Report

Topic:

"Bankruptcy Prediction for Public Companies: Identifying Key Predictive Features from NYSE and NASDAQ Data."

Business Problem:

Public companies facing bankruptcy pose significant risks to investors, creditors, and the overall market. This project seeks to solve the problem of early bankruptcy prediction by identifying the most important financial and operational features that signal impending bankruptcy. The goal is to improve risk assessment models for investors and financial institutions by answering key research questions such as: "Which financial or operational features are the strongest predictors of bankruptcy?" and "How early can we predict bankruptcy with high accuracy?"

Background/History:

The origin of the word "bankruptcy" is widely believed to derive from the Latin words bancus (bench or table) and ruptus (broken). In Medieval Italy, when a banker couldn't fulfill their financial obligations, their bench would be broken as a symbol of their failure, leading to the term banco rotto ("broken bank"). An alternative theory suggests it may come from the French phrase banque route, referring to a table left abandoned when a banker fled with entrusted funds.

In the U.S., modern bankruptcy laws focus more on helping debtors reorganize rather than punishing them. The Bankruptcy Act of 1898 first allowed distressed companies to be shielded from creditors. During the Great Depression, further legislation solidified the idea of offering a "fresh start" for debtors, as confirmed by a 1934 Supreme Court decision. According to the SEC, bankruptcy in the U.S. can occur under two conditions: Chapter 11, where a company reorganizes while continuing operations under court supervision, or Chapter 7, where the company ceases operations and liquidates its assets.

Data Explanation:

The resulting dataset was obtained from Kaggle. The owner/author of the dataset is Utkarsh Singh. He collected the data from companies listed on the New York Stock Exchange and NASDAQ. The dataset comprises accounting data from 8,262 distinct companies recorded during the period spanning from 1999 to 2018. Overall, the resulting dataset comprises a total of 78,682 observations. To make the data easier to work with, I removed both the “company name” and the “year” columns. I then converted the values in the “status_label” column to numerical values. Alive = 0 and failed = 1. These were the only changes made to the data.

Methods:

First I split the data into features (X) and target (y). Then, I split the dataset into training and testing sets using an 80/20 split. I then initialized a LightGBMClassifier and trained it on the training data. The model is tested on the test data, and evaluation metrics like accuracy, confusion matrix, and classification report are printed.

Report

Analysis:

The model achieved an overall accuracy of 93.44%, meaning it correctly predicted the bankruptcy status for 93.44% of the companies in the test set. The confusion matrix shows that the model correctly classified 14,658 "alive" companies (true negatives) and 46 "failed" companies (true positives). However, it misclassified 1,022 "failed" companies as "alive" (false negatives), which indicates the model struggles with identifying bankrupt companies. This is further reflected in the classification report, where the model shows a high precision of 0.93 for predicting non-bankrupt companies (class 0), but very low recall (0.04) for predicting bankrupt companies (class 1). The low recall means the model missed most of the bankruptcies, capturing only a small fraction. While the weighted averages are strong, the performance for the minority class ("failed" companies) suggests that the model might be biased toward predicting companies as "alive."

Conclusion:

In conclusion, the model demonstrated a strong overall accuracy of 93.44% in predicting the bankruptcy status of companies within the test set. While it effectively identified a majority of non-bankrupt companies, as indicated by the high true negative count, its performance in detecting bankrupt companies was notably weaker. The confusion matrix highlighted a concerning number of false negatives, with 1,022 bankrupt companies misclassified as "alive." This low recall rate of 0.04 for the "failed" class suggests that the model is biased towards predicting companies as non-bankrupt, potentially limiting its utility in practical applications.

Feature importance analysis revealed that several financial metrics, such as retained earnings (X15), inventory (X5), and market value (X8), were the most significant predictors of bankruptcy. However, the negligible importance of some features, particularly total revenue (X16), raises questions about their relevance in this predictive context.

Assumptions:

In this project, several assumptions were made regarding the dataset and the bankruptcy prediction model. It was assumed that the financial metrics included in the dataset are sufficient to capture the complexities of a company's operational health. Additionally, the model assumes that past financial performance is a reliable indicator of future outcomes, particularly regarding bankruptcy risk. The binary classification of the status label as "alive" or "failed" also assumes that these categories adequately represent the range of a company's financial status. Lastly, it is assumed that the historical data reflects relevant economic conditions, making it applicable to current market scenarios.

Limitations:

This project has inherent limitations that may impact the robustness of its findings. One significant limitation is the potential for class imbalance, as the dataset may contain more instances of non-bankrupt companies than bankrupt ones, which could skew the model's performance metrics. Additionally, the dataset is confined to a specific time frame (1999-2018) and geographical context (NYSE and NASDAQ), limiting the generalizability of the results to other markets or more recent

financial conditions. Moreover, the model does not account for qualitative factors, such as market sentiment or management decisions, which can also influence a company's bankruptcy risk.

Challenges:

Throughout the project, several challenges were encountered. One primary challenge was dealing with the class imbalance in the dataset, which complicated the model's ability to accurately predict the minority class (bankrupt companies). This required careful consideration of evaluation metrics beyond accuracy, such as precision and recall. Additionally, ensuring that the selected features effectively captured the nuances of bankruptcy risk posed a challenge, as some features had negligible importance in the final model. Another challenge was the interpretability of the model; while LightGBM provides feature importance, translating these insights into actionable strategies for stakeholders can be complex.

Future Uses/Additional Applications:

The insights gained from this project can be applied in various ways beyond predicting bankruptcy. Financial institutions could leverage the model to enhance their risk assessment frameworks, helping them identify at-risk clients and adjust lending practices accordingly. Additionally, the methodology could be adapted for different industries or markets, allowing for broader applications in financial forecasting. Beyond traditional banking, this predictive framework could also inform investment strategies and portfolio management, providing investors with a tool to evaluate the long-term viability of public companies.

Recommendations:

Moving forward, addressing the model's bias toward the majority class and enhancing its ability to identify bankrupt companies will be critical for improving its effectiveness in real-world applications. Strategies such as resampling techniques or employing different thresholds may be beneficial in mitigating these issues and optimizing overall predictive performance.

Implementation Plan:

To implement the recommendations for improving the model, hyperparameter tuning should be conducted to optimize model performance further. The model should then be retrained and reevaluated using the adjusted dataset, focusing on improving recall for the bankrupt class. Continuous monitoring of model performance over time will be necessary to ensure it adapts to changing economic conditions. Lastly, stakeholders should be trained on how to utilize the model effectively within their decision-making processes.

Ethical Assessment:

An ethical assessment of this project involves considering the implications of using automated models for bankruptcy prediction. It's essential to ensure that the model does not inadvertently discriminate against specific groups, such as smaller businesses that may have different risk profiles. Transparency in how predictions are made is crucial, as stakeholders need to understand the model's limitations and the data it relies upon. Additionally, using such predictive models responsibly is important; companies and investors must ensure that decisions based on these

predictions do not lead to unfair practices, such as disproportionately withdrawing support from vulnerable companies without exploring alternatives.

References:

- US Company Bankruptcy Prediction Dataset. (2023, May 27). Kaggle. https://www.kaggle.com/datasets/utkarshx27/american-companies-bankruptcy-prediction-dataset
- History of Bankruptcy — BankruptcyData. (n.d.). BankruptcyData. https://www.bankruptcydata.com/a-history-of-bankruptcy

Illustrations:

```
Accuracy: 0.9343585181419585
Confusion Matrix:
[[14658    11]
 [ 1022    46]]
Classification Report:
              precision    recall  f1-score   support

           0       0.93      1.00      0.97     14669
           1       0.81      0.04      0.08      1068

    accuracy                           0.93     15737
   macro avg       0.87      0.52      0.52     15737
weighted avg       0.93      0.93      0.91     15737
```

Fig 1

	Feature	Importance
14	X15	0.092667
4	X5	0.091667
7	X8	0.077000
2	X3	0.076667
6	X7	0.074667
1	X2	0.065667
0	X1	0.064000
16	X17	0.060000
12	X13	0.059000
10	X11	0.055667
17	X18	0.050667
13	X14	0.049667
9	X10	0.045667
5	X6	0.038333
3	X4	0.037333
11	X12	0.031000
8	X9	0.030333
15	X16	0.000000

Fig 2

Report

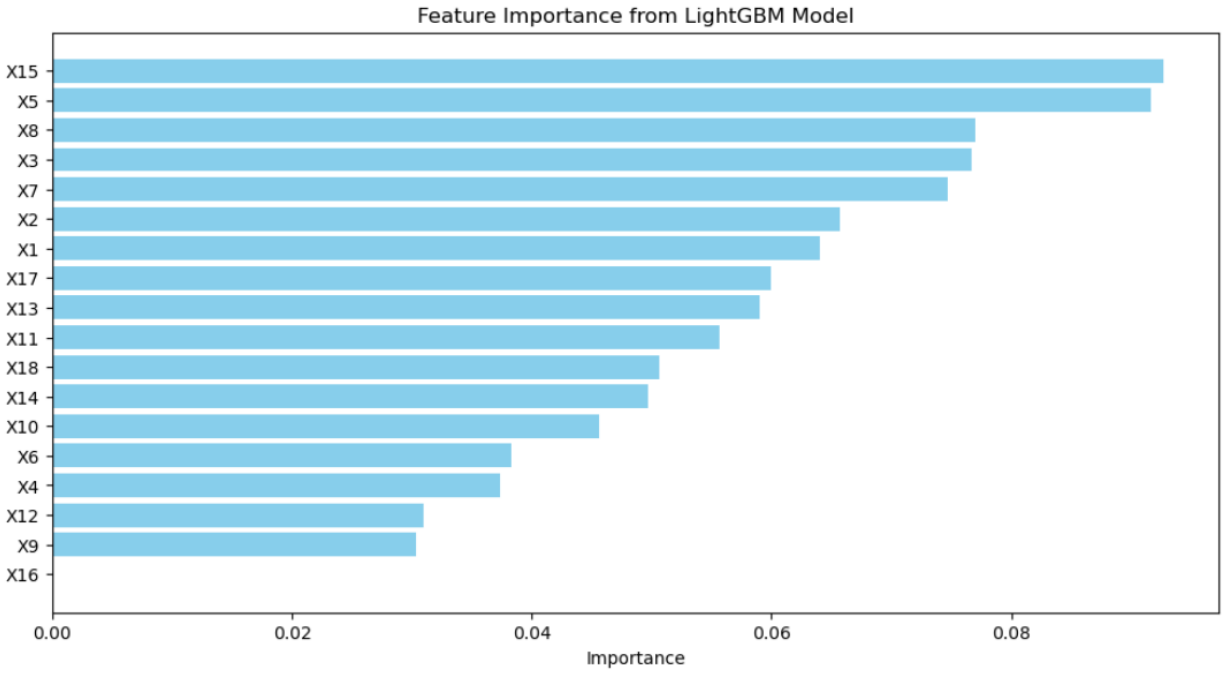
Feature Importance from LightGBM Model
X15
X5
X8
X3
X7
X2
X1
X17
X13
X11
X18
X14
X10
X6
X4
X12
X9
X16
0.00
0.02
0.04
0.06
0.08
Importance

Fig 3