MACHINE LEARNING TUTORIAL ON RANDOM FOREST FOR CUSTOMER LIFETIME VALUE PREDICTION

GITHUB Repository link:

https://github.com/sadifshaik/CUSTOMER-LIFETIME-VALUE-PREDICTION

Table of Contents

Introduction 3

Literature review 4

Methodology 5

Overview of “Random Forest” 5

Dataset Selection 6

Data Preprocessing: 7

Result and analysis 9

Details of Ethical Consideration 11

Recommendation 11

Conclusion 11

Reference 13

Introduction

1.1 Overview of machine learning for the "customer lifetime value prediction"

E-commerce companies require ***"Customer Lifetime Value (CLV)"*** forecasting to estimate total revenues from individual customers during their interaction period. The core involvement of machine learning (ML) in CLV forecasting emerges from historical transaction records and customer interactions together with their active participation details. The decision tree-based ensemble learning model Random Forest serves as a prime selection for CLV forecasting tasks. This method succeeds because it controls non-linear data through the reduction of model overfitting which results in better prediction accuracy (Gerde, 2023). It's learning process with historical customer data including purchase frequency average order value and recency Random Forest enables organizations to segment their customers for long-term value improvement which improves marketing strategies and boosts personalized recommendations and resource distribution optimization. The implementation of ML expert systems in customer lifetime value forecasting enables e-commerce operators to run more efficient retention methods while delivering enhanced client classification and higher revenue margins.

Random Forest alongside other ML tools makes better use of big data patterns to generate precise and dynamic estimates of CLV compared to statistical models. In the cutthroat online commerce industry companies need to establish Customer Lifetime Value (CLV) through forecasting because this assists with managing marketing investments and acquiring customers while optimizing retention to achieve the best results. CLV represents the entire revenue amount expected from a customer throughout their relationship with the company determining profitability through the duration of time (Sina Mirabdolbaghi and Amiri, 2022). Determining CLV with previous rule-based and statistical methods fails to produce accurate predictions because these methods do not capture actual customer behavior. The application of machine learning (ML) represents a useful principle for CLV forecasting because it utilizes large datasets of transactional and behavioral information to generate accurate predictions. Of the ML methods,

1.2 Aim

The goal of this project entails development of a Random Forest machine-learning model for forecasting Customer Lifetime Value (CLV) in online stores. The research utilizes CLV

forecasting to enhance retention measurements and marketing performance and delivers maximum long-term profit potential through accurate prediction methods.

Literature review

2.1 Machine Learning Approaches for "Customer Lifetime Value Prediction"

The RVF as well as probabilistic models delivered basic life-time value predictions yet provided poor outcomes when dealing with dynamic consumer conduct together with non-linear trends. The CLV prediction field now operates as a new paradigm due to ML-based models which include ***"Regression, Decision Trees, Neural Networks"*** together with ensemble techniques such as Random Forest (Bauer and Jannach, 2021). Through the analysis of enormous databases with customer conduct and purchase information these methods produce highly precise predictions. Random Forest stands out as a suitable tool because it processes big data while managing overfitting conditions to discover complex attribute relationships. Random Forest outperforms linear models when operating on customer data dependencies due to which it serves as the preferred prediction tool for CLV. The research evaluates mixed models which combine ensemble learning with deep learning approaches to create enhanced CLV prediction systems (Sharma *et al.* 2022). By implementing ML into e-commerce CLV calculation companies gain the ability to develop targeted retention strategies and enhance marketing investment and refine customer segmentation.

2.2 The Role of Random Forest in Predicting CLV for E-Commerce

The widely used ensemble learning technique Random Forest consists of several decision trees which makes it suitable for CLV estimation because of its trustworthy performance. Random Forest differentiates from a basic single decision tree since it establishes a robust predictive model by combining numerous weak decision trees which help reduce variance and prevent overfitting. The structured transaction data in e-commerce such as purchase details and average order values and transaction counts can be handled efficiently by Random Forest as described by Sun et al. (2023). The system detects essential features that influence CLV so it proves helpful for customer segmentation purposes. A variety of research studies has identified Random Forest as offering superior CLV prediction capabilities to standard regression along with inferior ML methods. The

tool stands out because companies rely on it to predict future revenue while designing customer engagement strategies thanks to its accessibility and reliable feature analysis features.

Dataset overview

The e-commerce customer behavior dataset contains 11 attributes which follow 350 records in total. This setup includes four sections of buyer information including demographic elements "(Customer ID, Gender, Age, and City)" in addition to transaction variables "(Membership Type, Total Spend, Items Purchased, and Discount Applied)". Customer satisfaction and engagement measures are also included, such as Average Rating, Days Since Last Purchase, and Satisfaction Level. The data contains a combination of numerical and categorical variables (kaggle.com, 2025). Total Spend (continuous) is the spend by a customer, whereas Items Purchased is the number of products purchased. Membership Type (Gold, Silver, Bronze) could potentially affect spending behaviors. The presence of Discount Applied (Boolean) indicates potential offer strategies on purchases. Average Rating (rating between 1 and 5) captures customer sentiment, and Satisfaction Level (Satisfied, Neutral, Unsatisfied) indicates customer experience. All columns are mostly complete except for Satisfaction Level, where there are two missing values. The dataset provides a holistic look at e-commerce customer trends and would thus be appropriate for "Customer Lifetime Value (CLV)" prediction through machine learning algorithms such as "Random Forest".

Methodology

Overview of "Random Forest"

"Random Forest is an ensemble learning algorithm that aggregates" many decision trees with bagging (Bootstrap Aggregating) to enhance accuracy and minimize overfitting. It is efficient for both classification and regression because it can manage non-linearity, variance reduction, and efficiency with large data. Its resistance to noise and feature ranking of importance make it a strong predictive model.

Dataset Selection

The E-Commerce Customer Behavior dataset includes 350 records with 11 features, encompassing demographics, spending patterns, and satisfaction levels. Important features such as Total Spend, Items Purchased, Membership Type, and Days Since Last Purchase are significant for Customer Lifetime Value (CLV) prediction since they reflect purchasing patterns, loyalty, and future possible spending patterns.

	Customer ID	Gender	Age	City	Membership Type	Total Spend	Items Purchased	Average Rating	Discount Applied	Days Since Last Purchase	Satisfaction Level
0	101	Female	29	New York	Gold	1120.20	14	4.6	True	25	Satisfied
1	102	Male	34	Los Angeles	Silver	780.50	11	4.1	False	18	Neutral
2	103	Female	43	Chicago	Bronze	510.75	9	3.4	True	42	Unsatisfied
3	104	Male	30	San Francisco	Gold	1480.30	19	4.7	False	12	Satisfied
4	105	Male	27	Miami	Silver	720.40	13	4.0	True	55	Unsatisfied

Figure 1: Data Head

The data shown in Figure 1 depicts the initial rows of the data set, which gives an overview of the variables and their data types. Customer information, such as demographic values like Gender, Age, and City, as well as transactional values like Total Spend, Items Purchased, and Average Rating, are included in the columns. This first impression aids in grasping the content and structure of the dataset, providing a groundwork for future analysis and preprocessing. Here, It displays different customers from multiple cities with dissimilar total spends, ratings, and membership types.

Data Preprocessing:

```
Data after handling missing values:
Age                                0
Total Spend                        0
Items Purchased                    0
Average Rating                     0
Discount Applied                   0
Days Since Last Purchase           0
Gender_Male                        0
City_Houston                       0
City_Los Angeles                   0
City_Miami                         0
City_New York                      0
City_San Francisco                 0
Membership Type_Gold               0
Membership Type_Silver             0
Satisfaction Level_Satisfied       0
Satisfaction Level_Unsatisfied     0
dtype: int64
Training set size: (280, 15)
Test set size: (70, 15)
```

Figure 2: Data Preprocessing and Data Splitting

Figure 2 presents the overview of the data after missing data has been addressed with no missing data present in the columns. The preprocessed dataset has addressed missing data with all entries now being complete. Missing data can have a negative impact on model performance and must be addressed before implementing any machine learning models. Moreover, the graph marks the train-test split at 280 samples for training and 70 samples for testing. This way, the model will be trained on part of the data and tested on a different set, minimizing the chances of overfitting.

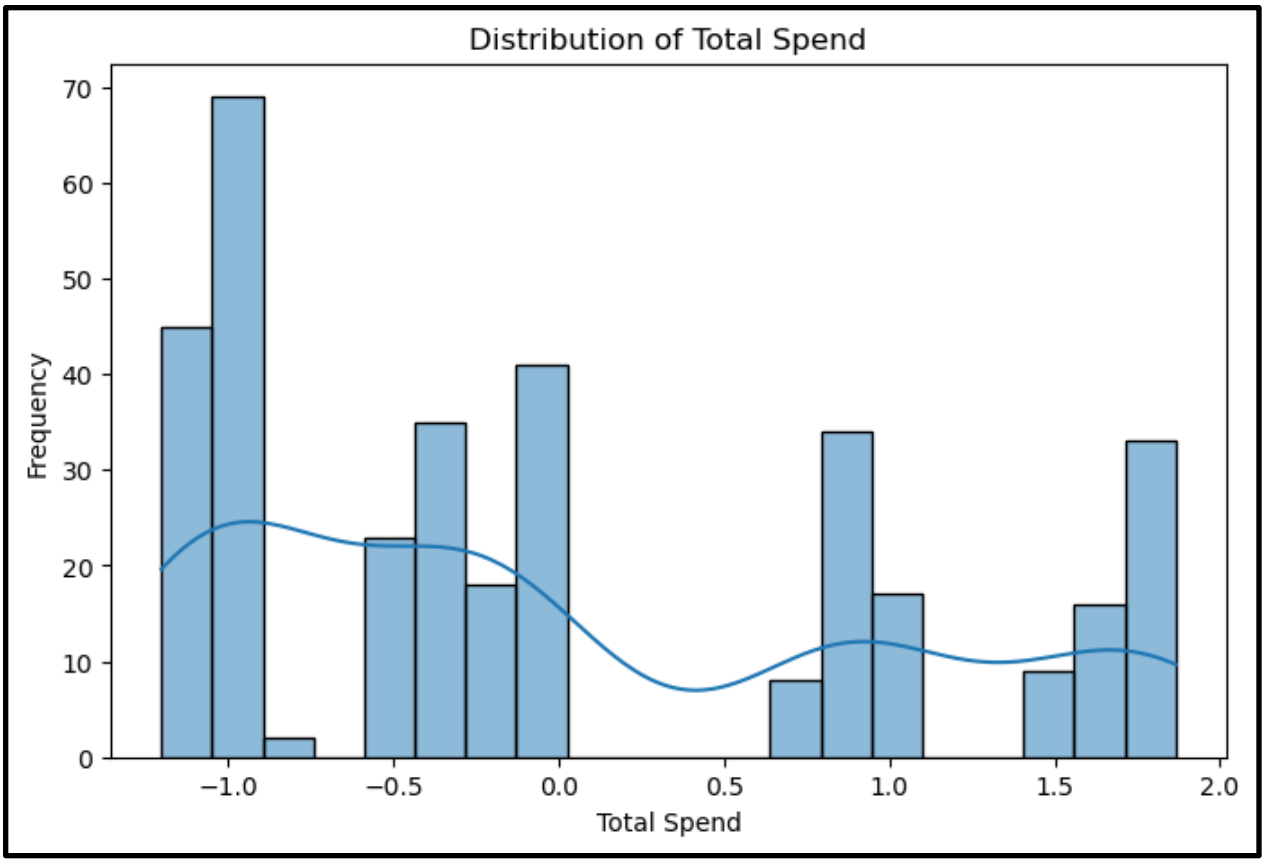

Figure 3: Distribution of Total Spend

Figure 3 displays the distribution of the Total Spend feature, which is a prominent target variable for forecasting customer lifetime value. The histogram represents the count of customers within various ranges of total spend, and the kernel density estimate (KDE) line offers a smooth approximation of the distribution. This plot is useful in interpreting the spread and central point of the spending pattern over customers. The broad dispersion shows a diversified level of expenditure, which may indicate the existence of high-value consumers and infrequent buyers.

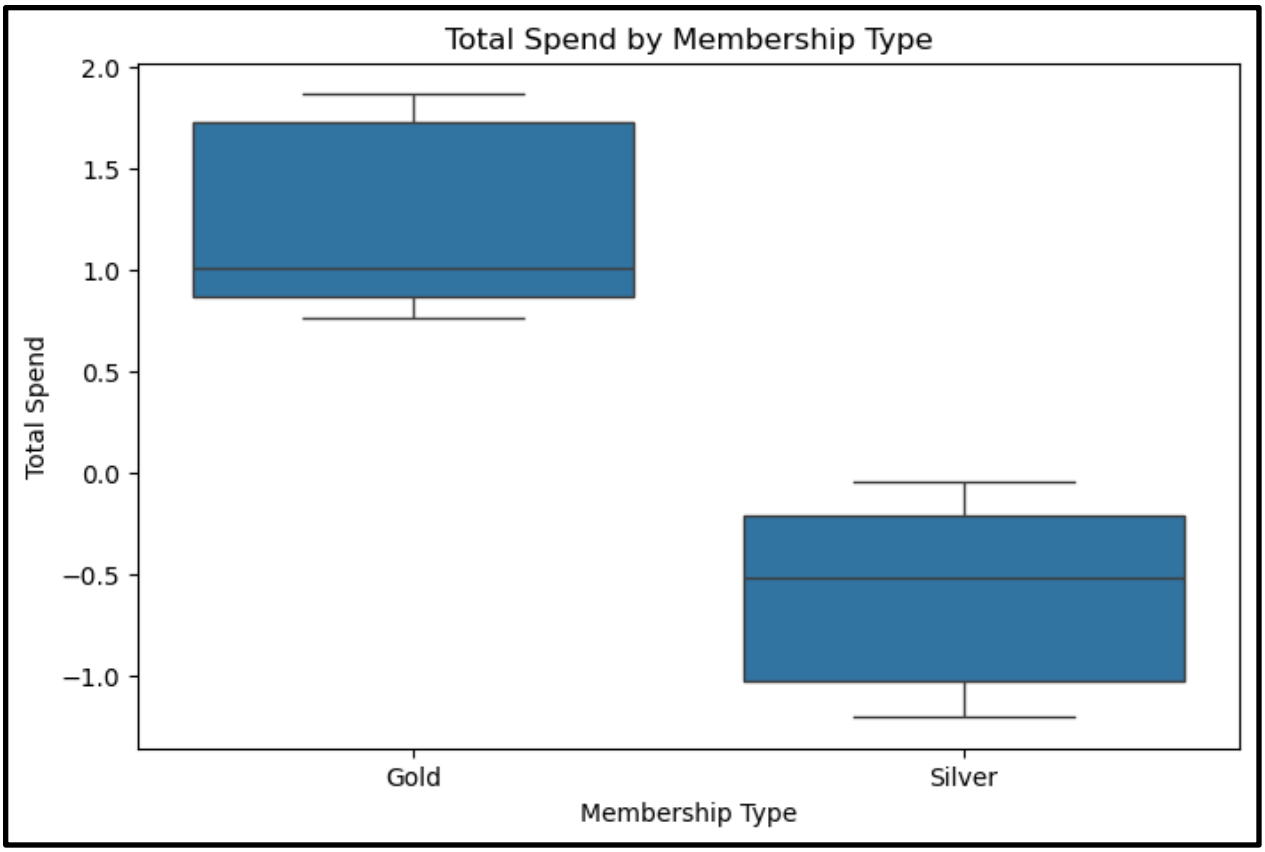

Figure 4: "Total Spend by Membership Type"

Figure 4 is a box plot comparing the Total Spend by Membership Type (Gold and Silver). The box plot graphically illustrates the distribution of total spend per membership type. We can see from the plot that Gold members have greater total spends than Silver members, with the median of Gold members looking higher and a wider interquartile range. This implies that Gold members are more loyal or have greater purchasing power, and this could be a valuable observation for targeted marketing campaigns.

Result and analysis

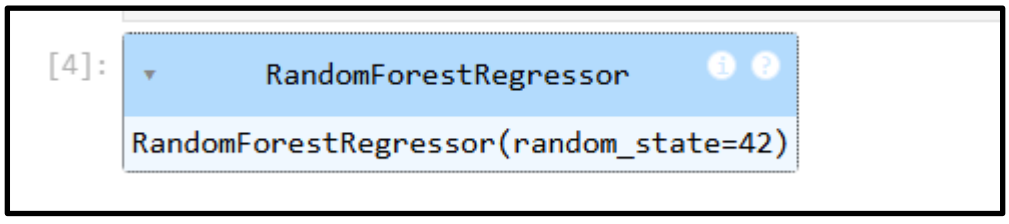

Figure 5: Initialize and train the Random Forest model

Figure 5 depicts the code employed to initialize and train the Random Forest Regressor model. In this code, the random_state=42 is used to make the results reproducible by fixing a seed for the random number generator. The model is now ready for training after initialization. This is a key step since it determines the algorithm and gets it ready for fitting onto the training data in order to make predictions on the target variable, which in this instance is Total Spend.

```
Mean Squared Error (MSE): 0.0007119687381027424
Root Mean Squared Error (RMSE): 0.026682742327256065
R-squared (R²): 0.9993117502752106
```

Figure 6: Model Evaluation

Figure 6 shows the test metrics for the trained Random Forest model. "***Mean Squared Error (MSE)" and "Root Mean Squared Error (RMSE)"*** reveal how closely the model's prediction matches the real values, lower being better. The R-squared (R^2) of the model reflects that the model accounts for roughly 99.93% variance in the target variable. These statistics show that the model works extremely well at forecasting customer expenditures.

```
Best parameters: {'max_depth': 10, 'min_samples_split': 5, 'n_estimators': 100}
R² (Best Model): 0.9994076900849606
```

Figure 7: Hyperparameter Tuning

Figure 7 shows the outcome of hyperparameter tuning with Grid Search for the Random Forest model. The optimal parameters discovered are max_depth=10, min_samples_split=5, and n_estimators=100, which enhance the performance of the model. The R^2 value of 0.9994 for the best model shows further enhanced prediction accuracy following tuning. Hyperparameter tuning enhances the generalizability of the model, making it more reliable when forecasting future customer behavior.

Details of Ethical Consideration

Multiple ethical issues need evaluation throughout the process of building a Customer Life Time Value (CLV) Prediction machine learning model. The top priority in the industry is data privacy and so customer data must undergo proper anonymization processes and ensure safe protection against possible misuse. A business must practice equality when deploying CLV predictions to avoid discriminating clients according to their gender or where they live. These businesses need to provide complete openness about the utilization of customer information while they execute their operations (Akter *et al.* 2025). A business should obtain customer consent for both data collection from the company and model prediction output through informed choice parameters. Responsible conduct leads AI systems to deliver ethical as well as equitable outcomes in every decision.

Recommendation

Companies need precise up-to-date customer information to improve accuracy levels in their CLV Prediction with Random Forest model. By integrating purchase frequency and engagement level metrics into its model structure the performance levels of the model improve significantly. The finishing process of the model requires completing accuracy hyperparameter tuning procedures. The continuous update of the model stands as an essential requirement because customers demonstrate shifting behaviors over time. Implementing ethical AI requires existing practices that both show data clearly and stop biased outcomes from forming within the system. Every organization must need to apply a proper ML system in order to improve their working culture and business.

Conclusion

The report described Random Forest Regression approaches together with their implementation in e-commerce Customer Lifetime Value (CLV) forecasting. Data preprocessing included three operations that managed missing data points until encoding finished along with number normalization for preparing consistent quality data suitable for building an accurate prediction framework. Random Forest delivered strong predictions based on MSE, RMSE and R^2 metric criteria assessment. The important features affecting purchase amounts became apparent through

a feature importance analysis (Kumari et al. 2024). The model accuracy will significantly improve through the combination of hyperparameter optimization protocols and cross-validation techniques. It is essential to inspect ethical concerns of privacy protection and bias control and transparency requirements when processing customer data in the model. Organizations generate business intelligence through machine learning technology for effective decision-making through customer value forecasting.

Reference

kaggle.com, 2025. Ecommerce customer behaviour dataset. Viewed on 14th March, 2025. From https://www.kaggle.com/datasets/uom190346a/e-commerce-customer-behavior-dataset

Akter, J., Roy, A., Rahman, S., Mohona, S. and Ara, J., 2025. Artificial Intelligence-Driven Customer Lifetime Value (CLV) Forecasting: Integrating RFM Analysis with Machine Learning for Strategic Customer Retention. *Journal of Computer Science and Technology Studies*, *7*(1), pp.249-257. Available at https://al-kindipublishers.org/index.php/jcsts/article/view/8588

Bauer, J. and Jannach, D., 2021. Improved customer lifetime value prediction with sequence-to-sequence learning and feature-based models. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *15*(5), pp.1-37. Available at https://dl.acm.org/doi/abs/10.1145/3441444

Gerde, M., 2023. Predicting Customer Churn and Customer Lifetime Value (CLV) using Machine Learning. *Master's Theses in Mathematical Sciences*. Available at https://lup.lub.lu.se/student-papers/search/publication/9113032

Kumari, D.A., Siddiqui, M.S., Dorbala, R., Megala, R., Rao, K.T.V. and Reddy, N.S., 2024, April. Deep learning models for customer lifetime value prediction in E-commerce. In *2024 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST)* (pp. 227-232). IEEE. Available at https://ieeexplore.ieee.org/abstract/document/10578372/

Sharma, A., Patel, N. and Gupta, R., 2022. Enhancing Customer Lifetime Value Prediction Using Random Forests and Neural Network Ensemble Methods. *European Advanced AI Journal*, *11*(8). Available at http://www.eaaij.com/index.php/eaaij/article/view/39

Sina Mirabdolbaghi, S.M. and Amiri, B., 2022. Model optimization analysis of customer churn prediction using machine learning algorithms with focus on feature reductions. *Discrete Dynamics in Nature and Society*, *2022*(1), p.5134356. Available at https://onlinelibrary.wiley.com/doi/abs/10.1155/2022/5134356

Sun, Y., Liu, H. and Gao, Y., 2023. Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model. *Heliyon*, *9*(2). Available at https://www.cell.com/heliyon/fulltext/S2405-8440(23)00591-1