AEA/ASSA 2020 Annual Meetings

# Use of Machine Learning Algorithms

VANDERBILT UNIVERSITY

# *Predicting Success among Female Entrepreneurs*

*Evidence from Three African Countries*

Joao Montalvao

Africa Gender Innovation Lab,
World Bank

Dario Sansone

Vanderbilt University
and University of Exeter

Friday January, 3rd 2020

# *Looking for high-growth female firms*

- **High-growth firms:** 20% of firms in manufacturing and service sectors
  - But contribute up to 80% to new sales and jobs in developing countries (Goswami et al., 2019)
- **Female entrepreneurship** seen as a way to stimulate economic growth and increase female economic empowerment (Hallward-Driemeier, 2013; Brixiová et al., 2019)
  - Access to capital key barrier limiting female entrepreneurs in poor countries (Delecourt and Ng, 2019)
- Previous attempts at finding high-growth firms based on observable info led to **lackluster results** (Goswami et al., 2019)

# *Research question*

- Can we identify successful female entrepreneurs?

- New rich and large data from **Ethiopia, Tanzania, and Togo**
- Compare simple models with heuristic models and **ML algorithms**

# *Preview of findings*

- **All models have low predicting power when focusing on profit levels in Tanzania**
  - Promising results when concentrating on top firms
  - Past profits, sales, and employment levels powerful predictors of future performance
  - ML algorithms often achieve higher performance, but results vary across algorithms, CI wide and overlapping
- **Substantially higher performance when combining data from all three countries**
  - ML algorithms can identify 45% of top firms

# *Related work*

- Almost no studies on predicting successful entrepreneurs in developing countries
  - Fafchamps and Woodruff (2017) judges' evaluations vs. survey-based measures
    - Both have some predictive power in Ghana
  - **McKenzie and Sansone (2019):** large business plan competition in Nigeria
    - Business plan scores from judges uncorrelated with business survival, employment, sales, profit levels
    - All models achieve low $R^2$ and accuracy rates
    - **No noticeable improvements from ML**

# *Contribution*

- Replicate most of the findings from McKenzie and Sansone (2019)

- Focus on female entrepreneurs
- **Larger sample size**
- **Richer data**
- Secondary use of data from multiple RCTs

# *Data*

- **First analysis: Tanzania**
  - 4,003 female microentrepreneurs (2016-2018)
  - Data on respondents’ mobile money (M-Pesa) and mobile savings/loans (M-Pawa) weekly transactions

- **Second analysis: Tanzania plus Ethiopia and Togo**
  - 2,369 female-owned middle-size firms in Ethiopia (2014-2017)
  - 789 female microentrepreneurs in Togo (2013-2016)

# *Basic models*

1. **Benchmark model with just a constant**

2. Age
3. Educations (Van Der Sluis et al., 2008; Queiro, 2016)
4. When the firm was founded (Agarwal and Gort, 2002)
5. **Baseline performance**: past profits, past sales, #employees

6. **Heuristic model** (Fafchamps and Woodruff, 2017; McKenzie and Sansone, 2019)
   - Age, marital status, education and ability, business knowledge, household wealth, risk aversion, business industry, access to credit, life satisfaction and optimism
7. Heuristic model with past performance

# *ML algorithms*

- **LASSO**
- **Support Vector Machine**
- **Boosting**
- Combine ML algorithms with **Ensemble**

- Different levels of flexibility and interpretability
- Fully exploit rich set of possible predictors
    - # predictors becomes even larger after considering how responses to certain questions should be coded (e.g. which incorrect answer one chooses)

- 5-fold CV procedure (80% training, 20% hold-out)

# *Goodness-of-fit*

- Continuous outcomes
  - **MSE**
  - Pearson correlation coefficient ($\mathbf{R^2}$)

- Binary outcomes:
  - **Accuracy**: proportion of predictions that are correct out of all observations
  - **Recall**: proportion of top firms correctly identifies

# *Profit levels and growth. Tanzania*

- Age, education, firm age: **low predictive power**

| | | | Profit levels | | | Profit growth | | |
|---|---|---|---|---|---|---|---|---|
| | | | MSE | | $R^2$ | MSE | | $R^2$ |
| | Model | Predictors | Mean | C.I. | | Mean | C.I. | |
| 1 | OLS | Constant | 23.08 | [20.05; 26.11] | 0.0% | 28,494 | [24,437; 32,551] | 0.0% |
| 2 | OLS | Age | 23.09 | [20.06; 26.13] | 0.2% | 28,545 | [24,476; 32,614] | 0.1% |
| 3 | OLS | Education | 22.99 | [19.97; 26.02] | 0.4% | 28,512 | [24,462; 32,563] | 0.0% |
| 4 | OLS | Firm age | 22.68 | [19.71; 25.66] | 1.8% | 28,522 | [24,475; 32,569] | 0.0% |
| 5 | OLS | Past performance | 22.08 | [19.11; 25.05] | 4.4% | 26,111 | [22,635; 29,586] | 8.4% |
| 6 | OLS | Heuristic | 22.95 | [20.00; 25.90] | 1.2% | 27,765 | [23,796; 31,735] | 2.6% |
| 7 | OLS | Heuristic + Past | 22.21 | [19.27; 25.14] | 4.2% | 25,518 | [22,151; 28,884] | 10.4% |
| 8 | LASSO | All baseline | 22.26 | [19.34; 25.19] | 3.6% | 24,450 | [21,235; 27,665] | 15.0% |
| 9 | SVM | All baseline | 22.76 | [19.63; 25.89] | 3.4% | 27,697 | [23,824; 31,569] | 2.7% |
| 10 | Boosting | All baseline | 22.05 | [19.13; 24.97] | 4.5% | 24,524 | [21,350; 27,698] | 13.8% |
| 11 | Ensemble | All baseline | 21.94 | [19.05; 24.83] | 5.0% | 24,360 | [21,178; 27,541] | 14.5% |

# *Profit levels and growth. Tanzania*

- Past profits, sales, #employees **reliable predictors**

| | | | Profit levels | | | Profit growth | | |
|---|---|---|---|---|---|---|---|---|
| | | | MSE | | $R^2$ | MSE | | $R^2$ |
| | Model | Predictors | Mean | C.I. | | Mean | C.I. | |
| 1 | OLS | Constant | 23.08 | [20.05; 26.11] | 0.0% | 28,494 | [24,437; 32,551] | 0.0% |
| 2 | OLS | Age | 23.09 | [20.06; 26.13] | 0.2% | 28,545 | [24,476; 32,614] | 0.1% |
| 3 | OLS | Education | 22.99 | [19.97; 26.02] | 0.4% | 28,512 | [24,462; 32,563] | 0.0% |
| 4 | OLS | Firm age | 22.68 | [19.71; 25.66] | 1.8% | 28,522 | [24,475; 32,569] | 0.0% |
| 5 | OLS | Past performance | 22.08 | [19.11; 25.05] | 4.4% | 26,111 | [22,635; 29,586] | 8.4% |
| 6 | OLS | Heuristic | 22.95 | [20.00; 25.90] | 1.2% | 27,765 | [23,796; 31,735] | 2.6% |
| 7 | OLS | Heuristic + Past | 22.21 | [19.27; 25.14] | 4.2% | 25,518 | [22,151; 28,884] | 10.4% |
| 8 | LASSO | All baseline | 22.26 | [19.34; 25.19] | 3.6% | 24,450 | [21,235; 27,665] | 15.0% |
| 9 | SVM | All baseline | 22.76 | [19.63; 25.89] | 3.4% | 27,697 | [23,824; 31,569] | 2.7% |
| 10 | Boosting | All baseline | 22.05 | [19.13; 24.97] | 4.5% | 24,524 | [21,350; 27,698] | 13.8% |
| 11 | Ensemble | All baseline | 21.94 | [19.05; 24.83] | 5.0% | 24,360 | [21,178; 27,541] | 14.5% |

# *Profit levels and growth. Tanzania*

- Heurist model **underperforms**

| | | | Profit levels | | | Profit growth | | |
|---|---|---|---|---|---|---|---|---|
| | | | MSE | | $R^2$ | MSE | | $R^2$ |
| | Model | Predictors | Mean | C.I. | | Mean | C.I. | |
| 1 | OLS | Constant | 23.08 | [20.05; 26.11] | 0.0% | 28,494 | [24,437; 32,551] | 0.0% |
| 2 | OLS | Age | 23.09 | [20.06; 26.13] | 0.2% | 28,545 | [24,476; 32,614] | 0.1% |
| 3 | OLS | Education | 22.99 | [19.97; 26.02] | 0.4% | 28,512 | [24,462; 32,563] | 0.0% |
| 4 | OLS | Firm age | 22.68 | [19.71; 25.66] | 1.8% | 28,522 | [24,475; 32,569] | 0.0% |
| 5 | OLS | Past performance | 22.08 | [19.11; 25.05] | 4.4% | 26,111 | [22,635; 29,586] | 8.4% |
| 6 | OLS | Heuristic | 22.95 | [20.00; 25.90] | 1.2% | 27,765 | [23,796; 31,735] | 2.6% |
| 7 | OLS | Heuristic + Past | 22.21 | [19.27; 25.14] | 4.2% | 25,518 | [22,151; 28,884] | 10.4% |
| 8 | LASSO | All baseline | 22.26 | [19.34; 25.19] | 3.6% | 24,450 | [21,235; 27,665] | 15.0% |
| 9 | SVM | All baseline | 22.76 | [19.63; 25.89] | 3.4% | 27,697 | [23,824; 31,569] | 2.7% |
| 10 | Boosting | All baseline | 22.05 | [19.13; 24.97] | 4.5% | 24,524 | [21,350; 27,698] | 13.8% |
| 11 | Ensemble | All baseline | 21.94 | [19.05; 24.83] | 5.0% | 24,360 | [21,178; 27,541] | 14.5% |

# *Profit levels and growth. Tanzania*

- ML: small improvements, large CI (McKenzie and Sansone, 2019; Beattie et al., 2016; Goel et al., 2010)

| | | | Profit levels | | | Profit growth | | |
|---|---|---|---|---|---|---|---|---|
| | | | MSE | | $R^2$ | MSE | | $R^2$ |
| | Model | Predictors | Mean | C.I. | | Mean | C.I. | |
| 1 | OLS | Constant | 23.08 | [20.05; 26.11] | 0.0% | 28,494 | [24,437; 32,551] | 0.0% |
| 2 | OLS | Age | 23.09 | [20.06; 26.13] | 0.2% | 28,545 | [24,476; 32,614] | 0.1% |
| 3 | OLS | Education | 22.99 | [19.97; 26.02] | 0.4% | 28,512 | [24,462; 32,563] | 0.0% |
| 4 | OLS | Firm age | 22.68 | [19.71; 25.66] | 1.8% | 28,522 | [24,475; 32,569] | 0.0% |
| 5 | OLS | Past performance | 22.08 | [19.11; 25.05] | 4.4% | 26,111 | [22,635; 29,586] | 8.4% |
| 6 | OLS | Heuristic | 22.95 | [20.00; 25.90] | 1.2% | 27,765 | [23,796; 31,735] | 2.6% |
| 7 | OLS | Heuristic + Past | 22.21 | [19.27; 25.14] | 4.2% | 25,518 | [22,151; 28,884] | 10.4% |
| 8 | LASSO | All baseline | 22.26 | [19.34; 25.19] | 3.6% | 24,450 | [21,235; 27,665] | 15.0% |
| 9 | SVM | All baseline | 22.76 | [19.63; 25.89] | 3.4% | 27,697 | [23,824; 31,569] | 2.7% |
| 10 | Boosting | All baseline | 22.05 | [19.13; 24.97] | 4.5% | 24,524 | [21,350; 27,698] | 13.8% |
| 11 | Ensemble | All baseline | 21.94 | [19.05; 24.83] | 5.0% | 24,360 | [21,178; 27,541] | 14.5% |

- **Mobile data** among selected predictors (Björkegren and Grissen, 2019)

# *Top firms. Tanzania*

- Can we identify firms in the **top 10%** of the profit distribution?

| | | | Profit levels | | | Profit growth | | |
|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | Recall | Accuracy | | Recall |
| | Model | Predictors | Mean | C.I. | | Mean | C.I. | |
| 1 | OLS | Constant | 81.5% | [79.4%, 83.6%] | 11.8% | 80.1% | [77.9%, 82.3%] | 7.0% |
| 2 | OLS | Age | 79.9% | [77.7%; 82.2%] | 4.7% | 80.6% | [78.4%; 82.9%] | 9.3% |
| 3 | OLS | Education | 81.3% | [79.1%; 83.4%] | 10.6% | 80.4% | [78.0%; 82.7%] | 8.1% |
| 4 | OLS | Firm age | 81.5% | [79.5%, 83.5%] | 11.8% | 80.4% | [78.2%; 82.5%] | 8.1% |
| 5 | OLS | Past performance | 85.9% | [83.8%; 88.1%] | 31.8% | 84.9% | [82.6%; 87.2%] | 27.9% |
| 6 | OLS | Heuristic | 83.1% | [80.8%; 85.3%] | 18.8% | 84.6% | [82.2%; 87.1%] | 26.7% |
| 7 | OLS | Heuristic + Past | 85.2% | [83.0%; 87.3%] | 28.2% | 86.0% | [83.7%; 88.2%] | 32.6% |
| 8 | LASSO | All baseline | 84.6% | [82.4%; 86.9%] | 25.9% | 88.1% | [85.8%; 90.4%] | 41.9% |
| 9 | SVM | All baseline | 83.6% | [81.2%; 86.0%] | 21.2% | 83.8% | [81.6%; 86.1%] | 23.3% |
| 10 | Boosting | All baseline | 87.2% | [85.1%; 89.4%] | 37.6% | 85.7% | [83.4%; 88.0%] | 31.4% |
| 11 | Ensemble | All baseline | 85.2% | [82.9%; 87.4%] | 28.2% | 86.8% | [84.6%; 89.0%] | 36.0% |

- Promising results

# *Profit levels and growth. Pooled*

- Substantial improvement in ML performance, especially for profit growth

| | | | Profit levels | | | Profit growth | | |
|---|---|---|---|---|---|---|---|---|
| | | | MSE | | $R^2$ | MSE | | $R^2$ |
| | Model | Predictors | Mean | C.I. | | Mean | C.I. | |
| 1 | OLS | Constant | 5.27 | [4.85; 5.71] | 0.0% | 29,894 | [26,666; 33,122] | 0.0% |
| 2 | OLS | Age | 5.26 | [4.82; 5.69] | 0.5% | 29,866 | [26,638; 33,095] | 0.1% |
| 3 | OLS | Education | 5.26 | [4.83; 5.70] | 0.3% | 29,855 | [26,627; 33,084] | 0.1% |
| 4 | OLS | Firm age | 5.16 | [4.73; 5.59] | 2.2% | 29,828 | [26,612; 33,044] | 0.3% |
| 5 | OLS | Past performance | 4.96 | [4.49; 5.44] | 6.3% | 27,484 | [24,552; 30,415] | 8.8% |
| 6 | OLS | Heuristic | 5.23 | [4.79; 5.67] | 1.0% | 30,039 | [26,815; 33,263] | 0.0% |
| 7 | OLS | Heuristic + Past | 4.93 | [4.46; 5.40] | 7.0% | 27,544 | [24,621; 30,467] | 8.3% |
| 8 | LASSO | All baseline | 4.76 | [4.31; 5.20] | 9.9% | 26,152 | [23,456; 28,849] | 12.9% |
| 9 | SVM | All baseline | 4.91 | [4.39; 5.43] | 10.2% | 26,979 | [23,967; 29,992] | 10.7% |
| 10 | Boosting | All baseline | 4.76 | [4.30; 5.22] | 10.1% | 23,007 | [20,504; 25,510] | 23.7% |
| 11 | Ensemble | All baseline | 4.73 | [4.27; 5.18] | 10.7% | 23,043 | [20,551; 25,534] | 23.5% |

# *Top firms. Pooled*

- **Correctly identify 45% high-growth firms**

| | | | Profit levels | | | Profit growth | | |
|---|---|---|---|---|---|---|---|---|
| | | | Accuracy | | Recall | Accuracy | | Recall |
| | Model | Predictors | Mean | C.I. | | Mean | C.I. | |
| 1 | OLS | Constant | 81.5% | [79.8%; 83.1%] | 8.1% | 80.4% | [78.8%, 82.1%] | 7.5% |
| 2 | OLS | Age | 81.5% | [79.9%; 83.0%] | 8.1% | 81.4% | [79.7%, 83.0%] | 11.6% |
| 3 | OLS | Education | 82.3% | [80.8%; 83.9%] | 12.5% | 81.2% | [79.5%, 82.9%] | 11.0% |
| 4 | OLS | Firm age | 81.8% | [80.3%; 83.2%] | 9.6% | 81.0% | [79.3%, 82.8%] | 10.3% |
| 5 | OLS | Past performance | 88.2% | [86.6%; 89.7%] | 41.2% | 86.6% | [84.9%; 88.3%] | 34.9% |
| 6 | OLS | Heuristic | 83.5% | [82.0%; 85.1%] | 18.4% | 81.0% | [79.4%; 82.7%] | 10.3% |
| 7 | OLS | Heuristic + Past | 87.4% | [85.8%; 89.0%] | 37.5% | 85.1% | [83.2%; 86.9%] | 28.1% |
| 8 | LASSO | All baseline | 87.4% | [85.8%; 89.0%] | 37.5% | 86.6% | [85.0%; 88.2%] | 34.9% |
| 9 | SVM | All baseline | 88.2% | [86.5%; 89.8%] | 41.2% | 85.2% | [83.5%; 86.9%] | 28.8% |
| 10 | Boosting | All baseline | 87.6% | [86.0%; 89.1%] | 38.2% | 88.6% | [86.9%; 90.2%] | 43.8% |
| 11 | Ensemble | All baseline | 88.0% | [86.5%; 89.5%] | 40.4% | 88.8% | [87.1%; 90.4%] | 44.5% |

# *Investment simulation*

- **3x higher returns** than randomly picking firms

| | Model | Predictors | Investment Simulation | |
|---|---|---|---|---|
| | | | Mean | C.I. |
| 1 | OLS | Constant | 19,801 | [10,531; 29,071] |
| 2 | OLS | Age | 14,890 | [6,314; 23,465] |
| 3 | OLS | Education | 19,017 | [13,832; 24,202] |
| 4 | OLS | Firm age | 16,548 | [10,863; 22,232] |
| 5 | OLS | Past performance | 58,487 | [42,514; 74,460] |
| 6 | OLS | Heuristic | 29,012 | [18,338; 39,686] |
| 7 | OLS | Heuristic + Past | 56,208 | [41,119; 71,296] |
| 8 | LASSO | All baseline | 56,982 | [42,069; 71,894] |
| 9 | SVM | All baseline | 61,866 | [45,870; 77,863] |
| 10 | Boosting | All baseline | 55,319 | [39,950; 70,689] |
| 11 | Ensemble | All baseline | 59,705 | [42,960; 76,450] |

# *Conclusions*

- **Difficult** to predict successful entrepreneurs using survey data: $R^2$ always below 11% for profit levels
- **ML algorithms can do significantly better** than basic and heuristic models
  - Requirement: large and rich data

- Currently collaborating with fintech company to further develop and distribute a ML algorithm

- Future research:
  - Use predictions from ML algorithms as preliminary step in RCTs (Chandler et al., 2011)
  - Incorporate ML predictions in human decisions

# *Thank you!*

*Review ML literature on my website*

*@SansoneEcon*

VANDERBILT UNIVERSITY