FI505E_par- Coding and Data Science for Accounting and Finance

November 2022

# Corporate Credit Rating Forecasting

## *Introduction*

The financial rating of a company is a useful indicator for investors to know which companies will yield profitable investments for them. Corporate ratings are given by agencies such as Moody's, Standard & Poor's, and Fitch Ratings, where (slightly variating among the agencies) AAA is the highest a company can receive (being that they have a very low credit default risk), and D is the lowest (being that the company is currently in default or has an extremely high probability of default).

This project has been carried out to answer the question, can a company's credit rating be forecasted using machine learning tools and techniques? The team was motivated to answer this question because knowing in advance what a company's rating will be before the agencies actually rate them would lead them to have a significant advantage in terms of investment since they could buy stock early for a relatively low price in comparison to investors without this information and sell it later for a higher price.

To answer the proposed question, the team set out to build and evaluate three different types of machine learning models: a multinomial logistic regression, a decision tree, and a support vector machine model. For the latter, the team used three variations: linear, polynomial and radial basis function. Therefore, in total five machine learning models were compared. It was found that the decision tree model yielded the highest predicting accuracy, but the obtained level of accuracy was not significant enough for investors to heavily rely on this model for orientation on their financial investments.

## *Data*

The data for this project has come from a given comma separated values file with no missing values containing cross-sectional data of 30 features for 2021 companies, 25 of those features being financial ratios (the independent variables or predictors) that can be classified in 5 different categories: liquidity measurement ratios, profitability indicator ratios, debt ratios, operating performance ratios, and cash flow indicator ratios. Each of these companies has already received a corporate rating included in the dataset (this is the class dependent variable); the aim will be to, using this data, build, train, and test a machine learning model that may be able to predict the corporate credit rating a new company will receive. To homogenize the rating names across the agencies, new risk labels have been set (low risk, lowest risk, medium risk, high risk, highest risk, and in default). However there are very few observations in the lowest risk and in default categories so it has been decided that they should be removed from the dataset for simplification purposes.

### *Statistical Summary*

This is what the statistical summary looks like for the first seven of the financial ratios contained in the dataset (all of them can be seen in Appendix 1). The mean, standard deviation, quartiles,

minimum, and maximum provide basic statistical information that allows analysts to get a broad picture of the data they are dealing with.

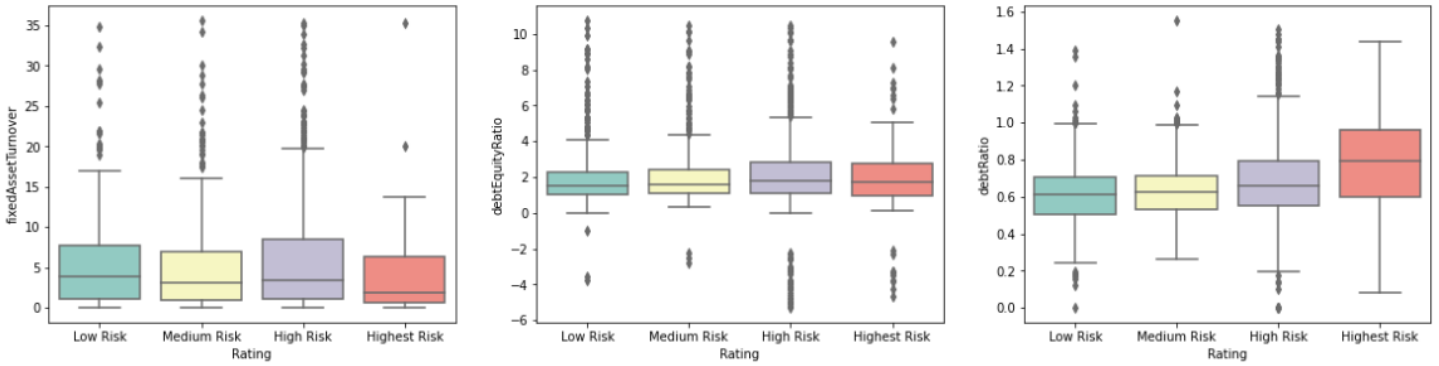
| | currentRatio | quickRatio | cashRatio | daysOfSalesOutstanding | netProfitMargin | pretaxProfitMargin | grossProfitMargin |
|---|---|---|---|---|---|---|---|
| count | 2021.000000 | 2021.000000 | 2021.000000 | 2021.000000 | 2021.000000 | 2021.000000 | 2021.000000 |
| mean | 3.535411 | 2.657150 | 0.669048 | 334.855415 | 0.278725 | 0.432721 | 0.496900 |
| std | 44.139386 | 33.009920 | 3.590902 | 4456.606352 | 6.076128 | 9.002733 | 0.525996 |
| min | -0.932005 | -1.893266 | -0.192736 | -811.845623 | -101.845815 | -124.343612 | -14.800817 |
| 25% | 1.071930 | 0.602298 | 0.131433 | 22.806507 | 0.020894 | 0.025649 | 0.232565 |
| 50% | 1.492804 | 0.979094 | 0.297859 | 42.281804 | 0.064323 | 0.084965 | 0.414217 |
| 75% | 2.160710 | 1.450457 | 0.625355 | 59.165369 | 0.113871 | 0.144763 | 0.849693 |
| max | 1725.505005 | 1139.541703 | 125.917417 | 115961.637400 | 198.517873 | 309.694856 | 2.702533 |

### *Correlation Heatmap*

The correlation heatmap (see Appendix 2) shows the degree of a positive or negative relationship among independent variables, where -1 or 1 signify a perfect (negative and positive respectively) correlation. From the one obtained, it is seen that most of the variables do not hold a significant correlation since the coefficients are close to 0, or in this case, to a brownish color. However, there are a few independent variables that have a strong correlation such as the operating profit margin and the net profit margin (0.97), the net profit margin and the EBIT per revenue (0.99), the current ratio and the cash ratio (0.74), etc. These coefficients are understandable since these pairs of variables are either calculated using similar financial information or variables, or because one is a part of the calculations to obtain the other. The fact that most of the independent variables are not strongly correlated among each other is a good sign that suggests that there is no strong presence of multicollinearity in the dataset.

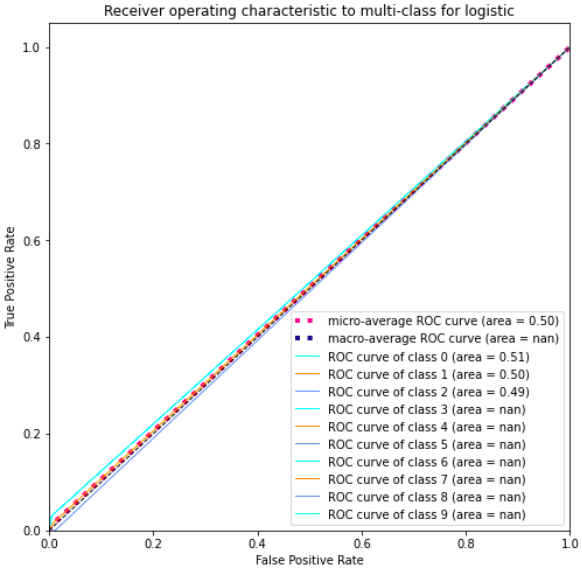
### *Box-Plots*

Box-plots give a visual statistical description of variables. They show the outliers, the maximum, the minimum, the interquartile range, and the median of any numeric variable. Box-plots for three ratios (fixed asset turnover, debt-to-equity ratio, and debt ratio) are seen below. All the box-plots can be seen in Appendix 3.




As it is seen, the medians (line inside the boxes) of the ratios variate according to the company rating which suggests the financial ratios are heterogeneous among the different ratings. This will help the machine learning models to be able to predict the ratings easier than if the ratios did not vary among companies with different ratings. However, some of these variations are very slight which might complicate the machine learning model to discern one class from another using the ratios.

## *The Machine Learning Model*




### *Data Splitting*

The first thing that must be done is to randomly (as it is not time series data) split the data into two. One part, 80% of the observations, will go towards training the machine to create a rating forecast model and the other part, the remaining 20%, will go towards testing the model by giving a prediction according to the ratios and comparing it with the actual ratings received; this will help measure the accuracy of the model.

### *Multinomial Logistic Regression*

The team decided to first see how a multinomial logistic regression model would work. This type of regression allows to identify an expected category (rating) based on a linear combination of predictor variables (ratios)[1]. After creating the model, its accuracy was tested and obtained that it made 40.95% accurate predictions. The receiver operating characteristic (ROC) curve for this model is presented at the left. The area under the curve indicates the model's predictions accuracy. It is easy to see that around half of the area falls under the curve for each of the classes or ratings. It is therefore not advised that investors rely in this model's predictions as it is not very accurate in predicting a company's rating. This is reason to continue to test other models in hope that they are able to do a better forecasting job than this multinomial logistic regression model.



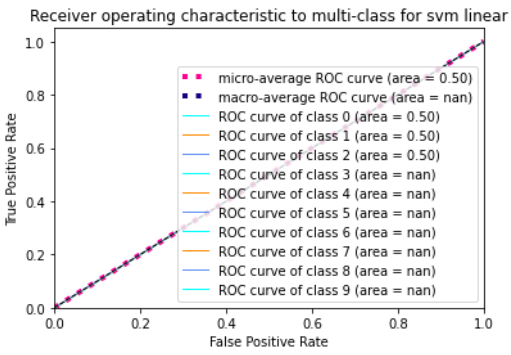
[1] UCLA. (2021). *Multinomial Logistic Regression | Stata Data Analysis Examples*. Retrieved from: https://stats.oarc.ucla.edu/stata/dae/multinomiallogistic-regression/

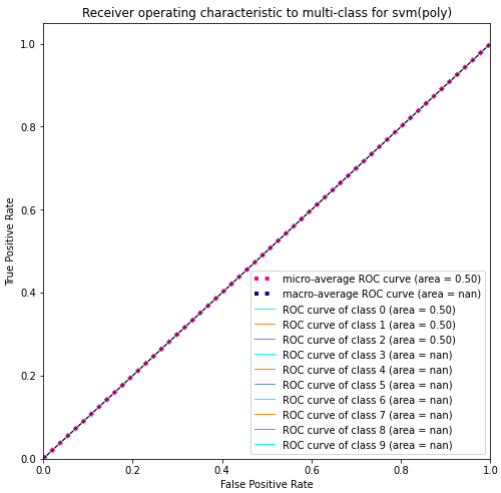
### *Decision Tree*

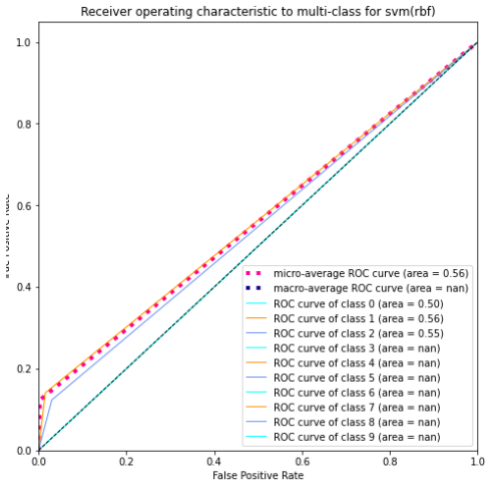
The team then decided to explore the option of a three-level decision tree. The obtained decision tree can be seen in Appendix 4. One of the main advantages of the decision tree it is that it is very easy to understand and follow, as it is very similar to how humans make decisions; it follows a yes/no pattern to get to a final result. However, decision trees are often not very accurate when making forecasts from a lot of data such as in this case. Indeed, a low accuracy value of 47.37% has been obtained. While this is better than the logistic regression model, it would still not be advised that investors base their financial investment decisions based on the rating predictions this model provides as it does not give enough confidence to trust in its predictions. One key takeaway from this model is that (according to the levels) the pre-tax profit margin is the most relevant variable among the ones in the dataset to predict the rating, followed by the debt ratio and the return on capital employed, and then by the operating cash flow per share and the net profit margin.

### *SVM Models*

In search for a more accurate model, the team decided to try out Support Vector Machine (SVM) linear, polynomial, and radial basis function (RBF) models. Nevertheless, this effort did not yield greater results. The linear SVM model had a 41.74% accuracy level, the polynomial 41.64%, and the RBF also 41.64%, making the linear the best out of the three SVM models by a very slight difference. Respective ROC plots are shown below.




Same as with the multinomial logistic regression, the ROC plots of the SVM linear and polynomial models have an area under the curve of 0.5. This means these models do not predict with great accuracy what a company's credit rating will be. On the other hand, the SVM RBF model has slightly more area under the curve (0.56). This accuracy exploring technique would suggest the SVM RBF model predicts more accurately than the previously mentioned models. In comparison with the decision tree, SVM models oftentimes work great with large amounts of data and yield more accurate results. However, a disadvantage of SVM models is that they are similar to a black box; they give good results, but analysts can't know for certain what was the process the machine followed to obtain them.

## *Results*

According to the performed accuracy tests, it is found that the decision tree is the best model for predicting a company's credit rating. However, while it has been proven that the machine can learn to take financial ratios information and give a rating prediction according to them, 47.37% accuracy does not give investors the confidence they need to rely on the given rating predictions and base their investment decisions on them.

## *Conclusion*

Machine learning models are useful in the area of finance since they simplify its users' decision-making processes and reduces to a certain extent their opportunity costs by helping them save time and effort when evaluating financial alternatives but still get accurate results. The intention of this project was to help investors get an informational advantage by building a model that could tell them, according to its financial ratios, what a new company's credit rating would be before the rating agencies announced it, and therefore give them the opportunity to take the upper hand in the market as these predictions would hint to them what a good or bad corporate investment would be.

At the end of this project, the question stated in the introduction has reached an answer; yes, it is possible to predict a company's rating according to its financial ratios using machine learning tools and techniques. This being said, it is important to note that the techniques the team tried out to carry this out were not satisfactory in terms of accuracy. This means that, yes, machine learning techniques can provide a prediction of a company's rating using financial ratios information, but the predictions' accuracy is not up to the standard of what a user would expect. As previously stated, the decision tree is the best machine learning model to predict a company's rating according to the financial ratios available in the given dataset, but with an accuracy level of 47.37%, could not possibly be relied upon.

An important limitation in this research that might have also caused low accuracy values is the fact that the dataset used for this project only contemplates financial ratios as independent variables when it is known that this is not the only thing that determines what a company's credit rating is. A suggestion for future exploration of this topic would therefore be to look closer at these ratios, remove the ones that are irrelevant, and add omitted variables such as share valuation, their position in corporate rankings, number of shares outstanding, a dummy variable of whether they have defaulted in credit payments in the past, and others that the researchers may deem necessary to account for.

Another reason why low accuracy values might have been obtained is that this project is dealing with a class dependent variable that has 4 different outcome possibilities (the ratings). It is oftentimes easier to classify when there are fewer categories than when there are various. For this, it is suggested that future researchers see if they obtain better results by creating only two rating categories; one that contains the low-risk ratings and one that contains the high-risk ratings. Hopefully, this would simplify the machine learning process and would give more accurate predictions. It is also suggested that future researchers explore other machine learning models such as random forest, or gradient boosting model and see if they provide more accurate predictions.

It is also important to consider that, if it were easy, everyone would be doing it. Rating agencies are believed to purposely not divulge exactly how they rate companies in order for them to have a sort of oligopoly in the rating department and for them and their functions to remain relevant. Therefore, it is not illogical that high accuracy values have not been obtained. Afterall, the machine learning models are trying to predict the rating given by these agencies without actually knowing how they do it. If any student or independent researcher could accurately predict a company's credit rating, there would soon not be a need for Moody's, Standard and Poor's and Fitch Ratings' rating services.

## *Appendix*

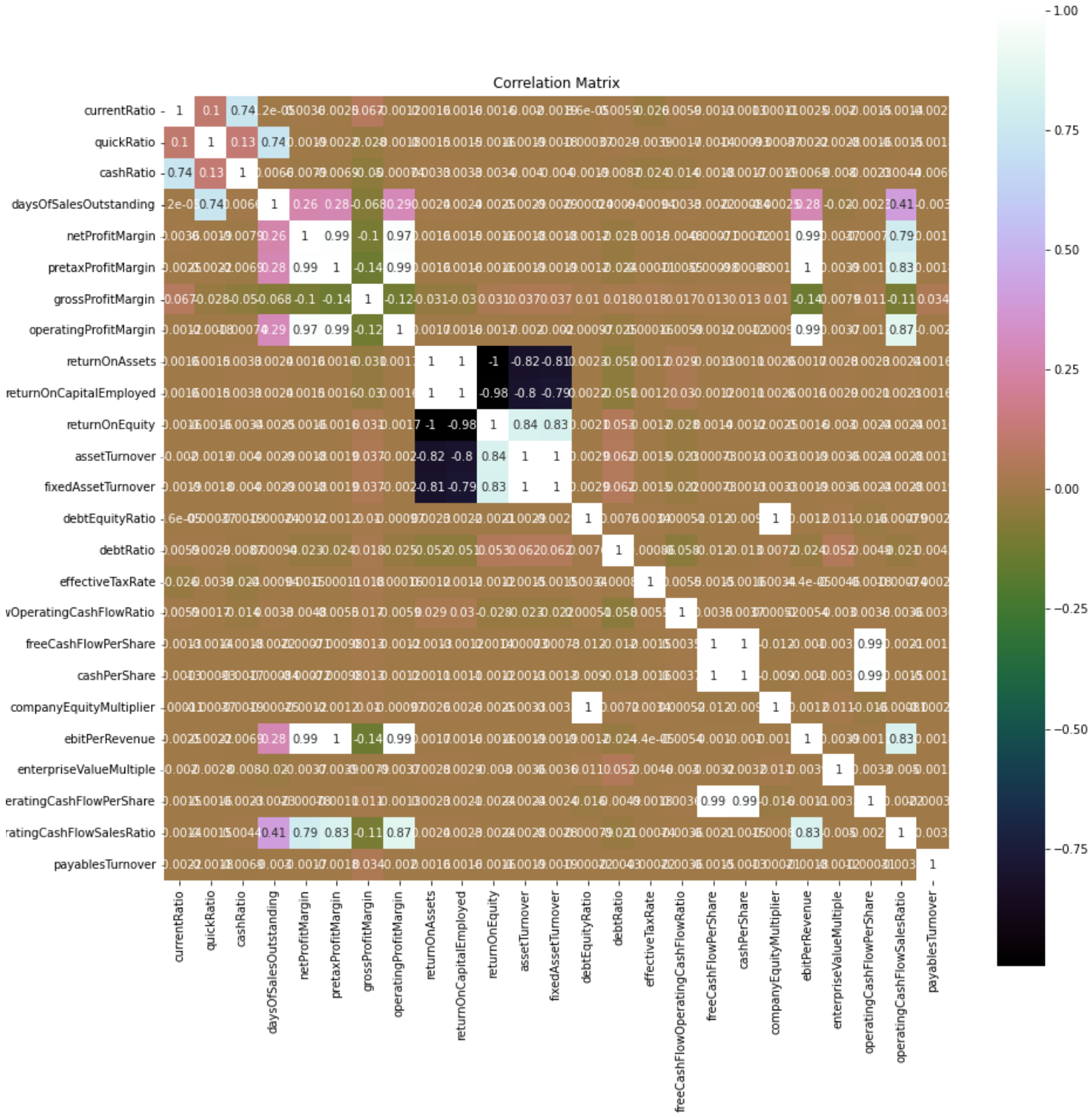
### Appendix 1: Statistical Summary

| | currentRatio | quickRatio | cashRatio | daysOfSalesOutstanding | netProfitMargin | pretaxProfitMargin | grossProfitMargin |
|---|---|---|---|---|---|---|---|
| count | 2021.000000 | 2021.000000 | 2021.000000 | 2021.000000 | 2021.000000 | 2021.000000 | 2021.000000 |
| mean | 3.535411 | 2.657150 | 0.669048 | 334.855415 | 0.278725 | 0.432721 | 0.496900 |
| std | 44.139386 | 33.009920 | 3.590902 | 4456.606352 | 6.076128 | 9.002733 | 0.525996 |
| min | -0.932005 | -1.893266 | -0.192736 | -811.845623 | -101.845815 | -124.343612 | -14.800817 |
| 25% | 1.071930 | 0.602298 | 0.131433 | 22.806507 | 0.020894 | 0.025649 | 0.232565 |
| 50% | 1.492804 | 0.979094 | 0.297859 | 42.281804 | 0.064323 | 0.084965 | 0.414217 |
| 75% | 2.160710 | 1.450457 | 0.625355 | 59.165369 | 0.113871 | 0.144763 | 0.849693 |
| max | 1725.505005 | 1139.541703 | 125.917417 | 115961.637400 | 198.517873 | 309.694856 | 2.702533 |

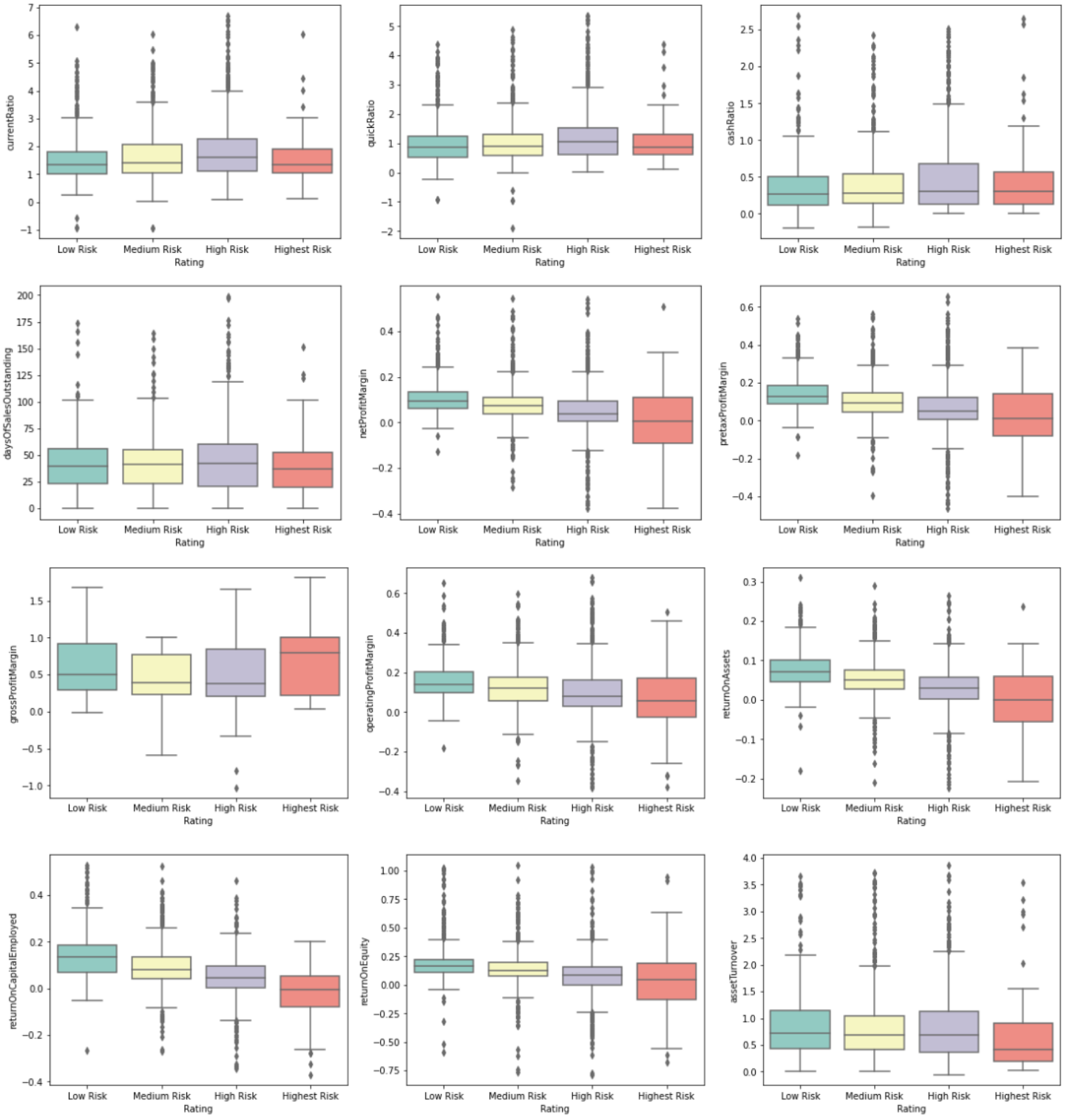
| operatingProfitMargin | returnOnAssets | returnOnCapitalEmployed | ... | effectiveTaxRate | freeCashFlowOperatingCashFlowRatio |
|---|---|---|---|---|---|
| 2021.000000 | 2021.000000 | 2021.000000 | ... | 2021.000000 | 2021.000000 |
| 0.588793 | -37.666843 | -74.267283 | ... | 0.400755 | 0.408272 |
| 11.246798 | 1168.476782 | 2354.920503 | ... | 10.613711 | 3.803929 |
| -124.343612 | -40213.178290 | -87162.162160 | ... | -100.611015 | -120.916010 |
| 0.044546 | 0.018757 | 0.028112 | ... | 0.147837 | 0.269616 |
| 0.107640 | 0.045417 | 0.074639 | ... | 0.300439 | 0.644265 |
| 0.175334 | 0.077159 | 0.135036 | ... | 0.370239 | 0.836949 |
| 410.182214 | 0.487826 | 2.439504 | ... | 429.926282 | 34.594086 |

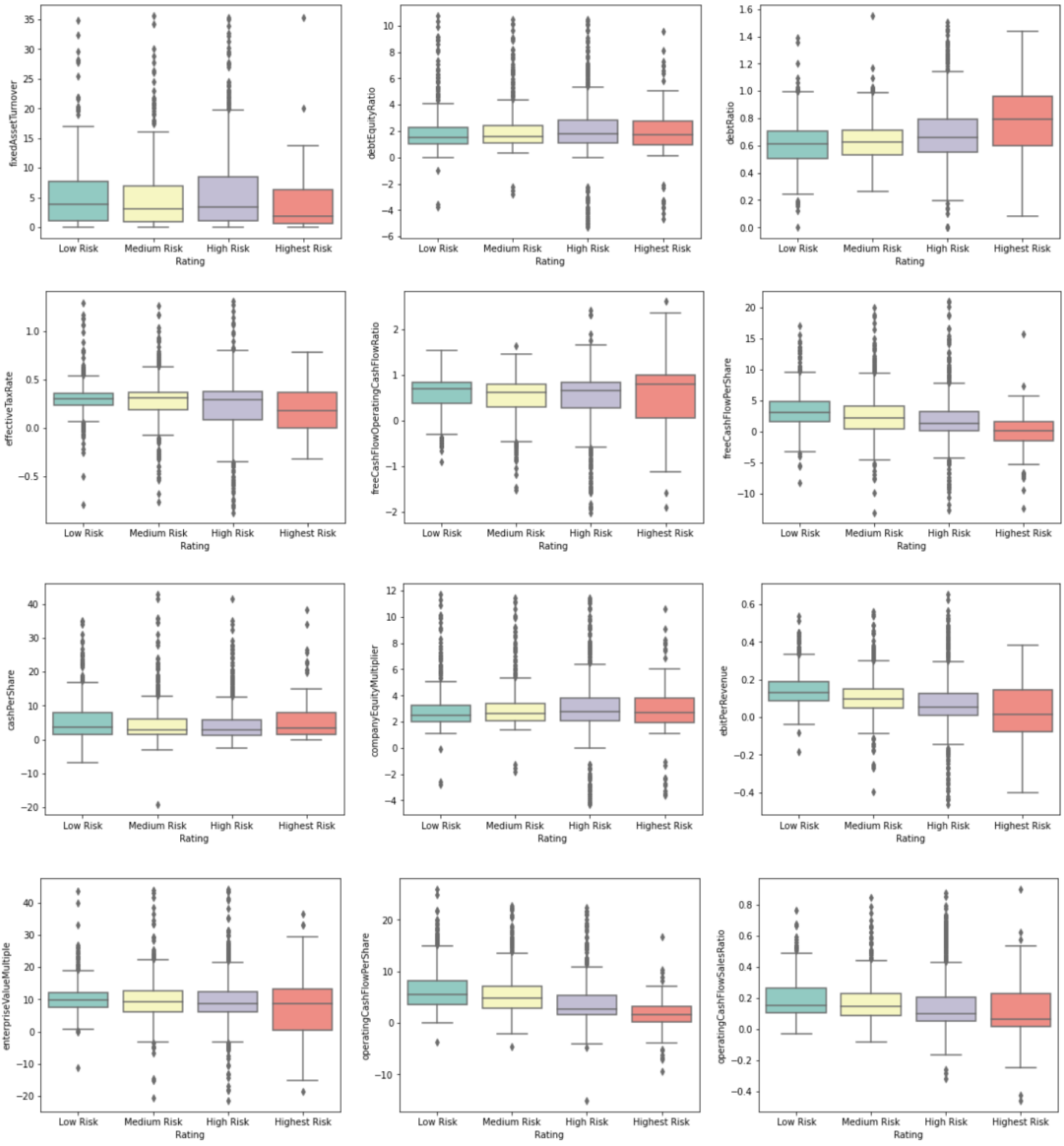
| freeCashFlowPerShare | cashPerShare | companyEquityMultiplier | ebitPerRevenue | enterpriseValueMultiple | operatingCashFlowPerShare |
|---|---|---|---|---|---|
| 2.021000e+03 | 2.021000e+03 | 2021.000000 | 2021.000000 | 2021.000000 | 2.021000e+03 |
| 5.114871e+03 | 4.244248e+03 | 3.335166 | 0.438715 | 48.426558 | 6.540891e+03 |
| 1.472059e+05 | 1.226418e+05 | 87.702375 | 9.002047 | 530.161001 | 1.778797e+05 |
| -4.912742e+03 | -1.915035e+01 | -2555.419643 | -124.343612 | -3749.921337 | -1.195049e+04 |
| 4.094118e-01 | 1.562116e+00 | 2.050249 | 0.028057 | 6.235759 | 2.348851e+00 |
| 2.123062e+00 | 3.680425e+00 | 2.657275 | 0.087424 | 9.269746 | 4.361649e+00 |
| 4.230253e+00 | 8.027524e+00 | 3.665438 | 0.149355 | 12.898855 | 7.322553e+00 |
| 5.753380e+06 | 4.786803e+06 | 2562.871795 | 309.694856 | 11153.607090 | 6.439270e+06 |

## Appendix 2: Correlation Heatmap



## Appendix 3: Box-Plots




fixedAssetTurnover
debtEquityRatio
debtRatio
effectiveTaxRate
freeCashFlowOperatingCashFlowRatio
freeCashFlowPerShare
cashPerShare
companyEquityMultiplier
ebitPerRevenue
enterpriseValueMultiple
operatingCashFlowPerShare
operatingCashFlowSalesRatio
Low Risk
Medium Risk
High Risk
Highest Risk
Rating

## Appendix 4: Decision Tree