ID 303

Special Event Item Prediction System for Retails - Using Machine Learning Approach

WHTM Alwis# and WPJ Pemarathne

Department of Computer Science, Faculty of Computing, General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka

#35-cs-0018@kdu.ac.lk

Abstract - In the modern era, marketing, which can be defined as selling and buying, has expanded in a number of technological fields. Marketing becomes fruitful when it achieves its key points, which are called sales and profits. A most common place to see this selling and buying process is retailing. Information technology involves in various marketing fields such as in prediction processes, data analysis, item designing and profit calculations. In this study, a prediction process is primarily developed using machine learning approaches. Sales item data is analyzed to predict which items give maximum or expected profit margins and those which satisfy the customer the most. There are various machine learning approaches for aspects such as sales item prediction, prediction for item features and item price prediction. The novelty of this research is that it mainly focuses on special event items, such as those available in the Christmas season, items specialized for mothers' day, lovers' day and Vesak festival. The research process is divided into two main sub-parts; item classification and item prediction, while both processes are carried out using several machine learning approaches. Item classification is done using four supervised learning classifiers: linear support vector machine (svc), logistic regression, multinomial Naïve Bayes, and random forest classifier. Results prove SVC has maximum accuracy for classification section, accomplished using SVC machine learning approach. The prediction process has been done using the linear regression approach and according to the preferred data set, its results prove that database attribute directly affects the prediction accuracy and precisions.

***Keywords**: item classification, item prediction, special event items, retail, machine learning.*

I. INTRODUCTION

Machine learning is a centralized technological field for accomplish several tasks: data prediction, data mining, data optimization, data classification, clustering, dimensionality reduction and etc. It is an accelerated growing area of computer science and it reached applications are very complex with other technologies (Dey, 2016). Among them Data Prediction is highly influenced with machine learning approaches. There are several areas of data prediction done using machine learning technology such as sales item prediction, item design prediction, sales profit predictions, medical diagnosis prediction using it symptoms (Cancer, Virus and etc.), Bankrupt's predictions, Personality prediction using textual data, Item popularity prediction (Ex; Car popularity prediction) and etc. This prediction process prescribes using past analyzed data. Major objective of this research is retail sales item prediction for items which are sold during special events, days, or seasons. Today sales items have become a key component in business industry and sales item data growing rate is very high. There are several types of retails: clothing retails, vegetable and fruit retails, retails with all essentials, cosmetic retails and etc. In this study, two types of retails are considered: retails only with gift items and retails with both essential and gift items. Sales items gain profit for retails or retailers.

Nowadays sales item prediction is an important process. It increases the sales profit, reduces the cost overruns, and helps to fulfill the customer expectations. Day by day, the customer expectations are updated with new features. In retailers, there is an item management team specialized for the identification and analysis of

customer expectations. Because of the busy lifestyle, people tend to buy everything from one place. Examples for most famous retailers in Sri Lanka are "Arpico Super Center", gift Shops such as "Vondy Party City". In this research study, items under the category of special events, seasonal festivals, and calendric days are considered. For each and every date mentioned above has a number of specific gift items such as cakes, chocolates, flowers, ornaments, teddy bears having various kinds of features with different colors, shape, taste, number of flowers in the bouquet, band and etc. Special events relevant to retails can be classified into three categories: calendric days, seasonal festivals, and special functions.

Retail special events can categorize in to three categories. calendric day : These are type of events with a unique date or month such as mothers' day, fathers' day, children's' day, lovers' day and etc. Most probably it is an internationally celebrate day but sometimes it is limit to the country such as in Sri Lanka May Day, Independence Day. Most of the people are celebrate these event days as their like or as tradition or as a habit. For such celebrations they used to share gifts or memorable items. It will become an opportunity for retails to increase their sales and profit rate in that time period by selling items relate with specific event. As well as retails can have more benefits by get ready to coming event using item prediction process using past few years data. This proposed "Special Event Item Prediction System" will be very useful for that process. Seasonal festivals : seasonal festivals can define as another type of calendric events because it also has unique date or month in the year. "Christmas", "Vesak", "Ester", "Ramazan", "Sinhala and Tamil New Year" are some examples for seasonal festivals. People are highly intended to by gift items, decorations, and meals in this event time period. Then these special event type also will be an opportunity to have above mentioned(In calendric day paragraph) advantages. Specially seasonal festivals are type of shopping time period. Then retails should have got ready with their special inventories. Special functions : special functions do not have unique or specific date or month in the year. Birthdays, wedding anniversaries are most common examples for special functions. "Retail event management team" should give more attention to such type events because each and every day they can have sales on that event items. As above mentioned by get ready early to these type events retails can full fill customer expectations and have expected profit margins.

In the proposed work, have proposed a special event item prediction system for retails using machine learning approaches. Item classification and prediction tasks are accomplished using machine learning techniques. This system will help to increase the expected profit margins in specific events and full fill the customer expectations. Key importance and requirements of this system gained by studying about existing systems. As a result of that identified the suitable algorithms and machine learning techniques. Prepare the dataset, design the systems, implement the system, training the model and evaluate the results are the other objectives of proposed system.

II. RELATED WORKS

Predicting future customer purchases is very important and support to planning the inventory of retail, shop, or warehouse (ndr«es Mart«õnez, Claudia Schmuck, Sergiy PereverzyevJr., Clemens Pirker, Clemens Pirker, Markus Haltmeier, 2018). In paper "A Machine Learning Framework for Customer Purchase Prediction in the Non-Contractual Setting" proposed an advanced analytics tools to perform above mentioned task. Their proposed application implemented through various machine learning algorithms for binary classification. They had used three types of classification methods called: logistic Lasso regression, extreme learning machine and gradient tree boosting. These methods are totally different one from another, reason to use such methods is to increase accuracy with reasonable computational effort. From the results they had proved gradient tree boosting has highest accuracy. This prediction done for before one month to get the inventory for next month.

In item prediction or forecasting applications mostly used historical data and most important thing is what are the product characteristics chose for prediction process (F. Jiménez, G. Sánchez, J.M. García, G. Sciavicco, L. Miralles, 2016). In paper "Multi-Objective Evolutionary Feature Selection for Online Sales Forecasting "

had focused on that. They had mentioned number of feature selection and decision-making methods such as ENORA, NSGA-II and RFE. They had done tests with different machine learning algorithms and provide most suitable feature selecting method for each and every algorithm.

In paper "House Price Prediction Using Machine Learning and Neural Networks": author has done extensive study on predicting housing prices with real factors (Ayush Varma, Abhijit Sarma, Sagar Doshi, Rohini Nair, 2020). The results of research proved that this approach provide minimum error and maximum accuracy than individual algorithms applied. It considered parameters are 'square feet area', 'no. of bedrooms', 'no of bathrooms', 'type of flooring', 'lift availability' ,'parking availability' and 'furnishing condition'. As a unique approach to increase accuracy, used that the actual real estate value also depends on nearby local amenities such as railway station, supermarket, school, hospital, temple, parks etc. Author has done this study using number of algorithms such as linear regression, forest regression, bootstrap regression, and neural network.

Regression and classification problems are types of problems in supervised learning which is a paradigm of Machine Learning algorithms (Sunakshi Mamgain, Srikant Kumar, Kabita Manjari Nayak, Swati Vipsita, 2018). In paper " Car Popularity Prediction: A Machine Learning Approach" solved a real-world problem of popularity prediction of a car. In this research training data set tested with three types of classification algorithms. Such as KNN (K-Nearest Neighbor), logistic regression, random forest, and support vector machine. Training data set contained number of attributes : buying price, maintenance cost, number of doors, number of seats etc. They successfully calculated the accuracy of above-mentioned algorithms and proved the SVM provide the best result.

In paper "Using textual data for Personality Prediction: A Machine Learning Approach" focused on linear discriminate analysis, multinomial Naïve Bayes, and AdaBoost over twitter standard dataset. Personality is categorized according to the "Big Five" psychological test. This research process has gone through real-time data which can have significance with real world (Aditi V.Kunte, Suja Panicker, 2019). After preprocessing step dataset trained with above mentioned three machine learning concepts. Author has obtained multinomial Naïve Bayes has highest accuracy, precision, and recall. For sentiment analysis can use logistic regression (LR), support vector machines (SVM), decision tree (DT), boosted tree (BT), and random forests (RF). Among them SVM performance is best and it provide highest accuracy. And also, can use the boosted regression tree model and improved regression model called multilayer perceptron neural network for the item index value prediction (Sai Vikram Kolasani, Rida Assaf, 2020).

In paper "A machine learning framework for sport result prediction" focused on different field with machine learning approaches and it is very useful for system clients (Rory P. Bunker a, Fadi Thabtah b,⇑, 2019). They had analyzed research gone through artificial neural network and proposed a new framework called SRP-CRISP-DM', for sports result prediction. They had obtained data online from publicly available sources. This framework focused on result prediction for team sports rather than individual sport. For classification problems they had used regression techniques which are most common machine learning approaches.

Wangwei presented, injury analysis based on machine learning in NBA data. This article proposed a machine learning model ; random forest method to analyses the injuries of players. As the training data set, he used injuries of two teams in NBA match There for gathered data at player's level and team's level. Purpose of this research is decreasing the uncertainty of the risk in the coming match (Wu, 2020). In here mentioned by training past seasons data can predict injury events in the future. As a future work this proposed method can use in one-to—one sports such as badminton and tennis.

In paper "Applying Machine Learning to Aviation Big Data for Flight Delay Prediction" represented a domain used machine learning and big data analytics. Two datasets were used for the research process based on time performance and control quality. They had predicted the flight arrivals delay by recognizing useful patterns of the flight delay from aviation data (Yushan Jiang,

Yongxin Liu, Dahai Liu, Houbing Song, 2020). This research followed machine learning approaches are support vector machine(SVM), decision tree, random forest models and multi-layer perceptron. After testing evaluation authors had obtained multilayer perceptron based on neural network method has better performance with highest accuracy and featuring scaling.

III. CASE STUDY

Introduction had mentioned several types of special events a retail can have. In every country in the world there are gift shops, decoration shops and retails with both gifts, decorations, and also essential items. In this study as the key physical area, we had chosen Sri Lanka. In Sri Lanka there are several types of retails but it is hard to find a retail which available any type of gift items and decorations. As the primary step we had done a survey on special event items and available shops. When going through more detailed with examples, in Arpico Super Center it has mostly decoration items and gift items. Using survey results we had chosen number of special events people mostly intended to celebrate in Sri Lanka and number of item types relate with that event. Most important fact is these selected items are not pick from same retail or shop due to the rareness of such kind of shop.

IV. METHODOLOGY

This research study had done in two sections. Section one had served for dataset creation item classification using machine learning approach and section two had served for item prediction using machine learning approach. Entire research study had gone through machine learning methods. Figure 1 had given below represent the top-level architecture diagram for this study.

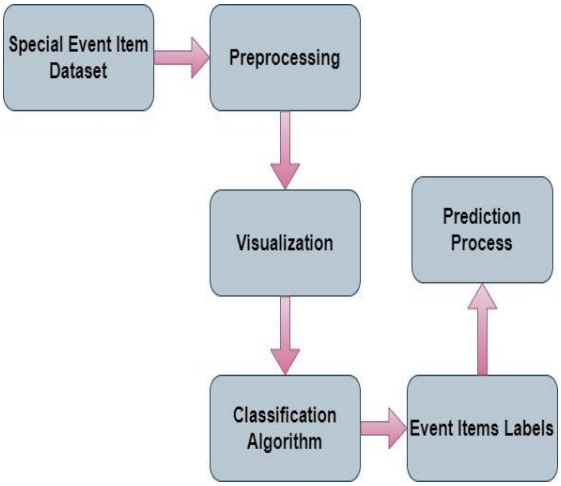

Figure 1 System Top Level Architecture Diagram

A. Section 1

1) Dataset creation.

Figure 2 given below is the created dataset using sales data of number of retails. It represented which date each and every item sold which is used to collect real-time item data and it will be useful for prediction part because special event items mostly relate with a unique date.

Item Id	Event Item Id	Item Name	Date	Quantity	Unit Price	Unit Profit
It0567	B001	Birthday Card type 1	1/2/2018	1	Rs. 85.00	Rs. 15.00
It0823	B002	Birthday CandleSet type1	1/2/2018	3	Rs. 65.00	Rs. 15.00
It1017	B004	Birthday Cake type 2	1/2/2018	1	Rs. 1,550.00	Rs. 250.00
It1024	B004	Birthday Cake type 8	1/2/2018	2	Rs. 3,250.00	Rs. 750.00
It1028	B004	Birthday Cake type 12	1/2/2018	1	Rs. 4,150.00	Rs. 900.00
It0826	B002	Birthday CandleSet type4	1/2/2018	4	Rs. 65.00	Rs. 15.00
It1022	B004	Birthday Cake type 6	1/3/2018	4	Rs. 2,600.00	Rs. 500.00
It1153	B005	Birthday BallonSet type5	1/3/2018	3	Rs. 100.00	Rs. 40.00
It0902	B003	Birthday DecoSet type4	1/3/2018	1	Rs. 1,550.00	Rs. 300.00
It0823	B002	Birthday CandleSet type1	1/5/2018	2	Rs. 65.00	Rs. 15.00
It0904	B003	Birthday DecoSet type6	1/5/2018	1	Rs. 1,550.00	Rs. 300.00
It0583	B001	Birthday Card type 17	1/6/2018	1	Rs. 85.00	Rs. 15.00
It0581	B001	Birthday Card type 15	1/6/2018	1	Rs. 85.00	Rs. 15.00
It1023	B004	Birthday Cake type 7	1/6/2018	1	Rs. 3,250.00	Rs. 750.00

Figure 2 Snapshot of Dataset

As given in figure 2 first column is “item Id” which provide unique Id for each and every item in dataset. Almost in retails also item named with unique Id which is easy for managing data. Second Column is “event Item Id”, a special event not only has one item type there are several numbers of item types and this column used to categorize item types in each and every event. As an example, Event name “Christmas Season” it contained items with event item Id “ Christmas Tree - Xm001”, “ Christmas Deco Balls - Xm002”. Rest of columns represent the “Item name”, “Date”, “Quantity”, “Unit Price” and “Unit Profit”.

2) Item Classification.

After creating the dataset, built up the item classification part. Classifying items is an important task in this study. For prediction process need to identify what are the items sold in specific event. There are number of machine learning algorithms for item classification which are almost contained under supervised learning algorithms. For the current work we had used four types of classification algorithms: linear support vector classification (SVC), logistic regression, multinomial Naïve Bayes, and random forest classifier. As input data for classification part used "Event item Id" and "Item name". Classify the each and every event item with its specific event is the output of this classification process.

a) Linear Support Vector Classification.

SVMs (Support Vector Machines) are a useful technique for data classification. If simply describe the classification process of SVM, transform data to the format of an SVM package (Martin Kappas , Phan Thanh Noi , 2017), randomly try a few kernels and parameters and Test. This method most effective in high dimensional spaces. This classification had gone through multi class classification. Briefly can talk with X and Y coordinates, X – number of samples and Y - number of labels.

b) Logistic Regression.

This method is very productive when the dependent variables are categorical (Mohammad Ali Mansournia, Angelika Geroldinger, Sander Greenland, Georg Heinze, 2018). In this study dependent variable is item name. It had done by naming an item with its event name. Example: event name "Birthday", item type "Birthday cake". In this method multiclass classification also same as binary classification. In classification specific event item Id denoted – "1" and other event item Id s denoted – "0".

c) Random forest Classifier

Random forests are enhancement of decision trees, those are consisting with bunch of independent decision trees (Dragutin Petkovic, Russ Altman, Mike Wong and Arthur Vigil, 2018). There are number of methods for get outcomes from this classifier. This study had focused on permutation feature importance. This is selecting a column (i.e., feature) in the validation set, then shuffling it randomly, for destroying the correlations between that feature and all the other features used by this model to make its predictions, and finally measuring the model's performance on this freshly shuffled validation set.

d) Multinomial Naïve Bayes.

This is a type of Naïve Bayes classifier. This is a probabilistic machine learning model and based on the Bayes theorem (NimaShiri Harzevili, Sasan H.Alizadeh, August 2018).

Bayes Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Figure 3 Equation for Bayes Theorem

According to this study A – class variable (Event Item Id). Variable B – represent the item name.

B. Section 11

After item classification process next step is item prediction. Prediction task feature effectiveness had discussed using linear regression machine learning technique.

Linear Regression Algorithm.

Regression is a way to modelling a target value based on independent predictors. This method is based on independent variables and find out relationships between independent and dependent variables (Amand F.Schmidt, ChrisFinan, 2018 June). If the relationship between these two variables is linear it is known as linear regression. In this study according to preferred data set "Unit Price" is independent variable and dependent variable is "Unit Profit". When give the amount of items want to select it provide items which are in first set.

V. RESULTS AND DISCUSSION

A. Section 1

For the given dataset focusing on above mentioned four classification algorithms. Results of these are discussed below.

a) Comparison of Accuracies.

By comparing the accuracy results can proved linear support vector classification(SVC) has the highest accuracy.

Linear SVC).: 0.895234, Logistic Regression.: 0.868313, Multinomial Naïve Bayes.: 0.855413, Random Forest Classifier.: 0.617169

Figure 4 given below represent the data visualization of accuracy with specific algorithm.

Given in the figure 5 is the graphical representation of accuracies for above mentioned four classification algorithms. Here represented only some selected item types.

Given in the figure 6 is the graphical representation of precisions for above mentioned four classification algorithms. Here represented only some selected item types.

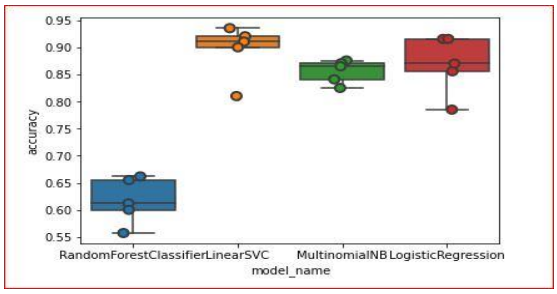

Figure 4 Accuracies of Applied Algorithms

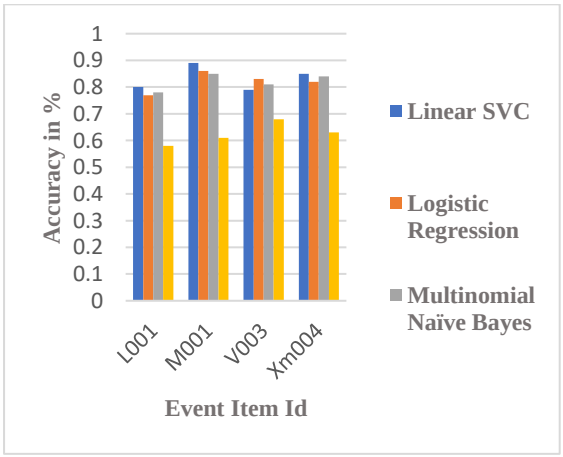

Figure 5 Accuracy Comparison of Classification Algorithms

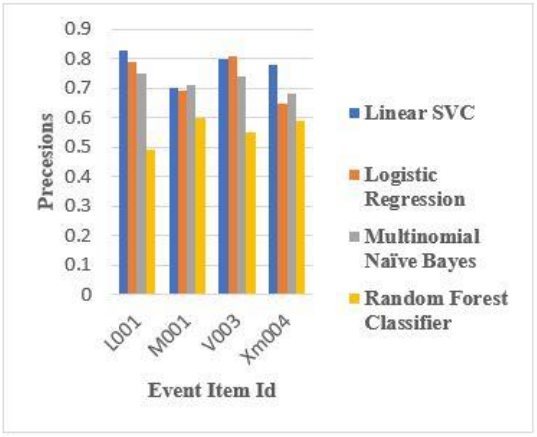

Figure 6 Precision Comparison of Classification Algorithms

a) Classification model evaluation.

When straight forward with best model (SVC) it can represent using confusion matrix and show the discrepancies between predicted and actual labels. The vast majority of the predictions end up on the diagonal (predicted label = actual label). If there are items which are touch more than one event then it caused to have misclassifications. In current data set there is no such kind of issue. But it can solve using python programming knowledge. Figure 4 given below represent the continuous matrices.

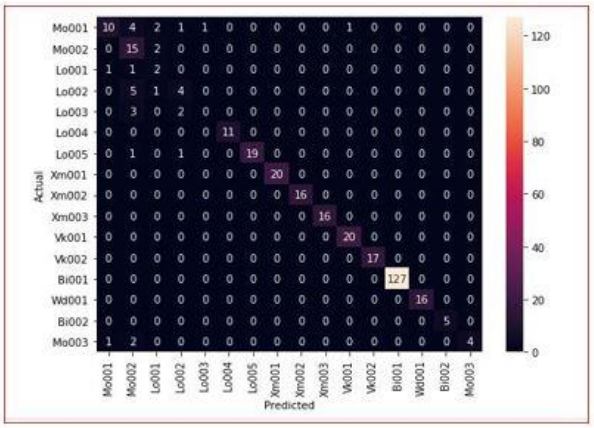

Figure 7 Continuous Matrices of Labels

B. Section 11

After classification next step is data prediction process. As machine learning technique linear regression algorithm has used and it implies numeric data is the most important feature in the dataset to achieve reliable results. For data testing process, 20% data has used and for data

training process, 80% has used. When consider the above-mentioned data set as the independent variable can used two columns which are contained numeric data called “Unit Price” and “Quantity”. But its accuracies are variant from each other. “Unit price” accuracy is 0.94% and “Quantity” accuracy is 0.005%. Then in this study “Unit price” use as the independent variables. Figure 6 provide a clear visualization about accuracies.

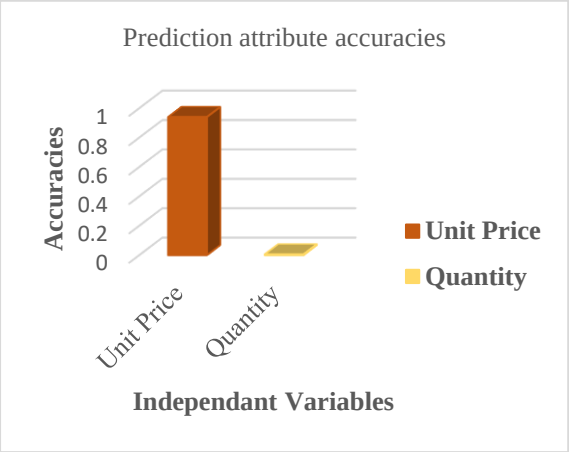

Figure 8 Comparison of Independent Variables accuracies.

After comparing the results can discuss about the accuracies, precisions using performance of algorithms. After executing the machine learning algorithms next tread is to find out the effectiveness of model based on various performance metrics. Different performance metrics are used for different machine learning algorithms. For example: For classification we use different performance metrics such as Accuracy, Precision, Recall, and f1 Score. From the above mentioned four algorithms SVC has the highest accuracy and precision. linear regression technique used for item prediction when use the “Unit Price” as its independent variable prediction accuracy become very high.

VI. CONCLUSION

Because of unfitted gift item or decorations mostly retails unable to achieve their profit margins and fulfill the customer expectations during special functions calendric days and seasonal festivals. This research study has presented a retail special event item prediction method using machine learning approach. First step is classifying the special event items for specific event. The item classification tasks accomplished using four supervised learning algorithms called liner support vector classifier(LSV), logistic regression, multinomial Naïve Bayes, and random forest classifier. According to results LSV algorithm gives the maximum accuracy(0.89) and precision(0.83). Second step is prediction process. Prediction Process had followed through linear regression technique and its accuracy highly depend on independent variables. Unit price and quantity are the two independent variables used to test the accuracy of prediction process. According to results unit price independent variable gives the highest accuracy(0.94). As a key result of this study can conclude linear regression algorithm which is use for prediction processes highly depend on numeric data in the dataset. Gained the expected profit margins and full fill customer expectations during special events are the main purposes and advantages of proposed system. As the future work for prediction process can do considering item features or geographical locations.

REFERENCES

Aditi V.Kunte, Suja Panicker, 2019. *Using textual data for Personality Prediction:A.* GLA University, Mathura, UP, India, s.n.

Amand F.Schmidt, ChrisFinan, 2018 June. Linear regression and the normality assumption. *Journa of Clinical Epidemiology,* Volume 98, pp. 146-151.

Ayush Varma, Abhijit Sarma, Sagar Doshi, Rohini Nair, 2020. *House Price Prediction Using Machine Learning.* s.l., s.n.

Dey, A., 2016. Machine Learning Algorithms: A Review. *International Journal of Computer Science and Information Technologies,,* Volume Vol. 7 (3), pp. 1174-1179.

Dragutin Petkovic, Russ Altman, Mike Wong and Arthur Vigil, 2018. Improving the explainability of Random Forest classifier – user centered approach. *Biocomputing,* pp. 204-215.

F. Jiménez, G. Sánchez, J.M. García, G. Sciavicco, L. Miralles, 2016. Multi-objective evolutionary feature selection for online sales forecasting. *Neurocomputing.*

Martin Kappas , Phan Thanh Noi , 2017. Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover

Classification Using Sentinel-2 Imagery. *Sensors,* p. article number 19.

Mohammad Ali Mansournia, Angelika Geroldinger, Sander Greenland, Georg Heinze, 2018. Separation in Logistic Regression: Causes, Consequences, and Control. *eparation in Logistic Regression: Causes, Consequences, and Control,* Volume 187, pp. 864-870.

ndr«es Mart«õnez, Claudia Schmuck, Sergiy PereverzyevJr., Clemens Pirker, Clemens Pirker, Markus Haltmeier, 2018. A Machine Learning Framework for Customer Purchase Prediction in the Non-Contractual Setting. *European Journal of Operational Research,* 239(3, 2014 Dec).

NimaShiri Harzevili, Sasan H.Alizadeh, August 2018. Mixture of latent multinomial naive Bayes classifier. *Applied Soft Computing,* Volume 69, pp. 516-527.

Rory P. Bunker a, Fadi Thabtah b,⇑, 2019. A machine learning framework for sport result prediction. *Applied Computing and Informatics,* Volume 15, pp. 27-33.

Sai Vikram Kolasani, Rida Assaf, 2020. Predicting Stock Movement Using Sentimen Analysis of Twitter Feed with Neural Networks. *Journal of Data Analysis and Information Processing,* Volume 8, pp. 309-319.

Sunakshi Mamgain, Srikant Kumar, Kabita Manjari Nayak, Swati Vipsita, n.d. Car Populairty Prediction Using Machine Learning Approach. *2018.*

Wu, W., 2020. Injury Analysis Based on Machine Learning in. *Journal of Data Analysis and Information Processing, 2020,* Volume 8, pp. 295-308.

Yushan Jiang, Yongxin Liu, Dahai Liu, Houbing Song, 2020. *Applying Machine Learning to Aviation Big Data.* s.l., s.n.

AUTHOR BIOGRAPHIES

Mrs. Punsisi Pemarathne is a Senior Lecturer/Researcher in Computer Science at the Department of Computer Science, General Sir John Kotelawala Defence University. She has obtained her BSc Computer Systems and Networks and her master's in computer and Network Engineering from Sheffield Hallam University in United Kingdom. She has completed her MPhil in Computer Science from University of Sri Jayewardenepura, Sri Lanka. Her research interests include, Swarm Robotics, Artificial Intelligence, Mobile and Wireless Communications, Network Security, Evolutionary Computing, and Internet of Things.

W.H.T.M. Alwis an undergraduate student in the General Sir John Kotelawala Defence University, Sri Lanka and will be graduating in 2022 with a BSc Hons. In Computer Science. Her research interest is in the field of Machine Learning.