E-commerce data analysis and visualization

Using Random Forest Regression and Prophet model E-commerce sales prediction.

Tasnim Rafia, Hasan Tanveer Mahmood, Billah Syed Mashkur, KM Zubair
Department of Computer Science, Kulliyyah of Information and Communication Technology,
International Islamic University Malaysia, Kuala Lumpur, Malaysia.
tasnim.rafia777@gmail.com, tanveer.mahmood2489@gmail.com, mashkurbillah@outlook.com , zubair.k@live.iium.edu.my

Abstract— The motivation behind this study is to better understand the ecommerce sector by examining public datasets. In our case, we'll work on Pakistan's Largest E-commerce dataset that we found from kaggle. We'll work on problems like finding the most sold products in the store to develop a better business model, most popular categories in terms of sale, demand for products that are in high rising phase, products that get cancelled the most and predicting upcoming year's sales condition to analyze the growth rate of sale in different timespan. We'll use Apache Spark, an open-source cluster-computing technology, as well as Tableau, which is also crucial for the tasks at hand. This research could be applied to commodity assessments of purchased things on e-commerce sites to do sentiment analysis.

Keyword: **E-commerce, Sales Analysis, Upcoming sales prediction, fbProphet model, RandomForestRegressor**

I. INTRODUCTION

This project is primarily trying to analyse one E-commerce dataset (Pakistan's Largest E-Commerce Dataset), where the product values will be analysed along with the popularity by comparing the sales history among themselves. According to the analysis, some suggestions to undertake are going to be provided in order to increase the popularity of the company by forecasting the sales condition of next year with the help of time series analysis in terms of the dataset history such as product status, the most sold product history etc. that is already accorded in the dataset. Much research has been already conducted using public datasets, still many business questions yet to be answered from realtime dataset. The purpose of this research is to understand the ecommerce domain better by analysing the public datasets. An overall depth analysis will be performed by comparing the product status history so far. The main aim of the project is to implement a solution to assist the company to enhance their popularity. Therefore, this research will propose a strategy on how they can improve their sales.

II. RELATED WORKS

Yang et al. [1] proposed a paper for Sentiment analysis for e-commerce product reviews. Sentiment analysis for product evaluations is a way of automatically estimating the customer's emotional inclination by analysing subjective remark language with the customer's emotional colour. They've used rule-based methods and lexicon-based methods to construct a sentiment lexicon using degree adverbs, sentimental intensity, words and polarity. They develop a novel sentiment analysis model based on the benefits of word vectors, sentiment lexicons, GRU, CNN and test it on a book review dataset from a genuine e-commerce website. To begin, the sentiment lexicon is employed to improve the sentiment elements in the reviews. The CNN and the Gated Recurrent Unit (GRU) networks are then utilised to extract the major sentiment and context characteristics in the reviews, which are then weighted using the attention method. Finally, categorise the weighted sentiment characteristics. Zhang et al. [2] proposed a Sentiment analysis-based Fuzzy decision support model for e-commerce item comparison. To describe online reviews, the model uses probability multivalued neutrosophic linguistic numbers (PMVNLNs). By taking into account neutral information and reluctance in text reviews, it overcomes the limitations of previous models. The suggested characterisation technique, unlike existing item rating algorithms, incorporates positive, neutral, and negative information in text evaluations and reflects review reticence. Because of the characteristics of item comparison difficulties in e-commerce, the model proposes QUALIFLEX. Using the integrated regret theory-QUALIFLEX, it addresses customers' constrained rational behaviour, unlike previous models.

III. DATA DESCRIPTION

The data was obtained via Kaggle's website, which offers complementary datasets that may be viewed and used for additional research. We choose Data analysis of Pakistan's largest e-commerce dataset that has 1048574 rows and 26 columns. From July 2016 to August 2018, the dataset contains sales information such as product quantity, payment method, order status, and so on.

IV. EXPERIMENTAL SETUP

This section of the paper covers the stages in detail, as selecting the right things to sell is not always simple. There are numerous elements that influence whether a product succeeds or fails, and one split choice could mean the difference between massive success and comparable disaster. Thus, e-commerce data analysis works by forecasting changes in a customer's buying behaviour as a result of qualitative products, seasonal sales and delivery performance, and a variety of other elements. The study will give effective monitoring analyses that will assist merchants in anticipating potential purchasing impulses and capitalising on trends while sustaining sales growth rates.

A. *Algorithm - Random Forest Regression / Prophet*

In this model Random Forest Regression machine algorithms will be applied. It is a supervised learning algorithm which uses the approach of ensemble learning for classification and regression. There is no inter relation between these trees while constructing the trees. It works by building a multitude of decision trees at training time and gives output which is mode of classification and regression. It can combine multiple predictions.

Prophet is a time series data forecasting process based on an additive model that fits non-linear trends with yearly, weekly, and daily seasonality, as well as holiday impacts. It can be implemented best with time series with substantial seasonal influences and historical data from multiple seasons. Prophet is forgiving of missing data and trend shifts, and it usually handles outliers as well.

We have used both methods and tried to figure out which methods provide the most accuracy and fits with datasets best. We have split data in a ratio of 7.5:2.5.

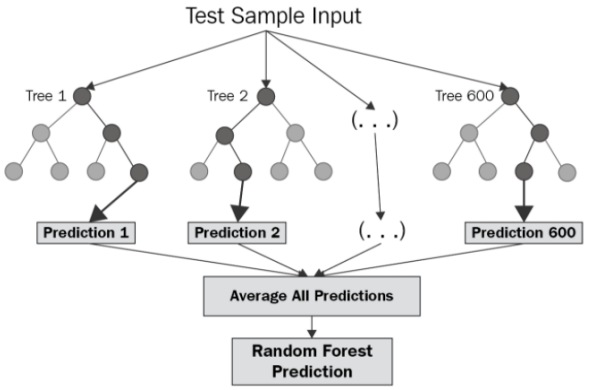

image source: Random forest regression

B. *Performance Evaluation*

We'll train our model using random forest regression, which will look for a correlation between prior sales rates and months. Predictive analysis of sales growth will be conducted using the trained model. It basically trained with the bagging method. The basic idea of this bagging method is a combination of learning models to increase overall results.

Prophet algorithm is used to conduct tasks on time series data and large volumes of data. From commonly purchased item sets, the prophet algorithm can build association rules. Following the analysis of all of these factors, a visual depiction of our study findings will be presented in order to make the research findings clear to people from all walks of life.

C. *Data Cleaning*

The datasets to be utilised are first reviewed and, if required, cleansed. Several data cleaning procedures were used to guarantee that the datasets were suitable for testing. The following graphic depicts the data cleaning procedures that were performed using Google collab:

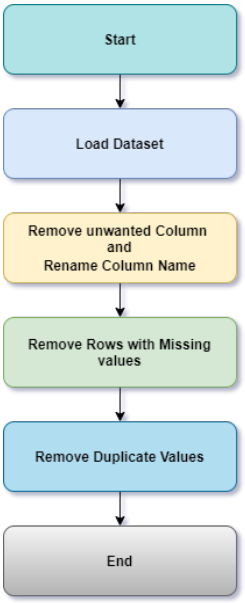

Figure 1: Flowchart for data Cleaning and Preprocessing.

To begin, the dataset's directory is specified. After that, the dataset is loaded into Google collab. For our project. Raw data will have missing values, columns containing information that we do not want, and some of them may be duplicated.To delete any columns that we do not need for our project, we have used the instruction datasets.drop (['column_name'], axis='columns', inplace=True) where datasets is the name of our data and [‘column_name’] is the column that we want to remove, in this case columns ‘sales_commission_code', 'Unnamed: 21', 'Unnamed: 22' and so on we’ve removed.

In addition, we tried to eliminate the rows with missing values, So, we used the python command .isnull().sum() to see the null values or missing values and .notna() to remove missing values from the dataset. We only want the completed version of the dataset with no missing values since we want to achieve an accurate result. By using

sns.heatmap we've visualized the null values column.

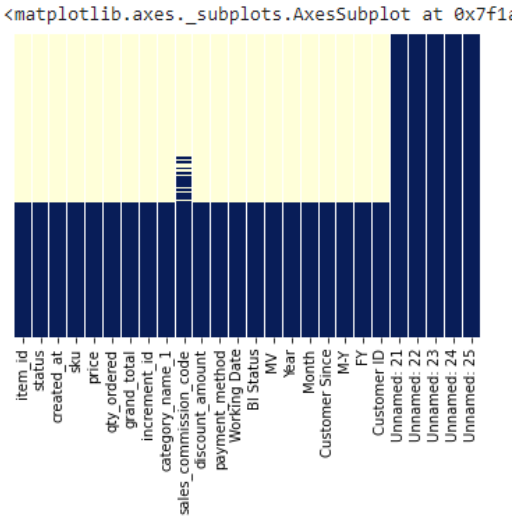

Figure 2: Heatmap before removing null values.

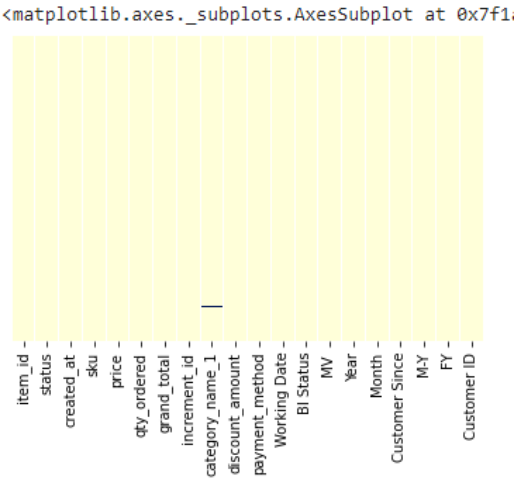

Figure 3: Heatmap after removing the null values and unnecessary column

V. MODELLING

We will move on to the modelling section after cleaning the datasets and selecting the sample to utilise. The Algorithm chosen for the dataset is "Random Forest And Prophet".

	count	mean	std	min	25%	50%	75%	max
Item ID	584524.0	565667.074218	200121.173648	211131.0	395000.75	568424.5	739106.25	905208.0
Price	584524.0	6348.747531	14949.269515	0.0	360.00	899.0	4070.00	1012625.9
Quantity	584524.0	1.296388	3.996061	1.0	1.00	1.0	1.00	1000.0
Grand Total	584524.0	8530.618571	61320.814625	-1594.0	945.00	1960.4	6999.00	17888000.0
Discount Amount	584524.0	499.492775	1506.943046	-599.5	0.00	0.0	160.50	90300.0
Year	584524.0	2017.044115	0.707355	2016.0	2017.00	2017.0	2018.00	2018.0
Month	584524.0	7.167654	3.486305	1.0	4.00	7.0	11.00	12.0
Customer ID	584513.0	45790.511965	34414.962389	1.0	13516.00	42856.0	73536.00	115326.0

Figure 4: Data Description

The modelling process is divided into several steps. We follow this flowchart below.

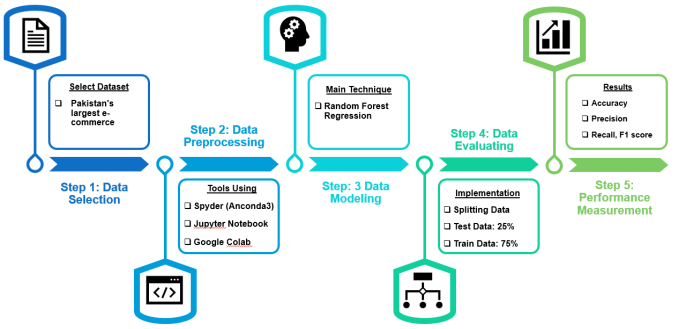

Figure 5: Flow on Modeling Steps

The first step, as shown in Fig. 3, is to import the dataset into Google Collab. Data preprocessing is required before modelling can begin in order to learn more about the dataset. The goal of exploratory data analysis is to find analytical graphs and the distribution of the data. After EDA, the dataset is randomly divided into a training set and a test set. The training set should account for 75% of the total dataset, whereas the test set should account for 25%. The Random Forest Regression is then applied to the training data. The number of clusters chosen is determined by the results of prior internal evaluation.

A. Exploratory Data Analysis (EDA)

a) Correlation Test

Following clustering on the training set, a correlation test is performed to determine whether there is a strong, weak, or no association between two variables. The direction of the association between the variables is also revealed by the correlation test heatmap. The correlation coefficient is usually between +1 and -1, with the positive side indicating a stronger degree of relationship. When the coefficient hits one, the correlation is said to be perfect. Pearson and Spearman Rank correlations are examined in this dataset.

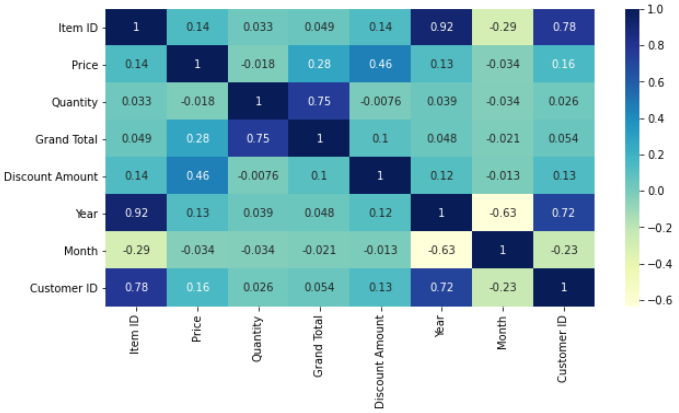

Figure 6: Correlation between rows and columns.

b) Filtering Columns for Visualization and Analysis:

category_name_1	no_of_item
cod	7
School & Education	3477
Entertainment	26326
Kids & Baby	16485
Computing	15933
\N	7850
null	175
Mobiles & Tablets	115710
Beauty & Grooming	41496
Health & Sports	17502
Soghaat	34011
Books	1870
Others	29218
Payaxis	4
Superstore	43613
Men's Fashion	92220
Appliances	52413
Women's Fashion	59721
Home & Living	26504

Figure 7: Filtered column (Name of the categories by summing up the item number for each category)

	customer_id	total_cancelled
0	39707	496
1	26527	355
2	28896	312
3	10654	309
4	12278	290

Figure 8: Customers with total cancelled order

	customer_id	total_Not_cancelled
0	5769	185
1	163	172
2	36	158
3	800	157
4	820	155

Figure 9: Customers with total confirmed order

c) Creating plots for visualization:

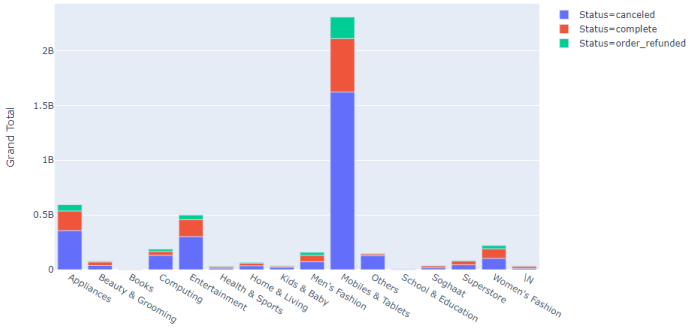

Figure 10 : The categories receiving the most cancellation.

B. Train Dataset

After EDA has been used to analyse the dataset, it is divided into two sets: one for the training phase and one for the testing phase. The data is divided into 7.5 : 2.5 ratios, with 75 percent of the data serving as the training set and the remainder serving as the test set. Random Forest Regression is used to implement the training set.

VI. ANALYSIS OF RESULTS

A. Hypothesis Testing

We assumed previously from the analysis that, *"One Particular Period will Provide a higher Sales"* The line graph shown in the prediction model agrees on the point that at a specific time, in the last quarter of 2018, the sale will increase.

B. Model Testing

We have used two models for the prediction in our project which are Random Forest Regression model and Prophet model. Both models have been considered in terms of their r2, mean squared error, mean absolute score. As the result of justification, we have preferred the Random Forest method to forecast the future sales condition.

C. Results and Analysis:

a) Descriptive Analysis:
The research question for descriptive analysis was, *"Which category has the most popularity in terms of sale? "*

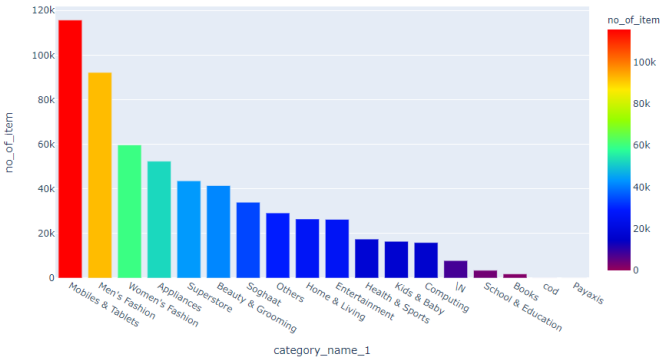

Figure 11 :Result of descriptive analysis.

From the analysis, we found that the most sold category is *'mobiles and tablets'*. To simply say, that is the most popular item according to the analysis.

b) *Diagnostic Analysis:*

Regarding the diagnostic analysis, the research question we selected is, *"How was the sales condition throughout the time from July, 2016 to July, 2018 ? "*

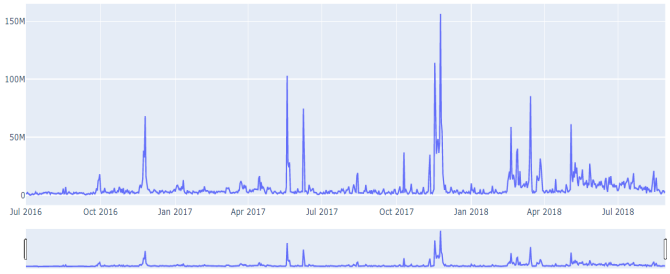

Figure 12: Result of diagnostic analysis

As from the investigation, the figure below shows the sales state of the past 3 years (2016 to 2018).

c) *Causality Analysis:*

For causality analysis, the research question was, *"Does the item cancellation get affected by time or customer ?"*

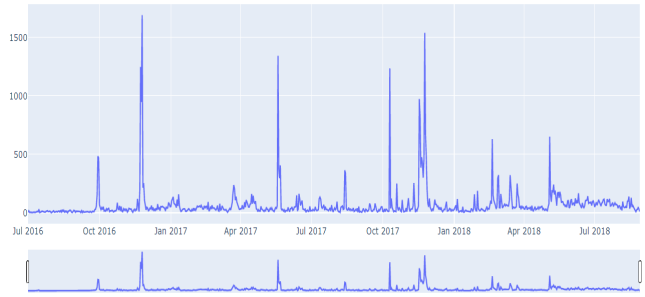

Figure 13: Item cancellation based on time

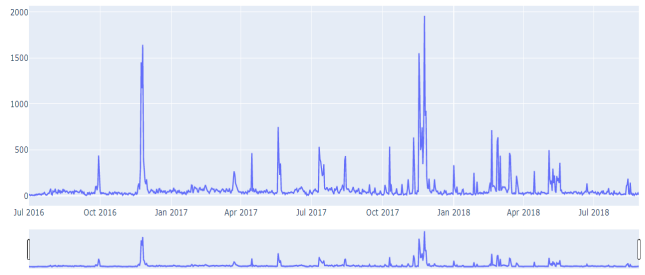

Figure 14: Items that not cancelled based on time

We were determined to discover the answer whether the item cancellation depends on some particular time of a year, or it is some other reason such as customers etc.

The result found that cancellation of items in the *last quarter* of year is really significantly high. So, *'time'* might be a reason for item negation.

	customer_id	total_cancelled	total_Not_cancelled
0	39707	496	18
1	26527	355	77
2	28896	312	55
3	10654	309	35
4	12278	290	21

Figure 15: Customers who order and cancel the most

Again, while scrutinizing the customer history of item negation, we found that there are some customers who order the most; whether they cancel the orders the most as well.

d) *Exploratory Analysis:*

"Which product demand is in a high-rising phase? "

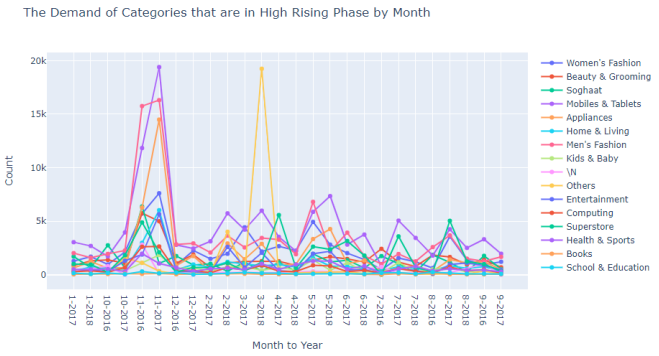

Figure 16: Product demanding phase by month

In this analysis, we found the popularity of products differ from month to month in terms of some products such as, mobile and tablets remaining at the peak of demand in the last quarter, when others are less noticeable. On the other hand, in the first quarter, other items are being sold out the most. Again, there are some variations as well like, school stationery, books are being bought at a continuous same rate over the year.

e) *Predictive Analysis:*

We have made some predictions with the next year's sales state condition which is based on this research question: *"How is the sales state going to be next year ?"* We have used two models for the prediction which are *'RandomForest Regression Model'* as well as *'Prophet Model'*

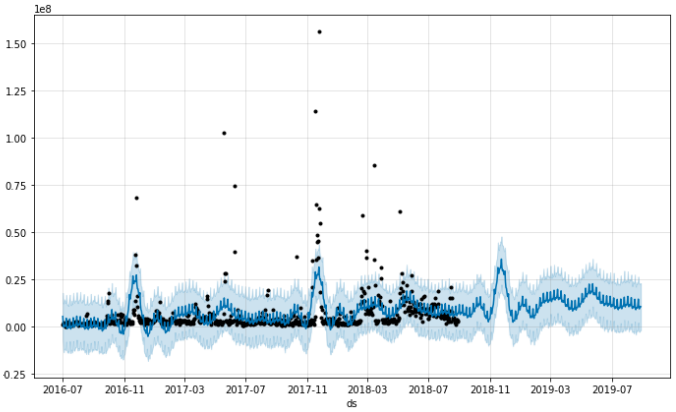

Figure 17: Predicted sales state of next year (2019)

```
r2_score: 0.4277885705891772,
mean_squared_error: 4.700391261310883e+16,
mean_absolute_error: 138643103.93385413
```

Figure 18: r2, MSE, MAE scores using RandomForestRegressor

Using the RandomForest regression model, we got the r2 score 0.4277885705891722, the mean squared error is 4.700391261310883e+16, and the mean absolute score is 138643103.93385413

```
r2_score: -0.6234392611467166,
mean_squared_error: 32187147347009.316,
mean_absolute_error: 4901326.587481598
```

Figure 19: r2, MSE, MAE scores using Prophet Model

Using the fbProphet model, we achieved the r2 score of -0.6234392611467166, the mean squared error of 32187147347009.316, and the mean absolute score of 4901326.587481598.

From the score (r2, MSE, MAE) stated above, after proper justification, it may be declared that, RandomForest regression model worked better in this case in terms of accuracy and nicer prediction.

VII. CONCLUSIONS

Our aim was to suggest a solution for the E-commerce sales site in order to increase their profit. We have focused on some certain analysis and results to discover the solution. At the end of the research, we have acknowledged a couple of solutions.

An important tip would be to start a reward and penalizing system. Because, the customers ordering most, are cancelling their orders at large. So, the customers who are ordering might get promotional offers as reward while the users cancelling more than the average or 100 times might have to accept some penalization. This technique, so far, will be able to reduce the amount of cancellation that is supposed to bring more profit for the E-commerce company.

Furthermore, the sales condition, at some particular period of year, is being defined by some certain products' popularity such as, demand for books and school stationery products remain almost same over the whole year. But in the last quarter, while mobile and tablets, men's fashion and appliances are altering the acceptance, it can be advised to the company to concentrate more on the items other than mobile and tablets in the last moments of years.

VIII. ACKNOWLEDGMENT

This project was conducted as part of the Big Data Analytics (CSC 3303) course offered by the Department of Computer Science at the International Islamic University Malaysia in Kuala Lumpur. The authors of the study would like to express their gratitude to Dr. Sharyar Wani for his assistance in making this work successful.

IX.REFERENCES

1. L. Yang, Y. Li, J. Wang and R. S. Sherratt, "Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning," in IEEE Access, vol. 8, pp. 23522-23530, 2020, doi: 10.1109/ACCESS.2020.2969854.

2. P. Ji, H. -Y. Zhang and J. -Q. Wang, "A Fuzzy Decision Support Model With Sentiment Analysis for Items Comparison in e-Commerce: The Case Study of http://PConline.com," in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 49, no. 10, pp. 1993-2004, Oct. 2019, doi: 10.1109/TSMC.2018.2875163.