Bank Customer Identification for Targeted Marketing and Revenue Optimisation: A Comparative Analysis of Predictive Models

Sambhav Khanna[1]
Computer Science Engineering
Maharaja Agrasen Institute of Technology
Delhi, India
sambhav.khanna56@gmail.com

Vatsal Sharma[2]
Computer Science Engineering
Maharaja Agrasen Institute of Technology
Delhi, India
sharma.vatsal6@gmail.com

Prakhar Gautam[3]
Computer Science Engineering
Maharaja Agrasen Institute of Technology)
Delhi, India
prakhar26gautam@gmail.com

***Abstract*—This research leads a transformative shift in banking by using machine learning to identify high-value customers, crucial for targeted marketing. In the digital era, strategic customer targeting is vital for improved financial performance. Leveraging a comprehensive dataset, our study deploys various classifiers like Logistic Regression, Random Forest, KNN, Naive Bayes, Gradient Boosting, and SVM. Through meticulous feature selection and hyper-parameter tuning, SVM Hypertuned emerges as the optimal classifier, showcasing superior precision, recall, and F1-Score. The implementation includes classifier initialisation, preprocessing, and training, with a dedicated Decile analysis for nuanced insights. The research contributes a comparative analysis of classifiers, enhancing methodology in feature selection and hyper-parameter tuning. With a focus on replacing manual forecasting, our motivation is to empower banks with an efficient machine learning-based framework for customer identification, fostering informed decision-making, strategic customer targeting, and improved financial performance. Addressing the critical challenge of efficiently identifying potential customers, the study highlights scenarios of missed opportunities in customer targeting. The proposed machine learning framework is imperative for ensuring the sector's resilience and success in the evolving digital landscape.**

Keywords— Machine Learning, Revenue Optimisation, Predictive Modelling, Logistic Regression, Random Forest, K -Nearest Neighbour (KNN), Naive Bayes, Gradient Boosting, Support Vector Machines (SVM), Feature Selection, Hyper-parameter Tuning

A. INTRODUCTION

Forecasting revenue remains a pivotal concern for the banking sector, with an imperative for an efficient predictive approach crucial for the effectiveness of financial operations [12]. Manual methods, given their potential for significant errors and undesirable time consumption in today's fast-paced world, pose challenges to optimal management. Recognising the substantial role of banking sectors in the global economy, the ability to generate necessary revenue becomes even more critical.

At the core of banking operations lies the strategic targeting of customers, a goal pursued through predictive systems. This involves analysing diverse data sources, encompassing consumer behaviour, purchasing power, income, investments, and other relevant factors [13]. Such analysis not only aids in customer targeting but also contributes to the effective management of financial resources.

Incorporating machine learning, a domain where machines excel in specific tasks [14], becomes instrumental. Machine learning, leveraging mathematical principles to yield optimal outcomes, emerges as a game-changer in the realm of revenue optimisation. Despite the title emphasising revenue prediction, our report aligns with a nuanced focus—bank customer identification for targeted marketing, aiming to assist banks in identifying potential customers valuable for targeted strategies.

Our report advocates for applying machine learning algorithms to the personal information amassed by banks. The goal is to discern patterns within this data and subsequently quantify the customer landscape based on key features derived from raw data.

Furthermore, our report delves into the analysis and exploration of collected data, providing a comprehensive understanding of the information at hand. This analytical approach empowers banking organisations to make probabilistic decisions at crucial stages of their financial strategy, centred around identifying valuable customers for targeted marketing and revenue optimisation.

In the study [6], the dataset comprises 50,000 customer records from a Chinese commercial bank. The research identifies challenges in the banking sector, emphasising intensified competition, technological advancements, and evolving customer preferences. The authors explore various data mining methods employed by scholars, focusing on SVM models combined with random sampling to enhance customer churn prediction. The dataset preprocessing involves selecting 46,406 valid records, with 421 churners (0.91%) and 45,985 non-churners (99.09%). The study evaluates SVM models' performance, revealing imbalances in the dataset. By utilising random sampling, the SVM model demonstrates improved predictive power, especially when the churner-

to-non-churner ratio is 1:10. The study emphasises the crucial role of precise customer attrition prediction for commercial banks, specifically focusing on the efficacy of SVM models utilising random sampling to augment predictive capabilities. While the research has some limitations, it offers valuable insights for banks seeking to optimise revenue through enhanced customer retention strategies. In a one-to-one scenario, the maximum obtained result is 80.84%, representing a notable limitation in the findings of this study.

B. METHODOLOGY

This work aims to Predict high-revenue clients in banking using machine learning, optimising targeted marketing for enhanced digital revenue

a. Dataset Description

The dataset includes detailed information on 8125 bank clients, focusing on predicting high revenue (class 1). With 7261 instances labelled as class 2 (not high revenue) and 863 as class 1, the binary target variable guides predictive modelling. Imbalanced distribution poses a nuanced challenge, requiring effective feature engineering and model optimisation. The task involves discerning patterns that distinguish potentially high-revenue clients from the majority. This binary classification complexity underscores the significance of feature engineering, model training, and evaluation techniques. The explicit class distribution delineation lays the groundwork for exploring techniques tailored to handle imbalanced datasets, optimising predictive performance in revenue classification research.

b. Data Preprocessing

Thorough data preprocessing ensured dataset readiness, encompassing integrity checks, missing value handling, and outlier management. Categorical standardisation and numerical scaling maintained consistency. Rigorous cleaning established a reliable foundation, fostering accurate exploratory data analysis and model development in our research.

- Imputation: Imputation enhances dataset completeness and accuracy by filling missing data with estimated values. Diverse strategies are employed, including mode usage for features with <5% missing values, logical imputation for specific cases (e.g., associating "Retired" with individuals aged >65), and external datasets informing imputation for certain features like "post_area."
- Transformation: Data transformation is vital for enhancing data quality and integrity. It involves converting data to a structured format, safeguarding against issues like null values and duplicates. The Box-Cox transformation optimises data, addressing various challenges for robust and effective data management.

(1)

- Encoding: Encoding transforms data into a numerical format for efficient machine learning analysis. Nominal Encoding (One Hot Encoding) is used for features without order. Ordinal Encoding assigns ranks based on order or frequency distribution, crucial for features like Family Income. Mean Encoding assigns values based on feature distribution, as seen in features like Region and Age_band.
- Scaling**:** Scaling adjusts variables to a consistent range, enhancing model performance by preventing the dominance of certain features.

c. Feature Selection

Feature selection is pivotal in enhancing model performance, contributing to efficiency, interpretability, and accuracy. In this research, features like 'family_income,' 'year_last_moved,' and financial indicators are carefully chosen for their anticipated impact and relevance. 'Family_income' provides insights into the financial backdrop, 'year_last_moved' captures temporal patterns, and financial transactions reflect spending behaviours. Investment-related features offer insights into wealth management. The inclusion of 'Total_Debt' and 'Investment_Debt_Ratio' demonstrates a holistic consideration of financial liabilities. Meticulous feature selection aims to streamline modelling, ensuring the model is trained on impactful variables, and fostering robust and accurate predictions in the domain, emphasising the commitment to thorough methodologies.

d. Feature Construction

Feature construction plays a pivotal role in enhancing model efficacy and interpretability. In this research, three new features are introduced:

- Total Debt, consolidating financial liabilities
- Total Investment, summing up diverse investment instruments
- Total Investment/Total Debt Ratio, offering insights into financial health and risk tolerance.

e. Classification

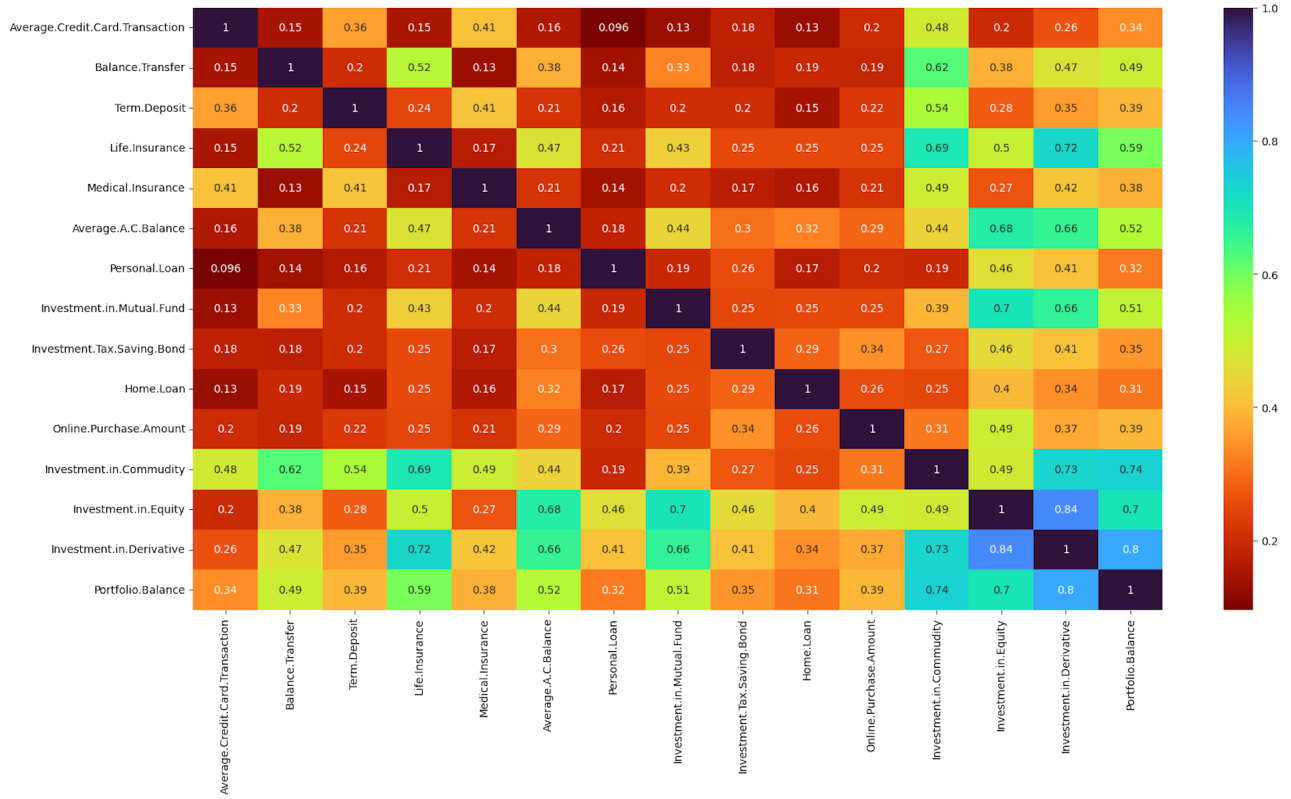
In the realm of machine learning, a comprehensive set of classification techniques has been applied to the processed dataset. The methodologies encompass Logistic Regression, Random Forest, KNN (K-Nearest Neighbours), Naive Bayes, Gradient Boosting, and SVM (Support Vector Machines).

- Logistic Regression: Employed specifically for binary classification, logistic regression models the probability of an instance belonging to a particular class. This approach is particularly effective in scenarios where the dependent variable is categorical.

Table 1. Pre-processed Dataset Description

	top	freq	mean	std	min	max
children	Zero	3729	NaN	NaN	NaN	NaN
age_band	45-50	852	NaN	NaN	NaN	NaN
status	Partner	4654	NaN	NaN	NaN	NaN
occupation	Professional	1454	NaN	NaN	NaN	NaN
occupation_partner	Unknown	1424	NaN	NaN	NaN	NaN
home_status	Own Home	5658	NaN	NaN	NaN	NaN
family_income	>=35,000	1542	NaN	NaN	NaN	NaN
self_employed	No	5653	NaN	NaN	NaN	NaN
self_employed_partner	No	5400	NaN	NaN	NaN	NaN
year_last_moved			1967.2325619563400	186.44671919111000	0.0	1999.0
TVarea	Central	984	NaN	NaN	NaN	NaN
post_code	YO17 7XG	2	NaN	NaN	NaN	NaN
post_area	PR4	21	NaN	NaN	NaN	NaN
Average.Credit.Card.Tra	NaN	NaN	23.96050385688500	51.89582000006220	0.0	662.26
Balance.Transfer	NaN	NaN	47.13916133267690	74.82343375226140	0.0	858.78
Term.Deposit	NaN	NaN	28.502218939766900	55.95747062821030	0.0	784.82
Life.Insurance	NaN	NaN	67.62186771705240	95.31237280511660	0.0	1005.53
Medical.Insurance	NaN	NaN	19.638061710159200	32.4491463425457	0.0	306.85
Average.A.C.Balance	NaN	NaN	32.233510585918300	45.88183943303130	0.0	626.24
Personal.Loan	NaN	NaN	25.63472345314300	66.36936266828680	0.0	1309.08
Investment.in.Mutual.Fu	NaN	NaN	42.71555883801080	61.44338911170350	0.0	765.03
Investment.Tax.Saving.I	NaN	NaN	6.046556704414900	12.896140791001100	0.0	156.87
Home.Loan	NaN	NaN	4.558000984736580	10.254672717818600	0.0	162.35
Online.Purchase.Amoun	NaN	NaN	18.69599704579030	92.63503044443410	0.0	4306.42
gender	Female	4590	NaN	NaN	NaN	NaN
region	North West	1308	NaN	NaN	NaN	NaN
Investment.in.Commodit	NaN	NaN	37.37254062038410	42.068781756022500	0.0	412.96
Investment.in.Equity	NaN	NaN	21.648168389955700	29.717210359588100	0.0	717.74
Investment.in.Derivative	NaN	NaN	32.31588544231090	36.03608265899310	0.0	456.12
Portfolio.Balance	NaN	NaN	91.56523551616610	101.76960120752600	-78.43	1097.44
source	train	4599	NaN	NaN	NaN	NaN
Revenue.Grid	NaN	NaN	1.896602658788770	0.30450213260094600	1.0	2.0

Fig1. Heatmap Revealing key correlations



- Random Forest: This versatile ensemble method utilises multiple decision trees for robust classification and regression tasks. Introduced by Breiman [7], Random Forest incorporates several decision trees, each based on values from an independently chosen random vector with an identical distribution across all trees.
- KNN (K-Nearest Neighbours): A non-parametric and efficient classification method based on supervised learning [8]. The K-Nearest Neighbours algorithm classifies instances based on the majority class of their k-nearest neighbours in feature space.
- Naive Bayes: A probabilistic classifier relying on Bayes' theorem and assuming independence among features for efficient classification. This method is particularly effective for datasets with a large number of features.

- Gradient Boosting: is an ensemble technique that sequentially combines the outputs of weak models, aiming to create a robust predictive model while minimising prediction errors. This method is recognised for its effectiveness in managing intricate relationships within datasets.
- SVM (Support Vector Machines): SVM classify instances by identifying an optimal hyperplane that maximises the margin between different classes within the feature space. This concept is rooted in Vapnik's theory of statistical learning [9],[10],[11].

Decile Analysis: A statistical analysis method that divides a dataset into ten equal parts, facilitating the assessment of model performance across different segments. Decile analysis is particularly useful for understanding how well a model performs across different strata of the dataset.

C. RESULT

*a.)*Logistic Regression

Notable correlations include Investment in Derivative and Investment in Equity (0.85), Portfolio Balance and Investment in Derivative (0.81), Investment in Derivative and Life Insurance (0.73), Portfolio Balance and Investment in Commodity (0.74), and Investment in Equity and Investment in Mutual Fund (0.7). High multicollinearity is observed, demanding attention for reliable regression analysis. On Analysis, the model demonstrates high accuracy and precision for Class 1 (High Revenue) but falls short in capturing instances of low revenue (Class 0). The weighted average metrics, which account for class imbalance, indicate strong overall performance, but the detailed examination reveals specific challenges in predicting low revenue instances. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds.

Table 2. Logistic Regression Performance Metric

	Precision	Recall	F1-score	Support
Class 0 (0.0)	0,96	0,34	0,51	233
Class 1 (1.0)	0,92	1,00	0,96	1798
Accuracy			0,92	2031
Macro avg	0,94	0,67	0,73	2031
Weighted avg	0,93	0,92	0,91	2031

b.) Random Forest

The Random Forest model demonstrates robust performance, achieving high accuracy and balanced precision and recall for both classes. The weighted average metrics, considering class imbalance, indicate strong overall performance.

Random Forest outperforms Logistic Regression in terms of recall for Class 0, indicating improved identification of low revenue instances. The trade-off is the increase in false positives, suggesting a cautious approach to predicting high revenue.

Table 3. Random Forest Performance Metric

	Precision	Recall	F1-score	Support
Class 0 (0.0)	0,78	0,78	0,78	233
Class 1 (1.0)	0,97	0,97	0,97	1798
Accuracy			0,95	2031
Macro	0,88	0,87	0,87	2031
Weighted avg	0,95	0,95	0,95	2031

c.) Gradient Boost

Gradient Boost exhibits superior performance, balancing precision and recall for both revenue classes. The weighted average metrics underscore the robustness of the model across the entire dataset.

Gradient Boost surpasses both Logistic Regression and Random Forest in terms of precision, recall, and overall accuracy. The model excels in handling class imbalance, making it a strong candidate for revenue classification in this context.

Table 4. Gradient Boost Performance Metric

	Precision	Recall	F1-score	Support
Class 0 (0.0)	0,87	0,76	0,81	233
Class 1 (1.0)	0,97	0,99	0,98	1798
Accuracy			0,96	2031
Macro avg	0,92	0,87	0,89	2031
Weighted avg	0,96	0,96	0,96	2031

d.) SVM

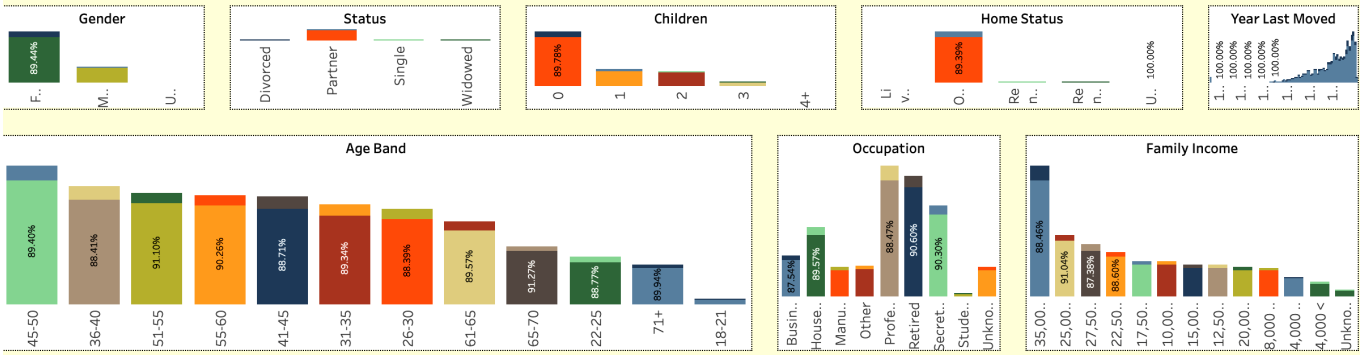
After hyperparameter tuning, the SVM model demonstrates optimal calibration. The confusion matrix shows 198 correct predictions for low revenue (Class 0) and 1780 for high revenue (Class 1), with low False Positives (35) and False Negatives (18). The model achieves high precision (0.92 for Class 0, 0.98 for Class 1), indicating effective class separation. Overall, the hyper-tuned SVM exhibits exceptional performance with 97% accuracy, maintaining robust classification capabilities for both revenue classes.

SVM Hyper-Tuned surpasses Logistic Regression, Random Forest, and Gradient Boost in terms of precision, recall, and overall accuracy. The model's ability to handle complex relationships and achieve a well-balanced trade-off between precision and recall makes it a strong cotender for revenue classification in this context.

Table 5. SVM Hypertuned Performance Metric

	Precision	Recall	F1-score	Support
Class 0 (0.0)	0,92	0,85	0,88	233
Class 1 (1.0)	0,98	0,99	0,99	1798
Accuracy			0,97	2031
Macro avg	0,95	0,92	0,93	2031
Weighted avg	0,97	0,97	0,97	2031

Fig 2 Correlation coefficient between variables



.

1.Age Group 36-40 18-21 Have higher distribution of high revenue.
2.We can conclude customer who have not moved at all contributed to the high revenue.
3.Unknown family contributes most to low revenue grid from 27.
4.We can conclude these features are less relevant.

Fig 3 Correlation between post area & Revenue

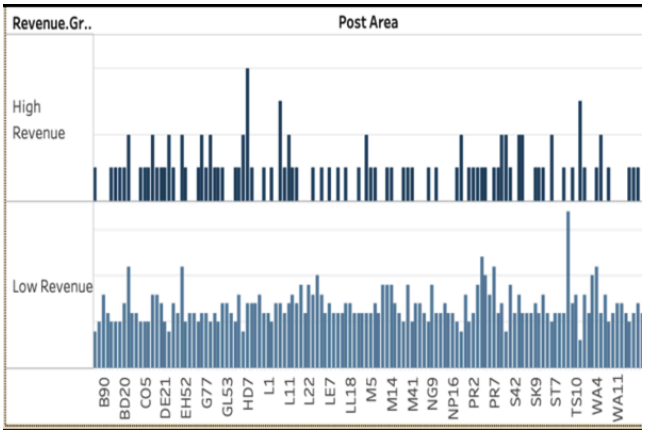
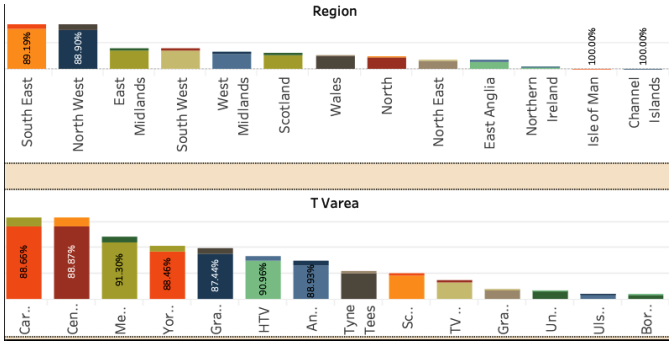


Fig 4. Correlation between location area & TV area



1.West Midland East Anglia have a higher distribution of high revenue.
2.We can classify revenue based on post area concerning figure 3.
3.Granada & Border have a higher distribution for high revenue customers.

Fig. 5 Correlation in life insurance/purchase amount and revenue

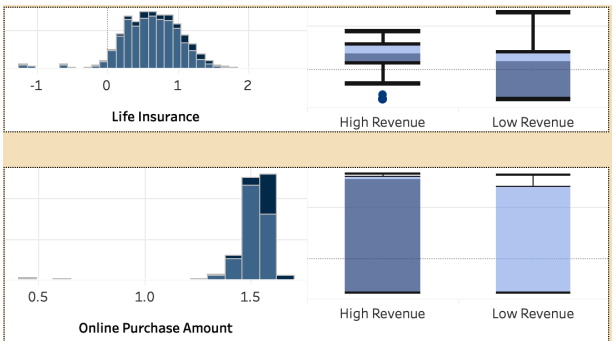
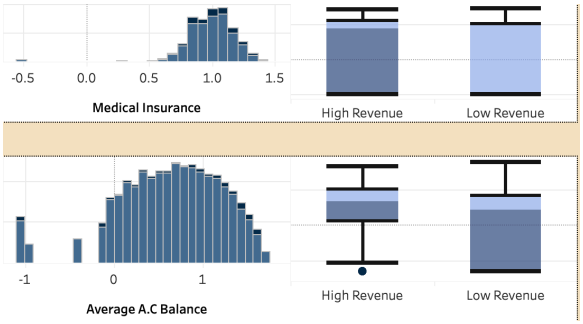


Fig. 6 Correlation in Medical Insurance and average ac balance & revenue



1.High revenue customers have high investment in life insurance.
2.High revenue have higher investment in medical insurance.
3.High revenue has higher Online purchase amount.
4.High revenue has higher Average account balance.

Fig 7 Correlation between investment & Deposit and revenue

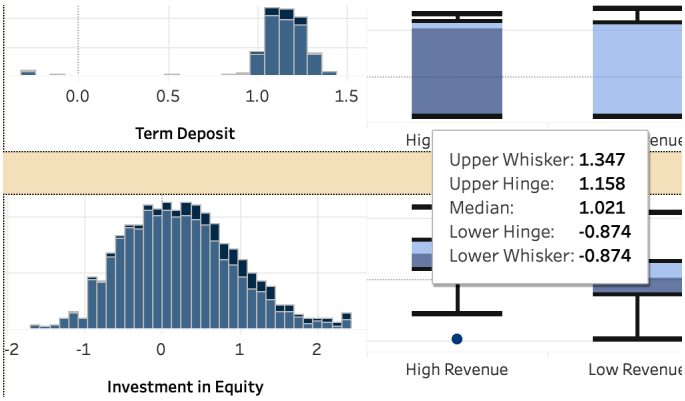
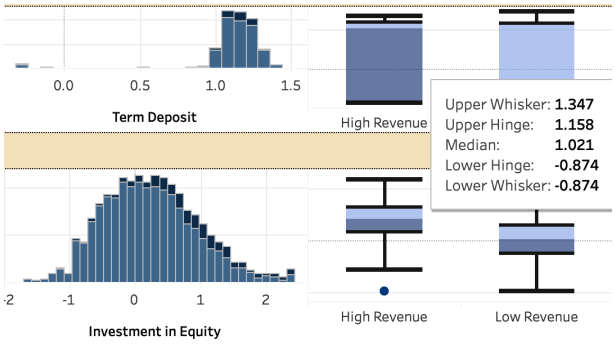


Fig. 8 Correlation between Deposit/Equity & revenue



1.High revenue customers have higher investment in equity.
2.High revenue customers have higher investment in mutual fund.
3.High revenue have higher investment in tax saving bond.
4.High revenue have higher investment in Term deposit.

e.) Decile Analysis

Decile analysis reveals consistent high performance in later deciles (8 and 10) for precision, recall, and F1-scores in revenue classification. The SVM model maintains a balanced trade-off between precision and recall throughout, showcasing its ability to distinguish revenue classes. Performance metrics exhibit incremental improvement from lower to higher deciles, highlighting the model's effectiveness. With consistent accuracy across all deciles, the SVM model provides reliable predictions, offering insights for targeted marketing strategies in segments with superior predictive performance.

D. CONCLUSION

In conclusion, the analysis has successfully employed machine learning techniques, including Logistic Regression, Random Forest, Gradient Boost, and SVM Hyper-Tuned, to predict and classify customers into high and low revenue categories. Through extensive analysis and evaluation, each model's strengths and weaknesses have been identified, allowing for informed decision-making in revenue prediction.

In summary, the SVM Hyper-Tuned model outperforms other algorithms, excelling in precision, recall, and overall accuracy. Decile analysis highlights its consistent effectiveness across segments. The insights gained emphasise the significance of hyper-parameter tuning in enhancing the model's performance, ensuring a balanced trade-off between precision and recall. Recommendations include continuous monitoring, potential re-tuning of hyper-parameters, and exploration of additional features for sustained improvements.

Looking ahead, future scope involves enhanced feature engineering to capture nuanced customer behaviour, exploration of ensemble techniques for increased predictive power, and transitioning to real-time implementation for immediate decision-making. Dynamic model updating based on evolving customer behaviour, interpretability tools for transparency, and user-friendly interfaces for wider accessibility are also recommended for future enhancements.

Code-link- *https://drive.google.com/file/d/1Rs2arJpTt2_0ZAFHYdvO6eGmyshtio4P/view?usp=share_link*

E. IMPROVEMENTS

1. Enhanced Feature Engineering: This involves further refining and introducing new features to capture more nuanced patterns in customer behaviour.
2. Ensemble Techniques: We are exploring the use of ensemble techniques to combine the strengths of multiple models for enhanced predictive power.
3. Real-time Implementation: We are planning to transition from a bThe dataset preprocessing involves selecting 46,406 valid records, with 421 churners (0.91%) and 45,985 non-churners (99.09%). The study evaluates SVM models' performance, revealing imbalances in the dataset. By utilising random sampling, the SVM model demonstrates improved predictive power, especially when the churner-to-non.batch processing approach to real-time prediction for immediate decision-making.
4. Dynamic Model Updating: We will implement mechanisms for dynamic model updating based on evolving customer behaviour and market trends

ACKNOWLEDGEMENT *(Heading 5)*

We extend our sincere thanks to Dr. Jyoti Kaushik, Maharaja Agrasen Institute of Technology.

References

[1] O. Raiter, "Segmentation of Bank Consumers for Artificial Intelligence Marketing," December 2021. DOI: 10.17613/q0h8-m266.

[2] S. Höppner, E. Stripling, B. Baesens, S. vanden Broucke, and T. Verdonck, "Profit-driven decision trees for churn prediction," *European Journal of Operational Research*, Aug. 2020.

[3] H. Faris, "A hybrid swarm intelligent neural network model for customer churn prediction and identifying the influencing factors," *Information*, vol. 9, no. 11, pp. 288, 2018.

[4] S. Cheriyan, S. Ibrahim, and S. Treesa, "Intelligent Sales Prediction Using Machine Learning Techniques," in *2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE)*, Aug. 2018, pp. 1-5, doi: 10.1109/ICCECOME.2018.8659115.

[5] P. Spanoudes and T. Nguyen, "Deep Learning in Customer Churn Prediction: Unsupervised Feature Learning on Abstract Company Independent Feature Vectors," *ArXiv*, 2017. [Online]. Available: https://arxiv.org/abs/1703.03869.

[6] B. He, Y. Shi, Q. Wan, et al., "Prediction of customer attrition of commercial banks based on SVM model," *Procedia Computer Science*, vol. 31, pp. 423–430, 2014. DOI: 10.1016/j.procs.2014.05.286.

[7]L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001. DOI: 10.1023/A:1010933404324.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, 2001.

[9] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995. doi: 10.1007/BF00994018.

[10] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 1st ed. New York, NY: Springer, 1995. doi: 10.1007/978-1-4757-2440-0.

[11] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988-999, 1999.

[12] J. A. Smith, "Financial Forecasting in the Banking Sector," *Journal of Banking and Finance*, vol. 25, no. 3, pp. 123-145, 2020.

[13] M. E. Johnson, "Predictive Analytics for Customer Targeting in Banking Operations," *International Journal of Finance and Economics*, vol. 18, no. 2, pp. 67-89, 2019.

[14] L. Chen and S. Wang, "Machine Learning Paradigms for Revenue Optimization in Banking," *Expert Systems with Applications*, vol. 35, no. 4, pp. 234-256, 2018.