Bankruptcy Prediction with Machine Learning

David Richter, Flatiron School, June 3, 2022

Business Problem

- Bankrupt companies typically have to liquidate their assets in order to repay creditors, causing the value of their shares to drop to zero.
- This represents a risk for investors taking long positions on such companies, but an opportunity for investors taking short positions.
- In this presentation I'll discuss two machine learning models trained to predict bankruptcies and how these models can be used to build a short investment strategy.

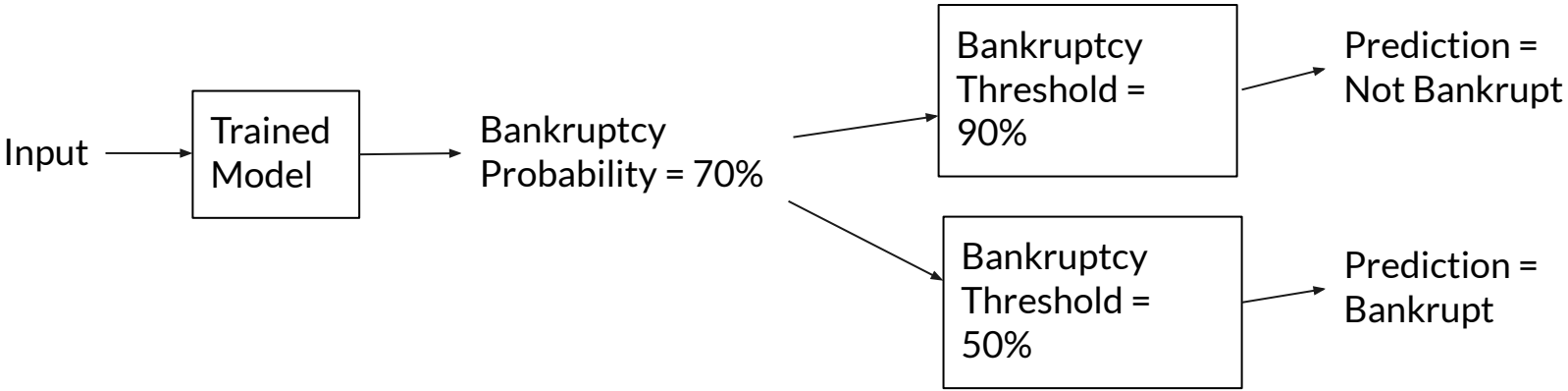
Data

- This study uses financial data for almost 7000 companies listed on Taiwan's stock exchange between 1999 and 2009.
- The original data set included 95 columns of financial data for each company.
- Only 3% of companies in the data set went bankrupt, so the Machine Learning models I trained take account of this class imbalance.

Output of a ML Classification Model

Machine learning classification models output both a **probabilities** and a **predictions**:

1. For each input, the model gives a **probability** that it belongs to a particular class.
2. If the probability of an input falls above a certain threshold (e.g. 50%) the model gives the **prediction** that the input does in fact belong to that class.



Performance Metrics for Classification Algorithms

- Two important metrics tell us how well a classification model performs on a set of testing data.
- **Precision** = True Bankruptcies Predicted by Model/Total Bankruptcies Predicted by Model
- **Recall** = True Bankruptcies Predicted by Model/Total Bankruptcies
- **Short strategy investors should care about Precision AND Recall**:
    - High Recall = more profits from shorting companies that do go bankrupt
    - High Precision = fewer losses from companies that don't go bankrupt when they were expected to.
- There is always a trade-off between precision and recall.
- This means we need to combine precision and recall into a single Expected Profit metric that will help investors choose the best model.

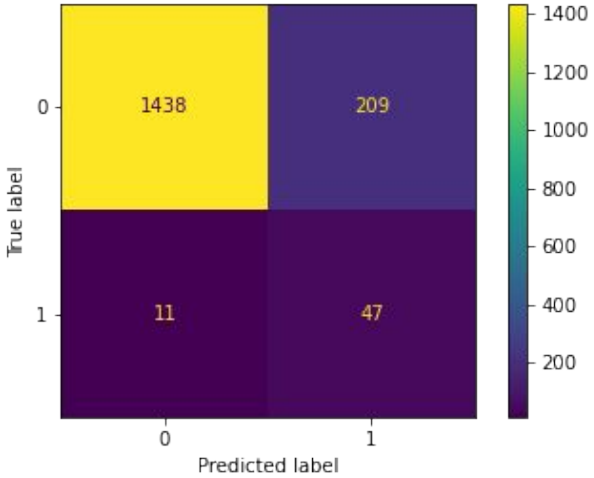
Expected Profit - a back of the envelope calculation

- For this presentation, I'll make the simplifying assumption that the profits from correctly predicting a single bankruptcy equal the losses from incorrectly predicting a single bankruptcy.
- Given this assumption:
  - Expected Profit is proportional to: recall * (precision - (1-precision) OR
  - Expected Profit is proportional to: recall * (2*precision - 1)
- If our bankruptcy guesses are half right and half wrong, precision = 0.5, and we would expect 0 profit, regardless of our recall.
- The recall term gives us the relative size of our profit or loss.

Two ML Classifiers: Logistic Regression and XGBoost

Logistic Regression:

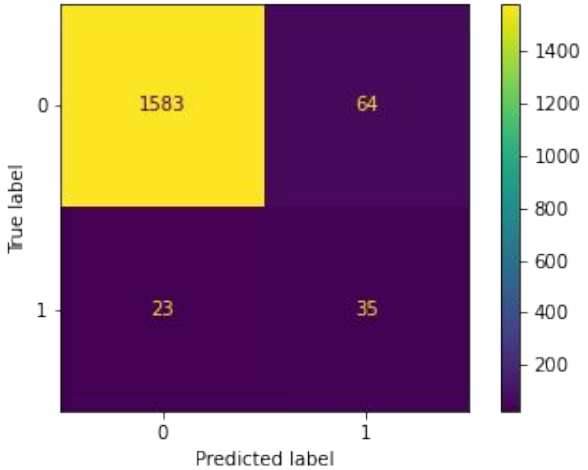
50% probability threshold



Recall= 0.81, Precision= 0.18, Expected Profit=-0.52

XGBoost:

50% probability threshold

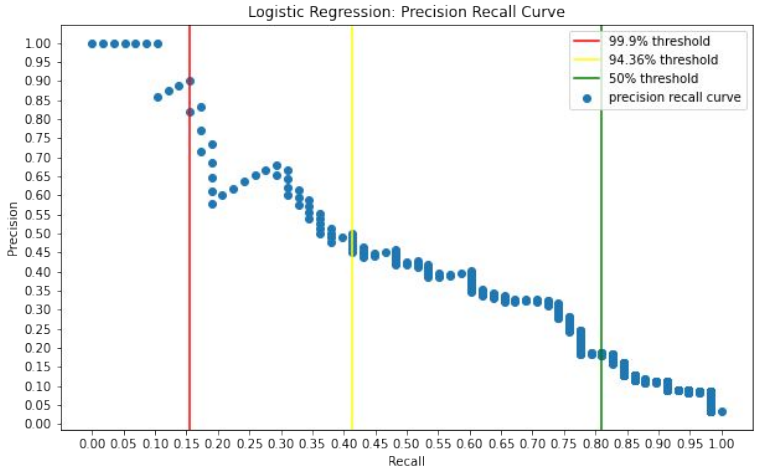


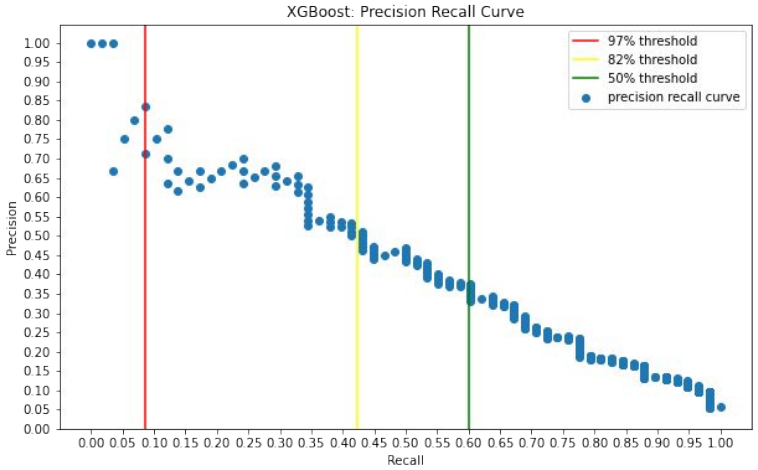
Recall=0.60, Precision = 0.35, Expected Profit=-0.18

The Problem with a 50% Threshold

- As the previous slide showed, we see losses if we use a 50% probability threshold
- In the next slide, we'll be able to see a precision recall curve, which shows the precision and recall at every probability threshold.

Logistic Regression vs. XGBoost: Precision/Recall Tradeoff



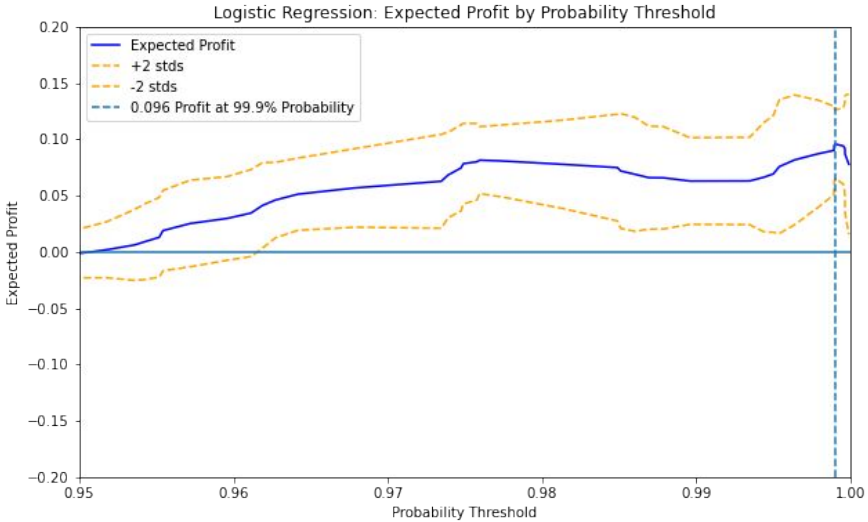


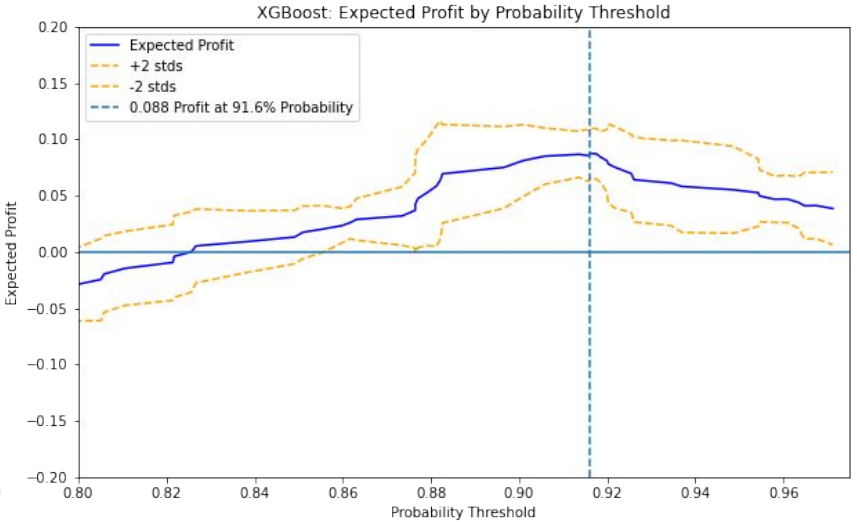
Left of the yellow lines: probability thresholds where precision > 50%

Finding Maximum Expected Profits

- The Precision Recall Curve is interesting, but it doesn't tell us which model is most profitable.
- To find Maximum Expected profits for both models, we can calculate an expected profits curve as a function of precision and recall.
- I used a smoothing function to generate this curve, calculating profits at a given probability threshold as the mean of the empirically determined profit at that threshold and the empirical profits of its five upper and lower neighbors.
- Smoothing function is necessary because the sharp rises and dips in precision we see are the result of noise in our test data and we want a measure of performance that's indifferent to the particular test data we use.
- The profits graph also shows an interval of two standard deviations above and below the curve.

Logistic Regression vs. XGBoost: Max Expected Profits





Conclusions, Further Analysis, and Limitations

- Given our assumptions, the maximum profit we can expect from a short strategy comes from using the Logistic Regression Model, which, based on its performance we would expect to yield 9% higher profits than the XGBoost Model.
- With further analysis, we might want to revise our assumptions about expected gains and losses from false and true bankruptcy predictions and revise our expected profit calculation accordingly.
- An even more important limitation of this study is that none of the data included in this data set has dates attached to it, so we don't know the extent to which variables incorporating share prices have been impacted by variables related to financial reporting. If much of the risk that we're identifying has already been priced in to share prices, then the profits to be realized through a short strategy will be limited accordingly.

Thank you for viewing my presentation!

Please see my github at
https://github.com/DavidKRichter/dsc-bankruptcy-project/blob/main/index.ipynb

and email me me at

d.richte@gmail.com

with any questions!