Network for Studies on Pensions, Aging and Retirement

NETSPAR ACADEMIC SERIES

Classification of imbalanced data sets with sampling methods for predicting online shopping intentions

Mustafa Bulca

MSc 01/2020-017

Classification of imbalanced data sets with sampling methods for predicting online shopping intentions

Mustafa Bulca

Student number: 2033710

THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE OR DATA SCIENCE & SOCIETY

DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE

SCHOOL OF HUMANITIES AND DIGITAL SCIENCES

TILBURG UNIVERSITY

Thesis committee:

Prof. Dr. Eric Postma

Dr. Henry Brighton

Tilburg University

School of Humanities and Digital Sciences

Department of Cognitive Science & Artificial Intelligence

Tilburg, the Netherlands

January 2020

Preface

As part of the Master's study Data Science & Society Business track of Tilburg University, this presented thesis is the final assignment to complete my Master's degree. You will read my last assignment where the knowledge and skills that I developed during my Master's period are combined into the last project. I can summarize this last period as a student as a period with a lot of pressure and independent work. I am very happy that I decided to study Data Science at Tilburg University. I developed valuable skills and networked with other students. In the end, I can say that I am proud of my work and I enjoyed my study career!

To fulfill the last assignment I had several persons that helped me with succeeding in this project. Firstly, I want to thank Prof. Dr. Eric Postma for supervising the thesis project and keeping me on track with the critical questions. Second, I want to thank my co-students who are supervised by Prof. Dr. Eric Postma, for sharing questions and difficulties while doing research and writing their thesis. Seeing where fellow students get stuck helped with making decisions and solving problems. Lastly, I want to thank my family, girlfriend, and friends who had the patients to live with me during this busy time.

Thank you!

Mustafa Bulca

Tilburg, December 2020

Abstract

With the increasing use of the internet, online shopping is getting more popular and these developments are opening new possibilities in the area of data science. Data from visitors that visit online webshops are more often collected for prediction purposes. One of the important properties of online webshop data is that the data is imbalanced. Due to the ease of gaining information and comparing products on the internet, data sets are getting imbalanced. Visits of online webshops are more often ending without a buy of the visit. Handling imbalanced data sets in machine learning is an important part when building classifiers. Several algorithms handle imbalanced data sets and these methods can be split into two categories. The first is algorithm-based approaches where the focus is mainly on improving the algorithm to enhance prediction performance. The second is sampling-based approaches where the data set is oversampled or downsampled for training classifiers. The data used in this study is obtained from the UCI machine learning repository site and this data set is previously used in the study of Sakar et al. (2018), where the researchers are randomly oversampling the data set. This study broadens the study of Sakar et al. (2018), by applying different sampling-methods to the same classifiers. In this research, sampling-based approaches are used to enhance the prediction of online shopping intentions. The data set used in this study is randomly downsampled, oversampled with the SMOTE algorithm, and a combination of both methods is used with a decision tree, support vector machine, and multilayer perceptron. The results are showing that the proposed combination of downsampling and the SMOTE algorithm is not outperforming the SMOTE algorithm. Further research on the combination of two sampling methods is needed to further develop or adapt this combination in other data sets and machine learning classifiers.

Contents

1.0 Introduction

Nowadays, online shopping is becoming increasingly popular and 40% of all internet users buy products on the internet (Rubin, Martins, Ilyuk, & Hildebrand, 2020). The growth of online shopping increased the possibilities for webshop owners to analyze their visitors. There is a large number of studies regarding online shopping behavior. One important main domain is predicting online shopping behavior. In these studies, data of visitors are used to predict whether a visitor will buy a certain product or not. The increase in online shopping and tools for analysis provides new opportunities in the area of data science.

Visitors of webshops are monitored by, for example, the duration of page visit, exit rate, landing page, bounce rate, and demographics. Data of visitors are used for descriptive and predictive modeling. With these studies, researchers have to deal with imbalanced data sets. Data of online webshops are moreover visitors that did not buy products. Because of the ease of looking for products on the internet, people use online webshops to gain information. Besides, people can compare the same kinds of products on other webshops. Information gain and comparing products on webshops causes imbalanced data sets that can make predictions of online shopping behavior hard, especially in the minority class.

An imbalanced data set can produce prediction problems for the minority class while scoring high on accuracy in the overall model (Khalilpour Darzi, Niaki, & Khedmati, 2019). Besides, misclassification in the minority class can have higher costs than misclassification in the majority class. In the case of this study, classifying a buyer as a not buyer produces more costs than classifying a not buyer as a buyer because potential buyers are missed. Solving the prediction problem of imbalanced data sets will improve the accuracy rate on the minority class and there will be less bias towards the minority class in predictions.

To improve the prediction power of imbalanced data sets, several algorithms are developed to deal with this problem. The common way to address these approaches is algorithm-based approaches and sampling-based approaches. The algorithm-based approach focuses more on improving the algorithm to increase predicting power. The sampling-based approach focuses on oversampling the minority class to the majority class or downsampling the majority class to the minority class. Oversampling is copying or creating new instances whereas downsampling eliminates instances (Liu, An, & Huang, 2006).

A common sampling-based approach for balancing data sets is oversampling the minority class. However, with random oversampling of the minority class, the model is sensitive to overfit the training data and not generalize well on unseen data. The synthetic minority oversampling technique (SMOTE) is

a sampling-based approach that oversamples the minority class by creating new observations into the training instances (Sun, Li, Fujita, Fu, & Ai, 2020). The model does not copy instances and creates new instances by using the nearest neighbors of the instances in the training set.

1.1 Aim of the study

Previous research of Sakar et al. (2018), experimented with online shopping intentions for webshops. For this prediction task, a decision tree, support vector machine, and a multilayer perceptron are compared. The study concludes that the multilayer perceptron produces the best prediction performance. The researchers compared the models with a data set of 10,422 of the negative classes and 1,908 positive classes. While experimenting with shopping intentions, the researchers faced the problem of an imbalanced data set. The researchers randomly oversampled the minority class to improve the prediction performance of the models. Finally, after randomly oversampling the data set, still, the multilayer perceptron produced the best performance. This study aims to improve the prediction performance of the study of Sakar et al. (2018) by applying sampling methods to the data set and creating a balanced data set that improves the prediction performance of a decision tree, support vector machine, and a multilayer perceptron. In this way, models can be more accurate in predicting the minority class. In this study, the results of the algorithms used in the study of Sakar et al. (2018) are used as baseline results for this study.

1.2 Scope of research

The scope of this research is to experiment with which sampling method fits best in which machine learning algorithm of the study of Sakar et al. (2018). This is done by experimenting with downsampling the data set and oversampling the data set with the SMOTE algorithm. Besides, a new approach to handling imbalanced data sets is proposed. In this research, there will be a study of the effect of combining the SMOTE algorithm with downsampling the data set. Furthermore, this research focuses on imbalanced data sets within the scope of online webshop data. Due to the limited time, besides a decision tree, support vector machine, and multilayer perceptron, no other machine learning classifiers are tested.

1.3 Theoretical and scientific relevance

From a theoretical perspective, this study contributes to experiments about predicting the online shopping intentions of webshop visitors with imbalanced data sets. With the increasing use of the internet, potential customers will more often compare products on the internet and gain information about products. Besides comparing and gaining information, there will be more people buying online rather than physically in stores (Constantinides & Holleschovsky, 2016). This study provides online webshop tools to gain more insight into their customers. In this way, owners of webshops are better able to fulfill the needs of a potential customer to end the visit to the webshop with a buy.

From a scientific perspective, this study broadens the study of Sakar et al. (2018) of the problem of imbalanced data sets. Working with imbalanced data sets is an important part of machine learning tasks. When building prediction models for classification tasks, facing an imbalanced data set often occurs. Machine learning classifiers are optimizing the overall accuracy of the prediction model and considering the importance of the minority class the same as the majority class, which leads to wrong predictions in the minority class (Ibrahim et al., 2019). This study emphasizes the importance of making the right predictions in the minority class. Furthermore, this study contributes to existing scientific research by experimenting with a new approach to handling imbalanced data sets by combining downsampling with a synthetic oversampling method.

1.4 Research question

By improving the prediction performance of the decision tree, support vector machine, and multilayer perceptron, used in the study of Sakar et al. (2018), this study aims to gain more insight into sampling methods that handle the problems of imbalanced data sets. Furthermore, in this study, there will be a new sampling approach to improve prediction performance. Therefore, the following research question is formulated:

To what extent can sampling methods enhance the prediction of online shopping intention?

2.0 Theoretical framework

In this section, there will be a description of the sampling methods to deal with imbalanced data sets and a brief overview of the machine learning algorithms used in this study. Furthermore, there will be an overview of previous scientific research.

2.1 Downsampling

In a binary classification task, the prediction value consists of two different values. The imbalance in data sets occurs when the amount of one class is more and when this difference biases the model results negatively. However, to train the model in a way that the model predicts correct instances in the minority class, downsampling the data is one of the methods. With downsampling the data set, instances from the majority class are randomly downsampled until there is a balanced data set. The downsampling is applied to the training data set, the test set is still imbalanced to generalize well on unseen data. Downsampling the data set is an effective solution to prevent the model from overfitting the training data set (Rustogi & Prasad, 2019). One big drawback of downsampling the training data set is that, when downsampling a large number of instances, many data are not used for training the model. Machine learning algorithms work well with more data, downsampling can lead to poor prediction performance on the test data (Barros et al., 2019).

2.2 SMOTE oversampling

Another sampling-based method to deal with imbalanced data sets is oversampling the minority class, the opposite of downsampling the majority class. However, with oversampling the minority class, another problem occurs. The oversampled instances from the minority class are affecting the algorithm by overfitting the minority class. The SMOTE algorithm is an extension to oversampling data in the minority class. The major advantage of SMOTE is that it oversamples the minority class, so in this way, no useful data is lost because of downsampling (Fernandez et al., 2018). As with downsampling, the SMOTE algorithm is applied to the training data set. The test seat data is used to test the trained model for the results of the model. In this way, the generalizability of the model can be ensured. Machine learning models are generally working better with a lot of data. Therefore, another advantage of the SMOTE algorithm is that it creates new training instances for the model. Classification algorithms usually perform poorly with fewer data (Barros et al., 2019).

The process that the SMOTE algorithm follows is first determining how many new instances to create as an integer value. After this value is determined, the algorithm randomly takes one instance from the minority class and searches by default to the five nearest neighbors. Concerning the nearest neighbors, a new training instance is created. The new training instance is then multiplied by a random factor between 0 and 1. Finally, this instance is added to the training data set. This is done repeatedly till the integer value of instances that was needed is reached (Rodriguez-Torres, Carrasco-Ochoa, & Martínez-Trinidad, 2019).

2.3 Combining downsampling and SMOTE

A new approach to handling imbalanced data sets proposed in this study is combining the SMOTE algorithm with downsampling. As mentioned in section 2.1, machine learning algorithms perform better with more data. However, when oversampling the data, the model can overfit the data because the created instances are based on the minority class. To avoid these problems and benefit from the advantages of the two sampling methods, a combination of the methods will experiment with. Therefore, in this way, there will be fewer data oversampled rather than only applying SMOTE and there will be less loss of data rather than only downsampling the data set.

2.4 Machine learning algorithms

In this section, there will be a brief introduction to the algorithms used in this research. First, one of the popular machine learning algorithms is the decision tree. The reason why decision trees are popular is that they are easy in understanding and decision trees do not need much computational power. Besides, when developing a decision tree, it is not necessary to set parameters (Han, Pei, & Kamber, 2011). In the case of this study, it would be interesting how decision trees perform on data that is balanced after downsampling or applying the SMOTE algorithm. The decision tree is an algorithm that works well with balanced data, this results in a balanced tree where the splitting of information is nearly equally divided.

The second algorithm used in the research is the support vector machine. Support vector machines are reliable classifiers with strong mathematical substantiation. Like the decision tree algorithm, support vector machines perform well on balanced data sets. When training a support vector machine on an imbalanced data set, the algorithm tends to bias towards the majority class and produces incorrect predictions in the minority class (Batuwita, & Palade, 2010). Regarding this study, studying how the support vector machine performs after handling the data set with downsampling and oversampling with the SMOTE algorithm can produce interesting results.

The last classifier used in this study is the multilayer perceptron. The multilayer perceptron is a more advanced classifier than the decision tree and support vector machine. Especially in extremely complex classification tasks, the multilayer perceptron performs very well (Oh, 2011). Multilayer perceptrons are widely used in multiclass classification tasks and are performing well on complex data (Lin et al., 2013). The multilayer perceptron learns his weights and biases while trying to make correct predictions. When the model makes mistakes in the predictions, it propagates back the information that the output is incorrect. In this way, weights and biases are trained for the model. Furthermore, the multilayer perceptron exists of different layers with different weights, and this makes the model good in predicting complex data. In the case of this study, the multilayer perceptron could produce good results because of the size of layers and neurons. Besides, the multilayer perceptron is a strong binary classifier that can lower bias in the majority class and make better predictions in the minority class (Díaz-Vico et al., 2018).

2.5 Related work

To address the problem of imbalanced data sets in predicting online shopping behavior on online webshops, there will be an overview of studies regarding imbalanced data sets that examined and experimented with methods against imbalanced data sets. In machine learning, making predictions in the minority class is hard, the imbalanced data sets bias the machine learning models in making correct predictions in the minority class (Haixiang et al., 2017).

In research to machine learning, there have been several studies regarding the problem of imbalanced data sets. Important to mention is that machine learning models are not just affected by the imbalanced data sets. Previous research of Zhang & Trubey (2018), studied the effects of imbalanced data sets on five different machine learning classifiers. In their study, the researchers made use of an artificial neural network, decision tree, support vector machine, logistic regression, and a random forest. The models are tested with a data set on money laundering detection. In the study, detecting the minority class was important, as those are the instances with a higher cost rate. The researchers concluded that the artificial neural network produced consistent outcomes and outperformed the other four classifiers. The decision tree had the lowest predicting scores. Furthermore, the researchers concluded that ANN is less affected by imbalanced data sets. In the former study of Zhang & Trubey (2018), the researchers did not make use of any sampling method that balances the data set.

Other previous research on imbalanced data sets studied randomly downsampling the data set in different proportions and feature selection for training models (Hasanin et al., 2019). The researchers

used a random forest, gradient boosted tree, and logistic regression to examine the two different methods that were tested on two different data sets. The main goal of the researchers was to yield high scores of correct predictions in the minority class. The result of the randomly downsampled experiment is interesting for this study, as random downsampling is used in this study for different machine learning classifiers. By randomly downsampling the data set, the gradient boosted tree yields the best results. The gradient boosted tree made more correct predictions in the minority class.

Another previous research about methods for imbalanced data sets is conducted by Ramezankhani et al. (2016). In their study, the researchers developed machine learning classifiers for predicting diabetes at an earlier stage. In this particular study, the researchers made use of the SMOTE algorithm to balance the data set for training the models. In total, there were three different models tested; probabilistic neural network, decision tree, and a naive bayes algorithm. The researchers used different proportions of synthetically oversampling, with a start from 100% step-wise to 700% with 100% per step. The researchers are concluding that the models, with a proportioned size of 700% of oversampling, are increasing in terms of sensitivity, with the probabilistic neural network as strongly increasing. However, the models are decreasing in terms of accuracy scores. Therefore, the models are predicting more correct minority instances and less correct majority instances.

Furthermore, there has been previous research on how to evaluate models that are trained and tested on highly imbalanced data sets. The study of Picek et al., (2019), emphasizes the evaluation metrics used for models affected by highly imbalanced data sets. In the study, the researchers experiment with different methods that deal with imbalanced data sets. The data set has been used for machine learning algorithms, to predict rare events. The researchers are concluding that the accuracy metric is distorted. In machine learning, and especially when dealing with imbalanced data sets, the goal of the model is important for choosing evaluation metrics. If the model is built for achieving high predicting the minority class, precision and recall are better metrics to evaluate the models. Besides, the study of Ramezankhani et al. (2016) shows an increase in incorrect predictions in the minority class while decreasing the overall accuracy of the model. Because the costs of wrong predictions in the minority class are higher, this is the class where the model has to predict the most correct instances. However, after applying sampling algorithms, the model should perform better in terms of accuracy.

3.0 Methods

In this section, first, the data set is described and from where it was retrieved. This is followed by preprocessing steps to analyze the data and train the predictive models. Finally, the evaluation steps used in this study are presented. Especially with an imbalanced data set, and the purpose of this research, evaluating models with the correct evaluation metrics is important. Due to the nature of imbalanced data sets, evaluation metrics can present a biased result by presenting high scores.

3.1 Data set

The data set used in this study is retrieved from the UCI machine learning repository. In total, the data set consists of 12,330 instances with 18 variables and all are collected in a time frame of 1 year. Furthermore, the data set is uploaded on the UCI machine learning repository website on 31-08-2018. The data that is collected from one specific webshop and to avoid copies of instances, all instances are belonging to a specific visitor of the online webshop. The target variable in the data set is a binary variable with two values (buyer versus not buyer). In total 10,422(84.53%) were negative class(not buyer) and 1,908(15.47%) were positive class(buyer). The other 17 variables are independent variables consisting of ten numerical and seven categorical variables. A detailed table of the variables is described in the appendix (Appendix A). The algorithms and data set used in this study can be retrieved via the link in the appendix (Appendix B).

3.2 Software

To complete this experiment, Python 3.6 is used as the programming language with coding in Jupyter Notebooks. In Python NumPy and pandas are used for data manipulation and exploring the data set. Scikit-learn is used to build decision trees, support vector machines, and a multilayer perceptron. Also, splitting the data set into a train-test set, and evaluating the models, is done with Scikit-learn. Furthermore, to handle imbalanced data sets, the package imbalance-learn is used for implementing downsampling and the SMOTE algorithm.

3.3 Preprocessing data

In this section, there is a detailed explanation of the preprocessing steps before training the classifiers. Since the data set was previously used in another research, the data set did not need many preprocessing steps. The rationale for choosing the same data set is to compare the result of the models in this study with the results of the models in the study of Sakar et al. (2018).

The first step in preprocessing the data set was checking the data set for possible NaN values. As mentioned before, as a result of using a data set that is used in previous research, the data set did not contain any NaN values.

The second important preprocessing step is checking for outliers. To train the prediction models and apply the algorithms that create new instances like SMOTE, it is important to identify and process outliers (e.g. in the case of this data set, a value of 60,000 seconds of a page visit). For identifying the outliers, the interquartile range of the continuous variables is calculated. While doing this, the instances with values in variables of the lower 5 percent and upper 5 percent are removed from the data set. As a result of removing outliers, the data set consisted of 11,743 instances and 18 variables. The target variable (buyer versus not buyer) in the cleaned data set consisted of 10,036 (85.46%) instances in the negative class (not buyer) and 1,707 (14.54%) in the positive class (buyer).

After removing outliers, two variables are converted from string values to integer values. The variable ‘Month’ is converted to an equal number of the month. The variable ‘VisitorType’ is converted from ‘Returning_Visitor’ to 0, ‘New_Visitor’ to 1, and ‘Other’ to 2. All other variables were noted as integers except for the variables ‘Weekend’ and ‘Revenue’, who were booleans (True or False).

3.4 Sampling the imbalanced data set

Finally, before training the machine learning models, the data set is downsampled and oversampled. The models are trained on three different downsampled data sets. Proportionally to the minority class, the majority class is downsampled 0.25, 0.5, 0.75, and 1.0 times the ratio of the minority class. Also, the models trained on the data set after applying the SMOTE algorithm, are trained on three different data sets. Proportionally to the minority class, the data sets are oversampled 0.25, 0.5, 0.75, and 1.0 times the ratio of the majority class. Finally, the combination of downsampling and applying the SMOTE algorithm to the data set was done. First oversampling the minority class to 50% of the majority class by applying the SMOTE algorithm. After, downsampling the majority class to the same amount of instances in the oversampled minority class

4.0 Experimental setup

In the following section, the experimental setup of the three machine learning classifiers is explained in detail. There will be a description of the approach of building and optimizing the models, and evaluation metrics of the models. In total, three different classifiers are built to predict potential shoppers for an online webshop. The classifiers are trained on the imbalanced data set, oversampled data set, and downsampled data sets. The oversampled and downsampled data sets are sampled in four different proportions. Finally, a combination of an oversampled and downsampled data set is used to train the three classifiers. In total, 30 unique test results are obtained with the three classifiers.

4.1 Train-test split

For training the models, a train-test split of 75% training data and 25% testing data is chosen. The total training data set consisted of 11,743 instances. The training data set consisted of 8,807 instances and the testing data set consisted of 2,936 instances. The reason for choosing a test set of 25% is to generalize well on unseen data. Especially with imbalanced data sets, too small test sets lower the minority class instances in the test set. The models are trained on the downsampled and oversampled train data sets and tested on the test set. The rationale for this is to enhance the generalizability of the model on unseen data. In this way, the prediction task that the model has to perform on unseen data stays imbalanced, and therefore the test data set is not processed.

4.2 Cross-validation

To prevent the models from overfitting, cross-validation is used. In the models, the training data set is split into 10 equal disjoint data sets. These data sets are all used to train and test the model on the validation set. After training and testing these disjoint data sets, the scores are summed and divided by 10, the number of data sets.

4.3 Decision tree

After preprocessing the data set, converting the variables, and splitting the data set into train and test sets, the first decision tree is trained on the imbalanced data set. The majority class of the imbalanced data set consisted of 7,516 instances while the minority class consisted of 1,291 instances. The decision tree is built with default parameters and optimized with Grid Search. Within this model, the following parameters are set: criterion="gini", max_depth=5.

After the first decision tree was trained, the second decision tree is trained with a downsampled data set. Table 1 presents the minority class and majority class instances for the models with different proportions of downsampling the training data set. The decision tree is built with default parameters and optimized with Grid Search. Within this model, the following parameters are set: criterion="entropy", max_depth=10.

Proportion of downsampling	Minority class instances	Majority class instances
0.25	1,291	5,164
0.50	1,291	2,582
0.75	1,291	1,721
1.0	1,291	1,291

Table 1: Decision tree instances in classes with different sampling proportions after downsampling.

The third decision three is trained on the data set after applying the SMOTE algorithm to the minority class. Table 2 is presenting the minority class and majority class instances with different sampling proportions in the training data set. The decision tree is built with default parameters and optimized with Grid Search. Within this model, the following parameters are set: criterion="gini", max_depth=10.

Proportion of oversampling	Minority class instances	Majority class instances
0.25	1,891	7,516
0.50	3,758	7,516
0.75	5,637	7,516
1.0	7,516	7,516

Table 2: Decision tree instances in classes with different sampling proportions after applying SMOTE.

Finally, the last decision three is trained on a data set that is first oversampled with the SMOTE algorithm to 50% of the majority class. After oversampling, the majority class is decreased to the same value as the oversampled minority class. The minority class increased to 3,758 and the majority class decreased to 3,758. The decision tree is built with default parameters and optimized with Grid Search. Within this model, the following parameters are set="entropy", max_depth=10.

4.4 Support vector machine

The first support vector machine is trained on the imbalanced data set. The majority class of the imbalanced training data set consisted of 7,516 instances while the minority class consisted of 1,291 instances. The support vector machine is built with default parameters and optimized with Grid Search. Within this model the following parameters are set: max_iter=1,000, alpha=0.1, loss="modified_huber", penalty="l1".

After the first support vector machine was trained, the second support vector machine is trained with a downsampled data set. Table 3 presents the minority class and majority class for the models with different proportions of downsampling in the training data set. The support vector machine is built with default parameters and optimized with Grid Search. Within this model the following parameters are changed to: max_iter=1,000, alpha = 0.1, loss="modified_huber", penalty="l1".

Proportion	Minority class instances	Majority class instances
0.25	1,291	5,164
0.50	1,291	2,582
0.75	1,291	1,721
1.0	1,291	1,291

Table 3: Support vector machine instances in classes with different sampling proportions after downsampling.

The third support vector machine is trained on the data set after applying the SMOTE algorithm to the minority class. Table 4 presents the different proportions of sampling within the training set for the classifier. The support vector machine is built with default parameters and optimized with Grid Search. Within this model the following parameters are set: max_iter=1,000, alpha = 0.1, loss="modified_huber", penalty="l1".

Proportion	Minority class instances	Majority class instances
0.25	1,291	7,516
0.50	3,758	7,516
0.75	5,637	7,516
1.0	7,516	7,516

Table 4: Support vector machine instances in classes with different sampling proportions after applying SMOTE.

Finally, the last support vector machine is trained on a data set that is first oversampled with the SMOTE algorithm to 50% of the majority class. After oversampling, the majority class is decreased to the same value as the minority class. The minority class increased to 3,758 and the majority class decreased to 3,758. The support vector machine is built with default parameters and optimized with Grid Search. Within this model the following parameters are set: max_iter=1,000, alpha = 0.0001, loss="modified_huber", penalty="l1".

4.5 Multilayer perceptron

The first multilayer perceptron is trained on the imbalanced data set. The majority class of the imbalanced data set consisted of 7,516 instances while the minority class consisted of 1,291 instances. The multilayer perceptron is built with default parameters and optimized with Grid Search. Within this model the following parameters are set: max_iter=100, activation = "relu", alpha= 0.05, hidden_layer_sizes=(20,), learning_rate='constant', solver='adam'.

After the first multilayer perceptron was trained, the second multilayer perceptron is trained with a downsampled data set. Table 5 presents the minority class and majority class for the models with different proportions of downsampling in the training data set. The multilayer perceptron is built with default parameters and optimized with Grid Search. Within this model the following parameters are set: max_iter=100, activation = "relu", alpha= 0.05, hidden_layer_sizes=(10, 30, 10), learning_rate='adaptive', solver='adam'.

Proportion	Minority class instances	Majority class instances
0.25	1,291	5,164
0.50	1,291	2,582
0.75	1,291	1,721
1.0	1,291	1,291

Table 5: Multilayer perceptron instances in classes with different sampling proportions after downsampling.

The third multilayer perceptron is trained on the data set after applying the SMOTE algorithm to the minority class. Table 6 presents the minority class and majority class for the models with different proportions of downsampling in the training data set. The multilayer perceptron is built with default parameters and optimized with Grid Search. Within this model the following parameters are set: max_iter=100, activation = "relu", alpha= 0.05, hidden_layer_sizes=(20,), learning_rate='adaptive', solver='adam'.

Proportion	Minority class instances	Majority class instances
0.25	1,879	7,516
0.50	3,758	7,516
0.75	5,637	7,516
1.0	7,516	7,516

Table 6: Multilayer perceptron instances in classes with different sampling proportions after applying SMOTE.

Finally, the last multilayer perceptron is trained on a data set that is first oversampled with the SMOTE algorithm to 50% of the majority class. After oversampling, the majority class is decreased to the same value as the minority class. The minority class increased to 3,758 and the majority class decreased to 3,758. The multilayer perceptron is built with default parameters and optimized with Grid Search. Within this model the following parameters are set: max_iter=100, activation = "tanh", alpha= 0.0001, hidden_layer_sizes=(20,), learning_rate='adaptive', solver='adam'.

4.6 Evaluation metrics

Choosing reliable evaluation metrics is important when dealing with imbalanced data sets because some metrics can provide good results while scoring low on the minority class. When choosing evaluation metrics, it is important to keep in mind that the metrics are presenting correct scores. Sakar et al. (2018) used the confusion table metrics to analyze the results. In our comparative evaluation, the same metrics will be adopted.

The accuracy of the model will display the overall prediction performance. When dealing with imbalanced data sets, the accuracy will be biased towards the majority class of the model. The model will perform high on accuracy because it will predict correctly the majority class. In this study, the accuracy scores will be used to compare it with the baseline study and to base the accuracy score of the model if all instances are predicted in the majority class. Precision will measure the correct positive classifications that were positive. Recall (true positive rate) will measure the positive rate of the instances. In the case of this study, the true positive rate reflects instances that are correctly classified in the majority class. Sensitivity (true negative rate) will measure the proportion of correct predictions in the minority class. Finally, the F1-score will present the harmonic mean of precision and recall and will provide a better overall measure than accuracy. Besides, the correct number of predictions in the minority class is included in the results to see whether this is increasing or decreasing.

5.0 Results

This study aims to improve the prediction performance of the models used in the study of Sakar et al. (2018) with sampling methods that handle imbalanced data sets. In the experiments, there has been using oversampling and downsampling of the data set in different proportions to improve prediction performance. In the next sections, the results of the decision tree, support vector machine, and multilayer perceptron are presented with the different sampling proportions.

When fitting models on imbalanced data sets, the accuracy score is high if the algorithm predicts all instances as the majority class. The data set used in this study contains 85.46% instances in the majority class and 14.54% instances in the minority class. If the models predict the outcome all as the majority class, the models will achieve an accuracy score of 85.46%. Thus, if accuracy is used as a metric to evaluate the model, the model has to perform better than 85.46%.

5.1 Baseline

The results of Sakar et al. (2018) provide our baseline. In their study, the researchers only made use of random oversampling to obtain better predictive performance on the machine learning classifiers. Table 7 is showing the baseline for this study. The decision tree classifier obtained the best accuracy score. However, the multilayer perceptron outperformed the decision tree in terms of F1-score and true-negative rate. The reason why the multilayer perceptron is obtaining a lower accuracy score is that the model is predicting more majority class instances wring, and more minority class instances right than the decision tree. The researchers are therefore concluding that the multilayer perceptron is the best performing model after randomly oversampling the data set

Baseline	**Accuracy (%)**	**F1-score**	**True-negative rate**	**True-positive rate**
Decision tree	88.92%	0.57	0.57	0.96
Support vector machine	88.25%	0.52	0.42	0.97
Multilayer perceptron	87.92%	0.58	0.58	0.96

Table 7: Baseline Sakar et al. (2018)

5.2 Decision tree

Table 8 shows the results of the decision tree classifier in terms of accuracy (%), precision, recall, F1-score, and the number of true negatives. The results are obtained on the imbalanced data set, downsampled training data set, applying SMOTE to the training data set, and a combination of these two methods on the training data set. Besides, the proportion of downsampling or oversampling is presented. The results are showing that downsampling the data set is not improving the accuracy score of the model. Furthermore, the proportion of downsampling affects the F1-score of the model. When the data set is downsampled to the amount of the minority class, the amount of true negatives increases. However, The F1 score is decreasing, which means that the model is predicting more wrong positive classes. With applying SMOTE to the data set, accuracy, precision, the F1-score, and the true negatives do not change a lot. However, with a higher proportion of oversampling, the true negatives are increasing which means that the model makes more correct predictions in the minority class. After applying the SMOTE algorithm and downsampling the data set, the decision tree is scoring better than downsampling. The classifier is also obtaining a higher number of true negatives concerning only applying the SMOTE algorithm to the data set.

Decision tree	**The proportion of oversampling or downsampling**	**Accuracy (%)**	**Precision**	**Recall**	**F1 Score**	**True Negatives**
Imbalanced data set	1.0	90.12%	0.93	0.95	0.94	247
Downsampling	0.25	81.00%	0.95	0.82	0.88	309
Downsampling	0.50	65.60%	0.97	0.62	0.75	375
Downsampling	0.75	59.95%	0.98	0.54	0.70	391
Downsampling	1.0	52.96%	0.98	0.46	0.63	394
SMOTE	0.25	88.62%	0.92	0.95	0.93	220
SMOTE	0.50	88.56%	0.93	0.93	0.93	252
SMOTE	0.75	88.08%	0.94	0.92	0.93	274
SMOTE	1.0	88.86%	0.94	0.93	0.93	277
SMOTE & downsampling	1.0	79.33%	0.96	0.79	0.87	343

Table 8: Results decision tree

5.2 Support vector machine

In table 9, the results of the support vector machine are presented. With the models trained on the imbalanced data set, the support vector machine obtains a high accuracy and F1-score. Furthermore, downsampling the data set is negatively affecting the model. With increasing the sampling proportion to the minority class, accuracy and F1-score is decreasing. The amount of correct classifications in the minority class is increasing. However, the model makes more mistakes in predicting the majority class right. After applying the SMOTE algorithm to the data set, the support vector machine performs stable within the range of sampled proportions. Furthermore, the correct predicted instances in the minority class also stay stable. By combining the two sampling methods, accuracy and F1-score are decreasing. However, the correct predictions in the minority class are higher than applying just the SMOTE algorithm to the data set. Thus, the model makes more mistakes in predictions in the majority class.

Support vector machine	**The proportion of oversampling or downsampling**	**Accuracy (%)**	**Precision**	**Recall**	**F1 Score**	**True negatives**
Imbalanced data set	1.0	89.65%	0.95	0.93	0.94	281
Downsampling	0.25	72.31%	0.94	0.72	0.82	300
Downsampling	0.50	54.22%	0.96	0.49	0.65	365
Downsampling	0.75	49.32%	0.96	0.43	0.59	369
Downsampling	1.0	47.07%	0.98	0.39	0.56	391
SMOTE	0.25	89.41%	0.96	0.92	0.94	308
SMOTE	0.50	88.83%	0.96	0.91	0.93	319
SMOTE	0.75	88.96%	0.95	0.92	0.93	303
SMOTE	1.0	89.00%	0.96	0.91	0.93	318
SMOTE & downsampling	1.0	65.77%	0.96	0.63	0.76	354

Table 9: Results support vector machine

5.3 Multilayer perceptron

In table 10, the results of downsampling the training set, applying SMOTE to the training set, and combining the SMOTE algorithm with downsampling are presented. The multilayer perceptron performs well on accuracy and F1-score trained on the imbalanced data set. However, the model is not predicting a lot correctly in the minority class. The imbalanced data set is affecting the predictions made in the minority class. After downsampling the data set, there is an increase in predictions in the minority class. However, there is also a decrease in accuracy and F1-score with the increase of sampling proportions. With the SMOTE algorithm, the model performs stable in accuracy and F1-scores with different sampling proportions. Furthermore, the model is predicting more minority classes correct with increasing proportions of sampling. The combination of the SMOTE algorithm with downsampling produces a low accuracy score. However, the model predicts more minority instances correctly than only applying the SMOTE algorithm.

Multilayer perceptron	**The proportion of oversampling or downsampling**	**Accuracy (%)**	**Precision**	**Recall**	**F1 Score**	**True negatives**
Imbalanced data set	1.0	89.41%	0.91	0.98	0.94	166
Downsampling	0.25	75.75%	0.96	0.75	0.84	335
Downsampling	0.50	60.59%	0.97	0.56	0.71	373
Downsampling	0.75	50.95%	0.98	0.44	0.61	390
Downsampling	1.0	57.02%	0.98	0.51	0.67	383
SMOTE	0.25	89.58%	0.91	0.98	0.94	164
SMOTE	0.50	89.37%	0.94	0.94	0.94	265
SMOTE	0.75	88.96%	0.94	0.93	0.94	278
SMOTE	1.0	85.59%	0.96	0.87	0.91	313
SMOTE & downsampling	1.0	78.54%	0.97	0.78	0.86	345

Table 10: Results multilayer perceptron

6.0 Discussion

This research aimed to enhance the prediction of online shopping behavior with the use of sampling methods to handle the imbalanced data set. In this study, there have been using a decision tree, support vector machine, and a multilayer perceptron to predict if a visit to an online webshop ends with a buy. To handle the problem of the imbalanced data set, the data set is randomly downsampled, oversampled with the SMOTE algorithm and a combination of the two methods are used.

In contrast to the research of Zhang & Trubey (2018), in this research, the decision tree classifier performed best with no methods that deal with imbalanced data sets. The decision tree classifier produced the best scores in terms of recall and F1-score. However, the model did not outperform the support vector machine in the number of correct predictions in the minority class. This result should be taken into account when not applying sampling-methods that deal with an imbalanced data set.

With downsampling the data set, it is clear that the models are performing better with many data. After downsampling the training data set, the results are showing a decrease in the overall model performance. This result is in line with the findings of Barros et al. (2019), where the researchers are concluding that machine learning algorithms are performing better with the increasing amounts of data. However, the correct predictions in the minority class are increasing with downsampling the proportion of the majority class to the minority class. This result is also in line with the result of Hasanin et al. (2019), where the researchers obtained more correct predictions in the minority class. However, the researchers did not state how this affects the predictions in the majority class. The increase of correct predictions in the minority class arises because after downsampling the model is less biased towards the majority class.

After applying the SMOTE algorithm to the data set, the models showed that by creating synthetic oversampled instances, the overall model performance is improved for the support vector machine and the multilayer perceptron. These results indicate that the models with the SMOTE algorithm are still biased towards the majority class instances. The amount of correct predictions in the minority class stays lower than downsampling the data set. The results obtained in this study are contrary to the results of Ramezankhani et al. (2016) where the researchers concluded that the SMOTE algorithm obtained the best results in handling imbalanced data sets. By creating new instances based on the nearest neighbors of the minority class, the algorithm is adding noise to the model that prevents the model from overfitting the training data (Chawla et al., 2002). Oversampling the minority class is not

reducing the bias towards the majority class. Therefore, the predictions in the majority class are not decreasing, and overall, the model is still scoring high.

The proposed approach of combining the SMOTE algorithm with downsampling is not improving the prediction performance of the models. The results show that the combined version of the sampling method is resulting in a balanced performance between correct predictions in the minority class and the overall performance of the model. These results are in line with the results of Barros et al. (2019), where the researchers conclude the loss of data affects the model.

It is important to mention that machine learning algorithms are influenced by other factors besides class imbalance. Factors like the size of the data set, choosing the predictor variables, and the quality of the data. In this research, the main goal was to improve the result of Sakar et al. (2018) with different sampling methods. Applying SMOTE, downsampling, or combining these sampling methods did not improve the results of Sakar et al. (2018). In their study, the researchers are randomly oversampling the data set, which causes more bias towards the oversampled data set. Therefore, the models are overall performing well with the oversampled data set.

This research contributes to scientific research by combining two sampling methods that handle imbalanced data sets. Overall, this research is showing that combining sampling methods can fit different purposes of handling imbalanced data sets. To achieve high overall scores or achieving more correct instances in the minority class, the proportion of oversampling of downsampling can be adjusted. The performed experiments are showing that there is still a need for improvement in the study of machine learning with imbalanced data sets.

6.1 limitations & further research

The main limitation of this study was experimenting with different proportions for the combinations of applying sampling methods. The combinations of downsampling and the SMOTE algorithm is tested by creating instances to 50% of the majority class and after downsampling the majority class to the same amount of instances as the oversampled minority class. Further research could experiment with different proportions of oversampling and downsampling. Another limitation due to the limit of time and the focus of previous research was the focus on three classifiers. Further research can investigate the prediction performance of imbalanced data sets within other machine learning classifiers like XGBoost and Random forest. Especially, with applying the SMOTE algorithm in combination with downsampling the majority class.

7.0 Conclusion

The aim of this study was to what extent can sampling methods enhance the prediction of online shopping intention? To answer this question the following machine learning algorithms are used; a decision tree, support vector machine, and a multilayer perceptron. To improve the prediction performance of the models, the research of Sakar et al. (2018) is taken as a baseline in this research. In the research of Sakar et al. (2018), the researchers conducted a study to predict whether a potential online webshop visitor will buy a product. Besides, the researchers oversampled the data after facing the problem of an imbalanced data set. In this study, the imbalanced data set problem is tried to improve by experimenting with downsampling the training data set and oversampling the training data set with the SMOTE algorithm. Lastly, the two sampling methods are used together to create a balanced training set for training the models.

The results are showing that randomly downsampling the data set is affecting the overall accuracy of the model. While increasing the proportions of downsampling, the model predicts more correct instances in the minority class. Applying the SMOTE algorithm provides a stable result in the overall score of the model, however, the correct predictions in the minority class are lower than downsampling the data set. Finally, the combination of downsampling and applying the SMOTE algorithm is resulting in a stable model. The model is better in predicting more correct instances in the minority class than applying SMOTE and is better in the overall performance of the model than downsampling the data set. The results are showing that downsampling the data set enhances predictions in the minority class and oversampling the data set with the SMOTE algorithm enhances the overall prediction performance.

After conducting this study, it can be concluded that applying SMOTE, downsampling, or a combination of these two methods, in the study of Sakar et al. (2018), does not improve the prediction performance of machine learning algorithms. In the scope of processing imbalanced data sets to better train models and perform better predictions in the minority class, this study contributes to a new approach of combining these sampling methods. As mentioned in section 6.1, further research is required to conclude this combination of sampling methods within the use of different machine learning classifiers. However, this research is providing a new direction with the combined sampling methods.

References

Barros, T. M., SouzaNeto, P. A., Silva, I., & Guedes, L. A. (2019). Predictive Models for Imbalanced Data: A School Dropout Perspective. *Education Sciences*, *9*(4), 275.

Batuwita, R., & Palade, V. (2010). FSVM-CIL: fuzzy support vector machines for class imbalance learning. *IEEE Transactions on Fuzzy Systems*, *18*(3), 558-571.

Constantinides, E., & Holleschovsky, N. I. (2016). Impact of Online Product Reviews on Purchasing Decisions. *Proceedings of the 12th International Conference on Web Information Systems and Technologies.*

Díaz-Vico, D., Figueiras-Vidal, A. R., & Dorronsoro, J. R. (2018, July). Deep MLPs for imbalanced classification. *In 2018 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-7). IEEE.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220–239.

Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, *61*, 863–905.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., & Seliya, N. (2019). Examining characteristics of predictive models with imbalanced big data. *Journal of Big Data*, *6*(1), 1–22.

Ibrahim, Z. M., Bader-El-Den, M., & Cocea, M. (2019). Improving Imbalanced Students' Text Feedback Classification Using Re-sampling Based Approach. *Advances in Intelligent Systems and Computing*, 262–267.

Khalilpour Darzi, M. R., Niaki, S. T. A., & Khedmati, M. (2019). Binary classification of imbalanced datasets: The case of CoIL challenge 2000. *Expert Systems with Applications*, *128*, 169–186.

Lin, M., Tang, K., & Yao, X. (2013). Dynamic sampling approach to training neural networks for multiclass imbalance classification. *IEEE Transactions on Neural Networks and Learning Systems*, *24*(4), 647-660.

Liu, Y., An, A., & Huang, X. (2006). Boosting Prediction Accuracy on Imbalanced Datasets with SVM Ensembles. Advances in Knowledge Discovery and Data Mining, 107–118.

Oh, S. H. (2011). Error back-propagation algorithm for classification of imbalanced data. *Neurocomputing*, *74*(6), 1058-1061.

Picek, S., Heuser, A., Jovic, A., Bhasin, S., & Regazzoni, F. (2019). The curse of class imbalance and conflicting metrics with machine learning for side-channel evaluations. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, *2019*(1), 1-29.

Ramezankhani, A., Pournik, O., Shahrabi, J., Azizi, F., Hadaegh, F., & Khalili, D. (2016). The impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes. *Medical decision making*, *36*(1), 137-144.

Rodriguez-Torres, F., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. (2019). Deterministic oversampling methods based on SMOTE. *Journal of Intelligent & Fuzzy Systems*, *36*(5), 4945–4955.

Rubin, D., Martins, C., Ilyuk, V., & Hildebrand, D. (2020). Online shopping cart abandonment: a consumer mindset perspective. *Journal of Consumer Marketing*, *37*(5), 487–499.

Rustogi, R., & Prasad, A. (2019). Swift Imbalance Data Classification using SMOTE and Extreme Learning Machine. *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, 234–240.

Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2018). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications*, *31*(10), 6893–6908.

Shatnawi, R. (2017). The application of ROC analysis in threshold identification, data imbalance and metrics selection for software fault prediction. *Innovations in Systems and Software Engineering*, *13*(2-3), 201-217.

Sun, J., Li, H., Fujita, H., Fu, B., & Ai, W. (2020). Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. *Information Fusion*, *54*, 128–144.

Zhang, Y., & Trubey, P. (2018). Machine Learning and Sampling Scheme: An Empirical Study of Money Laundering Detection. *Computational Economics*, *54*(3), 1043–1063.

Appendix A: Variables data set

Variable	Description	Values	Type
Administrative	Visit administrative page	continuous	Numerical
Administrative_Duration	Duration at the administrative page	continuous	Numerical
Informational	Visit informational page	continuous	Numerical
Informational_Duration	Duration at the informational page	continuous	Numerical
ProductRelated	Visit product-related page	continuous	Numerical
ProductRelated_Duration	Duration at the product-related page	continuous	Numerical
BounceRates	Enter and leave from the same page	continuous	Numerical
ExitRates	The percentage that was the last in the session	continuous	Numerical
PageValues	Value of page before completing a transaction	continuous	Numerical
SpecialDay	Closeness to a special day when visiting the site	continuous	Numerical
Month	The specific month of visit	12	Categorical
OperatingSystems	Visitors operating systems	8	Categorical
Browser	Visitors browsers	13	Categorical
Region	Region of visitors	9	Categorical
TrafficType	From which site the visitors enter the site	20	Categorical
VisitorType	Kind of visitor	3	Categorical
Weekend	If it was a weekend day while visiting	2	Boolean
Revenue	If visitor bought or not	2	Boolean

Appendix B: Link data set and algorithms

https://github.com/MustafaBulca/Thesis_DSS.git