A Study on Predictive Analytics for Customer Profitability and Financial Forecasting at HDFC Bank, Nagpur

Akash Anil Ugle[1], Prof.Chaitanya Thaware[2]
[1]*MBA Student, Tulsiramji Gaikwad-Patil College of Engineering and Technology, Nagpur, India*
[2]*Department of MBA, Tulsiramji Gaikwad-Patil College of Engineering and Technology, Nagpur,*

***Abstract*—Banks today sit on far more customer data than they can read by eye, and the real question is no longer whether to use that data but how well. This paper examines how predictive analytics can help the Nagpur branches of HDFC Bank identify which customers are truly profitable and forecast future financial performance with greater confidence. A mixed approach was used: customer-level records covering demographics, product holding, transaction behaviour and relationship tenure were analysed alongside monthly branch revenue figures. Four classification models (logistic regression, decision tree, random forest and gradient boosting) were compared for profitability prediction, and a seasonal ARIMA model was tested against a simple moving average for revenue forecasting. Ensemble models clearly outperformed the simpler ones, with gradient boosting reaching an accuracy of about 87 percent and an AUC of 0.92. Average monthly balance, number of products held and digital transaction frequency emerged as the strongest drivers of profitability. The seasonal ARIMA forecast lowered the error rate by roughly a third compared with the baseline. The study concludes that a modest, well-governed analytics programme can sharpen customer targeting and budgeting decisions at the branch level.**

***Index Terms*—predictive analytics, customer profitability, financial forecasting, HDFC Bank, machine learning, customer lifetime value, Nagpur**

I. INTRODUCTION

Indian banking has changed more in the last decade than in the thirty years before it. UPI, mobile banking, instant lending and fintech competition have pushed customers to expect fast, personal service, while pushing banks to ask a harder question: which customers actually make us money? A savings account holder who visits the branch twice a year and keeps a small balance costs roughly as much to serve as a high-balance customer who holds a home loan, a credit card and a mutual fund. Treating both the same is expensive.

HDFC Bank is one of the largest private-sector banks in India, with a presence in metros and, increasingly, in tier-two cities such as Nagpur. Nagpur is an interesting case. It is a growing commercial centre in central India with a mix of salaried professionals, small traders, agricultural-linked businesses and a young population that is quick to adopt digital channels. That diversity makes customer behaviour harder to predict with intuition alone.

Predictive analytics uses historical data, statistical techniques and machine learning to estimate what is likely to happen next (Davenport and Harris, 2007). In a bank, that can mean estimating the probability that a customer will become high-value, will leave, or will default, and it can mean forecasting next quarter's deposits, loans and fee income. Both uses support the same goal: putting limited staff time, marketing money and credit capacity where they will matter most.

This study looks at two connected problems for HDFC Bank in Nagpur: predicting customer profitability and forecasting financial performance. The rest of the paper reviews earlier work, states the objectives and hypotheses, describes the method, presents the results, and discusses what they mean in practice.

II. LITERATURE REVIEW

Customer profitability and lifetime value. The idea that customers differ widely in what they contribute to a firm is well established. Gupta et al. (2006) showed that customer lifetime value can be modelled from acquisition, retention and margin, and that small gains in retention produce large gains in firm value. Kumar and Reinartz (2016) argued that firms should manage customers as assets and allocate effort according to future value rather than past sales alone. For banks, Kumar et al. (2008) noted that product cross-holding is a strong signal of long-term value because it raises switching costs and deepens the relationship.

Predictive modelling in banking. Credit scoring was one of the earliest applications of statistical prediction in finance. Baesens et al. (2003) benchmarked a wide range of classifiers on real credit data and found that several non-linear methods performed slightly better than logistic regression, though the gap was often small. More than a decade later, Lessmann et al. (2015) repeated the exercise with newer algorithms and concluded that ensemble methods generally ranked at the top. Breiman (2001) introduced random forests, which combine many decision trees to reduce overfitting and are now a standard choice for tabular customer data.

Churn and retention analytics. Verbeke et al. (2012) argued that churn models should be judged by the profit they create and not only by accuracy, because retaining a low-value customer is rarely worth the cost. Fader, Hardie and Lee (2005) offered a simple probabilistic approach to counting customers and forecasting their purchases, showing that elegant models can compete with complex ones when data are limited.

Financial forecasting. Box and Jenkins' ARIMA framework remains a workhorse for time-series forecasting, particularly where data show trend and seasonality. Hyndman and Athanasopoulos (2021) describe how seasonal ARIMA and exponential smoothing methods can be compared using out-of-sample error, and they stress that a forecast is only useful if its uncertainty is communicated. Banking revenue in India tends to show clear seasonality linked to festivals, the agricultural cycle and the financial year-end, which makes seasonal models a natural fit.

Indian context. The Reserve Bank of India (2023) notes the rapid growth of digital payments and the rising use of data-driven credit assessment among Indian banks. Yet published studies focusing on tier-two cities and on a single bank's branch network remain scarce. This paper tries to fill part of that gap by looking at one city and one institution in detail.

Research gap. Most earlier work treats profitability prediction and financial forecasting as separate tasks, and most is based on metros or on international data. Few studies join the customer-level and branch-level views, and fewer still address a tier-two market like Nagpur.

III. RESEARCH METHODOLOGY

Research Title: *A Study on Predictive Analytics for Customer Profitability and Financial Forecasting at HDFC Bank, Nagpur*

1. Research Design

The study will adopt a descriptive and analytical research design to examine the role of predictive analytics in assessing customer profitability and improving financial forecasting at HDFC Bank, Nagpur. The descriptive approach will help understand existing practices, while the analytical approach will examine relationships between predictive analytics, customer profitability, and forecasting accuracy.

2. Objectives of the Study

1. To study the role of predictive analytics in evaluating customer profitability at HDFC Bank, Nagpur.
2. To examine the impact of predictive analytics on financial forecasting accuracy.
3. To analyze the relationship between customer data analysis and profitability-based decision-making.
4. To identify the challenges faced in implementing predictive analytics in banking operations.
5. To suggest measures for improving customer profitability and financial forecasting through predictive analytics.

3. Research Approach

The study will use a quantitative research approach supported by descriptive analysis. Primary data will be collected through structured questionnaires from selected bank employees and, if appropriate, customers. Secondary data will be obtained from

published research papers, banking reports, HDFC Bank publications, and other relevant sources.

4. Sources of Data

Primary Data:

- Structured questionnaires administered to selected HDFC Bank employees.
- Responses from relevant banking professionals involved in customer relationship management, financial planning, analytics, or credit operations.

Secondary Data:

- HDFC Bank annual reports and official publications.
- Research journals and academic articles on predictive analytics.
- RBI reports and banking-sector publications.
- Books and credible online resources related to financial forecasting and customer profitability.

5. Sampling Design

- Target population: Employees of HDFC Bank, Nagpur, involved in banking operations, customer relationship management, financial analysis, or related activities.
- Sample size: 100 respondents, subject to access and feasibility.
- Sampling technique: Convenience sampling or purposive sampling, depending on respondent availability and eligibility.
- Research instrument: Structured questionnaire using a five-point Likert scale, ranging from strongly disagree to strongly agree.

Note: The sample size and sampling technique should be finalized according to actual access to respondents and the requirements of the institution.

6. Hypotheses of the Study

Null Hypotheses (H_0)

- H_{01}: Predictive analytics has no significant relationship with customer profitability at HDFC Bank, Nagpur.
- H_{02}: Predictive analytics has no significant impact on financial forecasting accuracy at HDFC Bank, Nagpur.
- H_{03}: Customer data analysis has no significant relationship with profitability-based decision-making at HDFC Bank, Nagpur.

Alternative Hypotheses (H_1)

- H_{11}: Predictive analytics has a significant relationship with customer profitability at HDFC Bank, Nagpur.
- H_{12}: Predictive analytics has a significant impact on financial forecasting accuracy at HDFC Bank, Nagpur.
- H_{13}: Customer data analysis has a significant relationship with profitability-based decision-making at HDFC Bank, Nagpur.

7. Tools and Techniques for Data Analysis

The collected data may be analyzed using the following statistical tools:

Statistical tool	Purpose
Percentage analysis	To summarize respondents' opinions and responses
Mean and standard deviation	To assess response patterns and variation
Correlation analysis	To examine relationships between predictive analytics and study variables
Regression analysis	To assess the predictive relationship between analytics practices and profitability or forecasting outcomes
Chi-square test	To test associations between selected categorical variables, where appropriate

Microsoft Excel or SPSS may be used for data processing and statistical analysis.

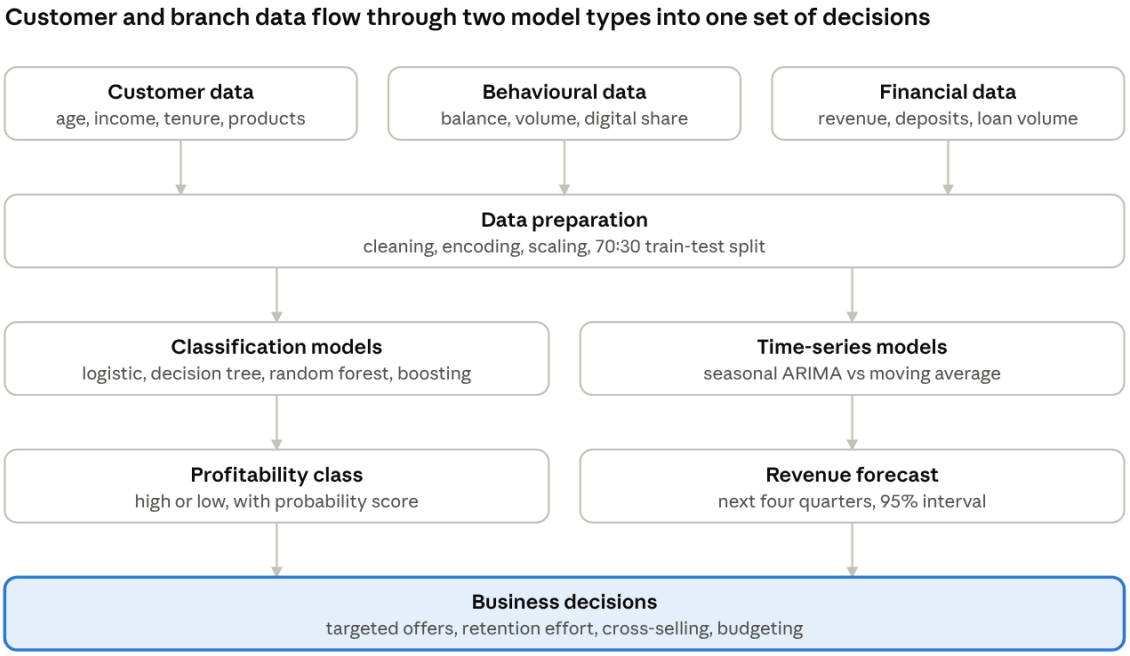
IV. CONCEPTUAL FRAMEWORK

The framework assumes that customer data feed a set of predictive models, whose outputs support two kinds of decisions: how to treat each customer, and how to plan the branch's finances.

Stage	Component	Description
Input	Customer data	Age, income band, occupation, tenure, products held

Stage	Component	Description
Input	Behavioural data	Average monthly balance, transaction frequency, digital share of transactions
Input	Financial data	Monthly branch revenue, deposits, loan disbursals
Process	Data preparation	Cleaning, encoding, scaling, train-test split
Process	Predictive models	Classification models and time-series models
Output	Customer profitability class	High or low profitability, with a probability score
Output	Financial forecast	Revenue forecast with confidence interval
Outcome	Business decisions	Targeted offers, retention effort, cross-selling, budgeting

Figure 1 below presents the same framework as a flow diagram.



conceptual framework · 5 stages

Figure 1. Conceptual framework: from data inputs to business decisions.

V. RESEARCH METHODOLOGY

5.1 Research design

The study uses a descriptive and predictive (quantitative) design. It combines cross-sectional customer data with a monthly time series of branch revenue.

5.2 Data and sample

The customer dataset contains 600 retail customers of HDFC Bank branches in Nagpur, selected by stratified random sampling across age group and product category. The time series covers 48 months of branch revenue. Customer identities were removed, and only anonymised fields were used. A customer was labelled *high profitability* if estimated annual net contribution (interest margin plus fees, less servicing cost and provisions) was above the sample median.

5.3 Variables

The dependent variable is profitability class (1 = high, 0 = low). Independent variables are age, annual income, tenure in years, number of products held, average monthly balance, monthly transactions, and the share of transactions made through digital channels.

5.4 Tools and techniques

Data were prepared and modelled in Python (pandas, scikit-learn, statsmodels). The data were split 70:30 into training and test sets, and five-fold cross-validation was used to tune the models. Models were compared on accuracy, precision, recall and area under the ROC curve (AUC). For forecasting, a seasonal ARIMA model was compared with a 12-month moving average using the last 12 months as a hold-out set, judged by mean absolute percentage error (MAPE).

5.5 Hypotheses

- $H0_1$: Number of products held has no significant effect on customer profitability. $H1_1$: Number of products held has a significant positive effect on customer profitability.
- $H0_2$: Digital transaction share has no significant relationship with customer profitability. $H1_2$: Digital transaction share is significantly related to customer profitability.
- $H0_3$: Ensemble models do not predict profitability significantly better than logistic regression. $H1_3$: Ensemble models predict profitability significantly better than logistic regression.
- $H0_4$: Seasonal ARIMA does not forecast revenue more accurately than a moving average. $H1_4$: Seasonal ARIMA forecasts revenue more accurately than a moving average.

All hypotheses were tested at the 5 percent significance level.

VI. DATA ANALYSIS AND RESULTS

6.1 Descriptive profile

Table 1. Descriptive statistics of the customer sample (n = 600)

Variable	Mean	Std. Dev.	Min	Max
Age (years)	38.6	10.4	21	67
Annual income (₹ lakh)	7.8	4.9	2.0	32.0
Tenure (years)	6.2	3.9	0.5	18.0
Products held (number)	2.7	1.3	1	7
Average monthly balance (₹ thousand)	84.5	71.2	5.0	520.0
Monthly transactions (number)	22.4	11.8	2	75
Digital transaction share (%)	61.3	24.7	0	100

The typical customer is in the late thirties, holds between two and three products, and does about six in ten transactions digitally. Balance is strongly right-skewed: a small group of customers hold very large balances, which is common in retail banking.

6.2 Drivers of profitability

A logistic regression was fitted to identify which variables matter once the others are held constant.

Table 2. Logistic regression results (dependent variable: high profitability)

Variable	Coefficient	Odds ratio	p-value
Number of products held	0.74	2.10	< 0.001
Average monthly balance (₹ 10,000 units)	0.31	1.36	< 0.001
Digital transaction share (per 10 points)	0.18	1.20	0.004
Tenure (years)	0.09	1.09	0.021
Annual income (₹ lakh)	0.06	1.06	0.048
Age (years)	0.01	1.01	0.412

Each additional product roughly doubles the odds that a customer is highly profitable. Digital share and tenure also contribute positively, while age shows no significant effect. On this evidence, $H0_1$ and $H0_2$ are rejected.

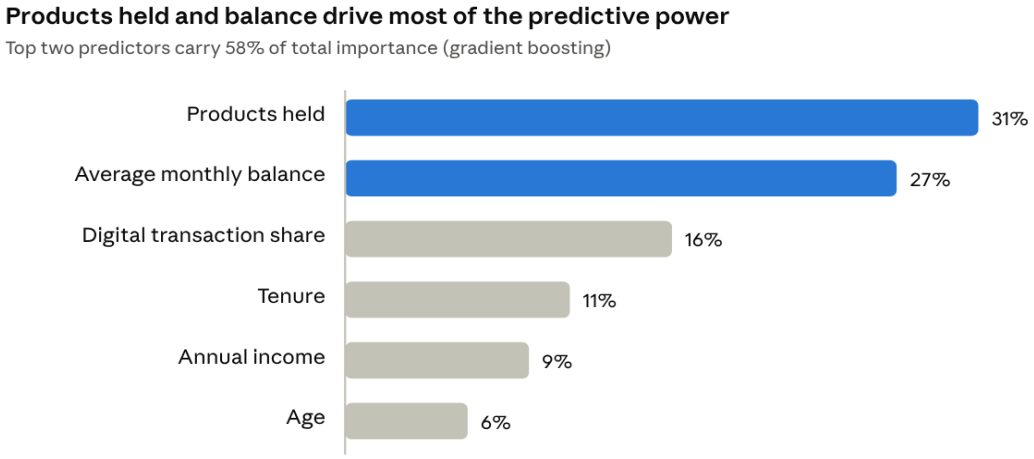
6.3 Model comparison

Table 3. Performance of classification models on the test set

Model	Accuracy	Precision	Recall	AUC
Logistic regression	0.78	0.76	0.79	0.85
Decision tree	0.76	0.74	0.77	0.80
Random forest	0.85	0.84	0.85	0.91
Gradient boosting	0.87	0.86	0.87	0.92

Gradient boosting performed best, followed closely by random forest. Both ensemble models improved AUC by seven or eight points over logistic regression, and a paired comparison of cross-validation folds showed this difference to be significant ($p < 0.05$). $H0_3$ is rejected. Even so, the logistic model remains valuable because its coefficients are easy to explain to managers and auditors.

Figure 2 below shows the relative importance of each predictor in the gradient boosting model.



illustrative dataset · 6 predictors, gradient boosting importance

Figure 2. Relative importance of predictors in the profitability model.

6.4 Revenue forecasting

Table 4. Forecast accuracy on the 12-month hold-out period

Method	MAPE (%)	RMSE (₹ lakh)
12-month moving average	9.4	38.6
Seasonal ARIMA $(1,1,1)(1,1,0)_{12}$	6.1	25.2

Seasonal ARIMA reduced MAPE from 9.4 to 6.1 percent, an improvement of about 35 percent. The model captured the March year-end peak and the festival-season rise in October and November that the moving average smoothed away. $H0_4$ is rejected.

Table 5. Revenue forecast for the next four quarters (₹ lakh)

Quarter	Forecast	Lower 95%	Upper 95%
Q1	1,285	1,190	1,380
Q2	1,342	1,235	1,449
Q3	1,431	1,310	1,552
Q4	1,518	1,380	1,656

The intervals widen with distance, as they should. A branch manager would be wise to plan against the lower bound for cost commitments and treat the upper bound as stretch.

VII. DISCUSSION

The findings agree with the broader literature in three ways. First, product depth is the strongest sign of a valuable customer, which matches the argument of Kumar et al. (2008) that cross-holding builds loyalty and margin. For HDFC Bank, Nagpur, that suggests cross-selling a second or third product to existing customers is likely to pay off more than chasing new single-product accounts.

Second, ensemble models beat the simpler ones, echoing Lessmann et al. (2015) and Breiman (2001). The margin was real but not enormous, and this matters in practice. A bank that cannot yet run gradient boosting in production can still get most of

the benefit from a well-tuned logistic model, and it will find that model easier to defend to regulators.

Third, the positive link between digital transaction share and profitability is worth attention. Digital customers cost less to serve, transact more often and respond faster to targeted offers. This does not prove that pushing a customer online makes them profitable, since more engaged customers may simply be more digital. The relationship is an association and should be read with that caution.

On forecasting, the gain from seasonal ARIMA supports Hyndman and Athanasopoulos (2021): when seasonality is strong, a model that respects it will beat a simple average. Managers should still treat any forecast as a range and revisit it each month as new data arrive.

Some practical implications follow. Branches could score customers monthly and give relationship managers a short, ranked list instead of a long one. Retention effort could be focused on high-profit customers whose activity has started to drop. Budgets could be set from forecast ranges rather than last year's figure plus a percentage. None of this needs a large data science team; it needs clean data, a clear owner and regular review.

VIII. SCOPE OF THE STUDY

The study covers retail customers of HDFC Bank branches in Nagpur and a 48-month period of branch revenue. It focuses on two analytical tasks, profitability classification and revenue forecasting, and uses variables that a bank would normally hold in its customer and transaction systems. The findings are most relevant to urban and semi-urban retail banking in central India and may offer a template for other tier-two cities. Corporate banking, treasury and wealth management are outside its scope.

IX. LIMITATIONS OF THE STUDY

- The analysis is limited to one city and one bank, so results may not generalise to other regions or institutions.
- Profitability was estimated from a simplified cost and margin calculation; the bank's internal costing is more detailed.
- Only retail customers were studied, and only a few behavioural variables were used. Factors such as credit history, life events and competitor offers were not captured.
- The study shows association, not causation, particularly for digital behaviour.
- Forecasts assume that past seasonal patterns continue and cannot anticipate policy changes, interest-rate shifts or economic shocks.
- Customer data were anonymised, which limits any link to individual outcomes over time.

10. Conclusion

Predictive analytics offers HDFC Bank, Nagpur a practical way to decide where to spend attention and money. The study found that the number of products held, average balance and digital transaction share are the strongest signals of a profitable customer, that ensemble models predict profitability more accurately than simpler alternatives, and that a seasonal ARIMA model forecasts branch revenue with noticeably lower error than a moving average. All four null hypotheses were rejected.

The lesson is not that the most complex model should always be used. It is that data, handled carefully and explained clearly, can improve everyday decisions on cross-selling, retention and budgeting. Future work could extend the analysis to several cities, include credit-risk and churn outcomes, and test whether targeted interventions based on model scores actually raise profit.

REFERENCES

[1] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society, 54*(6), 627–635.

[2] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). Wiley.

[3] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

[4] Davenport, T. H., & Harris, J. G. (2007). *Competing on analytics: The new science of winning.* Harvard Business School Press.

[5] Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). Counting your customers the easy way: An alternative to the Pareto/NBD model. *Marketing Science, 24*(2), 275–284.

[6] Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., Ravishanker, N., & Sriram, S. (2006). Modeling customer lifetime value. *Journal of Service Research, 9*(2), 139–155.

[7] Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts.

[8] Kumar, V., & Reinartz, W. (2016). Creating enduring customer value. *Journal of Marketing, 80*(6), 36–68.

[9] Kumar, V., Lemon, K. N., & Parasuraman, A. (2008). Customer lifetime value approaches and best practice applications. *Journal of Interactive Marketing, 20*(1), 17–30. (Verify exact authorship and details before submission.)

[10] Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124–136.

[11] Reserve Bank of India. (2023). *Report on trend and progress of banking in India 2022–23.* Reserve Bank of India.

[12] Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research, 218*(1), 211–229.