# Data Science Honors Thesis: Methodology

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

## 1 Data Collection

Data for this thesis will come almost exclusively from S&P 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 Data Processing

Data from Compustat will include the features, quarterly financial figures from several energy companies, and their respective long-term issuer ratings at those dates. Data processing will be performed entirely in Python 3 via the Jupyter Notebook environment using libraries such as NumPy, Pandas, SciPy, among others.

### 2.1 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.

### 2.2 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 Modeling

### 3.1 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.

### 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 Method Documentation

The methods I expect to employ in this thesis will be documented in the final write up for this project. Also, notebooks with code used for the above tasks will include instructions so that results may be reproduced by third parties, e.g. using random seeds whenever randomness is involved.

## 5 Current Problem Landscape

### 5.1 Literature

References whose methods I will utilize/build upon include but are not limited to:

- Pai, P.-F., Tan, Y.-S., and Hsu, M.-F. (2015). Credit rating analysis by the decision-tree support vector machine with ensemble strategies. International Journal of Fuzzy Systems, 17(4):521–530

- Petropoulos, A., Chatzis, S.-P., and Xanthopoulos, S. (2016). A novel corporate credit rating system based on Student's-t hidden Markov models.
- Golbayani, P., Florescu, I., and Chatterjee, R. (2020). A Comparative study of forecasting Corporate Credit Ratings using Neural Networks, Support Vector Machines, and Decision Trees.
- Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., and Wu, S. (2003). Credit rating analysis with support vector machines and neural networks: a market comparative study

### 5.2 Novelty and Contributions

The goal of this project is to build upon the current research in corporate credit rating prediction by leveraging unique feature selection. I will seek to advance this research field by improving prediction accuracy of models in the energy sector and better explain the subjective process of debtor assessment through quantifiable means.

## 6 Biases

As detailed in the data source documentation, there is room for potential bias when it comes to the credit ratings of companies. The 2008 financial crisis unearthed many wrongful practices from agencies such as Standard and Poor's in how they produce credit ratings. It is necessary to understand that the labels will be inherently "noisy" and may even have a specific distribution, i.e. companies in X country have a higher rating compared other companies with similar fundamentals in other locations.