Data Science Honors Thesis Prospectus: Corporate Credit Rating Prediction in the Energy Sector

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

1 Overview

To setup the context for this project, we will explore existing methods of supervised learning for classifying Standard & Poor's (S&P) credit ratings of individual, publicly-traded corporations within the energy sector as well as successful approaches in other industries. The goal of this work is to create improvements to current approaches with emphasis on novel implementations of feature engineering, dimensionality reduction, and feature selection. Through improving the feature selection by using a variety of techniques, we hope to increase prediction accuracy by focusing on important predictors, and increase the interpretability of the resulting model by discovering which original and engineered features contribute most to making predictions. As a corollary to this thesis, we will be improving on granularity of ratings since the model can compute credit ratings with higher frequency than the ratings agencies – limited only by feature granularity.

2 Guiding Questions and Goals

2.1 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?

2.2 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.

2.3 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.

3 Data Techniques

Techniques for this study can be separated in three groups:

3.1 Data Collection & Preprocessing

Web scraping applied to SEC quarterly filings and news feeds, normalization/scaling of features, imputing data for missing values, accounting for differences in frequency of data points (e.g. a feature with monthly data vs. one with quarterly outputs).

3.2 Feature Engineering

Analysis of correlation between features, recursive feature elimination, Linear Discriminant Analysis (LDA), Principal Component Analysis (PCA), Lasso, and CCA.

3.3 Prediction

Current models used thus far to tackle this task include, but are not limited to: Support Vector Machines (SVM), Student's-t Hidden Markov Models, Artificial Neural Networks (ANN), Bagged Decision Trees, Random Forest, and Boosting.

4 Dataset

Data to be used for this project would primarily come from the financial statements of energy companies. It is unlikely that this data will be formatted and readily downable in our necessary format, it is expected there may be an element of web scraping, as mentioned in section 3, to collect this data.
Given the extent of work done around credit rating prediction, this study will pursue utilizing features from datasets used in previous works as well. Lastly, as this study is largely focused on the impact of features, I expect to incorporate public financial data/ratios as well as non-financial data that may not have been

included in previous studies. The sources of such data will be varied and will require some data mining creativity.

5 Faculty Advisor

Currently looking for faculty advisors with experience in applying probabilistic or prediction models in finance.

6 EDA

Financial data falls under the class of Time Series data i.e. data that are indexed by time and often have a relationship with time, seasonality, etc. EDA approaches will keep this in mind and include analysis of autocorrelation. Also, I expect to work with features with different granularity so it will most likely be necessary to impute values to fill gaps.

7 Career Connection

As a Data Science major at UC Berkeley, coursework in probability, machine learning, and statistical methods allowed me to work with an assortment of tools for supervised learning which will be employed here. Furthermore, the applied mathematics and modelling Domain Emphasis provided me with the background to interpret the math content of research in this topic. Upon graduation, I will join Citi in a Quantitative Analysis capacity and seek to further my work in understanding the financial markets through mathematical modelling.

8 Human Contexts and Ethics

Energy companies provide a foundational service for our society. It is an industry with very high capital and infrastructural barriers of entry. As a result, often a handful of corporations dominate entire regional markets in this sector. If we can better understand the health of such a company via their credibility as a debtor, it can help us recognize vulnerabilities in the case of a natural disaster or emergency. The ethical concern posed by titans in the energy sector is deterred by the transparency of credit ratings among other public financial indicators.