Corporate Credit Rating Classification in the Energy Sector

Data Science Honors Thesis

Rodrigo Palmaka

A thesis presented for the degree of
Bachelor of Arts

Data Science
University of California, Berkeley
U.S.A
May, 2021

Contents

1 Abstract

2 Introduction

2.1 The Problem

How to better understand a subjective, potentially biased process such as rating a company's credit-worthiness?
Explain the S&P Long-Term Issuer rating and how analysts come up with this assessment. Include picture of ratings scale.
Touch on past ethical dilemmas caused by these ratings.

3 Literature Review

Discuss current research in the field. SMOTE (Chawla et al - 2002), SMOTE applied (Hajek, Olej - 2014), SVM (Huang et al - 2004), comparative study and deep NNs (golbayani et al - 2020), CCA papers. How does my work aim to build off these? What is lacking in these papers that I hope to bring (e.g. focus on feature selection + engineering, reproducible codebase).

4 Data Collection

- Overview of Compustat (ratings and financial ratios)
- Feature exploration process, combining ratios to make new features

5 EDA

5.1 Data Cleaning

- Standardization/Normalization
- discuss missing data (Paleologo - 2010)

5.2 Feature Statistics

- ANOVA testing features (see Huang et al - 2004)
- Feature correlation matrix
- Recursive Feature Elimination
  feature engineering discussed with respect to specific Models later

5.3 Train/Test Distributions

- Explore label distribution
- SMOTE
- Theo's idea for dividing train/test by companies
- dataset of ratings changes (EVD proposed)

5.4 Dimensionality Reduction

- PCA
- CCA

6 Models

Heavy use of plots, visualizations.
Include small amount of math regarding specific use of feature engineering or kernels.
Visualizations of custom Neural Net architecture. *See if I can use 189 widgets for Hyperparam tuning in notebooks.

- SVM - linear, rbf, poly. Include decision boundaries viz.
- LDA/QDA - explore the assumptions and explain how they are/aren't fulfill in this problem. Decision bound plot
- KNN
- Random Forest
- Boosting - AdaBoost
- Softmax
- Fully Connected NN
- RNN

7 Evaluation Methodology

- Discuss Golbayani comparative study section 4.2
- "One-off" score (Huang et al 2004)
- EVD's method for predictive ratings movement
- The case for predicting +/- distinctions or just letters (17 class vs 7 class problem)
- Mention notch distance from Golbayani and ponder custom performance metric
- Discuss effects of non-uniform labels distribution and effect on accuracy as metric
- Classification metrics and relevance (F1 score, precision, recall)
- ROC curves

8 Results

Tables for different datasets

9 Conclusion

Main focus should be on interpreting important features that can explain the process by which analysts come up with ratings.
Discuss my results in regards to those of other papers.

10 Future Considerations