

Modeling process between the validity of multiple linear regression results and the accuracy of prediction using artificial intelligence tools
A Study of a middle category of consumers

BOUTAGHANE Mohammed Abderrezak¹, YEKHLEF Abdellah²

¹ Mostaganem University (Algeria), MIM Laboratory, Djilali Liabes University (Algeria), mohammed.boutaghane@univ-mosta.dz

² Mostaganem University (Algeria), Abdellah.yekhlef@univ-mosta.dz

Received: 07/03/2026

Accepted: 25/03/2026

Published: 01/05/2026

Abstract:

This study provides a comprehensive evaluation of the quality of explanatory and predictive capabilities of traditional statistical approaches against advanced artificial intelligence tools in assessing the factors of purchasing decisions. It contrasts a multiple regression model for its interpretability and clear identification with a Random Forest as an AI tools recognized for its ability to capture non linear and complex data patterns. Using a medium size dataset with 164 observations. The analysis try to highlight testing both models MR and RF to proof accuracy and robustness. As a final results this study demonstrate that MR model does not lose its ability to explain variables in the case of medium size samples, also it can be fundamentally relied for future predictions. Conversely, when compared with AI tools (RF), it became evident that the AI generated model loses statistical credibility in its explanatory power, so the AI model does not exhibit the high predictive capacity that it typically demonstrates when applied to large datasets.

Keywords: modeling; multiple regression; accuracy; prediction; AI Tools.

JEL Classification Codes: M31, C13, D12, P36

Corresponding author: Mohammed Abderrezak BOUTAGHANE

E-mail : Mohammed.boutaghane@univ-mosta.dz

1. INTRODUCTION

Understanding purchasing decision of consumer under diverse influences remains a big challenge in marketing analytics. Multiple Regression as a traditional approach has served as a foundational statistical tool for modeling these behaviors. As highlighted by (Josef F Hair, 2018) regression model is regarded for its statistical interpretability truth, it excels at estimating precisely the impact of independent variables, establishing robust statistical significance and discover linear relationships, however its rigid parametric assumptions often fail to capture the complex, hidden non linear interactions within consumer data in order to limite its ultimate predicting role (Varian, 2014). To adress the predictivity of standard linear models, contemporary research has turned to artificial intelligence (AI) and machine learning, AI tools have proven efficient not only in evaluating the limitation of linear regression but also in detecting intricate non linear relationships that traditional models overlook (Alain CHONG, 2015). By capturing these non linear patterns, AI models can significantly reduce the error variance inherent in traditional regression and offer the distinct advantage of continuous training and learning, adapting to new data inputs to generate accurate and dynamic future forecasts (H.Saker, 2021). Consequently, this study advocates for the implementation of a hybrid modeling approach by integrating the inferential rigor of multiple regression with the advantage of error reduction and non linear mapping capabilities of AI, in order to achieve dual objectives, first maintaining statistical credibility, second maximizing prevision (Vijay Kotu, 2015), By the way the study try to examine the outputs of the proposed hybrid model and testing its efficacy to validate the recherche hypotheses regarding the detarminants on consumer purchasing decisions.

Our main question research is articulated as follows : **what is the most effective approach to modeling the purchacing behaviors of middle dataset consumers between robustness of multiple linear regression and the predictive effiience of AI tools ?**

In tackling the question research, we posit the following hypotheses :

Modeling process between the validity of multiple linear regression results and the accuracy of prediction using artificial intelligence tools a study of a middle category of consumers

- **H1**-Multiple regression maintains a high degree of statistical validity and explanatory power among middle category of consumer's dataset.
- **H2**- AI (Random Forest) proof superiority in prediction, when applied to middle category of consumer's dataset.
- **H3**-A hybrid modeling test in middle category of consumer's dataset can be both, statistically robust and estimately efficient.

Study objectives :

- To assess the ability of MR to explain the relation ship betwin variables powering the purchase decison.
- To examine the capacity of AI models to detect complex and non-linear facts in the other medels.
- To To examine whether AI improves prediction quality compares to traditional statistical models.
- To explore the potential of developing hybrid models that combine the interpretability of MR with predictive capability of AI techniques.
- To discover the methodological specification of use midle consumer's dataset against big or small Data.

Literature review :

Severa empirical studies have compared MR models and AI techniques in order to proof validity and prediction in purchase decisions, therefore we compiled the following studies :

First:The study of Thomas Davenport and ALL, 2020 : « How AI will change the future of marketing », the authors propose a multidimentional framework to understand AI impact focusing on three dimentions : the level of intelligence, the type of task performed and third AI viw as robotic form, the authors outline a future research agenda examining how AI will reshape parkering practices and consumer decisions, Importantly, they conclude tha AI will be most effective when used to human decision in marketing rather than replace managers (Thomas Davenport, 2020).

Second :The sdudy of Ming hui Huang and Roland T.Rust 2018 : « Artificial intelligente in service », the authors developed a theory of AI

job replacement to adress this double adged impact, this theory speciefies four intelligences required for service tasks, mechanical, analytical, intuitive and empethetic and lays out the way firms should decide between human and machines fore accomplishing those tasks, an important implication of this thory is that analytical skills will become less important, giving intuitive and empathetic skills more importance for service employees, the results still true when human employee give up innovation spirit in the work (Ming-Hui Huang, Roland T. Rust, 2018).

Third :The sdudy of Mickey Singh 2024 : « Machin Learning in marketing Analysis », this research delves into the significance of machine learning techniques in marketing analytics focusing on online advertising and sales, it highlights the transition of traditional indiscriminated data driven stratigies emphethizing the importance of machine learning models to target the promising customer segments and optimize marketing efforts, focusing on customer segmentation, predictive modeling and recomendation systems, the autor finally sugest to use his research as guide to understand the multifaced role of machine learning in analytics, offering insights, challenges, and future prospects for transforming th figure of marketing strategy and decision making (Singh, 2024).

Fourth :The sdudy of Sendhil Mullainathan and Jann Sapiess : « Machine learning : An Applied Econometric Approach », the authors present a way of thinking about machine learning that gives it a big place in the econometric toolbox, it solves a differente problems, especialy prediction because it's more technicaly esasier to use only when download convenient packages in R or Python, the authors precise that econometric approche need not to be in conflict with theoritical structure but respects data inputs changes in order to excell real outputs, also machine learning is particulary useful in high dimentional and complex data environements, finally authors precise that machine learning should not replace econometrics but complement it (Sendhil Mullainathan, 2017).

2. Conceptual framework of modeling

2.1 Fundamentals of modeling with Multiple Regression:

MR is the most famous classical statistical tools used to modeling relationship between variables, MR aims to explain the variance in a single dependent variable based on the influence of independent variables, this classical statistical tool allows researchers not only to predict data from actual inputs but, more importantly to explain and determine the relative weight and statistical significance of each influencing factor (DOUGLAS C. MONTGOMERY, 2021).

2.1.1 Types of Multiple regression in consumer behavior contexts:

Depending of the nature of research, multiple regression in divided into two pathways:

- Multiple Linear Regression: used when the dependent variable is quantitative or continuous.
- Multiple Logistic Regression: used when dependent variable is categorical or binary (Buy=1, Do not buy=0).

2.1.2 Mathematical Formulation: The mathematical foundation of Regression is based on the best fitting line that minimizes all the sum of squared errors in the linear regression or maximizes the likelihood function of estimation in the case of logistic regression.

- **Multiple Linear Regression Equation:**

$$Y = B_0 + B_1X_1 + B_2X_2 + \dots B_nX_n + E$$

- **Multiple Logistic Regression Equation:**

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1X_1 + \beta_2X_2 + \dots + \beta_nX_n)}}$$

Where:

- Y or $P(Y=1)$: the dependent variable
- $X_1, X_2, X_3, \dots, X_n$: the independent variables.
- B_0 : the base line value when $X=0$.
- $B_1, B_2, \dots, B_n$: the regression coefficients which measure the direction of each independent variable's effect.

- E : The random error.

2.1.3 Core statistical assumptions: To proof the statistical validity of the regression model, the data must respect several theoretical conditions as: **linearity** (linear relationship must exist between dependent and independent variables), **Independence of errors** (by using Durbin Watson test the errors must be independent of one another), **Absence of Multicollinearity** (measured by Variance inflation Factor VIF, which means the independent variables must not be highly correlated with one another (Nikam, 2024).

2.1.4 Evaluation Metrics: Explanatory Power (R^2 or Pseudo - R^2): Determines the proportion of variance in the dependent variable that the model explains successfully.

Overall Statistical Significance (F): Evaluates whether the model as a whole is statistically significant.

Individual Variable Significance (P-value): Significance is generally accepted if $P < 0.05$, it means we test hypothesis related to each variable separately (FARSHAD, 2024).

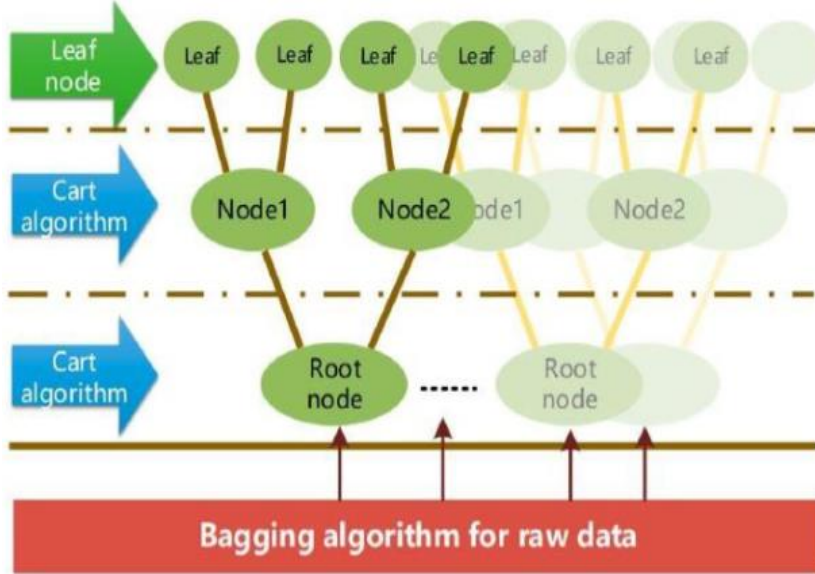
2.2 Fundamentals of modeling with AI (Random Forest):

2.2.1 How Random Forest Work? Against traditional models, the Random Forest excels in high dimensional data environments, its operational strength is optimizing predictive precision, its algorithm uses **Firstly** bootstrap aggregating, with creating multiple varied subsets (trees) of the training data through random sampling with replacement which means that some data set can be used many time within the same sub simple in order to left out the others non-desirable data; **Secondly**, for each sub set (tree) the algorithm considers only random ones when limited the correlation between trees to make model more diverse, now Random Forest (RF) excels in situations where certain variables dominate other variables and favoring the exploration of all data dimensions.

Thirdly, during prediction process, the algorithm aggregates the results of different trees according to their specialization, each tree votes and the majority score's class is chosen; with this system we can obtain accurate

results, minimizing the majority of errors and proof the reliability of predictions (Hasan Ahmed Salman, 2024).

Fig.1. Schematic Digram of random forest algorithm



Source: Hasan Ahmed Salman, 2024, P06.

2.2.2 Random Forest applications in marketing:As sited before RF is the most importante tool using in prediction and complex data analysis, at least it can help us in the marketing context, such as CRM data base fragment, behavioral studies and so..., marketing in general have adopted RF especially to inhance predictive performance in purchase intention modeling, therefore, we clarify some areas of use as follows:

- RF in consumer purchase prediction : it's the most prominent application when predicting purchase decision in consumer behavior due to its ability to capture nonlinear relationship and interation effects (Chen, 2025).
- RF in Customer segmentation and market profiling: Rf can incorporate supervised classification approaches to predict lifetime value of the customer, classify consumers into specific behavioral

segments (Zahara, 2025).

- RF in churn prediction and consumer retention : in this case RF has shown strong performance when it can model nonlinear patterns influenced by perception qhality service's, psycological facts, sensibility to prices (Payam Boozary, 2025).
- RF in mix modeling and demand forecasting : RF have been applied in MMM (marketing mix modeling) and demande forecasting, because consumer responses to advertising and promotions are non linear, also interaction effects between price and promotion (Friday O. Ugbebor, 2024).

2.2.3 Random Forest metrics evaluation: Out of Bag (OOB) Error : is derived from the subset not used in creating the decision tree of RF, serving as an internal error, after the prediction process use each training sample error to provide an unbiased estimate result in order to reduce error lore and more.

- **Accuracy** : is a proportionaly measure of total prediction that the model classified correctly, wich evaluate the performance of RF by combining predictions from many decision through voting process, it is calculated using the bellow formula :

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{TP + TN}{TP + TN + FP + FN}$$

The score variate between 0 (non correct prediction) and 1 (pefect model).

- **Confusion Matrix** : is a visualization method for classify results, it figure the number of ground truth instances of a specific class against the number of predicted class instances, by the way it's a table visulised four state (True positive, True negative, False positive, False negative)

Fig.2. Confusion matrix and common calculate from it

		<u>True class</u>			
		<u>P</u>	<u>N</u>		
<u>Hypothesized class</u>	<u>Y</u>	True Positives	False Positives	fp rate = $\frac{FP}{N}$	tp rate = $\frac{TP}{P}$
	<u>N</u>	False Negatives	True Negatives	precision = $\frac{TP}{TP+FP}$	recall = $\frac{TP}{P}$
Column totals:		P	N	accuracy = $\frac{TP+TN}{P+N}$	
				F-measure = $\frac{2}{1/precision+1/recall}$	

Source: (Fawcett, 2006, P862).

- **Classification report** : it deals with F1 score for a classification model, it range between 0 (denoting the lowest possible resutl) and 1 (denoting the flawless result), F1 crose can be calculate from the formula :

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN}$$

- **Feature importance** : FI in RF measures how feature contributes more to the model's prediction and identifying the most influential input variables, in order to inhance model performance, to faster training times and reduce errors (Utkal Khandelwal, 2021).
- **Validation Set/Cross-Validation** : is a method to test a model if it can handle new data correctly, ensuring it actually learned the patterns instead of just memorizing training dataset.

3.Methodology

The study aims to highly the possibility of combine important econometrics features such as statistical significance, linear relationships, explanation, prediction and testing hypothesis deals with Multiple Regression and Random Forest as an Artificial Intelligent tool powered by another features such as learning machine, accuracy, many decision trees, multiple combine modeling, predictive exactitude ; in order to describe

robust model purchase decision of a sample contains 164 respondents which is a middle data set through data gathered from a questionnaire distributed online, after code, we start build the multiple regression model and show its statistical features by the way, its mathematical formula; in second time we enter the previous database and treating it by Random Forest which is an Artificial Intelligence strong tool towards create a new model deals with an AI features, finally we will be suggested a hybrid model can integrate large number of preferences mainly relate to statistical significance and prediction accuracy.

3.1 Applied Multiple Regression and Random Forest comparison:

Table 1. MR vs RF key comparison

Comparison Criteria	Logistic Regression (LR)	Random Forest (RF)
Overall Performance	Good (Baseline Model)	Excellent (Superior Model)
Accuracy	~89% (0.89)	~96% (0.96)
AUC-ROC Score	~0.84	0.97 (High capability to distinguish classes)
Relationship Type	Linear (Assumes direct linear relationships between variables)	Non-Linear (Captures complex, multi-faceted interactions)
Precision & Recall	Lower precision/recall, higher rate of False Positives/Negatives.	High Precision (0.95) and Recall; superior at identifying actual churners.
F1-Score	Lower; suggests less balance between precision and recall.	Higher; indicates a robust balance and reliability.

Modeling process between the validity of multiple linear regression results and the accuracy of prediction using artificial intelligence tools a study of a middle category of consumers

Comparison Criteria	Logistic Regression (LR)	Random Forest (RF)
Interpretability	High. Easier to explain how a specific variable affects the outcome (Glass Box).	Moderate. Complex "Black Box" nature makes it harder to explain individual pathing, though Feature Importance helps.
Complexity Handling	Struggles with complex customer behaviors and interactions.	Excels at uncovering hidden patterns and intricate data dependencies.
Overfitting Risk	Prone to underfitting if data is complex.	Resistant to overfitting due to ensemble method (averaging multiple decision trees).
Best Use Case	When researchers need simple explanations of variables.	When the primary goal is maximum predictive accuracy.

Source: Study recap of (Po Abas Sunarya, 2024)

3.2 Multiple Regression results:

3.2.1 Correlation analysis: using Pearson Correlation Coefficient, we test and measure the strenght and direction of the linear relationship between the dependent variable Y (Purchase decision) and the independent variables: S (sociological), E (Economic), M (Marketing), P (Psychological), A (Physical). The results indicate a positive statistically significant correlations between dependent variable Y and all the explanatory variables based on middle sample size of 164 respondents.

Table 2. Correlation and significance results

Variable	Correlation Value	Significance Value
S	0.687	0.000
E	0.614	0.000
M	0.713	0.000
P	0.698	0.000
A	0.618	0.000

Source: Elaborate by the researchers

The matrix shown above reveals that all the independent variables have a strongest positive relationship with the purchase decision and statistically significant, as it appears in the table, R=71.3% refers to Marketing impact variable in first place, R=69.8% refers to Psychological impact variable secondly, R=68.7% refers to Sociological impact in the third place, after we find R=61.8% for the Physical impact and R=61.4% for the Economic impact. About signification we notice that Sig=0.000 wich means a good statistical signification at 5% significance level.

3.2.2 Multicollinearity Test: for this purpose, we use VIF (Variance Inflation Factor) to examine the problem of multicollinearity among independent variables, the table below indicate the test value.

Table 3. VIF table indicators

Variable	VIF
S	1.47
E	1.53
M	1.63
P	1.22
A	--

Source: Elaborate by the researchers

Vif value range between 1.22 and 1.63 wich means low multicollinearity among the independent variables, by the way the regression model is statistically reliable for explaining purchase decision.

3.2.3 Regression mathematical formula:

$$Y = 2.302 + 0.185S - 0.183E + 0.102P + 0.170M + 0.024A$$

Y (Purchase decision Dependent Variable), S (Sociological Independent

Variable), E (Economic IV), P (Psychological IV) M (Marketing IV), A (Physical IV).

3.3 Random Forest results:

3.3.1 Random Forest metrics:

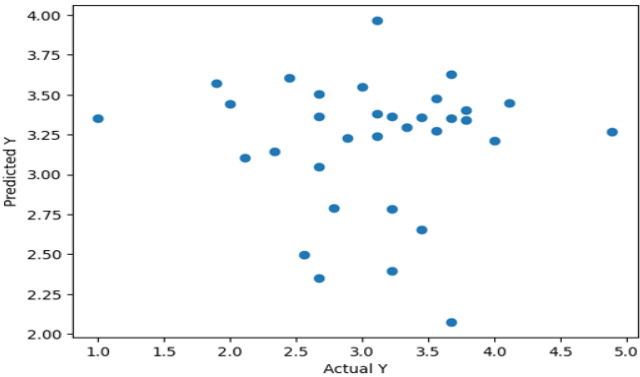
Table 4. Random Forest metrics with 164 observations

Metrics	Value
R ²	-0.359
MAE	0.648
RMSE	0.859

Source: Elaborate by the researchers

We notice that $R^2=-0.359$ which means that the AI model through the machine learning process predicts worse more than the primary model, and the prediction errors of the model are larger than the variance of the data used before. $MAE=0.648$ means that the RF model’s prediction varies from the real purchase decision with 0.359 point; based on the questionnaire scale range (1 to 5) this error considered as moderate in our case the AI model can approximate consumer decision but some prediction uncertainly remains. RMSE (Root Mean Square Error) =0.859 proof that many prediction errors are done in some observations, it could happen mostly in consumer behavior data set where decisions are influenced by non-captured factors.

Fig.3. Predected vs Actual Purchase Decision



Source: Elaborate by the researchers

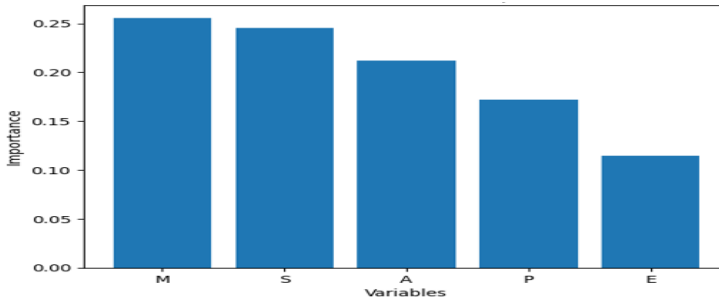
3.3.2 Random Forest Feature Importance:

Table 5. Random Forest Feature Importance with 164 observations

Variable	Importance
M	0.255
S	0.245
A	0.212
P	0.172
E	0.114

Source: Elaborate by the researchers

Fig.4. Random Forest Feature Importance Graph



Source: Elaborate by the researchers

The AI model indicates that: First, The strongest determinant of purchase decision is M (Marketing factor), Second, S (Social influence), Third (Physical influence), Fourth P (Psychological influence), Fifth A(Physical influence).

3.3.3 Random Forest Prediction Function:

$$\hat{Y} = \frac{1}{T} \sum_{t=1}^T f_t(S, E, P, M, A)$$

Where:

T represents the total number of trees in the forest.

f_t represents the prediction generated by tree t

4. Up grade to the hybrid model

Regarding to the results before we integrate both AI and traditional statistical modeling by developing a Hybrid Analytical Model. This

approach combines MLR (multiple Linear Regression) and RF (Random Forest) in order to benefit from interpretability, nonlinear patterns detection, prediction accuracy.

5. Experimental

5.1. Hybrid model with Middle stampe results (164 Observations):

When we use data base with 164 observations, combined MLR and RF model can extract metrics for both hybrid approach.

5.1.1. First approach : Stacking the both (MLR+RF) with 164 observations :

This approach opperates both models in parallel and uses a meta-model to prediction, the table below appears the results.

Table 4. Hybrid model in stacking approach with 164 observations

Metrics	Value
R^2	-0.223
MAE	0.600
RMSE	0.815

Source: Elaborate by the researchers

5.1.2. Sconde approach : Hybrid model based on residual errors with 164 observations :

This approach uses RF to predict the residual errors of data related to MLR model, the results were as follows.

Table 5. Hybrid model based on residual errors with 164 observations

Metrics	Value
R^2	-0.351
MAE	0.647
RMSE	0.857

Source: Elaborate by the researchers

Concerning the interpretation of results, we notice that R^2 scores for both approaches are negative (-0.223), (-0.351) using 164 observations which means that the model is favoring worse than the mean value of (Y) every time, so the deficiency is related to: **overfitting** (the model learned patterns that are different from 20% of the entire test data), and **Non-linear**

relationship (may be there are no predictive relationship between variables, S, E, P, M, A, and the dependent variable Y in this data base make the model more unstable).

5.2. Hybrid model results with 315 Obserations :

After training data from 164 observations to 315 observations the hybrid analytical model extracted the following metrics :

5.2.1. First approach : Stacking the both (MLR+RF) with 315 observations :

Table 6. Hybrid model based on residual errors with 315 observations

Metrics	Value
R^2	0.227
MAE	0.525
RMSE	0.665

Source: Elaborate by the researchers

5.2.2. Sconde approach : Hybrid model based on residual errors with 315 observations :

Table 7. Hybrid model based on residual errors with 315 observations

Metrics	Value
R^2	0.188
MAE	0.525
RMSE	0.682

Source: Elaborate by the researchers

Now, we notice that R^2 scores for both approaches are convert to positive value (0.227), (0.118) using 164 observations wich means that the model proves doing better than just guessing the average value of dependent variable (Y) every time, and there is a real pattern connecting the inputs data to the outputs.

5.3 Hybrid model results with 1315 observations :

5.3.1. First approach : Stacking the both (MLR+RF) with 1315 observations :

Table 8. Hybrid model based on residual errors with 1315 observations

Metrics	Value
R^2	0.150
MAE	0.520
RMSE	0.656

Source: Elaborate by the researchers

5.3.2. Second approach: Hybrid model based on residual errors with 1315 observations:

Table 9. Hybrid model based on residual errors with 1315 observations

Metrics	Value
R^2	0.147
MAE	0.520
RMSE	0.657

Source: Elaborate by the researchers

In this step, we still have a positive R^2 value (0.150), (0.147) but less than the experimental process including 1315 observations, this can happen when synthetic data generated via linear relationship introduces some noisy impact which hide patterns that Random Forest was previously exploiting. About MAE and RMSE variation, we can explain it as follow; MAE and RMSE are both ways to measure how far of a model's predictions are from the actual values, so, the treat large errors different mathematically (O.Hodson, 2022), in our case MAE=0.520, RMSE=0.657, so, RMSE is higher 26% than MAE which is normal.

5.3.3. Third approach : using GBR (Gardien Boosting Regressor) and SVR (Support Vector Regressor)

Table 10. Random Forest metrics with 315 observations, GBR, SVR, HM

Metrics	Value with 315 Observations	Gardien Boosting Regressor GBR	Support Vector Regressor SVR	Hubrid model
R ²	0.187	0.200	0.200	0.227
MAE	0.508	0.515	0.533	0.525
RMSE	0.683	0.677	0.677	0.665

Source: Elaborate by the researchers

Before discussing results, we must describe both GBR and SVR, so: **GBR** means Gardian Boosting Regression, its logic of work builds a team of simple decision trees sequentially where each tree is tasked with correcting the residuals errors of previous trees.

SVR means Support Vector Regressor, its way of work is to fit a tube (measurement field) of acceptable error around the data and ignored each point inside this tube in order to build its mathematical formula based on the outlier's data. In metrics, both GBR and DVR submit to R² value, MAE and RMSE.

In **table 10**, R² value for GBR, SVR is (0.200) wich is acceptable in social sciences (human behaviour, between 0.15 and 0.20), both RMSE value is higher 23.9% for GBR and 21.2% for SVR than MAE which is normal.

5. CONCLUSION

Finally, our study aimed to examine the purchase decision (Y) as a dependent variable by employing both traditional statistical techniques (Multiple Linear Regression) and Artificial intelligence (Random Forest) predictive model. The research was conducted in many steps in order to understanding the relationships between studying variables and the possibility of accurate prediction.

The relationships between variables were examined using Pearson Correlation Coefficient (0.666), the result revealed statically significant between purchase decision (Y) and independent variables (Socials,

Economics, Psychological, Marketing, Physicals). Variance Inflation Factor (1.46) confirms the absence of multicollinearity among independent variables, so, the model provides reliable coefficient estimates.

In order to complete analysis, we applied an AI tool by using the Random Forest process, it's capturing nonlinear relationships and complex interactions that may not captured by linear models, the predictive performance was tested using Mean Absolute Error and Root Mean Square Error to the average of prediction errors, the metrics values especially R^2 which is negative (-0.359) motivate us to deepen the research and make sure that the matter is related to middle sample (164 responders), by the way leveling up the number of responders contributes to raising R^2 positively, in data base of 315 responders R^2 up to 0.187, in data base of 1315 $R^2=0.147$, this caused confusion once again because decrease of R^2 , after repeating the experiment many times, the problem lies in synthetic data generation handling with AI.

A hybrid model was proposed to gain feature advantages from Classical statistical model (Multiple Linear Regression) and AI tool (Random Forest), the powerful model deals with combine important econometrics features such as statistical significance, linear relationships explanation, prediction and testing hypothesis, learning machine, accuracy, many decision trees, multiple combine modeling, predictive exactitude. In the first test hybrid model give up an $R^2 = 0.227$, $AME=0.525$, $RMSE=0.665$, so, $RMSE$ is higher 21% than MAE which is generally normal.

Finally we constat that the good metrics model is related to: **Sample size** which must exceed 300 respondents, and the **strength** of the hybrid model, which possesses a wide range of analytical properties.

6. Bibliography list

- Alain CHONG, e. a. (2015, Decembre). Predicting online product sales via online reviews, sentiments, and promotion strategies: A big data architecture and neural network approach. *International Journal of Operations & Production Management*. Récupéré sur https://www.researchgate.net/publication/283476126_Predicting_online_pr

- oduct_sales_via_online_reviews_sentiments_and_promotion_strategies_A_big_data_architecture_and_neural_network_approach#fullTextFileContent
- Chen, C. (2025, Novembrer). Machine Learning-Driven Analysis of Factors Influencing Consumer Purchase Decision. *Applied and Computational Engineering*.
- DOUGLAS C. MONTGOMERY, a. A. (2021). *Introduction to Linear Regression Analysis* (éd. 6). WILEY.
- FARSHAD. (2024, Octobre 31). *Essential Regression Evaluation Metrics: MSE, RMSE, MAE, R^2 , and Adjusted R^2* . Récupéré sur medium: <https://farshadabdulazeez.medium.com/essential-regression-evaluation-metrics-mse-rmse-mae-r%C2%B2-and-adjusted-r%C2%B2-0600daalc03a>
- Fawcett, T. (2006, December). An introduction to ROC analysis. *Pattern Recognition Letters*. Récupéré sur <https://people.inf.elte.hu/kiss/11dwhdm/roc.pdf>
- Friday O. Ugbebor, e. a. (2024, April). PREDICTIVE ANALYTICS MODELS FOR SMES TO FORECAST MARKET TRENDS, CUSTOMER BEHAVIOR, AND POTENTIAL BUSINESS RISKS. *Journal of Knowledge Learning and Science Technology*. Récupéré sur https://www.researchgate.net/publication/391628298_Measuring_the_forecast_accuracy_in_retail_MSMEs_A_comparative_analysis_between_AI_and_traditional_methods_in_the_era_of_digital_selling
- H.Saker, I. (2021, March). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*. Récupéré sur https://www.researchgate.net/publication/350297716_Machine_Learning_Algorithms_Real-World_Applications_and_Research_Directions/citations#fullTextFileContent
- Hasan Ahmed Salman, a. a. (2024). Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning*. Récupéré sur https://www.researchgate.net/publication/382419308_Random_Forest_Algorithm_Overview
- Josef F Hair, B. J. (2018). *Multivariate Data Analysis*. CENGAGE.
- Ming-Hui Huang, Roland T. Rust. (2018). *Journal of Service Research*.
- Nikam, G. (2024, Decembre 3). *Key Assumptions in Linear Regression*. Récupéré sur medium: <https://medium.com/@gayatrinikam5103/key-assumptions-in-linear-regression-0734088c90e9>
- Payam Boozary, a. a. (2025, June). Enhancing customer retention with machine learning: A comparative analysis of ensemble models for accurate churn prediction. *International Journal of Information Management Data*

Modeling process between the validity of multiple linear regression results and the accuracy of prediction using artificial intelligence tools a study of a middle category of consumers

Insights, 5. Récupéré sur <https://www.sciencedirect.com/science/article/pii/S2667096825000138>

- Sendhil Mullainathan, J. S. (2017). Machine Learning: An Applied. *Journal of Economic Perspectives*, 31(02). Récupéré sur <https://pubs.aeaweb.org/doi/pdfplus/10.1257/jep.31.2.87>
- Singh, M. (2024). Machine Learning in Marketing Analytics. *International Journal of Enhanced Research in Management & Computer Applications*.
- Thomas Davenport, a. A. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science*.
- Utkal Khandelwal, e. a. (2021). Measuring Online Consumer Conformity Among Indian Consumers. *International Conference on Information Systems and Computer Networks*.
- Varian, H. R. (2014). Big Data: New tricks for Econometrics. *JOURNAL OF ECONOMIC PERSPECTIVES*, 28(02). Récupéré sur <https://www.aeaweb.org/articles?id=10.1257/jep.28.2.3>
- Vijay Kotu, B. D. (2015). *Predictive Analytics and Data Mining Concepts and Practice*. Waltham: ELSEVIER. Récupéré sur <http://repo.darmajaya.ac.id/4317/1/Predictive%20Analytics%20and%20Data%20Mining.pdf>
- Zahara, M. (2025, October). Enhancing Customer Experience with AI-Powered Marketing. *IGI Global scientific publishing platform*. Récupéré sur https://www.researchgate.net/publication/396161323_Enhancing_Customer_Experience_with_AI-Powered_Marketing