the users did not complete the transaction, while 1908 are positive, indicating that the users completed the transaction. The dataset has an unequal class label due to the few positive transactions relative to the negative transactions, indicating that the dataset is imbalanced. This problem will be discussed in further detail later in this chapter.

4.2 *Preprocessing*

Sakar et al. (2018) cleaned the data before releasing it, including removing missing values. The dataset had no odds. Furthermore, no features are excluded from the dataset in order to determine the feature importance scores that account for all the features. As mentioned in the previous section, the dataset consists of numerical and categorical features. The categorical features are numerically transformed so that these features can be utilized to build and train the models. The Boolean datatypes revenue and weekend are transformed into integer datatypes. The other categorical features month and visitor type used ordinal labeling to establish an ordinal order. Finally, the outliers of the dataset were not removed since these outliers are realistic outliers that represent the few visitors who have visited the administrative pages and informational pages of the website. This indicates that throughout a visit, there is not a significant number of people looking at the website's profile page or the company's contact information.

4.3 *Exploratory Data Analysis*

Despite the increasing popularity of online shopping, the conversion rate has remained unchanged. The conversion rate is the total number of purchases completed (Behera et al., 2020; Liu et al., 2019; Zhou et al., 2019). The conversion rate of this dataset is 15%. This is compared to a variety of features that offer further information about the purchasing intention. The three features visitor type, month, and weekend will be further explained below.

In Figure 2 and Figure 3, the two histograms illustrate the three visitor types that make a purchase or not. Visitor type 0 corresponds to unknown visitors, 1 corresponds to new visitors and 2 corresponds to returning visitors. Figure 2 illustrates that the returning visitors make the most purchases. However, Figure 3 illustrates that even though the majority of purchasing visitors are returning visitors, the conversion rate is higher for the new visitors.

In Figure 4, the two histograms illustrate the months of the year. The histograms only show 10 months since the months of January and April are missing from the dataset. Figure 4 illustrates that the month of November has the highest number of transactions, whereas the month of May has the

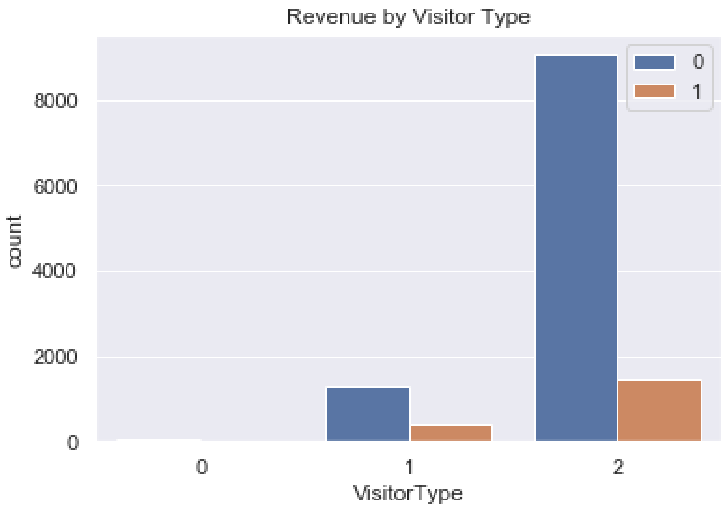

Figure 2: Revenue by Visitor Type

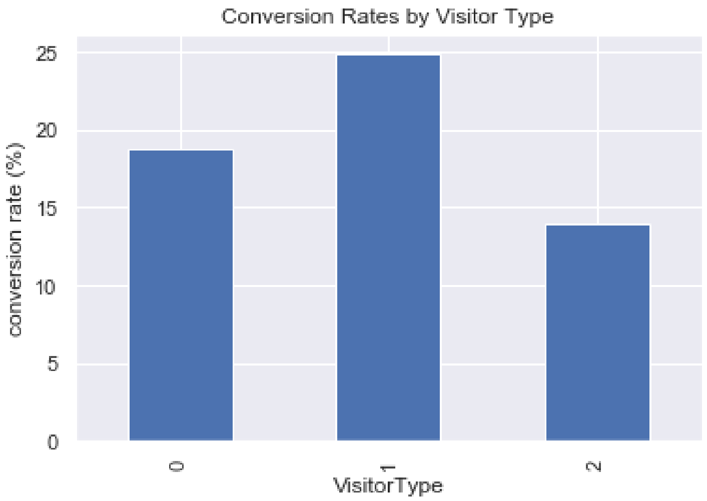

Figure 3: Conversion Rate by Visitor Type

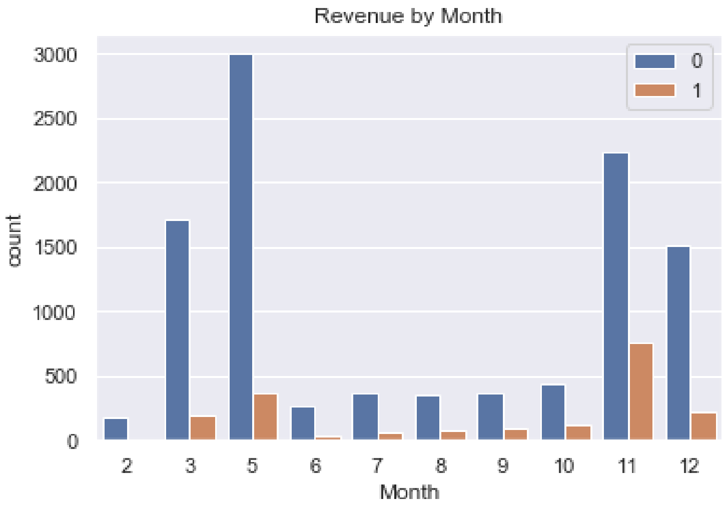

Figure 4: Revenue by Month

most visitors. The second graph of Figure 5 also illustrates that November is the month with the highest conversion rates.

In Figure 6, the two histograms illustrate whether the day of the week is on the weekend or not. Figure 6 illustrates that weekdays have the highest revenue. However, Figure 7 illustrates that even though the visitors are less active on the weekends, the weekends have the highest conversion rate.

4.4 *Experimental procedure*

The purpose of this study is to discover the best model and features for predicting customer purchase intent. As described in the method part of this study, machine learning employs ensemble models and deep learning employs a Multilayer Perceptron. These models are evaluated to determine which model is the most accurate in predicting customer purchase intent. Furthermore, the feature importance scores are used to study which features have the largest impact on predicting customer purchase intent.

4.4.1 *Models*

The models in this study are compared to Logistic Regression, which is the baseline method. The Bagging and Boosting ensemble learning models use the default parameters from Scikit-learn (Pedregosa et al., 2011). The parameters are briefly explained in Appendix B. The first model is Bagged Decision Trees, which has as base estimator the Decision Tree Classifier, and the number of trees is by default 10. Random Forest is the second

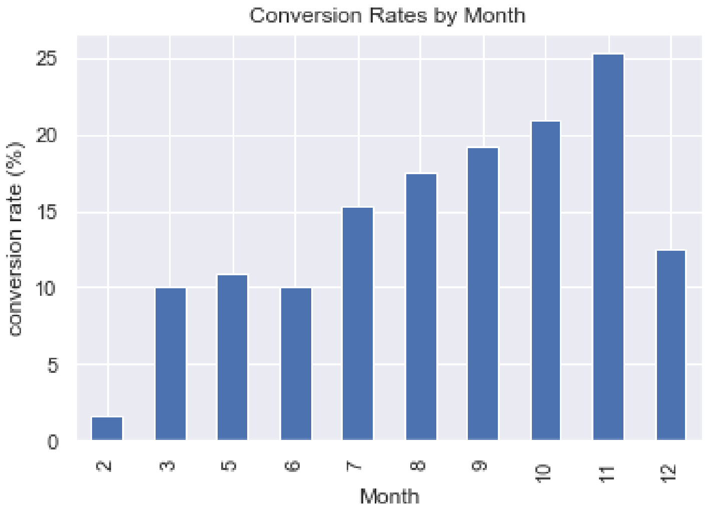

Figure 5: Conversion Rate by Month

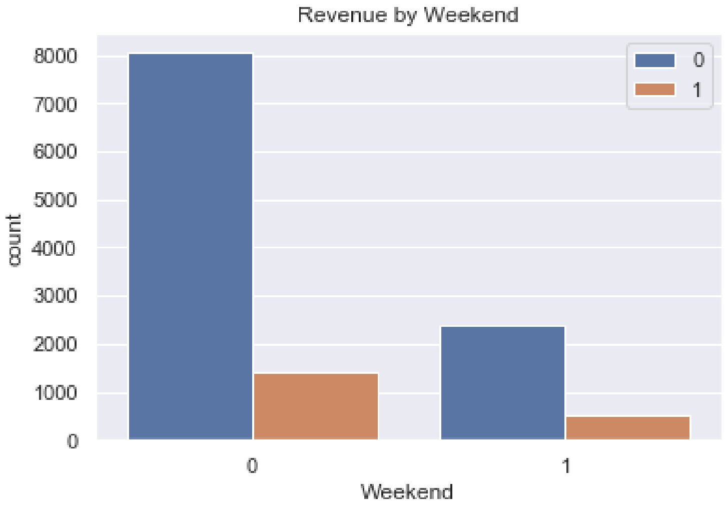

Figure 6: Revenue by Weekend

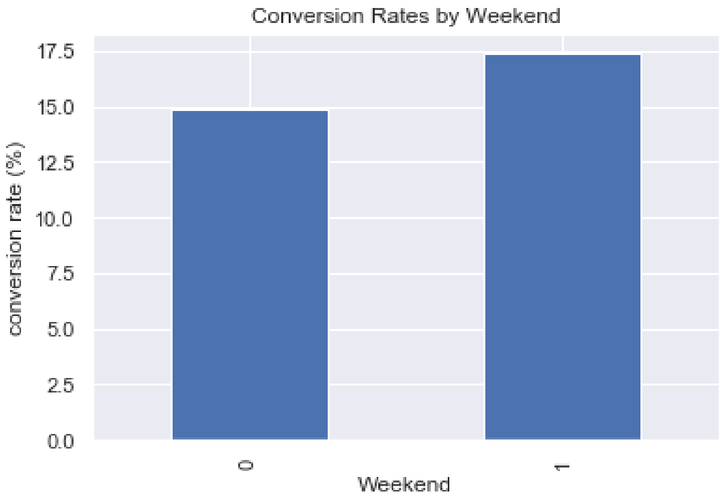

Figure 7: Conversion Rate by Weekend

model, with 100 trees by default and Gini as the function to measure the quality of the split. The third model is AdaBoost, with 50 trees by default and a learning rate of 1. The fourth model is Gradient Boosting, with 100 trees by default. The loss function is deviance, which refers to Logistic Regression, and the learning rate is 0.1.

The final model is Multilayer Perceptron, which is created using the functions of the TensorFlow Keras package (Abadi et al., 2016). The model is built using a sequential model with an input layer of 64 units and Relu as the activation function. The hidden layer has 32 units and Relu as the activation function. The output layer has Sigmoid as the activation function. A dropout layer has been added between the layers to prevent overfitting. Furthermore, Adam is the weight optimizer with a learning rate of 0.01 and the loss function binary cross entropy is used.

A test set of 20% is retained for final evaluation prior to training the data. This study used K-Fold cross-validation with k = 5, which reduces the amount of data required for training. In K-Fold cross validation, the dataset is randomly divided into 5 folds, also known as subgroups, of similar size. The model is trained on $k - 1$ folds, which represent the training set (Berrar, 2018). The remaining fold is used as the validation set. This guarantees that each instance in the dataset appears in both the training and validation sets. The 5-fold was chosen because 20% of the data is usually used as a validation set.

While training the dataset, the models tend to predict class 0 rather than class 1. This is due to the unequal class labels, which result in 10,422 negative instances and 1908 positive instances. To address this imbalanced problem, the oversampling method SMOTE will be used on the training set in each fold. Oversampling is a method to over-sample the minority class, in this case the positive class, so that the minority class is not ignored during prediction (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). The results will be compared to see if the models have improved after oversampling.

After oversampling the training data, the most appropriate hyperparameter values are tuned for each model. This provides the most accurate results using the validation data of each fold. The Bagging and Boosting models used GridSearchCV to accomplish this, which is a method that searches for the optimal combination of parameters to get the most accurate results (Pedregosa et al., 2011). In comparison to GridSearchCV, the deep learning model uses RandomSearchCV to accomplish this. Instead of attempting all possible values, RandomSearchCV samples a specified number of parameter settings from a given distribution (Pedregosa et al., 2011). This approach is useful when the deep learning model has a significant number of parameters and the training period is considerable. The results of various hyperparameter values are compared to see if the tuned models are improved. Finally, the performance of all models is evaluated, and an answer to sub-question 1 is provided based on the results.

4.4.2 *Feature importance*

All the features in the dataset are utilized to identify which feature has the largest impact on predicting customers' purchase intent. The feature importance score might be calculated in a variety of ways depending on the model. However, the same method will be used to compare the findings of the different models. This is the reason that the SHAP values are used to generate the feature importance scores (Lundberg & Lee, 2017). This method has the advantage of being suitable for both machine and deep learning models. Furthermore, this method works effectively with complex models like Gradient Boosting and Multilayer Perceptron (Lundberg et al., 2018; Lundberg & Lee, 2017). The feature importance scores are evaluated, and the results of the different models are compared to determine if any differences occur between these models, the baseline, and the findings from previous research. Furthermore, these findings will provide an answer to sub-question 2.

4.4.3 *Evaluation*

Four classification evaluation measures are used to evaluate the baseline method and the other models. These evaluation measures are Precision, Recall, F1-Score, and Accuracy. Because of the imbalanced dataset, it is crucial to consider the F1-score (Fatourechi et al., 2008).

4.5 *Software*

The models were created in the Jupyter Notebook environment using Python 3.7 (Rossum & Drake, 2009). The Python libraries Pandas and NumPy are used for data preprocessing and analysis (McKinney, 2012; Van Der Walt, Colbert, & Varoquaux, 2011). Scikit-learn was used for feature engineering, modeling the machine models, and evaluations (Pedregosa et al., 2011). Tensorflow Keras was used for modeling the deep learning model (Abadi et al., 2016). Visualizations were made using Seaborn and Matplotlib (Hunter, 2007; Waskom, 2021). The imbalanced dataset is oversampled using the package Imblearn (Chawla et al., 2002). Finally, the SHAP package is used to calculate the feature importance scores (Lundberg et al., 2018; Lundberg & Lee, 2017).

5 RESULTS

This chapter presents the performance of the models described in Section 3 on the Online Shoppers Purchasing Intention Dataset. The evaluation scores are analyzed and compared with the baseline model and with each other to determine the best model. Furthermore, the impact of the features is analyzed using the feature importance scores of SHAP. The results of this method are compared to the baseline model and the previous research findings. This provides a more detailed understanding of the important features of the ensemble learning models and the deep learning model.

5.1 *Results of the models*

In this section, the performance of the ensemble learning models and the deep learning model will be presented. As shown in Table 1, all the models outperform the baseline model Logistic Regression after oversampling. The model Random Forest has the best performance with an accuracy of 89.3%. The other models are Gradient Boosting with an accuracy of 89.1%, Bagged Decision Tree with an accuracy of 88.6%, Multilayer Perceptron with an accuracy of 88.4% and AdaBoost with an accuracy of 87.6%. The comparison of cross-validation and test scores show that the models are not overfitting the data.

Due to the imbalanced dataset, the F1-score is a more accurate evaluation metric to evaluate the performance of the models. The F1-score of the models without SMOTE is quite low, indicating that the models only predict the majority class rather than both classes. Therefore, oversampling was used on the training data. As a result, the minority class has been given more weight and the F1-scores have increased. According to the

Table 1: Results of the models

Evaluating models using k-fold	Accuracy CV	Accuracy Test	$F1$-score CV	$F1$-score Test
Logistic Regression (Baseline)	0.882	0.882	0.484	0.501
Logistic Regression SMOTE	0.868	0.873	0.632	0.642
Bagged Decision Tree	0.901	0.906	0.642	0.657
Bagged Decision Tree SMOTE	0.885	0.886	0.656	0.667
Random Forest	0.901	0.901	0.629	0.623
Random Forest SMOTE	0.890	0.893	0.671	0.688
AdaBoost	0.889	0.888	0.612	0.608
AdaBoost SMOTE	0.871	0.876	0.634	0.653
Gradient Boosting	0.903	0.903	0.651	0.652
Gradient Boosting SMOTE	**0.883**	**0.891**	**0.666**	**0.692**
Multilayer Perceptron	0.885	0.878	0.0386	0.396
Multilayer Perceptron SMOTE	0.878	0.884	0.629	0.666

results of the F1-score, Gradient Boosting outperforms the other models with an F1-score of 69.2%.

Furthermore, rather than using the model's default parameters, the hyperparameters were tuned to find the most optimal hyperparameter values for this study. Appendix C shows the hyperparameters that have been tuned for each model, including the tuned values. The results of the models after hyperparameter tuning are shown in Table 2. The findings show that the baseline model remained unchanged after tuning the hyperparameters. Moreover, as compared to the default parameters, the accuracy and F1-score of Random Forest decreased. This implies that increasing the number of trees has no significant impact on the performance of the Random Forest in this study.

The accuracy of the other models has improved, demonstrating that Gradient Boosting is equivalent to Random Forest in terms of accuracy. However, as previously indicated, the F1-score is the most critical evaluation metric due to the imbalanced dataset. After hyperparameter tuning of Bagged Decision Trees, the accuracy slightly improves but keeps the F1-score unchanged. This is because using a significant number of decision trees has an effect only when the dataset includes a lot of noise or numerous strong predictors (Boehmke & Greenwell, 2019). After tuning the hyperparameters of AdaBoost, using more decision trees significantly improves the F1-score. After tuning the hyperparameters for Gradient Boosting, the F1-score is decreased, indicating that the default parameters have a better number of splits and trees. This enables the model to predict the unbalanced dataset more accurately. After tuning the hyperparameters

Table 2: Hyperparameter results of the models

Evaluating models using k-fold	Accuracy CV	Accuracy Test	$F1$-score CV	$F1$-score Test
Logistic Regression SMOTE	0.868	0.873	0.632	0.642
Logistic Regression	0.868	0.873	0.632	0.642
Bagged Decision Tree SMOTE	0.885	0.886	0.656	0.667
Bagged Decision Tree	0.885	0.888	0.657	0.667
Random Forest SMOTE	0.890	0.893	0.671	0.688
Random Forest	0.891	0.892	0.675	0.686
AdaBoost SMOTE	0.871	0.876	0.634	0.653
AdaBoost	0.876	0.882	0.664	0.685
Gradient Boosting SMOTE	**0.883**	**0.891**	**0.666**	**0.692**
Gradient Boosting	0.890	0.893	0.673	0.669
Multilayer Perceptron SMOTE	0.878	0.884	0.629	0.669
Multilayer Perceptron	0.867	0.877	-	0.675

for Multilayer Perceptron, results of the F1-score are significantly increased. The highest F1-score from Table 2 is evaluated. As a conclusion, Gradient Boosting with the default parameters continues to outperform the other models with an F1-score of 69.2

5.2 *Feature importance scores*

In this section, the feature importance scores are obtained for each model. These scores show which features have a significant impact on predicting a customers' purchase intent. The visualizations of the feature importance scores of each model are presented in Appendix D. The results show that page value has the largest impact on predicting customers' purchase intention. This means that the amount of time a visitor spends on the website has an impact on whether the visitors make a purchase. As reported in prior research, the page values are positively correlated with customers' purchase intent (Shi, 2021).

The importance scores of the other features differ for each model, although there are some similarities between the models. For each model, the feature importance threshold of five features is used. Table 3 lists the five features with the highest feature importance scores for each model. The feature with the highest importance score, occurring in at least five models, is exit rates. According to the data, a lower exit rate has a positive impact on whether a visitor makes a purchase. The exit rate is defined as

Table 3: Features with highest importance scores

Models	The five features with the highest feature importance score
Logistic Regression	page values, month, exit rates, bounce rates, product related
Bagged Decision Tree	page values, product related, exit rates, product related duration, visitor type
Random Forest	page values, month, administrative duration, exit rates, visitor type
AdaBoost	page values, weekend, operating systems, bounce rates, administrative
Gradient Boosting	page values, month, exit rates, visitor type, administrative duration
Multilayer Perceptron	page values, month, exit rates, administrative, product related

the percentage of pages that were broken out of the total number of times the page was seen (Sakar et al., 2018).

Furthermore, the feature with the highest importance score, occurring in at least four models, is the month. The four models are the Baseline model Logistic Regression, Random Forest, Gradient Boosting, and Multilayer Perceptron. This is the third feature that the ensemble learning models and the deep learning model have in common with the baseline model. According to the data, the month in which the visit occurred has an impact on the prediction. The features with the highest importance score, occurring in at least three models, are product-related and visitor type. Whereas the product relevant feature is an important feature for the baseline and deep learning model, it is less important for the Bagging and Boosting ensemble models. According to the data, customers' visits to product-related web pages have a positive impact on their purchases. On the other hand, the feature visitor type is an important feature for the Bagging and Boosting ensemble models. According to the data, the type of visitor who visits the web page has an impact on the prediction.

The final results show that the feature administrative duration is also an important feature for the complex ensemble learning models. According to the data, spending less time on administrative-related web pages, such as the company's profile page, increased customers' purchase intention. Lastly, the feature bounce rate is an important feature for both the baseline model and AdaBoost. According to the data, a lower bounce rate results in an increase in the purchase intent of customers. As reported in prior research, the exit rates and bounce rates are negatively correlated with the customers' purchase intent, which is similar to previous studies (Shi, 2021).

6 DISCUSSION

This chapter evaluates the research findings of the research questions. The goal of the research is defined, and the findings are discussed in detail. Additionally, there are certain limitations to the study that are described. Finally, the contribution of the study is presented.

6.1 *The goal of the research*

The conversion rates of e-commerce websites have remained unchanged even though the use of these websites has increased rapidly. Due to the unchanged conversion rates, the purchase intents of online customers are significantly important to online decision makers. Therefore, the goal of this study is to find the best model and features for predicting the customers' purchase intentions. From a scientific point of view, this study aims to contribute to the field of research by examining the added value of a deep learning model in comparison with ensemble models. According to Kabir et al. (2019), using a deep learning model can increase the accuracy. Furthermore, it is significantly important to understand the impact of the model's features in order to comprehend the studies' predictions (Bugaj et al., 2021). Accordingly, this study fills a gap in the literature by determining the important features of ensemble models and a deep learning model when predicting the customers' purchase intention. As a result, this study addressed the following research question: *To what extent can the customers' purchase intention be predicted using machine and deep learning models?* Two sub-questions are answered to address the main research question, which is discussed in detail in the following sections.

6.2 *Findings of the models*

The first sub-question explores which model performs significantly better when predicting the customers' purchase intentions. In this study, machine learning used Bagging and Boosting ensemble learning algorithms. This is based on the findings that ensemble learning algorithms provide better forecasts in terms of accuracy and generalization (Dong et al., 2019). Bagging and Boosting learning models also address the issue of imbalance by relying on weak classifiers that raise the weight of incorrectly classified examples (Algawiaz et al., 2019). According to Kabir et al. (2019), accuracy can be improved with the use of a deep learning model. That is the reason that Multilayer Perceptron is included in this study, which has been successfully applied to the same dataset (Sakar et al., 2018).

As previously stated, the F1-score is used as an evaluation metric due to the imbalance in the Online Shoppers Purchasing Intention Dataset. The findings indicate that the F1-score of all the models improved, implying that the SMOTE oversampling method contributed to addressing the imbalanced difficulty. The models are compared to the baseline model Logistic Regression. The findings show that each model outperforms the baseline method.

As suggested by Kabir et al. (2019), the accuracy can be improved with the use of a deep learning model. The F1-score of the deep learning model performs better than the baseline model and the Bagged Decision Trees. Unlike the study of Sakar et al. (2018), the deep learning model Multilayer Perceptron does not perform significantly better than Random Forest. Furthermore, the findings of this study show that the Multilayer Perceptron does not perform significantly better than AdaBoost and Gradient Boosting. In this study, Gradient Boosting outperforms the other models with an F1-score of 69.2%. This is consistent with recent studies indicating that Gradient Boosting is the best model for predicting the unbalanced dataset due to the best number of splits and trees (Kabir et al., 2019).

6.3 *Findings of the features*

The second sub-question explores which features have the largest impact on the prediction of the customers' purchase intention. These features are measured using the feature importance scores of Shapley Additive Explanation (SHAP). This method was chosen because it works effectively with complex models like Gradient Boosting and Multilayer Perceptron (Lundberg et al., 2018; Lundberg & Lee, 2017). Moreover, it is significantly important to understand the impact of the features on these models in order to comprehend the predictions for this study Bugaj et al. (2021).

As a result, in both the different types of models and the previous research, the features page value and exit rate are referred to as important features that have a significant impact on the customers' purchase intention (Shi, 2021). The other features that have an impact differ from model to model and have not been studied previously. When the models were compared, a significant number of similarities were discovered. For each model, the feature importance threshold of five features was used in this study. As a result, the feature that the baseline model, ensemble learning models, and deep learning model share in common is the feature month. Future research could predict which month of the year has the largest impact on customers' purchasing intentions.

Furthermore, Random Forest and Gradient Boosting have the same five features as the most important features for predicting the customers' purchase intention. The other two features that these models have in common are visitor type and administrative duration. According to the data, the type of visitor who visits the web page has an impact on the prediction and spending less time on administrative-related web pages, such as the company's profile page, increased customers' purchase intention. Future research could predict which types of visitors have the largest impact on the customers' purchasing intentions. Finally, the findings show that the baseline model and the deep learning model differ by only one feature. The feature that the baseline and deep learning model have in common is product-related. According to the data, visits to product-related web pages have a positive impact on the purchase.

6.4 *Limitations*

The performance of the Multilayer Perceptron could be limited by the small size of the dataset. According to Botalb, Moinuddin, Al-Saggaf, and Ali (2018), the accuracy of the Multilayer Perceptron can be increased by doubling the dataset. Therefore, this limitation can be overcome by the use of a larger dataset to determine whether the accuracy of the deep learning model can be increased. Future research should be conducted to determine this. In this study, the results are still valid because the Multilayer Perceptron outperforms the baseline model.

Another limitation in the current study is that the SHAP method applies the TreeExplainer to the tree-models. This function, on the other hand, does not support the AdaBoost and the Bagged Decision trees, implying that the general KernelExplainer function was used. The difference between the two functions is that the TreeExplainer utilizes the tree structure while the KernelExplainer rejects decision paths due to a lack of data (Lundberg et al., 2018). Future research could overcome this limitation by adjusting the function of TreeExplainer with the information needed for AdaBoost and Bagged Decision Tree. In this study, the results remain valid since a comparison of Bagging and Boosting ensemble learning models between Gradient Boosting and Random Forest could be conducted. As a result, the difference between the baseline model, previous work, ensemble learning models, and a deep learning model can still be determined.

6.5 *Contribution to the field*

The findings of this study contribute to the current field of research by determining the additional value of a deep learning model. The findings show that a deep learning model provides a limited additional value, and the ensemble learning model Gradient Boosting remains the best model for predicting the customers' purchase intent. Gradient Boosting outperforms all other models, including the deep learning Multilayer Perceptron. For this reason, this study shows that it is still advantageous to employ traditional machine learning models for classification rather than depending on deep learning models (Kiki & Houndji, 2020).

Furthermore, the ensemble learning models and the deep learning model are used to find the most important features that have an impact on predicting the customers' purchase intentions. This increases the explainability of the ensemble learning models and the deep learning model used (Bugaj et al., 2021). Previous studies achieved no significant conclusions about the other features than page values, exit rates and bounce rates (Shi, 2021). Furthermore, there were significant distinctions between the important features of each model that had not been discussed previously. Therefore, this study fills a gap in the literature by combining the important features of ensemble models and a deep learning model when predicting the customers' purchase intention. Future research could utilize only the most important features of these models when predicting customers' purchase intentions. This enables researchers to determine if ensemble learning and deep learning models perform significantly better than the baseline model and previous research.

7 CONCLUSION

Given the unchanged conversion rates for e-commerce websites, it is significantly important to understand the purchase intentions of customers. This provides online decision-makers with a better knowledge of the customers' intentions, improving the customer experience and, as a result, increasing conversion rates. This study focused on finding the best model and features for predicting the customers' purchase intentions.

The best model for the prediction is found by comparing different Bagging and Boosting ensemble learning models with a deep learning model. This study examined the additional value of the deep learning model, Multilayer Perceptron. The models are evaluated with the use of an F1-score due to the imbalance in the Online Shoppers Purchasing Intention Dataset. The F1-score of all the models improved after applying oversampling to the training data to solve the imbalanced difficulty. Furthermore, the hyperparameters were tuned to find the most optimal hyperparameter values for this study. The findings indicate that the F1-score of the deep learning model performs better than the baseline model and the Bagged Decision Trees. However, the findings show that Multilayer Perceptron does not perform significantly better than Random Forest, AdaBoost, and Gradient Boosting. In this study, Gradient Boosting outperforms the other models with an F1-score of 69.2%. As a conclusion, Gradient Boosting is the best model for predicting the unbalanced dataset due to the best number of splits and trees (Kabir et al., 2019).

Finally, the features that have the largest impact on predicting customers' purchase intentions are found. The findings show that in both the different types of models and the previous research, the feature page value and exit rate are referred to as important features that have a significant impact on the customers' purchase intention (Shi, 2021)(Shi, 2021). This implies that the time a visitor spends on a web page influences whether they make a purchase, and a lower exit rate has a positive effect on whether a visitor makes a purchase. The other features that have an impact differ from model to model. The features month, visitor type, administrative duration, and product-related are features with high importance scores. The most significant findings indicate that Random Forest and Gradient Boosting have the same five features as the most important features for predicting the customers' purchase intention. Furthermore, the findings show that the baseline model and the deep learning model differ by only one feature.

In conclusion, the findings of this study contribute to the current field of research by determining the additional value of a deep learning model and identifying the significant features of ensemble learning models and a deep learning model in comparison with the baseline model and previous studies. Future research should focus on two aspects. First, future research is required to determine the additional value of a deep learning model used on a larger dataset of customer purchase intentions. Furthermore, the second aspect could utilize only the most important features of these models when predicting the customers' purchase intention. This allows researchers to see whether ensemble learning and deep learning models provide similar or significantly better results than the baseline model and previous research.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., . . . Zheng, X. (2016, may). TensorFlow: A system for large-scale machine learning. *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016*, 265–283. Retrieved from `https://arxiv.org/abs/1605.08695v2` doi: 10.5555/3026877.3026899

Algawiaz, D., Dobbie, G., & Alam, S. (2019, nov). Predicting a User's Purchase Intention Using AdaBoost. *Proceedings of IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering, ISKE 2019*, 324–328. doi: 10.1109/ISKE47853.2019.9170316

Baati, K., & Mohsil, M. (2020). Real-Time Prediction of Online Shoppers' Purchasing Intention Using Random Forest. *Artificial Intelligence Applications and Innovations*, *583*, 43. Retrieved from `/pmc/articles/PMC7256375//pmc/articles/PMC7256375/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7256375/` doi: 10.1007/978-3-030-49161-1_4

Bahad, P., & Saxena, P. (2020). Study of AdaBoost and Gradient Boosting Algorithms for Predictive Analytics. , 235–244. Retrieved from `https://link-springer-com.tilburguniversity.idm.oclc.org/chapter/10.1007/978-981-15-0633-8{_}22` doi: 10.1007/978-981-15-0633-8_22

Behera, R. K., Gunasekaran, A., Gupta, S., Kamboj, S., & Bala, P. K. (2020, mar). Personalized digital marketing recommender engine. *Journal of Retailing and Consumer Services*, *53*. doi: 10.1016/J.JRETCONSER.2019.03.026

Berrar, D. (2018, jan). Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, *1-3*, 542–545. doi: 10.1016/B978-0-12-809633-8.20349-X

Boehmke, B., & Greenwell, B. M. (2019). Hands-on machine learning with R.

Botalb, A., Moinuddin, M., Al-Saggaf, U. M., & Ali, S. S. (2018, nov). Contrasting Convolutional Neural Network (CNN) with Multi-Layer Perceptron (MLP) for Big Data Analysis. *International Conference on Intelligent and Advanced System, ICIAS 2018*. doi: 10.1109/ICIAS.2018.8540626

Breiman, L. (1996). Bagging predictors. *Machine Learning 1996 24:2*, *24*(2), 123–140. Retrieved from `https://link-springer-com.tilburguniversity.idm.oclc.org/article/10.1007/BF00058655` doi: 10.1007/BF00058655

Breiman, L. (2001, oct). Random Forests. *Machine Learning 2001 45:1*, *45*(1), 5–32. Retrieved from `https://link.springer.com/article/`

10.1023/A:1010933404324 doi: 10.1023/A:1010933404324

Bugaj, M., Wrobel, K., & Iwaniec, J. (2021, may). Model Explainability using SHAP Values for LightGBM Predictions. *International Conference on Perspective Technologies and Methods in MEMS Design*, *2021-May*, 102–106. doi: 10.1109/MEMSTECH53091.2021.9468078

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002, jun). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. Retrieved from https://www.jair.org/index.php/jair/article/view/10302 doi: 10.1613/JAIR.953

Ding, B., Qian, H., & Zhou, J. (2018, jul). Activation functions and their characteristics in deep neural networks. *Proceedings of the 30th Chinese Control and Decision Conference, CCDC 2018*, 1836–1841. doi: 10.1109/CCDC.2018.8407425

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2019, aug). A survey on ensemble learning. *Frontiers of Computer Science 2019 14:2*, *14*(2), 241–258. Retrieved from https://link-springer-com.tilburguniversity.idm.oclc.org/article/10.1007/s11704-019-8208-z doi: 10.1007/S11704-019-8208-Z

Esmeli, R., Bader-El-Den, M., & Abdullahi, H. (2020, dec). Towards early purchase intention prediction in online session based retailing systems. *Electronic Markets 2020*, 1–19. Retrieved from https://link-springer-com.tilburguniversity.idm.oclc.org/article/10.1007/s12525-020-00448-x doi: 10.1007/S12525-020-00448-X

Fatourechi, M., Ward, R. K., Mason, S. G., Huggins, J., Schlögl, A., & Birch, G. E. (2008). Comparison of evaluation metrics in classification applications with imbalanced datasets. *Proceedings - 7th International Conference on Machine Learning and Applications, ICMLA 2008*, 777–782. doi: 10.1109/ICMLA.2008.34

Freund, Y., & Schapire, R. E. (1997, aug). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. doi: 10.1006/JCSS.1997.1504

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. doi: 10.1214/AOS/1013203451

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. New York, NY: Springer New York. Retrieved from http://link.springer.com/10.1007/978-0-387-84858-7 doi: 10.1007/978-0-387-84858-7

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in*

Science and Engineering, *9*(3), 90–95. doi: 10.1109/MCSE.2007.55

Janocha, K., & Czarnecki, W. M. (2017, feb). On Loss Functions for Deep Neural Networks in Classification. *Schedae Informaticae*, *25*, 49–59. Retrieved from https://arxiv.org/abs/1702.05659v1 doi: 10.4467/20838476SI.16.004.6185

Kabir, M. R., Ashraf, F. B., & Ajwad, R. (2019, dec). Analysis of different predicting model for online shoppers' purchase intention from empirical data. *2019 22nd International Conference on Computer and Information Technology, ICCIT 2019*. doi: 10.1109/ICCIT48885.2019.9038521

Kiki, Y., & Houndji, V. R. (2020). Prediction of the Purchase Intention of Users on E-Commerce Platforms using Gradient Boosting. *International Journal of Engineering and Advanced Technology (IJEAT)*, 2249–8958. doi: 10.35940/ijeat.A1929.1010120

Linoff, G. S., & Berry, M. J. A. (2004). *Data mining techniques: for marketing, sales, and customer relationship management*. John Wiley & Sons, Inc.

Liu, X., Lee, D., & Srinivasan, K. (2019, aug). Large-Scale Cross-Category Analysis of Consumer Review Content on Sales Conversion Leveraging Deep Learning:. *https://doi-org.tilburguniversity.idm.oclc.org/10.1177/0022243719866690*, *56*(6), 918–943. Retrieved from https://journals-sagepub-com.tilburguniversity.idm.oclc.org/doi/10.1177/0022243719866690 doi: 10.1177/0022243719866690

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2018, feb). Consistent Individualized Feature Attribution for Tree Ensembles. Retrieved from https://arxiv.org/abs/1802.03888v3

Lundberg, S. M., & Lee, S. I. (2017, may). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *2017-Decem*, 4766–4775. Retrieved from https://arxiv.org/abs/1705.07874v2

Martínez, A., Schmuck, C., Pereverzyev, S., Pirker, C., & Haltmeier, M. (2020, mar). A machine learning framework for customer purchase prediction in the non-contractual setting. *European Journal of Operational Research*, *281*(3), 588–596. doi: 10.1016/J.EJOR.2018.04.034

McKinney, W. (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. Retrieved from https://books.google.nl/books?hl=en{&}lr={&}id=v3n4{_}AK8vu0C{&}oi=fnd{&}pg=PR3{&}dq=McKinney+2012+pandas{&}ots=rhGM4iAtnA{&}sig=7ujo-Jg1XabESLO9wKATDEplxXw

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. Retrieved from http://jmlr.org/papers/v12/pedregosa11a.html

Ramchoun, H., Amine, M., Idrissi, J., Ghanou, Y., & Ettaouil, M. (2016). Multilayer Perceptron: Architecture Optimization and Training. , *4*(1), 26. doi: 10.9781/IJIMAI.2016.415

Rossum, G., & Drake, F. (2009). *Python 3 Reference Manual*.

Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2018, may). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications 2018 31:10*, *31*(10), 6893–6908. Retrieved from https://link-springer-com.tilburguniversity.idm.oclc.org/article/10.1007/s00521-018-3523-0 doi: 10.1007/S00521-018-3523-0

Schapire, R. E., Freund, Y., Bartlett, P., & Lee, W. S. (1998, oct). Boosting the margin: a new explanation for the effectiveness of voting methods. *https://doi.org/10.1214/aos/1024691352*, *26*(5), 1651–1686. Retrieved from https://projecteuclid.org/journals/annals-of-statistics/volume-26/issue-5/Boosting-the-margin--a-new-explanation-for-the-effectiveness/10.1214/aos/1024691352.fullhttps://projecteuclid.org/journals/annals-of-statistics/volume-26/issue-5/Boosting-the-margi doi: 10.1214/AOS/1024691352

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, *20*(1), 3–29. doi: 10.1177/1536867X20909688

Shi, X. (2021, may). The Application of Machine Learning in Online Purchasing Intention Prediction. *ICBDC 2021: 2021 6th International Conference on Big Data and Computing*, 21–29. Retrieved from https://doi.org/10.1145/3469968.3469972 doi: 10.1145/3469968.3469972

Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. *IEEE Access*, *7*, 53040–53065. doi: 10.1109/ACCESS.2019.2912200

Sridhar, S., Mootha, S., & Kolagati, S. (2020, jul). A University Admission Prediction System using Stacked Ensemble Learning. *Proceedings - 2020 Advanced Computing and Communication Technologies for High Performance Applications, ACCTHPA 2020*, 162–167. doi: 10.1109/ACCTHPA49271.2020.9213205

Syarif, I., Zaluska, E., Prugel-Bennett, A., & Wills, G. (2012). Application of Bagging, Boosting and Stacking to Intrusion Detection. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7376 LNAI*, 593–602. Retrieved from https://link-springer-com.tilburguniversity.idm.oclc.org/chapter/10.1007/978-3-642-31537-4_46 doi: 10.1007/978-3-642-31537-4_46

Taud, H., & Mas, J. (2018). Multilayer Perceptron (MLP). , 451–455. Retrieved from https://link-springer-com.tilburguniversity.idm.oclc.org/chapter/10.1007/978-3-319-60801-3{_}27 doi: 10.1007/978-3-319-60801-3_27

Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011, feb). The NumPy array: a structure for efficient numerical computation. *Computing in Science and Engineering*, *13*(2), 22–30. Retrieved from http://arxiv.org/abs/1102.1523http://dx.doi.org/10.1109/MCSE.2011.37 doi: 10.1109/MCSE.2011.37

Wang, R. (2012, jan). AdaBoost for Feature Selection, Classification and Its Relation with SVM, A Review. *Physics Procedia*, *25*, 800–807. doi: 10.1016/J.PHPRO.2012.03.160

Waskom, M. (2021, apr). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. doi: 10.21105/JOSS.03021

Zhou, Y., Mishra, S., Gligorijevic, J., Bhatia, T., & Bhamidipati, N. (2019, jul). Understanding consumer journey using attention based recurrent neural networks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 3102–3111. Retrieved from https://doi.org/10.1145/3292500.3330753 doi: 10.1145/3292500.3330753

Appendix A: Features of the dataset

Table 4: Features of the dataset

Feature	Feature Description	Type
Administrative	The visitors of web pages related to account management such as a profile page.	Numerical
Administrative Duration	Time spent on account management-related pages by the visitor.	Numerical
Informational	The visits of web pages related to information about the website such as their address or contact information.	Numerical
Informational Duration	Time spent on information-related pages by the visitor.	Numerical
Product Related	The visits of web pages related to products.	Numerical
Product Related Duration	Time spent on product-related pages by the visitor.	Numerical
Bounce Rate	Percentage of visitors who enter the site from that page and then leave.	Numerical
Exit Rate	The proportion of pages that were broken out of the total number of times the page was seen.	Numerical
Page Value	The average value of a web page that a user visited before completing a transaction.	Numerical
Special Day	The closeness of the visiting time to a specific special day.	Numerical
Operating Systems	The operating system used by the visitor.	Categorical
Browser	The browser used by the visitor.	Categorical
Region	The geographic region from which the visitor begins the session.	Categorical
Traffic Type	The traffic type that sent the visitor to the website.	Categorical
Visitor Type	The type of visitor to the website, which could be a new visitor, a returning visitor, or others.	Categorical
Weekend	Whether or not the day is in the weekend.	Categorical
Month	The month in which the visit occurred.	Categorical
Revenue	Whether a visit to the website led to a purchase.	Categorical

Appendix B: Briefly a description of the hyperparameters

Table 5: Description of the hyperparameters

Models	Hyperparameters (Pedregosa et al., 2011)
N estimators	Total number of decision trees in the model.
Solver	This option specifies which optimization algorithm should be used.
Penalty	This option is used to indicate the penalization norm (L1 or L2) (regularization).
C	It is the inverse of regularization strength, which must be always a positive float.
Max features	Train each decision tree with the maximum number of features.
Learning rate	Each decision tree is given a certain amount of weight. The contribution of each decision tree grows as the rate rises.
Subsample (Gradient Boosting)	The percentage of samples that will be utilized to fit individual base learners. Stochastic Gradient Boosting occurs when the value is less than 1.0.
Max depth	The number of nodes in the tree is limited by the maximum depth.
Hidden layer sizes	The number of neurons in the hidden layer is represented by this value.
Max iter	Maximum number of iterations.
Activation	The activation function for each layer.
Alpha	L2 penalty

Appendix C: The tuned hyperparameters for each model

Table 6: The tuned hyperparameters for each model

Models	Hyperparameters
Logistic Regression	solver = lbfgs, penalty = L2, and C = 10
Bagged Decision Tree	number of estimators = 50
Random Forest	number of estimators = 1000, max features = sqrt
AdaBoost	number of estimators = 100, learning rate = 0.1
Gradient Boosting	number of estimators = 500, learning rate = 0.1, subsample = 1, max depth = 9
Multilayer Perceptron	hidden layer one = 128, hidden layer two = 32, dropout 0.3, batch size = 40, epoches = 50, learning rate = 0.001

Appendix D: Visualization of the feature importance score of each model

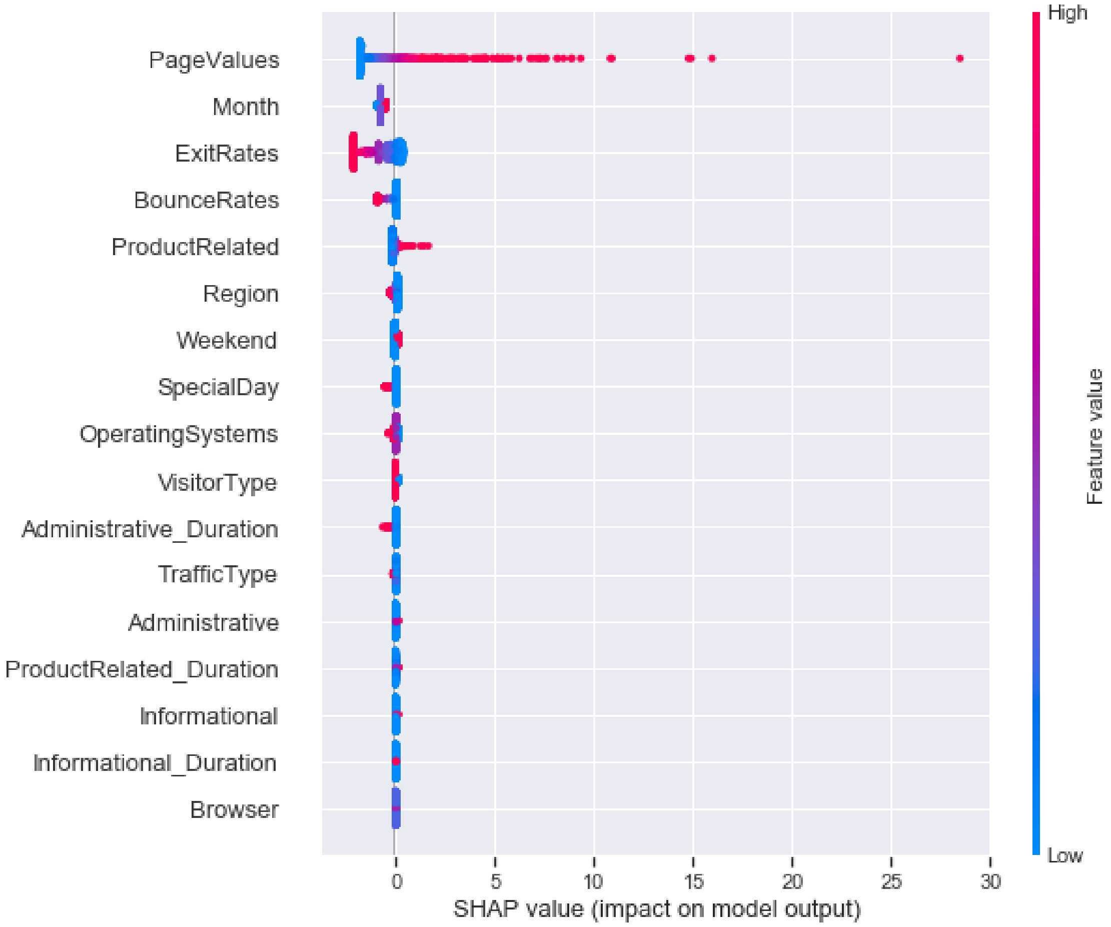

Figure 8: Feature importance score of Logistic Regression

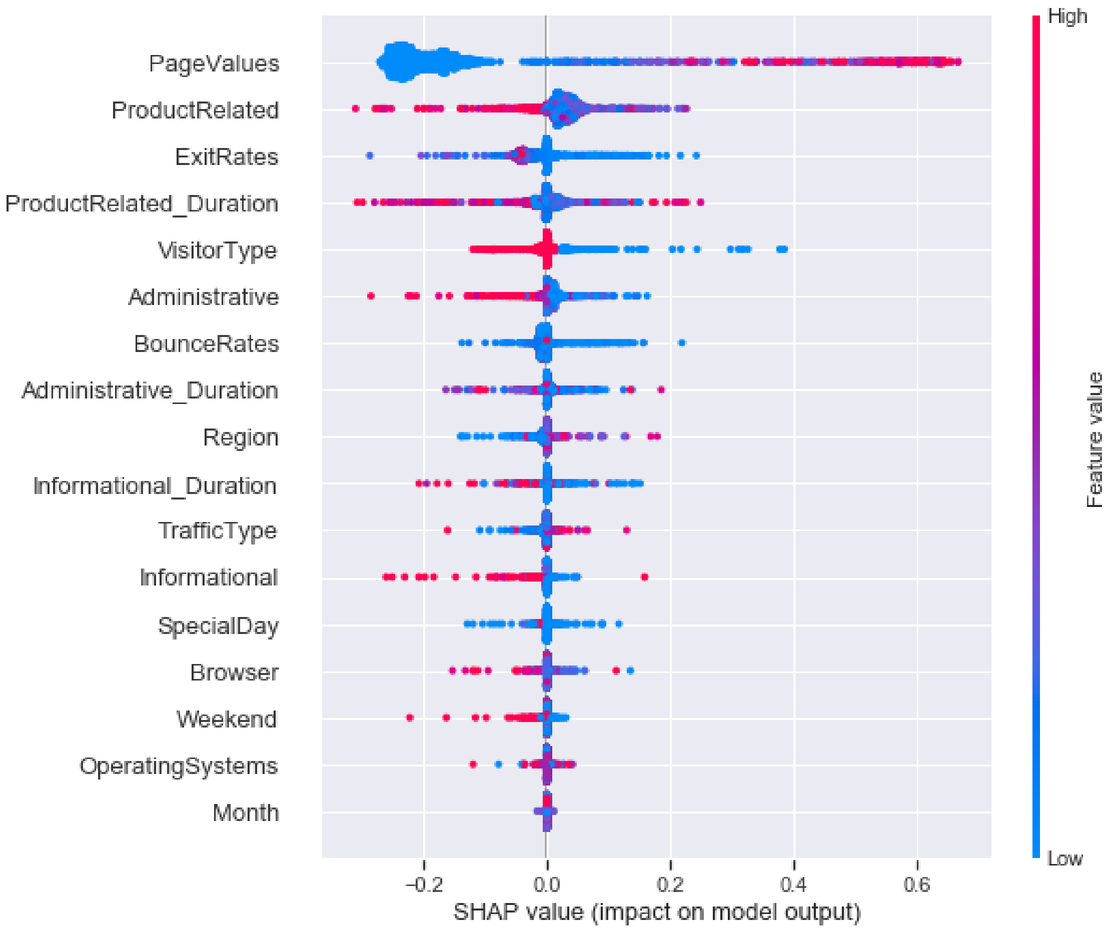

Figure 9: Feature importance score of Bagged Decision Tree

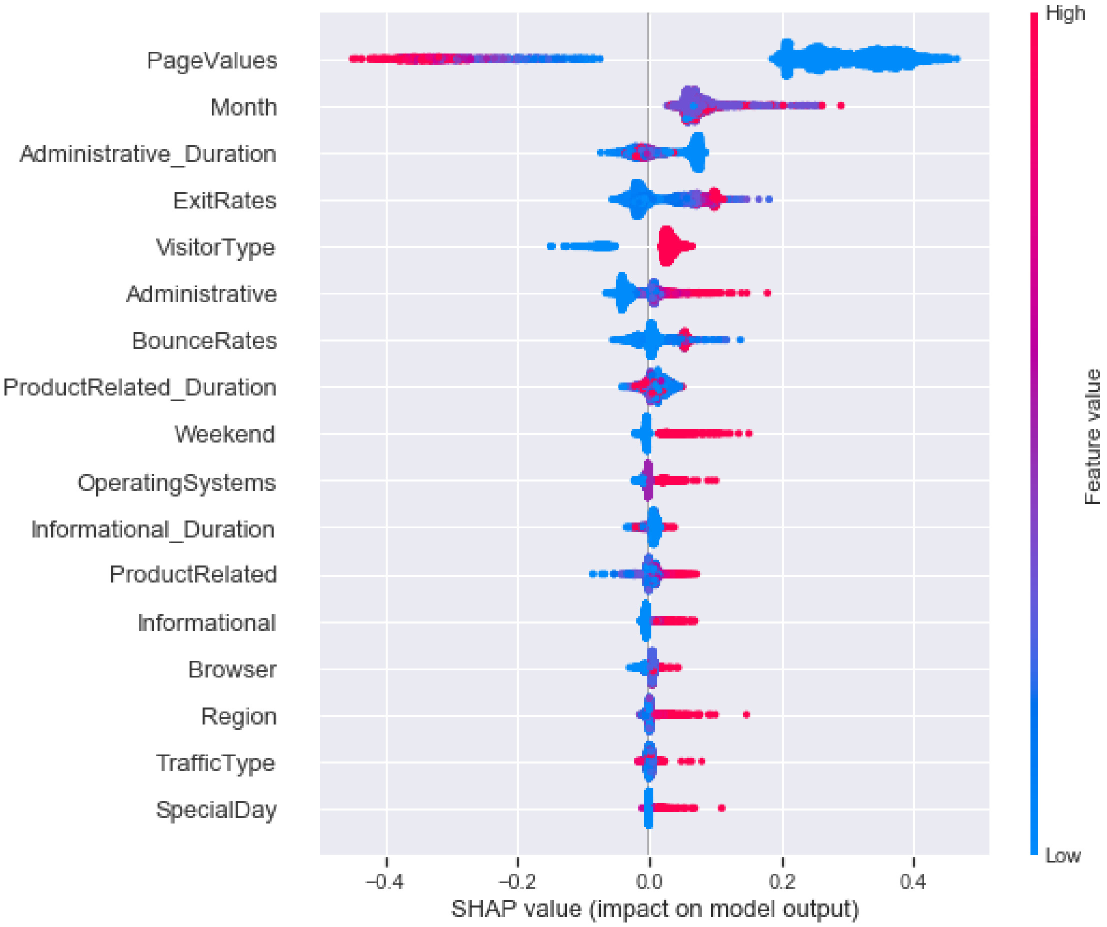

Figure 10: Feature importance score of Random Forest

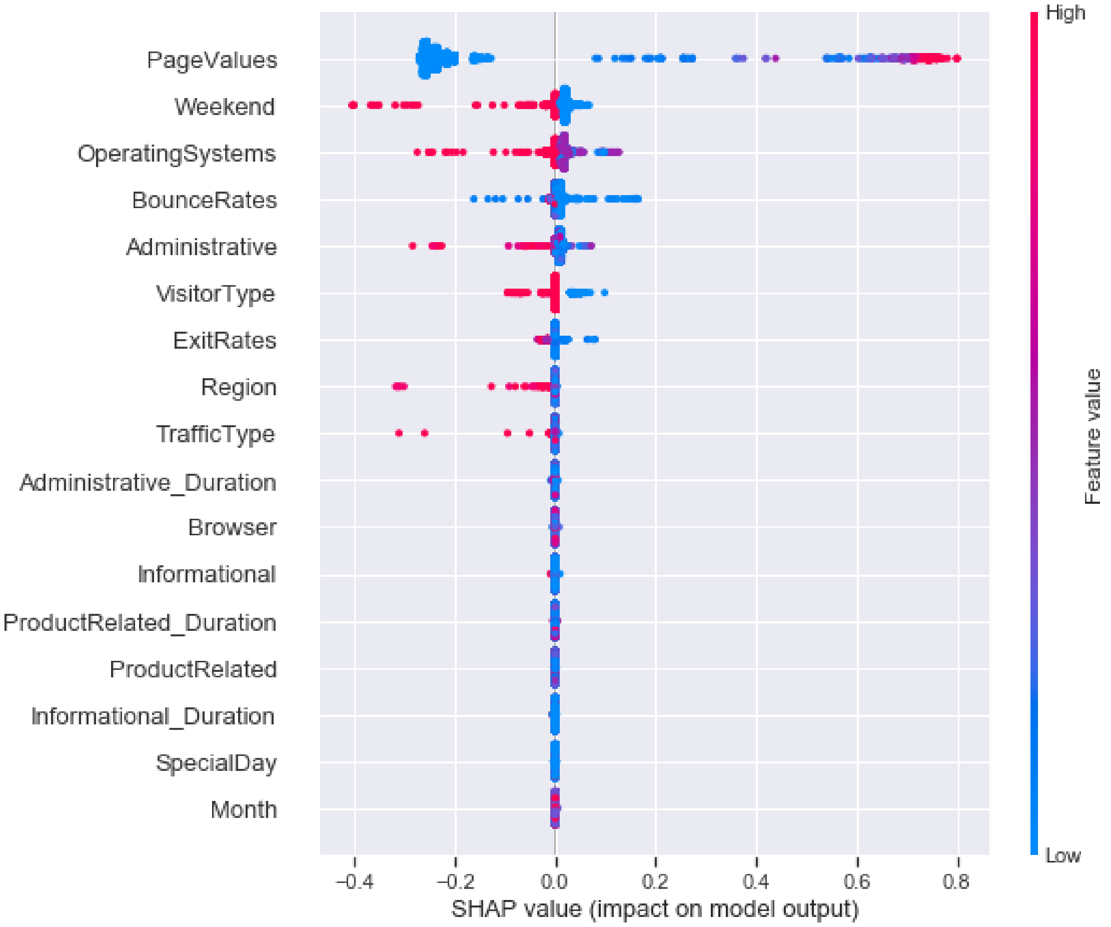

Figure 11: Feature importance score of AdaBoost

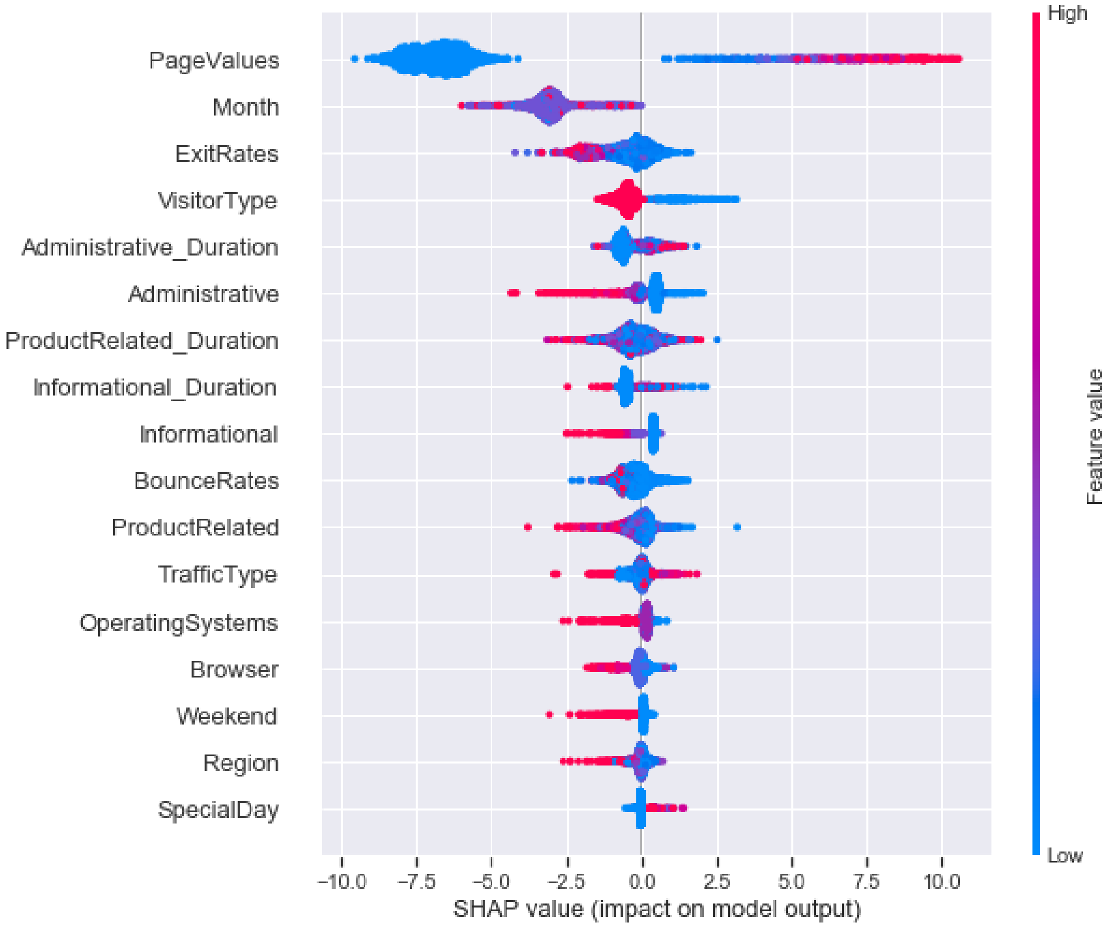

Figure 12: Feature importance score of Gradient Boosting

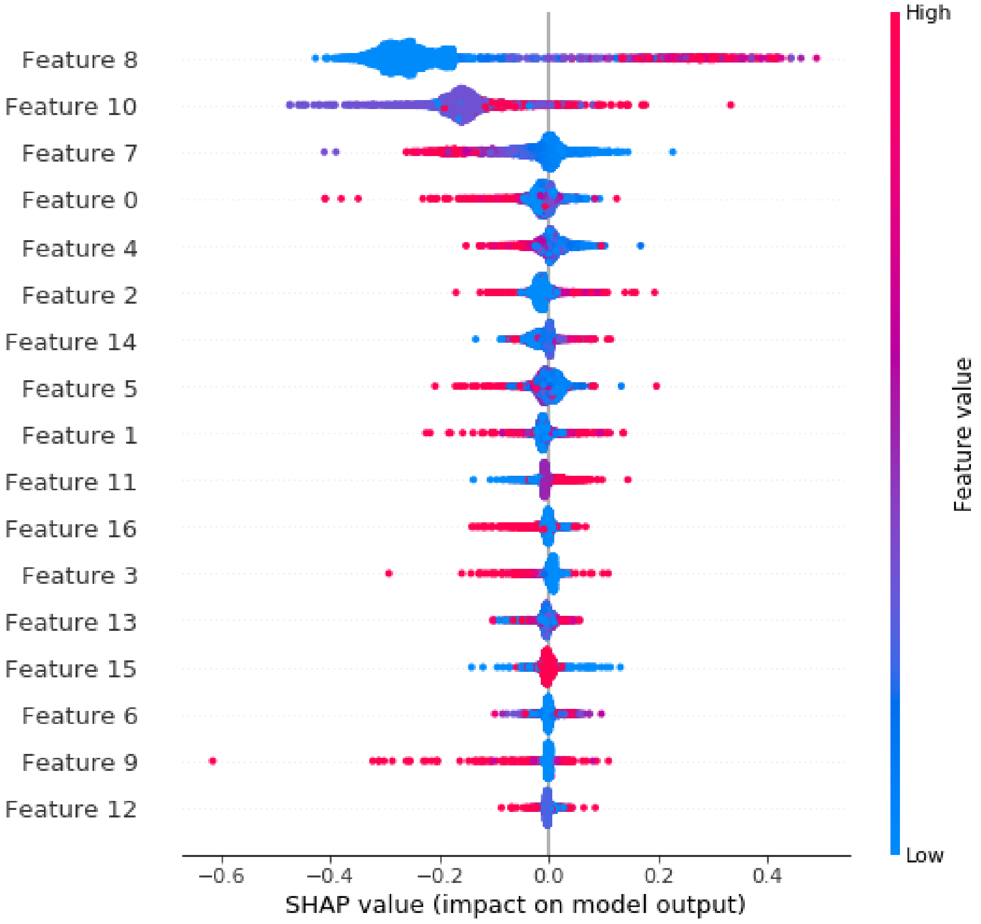

Figure 13: Feature importance score of Multilayer Perceptron