**ML Assignment**

Sri Venkatesh Subramaniam, Ankit Sharma

School of Business, St. Lawrence College

ADMN5006- Financial Analytics

Professor Maverick Ramsaran

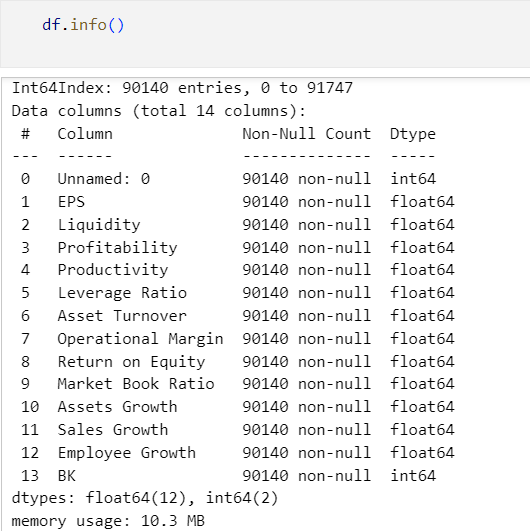
November 19, 2023

**Section 1: Introduction and Business Problem**

The primary objective of this analysis is to develop a predictive model for bankruptcy using financial data. The goal is to accurately predict whether a company will go bankrupt based on various financial indicators. This is crucial for investors, creditors, and other stakeholders to make informed decisions and manage risks effectively.

**Section 2: Exploratory Data Analysis (EDA)**

The dataset consists of financial data from various companies.



Initial exploration was performed for descriptive statistics to understand the nature of the numerical data patterns and trends in the data, as well as some issues like missing values.

Value Counts revealed the number of missing values in each column.

A screenshot of a computer

Description automatically generated

**Data Correlations:**

Both Pearson and Spearman correlation coefficients were used in our analysis to understand the relationships between variables in your dataset. Each of these methods has its own strengths and offers unique insights, particularly in a context like bankruptcy prediction where understanding these relationships is crucial.

In the context of bankruptcy prediction, where the dataset likely comprises various types of financial metrics, employing both Pearson and Spearman correlations provides a more thorough analysis. While Pearson's correlation helps in understanding linear relationships, Spearman's correlation complements this by capturing non-linear, monotonic relationships. This dual approach ensures a more comprehensive understanding of how different financial indicators relate to each other and to the likelihood of bankruptcy, which can be crucial for building effective predictive models.

Visualizations, including heatmaps, were employed to identify correlations between different financial variables. The correlations performed were both Pearson correlation and Spearman correlation.

The Correlations were also filtered for over .5 and under -.5

A diagram of a graph

Description automatically generated with medium confidence

A graph of a number of red and blue squares

Description automatically generated A graph with numbers and lines

Description automatically generated

A chart with numbers and symbols

Description automatically generated with medium confidence

A graph with numbers and text

Description automatically generated with medium confidenceA graph with text on it

Description automatically generated with medium confidence

Correlation matrix depicted as a heatmap, which visually represents the Pearson correlation coefficients between different financial metrics. The matrix includes various metrics such as Earnings Per Share (EPS), Liquidity, Profitability, Productivity, Leverage Ratio, Asset Turnover, Operational Margin, Return on Equity, Market Book Ratio, Assets Growth, Sales Growth, Employee Growth, and BK(Bankruptcy).

Here's a general interpretation of the correlations presented:

1. **EPS (Earnings Per Share):**
   * Strongly positively correlated with itself (as are all variables, hence the 1.0 in the diagonal).
   * No or negligible correlation with other metrics as indicated by 0.0 values.
2. **Liquidity:**
   * Moderately positively correlated with Profitability (0.6), suggesting that more profitable companies tend to have better liquidity.
   * Weakly to moderately positively correlated with Productivity (0.5).
3. **Profitability:**
   * Similarly, has a moderate positive correlation with Liquidity and Productivity.
4. **Productivity:**
   * Shows a moderate positive correlation with Liquidity (0.5) and Profitability (0.6), suggesting that higher productivity might be associated with better liquidity and profitability.
5. **Leverage Ratio:**
   * Shows very weak correlation with most metrics, indicating leverage is not linearly related to these other financial measures.
6. **Asset Turnover:**
   * Has a perfect correlation with itself and negligible correlations with most other metrics.
7. **Operational Margin, Return on Equity, Market Book Ratio, Assets Growth, Sales Growth, and Employee Growth:**
   * These metrics show very weak correlations with most other variables, as indicated by values close to 0.0.
8. **BK:**
   * Bankruptcy shows no correlation with other metrics.

The red squares indicate a positive correlation where one variable is likely to increase as the other increases. The shades of red indicate the strength of this positive correlation, with darker shades representing stronger correlations. Cells with a correlation of 1.0 are dark red because a variable is always perfectly correlated with itself. Most other cells show correlations close to 0, indicating very little to no linear relationship between the variables.

**Section 3: Data Pre-Processing**

Data processing involved several steps:

**Removing rows of Missing Values**

Rows with four or more missing values were dropped. As There were a total of 12 independent variable columns, 4 columns of missing values would mean 1/3 rd. of the data, so we removed the rows with 4 or more missing values. 1,123 rows accounted for rows with 4 or more missing values.

**Replacing missing values**

The remaining missing values were replaced with median values, as this was deemed the most appropriate method for this dataset. Outliers have relatively less effect on Median compared to mean.

**Outlier Detection and Removal**

Outliers were identified and removed using the Z-score method. This step was crucial to prevent extreme values from skewing the results and to train a better model. Z-score standardizes the dataset, meaning it puts the data points on a common scale, allowing for a straightforward comparison between different features. This is particularly useful in your case where different financial metrics might be on different scales.

Reason for choosing Z-Score:

Z-score is particularly effective when the data has a roughly symmetrical distribution. In financial datasets, after log transformation or other normalizing steps, using Z-scores becomes a robust method for identifying outliers.

In financial data analysis, outliers can skew the results, leading to biased predictions. By removing these outliers, models like Logistic Regression, Naive Bayes, and KNN can make more accurate predictions, as they might otherwise be overly influenced by extreme values.

Downside of using Z-Score: One downside of removing outliers is the potential loss of valuable information, especially if the outliers are not errors but represent rare, significant events. In financial data, extreme values could sometimes be indicators of important trends or warning signs.

**Data Normalization / standardization**

The data was scaled using Min-Max scaling, excluding the 'BK' column due to its categorical nature, and to ensure that all variables contributed equally to the analysis.

**Data Balance**

The data was critically imbalanced with 90,140 rows of value “1” and 543 rows of value “0”. The decision to use both SMOTE (Synthetic Minority Over-sampling Technique) for upscaling and down-sampling techniques for balancing the dataset in our bankruptcy prediction models is a strategic approach to tackle the challenges posed by imbalanced data.

We used SMOTE as Unlike simple over-sampling (replicating minority class samples), SMOTE generates new, synthetic samples that are similar yet slightly different, reducing the risk of overfitting.

We also decided to use down-sampling as It forces the model to focus on a smaller, more balanced dataset, potentially leading to more robust learning. While SMOTE introduces synthetic samples, down-sampling only uses actual, observed data, avoiding any biases or artifacts that might come from synthetic data. We believe that down-sampling was the best approach to the data imbalance problem. Using both techniques provides insights into how different balancing methods affect model performance. It allows for a more comprehensive evaluation of the models under different data conditions. Some models might perform better with up-sampled data, while others might benefit from down-sampling. Experimenting with both allows you to understand the sensitivity of different models to the way the data is balanced.

**Section 4: Model Development**

At First Logistic Regression was performed with default parameters. The team then decided to use Down-Sampling for testing the model as well.

Models chosen:

1. Logistic Regression
2. Naïve Bayes
3. K Nearest Neighbours

**Logistic Regression**

1. **Binary Classification**: Logistic Regression is a widely used algorithm for binary classification problems, such as predicting bankruptcy (a binary outcome: yes or no, represented by 1 or 0 in your dataset).
2. **Interpretability**: It provides coefficients that give insights into the relationship between each feature and the likelihood of bankruptcy, which is valuable for understanding which factors are most predictive.
3. **Performance**: It performs well when the relationship between the independent variables and the log-odds of the dependent variable is linear.
4. **Baseline Model**: Logistic Regression is often used as a baseline model for binary classification problems due to its simplicity and efficiency.

**Naive Bayes**

1. **Probabilistic Approach**: Naive Bayes is based on applying Bayes' theorem with the assumption of independence between every pair of features. This probabilistic approach is suitable for classification tasks.
2. **Efficiency**: It is known for being extremely fast and efficient, especially when dealing with a large number of features.
3. **Good Baseline**: Naive Bayes can serve as a good baseline model for classification problems and is particularly known for its effectiveness in text classification.
4. **Performance with Small Data**: Even with a smaller dataset, Naive Bayes can perform quite well and is not as prone to overfitting as some other models.

**K-Nearest Neighbors (KNN)**

1. **Non-Parametric and Lazy Learning**: KNN is a non-parametric method, meaning it makes no assumptions about the underlying data distribution. This flexibility allows it to adapt to real-time data.
2. **Versatility**: It can be a good choice for classification tasks where the decision boundary is not linear.
3. **Intuitiveness**: The concept of KNN is simple and intuitive—predict the class of a data point by looking at the 'K' closest labeled data points and taking a majority vote.
4. **Effectiveness in Certain Scenarios**: KNN can be very effective if the dataset has enough representative data for each class, and the distance metric is meaningful.

**Section 5: Model Comparison**

The models were compared with the help of confusion matrix and Classification reports.

Here is a table of different scores based on sampling method and different models.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Precision (Class 0) | Precision (Class 1) | Recall (Class 0) | Recall (Class 1) | F1-Score (Class 0) | F1-Score (Class 1) | Macro Avg F1-Score | Weighted Avg F1-Score |
| Logistic Regression with Up-Sampling | 1.00 | 0.01 | 0.71 | 0.50 | 0.83 | 0.02 | 0.42 | 0.82 |
| Logistic Regression with Down-Sampling | 0.99 | 0.01 | 0.69 | 0.45 | 0.81 | 0.02 | 0.41 | 0.81 |
| Naive Bayes with Up-Sampling | 1.00 | 0.01 | 0.17 | 0.91 | 0.29 | 0.01 | 0.15 | 0.29 |
| Naive Bayes with Down-Sampling | 0.99 | 0.03 | 0.94 | 0.24 | 0.97 | 0.05 | 0.51 | 0.96 |
| KNN with Up-Sampling | 1.00 | 0.04 | 0.94 | 0.32 | 0.97 | 0.06 | 0.52 | 0.96 |
| KNN with Down-Sampling | 1.00 | 0.02 | 0.81 | 0.61 | 0.90 | 0.04 | 0.47 | 0.89 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Macro Avg Precision | Weighted Avg Precision | Macro Avg Recall | Weighted Avg Recall |
| Logistic Regression with Up-Sampling | 0.50 | 0.99 | 0.60 | 0.70 |
| Logistic Regression with Down-Sampling | 0.50 | 0.99 | 0.57 | 0.68 |
| Naive Bayes with Up-Sampling | 0.50 | 0.99 | 0.54 | 0.17 |
| Naive Bayes with Down-Sampling | 0.51 | 0.99 | 0.59 | 0.94 |
| KNN with Up-Sampling | 0.52 | 0.99 | 0.63 | 0.94 |
| KNN with Down-Sampling | 0.51 | 0.99 | 0.71 | 0.81 |

**Hyper Parameter Tuning**

We used grid search for hyper parameter tuning. We used KNN model as Hyperparameter tuning for Naive Bayes models is somewhat limited compared to other machine learning algorithms because Naive Bayes models, by their nature, have very few hyperparameters to tune.

Our best KNN model had k =9, weights = ‘distance’ and metric = ‘manhattan’ as the parameters.

This model was trained using the down-sampled dataset.

**knn\_\_metric: 'manhattan'**

**Purpose**: This hyperparameter defines the distance metric used to measure the closeness between data points.

**Manhattan Distance**: Also known as the L1 distance, the Manhattan distance calculates the absolute differences between the coordinates of a pair of points. The formula is:

D(x,y)=∑i=1n∣xi−yi∣D(x,y)=∑i=1n​∣xi​−yi​∣

where xx and yy are two points in an n-dimensional space.

**Why Choose Manhattan Distance**:

* It is particularly effective in high-dimensional spaces.
* Can be more robust to outliers compared to Euclidean distance.
* Suitable for cases where the 'grid-like' difference matters, as in urban block distances.

**knn\_\_n\_neighbors: 9**

**Purpose**: This defines the number of nearest neighbors to consider when making a prediction.

**Why 9 Neighbors?**:

* The choice of 9 neighbors suggests a model that considers a relatively broader neighborhood, which can be useful in avoiding overfitting to the training data.
* An odd number helps in making a clear decision in binary classification tasks, as it avoids ties (e.g., an equal number of neighbors in each class).
* The specific value (like 9) would typically be reached after experimenting with different values and selecting the one that performs best during cross-validation.

**knn\_\_weights: 'distance'**

**Purpose**: This parameter determines the weight given to the neighbors based on their distance.

**Distance-Weighted Neighbors**:

* When set to 'distance', nearer neighbors have a greater influence on the prediction than those further away.
* This approach can yield better results, especially in cases where the closest neighbors might be more similar to the query point than distant ones.
* It helps in reducing the impact of a majority class in an imbalanced dataset, as closer minority class neighbors will have a higher influence.

**Section 6: Conclusion**

**Our choice for the best model:**

We are comparing the effectiveness of a model with the Recall score for Class – 1. The outcome of wrongly predicting a business might go bankrupt is not as significant as the mistake of predicting that the business wouldn’t file for bankruptcy, but it does.

When the model would predict a company would go bankrupt, the stakeholders can be more aware and careful with the finances, even though it might not be bankrupt. But the Reverse of predicting that the company wouldn’t be bankrupt and, it might be is that it would be devastating to all the stake holders.

Therefore, we have picked the model having the best Recall score for Class 1, even if it might have a lower macro and weighted average Recall score.

We see that the Naïve Bayes with down sampling and KNN have higher Macro average and weighted average for recall, but we believe we must use the model with highest Recall for a particular class – class 1 (Bankrupt) than higher average recall, due to the consequences of having false negative. The improvement for the model would be to primarily to focus on reducing the false negatives.

Given time constraints, compute power and other resources we couldn’t implement SVM and random forests, they might’ve given better results.

**Risk Assessment and Management:**

* Lending Institutions: Banks and other lending institutions can use the model to assess the risk of loan defaults by evaluating the financial health of borrowing entities. This can inform credit decisions, loan pricing, and conditions.
* Investment Decisions: Investors can utilize the model to identify potentially risky investments and adjust their portfolios accordingly to mitigate losses from bankruptcies.

**Results and Recommendations:** The Naïve Bayes model showed promising results in predicting bankruptcy based on class 1 recall rate. However, given the complexity of financial data, a combination of models or more advanced techniques might yield better results.

**Future Improvements:** Future improvements could include the use of PCA – Principal component analysis and more sophisticated models like Random Forest or Neural Networks, which might capture complex patterns in the data more effectively. Additionally, incorporating macroeconomic factors and industry-specific variables could enhance the model's predictive power.