E*(bank)*RUPT – Predicting the potential of non-solvency for American public companies.

Business Understanding

- What problem are you trying to solve, or what question are you trying to answer?
  - **Business Problem**: Using publicly provided, presumably audited financial data, can we predict if a business can remain solvent and continue to operate successfully under normal business conditions?
  - A predictive model like this would help provide sound recommendations for clients looking for near-term investment opportunities. We strive to mitigate risk by identifying, assessing, and tracking potential dangers with businesses using their financial data. This mitigation comes from understanding what components could lead to a problem ending with a filed bankruptcy.

- What industry/realm/domain does this apply to?
  - Our analytics model will most closely apply to the Business / Financial services public sector.

- What is the motivation behind your project?
  - Given our passion for business, finances, and newly acquired knowledge of AI and Machine Learning, we wanted to find a way to combine these to focus on creating something meaningful that could apply to any real-world business problem.

Data Understanding

- What data will you collect?
  - We have chosen to use a Kaggle.com provided dataset for bankruptcy prediction related to American public companies. The dataset comprises accounting data from 8,262 distinct companies, for a total of 78,682 observations of year-end filings, for a period spanning up to 18 consecutive years.

- Is there a plan for how to get the data (API request, direct download, etc.)?
  - The Kaggle dataset is provided as a one-time CSV direct file download.

- Are the features that will be used described clearly?
  - All features provided are common FASB required data elements used in accounting and financial reporting. We can use the data as provide to produce our Target variables.

Current Assets	Total Receivables	EBITDA	Net Income	Market Value
Net Sales	Total Assets	Total Long-Term Debt	Gross Profit	Total Revenue
Total Liabilities	Total Operating Expenses	COGS	Depreciation	Amortization

Data Preparation

- What kind of preprocessing steps do you foresee (encoding, matrix transformations, etc.)?
    - Importing needed Python Libraries (NumPy, Pandas, Matplotlib, ski-kit learn) needed for data analysis and cleaning.
    - Loading the dataset (data file) into a dataframe using Pandas.
    - Identify and handle missing, incomplete, and incorrect data elements.
    - Perform any encoding steps to make sure all the key features are in a numeric form
    - Splitting the data into Training, Validation and Test datasets.

- What are some of the cleaning/pre-processing challenges for this data?
    - Make sure that all the key measures (data elements from file) needed to derive the target variable are populated, are in numeric form, and are shown on the same scale as the other data elements and rows.

Modeling

- What modeling techniques are most appropriate for your problem?
    - We will be using Supervised Learning as we will need to train a model to make a prediction of a target (z-score) using data from continuously changing measure/data values.

- What is your target variable?
    - There is no perfect way to determine a company's financial health, let alone have a full-proof prediction of sustainability. However, four critical areas of financial well-being can be modeled and scrutinized closely. Liquidity, solvency, profitability, and operating efficiency (D/E Ratio) are all important indicators to consider in a company's "Bottom Line" for sustainability and overall financial health. We intend to use a calculated measure known as the **Altman Z-Score**, which combines our identified business indicators and produces a probability score we can use as our **target variable** to identify each business's sustainability.

      Target Z-Score Interpretation:

3.0 and above	Safe - based on key financial measures.
2.7 - 2.99	On Alert – should exercise caution before investing.
1.80 – 2.69	High probability of bankruptcy within 2 years
Below 1.80	Avoid

- Is this a regression or classification problem?
    - We believe this will be a classification problem as we try to assign a business into a group (shown above) that will predict the probability of business insolvency.

Evaluation

- What metrics will you use to determine success (MAE, RMSE, etc.)?
    - To Be Determined

Tools/Methodologies

- What modeling algorithms are you planning to use (i.e., decision trees, random forests, etc.)?
    - We plan on utilizing multiple modeling algorithms for our model. We will use decision trees, random forests, and multiple regression. It is important for us to experiment with these different models to ensure the greatest accuracy.