Enhanced Customer Behavior Analysis in Retail Using AWS

Suryamritha M, Varshini Balaji, Sruthi S, K Dinesh Kumar
Department of Computer Science and Engineering,
Amrita School of Computing, Bengaluru, Amrita Vishwa Vidyapeetham, India
suryamritha.manoj@gmail.com, bl.en.u4aie21139@bl.students.amrita.edu, bl.en.u4aie21124@bl.students.amrita.edu,
kk_dinesh@blr.amrita.edu

***Abstract*—In the current advanced world, e-commerce comes into the limelight by providing goods to customers at their doorsteps without stepping out. Now that so many shoppers use the internet to make purchases, customer opinions have increased in significance when deciding to buy an item. Yet, sorting through hundreds of customer reviews to comprehend a product or its features is a lengthy process. However, using machine learning, especially supervised learning methods, we can do this task efficiently. This project uses the the Amazon Customer Reviews Electronics dataset available on Amazon S3 which makes it easy to obtain massive product review datasets. Landscape sentiment analysis and classifiers include Random Forest, Gradient Boosting, and Support Vector Machines – which we compare as to their effectiveness in capturing and predicting the polarity of a sample. This combination of sentiment analysis and the provision of remote commerce facilities improves the way people shop and enables them to glean crucial insights concerning goods in more simplified terms as they are available with lots of reviews.**

***Index Terms*—e-commerce, machine learning, sentiment analysis, supervised learning, Amazon Customer Reviews Electronics dataset, cloud-dataset**

I. Introduction

Nowadays, buyers have their needs satisfied online, with e-commerce being the most convenient approach most 'online shoppers' use [1]. They browse the Internet for the types of items they want to buy, especially electronics where product options are constantly expounding. With this transition, potential customers are above all searching for feedback to determine how other users see the products- so now reviews are very important. However, many reviews for popular items can be quite a lot for customers and not useful as they are virtually impossible to work through or find needed information.

Machine learning is an effective way to address this by providing automated processes for classifying and analyzing customer feedback. Large-scale datasets can be used to train models that're able to distinguish reviewers' attitudes toward products (positive, negative, or neutral) within reviews, enabling customers to get a sense of the product's overall favorability. This paper examines the Amazon Customer Reviews Electronics dataset that contains customer reviews for various electronic products stored in the cloud to develop a sentiment analysis model. The dataset includes essential characteristics such as product ID, review title, review content, and number of stars received, which are all useful for model training.

To address this issue, we use several machine learning algorithms such as logistic regression, support vector machines (SVM), and deep learning models like BERT (Bidirectional Encoder Representations from Transformers). These models are trained and validated on the cloud dataset in Amazon S3, with Amazon SageMaker being employed for the training and deployment of the model respectively. With the combination of SageMaker's infrastructure and algorithms for machine learning, we are able to carry out the task of classification of the reviews of a particular product basing on their sentiment and this process is way faster as compared to manual review evaluation.

The output of this classification task is beneficial to both, the customers and the businesses. With the availability of this technique, customers will no longer take time to search for a product as they now have all the information they need. On the business point of view, businesses are in a position to analyse customers satisfaction and therefore enhance the products being offered. This project emphasizes the need for deployment of ecommerce cloud services and machine learning in modern days, showing how machine learning and cloud technologies can promote users' experience by reducing time taken to complete essential tasks such as reviewing items.

In this area too, there's a gap in the communication strategy as they fail to identify and analyze customer sentiment and its importance. Businesses looking to stay relevant as the volume of customer reviews keeps improving need to harness the power of sentiment analysis aimed at improving competitiveness in the e-commerce landscape. In addition, analyzing the customer's voice prevents the possible areas of discomfort during the use of the product. As soon as customers express dissatisfaction, they can address the problem faster minimizing its escalation. More than improving customer experience and enhancing customer retention that results from trust that their problem is being fixed, is better usage of such an approach. Additionally, the analysis improves on how businesses understand their customers as they take a more targeted marketing approach which improves overall sales experience.

Additionally, the process of classification can further be enhanced by incorporating more complex and advanced models leading to integration of more consistent frameworks. We can come closer to this goal using models like BERT that allows us to focus on the context in the text and achieve higher integration. This feature makes it possible to conduct a more in-depth review of emotions and feelings attached with consumer products and portray sarcastic emotions through strong emotional impulses that may not be captured through conventional approaches. Consequently, the knowledge gathered from this study can guide the creation of new products, consumer engagement plans, and general corporate expansion. However, this unification of machine learning and cloud computing also helps make customer feedback analysis more efficient, putting companies in a very favorable position to succeed in a market that is quickly becoming increasingly informed by data and where success depends on accessing insights more quickly and with greater relevance.

II. Literature Review

The focus has been on sentiment analysis, as outlined in Kousik et al's study [2] in the field of NLP and text mining for more effective analysis of clients' feedback. The reason why many firms apply this technique is to measure the level of satisfaction among customers or the value of products offered in the market. Feedback gathered through websites is a common way feedback is collected and data can be stored in databases and MySQL or as CSV files for analysis. However, Naïve Bayes, sentiment classification commonly used technique, has some limitations for example use non-sensitive to word position that causes inaccurate sentiment prediction and context. Thus, indeed higher level methods, such as deep learning models as LSTMs and transformers are being considered to capture further nuances of the customers' feedback. It is the intention of this study to determine the effectiveness of these methods and determine the best approach to use when it comes to analyzing customer sentiment.

Hardt et al. have offered Amazon SageMaker Clarify [3] that which was launched December 2020 aims at meeting the increasing needs for interpretability and fairness in ML models for bias identification and prediction explanation. Clarify is always a feature built into Amazon SageMaker; it can remove bias and showcase the significance of features from data preprocessing through evaluation and even after the model is deployed. Its premise of modularity is versatility through implementation of features undertaken by customers. Some of the big challenges include compromise between computation complexity and those compromises inherent in bias detection methodologies. Clarify has been successfully deployed in client use cases and has received great feedback for improving openness and accountability in machine learning workflows. These experiences also helped to highlight best practices for incorporating fairness tools into real-world applications.

Chougule et al. researched on Credit card fraud detection [4], with classic methods such as rule-based systems and statistical analysis frequently proving insufficient to handle the complexity and volume of modern transactions. Recent work highlights the expanding use of machine learning for fraud detection due to its capacity to automatically find patterns and abnormalities in huge datasets. The imbalanced nature of fraud datasets, in which fraudulent transactions make up a small fraction of the data, necessitates advanced techniques such as oversampling, undersampling, and anomaly detection. Decision trees, support vector machines (SVM), and neural networks are among the algorithms investigated in the studies, with ensemble techniques frequently producing superior accuracy. While machine learning models have demonstrated promise, scalability and deployment remain challenges in practical use. Platforms like AWS SageMaker help address by offering powerful cloud-based solutions that streamline the process from model training to real-world deployment.

Ravindranathan et al. talks [5] about how Cloud computing has transformed access to IT resources, with Amazon Web Services (AWS) pioneering full cloud-based Machine Learning solutions. From the AWS's perspective, there are several services that facilitate smoother design, training and deployment of the ML models and therefore suggest them to be more scalable. This evaluation specifically concerns AWS's Machine as Learning as a Service (MLaaS) categorizing various services it offers including SageMaker for model building, S3 for data management, and Lambda for execution. Earlier studies have indicated that AWS is capable of handling variety of ML application such as predictive analytics and deep learning. All in all, AWS's cloud solutions are changing the way in which machine learning is done providing handy tools for cloud connectivity and sizablity. Suitable approach to identify Customer Sentiment precisely.

The research done by Liberty et al. [6] titled scalable machine learning deploy a lot of emphasis on the fact that much challenges that companies face when training models from big, dynamically changing datasets are the high cost of computation and finances that are required. In view of all these complexities, true large-scale machine learning platforms will need to contain features like incremental learning capability as well as flexibility and hyperparameter tuning. Amazon SageMaker, a service provided by Amazon on AWS, is precisely designed to overcome these challenges. They can be continued and are also easily scalable with additional features that include auto-optimization. The work also discovers in what way various ML schemes might be utilized together with SageMaker and unveils that, on big datasets, assessment of time and cost, it excels many other versions of JVM-based ML.

Sharma and Waoo used machine learning to analyse customer behaviour in e-commerce [9]. The goal was to use machine learning to analyse customer behaviour in order to improve e-commerce systems. The strategy centred on using data mining and machine learning methods to forecast customer behaviour. According to their findings, machine learning models enhanced the precision of customer behaviour predictions, supporting business decision-making.

Singh et al.'s research focuses on the summarization of factors enhancing customer engagement and retention in e-commerce by reviewing segmentation techniques used for personalized marketing [10]. They utilize K-means clustering and RFM analysis, and apply decision trees, contrasting these methods. With the help of algorithms and machine learning, the target market is drawn to strategies developed on segmentation patterns. It has been proven that with these segmentation methods, people are more likely to convert, thus increasing their satisfaction with the services provided. Consequently, these approaches are important in modern marketing.

Sharma et al.'s objective focuses on recommending refined pricing strategies that are better suited to the current e-commerce environment [11]. This is achieved by using a dynamic pricing strategy, concentrating on statistical forecasting. Sharma et al. have adopted a weight-optimized LSTM model to dynamically change prices. In their case, an LSTM model was trained to adjust prices on the fly depending on customer demand and buying patterns. Their findings demonstrate that such methods outperform classic pricing strategies, improving competition, adaptability, and overall customer satisfaction on e-commerce platforms.

The tool for monitoring deployed machine learning models in real time, Amazon SageMaker Model Monitor, is discussed by Nigenda et al. [7]. The aim was to resolve challenges such as data drift and concept drift, which are encountered when maintaining the performance of a model post-deployment. To tackle biases and drifts, the method uses customizable alarms and continuous monitoring with specified statistical criteria. More than two years of actual production deployment clearly demonstrated the system's ability to ensure high-quality machine learning models in production.

Focusing on Amazon SageMaker, Yeung et al. sought to improve e-commerce activities through the use of cloud-based machine learning [8]. The strategy included cloud-oriented stages for sequencing, gathering, accumulating, and utilizing data, which involved constructing and implementing machine learning models using SageMaker. As illustrated in a real-world e-commerce case study, this integration enhanced data analysis and provided customized customer experiences.

III. METHODOLOGY

A. AWS IAM: Secure Access Configuration

he first part of the project was to create a secure access provision using AWS Identity and Access Management (IAM). For this particular project an IAM user was created with programmatic access. This enabled generation of access key ID and secret access key which were securely recorded and stored for use in future. Subsequently, a new custom IAM IAM policy was created that was intended to provide level of access to the primitive to this new user. It was possible to give permission to use Amazon S3 for data storage, Amazon Athena for executing queries on datasets, and Amazon SageMaker for training and deploying the model, Amazon CloudWatch for monitoring logs and metrics. Full access permissions were issued to facilitate testing and integration of services. This made it possible to perform unrestricted actions on the assigned services such as creating, modifying, and deleting resources. Further activities such as dataset management and service configurations must adhere to best practices. All these measures were geared toward meeting the requirements of best practice.

B. Amazon S3: Data Storage and Management

The main storage address used throughout the project was Amazon S3. The raw dataset was placed into a dedicated S3 bucket before the completion of the project. Another bucket was converted to collect the Athena query outputs, consisting of a balanced dataset created using queries. After the balanced dataset was achieved, this file was downloaded locally, verified, and re-uploaded to a separate S3 bucket.

C. Amazon Athena: Data Querying and Balancing

query and preprocess the colossal dataset in Amazon S3, Amazon Athena was used. With a SQL-like query language, outliers in the distribution of the review scores (1–5) were identified and handled. To make later queries easier, an external table `new_reviews_table` was created to structure the reviews and point to the specified S3 location. For each review sentiment score, a sample of 100 reviews was selected and placed into one of the 5 temporary tables. After that, the project proceeded by combining these 5 tables into a single balanced dataset using the `UNION ALL` operator. A final set was prepared and posted to S3, while the temporary tables were removed to save space. This task was made more manageable due to Athena's efficiency, and it allowed for the entire operation to be performed in a single interaction, without additional storage costs. This preprocessing step was crucial in transforming the data for the subsequent machine learning workflows.

D. Amazon SageMaker: Model Training, Evaluation, and Deployment

Amazon SageMaker created an end-to-end solution that encompassed the artificial intelligence project development cycle. The scope of work implemented covered everything from data preparation to model deployment, with SageMaker being used throughout the entire process, including model training, evaluation, and real-time inference.

1) Data Preparation in SageMaker: The first task was to upload the processed dataset to the `mynewreviews` S3 bucket. This dataset, which consisted of ratings and several text and numerical features, was uploaded to SageMaker for model training. Activities involved the cleanup of the data, including textual data preparation. The textual content of the reviews was processed using certain Python libraries, which also involved tokenization and vectorization of content via TF-IDF (Term Frequency-Inverse Document Frequency). This method was used to transform text data into machine-recognizable numerical formats suitable for machine learning processes. Additionally, normalization of the numeric data was performed, and the numeric data was integrated with text-based measures to form a complete feature set.

2) Model Training: Several machine learning algorithms were employed to predict the review rating, which could range between 1 and 5 based on the features:

a) Random Forest: This algorithm was selected due to its strength in handling highly complex, large-density data. It was trained with 100 estimators, allowing it to provide reliable predictions across various scenarios.

b) Logistic Regression: Regularization was applied to the logistic regression algorithm to reduce overfitting, especially given the large number of variables caused by the many features in the dataset. This made logistic regression suitable for this type of data.

c) XGBoost: XGBoost, known for its efficiency in gradient-boosting and cross-validation, was utilized to create an optimized model. The algorithm focused on improving performance and accuracy, making it the most suitable choice for this task.

The dataset was divided into training and testing subsets. The models were trained using the training data and then evaluated on the testing data to measure their performance. After assessing all three models, Random Forest stood out as the best performer, demonstrating the highest accuracy and robustness for the dataset.

3) Model Deployement: Once the Random Forest model was selected, it was serialized for deployment. The trained model was saved using joblib, creating a file called `best_model.pkl`. The model was then compressed into an archive file (best_model.tar.gz) for easy uploading. This archived model was uploaded to the mydeploy18 S3 bucket for deployment. Using SageMaker's SKLearnModel class, the model was deployed to an ml.t2.medium instance, where an endpoint was configured to handle real-time predictions. This deployment allowed the model to make predictions on new review data as it was received, enabling on-the-fly analysis and scoring of product reviews.

E. Amazon CloudWatch: Monitoring and Logging

Amazon CloudWatch was used to monitor the deployment and operations of the model, along with other AWS services such as Athena and S3. While custom dashboards weren't created, CloudWatch Logs captured essential information about the model's deployment, including endpoint status and any errors that occurred. This allowed for quick identification and resolution of issues, ensuring smooth operations. CloudWatch also monitored the execution of Athena queries and S3 bucket activities, offering critical logs and metrics for troubleshooting and optimization throughout the project.

IV. Results

The implementation successfully executed a machine learning pipeline on Amazon Web Services (AWS) cloud to give a feel for product review analyses. The project began with the setup of secure access through AWS Identity and Access Management (IAM) and went further to emphasize the need for granting permission based on the least privilege principle. To conduct queries including sample balancing, Amazon Athena was used in the preprocessing and querying of the large dataset stored in Amazon S3. This processed and balanced dataset was exported into Amazon SageMaker where multiple ML models such as Random Forest, Logistic Regression and XGBoost were trained and evaluated. In the end, the best performing algorithm, Random Forest, was trained and the model parameters were serialized, zipped and deployed on SageMaker real-time endpoint on ml.t2.medium instance. The model performed well as it was able to provide real time prediction of product review ratings based on the features of the new product as requested by the users. AWS CloudWatch was utilized in the task of collecting all the logs and for tracer monitoring of the model performance level in order to facilitate better operational efficiency and error management. In conclusion, the project utilized the services of AWS to ensure that data was readily available for modeling training and deployment describing a complete end to end to be real time analytics of product reviews.

References

[1] T. U. Haque, N. N. Saber, and F. M. Shah, ”Sentiment analysis on large scale Amazon product reviews,” in 2018 IEEE International Conference on Innovative Research and Development (ICIRD), Bangkok, Thailand, 2018, pp. 1-6.

[2] K.V. Kousik, M. AsishTony, K.S. Krishna, S. Narisety, and P. Yellamma, 2023, April. ”An E-Commerce Product Feedback Review using Sentimental Analysis,” In 2023 International Conference on Inventive Computation Technologies (ICICT) (pp. 608-613). IEEE.

[3] M. Hardt, X. Chen, X. Cheng, M. Donini, J. Gelman, S. Gollaprolu, J. He, P. Larroy, X. Liu, N. McCarthy and A. Rathi, 2021, August. ”Amazon sagemaker clarify: Machine learning bias detection and explainability in the cloud,” In Proceedings of the 27th ACM SIGKDD conference on knowledge discovery data mining (pp. 2974-2983).

[4] N.S. Chougule, C.J. Awati and R. Deshmukh, 2024, January. ”Using AWS SageMaker to Deploy ML Credit Card Fraud Detection Model,” In 2024 5th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI) (pp. 150-156). IEEE.

[5] M.K. Ravindranathan, D.S. Vadivu and N. Rajagopalan, 2024, January. ”Cloud-Driven Machine Learning with AWS: A Comprehensive Review of Services,” In 2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE) (pp. 1-8). IEEE.

[6] E. Liberty, Z. Karnin, B. Xiang, L. Rouesnel, B. Coskun, R. Nallapati, J. Delgado, A. Sadoughi, Y. Astashonok, P. Das and C. Balioglu, 2020, June. ”Elastic machine learning algorithms in amazon sagemaker,” In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (pp. 731-737).

[7] D. Nigenda, Z. Karnin, M. B. Zafar, R. Ramesha, A. Tan, M. Donini, and K. Kenthapadi, ”Amazon SageMaker Model Monitor: A System for Real-Time Insights into Deployed Machine Learning Models,” *28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Washington, DC, USA, 2022, pp. 1-11.

[8] J. Yeung, S. Wong, A. Tam, and J. So, ”Integrating Machine Learning Technology to Data Analytics for E-Commerce on Cloud,” *Third World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, 2019, pp. 105-109.

[9] S. Sharma and A. A. Waoo, ”Customer Behavior Analysis in E-Commerce using Machine Learning Approach: A Survey,” *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, vol. 10, no. 2, pp. 163-170.

[10] A. Singh, P. Kumar, and R. Verma, ”A Review on Customer Segmentation Methods for Personalized Customer Targeting in E-Commerce Use Cases,” *Proceedings of the International Conference on Artificial Intelligence and Applications*, pp. 450-456.

[11] P. Sharma, K. Gupta, and M. Rao, ”Novel Weight-Optimized LSTM for Dynamic Pricing Solutions in E-Commerce Platforms Based on Customer Buying Behavior,” *Proceedings of the International Conference on Machine Learning and Data Engineering*, pp. 120-125.