William Line and Peter Emanuel
wl434 and pae54

# CDS Pricing: Midterm Report

**Description of Data Set:**

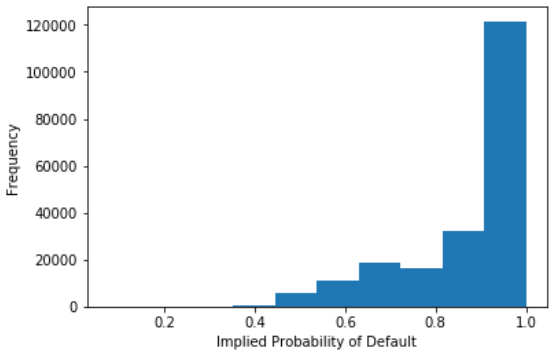
We started to build our dataset based on the tickers included in the iShares iBoxx High Yield Corporate Bond ETF HYG (1338 bonds). We narrowed this universe down to bonds for those tickers maturing between 2022 and 2026 (646, they may not necessarily be in the index), because these bonds could have, at some point in the life of our dataset, been considered as 5-year benchmark bonds. We further narrowed down this dataset to include only the bonds with high enough liquidity (in terms of amount outstanding and issue date) to analyze. Finally, we used this subset of bonds to pull 5-Year Credit Default Swap Spreads and Stock Price data for the corresponding companies.

- Summary:
  - Bonds: Bloomberg (1338 - 646 maturing between 2022 and 2026)
    - Ticker: unique identifier
    - Issue Date
    - Maturity Date
    - Amount Outstanding
    - Sector: business sector in which the company operates.
    - Bloomberg Composite Rating: ranging from CC+ to BBB- (“crossover”) in our universe, identifies “quality”, the mode of Moody’s, S&P, and Fitch ratings.
  - Stocks: Bloomberg (144)
    - Ticker: unique identifier
    - Stock Price: stock prices for each ticker symbol from 9/1/2016 to 10/28/21
  - CDS: Bloomberg (150)
    - 5-year CDS Spreads: 9/1/2016 to 10/28/2021
    - Implied 5-year Default Probabilities: Calculated from formula in proposal, using CDS spreads and market convention recovery rates (primarily 0.4).

Due to the illiquid nature of the High Yield CDS market, our 5-year CDS spread data was missing or unchanged for all days without a transaction. We decided to backfill any data missing at the start of the time frame and frontfill the rest. We did the same for missing stock data.

**Data Exploration:**

To better understand the data, we first made a histogram of the 5-year implied probability of default calculated using the equation: $P(0,t) = 1 - e^{\frac{S_t}{1-R}}$, where $S_t$ is the 5-year CDS spread in percentage points, not basis points, and R is the recovery rate.



From the histogram, we can see that the majority of our 5-year implied probability of defaults are between 90% and 100%, which is expected for many High Yield bonds. About 40% of the total mass lies between 40% and 90%, however.

Further exploration of our initial data was done looking at the summary statistics of our distribution of implied 5-year probability of default, first by bond rating, and then by business sector. When we look at the summary statistics (shown below), we can see that, in general, bonds with higher ratings have lower 5-year probabilities of default, trending to higher probabilities the worse the rating gets. Stable sectors such as Utilities and Consumer Non-Cyclical have, on average, the lowest 5-year probabilities of default.

Summary Statistics: 5-Year Probability of Default by Bond Rating

| rating | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| BBB- | 11 | 85.0% | 14.1% | 51.1% | 78.7% | 87.0% | 94.2% | 99.9% |
| BB+ | 31 | 81.3% | 12.0% | 56.4% | 72.4% | 83.0% | 90.4% | 98.8% |
| BB | 26 | 87.4% | 11.8% | 59.2% | 79.4% | 92.7% | 96.3% | 97.9% |
| BB- | 30 | 88.4% | 9.7% | 63.3% | 82.8% | 92.5% | 96.9% | 99.3% |
| B+ | 21 | 91.2% | 13.1% | 59.0% | 88.4% | 98.3% | 99.2% | 100.0% |
| B | 13 | 95.4% | 9.9% | 63.5% | 94.7% | 99.2% | 99.5% | 99.9% |
| B- | 9 | 96.6% | 4.8% | 84.3% | 96.6% | 98.6% | 99.0% | 99.5% |
| CC+ | 1 | 99.0% | | 99.0% | 99.0% | 99.0% | 99.0% | 99.0% |
| CCC+ | 7 | 93.7% | 10.3% | 71.0% | 94.5% | 97.7% | 99.3% | 99.9% |
| CCC | 1 | 97.0% | | 97.0% | 97.0% | 97.0% | 97.0% | 97.0% |
| NR | 3 | 76.3% | 16.5% | 57.4% | 70.4% | 83.5% | 85.7% | 88.0% |

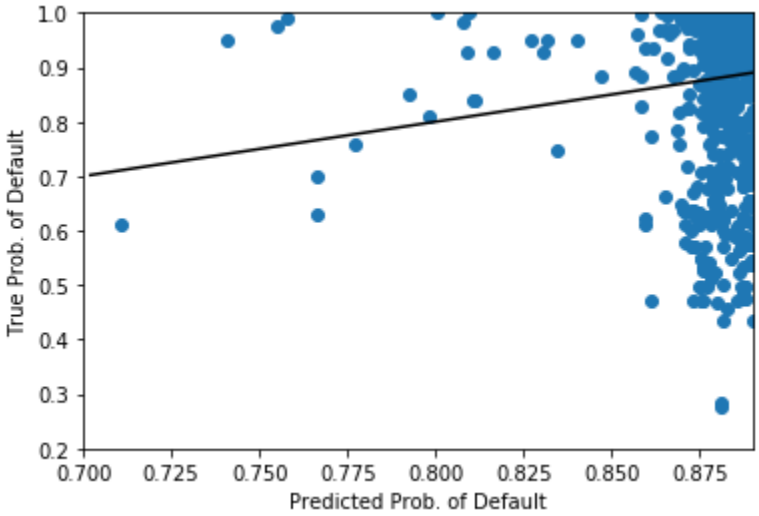
Summary Statistics: 5-Year Probability of Default by Economic Sector

| sector | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Basic Materials | 9 | 90.8% | 7.8% | 71.7% | 89.2% | 93.7% | 93.9% | 99.4% |
| Communications | 31 | 85.9% | 13.0% | 57.4% | 77.3% | 92.0% | 97.8% | 99.7% |
| Consumer, Cyclical | 35 | 93.5% | 6.1% | 78.8% | 91.5% | 96.0% | 97.6% | 99.6% |
| Consumer, Non-cyclical | 18 | 83.7% | 15.6% | 59.0% | 68.5% | 89.0% | 98.4% | 99.9% |
| Energy | 15 | 89.2% | 9.1% | 71.0% | 85.7% | 89.2% | 97.3% | 99.9% |
| Financial | 20 | 86.9% | 11.3% | 63.0% | 76.5% | 89.7% | 97.1% | 99.3% |
| Industrial | 12 | 90.4% | 12.7% | 61.6% | 83.2% | 97.1% | 99.2% | 100.0% |
| Technology | 5 | 85.3% | 17.1% | 56.4% | 83.0% | 93.4% | 94.6% | 98.8% |
| Utilities | 8 | 79.8% | 19.4% | 51.1% | 63.5% | 83.3% | 97.2% | 99.9% |

**Initial Modeling:**

As a preliminary model for predicting 5-year probability of default (Y), we first ran a linear regression using only one feature, stock returns. We did not find this model to be useful as it trivially predicted almost all 5-year probabilities of default to be 1. We ran the regression again using stock price (X) and the output is shown below. The data was split into training and test sets (80/20).

$$\hat{Y} = 89\% - (0.02\%)*X$$




The mean squared errors were low at .02 for both our training (n = 164,628) and test (n = 41,310) sets. We believe that the model yields a powerful predictive ability, not because of correlation between stock price and 5-year probability of default (-0.1) but due to high levels of autocorrelation in our stock price data (Durbin-Watson of 0.002). Another downside is that strictly positive stock prices result in 5-year probabilities of default no higher than the intercept of 89% when fit to a negative-sloping equation. We tend to overestimate the probability of default, especially for companies with low stock prices, but we still do capture the fact that "penny" stocks should have higher probabilities of default. The residuals themselves were not significantly autocorrelated (Durbin-Watson of 1.75), telling us that a linear model is appropriate to use on this dataset. In order to expand our model, we looked into adding a many-hot encoding of bond rating to the regression. When we did so, we saw virtually no change in our MSE. Normalizing our stock price data may change the impact of new features on predictive power.

Overfitting is always a concern when creating a predictive model. In order to avoid overfitting, we will continue to create future models with certain precautions in mind. We will keep the ratio of features to data points low, keep using a linear model for regression, and select our set of features based on lowest cross-validation error; this is particularly important when looking at adding possibly correlated, redundant macro-indicators or when looking at how to best standardize our features.

**Further Work:**

- Regression:
    - First, we plan to normalize the stock price data by ticker in order to eliminate some of the autocorrelation in our model. We will then use our universe of liquid bonds to find a best estimate for each ticker's 5-year corporate bond yield on each day in our history, and add this data to our model. Finally, we will examine more ways to improve cross-validation error by adding features for bond rating, economic sector, and macro-indicators such as Consumer Price Index, Volatility (VIX), US GDP, and 5-year US Treasury yields.
- Classification:
    - With our predicted 5-year probabilities of default and the observed 5-year CDS spreads, we can then solve for the recovery rate to get a range of estimates, rather than a constant 0.4. To test the reliability of our new recovery rate estimates, we will train a classifier to categorize recovery rates for companies as either over or under the standard recovery rate of 0.4. This can tell us when each CDS was overpriced or underpriced throughout the life of our dataset and if each CDS is overpriced or underpriced at the present day.