Credit Risk Analysis

Project Report - Part A

By: Mustafa Assem Mubarak, MDS- December 2023

Table of Content

1. Data Overview
2. Problem Statement
3. Approach
4. First Model
5. First Model results
6. Best Features
7. Univariate
8. Bivariate
9. Other Models Results
10. Insights & Recommendations

1 - Data Overview

Total Companies	Time Frame	Defaulter	Non-Defaulter
2,058	One Year	220	1,838

Analysis Overview

No. of Models	Best Model	Recall	Precision
Three	Logistic Reg.	68.7%	20.7%

2 - Problem Statement

Summary:

Staying financially healthy is key for businesses. Companies with debt troubles can face lower credit scores, making it harder and more expensive to borrow money in the future. Investors, on the other hand, prefer companies that manage debt well, can grow quickly, and handle that growth. To assess a company's overall health, we use its balance sheet, a financial statement that shows what the company owns, owes, and what investors have put in. This information, from past financial reports, is the foundation of our analysis.

My role in this is to build a model that predicts which companies might have trouble meeting their financial obligations. This model uses the balance sheet data to estimate a company's financial health and the risk of default.

This report explains how we built the model and how it can be used to identify companies that might have financial difficulties

3 - Approach

Approach:

To assess the financial health of companies in our dataset, we compared three different models: logistic regression, linear discrimination analysis (LDA), and random forest. Our primary goal was to identify companies at higher risk of financial difficulties. Therefore, we focused on a metric called "recall," which tells us how well the model finds these high-risk companies. We also considered "precision" to ensure a good balance between identifying high-risk companies and avoiding false positives. We'll dive deeper into each model - logistic regression, LDA, and random forest - and their results in the following slides.

4 - First Model – Best Model

Logistic Regression:

Logistic regression is a common technique used to analyze financial data and predict the likelihood of events, such as a company experiencing financial difficulties in this case.

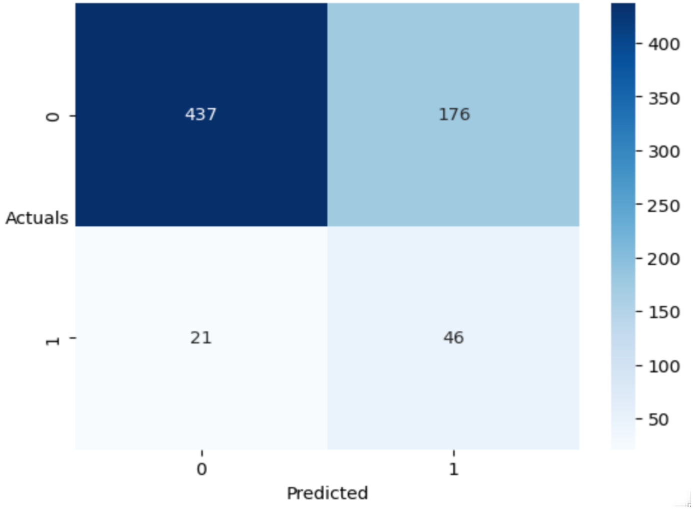
1- Data Preparation

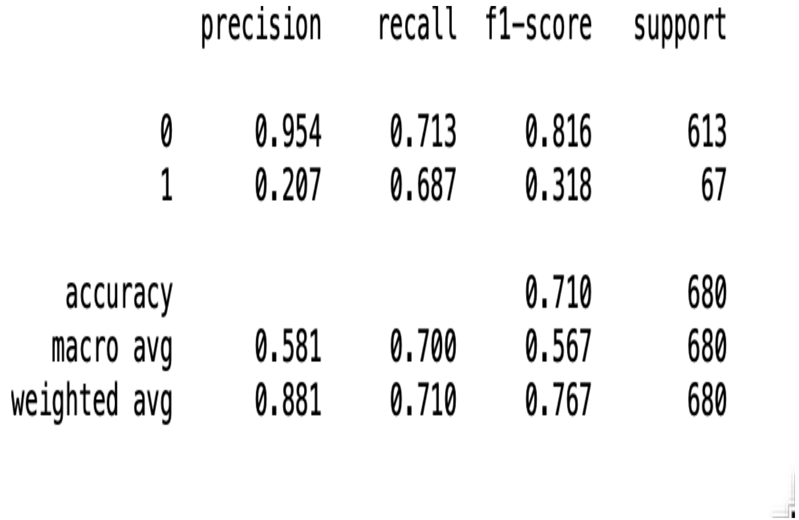
Before applying logistic regression, we had to ensure the data met its assumptions. We addressed missing data by replacing them with medians, as there were outliers. We also dealt with outliers using capping to avoid them influencing the model.

2 -Tackling Multicollinearity

Financial data often exhibits high correlations between features. With 58 features, this presented a challenge. We used the VIF test to identify and remove highly correlated features, resulting in a set of only 5 significant features. We'll focus on interpreting these key features and their impact on financial health in the next slides

5 - First Model - Logit (Best results on Test Data)





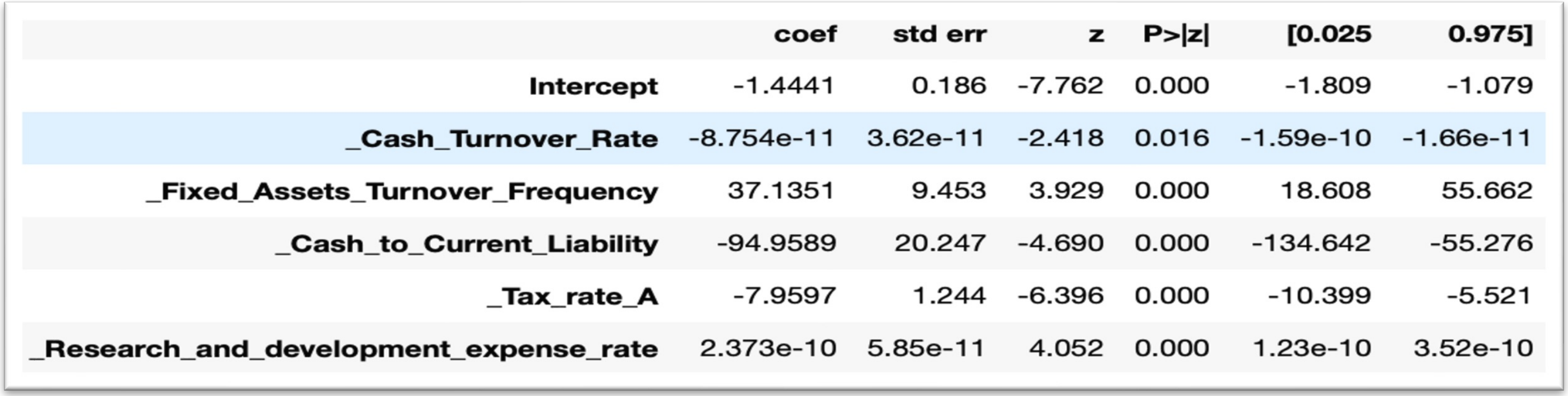
```
              precision    recall  f1-score   support

           0      0.954     0.713     0.816       613
           1      0.207     0.687     0.318        67

    accuracy                          0.710       680
   macro avg      0.581     0.700     0.567       680
weighted avg      0.881     0.710     0.767       680
```

Model Performance

Our Best logistic regression model with optimal threshod of 0.14 (ROC-Curve) demonstrates a trade-off between precision and recall in predicting defaulter companies on the test data.

- Strengths:
  - The model identified a substantial number of true negatives (437 cases) - companies correctly classified as non-defaulters.
  - It achieved a reasonable recall of 68.7%, indicating it captured a good proportion of actual defaulters in the test data.
  - The model's precision, which reflects the proportion of predicted defaulters who were actually defaulters, is lower at 20.7%. This suggests that the model identified a relatively high number of false positives (176 cases) companies predicted as defaulters but not actually defaulting in the test data.

6 - First Model - (Best Features)

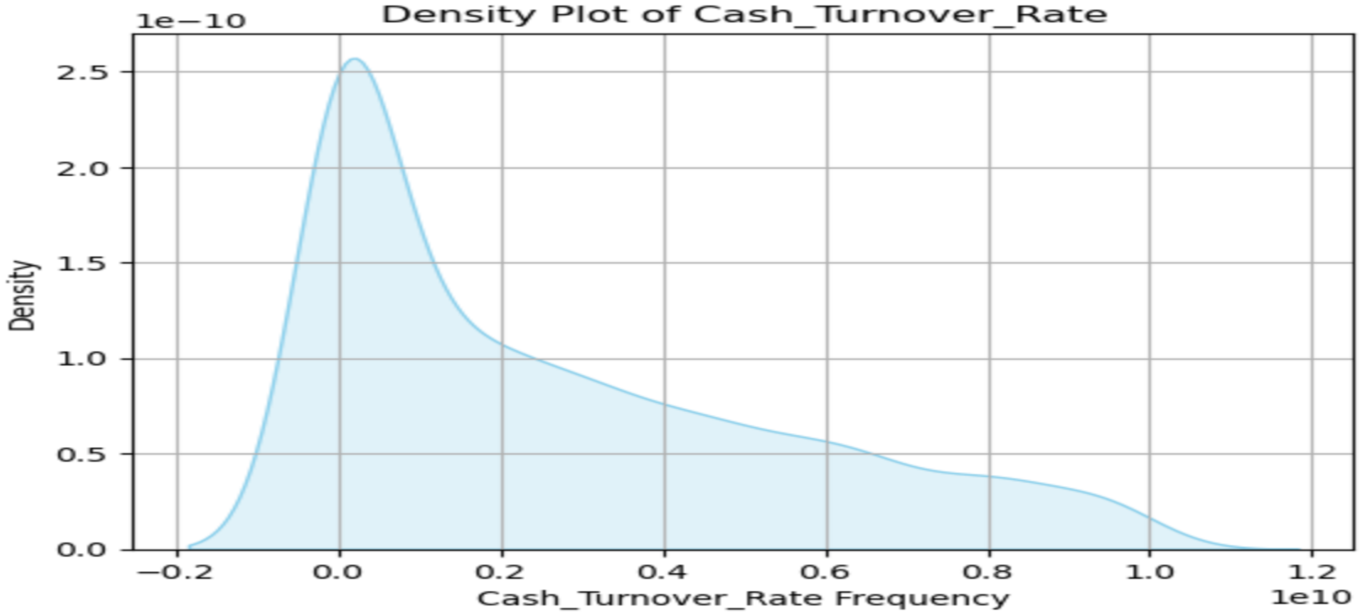


	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1.4441	0.186	-7.762	0.000	-1.809	-1.079
_Cash_Turnover_Rate	-8.754e-11	3.62e-11	-2.418	0.016	-1.59e-10	-1.66e-11
_Fixed_Assets_Turnover_Frequency	37.1351	9.453	3.929	0.000	18.608	55.662
_Cash_to_Current_Liability	-94.9589	20.247	-4.690	0.000	-134.642	-55.276
_Tax_rate_A	-7.9597	1.244	-6.396	0.000	-10.399	-5.521
_Research_and_development_expense_rate	2.373e-10	5.85e-11	4.052	0.000	1.23e-10	3.52e-10

1. Cash_Turnover_Rate (coef = -8.754e-11, p-value = 0.0016): A higher cash turnover rate (which means cash is cycled through the business faster) is associated with a significantly **decreased** likelihood of default.
2. **Fixed_Assets_Turnover_Frequency** (coef = 37.1351, p-value = 0.000): A higher fixed-asset turnover frequency (meaning fixed assets generate more revenue) is associated with an **increased** likelihood of default.
3. **Cash_to_Current_Liability** (coef = -94.9589, p-value = 0.000):A higher ratio of cash to current liabilities suggests a stronger financial position. This is expected, as companies with more cash relative to short-term debts are less likely to default.
4. Tax_rate_A (coef = -7.9597, p-value = 0.000): A higher tax rate is associated with a **decreased** likelihood of default.
5. Research_and_development_expense_rate (coef = 2.373e-10, p-value = 0.000): This feature has a positive and statistically significant coefficient but look **weak positive**.

7 - Best Features Univariate (EDA)

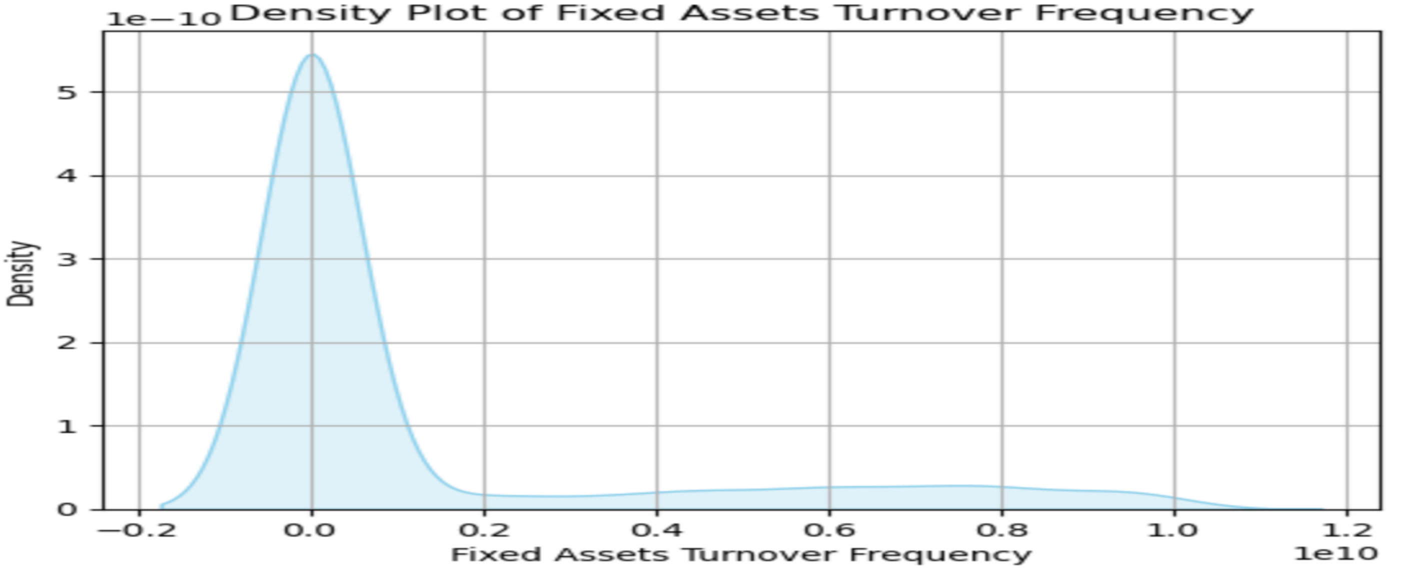
Cash Turnover Rate



Observations / Findings

- The the data tends to be clustered on the lower end, with a few companies having much higher values.

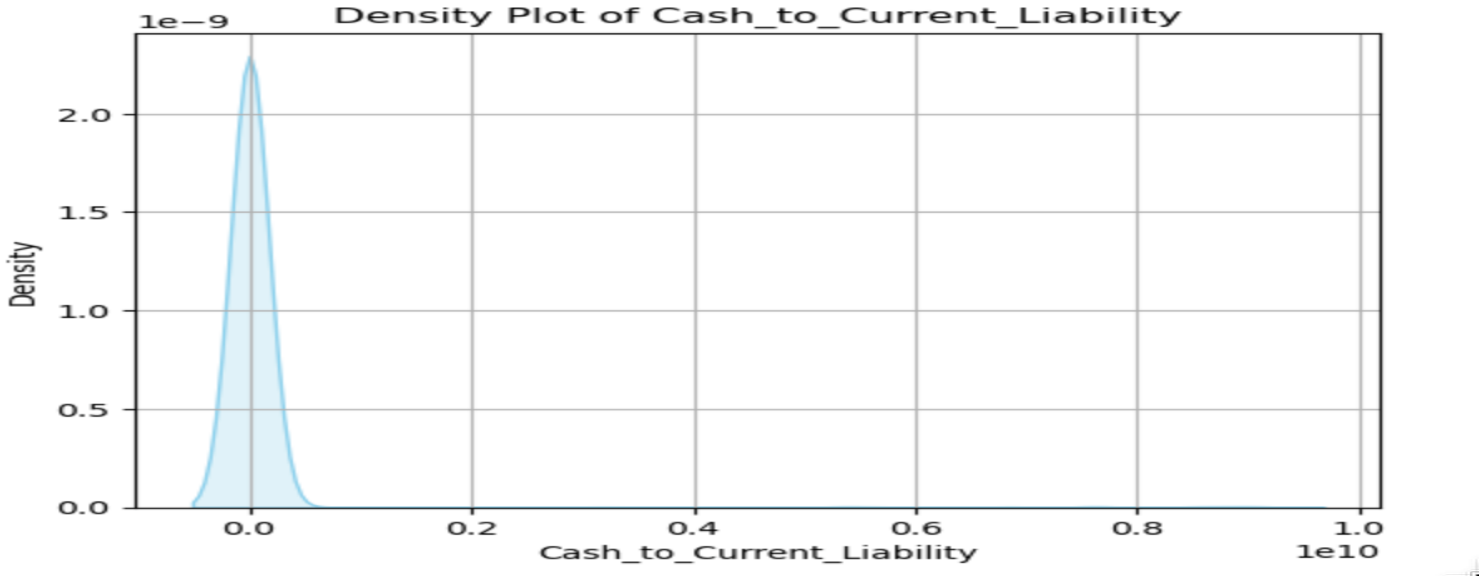
Fixed Assets Turnover Frequency



Observations / Findings

- The data appears to follow a normal distribution overall, there are a few companies with significantly higher values..

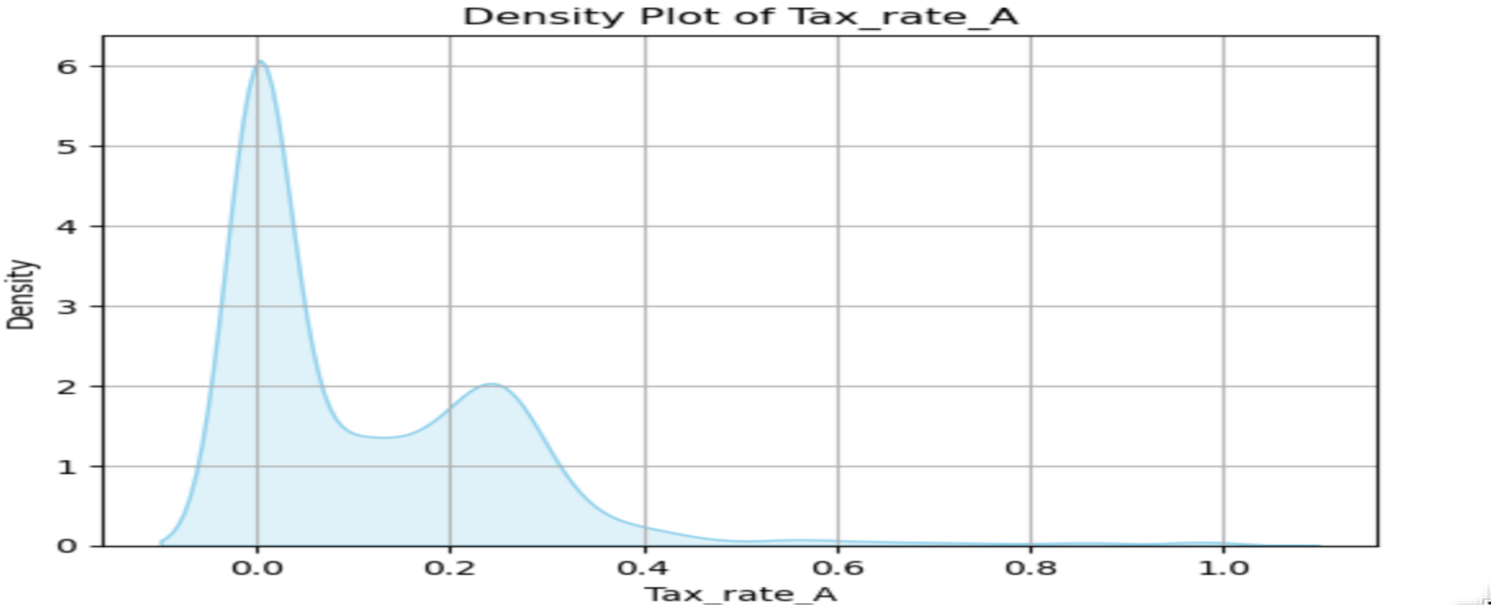
Cash to Current Liability



Observations / Findings

- The data tend to be in perfect normal distribution without tails.

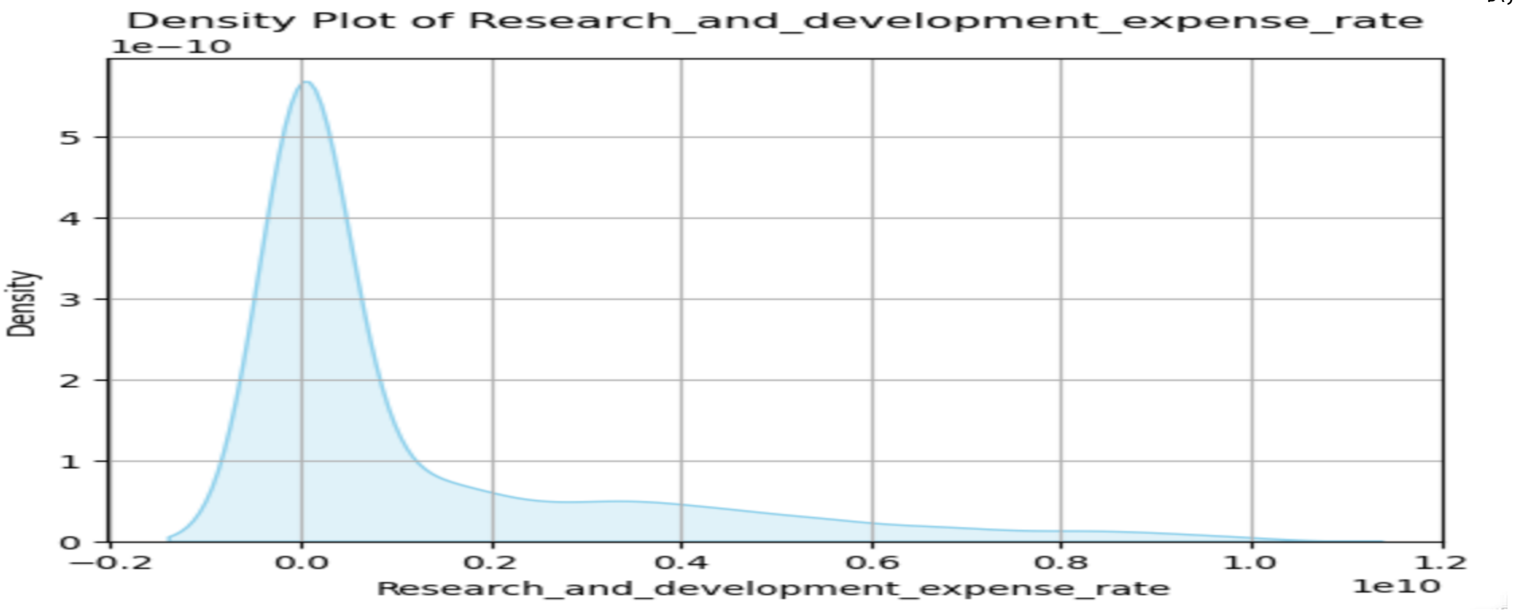
Tax_Rate



Observations / Findings

- The data exhibits a multimodal distribution, with distinct clusters of companies around several specific valuesThere

Research and Development Expense Rate

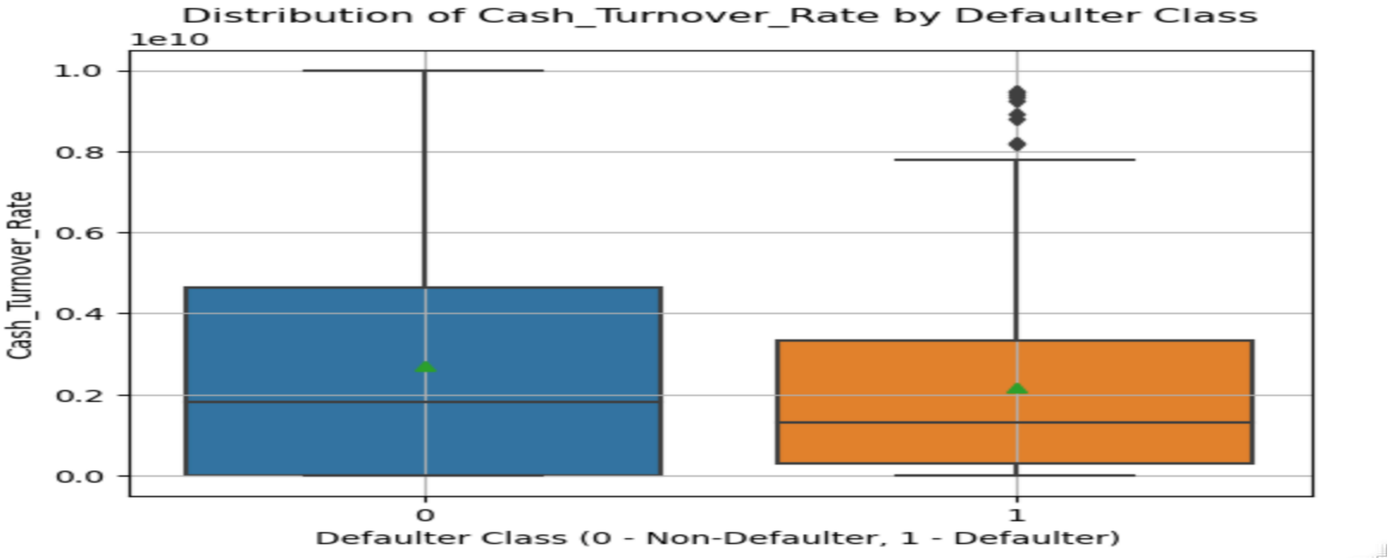


Observations / Findings

- The data appears to follow a normal distribution overall, there are a few companies with significantly higher values..

8 - Best Features Bivariate (EDA)

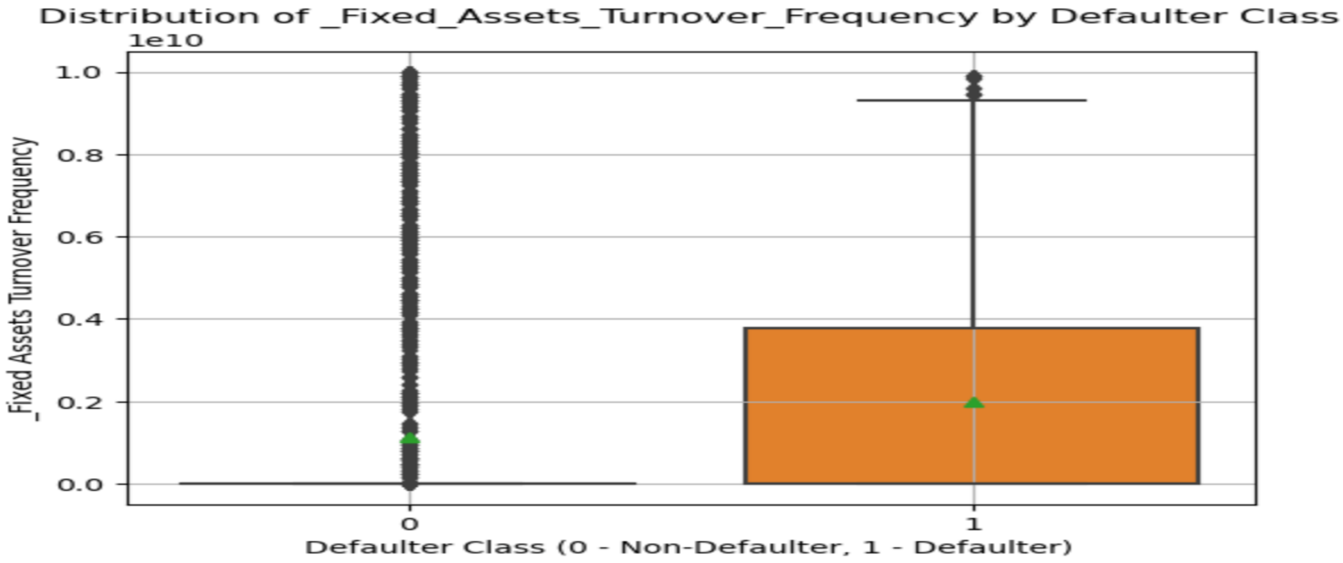
Cash Turnover Rate by Default



Observations / Findings

- The companies classified as non-defaulters tend to have a slightly higher median and mean cash turnover rate compared to defaulters.
- larger number of companies in the non-defaulter category. However, note the presence of outliers within the non-defaulter group.
- These outliers represent companies with a high cash turnover rate despite experiencing defaults..

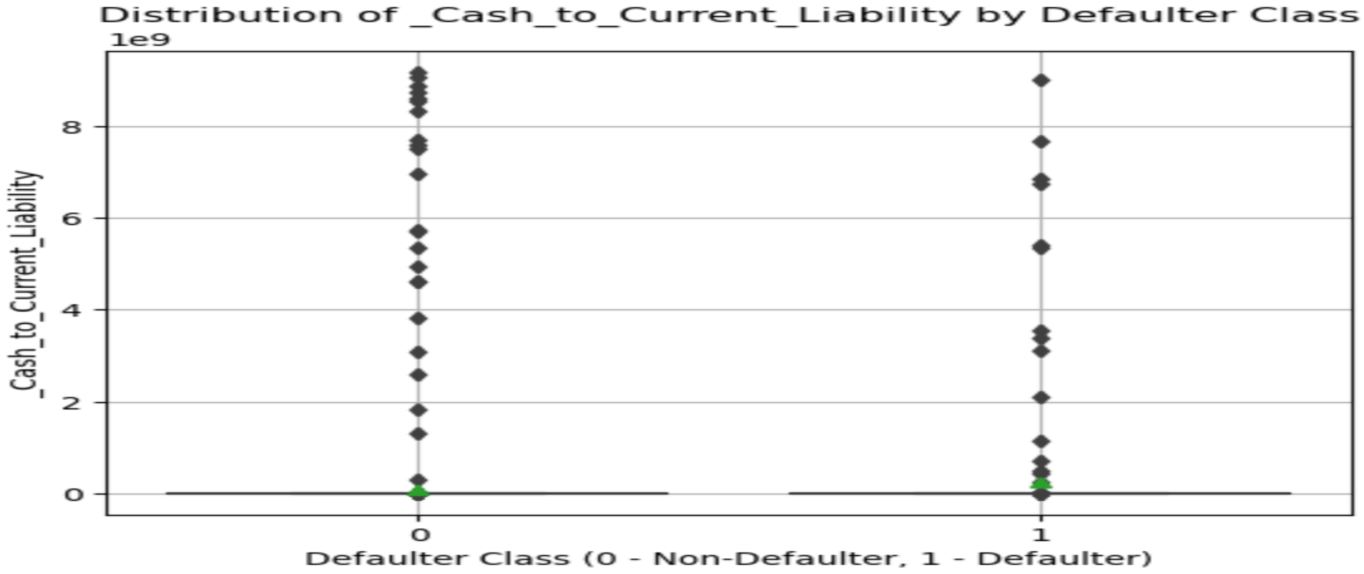
Fixed Assets Turnover Frequency by Default



Observations / Findings

- While the defaulter group is larger, their average fixed asset turnover frequency appears to be higher compared to non-defaulters. This suggests that despite being a larger group, defaulters might exhibit a higher average rate of asset utilization on average. however, there are outliers in both groups. The presence of outliers, particularly a potentially higher number with exceptionally high turnover frequencies in the non-defaulter group, need further investigation.

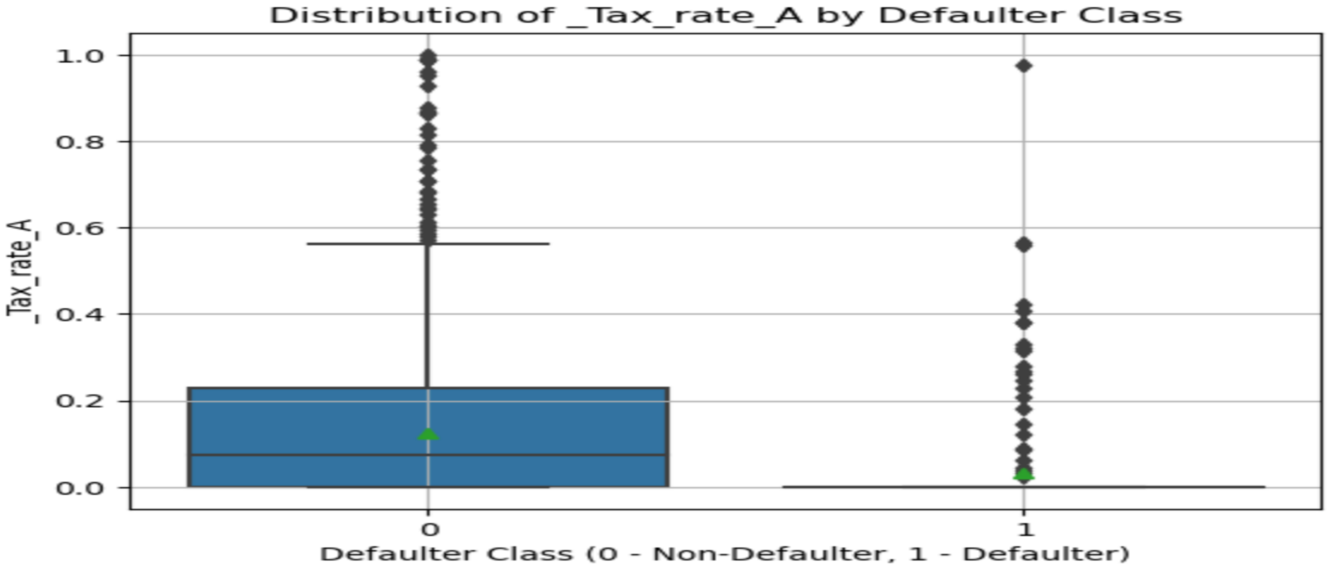
Cash to Current Liability by Default



Observations / Findings

- The defaulter group exhibits a slight higher average ratio compared to non-defaulters, as indicated by the mean value. Which on average, defaulter companies might have a slightly higher level of cash relative to their current liabilities. However, it's important to consider the presence of outliers in both groups. Notably, the non-defaulter group, while having a lower average ratio, might also have a larger number of outliers
- .

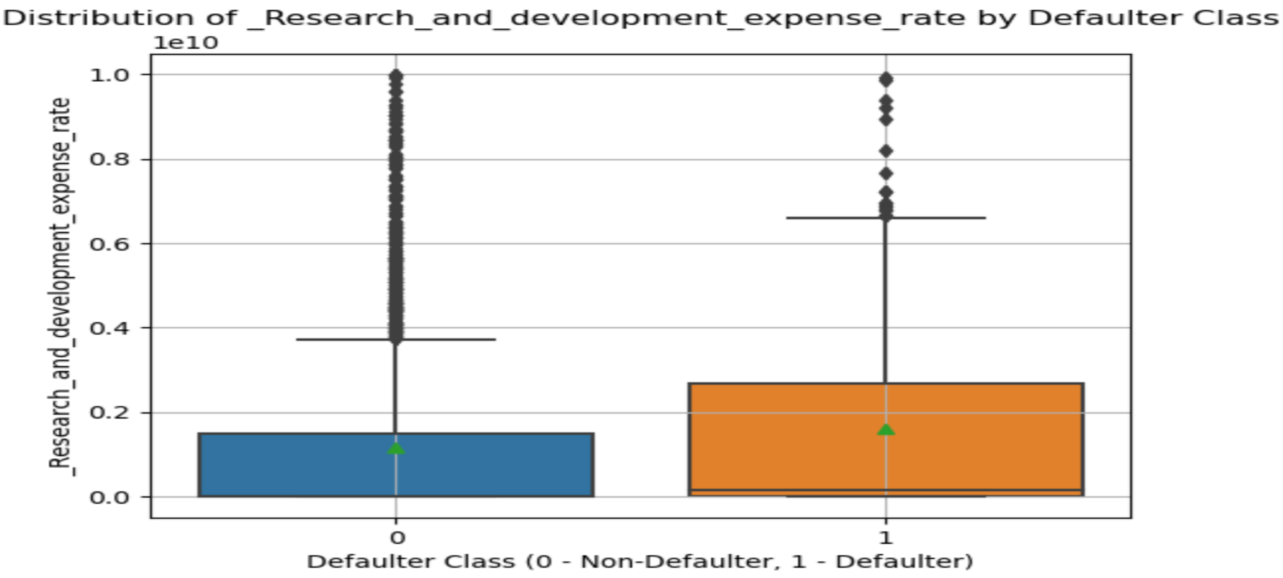
Tax Rate by default



Observations / Findings

- The non-defaulter group appears to have a higher average tax rate as reflected by the mean and median values. This suggests that, on average, non-defaulter companies might be paying a larger portion of their income in taxes. However, note the presence of outliers in both groups. Companies in both defaulter and non-defaulter categories might have outliers with exceptionally high tax rates.

Research and Development Expense Rate by Def.



Observations / Findings

- The defaulter group exhibits a higher average R&D expense rate compared to non-defaulters, as indicated by the mean and median values. which, on average, defaulter companies might be investing a larger portion of their revenue in R&D activities. However, consider the presence of outliers in both groups. The non-defaulter group, despite having a lower average R&D expense rate, might also have a larger number of outliers

9 - Other Models Results
(Not favorable)

Second Model – Not Considered

Linear Discriminant Analysis (LDA):

Linear Discriminant Analysis (LDA) is a supervised machine learning technique commonly used for dimensionality reduction and classification, aims to differentiate between predefined groups by separation between these groups. In our case, it will clearer distinction between defaulters and non-defaulters based on the chosen features.

1- Data Preparation

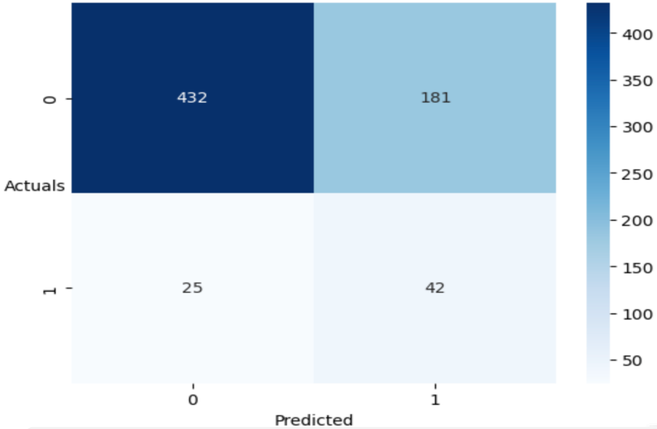
As logistic regression we did data cleaning and removed outliers and missing values. We fit the model to our spitted data, X_Train and Y_Train.

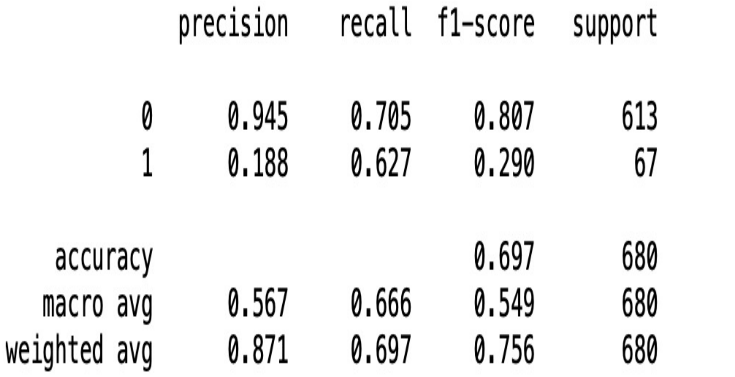
2 - Identifying the Best Threshold

By using ROC technique, which a trade-off between true positive rate (TPR) and false positive rate (FPR) we determined the optimal threshold of 0.12 which the max difference.

3 - **Applying the Model and Threshold to Test Data**

Second Model - LDA (Model results on Test Data)





```
              precision    recall  f1-score   support

           0      0.945     0.705     0.807       613
           1      0.188     0.627     0.290        67

    accuracy                          0.697       680
   macro avg      0.567     0.666     0.549       680
weighted avg      0.871     0.697     0.756       680
```

Our second model, Linear Discriminant Analysis (LDA), exhibits a similar trade-off between precision and recall compared to the initial logistic regression model, but with a less favorable precision /recall rate on the test data.

- ○ The model identified a substantial number of true negatives (432 cases) - companies correctly classified as non-defaulters.
- ○ It achieved a moderate recall of 62.7%, indicating it captured a reasonable proportion of actual defaulters in the test data.
- ○ The model's precision, which reflects the proportion of predicted defaulters who were actually defaulters, is even lower at 18.8% compared to the logistic regression model. This suggests that the model identified a significantly higher number of false positives (181 cases) - companies predicted as defaulters but not actually defaulting in the test data.

- **Weakness**:
  - ○ While the model's recall of 62.7% indicates it captured some defaulters, it's lower than the recall achieved by the logistic regression model. This suggests that LDA might be missing a larger proportion of actual defaulters in the test data compared to the first model.
  - ○ The model's precision, although lower than logistic regression at 18.8%, might not be the primary concern here.

Third Model – Rejected

Random Forest Classifier:

Our third model used a Random Forest classifier, an ensemble learning technique that combines multiple decision trees for improved prediction accuracy and robustness..

1- Data Preparation

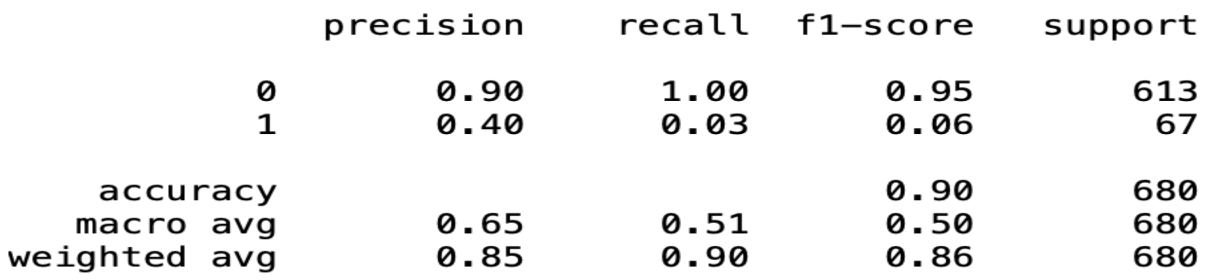
As logistic regression and LDA, we did data cleaning and removed outliers and missing values. We fit the model to our spitted data, X_Train and Y_Train.

2 - Model Training and Grid Search

We utilized a grid search approach to identify the optimal hyperparameter settings for the Random Forest model. This involved training the model with various combinations of hyperparameters defined in the param_grid i.e. (max_depth (7), min_samples_leaf (5), min_samples_split (15), n_estimators(50)).

3- The best performing combination was then selected for the final model.

Third Model - RFC (Model results on Test Data)



```
              precision    recall  f1-score   support

           0       0.90      1.00      0.95       613
           1       0.40      0.03      0.06        67

    accuracy                           0.90       680
   macro avg       0.65      0.51      0.50       680
weighted avg       0.85      0.90      0.86       680
```

Model Performance

Our third model, a Random Forest classifier, exhibited the lowest overall performance compared to the logistic regression and LDA models on the test data.

- **Strengths:**
  - The model achieved a moderate precision of 40% on the test data, indicating that a reasonable proportion of companies predicted as defaulters were actually defaulters.
- **Major Weakness:**
  - The significant drawback of the Random Forest model is its extremely low recall of only 3%. This suggests that the model missed a large number of actual defaulters in the test data. In the context of default prediction, where identifying potential defaulters is crucial, this the reason of rejecting this model.

10 - Insights and Recommendations

Bivariate Analysis and Coefficient Reading:

- **A positive correlation between tax rate and default status**, suggesting companies with higher tax rates might have a higher chance of defaulting.

- **A negative correlation between R&D expense rate and default status**, indicating companies investing more in R&D might have a lower chance of defaulting

- **Efficient Cash Flow Management**: Companies that cycle cash through the business faster (higher cash turnover rate) are less likely to default. This suggests efficient cash flow management as a positive indicator for financial health

- **Asset Utilization:** A high fixed-asset turnover frequency, where fixed assets generate more revenue, might be a double-edged sword. It could indicate efficient asset use, but also potential strain on the assets, requiring further investigation

- **Financial Strength Matters**: Companies with a stronger financial position, as indicated by a higher cash-to-current liability ratio, are less likely to default. Additionally, companies with higher tax rates (potentially reflecting higher profits) also show a lower default risk.

10 (a) - Insights and Recommendations

Model results insights:

- **Logistic Regression** as a Baseline: The initial logistic regression model provided a reasonable baseline for identifying defaulters and non-defaulters. **While not perfect, it offered a balance between precision and recall.**
- **LDA** for Feature Importance: While LDA's direct prediction performance wasn't the strongest, it can be valuable for understanding the underlying data structure. Analyzing the features that contribute most to the separation of defaulters and non-defaulters in LDA's lower-dimensional space can provide insights for improving other models.
- **Random Forest** Potential for Non-Defaulter Identification: Random Forest, despite its low recall for defaulters, achieved a relatively high precision. This suggests it might be a valuable tool for scenarios where accurately identifying non-defaulters is the primary objective.

— END —