Spring 5-29-2021

Modelling User Behaviour in Market Attribution: finding novel data features using machine learning

Tanisha Naomi Thornton
Cardiff School of Technologies, Cardiff Metropolitan University, Wales, United Kingdom,
st20111334@outlook.cardiffmet.ac.uk

Simon Thorne
Cardiff School of Technologies, Cardiff Metropolitan University, Wales, United Kingdom,
SThorne@cardiffmet.ac.uk

Ana Calderon
Cardiff School of Technologies, Cardiff Metropolitan University, Wales, United Kingdom,
acalderon@cardiffmet.ac.uk

Modelling User Behaviour in Market Attribution:
finding novel data features using machine learning

Tanisha N. Thornton
St20111444@outlook.cardiffmet.ac.uk

Dr Simon Thorne
SThorne@cardiffmet.ac.uk

Dr Ana Calderon
acalderon@cardiffmet.ac.uk

Cardiff School of Technologies, Cardiff Metropolitan University, Wales, United Kingdom

Abstract –

This paper presents an exploration of market attribution methods and the integration of user behaviour. Attribution is the measurement of interaction between marketing touchpoints and channels along the customer journey, improving customer insights and driving smarter business decisions. Improving the accuracy of attribution requires a deeper understanding of user behaviour, not just marketing channel credit assignment. Evidence has been provided regarding the problems in the standardized approach to behavioural modelling and alternatives have been presented. The study explores data provided by a British based jewellery company with an investigation into pre-existing data features that can aid with the analysis of user behaviour. The study contains over 10 million rows collected over 2 years and presents the initial findings made in the first 15 months of a PhD study.

Keywords: Marketing, Attribution, Modelling, Behaviour, Psychographics, E-commerce, Segmentation

1.0 Introduction

In recent years the popularity of online e-commerce has grown significantly in the United Kingdom. This offers retailers the opportunity to engage with larger audiences, encouraging visitors to purchase online through the use of website-only deals and attractive shipping costs. The rise in retailers offering online shopping shifts the focus onto the use of marketing channels and the assignment of credit to these channels using an attribution model that should complement the goals of the retailer. These models range in simplicity from single-touch to multi-touch attribution and include channels such as Search Engine Optimisation (SEO), social media marketing, direct marketing and Pay-Per-Click (PPC). The pre-existing, popular attribution models fail to take one key factor into consideration, user behaviour. More in-depth

research should be conducted into user behaviour and the factors that drive a visitor to conversion and through this understanding of behaviour, businesses do not just have a 'where', but a 'why'. This study makes a comprehensible effort to improve the understanding of user behaviour through the analysis of features within the e-commerce data provided.

The presentation of this study is as follows; section 2 presents the background of the study. Section 3 outlines the definitions of digital marketing, market attribution, user behaviour, behavioural modelling and psychometrics with discussions on each. Section 4 presents the methodologies to be employed for the duration of the study and the hypotheses we can form based on the research. Section 5 contains explanations into the construction of a useable dataset/s. Section 6 discusses the approaches to the experimentation process and machine learning algorithms. Section 7 contains information encompassing the optimisation of the dataset/s and the recursive processes involved. Section 8 concludes the study.

The implementation of standardised market attribution models has become a pervasive issue in this technological age; the ease of selecting a static, unbending model that will assign credit to marketing channels has become the norm. As convenient as this solution may be, it pays little to no regard for real user behaviour, intention and exploration, providing few insights to the thought processes conducted by an individual users' path-to-purchase. Current attribution models treat every user the same despite every users' journey being complex and individual, therefore it should be treated as such. This paper outlines the future processes that will be employed and explored to provide a solution to this issue.

2.0 Digital Marketing

The concept of digital marketing first emerged in the 1990s and is a term is used to describe the commercial activities of the buying and selling of goods and/or services in a digital landscape. It has evolved significantly over the past three decades, from the first e-commerce transaction in 1994 to the development of tech giant Google, the birth of social media sites such as LinkedIn and Facebook and the opportunities for businesses to reach larger and more diverse audiences expanded. This technological growth paved the way for the development of marketing channels, the innovation of customer relationship management (CRM), direct correspondence via email, the development of search engine optimisation (SEO) and pay-per-click advertising (PPC).

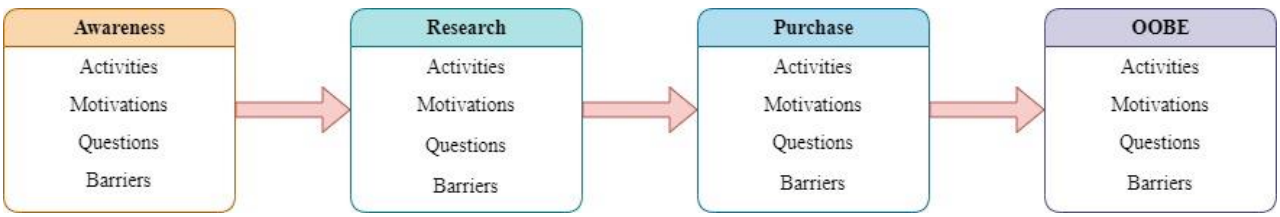
An abundance of research has been conducted in the field of digital marketing and marketing channels. Libai, Naravandas & Humby (2002), Dost *et al*., (2014) and Ariker *et al*., (2015) discuss the importance of marketing for profitability and tailored marketing actions.

These include personalised correspondence and the distribution of rewards or discounts to the right customers, a form of customer segmentation (those who are susceptible to discounts and potentially on the edge of buying and those who aren't). This could also be referred to as the development of an effective marketing strategy. Wang & Wang (2006) confirm the effectiveness of customer segmentation for the development of effecting marketing strategies adding that behavioural patterns can help with the identification of potential segments. Giddings (2011) coined a study encompassing the measurement of customer engagement across marketing channels and enforces the importance of customer segmentation to understand the needs and behaviours presented by visitors and that customization of marketing efforts are more effective.

2.1 Market attribution

The development of marketing channels led to market attribution (the evaluations of customer touchpoints), forming the backbone of attribution models and the optimisation of the marketing process. Touchpoints are the points of contact a visitor has with a business; including social media advertisement, digital marketing and peer referral among others. These are placed by businesses at each stage of the customer journey and can be categorised as digital or physical. Digital touchpoints include online advertisements; these can be via a search engine or on third-party websites (PPC), online correspondence such as email and newsletters (CRM) and website optimisation, which includes the positioning of a result on a search engine (SEO). Physical touchpoints are more direct in their implementation and include direct mail, marketing calls, in-store promotions and word-of-mouth interactions.

These touchpoints are located at various junctures in the customer journey and can be used to map interactions. There are many interpretations of customer journey maps and these will vary from business to business. Richardson (2010) outlines the customer journey process as 4 categories; Awareness, Research, Purchase and OOBE (out-of-box-experience); each of these is reduced further, including the subcategories Activities, Motivations, Questions and Barriers. These are utilized to understand user mentality are they traverse through the stages of the customer journey map. Acknowledgements are also made to the journey not always following a linear process; some users spending longer in the awareness and research phases



for expensive or emotionally significant purchases. In alternative industries, users may go straight from awareness to purchase depending on recommendations they have received from friends or family. The creation of customer journey maps helps to outline the general purchasing processes, enabling the optimisation of customer touchpoints.

Figure 1. A diagram of the generalised customer journey processes discussed by Richardson (2010), including the 4 categories and 4 subcategories included within each process (Source: Personal collection)

Attribution models come in all different shapes and sizes, but the main two categories of models are referred to as single-touch and multi-touch. The key differences between the two approaches are the distribution of credit to marketing channels. Single-touch attribution aims to distribute credit to a single touchpoint (this commonly either the first-touch or last-touch) Multi-touch attribution distributes credit across several channels but is still dependent on the chosen model. For example, time decay (distributing credit based on the assumption that the further back it was interacted with, the less important it was) and linear attribution (the distribution of credit spread evenly across all channels).

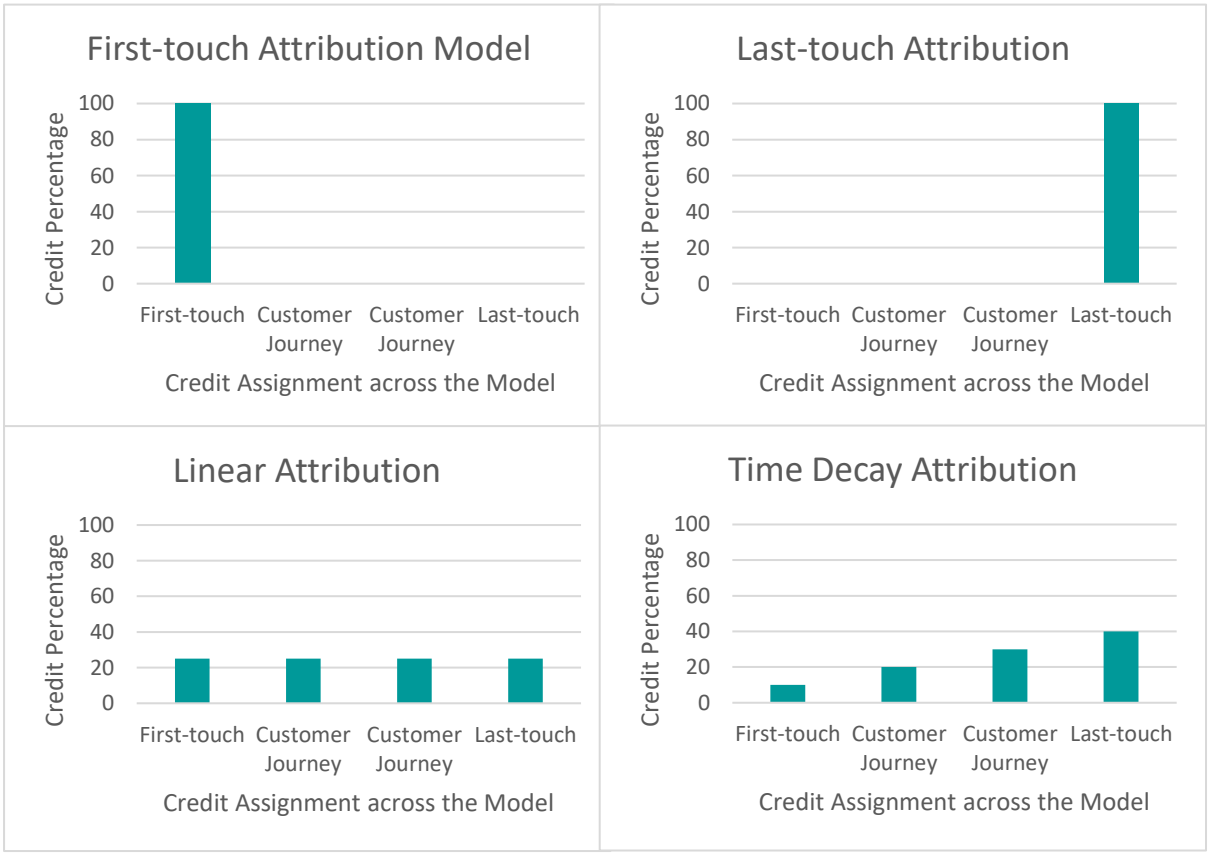


Figure 2. A visual representation of the credit distributions between single-touch attribution (first and last touch) and multi-touch attribution models (linear and time decay attribution) (Source: Personal collection)

Shao & Li (2011) state that the goal of attribution is to pin-point the credit assignment of each positive user to one or more touchpoint and suggest that attribution modelling should be easy to interpret and be used to derive insights for businesses to optimise their marketing strategies. They also stress the criticality of choosing the right attribution model as it drives the performance metric, produces insights regarding advertising and helps to optimise marketing strategies. Anderl *et al*., (2016) agree that the insights derived from attribution are valuable, adding that some insights can be generalized across businesses and some are company-specific. These insights can help to analyse the effectiveness of a marketing channel and shed light on the interplay of channels for more accurate attribution. Ren *et al*., (2018) continue with this concept adding that aggregated attribution across marketing channels can guide advertisers for the allocation of budgets.

Despite all the effective research conducted in the field of attribution modelling, the pre-built models may not provide a solution for every business, the generalisation of credit assignment may not provide the most accurate representation of the customer base and therefore accurate marketing strategies cannot be formed. A solution to this problem would be for businesses to build custom multi-touch attribution models, tailored to their personal goals, based on real customers and real behaviour, enabling them to create more accurate and effective marketing strategies. As suggested by Sato *et al*., (2013), continuous monitoring of the model should be conducted, this will provide the business with valuable data regarding the customer base and the model can be adapted accordingly.

2.2 User Behaviour

User behaviour is the term used to describe how users of an online service interact with a website or product. To analyse the behaviour of users on a website, user metrics must be built into the database that can measure behaviour, i.e. clicks, the customer journey, events and conversion tracking. The chain of behaviours displayed by a user is referred to as a purchasing sequence, enabling the business to identify patterns within the data that can be used to improve the attribution model, credit analysis and marketing touchpoints.

Clicks or clickstream data can be used to track page visits, determine the length of the journey a customer has made, how long they spent looking at a particular page and what page they visited next. This can be useful in helping businesses understand the impact of certain webpages on users and how they choose to interact with them can lead to improved marketing strategies and enhance the user experience.

User journey data provides businesses with the ability to measure the intent of their users and determine where they are in the customer journey (see figure 1), aiding businesses to adapt the way they interact with customers to improve conversion rates, perceived trust, customer satisfaction or a combination of these. A feature such as 'visitor_journey' can also help identify a particular product a user has viewed. This information can be used to determine whether they could be persuaded to convert through the use of tailored advertising or offering a reminder or discount for the item via direct correspondence.

Event data is produced when certain events are fired, these are often small goals, set by the business and can be as simple as a customer signing up for a newsletter. Actions such as favouriting an item, adding an item to a basket or taking an item out are all examples of fired events. Similar to user journey data, this metric can help businesses measure the intent of each customer and make predictions based on their behaviour, aiding the predictive process regarding the likelihood to convert amongst other goals. Rendle *et al*., (2010) add that sequential patterns within the user browsing behaviour are of great value to predictive analytics and decision-making processes.

Hoffman & Novak (1997) suggest that navigational behaviour (such as journey data) can be divided into two main categories: experimental behaviour which is often unstructured with either no goal or a frequently changing goal, and goal-oriented behaviour by which the customer will know exactly what they are looking to buy or looking to achieve. Moe & Fader (2002) add that the behaviours presented by new customers will differ from that of returning customers, this should be taken into consideration as user goals my not have been defined straight away but goals may form over time. Ven den Poel *et al*., (2004) go on to suggest that the monitoring of user behaviour can be used to adjust marketing techniques, influence shopping behaviours and stress the importance of clickstream data and the improvement it poses for predictive performance.

More recently, Othman *et al*., (2013) researched the perceived value and satisfaction of online purchasing. They discovered that the level of trust a user has in a business positively affects the relationship between customer loyalty and commitment; positively affecting user purchasing behaviour. O'Flaherty & Heavin (2015) link the prediction of user behaviours to the formation of proactive data-driven decisions and the improvement of predictive analytics, which in turn, improves the process of decision making. Andreeva, Ansell & Crook (2017) agree and add that the monitoring of purchase and behavioural characteristics over time becomes more important as it increases in accuracy and provides more predictive power.

Research conducted by Anderlová & Pšurný (2020), explored the emotional connections formed by users within the Czech luxury cosmetic market and proposed a model consisting of 2 predominant segments. The first segment consisted of those who responded well to emotive language and have high regard for social status. The second were those who were described as "novelty lovers", not emotionally influenced and not attached to the brand. This could hold true for other businesses labelling themselves as being ‘luxury’ including the automotive industry, homewares, jewellery and clothing.

Considering the research presented, we must decide whether all of these functions can truly be achieved by applying a static attribution model such as last-touch or linear to data. These models leave very little room for improvement, customer goal-tracking, behavioural analysis and service customisation; in addition to this, they treat every user as the same even though the intentions, personality and behaviours will differ from user to user.

2.3 Behavioural Modelling

Behavioural modelling is the analysis of past behaviour such as consumer and business spending data to predict future actions. Modelling the behaviour of customers helps to establish their profitability, in turn benefitting marketing processes for that business (Schroder & Hruschka, 2017). Behavioural modelling is used by sectors such as finance and e-commerce to analyse customer risk, track spending behaviours, analyse purchase intentions and implement loyalty programs. Standard economic theory (the sole seeking of profit) and economic assumptions (assumptions made about the general marketing environment) can be used as the starting point for the modelling of human behaviour; choices such as perceived cost and user benefits are suited to analysis based around economic theory.

Numerous models have been proposed by researchers for the analysis of user behaviour, including Behavioural Targeting Advertising (BTA) presented by Sato *et al.*, (2013); a marketing method used to attract users through targeted ads based on previous browsing behaviours. The behavioural data can then be analysed and used to classify users into specific segments depending on their actions; this can be used to further monitor the behaviours and attitudes of users within each segment. Andreeva, Ansell & Crook (2017) employed the use of the AUROC model, Area Under the Curve (AUC) and Receiver Operating Characteristics (ROC); a performance method used for classification problems. This is most commonly plotted on a graph and helps to determine the accuracy of the model when distinguishing between classes (see Figure 3). For this particular study, it was used to model behavioural variables relating to the use of a store loyalty card, discovering that over time, behavioural information

increased the predictive power of the model. The best results were achieved by incorporating all of the behavioural information, reinforcing the view that over time, behavioural data becomes more important in determining the actions of individual users and establishing individual customer profitability.

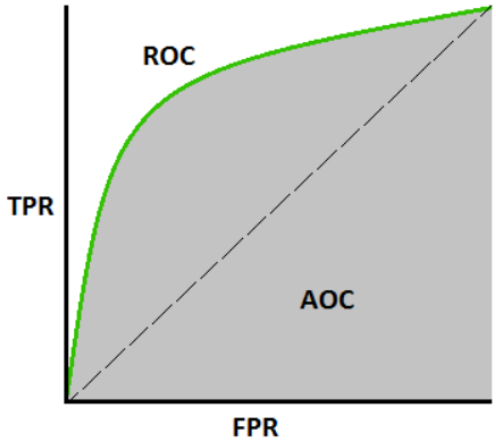
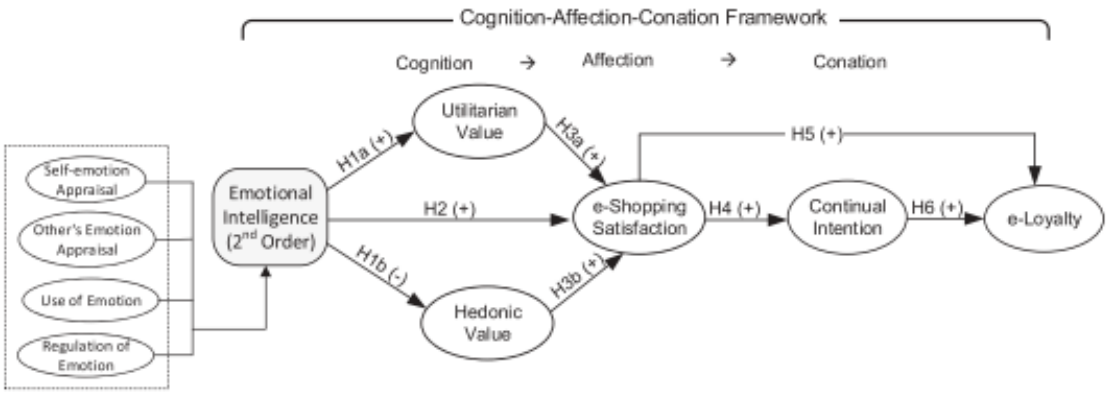


Figure 3. Graph displaying the AUROC model (Narkhede, S, 2018)

Ha & Lee (2015) propose a behavioural model based on the affordances and emotional responses of users, producing a task tree involving the processes of users, the significances and the affordance probabilities based on the behaviours displayed. They produced the model based on four aspects; design suitability of a product from a users' point of view, the prediction of the order of behaviour sequences, prediction of emotional losses/changes caused by usage failures and the optimal behaviour sequence predictions based on the emotional response. Similarly, Lim & Kim (2020) formed a study encompassing the emotional intelligence of users' and the effects this had on their shopping behaviour, providing a cognitive-affective-conative framework (see Figure 4). The results confirm the hypotheses that Emotional Intelligence (EI)



could influence the value perceived by online customers and shopping behaviours. EI made a positive impact on the value perception of every day/utilitarian items but provided no significant relationship between EI and luxury/hedonic purchases.

Figure 4. A general outline of a cognitive-affective-conative framework (Lim, S., Kim, D, 2020)

2.4 Psychographics

Psychographics is the study and classification of people, based on attitudes, habits, hobbies, aspirations, personality and other psychological criteria. These measures can be particularly useful during market research, model development and customer segmentation (Barnes *et al*., 2007; Oliver & Rosen, 2010). Psychographic profiles can be built by analysing the data left behind by users' online, including their engagement with social media, online purchasing behaviours and cross-device tracking. The information gathered can help to improve the predictive measures put in place by a business, optimise online services and target users more effectively using customised advertisements or website homepages.

In a study conducted by Mauri, Maira & Turci (2015), psychographics were used to differentiate the behaviours between private labels and national brand promotions; they found that both of these options were driven by different psychographics and that customers commonly weigh-up the options between cost vs. benefit. They discovered three main significant coefficients: Quality Consciousness, Price Consciousness and Innovativeness and all of these are related directly to utilitarian purchases.

France & Ghose (2019) suggest that psychographic tests that are adapted to marketing can be implemented to discover underlying psycho-social traits amongst users' but they have limited ability when it comes to the prediction of specific product purchases. They add that psychographic segmentation has limited usefulness regarding brand-specific behaviours, however, when it's combined with traditional segmentation such as demographic and geographic, it can prove to be a powerful and insightful tool.

Developed by John, Donahue & Kentle in 1991, one of the most comprehensive personality tools of its time was built. Referred to as the Big Five Inventory (BFI) or Five-Factor Model (FFM), it was the original measurement of personality, consisting of 44 items used to measure psychographic responses and personality traits based on 5 domains; neuroticism, extraversion, openness, agreeableness and conscientiousness. Several revisions of the original model have been made since including developments in the Big Five Inventories (BFI, BFI-44, BFI-10, BFI-2) and the development of the Neuroticism, Extraversion and Openness (NEO) Inventories (NEO-PI, NEO-FFI, NEO-FFI-3). Several improvements have

been made during this development process including the ease of readability, translations into alternative languages and added personality measures. Barnes *et al*., (2007) suggest that the integration of psychographics contribute to more insightful and accurate visitor information, helping to segment visitors, understand the values held by visitors and form improved marketing strategies.

The metrics (neuroticism, extroversion, openness, agreeableness and conscientiousness) are not currently available in the dataset. These will be derived from the data based on the behaviours presented by visitors and measured using machine learning, providing successful behavioural segmentation. This will aid the analysis of purchasing behaviour observation and the process of events that visitors perform leading them to conversion. A by-product of these applications is the improved accuracy of purchasing predictions, clarification of visitor attitudes, perceived value, purchase intent and deeper understanding of behavioural loyalty (Thatcher & George, 2004; Bailey *et al*., 2009; Sato *et al*., 2013; Pura, 2005; Curtis *et al*., 2017; Andreeva, Ansell & Crook, 2017). These findings can help to improve the segmentation methods employed by businesses, delivering smarter marketing correspondence and forming deeper connections with users.

3.0 Methodology and Hypotheses

Based on the findings presented in the numerous studies, we hypothesise that a few of the 5 domains (neuroticism, extraversion, openness, agreeableness and conscientiousness) will prove easier to measure using specific data features. Factors such as trust can be measured with ease using features such as revenue and device type, helping with the measurements relating to neuroticism, extraversion and openness. Agreeableness will be much harder to analyse as it is characterised as having a warm, compassionate and empathetic personality; a visitor with an agreeable personality, however, may be more susceptible to discounts, promotional emails and product offers. Similarly, conscientiousness may be harder to trace as it can characterise itself as focused and stubborn (high conscientiousness) or spontaneity and flexibility (low conscientiousness). It may be possible to attribute features such as event journey, basket items and visit data to the display of contentious behaviour.

Lissitsa & Kol (2019) found a positive correlation between mobile device shopping and extraversion, agreeableness and openness, those who were less likely to shop via mobile device displayed higher levels of neuroticism and conscientiousness. Therefore, we can hypothesize the following:

H1a: Visitors using mobile devices such as smartphones and tablets to purchase items will score higher in extraversion, agreeableness and openness.
H1b: Visitors producing higher rates of revenue using mobile devices will score even higher in these personality traits and significantly lower in neuroticism and conscientiousness.
H1c: Visitors using desktop PCs and laptops to purchase items will score lower in extraversion, agreeableness and openness and higher in neuroticism and conscientiousness.

Zhou & Lu (2011) found that all of the 5 personality measures; neuroticism, extroversion, openness, agreeableness and conscientiousness significantly affect trust, however, only neuroticism and agreeableness affect perceived usefulness. Therefore, we can hypothesize:

H2a: Visitors spending more money under the feature 'revenue' will have higher levels of trust and score higher in agreeableness and openness.
H2b: Visitors spending little to no money and continuously provide only browsing data will score lower in openness and agreeableness and higher in conscientiousness and neuroticism than those who generate more revenue.

Li *et al*., (2020) discuss path to purchase data and how these patterns change between hedonic purchases (for enjoyment and pleasure) and utilitarian purchases (for practical use or is a necessity). The results show that hedonic purchasers are more likely to browse product pages on the retailers' website and those displaying more utilitarian behaviours tend to perform efficient searches across alternatives; comparing prices, deals and reading reviews. Although this study has no direct link to the NEO-FFI personality model and the primary product type is hedonic based on its nature and luxury branding, we can hypothesize based on the study by Li *et al*., (2020):

H3a: Visitors with longer journeys ending with a purchase will display higher levels of extraversion, agreeableness and openness, with lower levels of conscientiousness and neuroticism.
H3b: Visitors with short journeys ending with a purchase will score significantly higher in extraversion, agreeableness and openness and significantly lower in conscientiousness and neuroticism.
H3c: Visitors with longer journeys not ending with a purchase will display higher levels of conscientiousness and neuroticism and lower levels of extraversion, agreeableness and openness.

Correlations have also been found regarding generational cohorts, online purchasing and device usage (Lissitsa & Kol, 2016; Lissitsa & Kol, 2019), finding that older generations are more inclined to use a desktop or laptop as they are less familiar with portable devices and concerned with the privacy of these. Donnellan & Lucas (2009) published research outlining the deviations between age groups (ranging from 16 to 85) and the scoring for each of the 5 personality measures. As the age groups progressed, the levels of extraversion, neuroticism and openness declined. Agreeableness was low with age group 16-19 and relatively level with those over the age of 20, this may be attributed to the research phase outlined by Richardson (2010). Conscientiousness rose significantly from ages 16-19 to 20-29 and began to fall from age 60; this pattern has been attributed to the development of conscientiousness throughout the transition from adolescence to adulthood (Donnellan, Conger & Burzette, 2007). Unfortunately, these are not hypotheses we can elude to as the data is fully anonymised and no personal information about individual visitors is available.

We have not formed a hypothesis for visitors with short journeys and no purchase as this includes single-page sessions and can therefore be interpreted as a website bounce.

4.0 Constructing a Dataset

The data to be explored for the duration of the study is provided by a UK based high-end luxury jewellery brand providing worldwide shipping to customers. The database is substantial in size, containing over 1 million records, across an average of 116 tables made up of 319 original data features. It is important to note that demographic and geographical information about individual visitors is present within the extended database however these metrics will not be used for this study to avoid and eliminate bias. All data is entirely anonymous and no identifiable personal information such as name, age, occupation or otherwise are available. The software used to initially explore the data and database structure was HeidiSQL and JupyterLab will be used to extract tables, further exploration, data cleaning, the formation of tables suitable for the measurement of user behaviour and model building.

4.1 Feature Selection

The process of feature selection began with a standard data cleaning practice; the elimination of any table containing no viewable data or tables that were not passed any data; the original table count totalled 133 however this process scaled the count down to 116. Following this process, a list of all features was created from all the tables containing data and all duplicates were removed. 319 features remained however not all of these features provided

enough data to be considered viable, therefore tables containing features with less than 500 rows of useable data were removed.

Following the initial cleaning process, a feature dictionary was created to provide a detailed description of each feature, helping to determine its usability and relevance regarding user behaviour. The understanding of each feature was paramount in the process of feature reduction; 319 was condensed down to 18 individual features, all containing enough data; some of the features were too similar and so these were reduced accordingly (see Figure 5). The 18 features were then tracked back to tables and where possible, tables containing multiple features were chosen, helping to reduce the number of tables from 116 down to 22. Following this, the tables were organised from the least to the most populated including size (MiB) and the number of rows and columns in each. Completing this process helped to reduce the number of tables from 22 down to 9 which contained all the desired features with as much information as possible to make the data pulling process simple and save on any unnecessary processing time and power.

4.2 Dataset Construction and Experimentation

The study will be conducted using quasi-experimental design, users will not be assigned to random groups for testing purposes. Groups will be pre-formed based on a range of features or dependent on the behaviours displayed; for example, the analysis of behaviours based around conversion, in this case, users could be segmented based on their propensity score (propensity 1 – converters, propensity 75 to 99 – users on-the-edge of converting, propensity 0 – non-converting users). The random assignment of users to groups when looking at specifics such as conversion data would not form the basis of a comprehensive study, making quasi-experimentation a more viable and logical approach to this research.

Following the implementation of experiment design and feature selection, the data will be combined into one table with three iterations; Table 1: raw data, Table 2: psychographic measures and Table 3: a blend of Table 1 and 2 (see Figure 5). This process will help us to achieve the best possible performance and leave room for the adjustment of experimental conditions and the implementation of alternative machine learning algorithms to improve processing time and accuracy. The splitting of datasets will help to evaluate the employed machine learning algorithm/s and explore a broader range of results, forming justified and accurate reasonings. The table structure may change pending experimentation but this table structure is representative of our initial thoughts about the features.

The implementation of machine learning algorithms requires careful consideration when factors such as cost and performance are involved. If the performance of a model is highly accurate but the cost of running the data through the model built using a specific algorithm is very high, this may not be the most efficient approach to business. Building alternatives and reducing the processing time whilst retaining accuracy is paramount to the success of the study and a balance between the two should be considered.

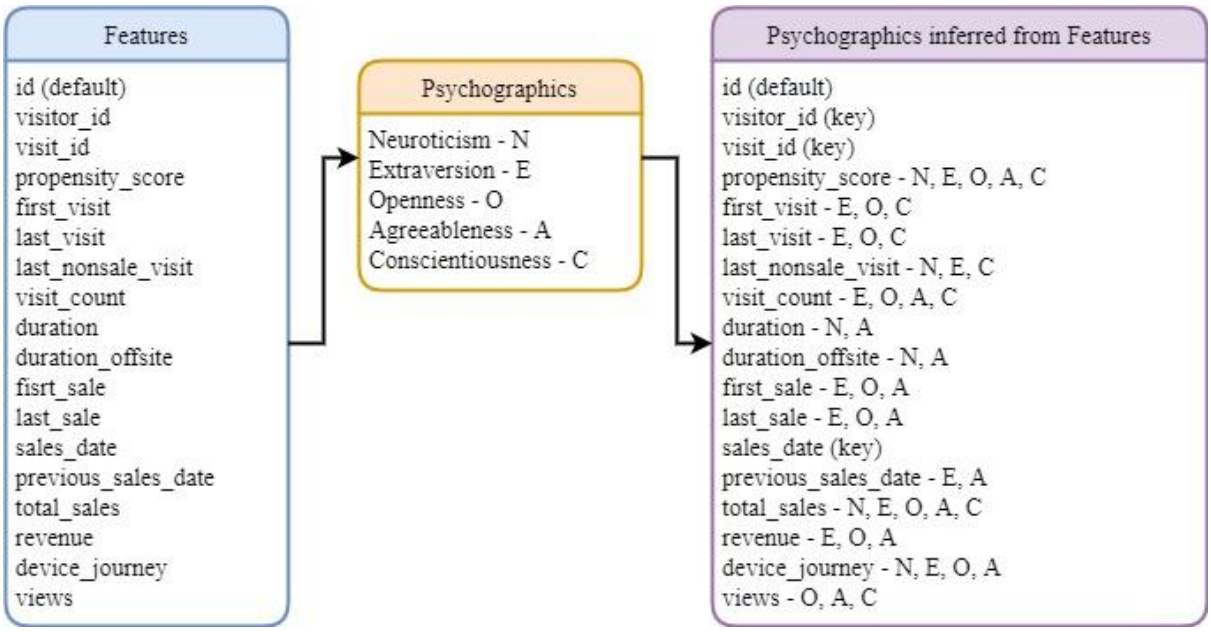


Figure 5. A diagram showing the initial 3 table structure: feature list, psychographics and the psychographics we may infer from each feature. Features labelled with *key* are necessary for the structure of the database (Source: Personal collection)

4.4 Machine Learning Implementation Methods

Machine learning has been successfully utilised by previous studies to understand customer journeys, analyse and predict user behaviour and successfully segment customers (Sato, *et al.*, 2013; Andreeva, Ansell & Crook, 2017; Li, *et al.*, 2020; Anderlová & Pšurný, 2020). Machine learning provides advantages and opportunities for a more profound understanding of the available data. These include a flexible and customisable modelling environment that works successfully with large and complex datasets. The relationships between features are not always initially apparent and machine learning algorithms accommodate this learning process and allow for feature and data experimentation. Many algorithms are easy to implement, computationally cheap to run and can produce strikingly accurate results.

4.4.1 Logistic Regression

Logistic Regression (also referred to as Logit Regression or Logit Modelling) is a machine learning algorithm used in statistics to estimate the probability of an event occurring based on previous data. It is a form of regression analysis whereby the dependent variable (a value that depends on that of another) is dichotomous, a successful way of determining the

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

relationship between variables. This approach to machine learning can be applied to a broad range of research problems and can be altered to improve results through overfitting (data containing more parameters). The standard equation for the calculation of logistic regression is as follows:

Figure 6. The equation for the plotting of a logistic curve

To form a visualisation of the results achieved using Logistic Regression, the data can also be plotted onto a graph:

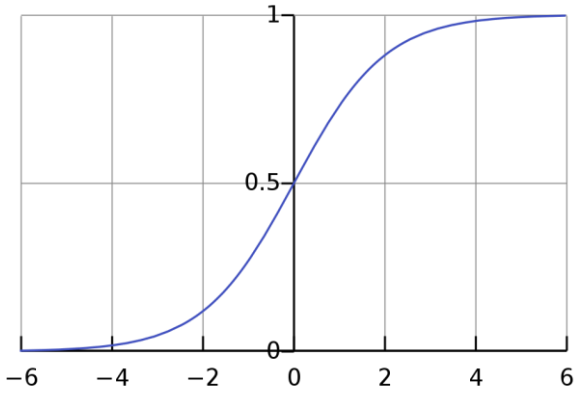


Figure 7. Logistic regression plotted onto a graph (Capeletti, D, 2018)

Logistic Regression can measure the relevance of a predictor but also the direction of association (positive or negative), it is easy to interpret, implement and efficient to train. This method provides disadvantages such as the assumption of linearity between the dependent variable and the independent, and we may discover that our data is much too complex for a model such as this. Despite this, the processing time on a Logistic Regression model will no doubt be much quicker than that of Neural Networks and it's important to explore all options relating to cost vs. performance.

4.4.2 Cluster Analysis

Cluster Analysis refers to the algorithms that are used to form groups based on similarities. This form of analysis is performed on raw datasets where each row signifies an object and each column represents quantitative characteristics or variables, for example:

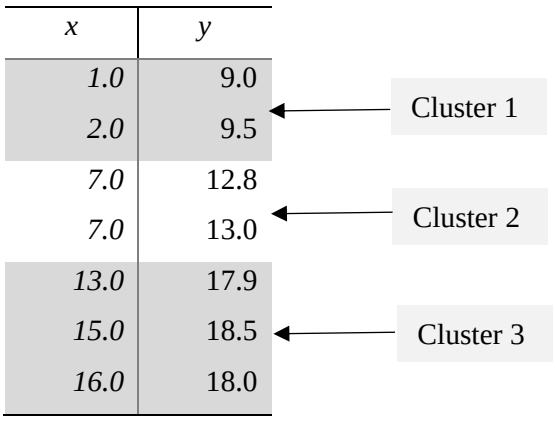


Figure 8. A diagram with 2 columns of values sectioned into generalized clusters (Source: Personal collection)

It is easy to identify the clusters using the data in the above example, this is because there are only two dimensions (x and y). Cluster analysis would provide more interesting results with high-dimensional data (i.e. 20 variables) where the relationships may not be easy to identify.

Several variations of clustering algorithms have been developed including *Hierarchical clustering, k-means cluster analysis and Latent class analysis*, all of which can provide solutions for different problems. Hierarchical clustering initially treats each object as its own cluster and follows an iterative process whereby is joins two objects that are the closest together and then merges the two most similar clusters together. K-means cluster analysis requires the user to determine the required number of clusters; observations are allocated to the specified clusters; cluster means are computed and then objects are joined with the nearest cluster. K-means also follows an iterative process until the clusters no longer change.

4.4.3 Neural Networks

Neural networks are algorithms primarily designed to recognise patterns, are loosely modelled on the human brain and there are two main uses; data clustering and classification. These algorithms can process various data types including sound, text and image into

recognisable patterns contained in vectors and help to group unlabelled data depending on their similarities.

These algorithmic networks are comprised of nodes and this is where the data is passed through and where computation takes place. Each node blends the input from the data with a set of coefficients (or weights) that either strengthen or dampen that input, improving the learning process by assigning significance to inputs. One node is compromised of five key stages (see Figure 9).

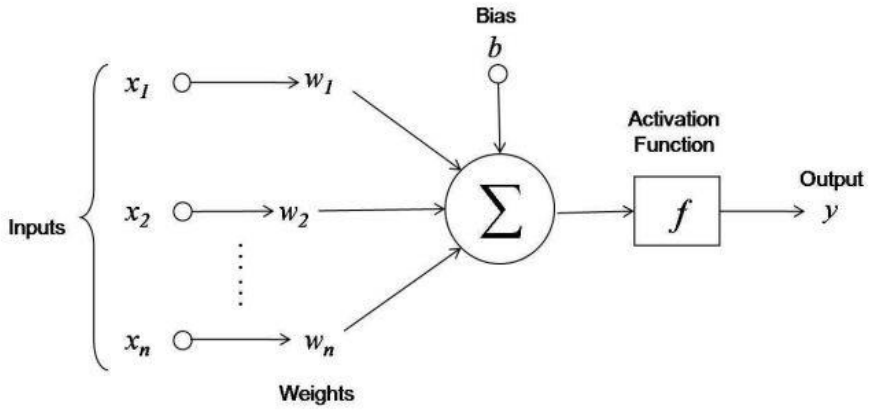
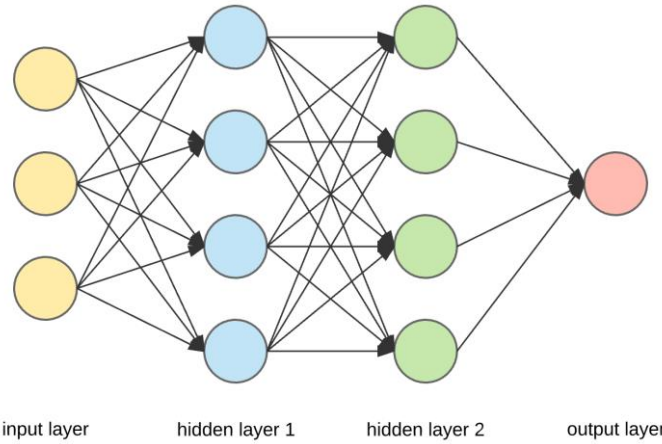


Figure 9. A visual representation of a single Neural Network node (Gupta, R, 2017)

A node layer will be compromised of several nodes in a row which act like switches when data is passed in and out. Similar to Cluster analysis, Neural networks run through several iterations or layers, the output of one layer will subsequently act as the input for another until results are achieved, for example.

Neural networks have several advantages to note, firstly, they can handle the processing of years' worth of data and therefore results are likely to be more accurate. Secondly, they are capable of processing incomplete datasets or unlabelled data with ease and have a tolerance for fault within the data which won't prevent it from processing through a node. Finally, they are



powerful enough to process in parallel and have more computational power than other machine learning algorithms among other advantages.

Figure 10. A diagram outlining the general structure of a Neural Network (Sorokina, K, 2017)

The main disadvantage of working with Neural networks is what is referred to as a 'Black box', this means that we don't know how or why the neural network produced a particular output. This makes predictions and hypothesis difficult to establish before running the data through the Neural network as the processes between input and output are unknown. There are simpler algorithms with a similar structure such as Decision trees where all the processes can be interpreted. If justification is required alongside the results, Neural networks may not be the best option. These algorithms are computationally expensive, meaning that the processing time could run into hours or days; this is dependent on the size of the dataset, the number of nodes and data complexity.

Deep learning is often referred to as a specialised form of machine learning, a subset of Artificial Intelligence (AI) that learns through experiences and performs tasks based on these. There are many variations of deep learning algorithms such as:

Long Short-Term Memory Networks (LSTMs); a type of recurrent neural network that memorises and learns long-term dependencies. They remember previous inputs and retain information over time. The architecture of LSTMs is chain-like and follows steps through collaborative layers (see Figure 11).

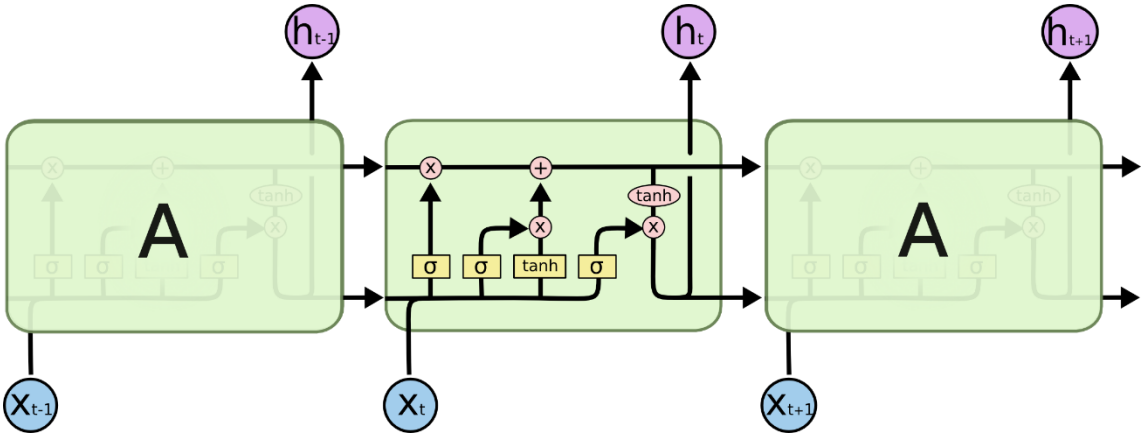


Figure 11. A visual representation of a Long Short-Term Memory Network (Olah, C, 2015)

Convolutional Neural Networks (CNNs); architectures consisting of multiple layers that process and extract data features (see Figure 12). CNN's are made up of 4 layers

(Convolution, Rectified Linear Unit, Pooling and Fully Connected) and are used mainly for the processing of images and object detection.

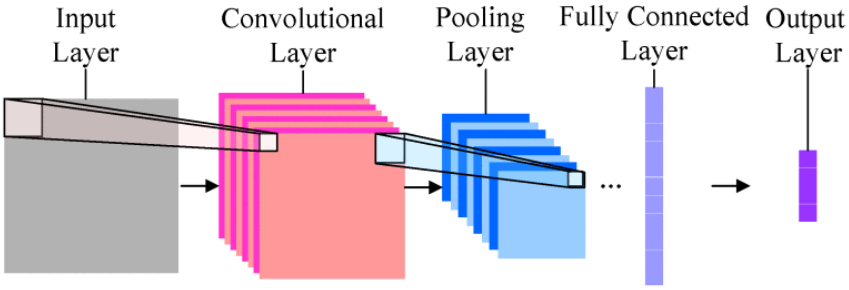


Figure 12. A diagram explaining the structure of a Convolution Neural Network (Peng, M., *et al*, 2016)

Recurrent Neural Networks (RNNs) follow a sequence of events and consider previous events. They are compromised of several hidden states which output vectors after they have been activated, similar to that of a standard neural network, however, RNNs go through several iterations before they have fully processed (see figure 13). They are commonly used for speech recognition, language translation, video recognition and sentiment classification.

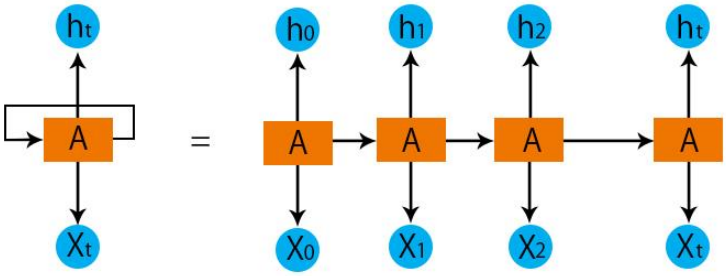


Figure 13. Visual representation of the structure of a Recurrent Neural Network (EDUCBA, 2020)

Deep learning algorithms perform differently from other machine learning methods as they do not require manual feature extraction and can process images as direct inputs. They require significantly more power and performance than that needed for logistic regression of cluster analysis and more data means increased accuracy. Deep learning learns variations in the data automatically and can process massive amounts of computations in parallel with ease. The drawbacks of deep learning include the hordes of data required for them to perform well and the expenses needed to train the models, cost vs. performance.

5.0 Optimization

To gauge a clear understanding of each algorithm, its abilities and the accuracy it provides, each of the machine learning algorithms discussed will be applied during the process of experimentation. A table of results will be curated stating the dataset used, the algorithm used, performance values and the time taken to process. Attempts to recreate the behavioural models proposed by other researchers in the field could provide interesting results and extend the experimental process to optimize the algorithms.

The evaluation of the machine learning algorithms employed is essential and metrics such as *Accuracy Score* may originally provide positive results, the skill of the model and the accuracy it produces is all relative; this problem has not been solved before with these exact features in this exact environment. It would be important to follow standard procedure and develop and baseline model to determine the lowest acceptable performance of a particular machine learning algorithm. This can help to determine and evaluate the success of alternative models that will be trained on the data. If a model produces results below the discovered baseline, it may be that the model is not appropriate for the problem and alternative considerations should be made based on these initial findings. Amazon Web Services (AWS) will be used the determine the cost vs. performance of the algorithms to determine the optimal algorithms for accuracy and cost vs. performance.

Accuracy is defined in different ways depending on the type of algorithm; a perfect score for a regression model is 0.0 error and the best score for classification is 100% accuracy. It is important to note that these scores are unachievable as all predictive modelling algorithms will contain prediction error. These errors can originate from missing or incomplete data samples and data noise (meaningless or corrupt data). Optimizing the results to fall somewhere between the upmost accuracy (i.e.100%) and the baseline would be considered a good score and the performance can be improved on using either the Start High or Exhaustive Search strategies. Start high requires the selection of a machine learning method that is known to perform well on a range of problems and is sophisticated. The results can be evaluated and be used to form an approximate top-end benchmark for results; following this, the simplest model that achieves a similar level of performance can be found. Exhaustive search evaluates all suitable and available machine learning algorithms for the problem and uses the results to select the best performance relative to the baseline. The employed algorithms should be further evaluated using metrics such as *Logarithmic Loss, F1 Score, Classification Accuracy* and

Confusion Matrix to clarify the true accuracy, algorithmic justifications and a deeper understanding of performance.

The experimental process will be run in phases to enable us to reflect on and refine the results and eliminate methods that yield little to no results, the process will be recursive until the successful methods are determined and fully optimized.

5.1 Implications for alternative Industries

We acknowledge that the project thus far has taken place with one specific dataset in mind; luxury jewellery. This does not, however, limit the usability of the model based on the measurement of psychographics in luxury e-commerce, it can be considered for use among most e-commerce businesses. The list of features disclosed (see Figure 5) are moderately generic and would be available in most if not all e-commerce visitor, visit and sales databases. These may vary in name but businesses tracking their online sales will have this information to hand as well as more personal information (should they wish to include demographics and/or geographic information).

The behaviours observed will be different in alternative industries but this does not restrict the uses the final model may provide to businesses. The emotional connections users may have with products will differ depending on the product archetype (luxury vs. utilitarian), but the individual user behaviours are still important, applicable and worth considering. These all attribute to the accuracy of the customer segmentation that leads to more effective sales, marketing and attribution procedures.

6.0 Conclusions

This paper has outlined the issues regarding the impersonal and standardised approach to behavioural modelling in industry and the weaknesses they pose for future development. A quasi-experimental research design has been proposed for the study and improvement of models going forward to develop personal, forward-thinking alternatives, contributing innovative solutions to the analysis of user behaviour.

Research has been conducted encompassing digital marketing, user behaviour, modelling for behaviour and the application of psychographics for user behaviour analysis. Explorations have been made into the models presented by researchers in these fields and considerations will be made regarding the replication of these to determine their success and suitability for this dataset. A methodology and hypotheses have been formed for the future development of the project alongside a guide for the construction of a suitable dataset.

Optimization has been discussed alongside cost vs. performance and the measurement of accuracy, enabling the best possible results to be presented as a solution to the impersonal models currently available to businesses.

Using carefully selected data features we will analyse the purchasing behaviours of customers, linking these to the NEO-FFI attributes: Neuroticism, Extroversion, Openness, Agreeableness and Conscientiousness. Through the experimentation of extensive algorithmic approaches, we will create a model that will analyse the data and apply psychographic markers to individual users. The results will be used to establish customer segments that are reflective of true user purchasing behaviours and individual visitor attitudes. We will aim for the customer segments to be processed in real-time to deliver the best results with the highest levels of accuracy and recency possible. The implications of this study will have a significant impact on the accuracy of user segmentation, improvements to marketing and help to advance the personalisation of services to users. Businesses will be provided with an alternative to standardised, out-of-date attribution models which will constantly learn from real users to improve experiences and drive sales.

References

Anderl, E., Becker. I., Wangenheim, F., Schumann, J, (2016). 'Mapping the customer journey: Lessons learned from graph-based online attribution modelling', *International Journal of Research in Marketing,* 33 (3), pp. 457-474, DOI: 10.1016/j.ijresmar.2016.03.001

Anderlová, D., Pšurný, M, (2020). 'Exploring the Importance of Emotions Within Consumer Behaviour on the Czech Luxury Cosmetic Market', *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis,* 68 (2), pp. 363-372, DOI: 10.11118/actaun202068020363

Andersen, J., Larsen, R. S., Giversen, A., Pedersen, T. B., Jensen, A. H., Skyt, J, (2000). 'Analyzing Clickstreams Using Subsessions', *Proceedings of the 3rd ACM international workshop on Data Warehousing and OLAP,* Virginia: United States of America, 1st November 2000

Andreeva, J., Ansell, J., Crook, J, N, (2005). 'Modelling the purchase propensity: analysis of a revolving store card', *Journal of the Operational Research Society,* 56 (9), pp.1041-1050, DOI: 10.1057/palgrave.jors.2601933

Ariker, M., Heller, J., Diaz, A., Perrey, J, (2015). 'How Marketers Can Personalize at Scale', Harvard Business Review

Bailey, C., Baines, P. R., Wilson, H., Clark, M, (2009). 'Segmentation and customer insight in contemporary services marketing practice: why grouping customers is no longer enough', *Journal of Marketing Management,* 25 (3-4), pp. 227-252, DOI: 10.1362/026725709X429737

Barnes, S. J., Bauer, H. H., Neumann, M. M., Huber. F, (2007). 'Segmenting cyberspace: a customer typology for the internet', *European Journal of Marketing,* 41 (1-2), pp. 71-93, DOI: 10.1108/03090560710718120

Curtis, T., Arnaud, A., Waguespack, B, (2017). 'Advertising Effect on Consumer Emotions, Judgements, and Purchase Intent', *Asian Journal of Business Research,* 7 (2), pp. 57-73, DOI: 10.14707/ajbr.170037

Donnellan, M. B., Conger, R. D., & Burzette, R. G, (2007). 'Personality Development From Late Adolescence to Young Adulthood: Differential Stability, Normative Maturity, and Evidence for the Maturity-Stability Hypothesis', *Journal of Personality,* 75 (2), pp. 237-264, DOI: 10.1111/j.1467-6494.2007.00438.x

Donnellan, M. B., Lucas, R. E, (2009). 'Age Differences in the Big Five Across the Life Span: Evidence from Two National Samples', *Psychol Aging,* 23 (3), pp. 558-566, DOI: 10.1037/a0012897

Dost, F., Wilken, R., Eisenbeiss, M., Skiera, B, (2014). 'On the Edge of Buying: A Targeting Approach for Indecisive Buyers Based on Willingness-to-Pay Ranges', *Journal of Retailing,* 90 (3, 2014), pp. 393-407, DOI: 10.1016/j.jretai.2014.03.007

France, S. L., Ghose, S, (2019). 'Marketing analytics: Methods, practice, implementation, and links to other fields', *Expert Systems with Applications,* 119 (2019), pp.456-475, DOI: 10.1016/j.eswa.2018.11.002

Giddings, C. (2011). 'Measuring engagement across channels', *Multichannel Merchant,* 28 (7), pp.20-22

Ha, T., Lee, S, (2015). 'User Behaviour Model Based on Affordances and Emotions: A New Approach for an Optimal Use Method in Product-User Interactions', *International Journal of Human-Computer Interaction,* 31 (6), pp. 371-384, DOI: 10.1080/10447318.2014.986636

Hoffman, D., Novak, T, (1997). 'Measuring the Flow Experience Among Web Users', *Interval Research Corporation,* Palo Alto: United States of America, 31st July 1997

John, O. P., Donahue, E. M., Kentle, R. L, (1991). The Big Five Inventory – Versions 4a and 54. Berkley, CA: University of California, Berkeley, Institute of Personality and Social Research.

Li, Z., Sha, Y., Song, X., Yang, K., Zhao, K., Jiang, Z., Zhang, Q, (2020). 'Impact of rick perception on customer purchase behaviour: a meta-analysis', *Journal of Business & Industrial Marketing,* 35 (1), pp. 76-96, DOI: 10.1108/JBIM-12-2018-0381

Libai, B., Narayandas, D., Humby, C, (2002). 'Toward an Individual Customer Profitability Model: A Segment-Based Approach', *Journal of Service Research,* 5 (1), pp. 69-76, DOI: 10.1177/1094670502005001007

Lim, S. H., Kim, D. J, (2020). 'Does Emotional Intelligence of Online Shoppers Affect Their Shopping Behavior? From a Cognitive-Affective-Conative Framework Perspective', *International Journal of Human-Computer Interaction,* 36 (14), pp.1304-1313, DOI: 10.1080/10447318.2020.1739882

Lissitsa, S., Kol, O, (2016). 'Generation X vs. Generation Y – A decade of online shopping', *Journal of Retailing and Consumer Services,* 31, pp. 304-312, DOI: 10.1016/j.jretconser.2016.04.015

Lissitsa, S., Kol, O, (2019). 'Four generational cohorts and hedonic m-shopping: association between personality traits and purchase intention', *Electronic Commerce Research,* DOI: 10.1007/s10660-019-09381-4

Mauri, C., Maira, A., Turci, L, (2015). 'An empirical study of consumer behaviour related to private labels and national brand promotions', *The International Review of Retail,*

Distribution and Consumer Research, 25 (4), pp. 33-361, DOI: 10.1080/09593969.2015.1042494

Moe, W. W., Fader, P. S, (2002). 'Dynamic Conversion Behavior at E-Commerce Sites', *Management Science,* 50 (3), pp. 326-335

O'Flaherty, B., Heavin, C, (2015). 'Positioning predictive analytics for customer retention', *Journal of Decision Systems,* 24 (1), pp. 3-18, DOI: 10.1080/12460125.2015.994353

Oliver, J. D., Rosen, D. E, (2010). 'Applying the Environmental Propensity Framework: A Segmented Approach to Hybrid Electric Vehicle Marketing Strategies', *Journal of Marketing Theory and Practice,* 18 (4), pp. 377-393, DOI: 10.2753/MTP1069-6679180405

Othman, A. K., Jailani, S. F. A. K., Kassim, E. S., Hamzah, M. I, (2013). 'The influence of Supplier Characteristics, Customer Trust and Online Emotional Intelligence on Perceived Value and Satisfaction of Online Purchasing Behaviour', *International Journal of Business and Management,* 8 (24), pp. 37-47, DOI:10.5539/ijbm.v8n24p37

Pura, M, (2005). 'Linking perceived value and loyalty in location-based mobile systems', *Journal of Service Theory and Practice,* 15 (6), pp. 509-538, DOI: 10.1108/09604520510634005

Ren, K., Fang, Y., Zhang, W., Liu, S., Li, J., Zhang, Y., Wang, J, (2018). 'Learning Multi-touch Conversion Attribution with Dual-attention Mechanisms for Online Advertising', *The 27th ACM International Conference on Information and Knowledge Management (CIKM) 2018,* Torino: Italy, 22nd to 26th October 2018

Rendle, S., Freudenthaler, C., Schmidt-Thieme, L. (2010). 'Factorizing personalized Markov chains for next-basket recommendation', *Proceedings of the 19th International Conference on World Wide Web, WWW 2010,* North Carolina: United States of America, 26th to 30th April 2010

Richardson, A. (2010). Using Customer Journey Maps to Improve Customer Experience. [Online]. Harvard Business Review. Available at: https://hbr.org/2010/11/using-customer-journey-maps-to [Accessed: 01.03.2021]

Sato, A., Tamura, T., Huang, R., Ma, J., Yen, N. Y, (2013). 'Smart Business Services via Consumer Purchasing Behavioural Modeling', *2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing,* Beijing: China, 20th to 23rd August 2013

Schröder, N., Hruschka, H, (2017). 'Comparing alternatives to account for unobserved heterogeneity in direct marketing models', *Decision Support Systems,* 103 (2017), pp. 24-33, DOI: 10.1016/j.dss.2017.08.005

Shao, X., Li, L, (2011). 'Data-driven Multi-touch Attribution Models', *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* San Diego, California: United States of America, 21st to 24th August 2011

Thatcher, J. B., George, J. F, (2004). 'Commitment, Trust, and Social Involvement: An Exploratory Study of Antecedents to Web Shopper Loyalty', *Journal of Organizational Computing and Electronic Commerce,* 14 (4), pp. 243-268, DOI: 10.1207/s15327744joce1404_2

Ven den Poel, D., Buckinx, W, (2004). 'Predicting online-purchasing behaviour', *European Journal of Operational Research,* 166 (2005), pp. 557-575, DOI: 10.1016/j.ejor.2004.04.022

Wang, H., Wang, S. (2006.) 'A Purchasing Sequences Data Mining Method for Customer Segmentation' *2006 IEEE International Conference on Service Operations and Logistics, and Informatics,* Shanghai, China, 21st to 23rd June 2006.

Zhou, T., Lu, Y, (2011). 'The Effects of Personality Traits on User Acceptance of Mobile Commerce', *International Journal of Human-Computer Interaction,* 27 (6), pp. 545-561, DOI: 10.1080/10447318.2011.555298

Figures