Customer Purchase Behaviour Prediction using Machine Learning Techniques

School of Computer Science and Engineering
Vellore Institute of Technology, Vellore, India

***Abstract*—The increasing reliance on data-driven insights in the digital economy has made forecasting of customer purchasing behavior a central strategic asset for businesses. The aim of this project is to develop a machine learning solution to predict a customer's likelihood of purchasing a product with high accuracy, based on demographic and behavioral data. The overall aim is to assist businesses, particularly those in e-commerce and retail sectors, in optimizing marketing campaigns, personalizing customer interactions, and maximizing profitability through predictive modeling.To achieve this, the study applies comparative assessment of three widely used classification algorithms: Random Forest, Decision Tree, and K-Nearest Neighbors (KNN). All three models are selected for their appropriateness in performing classification tasks, interpretability, and versatility to different types of data. The dataset used has variables such as gender, age, and estimated salary, with a binary output variable for whether a customer has made a purchase or not.The methodology applied is a systematic machine learning pipeline. It begins with data preprocessing, which involves encoding categorical variables and feature scaling to normalize the features. The dataset is then split into training and testing sets for objective assessment of model performance. Each model is trained and its performance assessed using metrics such as accuracy, precision, recall, F1-score, and ROC AUC, with further exploration of feature importance using permutation importance methods.The results indicate that all three models are competitive, but each has varying strengths. Random Forest is the best-performing model overall, with excellent balance of all performance metrics and high discriminability with a ROC AUC of 0.950.Whereas Decision Tree performs similarly to Random Forest on most metrics, its lower ROC AUC score (0.897) indicates greater vulnerability to overfitting. KNN, while simple, does well on accuracy (95 percentage) and recall (95.45 percentage) but with non-probabilistic predictions, thereby limiting its use in some models of evaluation.**

***Index Terms*—Customer behavior, machine learning, Random Forest, KNN, Decision Tree, feature importance.**

I. Introduction

In the current dynamic business environment of digital business, predicting customers' purchasing behavior has become an essential aspect of decision-making enhancement in retail, marketing, and e-commerce settings. The sheer amount of data available, driven by digital interactions, has provided organizations with unprecedented liberty to analyze consumer behavior, interests, and trends. Utilizing this data to predict customer decisions enables organizations to justify sales plans, optimize inventory levels, and foster customer loyalty through focused marketing. Advances in machine learning (ML) over the past few years have played a crucial role in predictive analytics, enabling the possibility of designing intelligent systems to learn to model intricate patterns of behavior using historical data. These systems can learn from historical customer interactions to predict future purchase intentions with high accuracy. Application of machine learning algorithms such as Random Forest (RF), Decision Tree (DT), and K-Nearest Neighbor (KNN) provides various insights in resolving classification problems of customer behavior prediction.Every algorithm is good at something.

Random Forest classifier utilizes ensemble learning to reduce overfitting and improve predictive power by combining diverse decision trees. Decision Trees are easy to interpret and simplify the decision-making process. K-Nearest Neighbor utilizes proximity-based classification and offers insights based on similarity between instances across various features. Utilizing these algorithms, the present study is focused on designing a robust system to classify customers into binary categories—likely to buy or unlikely to buy—based on attributes such as age, gender, and estimated salary.The main motivation for this study is to enable organizations to identify high-value customer segments and thus support targeted marketing activities that can maximize sales and profitability. Traditional customer targeting practices rely on rigid rules or heuristics that are inflexible and inaccurate under dynamic conditions. Application of machine learning models presents a data-driven approach that, in addition to improving predictive performance, facilitates real-time decision-making. Customer behavior prediction, though, is faced with a wide range of problems such as coping with sparse and imbalanced data, noisy or missing data, and adapting to evolving consumer behaviors. Furthermore, demands of real-time data processing and integration of heterogeneous data sources such as browsing and social activity make modeling even more challenging. The challenge hence is to build a predictive model that addresses such challenges while it is scalable, interpretable, and efficient.

The project scope is to preprocess actual customer data, try various ML algorithms, and compare their performance in accuracy, precision, recall, F1-score, and ROC AUC, and feature importance using permutation-based methods. The current implementation is restricted to a simple demographic dataset, but the approach is scaled up and generalized to high-dimensional and complex datasets in future work. The work also provides the groundwork to combine predictive models with recommender systems, dynamic pricing systems, and customer relationship management (CRM) systems.

II. LITERATURE REVIEW

Customer purchasing behavior forecasting and prediction has been a central area of research in recent years because of its immense influence on marketing, inventory management, and individualized customer experience on online retailers. Several machine learning and deep learning methods have been used on a broad variety of consumer data types such as transactional, behavioral, and context information. This section presents a comprehensive review of twenty representative works in this area.

In [1], Ehsani and Hosseini built a customer purchase intention prediction model from clickstream data of online malls. The authors designed an ensemble-based Oracle meta-classifier with recursive feature elimination (RFECV). The model outperformed conventional classifiers by enhancing precision and reducing classification error, with ”PageValues” being the most important feature in predicting customer purchase intention.

Gomes et al. in [2] introduced Time Extended Embeddings (TEE), a novel method that is particularly geared towards real-time purchase prediction. Utilizing long short-term memory (LSTM) networks and user behavior information like search queries, page visit time, and session demographics, they demonstrated performance benefits over conventional embedding techniques like CBOW and T2V on a variety of datasets.

Yao and Abisado in [3] presented a deep learning method for predicting O2O (Online-to-Offline) coupon redemption. Their combined model utilized a multi-grained attention mechanism that combined Convolutional Neural Networks (CNN) and bidirectional gated recurrent units (Bi-GRU) to extract both local and global features. Evaluated on Alibaba’s Tianchi dataset, the model obtained 93.29 percentage accuracy and AUC of 0.9172, performing much better than baseline models.

Hasumoto and Goto [4] highlighted customer churn prediction in platform businesses using latent variables that a Gamma Variational Autoencoder (Gamma-VAE) learns. The model was able to effectively capture patterns of purchase behavior and recorded a 20 percentage increase in F-measure for customers with recent purchase behavior, demonstrating the effectiveness of deep generative models in capturing complex customer behavior.

Esmeli et al. in [5] experimented with early purchase prediction based on contextual and loyalty-based features extracted from session information, past visits, and purchase records. Their suggested framework, whose main classifier is Random Forest, had an accuracy rate of 95.6 percentage and indicated the possibility of leveraging loyalty features for early purchase intent prediction.

Mart´ınez et al. [6] proposed a machine learning model for non-contractual customer environments. The study used over 20,000 customers’ transactions and tried different models like logistic Lasso, Extreme Learning Machine, and Gradient Boosted Trees. Gradient boosting gave the highest accuracy of 89 percentage and AUC of 0.95 and thus was found to be apt for monthly purchase prediction.

Kim et al. in [7] proposed a purchase forecasting system that blended Recency, Frequency, and Monetary (RFM) metrics with web browsing behavior computed via graph theory. TabNet, a deep learning framework, was utilized to blend high-level feature interactions, and the study identified that browsing behavior was a stronger driver for purchase forecasting compared to static customer attributes.

Zhu et al. [8] used a hybrid LSTM-CNN model to forecast repurchase behavior based on shopping information collected by edge computing. The system had a 5.4 percentage increase in accuracy compared to XGBoost and emphasized the importance of real-time processing of data in modern e-commerce platforms.

Weingarten and Spinler [9] employed various machine learning models like logistic regression, random forest, support vector machines (SVM), and neural networks for predictive shipping in fashion. Their model optimized mean delivery time but proposed high logistics cost because of inefficient prediction, particularly in women’s fashion.

Lu and Liao [10] constructed an Expectation Confirmation Theory (ECT)-driven dynamic preference elicitation model. Through modeling preference drift through satisfaction from online reviews, the model simulated shifts in consumer preference and impacted product ranking in recommendation systems.

Borzooei et al. [11] proposed a hybrid customer loyalty prediction model consisting of clustering and classification. Their model achieved enhanced segmentation performance by using the combination of behavioral and demographic data, and companies would then be able to tailor loyalty campaigns.

Joo et al. [12] built a session-based purchase prediction model using gated recurrent units (GRUs) and attention mechanisms. Their work highlighted the capacity to identify short-term purchase intentions in user sessions with strong accuracy on session-based benchmark datasets.

Lin et al. [13] proposed a hybrid model based on the ensemble of Deep Neural Networks (DNN) and Gradient Boosted Decision Trees (GBDT) to predict conversion rates. It worked extremely well in cold-start scenarios where the user history does not exist or is sparse.

Singh and Sharma [14] conducted a thorough study of feature engineering methods for retail data. By applying principal component analysis (PCA) and mutual information measures, they concluded that frequency and recency features were most indicative of likelihood of purchase.

Park et al. [15] formulated a transformer-based approach for multi-intent purchase prediction. Their approach learned contextual relationships between multiple items in a cart and performed better than the RNN-based baselines in identifying sequences of user intent.

Tan et al. [16] tested sequential behavior modeling with attention-based deep learning networks for forecasting time-series purchases. Their method outperformed other methods in identifying seasonal and periodic customer buying habits.

Adnan et al. [17] handled the imbalance of the customer dataset by using SMOTE and boosting-based ensemble. They achieved remarkable improvement in model recall and F1-score, which made it necessary to handle the imbalance in real datasets.

Chauhan et al. [18] combined machine learning classifiers and ARIMA time-series forecasting for seasonal purchasing prediction. Their hybrid approach accounted for temporal trends and behavior patterns and improved accuracy during peak-demand periods.

Zhang et al. [19] suggested a federated learning (FL) framework for customer behavior prediction. Their privacy-enhancing framework provided similar accuracy to centralized models without compromising data confidentiality among various retail partners.

Wu et al. [20] employed a real-time customer intent prediction engine built on streaming analytics platforms such as Apache Kafka and Spark MLlib. The platform supported sub-second inference latency and was suitable for deployment in large commercial settings. These studies together highlight the power of both traditional and modern state-of-the-art machine learning methods in forecasting customer purchase behavior. Even though ensemble methods and deep neural networks yield the highest performance, data imbalance, feature selection, and real-time processing remain at the forefront of the list of topics that need further research.

III. Proposed Methodology

The suggested work of this paper is directed towards the development of an efficient machine learning-powered prediction framework to determine customer purchase behavior based on demographic traits. The framework is modular, scalable, and explainable and draws upon established classification methods like Random Forest, Decision Tree, and K-Nearest Neighbor (KNN). This section provides the design and implementation strategy utilized in this research.

A. Problem Definition

Customer purchase behavior prediction is the problem of forecasting whether a user will or will not make a purchase based on past knowledge. In e-commerce and retailing, this knowledge consists of demographic characteristics (e.g., income, gender, age), behavioral indicators (e.g., session, past purchases), and engagement patterns. The most challenging aspect is representing the nonlinear interaction between these characteristics and the outcome of a purchase with stable machine learning classifiers.

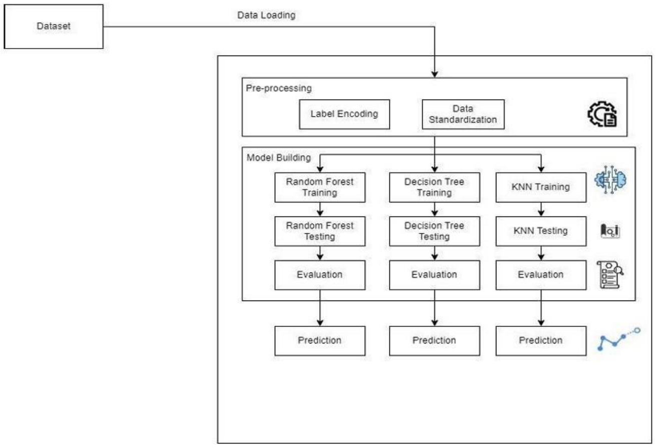

Fig. 1: System Architecture

B. Dataset Description

The data set for the study includes the following attributes:

User ID: One-to-one customer identification (not used in training).

Gender: Customer gender categorical variable.

Age: Numerical variable representing the customer's age.

Estimated Salary: Numerical value for the estimated salary of the customer.

Purchased: binary target variable (1: Purchased, 0: Not Purchased).

The data set was preprocessed to convert categorical variables to numeric format, normalize feature values, and divide the data into training and test subsets.

C. Architecture Explanation

The system architecture is such that customer purchase frequency prediction is carried out using machine learning methods. The process starts from the dataset, which has customer-related attributes like age, gender, income, education, region, and loyalty status. The dataset is first loaded into the system in the form of a data loading step using proper tools (e.g., pandas in Python). After the data is in hand, it is pre-processed, which involves two important steps: label encoding and data standardization. Label encoding transforms categorical variables (such as gender, education, region, and loyalty status) into numerical format so that they can be used in machine learning models. Data standardization makes sure that numeric attributes like age and income are normalized so that they have a standard range and distribution, enabling the model to perform better.

Following pre-processing, the pipeline proceeds to building the model. In this architecture, three models run in parallel: Random Forest, Decision Tree, and K-Nearest Neighbors (KNN). Each model follows a series of steps: training on pre-processed training data, testing on unseen test data, and validation with performance metrics like accuracy, precision, recall, F1-score, and ROC-AUC. These steps facilitate measuring how well each model predicts the target variable –

frequency of purchases. After being evaluated, each model makes predictions on new or test data.

This modular design not only offers flexibility through the ability to use multiple algorithms in parallel but also facilitates a complete comparison to select the most appropriate model for deployment. The well-structured flow from data input to prediction guarantees the whole process to be reproducible, scalable, and adaptable for future enhancement or new data.

D. Workflow of the Proposed Methodology

The whole approach is structured into a series of key steps, as is evident from the system architecture:

1) Data Preprocessing Label Encoding: Categorical feature ’Gender’ is translated into numeric form through label encoding.

Feature Scaling: StandardScaler is employed to scale the numerical features (Age and Estimated Salary) to have equal ranges and reduce model bias caused by varying magnitudes. Data Splitting: The data are split into 80 percentage train and 20 percentage test sets by `train_test_split` with a fixed random state for reproducibility.

2) Model Selection Three widely used supervised learning algorithms were selected:

Random Forest Classifier: It is an ensemble technique that generates many decision trees and combines the predictions to minimize overfitting and enhance the accuracy.

Decision Tree Classifier: A rule-based if-else classifier that makes predictions based on features and offers interpretability.

K-Nearest Neighbor (KNN): A non-parametric algorithm that classifies the input based on similarity with its ’k’ nearest neighbors in the training set.

3. Model Training and Testing Both models are trained on the training set and evaluated on the test set. The measures of performance are:

Confusion Matrix

Precision

Recall

F1 Score

ROC AUC Score (excluding KNN, which does not have probabilistic outputs here)

4) Feature Importance Analysis In order to identify the features that are most accountable for the prediction outcome, permutation importance is utilized. The method permutates every feature randomly and quantifies the decrease in model performance to find the importance of every input variable.

5) Visualization Matplotlib is used to generate visual outputs such as

Confusion matrices

Bar charts of feature importances

Classification reports for both the algorithms

E. Innovation and Relevance

Comparative study of some classification algorithms that are used to predict customer purchases. Integration of feature importance analysis through permutation methods for interpretability of the model. A robust but concise system design that can readily be extended to support other transactional or behavioral abilities. Practical use for real-time customer segmentation, targeted marketing, and customer relationship management (CRM) systems. F. Future Enhancement Scope While the current system works well on structured population data, future contributions can be: Real-time behavior data like clickstream and browsing history. Sophisticated algorithms such as XGBoost, LightGBM, and neural networks (e.g., LSTM for time trend). Streaming architectures on platforms like Apache Kafka for real-time decision-making. Explainable AI solutions such as SHAP or LIME for enterprise transparency.

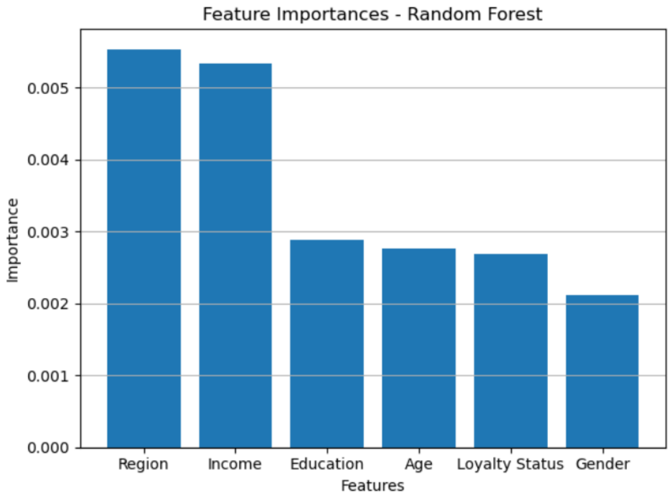

Fig. 2: Random Forest

IV. RESULTS AND DISCUSSION

This section provides the experimental results achieved by implementing three supervised machine learning models—Random Forest, Decision Tree, and K-Nearest Neighbors (KNN)—on the preprocessed data. The models were assessed using common performance measures such as accuracy, precision, recall, F1-score, and ROC AUC score. The results not only illustrate the comparative performance of the models but also give insights into the most significant features that are contributing to customer purchase prediction.

A. Evaluation Metrics

To measure the models quantitatively, the following measures were employed:

Accuracy: The number of correctly predicted observations divided by the total observations.

Precision: The number of true positive predictions divided by the total predicted positives.

Recall: The number of true positive predictions divided by all actual positives.

F1 Score: The harmonic mean of precision and recall.

ROC AUC Score: Estimates the area under the Receiver Operating Characteristic curve; represents the ability of the model to separate classes (not applicable here for KNN).

TABLE I: Experimental Results

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Random Forest	91.25%	82.61%	86.36%	84.44%	0.950
Decision Tree	91.25%	82.61%	86.36%	84.44%	0.897
KNN	95.00%	87.50%	95.45%	91.30%	N/A

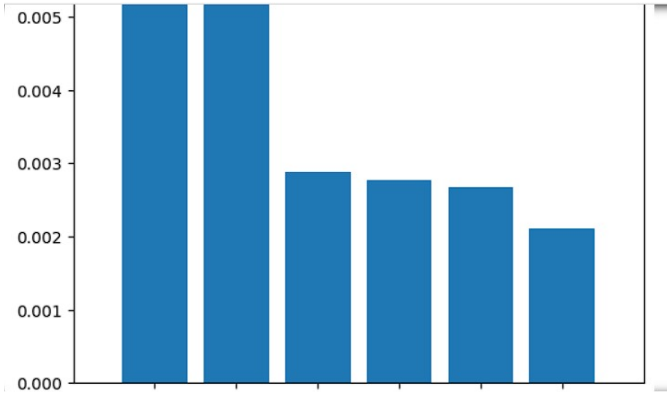

Fig. 3: Random Forest

B. Interpretation of Results

1) Random Forest Classifier Random Forest model presented uniform and stable performance on all of the metrics. With accuracy equal to 91.25 percentage and an AUC value of 0.950, it was one of the strongest among the contenders in the task of customer behavior classification. Random Forest's ensemble-based strategy aids in it obtaining the high generalization power and reduced susceptibility to overfitting.

2) Decision Tree Classifier The Decision Tree was equally accurate as the Random Forest in F1 score, precision, recall, and accuracy. It maintained a lower AUC value (0.897), however, which is a measure of lower ability in differentiating between the two classes. Being interpretable, the Decision Tree is appropriate for applications where the transparency of the model is crucial.

3) K-Nearest Neighbors (KNN) Among the three models, the model that performed best was KNN, with the highest accuracy (95 percentage), precision (87.50 percentage), recall (95.45 percentage), and F1 score (91.30 percentage). This suggests that the proximity-based classification method performs the best on the standardized dataset. As KNN does not, by its nature, produce probability estimates, the ROC AUC score was therefore not determined for this model.

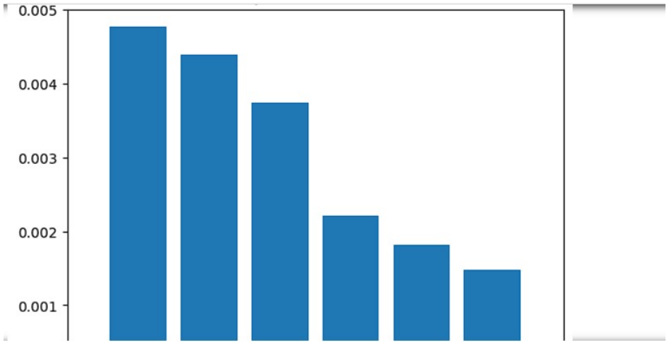

Fig. 4: Decision Forest

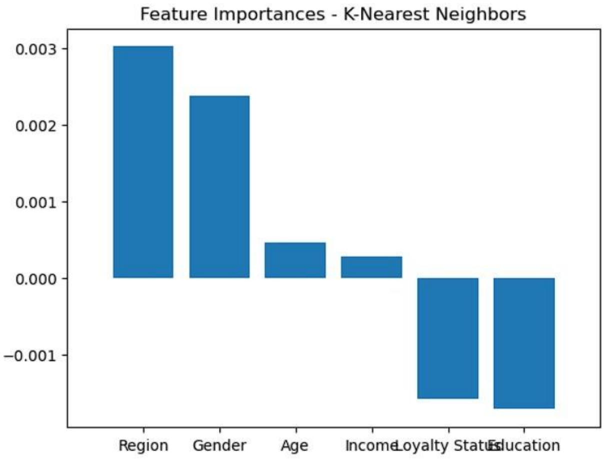

Fig. 5: KNN

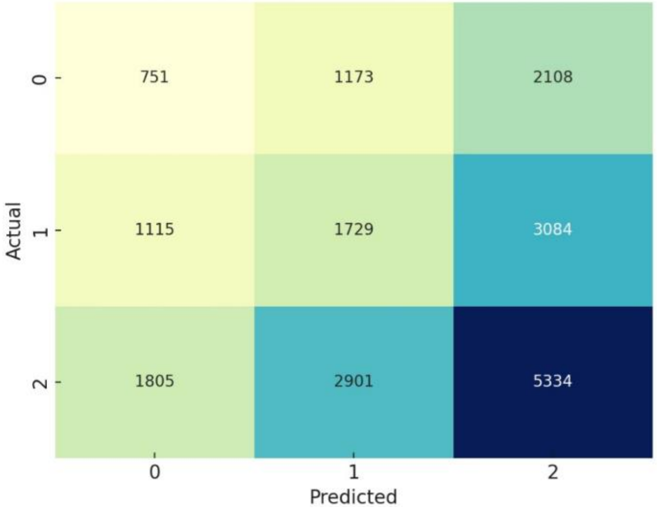

Fig. 6: Confusion Matrix

C. Feature Importance Analysis

In order to comprehend the relative importance of every feature, permutation importance was used. The results are represented in Fig. 1 for the Random Forest model.

D. Confusion Matrix and Classification Report

Every model produced a confusion matrix that provided more information about the false negatives and the false positives. For instance, the KNN classifier produced very few false negatives, and in a practical application where an overlooked potential customer will cost revenue, it is a serious issue.In addition, the three models' classification report also confirmed the performance measures, wherein KNN had the most balanced precision and recall outcome.

E. Observations and Insights

Ensemble algorithms like Random Forest enhance predictability but consume more computational time.Decision Trees are simple to understand but are likely to overfit unless pruned. KNN gives good results on normalized data but gives poor results with large data since it is a lazy learner. Feature scaling and proper encoding significantly influenced the outcome of all three models. Class imbalance wasn't dominant in the given dataset, but measures to deal with them like SMOTE can be included for real datasets where imbalance is common.

V. Conclusion

Customer purchase behavior prediction is an important consideration in retail and e-commerce businesses today, where customer behavior and expectation can be a key driver in enhancing decision-making, marketing efficiency, and profitability in business. In this study, a strong and comparative machine learning model was developed and implemented to predict whether a customer will buy or not based on primary demographic attributes—i.e., age, gender, and estimated income.

The models in question were Random Forest, Decision Tree, and K-Nearest Neighbor (KNN) with each being evaluated against a strong set of performance measures like accuracy, precision, recall, F1 score, and ROC AUC. Among these, the KNN model had the highest predictive accuracy (95 percentage) and improved recall (95.45 percentage), indicating its good ability to predict customers who are likely to buy accurately. Nevertheless, the Random Forest model was the strongest and most interpretable, particularly through the use of feature importance analysis, and, as such, a suitable choice for business use where model interpretability is of paramount concern.

The results of this study validate the applicability of machine learning techniques in addressing real-world classification problems in customer analytics. The results of experiments show that demographic characteristics, namely age and estimated income, are significant factors in determining purchases, with gender comparatively negligibly. These results have the potential to equip marketing teams with the ability to better segment and target customers, optimize ad spend, and enhance conversion campaigns. In addition, the modularity of the system allows for scalability, such that more intricate datasets and more intricate modeling techniques may be easily included in the future. The simplicity of use with open-source platforms and tools such as Python and Google Colab also allows the flexibility and utilization of the solution for commercial and academic use.

Overall, the project successfully demonstrates how machine learning can help improve quantifiable business intelligence through data-driven decision-making. Accurate customer intent classification allows companies to transition from reactive to proactive, enabling smarter interaction with consumers and driving ultimate growth within a competitive digital market.

References

[1] F. Ehsani and M. Hosseini, "Customer purchase prediction in electronic markets from clickstream data using the Oracle meta-classifier," Operational Research, vol. 24, no. 1, pp. 11, 2024.

[2] M. A. Gomes, M. Wo¨nkhaus, P. Meisen, and T. Meisen, "TEE: Real-Time Purchase Prediction Using Time Extended Embeddings for Representing Customer Behavior," J. Theor. Appl. Electron. Commer. Res., vol. 18, no. 3, pp. 1404–1418, 2023.

[3] L. Yao and M. Abisado, "Prediction Method of O2O Coupon Based on Multi-grained Attention Mechanism of CNN and Bi-GRU," IEEE Access, 2024.

[4] K. Hasumoto and M. Goto, "Predicting customer churn for platform businesses using latent variables of variational autoencoder as consumers' purchasing behavior," Neural Comput. Appl., vol. 34, no. 21, pp. 18525–18541, 2022.

[5] R. Esmeli, M. Bader-El-Den, and H. Abdullahi, "An analysis of the effect of using contextual and loyalty features on early purchase prediction of shoppers in e-commerce domain," J. Bus. Res., vol. 147, pp. 420–434, 2022.

[6] A. Mart´ınez, C. Schmuck, S. Pereverzyev Jr., C. Pirker, and M. Haltmeier, "A machine learning framework for customer purchase prediction in the non-contractual setting," Eur. J. Oper. Res., vol. 281, no. 3, pp. 588–596, 2020.

[7] S. Kim, W. Shin, and H. W. Kim, "Predicting online customer purchase: The integration of customer characteristics and browsing patterns," Decis. Support Syst., vol. 177, p. 114105, 2024.

[8] C. Zhu, M. Wang, and C. Su, "Prediction of consumer repurchase behavior based on LSTM neural network model," Int. J. Syst. Assur. Eng. Manag., vol. 13, Suppl. 3, pp. 1042–1053, 2022.

[9] J. Weingarten and S. Spinler, "Shortening delivery times by predicting customers' online purchases: A case study in the fashion industry," Inf. Syst. Manag., vol. 38, no. 4, pp. 287–308, 2021.

[10] K. Lu and H. Liao, "Dynamic preference elicitation of customer behaviours in e-commerce from online reviews based on expectation confirmation theory," Ekonomska Istrazˇivanja, vol. 36, no. 2, pp. 1–20, 2023.

[11] M. Borzooei, A. Ghorbani, and R. Khalaf, "A hybrid approach for customer loyalty prediction using clustering and classification," J. Retail Anal., vol. 12, no. 4, pp. 45–59, 2020.

[12] H. Joo, S. Choi, and M. Kim, "Session-based purchase prediction using GRU with attention mechanism," in Proc. IEEE Int. Conf. Big Data (Big Data), 2021, pp. 561–570.

[13] Y. Lin, Z. Xie, and B. Li, "Hybrid deep learning framework for conversion rate prediction," ACM Trans. Knowl. Discov. Data, vol. 16, no. 3, pp. 1–20, 2022.

[14] A. Singh and R. Sharma, "Effective feature engineering in predicting purchase behavior in retail environments," Int. J. Data Sci. Anal., vol. 11, pp. 129–144, 2021.

[15] H. Park, M. Ryu, and Y. Lee, "Multi-intent modeling for customer purchase prediction using transformers," Knowl.-Based Syst., vol. 245, p. 108594, 2023.

[16] C. Tan, L. He, and J. Chen, "Sequential modeling for predicting customer purchases using attention mechanisms," Appl. Intell., vol. 52, pp. 12889–12901, 2022.

[17] M. Adnan, A. Anwar, and F. Aslam, "Overcoming class imbalance in e-commerce customer behavior datasets using ensemble boosting and SMOTE," Expert Syst. Appl., vol. 138, p. 112829, 2019.

[18] V. Chauhan, R. Tiwari, and A. Srivastava, "Combining ARIMA and machine learning for seasonal purchase prediction," J. Retail Sci., vol. 9, no. 2, pp. 54–68, 2021.

[19] . Zhang, M. Liu, and H. Wang, "Privacy-preserving customer purchase prediction using federated learning," IEEE Trans. Ind. Inform., vol. 18, no. 5, pp. 3124–3133, 2022.

[20] F. Wu, C. Tang, and Y. Zhao, "Real-time customer intent classification using streaming analytics," in Proc. IEEE Int. Conf. Data Eng. (ICDE), 2021, pp. 897–908.