JOURNAL OF INTERNATIONAL CRISIS AND RISK COMMUNICATION RESEARCH
ISSN: 2576-0017
2020, VOL 3, NO 1

The Normalized Mean Value Model: A Novel Approach For Propensity Score Prediction And Its Comparative Performance

Harish Janardhanan

Independent Researcher , harishjan@gmail.com

Abstract

This research paper compares Logistic Regression, Random Forest, and a new proposed algorithm ”Normalized Mean Value Model” to see how well they predict propensity scores for different events based on website activity data. The goal is to find the model that gives the most accurate propensity scores for website product purchases. This could then be used to predict other likely events. The paper explains what propensity scores are and what data is needed to calculate them from website activity and compares how well Logistic Regression and Random Forest models perform against the proposed models, termed ”Normalized Mean Value Model” built using correlation strength of independent variables. This research aims to validate the effectiveness of the proposed model propensity score calculation and ascertain whether the distribution of the calculated propensity scores is solely influenced by specific correlated independent variables. Additionally, this research seeks to uncover potential improvements and insights that could enhance existing prediction models.

INDEX TERMS Classification, Predictive Modeling, Propensity Scores, Machine Learning Algorithms, Logistic Regression, Random Forest, Normalized Mean Value Model, Event Prediction, Data Analysis.

1 Introduction

Predicting propensity scores is important in healthcare, education, finance, and e-commerce. It helps us determine how likely certain events will happen based on what we already know [1]. This study compares different prediction models like Logistic Regression [2] and Random Forest [3], as well as the proposed logic model, to see if I can get better results by using data points based on closely related variables. It’s essential to understand how these models work so that we can make better decisions in situations where accurately predicting events is key [4].

In this paper, I have compared popular Machine learning models with the proposed logic, aka” Normalized Mean Value Model”, developed to see if this logic can give us a better prediction. This paper may help further enhance this methodology, which can be used for other use cases and help enhance research in predicting outcomes for other real-world scenarios as well.

This paper is written to fill the gap where existing models are compared with the proposed logic and models and motivate the scope for further research in the methodology discussed in this paper.

1.1 Abbreviations and Acronyms

Table 1: Abbreviations

Abbreviation	Meaning
TP	True positive
FP	False positive
TN	True negative
FN	False negative
TPR	True positive rate
TNR	True negative rate

Table 2: Formulas

Formula	Description
TNR = TN / (TN + FP)	Specificity is the fraction of negative examples p model [5]
TPR = TP / (TP + FN)	Sensitivity the fraction of positive examples pr model [5]
Accuracy = (TP + TN) / (TP + TN + FP + FN)	Accuracy is the share of correctly classified objects in the total number of objects. In other words, it shows how often the model is right overall [6].

2 Literature Review

The propensity score is basically a number that tells you how likely something is to happen based on what you already know [7]. Think of it like trying to guess if someone will buy a product online based on their browsing history and past purchases. Now, to make sure our predictions are as accurate as possible, I used a technique called propensity score matching. This involves calculating the propensity score using different methods, like Logistic Regression and Random Forest. The idea is to make sure that the groups I am comparing (like those who bought a product and those who didn't) are as similar as possible in terms of their characteristics. This helps us get a clearer picture of what's really driving the difference between the groups [8] [9] [10].

2.1 Logistic Regression

Logistic Regression is a go-to method for calculating propensity scores, especially in observational studies. It works by modeling the probability of an event happening based on a set of known factors. Essentially, it transforms these factors into a number between 0 and 1, reflecting the likelihood of the event occurring. Formula (1) explains Logistic Regression Model [11].

$$P(Y = 1|X) = \frac{1}{(1 + e^{-\beta X})} \qquad (1)$$

Where Y represents the event indicator variable, X denotes the predictor variables, and β signifies the coefficients in the Logistic Regression model that tell us how much each factor influences the outcome [11]. Logistic Regression shines in situations where we deal with binary outcomes–like whether someone clicks on an ad or not. It helps ensure that the groups we compare are balanced, making it easier to draw conclusions about cause and effect.

Furthermore, a study by Austin and Stuart [12] provided insights into the practical application of Logistic Regression in propensity score analysis for causal inference, demonstrating its effectiveness in balancing covariance across treatment groups. The robustness and interpretability of Logistic Regression

make it a valuable tool for researchers aiming to estimate propensity scores accurately and facilitate causal inference in observational studies [13] [14].

2.2 Random Forest Model

Random Forest, another popular method for propensity score calculation, utilizes an ensemble of decision trees to predict the propensity score. The propensity score in a Random Forest model can be computed by averaging the predictions of all decision trees in the forest. Formula (2) explains Random Forest Model [15].

$$Propensity\_Score_{RF} = \frac{1}{N}\sum_{i=1}^{N} Prediction_i \qquad (2)$$

Where N is the number of decision trees in the Random Forest model, and $Prediction_i$ represents the prediction of the ith decision tree [15].

Random Forest offers several advantages in propensity score calculation, including robustness against over-fitting and the ability to handle complex relationships in the data. It is particularly effective when dealing with high-dimensional data and interactions among variables. Numerous studies have emphasized the utility of Random Forest in propensity score analysis. For instance, Breiman [15] introduced Random Forest as an ensemble learning method that effectively improves prediction accuracy and handles high-dimensional data, making it suitable for propensity score estimation. Additionally, a study by Wager and Athey [16] explored the applications of Random Forest in causal inference, highlighting its flexibility and performance in estimating treatment effects based on propensity scores.

Propensity score production is used in many areas of science and technology like Health-care [17], educational research [18], and financial risk assessment [1], to name a few. This Paper will focus on a comparative study between two popular models, Random Forest, and Logistic Regression, along with the proposed model formulated for this paper for a comparative study to see if it improves the accuracy of the prediction.

3 Methodology

The main focus of this paper is to analyze a sample e-commerce website traffic data and do a comparative study of different propensity score classifications for a certain outcome to occur. Before getting into the comparison, first, let us look at the data schema and how the data will be used in this experiment.

3.1 Data Set

As retrieved from the original dataset source from Kaggle [19], this dataset contains one day's user visit data from a fictional website, with each row being a unique user identified by their unique user ID. The columns represent features of the user's visit, like the device and the actions or events the user did on the website on that day.

```
[UserID] A unique identifier for the visitor
[basket_icon_click] Did the visitor click on the shopping basket icon?
[basket_add_list] Did the visitor add a product to their shopping cart on the 'list' page?
[basket_add_detail] Did the visitor add a product to their shopping cart on the 'detail' page?
[sort_by] Did the visitor sort products on a page?
[image_picker] Did the visitor use the image picker?
[account_page_click] Did the visitor visit their account page?
[promo_banner_click] Did the visitor click on a promo banner?
[detail_wishlist_add] Did the visitor add a product to their wishlist from the 'detail' page?
[list_size_dropdown] Did the visitor interact with a product dropdown?
[closed_minibasket_click] Did the visitor close their mini shopping basket?
[checked_delivery_detail] Did the visitor view the delivery FAQ area on a product page?
[checked_returns_detail] Did the visitor check the returns FAQ area on a product page?
[sign_in] Did the visitor sign in to the website?
[saw_checkout] Did the visitor view the checkout?
[saw_sizecharts] Did the visitor view a product size chart?
[saw_delivery] Did the visitor view the delivery FAQ page?
[saw_account_upgrade] Did the visitor view the account upgrade page?
[saw_homepage] Did the visitor view the website homepage?
[device_mobile] Was the visitor on a mobile device?
[device_computer] Was the visitor on a desktop device?
[device_tablet] Was the visitor on a table device?
[returning_user] Was the visitor new or returning?
[loc_uk] Was the visitor located in the UK, based on their IP address?
[ordered] Did the customer place an order?
```

Figure 1. User visit data schema.

3.2 Data Cleansing and Analysis

To ensure that the data is evenly distributed and that the dataset is more random in nature, I have merged the training and test data that I got from the source, after which I have further split the data into training and test data. Furthermore, I have also removed columns that do not have values.

From the data analysis, the dependent variable, ”ordered,” is categorical. This means, I will need to use a classification model that can predict based on the independent variables whether a user will order or not.

3.3 Correlation and Categorization

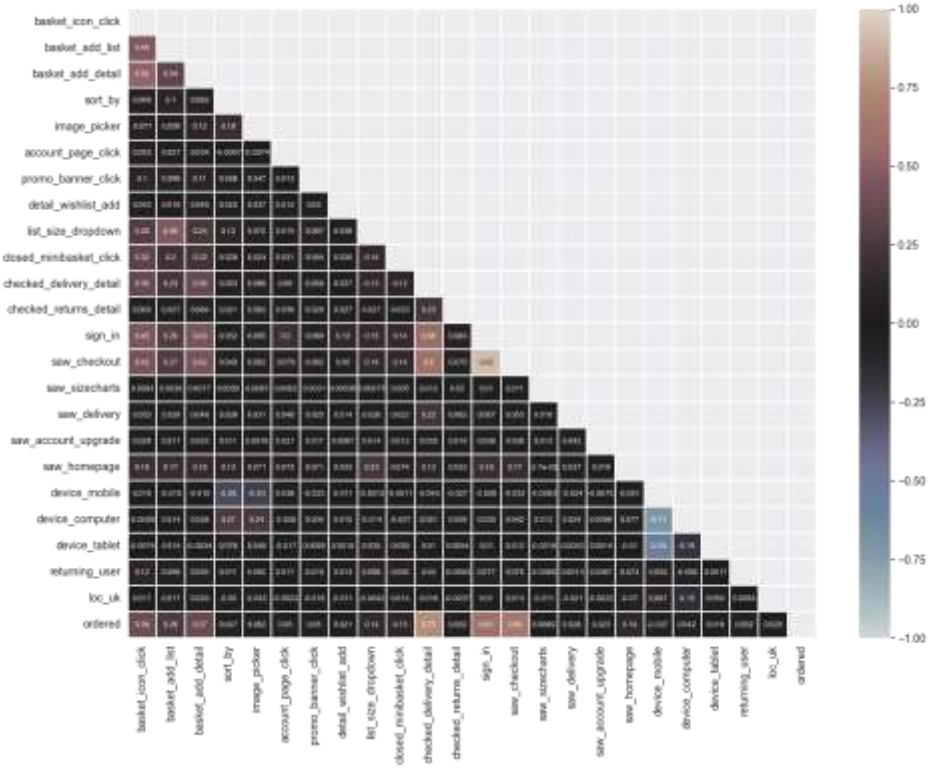

Figure 2. Correlation Matrix

This experiment aims to predict purchase behavior based on user interactions with an e-commerce platform. To achieve this, a comprehensive set of independent variables was identified and categorized using correlation matrices based on their potential influence on purchase events. This set of independent variables is categorized as follows.

1. **Highly Correlated Independent Variables**: An initial analysis of the correlation matrix revealed a set of variables exhibiting strong correlations with the target variable (Ordered). These variables are deemed highly predictive of purchase behavior:
    - **’checked delivery detail’**: This variable reflects user engagement with delivery information, suggesting a higher likelihood of purchase completion.
    - **’sign in’**: Users who sign in demonstrate a greater commitment to the purchase process, potentially indicating a higher conversion rate.
    - **’saw checkout’**: Viewing the checkout page is a strong indicator of purchase intent, as users are actively considering finalizing their purchase.

- **'basket icon click'**, **'basket add list', 'basket add detail'**: These actions directly reflect the addition of items to the shopping cart, signifying a strong intention to purchase.
- **'list size dropdown', 'closed minibasket click'**: These variables suggest active product exploration and customization, potentially leading to a higher likelihood of purchase completion.

2. **"Most Probable" Independent Variables**: In addition to the highly correlated variables, a set of "most probable" variables were identified. While these variables may not exhibit strong correlations, they are considered to have a significant impact on purchase behavior in a practical context:
   - **'returning user'**: Returning users are more likely to purchase due to familiarity with the platform and potentially positive past experiences.
   - **'checked returns detail'**: Users who check return details might be more confident in their purchase decision, indicating a higher likelihood of conversion.
   - **'detail wishlist add'**: Adding items to a wish list suggests potential future purchase intent, highlighting a user's interest in the product.
   - **'promo banner click'**: Users who click on promo banners are often drawn to special offers, which can significantly influence their purchase decision.

3. **All Independent Variables**: For a complete analysis, all independent variables, including those with even a minor correlation, will be included in the study. This approach ensures a complete exploration of user behavior and potentially uncovers unexpected patterns that might not be captured by focusing solely on highly correlated variables.

The inclusion of both highly correlated and "most probable" variables allow a better understanding of purchase behavior. This approach recognizes the importance of direct and indirect purchase intent indicators. Furthermore, incorporating all independent variables ensures a comprehensive analysis, potentially revealing hidden relationships and patterns that might otherwise be missed. The selected variables will be used to develop and evaluate predictive models for purchase behavior. The performance of these models will be evaluated using appropriate metrics, and the results will be analyzed to identify the most influential factors driving purchase decisions.

3.4 Result From Logistic Regression

A logistic regression model was implemented to predict purchase behavior using the three categories of variables: highly correlated, "most probable," and all variables. A confusion matrix was generated for each set of independent variables to assess the model's accuracy. Table 3 summarizes the results.

Table 3: Comparison of Accuracy: Logistic Regression

Variables	TP	FP	TN	FN	Accuracy	TPR	TNR
All variables	3759	793	116788	72	99.29	98.12	99.33
High correlated	3794	860	116721	37	99.26	99.03	99.27
Most probable	3794	859	116722	37	99.26	99.03	99.27

Here are the observations based on the experiment run using Logistic Regression:

1. Accuracy is overall high for the model run with all independent variables.
2. While true positive (predicting that user would order) is high with model ran with "Highly correlated" and "Most probable" independent variables.
3. True negative (predicting that user would not order) is high with all variables.
4. While false negative is high for a model with all variables, false positive is high with the other two models.

5. There is no statistically significant difference between highly correlated and most probable variables.
6. Propensity score for predicted value 1 is in-between 50 and 100 and for predicted value 0 is below 50, which looks like a good distribution, which means higher propensity values users are ordering more.

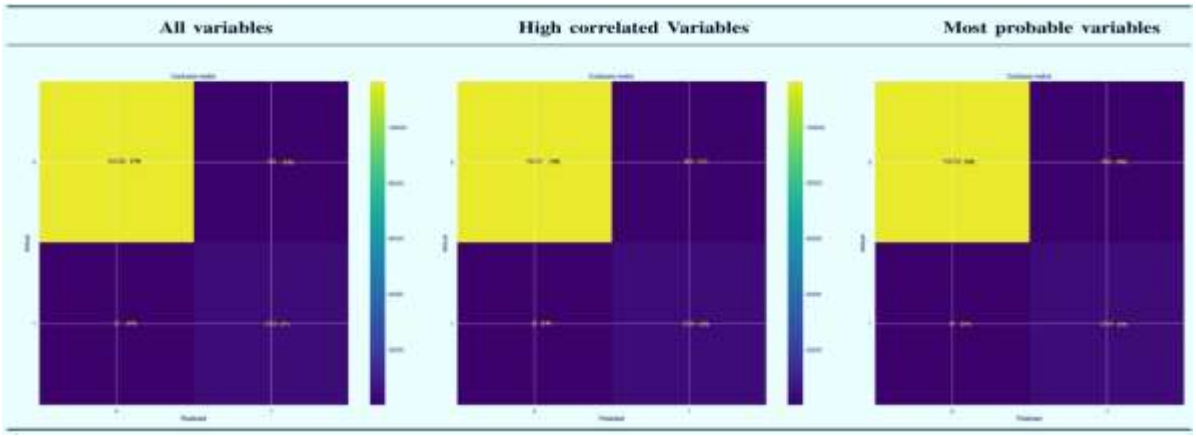

Figure 3. Comparison of the Confusion matrix: Logistic Regression

3.5 Result From Random Forest

The experiment was run with 'depth' ranging from 1 to 5, and in each depth, the test was run with 'n estimator' ranging from 1 to 9. Table 4 shows the TP, FP, TN, and FN, which have given the best accuracy in the three categories of independent variables.

Table 4: Comparison of Accuracy: Random Forest

Variables	TP	FP	TN	FN	Accuracy	TPR	TNR
All variables	3762	841	116740	69	99.25	98.20	99.28
High correlated	3794	860	116721	37	99.26	99.03	99.27
Most probable	3794	854	116727	37	99.27	99.03	99.27

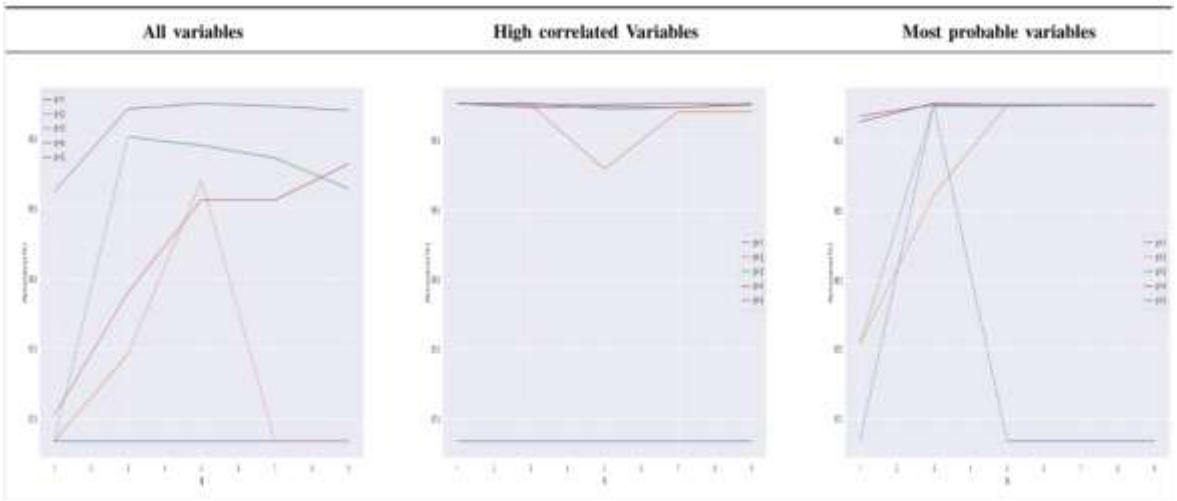

Figure 4. Random Forest: 'Depth' and 'n estimator' vs Accuracy

The propensity score distribution from 1 to 100 is shown in Figure 5.

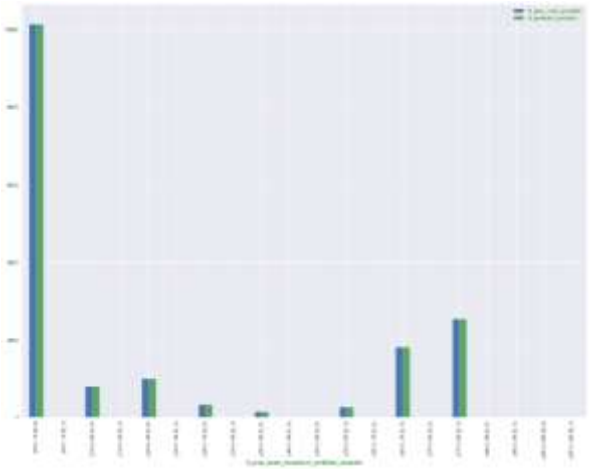

Figure 5. Random Forest propensity score distribution.

Here are the observations based on the experiment run using the Random Forest model:

1. The experiment output shows that the highest Accuracy was for the test done on estimator = 3 with Depth = 4.
2. With the Random Forest model experiment, the propensity distribution for predicting a purchase would be 1 is between 50 and 80, Compared to logistic regression this is not a great distribution.

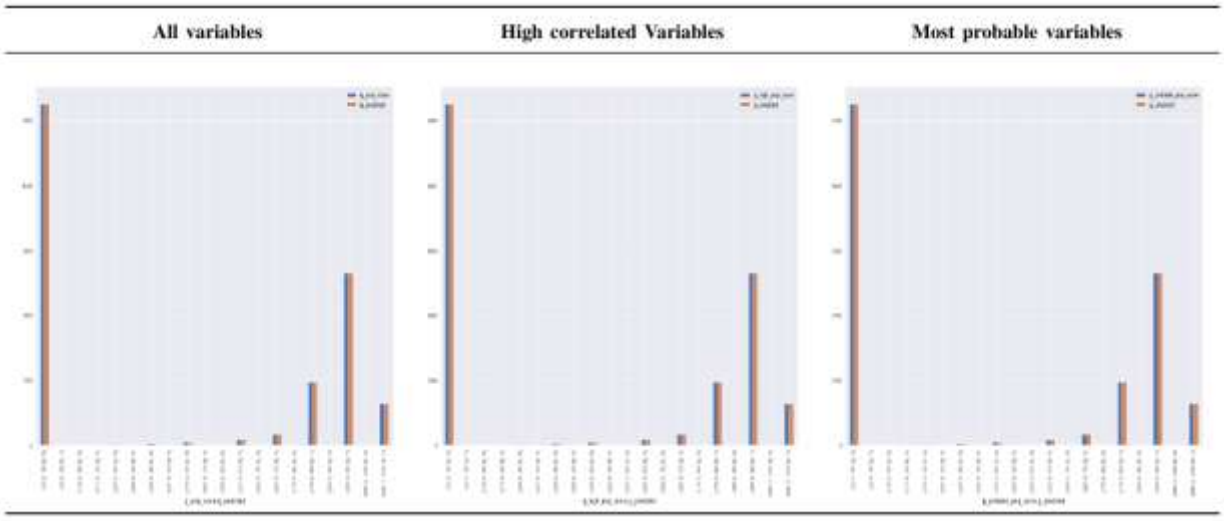

Figure 6: Comparison of the predicted values grouped for values from 1 to 100 .

3.6 Proposed Logic: Normalized Mean Value Model

For this experiment, I want to validate whether the calculated propensity score and its distribution that is achieved are only due to certain correlated independent variables along with that, I also want to check if there are any new improvements or insights that can be brought to enhance our prediction models. To do so, I will discuss here a new algorithm propensity score calculation logic, which I call the ”Normalized Mean Value Model”. By running this Novel Algorithmic Approach, I will be able to identify if a new methodology can be beneficial along with the popular Models and if there are gaps in the study that this Model can fill as a complementary approach alongside Logistic Regression and Random Forest Models. Let us examine how this model was designed.

First, I have divided the independent variables into three categories:

- Highly correlated variables like ’checked delivery detail’, ’sign in’, ’saw checkout’
- Medium correlated variables ’basket icon click’, ’basket add list’, ’basket add detail’, ’list size dropdown’
- Low correlated variables: which includes all independent variables other than what is there in High and Medium categories.

Next, I have assigned an arbitrary propensity score range for each category based on its relevance to the outcome of this test, which is the 'ordered' field. I assigned max values as 100 for high, 80 for medium and 60 for low correlated variables, and these values are called the normalized max values or 'nor_max.' In addition, I will define the Min value, which will be the mean of the variables multiplied by the 'nor_max' value for each category.

I have the following parameters for this formula (3) that I have defined for this model to calculate the propensity score:

- **mean**: This is the mean value calculated for the independent variables.
- **act min**: This is the actual output min value possible for the propensity, which is 0.
- **act max**: This is the maximum actual output value possible for the propensity,and this value is 1.
- **nor max**: This is the normalized maximum values assigned to each category of variables (100, 80, 60).
- **nor min**: This is the normalized output min value calculated for each category of independent variable, which is the mean value of the variables multiplied by the nor_max value assigned to each category of variables.

$$\text{Propensity} = \frac{(\text{mean} - \text{act min}) \times (\text{nor max} - \text{nor min})}{(\text{act max} - \text{act min}) + \text{nor min}} \quad (3)$$

3.7 Normalized Mean Value Model Results

The propensity score calculated with Normalized Mean Value Model logic was further tested to identify the propensity score range that should be picked to get the best prediction. Based on this test, Propensity scores greater than 60 and greater than 80 had more False Positives (FP) results, and Values greater than 90 had the best accuracy overall. Figure 7 shows the confusion matrix for scores above 90.

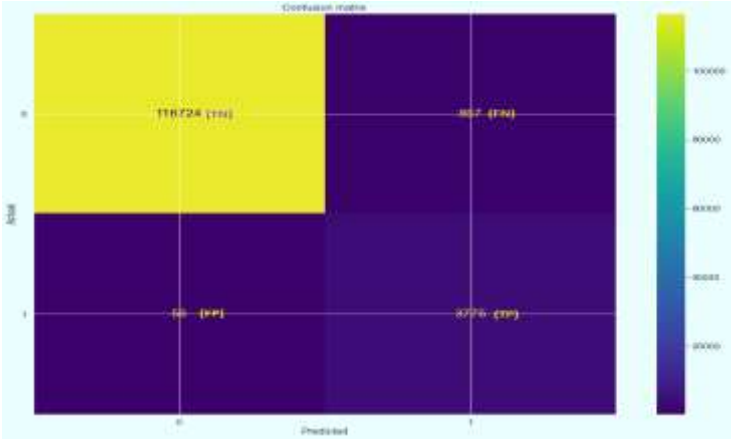

Figure 7. Confusion matrix: Normalized Mean Value Model Logic

The Accuracy calculated is given in Table 5.

Table 5: Accuracy: Normalized Mean Value Model Logic: Normalized Mean Value Model

Variables	TP	FP	TN	FN	Accuracy	TPR	TNR
Score > 90	3775	857	116724	56	99.25	98.54	99.27
Score > 80	3827	4779	112802	4	96.06	99.90	95.94
Score > 60	3828	9162	108419	3	92.45	99.92	92.21

The distribution of propensity based on the Normalized Mean Value Model logic is given in Figure 8.

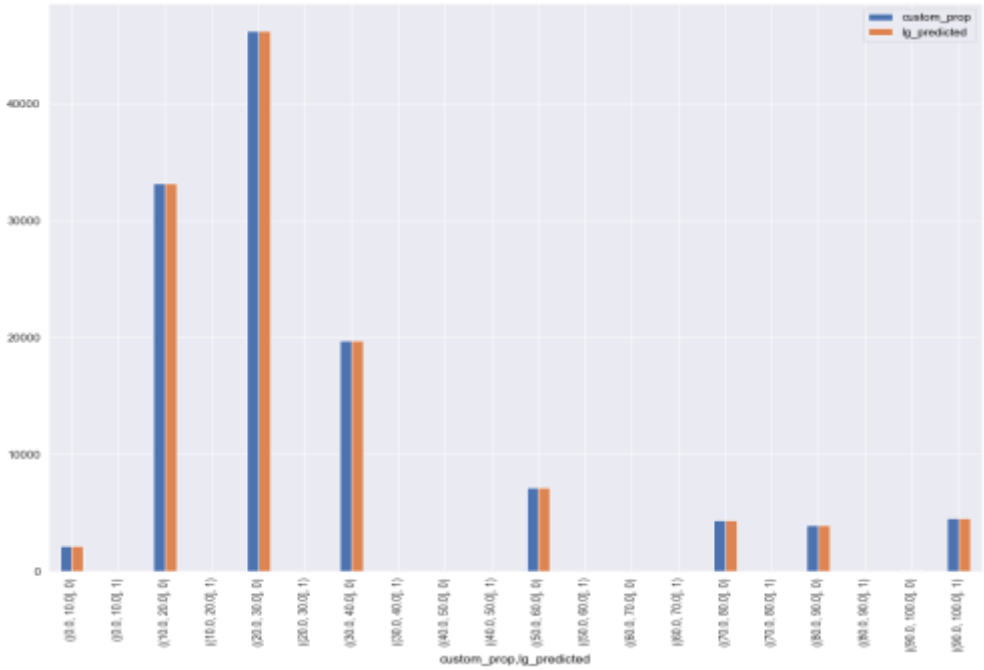

Figure 8. ”Normalized Mean Value Model” Propensity Score distribution.

Here is a summary of the result:

1. The test conducted with Normalized Mean Value Model Logic gave the insight that the propensity score calculation heavily depends on the High correlated variables.
2. The propensity score calculated has a distribution where most predictions where the value is 1 fall within the 90 and 100 buckets.
3. Comparing with the Logistic Regression model, the Normalized Mean Value Model logic accuracy was only slightly lower; this may be further improved by adjusting the range of propensity scores.

4 Observations from the Results

Here are some key observations from the results:

1. Accuracy-wise, Random Forest gave a slightly better prediction, but Logistic regression gives better propensity score matching. Normalized Mean Value Model Logic also gave a better accuracy rate and prediction than the two models [20] [21].

2. As shown in Figure 9, most matching values are very close to the predicted score for Logistic regression.
3. In the case of Random Forest Figure 10, we can see a few places where the match is a bit away from the prediction.
4. Normalized Mean Value Model logic has good accuracy and can be a great way to check the impact of variable correlation.
5. Normalized Mean Value Model logic can be complemented with other traditional models to validate and refine the result and aligns with heuristic methods for score estimation that have been explored in various contexts [22], though our specific implementation represents a novel application to propensity scoring

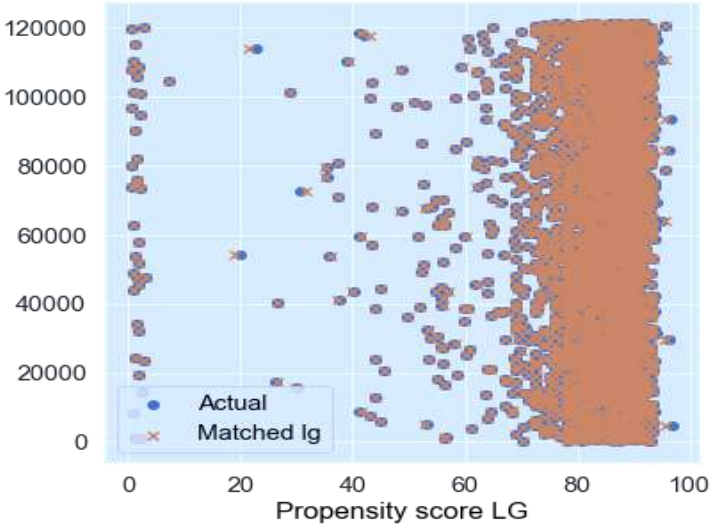

Figure 9. Comparing Logistic Regression vs Actual

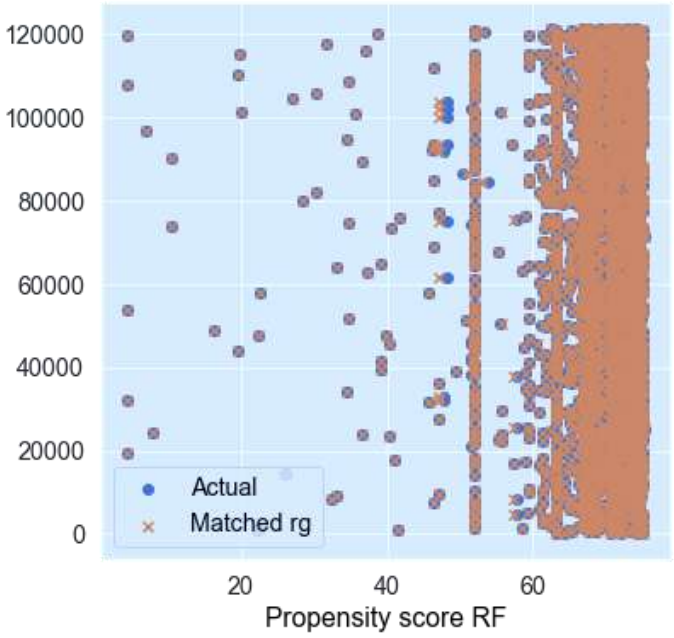

Figure 10. Comparing Random Forest vs Actual

5 Conclusion

In this study, I explored the utilization of Logistic Regression and Random Forest in estimating propensity scores for observational studies [23]. Our findings suggest that while Random Forest provided slightly superior predictive accuracy, propensity score matching using Logistic Regression yielded more favorable outcomes [12] [24].

The analysis revealed that Random Forest, known for its robustness and ability to handle complex data relationships, excelled in prediction accuracy. However, when achieving optimal results in propensity score matching, Logistic Regression emerged as the preferred choice [25] [26].

Furthermore, this paper illustrates that other Normalized Mean Value Model logic, like the ”Normalized Mean Value Model,” can be developed to predict the outcome based on independent variables, which can give good accuracy in prediction along with other insights to improve the tests. Also, a broader area of research and development scope can be opened up to improve the accuracy of this Normalized Mean Value Model logic; by doing so, the Complementary strengths of all three approaches can be used to enhance this area of study.

6 References

[1] D. Zhao, J. Dai and C. Song, "Systemic financial risk prediction using least squares support vector machines," Modern Physics Letters B, vol. 32, no. 17, p. 1850183, 2018.

[2] M. Soledad Cepeda, R. B. Boston, J. T. Farrar, F. B. L. Strom and S. M. Hennessy, "Comparison of Logistic Regression versus Propensity Score When the Number of Events Is Low and There Are Multiple Confounders," American Journal of Epidemiology, vol. 158, no. 3, pp. 280-287, 2003.

[3] R. Louronne, P. Pampel and A. Beyersmann, "Random forest versus logistic regression: a large-scale benchmark experiment," BMC Bioinformatics, vol. 19, no. 270, 2018.

[4] "Propensity Score Matching," 2019. [Online]. Available: https://dimewiki.worldbank.org/Propensity_Score_Matching.

[5] K. Ting, "Confusion Matrix," in Encyclopedia of Machine Learning, Springer, 2011, p. 209.

[6] O. Caelen, " A Bayesian interpretation of the confusion matrix. Annals of Mathematics and Artificial Intelligence," Annals of Mathematics and Artificial Intelligence, vol. 81, no. 3, pp. 429-450, 2017.

[7] D. B. Rubin, "Causal inference using potential outcomes: Design, modeling, decisions," Journal of the American Statistical Association, vol. 100, no. 469, pp. 322-331, 2005.

[8] J. S. Haukoos and R. J. Lewis, "The Propensity Score," JAMA, vol. 314, no. 15, pp. 1637-1638, 2015.

[9] D. Muchlinski, D. Siroky, J. He and M. Kocher, "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data," Political Analysis, vol. 24, no. 1, pp. 87-103, 2016.

[10] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, pp. 273-297, 1995.

[11] D. W. Hosmer, S. Lemeshow and R. X. Sturdivant, Applied Logistic Regression, 3rd ed., Hoboken, NJ: John Wiley & Sons, 2013.

[12] P. C. Austin, "An introduction to propensity score methods for reducing the effects of confounding in observational studies," Multivariate Behavioral Research, vol. 46, no. 3, pp. 399-424, 2011.

[13] G. W. Imbens, "Nonparametric estimation of average treatment effects under exogeneity: A review," Review of Economics and Statistics, vol. 86, no. 1, pp. 4-29, 2004.

[14] J. Pearl, "Causal inference in statistics: An overview," Statistics Surveys, vol. 3, pp. 96-146, 2009.

[15] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5-32, 2001.

[16] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," Journal of the American Statistical Association, vol. 113, no. 523, pp. 1228-1242, 2018.

[17] R. B. D'Agostino, "Propensity Scores in Cardiovascular Research," Circulation, vol. 115, no. 17, 2007.
[18] F. Lane, Y. To, K. Shelley and R. Henson, "An Illustrative Example of Propensity Score Matching with Education Research," Career and Technical Education Research, vol. 37, no. 3, pp. 187-212, 2012.
[19] B. Powis, "Customer propensity to purchase dataset," 2018. [Online]. Available: https://www.kaggle.com/datasets/benpowis/customer-propensity-to-purchase-data.
[20] D. E. Ho, K. Imai, G. King and E. A. Stuart, "MatchIt: Nonparametric preprocessing for parametric causal inference," Journal of Statistical Software, vol. 42, no. 8, pp. 1-28, 2011.
[21] R. H. Dehejia and S. Wahba, "Propensity score matching methods for nonexperimental causal studies," Review of Economics and Statistics, vol. 84, no. 1, pp. 151-161, 2002.
[22] J. Hainmueller, "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies," Political Analysis, vol. 20, no. 1, pp. 25-46, 2012.
[23] E. A. Stuart, "Matching methods for causal inference: A review and a look forward," Statistical Science, vol. 25, no. 1, pp. 1-21, 2010.
[24] P. C. Austin and E. A. Stuart, "Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies," Statistics in Medicine, vol. 34, no. 28, pp. 3661-3679, 2015.
[25] B. K. Lee, J. Lessler and E. A. Stuart, "Improving propensity score weighting using machine learning," Statistics in Medicine, vol. 29, no. 3, pp. 337-346, 2010.
[26] V. Chernozhukov, D. Chetyrkov, M. Demirer, E. Duflo and C. Hansen, "Double/debiased machine learning for treatment and structural parameters," Econometrics Journal, vol. 21, no. 1, pp. C1-C68, 2018.
[27] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," Biometrika, vol. 70, no. 1, pp. 41-55, 1983.