Machine learning and job matching

Neil Rankin

August 21, 2017

Goal

- Can we build a ‘machine learning’ algorithm to identify young people who are likely to become employed in South Africa?

Outline

- What is machine learning (ML)?
- Can we apply this in a South African labour market environment?
- Building a prototype ML algorithm

What is machine learning (ML)?

‘Machine learning is the subfield of computer science that, according to Arthur Samuel, gives “computers the ability to learn without being explicitly programmed”. . . Evolved from the study of pattern recognition and computational learning theory in artificial intelligence, machine learning explores the study and construction of algorithms that can learn from and make predictions on data – such algorithms overcome following strictly static program instructions by making data-driven predictions or decisions, through building a model from sample inputs. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms with good performance is difficult or infeasible; example applications include email filtering, detection of network intruders or malicious insiders working towards a data breach, optical character recognition (OCR), learning to rank, and computer vision.’

Wikipedia

Applying this in a South African labour market environment

Harambee Youth Employment Accelerator is an organisation which brings together young people, who are marginalised in the labour market, and employers looking for entry-level recruits.

Their reputation depends on finding good, but overlooked, candidates. They need a mechanism to find these candidates? Can we help?

The Harambee process

1. 'Source' people either through actively recruiting or through sign-ups trhough a mobile site
2. Invite people to further assessments and support:
   - 'Work-seeker support' which includes interview preparation, help with job search etc
   - 'Learning Potential' or CFT. A non-verbal reasoning test.
   - Communications and numeracy tests (not all candidates get this, and these can vary between candidates)
3. 'Bridge' with job-specific skills
4. Facilitate placement (for those going through the whole process), or self-placement (who find a job after work-seeker support)
5. 'Employment Journey' survey every 4 months after initial contact

Can we find characteristics which can help to pick candidates to go through this process?

Building a prototype ML algorithm (for employment)

Steps:

1. Split the dataset between training and testing
2. Train a model on the training data
   - OLS
   - Logit (?)
   - Random forest (?)
3. Evaluate it on the test data

See the do file, R script

ML algorithm, step 1

Split the data into a training and testing dataset

1. Set the random number seed
2. Generate a random number
3. Split into training and testing

ML algorithm, step 1 (STATA)

```
set seed 12345 ssc install randomtag randomtag
```

ML algorithm, step 1 (R)

```r
### split out training and test data
set.seed(1234)
## Using 70% as training currently

# check dimensions of data
dim(ej_data)
```

```
## [1] 22304    11
```

```r
#Sample Indexes
indexes = sample(1:nrow(ej_data), size=0.3*nrow(ej_data))

# Split data and just checking dimensions
ej_test = ej_data[indexes,]
dim(ej_test)
```

```
## [1] 6691   11
```

Training the model

===

employed

english_mark

0.002***
(0.0005)

as.numeric(CFT_answer_ODS) -0.009***
(0.002)

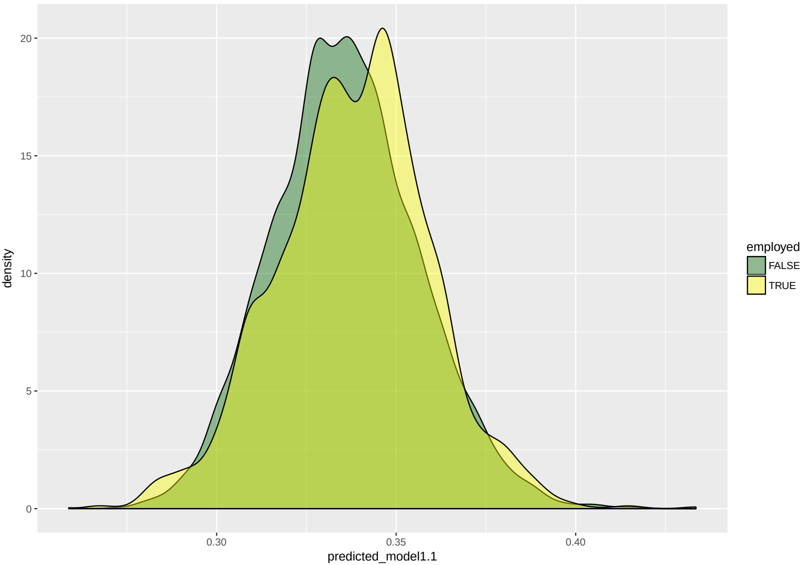
Constant 0.296***
(0.025)

N 11,584
R2 0.002
Adjusted R2 0.002
Residual Std. Error 0.472 (df = 11581)
F Statistic 11.176*** (df = 2; 11581)

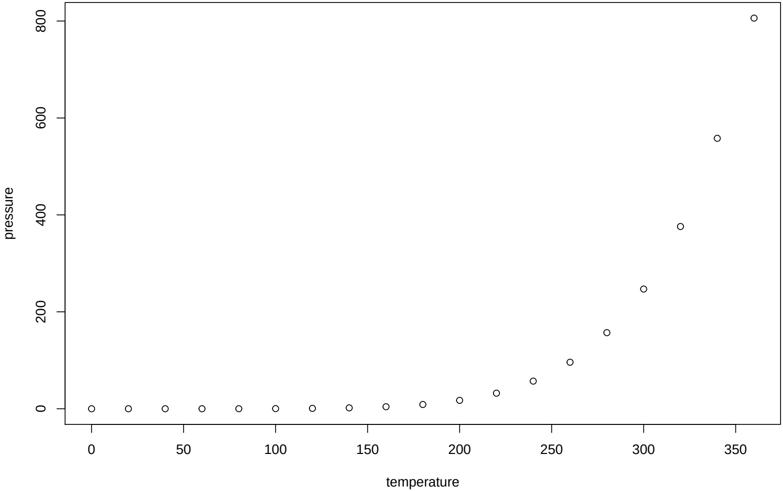
===

Notes: ***Significant at the 1 percent level.*** *Significant at the 5 percent level.* Significant at the 10 percent level.

Testing the model (and a picture)



Slide with Plot


pressure
temperature