Clickstream Data Classification for Online Shopping

Abstract

This project explores a comprehensive dataset from an online clothing store for pregnant women, capturing intricate clickstream data over five months in 2008. The dataset's multivariate and sequential nature provides a rich ground for classification, with a focus on understanding if the price of a particular product is higher than the average price for the entire product category. It comprises 165,474 instances across 14 distinct features, including product categories, photo locations, and pricing details. Our analysis aims to uncover price classification among different product categories being bought by pregnant women in Poland and other European countries. The insights derived could serve as a strategic tool for e-commerce platforms to optimise user experience and enhance business performance. The dataset, licensed under CC BY 4.0, encourages further research and application in the realm of business analytics.

Keywords

Clickstream Analysis, E-commerce Behavior, Price Classification, Consumer Buying Patterns, Consumer Buying behaviour, E-Shop

1. Introduction

The dataset comprises information based on clickstream from an online store which offers clothing for pregnant women. Data are from five months of 2008 and are multivariate and sequential in nature. This combination of multivariate and sequential data allows for comprehensive analysis of user behaviour in the online shopping context, supporting tasks like classification, regression, and clustering.

The clickstream data is highly relevant in the business domain, especially for e-commerce analytics, due to its rich features that capture the user interaction with online stores. The data provides a sequential record of user actions, which can be analysed to understand customer journey and preferences. Information like product category, colour, and price helps in accessing product performance and planning inventory. The country of origin and location data can be useful in marketing campaigns[6].

The application of classification algorithms in the context of clickstream data involves predicting the 'price 2' variable, which is a binary variable indicating whether the price of a product is higher than the average price for the entire product category. The classification algorithm to predict the 'price 2' variable based on other relevant features in the dataset helps in understanding pricing strategies, customer behaviour and in making business decisions.

Moreover, machine learning models can be dynamically updated and refined over time as new data becomes available, ensuring their relevance and effectiveness in real-world clinical settings.

Despite the potential benefits of using classification algorithms for clickstream database prediction, several challenges and limitations must be addressed. One such challenge is the complexity of the data, the dataset is multivariate and sequential which require complex algorithms to capture the patterns effectively. With 14 numbers of features, it is very important to select the relevant features for the classification task. Additionally, given the business importance of the dataset it is important to use models that provide interpretable results for better decision-making. Concerns regarding biased classification, transparency, and generalizability must be carefully considered to ensure the trustworthiness of predictive models in decision making [7] . By addressing these challenges and leveraging the strengths of classification algorithms, researchers can advance our understanding and develop more accurate and reliable tools for better analysis and business decisions.

2. Related Work

Clickstream data is particularly very useful for deriving insights that will be considered for the business decisions. Mariusz Łapczyński and S. Białowąs dive deeper from e-commerce point of view: attracting new visitors to the site, ensuring that the visitors will find the product they need and supporting the sales process[1]. Rainer Olbrich &Christian Holsing finds that community members are more likely to make a click-out than ordinary users. This implies that community members are more profitable [2]. Melina Zavali, Ewelina Lacka & Johannes de Smedt shows that although the “mobile window shoppers” segment consists of the largest consumer segment, it attracts the lowest revenue[3]. Gökhan Sılahtaroğlu & Hale Dönertaşli works on whether customers will or will not buy their items added to shopping baskets on a digital marketplace[4]. Randolph E. Bucklin and Catarina Sismeiro’s paper reviews major developments from the analysis of these data, covering advances in understanding (1) browsing and site usage behavior on the Internet, (2) the Internet's role and efficacy as a new medium for advertising and persuasion, and (3) shopping behavior on the Internet [5].

3. Proposed Methodology

3.1 Data Description

The dataset features clickstream data from an online store specialising in clothing for pregnant women. It captures user behaviour over five months of 2008. It is a multivariate and sequential dataset with 14 features and 165,474 instances. The data includes variables like year, month, session ID, order, model photography, price 2 and page. The ‘session ID’ is a unique identifier for each session. ‘Order’ the sequence of clicks during a single session. ‘Model Photography’ indicates if the product was displayed with model photography. ‘Price’ is the price in USD. ‘Price 2’ is a binary variable indicating if the price is higher than the

average for the category. This comprehensive dataset you're viewing offers detailed insights into online shopping behaviour, specifically for a store catering to pregnant women

	year	month	day	order	country	session ID	page 1 (main category)	page 2 (clothing model)	colour	location	model photography	price	price 2	page
0	2008	4	1	1	29	1	1	A13	1	5	1	28	2	1
1	2008	4	1	2	29	1	1	A16	1	6	1	33	2	1
2	2008	4	1	3	29	1	2	B4	10	2	1	52	1	1
3	2008	4	1	4	29	1	2	B17	6	6	2	38	2	1
4	2008	4	1	5	29	1	2	B8	4	3	2	52	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
165469	2008	8	13	1	29	24024	2	B10	2	4	1	67	1	1
165470	2008	8	13	1	9	24025	1	A11	3	4	1	62	1	1
165471	2008	8	13	1	34	24026	1	A2	3	1	1	43	2	1
165472	2008	8	13	2	34	24026	3	C2	12	1	1	43	1	1
165473	2008	8	13	3	34	24026	2	B2	3	1	2	57	1	1

3.2 Data Pre-processing

The data preprocessing steps performed on the dataset aims to prepare it for analysis and modelling by addressing missing values and handling categorical variables [6]. No missing or duplicate values were found in the dataset. Next, unnecessary features such as 'session ID' and 'year' are removed to focus more on significant variables.'session ID' is a unique identifier that could lead to overfitting [7]. Since the value of 'year' does not vary (e.g., all data is from 2008), it won't affect the model's outcome and hence is dropped. In order to reduce the dimensionality of the data, only 6 features are selected based on the mutual information between each feature and the target variable. Top 6 features with highest mutual information are considered for final analysis [8]. Lastly, numerical features are standardised using StandardScaler to ensure consistent scaling across different features, resulting in effective model training and interpretation. Overall, the data preprocessing pipeline ensures that the dataset is properly formatted and ready for further analysis [9].

3.3 Model Description

1. **Random Forest:-** Random Forest is a versatile and powerful ensemble learning algorithm widely used in both classification and regression tasks. It operates by constructing a multitude of decision trees during the training phase. Each decision tree is built using a subset of the features and a bootstrapped sample of the data, ensuring diversity among the trees. In classification tasks, the final prediction is determined by aggregating the predictions of individual trees through either voting or averaging, while in regression tasks, it's typically the mean prediction of all trees. Moreover, it can handle high-dimensional data and interactions between features effectively.
2. **Support Vector Machine (SVM):** Support vector machine is a powerful supervised learning algorithm used for both classification and regression tasks. Its primary objective in classification is to find the optimal hyperplane that best separates different classes in the feature space, maximising the margin between the classes. SVM achieves this by identifying support vectors, which are data points closest to the decision boundary or hyperplane. These support vectors play a crucial role in defining the decision boundary and are instrumental in achieving the algorithm's robustness to outliers.

3. **Logistic regression:-** Logistic Regression is a fundamental technique for binary classification tasks.It models the probability of a binary outcome based on one or more predictor variables by fitting a logistic function to the observed data. The logistic function, also known as the sigmoid function, maps any real-valued input to the range [0, 1], allowing it to output probabilities.
   Despite its simplicity, logistic regression offers several advantages, including interpretability, efficiency in handling linearly separable data, and resistance to overfitting.
4. **The K-Nearest Neighbors (KNN) algorithm:** KNN is a simple yet effective non-parametric method used for both classification and regression tasks. In KNN, predictions for a new data point are made based on the majority class or average value of its nearest neighbours in the feature space. The algorithm calculates the distance between the new data point and all other points in the training dataset, typically using Euclidean distance, and selects the K nearest neighbours. The class or value of the new data point is then determined based on the most common class or average value among its neighbours.
5. **Decision Tree:** Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

4. Results and Analysis

`

The evaluation metrics considered for each model are accuracy, loss, recall, precision, AUC (area under the curve), and [11].
Accuracy measures the ratio of correctly predicted observations to the total observations.
Precision is the ratio of true positives (TP) to the sum of true positives and false positives (FP).
Recall is the ratio of true positives to the sum of true positives and false negatives (FN).

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (1)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (2)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (3)$$

Table 1: Comparison Table Of Various Models Applied

Models	Precision	Recall	Accuracy
Random forest	1.000000	1.000000	1.000000
Support vector machine	0.999882	0.999949	0.999914
logistic regression	0.996582	1.000000	0.998247
KNN	0.893236	0.955659	0.927738
Decision tree	1.000000	1.000000	1.000000

From the provided table, we can observe the precision, recall and support metrics for five different classification models: Random Forest, Support Vector Machine (SVM), Logistic Regression, decision tree, and K-Nearest Neighbors (KNN).

6. Conclusion

In conclusion, the comparative analysis of five classification models - Random Forest, Support Vector Machine (SVM), Logistic Regression, decision tree, and K-Nearest Neighbors (KNN) - reveals varying degrees of performance in predicting the target variable. Among these models, Random Forest and decision tree stand out as the top performers, boasting precision, recall, and accuracy of 1.0000000, indicating their robustness and high accuracy in classifying instances. Both models outshine SVM, Logistic Regression, and KNN, which exhibit slightly lower performance metrics.Overall, the results highlight Random Forest and decision tree are the preferred choices for this classification task due to their superior performance [10].

References

[1] Discovering Patterns of Users' Behaviour in an E-shop - Comparison of Consumer Buying Behaviours in Poland and Other European Countries
By Mariusz Łapczyński, S. Białowąs. 2013

[2] Modeling Consumer Purchasing Behavior in Social Shopping Communities with Clickstream Data
Rainer Olbrich & Christian Holsing

[3] Shopping Hard or Hardly Shopping: Revealing Consumer Segments Using Clickstream Data
Melina Zavali, Ewelina Lacka & Johannes de Smedt

[4] Analysis and prediction of E-customers' behavior by mining clickstream data
Gökhan Sılahtaroğlu & Hale Dönertaşli

[5] Click Here for Internet Insight: Advances in Clickstream Data Analysis in Marketing
Randolph E. Bucklin and Catarina Sismeiro

[6] Consumers' decision-making process and their online shopping behavior: a clickstream analysis
Sylvain Senecal , Pawel J. Kalczynski , Jacques Nantel

[7] Visualization and Analysis of Clickstream Data of Online Stores for Understanding Web Merchandising

[8] Using clickstream data to analyze online purchase intentions
Ricardo Filipe Fernandes e Costa Magalhães Teixeira

[9] Shopper intent prediction from clickstream e-commerce data with minimal browsing information
Borja Requena, Giovanni Cassani, Jacopo Tagliabue, Ciro Greco & Lucas Lacasa

[10] A Deep Markov Model for Clickstream Analytics in Online Shopping
Yilmazcan Ozyurt, Tobias Hatt, Ce Zhang & Stefan Feuerriegel

[11] Clickstream Data and Inventory Management: Model and Empirical Analysis
Tingliang Huang and Jan A. Van MieghemView all authors and affiliations