Finance and Risk Analytics Business Report

Credit Risk Analysis

Ayush Sharma

Contents

Topic	Page Numbers
Data Summary	3
Outlier Treatment	4 - 5
Missing Values Treatment	6 - 7
Correlation Plot	8
Train-Test Split	9
Logistic Regression Model	10 - 14
Random Forest Model	15 - 16
Linear Discriminant Analysis Model	17
Model Performances Comparison	18
Conclusions and Recommendations	19

Data Summary

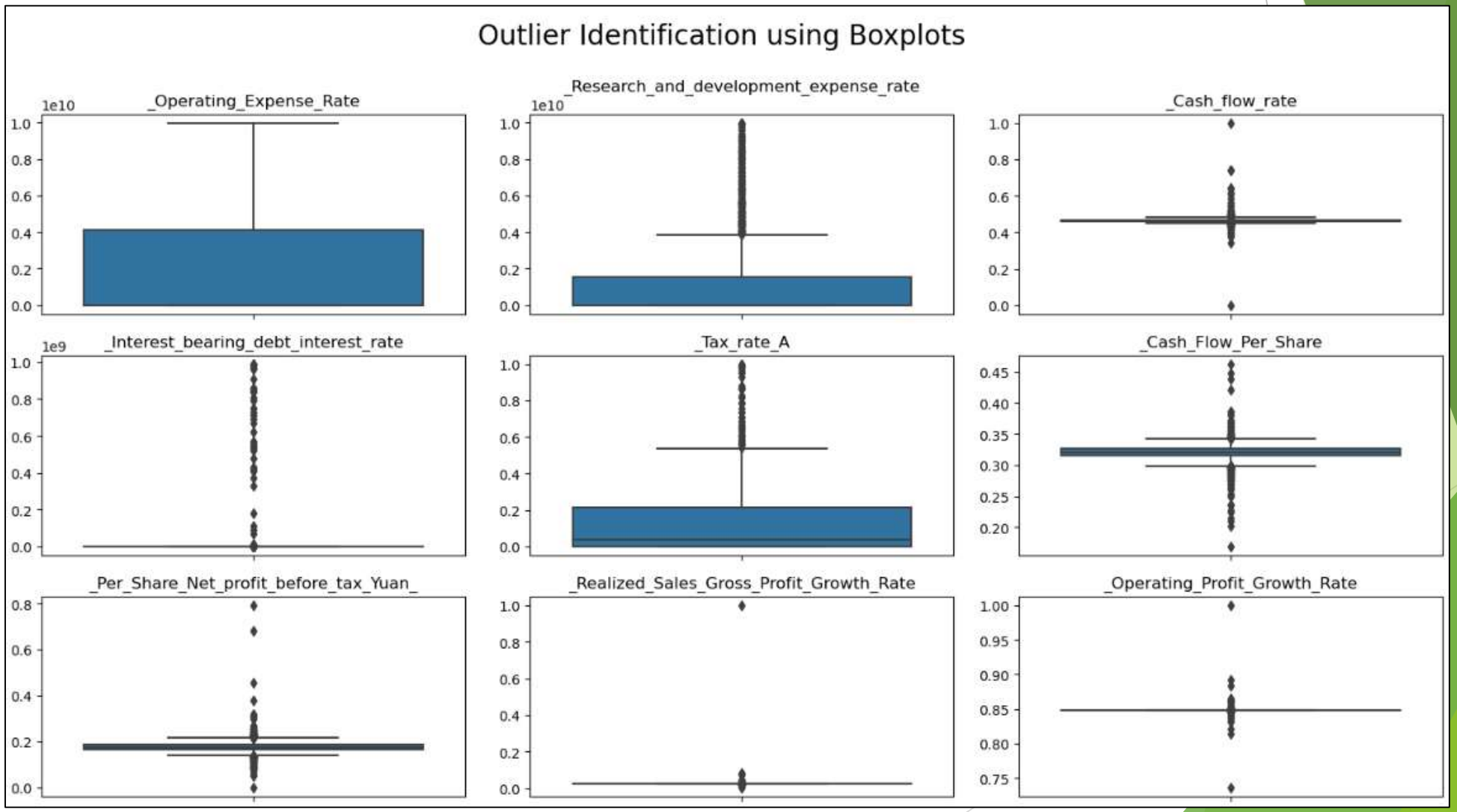
- The data consists of 2058 rows and 58 columns.
- Null values are present in a few columns the dataset.
- No duplicate values are present in the dataset.
- Most of the variables are of the "float" datatype.
- Only 220 companies which accounts for 10.6% of the entire dataset have defaulted which represents a major class imbalance which should be rectified while model building.

	Co_Code	Co_Name	_Operating_Expense_Rate	_Research_and_development_expense_rate	_Cash_flow_rate	_Interest_bearing_debt_interest_rate	_Tax_rate_A
0	16974	Hind.Cables	8.820000e+09	0.000000e+00	0.462045	0.000352	0.00141
1	21214	Tata Tele. Mah.	9.380000e+09	4.230000e+09	0.460116	0.000716	0.00000
2	14852	ABG Shipyard	3.800000e+09	8.150000e+08	0.449893	0.000496	0.00000
3	2439	GTL	6.440000e+09	0.000000e+00	0.462731	0.000592	0.00931
4	23505	Bharati Defence	3.680000e+09	0.000000e+00	0.463117	0.000782	0.40024

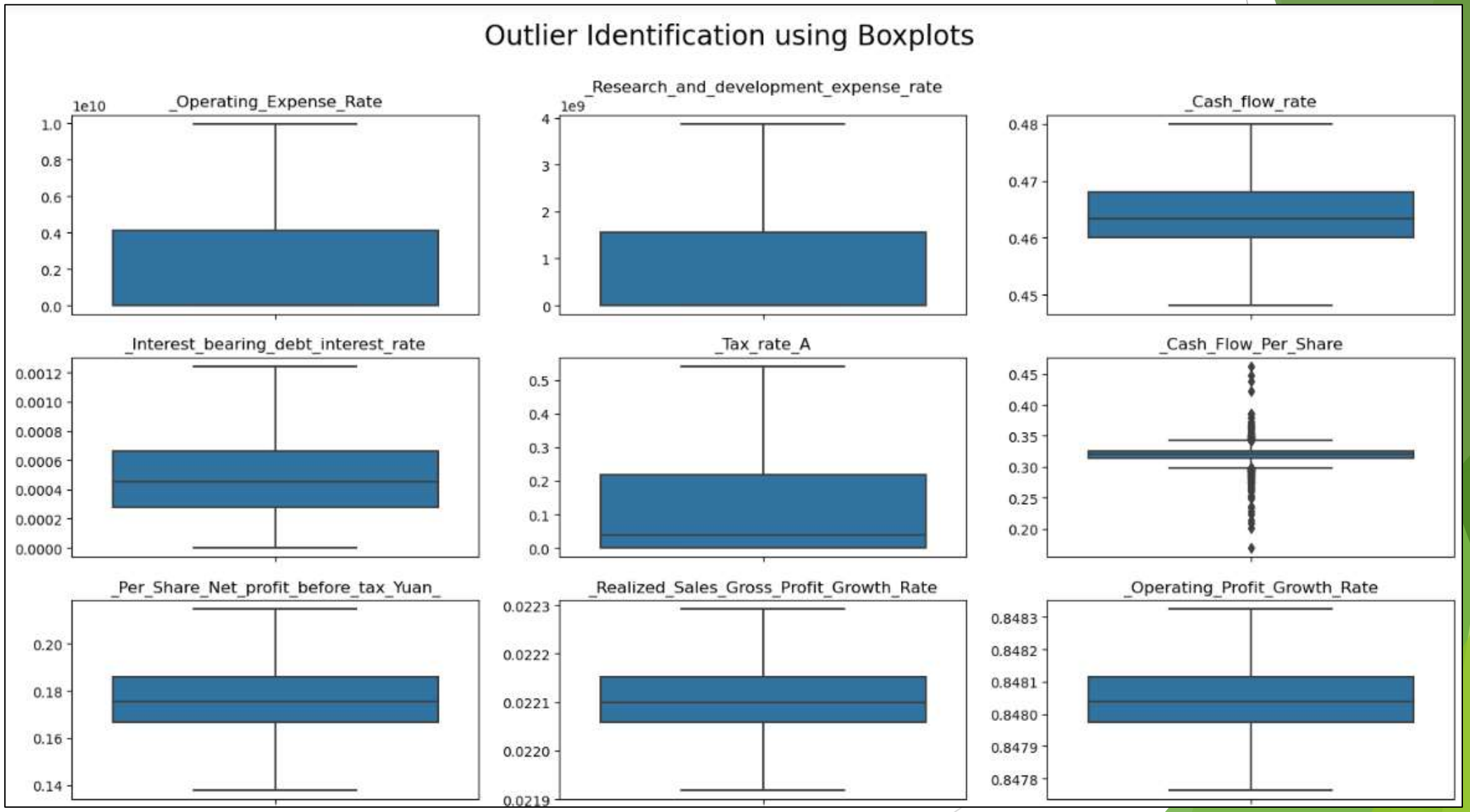
	Co_Code	Co_Name	_Operating_Expense_Rate	_Research_and_development_expense_rate	_Cash_flow_rate	_Interest_bearing_debt_interest_rate	_Tax_r
2053	2743	Kothari Ferment.	3.021580e-04	6.490000e+09	0.477066	0.000000	0.1
2054	21216	Firstobj.Tech.	1.371450e-04	0.000000e+00	0.465211	0.000658	0.0
2055	142	Diamines & Chem.	2.114990e-04	8.370000e+09	0.480248	0.000502	0.0
2056	18014	IL&FS Engg.	3.750000e+09	0.000000e+00	0.474670	0.000578	0.3
2057	43229	Channel Nine	2.981110e-04	0.000000e+00	0.467203	0.000826	0.0

Outlier Treatment

Many Outliers can be observed in different columns and these need to be treated. The IQR technique has been used to cap these outliers.



The outliers have been successfully treated as seen in the boxplots.



Missing Values Treatment

A total of 298 null values are present which roughly accounts for 14% of the dataset. Out of these, 30 null observations are present for companies that have defaulted. This accounts for 13% of the total default companies present in the dataset.

Dropping these rows would mean that we'll lose meaningful data for the default companies. Hence it would be wise to impute these null values with their respective mean observations.

```
Null Values for _Cash_Flow_Per_Share: 167
Total defaulters for these null values: 12

Null Values for _Total_debt_to_Total_net_worth: 21
Total defaulters for these null values: 3

Null Values for _Cash_to_Total_Assets: 96
Total defaulters for these null values: 14

Null Values for _Current_Liability_to_Current_Assets: 14
Total defaulters for these null values: 1
```

Null Values have been removed from all of the columns.

_Operating_Expense_Rate	0
_Research_and_development_expense_rate	0
_Cash_flow_rate	0
_Interest_bearing_debt_interest_rate	0
_Tax_rate_A	0
_Cash_Flow_Per_Share	0
_Per_Share_Net_profit_before_tax_Yuan_	0
_Realized_Sales_Gross_Profit_Growth_Rate	0
_Operating_Profit_Growth_Rate	0
_Continuous_Net_Profit_Growth_Rate	0
_Total_Asset_Growth_Rate	0
_Net_Value_Growth_Rate	0
_Total_Asset_Return_Growth_Rate_Ratio	0
_Cash_Reinvestment_perc	0
_Current_Ratio	0
_Quick_Ratio	0
_Interest_Expense_Ratio	0
_Total_debt_to_Total_net_worth	0
_Long_term_fund_suitability_ratio_A	0
_Net_profit_before_tax_to_Paid_in_capital	0
_Total_Asset_Turnover	0
_Accounts_Receivable_Turnover	0
_Average_Collection_Days	0
_Inventory_Turnover_Rate_times	0
_Fixed_Assets_Turnover_Frequency	0
_Net_Worth_Turnover_Rate_times	0
_Operating_profit_per_person	0
_Allocation_rate_per_person	0
_Quick_Assets_to_Total_Assets	0
_Cash_to_Total_Assets	0
_Quick_Assets_to_Current_Liability	0

_Cash_to_Current_Liability	0
_Operating_Funds_to_Liability	0
_Inventory_to_Working_Capital	0
_Inventory_to_Current_Liability	0
_Long_term_Liability_to_Current_Assets	0
_Retained_Earnings_to_Total_Assets	0
_Total_income_to_Total_expense	0
_Total_expense_to_Assets	0
_Current_Asset_Turnover_Rate	0
_Quick_Asset_Turnover_Rate	0
_Cash_Turnover_Rate	0
_Fixed_Assets_to_Assets	0
_Cash_Flow_to_Total_Assets	0
_Cash_Flow_to_Liability	0
_CFO_to_Assets	0
_Cash_Flow_to_Equity	0
_Current_Liability_to_Current_Assets	0
_Total_assets_to_GNP_price	0
_No_credit_Interval	0
_Degree_of_Financial_Leverage_DFL	0
_Interest_Coverage_Ratio_Interest_expense_to_EBIT	0
_Equity_to_Liability	0
Default	0
_Liability_Assets_Flag	0

Correlation Plot

High correlation can be observed in multiple columns:

- _CFO_to_Assets
- _Quick_Ratio
- _Net_Profit_before_tax
- _Net_Worth_Turnover_Rate
- _Operating_Funds_to_Liability

Train-Test Split

The model is split into Train and Test datasets in the proportion of 67:33. The parameter '*stratify*' has been passed to ensure that both the train and test datasets consist of approximately equal distribution of the default variable.

```
Train Size:(1378, 54)
Test Size:(680, 54)
```

```
Train Default Distribution:
0    0.893324
1    0.106676
Name: Default, dtype: float64

Test Default Distribution:
0    0.892647
1    0.107353
Name: Default, dtype: float64
```

Logistic Regression Model

As previously observed in the correlation plot, the dataset consists of high multicollinearity. Multicollinearity affects the performance of a Logistic Regression model hence it needs to be rectified. This has been achieved by determining the Variance Inflation Factor (VIF) values for different variables and removing the variable with the highest variance. This is a reiterative process and is carried out until the VIF values of all the remaining variables is less than 5.

variables	VIF
_Per_Share_Net_profit_before_tax_Yuan_	98.438000
_Net_profit_before_tax_to_Paid_in_capital	98.402854
_Cash_Flow_to_Total_Assets	46.443945
_CFO_to_Assets	28.304223
_Operating_Funds_to_Liability	21.146109
_Quick_Assets_to_Current_Liability	20.070249
_Cash_Flow_to_Liability	18.088591
_Cash_flow_rate	16.560195
_Cash_Flow_to_Equity	16.072751
_Quick_Ratio	12.476794

Logistic Regression Model

After eliminating the correlated variables, different models have been built using the remaining variables and the most insignificant variable (p-value>0.05) is removed.

In the provided model report, the p-value of the variable '*_Liability_Assets_Flag*' is of 0.99 which means that the contribution of this variable is not significant. Hence this variable can be dropped and the model is built using the other variables.

This process is repeated until only the significant variables are remaining for the model building.

	coef	std err	z	P>|z|
Intercept	-3.4692	0.529	-6.557	0.000
_Accounts_Receivable_Turnover	-620.4251	142.868	-4.343	0.000
_Allocation_rate_per_person	38.7310	9.214	4.204	0.000
_Total_Asset_Growth_Rate	-4.524e-11	3.76e-11	-1.204	0.229
_Cash_to_Current_Liability	-39.5344	28.285	-1.398	0.162
_Total_expense_to_Assets	50.4959	6.729	7.504	0.000
_Cash_to_Total_Assets	-4.4968	2.098	-2.143	0.032
_Interest_bearing_debt_interest_rate	988.1986	366.753	2.694	0.007
_Inventory_to_Current_Liability	-22.2366	17.208	-1.292	0.196
_Total_assets_to_GNP_price	39.1661	23.375	1.676	0.094
_Current_Asset_Turnover_Rate	-96.5336	104.891	-0.920	0.357
_Current_Liability_to_Current_Assets	7.3226	2.070	3.537	0.000
_Long_term_Liability_to_Current_Assets	-8.6165	12.685	-0.679	0.497
_Fixed_Assets_Turnover_Frequency	15.5592	13.701	1.136	0.256
_Cash_Turnover_Rate	-1.034e-10	4.25e-11	-2.432	0.015
_Quick_Asset_Turnover_Rate	1.595e-11	3.12e-11	0.510	0.610
_Tax_rate_A	-6.6870	1.317	-5.079	0.000
_Inventory_Turnover_Rate_times	2.891e-11	3.43e-11	0.842	0.400
_Research_and_development_expense_rate	2.181e-10	6.67e-11	3.270	0.001
_Operating_Expense_Rate	3.284e-11	3.44e-11	0.955	0.340
_Liability_Assets_Flag	15.7957	6368.868	0.002	0.998
_Total_debt_to_Total_net_worth	3.935e-10	2.4e-10	1.641	0.101

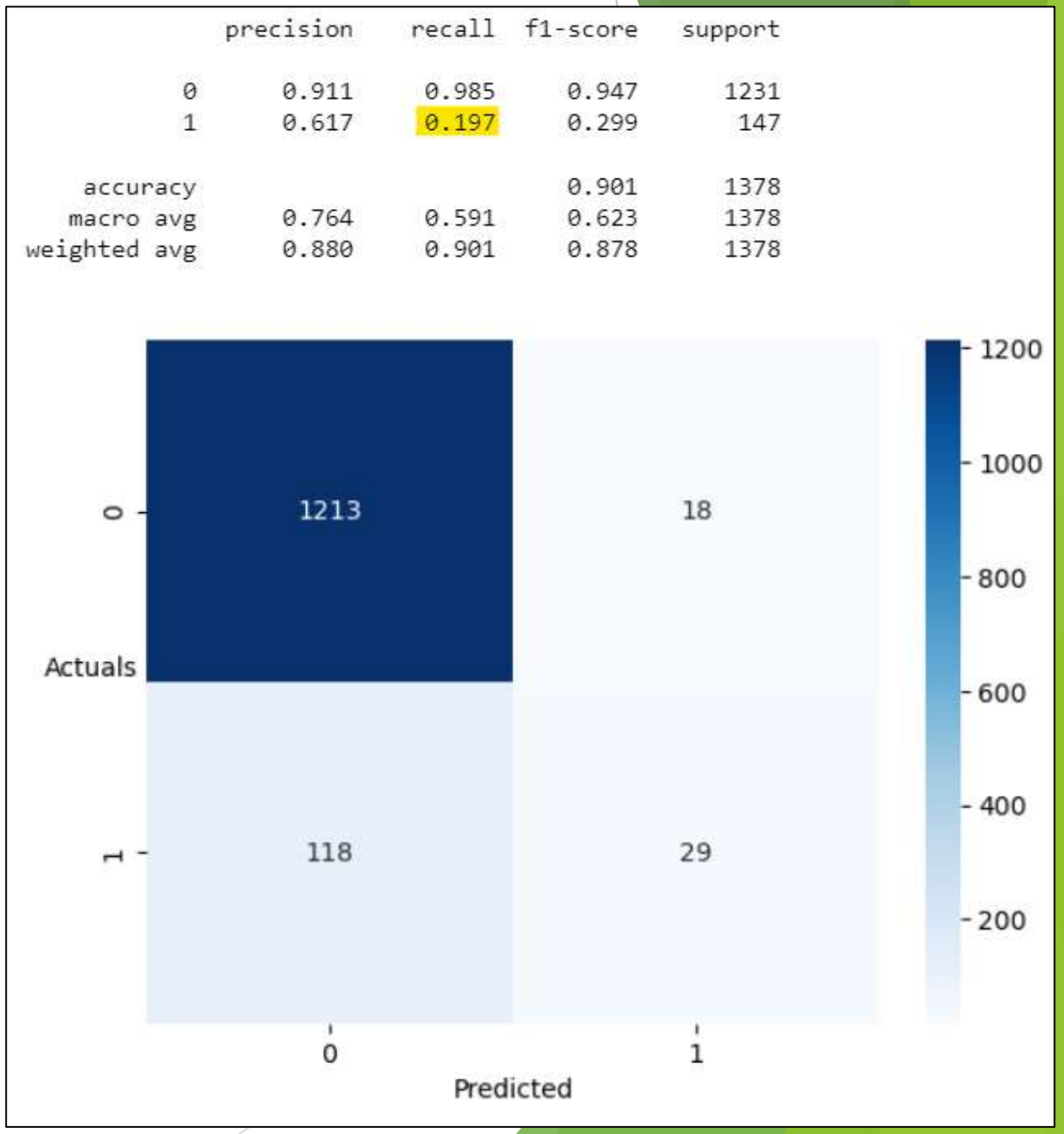
Logistic Regression Model

Once the model comprises of only the significant variables, its performance is evaluated on the training dataset.

The classification report and confusion matrix helps evaluate the performance of the final logistic regression model and it can be observed that the recall value for the default companies is very low. This can be attributed to the considerable class imbalance between the default and non-default companies which was observed earlier

This model can thus be optimized by redefining the optimum threshold frequency to efficiently distinguish between the actual default companies.

	precision	recall	f1-score	support
0	0.911	0.985	0.947	1231
1	0.617	0.197	0.299	147
accuracy			0.901	1378
macro avg	0.764	0.591	0.623	1378
weighted avg	0.880	0.901	0.878	1378



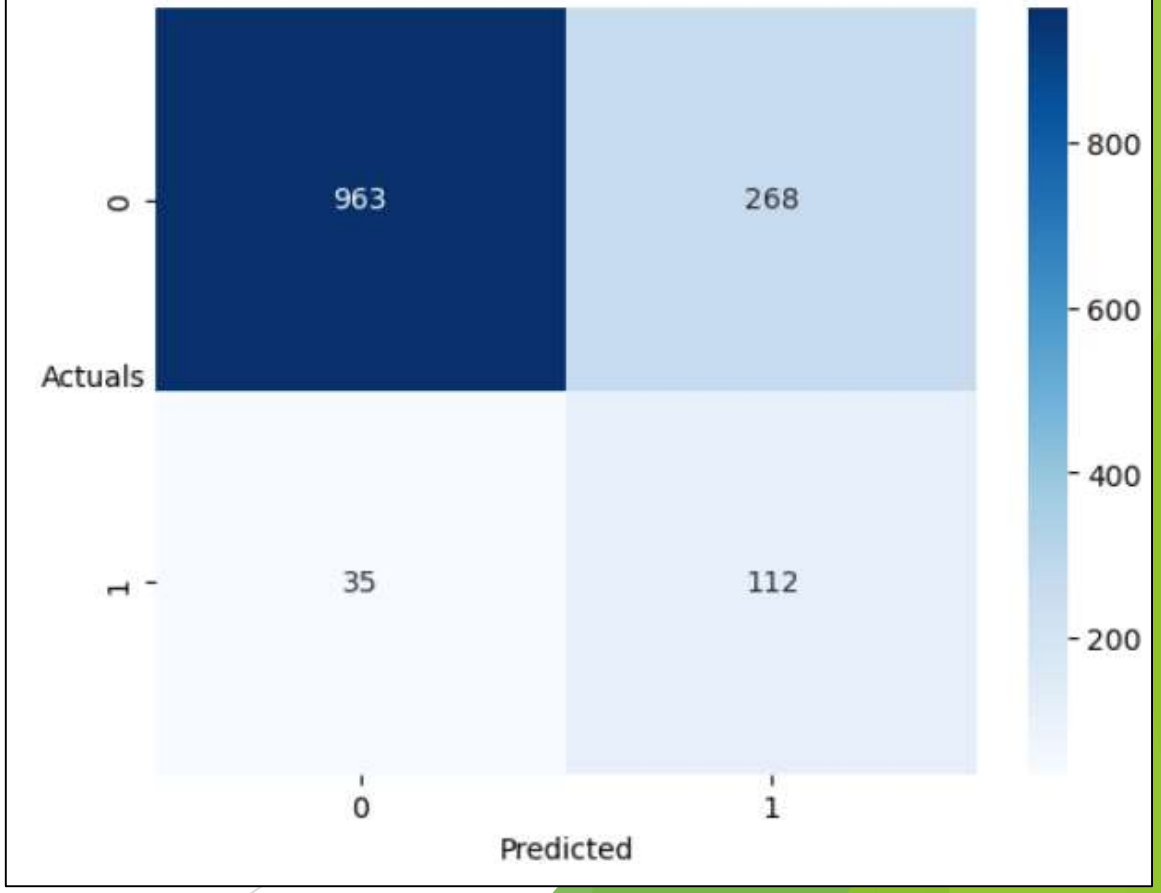
Logistic Regression Model

After redefining the optimum threshold frequency, it can be observed that the recall value has significantly increased for the train dataset.

The Logistic Regression model is thus able to correctly predict 76% of the actual default companies with an accuracy of 78% for the train dataset.

We'll now proceed by applying this model on our test dataset followed by evaluating its performance.

	precision	recall	f1-score	support
0	0.965	0.782	0.864	1231
1	0.295	0.762	0.425	147
accuracy			0.780	1378
macro avg	0.630	0.772	0.645	1378
weighted avg	0.893	0.780	0.817	1378



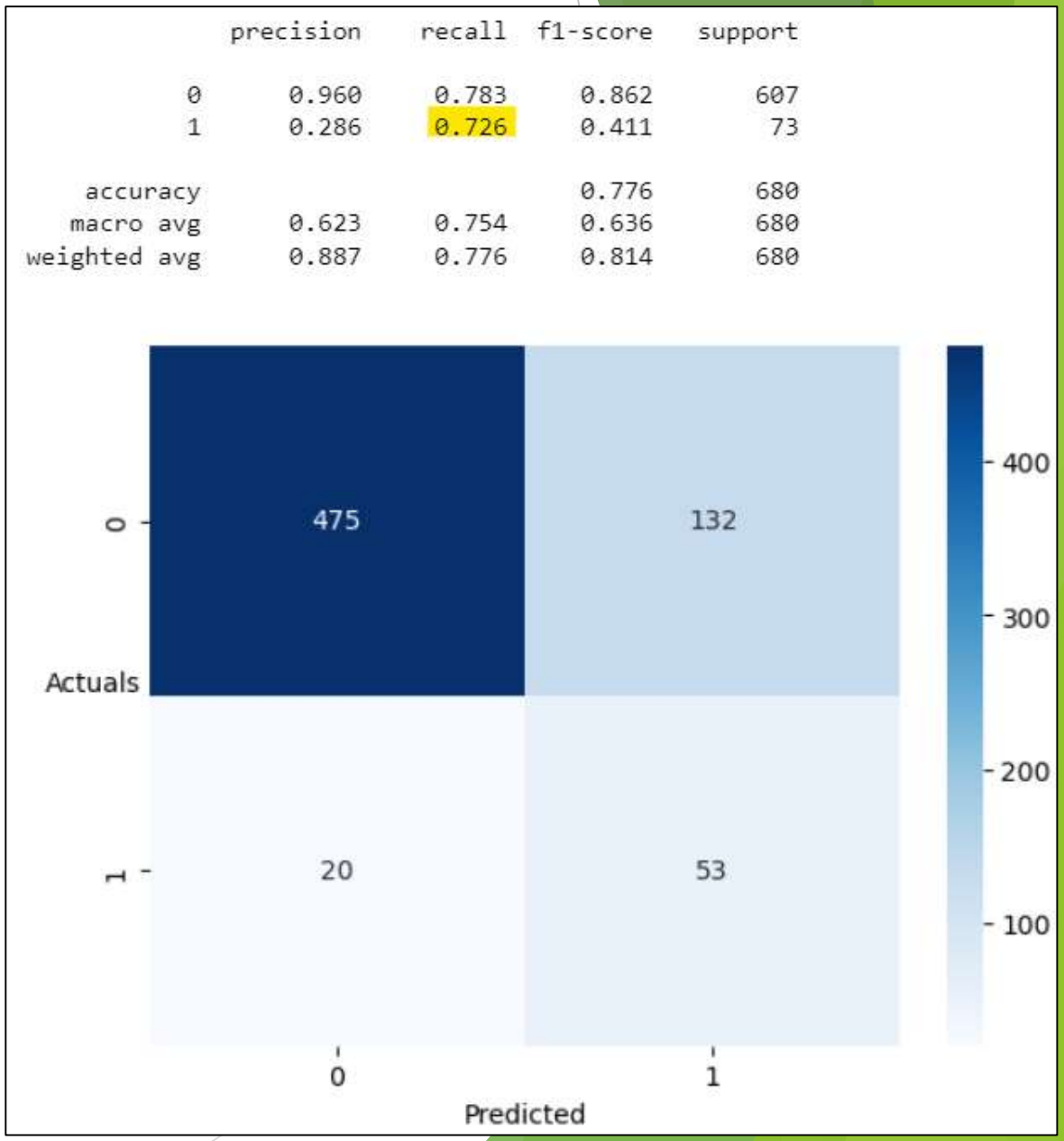
Logistic Regression Model

When evaluating the model performance on the test dataset, similar values for recall and model accuracy can be observed.

The Logistic Regression model is able to correctly predict 72% of the actual default companies with an accuracy of 78% for the train dataset.

All in all, this model predicts the default variable well. We'll build other models and compare their performance metrics to identify the best model.

	precision	recall	f1-score	support
0	0.960	0.783	0.862	607
1	0.286	0.726	0.411	73
accuracy			0.776	680
macro avg	0.623	0.754	0.636	680
weighted avg	0.887	0.776	0.814	680



Random Forest Model

The Random Forest model has been built by using the Grid Search function which helps in identifying the best hyperparameters for building the model.

The Random Forest model is an ensemble modelling technique consisting of multiple decision trees and the model built is an aggregate of these individual trees. Hence it becomes very important to ensure that the correct hyper-parameters are used for model building to ensure proper pruning of all the decision trees involved in the model.

The following hyper-parameters have been identified as appropriate for building the Random Forest Model:

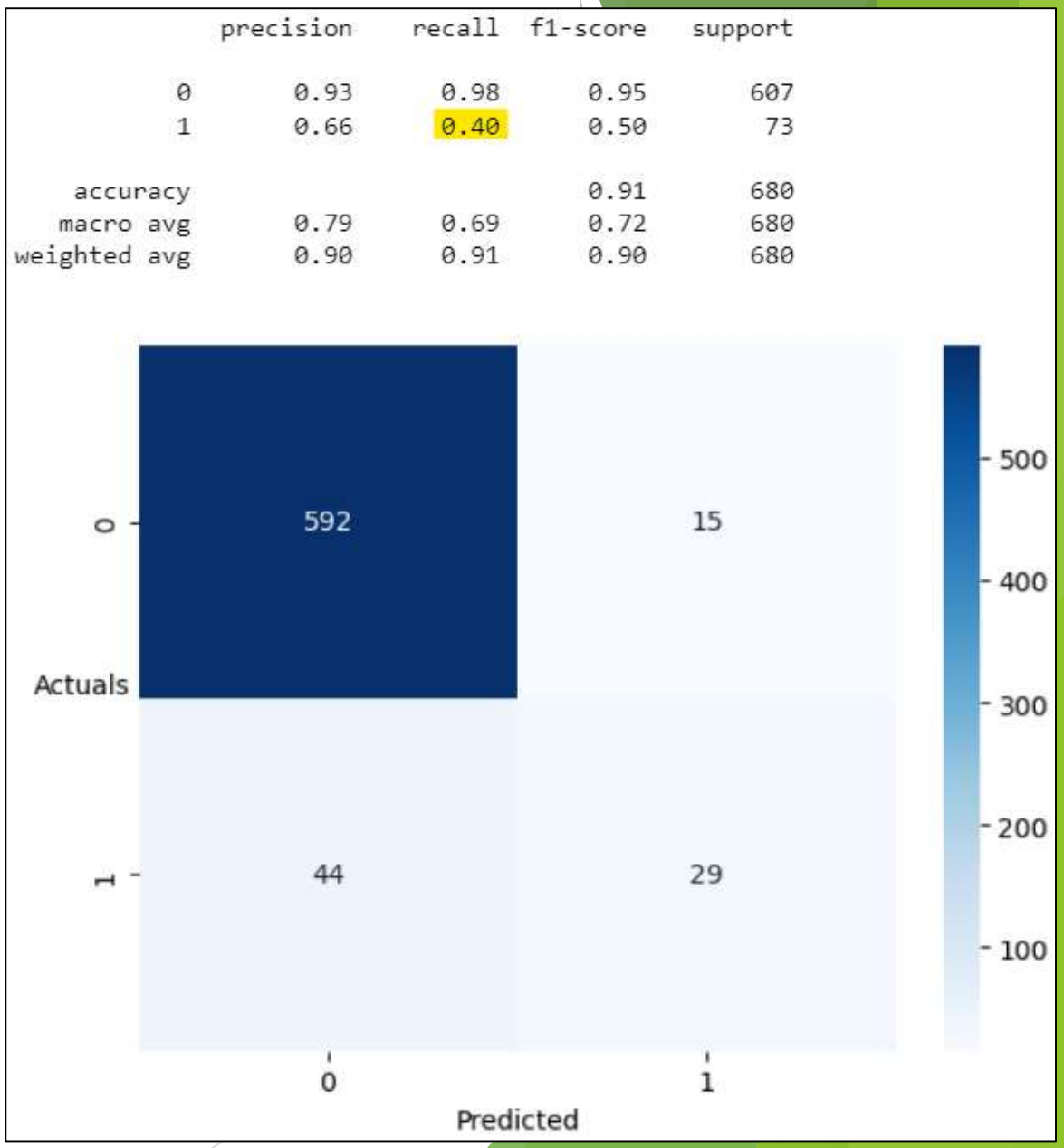
```
{'max_depth': 9,
 'min_samples_leaf': 15,
 'min_samples_split': 45,
 'n_estimators': 25}
```

Random Forest Model

When evaluating the Random Forest model performance on the test dataset, it can be clearly observed that the model is underperforming.

The model is able to correctly predict 40% of the actual default companies with an accuracy of 91% for the test dataset. The overall accuracy of the model is high however due to its low recall value, it cannot be considered an efficient model.

	precision	recall	f1-score	support
0	0.93	0.98	0.95	607
1	0.66	0.40	0.50	73
accuracy			0.91	680
macro avg	0.79	0.69	0.72	680
weighted avg	0.90	0.91	0.90	680



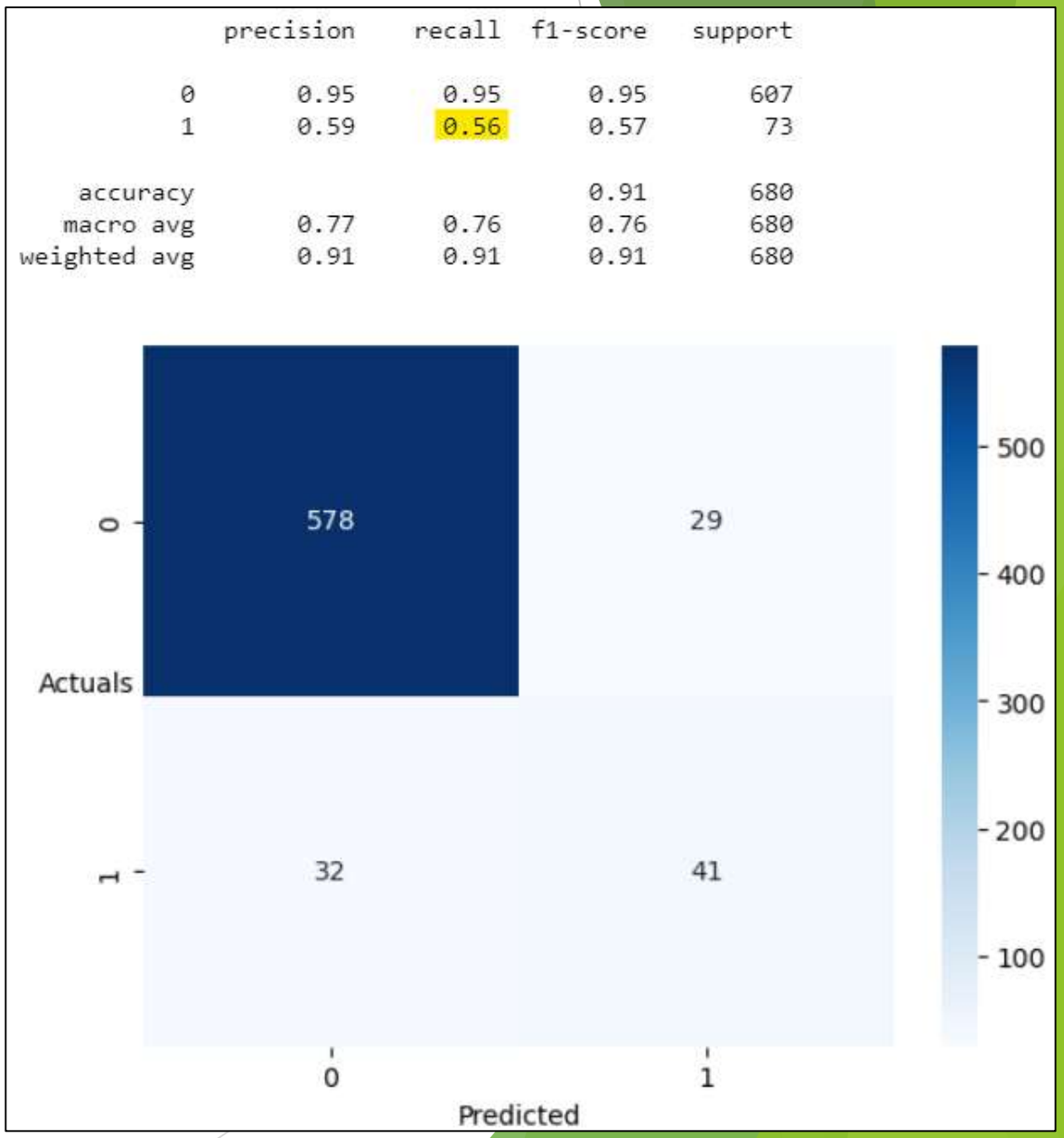
Linear Discriminant Analysis Model

The Linear Discriminant Analysis (LDA) Model focuses on finding linear combinations of the best features that are available in available in our dataset.

When evaluating the LDA model performance on the test dataset, it can be observed that the model performs well in comparison to the Random Forest model however it still underperforms in comparison to the Logistic Regression model.

The model is able to correctly predict 56% of the actual default companies with an accuracy of 91% for the test dataset. This makes it a better model.

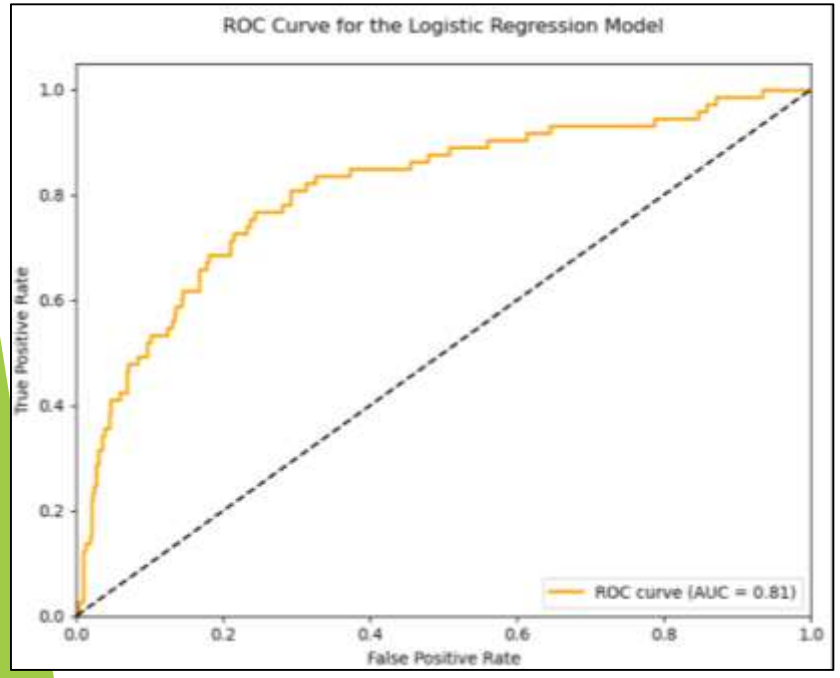
	precision	recall	f1-score	support
0	0.95	0.95	0.95	607
1	0.59	0.56	0.57	73
accuracy			0.91	680
macro avg	0.77	0.76	0.76	680
weighted avg	0.91	0.91	0.91	680

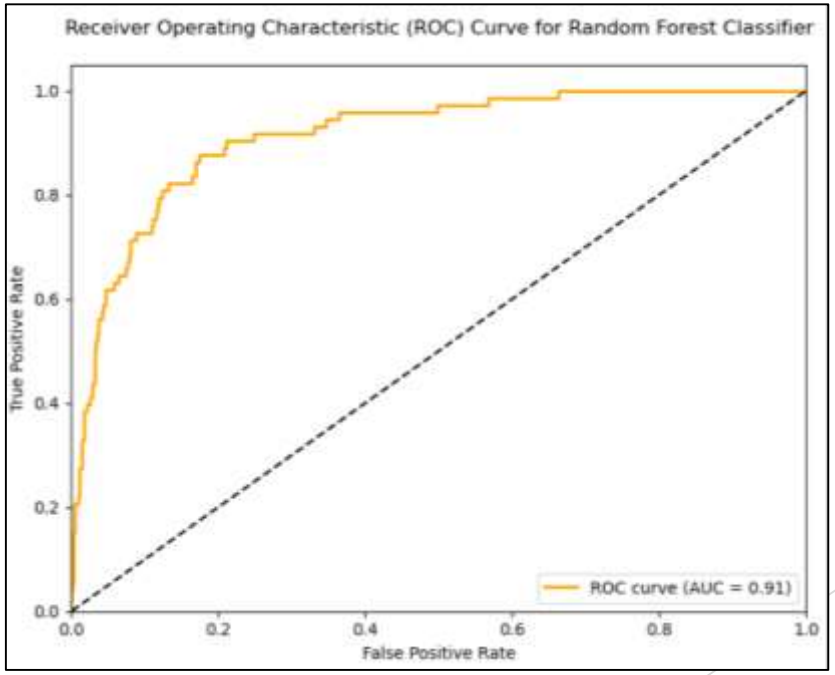


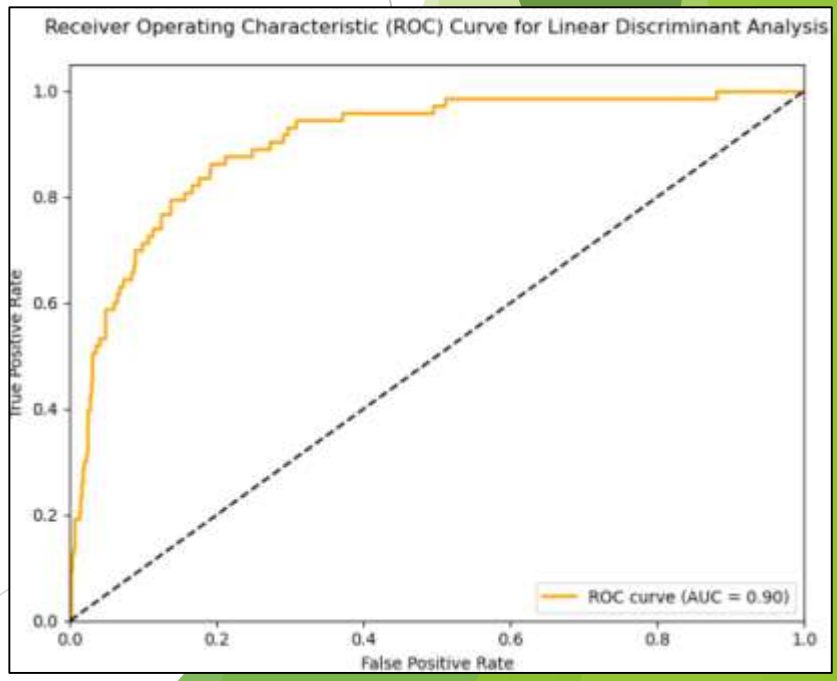
Model Performances Comparison

When comparing the model performances and ROC curves of all three models, it can be said that the LDA model outperforms the other models in terms of overall model accuracy and AUC scores. However, the goal of our model is to predict the maximum number of default companies. The LDA model fails to capture this in comparison to the Logistic Regression model.

The Logistic Regression model correctly predicts most of the default companies which is the main objective of our problem statement hence we'll not consider the other models appropriate despite their high values of accuracy and AUC scores.







Conclusions and Recommendations

The following recommendations can be given to investors and stockholders for evaluating a company's financial performance:

- Despite the large number of metrics available, a company's financial performance can be attributed to a few important features and the key is to correctly identify them for a company's efficient evaluation.

- The features '*Total_expense_to_Assets*', '*Allocation_rate_per_person*' and '*Tax_Rate_A*' have been identified as the most important features when determining whether or not a company would default hence they should always be included in times of important decision making.

- Including all of the available features to evaluate a company's financial position is redundant as its highly likely that these features are related which would lead to a discrepant evaluation.