RES MILITARIS
REVUE EUROPEENNE D ETUDES EUROPEAN JOURNAL OF MILITARY STUDIES
Social Science Journal

Enhancing Customer Retention in E-commerce Through Data-Driven Insights and Predictive Analytics

Ch. Aruna[1*], G. Vital[1], K. Mahesh[1], K. Venkatesh[1], Md. Umez[1]

[1]Department of CSE, Sree Dattha Institute of Engineering and Science, Sheriguda, Hyderabad, Telangana, India

Corresponding E-mail: aruna@sreedattha.ac.in

Abstract

In recent years, the expansion of e-commerce has been rapid and it has become a dominant force in the retail industry. Nevertheless, the rapid expansion of e-commerce has led to a fierce rivalry among enterprises, hence emphasizing the utmost importance of customer retention. Retaining clients is not only economically efficient but also plays a substantial role in achieving long-term business success and profitability. The notion of client retention has been essential to company operations in several industries for centuries. The difficulty lies in creating efficient customer retention methods customized to the distinct dynamics of online commerce. One must possess a profound comprehension of customer behavior, preferences, and purchasing intents in order to do this. Examining data on consumers' desire to make purchases yields useful information that may be used to develop focused tactics aimed at maintaining client engagement and loyalty. In the past, customer retention strategies in e-commerce were mostly based on email marketing, loyalty programs, and occasional discounts. Although these tactics continue to be effective, they frequently lack the personalization and data-driven insights necessary to fully comprehend and adapt to particular customer preferences. Thus, the suggested method utilizes sophisticated data analytics techniques to extract valuable insights from purchasing intention data. Through the examination of client behavior, browsing habits, and interactions on the e-commerce platform, firms can discern crucial indicators of purchasing intent. This information can be utilized to customize marketing strategies, deliver personalized suggestions, and provide incentives that strongly appeal to individual clients. Moreover, the suggested algorithms are utilized to forecast forthcoming purchasing intentions by leveraging past data, allowing organizations to actively interact with customers prior to their purchase determination. This data-centric approach not only improves customer happiness but also optimizes marketing endeavors and resource distribution.

Keywords: Machin Learning, E-commerce growth, Customer retention, Data analytics, Personalized marketing.

1. Introduction

The e-commerce sector in India has experienced rapid growth, driven by increasing internet penetration, smartphone adoption, and a burgeoning middle class. The convenience of online shopping, coupled with a wide range of products and services, has made e-commerce a preferred choice for many consumers. As the sector expands, the focus has increasingly shifted towards customer retention, a key factor in ensuring long-term success and profitability for e-commerce businesses. E-commerce in India began gaining traction in the early 2000s with the entry of players like Rediff and Indiatimes. However, the real transformation started around 2007 with the launch of Flipkart, which revolutionized the online shopping experience in India.

- 2007-2012: Flipkart's introduction of a customer-centric approach, including cash on delivery (COD) and a liberal return policy, set new benchmarks in customer service. Other players like Snapdeal and Myntra also emerged during this period.
- 2013-2016: Amazon entered the Indian market in 2013, intensifying competition and leading to significant innovations in logistics and supply chain management. The e-commerce sector saw explosive growth, with the gross merchandise value (GMV) of the market reaching $14 billion in 2015.
- 2017-Present: The entry of Reliance Jio in 2016 with affordable data plans further accelerated internet penetration, bringing millions of new users online. By 2020, the e-commerce market in India was valued at approximately $64 billion and is projected to reach $200 billion by 2026.

Internet Penetration and E-commerce Usage

- Internet Users: As of 2023, India has over 850 million internet users, making it the second-largest online market globally.
- Smartphone Users: There are around 760 million smartphone users in India, which is a significant driver of mobile e-commerce.
- E-commerce Growth: The Indian e-commerce market is expected to grow at a compound annual growth rate (CAGR) of 27% from 2021 to 2026, reaching a market size of $200 billion by 2026.

Consumer Behavior

- Online Shoppers: Approximately 150 million Indians made at least one online purchase in 2021, and this number is expected to rise as internet and smartphone penetration continue to increase.
- Shopping Preferences: Indian consumers are increasingly preferring online shopping for its convenience, variety, and competitive pricing. Categories like electronics, fashion, groceries, and personal care products are particularly popular.

Government Initiatives

- Digital India: Launched in 2015, this initiative aims to transform India into a digitally empowered society and knowledge economy, facilitating greater e-commerce adoption.
- FDI Policies: The Indian government has eased foreign direct investment (FDI) norms for the e-commerce sector, attracting significant investments from global players like Amazon and Walmart (which acquired a majority stake in Flipkart).

Challenges

- Competition: The intense competition among e-commerce platforms requires continuous innovation and investment in customer acquisition and retention strategies.
- Infrastructure: While urban areas are well-served, rural and remote regions still face challenges in terms of logistics and last-mile delivery.
- Regulatory Environment: E-commerce companies must navigate a complex regulatory environment, including data protection laws and foreign investment regulations.

The e-commerce sector in India has come a long way from its early beginnings, evolving into a dynamic and rapidly growing market. With increasing internet and smartphone penetration, favorable government initiatives, and a large consumer base, the potential for further growth is immense. However, the key to sustaining this growth lies in effective customer retention strategies that leverage data-driven insights to understand and meet the evolving needs of consumers. By focusing on

personalization, predictive analytics, and proactive engagement, e-commerce businesses in India can enhance customer satisfaction and loyalty, driving long-term success in an increasingly competitive market.

2. Literature survey

Research shows that companies applying big data analytics outperformed their peers by 5% in productivity and 6% in profitability and 92% of company executives from across 19 countries were satisfied with the results produced by big data analytics and they expected big data to have bigger impacts in their organizations specifically on customer relationship management [1]. However, there are still findings that 70% of the customer data is never used for making improvements and only 30% of organizations use customer experience to help succeed in today's market [2]

There are concerns over big data analytics that customer behavior is changing so fast that the historical data and model cannot be trusted, and the patterns are hard to discern. This can be remedied by tapping high quality data such as company's own in-house data or third-party data and employing and iterating a better model in an agile setting [3]. Santana can predict the problem even before engaging the customer and ends up with a better customer relationship management [4]. The barrier to establish a digital brand is no longer high and there are so many brands that customers are easily attracted by other brands and tend to be less loyal [5]. The firm's cash back fuel reward card builds the customer loyalty as well as retention [6]. This loyalty program has proved to work well and poses a threat to existing business rivals such as Amazon's loyalty program Prime [7]. This system is utilized to analyze around 100TB of data generated daily over 11000+ stores and 12 online websites [8] globally in order to ensure customers can always buy the desired product. Customer checkout experience is thus greatly improved [9]. the company expects to experience huge gains in grocery pickup and delivery programs and expects to surpass $28B in 2020 [10]. The new Walmart Grocery App was developed to assist customers in paying for online groceries. This app recorded a rise in 2020, following Covid-19, surpassing Amazon by 20% [11]. The Walmart app is used by more than 22 million active users who rank the company as one of the best worldwide [12].

3. Proposed methodology

Figure 1 shows the proposed system architecture.

Dataset Acquisition: The first step involves acquiring a comprehensive dataset that includes relevant information on customer interactions, purchase history, browsing patterns, and other relevant data within the e-commerce platform. This dataset serves as the foundation for deriving insights into purchasing intentions.

Data Preprocessing: Raw datasets often contain noise, missing values, and outliers. Data preprocessing is essential to clean and transform the data into a usable format. This step includes handling missing values, removing outliers, and standardizing data to ensure the accuracy and reliability of subsequent analyses.

Data Splitting: To effectively evaluate the performance of the developed strategies, the dataset is divided into training and testing sets. The training set is used to build and train the predictive models, while the testing set is employed to assess the models' performance and generalizability.

Performance Evaluation: Various performance metrics are employed to evaluate the effectiveness of the developed customer retention strategies. Metrics such as accuracy, precision, recall, and F1 score are calculated to gauge the models' ability to predict purchasing intentions accurately.

Prediction from Test Data: The predictive models developed from the training set are applied to the test set to predict customer purchasing intentions. This step assesses the models' performance in real-world scenarios, providing valuable insights into their practical utility.

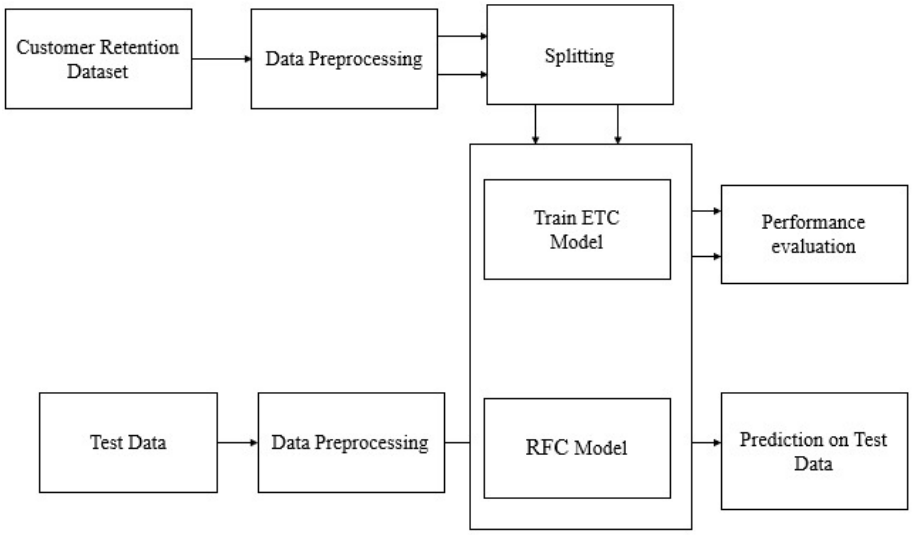

Figure 1. Proposed System Architecture.

3.1 Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

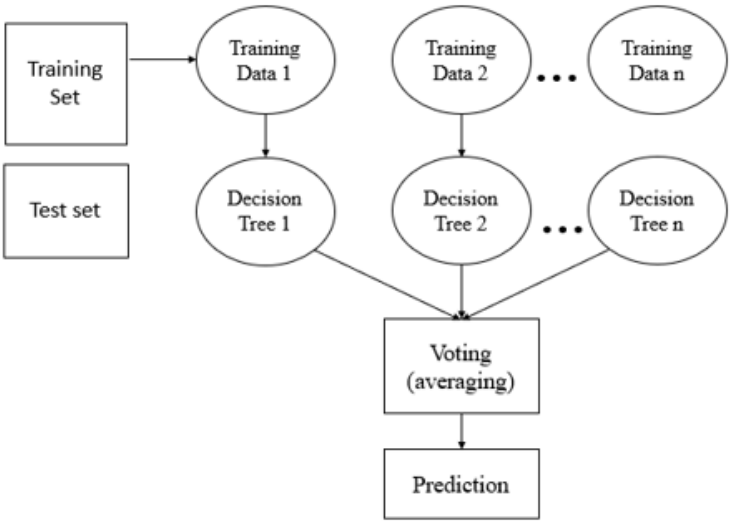

Fig. 2: Random Forest algorithm.

Random Forest algorithm

Step 1: In Random Forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

Important Features of Random Forest

- Diversity- Not all attributes/variables/features are considered while making an individual tree, each tree is different.
- Immune to the curse of dimensionality- Since each tree does not consider all the features, the feature space is reduced.
- Parallelization-Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.
- Train-Test split- In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.
- Stability- Stability arises because the result is based on majority voting/ averaging.

Types of Ensembles

Before understanding the working of the random forest, we must look into the ensemble technique. Ensemble simply means combining multiple models. Thus, a collection of models is used to make predictions rather than an individual model.

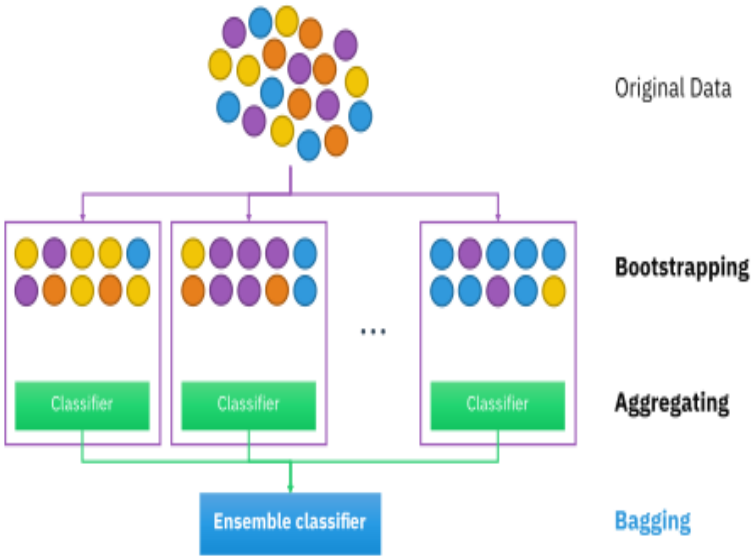

Fig. 3: RF Classifier analysis.

Ensemble uses two types of methods:

Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest. Bagging, also known as Bootstrap

Aggregation, is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as aggregation.

Boosting– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.

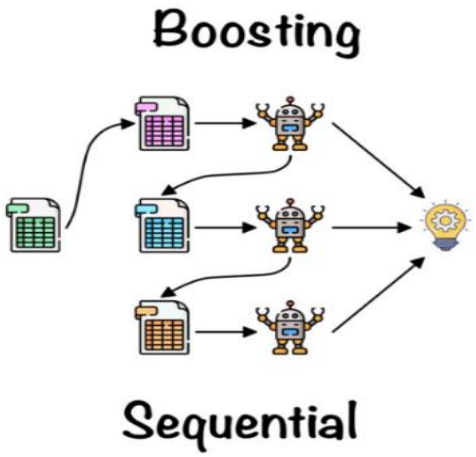

Fig. 4: Boosting RF Classifier.

4. Results and Discussion

Figure 5 shows that Confusion matrix of Extra Tree Classification (ETC). A confusion matrix is a table used to visualize the performance of an image classification model. The rows represent the actual classes and the columns represent the predicted classes. The diagonal elements represent the number of correctly classified instances, while the off-diagonal elements represent the number of misclassified instances.

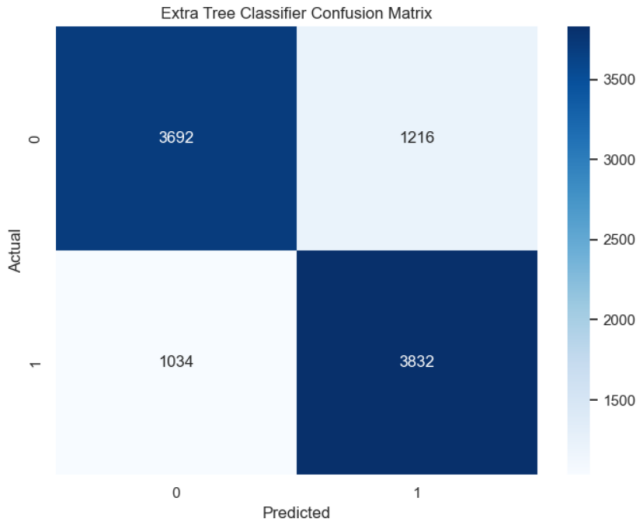

Figure 5: Figure illustrates the performance evaluation of Extra Tree classifier.

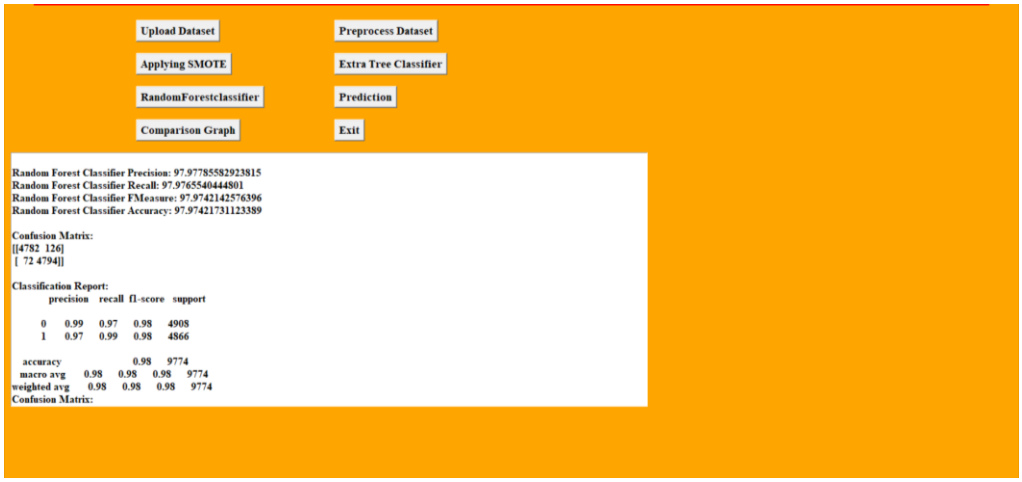

Figure 6: Figure illustrates the performance evaluation of Random Forest classifier.

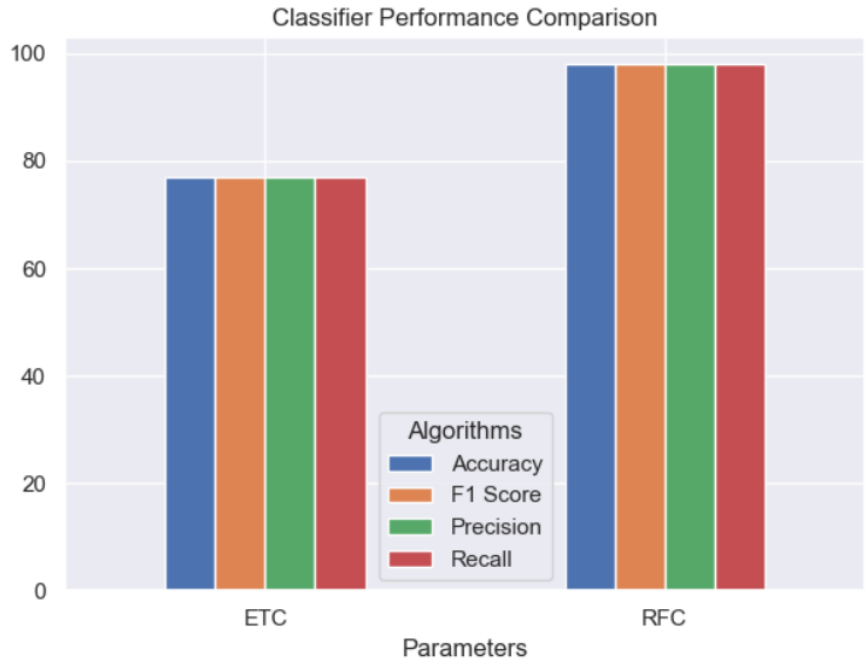

Figure 7: Performance comparison of existing and proposed models.

Figure 6 and Figure 7 shows the RFC is best accuracy, precision and recall and F1 score that ETC.

- **Precision:** How many of the items the model classified as positive actually were positive. In this case, the precision is 97.98 for both class 0 and class 1. This means that out of 100 items the model classified as positive, 98 of them were actually positive.
- **Recall:** How many of the actual positive items did the model classify correctly? The recall is 97.97 for class 0 and 99 for class 1. This means that out of 100 actual positive items, the model classified 98 of them correctly for class 0 and 99 of them correctly for class 1.
- **F1-Score:** The harmonic mean between precision and recall. It's a way of looking at both precision and recall at the same time. A value of 1 means the classifier performs perfectly with both precision and recall being 1. In this case, the F1 score is 0.98 for both class 0 and class 1.

- **Accuracy:** How many of the total items did the model classify correctly? The overall accuracy of the model is 97.98%. This means that out of 100 items, the model classified 98 of them correctly.
- **Confusion Matrix:** This table shows how many items were classified into each category. The rows represent the actual classes, and the columns represent the predicted classes. So, for example, the value 4782 in the top left corner of the confusion matrix means that 4782 items were actually class 0 and the model also predicted that they were class 0.

5. Conclusion

This project presents a comprehensive approach to analyzing customer retention strategies in e-commerce by leveraging advanced data analytics and machine learning techniques. By utilizing a user-friendly GUI application, the project enables users to upload, preprocess, analyze, and predict customer retention based on purchasing intention data. The integration of algorithms like Extra Trees Classifier and Random Forest Classifier, along with techniques such as SMOTE for handling imbalanced datasets, ensures robust and reliable analysis. Key insights gained from this analysis can help e-commerce businesses understand customer behavior, preferences, and purchasing intentions more deeply. This understanding allows for the development of personalized marketing strategies, improving customer engagement and loyalty. By predicting future purchasing intentions, businesses can proactively engage with customers, enhancing satisfaction and optimizing resource allocation.

REFERENCES

[1] Ghandour, A. (2015). Big Data Driven E-Commerce Architecture. International Journal of Economics, Commerce & Management (IJECM) [online]. 3. [Viewed 5 April 2024]. Available from: https://www.researchgate.net/publication/277018506_Big_Data_Driven_E-Commerce_Architecture

[2] Satish, L. and Yusof, N. (2017). A Review: Big Data Analytics for enhanced Customer Experiences with Crowd Sourcing. Procedia Computer Science [online]. 116, 274-283. [Viewed 6 April 2024]. Available from: doi: 10.1016/j.procs.2017.10.058

[3] Bibby, C., Gordon, J., Schuler, G. and Stein, E. (2021). The big reset: Data-driven marketing in the next normal [online]. Mckinsey & Company. [Viewed 8 April 2024]. Available from: https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/the-big-reset-data-driven-marketing-in-the-next-normal

[4] Santana, J. (2019). How companies will use big data for customer retention [online]. ALTA. [Viewed 8 April 2024].

[6] Mckinsey & Company. (2020). Customer loyalty: The new generation [online]. Mckinsey & Company. [Viewed 12 April 2024]. Available from: https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/customer-loyalty-the-new-generation

[7] Gallagher, B. (2020). Walmart's rise of e-commerce & brand loyalty [online]. ANNEX CLOUD. [Viewed 12 April 2024]. Available from: https://www.annexcloud.com/blog/walmarts-rise-to-ecommerce-brand-loyalty

[8] Springer, J. (2020). Walmart Reveals Paid Loyalty Program [online]. Insight Grocery Business. [Viewed 12 April 2024]. Available from: https://www.winsightgrocerybusiness.com/retailers

/walmart-reveals-paid- loyalty-program

[9] Malur, R. "Pillars of Walmart's Demand Forecasting." Medium, 1 Aug, 2019, https://medium.com/walmartglobaltech/pillars-of-walmarts-demand-forecasting- f6722de86e1a Accessed 30 April, 2024.

[11] Kaziukenas, J. (2020). Walmart's Online Sales to Reach $28 Billion in 2020 [online]. Marketplace Pulse. [Viewed 12 April 2024]. Available from: https://www.marketplacepulse.com/articles/walmarts-online-sales-to-reach-28-billion-in- 2020

[12] Perez, S. (2020). Walmart Grocery app sees record downloads amid COVID-19, surpasses Amazon by 20% [online]. TechCrunch. [Viewed 13 April 2024]. Available from: https://techcrunch.com/2020/04/09/walmart-grocery-app-sees-record-downloads-amid-covid-19-surpasses-amazon-by-20

[13] SAS. (N.d.) How Walmart makes data work for its customers [online]. SAS. [Viewed 14 April 2024]. Available from: https://www.sas.com/en_us/insights/articles/analytics/how-walmart-makes-data-work-for-its-customers.html