Data Science Honors Thesis: Final Prospectus

Corporate Credit Rating Prediction in the Energy Sector

Rodrigo Palmaka
University of California, Berkeley, California, USA.

1 Motivation

1.1 Industry Landscape

The field of quantitative finance is a fast paced, competitive landscape where big players fight for narrow advantages and nearly endless alpha opportunities allow even small firms to make an impact if they're able to creatively detect market inefficiencies. The quantitative approach to trading is very intriguing since, at it's core, the mission is to understand a chaotic, seemingly random process using mathematically rigorous explanations. Investing is an important part of the American way of life. Currently, about 55% of Americans own stock[1]. Less than two decades ago, this rate was as high as 62%. The relative strength and consistent growth of the US market over time despite crises has made it the ideal playing field for investors all over the world. The importance of the stock and bond markets for economics, politics, and social wellbeing make advances in trading techniques highly lucrative.

By nature of the stock market (and basic economics), effective trading strategies are not profitable if everyone can take advantage of them. For this reason, firms in this industry must be highly secretive with their models as any exposure may ruin their entire strategy. The siloed environment of quantitative finance drives the need for novel approaches of applying cutting edge mathematical modeling and machine learning. Due to this need for fresh statistical methods, quantitative finance is closely tied to academia.

1.2 Academic

My personal interests in quant finance are supported by my academic background. While at Berkeley, I have immersed myself in a medley of Computer Science, Mathematics, and Statistics courses with the goal of being able to tackle data science challenges in finance and other industries. Focusing on developing

[1] https://news.gallup.com/poll/266807/percentage-americans-owns-stock.aspx

fundamentals, I have done coursework in Multivariable Calculus[2], Probability[3], Algorithms[4], Linear Algebra[5], and Real Analysis[6]. Considering the skills necessary to work in the realm of quantitative finance I have also taken courses in Stochastic Processes[7], Machine Learning[8], and Deep Neural Networks[9].

1.3 Problem Statement

For this thesis project, I sought to pursue my interests in quantitative finance and apply theory from my coursework at Berkeley to a project that could be impactful and developed within the timeframe of two semesters. I became interested in the literature involving corporate credit rating prediction as I saw a straightforward classification task that could disrupt a rating process currently done by hand. Ratings agencies such as Standard & Poors, Moody's, and Fitch dominate the market for corporate and bond ratings. Given the subjective nature of these ratings, they can be influenced by third parties and have been prone to ethical issues in the past[10], these issues are covered in more detail later in "Ethical Considerations".

My goal is to develop a better classification model by using financial features and energy sector specific metrics, then applying novel feature engineering approaches to existing research methods.

2 Dataset

Data for this project will be sourced from Standard & Poors Global's Compustat database. This data source contains quantitative information from companies' quarterly and annual filings as well as the necessary labels for the classification task, the long-term issuer ratings from S&P. Data can be queried from this database and features over 600 financial features for listed companies.

2.1 Features

The bulk of the features used for this task will come from quarterly financial numbers and ratios as well as sector specific indicators from approximately 100 of the largest energy companies by market cap. This data will come directly from Compustat. I will also attempt to find non-financial sector specific features that

[2] MATH 53
[3] STAT 140: http://prob140.org/
[4] CS 170: https://cs170.org/
[5] MATH 110
[6] MATH 104
[7] STAT 150
[8] EECS 189/289A
[9] EECS W182
[10] https://www.reuters.com/article/us-mcgrawhill-sandp-civilcharges/u-s-government-slams-sp-with-5-billion-fraud-lawsuit-idUSBRE9130U120130206

may give more insight beyond the variables commonly used for sector agnostic credit prediction. An important consideration when using financial time series data is having a similarly distributed train/test sets, i.e. training and testing in time periods within the same latent market regime.

2.2 Labels

The target variables for this problem are the Long Term Issuer Credit Ratings produced by S&P for individual companies. S&P is known for their bond ratings which provide investors with insight into the health of a specific issue of debt. In this project, we will focus on the credit ratings associated with the companies themselves. Issuer ratings are correlated with bond ratings but make for a more straightforward classification task as they are not dependent on nuances of specific bond issues. These ratings range from AAA to BBB- for investment grade and BB+ to C for junk. The D rating represents a company in default.

3 Methodology

3.1 Data Processing

3.1.1 Missing Data

For missing features, my general approach will be to impute missing data with the mean of that feature. Data points with missing labels, will be omitted. I will also employ methods in the processing phase to eliminate companies with too high a ratio of missing feature data. Some features themselves are often sparse and will not be included in modeling.

3.1.2 Standardizing/Normalizing Data

In this phase, I will preserve the original data and produce a separate data set with features that are standardized if they are normally distributed or normalized if they follow some other distribution.

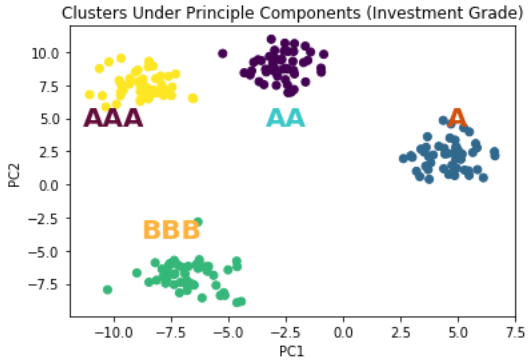
3.1.3 Unsupervised Learning and Dimensionality Reduction

The first approach to modeling credit ratings will be to better understand the latent variable space of my data. PCA and CCA will be powerful tools to visualize the directions of this high dimensional data set. Ideally, at this stage I would be able to see separate clusters respective to each rating class. It may be useful to perform K-means clustering to enhance this low-dimensional analysis.

3.2 Feature Selection

A pivotal part of this project, finding relevant features can be an impactful method of exploring what factors contribute to an issuer's credit rating. Methods

in this stage will include ANOVA testing, correlation between variables, and feature selection techniques like Recursive Feature Elimination. Dimensionality reduction also plays a key part in selecting important features so this part of the modeling process will work together with the former.



(Figure 1: example of a relationship we may find in data)

3.3 Supervised Learning

Finally, with a candidate set of features I will train different classification models on the data and score their predictions using ROC curves, and confusion matrices. For applicable models, I will run grid search hyperparameter tuning. The models to be used at this stage will highly depend on the current literature. I will seek to explore traditional methods of multiclass classification via SVM and Feedforward Neural Networks as well as niche methods from other research like applying Students-t HMMs to this task. The goal is to predict a credit rating most accurately from AAA to C or even in default. The models will be trained with many different combinations of features as understanding the underlying information that comprises ratings is the key to this thesis.

4 Expected Outcomes and Relevance

4.1 Guiding Questions

In this project, we will be addressing the questions:

(1) What are the current state-of-the-art models used for predicting corporate credit ratings?

(2) Can we improve upon these methods via creative approaches in feature engineering and selection before the model fitting stage?

4.2 Target Goals

Realistically, this study should provide some color to the learning models academics currently use in determining a company's health as a debtor and give insight into what quantifiable data is genuinely valuable for computing credit ratings of energy companies.

4.3 Reach Goals

(1) Improved classification errors for out-of-sample tests when applying a specific model from prior work in this problem, e.g. improving upon a SVM based rating system through novel feature engineering.
(2) Successfully using an existing supervised learning model for credit prediction for energy companies that was not previously successful solving this task generally or in the energy sector specifically.
(3) Given that the novel method can bring greater accuracy to current models and/or holds some value in better understanding how to replicate actual S&P credit ratings, it would be useful to develop a platform that can be accessed by laymen and continuously updates credit ratings.

5 Ethical Considerations

Regarding the reliability of the data, understanding bias will be necessary. The quarterly financial data I will work with from the Compustat Data Sets include data that was submitted by publicly traded companies to the SEC per regulatory demands. This data is publicly available for all to analyze and it is scrutinized by the regulatory agency itself to detect discrepancies and fraud. Therefore, I believe it is fair to trust the data published by the SEC and assume their due diligence process would detect fraudulent data. The credit ratings themselves, however, aren't as impartial.

During the great recession, for example, the "big three" rating agencies faced a barrage of criticism for propping up credit ratings of institutions in distress and downplaying the risk exposure of products such as Mortgage-Backed Securities. The questionable validity of ratings given by the ratings organizations misled investors and may have contributed greatly to the 2008 recession. It is necessary to understand that the labels will be inherently "noisy" and may even have a skewed distribution, i.e. companies in X country have a higher rating compared other companies with similar fundamentals in other locations.