PREDICTION USING SOCIAL FEATURES INFLUENCE IN C2C E-COMMERCE

ANGELOS ZOIS

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2110596

COMMITTEE

dr. Giovanni Cassani
PhD student Sara Østergaard

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 24th, 2024

WORD COUNT

8349

ACKNOWLEDGMENTS

I would like to extend my heartfelt gratitude to my supervisor, Dr. Giovanni Cassani, for his invaluable guidance, support, and encouragement throughout the course of my thesis. His expertise and insights have been instrumental in shaping this work, especially as I was writing a thesis for the first time.

I am also deeply thankful to my family for their unwavering support and understanding. Their constant encouragement and belief in my abilities have been a source of immense strength and motivation. This achievement would not have been possible without their love and support.

Finally, thank you to everyone who contributed to this journey, directly or indirectly. Your support has been greatly appreciated.

PREDICTION USING SOCIAL FEATURES INFLUENCE IN C2C E-COMMERCE

ANGELOS ZOIS

Abstract

This thesis explores the role of social features in predicting the number of products sold in a Consumer-to-Consumer (C2C) environment using machine learning techniques. Leveraging a dataset from a French C2C fashion e-commerce, this study employs Random Forest Regressor and XGBoost models to assess the predictive power of social features such as follows, followers, and likes. The methodology includes data preprocessing, the application of the Synthetic Minority Over-sampling Technique for Regression (SMOTER) to address data imbalance, and the use of Shapley values to analyze feature importance.

The results indicate that incorporating social features significantly enhances the predictive performance of tree-based models. Both Random Forest Regressor and XGBoost models show improved accuracy with the inclusion of social variables, with the quality of product decscription ('ProductsPassRate') and the number of followers of the user ('SocialNbFollowers') emerging as the most impactful features. Random Forest Regressor was the best performing model with a Mean Absolute Error of 0.0671. Although the application of SMOTER slightly increased the Mean Absolute Error, it provided a more balanced dataset, enhancing the models' generalizability. Notably, using SMOTER enabled the models to make accurate predictions for users with up to 25 sales compared to when SMOTER was not used.

These findings have substantial societal and business relevance. For society, the results suggest that improving user interfaces to highlight social metrics can boost consumer trust and engagement. For businesses, the insights can optimize marketing strategies and refine predictive models for better sales forecasting.

1 SOURCE/CODE/ETHICS/TECHNOLOGY STATEMENT

Data Source: The data has been acquired from Mabilama, 2019 through Kaggle.com. The obtained data is anonymised. Work on this thesis did not involve collecting data from human participants or animals. The original owner of the data and code used in this thesis retains ownership of the data during and after the completion of this thesis. However, the institution was informed about the use of this data for this thesis and potential research publications. All the figures belong to the author. The thesis code can be accessed through the GitHub repository following the link [1]. In terms of writing, the author used assistance with the language of the paper. A generative language model (ChatGPT, OpenAI, 2024) was used to improve the author's original content, for paraphrasing and Grammarly was used for spell checking and grammar. No other typesetting tools or services were used.

2 INTRODUCTION

This study explores the impact of social factors on predicting product sales in a Consumer-to-Consumer (C2C) setting using machine learning techniques. The investigation aims to enhance our understanding of social media-like characteristics such as comments, likes, and interactions between users. These characteristics could improve the accuracy of machine learning model predictions and uncover potential additional capabilities beyond facilitating social interactions.

2.1 *Consumer-to-Consumer & Social Commerce*

In recent years, the electronic commerce industry (e-Commerce) has been recognised as the principal catalyst behind the growing trend of online shopping sales. E-Commerce, commonly known as EC, involves leveraging the Internet and other related networks, such as intranets, for the buying, selling, delivery, or exchange of information, products, and services (Turban et al., 2017). The rise of e-commerce has primarily been fueled by the adoption of Business-to-Consumer and Consumer-to-Consumer business models, resulting in a surge in online shopping activity (Lone et al., 2023).

As Sukrat et al. (2016) explained, the C2C business model enables individuals to act both as buyers and sellers, offering products or services. This model is intrinsically linked with social networking sites, facilitating sellers in locating potential buyers and promoting their offerings effectively. Typically, an online marketplace builds the necessary platform for buyers and sellers to interact and conduct transactions.

C2C platforms created social features to increase user interaction and ultimately drive conversions. Conversion refers to the desired action taken

[1] https://github.com/angelzois/Thesis-code/blob/main/thesis_code.ipynb

by a user, typically a purchase, as a result of interacting with the platform. This study will focus on C2C platforms designed to facilitate consumers acting as both sellers and buyers. However, businesses can also derive benefits from operating on C2C platforms. For example, businesses could sell surplus stock from previous collections in a fashion marketplace to limit losses. As D'Adamo et al. (2022) discussed, companies and consumers will gain benefits and opportunities, such as accessing new markets for the former and increasing the accessibility of products for the latter.

On the other hand, social commerce encompasses transactions conducted on social platforms, such as Facebook Marketplace (Turban et al., 2017). As a burgeoning field within the digital marketplace, social commerce has attracted significant academic and commercial interest, particularly for its integration of social media and e-commerce dynamics. Also, extensive research has been conducted, notably through the amalgamation of datasets from social media platforms and e-commerce systems. This synthesis provides a unique lens to examine consumer interactions and purchasing behaviors.

An essential aspect of such studies is the analysis of social features, which significantly impact user engagement and transactional outcomes. In this research, social features encompass metrics such as follows, followers, and likes associated with a user within our dataset. The investigation of these features seeks to uncover deeper insights into their direct and indirect effects on commercial success in a social commerce context. These social variables were chosen because they are the most common metrics on social media platforms.

2.2 *Relevance*

This research investigates the influence of social variables on consumer behavior, specifically within the context of online retail environments. The scientific significance of this study lies in its goal to deepen our understanding of social dynamics, such as the correlation between the number of followers and sales, within C2C marketplaces—a domain that has received relatively little empirical attention. The exploration of these dynamics not only enhances theoretical frameworks regarding consumer interactions but also offers tangible insights for e-commerce platforms to enhance user engagement and sales strategies.

From a practical standpoint, the findings of this study are poised to offer significant benefits to industry stakeholders. By integrating machine learning methodologies to analyze the importance of social features, the research aims to uncover the key drivers behind user engagement and conversion rates in online marketplaces. These insights could be pivotal in optimizing business strategies and refining customer experiences, ultimately leading to increased effectiveness in marketing and sales tactics.

Adopting and adapting the methodology proposed by Semerádová and Weinlich (2022), this research shifts the focus from merely predicting consumer intent to quantifying the actual sales outcomes in C2C e-commerce environments. This adjustment aims to bridge the current gap between generic e-commerce analyses and targeted social feature studies by investigating the importance of social features in a crucial aspect of e-commerce analysis, which is sales prediction.

This endeavor is particularly innovative as it represents one of the first attempts to systematically assess the impact of social features on sales within an e-commerce website that sources its data from the same platform. It not only adds to existing knowledge but also paves the way for future research at the intersection of social science and e-commerce.

2.3 *Research Questions*

The objective of this study is to explore the identified scientific gaps, which gives rise to the following research question:

Main RQ: *What is the impact of social features on predicting the number of products sold by a user in C2C through the application of machine learning algorithms?*

This study aims to understand the importance of social characteristics (follows, followers, and likes) within C2C marketplaces, with a particular emphasis on investigating whether the incorporation of these variables enhances the predictive accuracy of machine learning algorithms, compared to scenarios where they are not utilised.

RQ1: *What is the comparative impact of Random Forest and XGBoost on the prediction accuracy of number of sales, measured in terms of Mean Absolute Error, in comparison to Ridge model as the baseline?*

To achieve this objective, a comparative analysis is conducted between the Random Forest Regressor (RFR), XGBoost (XGB), and the Ridge model. The models will be evaluated using the Mean Absolute Error (MAE) with all the features available after preprocessing.

RQ2: *How does the Mean Absolute Error vary based on the incorporation or exclusion of social features?*

The models will undergo training both with and without the social variables, followed by a comparative analysis. This process aims to determine any variations in the models' predictive accuracy and clarify the influence of these features on the predictions.

RQ3: *How does the application of SMOTE for Regression influence model predictions compared to scenarios where it is not utilized?*

The Synthetic Minority Over-sampling Technique for Regression (SMOTER) will be employed to address the imbalance in the dataset, aiming to enhance the generalizability of the predictions. Ensuring the generalizability of the predictions is crucial because otherwise, the predictions could be considered unreliable. Also, to assess the effectiveness of this technique, we will conduct a comparative analysis of the results before and after its application. This step will involve using both sets of features: with and without the social variables. Additionally, we will test SMOTER on each of the models selected for this research.

RQ4: *What is the feature importance, as assessed by Shapley values analysis, for the models incorporating social features?*

A feature importance analysis will be conducted using Shapley values (SHAP) to measure the importance of the variables. This step is crucial in the study as it will determine whether the social features significantly contribute to sales predictions or not.

RQ5: *How does the application of SMOTE for Regression influence the feature importance analysis?*

This research question aims to assess whether the introduction of new synthetic entries alters the feature importance analysis within the model. SMOTER should not affect the importance of variables, as it merely creates new entries based on the existing ones.

3 RELATED WORK

The exploration of social features' influence within marketplaces is a relatively under-researched area to date. Until now, research has predominantly concentrated on the theoretical aspects of customer behavior or on the integration of datasets from diverse sources for conducting studies. In this section, we will discuss related work from three areas: theoretical studies (such as psychological analyses of social features), technical background (data science and machine learning research in this domain), and studies centered on social media and e-commerce.

3.1 *Theoretical background*

Social characteristics have been extensively studied in relation to customer behavior, primarily by psychologists aiming to comprehend the dynamics underlying these traits. Cialdini (2007) discusses in his book that individuals are often influenced by the actions of others largely due to social influence. This phenomenon extends to the realm of social media, where the presence of hundreds of thousands of likes on a post can lead people to assume the content is worthwhile or the person is admirable, often without critical evaluation. In addition, Hoyer et al. (2017) observe that individuals often consult online reviews and opinions to guide their purchasing decisions. Furthermore, the number of followers a user has can significantly influence the shopping intentions of potential buyers, as these followers can set trends (Talib & Saat, 2017). The authors recommend that online businesses leverage this form of social proof to boost sales by transforming their social media accounts into trendsetters.

3.2 *Technical background*

Data science and machine learning are pivotal for companies, enabling them to utilize these technologies to improve business operations and foster innovation. In this context, extensive research has been undertaken specifically in e-commerce and sales forecasting, aiming to refine and advance this sector further.

Lee et al. (2021) evaluated eight different models for predicting online conversion rates, identifying XGB as the most effective. Furthermore, research has shown that machine learning, particularly when utilizing XGB, achieves high accuracy in sales predictions. Moreover, the integration of SHAP with XGB was recommended to improve the interpretability of results. This methodology aids stakeholders in understanding which features most significantly impact predictions, thus enabling more informed decision-making. SHAP has also been employed in Y. Chen et al. (2023) to determine critical variables in purchase behavior, with the goal of assessing various models' predictive and explanatory efficiency. One of the key findings was the versatility of the RFR, which can be used both for prediction and interpretation, especially when combined with SHAP, due to its exceptional performance.

Additionally, the RFR is renowned for its capability to manage nonlinear relationships and its robustness against overfitting, making it highly valuable for large datasets (Géron, 2022; Müller & Guido, 2016). This durability makes RFR an appealing option for complex predictive tasks across various business contexts. Moreover, Bajaj et al. (2020) evaluated numerous machine learning algorithms to predict sales for Big Mart Companies, finding that the RFR was the best-performing model for these types of predictions. This consistency in findings underscores the efficacy of RFR in commercial applications where predictive accuracy and reliability are paramount.

3.3 *Social Media & e-Commerce*

Thus far, research has primarily concentrated on utilizing datasets from both social media and e-commerce websites, frequently attempting to integrate them to create comprehensive datasets necessary for detailed analysis. Zhang and Pennacchiotti (2013) employed a combination of datasets from Facebook and eBay to analyze consumer purchasing behaviors, identifying significant correlations between specific features and purchasing patterns. While merging datasets from different sources is a common practice, using data from a single source offers the advantage of consistency in data collection, thereby enhancing the reliability of the findings. Merging disparate datasets can compromise the overall quality of the resulting dataset due to potential discrepancies in the motivations behind data collection.

E-commerce businesses can capitalize on the relatively new technologies of machine learning to enhance their online platforms (Iqbal, 2022). Applications such as trend analysis, recommendation systems, and chatbots are just a few examples of how machine learning can be applied. Furthermore, Iqbal (2022) suggests that leveraging freely available online user data—such as posts, comments, and reactions—and applying machine learning techniques can provide valuable insights for businesses. This enables them to make more informed decisions and tailor their offerings to meet consumer demands effectively.

Online social networking has shifted business strategies from being product-centric to customer-centric (Attar et al., 2022). The authors note that by understanding social dynamics, businesses and sellers can maximize their profits. This approach highlights the importance of integrating social media insights into business strategies to enhance customer engagement and profitability in the modern digital economy.

The hypothesis of this research could be validated by synthesising insights from the papers mentioned above, as each contributes significantly to the foundation of this thesis.

4 METHODS

The methodology pipeline implemented in this research is depicted in Figure 1. The subsequent section elaborates on each step outlined in the diagram and also discusses the models and algorithms that are utilised.

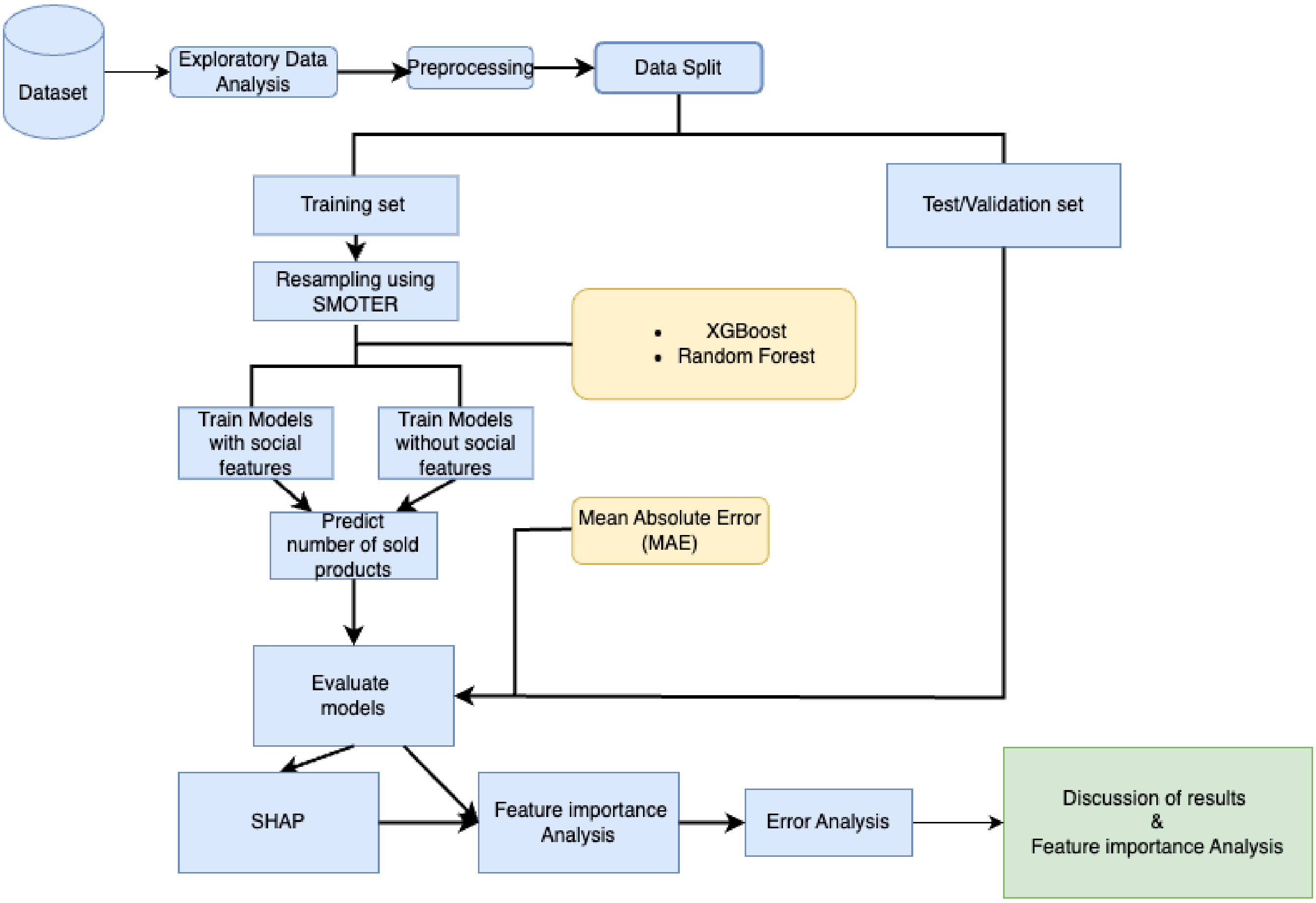
Dataset
Exploratory Data Analysis
Preprocessing
Data Split
Training set
Test/Validation set
Resampling using SMOTER
XGBoost
Random Forest
Train Models with social features
Train Models without social features
Predict number of sold products
Mean Absolute Error (MAE)
Evaluate models
SHAP
Feature importance Analysis
Error Analysis
Discussion of results & Feature importance Analysis

Figure 1: Workflow of the methodology pipeline.

4.1 *Dataset Description*

The dataset used in this research was obtained from a French C2C online fashion application and is available on Kaggle under the CC BY 4.0 DEED license. It consists of 98,913 entries and 24 attributes in CSV format (Mabilama, 2019). The attributes are categorized into demographics, social features, and sales metrics. It is important to note that the dataset is limited to registered users which indicates that the dataset does not contain missing values. Given the scarcity of social features in e-commerce datasets, this particular dataset provides a unique opportunity to investigate their impact on sales performance. A detailed overview of the dataset variables is given in Appendix A (page 31).

4.2 *Exploratory Data Analysis*

The application of Exploratory Data Analysis (EDA) helps to better understand the dataset and identify significant anomalies. In e-commerce, it is well-known that most users browse without any intention of making a purchase. Consequently, datasets related to e-commerce are significantly imbalanced when it comes to sales. This is the case with the dataset used in this study, as the target variable indicating the number of products sold predominantly consists of entries recording zero sales. Specifically, the dataset used for evaluating the impact of social features is heavily skewed, with over 90% of the entries recording zero sales.

Additionally, attributes deemed irrelevant to the study, such as those measuring seniority in days, months, and years, have been omitted to avoid redundancy. Furthermore, the presence of high correlation among some features could pose issues for regression analysis due to multicollinearity, rendering these features redundant. This understanding is crucial for refining the dataset and selecting appropriate modeling techniques.

Lastly, Figure 2 represents the correlation matrix of the social features used in this research.

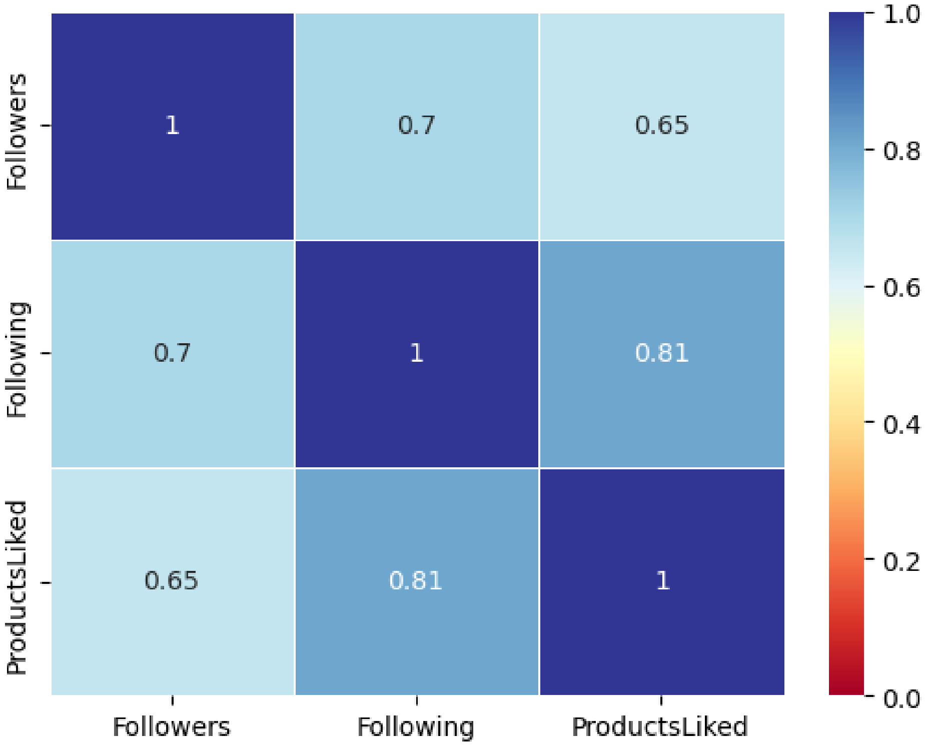

Figure 2: Correlation matrix of the social features

From Figure 2, it is clear that the features are highly correlated and this could introduce the problem of multicolinearity, affecting negatively the performance of the models.

4.3 *Preprocessing*

The dataset comprises both categorical and continuous variables, making preprocessing a critical step to ensure the effectiveness of the models. Through multiple iterations of running the analysis, supplemented by SHAP feature importance assessment and EDA, several variables were deemed irrelevant and subsequently removed from further consideration. Specifically, the variable describing the type of customers was excluded as it held a single value across all entries. Similarly, various representations of seniority were consolidated, retaining only the seniority calculated in days. The variables representing the country and country code were also removed following initial SHAP analysis which indicated their limited impact on the model's performance.
Additionally, the feature expressing if a user has any application was considered redundant, owing to the existence of two specific attributes, a

user has either Ios or Android application, which provides more direct information about the user's software environment. This decision was further supported by the fact that all data entries are from registered users who are required to download the application to register. Therefore, the feature indicating whether customers have any application is irrelevant for distinguishing user characteristics. These decisions in data preprocessing are aimed at streamlining the input feature set to enhance model accuracy. Ultimately, it was crucial to transform the categorical features into a format compatible with the models. This was achieved by employing the 'get_dummies' function from the Pandas library, which enabled the models to access the data effectively. In this preprocessing phase, all features subjected to the 'get_dummies' transformation are nominal.

4.4 *Data split*

Following the preprocessing, the dataset needs to be split into training, validation, and test sets. The training set comprises 60% of the dataset, while the validation and test sets each constitute 20%. To accomplish the final split, the original dataset should first be divided into training and test sets, and then the validation set should be created from the test set. The test set was isolated and only used for model comparison. Additionally, the dataset is split at random with the random state set to 42 to ensure the reproducibility of the results.

4.5 *Resampling with Smote for regression*

In this section, a comprehensive explanation of the SMOTER will be provided, drawing extensively on the seminal work by Torgo et al. (2013).

4.5.1 *SMOTER foundation*

SMOTER adapts the principles of its original classification-focused counterpart to address imbalances found in regression datasets. Such imbalances are often manifested as skewed distributions of continuous target variables, which can lead models to preferentially predict more frequently occurring outcomes. SMOTER addresses this issue by generating synthetic samples to enhance the representation of underrepresented target values within the dataset.

In this research, the application of SMOTER involves identifying the 5 nearest neighbors for each data point within the less represented areas of the target variable space, specifically the non-zero values. Synthetic data points are subsequently created through interpolation between these neighbors. This process not only pertains to the feature values but also extends to the target values, thereby aiding in the normalization of the target variable distribution across the dataset. Importantly, the generation of synthetic data points is confined to the training set to prevent data leakage.

Building upon the technique developed, an iterative approach proves essential. Here, synthetic data points are continually produced until either a predefined balance between non-zero and zero outcomes is achieved or a maximum number of iterations is reached. This strategy ensures adequate representation of the minority values, with the primary aim being to reduce model bias and enhance the accuracy of predictions for these less frequent, yet relevant, outcomes.

By incorporating SMOTER code modifications, the dataset is not merely increased in size; it also achieves a better representation of sales, particularly up to 25. This allows for the development of models that are effective across a broader range of the target variable. Consequently, this enhancement leads to insights that are both accurate and highly applicable to real-world situations.

4.5.2 *Addressing Data Imbalance with SmoteR*

To address the significant imbalance mentioned in subsection 4.2 and improve the predictive modelling process, the SMOTER technique, as detailed in subsection 4.5.1, is employed on the training set. The adjustments in distribution resulting from this method are illustrated in Figure 3.

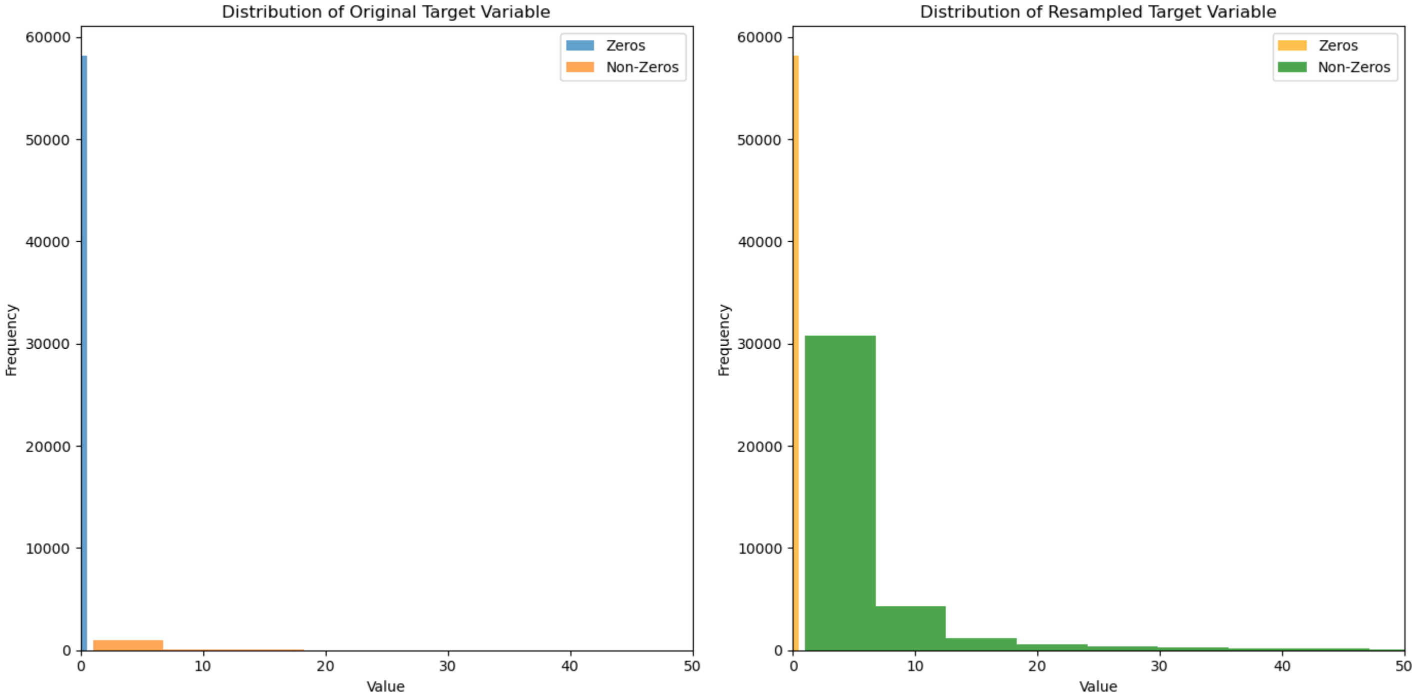

Figure 3: Distribution of the target variable before and after resampling with SMOTER. The left plot shows the distribution of the original target variable, with a high frequency of zero values (blue) and a minimal number of non-zero values (orange). The right plot displays the distribution after applying SMOTER, which resampled the target variable to increase the representation of non-zero values (green).

The distribution of the updated dataset with the synthetic entries demonstrates improved balance, making it more suitable for sales predictions. With the application of SMOTER, the non-zero ratio of the target variable increased to 0.4 due to the creation of more synthetic entries, mainly in the range of 1-25 sales. SMOTER was unable to generate synthetic entries for higher values because there were insufficient instances of these values to create unbiased entries.

4.6 *Models*

The selection of models for this study was guided by Section 3 of related work, which directed the choice of models used in the research. The primary aim of this research is not to optimise or develop new algorithms but rather to investigate the significance of social features. Therefore, models that have shown high performance in previous studies are deemed adequate for this research. After describing the models, we will explain the training and evaluation processes chosen for this study.

4.6.1 *Baseline Model*

For the baseline model, Ridge regression was chosen due to its simplicity, which aids in comparing its performance against more complex models. Ridge regression is particularly advantageous for handling multicollinearity among predictors, a common scenario in datasets with numerous features. By imposing a penalty on the size of coefficients, it helps reduce model complexity and prevent overfitting. As illustrated in Figure 2, the social features are highly correlated, which is a challenge that the Ridge model is well-suited to address in order to achieve good predictions.

4.6.2 *Random Forest & Extreme Gradient Boost Regressor*

RFR and XGB models were chosen for their ability to predict the number of products sold by users. Both models are recognized for their high accuracy and proficiency in modeling non-linear relationships within complex data structures. The use of bootstrapping generates more diverse trees, making Random Forest a robust algorithm (Breiman, 2001). RFR operates similarly to Random Forest in classification tasks, following the same steps to produce outcomes. It generates a result for each tree and then averages all tree results for the regression task (Breiman, 2001).

Likewise, XGB is an ensemble technique utilizing gradient boosting frameworks and it is well-known for its speed and performance, especially in structured or tabular datasets. This algorithm works by sequentially reducing the error of previous predictions until it achieves the final result (T. Chen & Guestrin, 2016).

4.6.3 *Training of the models*

To optimize these models, it is essential to perform hyperparameter tuning on RFR and XGB. The goal is to fine-tune the models by exploring a specified parameter grid. This method aids in identifying the optimal model settings and ensures robust results. The findings will be validated on the validation set to identify the best-performing hyperparameters for each model.

Table 1 outlines the range of hyperparameters tested for the RFR model. To determine the best combination, RFR was evaluated using every possible combination of these hyperparameters. Hyperparameters are vital for tuning RFR, as choosing the most suitable ones can significantly improve the algorithm's performance (Géron, 2022). In Table 1, 'max_depth' acts as a regularization parameter in RFR, with a default value of 'None'.

'n_estimators' denotes the number of trees used to construct the RFR, while 'min_samples_split' specifies the minimum number of samples required for a branch to split.

Hyperparameters	Range
max_depth	None, 5, 10, **20**,
N_estimators	50, 100, **200**, 300
min_samples_split	2, 5, 7, **10**

Table 1: Search Range for Random Forest Regressor

Similar to RFR, XGB is a tree-based algorithm and consequently shares similar hyperparameters, as shown in Table 2. XGB includes parameters like the number of estimators and maximum depth, with the addition of the learning rate. The 'learning_rate' parameter indicates the step size for the model training process; a smaller step size results in a more robust model but requires more trees.

Hyperparameters	Range
max_depth	3, 6, **9**, 12
N_estimators	50, 100, 200, **300**
learning_rate	0.1, **0.2**, 0.3, 0.4

Table 2: Search Range for XGBoost Regressor

The model selection and optimisation process is designed to robustly evaluate the influence of social features within the dataset, thereby allowing for precise and reliable insights into how these features affect product sales in the C2C e-commerce environment. Lastly, all models, including the baseline model, will be trained with and without the social variables to evaluate their impact on the MAE.

4.6.4 *Evaluation metric*

To evaluate the models' predictions, we utilize the MAE. MAE is particularly suitable for regression problems that may include outliers, as it remains unaffected by extreme values (Farnham et al., 2019; Müller & Guido, 2016; Pedregosa et al., 2011). Furthermore, MAE is a widely accepted metric in this field and complements the methodology effectively, providing measure of prediction accuracy without being skewed by anomalies. Finally, the MAE from all the models will serve as the comparative metric, used both when including and excluding social attributes. This comparison will help determine the influence of social attributes on the prediction accuracy.

4.7 *Shapley values*

The use of SHAP is crucial for understanding the importance of features, particularly the social variables, within the dataset. The feature importance analysis will be conducted on models that either include or exclude the social variables, depending on the performance outcomes of these models. This approach allows for a targeted analysis that assesses the impact of social variables by comparing their presence against their absence in the predictive models.

4.7.1 *Theoretical Foundations*

SHAP is based on additive feature attribution methods derived from cooperative game theory. It assesses the significance of each feature by examining the impact of a feature's presence or absence on the model's prediction. Crafted to align with principles like consistency and local accuracy, these attributions strive to offer a dependable representation of a feature's impact across different models (Lundberg & Lee, 2017).

The original research on SHAP used linear models to implement the explainable artificial intelligence technique but for the purposes of this research we will use the "TreeExplainer". The "TreeExplainer" is used to explain the outcomes of ensemble tree models like RFand XGB. The goal of the tree explanation is to combine many local explanations from the trees to create the global explanation (Lundberg et al., 2020). It is specially designed for tree-based models, leveraging the tree structure to efficiently calculate precise SHAP values. In contrast to the approximations needed for other types of models, TreeExplainer makes use of the decision tree paths, greatly decreasing computation time while preserving precision (Lundberg et al., 2019).

4.7.2 *Visualisation Tools*

TreeExplainer utilises a range of visualisation tools that assist in interpreting the model's decisions. For instance, SHAP summary plots offer a comprehensive perspective on the influence of features, whereas dependence plots reveal how features interact with each other (Lundberg et al., 2019).

4.7.3 *Integration in Current Research*

In this study, TreeExplainer will be employed to analyze the impact of social features on tree-based models such as RFR and XGB. Through its comprehensive and precise explanations of feature contributions, TreeExplainer facilitates a detailed comprehension of how social factors influence the predictive results within the dataset.

To better understand the importance of social features, we will use the summary plot and the dependency plot. First, the summary plot will be generated using the test set to create the Shapley values. This plot will illustrate the significance of each feature. Then, we will use the dependency plot, which also uses the Shapley values, to examine the relationships between specific features and other variables in the dataset. In our case, we will focus on the features showing the number of followers, follows, and likes a user have to measure their dependency levels with other features. These plots are a good representation of the importance of the social variables.

4.8 *Error Analysis*

The evaluation process involved both validation and test set assessments. The validation set was used to determine the best hyperparameters for the models. The dataset was partitioned, and the models were trained on the training set and validated on the validation set. This process ensured robustness, and the MAE on the validation set was calculated. Following hyperparameter tuning using the validation set, the models were tested on a separate test set to evaluate their performance on unseen data, with the test set MAE recorded.

A comparative analysis was conducted to compare the performance of Ridge Regression, RFR, and XGB models. This analysis highlighted the strengths and weaknesses of each model based on their MAE and residual distributions.

Additionally, a plot comparing actual values to predicted values was generated to better understand the predictions of the best performing model. This graph includes a red line representing the ideal predictions of the model and dots based on the model's predictions and actual values. The closer these points are to the red line, the lower the model's error.

Lastly, the impact of applying SMOTER was evaluated. Models trained with and without SMOTER were compared to determine its effect on MAE and the generalizability of predictions, particularly for less frequent outcomes.

5 RESULTS

5.1 *Model Performance & Evaluation*

The results of this research regarding the MAE on the test set are presented in Table 3. Three predictive models were evaluated, incorporating social features: Ridge Regression, Random Forest, and XGBoost.

Models	Mean Absolute Error
Ridge	0.4336
Random Forest Regressor	**0.0671**
XGBoost	0.0677

Table 3: Mean Absolute Error of Ridge, Random Forest Regressor, XGBoost models on the test set.

In this study, both tree-based models outperformed the baseline model in every scenario tested. Ridge model struggled to match the accuracy of the other algorithms in predicting the number of products sold by a user.

The RFR is the top-performing model in this research across all scenarios. It achieved low MAE scores on the test sets, indicating consistency and suggesting that it neither overfits nor underfits the data.

The XGB model also performed well, nearly matching the performance of the RFR. This suggests that both tree-based models are effective for predicting the number of sales in e-commerce. However, since the RFR achieved the lowest MAE, the feature importance and error analysis will be conducted on that algorithm.

5.2 *Influence of social features on Mean Absolute Error*

After identifying the best performing model, it is essential to assess the impact of the social features on MAE. Table 4 illustrates the variation in MAE when social features are included.

Social Features	Models	Mean Absolute Error
Without	Ridge	0.3826
	Random Forest Regressor	0.0856
	XGBoost	0.0937
With	Ridge	0.4336
	Random Forest Regressor	0.0671
	XGBoost	0.0677

Table 4: MAE by model and social feature inclusion for Ridge, Random Forest, and XGBoost models.The row "Without" represents the MAE without social fetures, and the row "With" represents the MAE with social features included.

From Table 4, it is evident that not all models benefit from the inclusion of social features. In particular, the Ridge model shows a significant increase in MAE when social features are added. Figure 2 indicates that the number of followers and the number of likes are highly correlated, introducing multicollinearity. Although Ridge regression can handle multicollinearity, the alpha value assigned to the model was insufficient to adequately penalize the correlation between these features. Consequently, the MAE is higher when these social features are included compared to when they are excluded.

Another interesting observation is that the MAE is lower for the tree-based models when social features are incorporated. This indicates that RFR and XGBoost can effectively leverage the additional information embedded in the social features. Specifically, RFR significantly reduced the MAE from 0.0856 to 0.0671 with the inclusion of social features. Similarly, XGBoost decreased the MAE from 0.0937 to 0.0677 by utilizing the social variables. These results highlight the importance of social features in tree-based models for predicting the number of products sold by a user.

5.3 *Influence of SMOTER on Mean Absolute Error*

After comparing the MAE for all the models with and without the social features, the next step is to assess the impact of SMOTER on the MAE. Figure 4 illustrates the MAE across all scenarios tested, both with and without the use of resampled data.

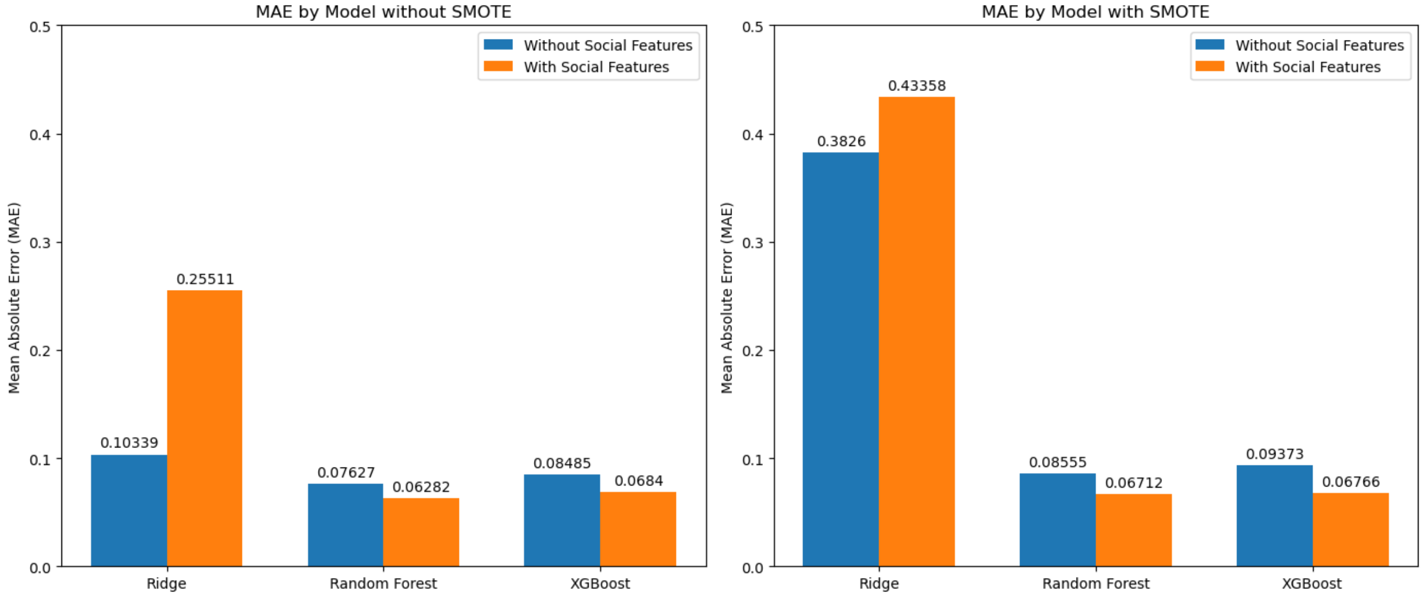

Figure 4: MAE by model and social feature inclusion for Ridge, Random Forest, and XGBoost models. The left plot shows the MAE without using SMOTER, while the right plot shows the MAE with SMOTER applied. Blue bars represent the MAE without social features, and orange bars represent the MAE with social features included.

In this phase of the research, each model utilized the synthetic data from SMOTER differently compared to the others. Starting with Ridge regression, it exhibited a significant increase in MAE when applied to the resampled dataset. Specifically, the MAE increased in both scenarios (with and without social features), indicating that the new synthetic entries had a negative effect on this model.

In contrast, the RFR showed consistent predictions in both cases. Although there was a slight increase in MAE when using the synthetic data, RFR maintained relatively robust performance. It is noteworthy that RFR was the best performing model in both scenarios.

Lastly, the XGBoost model effectively leveraged the new information from the synthetic entries, as evidenced by a slight reduction in MAE (from 0.0684 to 0.0677) when social features were included. In the scenario without social features, the algorithm exhibited an increase in MAE (from 0.08485 to 0.09373), but still maintained good performance.

5.4 *Feature Importance Analysis*

The feature importance analysis conducted using SHAP values has revealed significant findings, as shown in Figure 5 for the RFR, the best performing model. Additionally, Figure 6 displays the dependence plots of the social features, highlighting crucial patterns that are valuable for this research. Finally, the impact of data resampled using SMOTER will be discussed to address the final research question.

5.4.1 *Feature Importance Summary*

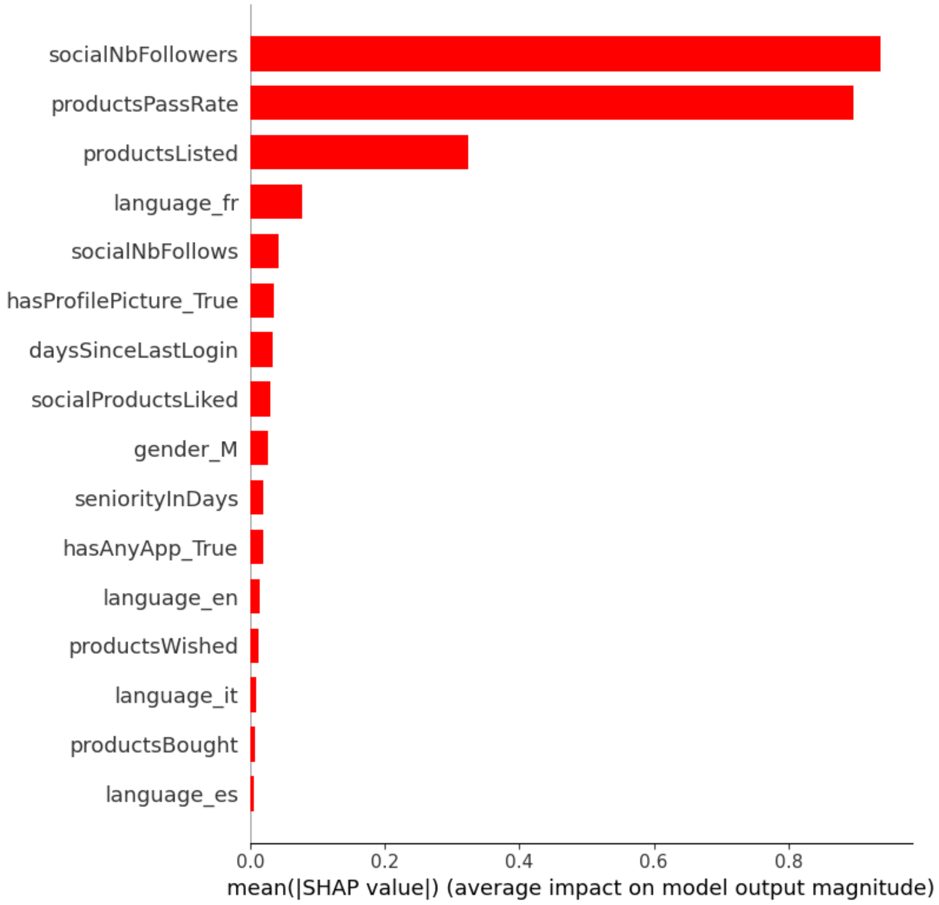

Figure 5: SHAP feature importance results for the Random Forest Regressor model.

The findings indicate that the description provided by the user and verified by marketplace employees ("ProductsPassRate") and the number of followers a user has ("SocialNbFollowers") are the most influential features in the decision-making process of the RFR. These two factors consistently exhibit a high impact, underscoring their essential role in determining sales outcomes. Following these, the number of products listed by a user ("productsListed") and the feature representing people speaking French ("language_fr") are also highly informative and valuable features for the model.

5.4.2 *Features Dependency*

After generating the summary plots, it was essential to gain a deeper understanding of the interactions between the social features and the other variables. To achieve this, dependence plots were utilized, and the results are presented in Figure 6.

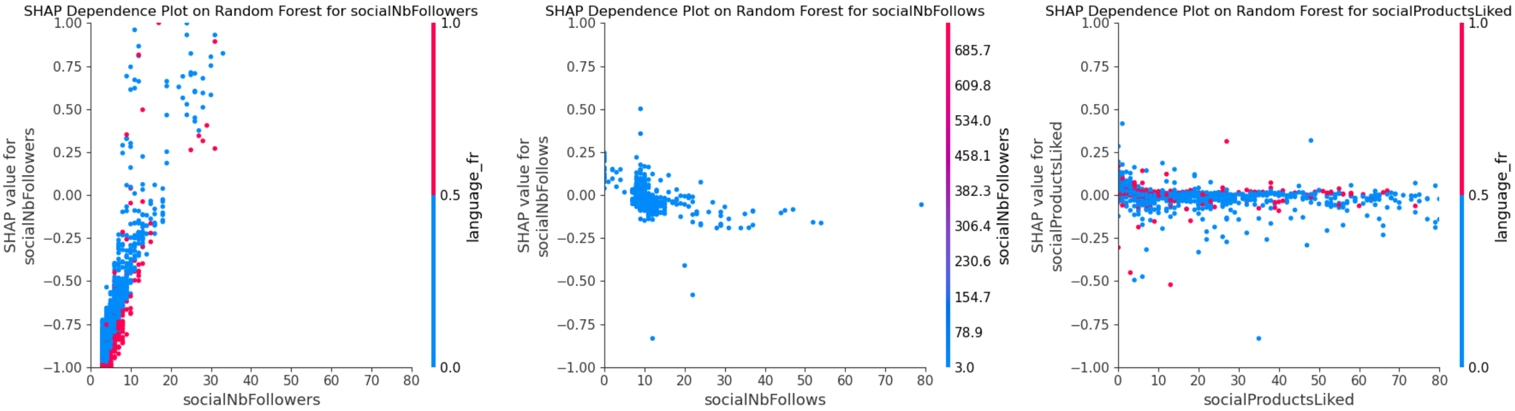

Figure 6: SHAP feature dependency plot illustrating the interaction with other features. The left column evaluates 'socialNbFollowers', the center column examines 'socialNbFollows', and the right column assesses 'socialProductsLiked'.

Figure 6 provides more detailed insights into the interactions between social features and other variables. Specifically, it illustrates the dependency plots where the x-axis represents the values of the tested feature, the y-axis (left side of each graph) displays the Shapley values for the feature. The color gradient indicates the values of the dependent feature. For instance, in the central graph of Figure 6, the different colors correspond to various values of the number of followers.

The number of followers ("socialNbFollowers") demonstrates a consistent influence on predicted outcomes for the RFR model across the various values shown on the x-axis. As the number of followers increases (x-axis), the importance of the feature (y-axis) also rises. The color gradient interaction with the "language_fr" feature indicates that the number of followers consistently affects model predictions, regardless of whether the user speaks French.

For the number of follows ("socialNbFollows"), most Shapley values are close to zero, indicating a negligible influence on the outcome. The interaction with the number of followers shows that the feature's impact remains consistent, likely because there are not many instances of a high number of followers in our dataset.

Last but not least, the number of products liked ("socialNbLiked") feature shows significant variation in its impact on the target variable. Similar to the first social feature discussed, its interaction with the French language ("language_fr") reveals that the feature's impact remains consistent regardless of the language spoken.

5.4.3 *Influence of SMOTER on feature importance analysis*

To evaluate the impact of SMOTER on the feature importance summary plot, two different figures were generated: one using the resampled data created by SMOTER and another without resampling. In both cases, the graphs were identical to Figure 5. This indicates that SMOTER did not influence the feature importance summary plot as the order and the importance of the features did not change.

5.5 *Error Analysis*

The final subsection details the error analysis of RFR when using data generated by SMOTER compared to when it was not utilized. Figure 7 presents the actual versus predicted values of the RFR in both scenarios. These graphs are essential tools as they visually represent how well the model's predictions align with the actual observed values.

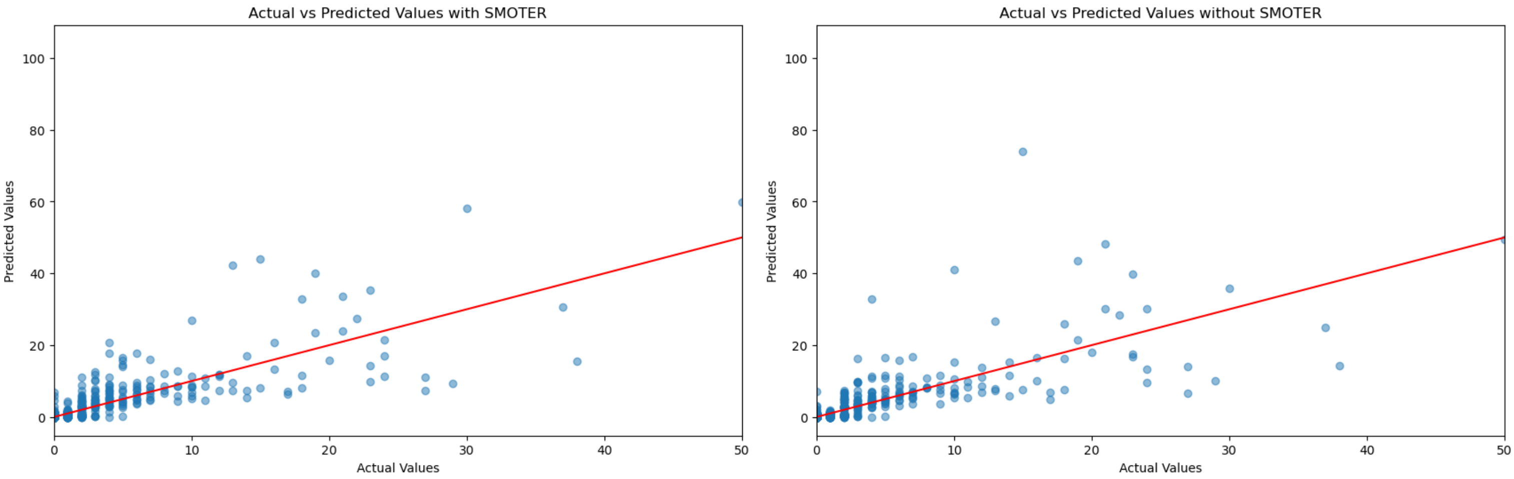

Figure 7: Comparison of actual versus predicted values for Random Forest with and without SMOTER. The left graph shows the actual versus the predicted values using the resampled data generated by SMOTER, and the right column shows the same without using SMOTER.

Figure 7 highlights two key patterns. The first pertains to the use of SMOTER. By comparing the two graphs, it is evident that when the generated data was utilized, the RFR was able to reduce the error, as the points are closer to the red line, indicating that the model's predictions are more closely aligned with the actual values. The second pattern focuses on the lower predicted values. When SMOTER is used, there appears to be more instances of lower predictions and more instances with error compared to when SMOTER is not used. This suggests that the model tends to make more errors in these lower value cases when SMOTER is applied, justifying the small increase in MAE in the RFR.

6 DISCUSSION

To date, research on the impact of social variables and their insights into sales dynamics has been somewhat limited. Therefore, the primary objective of this study was to enhance our understanding of how these features influence the prediction of the number of products sold.

This section offers a detailed discussion of the results presented in Section 5, demonstrating how the findings addressed each research question. Additionally, it explains how this research aligns with existing related work. Finally, the societal and business benefits will be explored, the study's limitations will be addressed, and potential directions for future research will be suggested.

6.1 *Interpreting Results*

RQ1: *What is the comparative impact of Random Forest and XGBoost on the prediction accuracy of number of sales, measured in terms of Mean Absolute Error, in comparison to Ridge model as the baseline?*

To address this research question, the three models were tested, and the results are shown in Table 3. The best performing model was RFR, with a score of **0.0671**. Notably, both tree-based models outperformed the baseline model. In response to the research question, the tree-based models proved to be superior to the baseline model in this study, with RFR outperforming XGB. Additionally, the good performance of RFR supports the findings of Bajaj et al. (2020), as this model was the best for sales prediction in both studies. Consequently, the feature importance analysis was conducted exclusively on RFR. The robust performance of RFR also reinforces the findings of Y. Chen et al. (2023), as RFR excelled in prediction and facilitated the interpretation of results through feature importance analysis.

RQ2: *How does the Mean Absolute Error vary based on the incorporation or exclusion of social features?*

Regarding research question 2, the models were tested with and without the social features, with the corresponding scores presented in Table 4. It is clear that the tree-based models benefited from the information embedded in the social features, as evidenced by the decreased MAE. Conversely, the Ridge model did not gain any value from these variables, with its MAE increasing. This suggests that the RFR model was better at handling the non-linear relationships of social features compared to Ridge, inline with the findings from Géron (2022) and Müller and Guido (2016).

RQ3: *How does the application of SMOTE for Regression influence model predictions compared to scenarios where it is not utilized?*

As disclosed in Figure 4, all models were tested with and without the resampled data generated by SMOTER. The new synthetic entries affected each model differently. The baseline model experienced a significant increase in MAE, likely due to insufficient regularization. The RFR saw a small increase in MAE but was able to reduce major prediction errors upon examining the error analysis. Meanwhile, XGBoost benefited from the new entries, as indicated by a slight decrease in MAE. Overall, each model reacted differently to the new entries, with SMOTER having a positive impact on the tree-based models, and a negative impact on Ridge model.

RQ4: *What is the feature importance, as assessed by Shapley values analysis, for the models incorporating social features?*

Addressing research question 4, Figures 5 and 6 were created to represent the feature importance summary plot and dependency plots, respectively.

The key findings indicate that the number of followers is the most important feature for the RFR, significantly enhancing prediction accuracy. As shown in Figure 6, an increase in the number of followers corresponds to an increase in the SHAP value, underscoring the feature's importance. Similarly noted by Iqbal (2022), incorporating freely available social related data can be beneficial. In this research, data from users' social interactions aided the tree-based models in predicting the number of products sold, supporting Iqbal (2022)'s claims.

Finally, the utilization of Shapley values was crucial, as it allowed to measure the importance of social features, endorsing the recommendation by Lee et al. (2021) to combine SHAP with tree-based models.

RQ5: *How does the application of SMOTE for Regression influence the feature importance analysis?*

Figure 5 illustrates the feature importance with and without using the resampled data generated by SMOTER. This indicates that in both scenarios, the feature importance for the RFR remained unchanged, and thus the graph is displayed once to avoid redundancy. In response to the research question, SMOTER did not affect the feature importance analysis of RFR because the graphs in both cases (with and without the resampled data) are identical. Consequently, it verifies the hypothesis that SMOTER does not affect the feature importance analysis.

Main RQ: *What is the impact of social features on predicting the number of products sold by a user in C2C through the application of machine learning algorithms?*

Regarding the main research question, social features confirmed to be relevant for sales prediction. Specifically, the number of followers was the most important feature, both among social features and overall, leading to relatively accurate predictions. This verifies the claim by Attar et al. (2022) that social variables are key drivers for a business, as they provide valuable information for predictions. This assertion is supported by the decreased MAE of the tree-based models when social variables were included. Also, the observed importance of social features like followers and product description quality proves the insights from Talib and Saat (2017), who emphasized the role of social proof in online shopping.

This study also addresses the gap identified by Semerádová and Weinlich (2022) by shifting the focus from predicting consumer intent to actual sales outcomes, thereby providing more actionable insights for e-commerce platforms. Consequently, businesses could leverage social features to enhance interactions on their websites and achieve more accurate sales predictions.

6.2 *Societal & Business Relevance*

The results of this study provide valuable insights into social variables, benefiting both businesses and society. Both sectors can leverage this research to advance their respective fields in their desired directions.

6.2.1 *Societal Relevance*

The societal relevance of this study lies in the discovery that social features are valuable for predicting sales. This indicates that enhancing consumer trust is achievable by improving user interfaces and interactions on online platforms through the display of social metrics. This aligns with the shift towards social commerce platforms like Facebook Marketplace, as discussed by Turban et al. (2017). Additionally, emphasizing follower counts can boost consumer confidence in sellers by providing additional social proof, while always considering the possibility of fraudulent follower counts.

Furthermore, the utility of social features can empower individual sellers on C2C platforms to leverage their social networks more effectively, improving their sales performance. Increased social engagement could lead to more successful transactions and a more vibrant online marketplace with trustworthy participants.

Additionally, by identifying key social features (similar to "socialNbFollowers") that aid in predicting sales, platforms can develop tools and resources to assist less experienced sellers or users with smaller social networks in enhancing their visibility and sales performance. This makes the platform more inclusive and accessible, democratizing access to the benefits of e-commerce.

Finally, platforms can utilize the importance of social interactions by fostering small communities within the platform. This approach benefits both the platform and society by increasing user satisfaction and loyalty while providing a sense of social inclusion.

6.2.2 *Business Relevance*

This study offers several benefits for businesses, providing crucial indicators for predicting sales numbers. One key finding is that tree-based models significantly benefit from the inclusion of social features. This suggests that businesses can improve their predictive models by integrating similar features, leading to more accurate sales forecasting and better strategic planning for the platform.

From the perspective of user-sellers, this study helps in understanding the dynamics of social variables in a C2C environment. By developing a social account that attracts engagement (followers, likes, and buyers adding products to their wishlist), user-sellers can gauge their potential sales based on these social dynamics.

Lastly, in a more balanced dataset, the SMOTER function could prove highly useful. Specifically, businesses using machine learning for sales forecasting can utilize the pipeline mentioned in section 1 to predict future sales and identify which features are most helpful in making those predictions.

6.3 *Limitations*

This research has multiple limitations that need to be addressed. First, while the results indicate an association between the number of follows of a user and the number of products sold, it is not a casual relationship, but a correlation. To establish this casual relationship, future research should test the number of sales with the social features to establish a casual relationship.

Second, while the application of SMOTER proved beneficial for this research, its scalability to larger datasets and different contexts requires further investigation. For instance, the effectiveness of SMOTER might be influenced by seasonal trends and temporal variations, suggesting a need to approach the generation of synthetic entries from a different perspective. Additionally, different variations of SMOTER should be tested to address the issue of not creating synthetic entries for higher sales values, which is the main limitation using this algorithm.

Third, the data utilized in this study represents a single snapshot in time, capturing the state of social interactions and sales metrics at one specific moment. This static view may not account for the dynamic nature of user behaviors, social interactions, and market trends, which can vary significantly over time. The potential impact of this limitation is that the results could be misleading and thus need to be cross-verified in different scenarios.

Last but not least, another limitation of this study is the generalisability of the findings across different e-commerce platforms or markets. The research was tailored to a specific dataset with its unique characteristics, which may not reflect the broader e-commerce environment. This limitation suggests the need for additional studies to verify these findings across various e-commerce settings to enhance their applicability and robustness.

6.4 *Recommendations for Further Research*

The findings of this study demonstrate that social variables are crucial for predicting sales, as all three tested features are included and specifically the number of followers ("socialNbFollowers") shows high significance in the feature importance analysis. To validate this hypothesis, future research could use the pipeline shown in Figure 1 to unseen data to cross-verify these indications.

Also, coming research could profitably expand upon this study by exploring additional social features, such as the Wishlist, which may provide valuable insights into user intentions. Understanding these aspects can enable online businesses to better anticipate customer purchasing patterns, thereby optimizing their logistics and inventory management.

Additionally, expanding this research to include a longitudinal dataset would help capture the temporal dynamics of social variables and their impact on sales. Combining this approach with time series analysis could help identify and analyze trends, seasonal effects, and cyclical patterns in the data having a better understanding of social features.

In a different research, the application of different oversampling and undersampling techniques in combination with newer techniques like deep learning could potentially enhance model performance and feature interpretation, especially in handling complex interactions and non-linear relationships within large datasets.

6.5 *Conclusion*

This study explored the predictive power of social features in a Consumer-to-Consumer (C2C) e-commerce setting, utilizing machine learning techniques. The research highlighted the significance of social variables, such as follows, followers, and likes, in enhancing the accuracy of sales predictions. Through a rigorous methodology that included Random Forest Regressor (RFR) and XGBoost (XGB) models, alongside data preprocessing and the application of the Synthetic Minority Over-sampling Technique for Regression (SMOTER), the findings demonstrated that social features substantially improve the performance of predictive models.

The Random Forest Regressor emerged as the most effective model, achieving the lowest Mean Absolute Error (MAE) and outperforming both XGBoost and the baseline Ridge Regression model. The study also emphasized the role of Shapley values in understanding feature importance, with social variables like the number of followers being identified as critical predictors.

From a societal perspective, the results suggest that enhancing user interfaces to prominently feature social metrics can boost consumer trust and engagement on C2C platforms. For businesses, these insights can inform marketing strategies and optimize predictive models, ultimately improving sales forecasting and operational efficiency.

REFERENCES

Attar, R. W., Almusharraf, A., Alfawaz, A., & Hajli, N. (2022). New trends in e-commerce research: Linking social commerce and sharing commerce: A systematic literature review. *Sustainability*, *14*(23), 16024.

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

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Chen, Y., Liu, H., Wen, Z., & Lin, W. (2023). How explainable machine learning enhances intelligence in explaining consumer purchase behavior: A random forest model with anchoring effects. *Systems*, *11*(6), 312.

Cialdini, R. B. (2007). *Influence: The psychology of persuasion* (Vol. 55). Collins New York.

D'Adamo, I., Lupi, G., Morone, P., & Settembre-Blundo, D. (2022). Towards the circular economy in the fashion industry: The second-hand market as a best practice of sustainable responsibility for businesses and consumers. *Environmental Science and Pollution Research*, *29*(31), 46620–46633.

Farnham, B., Tokyo, S., Boston, B., Sebastopol, F., & Beijing, T. (2019). *Hands-on machine learning with scikit-learn, keras, and tensorflow concepts, tools, and techniques to build intelligent systems second edition*.

Géron, A. (2022). *Hands-on machine learning with scikit-learn, keras, and tensorflow*. " O'Reilly Media, Inc.".

Hoyer, W. D., MacInnis, D. J., Pieters, R., Chan, E., & Northey, G. (2017). *Consumer behaviour: Asia-pacific edition*. Cengage AU.

Iqbal, M. (2022). Machine learning applications in e-commerce. *Organization, Business and Management*, *65*.

Lee, J., Jung, O., Lee, Y., Kim, O., & Park, C. (2021). A comparison and interpretation of machine learning algorithm for the prediction of online purchase conversion. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*, 1472–1491. https://doi.org/10.3390/jtaer16050083

Lone, S., Weltevreden, J., & Luharuwala, A. (2023). European e-commerce report 2023. www.ecommerce-europe.eu

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2019). Explainable ai for trees: From local explanations to global understanding. *arXiv preprint arXiv:1905.04610*.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, *2*(1), 56–67.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Mabilama, J. M. (2019). *E-commerce - users of a french c2c fashion store*. https://www.kaggle.com/datasets/jmmvutu/ecommerce-users-of-a-french-c2c-fashion-store

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with python: A guide for data scientists*. " O'Reilly Media, Inc.".

OpenAI. (2024). Openai [Available: https://www.openai.com].

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Semerádová, T., & Weinlich, P. (Eds.). (2022). *Achieving business competitiveness in a digital environment*. Springer International Publishing. https://doi.org/10.1007/978-3-030-93131-5

Sukrat, S., MahatananKoon, P., & Papasratorn, B. (2016). The evolution of c2c social commerce models. *2016 eleventh international conference on digital information management (ICDIM)*, 15–20.

Talib, Y. Y. A., & Saat, R. M. (2017). Social proof in social media shopping: An experimental design research. *SHS Web of Conferences*, *34*, 02005.

Torgo, L., Ribeiro, R. P., Pfahringer, B., & Branco, P. (2013). Smote for regression. *Portuguese conference on artificial intelligence*, 378–389.

Turban, E., Whiteside, J., King, D., & Outland, J. (2017). *Introduction to electronic commerce and social commerce*. Springer.

Zhang, Y., & Pennacchiotti, M. (2013). Predicting purchase behaviors from social media. *Proceedings of the 22nd international conference on World Wide Web*, 1521–1532. https://doi.org/10.1145/2488388.2488521

APPENDIX A

Table 5: Features

Name	Determinant Group	Data Description
Country	Demographics	origin of customer
CivilityGenderId	Demographics	1 = Mr & 2 = Mrs
CountryCode	Demographics	User's Country
Followers	Social features	Number of followers of the account
Follows	Social features	Number of follows of the account.
Gender	Demographics	Male (M) or Female (F)
Has App	Social features	has the app (True or False)
HasProfilePicture	Social Features	has a custom profile picture
Language	Demographics	preferred language
Last Login	Social Features	Number of days since the last login
Products Bought	Sales	Number of products bought

Continued on next page

Table 5 – Continued

Name	Category	Data Description
Products Liked	Social Features	Number of products liked
Products Listed	Sales	Number of unsold products uploaded.
Products Pass Rate (%)	Social features	% of products meeting the product description.
Products sold	Sales	Number of products sold
Products Wished	Social features	Number of products added to wishlist.
Seniority as days	Social features	Number of days since registered