Unveiling the features of successful eBay smartphone sellers

Ana Teresa Silva[1]

Sérgio Moro[2,*]

Paulo Rita[3]

Paulo Cortez[4]

Abstract

The present study adopts a data mining approach based on support vector machines (SVM) for modeling the number of sales of smartphone devices by eBay sellers. The data-based sensitivity analysis was adopted for extracting meaningful knowledge translated into the relevance of each input feature for the model. Such approach allowed unveiling that the number of items the seller also has on auctions, the price and the variety of products the seller offers are the three features that influence most the number of sales, in a total of almost 25%, surpassing the relevance of the features related to customers' feedback.

Keywords

Online sales; eBay sellers; data mining; sensitivity analysis; smartphones.

[1] E-mail: at.nbcs@gmail.com; ISCTE Business School, ISCTE – University Institute of Lisbon, Portugal
[2] E-mail: scmoro@gmail.com; Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR-IUL, Lisboa, Portugal, & ALGORITMI Research Centre, University of Minho, Portugal
[3] E-mail: paulo.rita@iscte.pt; Instituto Universitário de Lisboa (ISCTE-IUL), CIS-IUL, Lisboa, Portugal, & NOVA Information Management School (NOVA IMS), Campus de Campolide, 1070-312 Lisbon, Portugal
[4] E-mail: pcortez@dsi.uminho.pt; ALGORITMI Research Centre, University of Minho, Portugal
[*] Corresponding author. Postal address: Av.ª das Forças Armadas, 1649-026 Lisboa. Phone Nr.: +351 217 903 000

Unveiling the features of successful eBay smartphone sellers

Abstract

The present study adopts a data mining approach based on support vector machines (SVM) for modeling the number of sales of smartphone devices by eBay sellers. The data-based sensitivity analysis was adopted for extracting meaningful knowledge translated into the relevance of each input feature for the model. Such approach allowed unveiling that the number of items the seller also has on auctions, the price and the variety of products the seller offers are the three features that influence most the number of sales, in a total of almost 25%, surpassing the relevance of the features related to customers' feedback.

Keywords

Online sales; eBay sellers; data mining; sensitivity analysis; smartphones.

1. Introduction

With the advent of Web 2.0 and online shopping, an immensity of data is collected from myriad applications and devices. EBay is an excellent example of an online company boosting its way through the Web 2.0 era, being currently one of the largest online sales platforms, supplying online retailing services for any seller worldwide (Einav et al., 2014). Such a colossal player entails a large set of different means for users to contribute with feedback on the services provided and registered sellers. These feedback data plus other relevant data (e.g., data on the items being sold, and users' characteristics) is scattered throughout multiple sources, which inevitably asks for some form of further treatment that allows classification, discovery of patterns and trends or prediction of outcomes. Such treatment implies the usage of increasingly complex and combined statistical and machine learning tools as the size of datasets builds up (Amado et al., 2018). Nowadays, datasets may extend to several Exabytes, increasing the challenging task of transforming such loads of information into actionable knowledge using adequate methods (Canito et al., 2018).

Data mining is the process of discovering patterns of knowledge from raw data (Sharda et al., 2018). Its roots lie on statistics and data analysis, and have been greatly enhanced through machine learning techniques and methods. Data mining as an evolving process has been around for some time, but only since the 1990s, when the concept was coined, until today has it been gaining considerably more popularity and attention (Fayyad et al., 1996; Sharda et al., 2018). This is happening due to the large amounts of data, in what is known as big data, that are generated every second from several sources, such as sensors and devices (Pal et al., 2014) and also social media and smartphones' applications (Chen et al., 2012).

These themes are the stepping stones for this data mining study. Therefore, its goal is to generate the type of information that is able to leverage decision support through actionable knowledge. It might be of particular interest for online retail sellers, online marketplaces and marketing practitioners, who may use the insights provided by the analysis of how online features of sellers' influence sales. In fact, large online e-commerce websites represent the future of retailers (Clemes et al., 2014), and top players such as eBay, Amazon and Alibaba are among the most technologically innovative organizations worldwide (Liu and Lu, 2015). Therefore, research on improving customer service based on cutting edge technology can help cope with the challenges of tomorrow.

Traditional data mining projects are time-consuming as all the data is often manually extracted and with limited amount of resources, which usually leads to limitations in the scope of analysis. In this case, the research is narrowed to the extraction of knowledge in the form of features' relevance from sellers of smartphones on eBay, one of the largest e-tailers worldwide (Kornberger et al., 2017). The aim of this study is to provide insights about what it takes to be a successful eBay smartphones' seller by unveiling through data mining which seller features contribute the most to actual sales, i.e. which have the highest influence on the number of items sold. Previous literature has approached the subject mostly from customer and potential consumers' perspectives yet rarely from the sellers' point of view. As the number of registered sellers on online platforms rises worldwide, it becomes crucial to understand what drives the success of sellers within the different dimensions that can influence their results (Wu et al., 2015; Kannan, 2017). Such knowledge can be valuable both from a seller's perspective as well as for managing online platforms (e.g., offering premium services to the most prospective sellers or improving feedback services and information supplied to the registered users).

There is research focused on modeling the choice between auctions and posted prices (Einav et al., 2018), on online businesses emerging in the form of eBay ventures (Gregg & Parthasarathy, 2017), consumer trust in online purchases (Oghazi et al., 2018), e-satisfaction and consumer spending (Nisar & Prabhakar, 2017), and eBay sellers' reputation (Greenstein-Messica & Rokach, 2018). Yet a stream of research that grasps onto the conspicuous and measurable characteristics of online sellers combined with product attributes in order to determine their impact on sales using data mining predictive techniques is still scarce in the literature. Therefore, the immediate purpose of this paper is to fill in that research gap. In addition, the contributions for the literature are the following:

- Extraction of online seller features from an online sales renowned platform, eBay;
- Evaluation of the smartphones' online market through the analysis of eBay sellers' features and their impact on performance.

The next section dives deeply into the theoretical background, which supports the relevance of the subject along with the data mining techniques in use, followed by a detailed description of the chosen methodology and approach. Then, the results are discussed and interpreted in order to extract adequate knowledge out of the data. Finally, the conclusions are drawn in the last section.

2. Theory

2.1. Online sales

Web 2.0 is defined as a "set of applications and technologies that enable user-generated content, such as online social networks, blogs, video and photo sharing sites, and wikis" (Laudon and Traver, 2016, page 71). It is considered a new stage of development of the web and it differs

from the previous one by the drastic increase in information density, interactivity and level of customization. This new phase can be traced back to 2007, when the changes became evident. It is also relevant to draw attention to the associated shift from making online purchases to going shopping online (Hemp, 2006, page 1) as the online environment and virtual communities become vital elements in the consumer journey, in a phenomenon often named "social commerce" (Huang and Benyoucef, 2013). Thus, recommendations from other consumers, instead of friend/family advice are also becoming an increasingly important decision factor (Kotler and Keller, 2012, page 138). It is important to examine the e-tail environment since "electronic markets enable volumes and speeds that human middlemen could not accomplish" (Venkatesan et al., 2006). However, there is still plenty of research focused exclusively on brick and mortar retail context when compared with pure online players and bricks-and-clicks, which have been growing expressively in the last years (Grewal et al., 2010).

Looking from the consumers' perspective, Cheung et al. (2005) pointed out that the main determinants of online consumer behavior were related with consumer characteristics, environmental influences, product/service characteristics, medium characteristics and merchants and intermediates' characteristics, which would have a transversal impact through the online customer journey.

In early research about pricing it was often argued that the advent of the Internet would lead to heightened competition online, which would induce price reductions (Brynjolfsson et al., 2006). Nevertheless, other features of online markets were found to have more impact in the buying decision process such as variety and convenience. Hence, the trade-off between breadth and depth could stand a chance of being solved (Grewal et al., 2010). The turn up of Web 2.0

tools that speed up information sharing and networking has definitely been affirmative in self-generation of content.

Moreover, recent studies have devoted efforts in finding influencing features on the prices of online sales. Kocas and Akkan (2016) evaluated how online feedback and rating from customers affected the prices of books from twenty-four categories sold through Amazon.com. Their work has proved that customer ratings should be accounted for increasing profitability. Cao et al. (2015) presented a study on dynamic pricing of online shopping by dividing customers in patient and inpatient potential buyers, providing evidence that the optimal pricing policy should limit dynamic pricing when facing customers with little patience. Sellers' reputation has proven to be an effective influencer of the pricing policy followed, with highly reputed sellers having advantages in pricing, as shown by Xu and Ye (2015) through an analysis of TaoBao sellers. However, the same study also emphasizes that literature on pricing related to reputation is scarce. A previous article published by Ye et al. (2013) has reached a similar finding by analyzing both TaoBao and eBay sellers. Both studies are conclusive in that sellers' features do affect pricing, influencing sales performance, with the latter adopting a regression model for studying three sellers' attributes: reputation score, number of positive reviews, and score for "item as described". However, this study did not consider further features from sellers that are available on eBay, such as the neutral and negative reviews. Furthermore, both studies analyzed online sellers in a pricing perspective, not accounting for the number of sales derived from sellers' features, a research gap that the present study attempts to fill.

Managing the variety of products sold, i.e., the assortment has always been an essential element of business development (Ramdas, 2003). In today's environment where high levels of demand together with increased want for a personalized offer have become frequent, finding the

right balance between variety and the level of customization is often a challenge. High variety can be associated with increased variability and lead to errors in forecasting (Ramdas, 2003; Fisher, 1997). The adoption of niche versus mass strategies is another important aspect related with assortment management and it is inextricably linked with the level of variety and specialization of the products sold. It was discovered that, from the demand side, huge variety of inefficiently organized items can stagger consumers and hold back purchases due to forecasting errors and difficulty for consumers to find the products they are looking for (Brynjolfsson et al., 2006). If sellers choose marketing and assortment strategies that are not compatible in ensuring a smooth supply chain (Fisher, 1997), it can have a negative impact on consumer behavior and repurchase intention based on satisfaction (Yen et al., 2007) which will inevitably affect sales.

2.2. Smartphones

Technologies and telecommunications have become essential elements of everyday life and business. The need for increasingly fast and optimized devices has guaranteed a steady growth in the technological industry, although at due different regional paces (Kellerman, 2010). Mobile devices have also become one of the primary sources for online shopping (Pearce and Rice, 2013). In the UK, for example, mobile has already surpassed desktop by 44% (The Guardian, 2014). Such relevance can prove to be an effective driver for increasing sales of mobile devices (Bilgihan et al., 2016). Smartphones belong to this category since they are essentially "mobile phones with more advanced computing capabilities and connectivity than regular mobile phones" (Statista, 2016). Although they have been available in consumer markets since the 1990s, only became truly popular and mainstream when, in 2007, the iPhone's

introduction by Apple transformed the industry, leading also to the first Android based smartphone being released to consumer markets in late 2008 (Lee et al., 2015).

Within the 19 most popular online shopping categories, IT and mobile is ranked 5th, achieving 40% in global online purchase rate, which reveals the potential and relevance of the category for online shopping. Smartphones are mobile phones with operating systems similar to PCs and they are, therefore, included in the mentioned category. Their number of sales has been increasing sharply as in 2013 it already doubled compared to 2011. It is expected that by 2017 the market penetration of the devices will be of 65.8% in Europe and 62.2% in North America. Thus, it is clear that smartphones are gaining more and more popularity and, as such, it is foreseeable that their sales will increase and that gathering valuable market information will be a source of added value in marketing planning. Furthermore, it is important to mention that over 335 Exabyte of data are generated and stored on a yearly basis through smartphones only (Poelker, 2013). This immensity of data is transversal to all industries and can be extremely valuable if used to retrieve important information (Pal, 2013).

It is estimated that by 2017 a third of the population worldwide will own a smartphone, which will, according to forecasts, encompass 2.6 billion smartphones (Statista, 2016). In 2015, solely, the global smartphone industry was responsible for the generation of approximately 240.55 billion Euros although with a decrease by roughly 1.49% comparatively to the previous year.

Smartphones have become one of the most popular devices for purchasing products or services, social media activity or conducting research (Bilgihan et al., 2016) with levels of ubiquitous connectivity and convenience never experienced before (IBM, 2015).Those factors

along with the fact that the price of smartphones has been steadily decreasing (Statista, 2016) have fueled the growth in the industry

2.3. Data mining and support vector machines

According to Yu et al. (2012), online reviews and feedback have embedded in them the unique opportunity of extracting knowledge for leveraging business intelligence. The role of data mining becomes evident when wanting to derive information that generates actionable knowledge and that can be easily accessed and handled by decision makers. It is clear that mining is at the core of business intelligence (Han et al., 2012).

The potential of data mining extends to pretty much any scientific and business area. From astronomy to marketing, fraud detection, manufacturing or telecommunications (Fayyad et al., 1996; Hui and Jha, 2000), its usefulness transcends any field one might contemplate. Within business applicability, increasing customer intelligence, improvement of operational efficiencies and customer customization are only some of the broad possibilities for data mining (Pal et al., 2014). Data mining has been used for modeling tourist hotel scores (Moro et al., 2017), designing of products and information systems (Kusiak and Smith, 2007), predicting bank telemarketing successful contacts (Moro et al., 2015a) or measuring social media performance (Moro et al., 2016). Other examples include the application to e-learning domain (Hanna, 2004), customer response to direct mailing (Coussement et al., 2015), credit risk assessment (Moro et al., 2015b), or for discovering the helpfulness of online reviews (Lee and Choeh, 2014). These are only a few among a vast array of studies in which data mining was used. In sum, data mining and its techniques can be applied to any science and any industry as there is still a plethora of untapped opportunities. However, the particular application should always be taken into account

since there is arguably a universal data mining method so far. Therefore, selecting the most suitable one can be considered somewhat of an art (Fayyad et al., 1996).

In order to perform the inherent data mining tasks, numerous methods can be used. Such procedures usually entail machine learning algorithms, which resort to computational methods that allow learning information straight from the data without the need to have a pre-set equation serving as a model (Mathworks, 2016). This enables improvement of performance as more and more samples are added to the dataset, allowing the machine to learn. Decision trees, neural networks and support vector machines (SVM) are just a few of them. In the expanse of this project, only SVMs and sensitivity analysis will be explained in detail. Within these methods, there are several functionalities to handle the patterns which are found throughout data mining tasks. Those functionalities are fundamentally categorized into descriptive and predictive. The first ones are associated with the description of properties of the data in a target data set while the second ones use induction on the current data in order to make predictions (Han et al., 2012). Vapnik and Cortes (1995) are the "architects" behind the support-vector network learning machine in pivotal stages of SVMs. They presented the idea of mapping nonlinear input vectors into a high-dimension feature space where a linear decision surface would be built within a deeply widening scenario in which training data could be separated with errors, and therefore breaking ground to solving real problems, inspired by the initial discoveries of Fisher (1936) for pattern recognition algorithms.

"Support Vector Learning Machines are finding application in pattern recognition, regression estimation, and operator inversion for ill-posed problems" (Schölkopf et al., 1997), which was an early sign of their increasing popularity and applicability. They are often used for classification of linear and nonlinear data through the transformation of the original data into a

higher dimension using a kernel function that computes dot products in the transformed space (Friedman et al., 2001), from where it can find a hyper plane for data separation using essential training tuples called support vectors (Han et al., 2012). They can also be used for regression with the requirement of adding a loss function (Smola and Schölkopf, 1998).

A support vector machine is an algorithm which belongs to the same typology as other neural network classifiers, e.g., an SVM with radial-basis function displays a matching hyper plane to the neural network identified as radial basis function network (Han et al., 2012). The completeness of SVMs' algorithm enables the construction of models with enough complexity that, are, however, simple in a way which makes mathematical analysis possible. The algorithm comprehends a significant amount of neural nets, radial basis function network and also polynomial classifiers (Hearst et al., 1998).

One of the main benefits of SVMs is that they can attain good performance levels when applied to real problems just as they can be analyzed with higher complexity and employing theoretical concepts from computational learning theory (Hearst et al., 1998). The idea is often supported as SVMs and are considered an attractive approach to data modeling. They combine generalization control with a technique to address the curse of dimensionality (Gunn, 1998). However, more recent research has questioned the statements since in the presence of powers and products, by giving the same weights to terms in 2XjX′j form, a polynomial kernel of degree 2 in a 2-input feature space won't be able to adapt to subspace concentrations and will difficultly find structure by having many dimensions where to search (Friedman et al., 2001). Knowledge would have to be assembled into the model to solve the problem of multidimensionality. On the other hand, the same author backs the idea that, at the time, SVMs performed well when applied to real learning problems. Essentially, an SVM encompasses a set of techniques that allow

building a linear boundary in a large transformed version of the feature space in order to produce nonlinear boundaries (Friedman et al., 2001) and this simplifies analysis because it can be shown to correspond to a linear method in a high-dimensional feature space nonlinearly related to input space (Hearst et al., 1998). In other words, what occurs is that through the usage of a kernel, an approximation function, and addition of a loss function, the hinge loss function, the desired outcome is optimized. The jumbled data in the input space takes form into a separating hyper plane, the feature space, which will be easier to analyze because the data becomes structured and, therefore, further analysis is made possible until it develops into intelligible information.

2.4. Sensitivity analysis

When dealing with black box models, it is often a challenge to extract knowledge in a way that is easy to understand. That fact inspired a new stream of research to tackle the inherent problem. Consequently, methods such as extracting rules from networks and sensitivity analysis (SA) have emerged. Sensitivity analysis enables the assessment of the importance of input factors to a given model (Saltelli et al., 2000) and also their effects on the model's responses (Cortez and Embrechts, 2011). It is frequently employed in order to evaluate the coherence and attractiveness of a kernel-based and ensemble black box models such as SVMs or neural networks and, subsequently, facilitating their interpretation. This is a central element to any model since it will contribute for increased understanding by different audiences and trust in data mining. It can be disclosed using one among sensitivity analysis algorithms together with appropriate visualization techniques. It is pointed out that extraction of rules is rather simplistic and might fail at assessing the representativeness of the model due to disregard of relevant rules

and danger of generalization mainly resultant of discretization of the separating hyper planes (Cortez and Embrechts, 2013).

One of the advantages of SA is its broad applicability to almost any supervised learning model as the relationship goes straight to the bottom line of input-output relationship, i.e., the way that any variation in a given input changes the respective output. There are several types of methods, i.e., algorithms to choose from within SA. Among that group are included one-dimensional sensitivity analysis (1D-SA), global sensitivity analysis (GSA) and data-based sensitivity analysis (DSA). They diverge in their suitability to different goals. 1D-SA is very fast but cannot measure complex interactions among the features, whereas GSA is perfect SA method in terms of interaction measurement but it is computationally too costly. DSA is similar to 1D-SA but it uses training samples in detriment of a baseline vector. In effect, the main goal in the case of DSA is to harvest the possible interactions between inputs but in a faster manner than with GSA. If needed, DSA can even be speed up if a proportion of the training samples (randomly selected) are used instead of the whole training set. Such feature makes DSA computationally much more efficient than GSA, while having a better performance than 1D-SA due to its capability of detecting input variable interactions.

2.5. Regression performance metrics

One of the crucial steps in model building is assessing its adequacy in predicting what it is supposed to. Therefore, one can state that performance and adequacy in prediction models are inevitably connected, i.e., if one model fails to predict its output then it is inadequate (Diebold and Mariano, 2012). This brings out the importance of forecast accuracy since the derived forecasts are used to guide decisions.

Although there are plenty of performance metrics, error metrics are quite often chosen. Multivariate error distributions, which are produced by any forecasting method, enable this process. These metrics were created to assess the discrepancies between predicted and actual values. Additionally, they are relevant in model calibration and refining. As a consequence, choosing the most appropriate metrics for forecasting accuracy is critical (Armstrong and Collopy, 1992).

There are also performance evaluation techniques that contribute to model validity and subsequently, its overall accuracy, such as cross-validation. Models are frequently evaluated using this method and its estimates of a prediction error (Fayyad et al., 1996). According to Refaeilzadeh et al. (2009), cross-validation can be applied to estimate performance, model selection, and tuning learning model parameters. Moreover, it is also considered a reasonable technique to deal with overfitting. With k-fold cross-validation, all the observations are randomly split into k equal subsets, which are used as a test sample. The latter is used to assess model reaction to new data, constituting a realistic predictive testing approach (Berry et al., 2004). The remaining subsets are used as training data, which is used for model building. Essentially, each of the subsets will operate as both training and testing data but only once as the latter. The equally sized k testing subsets are gathered generating an estimate for the whole instances (i.e., cases) of the problem being addressed, which is the validation set.

As far as error metrics are concerned, mean absolute error (MAE) is one of the most frequently used metrics for assessing forecast accuracy and it consists of the mean of the absolute difference between the total of predicted values ($Pred_i$) for a given output variable and its actual values ($True_i$) for all its n observations. Thus, it assesses the deviation in predictive capacity of the model.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|True_i - Pred_i| \quad (1)$$

Mean absolute percentage error (*MAPE*) is fundamentally the ratio of the MAE divided by the total of true values. It is the relative variation to those values and it can only be applicable if $True_i > 0$, otherwise the calculation is impossible.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|True_i - Pred_i|}{True_i} \quad (2)$$

Due to the mentioned restriction of MAPE, other metrics such as the relative absolute error (RAE) came into the picture. The RAE is the difference between predicted and true total values as a fraction of the difference between predicted and average total values. This metric enables adjustment to the average values of the variable. The average value (Avg_i) is imputed into the single variations of each sample element which may contribute to increase accuracy for models with low dispersion. In models with high dispersion, this metric might not be the most suitable since pulling each set of values to its average instead of distributing the total of differences throughout a given n, allowing the weights of the differences to be offset within the model, widens enormously the gap in individual sets of values and ends up escalating the total difference. However, it allows assessing predictive capacity when MAPE cannot, since the average value imputation tackles the division by zero difficulty. The main advantage of *RAE* is the ease of interpretations and communication (Armstrong et al., 1992).

$$RAE = \frac{\sum_{i=1}^{n}|True_i - Pred_i|}{\sum_{i=1}^{n}|Avg_i - Pred_i|} \quad (3)$$

Other metrics may be computed to address the issue raised with the RAE. One of such possibilities is to compute a normalized mean absolute error (NMAE), entailing the distribution of the MAE through the difference between the maximum (R_{max}) and minimum (R_{min}) values of the output variable, as shown next:

$$NMAE = \frac{MAE}{R_{max} - R_{min}} \quad (4)$$

2.6. Online sellers' theoretical framework

The Internet gave rise to a new form of business where individual users compete side-by-side with corporate users as product sellers (Jeon et al., 2008). As the Internet moved forward and gained maturity, so did research in developing theories supporting it (Kaplan and Haenlein, 2010). Particularly, the role of sellers has been widely studied, reflecting a shift after the Web 2.0 emergence, with individual users gaining considerable weight within e-marketplaces (Li et al., 2015). Table 1 summarizes the main constructs of online sales success under two perspectives (seller and product), supported by recently published literature to help in framing online seller's theoretical background. It highlights the relevance of seller's reputation, particularly, in the face of online consumer reviews, which leverage such reputation and, subsequently, positively affect sales. Likewise, a positive product feedback will positively influence sales. The remaining features have an older background support, which dates to before the advent of the Internet. Yet, Table 1 shows that research has been prolific in recent years to update and adapt knowledge to recent developments on online sales. Specifically, consumer empowerment shows evidence of a distinct and more aware consumer behavior. The features highlighted in Table 1 have been individually studied; however, research is lacking a holistic updated view that encompasses all of

them in a unique data-driven model of online sales. This is a gap the present research addresses through a data mining approach, which offers the needed tools to build such holistic model (Moro et al., 2014). Particularly, recent studies have shown the power of data mining to build a model fed with features from different contexts and with a similar size to the collected dataset (e.g., Moro et al., 2017).

Table 1 - Theoretical framework of online sales.

Perspective	Feature	Theoretical contributions	References
Seller	Reviews	Sellers' reputation increases with a higher number of positive reviews. Typically, there are less negative reviews in eBay, but the existing ones have a negative impact on sales.	Tadelis (2016) Wang et al. (2016)
	Origin	In e-marketplaces, thanks to unrestricted boundaries, the seller's country of origin has little influence in sales.	Clemons et al. (2016)
	Nr. items sold	Popularity positively influences consumers' intention to purchase.	Kao et al. (2017)
Product	Condition (used / refurbished / new)	Customers are willing to pay a premium for seller refurbished products, and more for new products, when compared with used products.	Xu et al. (2017)
	Price	Price variations (e.g., promotions) have a greater impact on offline sales than on online sales.	Dinner et al. (2014)
	Brand	Brand loyalty is a known influencing feature of online purchase.	Moro and Rita (2018)
	Segment	Up-selling techniques are known to be effective in persuading consumers to purchase more expensive (higher segment) products.	Dawson and Kim (2009)
	Feedback	Sales tend to increase in the presence of positive feedback by satisfied consumers.	Floyd et al. (2014)

3. Materials and methods

3.1. Approach preamble

When pursuing the employment of data mining techniques, one must go through previous and subsequent technical stages in the knowledge discovery process from problem identification and translation into the data mining world to assess results and possibly repeating the process

(Berry et al., 2004). Prediction is a directed data mining task that requires performing all the tasks associated with those different stages, including business understanding, data preparation, modeling, validation and deployment into production or knowledge extraction for decision support (Han et al., 2012).

A comprehensive dataset, including characteristics of sellers and their items, was extracted manually to serve as the base set for the experiments. Sellers represent the problem instances and the set of characteristics comprises both nominal, ordinal and scale features. Moreover, data cleaning, data integration and data transformation, particularly some level of computation to produce new features, were carried out in order to improve the accuracy and efficiency of the mining algorithm (Han et al., 2012; page 83). Issues with missing values were barely registered since problem instances which did not fulfill all the features were immediately eliminated in the data cleaning phase; therefore, there was not a need to implement any method to tackle that problem. Different techniques, tools and metrics are used within the various stages of the process such as using SVM with RBF kernel for modeling, performing a k-fold cross-validation, computation of performance metrics MAE, RAE and NMAE, and DSA for assessing feature relevance.

For all the experiments conducted, the R statistical tool (https://cran.r-project.org/) was adopted. R is an open source framework for the development of data analysis solution, with a vast number of enthusiasts and contributors of packages in a wide number of fields of interest (Ihaka and Gentleman, 1996). Moreover, the “rminer” package was adopted as it provides a simple and coherent set of functions for performing data mining tasks such as modeling, model performance evaluation and sensitivity analysis (Cortez, 2010).

3.2. Data preparation

The problem at hand is linked with the shortage of information about which elements among seller and product characteristics have an impact on online sales of smartphones and how they affect them in a measurable way. The gathered data ensures the reliability of the predictions by means of extracting factual eBay features; yet, its internal validity is tested in the following stages of the process using error metrics.

Subsequently, one needs to select the appropriate data for the experiments. In this stage, data preparation was essential in compiling a coherent dataset characterized by features that could be used for modeling the number of sales, i.e., the features must provide to a certain degree a correlation with the outcome to predict. The dataset gathered for the experiments included 499 manually extracted reliable observations, which went through a transformation process prior to modeling.

Initially, 23 different features were collected, plus the output variable, which was the number of sales, "prodSales". Those features are listed in Table 2 and identified with source equals to "EBay" in the corresponding column, with Figures 1 to 4 showing the locations on the eBay webpages from where the features were extracted (the respective features' names are identified in the depictions). In the captions of each figure the URL link is also displayed to obtain the webpage identified in each of the figures, for easier reproducibility. Since the R tool was adopted for the experiments, the data types mentioned in Table 2 correspond to R data types (more details on those can be obtained from http://www.dataperspective.info/2016/02/basic-data-types-in-r.html).

Table 2 - List of features.

Feature name	Source	Data type	Description	Status
NameSeller	EBay	Character	Name of the seller on eBay	Removed
nrFollowers	EBay	Integer	Total number of followers of the seller	Approved
posR	EBay	Integer	Total number of positive reviews of the seller	Approved
negR	EBay	Integer	Total number of negative reviews of the seller	Approved
neuR	EBay	Integer	Total number of neutral reviews of the seller	Approved
Country	EBay	Factor	Country from which the product is sold	Approved
Continent	Computed	Factor	Continent of the country	Approved
frItem	EBay	Integer	Feedback rating for the items sold	Approved
frC	EBay	Integer	Feedback rating for the communication	Approved
frST	EBay	Integer	Feedback rating for the shipping time	Approved
frSC	EBay	Integer	Feedback rating for the shipping charges	Approved
diffProd	EBay	Integer	Total products available by the seller	Approved
nrViews	EBay	Integer	Total number of views of the seller	Approved
dateCollection	Computed	Date/time	Date in which data was collected	Removed/Converted
prodType	EBay	Factor	Model of the product	Removed/Converted
Segment	Computed	Factor	Assessment based on *prodType*, *Brand* and other sources	Approved
Brand	EBay	Factor	Brand of the product	Approved
priceMin	EBay	Numeric	Minimum price of the product (€)	Removed/Converted
priceMax	EBay	Numeric	Maximum price of the product (€)	Removed/Converted
nrItems4Sale	EBay	Integer	Number of items in "Buy it now" section	Approved
nrItemsAuction	EBay	Integer	Number of items in "Auction" section	Approved
nrResults4Phone	EBay	Integer	Similar to "diffProd", except it considers only items under the category of "Cell Phones & Smartphones"	Approved
Condition	EBay	Factor	Condition of the product (1=used, 2=refurbished, 3=new)	Approved
moreProdSales	EBay	Integer	Additional sales	Removed
memberSince	EBay	Date/time	Date of membership	Removed/Converted
diffToToday	Computed	Integer	Interval between *dateCollection* and 20th March 2016	Approved
priceAvg	Computed	Numeric	Average of minimum and maximum price (€)	Approved
memberDays	Computed	Integer	Interval between *dateCollection* and *memberSince*	Approved
prodSales	EBay	Integer	Total sales of the product	Approved

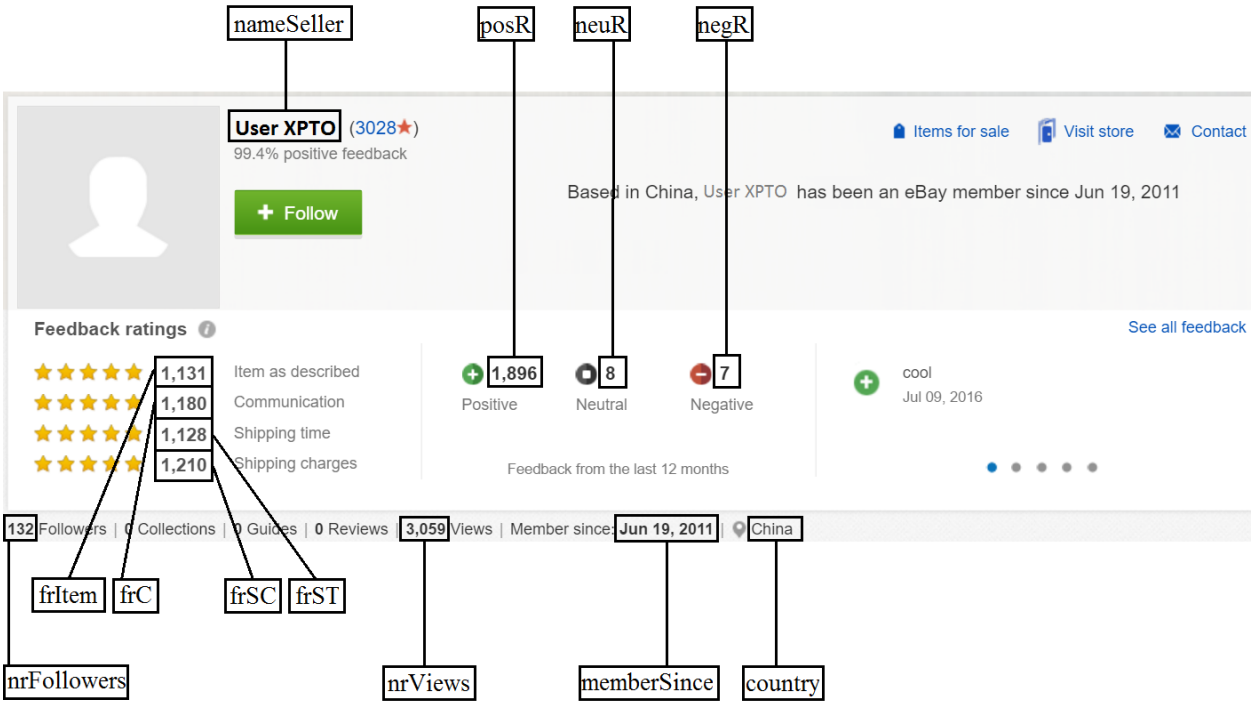

Figure 1 - Locations for the seller's features extracted from eBay (http://www.ebay.com/usr/<user>).

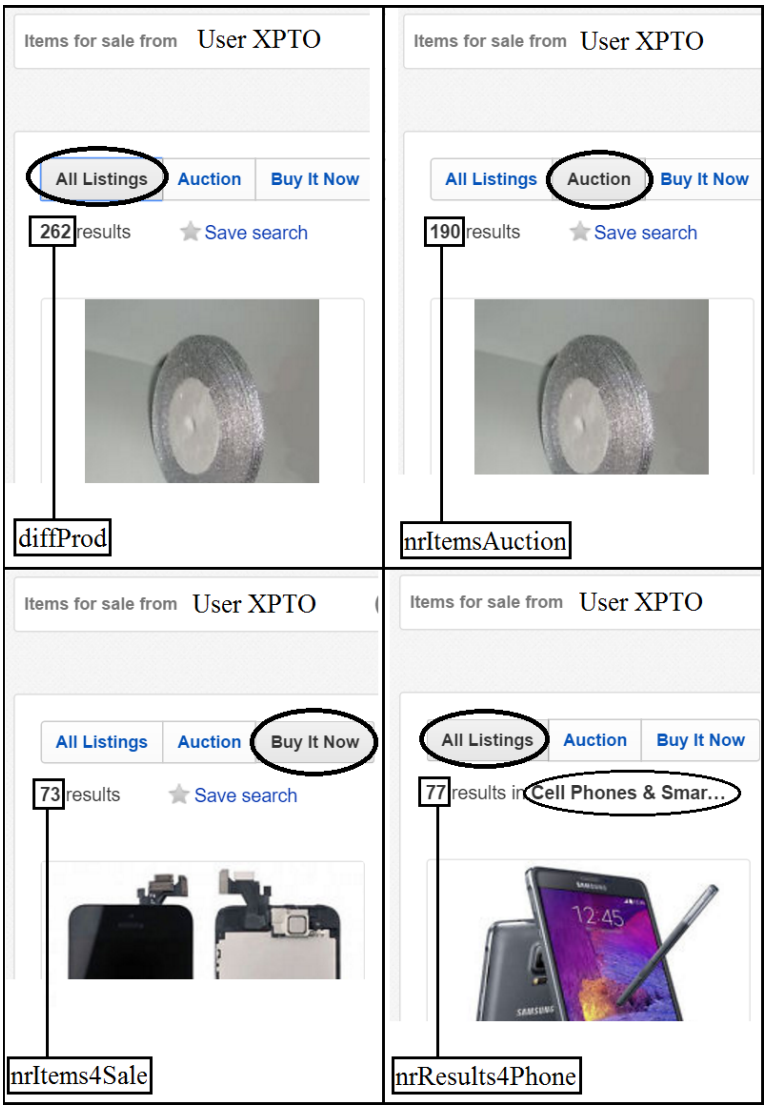

Figure 2 - Locations for the seller's products' features extracted from eBay (http://www.ebay.com/sch/<user>/m.html).

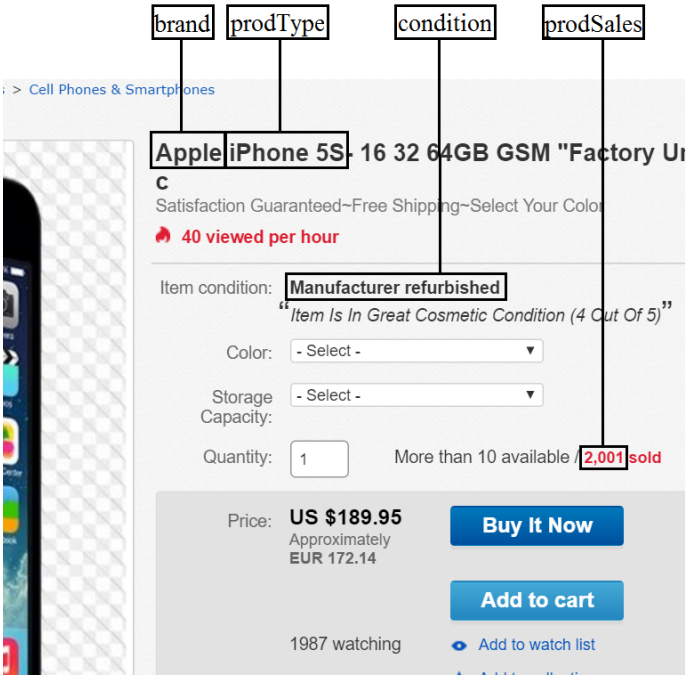

Figure 3 - Locations for the product's features extracted from eBay (http://www.ebay.com/itm/<product>).

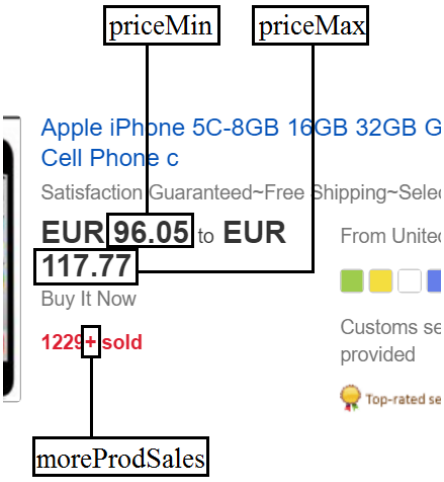

Figure 4 - Locations for the price's features extracted from eBay (http://www.ebay.com/sch/i.html?_nkw=<product>).

As part of the data preparation process, other additional features were calculated and added to the dataset. These features are also displayed in Table 2, marked with the source "computed". The "dateCollection" registers the date when the data for that record's seller was collected (between February 23 and March 15, 2016). The "continent" is another of the features included and was based on the seller's country and using the convention of seven prevailing continents as defined by Lewis (1997): Africa, Antarctica, Asia, Europe, North America, Oceania and South America. Despite the existence of several criteria, the former was chosen because it depicted more precisely the distinctions between Europe and Asia, North America and South America, which can be of value in understanding the effects of the geographical and cultural nature associated with the seller. The "segment" is the result of a categorization based on the smartphone brand and model, the model's release date for covering the issue of older and outdated models over the course of time and reviews from renowned sources such as CNET and GSMArena (Table 3).

Table 3 - Categorization of smartphones' segments.

segment	Brand	prodType	Release date[1]	CNET[2] review
1	Apple	iPhone 4S	October 2011	8.8
2	Apple	iPhone 5	September 2012	8.7
	ZTE	BoostMax	January 2014	6.7
	BlackBerry	Leap	April 2015	6.6
	Alcatel	OneTouchPopC9	June 2014	N.A.
3	Apple	iPhone 5S	September 2013	8.5
	Huawei	AscendP6	June 2013	N.A.
	Huawei	AscendP7	June 2014	6.7
	ZTE	AXONmini	November 2015	N.A.
	LG	G4	June 2015	8
	Microsoft Mobile	Lumia535	December 2014	6.3
	Microsoft Mobile	Lumia640	March 2015	7.2
	Microsoft Mobile	Lumia640XL	April 2015	N.A.

[1] Data retrieved from: http://www.gsmarena.com/
[2] Data retrieved from: http://www.cnet.com/ (N.A. – Not Available)

	Microsoft Mobile	Lumia650	February 2016	6.8
	Huawei	Mate2	January 2014	7.3
	Xiaomi	Mi4c	September 2015	8.7
	Xiaomi	Mi4i	April 2015	N.A.
	HTC	OneMini2	May 2014	7.3
	Alcatel	OneTouchIdol	May 2013	6
	Alcatel	OneTouchIdol3	June 2015	7.6
	Lenovo	VibeShot	June 2015	N.A.
	Lenovo	VibeZ2Pro	September 2014	N.A.
	Sony Mobile	XperiaC5Ultra	August 2015	N.A.
4	Apple	iPhone 6	September 2014	9
	ZTE	AXONElite	September 2015	N.A.
	HTC	Desire820	November 2014	N.A.
	LG	Gflex2	February 2015	8.3
	Samsung	Note4	October 2014	9
	HTC	OneM8	March 2014	8.7
	Huawei	P8lite	May 2015	6.9
	Samsung	S6	April 2015	8.9
	Sony Mobile	XperiaM5	September 2015	N.A.
	Motorola	XPlay	August 2015	8.4
	Motorola	XStyle	September 2015	7.8
5	Apple	iPhone 6+	September 2014	9
	Apple	iPhone 6S	September 2015	8.9
	Apple	iPhone 6S+	September 2015	9
	Huawei	AscendMate7	October 2014	7.7
	Microsoft Mobile	Lumia950XL	November 2015	7.2
	Huawei	Mate8	November 2015	7.4
	Huawei	MateS	October 2015	7
	Xiaomi	MiNote	January 2015	8
	Huawei	Nexus6P	August 2015	8.4
	HTC	OneA9	November 2015	6.9
	HTC	OneM9	March 2015	8
	Huawei	P8	April 2015	7.9
	BlackBerry	Passport	September 2014	7.3
	Samsung	S6edge	April 2015	9
	Samsung	S6edge+	August 2015	8.8
	Samsung	S7	March 2016	9
	Samsung	S7edge	March 2016	9.1
	LG	V10	October 2015	8.2
	Sony Mobile	XperiaZ5Compact	October 2015	8.7

	Sony Mobile	XperiaZ5Dual	October 2015	7.4
	Sony Mobile	XperiaZ5Premium	November 2015	7.4

Still at the same stage of data preparation, two different features that could not be directly linked to the output variable because they were not quantifiable as an interval were transformed. These are “memberSince” and “dateCollection”, which were converted into “memberDays” – interval between “memberSince” and “dateCollection” – and “diffToToday” – interval between March 20, 2016 (the date when modeling occurred) and dateCollection – respectively. Thus, those two features together with “nameSellers”, which was an identification feature, “moreProdSales”, which was not pertinent since the registered value was always the same (“N”) and “prodType”, which contained an unreasonable number of different categories, in a total of 56 from the 499 records, were removed.

Later on, the need for a new, more efficient, feature arose. It was “priceAvg” and it replaced “minPrice” and “maxPrice” through the computation of the average of both. This happened in order to avoid the creation of redundancies since, for most cases, the values registered were the same. Only 46 out of the 499 cases displayed a difference between both minimum and maximum prices. Column “status” from Table 2 reflects the actions taken for each feature, with only the “approved” being included for the modeling stage. Thus, the 21 different features approved plus the outcome to model (“prodSales”) were considered fully functional for proceeding to the next stage, the actual mining of the data.

3.3. Modeling and knowledge extraction

After gathering all the data with adequate methods, this stage is the pinnacle of a data mining project. It is the phase of discovery enabled by the application of suitable intelligent

methods, which will subsequently allow extracting knowledge. Figure 5 shows in a picture the approach followed for this stage. It comprises two main phases. First, the SVM's capabilities of correctly predicting the number of sales for each smartphone's seller are evaluated through a cross-validation scheme with 10-folds. For assuring even further the robustness of the model built on the data, the 10-fold cross-validation procedure is run twenty times. To evaluate prediction accuracy, three metrics were chosen: MAE, RAE, and NMAE. It should be noted that MAPE was ruled out from this procedure since the dataset contains five records with zero sales, meaning that MAPE cannot be computed for these cases. Furthermore, MAPE distorts the percentage deviation for low values of "prodSales", with this feature ranging from zero to 2,716. Since for each record there are twenty predicted values given the twenty runs of the procedure, the final prediction value for measuring performance is the average of these twenty results.

After assuring that SVM obtains reasonable prediction results, the knowledge extraction phase of the procedure follows. It uses the full dataset to take advantage of the maximum information possible and builds a model based on SVM on top of that data. The validation of fitting the whole dataset is achieved through the three metrics, MAE, RAE and NMAE. Also, to obtain a visual picture of the deviations of the predictions from the real results, a regression scatter plot is drawn.

Finally, the model built on this second phase is used for knowledge extraction through the DSA. DSA takes a sample from the dataset used for training the model and then performs an output sensitivity assessment based on varying the input features through their range of possible values. As a result, it makes possible to assess the influence each feature has on the number of sales. Two types of valuable knowledge are extracted: the percentage relevance that each feature

from the 21 has on the model; and how each of the features affects the number of sales. Such knowledge may provide valuable insights on understanding sellers' performance.

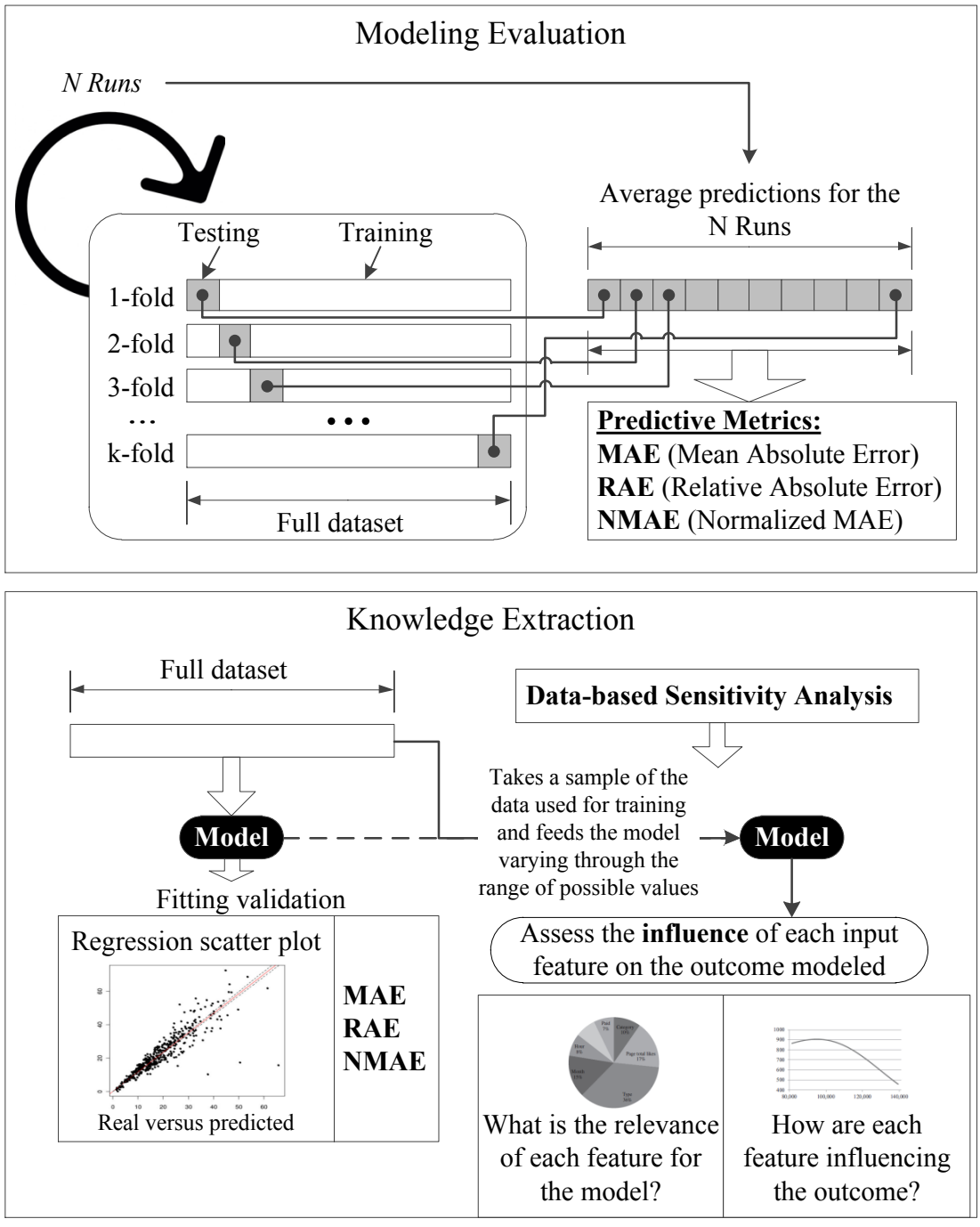

Figure 5 - Scheme with the Modeling Evaluation approach followed.

4. Results and discussion

4.1. Modeling

Ascertaining the adequacy of the model involves computing and gauging the performance metrics identified in the modeling evaluation phase from Figure 5. Accordingly, MAE was 60.84 of absolute difference in sales, whereas RAE was 74.48% and NMAE 2.24%. The substantial discrepancies between the values using different metrics bring out once again the seriousness about choosing the most fitting measure to a given model. NMAE was clearly successful in assessing model adequacy because it is adjusted to the reality of the model, i.e., the range values of the output variable. For instance, as the number of sales increases, it is possible that the difference between R_{max} and R_{min} increases and when using this metric that effect is accounted for without necessarily impairing the results. RAE was not as successful since in models with high dispersion or with a small concentrated cluster of exceptionally high or low values (e.g., 2,716 and zero, for the case addressed), the average does not reflect their broader spectrum. Basically, the more the numerator exceeds the denominator or the closer they are to each other when the numerator is inferior to the denominator, the less the metric will favor model adequacy. Evidently, in models with high dispersion this will lead to adverse results when using RAE. The MAE result is not directly comparable with RAE or NMAE, since this accounts for the absolute difference in the number of sales. However, as with RAE, such a dispersion of values, i.e., large difference between the maximum and minimum possible values does not reflect an accurate metric for evaluating performance as it happens with NMAE, where the breadth of values of the output variable is integrated in the formula.

Also, it becomes clear that this particular model is not suitable for higher number of sales' values due to the scarcity of comparable observations since there were only five

observations with sales over a thousand. Nonetheless, the values obtained using an SVM modeling technique provide evidences of an approximation of the predicted values to the real number of sales, confirming the usefulness of the model for knowledge extraction.

4.2. Knowledge extraction

In order to show how the employment of different metrics allows fitting validation, the charts below were drawn using MAE (Figure 6) and NMAE (Figure 7) as references for residuals on y-axis at against real sales ($True_i$) on x-axis. Both graphics show a linear relation between the corresponding metrics and the deviation to the real number of sales. The MAE was of 49.98, an improved result when compared with model evaluation. This happens because during model evaluation, the dataset was always divided in training and testing datasets, whereas in knowledge extraction, the model was built upon using the whole records from the dataset. Also aligned with this difference, RAE shows improvement at a value of 61.19% along with NMAE at 1.84%. Table 3 summarizes the three metrics for the two phases of the approach (Figure 5).

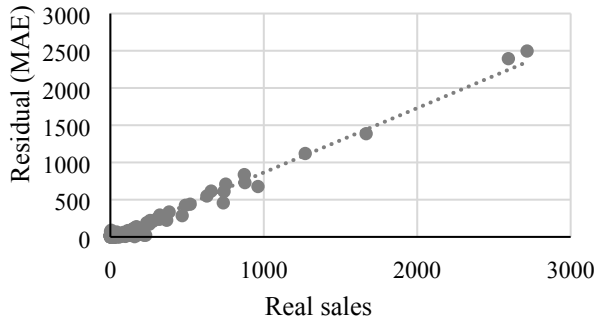

Figure 6 - Regression scatterplot with real sales (x) versus residual with MAE (y).

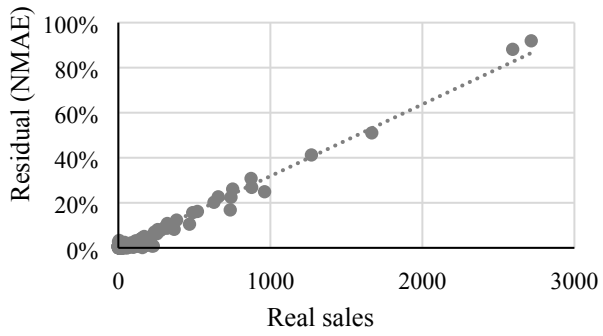

Figure 7 - Regression scatterplot with real sales (x) versus residual with NMAE (y).

Table 4 - Results for the three performance metrics.

	Modeling evaluation	**Knowledge extraction**
MAE	60.84	49.98
RAE	74.48%	61.19%
NMAE	2.24%	1.84%

DSA allowed understanding to which extent the features that fed the SVM algorithm explained the output variable. Figure 8 exhibits a visual picture of features' relevance, while Table 5 shows the percentage values for all the 21 features rounded to the hundredth. Correspondingly, it was discovered that 14 out of the 21 features had a particular influence in the output variable (above 5%), i.e., the number of sales of the smartphone. Their combined contribution to the model is of approximately 91%. The difference between the decomposed contributions of the 14 features was little. The least contributor, the "negR" had an influence of 5% whereas "nrItemsAuction" had an impact of 9%, the highest contribution.

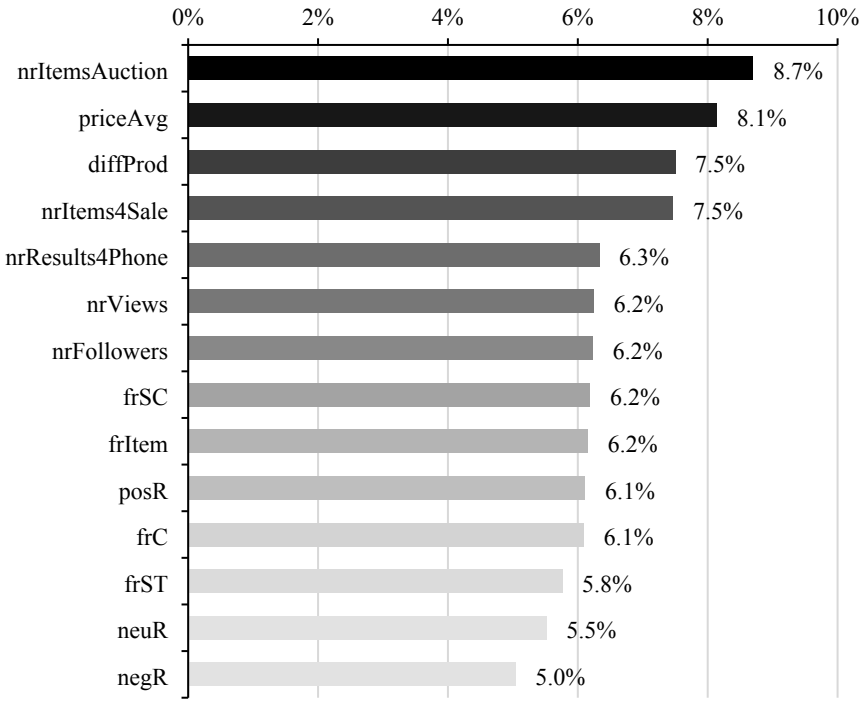

Figure 8 - Features' relevance for modeling sales (shows only values for features with relevance above 5%, rounded to the tenth).

Table 5 - Features' relevance for modeling sales.

Feature	**Relevance**
nrItemsAuction	8.70%
priceAvg	8.14%
diffProd	7.50%
nrItems4Sale	7.47%
nrResults4Phone	6.33%
nrViews	6.25%
nrFollowers	6.23%
frSC	6.19%
frItem	6.16%
posR	6.11%
frC	6.09%
frST	5.76%
neuR	5.52%

negR	5.04%
diffToToday	2.79%
segment	1.34%
country	1.32%
condition	0.98%
brand	0.82%
continent	0.65%
memberDays	0.64%

The most striking observation that both Figure 8 and Table 5 show is that the five most relevant features, comprising around 38% of influence, are all related to the assortment of products the seller offers and its management, i.e. showroom-related, including the average price and the range of different products offered by the seller. Organic reach and engagement through "nrViews" and "nrFollowers" respectively also play a role on the number of sales, even though with far less influence than the combined relevance of the assortment-related features. Such result is aligned with the findings of Moro et al. (2016), which concluded that organic reach and engagement have impact on brand building in social media. Interestingly, the next group of consecutive features in terms of relevance is constituted by seven customer feedback related features, with a combined weight of around 41% of relevance, in a total slightly above the showroom-related related features. This result is a confirmation of previous studies in terms of the influence that customer feedback has on sales (e.g., Kocas and Akkan, 2016). Nevertheless, the top five features remain all product related, relegating individual feedback features for an inferior level.

It was curious that specific product features such as brand and segment along with particular seller features such as country and membership time, barely influenced product sales when compared to the previous ones. The figures mirror the importance of a focused and carefully planned strategy in the background that incentives engagement, promotes reachability

and ensures customer satisfaction that is visible through feedback and ratings to the detriment of more specific seller and product features.

The number of items in auction can be traced back to behavior of bidders striving to get the best deal. As a typical auction website, it is natural that users refine their eBay search looking for items in this category even if they choose another selling format to buy the item. This is reflected in the observed relevance of the input on sales as shown in Figure 9.

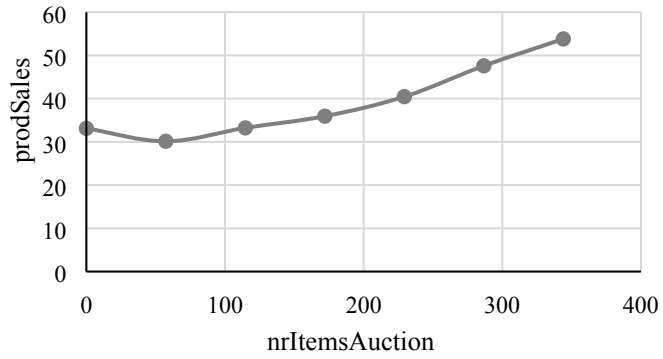

Figure 9 - Impact of number of items in auction on sales.

As expected, “priceAvg” described a fair share of the model as it was the second most significant feature. This happens because price is unquestionably one of the most important marketplace cues (Lichtenstein et al., 1993). Interestingly, it was found that in the particular case of smartphones, product sales plummet until the price level of approximately 1,000 €, after which they start rising (Figure 10). Although it happens many times that product positioning is focused on high-end markets and frequently price is used as a proxy of product quality, as Varian (2014) stated, it also happens often that there is not a direct relation between prices and sales, as sometimes high prices are related to high sales; therefore, the same author argues that continuous experiments with big data and data mining are in demand for obtaining an accurate model. The

data suggests that, regarding smartphones, customers are sensitive to the relationship between quality and price as they are willing to spend more on a smartphone in return for higher quality.

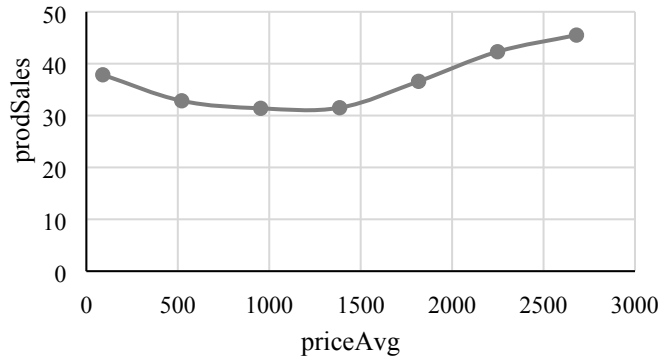

Figure 10 - Impact of average price on sales.

The variety of products sold by a given seller is linked with visibility. The more products are available for sale the more likely it is that views of the seller increase due to inherent exposure. However, if the products fall into many different categories, its assortment might not translate into sales of a particular smartphone. Thus, it makes sense that small groups of different products have less impact on boosting sales than having a larger variety and volume of products (Figure 11).

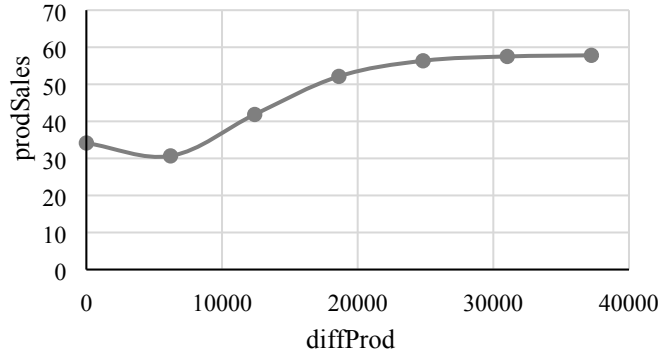

Figure 11 - Impact of assortment on product sales.

The number of items in the “buy it now” listing, “nrItems4Sale”, is also linked with visibility since it is tied with search filtering. Furthermore, if search is merely product-based, results for items on “buy it now” or in auction can both be presented. Shopping with this type of filtering might, however, be related to preferences for convenience and timeliness since instead of waiting for an auction to end or facing the possibility of losing to another bidder, one can simply buy the product straight away while having access to the same type of information. It is interesting to notice though that after 24,820 items, product sales stabilize at around 60 (Figure 12). However, it should be stressed that only two sellers offer more than that number of products, leading to hypothesize that additional data with sellers offering large number of products would be needed in order to confirm the curve drawn on Figure 12 after the threshold.

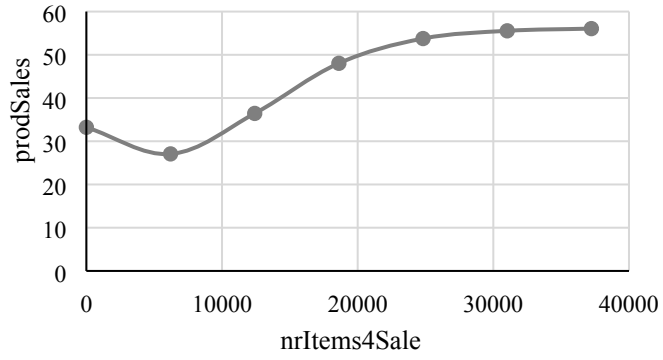

Figure 12 - Impact of number of items in “Buy it now” section on sales.

The influence of the number of results for smartphones shown in Figure 13 reflects the effect of specialization of sellers and therefore their commitment to the category. It is natural that people have more trust in specialized sellers of any category than in generalists, which might, in some cases, sell few smartphones in a multitude of products. There are 219 of the sellers within the dataset where the number of smartphones is more than half of the total number of items for sale, while the remaining 280 have a lesser portion of share of smartphones in their stock.

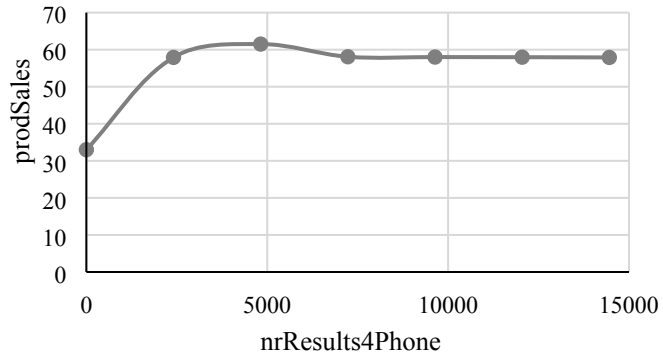

Figure 13 - Impact of specialization on product sales.

The number of views is intrinsically connected with organic reach and it was intriguing that after 186,160 views, product sales continuously dwindled until it reached a plateau (Figure 14). This may be caused by several different factors; for example, as Moro et al. (2016) pointed out for the case of social media, an increase in the number of views may also provoke some degree of erosion of the seller on eBay. Moreover, there is hardly a direct relation between reachability and market penetration, as other features should be accountable, as observed for the case of "nrResults4Phone".

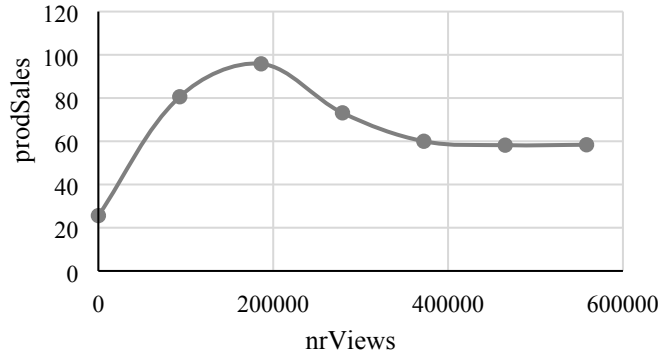

Figure 14 - Impact of number of views on product sales.

Number of followers is linked with reachability and engagement. It is an important source of partial estimates regarding past clients although there might be other reasons for

following a seller, which are not covered within the scope of this study. In this case, after 8,294 followers, product sales suffer a slight decay (Figure 15). This shows that after a certain number of followers, there is not any significant increment to sales.

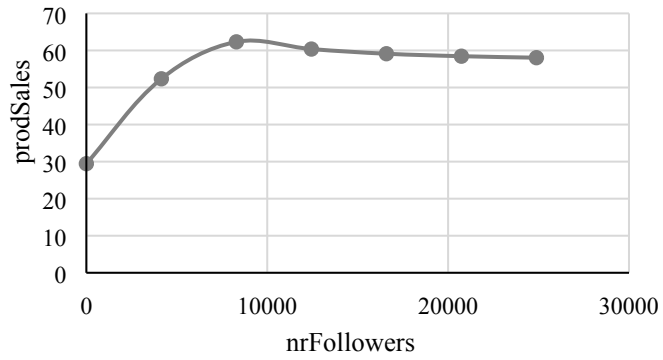

Figure 15 - Impact of number of followers on product sales.

Feedback ratings are often a quite accurate source of varied information about a seller since only after a transaction occurs can the members involved leave a feedback consisting of a short comment and ratings (in the case of the four features considered, only quantitative ratings were included). In this particular case, feedback rating regarding shipping charges was found to be the most significant out of all features within the typology, even though the difference between the most relevant feature ("frSC") and the least relevant, shipping time ("frST"), is just of 0.43% (Table 5). Such figure may be a result of the worldwide nature of eBay, with registered sellers shipping from around the world, raising the sensitivity that customers have to the values of shipping goods. Figure 16 shows that feedback for communication ("frC") and for the items sold ("frItem") have a similar influence on sales, while both shipping features previously mentioned have also a similar influence between each other. The latter group reveals that shipping feedback results in a more immediate impact on sales, even on a lower number of ratings.

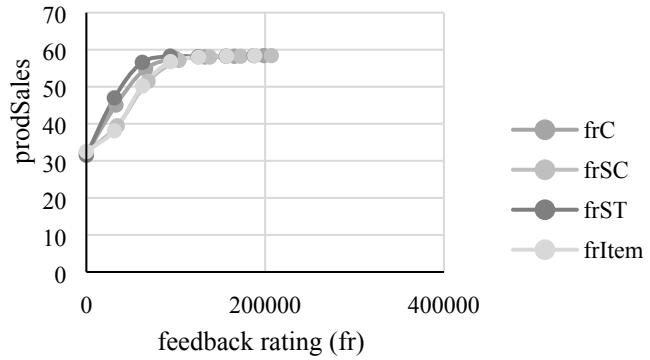

Figure 16 - Impact of feedback rating features on product sales.

The number of positive reviews is another source of valuable insights on successful transactions and it is reasonable to argue that product sales are highly influenced by their value since one can intuitively link a positive review with a positive future response (Hervas-Drane, 2015). It is common that as the number of transactions increases, positive reviews tend to be offset by both neutral and negative reviews, as it was observed throughout the dataset. However, if sellers manage to deliver consistently satisfying products along with the associated service, then product sales are expected to grow (Figure 17).

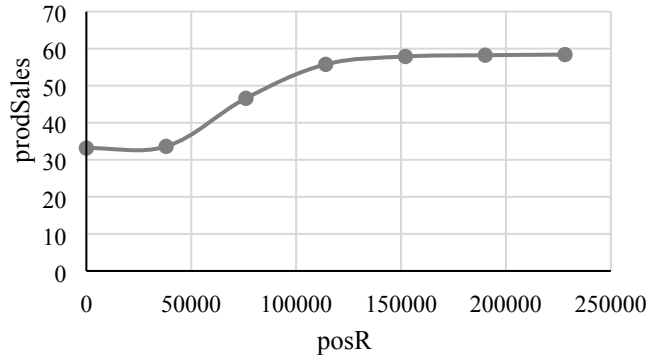

Figure 17 - Impact of the number of positive reviews on product sales.

Within the reviews' typology positive ones have a more gradual impact than neutral reviews and negative, which have higher impact on lower levels of product sales, as it becomes

clearly visible in Figure 18. The results reinforce the idea that for smartphones, product features play a more significant role than customer feedback and, subsequently, reviews that corroborate the initial perception regarding the product contribute more for sales than neutral or negative ones. On another note, it would be interesting to understand on further research the substantial difference in volume of positive reviews and remaining types which yield more comparable volumes.

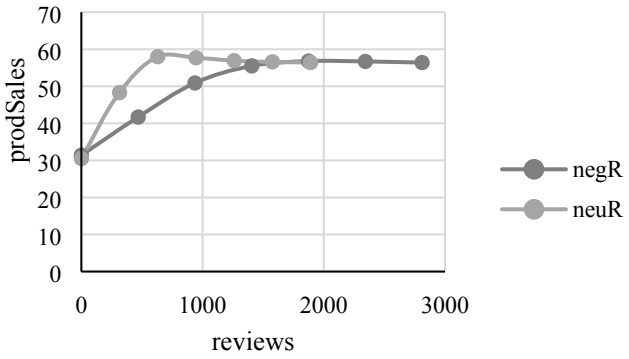

Figure 18 - Impact of the number of negative and neutral reviews on product sales.

4.3. Research findings and managerial implications

Table 6 summarizes the main findings stemming from previous sections and draws relevant managerial implications. The presented implications can serve as guidelines for sellers eager to increase smartphone sales. Visibility is a known key factor of sales success, and in the smartphone online market, this is no exception. Particularly relevant is the interesting relation between the number of products on auctions and the number of sales on the "buy it now" purchase model. This contribution has no precedents in the literature, but its value is reflected in the large number of eBay sellers which opt to choose both business models ("buy it now" and auctions). An also interesting and previous unforeseen behavior is the lower success of mid-range models when compared to both low and high-end models. This seems to be a specifically

smartphone sales behavior, justified by the polarized smartphone demand, with users looking for the cheapest on one hand, or for the most technologically advanced product on the other hand.

Although the case is drawn specifically from an empirical research on smartphone sellers, some of the findings suggest that potentially similar behavior may occur for other products, given the wide dissemination of eBay. Yet, future research on other products is required before any generalization.

Table 6 - Summary of findings.

#	**Research Findings**	**Managerial Implications**
1	The number of items on auction has a positive effect on the number of sales on the "buy it now" format.	Sellers willing to increase their "buy it now" sales can promote their products by also selling them on auctions.
2	Smartphones are products experiencing an effect where the low-end and high-end models are more success in terms of the number of sales, when compared to the middle range models.	Smartphone sellers are more success if they focus on a low or a high-end model.
3	A seller with a higher assortment is more likely to have a higher number of sales of a specific smartphone.	Visibility is linked with having a larger variety of products for sales, which translates into a higher number of sales.
4	Sellers specialized in smartphones tend to have more success than sellers that have a broader offer of other products.	Specialization is key to increase seller's reputation of a specific type of product, increasing its sales.
5	It is important to have many followers as it increases sales; however, having more than 8,000 does not have a significant impact.	The number of followers is linked with reach and engagement. However, the target audience of real prospective buyers is restricted to a certain threshold, as more followers above that do not directly translate into more sales.

5. Conclusions

Online marketplaces are currently one of the most thriving forces in retailing, with a huge impact in the global economy. Consequently, eBay, one of the largest online retailers, has a

worldwide visibility, making of it an adequate choice for sellers to promote their products. Furthermore, eBay provides numerous means for customers and users to submit feedback on products and sellers, information that helps to build sellers' reputation. While eBay allows selling any kind of product, the present study is focused on the sales of smartphones, which are sophisticated communication devices with computer capabilities. The smartphones' market is considered one of the most relevant in the information technology field, which has been growing with each new year since the first iPhone was launched by Apple in 2007.

Given the relevance of online sales and, in particular, of eBay and the smartphones' market, the present study focused on unveiling the features that best characterize the success of smartphones' eBay sellers, measured by the number of sales. In order to succeed in such goal, the approach adopted included a data mining project using support vector machines for modeling and the data-based sensitivity analysis for extracting knowledge in terms of the relevance of the input features used for modeling the number of sales. The contributions and novelty of the present study lie within two dimensions: on the management perspective, the focus on evaluating the features that best identify a successful seller in terms of the number of sales, as opposed to previous studies giving more emphasis on pricing; on the information science perspective, through the compilation of a previously non studied dataset including features related with distinct valences such as product (e.g., brand), reachability and engagement (e.g., number of followers), customer feedback (e.g., number of positive reviews), and seller information (e.g., the country of origin).

Modeling robustness was tested through a 10-fold cross-validation scheme, executed for twenty times. Model performance was evaluated using three performance metrics: the mean absolute error, the relative absolute error, and the normalized mean absolute error. The results

achieved during the model evaluation stage were considered good to proceed with knowledge extraction. Using the data-based sensitivity analysis, it was possible to unveil that the five most relevant features, in a total of around 38%, were all related to product information. The two features that followed in the relevance rank were related to reachability and engagement, namely the number of views and the number of followers. Next appeared seven customer feedback related features, concealing a total of around 41% of relevance, in the eighteenth to the fourteenth position. The discovery that the individual features related to customer feedback are less relevant than those related to the specific product and reachability/engagement is interesting; nevertheless, it should be noted that the total seven features for feedback also play a role, as these conceal as a whole the highest percentage of relevance (41%).

Essentially, sales of smartphones on eBay are mostly influenced by showroom-related features, which reflect the underlying marketing and assortment management strategies ("nrItemsAuction"; "priceAvg"; "diffProd"; "nrItems4Sale"; "nrResults4Phone") and are deeply enhanced by reach ("nrViews"), engagement ("nrFollowers") while being supported by access to several sources of feedback and reviews about the seller.

By further taking advantage of the sensitivity analysis, it was possible to observe how each of the most relevant features affected the number of sales. For example, the most relevant feature, i.e., the number of items the seller has in auction influences the number of sales in a linear proportion, i.e., the more items in auction, the higher the number of sales. Such result may derive from the fact that a seller with a lot of items in auction benefits from customers who do not want to wait for the outcome of an auction and instead choose to buy the item directly from the seller.

Online visibility is vital for a seller's success. This comes in several formats. First, a smartphone seller should invest in having a large number of products on auctions, to attract consumers to its "buy it now" showcase of smartphones. Second, the seller should have a large variety of smartphones for sales, and focus specifically on smartphones, not diverging to different products, which may lead consumers to view the seller as a generalist one as opposed to a specialized one, which reduces its credibility. Finally, a seller should have a large number of followers. Yet, there is a limit for the followers that will effectively translate into sales.

The present study has some limitations that may be addressed in future research. First, it used only eBay data for the experiments. While eBay is one of the largest online retailers, other huge players have risen in the past recent years; most notably, TaoBao, from China. Therefore, in the future, a much larger dataset could be compiled from different sources, namely through the usage of web scrapping tools that can automatically extract the features from the different webpages. Additionally, other features could be devised and tested, in order to enrich the model's knowledge about the number of sales.

References

Amado, A., Cortez, P., Rita, P., Moro, S., 2018. Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis. Eur. Res. Manage. Bus. Econ., 24(1), 1-7.

Armstrong, J.S., Collopy, F., 1992. Error measures for generalizing about forecasting methods: Empirical comparisons. Int. J. Forecasting, 8(1), 69-80.

Berry, T.A., McKeen, T.R., Pugsley, T.S., Dalai, A.K., 2004. Two-dimensional reaction engineering model of the riser section of a fluid catalytic cracking unit. Ind. Eng. Chem. Res., 43(18), 5571-5581.

Bilgihan, A., Kandampully, J., Zhang, T., 2016. Towards a unified customer experience in online shopping environments: Antecedents and outcomes. Int. J. Qual. Serv. Sci., 8(1), 102-119.

Brynjolfsson, E., Hu, Y.J., Smith, M.D., 2006. From niches to riches: Anatomy of the long tail. Sloan Manage. Rev., 47(4), 67-71.

Canito, J., Ramos, P., Moro, S., Rita, P., 2018. Unfolding the relations between companies and technologies under the Big Data umbrella. Comput. Ind., 99, 1-8.

Cao, P., Fan, M., Liu, K., 2015. Optimal dynamic pricing problem considering patient and impatient customers' purchasing behaviour. Int. J. Prod. Res., 53(22), 6719-6735.

Chen, H., Chiang, R.H., Storey, V.C., 2012. Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quart., 36(4), 1165-1188.

Cheung, C. M., Chan, G. W., Limayem, M. (2005). A critical review of online consumer behavior: Empirical research. J. Electron. Commer. Organ., 3(4), 1-19.

Clemes, M. D., Gan, C., Zhang, J., 2014. An empirical analysis of online shopping adoption in Beijing, China. J. Retailing Cons. Ser., 21(3), 364-375.

Clemons, E. K., Wilson, J., Matt, C., Hess, T., Ren, F., Jin, F., Koh, N. S., 2016. Global differences in online shopping behavior: Understanding factors leading to trust. J. Manage. Inform. Syst, 33(4), 1117-1148.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn., 20(3), 273-297.

Cortez, P., 2010. Data mining with neural networks and support vector machines using the R/rminer tool. In Industrial Conference on Data Mining, Springer Berlin Heidelberg, pp. 572-583.

Cortez, P., Embrechts, M.J., 2011. Opening black box data mining models using sensitivity analysis. In Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on. IEEE, pp. 341-348.

Cortez, P., Embrechts, M.J., 2013. Using sensitivity analysis and visualization techniques to open black box data mining models. Inform. Sciences, 225, 1-17.

Coussement, K., Harrigan, P., Benoit, D.F., 2015. Improving direct mail targeting through customer response modeling. Expert Syst. Appl., 42(22), 8403-8412.

Dawson, S., Kim, M., 2009. External and internal trigger cues of impulse buying online. Dir. Mark.: An Int. J., 3(1), 20-34.

Diebold, F.X., Mariano, R.S., 2012. Comparing predictive accuracy. J. Bus. Econ. Stat., 13(3), 253-263.

Dinner, I. M., Van Heerde, H. J., Neslin, S. A., 2014. Driving online and offline sales: The cross-channel effects of traditional, online display, and paid search advertising. J. Marketing Res., 51(5), 527-545.

Einav, L., Levin, J., Popov, I., Sundaresan, N., 2014. Growth, adoption, and use of mobile E-commerce. Am. Econ. Rev., 104(5), 489-494.

Einav, L., Farronato, C., Levin, J., Sundaresan, N., 2018. Auctions versus posted prices in online markets. J. Polit. Econ., 126(1), 178-215.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. From data mining to knowledge discovery in databases. AI Mag., 17(3), 37-54.

Fisher, M.L., 1997. What is the right supply chain for your product? Harvard Bus. Rev.

Floyd, K., Freling, R., Alhoqail, S., Cho, H. Y., Freling, T., 2014. How online product reviews affect retail sales: A meta-analysis. J. Retailing, 90(2), 217-232.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The elements of statistical learning (Vol. 1). Springer, Berlin: Springer series in statistics.

Greenstein-Messica, A., Rokach, L., 2018. Personal Price Aware Multi-Seller Recommender System: Evidence from eBay. Knowl.-Based Syst. DOI: 10.1016/j.knosys.2018.02.026.

Gregg, D.G., Scott, J.E., 2008. A typology of complaints about eBay sellers. Commun. ACM, 51(4), 69-74.

Gregg, D.G., Parthasarathy, M., 2017. Factors affecting the long-term survival of eBay ventures: a longitudinal study. Small Bus. Econ., 49(2), 405-419.

Grewal, D., Janakiraman, R., Kalyanam, K., Kannan, P.K., Ratchford, B., Song, R., Tolerico, S., 2010. Strategic online and offline retail pricing: a review and research agenda. J. Interact. Mark., 24(2), 138-154.

Gunn, S.R., 1998. Support vector machines for classification and regression. ISIS technical report, 14.

Han, J., Kamber, A., Pei, J., 2012. Data mining: Concepts and Techniques. 3rd Edition, Elsevier, USA.

Hanna, M., 2004. Data mining in the e-learning domain. Campus-wide information systems, 21(1), 29-34.

Hearst, M.A., Dumais, S.T., Osman, E., Platt, J., Scholkopf, B., 1998. Support vector machines. IEEE Intell. Syst. App., 13(4), 18-28.

Hemp, P., 2006. Are you ready for e-tailing 2.0. Harvard Bus. Rev. June 2006, 28.

Hervas-Drane, A., 2015. Recommended for you: The effect of word of mouth on sales concentration. Int. J. Res. Mark., 32(2), 207-218.

Huang, Z., Benyoucef, M., 2013. From e-commerce to social commerce: A close look at design features. Electron. Commer. R. A., 12(4), 246-259.

Hui, S.C., Jha, G., 2000. Data mining for customer service support. Inform. Manage., 38(1), 1-13.

Ihaka, R., Gentleman, R., 1996. R: a language for data analysis and graphics. J. Comput. Graph. Stat., 5(3), 299-314.

Jeon, S., Park, S. R., Digman, L. A., 2008. Strategic implications of the open-market paradigm under digital convergence: the case of small business C2C. Serv. Bus., 2(4), 321-334.

Kannan, P.K., 2017. Digital marketing: A framework, review and research agenda. Int. J. Res. Mark., 34, 22–45.

Kao, K.C., Rao Hill, S., Troshani, I., 2017. Online consumers' responses to deal popularity as an extrinsic cue. J. Comput. Inform. Syst., 57(4), 374-384.

Kaplan, A.M., Haenlein, M., 2010. Users of the world, unite! The challenges and opportunities of Social Media. Bus. Horizons, 53(1), 59-68.

Kellerman, A., 2010. Mobile broadband services and the availability of instant access to cyberspace. Environ. Plann. A, 42(12), 2990-3005.

Kocas, C., Akkan, C., 2016. How Trending Status and Online Ratings Affect Prices of Homogeneous Products. Int. J. Electron. Comm., 20(3), 384-407.

Kornberger, M., Pflueger, D., Mouritsen, J., 2017. Evaluative infrastructures: Accounting for platform organization. Account. Org. Soc., 60, 79-95.

Kotler, P., Keller, K., 2012. Marketing Management. 14th Edition, Prentice Hall, USA.

Kusiak, A., Smith, M., 2007. Data mining in design of products and production systems. Annu. Rev. Control, 31(1), 147-156.

Laudon, K.C., Traver, C.G., 2016. E-commerce: business, technology, society. 12th Edition, Pearson.

Lee, C.S., Ho, J.C., Hsu, C.F., 2015. Creating value in global innovation networks: A study of smartphone industry. In Management of Engineering and Technology (PICMET), 2015 Portland International Conference on. IEEE, pp. 755-760.

Lee, S., Choeh, J.Y., 2014. Predicting the helpfulness of online reviews using multilayer perceptron neural networks. Expert Syst. Appl., 41(6), 3041-3046.

Lewis, M.W., 1997. The myth of continents: A critique of metageography. Univ of California Press.

Li, H., Fang, Y., Wang, Y., Lim, K.H., Liang, L., 2015. Are all signals equal? Investigating the differential effects of online signals on the sales performance of e-marketplace sellers. Inform. Technol. Peopl., 28(3), 699-723.

Lichtenstein, D.R., Ridgway, N.M., Netemeyer, R.G., 1993. Price perceptions and consumer shopping behavior: a field study. J. Marketing Res., 30(2), 234-245.

Liu, S., Lu, C., 2015. Cultural tourism O2O business model innovation-case analysis based on CTRIP. In Logistics, Informatics and Service Sciences (LISS), 2015 International Conference on, IEEE, pp. 1-6.

Mathworks, 2016, Retrieved from: http://www.mathworks.com/solutions/machine-learning/index.html?s_tid=gn_loc_drop (26 June 2016).

Moro, S., Cortez, P., Rita, P., 2014. A data-driven approach to predict the success of bank telemarketing. Decis. Support Syst., 62, 22-31.

Moro, S., Cortez, P., Rita, P., 2015a. Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. Expert Syst. Appl., 42(3), 1314-1324.

Moro, S., Cortez, P., Rita, P., 2015b. Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns. Neural Comput. Appl., 26(1), 131-139.

Moro, S., Rita, P., 2018. Brand strategies in social media in hospitality and tourism. Int. J. Contemp. Hosp. M., 30(1), 343-364.

Moro, S., Rita, P., Coelho, J. (2017). Stripping customers' feedback on hotels through data mining: the case of Las Vegas Strip. Tourism Manage. Persp., 23, 41-52.

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

Nisar, T.M., Prabhakar, G. (2017). What factors determine e-satisfaction and consumer spending in e-commerce retailing?. J. Retailing Cons. Ser., 39, 135-144.

Oghazi, P., Karlsson, S., Hellström, D., Hjort, K., 2018. Online purchase return policy leniency and purchase decision: Mediating role of consumer trust. J. Retailing Cons. Ser., 41, 190-200.

Pal, K., Saini, J., 2014. A Study of Current State of Work and Challenges in Mining Big Data. Int. J. Adv. Netw. Appl., Special Issue, 73-76.

Pearce, K.E., Rice, R.E., 2013. Digital divides from access to activities: Comparing mobile and personal computer Internet users. J. Commun., 63(4), 721-744.

Poelker, 2013. Smartphones, big data, storage and you. Retrieved from ComputerWorld: http://www.computerworld.com/article/2473730/smartphones/smartphones--big-data--storage-and-you.html (26th June 2016).

Ramdas, K., 2003. Managing product variety: An integrative review and research directions. Prod. Oper. Manag., 12(1), 79-101.

Refaeilzadeh, P., Tang, L., Liu, H., 2009. Cross-validation. In Encyclopedia of database systems. Springer US, pp. 532-538.

Saltelli, A., Chan, K., Scott, E.M. (Eds.), 2000. Sensitivity analysis (Vol. 1). New York: Wiley.

Schölkopf, S.P., Vapnik, V., Smola, A.J., 1997. Improving the accuracy and speed of support vector machines. Adv. Neur. In., 9, 375-381.

Sharda, R., Delen, D., Turban, E., 2018. Business Intelligence, Analytics and Data Science: A Managerial Perspective. 4th Edition, Pearson Education.

Smola, A.J., Schölkopf, B., 1998. On a kernel-based method for pattern recognition, regression, approximation, and operator inversion. Algorithmica, 22(1-2), 211-231.

Statista. Number of internet users worldwide. Retrieved from: http://www.statista.com/statistics/273018/number-of-internet-users-worldwide/ (24th March 2016).

Tadelis, S., 2016. The economics of reputation and feedback systems in e-commerce marketplaces. IEEE Internet Comput., 20(1), 12-19.

The Guardian. Online shopping on mobiles overtakes desktop for first time. Retrieved from: http://www.theguardian.com/business/2014/sep/02/online-shopping-mobiles-overtakes-desktop (17th March 2016).

Varian, H.R., 2014. Beyond big data. Bus. Econ., 49(1), 27-31.

Venkatesan, R., Mehta, K., Bapna, R., 2006. Understanding the confluence of retailer characteristics, market characteristics and online pricing strategies. Decis. Support Syst., 42(3), 1759-1775.

Wang, M., Lu, Q., Ye, Q., 2016. The Impact of Different types of Online Reviews on Consumer Purchasing Decision-an Empirical Investigation with Online Marketplace Data. In PACIS (p. 300).

Wu, K., Vassileva, J., Noorian, Z., Zhao, Y., 2015. How do you feel when you see a list of prices? The interplay among price dispersion, perceived risk and initial trust in Chinese C2C market. J. Retailing Cons. Ser., 25, 36-46.

Xu, M., Ye, Q., 2015. Reputation and pricing strategies in online market. Proceedings on the Wuhan International Conference on e-Business (WHICEB), pp. 678-684.

Xu, X., Zeng, S., He, Y., 2017. The influence of e-services on customer online purchasing behavior toward remanufactured products. Int. J. Prod. Econ., 187, 113-125.

Ye, Q., Xu, M., Kiang, M., Wu, W., Sun, F., 2013. In-depth analysis of the seller reputation and price premium relationship: A comparison between ebay us and taobao china. J. Electron. Commer. Res., 14(1), 1-10.

Yen, C.H., Lu, H.P., 2008. Factors influencing online auction repurchase intention. Internet Res., 18(1), 7-25.

Yu, X., Liu, Y., Huang, X., An, A., 2012. Mining online reviews for predicting sales performance: A case study in the movie domain. IEEE T. Knowl. Data En., 24(4), 720-734.