BA 885 Advanced Analytics II - Project Proposal

Team Members: Yulong Gong, Scott McCoy, Antonio Moral, Yichi Zhang

Predicting Non-profit Terminations and Liquidations

Problem:

Can we use prior year financial information from non-profit tax return filings to predict whether an organization will terminate/liquidate all or part of its operations?

Definition of termination and partial liquidation per IRS instructions.

- Termination - “Liquidation, Termination, or Dissolution”
- Partial Liquidation - “Sale, Exchange, Disposition, or Other Transfer of More Than 25% of the Organization’s Assets”

Our analysis could help organizations determine the risk that a nonprofit stakeholder would cease to exist or substantially change it’s structure in the near future.

Dataset:
Dataset Links:

All Year CSV files from Kaggle - https://www.kaggle.com/irs/irs-990?select=irs_990_2017
IRS Schedule N and Instructions - https://www.irs.gov/pub/irs-pdf/f990sn.pdf
Data dictionary files from IRS -
https://www.irs.gov/statistics/soi-tax-stats-annual-extract-of-tax-exempt-organization-financial-data

The data comes from several years of IRS 990 annual tax return filings. For tax years ending between 2012 and 2016, there are more than 200,000 company observations and roughly 240 categorical and numeric features.

Some preprocessing is needed to get this data into a machine learning usable format. We start with all observations having a 2013 year end, and merge all columns with 2012 observations giving us 2 years worth of data as features. Next we convert categorical columns to numeric, and create dummy indicators for categorical features with more than 2 possibilities. This results in a feature space of 542 features, with all 2012 column labels having a suffix of “t-1” to indicate a prior year value.

To create the labels, we create a list of companies that had a termination, partial liquidation, or either between 2014 and 2016. Next, we create 3 indicator variables that are 0 for companies not present in the lists and 1 for companies that are. Of the 266,000 observations, 1.35% had either a termination or partial liquidation in the subsequent 3 years, giving us an imbalanced classification problem.

Proposed Methodology:

EDA

With the clean data, we plan on performing some initial exploratory analysis to understand the financial data at hand. Using financial documents from the IRS which explain how the fields in our data are calculated as well as other useful information about the data, we plan on showing the trends for different attributes of our data visually. Because of how our data is structured, it will be important to consider the relationship between categorical variables and the financial capabilities of each of these non-profit organizations, and see if we can uncover insights that will increase our effectiveness at labeling the data using a neural network.

To create our model, we plan to start with a feed-forward neural network. With the size and complexity of our input data, we can try many different combinations of layers of various sizes to find which architecture can best describe the relationships between our features and targets. We can also try various combinations of hyperparameters like different loss functions, optimization methods, or number of training loops.

Upsampling / downsampling.

The target variable is quite imbalanced in the dataset. Roughly 1.4% of observations have a termination or liquidation in the subsequent 3 years. We plan on trying the resampling methods to further improve the neural network performance. Since a neural network would need a lot of training data, the upsampling method might be more suitable. By synthetically generated data points corresponding to minority class (companies have a termination or liquidation), we can potentially prevent the model from inclining towards the majority class (companies do not have a termination or liquidation), and improve the overall performance.