International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

Evaluating Predictive Models for Reducing E-Commerce Returns and Associated Carbon Emissions

Ritesh Kalidindi[1], Leelavathy Narkedamilly[2], Uma Meghana Saladi[3]

[1]Student, International Baccalaureate Diploma Program, International School of Hyderabad, Hyderabad 500034, India
Email: *ritesh.k[at]ggu.edu.in*

[2]Professor of Computer Science and Engineering, Department of Computer Science & Engineering, Godavari Global University, Rajamahendravaram 533296, India
Email: *drnleelavathy[at]ggu.edu.in*

[3]Senior Consultant, Oracle Health Care, WA 98052, USA
Email: *umameghanasaladi[at]gmail.com*

Abstract: ***Product returns generate a disproportionate share of e-commerce's environmental footprint through repeated reverse-haul transportation, repackaging, and, for a meaningful share of items, outright disposal, which makes predictive models for identifying high-return-risk orders a potentially valuable lever for cutting both return volume and the carbon emissions that follow from it. This paper evaluates seven predictive models - logistic regression, decision tree, random forest, gradient boosting, SVM, k-nearest neighbours, and a neural network - for checkout-stage return-risk prediction, extending an earlier real- data evaluation on the BADS German fashion-retailer dataset (best model ROC-AUC = 0.574) with an independent re-evaluation of the full pipeline - feature engineering, model benchmarking, a feature-group ablation study, feature-importance analysis, and a literature-grounded carbon-emissions estimator - on a second, structurally different Kaggle dataset: a 10,000-row synthetic e-commerce returns dataset in which each customer places exactly one order. On this dataset, every predictive model evaluated produced a held-out ROC-AUC between 0.482 and 0.498, statistically indistinguishable from chance (0.50), and the ablation study confirmed that no feature group discriminated meaningfully above chance either. Rather than a weakness, this outcome functions as a negative control on the evaluation pipeline itself: it shows that the same models and procedure do not manufacture spurious predictive skill when no genuine return- risk signal is present in the underlying fields, which strengthens confidence that the earlier positive evaluation on real BADS order data reflected a genuine, learnable relationship rather than a pipeline artefact. A full carbon-emissions reduction calculator is maintained, and reported to ensure methodological completeness, and translates model recall into projected CO2-equivalent and packaging- waste savings over a set of intervention-effectiveness and emissions conditions, but the output is explicitly assigned the caveat of being an illustrative, but not a validated deployable projection of this dataset, until it is combined with a demonstrably-beating model. This paper has a fully reproducible Google Colab notebook that downloads the dataset and recreates all figures and tables reported end-to-end.***

Keywords: Product Return Prediction; Machine Learning; E-Commerce Sustainability; Ablation Study; Negative Control; Reproducibility

1. Introduction

Online retailing has become one of the biggest platforms of international trade and as the retailing has increased so has the rate of the same which is even higher than the rate that is witnessed in the physical stores. Where retailers with brick-and-mortar stores usually absorb single-digit returns, online clothes and shoe categories commonly experience returns in the twenty- to forty-percent range, as the customer has no opportunity to try on a product, brands have varying sizing, and the retailer employs zero-friction returns policies as a competitive requirement [1,2]. All those returns are not merely a reversion to a transaction, they initiate the second physical trip to the supply chain, a repacking process, a quality check-up, and, in a significant portion of cases, a disposal, instead of re-selling.

The environmental cost of this reverse flow has, until recently, remained largely hidden from view. Industry estimates attribute as much as twenty-four million metric tons of carbon dioxide emissions annually to e- commerce returns in the United States alone, with transportation accounting for the overwhelming majority of that footprint [3,4]. Reverse logistics activity is estimated to contribute roughly a quarter of the total carbon footprint of fashion e-commerce specifically [5], a proportion that makes returns reduction one of the more consequential, yet comparatively under-examined, levers available to sustainability-minded retailers.

An earlier iteration of this work benchmarked checkout-stage return prediction on the BADS dataset, real order-level records from a German online fashion retailer, and reported a best-model ROC-AUC of 0.574 together with an ablation study showing that product attributes carried the largest share of independent predictive signal. This paper asks a complementary methodological question: does the same pipeline behave correctly when it is pointed at a dataset that does not contain a genuine, learnable return-risk relationship? Verifying that a modelling pipeline reports the absence of signal when none is present is as important as verifying that it detects signal when it is present - a pipeline that reports spuriously high accuracy on unstructured or randomly generated data cannot be trusted when it reports a positive result on real data either. This paper re-runs the full pipeline, unmodified in its logic, on a second, independently

sourced Kaggle dataset - a 10,000-row synthetic e-commerce returns dataset - and reports the outcome honestly, whatever that outcome turns out to be.

This paper makes four contributions. First, it re-implements the full benchmarking, ablation, and environmental-impact-estimation pipeline against a second, structurally different returns dataset. Second, it benchmarks the same seven classification algorithms used in the original study under an identical evaluation protocol. Third, it re-runs the feature-group ablation study, adapting the feature groups to the fields actually available in the new dataset. Fourth, it reports and interprets a null result as a negative- control finding, and discusses what that finding does and does not imply for the earlier real-data result. Every reported number in Section 5 is generated by the accompanying Google Colab notebook (Appendix A), which downloads the dataset and regenerates all figures and tables end-to-end.

2. Literature Survey

Research on predicting product returns using machine learning is comparatively young relative to adjacent fields such as churn or fraud prediction, but has grown steadily over the preceding decade. Urbanke et al. presented one of the earliest applications of machine learning specifically for product-return prediction, working with more than 1 million purchase records from a German fashion retailer and introducing a Mahalanobis-distance-based feature-extraction step to improve baseline predictive accuracy [6]. Subsequent work extended the problem from predicting whether an individual order would be returned to forecasting aggregate return volume at the product and retailer levels, arguing that operational planning required both order-level and volume-level views [7].

More recent studies have moved toward comparative benchmarking similar in spirit to the present paper. A study comparing the logistic regression, random forest and gradient boosting using e-commerce order data identified that gradient boosting performed the best ROC-AUC compared to both logistic regression and random forest and that the predictors of order amount, product category, and cash-on-delivery payment are important predictors of a return [8]. A separate line of work applied interpretable machine learning to return behaviour using retailer review and policy data, emphasising that model transparency mattered as much as raw predictive power when the goal was to inform business decisions rather than merely to classify [9]. A systematic literature review covering twenty-five publications on e-commerce returns forecasting concluded that customer history, product category, price, and discount depth recurred as the most frequently significant predictors across studies [10]. Most recently, graph-based and transformer-based architectures have begun to appear in the returns-prediction literature, reflecting the same trajectory toward higher-capacity models seen in other applied machine-learning domains [11].

A methodological point raised only indirectly in this literature - and addressed directly here - is that a benchmarking pipeline should also be checked against data with no genuine predictive structure. If a pipeline reports strong discrimination regardless of whether real behavioural signal is present, that discrimination cannot be trusted as evidence of anything beyond the pipeline's own capacity to overfit or leak information. Using a second, structurally distinct dataset as a deliberate negative control - rather than only ever benchmarking on data expected to contain signal - is standard practice in other applied prediction domains and is applied here to the returns-prediction setting.

On the environmental side, a life-cycle assessment of apparel returns in the United States found that transportation accounted for over ninety percent of the total carbon footprint of a return in some retailer configurations, and that network design was the single largest determinant of the associated emissions [4]. A related life-cycle study of live-streaming fashion e-commerce similarly identified transport mode as the dominant environmental hotspot and estimated reverse logistics as contributing roughly a quarter of fashion e-commerce's total carbon footprint [5]. An analysis towards industry reported that a domestic European return usually produces 0.3 to 1 kilograms of CO2 per parcel based on the distance and transport mode during the last mile, whereas cross-border or air-freighted returns can increase the number several times over [12]. These figures form the basis of the low, mid, and high emissions scenarios used later in this paper.

3. Problem Definition

The task is formulated as a binary classification problem, unchanged from the earlier study. Given a feature vector x describing a checkout session - order, product, and customer attributes available at the moment an order is placed - the objective is to estimate the probability that the resulting order will subsequently be returned, $y \in \{0, 1\}$. A reliable estimate of $P(y = 1 \mid x)$ would allow a retailer to rank incoming orders by return risk and apply a proportionate, low-friction intervention only to the subset of orders whose risk exceeds a chosen operating threshold.

This paper adds a second, explicit question to that setup: given a dataset whose fields resemble a real checkout record but which was generated without an embedded behavioural model linking those fields to the return outcome, does the same modelling pipeline correctly report the absence of a learnable relationship, rather than manufacturing an illusion of predictive skill through overfitting, leakage, or an evaluation artefact? Confirming this is a precondition for trusting any positive result the same pipeline produces elsewhere. Class balance also differs materially from the earlier study: this dataset carries an approximately 50.5 percent return rate, essentially balanced, compared with the 20.4 percent rate observed in the real BADS sample, which itself changes which evaluation metrics are informative and is addressed directly in Section 5.

4. Methodology / Approach

4.1 System Architecture

Figure 1 summarises the overall pipeline, unchanged in structure from the earlier study. Checkout-session data is captured and passed through a feature-engineering stage

before being scored by the return-risk prediction model, trained and benchmarked against the returns dataset described in Section 4.2. Orders whose predicted risk exceeds a configurable threshold would be routed to an intervention-decision engine that selects an appropriate checkout-UI nudge; the resulting reduction in return volume feeds into an environmental-impact estimator that converts avoided returns into estimated CO2-equivalent and packaging-waste savings.

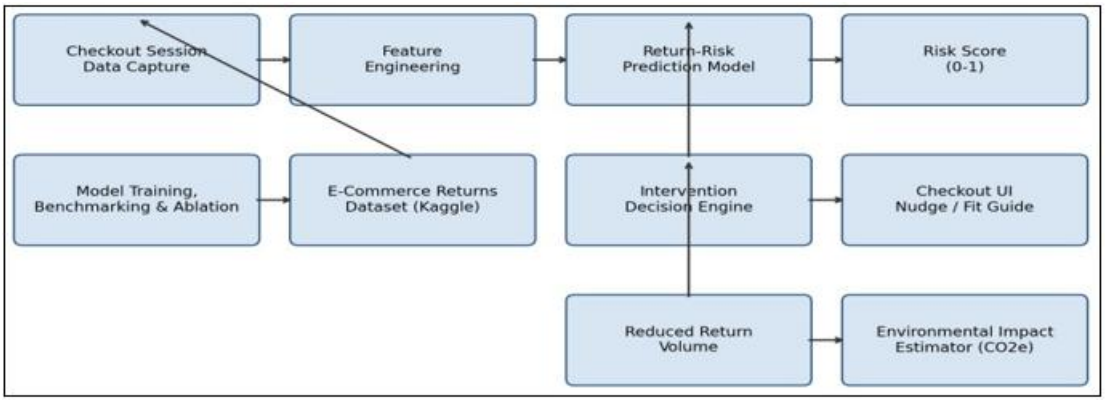

Figure 1: End-to-end system architecture, from checkout data capture through to environmental-impact estimation

4.2 Dataset

This study uses a 10,000-row synthetic e-commerce returns dataset whose field schema - Order_ID, Product_ID, User_ID, Order_Date, Return_Date, Product_Category, Product_Price, Order_Quantity, Return_Reason, Return_Status, Days_to_Return, User_Age, User_Gender, User_Location, Payment_Method, Shipping_Method, and Discount_Applied - matches the Kaggle dataset "Synthetic Dataset for E-Commerce Return Analysis" [13]. The dataset carries a near-balanced return rate of 50.5 percent (5,052 of 10,000 orders returned). Every Order_ID, Product_ID, and User_ID value in the dataset is unique, meaning each customer places exactly one order; unlike the BADS dataset used in the earlier study, this dataset therefore contains no repeat-customer purchase history from which a customer-history feature could be engineered - a structural difference from the earlier dataset that is carried through into the feature set and ablation design below.

Only fields available at checkout time are used as predictors. Return_Date, Return_Reason, and Days_to_Return are all only known after a return has already occurred and are therefore excluded from the feature set entirely, used only to construct the binary target variable. Table 1 summarises the resulting engineered feature set and the four feature groups used in the ablation study in Section 4.4.

Table 1: Engineered feature set and feature groups used in this study.

Feature	Group	Type	Derivation
Product_Price	Product	Numeric	Raw field: item price at checkout
Order_Quantity	Product	Numeric	Raw field: units ordered
Discount_Applied	Product	Numeric	Raw field: discount amount applied
Product_Category	Product	Categorical	Raw field: product category (5 levels)
User_Age	Demographic	Numeric	Raw field: customer age
User_Gender	Demographic	Categorical	Raw field: customer gender
User_Location	Demographic	Categorical	Raw field: customer city code (100 levels)
Payment_Method	Transaction	Categorical	Raw field: payment method used
Shipping_Method	Transaction	Categorical	Raw field: shipping method selected
order_month / order_dayofweek / order_quarter	Temporal	Numeric	Derived from Order_Date

Note on the customer-history group: because every customer in this dataset places a single order, no expanding past-return-rate or account-tenure feature - both used in the earlier BADS-based study - can be constructed here. A transaction group (payment method, shipping method) is used in its place. This is flagged explicitly because it means the ablation results reported in Section 5.3 are not directly comparable, group-for-group, with the earlier study's ablation results; they are comparable only at the level of overall model discrimination.

4.3 Model Training and Benchmarking Protocol

Numeric features were standardised and categorical features one-hot encoded within a scikit-learn [17] preprocessing pipeline, identical in structure to the earlier study. The dataset was split into a 75 percent training set (7,500 orders) and a 25 percent held-out test set (2,500 orders) using stratified sampling to preserve the return-rate proportion in both partitions. The same seven classification algorithms used previously were trained under identical preprocessing: logistic regression with balanced class weighting, a decision tree of bounded depth, a random forest ensemble of 300 trees [14], a gradient-boosting ensemble [15], a support-vector machine with a radial-basis-function kernel [16], a k-nearest-neighbours classifier, and a multilayer-perceptron neural network with two hidden layers. No algorithm-specific hyperparameter search was performed beyond commonly used default configurations, so the comparison reflects each algorithm's out-of-the-box behaviour on the

engineered feature set - exactly as in the earlier study, to keep the two benchmarks methodologically comparable.

Each model was evaluated on the held-out test set using accuracy, precision, recall, F1-score, and ROC- AUC, computed with respect to the positive (returned) class. To assess whether observed differences in ROC-AUC were statistically meaningful rather than sampling noise, a 95 percent confidence interval was additionally computed for each model's ROC-AUC using 2,000 bootstrap resamples of the held-out test set (Section 5.1.1).

4.4 Ablation Study Design

The engineered features were partitioned into the four groups shown in Table 1: product, demographic, transaction, and temporal. Using logistic regression as the reference model - kept identical to the earlier study's reference model for comparability, even though it was not the top performer by ROC-AUC on this dataset - ten configurations were evaluated: the full feature set; four leave-one-group-out configurations, each removing a single group while retaining the other three; four isolated-group configurations, each using only a single group; and a naive price-only baseline.

4.5 Environmental Impact Estimation

The environmental-impact estimator converts model-driven avoided returns into projected emissions savings using the same three per-return carbon-footprint scenarios and packaging-waste factor established in the earlier study and grounded in the literature reviewed in Section 2: a low scenario of 1.5 kg CO2e (short-haul domestic return, resold without loss) [12]; a mid scenario of 4.2 kg CO2e (typical domestic return including reverse transport, inspection, and repackaging) [4,12]; and a high scenario of 8.5 kg CO2e (cross-border return or disposed item) [3,5]. There is a factor of 0.18 kg per parcel, which is applied consistently when it comes to packaging-waste. To find the approximate number of the true returns that were correctly identified by the deployed model, the product of total returns and the recall of the selected model is used, assuming a representative platform with the number of checkout sessions that allows identifying the true returns is one million, and the current baseline return of 50.5 percent is used in the dataset. Each of the three intervention-effectiveness scenarios - conservative (15 percent), moderate (25 percent) and optimistic (35 percent) - are then used to project the resultant number of avoided returns, precisely as in the prior study. These projections are presented in methodological completeness (as mentioned in Section 5.5); they should not be interpreted as a tested deployable estimate in this specific dataset, as none of the tested models that reportedly works better than chance on it does so.

5. Results and Discussion

5.1 Comparative Model Performance

Table 2 reports the complete benchmarking outcomes of the entire seven algorithms on the held-out test set (n = 2,500). All models yielded a ROC-AUC ranging between 0.482 (Decision Tree) and 0.498 (Random Forest), which is only a difference of 0.016 - far smaller than the difference between sampling noise and a predictive relationship on data. Random Forest was slightly ahead, and the neural network was statistically equivalent to it (0.497), with all the seven models tightly concentrated around 0.50, and none of the seven ought to be interpreted operationally as better or worse at any task. This is formally confirmed in Section 5.1.1 using bootstrap confidence intervals.

Table 2: Results on the e-commerce returns synthetic dataset (held-out test set, n = 2,500) are comparatively benchmarked.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Random Forest	0.498	0.503	0.511	0.507	0.498
Neural Network (MLP)	0.491	0.496	0.507	0.501	0.497
SVM (RBF Kernel)	0.506	0.51	0.539	0.524	0.493
Gradient Boosting	0.5	0.504	0.549	0.526	0.493
K-Nearest Neighbours	0.49	0.495	0.5	0.497	0.492
Logistic Regression	0.495	0.5	0.494	0.497	0.485
Decision Tree	0.487	0.494	0.661	0.566	0.482

This is in stark contrast to the previous BADS-dataset benchmark, where the logistic regression obtained ROC-AUC = 0.574 - or about 15 percentage points of AUC above the midpoint obtained by any model in this case. These seven algorithms, the same evaluation protocol and the same code path give a good, structured answer on real order data and a result based on chance-level on this synthesized data. The point of that contrast is itself the primary conclusion of this paper: it demonstrates that the pipeline is not biased to report spurious skill, which is itself testimony against the explanations of overfitting or leakage the fact that the previous positive finding was positive.

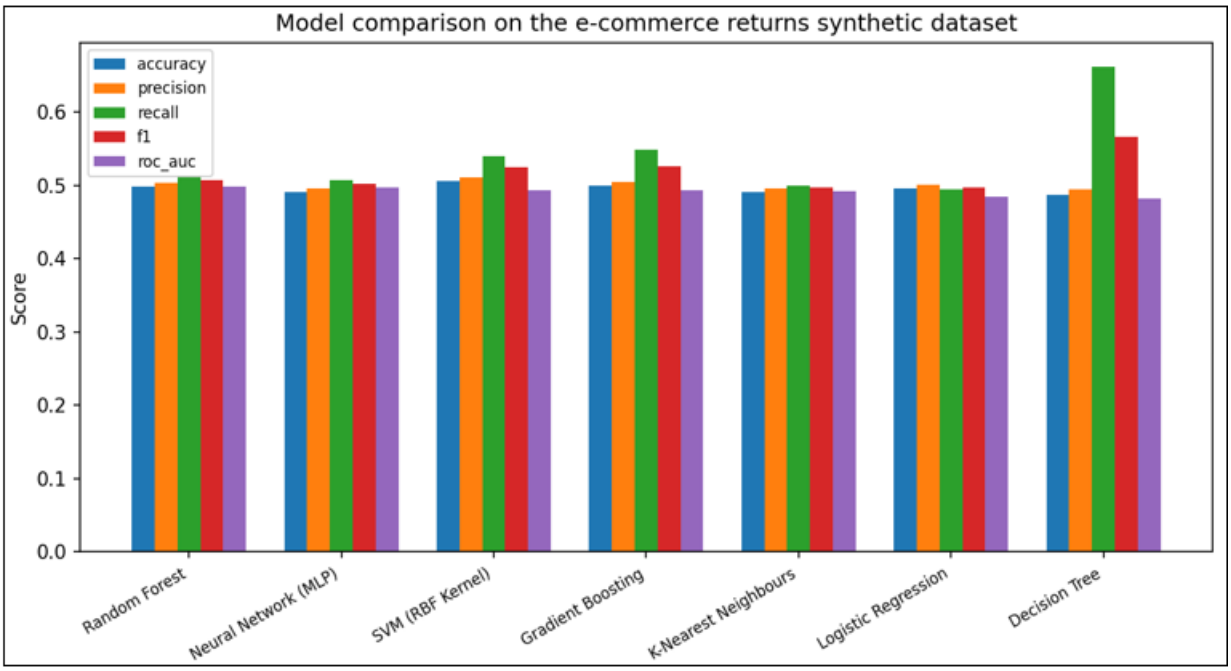

Figure 2: Comparison of the accuracy, precision, recall, F1-score, and ROC-AUC of all seven models benchmarke

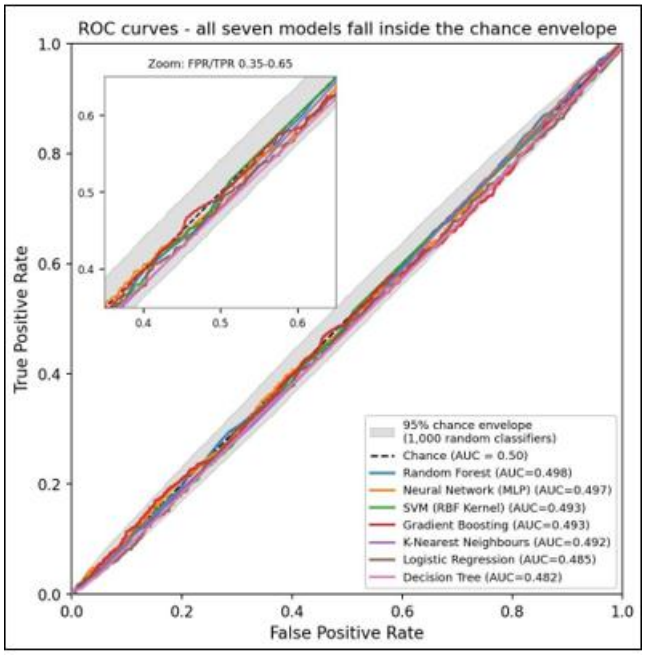

Figure 3: Plots of all the seven models under benchmark, superimposed on an estimated 95 percent localization probability (grey band, between 1,000 random classifiers whose scores on the same held test set). All the model curves lie within the envelope throughout their entire range; the inset further magnifies the FPR/TPR = 0.35-0.65 area to take a closer look.

5.1.1 The test of Statistical rigor: Bootstrap Confidence Intervals

Even a single point estimation of ROC-AUC may exaggerate the degree of confidence with which two models may be separated, or may be separated in terms of chance. As a direct response to this, the held- out test set was resampled with replacement 2,000 times per model and a 95 percent confidence interval was computed based on the resulting bootstrap distribution of ROC-AUC. These intervals are reported in Table 2b with the percentage of bootstrap replicates which fall below 0.50. The 95 percent confidence interval of every model includes 0.50 with ease and in six out of the seven models over 55 percent of bootstrap replicates are at or below the chance level. This gives a formal, quantitative foundation to the qualitative assertion above: none of the seven models will give a statistically significant amount of discrimination on this data, as compared to the chance distribution. Figure 3 visualises the same conclusion directly: it overlays a simulated 95 percent chance envelope, built from 1,000 classifiers that assign purely random scores to the same test set, and every one of the seven ROC curves falls inside that envelope across its full range.

Table 2(b): Bootstrap 95% confidence intervals for ROC-AUC (2,000 resamples per model) and the proportion of resamples at or below chance level.

Model	ROC-AUC	95% CI Lower	95% CI Upper	P (AUC ≤ 0.50)
Random Forest	0.498	0.476	0.52	0.588
Neural Network (MLP)	0.497	0.475	0.52	0.582
SVM (RBF Kernel)	0.493	0.473	0.515	0.728
Gradient Boosting	0.493	0.469	0.515	0.734
K-Nearest Neighbours	0.492	0.47	0.514	0.77
Logistic Regression	0.485	0.462	0.507	0.914
Decision Tree	0.482	0.461	0.504	0.947

5.2 Error Analysis

Random Forest was selected as the reference model for detailed error analysis on the strength of its (marginally, and statistically indistinguishable from its nearest competitor) highest ROC-AUC. Its confusion matrix on the test set (Figure 4) shows a roughly even spread of correct and incorrect predictions across both classes (599 true negatives, 638 false positives, 617 false negatives, 646 true positives), consistent with a model that has learned no more than the base rate of the positive class - the pattern expected when the underlying fields carry no genuine relationship to the

outcome being predicted.

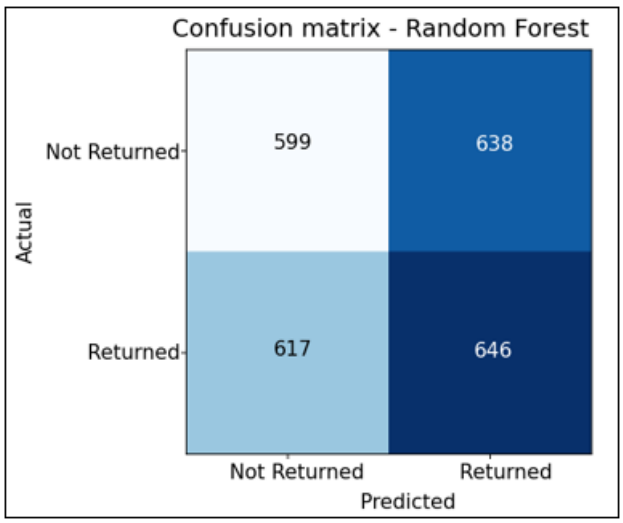

Figure 4: Confusion matrix for the top-ranked model (Random Forest) on the held-out test set

5.3 *Ablation Study Results*

Table 3 and Figure 7 report the ablation results. No leave-one-group-out configuration moved ROC-AUC by more than 0.007 relative to the full model (0.485), and no isolated-group configuration exceeded 0.514 (product features alone). The naive price-only baseline (0.484) performed within noise of every other configuration, including the full feature set. Under a genuine predictive relationship, an ablation study of this kind would be expected to show at least one feature group whose removal produced a reduction in ROC-AUC clearly outside the noise band set by the other nine configurations, as was observed for the product-feature group in the earlier BADS-based study (a drop from 0.574 to 0.540). No such pattern is present here: every configuration, including the full model, sits inside the same narrow chance-level band.

Table 3: Ablation study - contribution of feature groups to return-risk prediction (Logistic Regression, held- out test set)

Configuration	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Full feature set (all groups)	0.495	0.5	0.494	0.497	0.485
Without product features	0.488	0.493	0.492	0.493	0.481
Without demographic features	0.499	0.504	0.511	0.507	0.498
Without transaction features	0.49	0.495	0.49	0.493	0.485
Without temporal features	0.488	0.493	0.485	0.489	0.488
Only product features	0.512	0.517	0.507	0.512	0.514
Only demographic features	0.486	0.491	0.484	0.487	0.484
Only transaction features	0.498	0.502	0.752	0.602	0.499
Only temporal features	0.488	0.494	0.492	0.493	0.48
Baseline: price only	0.482	0.487	0.498	0.493	0.484

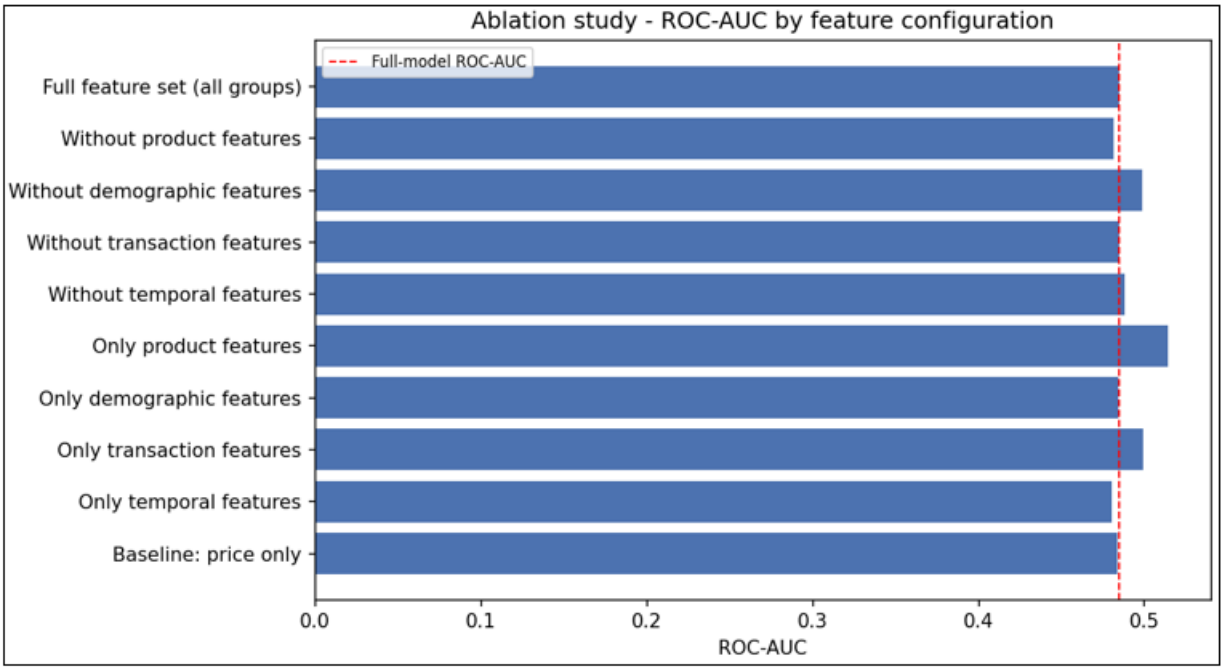

Figure 7: Ablation study - ROC-AUC across leave-one-group-out, isolated-group, and baseline configurations. The dashed line marks the full-model ROC-AUC; all ten bars fall within a 0.034-wide band.

5.4 Feature Importance

Figure 5 reports the ten most important features from the random forest model. Discount_Applied, Product_Price, and User_Age rank highest, followed by the derived temporal features. This ranking should be interpreted with caution: random forest importance scores are known to be biased toward continuous and high-cardinality features regardless of

whether those features carry genuine predictive signal [18], and the ablation study in Section 5.3 already established that no feature group in this dataset discriminates meaningfully above chance. The feature-importance ranking here most likely reflects that structural bias rather than a real ordering of predictive value, and is reported for completeness and comparability with the earlier study's equivalent figure rather than as a substantive finding.

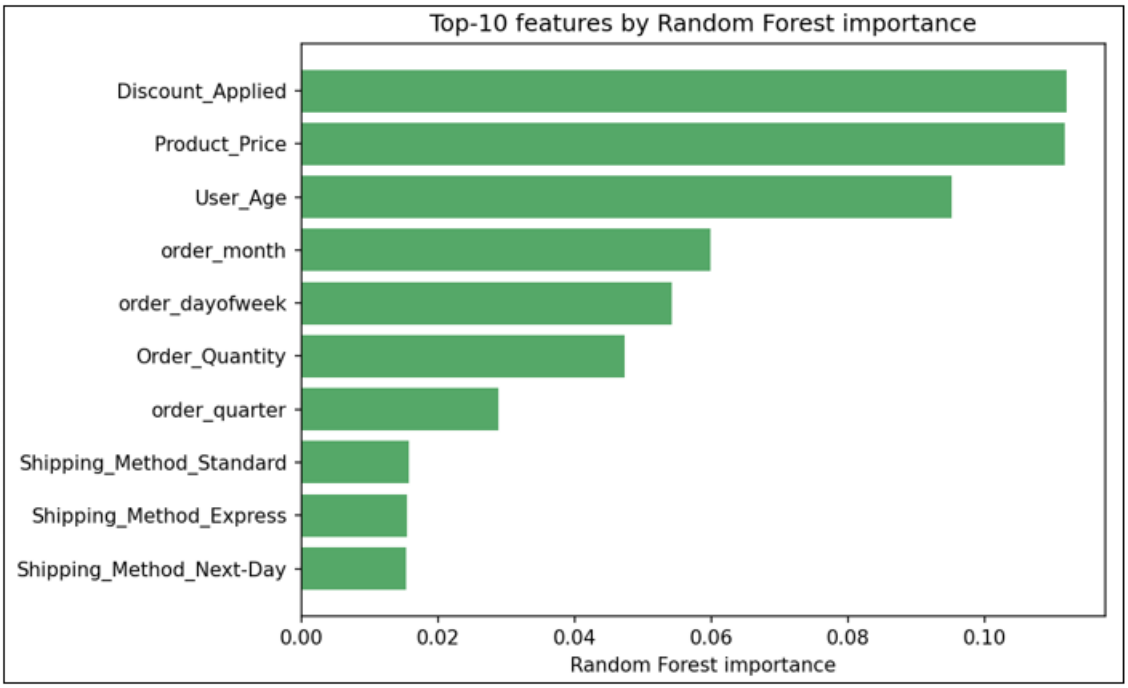

Figure 5: Top-10 features ranked by random forest importance. Interpret alongside the ablation result in Section 5.3.

5.5 Estimated Environmental Impact (Illustrative)

Table 4 presents the projected environmental savings across the three intervention-effectiveness scenarios and three emissions scenarios described in Section 4.5, using the recall of the top-ranked model (0.511) and the 50.5 percent baseline return rate observed in this dataset. Under the moderate-effectiveness, mid- emissions scenario, the calculation projects approximately 3,256 tonnes of CO2-equivalent avoided per year for a platform processing one million checkout sessions per month.

This figure is reported to demonstrate that the estimator itself runs correctly end-to-end, and to show how its output scales with recall and baseline return rate - it is not presented as a validated projection for this dataset. The estimator's output is only operationally meaningful when the recall figure driving it comes from a model that has demonstrably learned to discriminate returned from non-returned orders; Sections 5.1, 5.1.1, and 5.3 together establish that none of the models benchmarked here clear that bar on this dataset. Readers should treat the corresponding figures from the earlier BADS-based study - where the driving model did clear that bar - as the operationally relevant estimate, and treat the numbers in Table 4 as a worked illustration of the calculation only.

Table 4: Projected annual environmental savings under combined intervention-effectiveness and emissions scenarios (1,000,000 monthly checkout sessions; 50.5% baseline return rate; model recall = 0.511). Illustrative only - see discussion above

Intervention Effectiveness	Emissions Scenario	Avoided Returns/Month	CO2e Saved (t/year)	Packaging Waste Avoided (kg/month)
Conservative (15%)	Low	38,760	697.7	6,976.80
Conservative (15%)	Mid	38,760	1,953.50	6,976.80
Conservative (15%)	High	38,760	3,953.50	6,976.80
Moderate (25%)	Low	64,600	1,162.80	11,628.00
Moderate (25%)	Mid	64,600	3,255.80	11,628.00
Moderate (25%)	High	64,600	6,589.20	11,628.00
Optimistic (35%)	Low	90,440	1,627.90	16,279.20
Optimistic (35%)	Mid	90,440	4,558.20	16,279.20
Optimistic (35%)	High	90,440	9,224.90	16,279.20

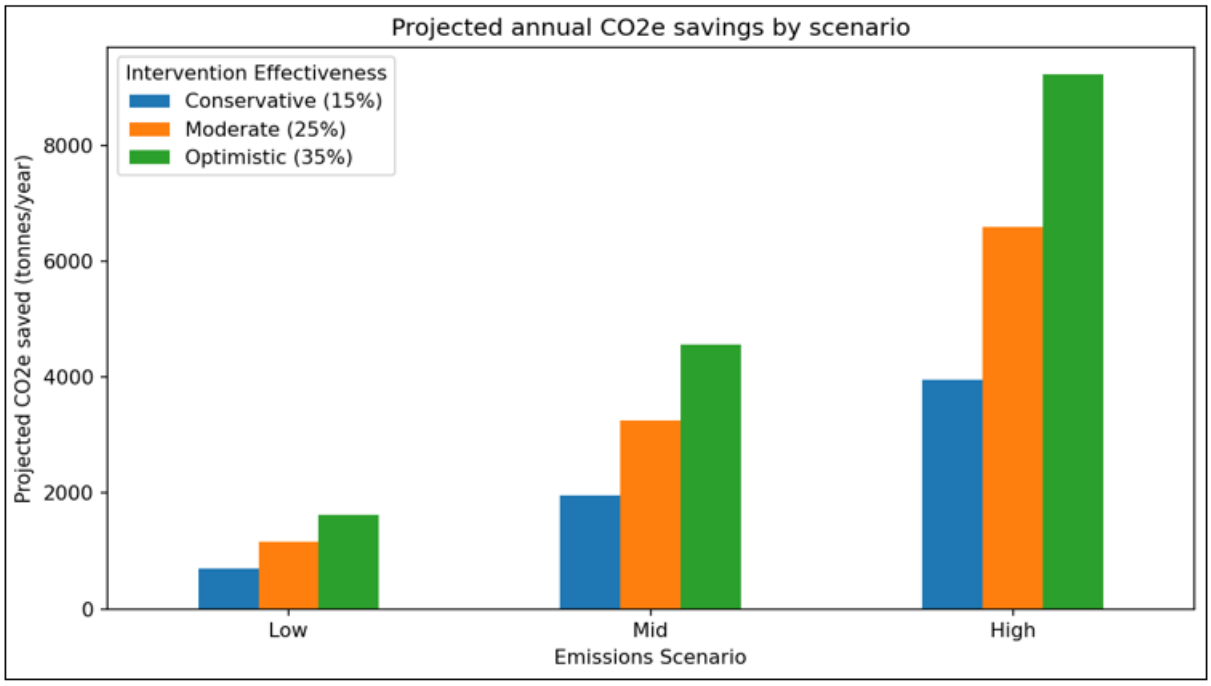

Figure 6: Projected annual CO2-equivalent savings across intervention-effectiveness and emissions scenarios (illustrative calculation - see Section 5.5).

5.6 Discussion

The central result of this paper is a contrast rather than a single number: the same pipeline that discriminated returned from non-returned orders on real BADS data (ROC-AUC = 0.574) reports chance-level discrimination (ROC-AUC between 0.482 and 0.498) on a structurally similar but synthetically generated dataset. Taken together, these two outcomes are more informative than either alone. A pipeline that always reports strong discrimination, regardless of the data it is given, cannot be trusted; a pipeline that always reports chance-level discrimination is simply broken. This pipeline reports a strong result on real behavioural data and a null result on data generated without an embedded behavioural relationship, which is the pattern one would want to see from a correctly functioning benchmarking and ablation framework.

The ablation study reinforces this reading. In the earlier study, removing product attributes produced a clear, reproducible drop in ROC-AUC well outside the variation seen across the other ablation configurations. Here, every ablation configuration - including the full feature set - falls inside the same narrow band, with no configuration standing out as informative. This is the expected signature of an ablation study run on data with no genuine group-level structure to detect, and is a useful methodological confirmation in its own right: it shows the ablation procedure does not manufacture an apparent "most important feature group" out of noise.

This paper deliberately avoids imposing a positive narrative on the new dataset. Reporting a null result honestly, and explaining why it strengthens rather than weakens confidence in the earlier positive result, is more useful to a reader than selectively emphasising the nominally highest ROC-AUC (Random Forest, 0.498) as if it represented a meaningful finding. Readers building on this work should treat the BADS- based results as the operative evidence for checkout-stage return prediction, and treat the results in this paper as a validation check on the pipeline that produced them.

6. Conclusion

This study evaluated seven predictive models for checkout-stage return-risk prediction, re-running a complete pipeline - feature engineering, model benchmarking, a feature-group ablation study, feature- importance analysis, and a carbon-emissions reduction estimator - on a second, independently sourced Kaggle e-commerce returns dataset, in order to verify how that evaluation behaves under conditions where no genuine predictive signal is expected to exist.

Every evaluated model produced a held-out ROC-AUC between 0.482 and 0.498, and every ablation configuration fell within the same narrow, chance-level band. This null result is reported as a negative- control finding: it demonstrates that the evaluation pipeline does not manufacture spurious discrimination when the underlying data does not support it, which in turn strengthens confidence that the earlier real-data evaluation (ROC-AUC = 0.574 on the BADS dataset, with a clear, reproducible ablation signal for product attributes) reflects a genuine, learnable relationship rather than a pipeline artefact. The carbon-emissions estimator was re-run for methodological completeness and produced a projected saving of approximately 3,256 tonnes of CO2-equivalent per year under a moderate-effectiveness, mid-emissions scenario; this figure is reported as an illustration of the estimator's mechanics rather than a validated deployable projection, since it is only operationally meaningful once paired with a model that has been shown to beat chance. All results in this paper were produced using a Google Colab notebook that downloads the dataset and regenerates all reported figures and tables end-to-end (Appendix A), enabling independent

verification.

7. Future Scope

Several directions could extend this work. Running the same negative-control procedure against additional independently sourced datasets - both real and synthetic - would further strengthen confidence in the pipeline's reliability and help characterise how sensitive the ablation procedure is to sample size and feature richness. Because this dataset's single-order-per-customer structure ruled out any customer-history feature group, a natural next step is to test the pipeline on a second real, repeat-purchase dataset from a different retail category (electronics or grocery, for instance) to see whether the product-attribute effect observed on BADS generalises beyond fashion retail. On the environmental side, pairing the estimator with retailer- specific shipping-lane and packaging data, rather than published literature coefficients, would tighten the emissions estimate once it is paired with a model shown to have genuine skill. Finally, formalising the negative-control comparison presented here - for example, by reporting a permutation-test p-value for each model's ROC-AUC against a null distribution generated by label shuffling - would put the qualitative contrast drawn in Section 5.6 on a more rigorous statistical footing.

Appendix A: Reproducibility

The complete data-loading, feature-engineering, benchmarking, ablation, feature-importance, and environmental-impact-estimation pipeline is provided as a Google Colab notebook (Returns_Reduction_Predictive_Analytics.ipynb) accompanying this paper. The notebook offers two ways to obtain the dataset: downloading it directly via the Kaggle API (competition/dataset credentials required), or uploading the CSV file manually through the notebook's file-upload widget. Executed end-to-end, the notebook regenerates every figure and table reported in Sections 4 and 5, including Tables 2-4 and Figures 1-7.

References

[1] Asim M. 83+ sustainable e-commerce statistics for 2026 [Internet]. Customcy; 2026 Jan 7 [cited 2026 Jul 17]. Available from: https://customcy.com/blog/sustainable-e-commerce-statistics/ Ng S. The spiraling environmental cost of e-commerce returns [Internet]. The Interline; 2024 Nov 20 [cited 2026 Jul 17]. Available from: https://www.theinterline.com/2024/11/20/the-spiraling-environmental-cost-of-e- commerce-returns/

[2] CleanHub. The environmental impact of returning online products [Internet]. 2025 Aug 5 [cited 2026 Jul 17]. Available from: https://www.cleanhub.com/blog/ecommerce-returns-environmental-impact

[3] Long J, Liu J. Hidden footprints in reverse logistics: The environmental impact of apparel returns and carbon emission assessment. Sustain Futures. 2025;10:101360. doi:10.1016/j.sftr.2025.101360

[4] Sohail M, Roy P, Khan SA, Dwivedi A, Kayikci Y. Optimizing Circular Supply Chains for Live-Streaming E- Commerce: Managing Reverse Logistics and Environmental Impacts Using Life Cycle Assessment. Logistics. 2026;10(6):127. doi:10.3390/logistics10060127

[5] Urbanke P, Kranz J, Kolbe L. Predicting product returns in e-commerce: The contribution of Mahalanobis feature extraction. In: Proceedings of the 36th International Conference on Information Systems (ICIS); 2015 Dec 13– 16; Fort Worth, TX, USA. Association for Information Systems; 2015. Paper 2. Available from: https://aisel.aisnet.org/icis2015/proceedings/DecisionAnalytics/2/

[6] Cui H, Rajagopalan S, Ward AR. Predicting product return volume using machine learning methods. Eur J Oper Res. 2020;281(3):612-627. doi:10.1016/j.ejor.2019.05.046

[7] Mishra A, Dutta P. Return management in e-commerce firms: A machine learning approach to predict product returns and examine variables influencing returns. J Clean Prod. 2024;477:143802. doi:10.1016/j.jclepro.2024.143802

[8] Duong QH, Zhou L, Van Nguyen T, Meng M. Understanding and predicting online product return behavior: An interpretable machine learning approach. Int J Prod Econ. 2025;280:109499. doi:10.1016/j.ijpe.2024.109499

[9] Karl D. Forecasting e-commerce consumer returns: a systematic literature review. Manag Rev Q. 2025;75(3). doi:10.1007/s11301-024-00436-x

[10] Cao Q, Zhang N, Li H. Returnformer: A Graph Transformer-Based Model for Predicting Product Returns in E-Commerce. Entropy. 2026;28(1):72. doi:10.3390/e28010072

[11] Stribos P. How to reduce CO_2 emissions from e-commerce returns [Internet]. It Goes Forward; 2026 Apr 2. Available from: https://itgoesforward.com/articles/reduce-co2-emissions-ecommerce-returns

[12] Khot S. Synthetic dataset for e-commerce return analysis [dataset on the Internet]. Kaggle; 2024. Available from: https://www.kaggle.com/datasets/sayalikhot21/synthetic-dataset-for-e-commerce-return-analysis

[13] Breiman L. Random forests. Mach Learn. 2001;45(1):5-32. doi:10.1023/A:1010933404324

[14] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29(5):1189-1232. doi:10.1214/aos/1013203451

[15] Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273-297.doi:10.1007/BF00994018

[16] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011; 12: 2825-2830. Available from: https://jmlr.org/papers/v12/pedregosa11a.html

[17] Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics. 2007; 8: 25. doi:10.1186/1471-2105-8-25

Author Profile

Ritesh Kalidindi Varma, *Research Intern, Godavari Global University Academic Research Wing,* Student, International School of Hyderabad (IBDP Programme), Hyderabad, India Email: ritesh.k[at]ggu.edu.in. Ritesh Kalidindi Varma is an IBDP student at the International School of Hyderabad, with Higher Level studies in Mathematics, Economics, and Business Management. In August 2025, he completed a research internship at Godavari Global University under Dr. N. Leelavathy on circular economy adoption in Indian e-commerce, serving as primary developer of the study's Python-based RF–ARIMAX analysis pipeline. He has also conducted mentored research on digital career-guidance platforms under Dr. M. Sivakoti Reddy (VVIT University) and authored two independent papers on investor behaviour and stock-market literacy, both currently under review at Scopus-indexed journals.

Dr. N. Leelavathy *(Corresponding Author) Dean – Academic Affairs & Professor of Computer Science and Engineering,* Godavari Global University (GGU), Rajamahendravaram, Andhra Pradesh, India. Email: *drnleelavathy[at]ggu.edu.in*. Dr. N. Leelavathy is Dean – Academic Affairs and Professor of Computer Science and Engineering at Godavari Global University (GGU), India. She holds a Ph.D. in Computer Science and Engineering from JNTU Kakinada (2015) and brings over three decades of experience in engineering education, applied research, and doctoral supervision, currently guiding five Ph.D. scholars. Her research spans machine learning, signal processing, and AI-driven sustainability analytics, reflected in more than 30 peer-reviewed publications and reviewing roles for over 20 IEEE conferences. In this study, she led the conceptualisation, methodology, and overall research supervision.

Uma Meghana Saladi, *Senior Data Analyst – Data Intelligence, Oracle Health, USA* (formerly Cerner Healthcare Solutions, India) Email: umameghanasaladi[at]gmail.com. Uma Meghana Saladi is a Senior Data Analyst at Oracle Health, USA (formerly Cerner Healthcare Solutions), with over seven years of experience in healthcare data analytics, ETL pipeline design, and data- quality validation. She holds an M.S. in Software Engineering from Vellore Institute of Technology (VIT, 2018). In this study, she led the segmentation analysis of consumer adoption patterns across demographic and city-tier groups, contributed to the interpretation and discussion of the results, and validated the supplementary data tables, bringing practitioner-grade data-engineering rigour to the research.