The Impact of Instagram Marketing on Sale in the Fashion Industry

Rachel Yi Shan Tang
School of Computing
Asia Pacific University of Technology and innovation (APU)
Kuala Lumpur, Malaysia
tp062404@mail.apu.edu.my

Mafas Raheem
School of Computing
Asia Pacific University of Technology and innovation (APU)
Kuala Lumpur, Malaysia
raheem@apu.edu.my

***Abstract*— There is an ongoing debate on Instagram's capabilities in sales generation as a social media-based marketing tool amongst business executives. Although an abundance of research was done to determine the impact of social media marketing on sales contribution, many of these studies have utilized purchase intention as a proxy for an actual purchase. This then creates a gap as purchase intention is merely the likelihood of an actual purchase and thus may be inaccurate when used to measure actual sales. This study then aims to investigate the impact of Instagram marketing on sales in the fashion domain via a data analytical approach to narrow the existing gap. A CRISP-DM framework is adopted, which includes a descriptive and predictive approach, in achieving the data mining goal of determining the impact of Instagram marketing on sales using Instagram and sales data of fashion retail in Klang Valley, Malaysia. The implementation of a data analytical approach in investigating the impact of Instagram marketing on sales was able to achieve all research objectives in determining the ability of Instagram marketing in influencing sales. In this line, both the XGBoost and LSTM models were able to predict sales using Instagram marketing factors whereas the LSTM model performed better with lower MAE and RMSE values. In future, studies can be conducted with more Instagram features with additional modelling techniques to gain better results than those obtained in this study.**

Keywords*— *Data Analytical, Instagram, machine learning, sales, social media marketing.

I. INTRODUCTION

INSTAGRAM is one of the famous social media platforms with immense popularity and ongoing growth. As of 2021, the social media platform recorded more than 1 billion active users with an average usage time of 30 minutes per day (Barnhart, 2021). 90% of Instagram users follow at least one business account and 81% of users utilize the platform as a source to gather information on products. Statistics show that Instagram is a potential and effective online marketing channel (Cui, et al, 2019). The visual focus is the key to Instagram which made it an ideal platform for businesses to showcase and market visual-based products such as fashion products which belong to one of the most popular business domains on Instagram. According to a statistic report on the top hashtags of Instagram, #fashion (812.7 million) stands third with #love (183.5 billion) and #instagood (1.150 billion) (Top-Hashtags, 2021), showing the popularity of fashion content, making it an ideal channel to market fashion products. Therefore, fashion was thus selected as the subject matter of this study.

Over the decade, Instagram has consistently developed different features (also known as products) that are beneficial to businesses in enhancing digital marketing activities. These features include HashTags (2011), video sharing (2013), Instagram Stories (2016), live videos (2016), Instagram TV (IGTV) (2018), and Reels (2020). Instagram has also catered to business use of the app by introducing different business tools such as advertising services (2013), business profile accounts, insights, analytics (2014), audience demographic, post impressions and post reach (2016). Despite the exclusive suite of business tools made available for businesses to analyze and measure the performances of marketing content, marketing managers may find it challenging to make sense of these metrics and understand their true revenue contributions (Gruber, et al, 2015) (Grave, 2019). The issue is especially apparent in measuring sales conversion from Instagram marketing, in other words, whether a sale is made under the influence of Instagram. As such, the purpose of this study is to identify the true impact of Instagram marketing on sales.

Along with other variations of social media platforms, Instagram marketing has been widely adopted as a vital digital marketing approach for businesses. Nearly 59% of business executives are keen to increase their social media marketing budget soon believing that a positive social media experience can be converted into sales (Globe Newswire, 2021). Bango (2021) found the perspectives of Chief Executive Officers (CEOs) proclaiming the unworthiness of metrics (i.e., impressions, reach) and that these measurements do not directly reflect on sales revenue. Although many studies were conducted to rectify the effects of social media marketing aspects (i.e., social media content, image quality) in increasing sales, purchase intentions are often used as a proxy for actual purchase decisions. The problem lies within the appropriateness of purchase intention in indicating actual purchases where purchase intentions are a vague measurement of the likelihood of an actual purchase (Tsai, 2020). Additionally, a customer's purchase intention may be subjected to change from the consistent search for information, evaluating alternatives or stumbling upon negative word-of-mouth. Therefore, businesses must identify the true impact of Instagram marketing on actual purchase decisions instead of purchase intentions which can help to make more efficient and effective marketing decisions.

This study aims to explore the impact of Instagram marketing on sales using Instagram metrics and features via a data analytical approach.

The scope of this study is limited to small enterprises in Klang Valley (Malaysia), that have an omnichannel (i.e., online store, physical store) of the fashion domain. Additionally, the findings of this research are limited to the variables that are present in the dataset used. This is such that businesses that are interested to determine the impact of Instagram features such as Instagram Live or IGTV on sales may not find this study relevant as the business case of this study does not include such Instagram features as part of their Instagram marketing strategy.

II. Related Work

A. Social Media Marketing

Social media is a highly influential communicative social networking technology in the contemporary digital era. The ever-growing popularity of social media complements an enormous user database, making it an ideal and attractive platform for businesses to leverage as a marketing platform and tool (Jacobson, et al., 2020). Businesses are willingly engaging on social media as an effort to communicate (i.e., information search, promotions) and build relationships with consumers (i.e., brand equity, brand engagement) (Zolo, et al., 2020) (Li, et al., 2020). The goal of these efforts is ultimately to encourage consumers to generate sales in a return (Dwivedi, et al., 2020). Although the wide adoption of social media as a marketing medium and tool for businesses, many marketers face challenges in evaluating social media marketing objectives due to the novelty and the constant change being made on social media (i.e., adding more features, policy changes) (Dwivedi, et al., 2020) (Misirlis, N., & Vlachopoulou, 2018).

There are many different social media marketing mixes and approaches that a business can implement. Past research has shown that social media marketing can be effective and efficient if the selected social media marketing platform complements the target audience, product, and marketing strategy of a business. A study found insisting on the need for different marketing strategies for different social media which indicates the importance of selecting the right social media platforms for businesses to be present on (Kusumasondjaja, 2018). The study has highlighted that marketing on Facebook and Instagram is more effective when the contents created are more interactive and entertaining whereas an informative appeal is the most effective on Twitter but not on Facebook and Instagram.

Digital content marketing (DCM) in recent years has also grown to be one of the most popular social media marketing approaches (Carlson et al., 2018). The DCM approach involves enhancing customer experience on social media by creating content that resonates with customers. The effectiveness of content can be determined through the willingness of consumers to engage, subsequently measured with tools provided (i.e., metrics, insights) by these social media platforms (Dwivedi, et al., 2020). The metrics and insights that are indications of the effectiveness of marketing content can be used to measure total customer engagement but may however have no relation to sales contribution. Therefore, it is vital to investigate the impact of these marketing metrics on sales to justify the use of these metrics and insights into social media content as the KPI for sales.

B. Consumer Behavior

The Consumer Decision-Making Process (CDMP) is a consumer behaviour model that describes the consumers' journey when making purchase decisions (Kotler, Armstrong & Opresnik, 2021). This model has been widely used in many consumer-behaviour studies as a framework to identify the impacts of social media marketing at each stage on the CDMP (Nash, 2018) (Voramontri & Klieb, 2019) (Oumayma, 2019). The CDMP model suggests 5 stages in the process such as (1) need recognition, (2) information search (3), evaluation of alternatives, (4) purchase decision, and (5) post-purchase behaviour. Oumayma (2019) yielded some insightful results useful for marketers to better understand how consumers behave at each stage of the CDMP when using social media. These results showed that 57.2% of respondents expressed that they were convinced to make a purchase of products via a photo or video reviewed on social media. Another finding of the study showed that 76% of the respondents were not motivated to provide feedback or reviews upon making a purchase. Findings help marketers better understand and prioritize efforts leading to sales. Based on the results, promoting customers to post pictures, and providing feedback and reviews may be useful to convince future purchases since respondents have expressed feedback as a motivator in purchasing a product.

The purchase decision stage of CDMP would be the most relevant to the context of this study. A proliferation of research has been done to identify social media marketing factors that are influential towards customers' purchase decisions. These findings showed that factors such as consumer cognitive appraisal (Sari, 2020) (Teo, et al., 2019), image quality (Sulthana & Vasantha, 2019), positive word-of-mouth (WOM) (Oumayma, 2019) (Park, et al., 2021), interactivity and content (Dabbous, & Barakat, 2020) are the influential factors to purchase intentions. These studies have however utilized consumer purchase intention as the substitute for purchase decision which is a problem.

Purchase intention is merely the likelihood or tendency of purchasing a product (Peña-García, 2020). A study on purchase intentions for online purchases showed that purchase intentions may subside when factors such as a *desire to shop online*, *frequency of online shopping* and *transaction convenience are*, hence, unable to convert into a sale (Indiani & Fahik, 2020). Another study on purchase intention listed instances where purchase intentions are correlated to an actual purchase decision, hence indicating that purchase intentions can only represent purchase decisions under certain circumstances or influence (Bianchi, et al., 2019). The true representation of purchase intention for purchase decisions becomes even more uncertain considering instances where the purchase intention of a consumer had faded due to certain influences (i.e., bad word of mouth, group influence). In this case, the use of purchase intention as a proxy for actual purchase decisions is then inaccurate. Therefore, the approach of this study in understanding the impact of Instagram marketing on sales will be evaluating real sales data that can more accurately measure the factors that influence the purchase decision.

C. Consumer Behaviour on Instagram

Consumer behaviour differs from one social media platform to another. Much research has been done on Instagram consumers' behaviours to understand the intentions behind Instagram usage as well as users' perspective on Instagram marketing. Each study on this matter has shown different findings which can be used to explain the data mining outcomes of this study.

Past research has shown that Instagram users are more committed and engaged as compared to Facebook, Twitter, and Snapchat (Phua et al., 2017). Furthermore, Instagram has also been shown to be a more ideal platform for developing a brand relationship with customers as compared to Facebook (Poulis et al., 2018). Huang & Su (2018) found that young university consumers expressed using Instagram as an application to look at posts for social interaction or to eliminate boredom. These findings suggested that Instagram may be more effective in building brand engagement as compared to influencing sales. Similarly, another study suggested that Instagram acts as an informational source for fashion enthusiasts to gain outfit inspiration and stay up to date with the latest fashion trends (Cooley & Parks-Yancy, 2019). This study also suggested that Instagram is effective in affecting the purchase intentions of millennials by information search through the consumption of Instagram and subsequently purchasing on the brand's website if interested. Therefore, the usage of Instagram by fashion enthusiasts showed the possibility of Instagram influencing sales.

The study by Djafarova & Bowes (2020) found different effects between Brand-Generated Content (BGC) and User-Generated Content (UGC) on Instagram. BGC was found to be more efficient in promoting consumer engagement whereas UGC has more influence on impulse purchases in the context of Generation Z of the United Kingdom on fashion products (Djafarova & Bowes, 2020). A study done by Astuti & Pramesthi (2018) showed the limited ability of Instagram in building consumer trust which suggests that Instagram marketing may not contribute to increasing sales in the context of Indonesia. These findings suggested the importance of context in determining the effectiveness of Instagram marketing in increasing brand engagement and sales which further encourages the purpose of this study.

D. Consumer Behaviour in Fashion on social media

Consumer behaviour varies between different demographic and industries. It is vital to investigate consumer behaviour, specifically the fashion industry, to understand how consumers of this market behave on social media. The fashion industry can be categorized into different categories where consumers of different fashion categories were found to show different purchasing behaviours (Nash, 2018) (Liu et al., 2019) (Bonilla et al., 2019). Every content published on Instagram has the potential to take users on a cognitive heuristic process in which a consumer may choose to react or not to react towards the content. In instances where a consumer is determined to respond, these responses may be in the form of likes, comments (Ric & Benazić, 2022) or purchase intentions (Sari, 2021). These responses are insightful that can be used to determine the effectiveness of a current Instagram marketing strategy or content.

The study by Samala & Katkam (2019) showed that the willingness of consumers to engage with a fashion brand is a response that signifies consumers' brand loyalty. As brand loyalty is a mediating factor of purchase intention (Ceyhan, 2019), it suggests a positive relationship between consumer engagement on Instagram and purchase intentions. Additionally, engagement in brand content such as shares and WOM are also indicators of brand equity where the brand is given a value premium by consumers which increases the chances of future purchases (Chu, 2019). Instagram metrics and insights represent the measurement of total customer engagement that can be used to study its impact on sales.

E. Machine Learning in Sales Prediction

Machine Learning (ML) has been recognized as an important and effective technique to predict future events using statistical and computational methods. As such, a proliferation of ML research has been conducted for different domains, including the development of sales prediction models. Sales prediction, otherwise known as sales forecasting, has been a practice of many businesses as a guide to making business decisions. The outcome of sales forecasting can often help enhance different business aspects such as marketing campaigns and strategies, planning future inventory, improving marketplace knowledge as well as making business expansion decisions (Behera & Nain, 2020). The businesses can lower the risks of failed investments and making a loss from ineffective funds allocation for marketing or overbudgeting for inventories (Knieriem, 2019) (Karaman, 2019). Table I compares the performance of multiple machine learning algorithms from past research in the sales forecast. Where the Root Mean Square Error (RMSE) or Mean Absolute Error (MAE) was used as the performance evaluation measures.

The selection of the machine learning algorithm depends on the problem at hand. Table I shows the frequent use of ensemble machine learning techniques such as Random Forrest, Gradient Boosted Tree and XGBoost in predicting the sales. Artificial Neural Network (ANN) on the other hand performed well in sales prediction too. The study by Dong, Li, and Zhao showed that Long-Short Term Memory (LSTM) outperformed ANN (Dong, et al., 2019).

TABLE I. A REVIEW OF DIFFERENT MACHINE LEARNING ALGORITHMS USED TO PREDICT SALES

Reference	Machine Learning Models	Data Description	Feature Selection	Parameter Tuning/ Optimization	Evaluation
Behera & Nain, 2020	• Linear Regression • Decision Tree • Ridge Regression • XGBoost	Big Mart Sales prediction	• Products • Outlet characteristics	• Cross Validation	**XGBoost** with the lowest RMSE and MAE
Bajaj et al., 2020	• Linear Regression • K-Neighbors Regressor • XGBoost • Random Forrest	Big Mart Sales prediction using products and outlets related variables	• Products • Outlet characteristics		**Random Forrest** with the highest accuracy
Cantón Croda et al., 2018	• Artificial Neural Network (ANN) • ARIMA	Chemical Company Sales prediction (2015-2016)	• Time Series (2015-2016)	• Multi-layer Perceptron • Feedforward Architecture • Sigmoid Function	**ANN** with a prediction error of less than 5%
Wu et al., 2018	• Linear Regression • Neural Network • Deep Learning • Decision Tree • Decision Tree with Bagging • XGBoost	Black Friday Sales prediction using past customers' information (i.e., demographic, past purchase)	• Customer demographic • Customer past spending	• Linear Regression: Data Transformation • XGBoost: Stepwise Ridge Regression	**XGBoost** with the lowest RMSE
Rincon-Patino et al., 2018	• Linear Regression • Multilayer Perceptron • Support Vector Machine • Regression Tree	Avocado Sales Prediction	• Weather • Time Series	• Data Normalization	**Regression Tree** with the lowest MAE
Cheriyan et al., 2018	• Generalized Linear Regression • Decision Tree • Gradient Boosted Tree	e-Fashion Store Sales	• Sales Volume • Price • Number of visits • Number of searches • Number of collectors • Time series		**Gradient Boosted Tree** with the lowest prediction error
Lin et al., 2019	• Linear Regression • Support Vector Machine • SARIMAX • ANN	Smartphone brand sales prediction	• Sentiment data • Time series		**ANN** with the lowest RMSE
Dong et al., 2019	• I-ARIMA • LSTM • ANN	E-commerce sales forecasting	• Time-series		**LSTM** with the lowest MAPE

III. METHODOLOGY

Fig. 1. CRSIP-DM Flowchart

As shown in Fig. 1, the CRISP-DM (Cross Industry Standard Process for Data Mining) model was adopted as the methodology framework for this study (Ayele, 2020).

F. Business Understanding

A real business case, Business A, has been used as the subject matter for this study. Business A is a fashion retailer that stocks multi-trendy international fashion labels falling under the Small Medium Enterprise (SME) business category in Klang Valley (Malaysia). Business A operates physically and online for its customers. Business A uses Instagram Posts and Stories as its main marketing feature where it posts 2-3 Instagram Posts each day and in a spontaneous manner.

G. Data Understanding

Data on Instagram features, Instagram metrics, and sales data were used to understand the relationship between Instagram Marketing and sales. A total of 4 different datasets were collected comprising 2 Instagram datasets and 2 sales datasets dated between July 2019 to May 2020. The 4 datasets were integrated as one used for analysis and modelling.

An Exploratory Data Analysis (EDA) was carried out using Tableau and Python for statistical EDA. Through data visualisation, the process of EDA was used to determine the (1) distribution of each variable, (2) the relationship between Instagram metrics, insights, features, and sales in a time series manner, and (3) the correlation or multicollinearity between

variables. Findings from statistical analysis were subsequently used to determine the tasks needed for data pre-processing.

H. Data Preparation

Data pre-processing tasks such as outlier treatment, feature selection, data scaling and feature engineering were performed to prepare the dataset for modelling.

I. Modeling

The XGBoost and LSTM algorithms were selected based on the literature review. Univariate and multivariate time series prediction models were developed using the proposed machine learning algorithms. The results of these models were used to determine the significance of Instagram marketing in impacting sales.

As stated in the literature, the XGBoost model can predict sales without using time-series data. A similar attempt was done in this study to compare the outcomes between time-series and non-time-series sales prediction models to determine the significance of time-series in sales prediction using Instagram data.

Here, the inputs of these models were labelled as X and the output as y. i represents the number of inputs, hence, i = 1, 2, 3… (i.e., 1 = likes, 2 = comments…). For time-series models, t represents the current time step and t-j is expressed as the number of previous time steps from the current time step, otherwise known as, the number of lags where j = 1, 2, 3… where t-1 = 1-day lag, t-2 = 2 days lag and so forth. A total of 5 models were developed and compared to determine the significance of time series and Instagram features and metrics in influencing sales.

Univariate Time-series Model

A univariate time-series model predicts sales (y) of the current timestep (y_t) using sales data of the previous timestep (y_{t-1}). Hence, the only input used for this model is the sales of the previous timestep (y_{t-1}) and the output would be the sales of the current timestep (y_t).

Multivariate Time-series Model

A multivariate time-series model predicts sales of current timestep (y_{t-1}) with Instagram data variables of current time-step (X_i), Instagram data variables of previous time-steps ($X_{i(t-j)}$; i = 1, 2, 3…; j = 1, 2, 3…), and sales data of previous time-step(s) (y_{t-j}) as inputs.

TABLE II. PREDICTIVE MODELS DEVELOPED

Algorithm	Model Type
XGBoost	Univariate Time-Series Model
	Multivariate Time-Series Model
	Multivariate Model Non-Time-Series Model
LSTM	Univariate Time-Series Model
	Multivariate Time-Series Model

XGBoost Algorithm

XGBoost is an ensemble machine-learning technique that involves building sequential decision trees to lower the loss function. The XGBoost algorithm uses the gradient descent method to sequentially build new tree models that can minimize the loss function. The individual model is ensembled to create a precise model used for prediction (Brownlee, 2016). The hyperparameters such as Max Depth, Lambda, Gamma, and Learning Rate can be tuned for model optimization.

Long-Short Term Memory (LSTM) Algorithm

Long-Short Term Memory (LSTM) is a variation of the Recurrent Neural Network (RNN) which can learn long-term dependencies. The LSTM of RNN is an algorithm designed for sequential data. Sequential data can be defined as using historical data to predict future data which indicates its suitability to be used on time-series data. According to IBM Cloud Education (2020) RNN as a Neural Network learns the patterns of training data to make future predictions (IBM Cloud Education, 2020). The "memory" component of RNN allows the model to use the output from previous steps to be used as input together with current inputs to produce current output. Each time a current input is passed into an RNN, the network then applies weights for current and previous inputs as opposed to applying weights for only the current inputs. The RNN then learns through backpropagation gradient descent where weights of each input are constantly being tweaked while training the model to obtain weights with the lowest loss function/error. In some cases, certain inputs that are given a weight too high or too small can result in exploding or vanishing gradient. This may cause an inability of RNNs to retrieve long-term past data. LSTM is designed to help eliminate the issue of vanishing gradient or exploding gradient as the algorithm is created to remember long-term memory by the gated cell of LSTM (Mittal, 2019). The hyperparameters such as hidden nodes, dropout, optimizer, batch size and epoch can be tuned when training the model to optimize the learning method of the model.

J. Model Evaluation

The selected performance metrics for this study were Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE). The MAE performance metrics is a measurement of forecast error in which it takes the summation of all forecast errors subsequently divided by the total sample used (Vandeput, 2021). As such, the MAE metric was used to compare the average prediction error and the average demand (sales) to determine how far off the average prediction values are.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\mathcal{X}_{predict} - \mathcal{X}_{actual}| \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(|\mathcal{X}_{predict} - \mathcal{X}_{actual}|)^2} \quad (2)$$

RMSE on the other hand is a performance metric that first sums up the squared residuals divided by the total sample used and subsequently applied a square root on the summarized value. This metric thus emphasized a larger prediction error which was used to determine if the model is biased. The equations to calculate the MAE and RMSE are shown in equations (1) and (2). Additionally, based on related works on machine learning for sales prediction, the model performances were evaluated based on RMSE or MAE. Therefore, the models of this study were evaluated using the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) values.

IV. Data Understanding & Preparation

A. Data Understanding by Visualization

Tableau platform was used for data visualization to achieve the research objectives of determining the relationship between Instagram metrics, features, and sales and to identify Instagram metrics and features that have a significant impact on sales. The independent variables were categorized into subcategories to better identify the impact of different Instagram features and their metrics on sales. The feature category consists of features that are present in this dataset whereas the metric category consists of metrics of each feature that were available on Instagram as shown in Table III.

TABLE III. INSTAGRAM FEATURE AND METRICS CATEGORIES

CATEGORY		
Feature	**Engagement**	**Variables**
Instagram Post	Active Engagement	*busi_add_taps*
		comment
		follows
		like
		product_button_click
		product_page_views
		profile_visits
		save
		share
		web_clicks
	Passive Engagement	*imp_explore*
		imp_hashtag
		imp_home
		imp_other
		imp_profile
		reach
		total_imp
Instagram Story	Active Engagement	*is_business_address_taps*
		is_follows
		is_get_direction_button_tap
		is_profile_visits
		is_replies
		is_shares
		is_sticker_taps
		is_websites_taps
	Passive Engagement	*is_impression*
		is_reach

The metrics were plotted as individual lines and then plotted against sales. The sales variable was plotted as an area plot so that independent variables (metrics) can be easily differentiated from the dependent variable (sales). Finally, 4 different variations of the proposed graphs were produced based on the 4 categories listed in Table III.

Log transformation was applied to the x-axis for categories that have metrics with significantly smaller distributions than other metrics of the same category.

B. Data Understanding by Statistical Method

Based on the statistical description of the dataset, *promotional_post* was removed as it does not provide information on the promotional Instagram feature. Similarly, the Quantity variable is merely sales recorded in another format, in other words, another target variable, which was removed as the research objective focused on sales revenue rather than quantities.

In terms of standard deviation, variables such as *like, comment, share, save, product_page_views, product_button_click, web_clicks, follows, reach, total_imp, imp_reach, imp_profile, imp_hashtag, imp_other, imp_explore, is_profile_visits, is_websites_taps, is_sticker_taps, is_reach, is_impression, Quantity, and sales* were found with a standard deviation of more than 3 which indicated outliers. The outlier treatment was discussed under data pre-processing.

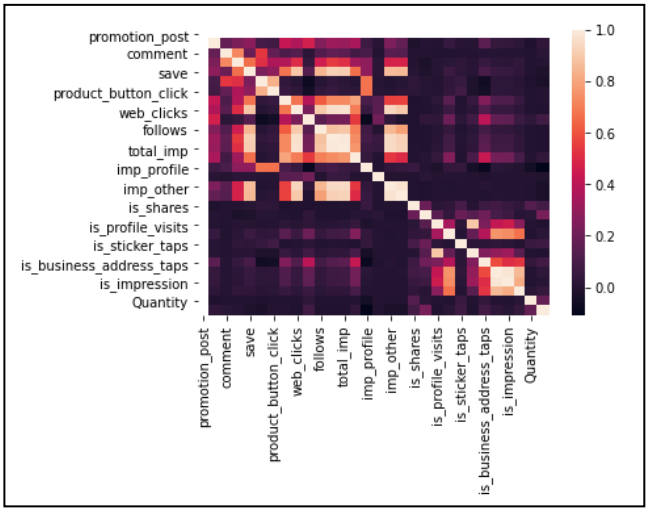


Fig 2. Heatmap of correlation between variables.

Additionally, variables such as comment, share, save, product_page_views, product_button_click, web_clicks, busi_add_taps, follows, imp_explore, is_shares, is_replies, is_profile_visits, is_website_taps, is_stickers_taps, is_get_direction_taps, and is_follows got a distribution of less than 10 on the 75th percentile. This might describe a consumer behaviour that is reluctant in engaging through these features. Additionally, the product_page_views and product_button_click were only made available in October 2020 by Instagram, hence might have led to a lack of distribution for such features.

Fig. 2 shows the correlation heatmap between different pairs of variables. Based on the heatmap, there are pairs of variables with a correlation coefficient of more than 0.5, indicating a strong association (Laerd Statistic, 2018). Based on the statistical descriptions of each variable, all the variables in the dataset were found with outliers. The outliers were visualized in boxplots as shown in Fig. 3. Additionally, the outliers showed a right skewness, indicating that these outliers were of higher values than the inliers. This could be due to external non-Instagram factors such as sales promotion (i.e., flash sales, black Friday). Another possibility could be the Instagram posts or stories that got boosted, promoted, or sponsored resulting in an outlier.

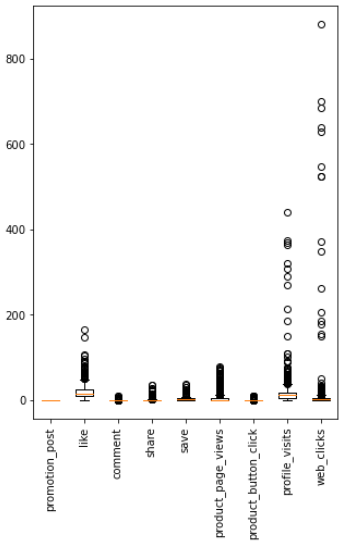


Fig. 3. Boxplots of selected variables.

The outlier treatment was done by setting the upper boundary (3rd quartile) as a cap limit to group all outliers as an inlier. The outlier of each variable was replaced with the interquartile upper boundary of their respective variables. The data point that fell outside the interquartile range was identified as an outlier and got replaced. Fig. 4 shows the boxplots of selected variables after outlier treatment.

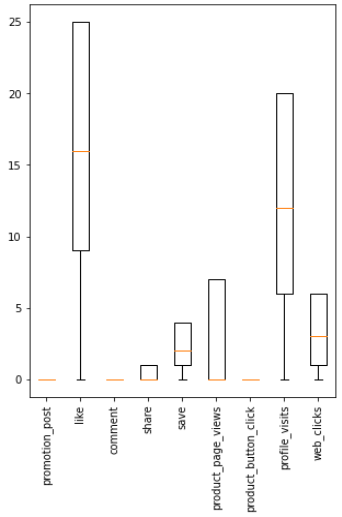


Fig. 4. Boxplots of selected variables after outlier treatment

C. Feature Selection

The variables selected for modelling were based on the significance of the Granger Causality test. Granger causality is a statistical hypothesis test used to determine the significance of one time series in forecasting another (Wei, 2019). The hypothesis is described as:

H0: The time series do not influence one another.
H1: The time series influence each other.

Additionally, the test here was applied up to a look back of 7 days lag of each variable in influencing the time-series of sales variable. A total of 15 out of 30 variables were identified as significant in influencing the time series of sales with a p-value less than 0.05 (Table IV). The variables namely share, save, product_page_views, profile_visits, reach, total_imp, imp_home, imp_profile, imp_hashtag, imp_other, imp_explore, is_profile_visit, is_websites_tap, is_reach, and is_impression was selected for modelling.

TABLE IV. GRANGER CASUALITY TEST P-VALUES ON SIGNIFICANT VARIABLES

Dependent Variable	Independent Variables	p-value
sales	*share*	0.0053
	save	0.0051
	product_page_views	0.0074
	profile_visits	0.0164
	reach	0.0057
	total_imp	0.0089
	imp_home	0.0400
	imp_profile	0.0000
	imp_hashtag	0.0025
	imp_other	0.0160
	imp_explore	0.0008
	is_profile_visit	0.0003
	is_websites_taps	0.0000
	is_reach	0.0010
	is_impression	0.0010

D. Data Normalisation

The normalisation method was used to undergo data scaling before modelling. Based on the statistical description of variables, variables such as like (mean = 22.95) and reach (mean = 2963.51) got significantly different ranges of values. This may result in an unequal measurement of variables when fitting a model as higher weight may be given to variables with larger values (Loukas, 2021). Therefore, the dataset was transformed into a scale of zero (0) to one (1) as shown in Fig. 5.

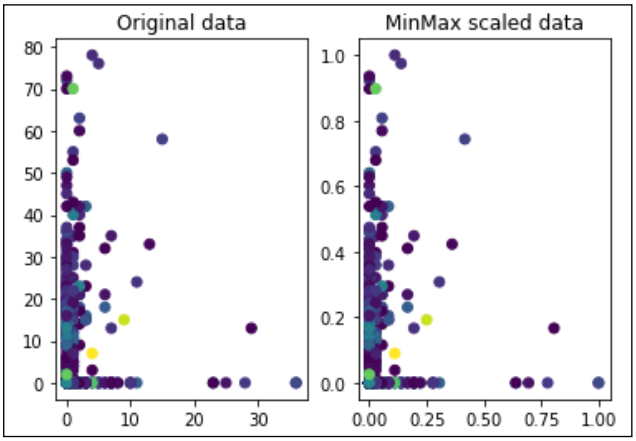


Fig. 5. Comparison of data before and after data normalization.

E. Feature Engineering

The dataset was transformed from a time series into supervised series. Each variable was shifted to a format where previous time-step(s) were used as inputs to predict the current time-step. Based on Fig. 6, with *var1(t)* being the output variable, *var1(t-1)* was the 1-day lag of *var1(t)* and *var(t-2)* was 2 days lag of *var1(t)* used to predict *var1(t)* (target variable). The shift was applied to all fifteen (15) independent variables as well as the target variable itself.

	var1(t-1)	var2(t-1)	var3(t-1)	var4(t-1)	var5(t-1)	var6(t-1)	\
1	0.706049	0.0	0.0	0.0	0.0	0.001232	
2	1.000000	0.0	0.0	0.0	0.0	0.000000	
3	1.000000	0.0	0.0	0.0	0.0	0.000000	
4	0.318228	0.0	0.0	0.0	0.0	0.000000	
5	1.000000	0.0	0.0	0.0	0.0	0.000000	

	var7(t-1)	var8(t-1)	var9(t-1)	var10(t-1)	var11(t-1)	var12(t-1)	\
1	0.001125	0.0	0.004926	0.0	0.0	0.0	
2	0.000000	0.0	0.000000	0.0	0.0	0.0	
3	0.000000	0.0	0.000000	0.0	0.0	0.0	
4	0.000000	0.0	0.000000	0.0	0.0	0.0	
5	0.000000	0.0	0.000000	0.0	0.0	0.0	

	var13(t-1)	var14(t-1)	var15(t-1)	var16(t-1)	var1(t)
1	0.0	0.0	0.0	0.0	1.000000
2	0.0	0.0	0.0	0.0	1.000000
3	0.0	0.0	0.0	0.0	0.318228
4	0.0	0.0	0.0	0.0	1.000000
5	0.0	0.0	0.0	0.0	1.000000

Fig. 6. Example of additional features after feature engineering.

The backshift of variables or the creation of time-lagged variables was a feature engineering effort in creating new variables derived from shifting existing variables based on the desired time-lag as inputs for modelling. Modelling was first trained by creating a 1-day lag variable for each variable. More models were trained with increased day-lags for each variable as optimization. For example, the input for the first model would be *var1(t-1)*, …, *var16(t-1)*; the input for the second model would be *var1(t-1), var2(t-1), …, var15(t-2), var16(t-2)* and so on. The purpose was to determine the optimal number of lags to be used as inputs based on model performance evaluation.

F. Data Partitioning

The dataset was partitioned into 70% training and 30% testing in which the training dataset was used for modelling and testing for validation. Due to the shifting of data to be used as input variables, the first few rows of the dataset were consisting of null values based on the number of time lags used as input. The number of observations for both training and testing datasets varied along with the number of time lags used to develop a model.

G. Model Development

The **LSTM** and **XGBoost** machine learning models were built using the datapoints of 1 day-lag and 2 days-lag of all variables. The decision to stop adding more models was based on whether additional time lags were able to optimize its previous model. The variations of the models developed are shown in Table IV.

TABLE V. FEATURES USED FOR DIFFERENT MODELS

Model	Time-lags	Features
XGBoost – Univariate	1	*sales(t-1)*
	2	*sales(t-1), sales(t-2)*
	3	*sales(t-1), sales(t-2), sales(t-3)*
XGBoost – Multivariate	1	*share(t-1), product_page_views(t-1), product_button_click(t-1), profile_visits(t-1), web_clicks(t-1), follows, reach(t-1), total_imp(t-1), imp_home(t-1), imp_profile(t-1), imp_hashtag(t-1), imp_other(t-1), imp_explore(t-1), is_shares(t-1), is_sticker_taps(t-1), is_business_address_taps(t-1), is_impression(t-1), sales(t-1)*
	2	All of the above + *share(t-2), product_page_views(t-2), … sales(t-2)*
	3	All of the above + *share(t-3), product_page_views(t-3), … sales(t-3)*
	4	All of the above + *share(t-4), product_page_views(t-4), … sales(t+4)*
LSTM - Univariate	1	*sales(t-1)*
	2	*sales(t-1), sales(t-2)*
LSTM - Multivariate	1	*share(t-1), product_page_views(t-1), product_button_click(t-1), profile_visits(t-1), web_clicks(t-1), follows, reach(t-1), total_imp(t-1), imp_home(t-1), imp_profile(t-1), imp_hashtag(t-1), imp_other(t-1), imp_explore(t-1), is_shares(t-1), is_sticker_taps(t-1), is_business_address_tap(t-1)s, is_impression(t-1), sales(t-1)*
	2	All of the above + *share(t-2), product_page_views(t-2), … sales(t-2)*
	3	All of the above + *share(t-3), product_page_views(t-3), … sales(t-3)*

V. RESULTS AND ANALYSIS

This section consists of the results and findings of data visualization and modelling. The findings were critically analyzed and discussed in detail to give meaning and context.

A. Data Visualization

This section aimed to identify variables with similar behaviours or distinctly opposite behaviours to determine the relationships between variables of the dataset. Findings from visualization were used to understand the correlation between different features and metrics. Additionally, the distribution of each variable was also used to describe consumer behaviour on Instagram in the context of fashion content which was discussed in the following section.

B. The relationship between different "Active Engagement" Metrics of Instagram Post

The metrics such as **"Like", "Web Clicks",** and **"Profile Visits"** showed larger values as compared to other metrics of this category. Additionally, **"Like"** and **"Web Clicks"** showed relatively weak positive relationships. The findings here could indicate that Instagram users were more succumbed to tapping the **"Like", "Web Clicks",** and **"Profile Visits"** buttons when being shown fashion content in the form of an Instagram Post. Additionally, a **"Like"** on an Instagram Post may motivate consumers to click on the website button depending on the interest of the consumer of the product or brand. The variables such as **"Profile Visit"**, **"Save"** and **"Follows"** showed relatively a weak positive relationship in which only the peaks of these variables follow a similar behaviour. Additionally, **"Busi Add Taps"** showed to diminish completely as the **"Product Button Click"** increases in distribution. **"Product Page Views"** and **"Profile Visit"** showed a negative relationship. The **"Product Page Views"** was the only available from January 2020 onwards, with more significant distributions from October 2020. The weak relationship between **"Profile**

Visit”, **“Save”** and **“Follows”** indicated that action on any of these buttons may lead to another. The diminishing distribution of **“Busi Add Taps”** can be explained by the popularity of **“Product Button”** in replacing the **“Busi Add Taps”** after being introduced as a new feature. This indicated that consumers prefer obtaining information about a product rather than the business. Finally, the negative relationship between **“Product Page Views”** and **“Profile Visit”** indicated a certain consumer behaviour when interested in a product or brand.

C. The relationship between different “Passive Engagement” Metrics of Instagram Posts

The metrics **“Total Imp”**, **“Imp Home”**, and **“Reach”** got significantly higher distributions compared to the distributions of other variables. Additionally, three variables with higher distributions showed a positive relationship altogether. This indicates that Instagram Posts of fashion content are more likely to appear on the home page of Instagram users. This indicates that contents produced by the business are more likely to be shown on a screen of a follower than a non-follower. The positive relationship between “**Reach**” and “**Imp Home**” indicated that the contents did not merely show up on the screen of an Instagram user but have been seen as “**Reach**” represents the number of unique users that have seen the Instagram Post.

The magnified variables from log transformation have shown a positive relationship between **“Imp Explore”** and the previously identified group of variables with positive relationships (i.e., **“Total Imp”**, **“Imp Home”**, and **“Reach”**). Although “**Imp Explore**” have shown a positive relationship with the group as mentioned, the distribution is significantly smaller. This indicates that the possibility of Instagram Posts showing on the explore page, in other words, to non-followers is very small. The positive relationship between “**Imp Explore**” and “**Reach**” however may suggest that Instagram Posts shown on explore page were seen rather than just being shown on screen.

D. The relationship between different “Active Engagement” Metrics of Instagram Story

The variables “**Is Profile Visits**”, “**Is Websites Taps**” and “**Is Sticker Taps**” have shown a significantly higher distribution compared to other active engagement metrics of Instagram Stories. Additionally, the variables “**Is Profile Visits**” and “**Is Websites Taps**” have shown a positive relationship. This could indicate that “**Is Profile Visits**”, “**Is Websites Taps**” and “**Is Sticker Taps**” buttons were the preferred method of engagement of consumers when consuming fashion content on Instagram. This could suggest that tapping on the website link of an Instagram Story may also lead to a profile visit.

After applying the logarithmic transformation, variables **“Is Replies”** and **“Is Shares”** showed a weak positive relationship. This could indicate the ability of a certain type of Instagram Story content in probing consumers to reply and share. However, the low distribution of the forms of active engagements indicates that consumers are usually reluctant to engage in such forms.

E. The relationship between different “Passive Engagement” Metrics of Instagram Story

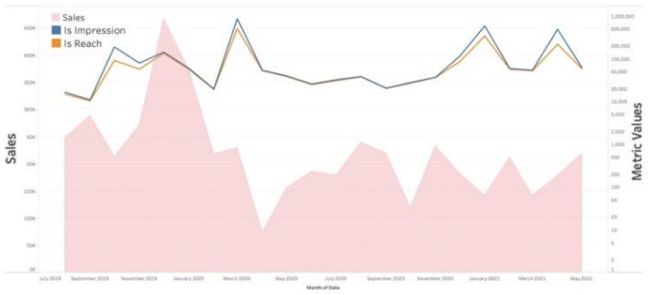


Fig.7. Passive engagements of Instagram stories against sales

Both the passive engagement variables of Instagram Story have shown a positive relationship with each other as shown in Fig. 7. Additionally, the two (2) variables have also shown a high distribution when compared to the passive engagement variables of Instagram. In terms of sales, no obvious relationships were shown between the passive engagements of Instagram stories and sales. The positive relationship between **“Is Impression”** and **“Is Reach”** here may indicate that an Instagram Story will be able to achieve similar measures of **Impressions** and **Reach**, in other words, an Instagram Story is likely to be seen when appeared on the screen.

F. To identify Instagram metrics and features that have significance in impacting the sale

An increase in the values of Instagram variables should lead to a significant increase in sales for instances where Instagram metrics and features are impactful towards sales. Therefore, this section aimed to identify such a relationship in which these relationships could either be positive or negative.

Based on Fig. 8, only a total of three (3) variables were shown identifiable relationships with **“Sales”**. The four variables were namely **“Profile Visits”**, **“Like”** and **“Share”**. These variables however have not shown a strong relationship with **“sales”**. The relationship here is determined by comparing the moving trend of each variable against the moving trend of “**sales”** as highlighted in Fig. 8.

Fig. 8. Line graphs between sales and Instagram variables that are determined to have a relationship with sales

G. Summary of the findings from Data Visualization

Although variables of different categories have been plotted against each other as well as on the target variable (sales), no obvious positive/negative relationships were identifiable between different Instagram features, metrics, and sales through data visualization. However, findings such as metrics with the highest distribution as well as metrics that share a similar behaviour were able to be identified through data visualization.

H. LSTM Models Results and Analysis

The LSTM architecture obtained the best results with 50 hidden layer neurons, 1 dense layer, and **tanh** activation function and was fitted using **20 epochs** and **batch size 72**. A total of 5 models were developed using LSTM architecture (Table VI). A learning curve model is a plot used to diagnose issues encountered by a machine learning model as well as the performance of the model (Brownlee, 2019). Based on table VI, both univariate and multivariate models have shown the tendency to overfit as more time-lags were used as inputs for training. This indicates that these models (LSTM-U1, LSTM-M2, LSTM-M3) were taking in more information than needed in predicting sales. In other words, these models were not generalizable as loss functions showed an increase in testing data.

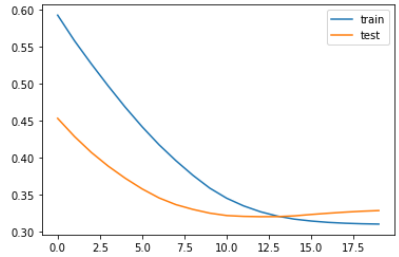
LSTM-U1 and **LSTM-M1** were the best-performing models of the univariate and multivariate variation respectively. In this case, both **LSTM-U1** and **LSTM-M1** have shown to underfit in the beginning, however, as the number of training cycles increased, the **LSTM-U1** model seemed overfitting. The **LSTM-M1** model showed to predict well for the testing dataset in which the loss function remained low proving that the testing dataset is easy for the developed model to predict than the training data (Brownlee, 2019).

The **LSTM-M1**, **LSTM-M2**, and **LSTM-M3** models showed the same MAE and RMSE values for the testing dataset. **LSTM-M2** and **LSTM- M3** however obtained lower MAE and RMSE for the training dataset. This further supported **LSTM-M2** and **LSMT-M3** in overfitting as testing datasets were not able to perform well. As such, an LSTM model trained with Instagram data of 1-day lag was able to produce the best results for this study.

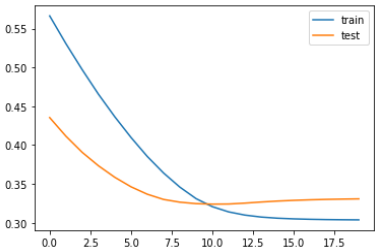
I. XGBoost Models Results and Analysis

Based on Table VII, all multivariate time-series **XGBoost** models performed better than univariate models. This indicates the ability of Instagram variables in providing the lowest loss function when predicting the future sales of 1 day. Additionally, the non-time-series **XGB-N-TS** model outperformed all univariate models. However, the **XGB-N-TS** failed to outperform any of the multivariate time-series models (XGB-M1, XGB-M2, XGB-M3, XGB-M4). This suggests the capability of Instagram variables in influencing sales, and hence, the ability to predict sales. Amongst the four (4) multivariate **XGBoost** and the **XGB-M3** models performed the best with the lowest loss function. This suggests that Instagram data with up to 3-day lags can be used to predict current sales. Based on **XGB-M3** model as shown in Fig. 9, variables such as *sales(t-3), reach(t-2), imp_home(t-2), imp_profile(t-2), imp_hashtag(t-2), sales(t-1), profile_visits(t-1), imp_profile(t-1), imp_other(t-1), and is_impressions(t-1)* were found as important features. Additionally, these variables include 5 passive engagement variables and 1 active engagement variable from Instagram Post whereas only 1 passive engagement variable was selected from Instagram Story. Other than that, only Instagram variables of 2-days were used to predict sales.

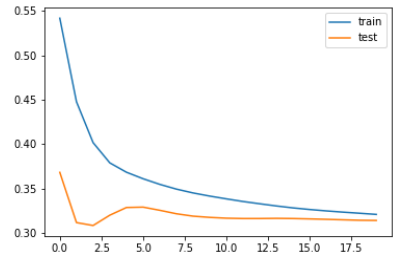
TABLE VI. LSTM MODELS RESULTS

<table>
<tr><th colspan="3">UNIVARIATE</th></tr>
<tr><td>1 time-step</td><td>2 time-steps</td><td>3 time-steps</td></tr>
<tr><td> 

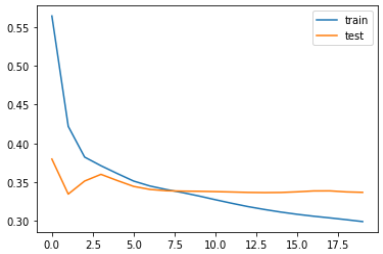
Learning Curve of LSTM-U1

Train – MAE: 1528.90 RMSE: 1747.65
Test – MAE: 1611.94 RMSE: 1843.38</td><td> 

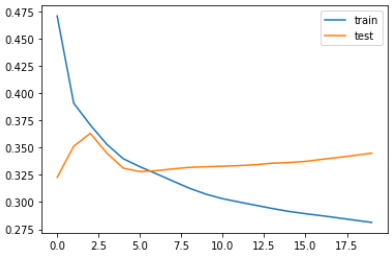
Learning Curve of LSTM-U2

Train – MAE: 1500.12 RMSE: 1711.77
Test – MAE: 1634.97 RMSE: 1861.24</td><td></td></tr>
<tr><th colspan="3">MULTIVARIATE</th></tr>
<tr><td>1 time-step</td><td>2 time-steps</td><td>3 time-steps</td></tr>
<tr><td> 

Learning Curve of LSTM-M1

Train – MAE: 1579.40 RMSE: 1837.11
Test – MAE: 1553.43 RMSE: 1828.59</td><td> 

Learning Curve of LSTM-M2

Train – MAE: 1467.13 RMSE: 1722.42
Test – MAE: 1553.43 RMSE: 1828.59</td><td> 

Learning Curve of LSTM-M3

Train – MAE: 1381.01 RMSE: 1669.15
Test – MAE: 1553.43 RMSE: 1828.59</td></tr>
</table>

TABLE VII. XGBOOST MODELS RESULTS

Model	Model Name	Time-lags	Results			
			MAE		RMSE	
			Train	Test	Train	Test
Univariate Time-series	XGB-U1	1	2819.68	2741.02	2819.68	2741.02
	XGB-U2	2	2805.50	2693.63	2820.99	2711.16
	XGB-U3	3	2799.09	2674.49	2817.70	2694.16
Multivariate Time-series	XGB-M1	1	2809.86	2729.17	2850.24	2303.54
	XGB-M2	2	2805.50	2280.79	2854.50	2291.31
	XGB-M3	3	2799.09	2230.54	2858.01	2244.11
	XGB-M4	4	2805.98	2321.14	2858.03	2332.62
Multivariate Non Time-series	XGB-N-TS	None	2811.24	2376.72	2819.68	2386.72

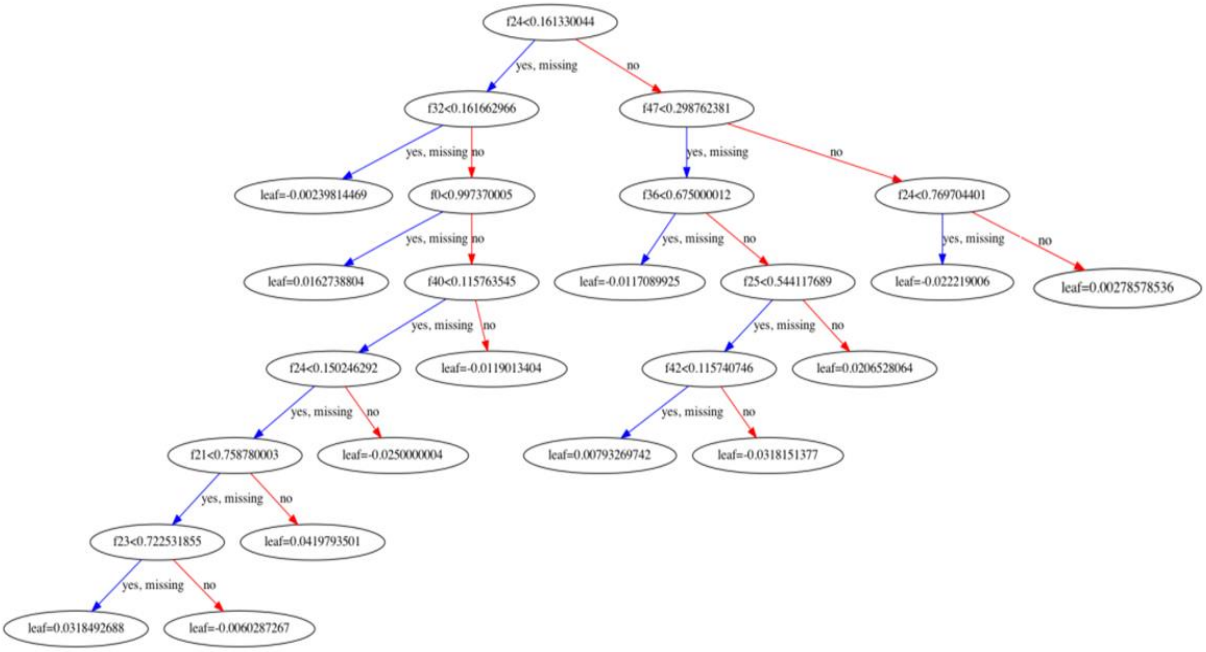


Fig. 9. Decision tree of XGB-M3 model

Performance evaluation of LSTM and XGBoost Models

Based on Table VI, all LSTM multivariate models (**LSTM-M1, LSTM-M2, LSTM-M3**) showed the lowest MAE and RMSE values on the testing dataset. However, in terms of best fit, the **LSTM-M1** model obtained the smallest MAE and RMSE values. Models **LSTM-M2** and **LSTM-M3** on the other hand showed higher values, thus indicating an overfit.

As **LSTM-M1** was identified as the best-performing model, a further analysis was done by plotting the predicted sales values against the actual sale value (Fig. 10). The plotted graph here showed a conservative **LSTM-M1** model in predicting sales as most of the predicted sales value were very small in variation as compared to the actual value. Additionally, predictions of LSTM-M1 showed to revolve around the median value.

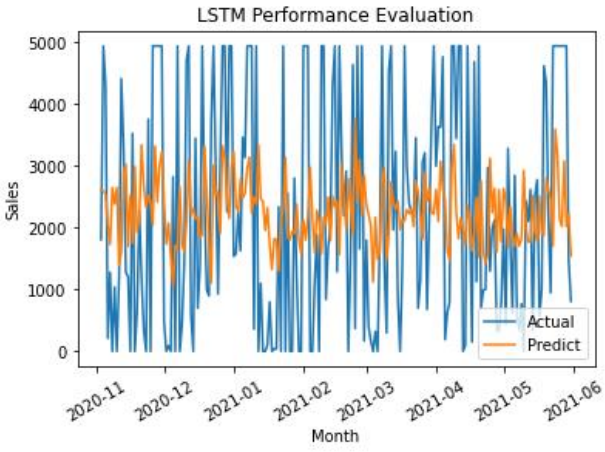


Fig. 10. LSTM-M1 performance evaluation line plot

Additionally, comparing between **MAE** value (1579.395) to the mean of sales (3642.214), the average forecast error is only 43% of the mean value, suggesting a relatively low accuracy of the model. However, when looking at the **RMSE** value of 1828.591, it can then suggest that the model does not tend to overpredict the actual sales value, which supports the statement on the model being one that is conservative when predicting. As such, there is still room for improvement for the current **LSTM-M1** model and hence determined as unsuitable for deployment.

VI. Discussions and Conclusions

Firstly, based on the findings of data visualisation, it can be determined that there is a drastic difference in distribution between different Instagram features (i.e., Instagram Posts, Instagram Stories (IS)), types of engagement (i.e., active engagement, passive engagement), and metrics. In terms of active engagement, metrics that have a higher distribution for Instagram Posts such as "**Like**", "**Web Clicks**" and "**Profile Visits**", whereas metrics such as "**Is Profile Visits**", "**Is Website Taps**", and "**Is Sticker Taps**" were of higher distribution for Instagram Stories. This thus describes the consumer behaviour on Instagram when consuming fashion content. As such, the findings had shown that consumers are more prone to visit Instagram business profiles and business websites when being shown content on Instagram Posts and Instagram Story as both features have the website visit (i.e., "**Web Clicks**", "**Is Website Taps**") and profile visit (i.e., "**Profile Visits**", "**Is Profile Visits**") buttons as the higher distributed metric.

In terms of passive engagements on Instagram, "**Total Imp**", "**Imp Home**" and "**Reach**" have significantly higher distributions for Instagram Posts whereas both metrics of

Instagram stories, "**Is Impression**" and "**Is Reach**", were equally high in distribution. The mean values of passive engagements range between 2000 to 4000, whereas the mean value of active engagements ranges only between 6 to 37. The drastic difference in mean values between active and passive engagements suggests that only approximately 0.15% to 0.93% of the viewed content (i.e., contents shown on screen, contents shown and seen) had led to active engagements. Furthermore, "**Imp Home**" as one of the passive engagement metrics has a higher distribution, indicating that most viewers of Instagram Posts were followers of the business's Instagram account. This is because "**Imp Home**" is a metric that measures the number of times content was shown on the screens of users' Instagram homepages. This can only happen if a user was following the Instagram business account. These findings then show the reluctance of consumers to actively engage (i.e., like a post, or comment on a post) when being shown Instagram content.

Secondly, based on the **XGBoost** and **LSTM** predictive models, both models had shown results in which a model trained with Instagram variables was able to perform better than a predictive model with only sales variable(s). Therefore, both models had the significance of Instagram data in predicting sales, in other words, indicating that Instagram data influence sales. However, not all metrics had shown significance in influencing sales. The Instagram metrics determined to have significance through the granger causality statistical test in influencing sales were *share, save, product_page_views, profile_visits, reach, total_imp, imp_home, imp_profile, imp_hashtag, imp_other, imp_explore, is_profile_visit, is_websites_tap, is_reach,* and *is_impression*. The use of machine learning technique had also yield insights that showed the significance of these metrics in impacting sales that could not be determined through data visualization.

Additionally, the **LSTM-M1** model being the best performing model suggested that the optimal lookback period in predicting sales was with a 1-day lag. In other words, the data from the previous day was the most suited to predict sales for the following day. This then indicates that Instagram users that consume fashion content react in a relatively fast manner when making a purchase decision. The type of engagements that may lead to purchase was identified as 15 variables as mentioned. In this case, businesses should include product buttons, product pages, and websites on Instagram Posts; stickers and business addresses on Instagram Stories as these features have shown significance in influencing sales. In terms of passive engagement, a total of 5 impressions from Instagram Posts and 1 impression metric from Instagram Story had shown significance in influencing sales. This thus suggests the types of strategies (Table VIII) a business can attempt to increase these impression metrics, which may lead to increased sales.

TABLE VIII. EXAMPLES OF MAKING STRATEGIES BASED ON FINDINGS

Metrics	Goal
imp_home	Increase more followers for Instagram content to show on the homepage of Instagram user
imp_profile	Create a call to action or road maps that lead to the business's Instagram account
imp_hashtag	Use relevant hashtags for each post
imp_other	Strategize on being tagged or featured in the Instagram of other accounts
imp_explore	Create more engagement to drive content to the explore tab
is_impression	Create more Instagram stories to increase impressions for Instagram Story

TABLE IX. MAE AND RMSE OF ALL DEVELOPED XGBOOST AND LSTM MODELS

Model	Model	MAE		RMSE	
		Train	Test	Train	Test
LSTM Univariate	LSTM-U1	1528.897	1611.935	1747.651	1843.381
	LSTM-U2	1500.124	1634.974	1711.765	1861.240
LSTM Multivariate	LSTM-M1	1579.395	1553.431	1837.114	1828.591
	LSTM-M2	1467.129	1553.431	1722.422	1828.591
	LSTM-M3	1381.011	1553.431	1669.154	1828.591
XGBoost Univariate	XGB-U1	2819.679	2741.016	2819.679	2741.016
	XGB-U2	2805.501	2693.6252	2820.992	2711.156
	XGB-U3	2799.092	2674.494	2817.696	2694.162
XGBoost Multivariate	XGB-M1	2809.861	2729.167	2850.240	2303.544
	XGB-M2	2805.500	2280.794	2854.502	2291.312
	XGB-M3	2799.092	2230.540	2858.010	2244.110
	XGB-M4	2805.982	2321.139	2858.029	2332.622
XGBoost Non-Time-Series	XGB-N-TS	2811.244	2376.719	2819.68	2386.717

In terms of all machine learning models built in this study, the LSTM-M1 multivariate model with 1-day lag time series data has shown to be the best-performing model based on performance evaluation metrics. As such, the LSTM model can be determined as suitable for sales prediction using Instagram data. However, the performance evaluation graph and performance evaluation metrics have also suggested that the LSTM-M1 model is biased and non-precise. This suggests that the best model developed for this study still has much room for improvement.

Finally, when comparing the results from data visualisation and modelling, 7 variables were identified as having high distribution and were significant based on the granger causality test (i.e., Web Clicks, Profile Visits, Total Impression, Impression from Home, IS Sticker Taps, IS Reach, and IS Impression). This indicates that a higher number of "likes", "IS profile visits", "IS websites taps", and "IS reach" may not lead to increased sales. However, based on Table X, variables "IS website taps" and "IS reach" were correlated to variables that have significance in influencing sales ("is_impression"). As such, these variables may influence the significant variables, which would subsequently lead to influencing sales. Therefore, variables such as "IS website taps" and "IS reach" should not be neglected. Variables "likes" and "IS profile visits" had shown no correlation with any significant variables that influence sales, hence, are suggested to be neglected when measuring the effectiveness of social media marketing.

TABLE X. PEARSON'S CORRELATION COEFFICIENTS BETWEEN NON-SIGNIFICANT AND SIGNIFICANT VARIABLES IN INFLUENCING SALES

Non-significant Variable in Influencing Sales	Significant Variables in Influencing Sales	Pearson's Correlation Coefficients
is_websites_taps	*is_impression*	0.737563
is_reach	*is_impression*	0.978768

When relating the results to materials of the reviewed literature, the results of this study had shown to both comply with and differ from certain studies. Firstly, results had shown that engagement may not lead to sales due to a lack of significance and correlation to sales (i.e., likes, comments). This slightly differs from the study of Coursaris, Van Osch, and Balogh (Chu et al., 2019) stating that shares, likes and comments have an influence on sales. This study also showed the significance of "shares" in influencing sales but not for likes and comments, therefore, concluding that not all forms of engagement can influence sales. The findings of the study from Djafarova & Bowes (2020) on brand-generated content being prompt to consumer engagement rather than a purchase have also differed from the results of this study. Most of the content from the business case used here was brand-generated content, the distribution of active engagements was found to be low but was able to influence sales. Finally, the results of this study agree with the study of Cooley and Parks-Yancy (Cooley & Parks-Yancy, 2019) in suggesting Instagram as an information hub for fashion enthusiasts that are prone to visiting the brand's website to make a purchase if interested. This can be seen with the significance of web clicks and product page views in influencing sales.

The findings identify from modelling that both Instagram Posts and Instagram Stories are significant in impacting sales as shown in Table XI. However, not all features and metrics from Instagram Posts and Instagram Stories were found significant.

TABLE XI. INSTAGRAM POST FEATURES & METRICES THAT HAS INFLUENCE ON SALES

Instagram Post	
Feature	**Metric**
Share Button	Product Page Views
Product Button	Profile Visits
Web Clicks	Reach
Follow	Total Impressions
	Impression from Home
	Impression from Hashtag
	Impression from Other
	Impression from Explore

Based on data visualization, it is determined that certain metrics that were not identified as significant variables in impacting sales were correlated to variables that were significant in influencing sales. Hence, this showed an insight into the relationship between different features and metrics in influencing sales. As pre-mentioned, both the **XGBoost** and **LSTM** models were able to predict sales using Instagram marketing factors. Additionally, the **LSTM** model showed a better performance in predicting sales.

VII. IMPORTANCE AND CONTRIBUTIONS OF THE STUDY

Based on all the discussions and conclusions drawn from the previous section, this study can suggest to the Instagram marketing team the features and metrics to focus on when marketing on Instagram. Furthermore, the relationship between different metrics can also give businesses an idea of the underlying meaning of metrics and features that can indirectly influence sales. Through these relationships, businesses can understand why certain Instagram metrics and features do not directly reflect on sales and consider a different strategy when marketing on Instagram. Besides all the findings on Instagram features, metrics and sales, this study also contributed to the novelty of building machine learning models using Instagram marketing factors. As such, businesses that are interested in using machine learning models for future sales prediction can consider the proposed models.

VIII. FUTURE RECOMMENDATIONS

The findings of this study were only to the business case used that limited the exploration of other Instagram features such as Instagram Live, Promotional Posts, Reels, etc... The business case used in this study got a relatively small number of followers (6000 followers) with only fashion content. The

number of followers may be the cause of lower active engagement. Hence, more studies can be conducted with different business cases to see how the findings differ on different natures of businesses. Further, an LSTM model was difficult to determine and gauge the weights given to each feature as the weights in each gate of the neural network are constantly being tweaked when learning the data for prediction. This created a limitation in identifying the most significant features. Future research can be conducted to propose a framework to decipher a neural network model in modelling and identifying weights given to each Instagram feature.

REFERENCES

Astuti, B., & Pramesthi, A. (2018). Analysis on the effect of Instagram use on consumer purchase intensity. *Review of Integrative Business and Economics Research,* 7(2), 24-38. http://buscompress.com/uploads/3/4/9/8/34980536/riber_7-s2_03k18-131_24-38.pdf

Ayele, W. Y. (2020). Adapting crisp-dm for idea mining: a data mining process for generating ideas using textual dataset. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 11(6), 20-32. http://dx.doi.org/10.14569/IJACSA.2020.0110603

Bajaj, P., Ray, R., Shedge, S., Vidhate, S., & Shardoor, N. (2020). Sales prediction using machine learning algorithms. *International Research Journal of Engineering and Technology*, 7(6), 3619-3625.

Bango. (2022). *Board to death: why digital marketing is failing to impress in the board room*. https://bango.com/board-to-death/

Barnhart, B. (2021). *The most important Instagram statistics you need to know for 2021*. Sprout Social. https://sproutsocial.com/insights/instagram-stats/

Behera, G., & Nain, N. (2020). A Comparative Study of Big Mart Sales Prediction. *Computer Vision and Image Processing*, 1147, 421-432. https://doi.org/10.1007/978-981-15-4015-8_37

Bianchi, E., Bruno, J. M., & Sarabia-Sanchez, F. J. (2019). The impact of perceived CSR on corporate reputation and purchase intention. *European Journal of Management and Business Economics*, 28(3), 206–221. https://doi.org/10.1108/ejmbe-12-2017-0068

Bonilla, M., Arriaga, J.L., & Andreu, D. (2019). The interaction of Instagram followers in the fast fashion sector: The case of Hennes and Mauritz (H&M). *Journal of Global Fashion Marketing*, 10(4), 342-357. https://doi.org/10.1080/20932685.2019.1649168

Brownlee, J. (2016). *A gentle introduction to XGBoost for applied machine learning.* Machine Learning Mastery. https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

Brownlee, J. (2019). *How to use learning curves to diagnose machine learning model performance*. Machine Learning Mastery. https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/

Cantón Croda, R., Damina, G., & Omar, C. (2018). Sales prediction through neural networks for a small dataset. *International Journal of Interactive Multimedia and Artificial Intelligence*, 5(4), 35-41. https://www.ijimai.org/journal/bibcite/reference/2668

Carlson, J., Gudergan, C., Gelhard, C., & Rahman, M. M. (2018). Customer engagement with brands in social media: Capturing innovation opportunities. *European Journal of Services Marketing*, 32(1), 83-94. https://doi.org/10.1108/JSM-02-2017-0059

Ceyhan, A. (2019). The impact of perception related social media marketing applications on consumers' brand loyalty and purchase intention. *EMAJ: Emerging Markets Journal*, 9(1), 88–100. https://doi.org/10.5195/emaj.2019.173

Cheriyan, S., Ibrahim, S., Mohanan, S. & Treesa, S. (2018) Intelligent sales prediction using machine learning techniques. *2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE'18),* 53-58. https://doi.org/10.1109/iCCECOME.2018.8659115

Chu, S. C., Kamal, S., & Kim, Y. (2019). Re-examining of consumers' responses toward social media advertising and purchase intention toward luxury products from 2013 to 2018: A retrospective commentary. *Journal of Global Fashion Marketing*, 10(1), 81–92. https://doi.org/10.1080/20932685.2018.1550008

Cooley, D., & Parks-Yancy, R. (2019). The effect of social media on perceived information credibility and decision making. *Journal of Internet Commerce*, 18(3), 249-269. https://doi.org/10.1080/15332861.2019.1595362

Cui, T., Wang, Y., & Namih, B. (2019). Build an intelligent online marketing system: an overview. *IEEE Internet Computing*, 23(4), 53-60. https://doi.ieeecomputersociety.org/10.1109/MIC.2019.2924637

Dabbous, A., & Barakat, K. A. (2020). Bridging the online offline gap: Assessing the impact of brands' social network content quality on brand awareness and purchase intention. *Journal of Retailing and Consumer Services*, 53(3). https://doi.org/10.1016/j.jretconser.2019.101966

Djafarova, E., & Bowes, T. (2020). 'Instagram made me buy it': generation z impulse purchases in fashion industry. *Journal of Retailing and Consumer Services*, 59. https://doi.org/10.1016/j.jretconser.2020.102345

Dong, W., Li, Q., & Zhao, H.V. (2019). Statistical and machine learning-based e-commerce sales forecasting. *ICCSE'19: Proceedings of the 4th International Conference on Crowd Science and Engineering*, 110-117.https://dl.acm.org/doi/pdf/10.1145/3371238.3371256

Dwivedi, et al. (2020). Setting the future of digital and social media marketing research: Perspectives and research propositions. *International Journal of Information Management*, 59. https://doi.org/10.1016/j.ijinfomgt.2020.102168

Globe Newswire. (2021). *Businesses anticipate significant increase in social media investment as it becomes essential for long-term success.* Global News Wire. https://www.globenewswire.com/news-release/2021/04/06/2205108/0/en/Businesses-Anticipate-Significant-Increase-in-Social-Media-Investments-as-it-Becomes-Essential-for-Long-Term-Success.html

Gräve, J. F. (2019). What KPIs are key? Evaluating Performance Metrics for Social Media Influencers. *Social Media + Society*, 5(3). https://doi.org/10.1177/2056305119865475

Gruber, D. A., Smerek, R. E., Thomas-Hunt, M. C., & James, E. H. (2015). The real-time power of Twitter: Crisis management and leadership in an age of social media. *Business Horizons*, 58(2), 163-172. https://doi.org/10.1016/j.bushor.2014.10.006

Huang, Y. T., & Su, S. F. (2018). Motives for Instagram use and topics of interest among young adults. *Future Internet*, 10(8), 77. https://doi.org/10.3390/fi10080077

IBM Cloud Education (2020). *Neural Networks.* IBM. https://www.ibm.com/my-en/cloud/learn/neural-networks#toc-how-do-neu-vMq6OP-P

IBM Cloud Education (2020). *Recurrent Neural Networks*. IBM. https://www.ibm.com/cloud/learn/recurrent-neural-networks

Indiani, N. L., & Fahik, G. A. (2020). Conversion of online purchase intention into actual purchase: The moderating role of transaction security and convenience. *Business: Theory and Practice*, 21(1), 18–29. https://doi.org/10.3846/btp.2020.11346

Jacobson, J., Gruzd, A., & Hernández-García, A. Á. (2020). Social media marketing: who is watching the watchers? *Journal of Retailing and Consumer Services*, 53. https://doi.org/10.1016/j.jretconser.2019.03.001

Karaman, B. (2019). *Predicting sales: forecasting monthly sales with lstm.* Towards Data Science. https://towardsdatascience.com/predicting-sales-611cb5a252de

Knieriem, K. (2019). *The Importance of Sales Forecasting & How It* Impacts *a Company*. Clari. https://www.clari.com/blog/the-importance-of-sales-forecasting

Kotler, P., Armstrong, G. M., & Opresnik, M. O. (2021). *Principles of Marketing*, 18th ed., Pearson Education.

Kusumasondjaja, S. (2018). The roles of message appeals and orientation on social media communication effectiveness. *Asia Pacific Journal of Marketing and Logistics*, 30(4), 1135-1158. https://doi.org/10.1108/APJML-10-2017-0267

Laerd Statistic. (2018). *Pearson product-moment correlation*. Pearson Product-Moment Correlation - When you should run this test, the range

of values the coefficient can take and how to measure strength of association.

https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

Li, F., Larimo, J., & Leonidou, L. (2020). Social media marketing strategy: Definition, conceptualization, taxonomy, validation, and future agenda. *Journal of the Academy of Marketing Science*, 49(1), 51–70.

https://doi.org/10.1007/s11747-020-00733-3

Lin, Z., Madotto, A., Winata, G.I., Liu, Z., Xu, Y., Gao, C., & Fung, P. (2019). Learning to learn sales prediction with social media sentiment. *Proceedings of the First Workshop on Financial Technology and Natural Language Processing (IJCAI'19)*.

https://aclanthology.org/volumes/W19-55/#page=57

Liu, X., Shin, H., & Burns, A. C. (2019). Examining the impact of luxury brand's social media marketing on customer engagement: using big data analytics and natural language processing. *Journal of Business Research*, 125(2021), 815-826.

https://doi.org/10.1016/j.jbusres.2019.04.042

Loukas, S. (2021). *Everything you need to know about min-max* normalization *in Python*. Medium. https://towardsdatascience.com/everything-you-need-to-know-about-min-max-normalization-in-python-b79592732b79

Misirlis, N., & Vlachopoulou, M. (2018). Social Media Metrics and analytics in marketing – S3M: A mapping literature review. *International Journal of Information Management*, 38(1), 270–276.

https://doi.org/10.1016/j.ijinfomgt.2017.10.005

Mittal, A. (2019). *Understanding RNN and LSTM*. Medium. https://aditi-mittal.medium.com/understanding-rnn-and-lstm-f7cdf6dfc14e

Nash, J. (2018). Exploring how social media platforms influence fashion consumer decisions in the UK retail sector. *Journal of Fashion Marketing and Management*, 23(1), 82-103.

https://doi.org/10.1108/JFMM-01-2018-0012

Oumayma, B. (2019). Social media made me buy it: the impact of social media on consumer purchase behavior. *SCA '19: Proceedings of the 4th International Conference on Smart City Applications*, 38, 1–7.

https://doi.org/10.1007/978-3-030-37629-1_18

Park, J., Hyun, H., & Thavisay, T. (2021). A study of antecedents and outcomes of social media wom towards luxury brand purchase intention. *Journal of Retailing and Consumer Services.* 58(2).

https://doi.org/10.1016/j.jretconser.2020.102272

Peña-García, N., Gil-Saura, I., Rodríguez-Orejuela, A., & Siqueira-Junior, J. R. (2020). Purchase intention and purchase behavior online: A cross-cultural approach. *Heliyon*, 6(6).

https://doi.org/10.1016/j.heliyon.2020.e04284

Phua, J., Jin, S.V., & Kim, J. J. (2017). Uses and gratification of social networking sites for bridging and bonding social capital: a comparison of Facebook, Twitter, Instagram, and Snapchat. *Computer in Human Behaviour*, 72, 115-122. https://doi.org/10.1016/j.chb.2017.02.041

Poulis, A., Rizomyliotis, I., & Konstantoulaki, K. (2018). Do firms still need to be social? Firm generated content in social media. *Information Technology & People*, 32(2), 387-404.

https://doi.org/10.1108/ITP-03-2018-0134

Ric, T., & Benazić, D. (2022). From social interactivity to buying: An Instagram user behaviour based on the S-O-R paradigm. *Economic Research-Ekonomska Istraživanja*, 1–19.

https://doi.org/10.1080/1331677x.2021.2025124

Rincon-Patino, J., Lasso, E., & Corrales, J.C. (2018). Estimating avocado sales using machine learning algorithms and weather data. *Sustainability*, 10(10), 1-12.

https://doi.org/10.3390/su10103498

Samala, N., & Katkam, B. S. (2019). Fashion brands are engaging the millennials: A moderated-mediation model of customer-brand engagement, participation, and involvement. *Young Consumers*, 21(2), 233–253.

https://doi.org/10.1108/yc-12-2018-0902

Sari, O. H. (2021). Theory of planned behaviour in marketing: Cognitive consideration on purchase decision. *Golden Ratio of Mapping Idea and Literature Format*, 2(1), 01–07.

https://doi.org/10.52970/grmilf.v2i1.90

Sulthana, N., & Vasantha, S. (2019). Influence Of Electronic Word Of Mouth eWOM On Purchase Intention. *International Journal of Scientific & Technology Research*, 8(10), 1-5.

Teo, L. X., Leong, H. K., & Phua, Y. X. P. (2019). Marketing on Instagram social influence and image quality on perception of quality and purchase intention. *International Journal of Sports Marketing and Sponsorship*, 20(2), 321-332.

https://doi.org/10.1108/IJSMS-04-2018-0028

Top-Hashtags (2021). *Top 100 HashTags on Instagram*. Top-Hashtags. https://top-hashtags.com/instagram/

Tsai, M. C. (2020). Storytelling Advertising Investment Profits in Marketing: From the Perspective of Consumers' Purchase Intention. *Mathematics*, 8(10), 1704.

http://dx.doi.org/10.3390/math8101704

Vandeput, N. (2021). *Forecast KPI: RMSE, Mae, Mape & Bias*. Medium. https://towardsdatascience.com/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d

Voramontri, D., & Klieb, L. (2019). Impact of social media on consumer behaviour. *Int. J. Inf. Decis. Sci.*, 11(3), 209-233.

Wei, W. W. S. (2019). *Multivariate time series analysis and applications*, John Wiley & Sons.

Wu, C.S., Patil, P., & Gunaseelan, S. (2018). *Comparison of different machine learning algorithms for multiple regression on Black Friday sales data*. 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS). Beijing, China.

https://doi.org/10.1109/ICSESS.2018.8663760

Zollo, L., Filieri, R., Rialti, R., & Yoon, S. (2020). Unpacking the relationship between social media marketing and brand equity: The mediating role of consumers' benefits and experience. *Journal of Business Research*, 117, 256–267.

https://doi.org/10.1016/j.jbusres.2020.05.001