**Broke or Nope? (A Bankruptcy Case Study)**

Brandon Thomas & Keegan Henderson,

[bthomas@bellarmine.edu](mailto:bthomas@bellarmine.edu) , [khenderson5@bellarmine.edu](mailto:khenderson5@bellarmine.edu)

**ABSTRACT**

The data used for this analysis project is taken from the UCI Machine Learning Library for use in learning amateur machine learning techniques. The data represent Polish companies over a forecasting period that included various measures of economic success to determine whether a specific company is bankrupt or not. After using data manipulation techniques, we combined the 5 data sets included in this repository into a large data frame for use with a variety of machine learning algorithms. To deal with missing data, we replaced any missing data with the mean of its corresponding column. In this analysis, we used logistic regression, decision trees, and random forest algorithms to create a model to predict the economic futures of these companies. Our analysis yielded results that showed that a decision tree was the best method of prediction out of the three models we used with an overall accuracy of 96%.

1. **INTRODUCTION**

The dataset we will be working with contains data about various economic measures of Polish companies that were analyzed to determine and classify which companies are predicted to become bankrupt or not. The data was collected from Emerging Markets Information Service a database containing information on many emerging markets around the world. The bankrupt companies were analyzed in the period 2000-2012, while the still operating companies were evaluated from 2007 to 2013.

1. **BACKGROUND**

The data for this project was originally collected as part of a larger database of information on emerging markets around the world, specifically the Emerging Markets Information Service (EMIS). The data in EMIS was then used by the creator of this repository, Sebastian Tomczak, to form this data set. Each row of data included in the data set represents a specific company’s financial statement for that specific year. Tomczak then used this data to, using machine learning methods and techniques, to predict whether a company would experience bankruptcy or not.

Table 1 – Dataset Information

|  |  |
| --- | --- |
| Dataset | Description |
| 1stYear | The data contains financial rates from the 1st year of the forecasting period and a corresponding class label that indicates the company’s bankruptcy status after 5 years. The data contains 7027 instances (financial statements), 271 representing the number of bankrupted companies, and 6756 firms that did not bankrupt in the forecasting period. |
| 2ndYear | The data contains financial rates from the 2nd year of the forecasting period and a corresponding class label that indicates the company’s bankruptcy status after 4 years. The data contains 10173 instances (financial statements), 400 representing the number of bankrupted companies, and 9773 firms that did not bankrupt in the forecasting period. |
| 3rdYear | The data contains financial rates from the 3rd year of the forecasting period and a corresponding class label that indicates the company’s bankruptcy status after 3 years. The data contains 10503 instances (financial statements), 495 representing the number of bankrupted companies, and 10008 firms that did not bankrupt in the forecasting period. |
| 4thYear | The data contains financial rates from the 4th year of the forecasting period and a corresponding class label that indicates the company’s bankruptcy status after 2 years. The data contains 9792 instances (financial statements), 515 representing the number of bankrupted companies, and 9277 firms that did not bankrupt in the forecasting period. |
| 5thYear | The data contains financial rates from the 5th year of the forecasting period and a corresponding class label that indicates the company’s bankruptcy status after 1 year. The data contains 5910 instances (financial statements), 410 representing the number of bankrupted companies, and 5500 firms that did not bankrupt in the forecasting period. |

Table 2 – Variable Descriptions

|  |  |
| --- | --- |
| Variable Number | Description |
| 1 | net profit / total assets |
| 2 | total liabilities / total assets |
| 3 | working capital / total assets |
| 4 | current assets / short-term liabilities |
| 5 | [(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] \* 365 |
| 6 | retained earnings / total assets |
| 7 | EBIT / total assets |
| 8 | book value of equity / total liabilities |
| 9 | sales / total assets |
| 10 | equity / total assets |
| 11 | (gross profit + extraordinary items + financial expenses) / total assets |
| 12 | gross profit / short-term liabilities |
| 13 | (gross profit + depreciation) / sales |
| 14 | (gross profit + interest) / total assets |
| 15 | (total liabilities \* 365) / (gross profit + depreciation) |
| 16 | (gross profit + depreciation) / total liabilities |
| 17 | total assets / total liabilities |
| 18 | gross profit / total assets |
| 19 | gross profit / sales |
| 20 | (inventory \* 365) / sales |
| 21 | sales (n) / sales (n-1) |
| 22 | profit on operating activities / total assets |
| 23 | net profit / sales |
| 24 | gross profit (in 3 years) / total assets |
| 25 | (equity - share capital) / total assets |
| 26 | (net profit + depreciation) / total liabilities |
| 27 | profit on operating activities / financial expenses |
| 28 | working capital / fixed assets |
| 29 | logarithm of total assets |
| 30 | (total liabilities - cash) / sales |
| 31 | (gross profit + interest) / sales |
| 32 | (current liabilities \* 365) / cost of products sold |
| 33 | operating expenses / short-term liabilities |
| 34 | operating expenses / total liabilities |
| 35 | profit on sales / total assets |
| 36 | total sales / total assets |
| 37 | (current assets - inventories) / long-term liabilities |
| 38 | constant capital / total assets |
| 39 | profit on sales / sales |
| 40 | (current assets - inventory - receivables) / short-term liabilities |
| 41 | total liabilities / ((profit on operating activities + depreciation) \* (12/365)) |
| 42 | profit on operating activities / sales |
| 43 | rotation receivables + inventory turnover in days |
| 44 | (receivables \* 365) / sales |
| 45 | net profit / inventory |
| 46 | (current assets - inventory) / short-term liabilities |
| 47 | (inventory \* 365) / cost of products sold |
| 48 | EBITDA (profit on operating activities - depreciation) / total assets |
| 49 | EBITDA (profit on operating activities - depreciation) / sales |
| 50 | current assets / total liabilities |
| 51 | short-term liabilities / total assets |
| 52 | (short-term liabilities \* 365) / cost of products sold) |
| 53 | equity / fixed assets |
| 54 | constant capital / fixed assets |
| 55 | working capital |
| 56 | (sales - cost of products sold) / sales |
| 57 | (current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation) |
| 58 | total costs /total sales |
| 59 | long-term liabilities / equity |
| 60 | sales / inventory |
| 61 | sales / receivables |
| 62 | (short-term liabilities \*365) / sales |
| 63 | sales / short-term liabilities |
| 64 | sales / fixed assets |

1. **EXPLORATORY ANALYSIS**

This data set contains 43,405 data samples with 65 columns, of mostly ratio data types.

**Table 3: Data Types**

|  |  |
| --- | --- |
| *Variable Name* | *Data Type* |
| 1 | Ratio |
| 2 | Ratio |
| 3 | Ratio |
| 4 | Ratio |
| 5 | Ratio |
| 6 | Ratio |
| 7 | Ratio |
| 8 | Ratio |
| 9 | Ratio |
| 10 | Ratio |
| 11 | Ratio |
| 12 | Ratio |
| 13 | Ratio |
| 14 | Ratio |
| 15 | Ratio |
| 16 | Ratio |
| 17 | Ratio |
| 18 | Ratio |
| 19 | Ratio |
| 20 | Ratio |
| 21 | Ratio |
| 22 | Ratio |
| 23 | Ratio |
| 24 | Ratio |
| 25 | Ratio |
| 26 | Ratio |
| 27 | Ratio |
| 28 | Ratio |
| 29 | Ratio |
| 30 | Ratio |
| 31 | Ratio |
| 32 | Ratio |
| 33 | Ratio |
| 34 | Ratio |
| 35 | Ratio |
| 36 | Ratio |
| 37 | Ratio |
| 38 | Ratio |
| 39 | Ratio |
| 40 | Ratio |
| 41 | Ratio |
| 42 | Ratio |
| 43 | Ratio |
| 44 | Ratio |
| 45 | Ratio |
| 46 | Ratio |
| 47 | Ratio |
| 48 | Ratio |
| 49 | Ratio |
| 50 | Ratio |
| 51 | Ratio |
| 52 | Ratio |
| 53 | Ratio |
| 54 | Ratio |
| 55 | Ratio |
| 56 | Ratio |
| 57 | Ratio |
| 58 | Ratio |
| 59 | Ratio |
| 60 | Ratio |
| 61 | Ratio |
| 62 | Ratio |
| 63 | Ratio |
| 64 | Ratio |
| 65 | Nominal |

1. *Unusual Plots*



Figure 1 – Comparison of the means for bankrupt (1) and not bankrupt (0) companies regarding [equity / total assets]

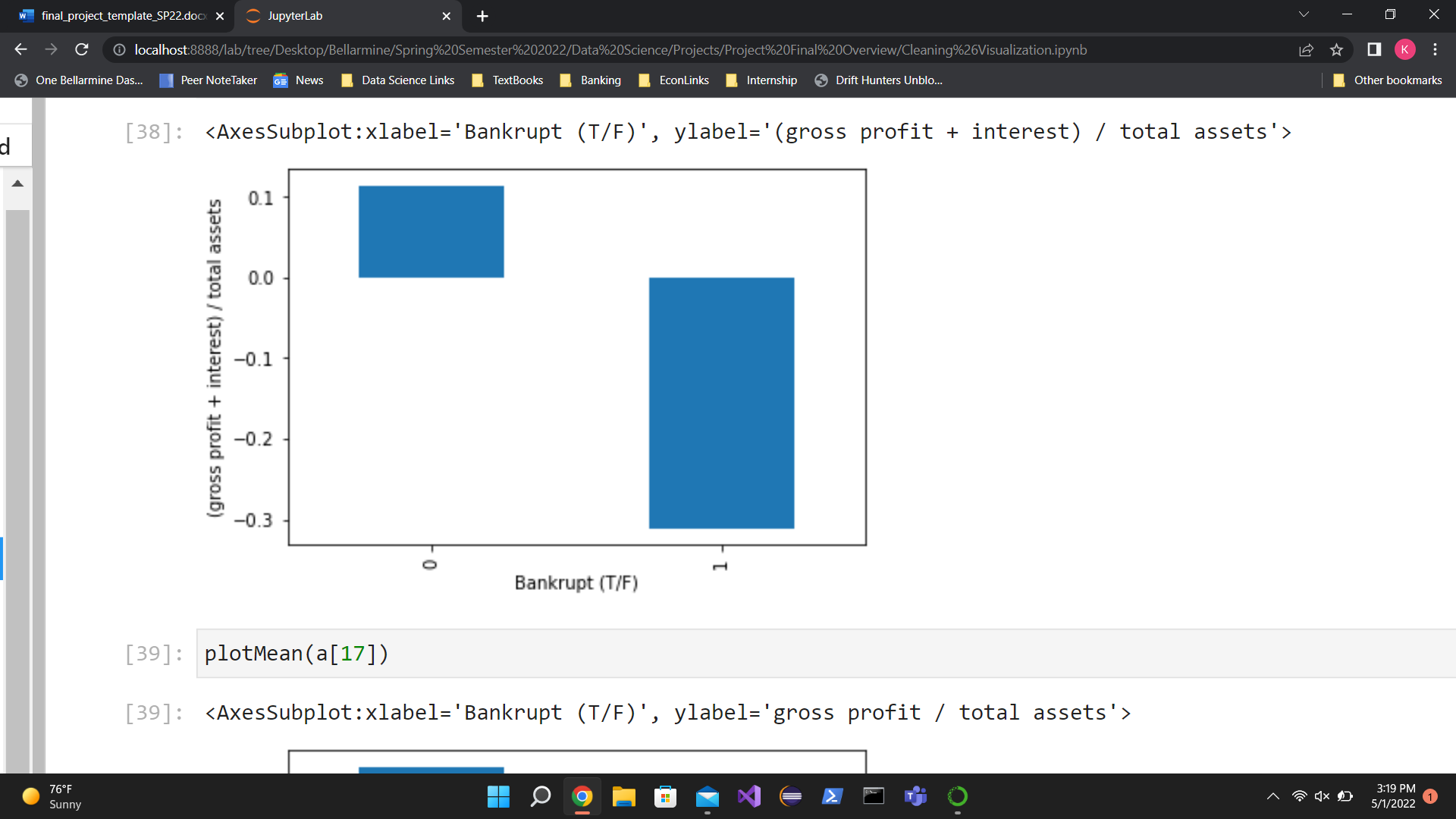


Figure 2 - Comparison of the means for bankrupt (1) and not bankrupt (0) companies regarding [(gross profit + interest) / total assets]

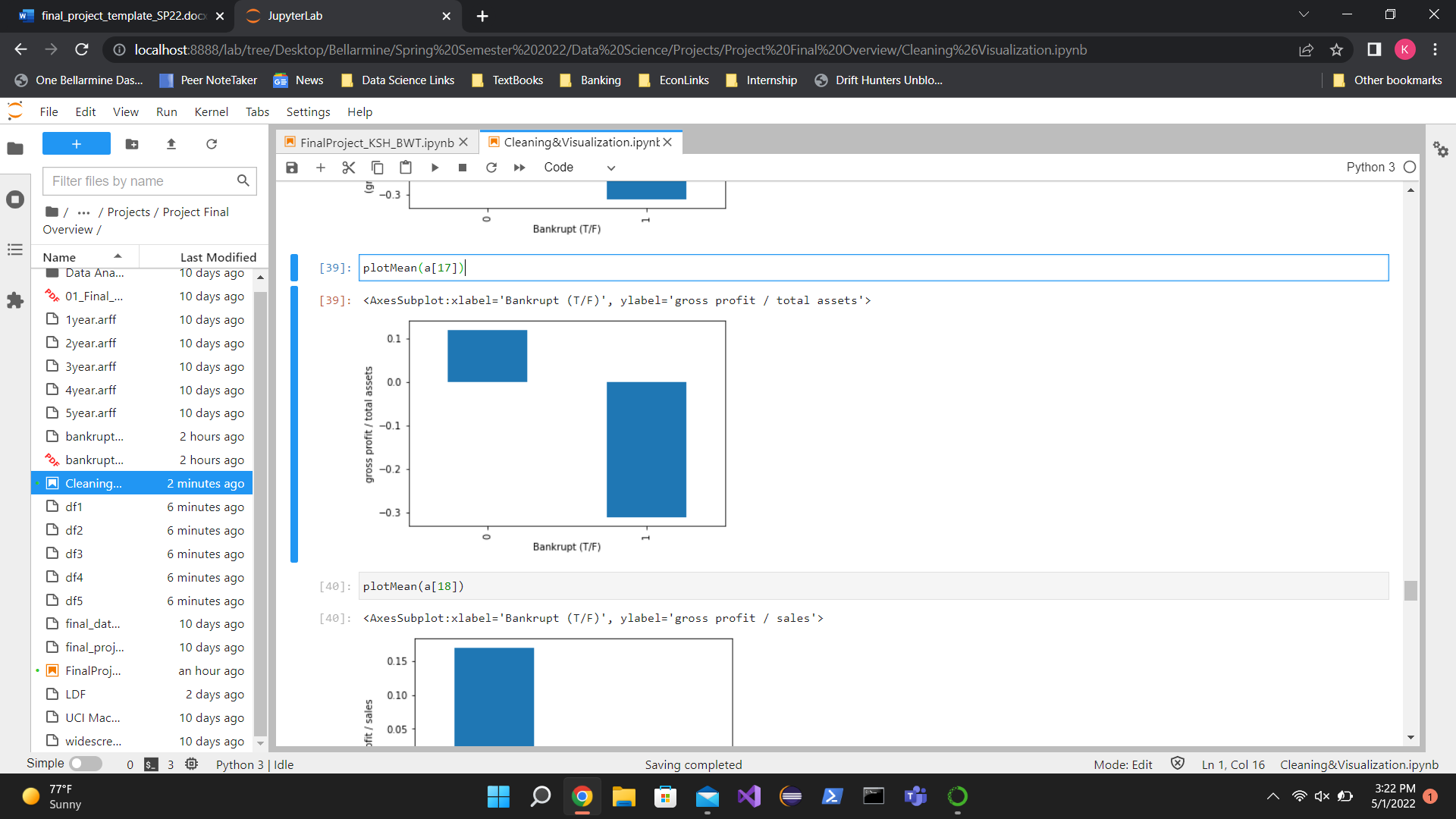


Figure 3 - Comparison of the means for bankrupt (1) and not bankrupt (0) companies regarding [gross profit / total assets]

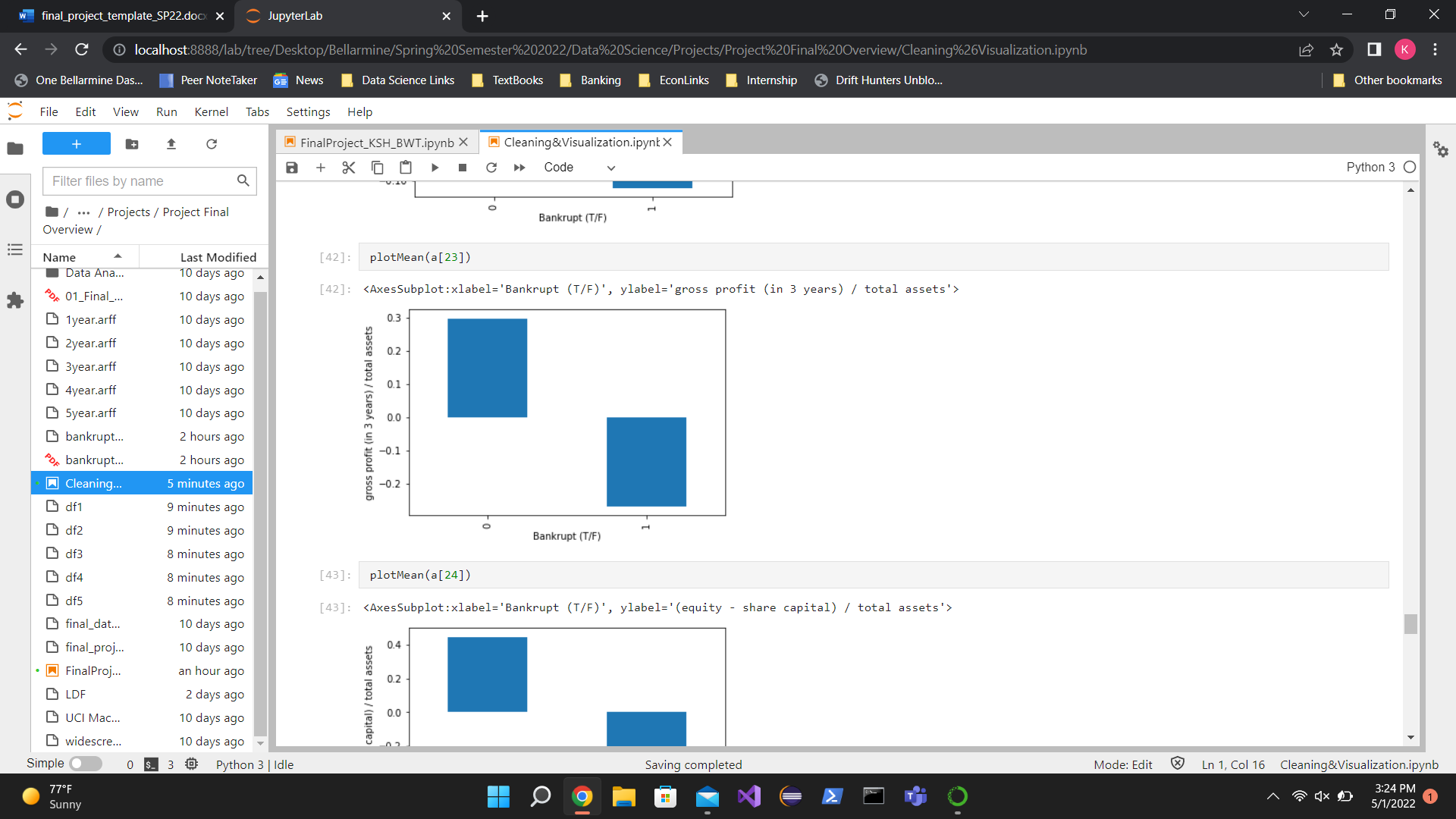


Figure 4 - Comparison of the means for bankrupt (1) and not bankrupt (0) companies regarding [gross profit (in 3 years) / total assets]

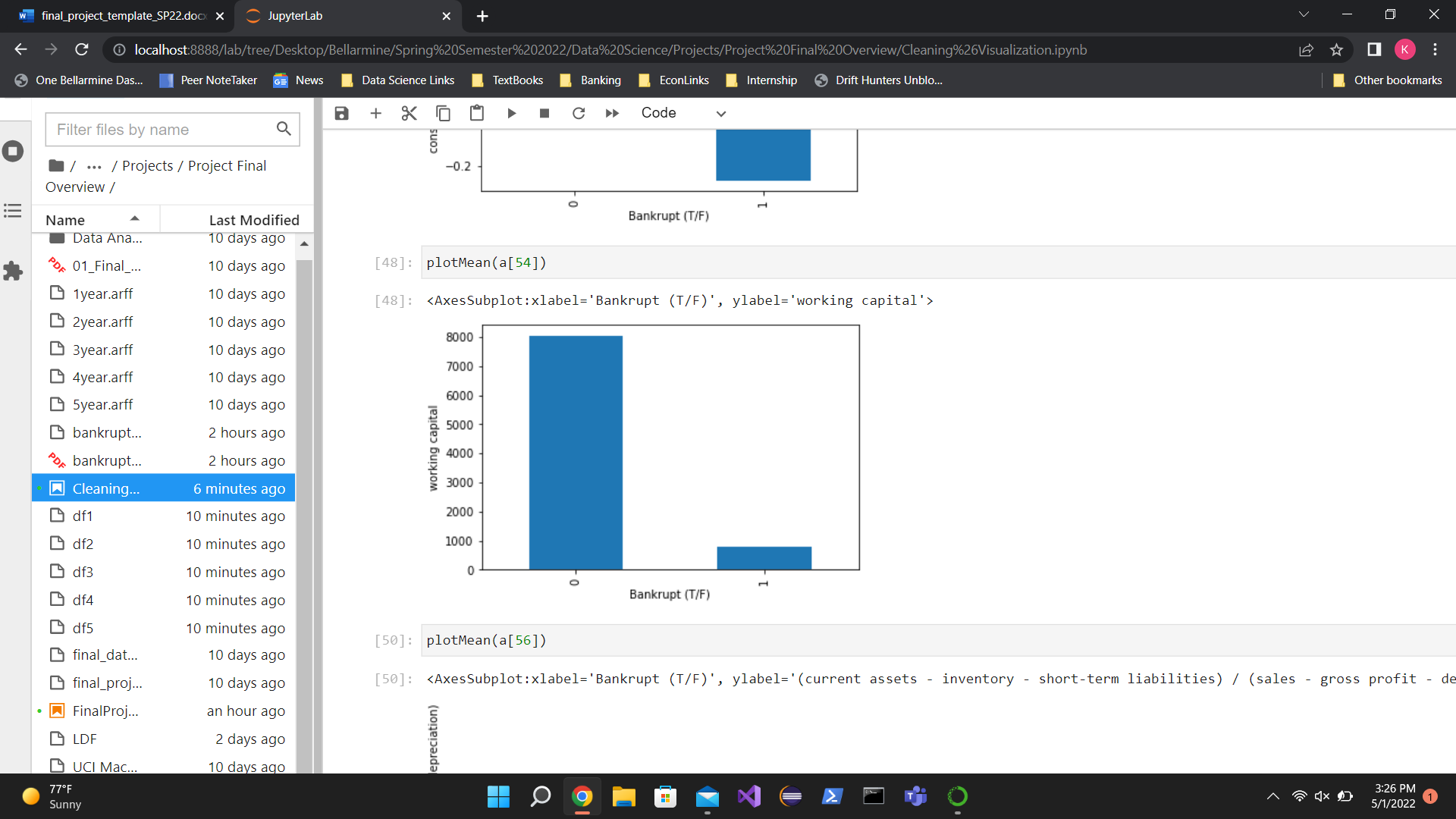


Figure 5 - Comparison of the means for bankrupt (1) and not bankrupt (0) companies regarding [working capital]

1. **METHODS**
   1. *Data Preparation*

The dataset we were given included five files with a file corresponding to a specific year of study. The files were given in a .arff file format and because of such, additional steps were required to transform them to a format usable in Python. For simplicity, during our data cleaning, we then renamed the columns 1-65, each column corresponding to an economic measure that is used in the bankruptcy analysis. Following this step, we addressed the missing data in each of the 5 data frames that we originally created. Our method of doing so was by filling in missing data with the mean of the corresponding column. Once this was finished, we then renamed our columns to the names of the economic measures used in the analysis for later use in the visualization phase of our project. After this, we then combined all 5 data frames into a larger data frame for our actual data analysis.

* 1. *Experimental Design*

When performing our experiment, our goal was to increase the recall of the bankrupt company category. This is because this category in the confusion matrix shows the number of companies that our models predicted to go bankrupt that did indeed go bankrupt.

Table X: Experiment Parameters

|  |  |
| --- | --- |
| **Experiment Number** | **Parameters** |
| 1 | Visualizing with Bar Graphs, comparing variables means of Bankrupt vs Non-Bankrupt |
| 2 | Logistic Regression: Single Variable   * 'total costs /total sales' * **Score: 0.9518258265176823** |
| 3 | Logistic Regression: Single Variable   * ‘retained earnings / total assets’ * **Score: 0.9518488653380947** |
| 4 | Logistic Regression: Single Variable   * 'working capital' * **Score: 0.9517336712360327** |
| 5 | Logistic Regression: Single Variable   * 'EBITDA (profit on operating activities - depreciation) / sales' * **Score: 0.9518258265176823** |
| 6 | Logistic Regression: Multiple Variables   * 'working capital', * 'retained earnings / total assets' * 'total costs /total sales' * **Score: 0.9517567100564451** |
| 7 | Decision Tree:   * Sample Split * **Test 1: 3** * Test 2: 5 |
| **8** | Decision Tree:   * Test Size * **Test 1: 0.80 Train, 0.20 Test** * Test 2: 0.75 Train, 0.25 Test |
| 9 | Decision Tree:   * Depth * **Test 1: 4** * Test 2: 7 |
| 10 | Random Forest:   * Test Size * Test 1: 0.90 Train, 0.10 Test * **Test 2: 0.85 Train, 0.15 Test** |
| 11 | Random Forest:   * Number of Estimators * **Test 1: 18 Estimators** * Test 2: 10 Estimators |
| 12 | Random Forest:   * Type (Entropy vs Gini) * **Test 1: Entropy** * Test 2: Gini |

**Bold Text** means Best Result

* 1. *Tools Used*

The following tools were used for this analysis: Python v3.10.0 running the Anaconda 2.0.3 environment for Windows (Lenovo/Dell) computer and was used for all analysis and implementation. In addition to base Python, the following libraries were also used: Pandas 1.4.2, Scipy 1.8.0, Seaborn v0.11.2, NumPy 1.19.2, Matplotlib 3.5.1, Sklearn 1.0.2, Graphviz 3.0.0, and Pydot 1.4.2. These packages were used for visualization and for their included machine learning techniques and algorithms.

1. **RESULTS**
   1. *Classification Measures*

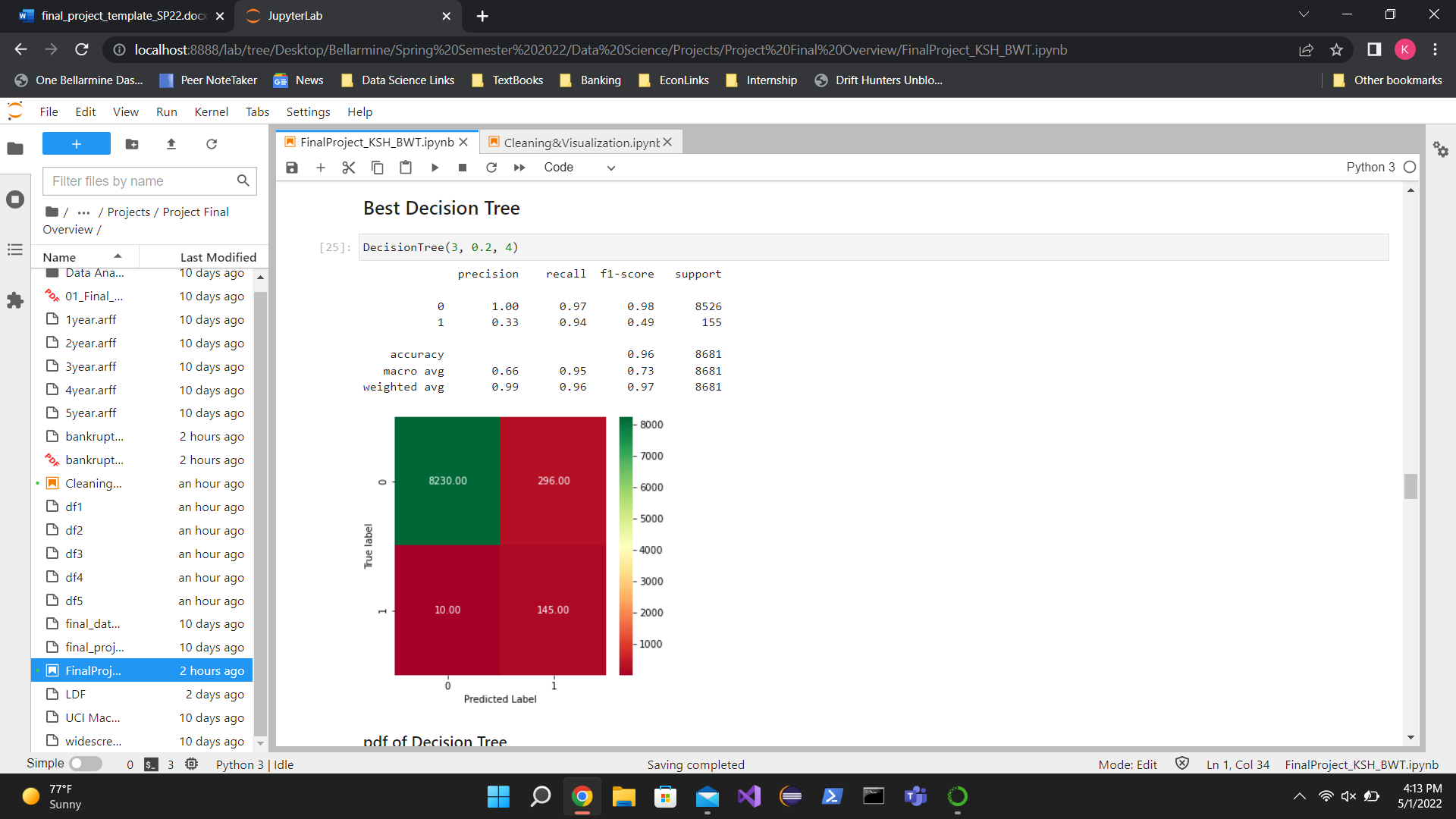


Fig. 6 – Best Decision Tree Confusion Matrix (Sample Split – 3, Test Size - 0.2, Depth - 4 ) and Classification Report

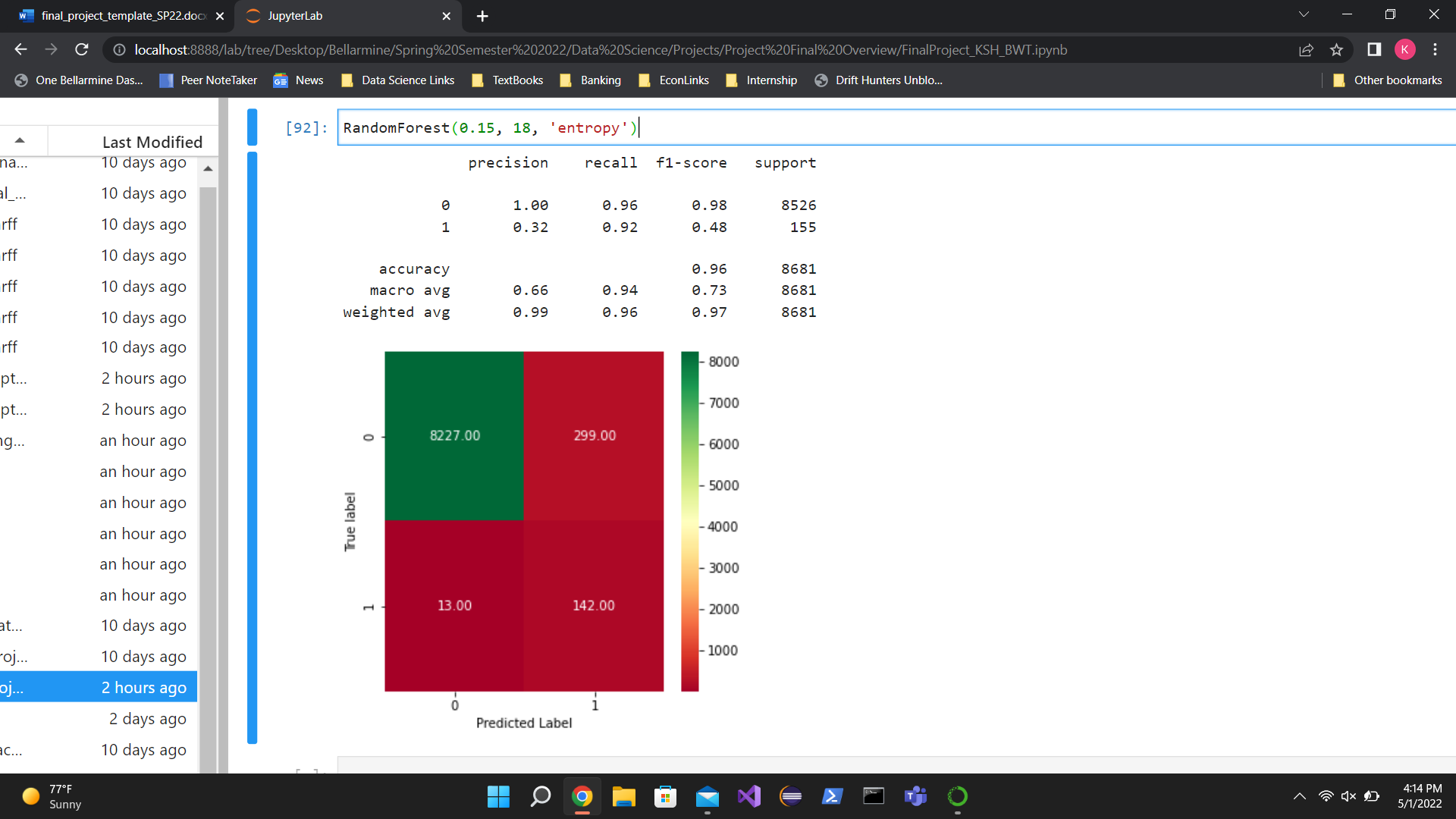


Fig. 7 – Best Random Forest Confusion Matrix (Test Size – 0.15, Number of Estimators – 18, Type - Entropy) and Classification Report

* 1. *Discussion of Results*

Our best model was the decision tree as it had a 96% overall accuracy. This is likely a result of utilizing more variables in making decisions, especially those that cause a split in the decision tree. Our worst model was logistic regression, but only slightly as it had an overall accuracy of 95%. It may be accurate with this number of variables but as more variables are added it may become less accurate.

* 1. *Comparison of Models*

The best classification method for our data was found by using decision trees, with a 96% overall accuracy, and a 94% recall for the bankrupt category. Our other methods of machine learning had similar overall accuracy scores, 96% for both the logistic regression and random forest but did not compare as favorably for the bankruptcy recall accuracy.

* 1. *Problems Encountered*

Through the duration of the project, we encountered a variety of problems, not unusual for a project involving data manipulation. We encountered our first issue right from the beginning. The data was given in a .arff format, something that we had not encountered before, so it required some research and trial and error to find a method that allowed us to convert that data into a usable format. Then, the classification column was given in bytes and required additional manipulation to both get over the initial confusion and use in the analysis. Next, in the data cleaning phase, we removed all rows that had missing data in any of their columns. However, we later realized that by removing all these rows, there were too few bankrupt companies left represented, so we then had to fill the missing data with the mean of the corresponding column. Then, our next set of problems occurred while using the machine learning algorithms. We sometimes struggled to figure out the syntax of the algorithms and often had to do some outside research to understand how to obtain the result we wanted. Finally, in most of the algorithms, we wanted to compare different trials and briefly struggled with finding an efficient way to go about doing that but quickly realized we could build a function to perform this, which we did.

* 1. *Limitations of Implementation*

The models that we used were limited in the fact that they both used similar classification methods. Random forest and decision tree classify data in similar ways and may not provide the best accuracy. KNN may have been a better predictor model because there was often a clear distinction between bankrupt and non-bankrupt companies. KNN may have been able to exploit this difference to produce a better accuracy score and thus a better model.

* 1. *Improvements/Future Work*

Our model could be improved in the future by experimenting with different machine learning algorithms, specifically KNN, to determine if other models performed better than the models we used in our predictions. Additionally, we could have further analyzed the correlations of the 64 variables and removed any that were not especially significant for an easier analysis process. Finally, there were a disproportionate number of bankrupt companies as compared to non-bankrupt companies which led to an accuracy imbalance. The models may have functioned better if there was a more proportionate split between bankrupt and non-bankrupt companies.

1. **CONCLUSION**

The bankruptcy data representing Polish companies over a period of a few years allowed us to make predictions about whether a company would go bankrupt or not over the forecasting period. Our models had very high accuracy scores, however, they were extremely overwhelmed by the amount of non-bankrupt data. The disproportionality of the data may have caused a misleading accuracy score and should be repeated with a more proportionate split of data for better and more accurate results.

**REFERENCES**

*UCI Machine Learning Repository: Polish Companies Bankruptcy Data Data Set*, https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data.

*Division of Labor*

Since there is currently no method to share a python script in real-time the project had to subsequently be divided to increase overall efficiency. Brandon focused on: the substitution of the mean of each column to replace the missing data (NAs in this case), then the renaming of the columns to the economic measures of corporate success, then highlighting the unusual distributions of the columns after concatenating the 5 data frames into a larger data frame, then built a correlation matrix and corresponding heatmap which turned out to not be useful, then built a decision tree for the data set. Keegan focused on: converting the .arff file to a data frame, then initially naming the columns 1-65 for simplicity while manipulating the data, then created a function to replace the values in the bankrupt column from their form in bytes to simply just 0s and 1s, then saved the concatenated data frame out as a .csv file for use in the machine learning notebook of our project, then he created a function for the mean comparison plots, then he performed the logistic regression and associated tests, followed by the random forest method and its associated tests.