Implementing Machine Learning Algorithms for Loan Approval Prediction

Samarth Kumar
STAT 6000
szk0187@auburn.edu

***Abstract*—Financial institutions evaluate several factors when determining an individual's eligibility for taking out a loan and predicting the approval remains a vital challenge. This project uses the following machine learning algorithms: Naive Bayes, Logistic Regression, and Support Vector Machines. Through a dataset containing approximately 4300 financial records, this project evaluates the influence of various parameters, such as income, employment status, loan amount, and credit score, on the eligibility of a loan for the given individual or organization. All three models achieved accuracy scores of over 90%, proving that machine learning can be used to effectively predict whether or not a loan will be approved across several combinations of parameters.**

***Index Terms*—Finance, Loan Approval Prediction, Machine Learning, Classification, Gaussian, Naive Bayes, Logistic Regression, Support Vector Machine.**

I. Introduction

Diabetes is a chronic disease that impacts millions worldwide. Due to its various causes and risk factors, early prediction and intervention remain critical to preventing severe effects on health despite advances in the field of medicine. Doctors currently predict cases of diabetes by examining a patient's medical history, family history, as well as by administering blood tests. However, traditional methods can often be challenging, thus, machine learning quickly became a popular and transformative tool for disease prediction and management. Among several machine learning algorithms, Naive Bayes is known to be highly efficient in handling large datasets with multiple predictors. The algorithm assumes independence between features, allowing for simpler calculations and rapid processing of complex datasets. Therefore, Naive Bayes would be ideal for training models with medical data.

By employing the Naive Bayes model to predict diabetes using a comprehensive dataset of 100,000 patients, I hope to fill this gap through my research. The dataset includes several demographic and medical parameters, like gender, age, body mass index (BMI), hemoglobin (HbA1c) levels, blood glucose, and smoking history. With these data points, I hope to develop a strong predictive model that is not only accurate, but also provides insights into the importance of the various risk factors that can increase the likelihood of having diabetes.

II. Methodology

A. Model Setup

1) Dataset: The dataset consists of financial information for roughly 4300 individuals. There are eleven features and the target variable indicates the loan eligibility, denoted by 0 for rejection and 1 for acceptance. The predictors include continuous features like annual income, loan amount, and credit score, as well as discrete features like number of dependents, self-employment status, and education status. The dataset consists of both discrete and continuous features, therefore pre-processing the data is essential before training a model.

2) Data Preprocessing: First, we begin with the discrete features, which are number of dependents, self-employed, and education. For consistency, each feature was set to use numerical categories. The number of dependents was the only feature that already had numerical values (0 through 5), so that feature was used as is. For education, the values "Not Graduate", and "Graduate" were replaced with integers 0 and 1 respectively. Similarly, the self-employed values "No" and "Yes" were set to 0 and 1. For the continuous features, the values were standardized using Z-score standardization, which uses the formula below, to ensure that each feature

has a mean (μ) of 0 and a standard deviation (σ) of 1.

$$Z = \frac{x - \mu}{\sigma}$$

B. Computation

1) Naive Bayes: When the Gaussian Naive Bayes algorithm is used, it is assumed that each feature is conditionally independent, which simplifies the calculation of the posterior probabilities. The conditional probability formula for events A, and B, is given below:

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

In the equation, $P(B|A)$ represents the likelihood of event B given the class A, P(A) represents the prior probability of class A while P(B) is the evidence. For continuous features, the probability is modeled by a Gaussian Distribution, with the following formula:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{2}$$

2) Logistic Regression:

3) Support Vector Machine:

III. Analysis

A. Data Description

When evaluating the performance of each algorithm, several metrics were used to provide greater insight. Accuracy describes the rate of correct predictions compared to the total number of predictions made. In the equation below, TP and TN refer to true positive and true negative cases while FP and FN refer to the false positive and false negative cases.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

However, relying on accuracy alone can sometimes be misleading, therefore using additional metrics is more beneficial. Next, precision measures the proportion of true positive cases and all predicted positive cases using the formula below. In other words, precision measures the accuracy of positive predictions.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Third, recall is used to measure the prediction of true positive cases and all cases that are actually positive, thus also accounting for false negative cases.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Lastly, the F1 score combines precision and recall in order to measure the balance between both metrics, accounting for false positive and false negative cases at once. The F1 score, also a weighted average of precision and recall, is represented by the equation below:

$$\text{F1-Score} = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right)$$

The model achieved an overall accuracy of 92% and it demonstrated an exceptional performance for negative cases of diabetes, where the precision, recall, and F1 score were 97%, 94%, and 96%, respectively. However, the model showed a poorer performance for the positive cases, where precision was 52% and F1 score was 61%. This is due to the dataset being imbalanced, where the positive cases accounted for a smaller portion of the samples.

TABLE I
NAIVE BAYES MODEL PERFORMANCE

Class	Precision	Recall	F1-Score	Support
0	0.90	0.94	0.92	318
1	0.96	0.94	0.95	536
Overall Scores				
			0.94	854
Macro Avg	0.93	0.94	0.93	854
Weighted Avg	0.94	0.94	0.94	854

TABLE II
LOGISTIC REGRESSION MODEL PERFORMANCE

Class	Precision	Recall	F1-Score	Support
0	0.88	0.86	0.87	318
1	0.92	0.93	0.92	536
Overall Scores				
			0.91	854
Macro Avg	0.90	0.90	0.90	854
Weighted Avg	0.90	0.91	0.90	854

TABLE III
SUPPORT VECTOR MACHINES MODEL PERFORMANCE

Class	Precision	Recall	F1-Score	Support
0	0.88	0.91	0.89	18300
1	0.94	0.92	0.93	1700
Overall Scores				
			0.92	20000
Macro Avg	0.91	0.91	0.91	20000
Weighted Avg	0.92	0.92	0.92	20000

B. Results

FROM CHATGPT - MUST CHANGE The numerical analysis of the model performance is presented through a confusion matrix, ROC curve, and precision-recall curve. The confusion matrix reveals that the model correctly identified 17,159 instances of Class 0 but misclassified 1,141 as Class 1, indicating a relatively high rate of false positives for this class. Conversely, the model identified 1,256 instances of Class 1 correctly, but misclassified 444 as Class 0, suggesting challenges in recognizing true Class 1 instances effectively.

confusion_matrix.png

Fig. 1. Confusion Matrix

The precision-recall curve shows a decline in precision as recall increases, which is typical for models handling imbalanced datasets. The curve starts with a high precision at low recall levels, reflecting effectiveness in identifying a subset of positive cases accurately. However, as the recall increases, precision drops significantly, indicating a compromise in precision to capture more true positives.

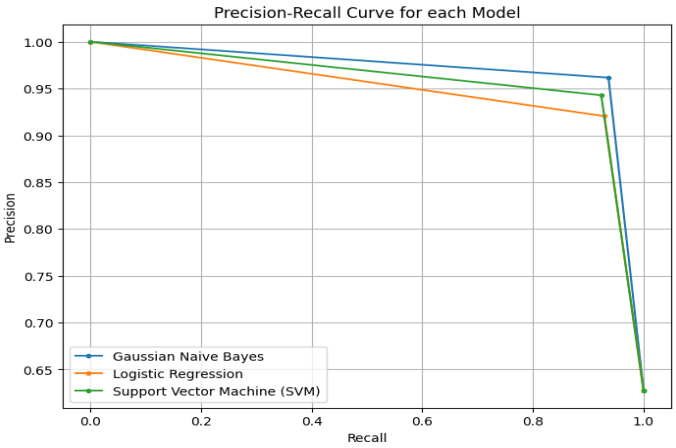

Fig. 2. Precision-Recall Curves

The ROC curve, with an area under the curve (AUC) of 0.84, demonstrates a strong capability of the model to distinguish between the classes across various thresholds. The curve's steep initial ascent and plateau before reaching a false positive rate of 0.2 is indicative of the model's effectiveness at achieving high true positive rates while maintaining a low rate of false positives.

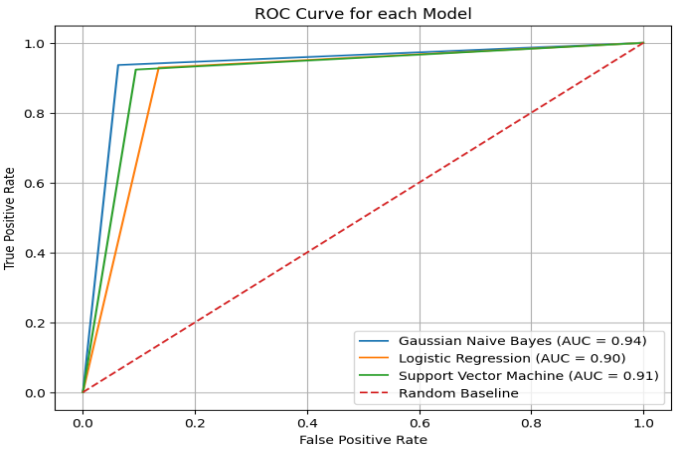

Fig. 3. ROC Curves with the AUC Values

As previously mentioned, the dataset is imbalanced, where 90% of the samples are negative cases

for diabetes. To account for this, the random oversampling technique was applied to the model and different threshold values were tested. By default, the threshold is 0.5. Adjusting the threshold along with random oversampling improved the recall score for the positive class while harming precision, recall, and the model's overall accuracy, thus we opted against the technique altogether to preserve the model's performance.

IV. Discussion

CHATGPT - MUST CHANGE The results discussed in this paper highlight the model's robust performance in distinguishing between the two classes, as evidenced by an ROC AUC of 0.84. However, the precision-recall curve indicates a trade-off between recall and precision, particularly as efforts are made to increase the true positive rate. This trade-off is further evidenced in the confusion matrix, where the model, despite achieving high overall accuracy, struggles with a notable number of false positives and false negatives. The disparity in performance metrics across different classes suggests that the model may benefit from additional tuning, possibly through techniques aimed at addressing class imbalance, such as SMOTE (Synthetic Minority Over-sampling Technique) or adjusted class weighting. Additionally, exploring more complex models or feature engineering strategies might improve the precision without significantly sacrificing recall. In conclusion, while the model performs adequately well in distinguishing between the classes, the observed trade-offs and the challenges in classifying the minority class (Class 1) underscore the need for continued model refinement. Future work should focus on enhancing the model's sensitivity to the minority class without adversely impacting the overall accuracy and precision, ensuring balanced performance across both classes. This balanced approach is crucial, especially in practical applications where the cost of misclassification can be significant.

V. MISC - Parts to Modify

References

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

References

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] GeeksforGeeks, "Handling Imbalanced Data for Classification," *GeeksforGeeks*. Available: https://www.geeksforgeeks.org/handling-imbalanced-data-for-classification/