Erasmus School of Economics

MSc Data Science and Marketing Analytics

Piro Kristina

Student Number: 523198

Supervisor: Radek Karpienko

How is binge consumption related to status-signalling on Instagram?

Table of Contents

Abstract

This master thesis aims to examine how binge consumption of social media is associated with posting expensive vs regular brands and which are the key factors that mainly contribute. Instagram users seem to have found a way to express themselves through their posts and more particularly the hashtags of various product/brands that they provide. Thus, a sample of 404 followers of famous Greek Instagram influencers who fulfil some selection criteria was scraped and analysed. Both logistic regression and a machine learning algorithm, random forest, were applied and compared for the prediction of the price level (high/low) of brands that users were most likely to tag in their posts. We defined the heavy usage of social media (binge consumption) the high posting frequency of the users. The results revealed that the number of posts that a user had uploaded all the years that his account was active, the followers and followings accounts as well as the likes that a previous post had gained, seem to impact the most. Users with low self-esteem in their way of seeking the approval, seem to be more likely to post expensive products (high price level) with the reward of getting more likes in their posts. Finally, an important managerial implication for a better segmentation strategy is proposed, through the influencer marketing (e.g. nano influencers) for promoting better their brands and reaching out their target audience.

1. Introduction

Social media networks are essential digital spaces, facilitating mainly the diffusion of communication and interaction between different people. ‘‘They virtually unify users from diverse backgrounds, nationalities, ways of thinking, statuses leading to a richer social structure’’ (Kapoor et al., 2018). Interestingly, due to their evolution over the past two decades they have attracted the attention of many scientists by taking the challenge of analysing large datasets (Langlois and Elmer, 2013) and deriving meaningful insights (Van Der Aalst, 2016) regarding user’s/consumer’s behaviour. Mining raw social media data aids to identify applicative patterns and trends regarding the users (Thakre et al., 2020).

Instagram is one of the most popular social media platforms for sharing photos of various rich content. While 50 million businesses are active on Facebook pages, more and more businesses use Twitter and Instagram for marketing purposes (Lister, 2017). The number of companies particularly in the fashion and apparel industry as well as big luxury brands active on Instagram is rapidly increasing as their main target group consists Instagram users (De Vries et al., 2012).

Social media unifies technology and social interaction, as it connects to a more intimate level the brand and the user/consumer. Users tend to signal various products through their posts indicating a way of expressing themselves, their preferences, and choices (Grewal et al., 2019), consisting an interesting angle of investigating consumer identities. Therefore, we assume that this signalling tendency of users is substituting the actual consumption of the product-brands that are posted.

In this master thesis, we are particularly interested in recognizing a potential a linkage between the binge consumption of social media (posting frequency of a user) with the signalling identity that a user reveals through his posts. This leads us to the main research question of this thesis:

“Which key factors have stronger impact on a binge consumption of Instagram users’ postings about expensive status brands vs. regular brands?”

In this master thesis we answer this question considering all the conceptual framework available following a different approach regarding the analysis. So far, the analysis taking place used either qualitative data by surveys conduction or taking the time that the user spends (Internet Usage Pattern) in a social media network. Our approach will be based on the available content of the Instagram posts and analyse the tags of various product brands depicted on the user's posts.

2. Related Research

2.1 Posting as a way of self-expressing

Consumption is a way for consumers to self-express and self-define themselves and their behaviour (Jensen and Gilly, 2003). What is more, consumers usually construct perceptible identities, created by sentimental attachments to material things. The choice of brands reveals information of their identity and even more what is ‘‘self-relevant’’ to them (Jensen and Gilly, 2003). *‘‘By wearing luxury brands, consumers demonstrate high levels of well-being, as luxury products can escalate their social status’’* (Kusumasondjaja, 2019). This exposure could be identified through their Instagram posts.

Brand companies tend to benefit from the positive effects of their sales by social media especially nowadays that digital marketing era is rapidly increasing (Leeflang et al., 2014; Järvinen et al., 2012). Precisely, in 2011 more than a half of social media’s population started following brands resulting to companies’ investments on social media reaching almost USD 4.3 billion (Leeflang et al., 2014; Williamson, 2011). There is an estimation of an increase (USD 5 to10 billion) of the social media advertising expenditures in the next two years (Gallagher, 2018; Kusumasondjaja, 2019).

In this way, users got influenced by companies’ brands, gradually ‘‘becoming fan’’ of the respective brands interacting with them through likes and comments, and finally tagging them in their own personal images. Users by tagging various brands of in social media, are showing except for their loyalty, their preferences which correspond to a certain status, price range etc, signalling in that way themselves.

Instagram is also broadly used, utilizing consumption related information. It has been observed that by posting photos or/and uploading stories of recent purchases, shopping experiences, or content of desired products, users indicate their purchase preferences of the (near) future (Zhang et al., 2017). Firms apparently take advantage of the marketing channel of social media to post information about their various promotion activities, advertise new products etc and generally expose their brand more.

Lastly, individuals who have extraversion and agreeableness personality traits behind selfie-posting, tend to post more as a need of communication and acceptance from other people in social media (Chaudhari et al., 2019). Therefore, communication consists the major motivation for posting selfies as it is followed by their need to gain more attention through their posts (Chaudhari et al., 2019).

2.2 Behavioural addiction

Lifted feelings of self-esteem can lead to an adoption of an impulsive/binge behaviour. Notably, extraverts are more vulnerable to binge consumption of social media (Wang et al., 2015). Gender consists also an important indicator that influences the impulsive (buying) behaviour with companies developing effective marketing strategies based on the gender group (Khan et al., 2016).

Nowadays, the exhibition of the individual's self-control in social media networks is quite concerning, given the increased people's spending time in a daily level. According to Wilson et al. (2010) personality characteristics such as extroversion, conscientiousness and self-esteem reveal individuals with more time spent on social networks and addictive tendencies towards them.

According to Khan et al. (2016), the existence of ''*lack of control, hedonic motive and materialism*'' stimulate impulsive (buying) behaviour. ''*Impulsiveness is related to a shortage of self-control*'' (Vohs and Faber, 2007) with the most recent generation- millenniums being more susceptible (Pelling et al., 2009; Hughes, 2008). Big luxurious firms such as Gucci, Dolce&Gabbana, Louis Vuitton already use social media to appeal millennial consumers with Gucci selling half of its products to them (Handley, 2018; Kusumasondjaja, 2019). As highlighted through many researches (Steinfield et al., 2008; Cramer et al., 2016; Raymer, 2015), an intime engagement in the virtual world of social media intensifies the self-esteem of an individual causing the lack of his self- control, resulting to a binge consumption of social media.

Binge behaviour is also considered an addiction as ''*individuals often engage in such behaviours to escape reality*'' (Gold et al., 2003). Therefore, binge behaviour has been characterized as a result coming from an "*excessive amount in a short time*" (Gold et al., 2003). Addictive behaviours have appeared and been examined in various ways in respect the field of study (Chou and Ting, 2003). According to rational addiction theory, the involvement of both reinforcement and tolerance do consist an addictive behaviour (Becker et al., 1991). Reinforcement includes by its definition, that a higher consumption of an addictive good of an individual in the past, leads to a greater increase of his desire for the same good in his current consumption. In this case, the meaning of reinforcement corresponds to the gratification through the number of likes and comments that a user can get in a post. Thus, a user who received in the past (e.g. his last post) a high number of likes or/and comments, is more eager to post more in short term.

Contrary to reinforcement, tolerance supports that the more utility you gain from an addictive good in the past, the lower in the future will be obtained (Gruber et al., 2001). In our case the ‘‘addictive good’’ is considered the binge consumption of social media (Instagram), quantifying it through the posting frequency of an Instagram user. Particularly, a user derives utility from signalling brands in his posts as a way of self-expressing and ‘‘consuming’’ the products he is posting.

The willingness of an individual to constantly desire the maximization of his utility over time is what rationality is defined by Becker et al. (1999). However, when the past consumption of a good affects in a high level the current consumption, the phenomenon of strong indication of addiction is distinguished. In respect to the rational model of Becker et al. (1999), each individual has a different level of utility depending on the consumption of two goods. The current utility is dependent on a measurement of the past consumption. Aim of every individual is to maximize his utility resulting in his highest satisfaction. A rational person maximizes his utility subject to a constraint on his expenditures, as well as his budget.

Furthermore, there is a positive significant impact on purchase expenditures between the engagement of a user and a brand in social media (Goh et al., 2013) as well as a ‘‘*higher participation and customer patronage*’’ (Rishika et al., 2013). The commutation from a strong preference of a product/brand or activity to become a habit is not far away but exciding a certain point (Chou et al., 2003). This comes to agree with the purpose of this paper, assuming that binge consumption of social media (defined by posting frequency of a user), leads to a higher participation of wanting to acquire an expensive brand (stated by the tags in a user’s posts).

Wilcox and Stephen (2013) showed that the usage of social media network tends to enhance the ‘‘*self-esteem of users who are focused on close friends*’’. Additionally, it was observed that ‘‘*greater social network usage is associated with higher levels of credit card debt for individuals whose ties with social media were strong enough*’’ (Wilcox and Stephen, 2013). Focusing more on the finding regarding higher levels of credit card debt, it can be clearly assumed that a user who tends to spent more time (e.g. by repeatedly posting) in social media (e.g. Instagram) does make less rational choices in a sense that he is making more impulsive purchases creating a debt in their credit cards. By taking into account the wide variety of the literature review we are going to address the following hypothesis:

H1: Users who want to boost their self-esteem use Instagram more intensively. *Thus, the amount of likes and comments of a post have an impact on the binge consumption of social media.*

H2: *Users in order to get more likes are more likely to post luxury expensive brands.*

Users are in a way encouraged to increase the posting frequency as they associate posts with a ‘‘reward’’ in forms of bigger amount of likes and comments. Consequently, they probably connect the more posts that will provide in their Instagram account with more interactions by their followers as the extroversion and self-esteem are the characteristics prone to a binge consumption of social media. The enhancement of their self-esteem and lack of self-control, is a reflection from the high exhibition to social media networks (Wilcox and Stephen, 2013).

3. Data Collection

For the purposes of the analysis, the sample data collection included the content (tags of brands) and reactions (e.g. likes, comments) of the user's Instagram posts in order to identify the impact of the brand signalling on Instagram users. Additionally, some information regarding the user's Instagram account were obtained, such as number of followers and following accounts as well as the number of posts and added as predictor variables in the dataset.

Particularly, we scraped the data from 404 Instagram followers of 10 Greek female influencers that were selected after meeting some specific selection criteria as it will be mentioned later. First, the users with private profiles were omitted as there is no possible way of having access to any further information. The reason behind choosing Greek influencers as a starting point of obtaining the raw data, is the fact that their followers are expected and assumed to be a representative audience. This is because the selected sample of users is more active in social media (Instagram), as they post more, interact by tagging various brands and participate in competitions/ challenges that take place from the influencers that they follow. Additionally, another reason of choosing Greek influencers is that they tend to collaborate with various brands (mainly local-Greek) and promote their products by creating various competitions known as giveaways and thus, obtaining a massive number of followers that keeps augmenting.

To obtain the users that consisted the final dataset, some selection criteria regarding the choice of the Instagram user's accounts were applied, as mentioned below.

- The number of posts per user should have been bigger than 150 within the time frame (2017-2019), excluding the year of 2020 where the global pandemic could influence.
- The users could also not exceed the number of 20 thousand followers, to avoid professional accounts (people who are getting paid by brands/companies to promote their products like influencers and thus they are expected to post more).

- Users who had clearly stated that consist business accounts were also omitted for the same reason. All the selection criteria that were applied, were aiming to obtain a sample that would be as much representative as possible.

Next, the tags of the brands that each user posted through his photos/videos were manually assigned into two levels: Low and High price level, which are mainly taking into account the average brands' cost. A protocol is provided below, explaining how the price levels for the various brand-tags were assigned.

3.1 Protocol of Price Level Definition

In this thesis the dependent variable that will occur in the analysis is a dummy variable defined as *PriceLevel* (Low/High). The price level is a variable that was created from the tags that the users used in their posts that correspond to various brands. Specifically, when only one brand occurs as a tag in a post, the average value of that brand is considered and assigned as a high or a low-price level brand. From the beginning of this thesis we assume that the tags of brands are products that the Instagram users desire, want to acquire or already have and show their satisfaction. In the case that more than one brand occurs in a post, the majority vote is taken into account, meaning that the brand that is tagged more times is selected. The price levels of the products are filled in a manual way and left blank when the user did not tag any brand.

The way of assigning the corresponding price level to a brand is mainly by visiting the website of the brand/company and looking at the average price of the products. For instance, brands of the INDITEX parent company such as ZARA, Bershka, Oysho, Stradivarius, Massimo Duty, and brands like adidas, nike, which occur quite often, are considered Low Price level brands as an average consumer can from an economic perspective afford to purchase. On the other hand, tags of brands like Dior, Luis Vuitton, Gucci, are considered as High Price level products, as their market price is relatively high, and they target a specific group.

Table 1: Coding price levels from brand tags

UserID	PubDate	BrandTag	PriceLevel	Mean Price Values
2	2019-07-17T16:16:09	pull&bear	Low	35
2	2018-09-16T15:28:40	Bershka Men	Low	30
5	2019-04-19T10:14:56	Adidas	Low	70
7	2019-12-25T18:58:49	Ducati Motor Holding	High	14000
7	2019-12-25T18:58:49	Lamborghini Squadra Corse	High	300000
7	2019-10-23T15:25:59	Adidas	Low	70
24	2018-05-08T05:12:26	michaelkors	High	270
207	2019-12-29T00:09:51	Louis Vuitton Official	High	1400
250	2017-05-15T21:22:31	Muacosmetics	Low	10
252	2019-09-08T18:13:12	bmw	High	50000
252	2019-09-08T13:02:40	Porsche	High	100000
314	2019-02-20T18:24:20	ZARA Official	Low	37
314	2019-01-23T17:40:11	Bershka	Low	30
314	2018-12-30T15:39:51	Swiss Watches since 1853	High	300
314	2018-12-30T14:13:57	tissot_official	High	570

3.2 Data Description

The final dataset that will be analysed consists 404 unique Instagram users and their 19970 posts, considering only the posts that took place between the time frame of 2017 to 2019 and those that people tagged a brand (so a price level of that brand was assigned). The reason for the choice of this specific time frame is because Instagram began to gain more popularity the last couple of years and thus, we wanted the user's posts to be comparable with each other. Therefore, each unique user had posts for each of these years excluding the year of 2020 as users may have been affected by the global pandemic COVID-19.

The *username* and *postUrl* have been removed for privacy reasons and replaced by a unique *UserId* variable. Furthermore, the variables *description*, *location* and *brandtags* (the name of the brands that users tagged) were also omitted due to no additional value as well as the fact that a lot of missing values occurred. The final dataset consists of 13 variables were 7 are numeric, 6 categorical and 19970 observations. The unit of the analysis is posting level data. The descriptive statistics (Table 2) with all the features of the dataset is indicated below, while the extent description of the variables can be found in Appendix Table 3.

Table 2: Descriptive statistics

Variable	**Type**	**Mean**	**Sd**	**Min**	**Max**	**Majority Class**
TotalPosts	numeric	874.7	988.81	150	5618	-
likeCount	numeric	109.6	160.8	0	2897	-
commentCount	numeric	2.53	37.4	0	2679	-
followersCount	numeric	2401	3927.64	78	19595	-
followingsCount	numeric	2204	1941.89	68	7500	-
viewVideocount	numeric	17.58	158.02	0	8711	-
TFPosts	numeric	2.1	2.91	1	96	-
hour	numeric	14.08	5.17	0	23	-
day	categorical	-	-	-	-	Friday=3218
month	categorical	-	-	-	-	August=2180
year	categorical	-	-	-	-	2019=7747
PeopleTags	categorical	-	-	-	-	No=16598
type	categorical	-	-	-	-	Photo=18731
PriceLevel	categorical	-	-	-	-	High=11775

The *TFPosts* is the variable that had been created to count the actual number of posts of each user for these 3 years that are analysed, for a certain period of time (day, month and hour). Regarding the variables *followersCount*, *followingsCount* and *TotalPosts*, they provide information though all the years that a user's account was active. Feature extraction from the time stamp, obtaining the year, month, day, hour was implemented and added to enhance the analysis. In the case of the *day* and *month* variable, days from Monday to Sunday and the 12 months respectively were obtained.

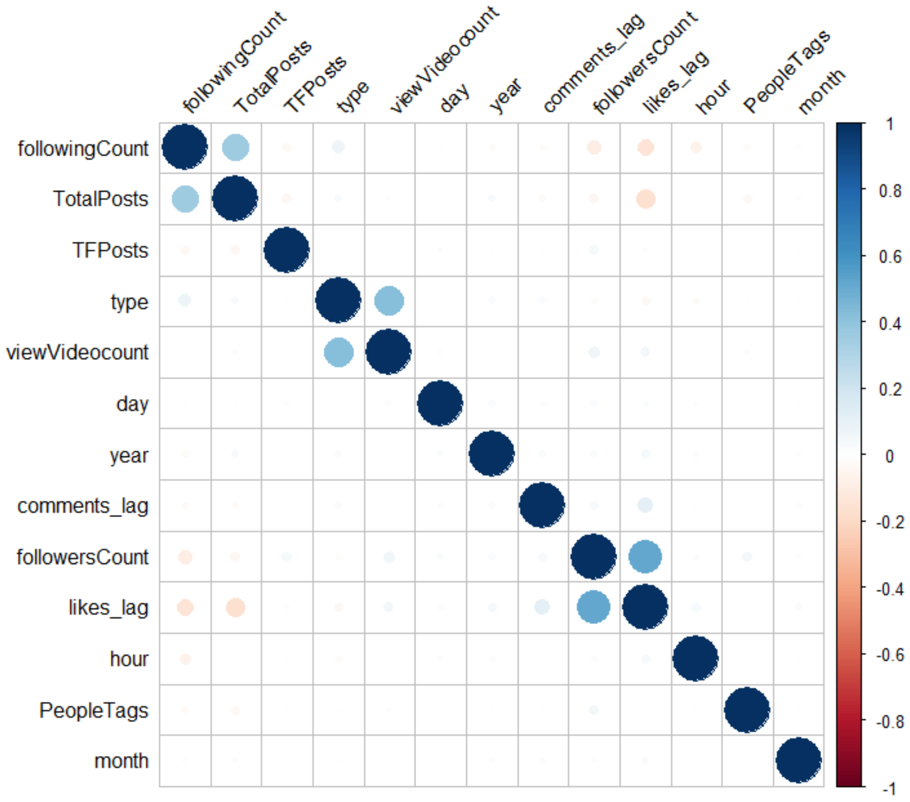
From Table 2, some interesting distributions are observed. The average number of *commentCount* is only 2.5 while the *likesCount* is 109.6 with the maximum number being almost 2897 likes for a certain post. Similarities in the mean of *followersCount* and *followingCounts* are detected, 68 and 78 respectively while the maximum value of the *followersCount* is 19595.

Regarding the dummy variable *type*, photos occur more often compared to video posts, in this dataset. For another dichotomous variable *PeopleTags*, it seems that in most of the posts, users did not tag another person. Additionally, as far as the posting behaviour of the users is concerned, most of the posts took place on Fridays and users used in their posts more tags corresponding to High price level product/brands. The year of 2019 had the most posts which can be explained from the fact that the popularity of Instagram has rapidly increased the last couple of years compared to other social media networks (e.g. Facebook, Twitter, Snapchat). Finally, the months that a greater number of posts were taken were August and December probably as they are considered important months of the summer and winter season respectively were celebrations and vacations usually occur. So, people are also more active generally in social media and in posting more often. The months of July, September and October are following with February being the least popular month.

The missing values that occurred in the dataset were mainly coming from the *PriceLevel* variable were people who did not tag any brand were left blank. Specifically, 105246 missing values were in the *PriceLevel* variable and 3 in the *commentCount*. The removal of the missing values took place since filling the NAs (e.g. with an average value) was not possible and could probably affect the results of the analysis. Finally, another reason for removing the NAs is due to the model being used. Random forest cannot handle any missing values for both dependent and independent variables. After the removal of NAs, the final number of observations is 19970.

Additionally, based on the literature review, we assume that the likes and the comments play an important role in the level of confidence of a user. Instead of analysing the actual number of likes and comments that each post received, it would be more interesting to investigate the value of the lagged values of likes and comments. The lagged values of likes and comments of a post, count the number of likes and comments respectively that the previous post of a user gathered. Both had been added in the dataset and will be used in the models instead of the actual likes and comments. Before every analysis, an important step is to check the correlation of the predictor variables for multicollinearity. The correlation matrix is showed in Graph 1.

Graph 1: Correlation of numeric variables



The *followersCount* and *likes_lag* seem to have the strongest positive correlation with each other, implying that when the number of followers of an Instagram user increases, the number of likes that the previous post gained, also increases. Another positive high correlation occurs for the variables *TotalPosts* and *followingCount*. Correspondingly, almost the same correlation for the *type* of the post and the *viewVideocount*. A small negative correlation between the total number of posts that a user had through all the years that his account was active with the lagged number of likes was also observed. Finally, a minor negative correlation seems to exist between the *likes_lag* the *followingCount*. Since we do not have to deal with very strong correlation between certain variables, all the variables will be kept and proceeded in the analysis.

4. Methodology

In this section the two models that will be used for the analysis part of this thesis paper are going to be discussed in detail. First the logistic regression (the most common approach for binary classification problems) and next, a machine learning /black box model, the random forest.

The usage of logistic regression is due to the dependent variable (*Pricelevel*) which gives a dichotomous outcome (whether an Instagram user tagged in his post a brand that was high vs low price). We will compare the accuracy of the two models (logistic regression and random forest) to see which model performs better and which predictor variables influence more according both models, answering in this way the research question.

4.1 Logistic Regression Model

Logistic regression is a statistical method which models binary (when the dependent variable has two classes) and multinomial (with more than two classes) outcome(s). In contrast to linear regression which works and handles only numeric outputs, logistic regression estimates probabilities by measuring the relation between the categorical dependent variable and the independent ones. All the formulas that will be used in this thesis are following the notation of (James et al., 2013). Formula (1) is indicated below:

$$\mathrm{p}(X)=\frac{\mathrm{e}^{\beta_0+\beta_1 \mathrm{X}}}{1+\mathrm{e}^{\beta_0+\beta_1 \mathrm{X}}} \quad (1)$$

with $\mathrm{p}(X)$, giving a probabilistic output taking values between 0 and 1. Logistic regression transforms and assigns the probability of X to belong to a certain category. We use the logit function to avoid the prediction of negative values (as it assigns the values to be in the range between 0 and 1), so no matter the value of X, the prediction will be sensible (Boehmke and Greenwel, 2019). Thus, low values of $\mathrm{p}(X)$, predict the probability of a user tagging a low-price brand, as close to, but never below, zero. Respectively for high values we predict a probability close to, but never above 1 (James et al., 2013).

Additionally, regarding formula (1), $\mathrm{e}^{\beta_0+\beta_1 \mathrm{X}}$ is equal to $\frac{\mathrm{p}(X)}{1-\mathrm{p}(X)}$ and the log $(\frac{\mathrm{p}(X)}{1-\mathrm{p}(X)})$ is called log-odds, and is frequently used instead of the probabilities to interpret the coefficients.

The interpretation of the logistic model differentiates with the linear regression. For logistic regression an increase in X by one unit, changes the log odds by the exp of β_1 (James et al., 2013). In order to interpret the coefficients of the logistic regression we first look at the sign of the coefficient (positive or negative) to see if that predictor variable increases or decreases respectively the odds of a user tagging a high vs low price level brand (dependent variable). Next, to compute the exact magnitude of that predictor variable, we take the exponential of that coefficient (e.g. exp (β_1)). The interpretation for numeric and categorical variables is different as for the categorical predictors we interpret the variable compared to a level/ class which has been kept from the model as a baseline.

In the logistic regression we can adjust a threshold value in order to achieve higher predictive accuracy as well as to maximize both sensitivity and specificity. The threshold value that is most frequently used is 0.5 and if the probability is larger than 0.5 then the user is most likely to be tag in his post a high price level brand while smaller than that value is the opposite, respectively.

4.2 Random Forest Model

Decision trees is a nonparametric algorithm used for modelling prediction. It provides intuitive insights which are understandable as they are close enough to a ‘‘*human's being decision making approach*’’ (James et al., 2013). There are some advantages in terms of interpretation as the results can be explained to non-experts and the fact that can handle both qualitative and quantitative variables. Decision trees for classification problems are known as CART and their goal is to maximize as much as possible the Gini index. The Gini index measures the purity of each class and its formula is indicated below:

$$\mathrm{G} = \sum_{k=1}^{K} \hat{p}_{mk}(1-\hat{p}_{mk}) \text{ (2)}$$

In formula (2), K are considered the classes, and $\hat{p}_{mk}$ the ratio of the observations in the training set in the m node from a certain k class. This index captures the total variance across the K classes. The smallest the value of Gini (when $\hat{p}_{mk}$ gets close to zero or one) the purest the predictive class will be.

The aggregation of many decision trees consists a powerful machine learning model known as random forest, and its special case bagging. Random forest is considered an extension of decision trees and provides an improved predictive performance compared to the traditional trees. This model comes to overpass the main disadvantage of decision trees which is high correlation, by decorrelating the trees and taking each time a random sample of predictor variables by the bootstrapping method. This results to a better predictive accuracy as well as the decrease of the error rate. Similar fundamental principles as the ones of decision trees are used to construct random forest (James et al., 2013).

Random forest contains several hyperparameters that after being tuned, the model performance also increases. The most important hyperparameters are the number of trees, the number of features, indicated as m_{try}, the number of nodes etc., that should be included in each split of the training set (Boehmke and Greenwel, 2019). It is also important to note that the ''running time'' augments linearly with the number of trees. A large number of trees may decrease the error rate and provide more robust results, with a computational expense.

The optimum number of trees is the one that gives the lowest out-of-bag (OOB) error or after that the error is continuing to be stable and low. The default value in R for the number of trees is 500 however may not be the optimal one. The number of features m_{try} determines how many predictor variables should be used in each split. A special case of random forest consists the bagging method, where m_{try} is equal to the number of predictor variables, p. This means that the model includes simply all the predictor variables in each split. Thus, the m_{try} in random forests ''*controls the split-variable randomization features*'' (Boehmke and Greenwel, 2019) helping to maintain a low correlation between the trees. The default value m_{try} of random forest (for classification), is equal to$\sqrt{p}$. Similarly, optimum values of m_{try} can be chosen according the lowest OOB error rate.

Random forests are constituted from several individual decision trees where the depth and complexity of the individual trees need to be controlled. Hyperparameters that occur the depth and the complexity are the node size, max depth, and number of terminal nodes (Boehmke and Greenwel, 2019). The node size needs to be adjusted taking into account the trade-off between accuracy and ‘‘running time’’ of the model.

4.3 Out of bag (OOB)

For classification trees in bagging/random forest we use *K* bootstrapped training sets and track the class for a given test of observations. Next, we take the majority vote that occurs most common in these *K* predictions that we get from these *K* bootstrapped training sets.

On average, the 75% of the observations is used for each bagged tree while the remaining 25% is predestined to fit each bagged tree and is what we have been referring to, the out-of-bag (OOB) observations. Thus, since we want to acquire a single prediction, we take the majority vote as it serves the classification goal.

4.4 Bootstrapping

Random forest builds the trees by using bootstrapped sets based on training data which are different with each other. The method of bootstrapping yields to low variance, better prediction and more accurate results. Since we have been referring to the use of bootstrapping samples in random forest and bagging, we need to clarify what consists a bootstrap method.

The bootstrapping is a resampling statistical technique which is used to improve the high correlation caused by the decision trees. From the original dataset, the method generates several *B* bootstrap training samples (e.g. 1000 or 5000) $Z_1, Z_2..., Z_B$ by sampling with replacement. This means that within the original sample the same observation may appear in the bootstrap sample more than once. Each of these generated bootstrap sample has the same size with the original dataset. Due to the fact that we deal with a classification method, by each of the *B* samples (trees) we get a single prediction, which is the most commonly occurring class among the *B* predictions. The advantage of this statistical technique is that it results to a better estimation of predictive accuracy and avoids dealing with the problem of overfitting even in an increase of the number of trees.

4.5 Modelling Random forest

Random forest helps decreasing the high correlation between the trees as they infuse the element of ‘‘randomness’’ into the tree-growing process. The trees constructed are not pruned, and each individual one has high variance, but low bias. The aim of random forest is the variance reduction which is accomplished by aggregating all these trees. For the construction of each tree, a random set of predictor features is selected in each split where a subset of the predictors determined by the m_{try}, is used. This has as a result the same equal chance for each predictor variable to be selected and not only the ones that were highly correlated. As the vote class that occurs more often is taken for the overall prediction, both elimination of bias and minimization of variance are accomplished.

Finally, random forest as a black-box model is quite difficult to interpret, however the visualization of the variable importance plot provides the most significant information. In the variable importance plot, the Mean Decrease Accuracy and Mean Decrease Gini are showed. The Mean Decrease Accuracy refers to how the prediction accuracy of the model decreases each time that a certain predictor variable is being permuted. The Mean Decrease Gini gives us information about how pure each class is. A predictor variable with a large value denotes a high importance for the model.

4.6 Performance Metrics

For the performance of random forest, the confusion matrix will be used were the accuracy, sensitivity, specificity will be discussed and evaluated as well. A confusion matrix is used to compare the actual outputs with the predicted ones. The original dataset is usually split into a training and evaluation (test) set. The terms that are included in a confusion matrix and used for the accuracy, precision and sensitivity (recall) are the following ones:

True positives (*TP*): These include the cases where the model predicted ‘Low’ (Instagram users that are more likely to post a low-price level brand) and they did post.

True negatives (*TN*): These include the cases where the model predicted ‘High’ (Instagram users that are likely to post a high-price level brand), and that users did not post a high price level brand.

False positives (*FP*): These are the cases where the model predicted as more likely to post a ‘Low’ price level brand, but these users did not post a low-price level brand.

False negatives (*FN*): These are the cases where the model predicted as not likely to post a ‘Low’ price level brand, but these users actually posted.

The accuracy in the confusion matrix gives an overall description of how often the classifier was correct regarding their predictions while the sensitivity and specificity are considered statistical measures of the performance. The equations for accuracy, precision and recall are indicated below.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \text{ Precision} = \frac{TP}{TP + FP}, \text{ Recall} = \frac{TP}{TP + FN}$$

5. Results

In this section the results generated from the models will be explained in detail. There will be discussed first the results of the logistic regression, next the ones of the default random forest that had been created and finally the optimized one. They will be compared in terms of accuracy and the other performance metrics. After accomplishing the appropriate pre-processing data preparation as well as some variable transformations the data was split into training (70%) and (30%) test or validation set. First the *UserId* variable was omitted from the dataset as it would not add any value as well as the *likeCount* and *commentCount* were excluded and instead of them, the lagged values of these variables were used.

From Table 4 below, variables like type = video, the year of 2019, months like January and June, the lagged values of likes etc., have a positive statistically significant effect in the dependent variable. Particularly when a user is posting a video, the odds of tagging an expensive brand, are enlarged by a factor of (exp 2.437e-^{01}) =1.27, compared to post a photo, with the rest of features remaining the same.

Additionally, the months of January and June also seem to have a positive significant effect increasing the odds by a factor of 1.30 and 1.35 respectively compared to a post taking place in the month of April (which the model kept as a baseline) and ceteris paribus. This practically means that a user is more likely to tag an expensive (high price level) brand in January and June. Interestingly, the total number of posts and the number of followers according the model seem to have a negative impact on someone posting an expensive brand. Thus, an increase of the total posts or followers will decrease the odds of posting a high price level brand by a factor of 0.99 respectively (ceteris paribus).

The logit model also does not consider statistically significant the lagged number of comments that a previous post might have taken. However, as expected, the lagged number of likes of a user's post seem to increase the odds of a person tagging an expensive brand by a factor of 1 (ceteris paribus). This means that the user is highly influenced by the fact that his previous post gained a lot likes that could have been due to an expensive brand/firm being tagged in his post. Extensive results of all the coefficients and the respective level of significance can be found in Appendix Table 5.

Table 4

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	7.866e-01	1.103e-01	7.131	9.93e-13	***
typeVideo	2.437e-01	9.135e-02	2.668	0.007626	**
followersCount	-7.369e-05	6.056e-06	-12.168	< 2e-16	***
TotalPosts	-1.744e-03	4.950e-05	-35.224	< 2e-16	***
year2019	-1.435e-01	4.895e-02	-2.931	0.003375	**
monthJan	2.676e-01	1.010e-01	2.650	0.008046	**
monthJun	3.049e-01	9.799e-02	3.112	0.001860	**
dayMonday	2.258e-01	7.133e-02	3.165	0.001548	**
daySunday	1.564e-01	7.070e-02	2.212	0.026961	*
dayThursday	2.193e-01	7.215e-02	3.039	0.002371	**
TFPosts	-3.779e-02	8.499e-03	-4.446	8.74e-06	***
likes_lag	6.052e-04	1.452e-04	4.169	3.06e-05	***
comments_lag	-2.353e-03	1.612e-03	-1.459	0.144493	

*Note : Signif. codes : 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘’ 1

Next, a default random forest was created to see the performance in terms of accuracy. For that reason, without tuning any hyperparameter, the *mtry* (the number of predictor variables used in every split of the trees in the model) was by default the square root of the sum of the predictor variables (3). As far as the number of trees are concerned, the default number was also used, 500 trees. As it was already mentioned, the training set was used to train the model while for testing it, the validation set was considered. The accuracy of the default model is 90% (see Table 6). The Out-of-bag error rate of the default model is 8.45% (see Table 6) which is relatively low. Regarding the rest performance metrices, the sensitivity is quite high 92% while the specificity was approximately 90%. In this classification problem we are more focusing on the accuracy performance. Even from the default random forest model it is observed a 20% higher accuracy compared to the logistic regression.

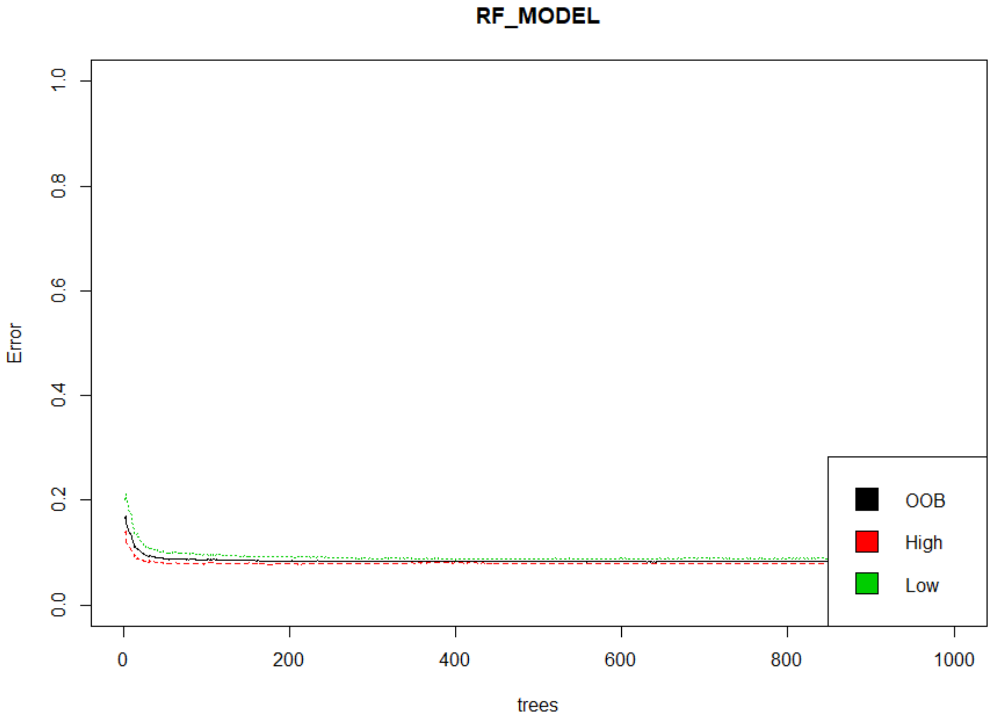
The next step was the hyperparameter tuning and optimization of the random forest model. The optimum number of *mtry* was obtained by performing a 10 *k*-fold cross validation in the training set. Additionally, because we deal with a classification problem and want to obtain the higher predicting accuracy, the metric= ‘Accuracy’ was chosen and the method= ‘rf’ during the grid search of 10 *k*-fold cross validation. The grid search provided a corresponding accuracy, sensitivity, and specificity for different values of *mtry* from 1 to 50.

The optimum number of *mtry* that was suggested by the grid search for the random forest model is 10, which presented the highest accuracy level as can be seen below in Graph 3. Another important hyperparameter that needs to be also tuned is the number of trees. Having a closer look at Graph 2 showed below, we observe that from 200 number of trees and after, the OBB error rate seems to be stable without any fluctuations until the 1000 number of trees. For that reason, 1000 number of trees was selected as optimal for the final model. Even in the case of choosing 5000 trees, the error wouldn’t decrease further but the computation time for the model to run would be quite higher. Therefore, as it was already mentioned in the methodology section, a higher number of trees does not necessarily mean that is the optimal one. What can also be seen from Graph 2, is the lowest and highest error rate as well. The model seems to predict slightly better the users who tagged a High-price level brand compared to the Low-Price level ones.

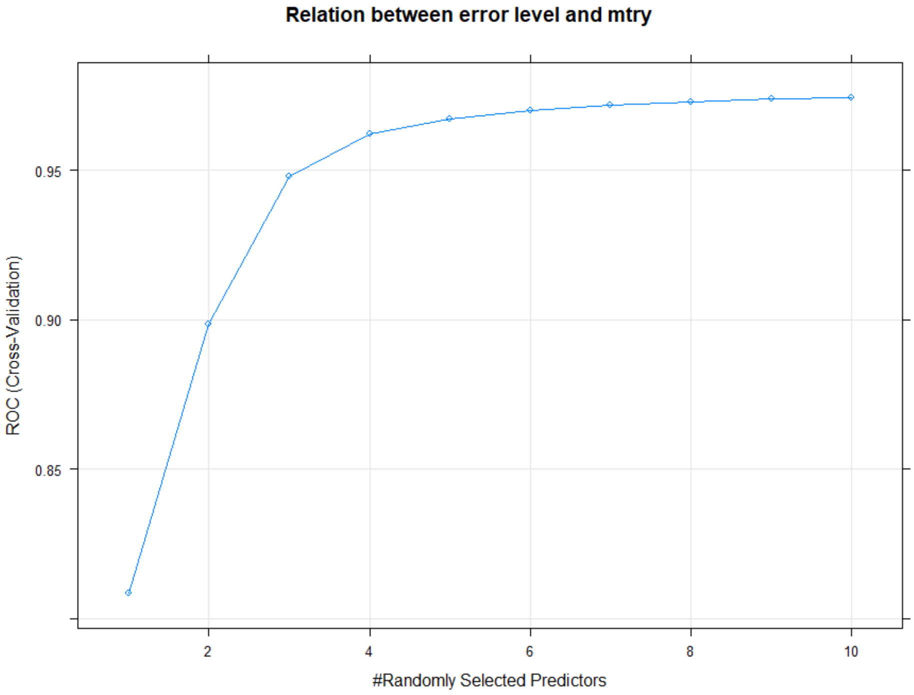
Table 6

Model	Accuracy	Sensitivity	Specificity	OOB Error Rate %
Logistic Regression	70%	68%	71%	-
Default Random Forest	90%	92%	90%	8.45%
Optimized Random Forest	92%	93%	90%	7.81%

Graph 2: OBB error rate *vs* Number of Trees



Graph 3: Distribution of accuracy with the *mtry*



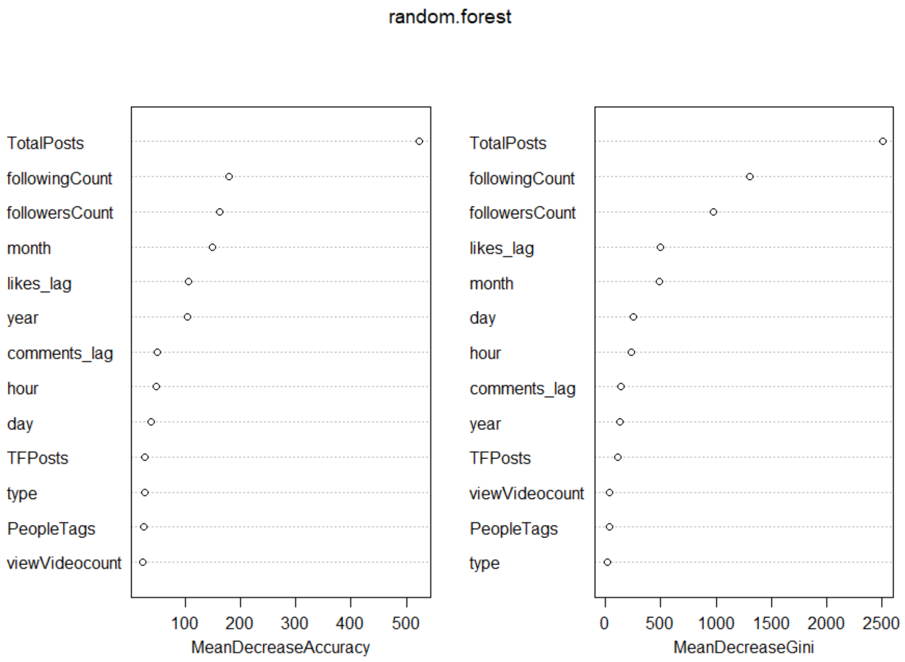
The optimized random forest model with the hyperparameters tuned, increased the accuracy to 92.4% (see Table 6). The sensitivity changed also to 93% while specificity remained stable to 90% respectively. The OOB error rate decreased to 7.81% which is not really a big decrease, but it was already quite low. While the accuracy level of a black box model such as random forest is relatively higher compare to the traditional regression models, their interpretability is quite hard. For that reason, the variable importance plot was generated.

Having a closer look at Graph 4 and Graph 5, it seems that the most important variable for a user to be more likely to post an expensive brand is the total number of posts that he had from when his account was active. This means that permutating this predictor variable, will have a negative impact on the prediction accuracy as well as in the purity (Gini Index). Additionally, the number of the accounts that a user follows and is followed by seem to be distinguished as important predictors. Thus, the more followers and followings a user have, the more likely to post an expensive versus a regular brand.

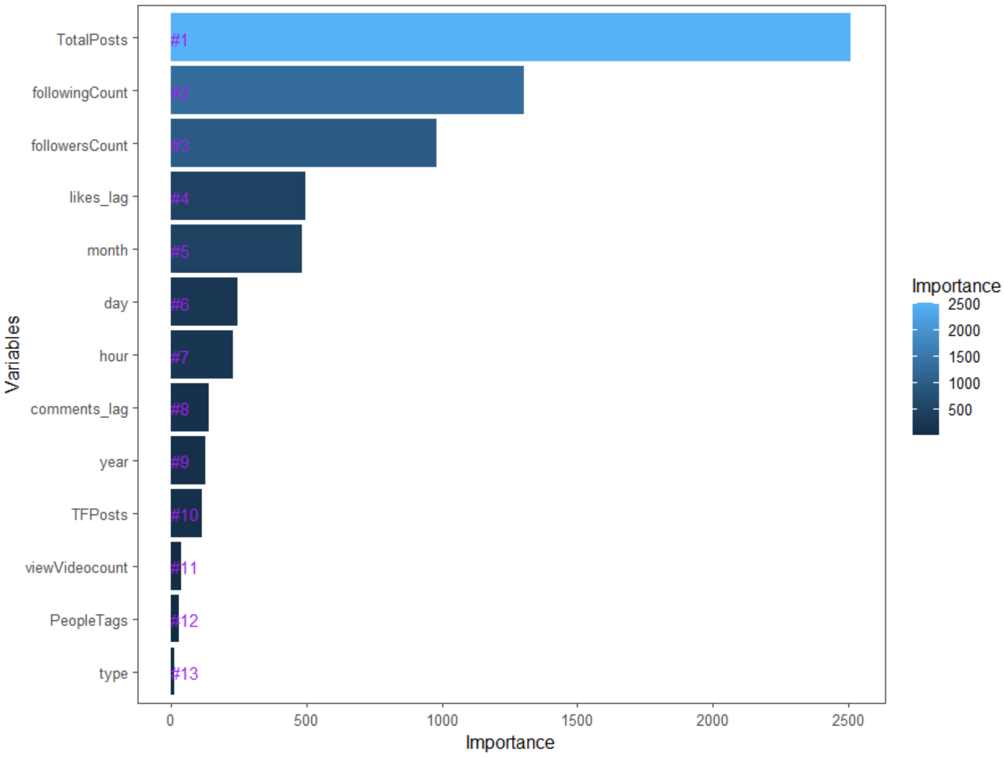
Likewise, in logistic regression, the number of the likes that the previous post took (*likes_lag*) is indicated important as well as the month of the posts. That means that a part of our first hypothesis as well as the second one is also accepted as we assumed that the number of likes act positively in boosting a user's self-esteem and leading him to more likely tag an expensive brand in his post. Thus, the user considers that by posting a 'big', well known and expensive product/ brand will enhance the number of likes he will get in his post.

On the other hand, the *comments_lag* seem to be less important in leading a user tag an expensive brand in his post, resulting to reject part of the first hypothesis. The predictor variable that indicate the less importance (approximately zero) is the type of the post and the dummy including whether in a post was tagged another person.

Graph 4: Variable Importance



Graph 5: Variable importance with rankings



To conclude, in Graph 5 the variable importance is depicted by ranking the variables according to their importance as the model showed, in terms of the Mean Decrease Accuracy. While we could clearly claim that random forest outperforms compared to logistic regression, its main disadvantage is that we are not able to compute the exact magnitude effect of each variable as in the logit model. Therefore, we see that it is a trade-off between accuracy and interpretability.

6. Managerial Implementation/ Educational Contribution

6.1 Managerial Implementation

From a managerial viewpoint, Instagram influencers consist the new powerful way of companies to promote through them their brands, the selection of those is quite vital. Nano-influencers, who are usually influencers promoting through various products/brands based on a specific philosophy and attitude that does represent them (e.g. vegan, luxury life, etc), tend to lead to higher engagement and conversion rates in social media. Thus, companies for better segmentation strategies, can rely on this type on influencers according the nature of their products, having as a result a fast and promising growing customer network, from their followers.

Additionally, influencer marketing has known quite some growth for luxurious expensive brands as the people who are promoting them are usually exposing their 'expensive' lifestyle, triggering while impressing their followers. Influencers seem to play a role model for their followers while the devotion that they gain from them springs from the inspiration that they create. According the literature review, users whose confidence is quite low, tend to be attractive by purchases of expensive brands as a way to satisfy themselves. Nowadays the fact that most successful marketing campaigns incorporate actions together with influencers, should make luxury firms allocate even a larger part of their budget to them. This means that influencers can encourage their followers to engage with the content of their post by posting themselves and tagging the corresponding brand (e.g. as a chance to win the product from a competition). Users who are more likely into binge consumption are expected to react quite positively.

Firms from the other hand, surely gain more audience and fellow customers as these actions still trigger them to purchase the product of the corresponding brand. According to Kim and Ko (2012), three out of the five aspects that marketing's actions in social media are based on, are the word of mouth, amusement and interaction between the brand and the user. All these attempts have been created to trigger the user's interest to ''consume'' the content via the respective social media platform.

Some more benefits from this segmentation strategy from the companies' perspective, is except for an increase in the sales revenue, a direct rise of organic traffic search, which automatically makes the company less dependent on Google. Finally, the companies and marketers, achieve a higher customer lifetime value and traffic-to-lead ratio which basically represent the new contact rate.

6.2 Educational Contribution

From an educational lookout, this thesis for the first time is connecting the tags of different brands that Instagram users used in their posts with a potential binge consumption towards social media through the posting frequency. So far, assumptions for the binge consumption of social media users have taken place through questionnaires or surveys that have been conducted. The attempt of this thesis was to gain some insights from the power of data by analysing the hashtags of the posts that the users provided in their accounts as well as extract as many information as possible. The price level of the brands is assumed to depict the user’s/consumers preferences as the fact that people who tagged different brands on their Instagram posts, showed somehow the price that are potentially willing to pay in order to obtain a certain brand. Finally, users that have an indication towards binge consumption are more likely to make less rational choices, meaning that their purchase volume will be higher.

7. Limitations/ Further Research

As every research that takes place within this, various limitations and perspectives for further future research occur and need to be mentioned. Theories within the Literature Review section do not determine or identify an addiction through a certain number of posts. For that reason, we cannot claim that a user who posts more than a specific number is addicted. However, we assumed that selecting users who posted more than 150 posts (as a benchmark) are more likely to get involved in a binge consumption of social media.

First, one of the limitations is the number of users and their posts which is quite limited as the scraping process was computational expensive. However, it is expected that increasing the number of observations of the dataset by adding more users, will improve the performance and the insights that could be derived. Additionally, since we did not define the panel type of data, logistic regression and random forest respectively treated each row as a separate observation. Thus, other methods would be also more appropriate taking into consideration the nature of the data. Time series could probably reveal also some insights of the data and on the binge consumption of the user as well as how they do differ from each other through the time. Also, we need to note that our data was quite unbalanced meaning that one user could have posted 3 times in a day while another one 40.

In case more variables were available (e.g. the comments of the posts as a text) various text analysis algorithms could be applied to get more meaningful insights. Firstly, we could do a sentiment analysis to see if the posts of each user correspond to positive or negative comments and how could these potentially influence the price level of the tagged brands. Another future research would be the topic modelling from the comments that would be available. That could let us know about which topics various Instagram users are mainly discussing through their comments. For these two-text analytics method approaches, a Greek dictionary would be necessary for the pre-processing part as the users being used are followers of Greek influencers, to be able to get the sentiment out of the comments. However, no scientific research has been done for the creation of such a dictionary. The method that we used in this thesis would be also enhanced by adding these features (topic modelling and sentiments) and potentially giving us higher predicting accuracy. Furthermore, the variables description (of the user's posts), if not so many missing values were not occurring, a text analysis could also have been done.

As far as the interpretation is concerned, a visualization method called LIME (Local Interpretable Model-agnostic Explanations) could be applied as an adding step of further research. LIME method can overcome the problem of interpretability that a lot of powerful machine learning methods face (this is the reason why they are often called black box models). This method can provide the impact (positive and negative) of the predictor variables regarding the dependent variable in a local scale (as it takes a certain number of observations, permuting them. Variable importance that was used in this thesis, captures in a global scale the importance of the predictor variables. Finally, a survey could be additionally conducted and displayed to the users that were analysed. In this way, more variables would be added in the dataset (e.g. gender, age, education, religion etc) increasing the credibility of the research simultaneously.

8. Conclusion

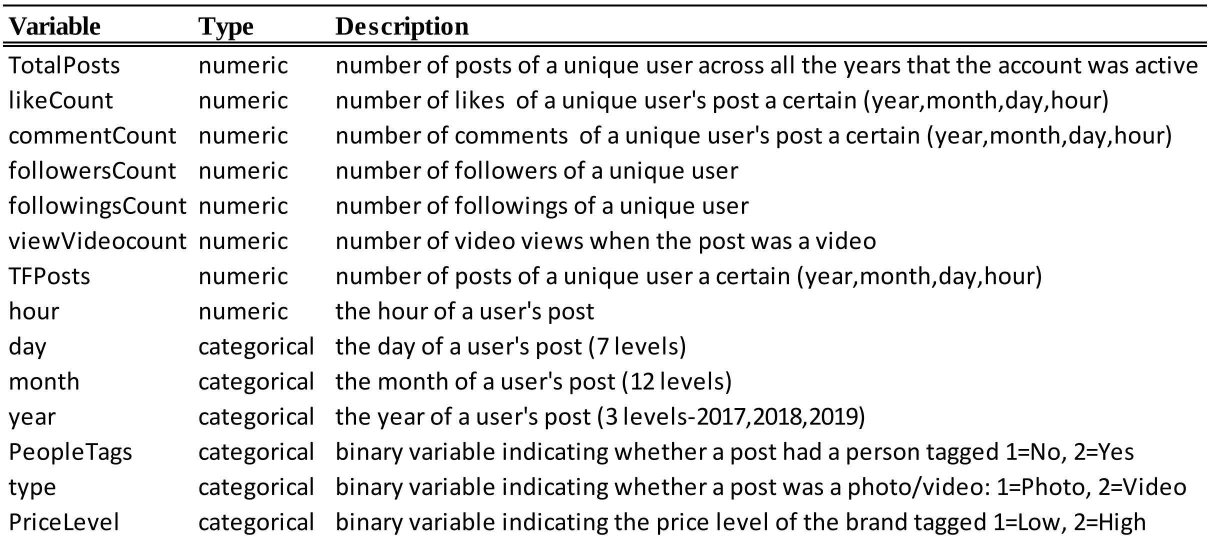
This master thesis managed to get some important insights regarding how the binge consumption towards social media is associated with a user's preferences in expensive brands (identified though the tags of his posts). The analysis that took place was by a random sample of Instagram users, selected based on some selection criteria. The aim of this thesis was to identify which factors tend to contribute a user posting an expensive versus a regular product. The model depicted as key contributors the number of posts, the number of likes that the user received in his previous post, certain months and days. Our hypothesis regarding the connection of the number of likes and the posts of expensive products was correct as we identified that a user with a low self-esteem, ''seeks the acceptance'' behind the number of likes he gets through his posts and even by tagging expensive luxury brands.

Interestingly, both models did not consider the number of comments that a previous post took or if a person was tagged in a post important for the dependent variable. Thus, we partially rejected the first hypothesis as the lagged value of comments seemed to not impact a user tagging an expensive vs a regular brand. The binge consumption of the users can be supported from various scientific researches which have brought into light that a user with a relatively low self-esteem has the need to get more likes through his post, consisting a boosting action in his self-esteem. This thesis, seeking from a different glance the existence of binge consumption of Instagram, was focused on the posting frequency of the users as one of the most important variables. Users by adopting a repeatedly posting tendency, they do spend more and more time in Instagram which can be an indication of binge consumption.

Finally, as it was also described in the Limitations and Future Research section, there are a lot of opportunities of new findings through an extension of the data collection. The power of the Big data combined with various powerful machine learning methods can bring numerous 'treasures' in a lot of sciences and food for thought for the researchers.

9. Appendix

Table 3



Variable	Type	Description
TotalPosts	numeric	number of posts of a unique user across all the years that the account was active
likeCount	numeric	number of likes of a unique user's post a certain (year,month,day,hour)
commentCount	numeric	number of comments of a unique user's post a certain (year,month,day,hour)
followersCount	numeric	number of followers of a unique user
followingsCount	numeric	number of followings of a unique user
viewVideocount	numeric	number of video views when the post was a video
TFPosts	numeric	number of posts of a unique user a certain (year,month,day,hour)
hour	numeric	the hour of a user's post
day	categorical	the day of a user's post (7 levels)
month	categorical	the month of a user's post (12 levels)
year	categorical	the year of a user's post (3 levels-2017,2018,2019)
PeopleTags	categorical	binary variable indicating whether a post had a person tagged 1=No, 2=Yes
type	categorical	binary variable indicating whether a post was a photo/video: 1=Photo, 2=Video
PriceLevel	categorical	binary variable indicating the price level of the brand tagged 1=Low, 2=High

Table 5

Coefficients	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	7.602e-01	1.092e-01	6.958	3.45e-12	***
typeVideo	2.513e-01	9.383e-02	2.678	0.00740	**
viewVideocount	-1.298e-04	1.765e-04	-0.735	0.46219	
PeopleTagsYES	-1.055e-01	5.116e-02	-2.063	0.03912	*
followersCount	-7.689e-05	6.061e-06	-12.686	< 2e-16	***
followingCount	-2.825e-05	1.111e-05	-2.543	0.01099	*
TotalPosts	-1.756e-03	4.971e-05	-35.336	< 2e-16	***
year2018	-2.463e-01	4.953e-02	-4.973	6.60e-07	***
year2019	-1.359e-01	4.890e-02	-2.780	0.00543	**
monthAug	8.600e-02	8.969e-02	0.959	0.33767	
monthDec	1.457e-01	9.269e-02	1.571	0.11607	
monthFeb	2.579e-01	1.035e-01	2.492	0.01270	*
monthJan	2.547e-01	1.001e-01	2.545	0.01093	*
monthJul	1.408e-01	9.356e-02	1.505	0.13241	
monthJun	2.120e-01	9.657e-02	2.196	0.02810	*
monthMar	1.460e-01	9.918e-02	1.473	0.14088	
monthMay	2.867e-01	9.769e-02	2.935	0.00334	**
monthNov	2.314e-01	9.488e-02	2.439	0.01472	*
monthOct	2.459e-01	9.322e-02	2.637	0.00835	**
monthSep	-1.102e-01	9.438e-02	-1.168	0.24292	
dayMonday	2.079e-01	7.062e-02	2.943	0.00325	**
daySaturday	7.518e-02	6.553e-02	1.147	0.25127	
daySunday	1.818e-01	7.070e-02	2.571	0.01013	*
dayThursday	1.917e-01	7.222e-02	2.654	0.00795	**
dayTuesday	1.282e-01	7.179e-02	1.786	0.07405	.
dayWednesday	8.772e-02	7.252e-02	1.210	0.22644	
hour	1.190e-02	3.740e-03	3.181	0.00147	**
TFPosts	-3.518e-02	8.522e-03	-4.129	3.65e-05	***
likes_lag	6.539e-04	1.450e-04	4.509	6.52e-06	***
comments_lag	-3.446e-03	2.309e-03	-1.493	0.13553	

10.References

Andreassen, C. S. (2015). Online social network site addiction: A comprehensive review. *Current Addiction Reports*, *2*(2), 175-184.

Balakrishnan, J., & Griffiths, M. D. (2018). An exploratory study of “selfitis” and the development of the Selfitis Behavior Scale. *International journal of mental health and addiction*, *16*(3), 722-736.

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: desire for interpersonal attachments as a fundamental human motivation. *Psychological bulletin*, *117*(3), 497.

Becker, G. S., & Murphy, K. M. (1999). A theory of rational addiction. *Journal of Political Economy*.

Becker, G. S. (1965). A Theory of the Allocation of Time. *The economic journal*, 493-517.

Becker, G. S., Grossman, M., & Murphy, K. M. (1991). Rational addiction and the effect of price on consumption. *The American economic review*, *81*(2), 237-241.

Boehmke, B., & Greenwell, B. M. (2019). *Hands-On Machine Learning with R*. CRC Press.

Brand, M., Young, K. S., Laier, C., Wölfling, K., & Potenza, M. N. (2016). Integrating psychological and neurobiological considerations regarding the development and maintenance of specific Internet-use disorders: An Interaction of Person-Affect-Cognition-Execution (I-PACE) model. *Neuroscience & Biobehavioral Reviews*, *71*, 252-266.

Can, L., & Kaya, N. (2016). Social networking sites addiction and the effect of attitude towards social network advertising. *Procedia-Social and Behavioral Sciences*, *235*, 484-492.

De Vries, L., Gensler, S., & Leeflang, P. S. (2012). Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing. *Journal of interactive marketing*, *26*(2), 83-91.

Chaudhari, B. L., Patil, J. K., Kadiani, A., Chaudhury, S., & Saldanha, D. (2019). Correlation of motivations for selfie-posting behavior with personality traits. *Industrial Psychiatry Journal*, *28*(1), 123.

Chou, T. J., & Ting, C. C. (2003). The role of flow experience in cyber-game addiction. *CyberPsychology & Behavior*, *6*(6), 663-675.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, *43*(3), 345-354.

Cramer, E. M., Song, H., & Drent, A. M. (2016). Social comparison on Facebook: Motivation, affective consequences, self-esteem, and Facebook fatigue. *Computers in Human Behavior*, *64*, 739-746.

Dhaha, I. S. Y. (2013). Predictors of Facebook addiction among youth: A structural equation modeling (SEM). *Journal of Social Sciences (COES&RJ-JSS)*, *2*(4), 186-195.

Dittmar, H., Beattie, J., & Friese, S. (1995). Gender identity and material symbols: Objects and decision considerations in impulse purchases. *Journal of economic psychology*, *16*(3), 491-511.

Durak, H. Y. (2018). Modeling of variables related to problematic internet usage and problematic social media usage in adolescents. *Current Psychology*, 1-13.

Ferguson, B. S. (2000). Interpreting the rational addiction model. *Health Economics*, *9*(7), 587-598.

Gallagher, K. (2018), “The influencer marketing report: the best practices and platforms brands use when tapping popular social media personalities for marketing”, www.businessinsider.com/the-influencer-marketing-report-2018-1/

Goldstein, B. A., Polley, E. C., & Briggs, F. B. (2011). Random forests for genetic association studies. *Statistical applications in genetics and molecular biology*, *10*(1).

Goldsmith, R. E., & Newell, S. J. (1997). Innovativeness and price sensitivity: managerial, theoretical and methodological issues. *Journal of Product & Brand Management*.

Goldsmith, R. E., Kim, D., Flynn, L. R., & Kim, W. M. (2005). Price sensitivity and innovativeness for fashion among Korean consumers. *The Journal of social psychology*, *145*(5), 501-508.

Gold, M. S., Frost-Pineda, K., & Jacobs, W. S. (2003). Overeating, binge eating, and eating disorders as addictions. *Psychiatric Annals*, *33*(2), 117-122.

Grewal, L., Stephen, A. T., & Coleman, N. V. (2019). When posting about products on social media backfires: the negative effects of consumer identity signaling on product interest. *Journal of Marketing Research*, *56*(2), 197-210.

Gruber, J., & Köszegi, B. (2001). Is addiction “rational”? Theory and evidence. *The Quarterly Journal of Economics*, *116*(4), 1261-1303.

Han, S., Gupta, S., & Lehmann, D. R. (2001). Consumer price sensitivity and price thresholds. *Journal of Retailing*, *77*(4), 435-456.

Handley, L. (2018), “How luxury fashion brands are increasingly using Instagram to win over millennials”, CNBC, 9 January, available at: www.cnbc.com/2018/01/09/how-luxury-fashion-brands-are-using-instagram-to-win-over-millennials.html

Irani, N., & Hanzaee, K. H. (2011). The effects of variety-seeking buying tendency and price sensitivity on utilitarian and hedonic value in apparel shopping satisfaction. *International Journal of Marketing Studies*, *3*(3), 89.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, pp. 3-7). New York: springer.

Järvinen, J., Tollinen, A., Karjaluoto, H., & Jayawardhena, C. (2012). Digital and social media marketing usage in B2B industrial section. *Marketing Management Journal*, *22*(2).

Jensen Schau, H., & Gilly, M. C. (2003). We are what we post? Self-presentation in personal web space. *Journal of consumer research*, *30*(3), 385-404.

Kacen, J. J., Hess, J. D., & Walker, D. (2012). Spontaneous selection: The influence of product and retailing factors on consumer impulse purchases. *Journal of Retailing and Consumer Services*, *19*(6), 578-588.

Karim, N. S. A., Zamzuri, N. H. A., & Nor, Y. M. (2009). Exploring the relationship between Internet ethics in university students and the big five model of personality. *Computers & Education*, *53*(1), 86-93.

Kapoor, K. K., Tamilmani, K., Rana, N. P., Patil, P., Dwivedi, Y. K., & Nerur, S. (2018). Advances in social media research: Past, present and future. *Information Systems Frontiers*, *20*(3), 531-558.

Karadağ, E., Tosuntaş, Ş. B., Erzen, E., Duru, P., Bostan, N., Şahin, B. M., ... & Babadağ, B. (2015). Determinants of phubbing, which is the sum of many virtual addictions: A structural equation model. *Journal of behavioral addictions*, *4*(2), 60-74.

Khan, N., Hui, L. H., Chen, T. B., & Hoe, H. Y. (2016). Impulse buying behaviour of generation Y in fashion retail. *International Journal of Business and Management*, *11*(1), 144.

Kircaburun, K., & Griffiths, M. D. (2018). Instagram addiction and the Big Five of personality: The mediating role of self-liking. *Journal of behavioral addictions*, *7*(1), 158-170.

Kim, A. J., & Ko, E. (2012). Do social media marketing activities enhance customer equity? An empirical study of luxury fashion brand. *Journal of Business research*, *65*(10), 1480-1486.

Kusumasondjaja, S. (2019). Exploring the role of visual aesthetics and presentation modality in luxury fashion brand communication on Instagram. Journal of Fashion Marketing and Management: An International Journal.

Kwon, H. E., So, H., Han, S. P., & Oh, W. (2016). Excessive dependence on mobile social apps: A rational addiction perspective. *Information Systems Research*, *27*(4), 919-939.

Langlois, G., & Elmer, G. (2013). The research politics of social media platforms. *Culture machine*, *14*.

Leeflang, P. S., Verhoef, P. C., Dahlström, P., & Freundt, T. (2014). Challenges and solutions for marketing in a digital era. *European management journal*, *32*(1), 1-12.

Lister, M. (2017). Essential Social Media Marketing Statistics For 2017," Wordstream.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of personality*, *60*(2), 175-215.

Moro, S., Rita, P., & Vala, B. (2016). Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. *Journal of Business Research*, *69*(9), 3341-3351.

Moe, W. W., & Trusov, M. (2011). The value of social dynamics in online product ratings forums. *Journal of Marketing Research*, *48*(3), 444-456.

Muratore, I. (2016). Teens as impulsive buyers: what is the role of price?. *International Journal of Retail & Distribution Management*.

Ng, B. D., & Wiemer-Hastins, P. (2005). Addiction to the internet and online gaming. *Cyberpsychology & behavior*, *8*(2), 110-113.

Orphanides, A., & Zervos, D. (1995). Rational addiction with learning and regret. *Journal of Political Economy*, *103*(4), 739-758.

Pelling, E. L., & White, K. M. (2009). The theory of planned behavior applied to young people's use of social networking web sites. *CyberPsychology & Behavior*, *12*(6), 755-759.

Ponnusamy, S., Iranmanesh, M., Foroughi, B., & Hyun, S. S. (2020). Drivers and outcomes of Instagram Addiction: Psychological well-being as moderator. *Computers in Human Behavior*, *107*, 106294.

Raymer, K. (2015). The effects of social media sites on self-esteem.

Seidman, G. (2013). Self-presentation and belonging on Facebook: How personality influences social media use and motivations. *Personality and individual differences*, *54*(3), 402-407.

Schweidel, D. A., & Moe, W. W. (2016). Binge watching and advertising. *Journal of Marketing*, *80*(5), 1-19.

Shachar, R. (2003). Party loyalty as habit formation. *Journal of Applied Econometrics*, *18*(3), 251-269.

Stephen, A. T., & Galak, J. (2012). The effects of traditional and social earned media on sales: A study of a microlending marketplace. *Journal of marketing research*, *49*(5), 624-639.

Steinfield, C., Ellison, N. B., & Lampe, C. (2008). Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *Journal of applied developmental psychology*, *29*(6), 434-445.

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics–Challenges in topic discovery, data collection, and data preparation. *International journal of information management*, *39*, 156-168.

Thakre, K. S., Dawande, D., & Thakre, V. S. (2020, January). Identifying Social Network Delusion to Investigate Addiction Ratio using Data Mining. In *Proceedings of the 2020 2nd Asia Pacific Information Technology Conference* (pp. 1-8).

Thompson, C. J., & Hirschman, E. C. (1995). Understanding the socialized body: A poststructuralist analysis of consumers' self-conceptions, body images, and self-care practices. *Journal of consumer research*, *22*(2), 139-153.

Turel, O. (2015). An empirical examination of the "vicious cycle" of Facebook addiction. *Journal of Computer Information Systems*, *55*(3), 83-91.

Van Der Aalst, W. (2016). Data science in action. In *Process mining* (pp. 3-23). Springer, Berlin, Heidelberg.

Virvilaitė, R., Saladienė, V., & Bagdonaitė, R. (2009). Peculiarities of impulsive purchasing in the market of consumer goods. *Inžinerinė ekonomika*, (2), 101-108.

Vohs, K. D., & Faber, R. J. (2007). Spent resources: Self-regulatory resource availability affects impulse buying. *Journal of consumer research*, *33*(4), 537-547.

Wang, C. W., Ho, R. T., Chan, C. L., & Tse, S. (2015). Exploring personality characteristics of Chinese adolescents with internet-related addictive behaviors: Trait differences for gaming addiction and social networking addiction. *Addictive behaviors*, *42*, 32-35.

Wilson, K., Fornasier, S., & White, K. M. (2010). Psychological predictors of young adults' use of social networking sites. *Cyberpsychology, behavior, and social networking*, *13*(2), 173-177.

Wilcox, K., & Stephen, A. T. (2013). Are close friends the enemy? Online social networks, self-esteem, and self-control. *Journal of Consumer research*, *40*(1), 90-103.

Williamson, D. A. (2011). Worldwide social network ad spending: 2011 outlook. *Retrieved February*, *10*, 2011.

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, *7*(6), 464-476.

Zhang, Y., Trusov, M., Stephen, A. T., & Jamal, Z. (2017). Online shopping and social media: friends or foes?. *Journal of Marketing*, *81*(6), 24-41.