Festival day arrival prediction using Random Forest, Support Vector Machine and Deep Learning Techniques

MSc Research Project
Data Analytics

Rajbharath Jothimani
Student ID: x21133000

School of Computing
National College of Ireland

Supervisor: Cristina Hava Muntean

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Rajbharath Jothimani
Student ID:	x21133000
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Cristina Hava Muntean
Submission Due Date:	01/02/2023
Project Title:	Festival day arrival prediction using Random Forest, Support Vector Machine and Deep Learning Techniques
Word Count:	8192
Page Count:	22

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:	Rajbharath Jothimani
Date:	31st January 2023

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):	

Festival day arrival prediction using Random Forest, Support Vector Machine and Deep Learning Techniques

Rajbharath Jothimani
x21133000

Abstract

Customers of e-commerce business exhibits a varied shopping pattern on any e-commerce portal related to general retail during festival season. Festival sales have begun to make a bigger contribution to the expansion of e-commerce business. With the help of variations in purchase patterns discovered from consumer session information on e-commerce portals, this study aims to predict the arrival of festival using user session data. The main benefit of this study is that it will help e-commerce companies increase their sales on festival days since it can anticipate festival day arrival using the user session details of the e-commerce users. Festival day arrival prediction using machine learning algorithms is made not only to help e-commerce businesses plan their marketing strategies well in advance of the festival, but also serves a number of other purposes, such as restocking specific products with a higher likelihood of selling during that festival and assisting the business in making decisions about improvements to the user interface of the e-commerce portal for obtaining better user traffic. This study will thus empower e-commerce companies to increase their revenue during festivals and initiate action plans towards marketing strategies and product restocking by predicting the arrival of the festival day. Machine learning models such as Random forest classifier, Support vector machine classifier and Deep learning using Keras are used in this research for predicting the festival day arrival since it falls under classification problem. Hyper parameter tuning were performed in these models to optimize the prediction performance. Among the three models chosen, Random forest classifier outperformed both support vector machine classifier and Deep learning using Keras with a higher accuracy of 85.47%. These models are also evaluated based on various other evaluation metrics such as Log loss, Matthew correlation coefficient, precision score, ROC-AUC score, Recall score and F1-score. Random forest classifier not only outperformed the other two models based on accuracy, but also on all the evaluation metrics being mentioned, thus making random forest classifier as the best model for performing the prediction of festival arrival using user session data of e-commerce.

Keywords: coefficient, precision, vector, session, recall, strategies, restocking

1 Introduction

1.1 Background Scope and Motivation

E-commerce industry has achieved more growth over the past decade and has seen a significant increase in customers recently, since individuals have grown accustomed to using e-commerce platforms to make purchases. The behavior of users of e-commerce platforms varies, which directly affects the sale of the products on e-commerce platforms. Several factors that directly affect customer's purchasing decisions can be the reason for difference in consumer shopping behavior. One crucial element to take into account is festival day, which is observed to have a higher impact on a customers purchasing decisions. Due to the anticipated significant shift in consumer purchasing behavior during festival seasons, e-commerce platforms use flash sales with extreme discounts to generate enormous sales. The primary goal of this study is to determine how e-commerce businesses may better profit from the prediction of festival day arrival by taking advantage of the fluctuating shopping habits of consumers obtained from user session data of e-commerce portal. With the aid of this prediction model, e-commerce businesses may launch better marketing campaigns far sooner than their rivals, assisting them in replenishing inventory as needed during the festival season.

A lot has been stated in previously published articles regarding how consumer purchasing behavior alter during the festival season. Additionally, mentions of the better product sales that often take place during festival time has also been found in few articles. The prediction model for festival arrival by utilizing consumer behavior obtained from user session data on the e-commerce portal has not been developed in any of the existing articles, despite the fact that it is crucial for an e-commerce company to do so in order to increase revenue. Therefore, predicting the arrival of the festival day is crucial and can increase the company's revenue.

Machine learning models such as Random Forest, Support Vector Machine and Deep Learning using Keras are used in this research. This research will predict the arrival of a festival using the user session details present in the dataset. The need to research using these prediction models is that, since the target variable to be predicted falls under classification problem, models such as Random Forest, Support Vector Machine and Deep Learning using Keras are used to determine the value of the target variable for predicting the arrival of a festival. Since the goal of our study is to predict the arrival of festival, Random forest is a well preferred method , because it requires creating numerous decision trees during training, which basically outperforms other classifier methods. Despite offering more precise results, they lack decision tree's inherent interpretability. Keras is a open source Python library that is frequently used for creating and analyzing deep learning models . It is a part of the Tensorflow library. In this study, neural networks are utilized in the classification problem to predict the arrival of festival based on attributes contained in the dataset. Neural networks can be simply trained and created with few lines of code using Keras. Support Vector Machine (SVM) is a discriminative classifier and is normally defined through a separating hyperplane. .Given a collection of training sample, with each designated as belonging to one or two categories, an SVM machine learning model assigns future examples to one category or the other thus functioning as a non-probabilistic binary linear classifier

1.2 Research Question

To what extent machine learning models such as Random forest, Support vector machine and Deep learning using Keras can be applied on user session details from e-commerce portal in predicting the arrival of festival day ?

1.3 Research Objectives

The following research objectives are created in this study in order to successfully complete the research work.

- Review the literature cited in the literature review section and critically analyze it.
- Create a subscription in azure cloud for creating azure data lake storage. Once the container is created inside azure data lake storage, load the input file as a Blob inside the container.
- Exploratory data analysis to be done on the data to get a better understanding on the data.
- Select and implement the preferred machine learning models and perform evaluation through evaluation metrics to choose the best model suitable for this prediction problem.
- Discuss on the results and conclusion and future works that can be done from this research.

1.4 Contributions

The major contribution of this research is a new approach which compares supervised machine learning models such random forest classifier and support vector machine classifier with deep learning model using keras which runs on tensor-flow framework for the prediction of festival day arrival using e-commerce user session data. This research can help organization is deciding on the proper approach to go for, while performing this prediction using user session data to predict the arrival of festival .

1.5 Structure of the paper

This paper discusses machine learning models used for classification with regards to e-commerce customer behaviour and factors influencing purchase decisions and purchase behaviour during festive season in section 2 related work . The research methodology being used in the research is discussed in detail in section 3. Section 4 discusses the design specifications of the various components used in the machine learning framework being used in the research. Section 5 gives a better overview on the implementation of this research in detail . Section 6 discusses about the evaluation metrics and the results obtained. Section 7 provides a conclusion on the research and discusses about the future works that can be performed on top of this research.

2 Related Work

2.1 Introduction

A lot of articles over the internet have discussed over the consumer behaviour impact on the revenue of a e-commerce business. They have mostly described on whether a customer will make purchase or not, depending on the shopping pattern exhibited by the customer. This research establishes a different concept, where it will predict the arrival of festival based on the shopping pattern exhibited by the customer using the user session data of e-commerce. Also various algorithms which are suitable for this prediction problem are also discussed in this section.

2.2 Impulsive Purchase Behaviour during Festive Season

Consumers typically spend more throughout the festive season. E-commerce companies have started taking advantage of this opportunity to push more sales at this time. According to this study (Yulianto et al.; 2021), the primary causes of impulsive shopping behavior during shopping festivals are sales promotion and hedonic shopping motivation. Additionally, it discusses how businesses use this impulsive behavior to boost sales at this time. On the other hand, it was also shown that throughout this time period, attitudes toward impulsive shopping were unaffected by the perishability of the products and scarcity. However, it did not address the question of whether or not impulsive shopping will ultimately benefit e-commerce companies.This study (Lim et al.; 2017) discusses impulsive buying benefits and drawbacks, which were not included in earlier research. Previous research largely neglected to discuss an empirical approach to difficulties that could occur from impulsive buying and instead focused exclusively on improving sales through customers impulsive shopping behavior during shopping festivals. The results of this study, which describe a customer purpose after making an online purchase and also discuss their behavioral tendency to return products, imply that impulsive buying behavior could have a significant impact after the sale of the product.This research (Chen and Li; 2020) states that understanding the impact of product promotion tactics used on customers and other promotion strategies to boost their intention to participate in online shopping festivals will leads to impulsive purchase behaviour. According to the study's findings, consumers intentions to participate in events can be considerably increased by constant temptation regarding discounts and offers, category-based promotions, and the promotion of functional activities. However, it did not go into great detail about the kinds of techniques used to collect the data for this study.This paper (Parmar and Chauhan; 2018) clarifies that, in order to fulfill objectives, a structured questionnaire was utilized to conduct the research and a count of 106 consumers were surveyed by using convenience sample approach, resolving the limitation in the previous paper described. Additionally, the paper aims to go deeper into the variables influencing online impulsive purchases. According to the study's findings, holiday deals and end-of-season sales are two of the primary things that make people more likely to use an e-commerce site.As described by the research(Tzeng et al.; 2021), China's Singles Day is the biggest online shopping day in the world, however despite higher sales during this time, consumer unhappiness is also on the rise. Additionally, it explores how after-sales services can raise customer happiness, which is typically low when making impulsive purchases during a shopping festival.

2.3 Factors influencing E-commerce users purchase decisions

According to this study (Quan; 2021), sales on the Alibaba online store during the Singles Day shopping frenzy are around three times as high as those during Black Friday. Festivals play a significant role in determining an e-commerce user inclination to make a purchase. It illustrates how the business uses this time to comprehend customer preferences, needs, and pricing discrimination. Additionally, it notes that during the festival sales season, the organization turnover has been steadily rising over the years. Festivals are one of the primary elements influencing a user buying decision, however this study did not go into great detail regarding other factors that can affect a customer purchase decisions.As per this study (Esmeli et al.; 2022), a novel framework to do early purchase prediction utilizing the online user sessions details for registered and unregistered consumer, once they visit an e-commerce platform is developed, thus addressing the limitation in the prior paper. The research focuses on how context characteristics and user loyalty-related features can be understood as aspects that can assist businesses in developing marketing plans that can influence consumer purchasing behavior. It also discusses about improving the user experience in the portal so that it may provide personalized offers and discounts to customers, boosting sales and influencing their buying decisions. By utilizing information from users of Lithuanian e-commerce platforms, this research (Šneiderienė and Beniušis; 2022) was conducted. According to this article, demographic parameters including age, gender, location, and monthly income are important in determining a customer's purchasing intention. Additionally, it asserts that given the rise in e-commerce companies, it is crucial to concentrate on elements that may influence customers purchase decisions.The product rating found on websites explains the attitude, customers have toward a product. A study (Johan et al.; 2021) was conducted to see whether it actually influences a customer decision to make a purchase on an e-commerce site. The participants in this study were students at private universities in Bandung. The findings demonstrated that determining whether or not to make a purchase depends significantly on both product reviews and the website's overall quality. As per the study (Astuti and Pulungan; 2022), a research was conducted to identify and evaluate the impact of various variables on the purchasing behavior of online shoppers in Medan City. A questionnaire was employed in this study to collect data, and close to 100 people were deemed as the target population. According to the study's findings, customers purchase decisions are positively and significantly impacted by promotions, ease of use, and trust.

2.4 Application of User Session Data in E-commerce Predictions

According to this report (Kao et al.; 2021), quick changes in consumer behavior have compelled many e-commerce companies to implement strategies to increase consumers time spent on their portals browsing. It may also increase the likelihood that people will click on adverts. This study uses, user click-stream data, which includes user session information, to forecast how long a user would stay on an e-commerce portal. One of the key outcomes that may be extracted from user session data is the prediction of customer shopping behavior, which this research failed to mention. This report (Dong et al.; 2022) addressed every point that the preceding research, which claimed that customer behavior prediction is a major issue for many e-commerce firms, had overlooked. In this study, user id and user session data were gathered, and a prediction model was developed that could forecast and offer personalized products for each individual based on their preferences and shopping habits on an e-commerce site.

According to this studyDiwandari and Hidayat (2022), an organization can greatly benefit from a proper examination of click-flow data that includes user session information in gauging user interest. This may also result in the creation of marketing strategies. User session data can help businesses evaluate customer satisfaction and increase productivity by helping them better understand consumer preferences.In this study (Virk; 2021), product recommendations are made using click sequential data along with user session information for sessions where no purchases have been made. Not only are these systems found to recommend products, but it was also noticed that they boost sales and customer loyalty for the e-commerce company and produce better outcomes than the current systems.Organizations use a variety of techniques nowadays to get to know their customers. The goal of this research Misra et al. (2021) is to understand consumer behavior so that it may be translated into financial action. Click-stream analysis has been suggested as a feasible method for performing automated behavioral studies at scale over the past few years. The data appears to be significant for the aim of classification as evidenced by the improved precision and accuracy found for the model generated.

2.5 Evaluation of Machine Learning Models

This article (Baati and Mohsil; 2020) describes a prediction system that can categorizes the visitor's purchase intent, once a customer visits a webpage. Customer and session data are taken into account while making this prediction. For this prediction, several algorithms are employed. The greater accuracy and f1-score of 83.64 and 0.10 on an unbalanced dataset, however, led to the selection of Random Forest. This research shows that random forest is one of the best algorithms for classifying e-commerce data based on user session information, but it is also crucial to know which algorithm may produce better outcomes in terms of data interpretation.Therefore, in this study (Ekelik and Şenol; 2021) user session data from an online store that operates in Turkey is taken into consideration for classification. Random forest is one of the classification models utilized in this study, and it produces findings with an accuracy rate of 80.86 percent. This study reveals that no other algorithm was regarded to be ranked top since each varied in different evaluation measures, even though accuracy of random forest classifier was determined to rank slightly lower than an artificial neural network. Additionally, the adoption of Random forest, which consists of several decision trees capable of better interpretation, was advised.

In order to find machine learning models other than the random forest classifier that can aid in this prediction, more research was done in this area. This study (Xiahou and Harada; 2022) claims that consumer behavior information from a B2C e-commerce company is used to forecast client loss and to gauge the effectiveness of classification models like SVM abbreviated as Support vector machines and LR abbreviated as Logistic regression. The results of the tests show that SVM outperformed LR with an accuracy of 92.56 percent. In a subsequent study (Gordini and Veglio; 2017) on churn prediction using e-commerce session data, SVM was found to beat all other models, with the maximum prediction accuracy of 89.67 percent, making it the best model for the study. However, it was unable to make many recommendations for the kernel to be utilized, leaving the experiment to the further investigation, that can be conducted on this subject. In this study (Chaudhuri et al.; 2021), machine learning models are created using deep learning from the Keras 2.3.0 Python package. It focuses mostly on forecasting a customer's e-commerce platform purchase intent using information about their online portal interactions. It was able to generate data with a greater accuracy of 89 percentage, making it one of the best models for classifying e-commerce data.

2.6 Conclusion

Thus by exploring various articles related to the research topic, a better understanding on the research to be conducted was identified . Also, a greater understanding on the machine learning models to be used for the purpose of research was selected such as Random forest classifier, Support vector machine classifier and deep learning using keras. This leads to a conclusion of how further research should be conducted and what factors to emphasize upon for building the models to perform the prediction of festival arrival using e-commerce session data.

3 Methodology

In this study, Knowledge discovery in databases (KDD) methodology is employed. After considering the many research methods employed as well as the concepts and guiding principles that supplied the factual basis for the study, this approach is chosen.

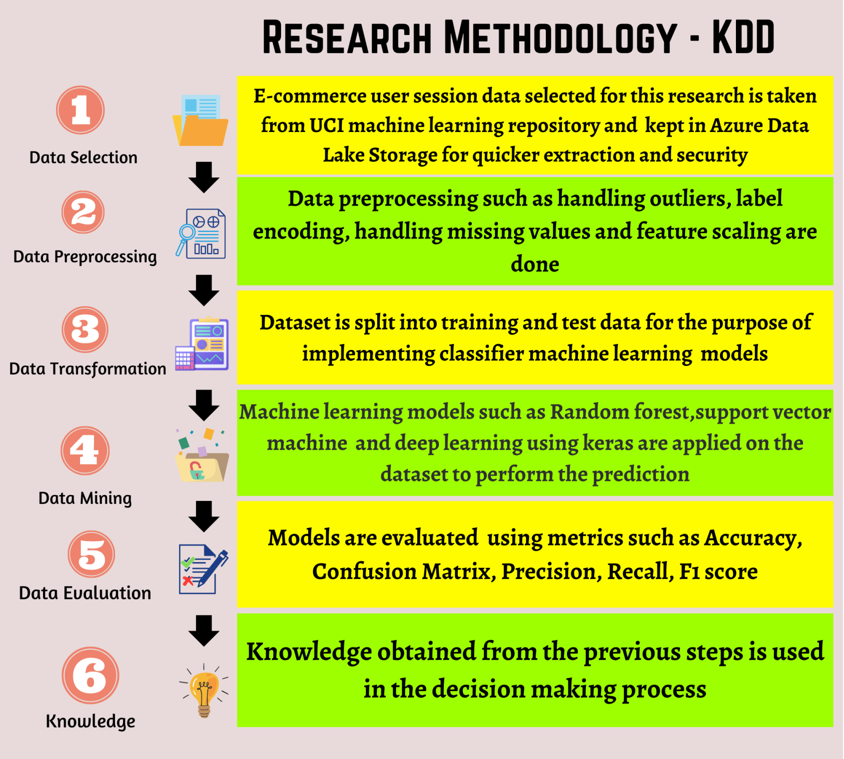

RESEARCH METHODOLOGY - KDD
1
Data Selection
E-commerce user session data selected for this research is taken from UCI machine learning repository and kept in Azure Data Lake Storage for quicker extraction and security
2
Data Preprocessing
Data preprocessing such as handling outliers, label encoding, handling missing values and feature scaling are done
3
Data Transformation
Dataset is split into training and test data for the purpose of implementing classifier machine learning models
4
Data Mining
Machine learning models such as Random forest,support vector machine and deep learning using keras are applied on the dataset to perform the prediction
5
Data Evaluation
Models are evaluated using metrics such as Accuracy, Confusion Matrix, Precision, Recall, F1 score
6
Knowledge
Knowledge obtained from the previous steps is used in the decision making process

Figure 1: Research Methodology

As seen in Figure 1, Steps in KDD methodology are carried one after the other to obtain the best results. The process starts with data selection where the data to be used in the research is extracted from UCI machine learning repository[1]. The dataset extracted contains user session details of e-commerce portal but no specific information is available about the name of the portal. Feature vectors present in the dataset contains detail of 12,330 sessions from different users. The dataset was formed in such a way that, each session information in the dataset belongs to a different user in a 1-year period so as to avoid any tendency to a specific campaign, special day, user profile, or period. It consist of totally 18 columns with 9 continuous and 9 categorical columns. Special day is decided as the target variable to be predicted.

Initially, the data is retrieved directly from the UCI machine learning repository using urllib package from the request library in python. Later, in-order to improve the fetching time and to secure the data , azure data lake storage is used for storing and retrieving the data.It was experimented and found that the retrieval from azure data lake storage takes lesser time than retrieving directly from the UCI machine learning repository on the web. The input data is stored in Azure data lake storage in the form of a blob. Once the data selection is completed, the data now goes through pre-processing stages where it is checked for missing values and outliers. The dataset is checked for missing values to identify if there are any missing detail present in the dataset. But no missing values were found in the dataset. The dataset was checked for outliers by plotting continuous variables using box-plots as seen in Figure 2 and it was found that outliers were present only in 3 columns such as Administrative_Duration,Informational_Duration, ProductRelated_Duration . All the extreme outliers present in the dataset were removed. The count of rows in the dataset is found to be 10527 after the removal of the outliers.

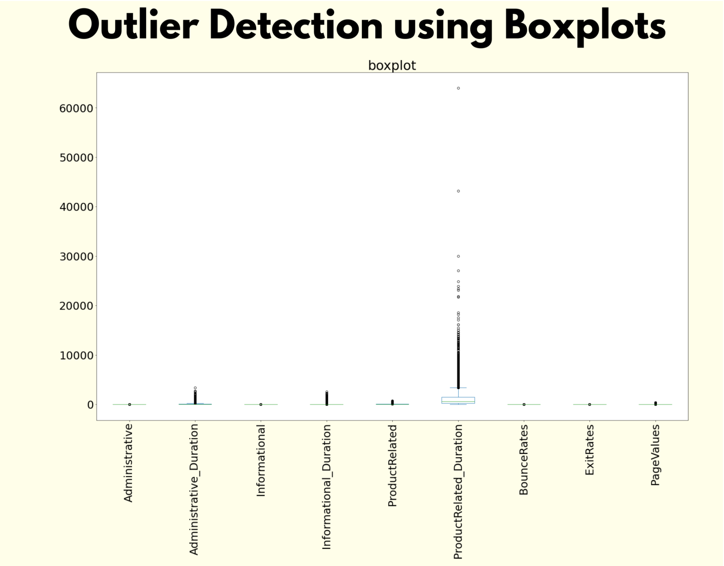


Figure 2: Outlier detection using box plots

Label encoding is done on all the columns with string data type to convert the data to numerical format. Feature scaling is also done to normalize the range of independent feature in order to obtain better results.Data pre-processing is followed by data transformation process. As seen in Figure 3, the target variable SpecialDay which was having 6 classes is transformed to have only two classes 0 and 1, with 1 denoting the arrival of

[1] https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

festival and 0 as not the arrival of festival.This is done to avoid the imbalance in the target variable while applying prediction models. Also, dataset is split into training set and test set with 80 percent of the data in training set and remaining 20 percent in the test set. Later, both training and test set are separated into two parts, with one part containing independent variables and the other part containing the dependent variable. In our case, the target variable SpecialDay moves into one part and the rest of the independent variables moves into the other.

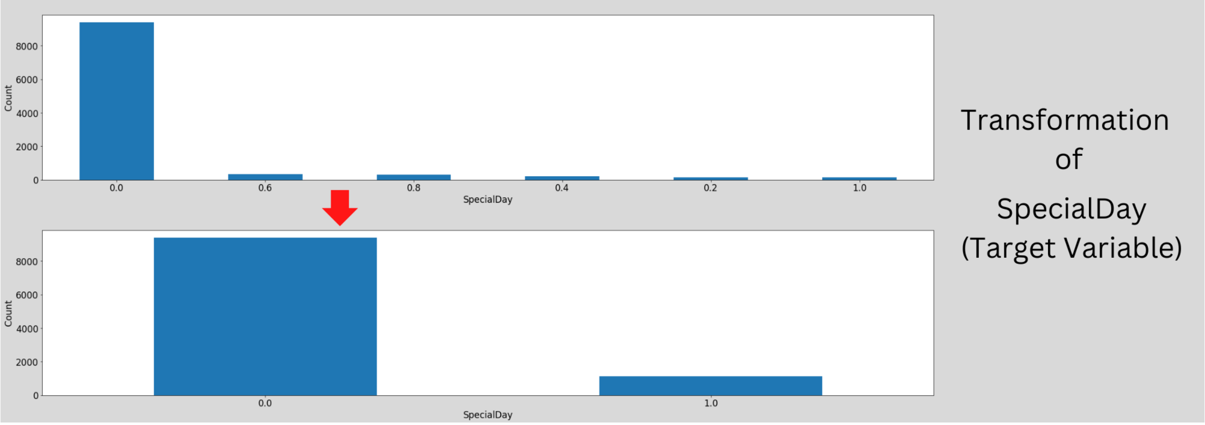


Figure 3: Target variable SpecialDay after data transformation

Now, The data mining process happens where the machine learning models related to classification problems such as Random Forest, Support Vector Machine and Deep Learning using Keras are applied on the dataset to predict the value of the target variable. As per the study (Piskunova and Klochko; 2020) , while classifying customers based on the e-commerce dataset, random forest gave a high accuracy of 99 percent when compared with classification algorithms such as Linear discriminant analysis, Classification and regression trees, Support vector machine and k - nearest neighbors thus making it as one of the suitable algorithm to work with classification data on e-commerce data.

4 Design Specification

This research follows the architecture design as shown in figure 4

4.1 Data Flow Architecture Design

The dataset to be used in the research is loaded into Azure Data Lake Storage for quicker fetch and data security. The data is retrieved from the data lake storage using blob service method in python into Google colab platform and converted into a pandas dataframe. In Google colab, all the data explorations and transformations are carried out and the data is split into training and test set for the purpose of modelling. Now, the data is processed by machine learning models which are related to classification problems such as random forest, support vector machine and deep learning using keras. The machine learning models are then evaluated based on the evaluation metrics such as Confusion Matrix, Accuracy score, Recall score, Precision score, F1-score log loss, Matthews correlation coefficient and ROC-AUC score.

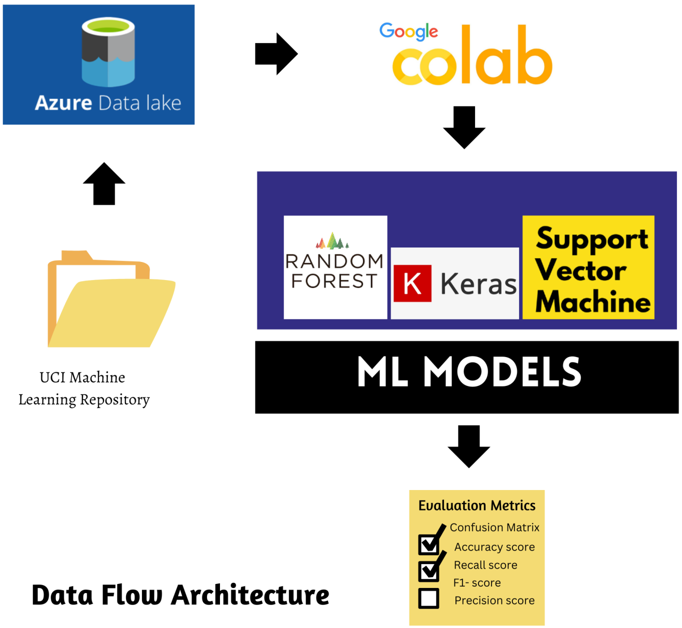


Figure 4: Data Flow Architecture Design

The Modelling techniques used in this research involves those models which are best suited for solving a binary classification problem. The target variable special day consist of two values 0 and 1 after transformation is being done, where 1 represents the arrival of a festival and 0 represents not the arrival of festival. Hence, models such as random forest, support vector machines and deep learning using keras are the selected techniques to be used in the research.

4.2 Work Flow Design used in Modelling techniques

As discussed in Figure 5, Cross validation is carried out for each of the model used in the research. This is done to make sure that the model does not falls under over-fitting or under-fitting scenarios. Hyper parameter tuning is carried out to optimize the performance of the model.Confusion matrix is also plotted to understand the proportion between the actual and the predicted values. Training and test accuracy are calculated along with Recall score, Precision score, F1-score log loss, Matthews correlation coefficient and ROC-AUC score.

4.3 Design specifications for Random Forest Classifier

Random forest design used in the modelling involves the following hyper-parameters to be tuned to obtain a better prediction with good accuracy score. RandomForestClassifier method is used in this binary classification problem. Parameter criterion is set to gini since it is faster when compared with entropy and is much less expensive. Since Random forest uses a multitude of decision trees to arrive at the result it is a best criterion to use in this research . Parameter bootstrap is set to True to enable each tree in random forest to train on different subset of observations instead of all the observations for each tree. Cross validation for the model is carried out with 7 folds . Training and Testing

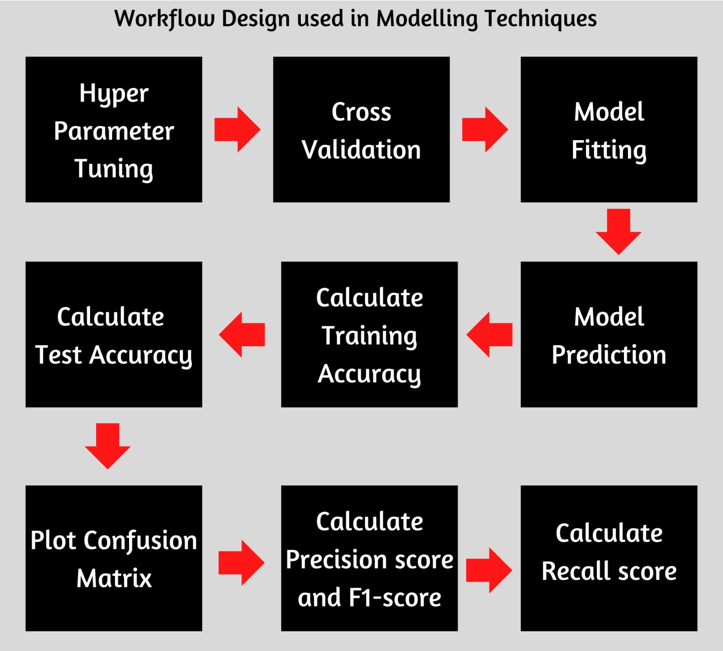


Figure 5: Work Flow Design

accuracy is measured by doing predictions on the training and test features using the model developed. Using matplotlib package of python, the confusion matrix is plotted. Cmap parameter in the plot_confusion_matrix method is set to PiYG to determine the color pattern for the confusion matrix. Figsize parameter is set to (10,10) to have a confusion matrix plot of width 10 pixels and Height 10 pixels.

4.4 Design specifications for Support Vector Machine Classifier

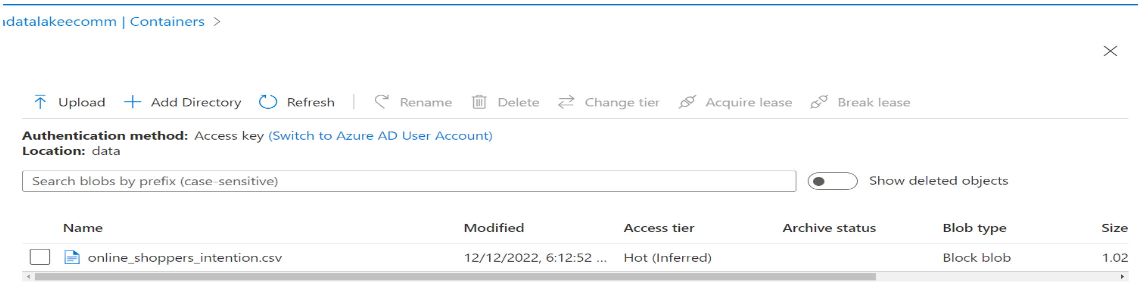
Support Vector Machine design used in the modelling involves the following hyper-parameters to be tuned to obtain a better prediction with good accuracy score with regard to classification problem. SVC is a classifier method in python for SVM. Cross validation is done using GridSearchCV method which takes SVC as a input along with parm_grid parameter. All the feasible values for input parameters in SVC modelling is given as feed to the GridSearchCV, so that it performs cross validation and picks the best parameters suitable for the model. Now the Support vector classifier model is created with the selected parameters. Training and testing accurracy are created for the model. Confusion matrix is plotted which is followed by the calculation of Recall score, Precision score, F1-score log loss, Matthews correlation coefficient and ROC-AUC score.

4.5 Design specifications for deep learning using Keras

In this research, Deep learning using keras involves designing a deep learning neural network for binary classification using keras python library which is a part of Tensor flow library . Sequential model is used where the layers are arranged in sequential order. Three different dense layers are used with activation functions such as sigmoid and relu. Sigmoid function involves calculating a exponential component which is more complex and provides high performance than a Relu function. Relu functions can generate quick results since it performs faster.

4.6 Design specifications for Azure Data Lake Storage

With regards to the storage design specification, Azure Subscription is obtained for storing the CSV data to be processed. In this project, pay-as-you-go subscription is used. Azure datalake storage is created and the CSV file is stored in the form of BLOB as seen in Figure 6.



Authentication method: Access key (Switch to Azure AD User Account)
Location: data

	Name	Modified	Access tier	Archive status	Blob type	Size
□	online_shoppers_intention.csv	12/12/2022, 6:12:52 ...	Hot (Inferred)		Block blob	1.02

Figure 6: Azure Data Lake Storage with Input file stored in the Container as Blob

BLOB data has to be retrieved from the Azure data lake storage and stored as a dataframe for further processing in the program. Hence, Access key details for the container is used to create a connection to the container in Azure data lake storage from python notebook in google colab environment.

5 Implementation

This section discusses in detail about the Implementation procedures used in this research. It focuses more on outputs produced from processes such as procedures used in data storage, retrieving the data from the data storage, transformations done on the data, models developed for prediction on classification problem, optimization done on improving the performance of the model and the type of evaluation metrics calculated based on the models developed.

5.1 Data Storage using ADLS

Azure Data Lake Storage (ADLS) is used in this research for storing the Input data to be used. The implementation process for ADLS is mentioned in the Figure 7.Initially, an Azure subscription is created. In order to create a Azure Data Lake Storage, a resource group is essential. Hence Resource group is created before the creation of ADLS. Later, ADLS is created by creating storage account under the research group created .

Hierarchial name space must be enabled for creating a Azure Data Lake Storage. A Container is created inside the ADLS to hold blob data. The Input File is then uploaded into the container created, as a Binary Large Object (BLOB). Once this process is established, the data to be retrieved will be available in the ADLS in form the of blob and can be accessed using the access keys of the container.

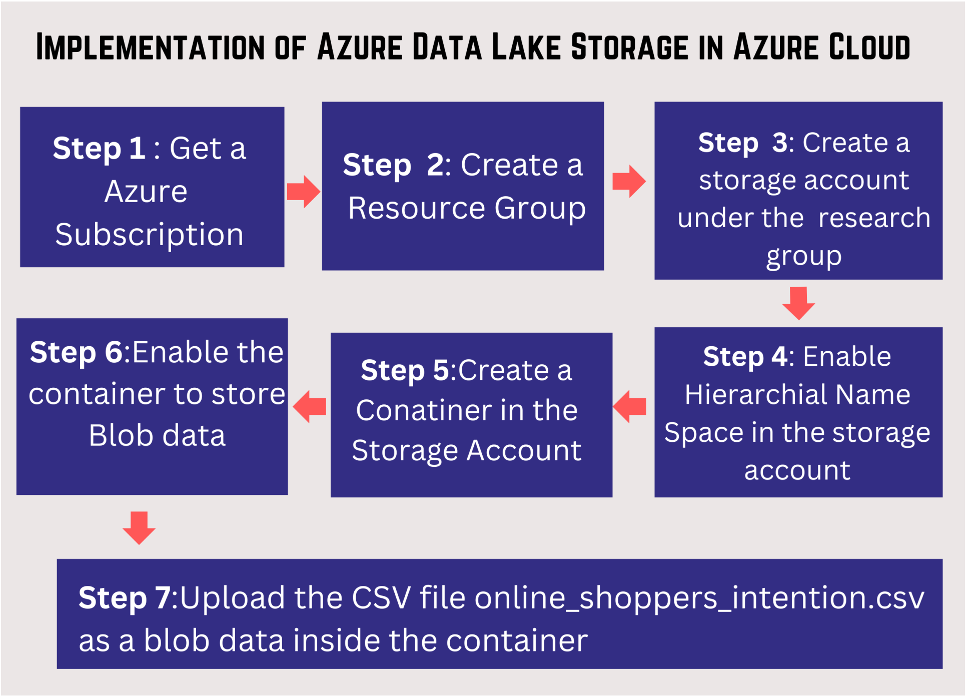
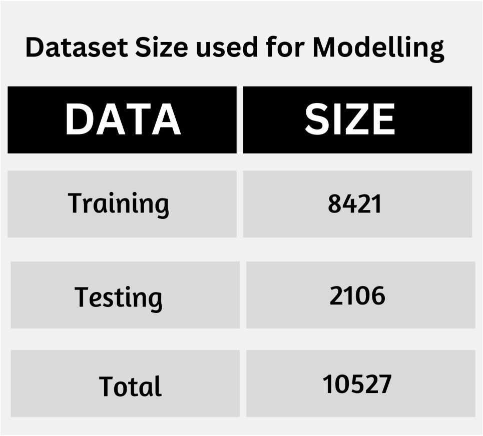

Implementation of Azure Data Lake Storage in Azure Cloud
Step 1 : Get a Azure Subscription
Step 2: Create a Resource Group
Step 3: Create a storage account under the research group
Step 4: Enable Hierarchial Name Space in the storage account
Step 5:Create a Conatiner in the Storage Account
Step 6:Enable the container to store Blob data
Step 7:Upload the CSV file online_shoppers_intention.csv as a blob data inside the container

Figure 7: Implementation of Azure Data Lake Storage in the cloud

5.2 Data Preparation for Implementation

: The transformed data after the pre-processing and transformation phase has to be prepared for the modelling technique. Hence the data is split into two different sets such as training and test for the purpose of implementing modelling techniques on the data. The total number of records in the dataset after the removal of outliers is 10527. This is now split into two with training dataset having 80 percent of the data and the test dataset with 20 percent of the data as seen in Figure 8. Later, cross validation is also implemented to enable unbiased sampling while performing the modelling techniques.



Dataset Size used for Modelling

DATA	SIZE
Training	8421
Testing	2106
Total	10527

Figure 8: Record count in training and test dataset

5.3 Implementation of Random Forest Classifier Model

RandomForestClassifier package from python library sklearn.ensemble is used in the modelling technique for building Random Forest classifier model. A seven fold Cross validation is done to perform hyper-parameter tuning. Cross validation resulted in the choosing on various values for the different parameter which resulted in a better prediction results. Some of the important parameters which were chosen after cross validation includes n_estimators, max_depth, criterion, max_features and max_leaf_nodes. n_estimators is chosen as 15 which implies that a multitude of 15 decision trees will be used in the random forest modelling. Also, the max_depth is chosen as 30 so that the trees in the random forest will be allowed to make a maximum split of 30. Gini criterion enables faster computation of the results with usage of less expensive resources. The value sqrt in max_features is used to determine the number of features to take into account for the purpose of modelling. max_leaf_nodes is set to 30 which states that splitting of the tree can go only upto 30 leaf nodes and the tree cannot grow after that. The mean accuracy score of the model has improved from 67.5 percentage to 84.7 percentage after the hyper-parameter tuning was done through cross validation. Now the model is built using the training data.

5.4 Implementation of Support Vector Machine Classifier Model

SVC package from python library is used in the implementation of Support Vector Machine Classifier model. In order to perform the cross validation on Support Vector Machine Classifier , GridSearchCV package from python library sklearn.model_selection is used. For the purpose of hyper-tuning the parameters, ranges are created for different parameters to be used in Support Vector Machine Classifier and provided to the GridSearchCV for performing the modelling. GridSearchCV runs a scan through the range and come up with optimized value for different parameters to be used in the modelling. In this modelling, the values obtained for the parameters are as follows. The value of C is selected as 10, gamma as 0.001 and kernel as rbf. The value of C is used to avoid the model from over-fitting scenario. Gamma controls the influence of a every training data point by managing the distance of influence of every single training point. RBF kernel is used in this model to handle space complexity issue. Using Fit method from sklearn python library, the model is built on the training data.

5.5 Deep Learning Using Keras

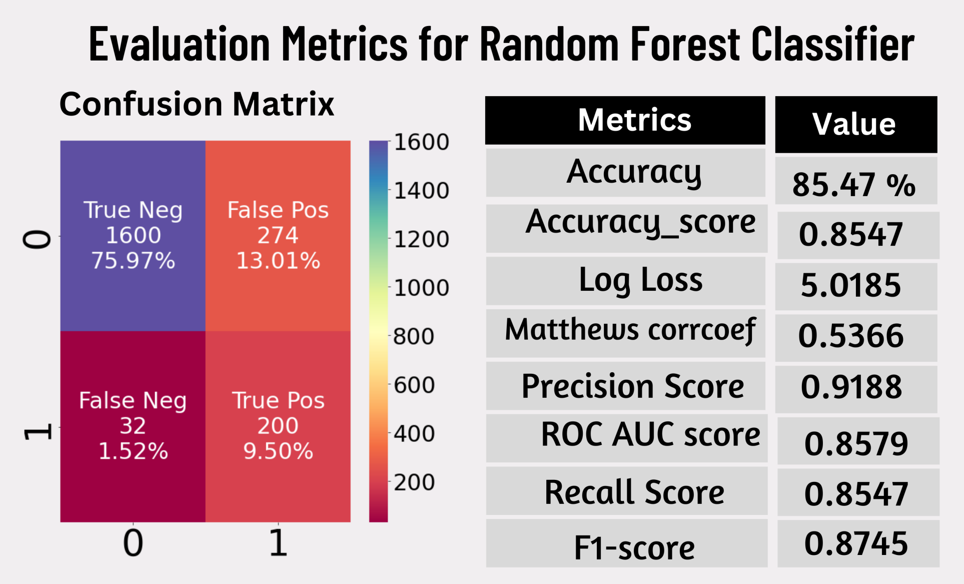
Tensorflow package from python is installed in-order to use keras modelling. Sequential model and Dense layer are selected from Tensorflow.keras package for the purpose of modelling. Sequential model involve using a stack of layers which gets executed in sequential order . Dense layer is used, since it is deeply connected in the neural network and hence can be used for a better prediction with regard to classification problem. Both Sigmoid and Relu activation function are used with regards to different dense layers involved in the modelling along with adam optimizer and binary_crossentropy loss function. For quick convergence and lesser computation time, Relu is used . Even though Sigmoid is computationally expensive, it has a exponential operation that can help in better prediction. The model is run with 5 epochs and the accuracy obtained is found to 59.50 percentage.

6 Evaluation

The results of the classification models used in the research are evaluated using evaluation metrics such as accuracy_score, confusion matrix, Recall score, Precision score, F1-score log loss, Matthews correlation coefficient and ROC-AUC score. The most crucial metric for this research is accuracy. This is because data transformation was used to balance the imbalanced data that made up the target variable SpecialDay in order to enhance prediction accuracy. Experiments are conducted for each of the model used in the modelling such as Random Forest Classifier, Support Vector Machine Classifier and Deep Learning using Keras. A comparison between these results are also done to understand which model performs best in accurately and precisely predicting the arrival of festival day using user session data of e-commerce. Mathews correlation coefficient is used to determine the statistical significance of the prediction done by the models. In the confusion matrix created for this evaluation, False Positive refers to the scenario where the festival is predicted to arrive based on user session data, but it does not arrive as predicted, whereas False Negative refers to the scenario where the festival is predicted to not arrive, but arrives as predicted, according to the actual data.

6.1 Experiment using Random Forest Classifier

Random Forest Classifier model developed is evaluated to understand it's performance in predicting the arrival of festival using user session data of e-commerce. From figure 9, it is clearly evident that it performed much better than expected. This is due to the fact that the confusion matrix resulted in 1600 true negatives which accounts for 75.97 percent of the test data and 200 true positives which accounts for 9.50 percentage of the test data. False positive and False negative accounted for 13.01 percentage and 1.52 percentage respectively. Also , the accuracy obtained was found to be 85.47 percentage which is comparatively a better accuracy for prediction.



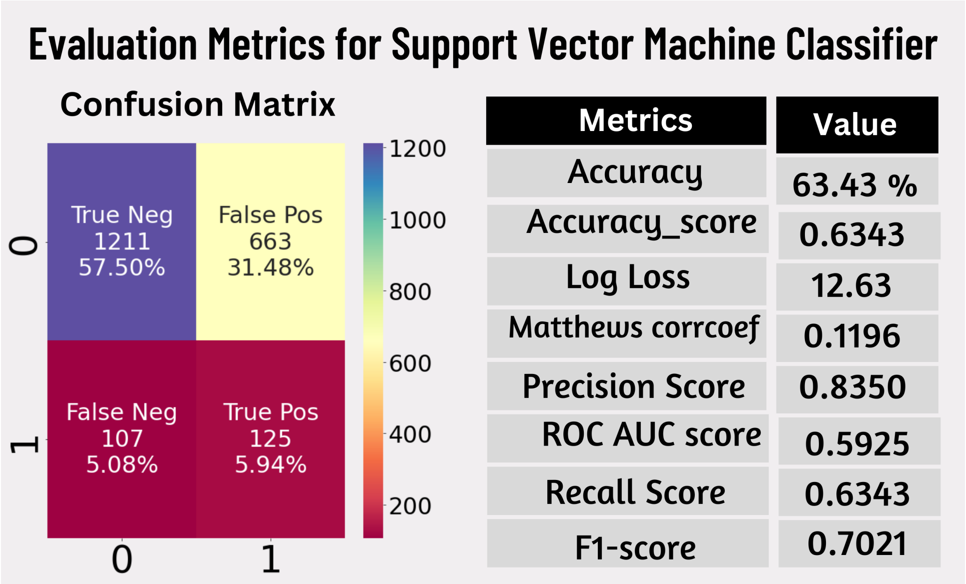
Metrics	Value
Accuracy	85.47 %
Accuracy_score	0.8547
Log Loss	5.0185
Matthews corrcoef	0.5366
Precision Score	0.9188
ROC AUC score	0.8579
Recall Score	0.8547
F1-score	0.8745

Figure 9: Evaluation metrics of Random Forest Classifier

The value of Log loss obtained was 5.0185, which is a lower value, thus proving that the prediction is performing better.The statistical significance of the model was calculated using Matthews correlation coefficient and it was found to be greater than 0.5 with a value of 0.5366, Hence, this prediction model is found to be statistically significant with better prediction results. Precision score and Recall score of the model was found to be 0.9188 and 0.8547 respectively which implies that result obtained were relevant and the relevant data were used in the prediction respectively. ROC-AUC score was calculated as 0.8579 which is defined as an excellent prediction and also suggests that no discrimination has been done based on the sample used. F1-score obtained was found to be 0.8745, which clearly states that model is more accurate while performing the data prediction since it is one of the important metrics used in validating the model's accuracy on a data-set.

6.2 Experiment using Support Vector Machine Classifier

Support Vector Machine Classifier model developed is evaluated and its performance is measured in predicting the arrival of festival. As seen in figure 10 , it is clear that it performs better but not as much as random forest classifier described above.



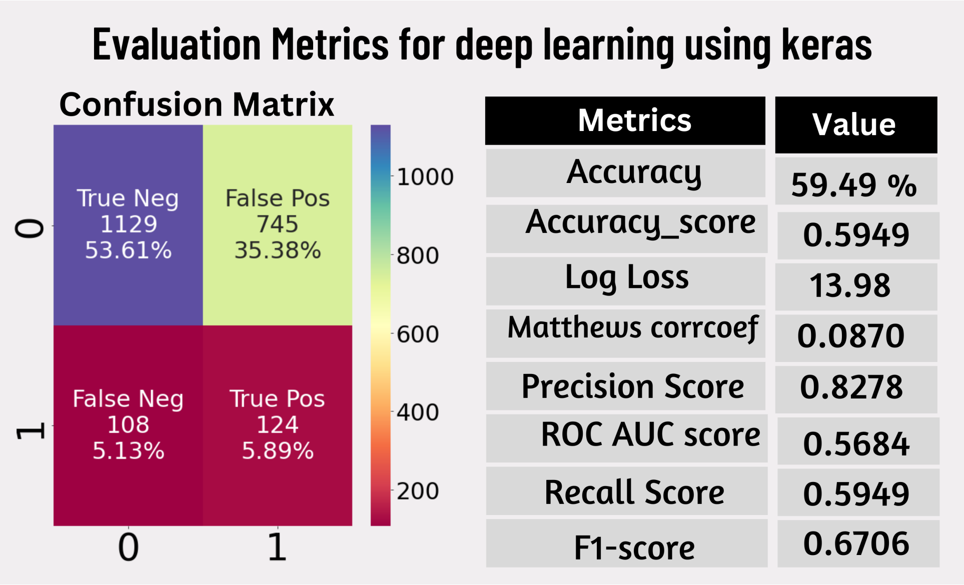
Metrics	Value
Accuracy	63.43 %
Accuracy_score	0.6343
Log Loss	12.63
Matthews corrcoef	0.1196
Precision Score	0.8350
ROC AUC score	0.5925
Recall Score	0.6343
F1-score	0.7021

Figure 10: Evaluation metrics of Support Vector Machine Classifier

This is due to the fact that the confusion matrix resulted in 1211 true negatives thus accounting for 57.50 percent of the data used in testing and 125 true positives which accounts for 5.94 percentage of the test data. False positive and False negative values accounted for 31.48 percentage and 5.08 percentage respectively. Also , the accuracy obtained was found to be 63.43 percentage which is good but comparatively lesser than Random forest classifier.The value of Log loss was found to be 12.63 , which is a moderate value thus proving that the prediction is performing at a moderate level.The statistical significance of the model was calculated using Matthews correlation coefficient and it was found to be 0.1196, Hence, this prediction model is found to perform good but not better than Random forest classifier. Precision score and Recall score of the model was found to be 0.8350 and 0.6343 respectively which implies that result obtained were closer to relevant and some relevant data were used in the prediction samples respectively. ROC-AUC score was calculated as 0.5925 which is defined as good prediction and also suggests that only less discrimination has been done based on the sample used. F1-score obtained was found to be 0.7021, which clearly states that model is accurate while performing the data prediction.

6.3 Experiment using Deep Learning with Keras

Deep Learning using Keras is done by developing a sequential model which is evaluated and it's performance is measured in predicting the arrival of festival through user session data features. As mentioned in figure 11 , it performs moderate but not as much as random forest classifier or Support Vector Machine Classifier model described above. The confusion matrix developed through this model resulted in 1129 true negatives thus accounting for 53.61 percent of the data used in testing and 124 true positives which accounts for 5.89 percentage of the test data. False positive and False negative values were found to be accounting for 35.38 percentage and 5.13 percentage of the test data respectively. Also , the accuracy obtained was found to be lesser than both Random Forest Classifier and Support Vector Machine Classifier with the value of 63.43 percentage which is comparatively lesser than both Random forest classifier and and Support Vector Machine Classifier.The value of Log loss was found to be 13.98 , which is an average value thus proving that the prediction is performing at a average level.The statistical significance of the model was calculated using Matthews correlation coefficient and it was found to be 0.0870, Hence, this prediction model is found to perform average and lesser than Random forest classifier and and Support Vector Machine Classifier. Precision score and Recall score of the model was found to be 0.8278 and 0.5949 respectively which implies that results obtained were closer to the relevant and not much relevant data were used in the prediction samples respectively. ROC-AUC score was calculated as 0.5684 which is defined as a good prediction and also suggests that only average discrimination has been done based on the sample used. F1-score obtained was found to be 0.6706, which clearly states that model is accurate and performed average while doing the data prediction.

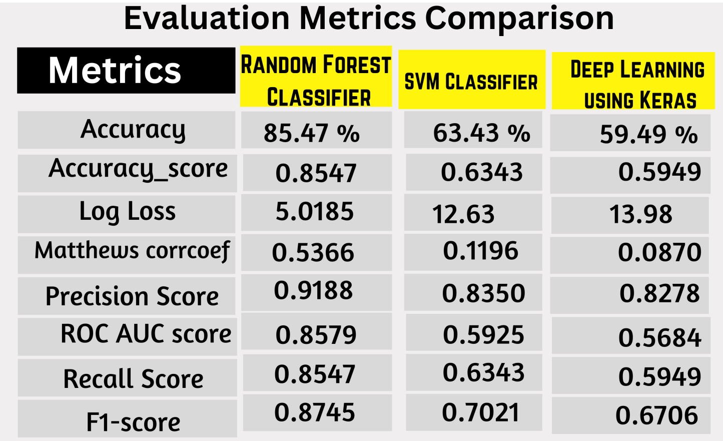

Evaluation Metrics for deep learning using keras

Metrics	Value
Accuracy	59.49 %
Accuracy_score	0.5949
Log Loss	13.98
Matthews corrcoef	0.0870
Precision Score	0.8278
ROC AUC score	0.5684
Recall Score	0.5949
F1-score	0.6706

Figure 11: Evaluation metrics of Deep Learning using Keras

6.4 Results and Discussion

A comparison between the evaluation metrics obtained for all the three different models is shown in Figure 12, where details about most of the metrics are described already in the above subsections, which describes about the conducted experiments using various classification models.



Evaluation Metrics Comparison

Metrics	RANDOM FOREST CLASSIFIER	SVM CLASSIFIER	DEEP LEARNING USING KERAS
Accuracy	85.47 %	63.43 %	59.49 %
Accuracy_score	0.8547	0.6343	0.5949
Log Loss	5.0185	12.63	13.98
Matthews corrcoef	0.5366	0.1196	0.0870
Precision Score	0.9188	0.8350	0.8278
ROC AUC score	0.8579	0.5925	0.5684
Recall Score	0.8547	0.6343	0.5949
F1-score	0.8745	0.7021	0.6706

Figure 12: Comparison of Evaluation Metrics

As seen in Fig 13 mentioned below, The ROC curve comparison between all the three models created shows that random forest gave the best performance in predicting the true positive values when compared with support vector machine classifier and deep learning using keras.

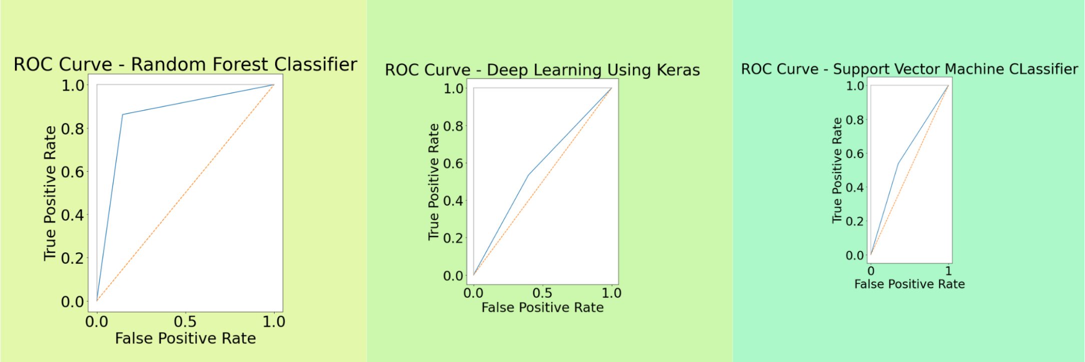


Figure 13: Comparison of ROC curves between models created

From all the above experiments conducted using machine learning models related to classification problem, it is clearly evident that Random Forest Classifier Model surpassed both Support Vector Machine Classifier and Deep Learning using Keras. This is similar to what is being discussed in the related works since Random forest was considered as the chosen model for performing classification prediction on e-commerce data with a higher accuracy of 83.64. In this experiment accuracy obtained for random forest classifier was 85.47 percentage, which is almost 22.43 percentage greater than accuracy of Support Vector Machine Classifier and 25.98 percentage greater than accuracy of Deep learning using Keras.Thus it shows an overall excellent accuracy while predicting the arrival of festival day using the user session data of e-commerce. Support vector machine classifier occupies the second position with an accuracy of 63.43 percentage. But a good precision score of 0.8350 denotes that the most of the results obtained from this prediction were found to be relevant, even-though not as much as relevant as Random forest classifier. But SVM was not able to generate a similar higher accuracy in classification as described in the related works. This is mainly due to the difference in data between the researches discussed. Also, among all the 3 models used in the prediction, deep learning using keras resulted in the lowest accuracy with a value of 59.49 percentage which is also somewhat lower than what is discussed in the related work . As already discussed in the earlier sections, the target variable SpecialDay used in this prediction has transformed to have two values 0 and 1 for the purpose of avoiding imbalance in the prediction. Also, evaluation metrics discussed in this paragraph are obtained only after performing the hyper parameter tuning through cross validation for each of the models experimented. Hence the scores obtained for each of the experimented models are more accurate with regard to this classification prediction problem.

7 Conclusion and Future Work

Machine learning algorithm plays a major role in organization nowadays in order to perform predictions . In this research, a detailed study was done to understand the extend to which classification machine learning models such as Random forest classifier, Support vector machine classifier and Deep learning using Keras can be used in predicting the arrival of festival day using e-commerce user session data. Among all the three machine learning models experimented, Random forest classifier exhibited an excellent performance by predicting the results with a higher accuracy of 85.47 percentage. Even though Support vector machine classifier and Deep learning using Keras was able to predict the results, their accuracies are found to be somewhat lower than Random Forest Classifier. Also, by performing various operation on the data through phases such as data extraction, preparation, transformation, optimization, modelling and evaluation, the objective of this research is achieved since this resulted in three different machine learning models that can help in this prediction. One of the key finding in this research is that the Random forest algorithm started to perform much better when hyper parameter tuning was done through cross validation when compared with the other modelling techniques used in this research. The limitation identified in this research is that, it could have performed much better if there are more correlated features being present in the dataset, that could have helped in improving the accuracy of the models to a better extent.

A little user demographic data, which was not included in the dataset provided, could have had a greater impact on this prediction even though the user session data of e-commerce employed in this research provides a better insight of a customers purchasing intention. Less correlation was detected between the dataset features and the target variable; the greatest positive and negative correlation values were 0.095229 and -0.102263, respectively. As a result, it is not viable to use feature selection algorithms based on correlation coefficient values.The Random Forest Classifier Model employed in the study has a n estimators parameter value of 15. Even though adding more estimators might have marginally increased expected accuracy, it also might have significantly slowed down code processing. This restriction is used in this research taking into account the resources at hand. If a better resource is made available for processing, future research initiatives might decide to employ a higher value for n estimators.

This research utilizes user session data of e-commerce in predicting the arrival of festival. Even-though it has performed a better prediction with a higher accuracy of 85.47 percentage after hyper parameter tuning, it also opens the door for various future works that can be done on top of this research. This research predicts the arrival of festivals but does not predicts the products which are sold more during the festival. Hence by developing a model that can predicts the highly sold products during a festival, it can help the company to specifically focus on products of importance thus adding more to the revenue of the organization. Moreover, a research could also be made on different factors to be considered on improving the User interface of an e-commerce platform since this prediction focuses on user session data which is completely related to the e-commerce platform. Hence, a better user interface can help the company to retain it's users for a longer time in their platform thus increasing the chances of purchase. As of now this research helps in predicting whether a festival is about to arrive or not using a target variable with binary values of 1 which says a festival is about to arrive and 0 which says there are no festivals which are about to arrive in the near future. But if this can be extended or improved in such a way that it can predict the number of days before the festival arrives, then it can be a value addition to the current research. This prediction could be achieved through regression algorithms instead of classification algorithm.

Acknowledgement

I would like to acknowledge and sincerely thank my supervisor, Prof. Cristina Hava Muntean, for her unwavering encouragement, support, and help during the course of this research. I would also like to thank her for all her answers to all the questions and clarifications asked regarding this research. I want to express my gratitude to National College of Ireland for providing all the materials required for this research through moodle and library facilities.

References

Astuti, R. and Pulungan, D. R. (2022). Analysis of factors affecting e-commerce customer purchase decisions, *MORFAI JOURNAL* **2**(1): pp. 1–20.

Baati, K. and Mohsil, M. (2020). Real-time prediction of online shoppers' purchasing intention using random forest, *IFIP International Conference on Artificial Intelligence Applications and Innovations*, Springer, pp. 43–51.

Chaudhuri, N., Gupta, G., Vamsi, V. and Bose, I. (2021). On the platform but will they buy? predicting customers' purchase behavior using deep learning, *Decision Support Systems* **149**: pp. 113–622.

Chen, C. and Li, X. (2020). The effect of online shopping festival promotion strategies on consumer participation intention, *Industrial Management & Data Systems* pp. 6–13.

Diwandari, S. and Hidayat, A. T. (2022). Predicting analysis of user's interest from web log data in e-commerce using classification algorithms, *Jurnal Ilmu Komputer dan Informasi* **15**(1): pp. 33–38.

Dong, Y., Tang, J. and Zhang, Z. (2022). Integrated machine learning approaches for e-commerce customer behavior prediction, *2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022)*, Atlantis Press, pp. 1008–1015.

Ekelik, H. and Şenol, E. (2021). A comparison of machine learning classifiers for evaluation of remarketing audiences in e-commerce, *Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi* **16**(2): pp. 341–359.

Esmeli, R., Bader-El-Den, M. and Abdullahi, H. (2022). An analyses of the effect of using contextual and loyalty features on early purchase prediction of shoppers in e-commerce domain, *Journal of Business Research* **147**: pp. 420–434.

Gordini, N. and Veglio, V. (2017). Customers churn prediction and marketing retention strategies. an application of support vector machines based on the auc parameter-selection technique in b2b e-commerce industry, *Industrial Marketing Management* **62**: pp. 100–107.

Johan, A., Rosadi, B. and Anwar, T. A. (2021). Product ranking: Measuring product reviews on the purchase decision, *Journal of Business Studies and Management Review* **4**(2): pp. 105–110.

Kao, L.-J., Chiu, C.-C., Wang, H.-J. and Ko, C. Y. (2021). Prediction of remaining time on site for e-commerce users: A som and long short-term memory study, *Journal of Forecasting* **40**(7): pp. 1274–1290.

Lim, S. H., Lee, S. and Kim, D. J. (2017). Is online consumers' impulsive buying beneficial for e-commerce companies? an empirical investigation of online consumers' past impulsive buying behaviors, *Information Systems Management* **34**(1): pp. 85–100.

Misra, G., Migliavacca, M. and Otero, F. E. (2021). Behavioural user identification from clickstream data for business improvement, *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, Springer, pp. 341–354.

Parmar, G. and Chauhan, J. (2018). Factors affecting online impulse buying behaviour, *International Journal of Education and Management Studies* **8**(2): pp. 328–331.

Piskunova, O. and Klochko, R. (2020). Classification of e-commerce customers based on data science techniques, *CEUR Workshop Proc*, Vol. 2649, pp. 6–20.

Quan, Y. (2021). Analysis of taobao single's day shopping festival from the perspective of managerial economics, *2021 International Conference on Enterprise Management and Economic Development (ICEMED 2021)*, Vol. 2, Atlantis Press, pp. 226–230.

Šneiderienė, A. and Beniušis, A. (2022). Factors influencing the decision-making of users of lithuanian e-commerce platforms, *Management theory and studies for rural business and infrastructure development: scientific journal* **44**(1): pp. 72–83.

Tzeng, S.-Y., Ertz, M., Jo, M.-S. and Sarigöllü, E. (2021). Factors affecting customer satisfaction on online shopping holiday, *Marketing Intelligence & Planning* pp. 8–18.

Virk, K. (2021). Improving e-commerce recommendations using high utility sequential patterns of historical purchase and click stream data, pp. 87–98.

Xiahou, X. and Harada, Y. (2022). B2c e-commerce customer churn prediction based on k-means and svm, *Journal of Theoretical and Applied Electronic Commerce Research* **17**(2): pp. 458–475.

Yulianto, Y., Sisko, A. and Hendriana, E. (2021). The stimulus of impulse buying behavior on e-commerce shopping festival: A moderated-mediated analysis, *Journal of Business and Management Review* **2**(10): pp. 692–714.