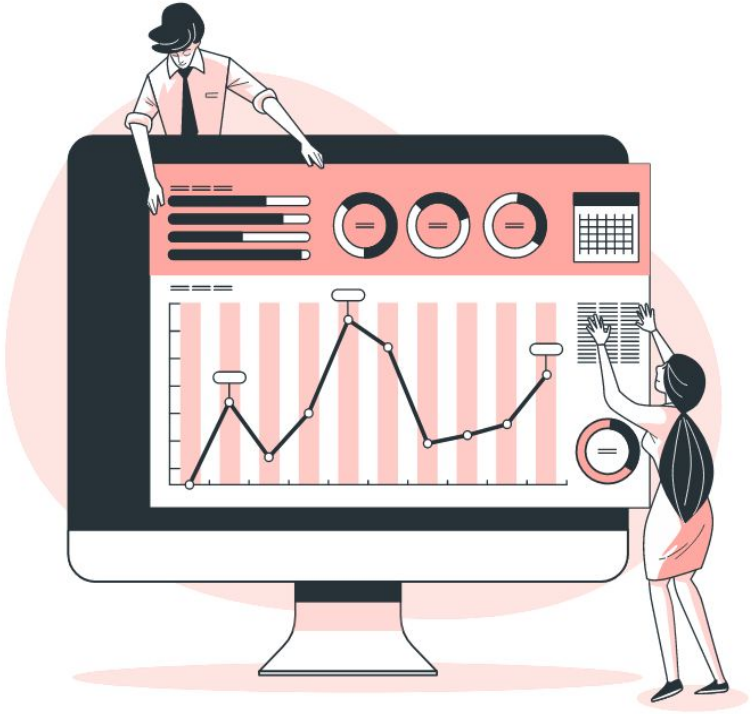
Predicting Non-Profit Terminations or Liquidations

Team 1: Yulong Gong, Scott McCoy, Antonio Moral, Yichi Zhang



Our team

Yulong Gong

Scott McCoy

Antonio Moral

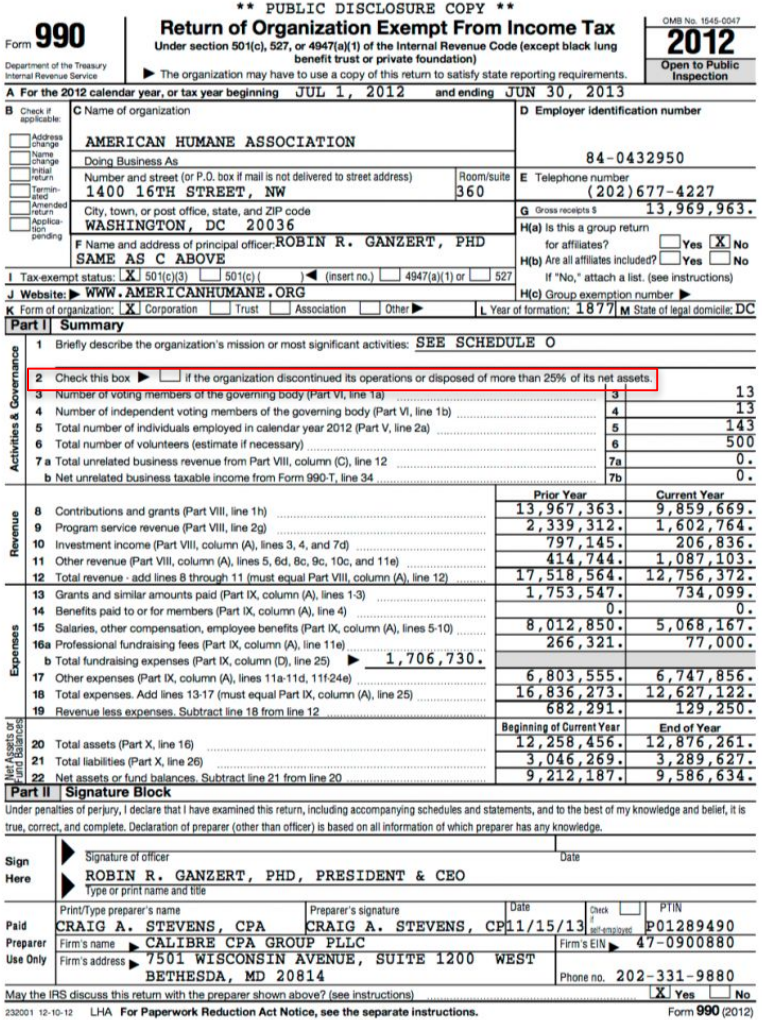
Yichi Zhang

Overview

- Dataset overview
- Models
  - Baseline model
    - Logistic regression
    - Random forest
  - Multi-classes neural network
  - Binary fc neural network
- Conclusions, limitations, future research
- Q & A

Dataset

- Non-profit tax return data
- Huge dimensions
  - 5 years of data
  - ~240 features
  - ~230,000 Companies
- The dataset is handled in GCP



** PUBLIC DISCLOSURE COPY **

Form **990** — Department of the Treasury, Internal Revenue Service

Return of Organization Exempt From Income Tax

Under section 501(c), 527, or 4947(a)(1) of the Internal Revenue Code (except black lung benefit trust or private foundation)

▶ The organization may have to use a copy of this return to satisfy state reporting requirements.

OMB No. 1545-0047 — **2012** — Open to Public Inspection

A For the 2012 calendar year, or tax year beginning JUL 1, 2012 and ending JUN 30, 2013

B Check if applicable: Address change; Name change; Initial return; Termin-ated; Amended return; Applica-tion pending

C Name of organization: AMERICAN HUMANE ASSOCIATION

Doing Business As:

Number and street (or P.O. box if mail is not delivered to street address): 1400 16TH STREET, NW — Room/suite: 360

City, town, or post office, state, and ZIP code: WASHINGTON, DC 20036

F Name and address of principal officer: ROBIN R. GANZERT, PHD SAME AS C ABOVE

D Employer identification number: 84-0432950

E Telephone number: (202)677-4227

G Gross receipts $: 13,969,963.

H(a) Is this a group return for affiliates? Yes [X] No

H(b) Are all affiliates included? Yes No — If "No," attach a list. (see instructions)

H(c) Group exemption number ▶

I Tax-exempt status: [X] 501(c)(3) [ ] 501(c) ( ) ◀ (insert no.) [ ] 4947(a)(1) or [ ] 527

J Website: ▶ WWW.AMERICANHUMANE.ORG

K Form of organization: [X] Corporation [ ] Trust [ ] Association [ ] Other ▶

L Year of formation: 1877 M State of legal domicile: DC

Part I Summary

1 Briefly describe the organization's mission or most significant activities: SEE SCHEDULE O

2 Check this box ▶ [ ] if the organization discontinued its operations or disposed of more than 25% of its net assets.

Activities & Governance	3 Number of voting members of the governing body (Part VI, line 1a)	3	13
	4 Number of independent voting members of the governing body (Part VI, line 1b)	4	13
	5 Total number of individuals employed in calendar year 2012 (Part V, line 2a)	5	143
	6 Total number of volunteers (estimate if necessary)	6	500
	7a Total unrelated business revenue from Part VIII, column (C), line 12	7a	0.
	b Net unrelated business taxable income from Form 990-T, line 34	7b	0.

		Prior Year	Current Year
Revenue	8 Contributions and grants (Part VIII, line 1h)	13,967,363.	9,859,669.
	9 Program service revenue (Part VIII, line 2g)	2,339,312.	1,602,764.
	10 Investment income (Part VIII, column (A), lines 3, 4, and 7d)	797,145.	206,836.
	11 Other revenue (Part VIII, column (A), lines 5, 6d, 8c, 9c, 10c, and 11e)	414,744.	1,087,103.
	12 Total revenue - add lines 8 through 11 (must equal Part VIII, column (A), line 12)	17,518,564.	12,756,372.
Expenses	13 Grants and similar amounts paid (Part IX, column (A), lines 1-3)	1,753,547.	734,099.
	14 Benefits paid to or for members (Part IX, column (A), line 4)	0.	0.
	15 Salaries, other compensation, employee benefits (Part IX, column (A), lines 5-10)	8,012,850.	5,068,167.
	16a Professional fundraising fees (Part IX, column (A), line 11e)	266,321.	77,000.
	b Total fundraising expenses (Part IX, column (D), line 25) ▶ 1,706,730.		
	17 Other expenses (Part IX, column (A), lines 11a-11d, 11f-24e)	6,803,555.	6,747,856.
	18 Total expenses. Add lines 13-17 (must equal Part IX, column (A), line 25)	16,836,273.	12,627,122.
	19 Revenue less expenses. Subtract line 18 from line 12	682,291.	129,250.

		Beginning of Current Year	End of Year
Net Assets or Fund Balances	20 Total assets (Part X, line 16)	12,258,456.	12,876,261.
	21 Total liabilities (Part X, line 26)	3,046,269.	3,289,627.
	22 Net assets or fund balances. Subtract line 21 from line 20	9,212,187.	9,586,634.

Part II Signature Block

Under penalties of perjury, I declare that I have examined this return, including accompanying schedules and statements, and to the best of my knowledge and belief, it is true, correct, and complete. Declaration of preparer (other than officer) is based on all information of which preparer has any knowledge.

Sign Here — Signature of officer — Date

ROBIN R. GANZERT, PHD, PRESIDENT & CEO — Type or print name and title

Paid Preparer Use Only

Print/Type preparer's name	Preparer's signature	Date	Check	PTIN
CRAIG A. STEVENS, CPA	CRAIG A. STEVENS, CP	11/15/13	self-employed [ ]	P01289490

Firm's name ▶ CALIBRE CPA GROUP PLLC — Firm's EIN ▶ 47-0900880

Firm's address ▶ 7501 WISCONSIN AVENUE, SUITE 1200 WEST BETHESDA, MD 20814 — Phone no. 202-331-9880

May the IRS discuss this return with the preparer shown above? (see instructions) [X] Yes [ ] No

232001 12-10-12 LHA **For Paperwork Reduction Act Notice, see the separate instructions.** Form **990** (2012)

Problem

Can we use tax return figures to predict whether an organization will terminate all or part of its operations in subsequent 3 years?

Line 31 - Did the organization liquidate, terminate, or dissolve and cease operations?

Line 32 - Did the organization sell, exchange, dispose of, or transfer more than 25% of its net assets?

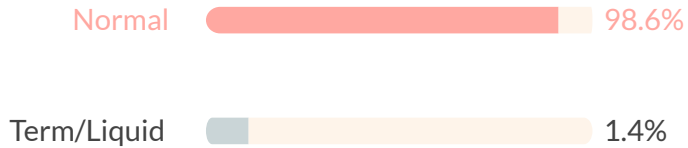
Data Pre-Processing

- ~230,000 returns from 2013
- Merge with 2012 data (yr-1)
- Create target variables
  - y_term = 1 if full termination (14-16)
  - y_liq = 1 if partial liquidation (14-16)
  - y_TL = 1 if full OR partial term/liquid (14-16)
- Processed Dataset:
  - 545 features
  - 228,181 observations
    - 3181 positive class (1.4%)

2012	Train (yr-1)
2013	train
2014	test
2015	test
2016	test

Imbalanced Classification

Imbalanced-learn is an open source, MIT-licensed library relying on scikit-learn and provides tools when dealing with classification with imbalanced classes



Resampling

downsample

PROS : - Taking less time to train

\- Giving reasonable results

CONS : - Losing valuable information

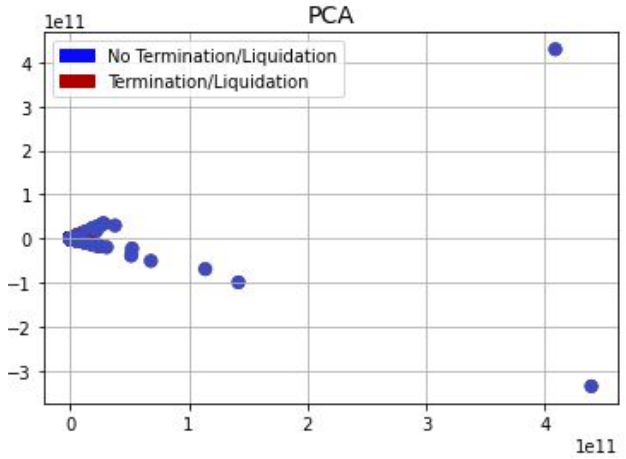
upsample

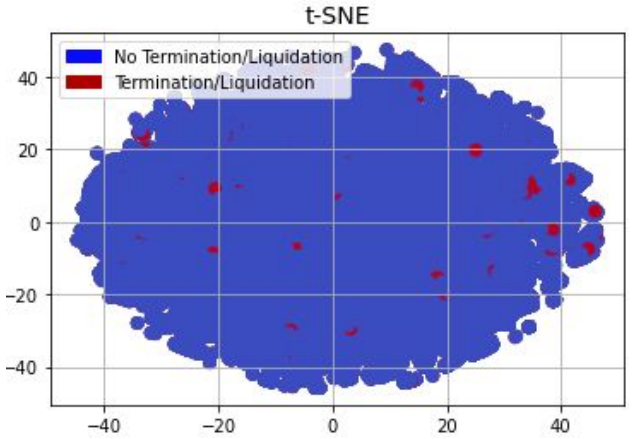
PROS : - Having enough data to train

\- Giving more accurate results

CONS : - Taking longer time to train

EDA





After dimension reduction, the positive class is hard to observe in the PCA plot

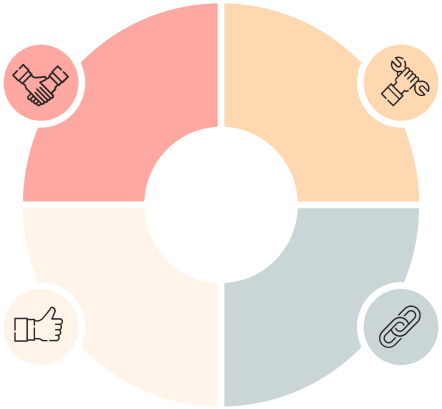
Two classes are all mixed together, and not naturally separable in the TSNE plot

Baseline Models

Multi-classes approach

Termination and Liquidation
130 records

Termination
1076 records



Normal
225000 records

Liquidation
1975 records

Logistic Regression

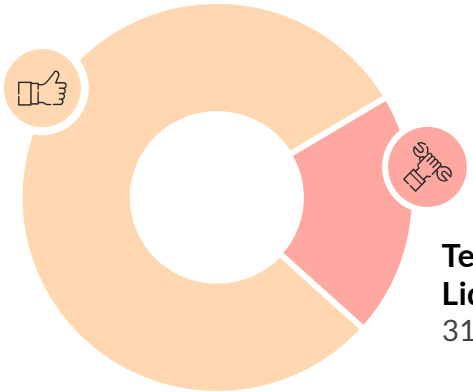
- Test accuracy : 0.5
- Test AUC: 0.64

Random Forest

- Test accuracy : 0.78
- Test AUC: 0.67

Baseline Models

Binary approach



Normal
225000 records

Termination or Liquidation
3181 records

Logistic Regression

- Test accuracy : 0.68
- Test AUC: 0.64

Random Forest

- Test accuracy : 0.71
- Test AUC: 0.66

Multi-Class classification NN

Data preprocessing & NN structure

- Unbalanced dataset handling:
  - Use resampling method to make the distribution relatively even.
  - Didn't modify the test set.

target			
3.0	179999	3.0	2000
2.0	1580	2.0	1800
1.0	861	1.0	1000
0.0	104	0.0	800

- NN structure:
  - Dense layer with 100 units & ReLU activation.
  - Dense layer with 4 units & softmax activation.
  - Compile with Adam as optimizer and categorical cross entropy as loss.

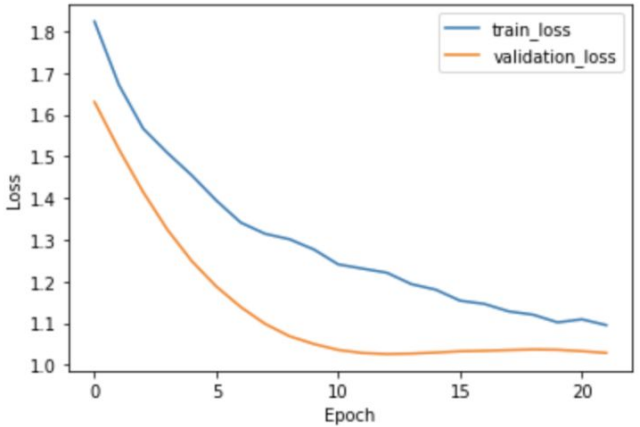
Layer (type)	Output Shape	Param #
dense_10 (Dense)	(None, 100)	54200
dropout_4 (Dropout)	(None, 100)	0
dense_11 (Dense)	(None, 4)	404

Total params: 54,604
Trainable params: 54,604
Non-trainable params: 0

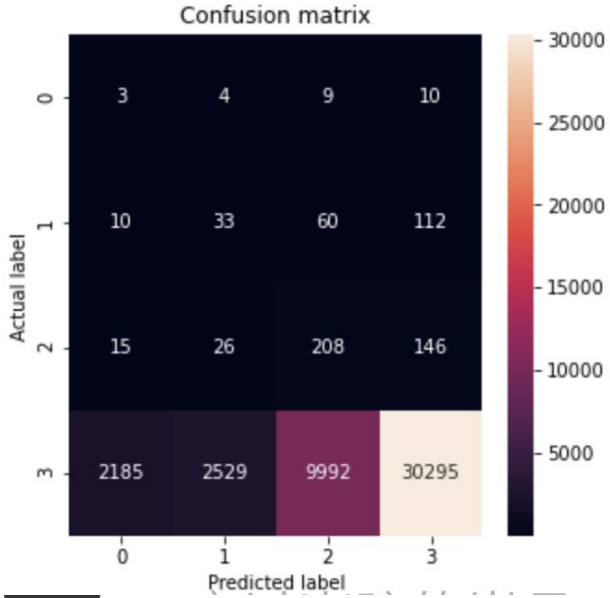
Multi-Class classification NN

NN performance

- Training:
  - Learning rate: 0.001.
  - Stopped at 22 epochs.
  - 0.78 training auc.
  - 0.85 validation auc.
- Evaluation performance:
  - 0.86 test auc



Multi-Class classification NN



True target distribution

3.0	45001
2.0	395
1.0	215
0.0	26

- Prediction skew towards 3.
- Serve misclassification especially for 0 & 1.
- Even for 3, around 25% of the observations were misclassified.

Fully-Connected Full-Sample

	AUC	TP	TN	FP	FN	Precision	Recall
10-15-Adam-relu-cw:1-72	0.620049	471.0	50518.0	16979.0	487.0	0.026991	0.491649
10-15-Adam-relu-cw:1-150	0.589754	816.0	22121.0	45376.0	142.0	0.017665	0.851775
10-15-Adam-relu-cw:1-100	0.629661	652.0	39063.0	28434.0	306.0	0.022416	0.680585
10-15-Adam-relu-cw:1-50	0.608824	352.0	57387.0	10110.0	606.0	0.033646	0.367432
10-15-20-15-10-Adam-relu-cw:1-100-dropout.5	0.632080	627.0	41151.0	26346.0	331.0	0.023245	0.654489
10-15-20-15-10-Adam-relu-cw:1-100-dropout.5-log_data	0.645021	572.0	46773.0	20724.0	386.0	0.026860	0.597077
100-15-20-15-10-Adam-relu-cw:1-100-dropout.5-log_data	0.626001	426.0	54492.0	13005.0	532.0	0.031718	0.444676
100-15-20-15-10-Adam-relu-cw:1-100-dropout.5-log_data-l1reg	0.646614	651.0	41422.0	26075.0	307.0	0.024358	0.679541
100-15-20-15-10-Adam-sigmoid-cw:1-100-dropout.5-log_data-l1reg	0.500000	958.0	0.0	67497.0	0.0	0.013995	1.000000
10-20-15-Adam-relu-cw:1-72-lr.00001-bs:500	0.644247	547.0	48430.0	19067.0	411.0	0.027888	0.570981
10-20-15-Adam-relu-cw:1-72-lr.00001-bs:50	0.636895	559.0	46592.0	20905.0	399.0	0.026044	0.583507
10-20-15-Adam-relu-cw:1-72-lr.0001-bs:150	0.607059	415.0	52710.0	14787.0	543.0	0.027299	0.433194
10-20-15-Adam-relu-cw:1-72-lr.001-bs:500	0.602207	356.0	56212.0	11285.0	602.0	0.030582	0.371608
baseline_RandomForest	0.662418	584.0	48276.0	19221.0	374.0	0.029488	0.609603

Conclusions and Limitations

Model Chosen

FC NN with 5 Hidden Layers and defined class weight

Model Performance

AUC: 0.65

Recall: 0.68

Data Imbalance

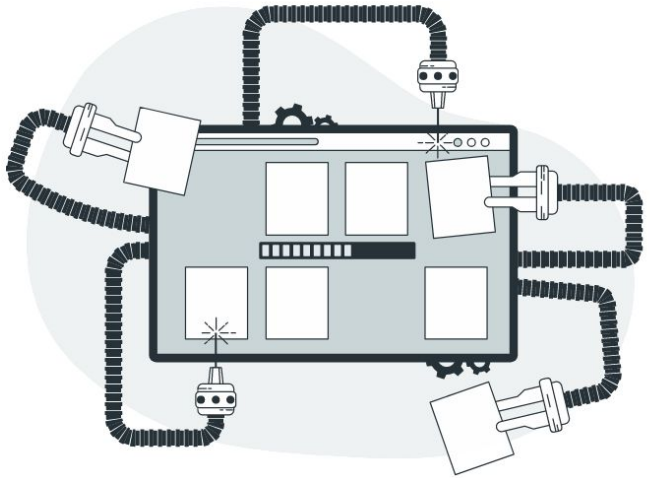
Only 1.4% of records in the positive class

Attribute Distribution

Huge variation within features

Lessons learned

- Importance of Data Preparation
- Increase in complexity does not equal improvement
- Effectiveness of the correct metrics



Q & A

Thanks!