Predicting revenue generation in an online retail website using machine learning algorithm

MSc Research Project
in Data Analytics

Annadurai Srinivasan
Student ID: x18130879

School of Computing
National College of Ireland

Supervisor: Prof. Jorge Basilio

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Annadurai Srinivasan
Student ID:	x18130879
Programme:	Msc in Data Analytics
Year:	2019-2020
Module:	Msc research project
Supervisor:	Jorge Basilio
Submission Due Date:	23/04/2020
Project Title:	Predicting revenue generation in an online retail website using machine learning algorithm
Word Count:	5921
Page Count:	18

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	26th May 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting revenue generation in an online retail website using machine learning algorithm

Annadurai Srinivasan
x18130879

Abstract

E-commerce has been the hot spot for business in today's world.Every day millions of people spend their time buying, selling and surfing on the internet. The main motive of any business is to generate optimal revenue. The most vital aspect in e-commerce is to track any potential buyer and try to generate income. The previous studies conducted were focused more on the deep learning. The biggest drawback in using deep learning models is that, it consumes a lot of time to run the model, which is not optimal in the real-world scenario. In this research, both the techniques of undersampling and oversampling were implemented. ANOVA was used for feature selection which aided in reducing the noise in the dataset by removing unwanted features, which in return increased the performance of machine learning models. A significant increase in accuracy was observe by combining XG Boost classifier with oversampling technique.

Contents

1 Introduction

Online business is now an emerging business around the world. Nowadays nearly half of the customers shop online products rather than buying in stores. E-commerce business enterprise has gained quite an amount of data containing interactions, transactions, data observation and also social association, location of the user, and the data which are online and offline. This data causes individualized to appeals more and more distinct. This enormous growth of e-commerce enterprises gained an interest for the researchers. Industrial developments offer diverse platforms for e-commerce businesses to get the maximum advantages for the higher revenue generation. Product pages/applications and websites are the most typically used platforms by the customers to find out various unique elements of the product while also contributing this to the mode of purchase. E-commerce enterprises target the most ultimate platform for the business. All these data of the user are collected via user registration credentials. Since the data collecting is large, it becomes a tough task to predict. However, standard tools have been implemented by a few industries whose results were not satisfactory. Even after employing some advanced tools, it became too complex. Computing these complex tasks, this research tries to use machine learning models, a useful tool for the prediction of the popular platform which can generate the highest revenue.

Researches that was previously performed were mostly based on deep learning. Deep learning is also a part of machine learning. Authors **Sakar et al. (2018)** used two modules in this research that concurrently predict a visitor's intent of shopping. In their first module random forest (RF), support vector machine (SVM), and multilayer perception(MLP) classifiers an input with the extracted features. While the second module uses the clickstream serial data (deep learning), here the memory-based neural network is used to produce the sigmoid output. Then the two modules are used together to predict the visitor's intention. One of the major issues in this study was, they used two modules (with deep learning) which consumes a lot of time to run a model that might be not optimal in the real-world scenario. So, therefore, it's better to use only machine learning models instead of deep learning. This brings us to the research question of this research.

To what extent a potential customer can generate revenue in an online retail site? Any machine learning model is known as a good model when it provides high accuracy with minimum time complexity. Gaussian naïve Bayes, XG boost, support vector machine, Random Forest classifiers are used in this study, all these models are well known for high accuracies. To minimize the time and complexity of space we have used ANOVA (Analysis of variance) feature selection technique in this research. Hyper-parameter tuning was done to fine-tune the model and achieve better accuracy. For Evaluation, different metrics were used such as accuracy, F1 Score, AUC score, kappa score.

2 Literature Review

Random Forest gives more accuracy in comparison with other classifiers **Valecha et al. (2018)**. In this research authors, the main motto was to analyze the relationship between customer behavior factors and the willingness to buy online. First, they started with finding out the connection among the consumer behavior factors to purchase any products on the changing parameters like environmental, organizational, interpersonal, and individual factors. And then the random forest classifier is used for the prediction of customer behavior, which involves the selection of buying the product significantly. In this research, they have compared the random forest with the other two models named Logistic Regression and K-nearest neighbor, which gave them an accuracy of 55% and 68% respectively. Random forest gave an accuracy of 94%. From this, it was known that the Random forest performs excellently in terms of accuracy in machine learning models.

Over-sampling can be used for the better performance of the model and scalability of classifiers **Sakar et al. (2018)**. To predict the customers buying intent and for website abandonment, authors have used two modules. In the first module using aggregated page view data they predicted the user's intention of purchase and then the features were fed to the classifiers named random forest, support vector machine, and MLP (multilayer perception). The second module uses the deep learning technique i.e. by applying only sequential clickstream data they trained memory-based recurrent neural network that produces a sigmoid output which shows the users intention to leave the site without completing the transactions. Both these modules are combined to predict the customer's intention for buying the products but most likely can leave the site in the prediction horizon. From this research, we got to know that machine learning and deep learning can perform better in terms of accuracy and scalability for a virtual shopping environment.

The software of data mining analyses the relationship between the patterns based on customers' request **Maheswari and Priya (2017)**. Using Support vector machine (SVM) classifier authors have predicted the customer behavior in online shopping. Inventory and sales data set which are available open on the internet has been used in this work. With the help of the neural network techniques, a support vector machine (SVM) was constructed. The function of kernel sigmoid was used by the SVM model in this research. As a result, they said that using a Support vector machine (SVM) is one of the better options for analyzing customer behavior in a better way.

Sophisticated machine learning libraries can be used for predicting the best revenue-making platform **Kamal et al. (2019)**. E-commerce enterprises are one of the most ideal platforms as a crucial task. The authors said that all the data related to the customers are gathered through their user credentials. They also mentioned that these data are large, so it's difficult to predict accurately. Some standard tools have been introduced by the industries whose results were not up to the mark. So, computing to complex tasks, supervised machine learning algorithms can be used for the prediction of popular platforms which can generate maximum revenue. In this work, they have used linear regression for the prediction of the famous platform that generates maximum revenue.

Machine learning can significantly enhance the performance in data mining while also can achieve precise marketing **Rao et al. (2018)**. In this research, authors have used tea device organization data from the e-commerce enterprise. To get the frequent itemsets Frequent pattern to grow algorithm is used and then they have used the naïve Bayes algorithm on feature vector for implementing cluster learning for the online referral benefits along with precision marketing. Then the evaluation begins with the feasibility of the data mining with machine learning algorithms across benefits generated by goods sale.

Recommender system is very important in E-commerce business **Kaur et al. (2019)**. The recommender system concept is basically to advise a customer a new product instead of manual search because whenever any user needs any new product, the user gets confused about which product will suit him better. For this in this study authors have performed an experiment on the intention of online shoppers using Machine learning algorithms. Several machine learning approaches were implemented. Machine learning algorithm used, namely: Naïve Bayes, Decision Stump, Random Forest, Multilayer perception, and J48. All these evaluations of algorithms were performed based on Mean square Error(MAE), Recall, Precision F Score, and Receiver Operating Characteristics (ROC). From the result, it was seen that random forest was the only classifier that performed extremely well in all parameters. It gave a ROC value of 92

From the above section, it can be understood that the machine learning algorithm works best for the e-commerce business in comparison with other standard tools. Also, classifiers like random forest, Support vector machine, naïve Bayes, works best in these models.

2.1 Sampling:

Class imbalance issues are the most typical problems in any real-world applications plus it affects the performance of the model significantly in terms of prediction, accuracy **Liu et al. (2014)**. In this research, a study on better handling class imbalance issues in customer behavior, the prediction is executed. Three methods have been used named weighted random forest, RUS boost, and random forest under-sampling. With the help of the evaluation metric, these three methods were investigated. Out of these three methods, weighted Random forest only increases the AUC for the appetency, and for further two classification issues, it does not perform well. Random forest with under-sampling performed best in all the three classifications problems, similarly, the RUS boost also improved all three classification issues but not as much as a random forest with under-sampling. Results showed that under-sampling works best for the class imbalance issues, and furthermore they concluded on future work that using the oversampling methods can also be more effective.

Machine learning algorithms are mostly expected to generate unsatisfactory classifiers with imbalanced data sets **Zhihao et al. (2019)**. A classifier with an accuracy of 98% with a 2% event rate is not accurate if it classifies all cases as majority class and eliminating the 2% minority class observations as noise. The approach used in this research is to define different approaches by using various sampling techniques and then weighing every technique for its pros and cons. The sampling techniques used in this study were Random downsampling, oversampling, k-Means clustering algorithm, Synthetic minority over-sampling. From all the above techniques, the result showed that balancing the data approach with SMOTE (oversampling), training a gradient boost algorithm on a balanced set can improve the accuracy significantly of a predictive model. They concluded this research by saying that whenever we face any imbalance data sets, there is not any proper solution to improve the accuracy of the prediction model. Instead, one should try various sampling techniques to figure out which techniques suit the best for the data set.

From the above papers, it says why we have used both up and sampling techniques in our research.

2.2 Label encoding:

Machine learning algorithms do not support text values without further exploitation so the feature encoding techniques should be applied **Jackson and Agrawal (2019)**. Encoding is converting any textual data into numerical values which can be understood by machine learning models. Authors have used three encoding techniques in this study named label encoder, one hot encoder and binary encoder. All these were useful encoding schemes when dealing with large text data. All these encoding techniques applied to the cybersecurity datasets to examine which encoding works the best with a machine learning algorithm in identifying the interference detections. In this study they concluded that using label encoder to machine learning algorithm works best in comparison with the other two encoders and the best performing ML model with label encoder is CART (used in this research).

To improve the overall performance of a system selection of a good encoding technique is important **Armano and Tamponi (2012)**. In tuning a classifier, encoding plays a key role. As the selection of the correct encoder may significantly improve the performance of the model. In this study, to test the performance of the encoding technique used in classification, they recommended different approaches. These approaches offer correlation-based metrics developed through the primary goal of focusing on encoding techniques. They mentioned that the proposed approaches allow saving computational time to a great extent since it requires only a small fraction of the time needed by standard methods. From this study, we got to know why encoding is necessary for machine learning models.

2.3 Feature selection: ANOVA

Authors **Yakub et al. (2016)** carried out research on feature selection, which is based on ANOVA for the microarray data classification. They have used SVM (Support Vector Machine) in the classification of microarray datasets. In this research, gene expression data is reduced to the minimum subset of genes with the help of a feature selection technique called an ANOVA. They mentioned that ANOVA can reduce the noise and computational burden occurring from irrelevant genes. Different machine learning algorithms and statistical theories were used in previous researches to eliminate the irrelevant and redundant features but still it is unclear about how these algorithms react to a condition called a small sample size. So, after the evaluation with SVM based method, they said that ANOVA (analysis of variance) works best to reduce the high dimensions and noise arise from irrelevant and redundant features.

Feature selection is very vital in machine learning because it improves the performance of the classifiers and reduces the dimensionality **Elssied et al. (2014)**. This research is based on a one-way ANOVA F-test for E-mail Spam classification. In e-mail spam classification, the Support vector machine (SVM) is typically used but still, there is an issue of the high dimensionality of feature space because of a huge number of e-mail data set. So, to enhance the limitations of SVM, improving classification accuracy and decreasing computational complexity, a one-way ANOVA f-test was carried out. It reduces the high data dimensionality of feature space ahead of the classification process. This proposed scheme test was conducted through a spam-base well known benchmarking data set for the estimation of feasibility proposed method. The result showed that after the feature selection technique was conducted, the classification accuracy increased to a great extent i.e. 93%.

2.4 Naïve Bayes:

SMOTE (Synthetic Minority Over Sampling Technique) technique is effective in increasing the overall performance of the unbalance classification of data **Saputra and Suharjito (2019)**. Authors (Adi and suharjito.et.2019) carried out research on Fraud detection using machine learning in e-commerce. The Machine learning models used in this study were named as Naïve Bayes, random forest, decision tree, and neural network. They said that the data which was selected was unbalanced. So, the Oversampling technique (SMOTE) was used to create balance data. Results were tested using confusion matrix, 96% by the neural network, 95% by naïve Bayes and random forest, and 91% by a decision tree. Based on the results, it was concluded that the neural networks and the classifier like naïve Bayes and random forest work best in fraud detection systems. They also mentioned that SMOTE also increases the recall value, f1 score of the model.

An intrusion method of detection using Gaussian naïve bayes based on improved PCA was proposed by the authors **Choraś (2018)**. Data pollution was reduced by weighing the first few feature vectors of traditional PCA. Then these final weighted components are given by sequential selection. PCA improves the dimensional reduction of data. Finally, the Gaussian naïve Bayes classifier detects the intrusion behavior of the model. For the evaluation of results detection time, precision rate, recall rate, indexes of detection accuracy were applied. The result showed that the method proposed in this study reduces the detection time by 60% and detection rate by 91% when compared with the traditional Bayes method. They also said that when comparing with other classifiers Bayes classifiers have the fast speed for classification.

2.5 Hyperparameter:

The selection of hyper-parameters can drastically affect the performance of the model **Claesen and De Moor (2015)**. In this study, researchers establish the hyper-parameter search issues in machine learning and examination about its major challenges from the optimization view. Models built by the machine learning systems depict few elements of interest-based on given data. Some typical machine learning algorithms feature a set of hyper-parameters which should be determined before training starts.

Random forest works relatively well with the default principles of hyper-parameter **Probst et al. (2018)**. This study is divided into 2 parts wherein the first part they provide a literature review of parameters that influences the important variable measures and on prediction performance along with the interactions among hyper-parameters. In the second part of the study after the overview of strategies of tuning, they showed the use of one of the highly established strategies of tuning model-based optimization. To make it simple tune, Ranger R package was provided which tunes RF along with model-based optimization. Results and the previous study showed that the tuning of random forest can improve the performance of the model. The authors also suggested that the use of sequential model-based optimization can use to tune the parameters entry, node, and sample size simultaneously.

3 Methodology

CRISP-DM is the method used for this research, which stands for a cross-industry process for data mining. The reason for selecting this method is, it gives us a fixed approach structure for this research. It's also flexible and used to solve mostly analytics for business challenges.

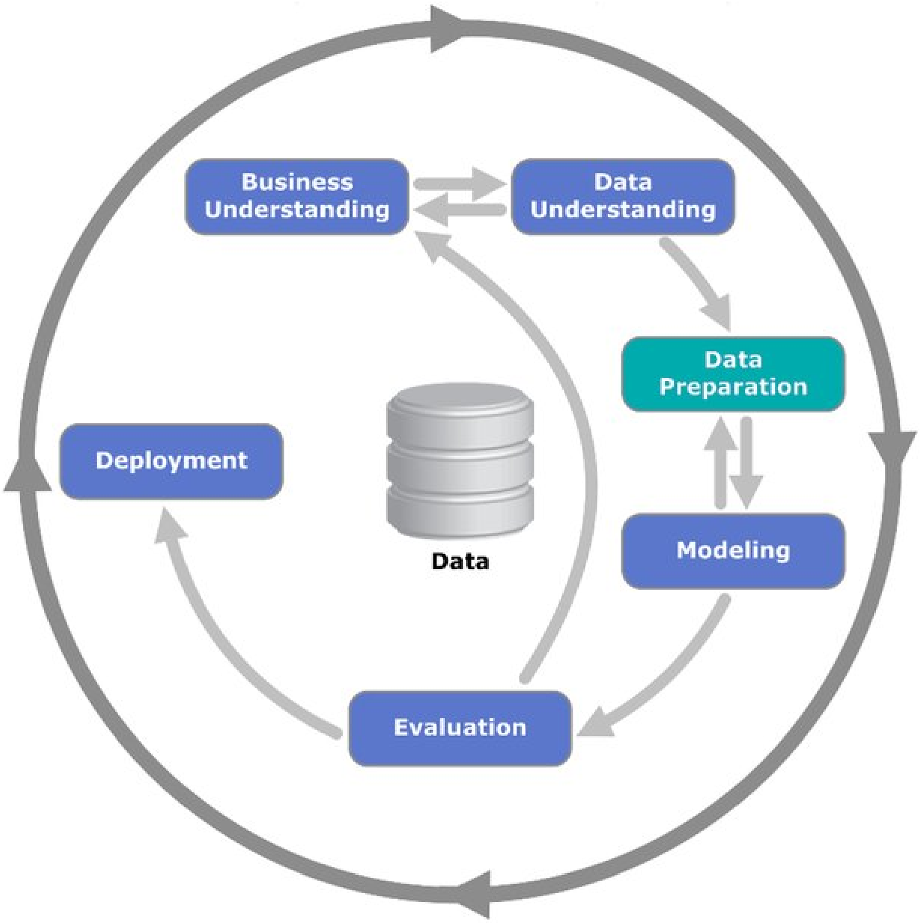

Figure 1: Cross-Industry Standard Process for Data Mining (CRISP-DM) Kenda and Mladenić (2017)

3.1 Business Understanding:

Understanding the base knowledge of any research is the first step taken in the machine learning research project. This research starts with a research question, whether a customer generates revenue in an online retail site? Before this, only a few types of research were done in online revenue generation using machine learning algorithms. In this researches, it was found that mostly they have concentrated on deep learning which uses two modules for the prediction purpose. Deep learning models are mostly time-consuming, which might be not suitable for real-world scenarios. Our research is mainly concentrated on machine learning models. With machine learning models (in real-world scenarios) not only accuracy is important but also few other factors are important like F1, AUC, kappa score, and accuracy with low latency. In this research, our main aim is to build a model with good accuracy (low latency) considering F1, AUC, and kappa score.

3.2 Data understanding:

Figuring out and identifying the data is one of the key aspects considering the machine learning models. The data set used The dataset used in this research comprises 12331 rows and 18features in which only a few features have been taken (using ANOVA) for implementing the models.

For the betterment of the data, we first performed the encoding because machine learning algorithms only performs with numerical values. So, it is very essential to encode data to be understood by the machine learning models. A label encoder is used for this research.

3.3 Data preparations:

Once the encoding was performed. It was found that the dependent variable had the class imbalance issues which are the typical problems in classification. So, the sampling techniques were used to avoid the class imbalance issues. Two sampling techniques were performed named as upsampling and downsampling. When the majority class of a dataset (dependent variable) is brought down to minority class it is called random downsampling. When the artificial data is added to the minority class it is known as the upsampling technique.

In this data set, we can use the SVM- Support vector machine as a main source of output, because SVM provides an accuracy of almost 75percent in any circumstances which means we could say that this model can predict 4out 5times correctly. So Hence we can use this model for one of the base classifiers. In-fact we can compare the results of the different classifiers regarding the SVM model. So we can get a better accuracy

3.4 Modelling:

In this section, the first step is to build a model with minimum numbers of features to avoid high dimensionality and correlation issues. For this, we use the technique called ANOVA (Analysis of variance) for feature selection. We have 18 features in the dataset out of which 17features are independent and 1 is dependent.

After using the ANOVA technique, machine learning models were implemented. Models like Gaussian naive Bayes, XG Boost are used in this research which is popularly known for the high accuracy, low latency. Support vector machines and random forest classifiers are used to deal with high dimensions. Hyper-parameter tuning was done at the end to fine-tune the machine learning model. Since Gaussian naive Bayes had no parameter to optimize while XG Boost, Random Forest and Support Vector Machine had hyper-parameters which were used for the tuning by Random Search.

3.5 Evaluation:

For the evaluation, different metrics were used. This metrics comprises F1 score, kappa score, AUC score. All these were used to evaluate the machine learning models. Since we have used upsampling and downsampling with Hyperparameter tuning at the end. All the results are got with both the sampling techniques and hypo parameter tuning.

4 Implementation and Design Specification

- **Data Selection:** The data used in this research is got from [1]. The dataset used in this research has no ethical issue as there is no personal or private data involved. The data comprises 18 features and 12,330 rows. The dependent variable used

[1] https://www.kaggle.com/henrysue/online-shoppers-intention

in this research is binary (True or False). The dataset was found to have a huge class imbalance. The majority class (False) has 10.4k rows when compared to 1908 rows of minority class (True). The dataset contains 4 categorical variables and 14 continuous variables.

- **Administrative:** It represents the number of administrative pages that the customer has visited.
- **Administrative_Duration:** It represents the time spent in the administrative pages.
- **Informational:** It represents the number of informational pages that the customer has visited.
- **Informational_Duration:**It represents the time spent in the Informational pages.
- **ProductRelated:**It represents the number of product-related pages that the customer has visited.
- **ProductRelated_Duration:**It represents the time spent on the product-related pages.
- **BounceRates:**This represents the percentage of people who enter the website and exit generating no added tasks.
- **ExitRates:**It denotes the percentage of page views on the website that concludes at that specific page.
- **SpecialDay:**This represents the value of the nearness of the browsing day to special days or holidays (eg Father39;s Day or Christmas) in which the transaction is expected to be concluded.
- **Month:**It denotes the month the page view occurred.
- **OperatingSystems:**It represents the Operating System that was used by the user for viewing the page.
- **Browser:**It represents the browser that was used by the user for viewing the page.
- **Region:**It is a numerical value that represents the region in which the user is located.
- **TrafficType:**It is a numerical value that represents the traffic the user is considered into.
- **VisitorType:**It is a categorical representation of a visitor, which shows whether the visitor is New Visitor, Returning Visitor or Other.
- **Weekend:**It is a categorical representation of whether the session is on a weekend.
- **Revenue:**It is a categorical representation of whether the user completed the procurement.

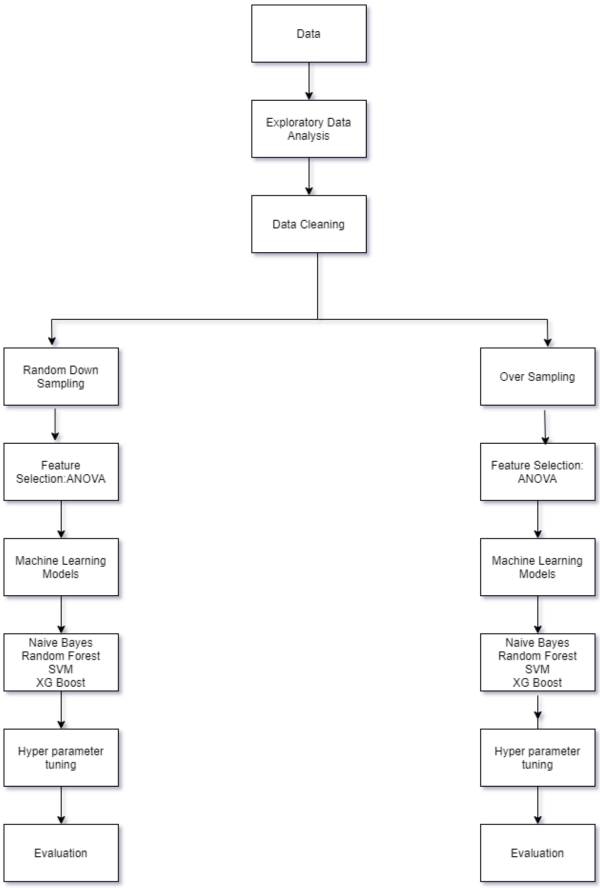

Figure 2: Flow Diagram

- **Exploratory Data analysis:**The critical process of investigating on any data is called as Exploratory data analysis. It is used for finding irregularities and exploring the underlying pattern in a dataset. It is the first step to understand the data before diving into further steps of data mining. In this research, the Pandas Profiling package is used for EDA. Pandas profiling offers a verity of exploratory analyses such as finding the values in each column, the number of missing values in each column, and also delivers a correlation plot of Pearson and Spearman correlation. From the use of Exploratory data analysis, the following 3 anomalies were found. in the dataset:

  1. Finding NA: By using Pandas profiling, it was found that the dataset Contained NA values which were needed to be eliminated. The removal of NA's was done in the data cleaning section.

  2. Finding categorical variables present: By using the Pandas profiling, it was found that the dataset contained categorical values which were needed to be encoded. The encoding of categorical values was done in the data cleaning section.

  3. Finding class imbalance: By using Pandas profiling, it was found that the dataset had a class imbalance issue. The class imbalance was eliminated by using two technique namely under-sampling and over-sampling.

- **Data Cleaning:** Data Cleaning is the most vital step in Data mining projects. If the data is not properly cleaned, it may add noise to the dataset and may have a negative impact on the performance of the machine learning model. The steps used for Data Cleaning in this research are:

1. **Removing NAs:** Existence of null values in a dataset can adversely affect the machine learning model. Hence it is very important to eliminate the NAs. In this research, dropna() is used to remove the NAs from the dataset.
2. **Encoding categorical variables**: Machine learning algorithms can only work with numerical values. Hence, it is important to encode the data in a way that is understood by the Machine learning algorithms. Variables can be encoded in several ways. In this research, Label encoding is used.
3. **Label encoding:** Label encoder assigns a value between 0 to n-1 where n represents the number of categories. In case a class reoccurs, it encodes the same value allocated previously. In this research, along with the dependent variable, three other variables are encoded. The reason for using Label encoder in this research is because the categories contain only binary values and hence there it is no need to use a one-hot encoder.
4. **Class imbalance:** refers to the imbalance in the Dependent Variable. One of the biggest problems in Classification is Class Imbalance. Two techniques can handle the Class Imbalance problem, Upsampling, and down-sampling.
   - **Upsampling:** is the technique that adds Artificial data to the minority class. In this research, the SMOTE package is used for Upsampling. SMOTE refers to Synthetic Minority Over-sampling Technique. The Upsampling approach is optimal when minority class is negligible and is not enough for training any machine learning algorithm, which holds true in this research. Hence, the upsampling technique has performed better than down-sampling in this research.
   - **Downsampling:**Random down-sampling is a technique where the majority class is brought down to the size of the minority class. This technique aids to solve the problem of Class Imbalance. To perform random down-sampling, the 'RandomUnderSampler' package is used in this research. Even tough down-sampling performed well, but in comparison, the Upsampling technique performed better.

- **ANOVA:** ANOVA is a statistical analysis of variance which aids to find the best features in a dataset. It is an advanced technique, to examine each feature in a dataset and assign a score to each feature. It performs F-test to check whether there are any significant diversities that exist amongst the features. According to the scores assigned to each feature in the dataset, it keeps only those with higher significant values.

- **Hyper parameter Tuning**:

  1. **Random Search:**

     Hyper-parameter tuning is a technique that optimizes the hyper-parameters of a machine algorithm. The technique used for Hyper-parameter tuning in this research project is Random search CV. It uses random combinations of parameters for conducting Hyper-parameter tuning. It provides the best combinations of the parameter to build a machine learning model. It is like Grid

Search CV, although it is well known to perform better than Grid Search CV in relation to computation. In this research, Random Search, CV is used to Hyper-parameter tune XGBoost, Random forest and SVC.

- **Evaluation Index**:

  Evaluation plays a vital role in the Data mining project. It shows how well a machine learning model is performing in relation to different metrics. The methods used for Evaluation in this project are:

  1. **Accuracy:** It is the most commonly used metrics to test the performance of any machine learning models. Accuracy is calculated using the following formula.Accuracy can be defined as the percentage of correctly classified instances. Accuracy is the most frequently used metrics to assess the performance of any model. The formula to calculate Accuracy is:

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

  2. **F1 Score:** The F1 score shows the balance between the precision and the recall. It is also known as the F Score or the F Measure. The value of F1-score lies between 0 to 1. The following formula is used to calculate the F1-score:

     $$F1 - Score = 2 * \frac{precision*recall}{precision+recall}$$

  3. **AUC**: The abbreviation of AUC is an area under the curve. AUC is used in classification problems to find the model that predicts the classes better. AUC is measured based on specificity and sensitivity. The range of AUC lies between **0 to 1**.

  4. **Kappa Score:** Kappa score is also known as Cohen's kappa. The inter-rater reliability can be calculated using the Kappa score. It is always a better measure than a simple percentage agreement calculation. In simple words, it determines whether or not the model is randomly guessing the output. The value of the Kappa score lies between -1 to 1, where 1 shows complete agreement and -1 shows complete disagreement.

- **Machine Learning Algorithms**

  1. **XGBoost:**XGBoost refers to extreme Gradient Boosting. As the name shows, XGBoost is the gradient boosting algorithm that works on tree-based learning. The tree-method parameter is used to 'GPU-hist' for quicker computation by utilising GPU resources. This alteration in parameter aids to boost up the performance of XGBoost significantly. XGBoost is used because:

     - XGBoost supports parallelization. XGBoost works faster compared to other machine learning models as it uses all the threads, by default.
     - XGBoost algorithm is very well known in the classification problem in the research community.

     Parameters used for XGBoost are:

     - min_child_weight;
     - subsample;

  - max_depth;
  - n_estimators;
  - learning_rate;
  - eta;

2. **Random Forest:** Random Forest is a combination of many individual decision trees working together. Every tree in this model shows a category of prediction, and the decision is made based on the votes. Similar to any tree-based model, random forest works best with continuous variables. The vital features of Random forest are comparatively decent accuracy, robustness, and ease of use. Random Forest is well known in the Research community for its Accuracy in the classification problems.

   Parameters used for Random Forest are as follows

   - n_estimators;
   - max_depth;
   - min_samples_leaf;

3. **Support Vector Machine (SVM):** SVM is commonly known as Support Machine Vector. It is a supervised Machine learning model which is opted by many data experts as it provides improved accuracy with low computational power. SVM can be used for Regression and Classification problems. The variant of SVM used in this research is SVC (Support Vector Classifier). SVM uses a method named the Kernel trick to convert the data and then based on these conversions it finds an ideal border between the possible outputs.

   Parameters used for SVM are:

   - C;
   - gamma;

4. **Gaussian Naïve Bayes:** Gaussian Naïve Bayes is a powerful algorithm used for model prediction. It is a variant of naïve Bayes which works with continuous variables. The dataset used in this research mostly comprises continuous variables and hence Gaussian Naïve Bayes was implemented. Gaussian Naïve Bayes is well known for its simplicity with low time and space complexity, as it only has to store probabilistic values that consume less space and computers faster. Because of the above features, it is optimal to use in a real-world scenario.

- **Design Specification** This research project is built on Google Colab. Google colab is a cloud-based data science work space similar to the jupyter notebook. Colab can be accessed by the following link https://colab.research.google.com. Run time was set to GUP for faster execution. Google Cloud Platform (GCP) machine having basic configurations as follows:

  CPU : 8 core

  RAM : 32 GB

  The language used in this project is Python.

5 Evaluation

5.1 Without hyper parameter tuning(down sampling)

Classifier	Accuracy	F1 Score	AUC Score	Kappa Score
Gaussian Naive Bayes	0.7618	0.7465	0.7667	0.5276
XG Boost	0.8325	0.8346	0.8339	0.6653
Random Forest	0.8259	0.831	0.8265	0.6517
Support Vector Machine	0.644	.6785	0.6402	0.2821

In this section, XG Boost gave the highest accuracy of 83.25 and an F1 score of 0.83 compared to the other models. Random Forest also performed well in terms of accuracy of 82.59 with an F1 score of 0.83. All these terms were achieved without hyper-parameter tuning.

5.2 Without hyper parameter tuning(Up Sampling)

Classifier	Accuracy	F1 score	AUC Score	Kappa Score
Gaussian Naive Bayes	0.7661	0.751	0.7654	0.5315
XG Boost	0.9122	0.9128	0.9125	0.8246
Random Forest	0.9214	0.9223	0.9216	0.8428
Support Vector Machine	0.7182	0.726	0.7186	0.4386

After performing Upsampling, Random Forest gave the maximum accuracy of 92.14 with F1 score of 0.92, AUC and Kappa Score of 0.92 and 0.84 respectively. XG Boost classifier also performed equally well in terms every performance metric as shown in the table above.

5.3 With hyper parameter tuning(Down Sampling)

Classifier	Accuracy	F1 Score	Auc Score	Kappa Score
XG Boost	0.8338	0.8349	0.8355	0.6681
Random Forest	0.8285	0.8323	0.8295	0.6572
Support Vector Machine	0.7055	0.7678	0.6939	0.3964

Hyperparameter tuning fine-tunes the model.After executing the tuning(downsampling) we had some changes in the performance metric as shown in the table. XG Boost classifier really good with an accuracy of 83.38, F1 score of 0.8349, AUC score of 0.8355, Kappa Score of 0.6681. Support Vector Machine classifier gave the lowest accuracy of 70.55 with a decent F1 score of 0.7678.

5.4 With hyper parameter tuning(Up Sampling)

Classifier	Accuracy	F1 score	AUC Score	Kappa Score
XG Boost	0.9287	0.9275	0.9287	0.8573
Random Forest	0.9222	0.9227	0.9226	0.8445
Support Vector Machine	0.7949	0.8211	0.7978	0.5921

Hyperparameter tuning with Upsampling performed exceptionally good in terms every performance matrix when compared with other tables above. XG Boost gave the highest accuracy of 92.87 with an F1 score of 0.9275, AUC gave 0.9287 and kappa Score of 0.8573.

5.5 Discussion

In this research four models were implemented in order to predict the revenue generation in an online retail site. Each trained model in this study was evaluated with hyper parameter tuning (up sampling and down sampling) and without hyperparameter tuning for the betterment of the results. From the results obtained we can conclude that the models trained with hyper parameter tuning with up sampling provided the best performance. This can be seen in the tables shown in the results section. In up sampling, XG Boost classifier gave the maximum accuracy of 92.87. Random forest classifier also performed equally good with an accuracy of 92.22. After the execution of the results we can say that, using hyper parameter tuning can improve the performance of the model to good extent.

Few of the main challenges were faced in this research. One of the major issues faced was the class imbalance while working on the data set. Class imbalance could result in the significant decrease in the size of the data. Due to this, we can face adverse effect on the performance and the understanding of the machine learning model. Since in this research we had only the small data set to work with, so the credibility can be questionable. So, to overcome this issue, various data sets can be used to train, test the data which can increase the integrity of this research. Various types of classifiers were used in this research named as Random Forest, Support vector machine, XG Boost, gaussian naïve Bayes. But when working with the classifier like XG Boost it was very crucial to know the parameters and the ranges of every parameter. Since this model is new, which was little difficult to examine and tune this model.

6 Conclusion and Future Work

The main objective of this research is to build a machine learning model with high accuracy, F1-Score, and AUC. Several steps were implemented to achieve the results. Firstly, Exploratory Data Analysis (EDA) was used to find the anomalies present in the data using Pandas_Profiling(). EDA gives the number of null values present in the dataset and the variables which need to be encoded. In the data cleaning process, all the null values are eliminated and all the categorical variables are encoded. The next step includes creating functions for Undersampling and Oversampling, to find the best technique that is apt for the research. Once this step was done, ANOVA was implemented for feature selection. This step helped to reduce the noise in the data by removing unwanted columns, which aided in improving the performance of the machine learning model. The final step was to implement the machine learning models and Hyper Parameter Tuning (HPT) them. HPT helped in achieving an overall better accuracy.

This research was successful as the tested hypothesis did hold true. The accuracy achieved can be considered as a good performance and hence can be implemented in the real-world scenario. This can be backed by the outputs shown in the results section. XG-Boost when combined with oversampling performed the best with an impressive accuracy of 93%. Random Forest also gave a good performance when combined with oversampling, with an accuracy of up to 92%.

6.1 Future Work:

1. Other improved machine learning models can be implemented to gain a better Accuracy.
2. Other advanced feature selection techniques can be used for better performance.
3. In future, better accuracy can be gained by using the bigger and better dataset

References

Armano, P. G. and Tamponi, D. E. (2012). Assessing encoding techniques through correlation-based metrics, *2012 11th International Conference on Machine Learning and Applications*, Vol. 1, pp. 634–639.

Choraś, Michał, Z. B. L. Z. (2018). Network intrusion detection method based on pca and bayes algorithm, *Security and Communication Networks* .
URL: *https://doi.org/10.1155/2018/1914980*

Claesen, M. and De Moor, B. (2015). Hyperparameter search in machine learning.

Elssied, N., Ibrahim, A. P. D. O. and Hamza Osman, A. (2014). A novel feature selection based on one-way anova f-test for e-mail spam classification, *Research Journal of Applied Sciences, Engineering and Technology* **7**: 625–638.

Jackson, E. and Agrawal, R. (2019). Performance evaluation of different feature encoding schemes on cybersecurity logs, *2019 SoutheastCon*, pp. 1–9.

Kamal, R., Karan, A. and Arungalai, V. S. (2019). Investigations on e-commerce data for forecasting the efficient promotional platform using supervised machine learning, *2019 4th International Conference on Recent Trends on Electronics, Information, Communication Technology (RTEICT)*, pp. 939–943.

Kaur, B., Sharma, R., Rani, S. and Deepali, G. (2019). Recommender system: Towards classification of human intentions in e-shopping using machine learning, *Journal of Computational and Theoretical Nanoscience* **16**: 4280–4285.

Kenda, K. and Mladenić, D. (2017). Autonomous on-line outlier detection framework for streaming sensor data, *Proceedings - International Symposium on Computers and Communications* .

Liu, N., Woon, W. L., Aung, Z. and Afshari, A. (2014). Handling class imbalance in customer behavior prediction, *2014 International Conference on Collaboration Technologies and Systems (CTS)*, pp. 100–103.

Maheswari, K. and Priya, P. P. A. (2017). Predicting customer behavior in online shopping using svm classifier, *2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, pp. 1–5.

Probst, P., Boulesteix, A.-L. and Wright, M. (2018). Hyperparameters and tuning strategies for random forest.

Rao, H., Zeng, Z. and Liu, A. (2018). Research on personalized referral service and big data mining for e-commerce with machine learning, *2018 4th International Conference on Computer and Technology Applications (ICCTA)*, pp. 35–38.

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

Saputra, A. and Suharjito, S. (2019). Fraud detection using machine learning in e-commerce.

Valecha, H., Varma, A., Khare, I., Sachdeva, A. and Goyal, M. (2018). Prediction of consumer behaviour using random forest algorithm, *2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, pp. 1–6.

Yakub, S., Arowolo, M., S.O., A. and M.D., S. (2016). A feature selection based on one way anova for microarray data classification, pp. 30–35.

Zhihao, P., Fenglong, Y. and Xucheng, L. (2019). Comparison of the different sampling techniques for imbalanced classification problems in machine learning, *2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, pp. 431–434.