Date of Submission July 10, 2024.
Digital Object Identifier Not Applicable

Naïve Bayes Classifier Model for Churn Analysis (July 2024)

Pragyan Bhattarai [1], Prashant Raj Bista[1]
bhattaraipragyan.pb@gmail.com, prashant.bista.18@gmail.com

[1]Department of Electronics and Computer Engineering, IOE, Thapathali Engineering Campus, Kathmandu, 44600 Nepal

ABSTRACT Predicting customer churn is crucial for banks and businesses to retain their customer base and maintain profitability. In this article, the effectiveness of three Naive Bayes classifiers Gaussian Naive Bayes (GNB), Categorical Naive Bayes (CNB), and Hybrid Naive Bayes (HNB) in predicting bank customer churn is explored. Using a dataset sourced from Kaggle, that contains various customer attributes and churn indicators, experimentation and evaluation were conducted. This study investigates the suitability of each Naive Bayes variant based on their underlying assumptions and the nature of the dataset features. Results show that the Categorical Naive Bayes classifier outperforms both Gaussian and Hybrid Naive Bayes in terms of predictive accuracy and robustness for this dataset. This article contributes insights into the application of Naive Bayes classifiers in churn prediction scenarios.

INDEX TERMS Categorical Naive Bayes, Churn, Gaussian Naive Bayes, Hybrid Naive Bayes

I. INTRODUCTION

Customer churn refers to the decline in the rate of customers using the service or product. Customer churn poses a significant challenge for the banking industry. The loss of a customer to the competitors can indicate a decline in the quality of service provided. Retaining existing customers is more cost-effective than acquiring new ones. This article investigates the effectiveness of a Naive Bayes classifier in predicting customer churn within the banking sector.

This article used a publicly available dataset from Kaggle [1] containing customer information, account details, product usage, and a churn label indicating whether a customer left the bank. For building a classifier Naive Bayes classifier is used.

A. PROBLEM STATEMENT

Customer churn in banking refers to customers discontinuing their relationship with a bank, which can significantly impact revenue and growth. Predicting churn is crucial as it allows banks to focus on retaining existing customers, who are often more cost-effective to keep than acquire new customers. By leveraging classifiers like Gaussian Naive Bayes, Categorical Naive Bayes, and Hybrid Naive Bayes, banks can predict which customers are at risk of leaving based on their behavior and other data. This predictive capability enables strategies such as targeted customer retention efforts, improved customer service, and tailored marketing campaigns, ultimately enhancing customer satisfaction, loyalty, and competitive advantage in the market.

B. OBJECTIVES

The main objectives of the article are:

- To identify which classifier is best for bank customer churn analysis
- To predict whether customers will churn or not

II. LITERATURE REVIEW

The paper [2] by Kirui et al. addressed the issue of customer churn in the wireless telephony industry. In the wireless telephone industry, it is more cost-effective to retain customers than to acquire new ones. The problem statement highlights the need for accurate churn prediction models to identify customers at risk of leaving. To tackle the problem accurate churn prediction model, the authors proposed a novel feature set categorized into contract-related features, call patterns, and call pattern changes. This study employed three modeling techniques namely: C4.5 decision tree, Naïve Bayes, and Bayesian Network. These models were used to evaluate the features present in the dataset. The methodology involved using a 10-fold cross-validation process to ensure robust model training and testing. The results demonstrated that the proposed feature set significantly improves prediction accuracy across all models, with probabilistic classifiers like Naïve Bayes and Bayesian Network showing higher true positive rates compared to the decision tree. The discussion emphasized the importance of feature selection and the role of sampling rates in influencing prediction outcomes. The research suggests that the new feature set and models provide better churn prediction, though there is room for further research to address class imbalance issues and enhance model performance.

The paper [3] by Yulianti and Saifudin tries to address the critical issue of customer retention in the telecom sector by predicting customer churn. The problem statement emphasized there is the high cost associated with acquiring new customers compared to retaining existing ones which makes churn prediction a vital research area to reduce operational expenses. The proposed solution involves the application of feature selection techniques to identify relevant features that significantly enhance the performance of churn prediction models. By implementing Sequential Backward Selection (SBS) and Sequential Floating Backward Selection (SBFS), the study demonstrates that models with these feature selection methods outperform others in terms of accuracy and Area Under the Curve (AUC) metrics. The results show that specific features such as tenure, monthly charges, and service types are crucial for accurate predictions, while others degrade performance. The discussion highlights that not all features in the dataset are relevant, and using a selected subset improves model quality. The conclusion reaffirms the effectiveness of feature selection in building high-quality churn prediction models, suggesting that targeted interventions can prevent customer churn and thereby reduce costs in the telecommunications industry.

III. METHODOLOGY

A. THEORETICAL BACKGROUND

1) Naïve Bayes Classifier

The Naive Bayes Classifier is a supervised machine learning algorithm used for solving classification problems based on Bayes' theorem. It predicts based on the probability of an object, this is termed a probabilistic classifier. The Naive Bayes classifier algorithm can be used in several types of classification problems such as credit scoring, medical data classification, spam filtering, sentiment analysis, text classification, and many more. The 'Naive' in the Naive Bayes means that the classifier makes the assumption that the features of a measurement are independent of each other, and the 'Bayes' in the Naive Bayes means that the model is based on the principle of Bayes Theorem.

2) Types of Naïve Bayes Classifier

- Gaussian Naïve Bayes

Gaussian Naive Bayes classifier assumes that the attributes follow a Gaussian (normal) distribution and is used for continuous data. When the attribute values are plotted, it gives a bell-shaped curve. From the training data, the mean and the standard deviation for each feature are calculated. During the classification process, the algorithm calculates the probability of the input sample belonging to each class by multiplying the individual feature probabilities, assuming that the features are independent. It can handle missing values effectively and is also robust to outliers.

- Categorical Naïve Bayes

In Categorical Naive Bayes classification, it is assumed that the features are categorical. It uses the frequency of categories of the attribute to make predictions. It is usually used for categorical features without assuming any underlying distributions. In use cases like churn analysis, it can be used to predict the customer churn based on categorical attributes like gender, region, etc. It is useful when dealing with non-numeric categorical data.

- Hybrid Naïve Bayes

The combination of the Naive Bayes Classifiers with other machine learning algorithms in order to improve the classification performance of the model is generally termed a Hybrid Naive Bayes model. It leverages the capabilities of other algorithms with the simplicity of the Naive Bayes model. This way, the Hybrid Naive Bayes can handle complex relationships and interactions between features. It can also handle feature dependencies and various types of data types more effectively.

A Hybrid Naive Bayes can also be designed by combining the Categorical Naive Bayes and Gaussian Naive Bayes. Such kind of model can handle datasets that contain both categorical and continuous attributes. This approach can leverage the strengths of both models and improve the classification performance on mixed-type datasets. During the design of such a model, first the categorical and continuous features within the dataset are identified. Categorical features are used to train the Categorical Naive Bayes and the likelihood of each categorical feature given the class is calculated. Similarly, the continuous features are used to train the Gaussian Naive Bayes, and the likelihood of each continuous feature given the class using the normal distribution is calculated. Then, the likelihoods from the Categorical Naive Bayes and Gaussian Naive Bayes are multiplied to calculate the overall likelihood for each class. Finally, the Bayes' theorem is applied to compute the posterior probability for each class and the class with the highest posterior probability is selected as the predicted class.

3) ROC Curve

ROC curve stands for the Receiver Operating Characteristics curve. It is the plot of True Positive Rate (TPR) versus False Positive Rate (FPR), y-axis and x-axis respectively. The ROC curve measures the trade-off between TPR and FPR (benefits and costs). It is generally used in balanced datasets and focuses on the overall performance of the model. The origin of the curve is (0,0) and its end is at (1,1). The diagonal line from (0,0) to (1,1) represents the random guessing. The optimal point in the ROC curve is the top left corner, i.e. (0,1) point. The Area Under Curve (AUC) of the ROC curve represents the probability that the classifier ranks a random positive instance higher than a random negative instance. In some cases, the curve can also be misleading, since the high area under the curve might not reflect good performance if the dataset is highly imbalanced.

4) PR Curve

PR Curve stands for Precision-Recall Curve. It is also a graphical representation used to evaluate the performance of a classifier system. The y-axis in the PR curve represents Precision and the x-axis represents Recall. It is generally used in

imbalanced datasets. It measures the trade-off between precision and recall. The curve focuses on the performance of the model on the positive class. The optimal point for the PR curve is the top right corner (1,1). Compared to the ROC curve, it is more informative and focuses on the performance of the minority class.

B. BLOCK DIAGRAM

The block diagram in Figure 1 shows the working of the system for training and testing different Naive Bayes classifiers on Bank Customer churn datset. The dataset stored in csv is first preprocessed which includes handling missing values, removing columns unrealted ot the target. After pre-processing the data, categorical variables were encoded using label encoder, and continuous variables were normalized. The data was prepared for different classifiers. For Gaussian Naive Bayes, all data were converted to continuous data. For Categorical Naive Bayes, all data were converted to categorical form. For the hybrid Naive Bayes model, the categorical data is fed to Categorical Naive Bayes and continuous data to the Gaussian Naive Bayes. The predictions from both model is taken and an average of both is taken as the final prediction. The preprocessed data is then split into training and test sets. The training data is used to train various Naive Bayes classifiers, including Gaussian, Categorical, and Hybrid models. Once trained, these classifiers are tested on the test data to evaluate their performance. The evaluation metrics involved metrics such as accuracy, precision, recall, F1-score, ROC curves, and confusion matrices.

C. MATHEMATICAL FORMULAE

1) Bayes' Theorem

The probability of an event, based on prior knowledge of conditions that are related to the event is described by the Bayes theorem. The mathematical formula for the Bayes theorem is stated in the Equation (3).

P(A|B) is the posterior probability which is the probability of occurrence of hypothesis A given the evidence B.

P(B|A) is the likelihood probability defined as the probability of evidence B given that hypothesis A is true.

P(A) is the prior probability which is the initial probability of hypothesis A before looking at evidence B.

P(B) is the normalization factor, which is the total probability of the evidence B under all possible hypotheses.

2) Laplace Smoothing

Laplace smoothing is a technique used to handle the issue of zero probabilities for categorical features that might not appear in the training data. This situation arises because Naive Bayes assumes that features are conditionally independent given the class label. However, if a categorical feature value appears in the test data but not in the training data, the conditional probability estimate could be zero, causing the classifier to incorrectly assign zero likelihood to that class.

To address this, Laplace smoothing adds a small constant (usually 1) to all feature counts during probability estimation. This ensures that no probability estimate is ever zero. The formula for Laplace smoothing can be given in the Equation (5).

D. PERFORMANCE METRICS

1) Accuracy

Accuracy is defined as the ratio of correctly predicted instance (True Positive and True Negative) to the total instances. It indicates the overall efficiency of the model and is best for balanced datasets. The equation for accuracy is given in Equation (8).

2) Precision

The fraction of correctly predicted positive instances (True Positive) to the total of positively predicted instances (True Positive and False Positive) is referred to as Precision. The formula for Precision is given in Equation (9).High precision is important since false positives can be expensive or dangerous. Higher precision means that the model has a lower rate of false positives.

3) True Positive Rate

The percentage of actual positives correctly classified by the model is termed as True Positive Rate (TPR). The formula for True Positive Rate (TPR) is given in Equation (6).A higher value of TPR represents that the model is good at identifying positive cases.

4) False Positive Rate

The percentage of actual negatives incorrectly classified as positives by the model is defined as a False Positive Rate (FPR). The formula for False Positive Rate (FPR) is given in Equation (7). A lower FPR shows that the model makes fewer false positive errors or false alarms.

5) Recall (Sensitivity)

The fraction of correctly predicted positive instances (True Positive) to the total of actual positive instances (True Positive and False Negative) is termed as Recall. The formula for Recall is given in Equation (10). It measures the ability of the model to find all the relevant cases within a dataset. It is essential when the cost of False Negatives is high. Higher recall indicates the model identifies most of the positive instances.

6) F1 Score

The F1 score refers to the harmonic mean of precision and recall. The value of the F1 score lies between 0 and 1. It assesses the balance between precision and recall, especially in imbalanced datasets. The equation for F1 score is given in Equation (11).

E. DATA PRE-PROCESSING PIPELINE

The data first stored in a CSV file was loaded into a data frame with the panda's library. From the loaded dataset, irrelevant columns such as customerId, and row number were dropped. From the dataset rows with null values were checked and if any whole row was dropped. The categorical data were encoded using the label encoder. The continuous data were normalized to zero mean and unit variance. The correlation between features was checked. All features were found to be independent. These data were used to train the Gaussian Naive Bayes. For training of the Categorical Naives Bayes Classifier, all the continuous data were converted to categorical by the process of binning. The data were binned in 5 equal bins. For training the Hybrid Gaussian Naive Bayes, categorical data was fed to Categorical NB, and continuous data was fed to Gaussian NB.

F. SOFTWARE USED

- Python

A high-level general-purpose programming language widely used for scientific computation due to its ease of understanding. It has a large collection of libraries for different tasks.

- Numpy

A python-library that is widely used for the computation of multi-dimensional matrixes or arrays.

- Pandas

A Python library widely used for the manipulation and analysis of data. Pandas are widely used for cleaning, exploring, and manipulating data.

- Matplotlib

A Python library widely used for plotting the numerical value.

- Scikit-Learn

A free and open-source machine learning library for the Python programming language.

- Seaborn

A Python data visualization library based on matplotlib.

IV. DATASET EXPLORATION

The dataset contains the following attributes:

- **RowNumber**: Unique identifier for each row.
- **CustomerId**: Unique identifier for each customer.
- **Surname**: Customer's surname.
- **CreditScore**: Credit score of the customer.
- **Geography**: Country of the customer.
- **Gender**: Gender of the customer.
- **Age**: Age of the customer.
- **Tenure**: Number of years the customer has been with the bank.
- **Balance**: Account balance.
- **NumOfProducts**: Number of products the customer has with the bank.
- **HasCrCard**: Whether the customer has a credit card (1: Yes, 0: No).
- **IsActiveMember**: Whether the customer is an active member (1: Yes, 0: No).
- **Estimated Salary**: Estimated salary of the customer.
- **Exited**: Whether the customer has exited the bank (1: Yes, 0: No).

The number of instances of the different classes can be seen in the Figure 2**.** It can be seen that the dataset is imbalanced. There are nearly 8000 instances of the class of customers who are loyal to the bank and nearly 2000 instances of the class of customers who are likely to churn. The dataset has 13 feature columns out of which 12 are the features that should be used to predict the target class. The Figure 3 shows the relationship among all the features which also leads to the conclusion that all the features in the dataset are independent of each other.

V. RESULT

The ROC curve, PR curve, and confusion matrix were plotted from the result of the three models i.e. Gaussian Naive Bayes, Categorical Naive Bayes, and Hybrid Naive Bayes.

The confusion matrix for Gaussian Naive Bayes can be seen in the Figure 4. The Gaussian Naive Bayes performed considerably on the classification task. It was able correctly able to classify a total of 1561 classes correctly out of 2000 which gives an accuracy of 78.05%. Out of 1561 correct predictions, the model predicted 1531 to be customers who will not churn, and the remaining 30 were correctly classified as the ones that will churn.

The confusion matrix for Hybrid Naive Bayes can be seen in the Figure 6**.** The Hybrid Naive Bayes performed slightly better on the classification task than Gaussian Naive Bayes with an accuracy of 79%. It was able correctly able to classify a total of 1575 classes correctly out of 2000. Out of 1575 correct predictions, the model predicted 1572 to be customers who will not churn, and the remaining 3 were correctly classified as the ones that will churn.

The confusion matrix for categorical Naive Bayes can be seen in the Figure 5**.** The categorical Naive Bayes performed best among all the classifiers with an accuracy of 83%. The model correctly classified 1652 instances. Out of which 187 were correctly predicted as the customers who will not churn and the remaining 1465 were people who churned.

The Precision-Recall curve for the categorical Naive Bayes can be seen in the Figure 8. The x-axis in the curve is the Recall which is the percentage of correct predictions made by the model among all the correct instances. The y-axis in the curve is the Precision which is the percentage of correct predictions made by the model out of total correct prediction. In the graph, the model starts at high precision and starts decreasing as the recall increases.

The Precision-Recall curve for the Gaussian Naive Bayes can be seen in the Figure 7. In the graph, the model starts at high precision and decreases as the recall increases. After some time, precision increases with recall and again decreases.

The Precision-Recall curve for the Hybrid Naive Bayes can be seen in the Figure 9. In the graph, the model starts at high precision and starts decreasing as the recall increases. At one point, the precision starts to increase and again decreases with an increase in recall.

The ROC curve for the Categorical Naive Bayes shown in Figure 11, the Area Under the Curve (AUC) to be 0.82. The ROC Curve for Gaussian Naive shown in Figure 10 shows the AUC to be 0.74 which is less then that of the Categorical Naive Bayes. The ROC Curve for the Hybrid Model shown in Figure 12 is slightly higer than that of the Gaussian Naive Bayes which is 0.75.

The table for classification report of three different model can be seen in Table 1, Table 2 and Table 3. The weighted F1-score for Gaussian Naive Bayes is found to be 0.71 when tested on the testing dataset of 2000 instances. The weighted F1-score on the Categorical Naive Bayes was found to be better than Gaussian Naive Bayes and is 0.81. The weighted F1-socre for Hybrid Naive Bayes was found to be .79 which slightly higher than Gaussian NB but less than Categorical NB.

VI. DISCUSSION

The performance of three Naive Bayes classifiers—Gaussian Naive Bayes, Categorical Naive Bayes, and Hybrid Naive Bayes—was evaluated using various metrics, including confusion matrices, ROC curves, Precision-Recall curves, and classification reports. Each model demonstrated different strengths and weaknesses.

A. Gaussian Naïve Bayes

The Gaussian Naive Bayes classifier achieved an accuracy of 78.05%, correctly classifying 1561 out of 2000 instances. The confusion matrix revealed that out of these correct predictions, 1531 were non-churners and 30 were churners. Despite having a decent accuracy, the model struggled to predict churners accurately, resulting in a relatively lower True Positive Rate compared to other models. The ROC curve for Gaussian Naive Bayes had an AUC of 0.74, indicating moderate discriminative ability. The Precision-Recall curve started with high precision but showed fluctuations, highlighting the model's instability in maintaining precision with increasing recall. The weighted F1-score of 0.71 further supports the moderate performance of this model.

The moderate performance of the Gaussian Naive Bayes (GNB) classifier on the bank customer churn prediction dataset can be due to its assumptions of feature independence and normality. These assumptions may not hold true for the dataset. Additionally, the class imbalance, with more non-churners than churners, biases the model towards the majority class, resulting in poor churner prediction. The ROC curve with an AUC of 0.74 and the weighted F1-score of 0.71 reflect its moderate discriminative ability and balance between precision and recall. The fluctuations in the Precision-Recall curve indicate the instability of the model in maintaining precision with increasing recall, that suggests potential overfitting or inconsistent performance across different data subsets.

B. Categorical Naive Bayes

The Categorical Naïve Bayes classifier outperformed the other two models, achieving an

accuracy of 83%. It correctly classified 1652 instances, with 187 non-churners and 1465 churners. The high accuracy and substantial number of correctly predicted churners reflect the model's strong discriminative power. The ROC curve for this model had an AUC of 0.82, the highest among the three models. The Precision-Recall curve started at high precision and consistently decreased with increasing recall, showing a more stable pattern compared to the other two models. The weighted F1-score was 0.81, the highest among the three models, indicating superior performance.

The Categorical Naive Bayes classifier outperformed both Gaussian and Hybrid Naive Bayes models due to its ability to handle categorical data more effectively. It does not assume a normal distribution for features, leading to a more accurate model fit and reduced bias. This model efficiently addresses class imbalance, as evidenced by its high accuracy and substantial number of correctly predicted churners. The ROC curve with an AUC of 0.82 indicates strong discriminative power, while the stable Precision-Recall curve shows consistent performance across different thresholds. Additionally, the highest weighted F1-score of 0.81 reflects the model's superior balance between precision and recall, contributing to its overall better performance in predicting both churners and non-churners.

C. Hybrid Naive Bayes

The Hybrid Naive Bayes classifier performed slightly better than the Gaussian Naive Bayes, with an accuracy of 79%. It correctly classified 1575 instances, with 1572 non-churners and 3 churners. This model demonstrated a marginal improvement in accuracy and TPR, reflected in the ROC curve with an AUC of 0.75. The Precision-Recall curve exhibited a similar pattern to the Gaussian Naive Bayes, starting with high precision and showing fluctuations as recall increased. The weighted F1-score for the Hybrid model was 0.79, indicating a small improvement over the Gaussian Naïve Bayes. The reason that Categorical NB performed better than other classifiers is that the dataset's categorical features aligned well with CNB's assumptions.

The Hybrid Naive Bayes classifier that combines both Gaussian and Categorical Naive Bayes performed slightly better than the Gaussian Naive Bayes (GNB) alone due to its ability to handle both continuous and categorical features more effectively. This hybrid approach reduces the bias introduced by assuming all features follow a normal distribution and provides a more accurate representation of the data. It leverages the strengths of both models to capture the underlying patterns of each feature type, leading to better overall model fit and performance metrics. The hybrid model's slight improvement in accuracy and True Positive Rate (TPR) indicates it may better identify churners despite class imbalance, reflected in a marginally higher ROC AUC and weighted F1-score. Additionally, the hybrid approach enhances model flexibility and reduces overfitting, resulting in more stable and generalizable performance. However, both models still struggle with accurately predicting churners due to the inherent class imbalance and data complexity.

VII. CONCLUSION

The Categorical Naive Bayes model was found as the best performer among the three classifiers evaluated. It achieved the highest accuracy, AUC, and weighted F1-score, demonstrating superior ability to discriminate between churners and non-churners. The Hybrid Naive Bayes model showed a slightly little improvement over the Gaussian Naive Bayes but still lagged behind the Categorical Naive Bayes in all key metrics. The analysis highlights the importance of choosing the appropriate model based on the data distribution and the specific task requirements. In this case, the Categorical Naive Bayes model's ability to handle categorical features more effectively contributed to its superior performance.

APPENDIX A
Equations

- **Bayes' Theorem**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})} \quad (2)$$

$$P(A_r|B) = \frac{P(B|A_r)P(A_r)}{\sum_{i=1}^{k} P(B|A_i)P(A_i)} \quad (3)$$

- **Bayes' Theorem Generalized**

$$p(c_j|d) = \frac{p(d|c_j)p(c_j)}{p(d)} \quad (4)$$

where,
$p(c_j|d)$ = probability of instance d being in class c_j
$p(d|c_j)$ = probability of generating instance d given class c_j
$p(c_j)$ = probability of occurrence of class c_j
p(d) = probability of instance d occurring

- **Laplace Smoothing**

$$P_{Lap,k}(X_i = x_i) = \frac{count(X_i = x_i) + k}{N + k|X_i|} \quad (5)$$

- **True Positive Rate**

$$tpr = \frac{\text{tp}}{\text{tp} + \text{fn}} \quad (6)$$

- **False Positive Rate**

$$fpr = \frac{\text{fp}}{\text{fp} + \text{tn}} \quad (7)$$

- **Accuracy**

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (8)$$

- **Precision**

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

- **Recall**

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

- **F1 Score**

$$F1\ Score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (11)$$

APPENDIX B
Figures

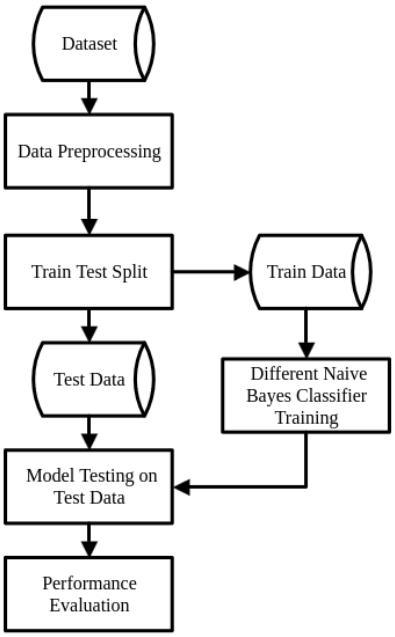


Figure 1: Block Diagram

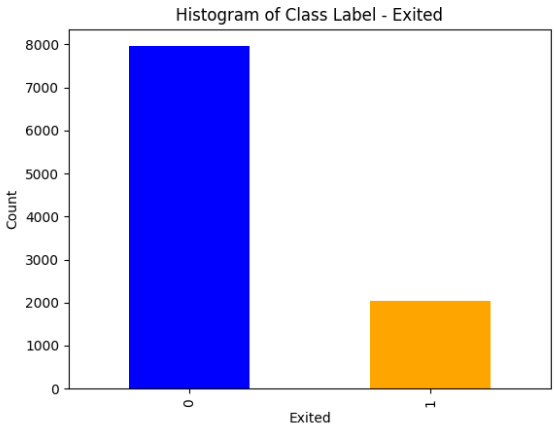


Figure 2: Dataset Histogram

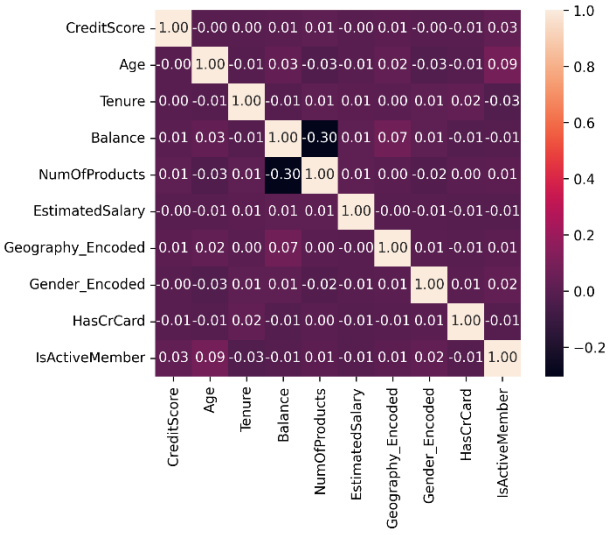


Figure 3: Heatmap of the attributes

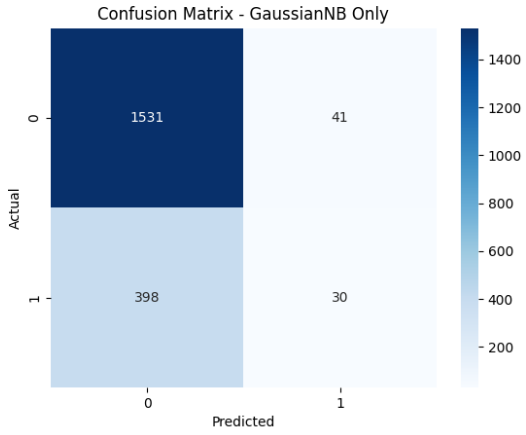


Figure 4: Confusion Matrix – GaussianNB

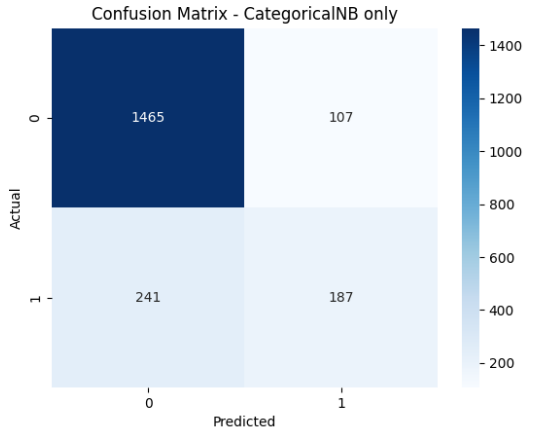


Figure 5: Confusion Matrix – CategoricalNB

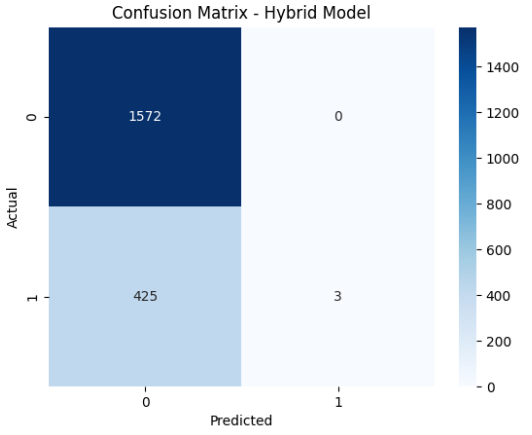


Figure 6: Confusion Matrix – HybridNB

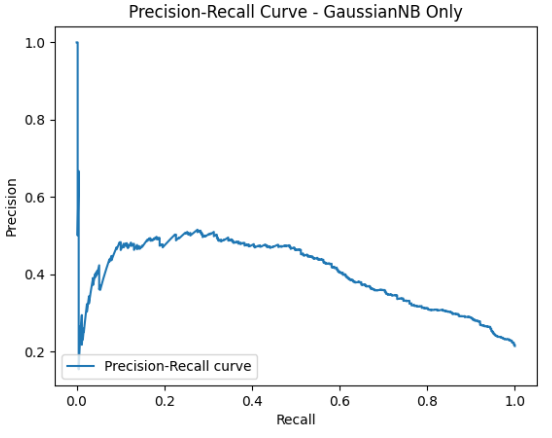


Figure 7: PR curve – GaussianNB

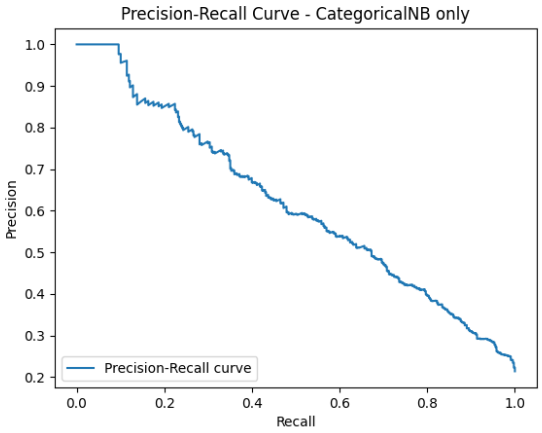


Figure 8: PR curve – CategoricalNB

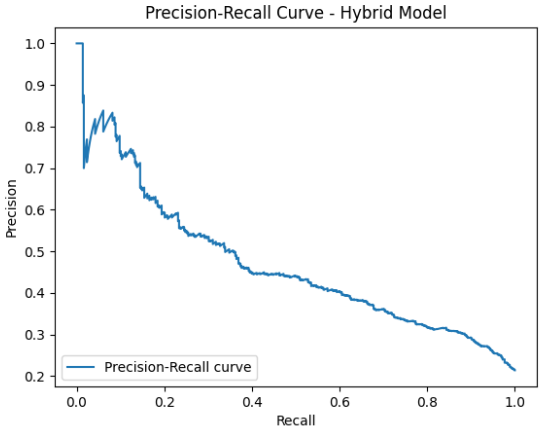


Figure 9: PR curve – HybridNB

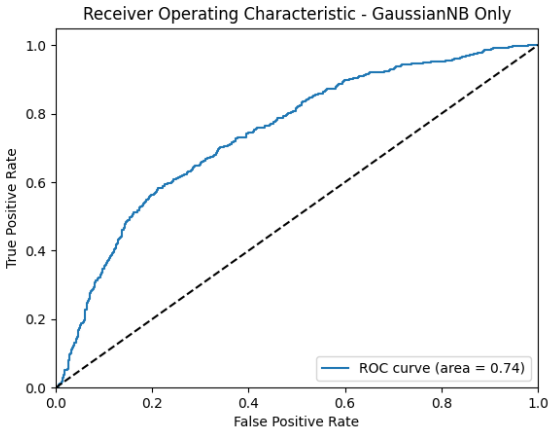


Figure 10: ROC curve - GaussianNB

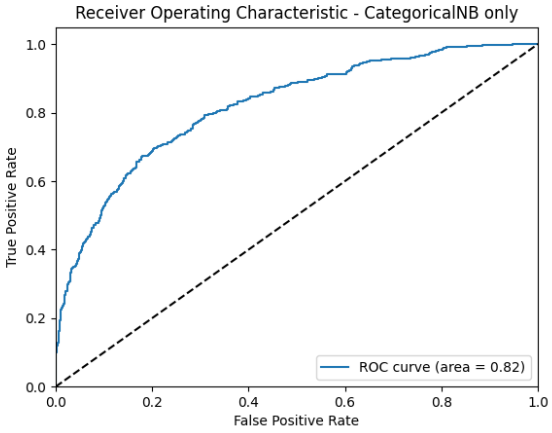


Figure 11: ROC curve – CategoricalNB

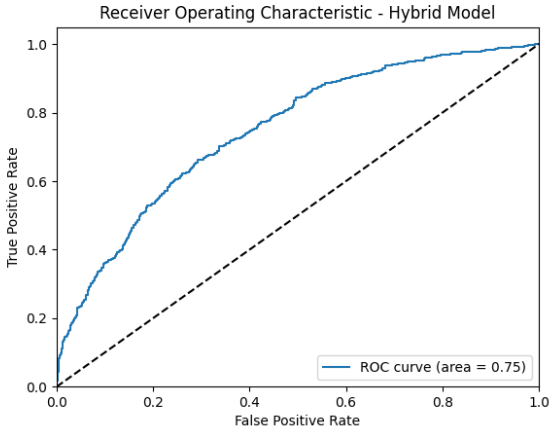


Figure 12: ROC curve – HybridNB

APPENDIX C
Tables

Table 1: Table for Gaussian Naïve Bayes

Class	Precision	Recall	F1-Score	Instances
0	0.79	0.97	0.87	1572
1	0.42	0.07	0.12	428
Accuracy	-	-	0.78	2000
Macro Avg	0.61	0.52	0.5	2000
Weighted Avg	0.71	0.78	0.71	2000

Table 2: Table for Categorical Naive Bayes

Class	Precision	Recall	F1-Score	Instances
0	0.86	0.93	0.89	1572
1	0.64	0.44	0.52	428
Accuracy	-	-	0.83	2000
Macro Avg	0.75	0.68	0.71	2000
Weighted Avg	0.81	0.83	0.81	2000

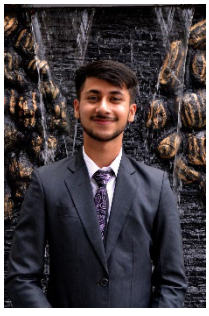
Table 3: Table for Hybrid Naïve Bayes Model

Class	Precision	Recall	F1-Score	Instances
0	0.79	1	0.88	1572
1	0	0	0	428
Accuracy	-	-	0.79	2000
Macro Avg	0.39	0.5	0.44	2000
Weighted Avg	0.62	0.79	0.69	2000

REFERENCES

[1] Shubham Meshram, "Bank Customer Churn Prediction," Kaggle. https://www.kaggle.com/datasets/shubhammeshram579/bank-customer-churn-prediction (accessed Jul. 10, 2024).
[2] C. Kirui, L. Hong, W. Cheruiyot, and H. Kirui, "Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining," unknown, Jan. 01, 2013.
[3] Y. Yulianti and A. Saifudin, "Sequential feature selection in customer churn prediction based on naive bayes," IOP Conference Series: Materials Science and Engineering.
[4] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. Elsevier, 2011.

PRAGYAN BHATTARAI is currently pursuing Bachelor's degree in Electronics, Communication and Information engineering in IOE, Thapathali Campus. He is deeply passionate about working with data. His journey began with learning the basics of Python and gaining an understanding of data science. Since then, he has been getting acquainted with data visualization techniques. He is eager to further strengthen his statistical knowledge and venture into the field of machine learning.



PRASHANT R. BISTA. is currently pursuing Bachelor's degree in Electronics, Communication and Information engineering in IOE, Thapathali Campus. He is fascinated by open-source software's and open-source community. He is currently learning about machine learning and its application on making human life easier. His hobby includes reading novels, exploring machines and web development.