# Machine Learning Algorithms for sourcing and evaluating VC and PE investment deals

*Yijie Sun, Prerna Khullar, Andrew Matangaidze*

yijiesun@stanford.edu
pkhullar@stanford.edu
abmatan@stanford.edu

Stanford CS229

## Motivation

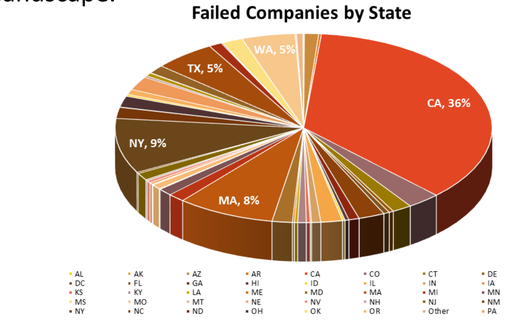
- In private investment markets, most investment decisions rely on simple ratio analysis and cash flow valuation based on financial metrics.
- By applying machine learning techniques to predict startup success, our research aims to help VC and PE investors identify deals that can potentially unlock value.

## Data

- We analyzed Pitchbook data containing 21403 companies founded in the United States between 1997-2019.
- Due to selection bias, our data is highly imbalanced, with a ratio of success/failure = 92/8.
- Growth investors target potentially successful companies, so we track the AUROC and precision metrics instead of accuracy.

## Feature Engineering

- Total 14 standardized features (excl. dummy levels): employee count, # board members, gender*, total funds raised to date*, total debt*, deal number, # tranches, # sellers, financing status*, stock type*, year founded, state, # competitors* and industry.
- Selected features represent team diversity, cash availability, regime dynamics, geography and market landscape.




* derived features

## Models

Logistic Regression
- Hypothesis: $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$
- Optimized the log loss through coordinate gradient descent.

Regularized Logistic
- L1 penalty with optimal $\lambda$ = 8.33
- L2 penalty with optimal $\lambda$ = 0.18

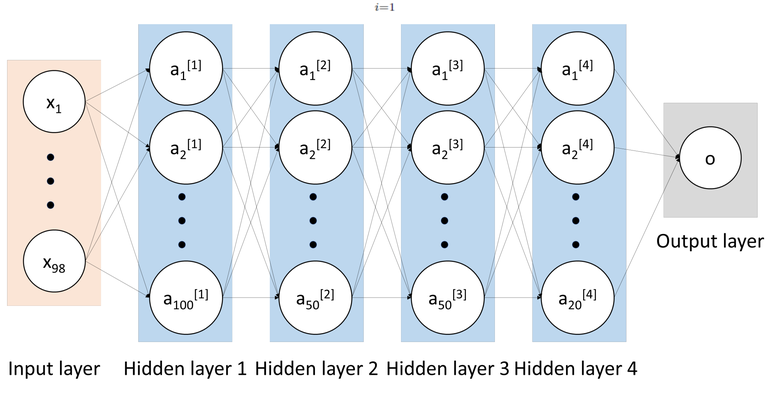
$$l(\theta) = \frac{1}{n}\sum_{i=1}^{n} y^{(i)} \log h(x^{(i)}) + (1-y^{(i)}) \log(1-h(x^{(i)})) + \lambda||\theta||_2^2$$

SVM - To speed up the computation, applied Nystroem approximation to the Laplacian kernel: $K(x,y) = exp(-\gamma||x-y||_1)$

Random Forest - Bagged 600 decision trees by considering the sqrt of total number of features at each split, with optimal max depth = 20.

NN (Multi-Layer Perceptron) - Trained a fully-connected regularized network (lambda = 0.05) with logistic activation, for 500 epochs. Hidden layers have size (100, 50, 50, 20). Weights were optimized using SGD.

$$a^{[i]} = g(W^{[i]}x + b^{[i]}) \qquad J = CE(y^{(i)}, \hat{y}^{(i)}) + \lambda\sum_{i=1}^{4}||W^{[i]}||^2$$




Note: To account for data imbalance, we used Synthetic Minority Oversampling Technique for the NN and penalized misclassifications of the minority class for all other methods. All model parameters were tuned by cross-validated grid search over the parameter grid.

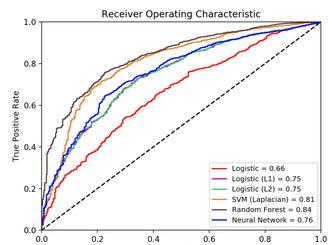
## References

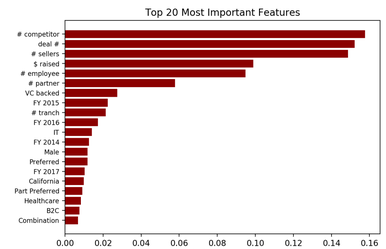
[1] J. Arroyo et al., "Assessment of Machine Learning Performance for Decision Support in Venture Capital Investments," in Ieee Access 7, 2019, pp.124233-124243.
[2] T. Prive. " Top 11 Reasons Startups Succeed." Forbes.
[3] "The Top 20 Reasons Startups Fail." CB Insights.

## Results

Data was split using stratified sampling in accordance with the 80/20 rule.

| Models \ Metrics | Train | Test | | | |
|---|---|---|---|---|---|
| | AUROC | AUROC | Precision | Recall | $F_{0.5}$ |
| Logistic | 0.63 | 0.66 | 0.92 | 1.00 | 0.94 |
| Logistic (L1) | 0.78 | 0.75 | 0.96 | 0.66 | 0.88 |
| Logistic (L2) | 0.78 | 0.75 | 0.96 | 0.66 | 0.88 |
| SVM (Laplacian) | 0.85 | 0.81 | 0.97 | 0.75 | 0.92 |
| Random Forest | 0.95 | 0.84 | 0.97 | 0.82 | 0.93 |
| Neural Network | 0.81 | 0.76 | 0.97 | 0.69 | 0.89 |




## Discussion

- Random Forest has the best performance possibly because it effectively handles categorical variables.
- SVM with a Laplacian kernel performs well due to the nonlinear transformation which makes the prediction more local.
- Oversampling the minority class in the training set may have led to a generalization gap in the test set for the neural network.
- Top numerical features: funds raised to date, # competitors, deal number, # board members and employee count.
- Top categorical features: being in California, having a female co-founder and launching in FY 2014-2017.

## Next Steps

- Acquire financial metrics and statistics on social media presence.
- Try other tree-based methods like ERT and GTB.
- Perform NLP on textual features like company description and competitors to extract in-depth information.
- Restore the detailed business status and run multi-class classification algorithms.
- Starting with a base set of features, use generative models to yield feature values that will make a startup successful.