# MKT 5566 - Predictive Analytics in R

# Group Project Report

Jacob Straub Nithin Songala Rayhan Kabir

## Business Objectives and Problem Statement

The primary objective of this project is to predict whether a business will go bankrupt within the next year based on financial statement data. This prediction is crucial for a local bank that is expanding into business lending. By accurately identifying businesses at high risk of bankruptcy, the bank can avoid originating loans to these entities, thereby minimizing financial risk. Additionally, this predictive capability has broader applications in underwriting risk for insurance, investing, accounting, and evaluating potential mergers and acquisitions. The business problem is to develop a reliable model that can classify businesses as likely or unlikely to go bankrupt, enabling informed decision-making in lending and other financial activities.

## Data Description

The dataset for this project was sourced from Kaggle (<https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction>) and comprises financial data from businesses, featuring over 100 potential predictor variables derived from financial statements. The target variable, “Bankrupt?”, is a binary indicator where 1 denotes bankruptcy.

To prepare the data for modeling, we undertook the following cleaning steps:

* Relabeled column names for clarity by simplifying punctuation, capitalization, and spacing.
* Calculated correlations to mitigate multicollinearity, selecting variables with unique accounting metrics.
* Applied financial and technical expertise to choose 15 numerical variables, ensuring they represent distinct financial ratios and metrics. *(Appendix A.3)*

**Table 1**, detailed in Appendix A.1, presents a comprehensive list of the 15 predictor variables utilized in the analysis. Each variable is accompanied by its meaning, derived from standard financial interpretations and corroborated through online research, along with an explanation of its relevance to bankruptcy prediction. This table serves as a critical reference for understanding the financial metrics driving the bankruptcy prediction models developed in this study.

**Data Summary**

**Table 2**, detailed in Appendix A.1, provides descriptive statistics for the 15 predictor variables, offering insights into their distributions across the dataset.

**Distribution Discussion**

The predictor variables exhibit varied distributions. For example, Debt\_ratio has a mean (0.11318) close to its median (0.11141) but a maximum of 1.00000, indicating right skewness with some highly leveraged companies. Operating\_Expense\_Rate and Interest\_bearing\_debt\_interest\_rate show extreme right skewness, with medians at 0 and means at 1.995e+09 and 16448013, respectively, due to a few companies with exceptionally large values, possibly reflecting unnormalized data or outliers. Conversely, Net\_Value\_Per\_Share is more symmetric, with a mean (0.1906) and median (0.1844) closely aligned. These patterns highlight the diversity in financial conditions, with skewness suggesting the presence of distressed firms, a critical aspect for bankruptcy prediction.

**Correlation Matrix**

The predictor variables exhibit varied distributions. For example, Debt\_ratio has a mean (0.11318) close to its median (0.11141) but a maximum of 1.00000, indicating right skewness with some highly leveraged companies. Operating\_Expense\_Rate and Interest\_bearing\_debt\_interest\_rate show extreme right skewness, with medians at 0 and means at 1.995e+09 and 16448013, respectively, due to a few companies with exceptionally large values, possibly reflecting unnormalized data or outliers. Conversely, Net\_Value\_Per\_Share is more symmetric, with a mean (0.1906) and median (0.1844) closely aligned. These patterns highlight the diversity in financial conditions, with skewness suggesting the presence of distressed firms, a critical aspect for bankruptcy prediction.

*Note for Figure 1 (Appendix A.2): Figure 1 illustrates the correlation matrix for the 15 selected predictor variables, highlighting the pairwise correlations among them to assess potential multicollinearity. This visual representation aids in understanding the relationships between the financial metrics and supports the variable selection process for the bankruptcy prediction models.*

**Target Variable**

The target variable, "Bankrupt," is binary, with descriptive statistics as follows: Min: 0.00000, 1st Qu.: 0.00000, Median: 0.00000, Mean: 0.03226, 3rd Qu.: 0.00000, Max: 1.00000. The mean of 0.03226 indicates that approximately 3.23% of companies in the dataset went bankrupt, reflecting a significant class imbalance. This imbalance may require specialized modeling techniques, such as oversampling or weighted loss functions, to enhance prediction accuracy.

## Model Selection, Analysis, and Results

**Introduction to Model Selection : Logistic Regression**

Logistic regression was **chosen** (initially) for bankruptcy prediction due to its suitability for binary classification tasks and its ability to provide interpretable results. In financial applications like bankruptcy prediction, understanding the relationship between predictor variables and the outcome is essential, making logistic regression an appropriate choice.

**Description of the Models**

Three logistic regression models were developed:

* **Model 1 (logit\_1)**: Utilized all available variables to create a comprehensive baseline model.
* **Model 2 (logit\_2)**: Included 15 selected variables to simplify the model and potentially improve performance by reducing noise.
* **Model 3 (logit\_3)**: Refined to include only 6 statistically significant variables from Model 2, aiming for a more focused and efficient model.

**Evaluation Metrics**

The models were assessed using the following metrics derived from their confusion matrices:

* **Accuracy:** The proportion of total correct predictions.
* **Sensitivity:** The percentage of bankrupt companies correctly identified (True Positive Rate).
* **Specificity:** The percentage of non-bankrupt companies correctly identified (True Negative Rate).
* **Precision:** The proportion of predicted bankruptcies that are correct.

**Analysis of Confusion Matrices**

Below are the confusion matrices and performance metrics for each model, presented in tables similar to the provided sample.

**Table 3: Model 1 Confusion Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
| **Confusion Matrix** | **Predict 0** | **Predict 1** | **Predicted Values** |
| Not Bankrupt (0) | TN: 6411 | FP: 188 | **Specificity**: 97.15% |
| Bankrupt (1) | FN: 96 | TP: 124 | **Sensitivity**: 56.36% |
| Actual Values | Predict 0 acc: 98.52% | Precision: 39.74% | **Accuracy**: 95.84% |

**Analysis**: Model 1 correctly identifies 56.4% of bankrupt companies (sensitivity) and 97.2% of non-bankrupt companies (specificity). However, its precision is relatively low at 39.7%, indicating a higher number of false positives.

**Table 4: Model 2 Confusion Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
| **Confusion Matrix** | **Predict 0** | **Predict 1** | **Predicted Values** |
| Not Bankrupt (0) | TN: 6574 | FP: 25 | **Specificity**: 99.62% |
| Bankrupt (1) | FN: 193 | TP: 27 | **Sensitivity**: 12.27% |
| Actual Values | Predict 0 acc: 97.15% | Precision: 51.92% | **Accuracy**: 96.80% |

**Analysis**: Model 2 achieves the highest accuracy (96.8%) and specificity (99.6%), with very few non-bankrupt companies misclassified. However, its sensitivity is low at 12.3%, meaning it misses most bankrupt companies.

**Table 4: Model 3 Confusion Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
| **Confusion Matrix** | **Predict 0** | **Predict 1** | **Predicted Values** |
| Not Bankrupt (0) | TN: 6572 | FP: 27 | **Specificity**: 99.59% |
| Bankrupt (1) | FN: 198 | TP: 22 | **Sensitivity**: 10.00% |
| Actual Values | Predict 0 acc: 97.08% | Precision: 44.90% | **Accuracy**: 96.70% |

**Analysis**: Model 3 performs similarly to Model 2, with high specificity (99.6%) and accuracy (96.7%), but its sensitivity drops further to 10.0%, indicating even fewer bankrupt companies are correctly identified.

**Results and Discussion**

The three models present a trade-off between sensitivity and specificity:

**Model 1** excels in detecting bankrupt companies (sensitivity: 56.4%), making it the most effective at identifying actual bankruptcies. However, its lower precision (39.7%) means it also flags more non-bankrupt companies as bankrupt, increasing false positives.

**Model 2** offers the highest accuracy (96.8%) and specificity (99.6%), minimizing false positives (precision: 51.9%). Yet, its sensitivity (12.3%) is poor, missing most bankrupt companies.

**Model 3** mirrors Model 2’s high specificity (99.6%) and accuracy (96.7%), but its sensitivity (10.0%) is the lowest, making it the least effective at identifying bankruptcies.

In bankruptcy prediction, failing to identify a bankrupt company (false negative) is often more costly than incorrectly predicting bankruptcy for a non-bankrupt company (false positive). Thus, Model 1 may be the most suitable due to its higher sensitivity, despite its lower precision. However, if the priority is to minimize false positives, Model 2 could be preferred for its balance of high accuracy and precision.

The choice of model depends on the specific goals and tolerance for different types of errors. Future improvements could involve adjusting the classification threshold or exploring alternative models to enhance both sensitivity and specificity.

**Introduction to Naive Bayes Classifier (NBC) Section**

After evaluating the logistic regression model results, which highlighted a persistent trade-off between sensitivity and specificity in predicting bankruptcy, we opted to **explore Naive Bayes Classifiers (NBC)** to assess their performance as an alternative approach. This decision was driven by the need to address the challenges posed by our dataset, particularly the class imbalance where bankruptcies represent a small fraction of the total observations. To tackle this, we implemented an NBC with Laplace smoothing, a technique that reduces the model's sensitivity to class imbalance by handling zero probabilities effectively.

For this analysis, the dataset was split into 70% training and 30% testing sets, ensuring a robust framework for model evaluation. The NBC model yielded promising results, achieving a significant improvement in sensitivity, correctly identifying 50 out of 57 bankruptcies in the test set, which translates to a sensitivity of 89.5%. However, this enhancement came at the expense of precision, which fell to 9.7%, reflecting a higher rate of false positives. These key results underscore the trade-offs inherent in the NBC approach and set the stage for a deeper examination of its performance.

The following sections delve into a detailed analysis of the NBC model's outcomes, with a focus on the confusion matrix, which provides a comprehensive view of its predictive capabilities. This exploration allows us to better understand how NBC compares to the logistic regression model and its suitability for bankruptcy prediction in the context of our data.

**Table 5: NBC Model 1 Confusion Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
| **Confusion Matrix** | **Predict 0** | **Predict 1** | **Predicted Values** |
| Not Bankrupt (0) | TN: 285 | FP: 1700 | **Specificity**: 14.36% |
| Bankrupt (1) | FN: 5 | TP: 56 | **Sensitivity**: 91.80% |
| Actual Values | Predict 0 acc: 98.27% | Precision: 3.19% | **Accuracy**: 16.67% |

**NBC Model 1**:

* **Strength**: High sensitivity (91.80%) and Predict 0 accuracy (98.27%), meaning it correctly identifies most bankrupt companies and non-bankrupt predictions are reliable.
* **Weakness**: Very low specificity (14.36%) and precision (3.19%), indicating many non-bankrupt companies are incorrectly classified as bankrupt, leading to a low overall accuracy (16.67%).

**Table 6: NBC Model 2 Confusion Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
| **Confusion Matrix** | **Predict 0** | **Predict 1** | **Predicted Values** |
| Not Bankrupt (0) | TN: 515 | FP: 1474 | **Specificity**: 25.89% |
| Bankrupt (1) | FN: 3 | TP: 54 | **Sensitivity**: 94.74% |
| Actual Values | Predict 0 acc: 99.42% | Precision: 3.53% | **Accuracy**: 27.81% |

**NBC Model 2**:

* **Strength**: Improved specificity (25.89%) and even higher sensitivity (94.74%), with a strong Predict 0 accuracy (99.42%).
* **Weakness**: Precision remains low (3.53%), and accuracy (27.81%) is still modest, suggesting a high rate of false positives persists.

**Table 7: NBC Model 3 Confusion Matrix Table**

|  |  |  |  |
| --- | --- | --- | --- |
| **Confusion Matrix** | **Predict 0** | **Predict 1** | **Predicted Values** |
| Not Bankrupt (0) | TN: 1407 | FP: 582 | **Specificity**: 70.74% |
| Bankrupt (1) | FN: 7 | TP: 50 | **Sensitivity**: 87.72% |
| Actual Values | Predict 0 acc: 99.50% | Precision: 7.91% | **Accuracy**: 71.21% |

**NBC Model 3**:

* **Strength**: Best balance with high specificity (70.74%), good sensitivity (87.72%), and the highest accuracy (71.21%). Predict 0 accuracy is excellent (99.50%), and precision (7.91%) is the highest among the three.
* **Weakness**: Slightly lower sensitivity compared to the other models, missing a few bankrupt cases, but this is offset by a significant reduction in false positives.

**Conclusion**

Among the three Naive Bayes Classifier models, **NBC Model 3** stands out as the most effective for bankruptcy prediction. It achieves the highest accuracy (71.21%) and a balanced performance between sensitivity (87.72%) and specificity (70.74%), while also improving precision (7.91%). This suggests it is better at minimizing false positives compared to Models 1 and 2, while still detecting most bankrupt companies. If the goal is to prioritize overall accuracy and a reasonable trade-off between identifying bankruptcies and avoiding misclassification of non-bankrupt companies, NBC Model 3 is the preferred choice. However, if maximizing sensitivity (catching every bankrupt company) is critical, NBC Model 2’s 94.74% sensitivity might be considered, though at the cost of lower specificity and precision.

## Managerial Interpretation

The analysis of both logistic regression and Naive Bayes Classifier (NBC) models provides valuable insights for managerial decision-making in bankruptcy prediction, enabling financial institutions, credit analysts, and business strategists to optimize risk management and resource allocation. Below, we interpret the results from both modeling approaches in a managerial context, focusing on their practical implications, trade-offs, and recommended actions.

**Logistic Regression Models**

The three logistic regression models (logit\_1, logit\_2, logit\_3) offer a range of performance profiles that cater to different managerial priorities:

* **Model 1 (logit\_model\_1)**:
  + **Interpretation**: With a sensitivity of 56.36% and an accuracy of 95.84%, this model is effective at identifying over half of the bankrupt companies while maintaining a high overall prediction accuracy. However, the precision of 39.74% indicates that nearly 60% of predicted bankruptcies are false positives, which could lead to unnecessary interventions or costs (e.g., increased scrutiny or loan rejections for solvent firms).
  + **Managerial Implication**: This model is suitable for scenarios where missing a bankrupt company (false negative) is costly—such as in high-stakes lending or investment decisions. Managers should allocate resources to investigate predicted bankruptcies thoroughly, accepting the higher false positive rate as a trade-off for better bankruptcy detection. This could involve targeted audits or credit reviews to filter out false positives.
* **Model 2 (logit\_model\_2)** and **Model 3 (logit\_model\_3)**:
  + **Interpretation**: Both models achieve high accuracy (96.80% and 96.70%, respectively) and specificity (99.62% and 99.59%), minimizing false positives (precision: 51.92% and 44.90%). However, their low sensitivities (12.27% and 10.00%) mean they miss the majority of bankruptcies, with only a small fraction correctly identified.
  + **Managerial Implication**: These models are ideal for conservative risk management strategies, such as maintaining portfolio stability or avoiding overreaction to bankruptcy risks. Managers can use these models to confidently rule out non-bankrupt companies, reducing operational costs associated with unnecessary monitoring. However, they are less effective for proactive bankruptcy prevention, as the high false negative rate could lead to undetected financial distress, potentially resulting in significant losses.
* **Strategic Recommendation**: Given the financial sector’s need to balance risk and cost, **Model 1** is the most actionable for managers. Its ability to detect over half of bankruptcies justifies the effort to manage false positives through additional validation steps. If minimizing false positives is the priority (e.g., in regulatory compliance or customer retention), **Model 2** could be adopted, but with a clear understanding that many bankruptcies will go unnoticed, necessitating supplementary monitoring systems.

**Naive Bayes Classifier Models**

The three NBC models (NBC Model 1, NBC Model 2, NBC Model 3) provide an alternative perspective, leveraging their robustness against class imbalance to enhance bankruptcy detection:

* **NBC Model 1**:
  + **Interpretation**: With a sensitivity of 91.80% and accuracy of 16.67%, this model excels at identifying bankrupt companies (56 out of 61), but its low specificity (14.36%) and precision (3.19%) result in a high number of false positives (1700). This suggests that while it captures nearly all bankruptcies, it also flags many solvent firms incorrectly.
  + **Managerial Implication**: This model is best suited for high-risk environments where detecting every potential bankruptcy is critical, such as in distressed asset management or turnaround planning. Managers should implement a rigorous follow-up process (e.g., financial audits or expert reviews) to sift through the large number of false positives, ensuring resources are not wasted on stable companies.
* **NBC Model 2**:
  + **Interpretation**: This model improves sensitivity to 94.74% and accuracy to 27.81%, with a slight increase in specificity (25.89%) and precision (3.53%). It continues to identify nearly all bankruptcies (54 out of 57) but still generates significant false positives (1474).
  + **Managerial Implication**: Similar to Model 1, this model supports a proactive risk mitigation strategy, ideal for industries with high bankruptcy stakes (e.g., banking or insurance). Managers should pair it with a secondary screening mechanism to reduce false positives, focusing resources on the most vulnerable firms identified.
* **NBC Model 3**:
  + **Interpretation**: Achieving the best balance with an accuracy of 71.21%, sensitivity of 87.72%, specificity of 70.74%, and precision of 7.91%, this model correctly identifies 50 out of 57 bankruptcies while significantly reducing false positives (582). It offers a more reliable trade-off compared to the earlier NBC models.
  + **Managerial Implication**: This model is highly actionable for general risk management across various business units. Managers can confidently use it to prioritize monitoring and intervention for predicted bankruptcies, while its higher specificity minimizes unnecessary actions on non-bankrupt firms. It supports cost-effective decision-making, such as allocating credit limits or initiating early warning systems, with a lower burden of false positives.
* **Strategic Recommendation**: **NBC Model 3** is the most managerially viable option due to its balanced performance. It enables managers to proactively address bankruptcy risks with a sensitivity of 87.72%, while its accuracy (71.21%) and specificity (70.74%) ensure efficient resource use. For critical applications requiring near-perfect bankruptcy detection (e.g., regulatory compliance or high-value portfolios), **NBC Model 2** could be considered, with additional validation steps to manage false positives.

**Comparative Managerial Insights**

* **Logistic vs. NBC**: Logistic regression (Model 1) provides a moderate sensitivity (56.36%) with higher precision (39.74%) compared to NBC models, making it more conservative and interpretable for managers who value clear variable relationships. In contrast, NBC (Model 3) offers superior sensitivity (87.72%) and a broader applicability due to its handling of class imbalance, though with lower precision (7.91%). This suggests NBC is better for detecting rare bankruptcies, while logistic regression suits scenarios requiring fewer false positives.
* **Decision Framework**: Managers should select models based on organizational goals:
  + **Risk Aversion**: Use logistic Model 2 or 3 to minimize false positives and maintain stability.
  + **Proactive Risk Management**: Adopt NBC Model 3 or logistic Model 1 to maximize bankruptcy detection, supplemented by validation processes.
  + **Resource Optimization**: NBC Model 3 strikes the best balance, enabling efficient allocation of monitoring and intervention resources.

**Actionable Steps**

1. **Implement NBC Model 3** as the primary tool, integrating it into early warning systems to flag companies for review.
2. **Develop a Two-Stage Process**: Use NBC Model 3 for initial screening, followed by logistic Model 1 or expert analysis to confirm bankruptcies and reduce false positives.
3. **Monitor Model Performance**: Regularly update models with new data and adjust thresholds to align with evolving business priorities (e.g., increasing sensitivity during economic downturns).
4. **Train Staff**: Equip credit analysts with training on interpreting model outputs, focusing on balancing sensitivity and precision to optimize decision-making.

In conclusion, the choice between logistic regression and NBC models depends on the manager’s tolerance for false positives versus the need to detect bankruptcies. NBC Model 3, with its balanced metrics, offers the most practical solution for effective bankruptcy prediction and resource management in today’s financial landscape.

## Future Work

To enhance the bankruptcy prediction model, we propose the following improvements:

* **Incorporating Additional Variables**: Test more variables from the original dataset to potentially uncover new financial indicators that could improve the model’s ability to predict bankruptcies more accurately.
* **Addressing Class Imbalance**: Explore oversampling methods like SMOTE or adjust class weights during training to better handle the dataset’s imbalance, ensuring the model learns effectively from the minority class of bankruptcies.
* **Experimenting with Advanced Models**: Evaluate other machine learning approaches, such as Random Forest or Gradient Boosting, which are often more effective for imbalanced datasets and can capture complex patterns in the data.
* **Conducting Cost-Benefit Analysis**: Incorporate a cost-benefit analysis to assess the financial impact of false positives and false negatives, allowing for model optimization that aligns with the bank’s specific lending priorities.

These enhancements aim to create a more balanced model that sustains high sensitivity for detecting bankruptcies while also improving precision, ultimately supporting the bank’s lending decisions more effectively.

## APPENDIX

**A.1 Tables**

**Table 1: Selected Variables and Their Relevance**

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Meaning** | **Relevance to Bankruptcy Prediction** |
| Operating\_Gross\_Margin | Gross profit margin: (Revenue - Cost of Goods Sold) / Revenue | Lower margins indicate reduced profitability, increasing bankruptcy risk. |
| Aftertax\_net\_Interest\_Rate | After-tax interest rate on debt, reflecting cost of borrowing post-tax | Higher rates increase debt burden, potentially leading to financial distress. |
| Operating\_Expense\_Rate | Operating expenses / Revenue, measuring operational cost efficiency | Higher rates reduce net profit, elevating bankruptcy likelihood. |
| Cash\_flow\_rate | Operating cash flow / Revenue, indicating cash generation efficiency | Lower cash flow suggests liquidity problems, a key bankruptcy indicator. |
| Interest\_bearing\_debt\_interest\_rate | Interest rate on interest-bearing debt, cost of debt financing | High rates increase interest expenses, straining financial stability. |
| Tax\_rate | Effective tax rate: Taxes Paid / Pre-tax Income | Higher taxes decrease net income, impacting ability to meet obligations. |
| Net\_Value\_Per\_Share | Book value per share: (Total Assets - Total Liabilities) / Shares | Declining values signal weakening equity, a bankruptcy risk factor. |
| Cash\_Reinvestment | Cash flow reinvested / Total Cash Flow, showing reinvestment capacity | Low reinvestment may indicate inability to sustain operations long-term. |
| Debt\_ratio | Total Debt / Total Assets, measuring leverage | Higher ratios reflect greater financial risk and bankruptcy potential. |
| Total\_Asset\_Turnover | Net Sales / Total Assets, indicating asset utilization efficiency | Lower turnover suggests inefficient asset use, linked to financial trouble. |
| Operating\_profit\_per\_person | Operating Profit / Number of Employees, a productivity measure | Lower productivity may indicate operational inefficiencies, raising risk. |
| Cash\_to\_Total\_Assets | Cash and Equivalents / Total Assets, a liquidity metric | Higher liquidity buffers against financial shocks, reducing bankruptcy risk. |
| Total\_income\_to\_Total\_expense | Total Income / Total Expenses, assessing overall profitability | Ratios below 1 indicate losses, strongly predicting bankruptcy. |
| Cash\_Flow\_to\_Sales | Operating Cash Flow / Sales Revenue, cash efficiency per sale | Lower values suggest poor cash conversion, increasing financial vulnerability. |
| Operating.Profit.Rate | Operating Profit Margin: Operating Income / Revenue | Lower margins reduce profitability, heightening bankruptcy risk. |

**Table 2:Descriptive Statistics of Selected Variables**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable Name** | **Min** | **Median** | **Mean** | **Max** |
| **Operating\_Gross\_Margin** | 0.0000 | 0.6060 | 0.6079 | 1.0000 |
| **Aftertax\_net\_Interest\_Rate** | 0.0000 | 0.8094 | 0.8091 | 1.0000 |
| **Operating\_Expense\_Rate** | 0.000e+00 | 0.000e+00 | 1.995e+09 | 9.990e+09 |
| **Cash\_flow\_rate** | 0.0000 | 0.4651 | 0.4674 | 1.0000 |
| **Interest\_bearing\_debt\_interest\_rate** | 0 | 0 | 16448013 | 990000000 |
| **Tax\_rate** | 0.00000 | 0.07349 | 0.11500 | 1.00000 |
| **Net\_Value\_Per\_Share** | 0.0000 | 0.1844 | 0.1906 | 1.0000 |
| **Cash\_Reinvestment** | 0.0000 | 0.3804 | 0.3797 | 1.0000 |
| **Debt\_ratio** | 0.00000 | 0.11141 | 0.11318 | 1.00000 |
| **Total\_Asset\_Turnover** | 0.00000 | 0.11844 | 0.14161 | 1.00000 |
| **Operating\_profit\_per\_person** | 0.0000 | 0.3959 | 0.4007 | 1.0000 |
| **Cash\_to\_Total\_Assets** | 0.00000 | 0.07489 | 0.12409 | 1.00000 |
| **Total\_income\_to\_Total\_expense** | 0.000000 | 0.002336 | 0.002549 | 1.000000 |
| **Cash\_Flow\_to\_Sales** | 0.0000 | 0.6716 | 0.6715 | 1.0000 |
| **Operating.Profit.Rate** | 0.0000 | 0.9990 | 0.9988 | 1.0000 |

**A.2 Plot**

**Fig 1: Correlation Matrix of the selected 15 variable**

A diagram of a graph

AI-generated content may be incorrect.

**A.3 Citations**

In summary, the 15 selected variables were chosen for their alignment with established bankruptcy prediction literature, covering profitability, liquidity, leverage, and efficiency, as supported by Altman (1968), Ohlson (1980), and Ogashi et al. (2020). Their inclusion ensures a comprehensive assessment of financial distress, while exclusions were based on lower relevance, high multicollinearity, or redundancy, validated by the correlation matrix in Figure 1 (Appendix A.2). This approach provides managers with a robust and interpretable model for risk management and decision-making.

**Key Citations**

* Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy [<https://www.jstor.org/stable/2978933>]
* Financial Ratios and the Probabilistic Prediction of Bankruptcy [<https://www.jstor.org/stable/2490395>]
* Corporate Bankruptcy Prediction Model, a Special Focus on Listed Companies in Kenya [<https://www.mdpi.com/1911-8074/13/3/47>]

**A.4 Code**

#Bankrupty Prediction - Group Project

# @author J. Straub and Nithin Songala

#wd

setwd("/users/nithinreddy/R predictive")

# load dataset from external source about online store

dataset <- read.csv("dataset.csv")

nb\_dataset <- read.csv("dataset.csv")

clean\_dataset <- read.csv("Dataset\_clean 1.csv")

nb\_clean\_dataset <- read.csv("Dataset\_clean 1.csv")

#summarize

summary(dataset)

View(dataset)

summary(clean\_dataset)

View(clean\_dataset)

#----------------------------------- Logistic Regression -----------------------------#

#Logit Model -1

logit\_model\_1 <- glm(Bankrupt. ~ .,

family = binomial(link = "logit"), data = dataset)

summary(logit\_model\_1)

dataset$probability\_predicted\_1 <- predict(logit\_model\_1, dataset, type = "response")

dataset$choice\_predicted\_1 <- 0

dataset$choice\_predicted\_1[which(dataset$probability\_predicted\_1 > 0.5)] <- 1

(confusion\_matrix\_1 <- table(dataset$Bankrupt, dataset$choice\_predicted\_1))

(accuracy\_1 <- sum(diag(confusion\_matrix\_1))/sum(confusion\_matrix\_1))

#------------------------------------------------------------------------------------------

#logit model 2: all of the 15 selected variables

logit\_model\_2 <- glm(Bankrupt ~ ., family = binomial(link = "logit"), data = clean\_dataset)

summary(logit\_model\_2)

#marginal/ multiplicative effects

(marginal\_effects\_2 <- coef(logit\_model\_2))

(multi\_effects\_2 <- exp(marginal\_effects\_2))

(multi\_effects\_percent\_2 <- exp(multi\_effects\_2-1)\*100)

#estimate logit model

#except promotion type

cbind(marginal\_effects\_2,multi\_effects\_2,multi\_effects\_percent\_2)

#making predictions

clean\_dataset$probability\_predicted\_2 <- predict(logit\_model\_2, clean\_dataset, type = "response")

clean\_dataset$choice\_predicted\_2 <- 0

clean\_dataset$choice\_predicted\_2[which(clean\_dataset$probability\_predicted\_2 > 0.5)] <- 1

(confusion\_matrix\_2 <- table(clean\_dataset$Bankrupt, clean\_dataset$choice\_predicted\_2))

(accuracy\_2 <- sum(diag(confusion\_matrix\_2))/sum(confusion\_matrix\_2))

#------------------------------------------------------------------------------------------

#Logit model 3 : 6 out of the 15 selected varaibles

logit\_model\_3 <- glm(Bankrupt ~ Tax\_rate+Net\_Value\_Per\_Share+Debt\_ratio+Total\_Asset\_Turnover+Cash\_to\_Total\_Assets+Total\_income\_to\_Total\_expense,

family = binomial(link = "logit"), data = clean\_dataset)

summary(logit\_model\_3)

#marginal/ multiplicative effects

(marginal\_effects\_3 <- coef(logit\_model\_3))

(multi\_effects\_3 <- exp(marginal\_effects\_3))

(multi\_effects\_percent\_3 <- exp(multi\_effects\_3-1)\*100)

#estimate logit model

#except promotion type

cbind(marginal\_effects\_3,multi\_effects\_3,multi\_effects\_percent\_3)

#making predictions

clean\_dataset$probability\_predicted\_3 <- predict(logit\_model\_3, clean\_dataset[ , c("Tax\_rate",

"Net\_Value\_Per\_Share",

"Debt\_ratio",

"Total\_Asset\_Turnover",

"Cash\_to\_Total\_Assets",

"Total\_income\_to\_Total\_expense")],

type = "response")

clean\_dataset$choice\_predicted\_3 <- 0

clean\_dataset$choice\_predicted\_3[which(clean\_dataset$probability\_predicted\_3 > 0.5)] <- 1

(confusion\_matrix\_3 <- table(clean\_dataset$Bankrupt, clean\_dataset$choice\_predicted\_3))

(accuracy\_3 <- sum(diag(confusion\_matrix\_3))/sum(confusion\_matrix\_3))

# ------------------------------- Navie Bayes --------------------------------#

library(naivebayes)

#Convert Bankrupt from int to facotr for NBC

nb\_dataset$Bankrupt <- as.factor(nb\_dataset$Bankrupt)

nb\_clean\_dataset$Bankrupt <- as.factor(nb\_clean\_dataset$Bankrupt)

#Split the dataset into training and validation sample

set.seed(1000)

proportion <- 0.70

indexes\_train <- sample(nrow(nb\_dataset), proportion\*nrow(nb\_dataset))

indexes\_train\_2 <- sample(nrow(nb\_clean\_dataset), proportion\*nrow(nb\_clean\_dataset))

head(indexes\_train,10)

data\_train <- nb\_dataset[indexes\_train, ]

data\_test <- nb\_dataset[-indexes\_train, ]

data\_train\_2 <- nb\_clean\_dataset[indexes\_train\_2, ]

data\_test\_2 <- nb\_clean\_dataset[-indexes\_train\_2, ]

#NBC Model 1: With all the predictors

nbc\_model\_1 <- naive\_bayes( Bankrupt ~ ., laplace = 1, data = data\_train)

#examine a priori probabilities and likelihoods

nbc\_model\_1$prior

nbc\_model\_1$tables$Tax\_rate

nbc\_model\_1$tables$Total\_Asset\_Turnover

# Making predictions using NBC model 1 for validation sample

data\_test$class\_predicted\_1 <- predict(nbc\_model\_1, data\_test)

#get posterior probabilities from nbc\_model\_1

post\_probabilities\_1 <- predict(nbc\_model\_1, data\_test, type = "prob")

data\_test$C1\_post\_prob\_1 <- post\_probabilities\_1[ , "0"]

data\_test$C1\_post\_prob\_1 <- post\_probabilities\_1[ , "1"]

View(data\_test)

#Compute predictive accuracy of NBC\_model\_1

(confusion\_matrix\_1 <- table(data\_test$Bankrupt, data\_test$class\_predicted\_1))

(predictive\_accuracy\_1 <- sum(diag(confusion\_matrix\_1))/sum(confusion\_matrix\_1))

#----------------------------------------------------------------------------------------

#NBC Model 2: With selected 15 predictors

nbc\_model\_2 <- naive\_bayes( Bankrupt ~ .,laplace = 1, data = data\_train\_2)

#examine a priori probabilities and likelihoods

nbc\_model\_2$prior

nbc\_model\_2$tables$Tax\_rate

nbc\_model\_2$tables$Total\_Asset\_Turnover

# Making predictions using NBC model 2 for validation sample

data\_test\_2$class\_predicted\_2 <- predict(nbc\_model\_2, data\_test\_2)

#get posterior probabilities from nbc\_model\_2

post\_probabilities\_2 <- predict(nbc\_model\_2, data\_test\_2, type = "prob")

data\_test\_2$C2\_post\_prob\_2 <- post\_probabilities\_2[ , "0"]

data\_test\_2$C2\_post\_prob\_2 <- post\_probabilities\_2[ , "1"]

View(data\_test\_2)

#Compute predictive accuracy of NBC\_model\_2

(confusion\_matrix\_2 <- table(data\_test\_2$Bankrupt, data\_test\_2$class\_predicted\_2))

(predictive\_accuracy\_2 <- sum(diag(confusion\_matrix\_2))/sum(confusion\_matrix\_2))

#----------------------------------------------------------------------------------------

#NBC 3: with same 6 variables used in logit model

nbc\_model\_3 <- naive\_bayes( Bankrupt ~ Tax\_rate +

Net\_Value\_Per\_Share +

Debt\_ratio +

Total\_Asset\_Turnover +

Cash\_to\_Total\_Assets +

Total\_income\_to\_Total\_expense,

laplace = 1, data = data\_train\_2)

#examine a priori probabilities and likelihoods

nbc\_model\_3$prior

nbc\_model\_3$tables$Tax\_rate

nbc\_model\_3$tables$Total\_Asset\_Turnover

# Making predictions using NBC model 3 for validation sample

data\_test\_2$class\_predicted\_3 <- predict(nbc\_model\_3, data\_test\_2[ ,c("Tax\_rate",

"Net\_Value\_Per\_Share",

"Debt\_ratio",

"Total\_Asset\_Turnover",

"Cash\_to\_Total\_Assets",

"Total\_income\_to\_Total\_expense")])

#get posterior probabilities from nbc\_model\_3

post\_probabilities\_3 <- predict(nbc\_model\_2, data\_test\_2[ ,c("Tax\_rate",

"Net\_Value\_Per\_Share",

"Debt\_ratio",

"Total\_Asset\_Turnover",

"Cash\_to\_Total\_Assets",

"Total\_income\_to\_Total\_expense")], type = "prob")

data\_test\_2$C3\_post\_prob\_3 <- post\_probabilities\_3[ , "0"]

data\_test$C3\_post\_prob\_3 <- post\_probabilities\_3[ , "1"]

View(data\_test)

#Compute predictive accuracy of NBC\_model\_3

(confusion\_matrix\_3 <- table(data\_test\_2$Bankrupt, data\_test\_2$class\_predicted\_3))

(predictive\_accuracy\_3 <- sum(diag(confusion\_matrix\_3))/sum(confusion\_matrix\_3))