Exploring business growth with machine learning

Cathy Atkinson
Data Science Team
Data Analytics and Business Statistics
2 October 2018

Department for Business, Energy & Industrial Strategy

Outline

- Project introduction
- Inter-Departmental Business Register data
- Machine learning
- Technique 1 – Logistic regression
- Technique 2 - Gradient boosted trees
- Technique 3 – Neural networks
- Results
- Questions

Project introduction

- Data Enabled Change Accelerator project
- Helping businesses to grow is a key aim of BEIS
- Exploring whether machine learning techniques can accurately predict which businesses will become High Growth Firms to enable targeting of firms for support
- OECD definition of a High Growth Firm is a business with 10 or more employees which grows 20% or more on average over 3 years either in turnover or employment

The wider project

- Collaboration between:
  - BEIS Business Growth team
  - BEIS Data Science team
  - HMRC
  - ONS Data Science Campus

- This presentation focuses solely on the work of the BEIS Data Science team and the techniques used

IDBR data

- The best source of data available within BEIS is the ONS Inter-Departmental Register (IDBR)
- BEIS hold quarterly extracts back to 2007
- Can only be used for statistical and research purposes
- Designed to be a sample frame for surveys, not for economic analysis
- Variety of data sources with differing timescales

Longitudinal IDBR

- Analysed the quality of source and date of each data entry for employment and turnover
- Produced longitudinal datasets of employment and turnover for each enterprise from 2006-2016
- Agreed the methodology with ONS
- Focussed on employees
- Used 2016 as outcome year and 2013 as base year

Variables

- Proxy for age of business
- No. employees or size-band
- Sector (from SIC 2007 code)
- Legal status (e.g. company, partnership, etc.)
- No. PAYE or VAT units (proxy for structure / complexity)
- Number of different premises (local units)
- Region of HQ & percentage of premises in each region
- Growth history

Question

Using what we knew about these businesses in 2013, can we predict which would meet the High Growth Firm criteria in 2016?

Data exploration

- Don’t leap straight into the machine learning
- Explore the data and the relationships in it
- Produce plots
- Think about the variables and impact on model

Machine learning

- Supervised machine learning classification problem approach
- Treated classification is binary – high growth at the end of the 3 year growth period or not
- Researched a range of techniques
- Short-listed 3 with different properties

Issues to be aware of

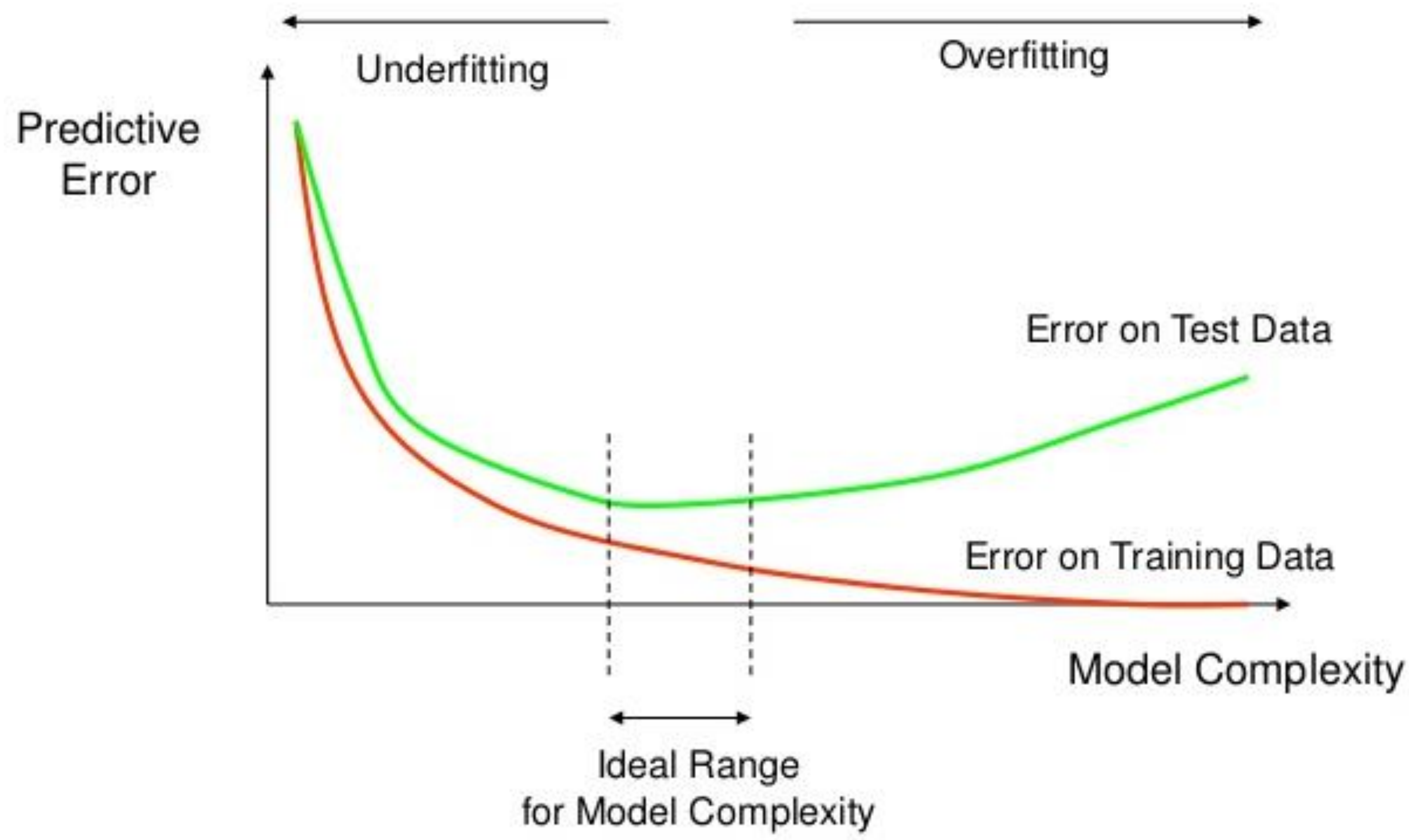
- Data is very unbalanced
- How to compare models
- Interpretability of models
- Overfitting

Data preparation

- Correct data format depends on technique, often need a design matrix:
  - One row per observation
  - Categorical predictor variables are ‘one-hot encoded’; each category a separate column with 1 to indicate the category the observation is in and 0 for all others
  - One category for each variable not included as this is the intercept
  - Continuous predictor variables usually standardised (scaled so all values between 0 and 1, or making mean 0 and SD 1)

Data splits

- Split data into 3:
  - **Training** - 65% - used to train the model
  - **Test** - 15% - used to assess the performance of the model on untrained data
  - **Validation** - 20% - used at the end to assess performance of final model
- For some techniques '**cross-validation**' used



***Source**: Penn State - Applied Data Mining and Statistical Learning*
https://onlinecourses.science.psu.edu/stat857/node/160/

Measuring success

- Models predict probability of the outcome being 1 (e.g. high growth); call this the **'score'**
- Choose a **'cut-off'** – a value for the score above which the outcome is predicted to be 1 (the rest being 0).
- Selected cut-off whereby top 20% scoring enterprises predicted to be high growth

Confusion matrix

		Predicted outcome	
		1	0
True outcome	1	True positive	False negative (Type II error)
	0	False positive (Type I error)	True negative

Measures

Accuracy

- Proportion of cases correctly classified
- Only valid when 50/50 split in actual outcomes

Recall

- Proportion of true positives predicted to be positive

Precision

- Proportion of those predicted to be positive that are truly positive

**** Important to look at both precision and recall ****

1. Logistic regression

- Easily interpretable models
- Model selection is difficult when many parameters
- Collinearity in the predictor variables can cause issues with model fit and estimates
- Outliers can lead to overfitting
- Regularisation is a way of introducing extra information (aka ‘hyperparameters’) into the model
- Ridge regression, lasso regression and elastic nets are forms of regularisation

Ridge & lasso regression

Ridge regression

- Predictor variables are ‘shrank’ rather than dropped

Lasso regression

- **L**east **A**bsolute **S**hrinkage and **S**election **O**perator
- With correlated predictors lasso will tend to pick one and discard the others
- Expected to perform well when there’s a small number of true predictors affecting the response variable

Elastic nets

- Elastic nets combine both ridge regression and lasso
- Run models with different weights between ridge and lasso regression, different amounts of shrinkages etc.
- Cross-validation
- Take model with minimum loss figure
- Calculate precision and recall

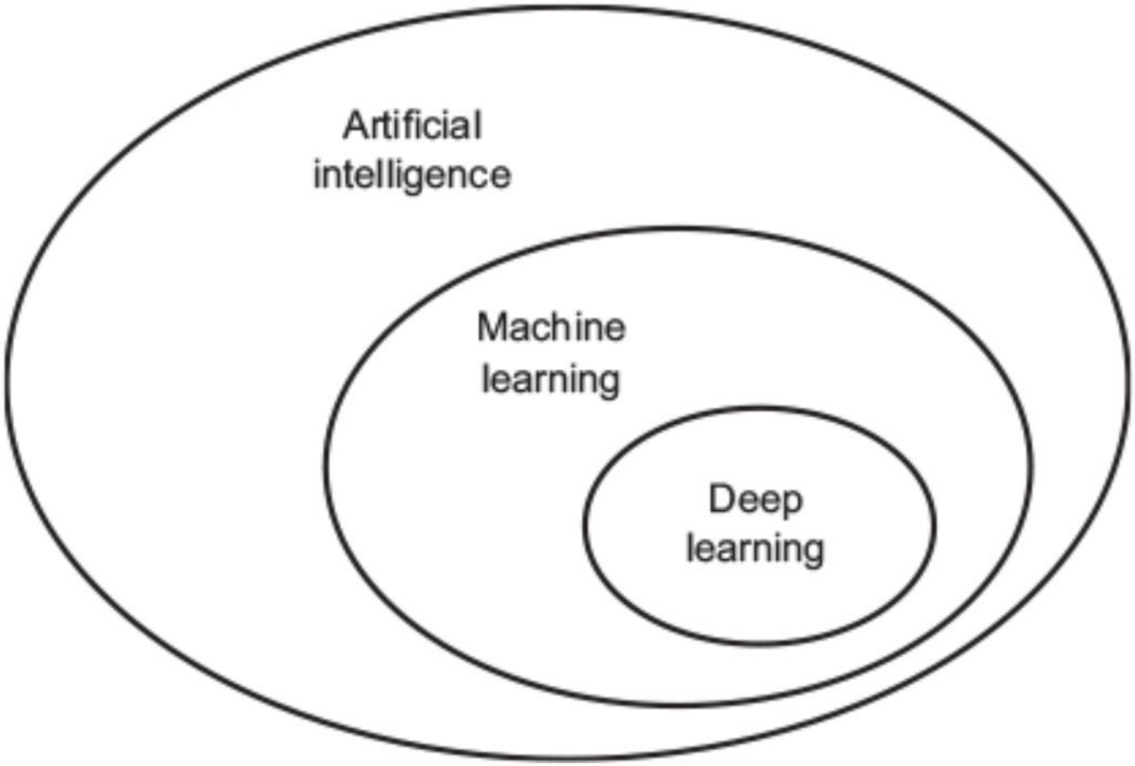
2. Gradient boosted trees

- Decision trees with gradient boosting in xgboost (e**X**treme **G**radient **Boost**ing package)
- Ensemble models that are trained sequentially
- Multiple decision trees with trees fitted to the errors of the previous trees
- Several parameters to tune
- Interpretability generally not very easy though depends on parameters used

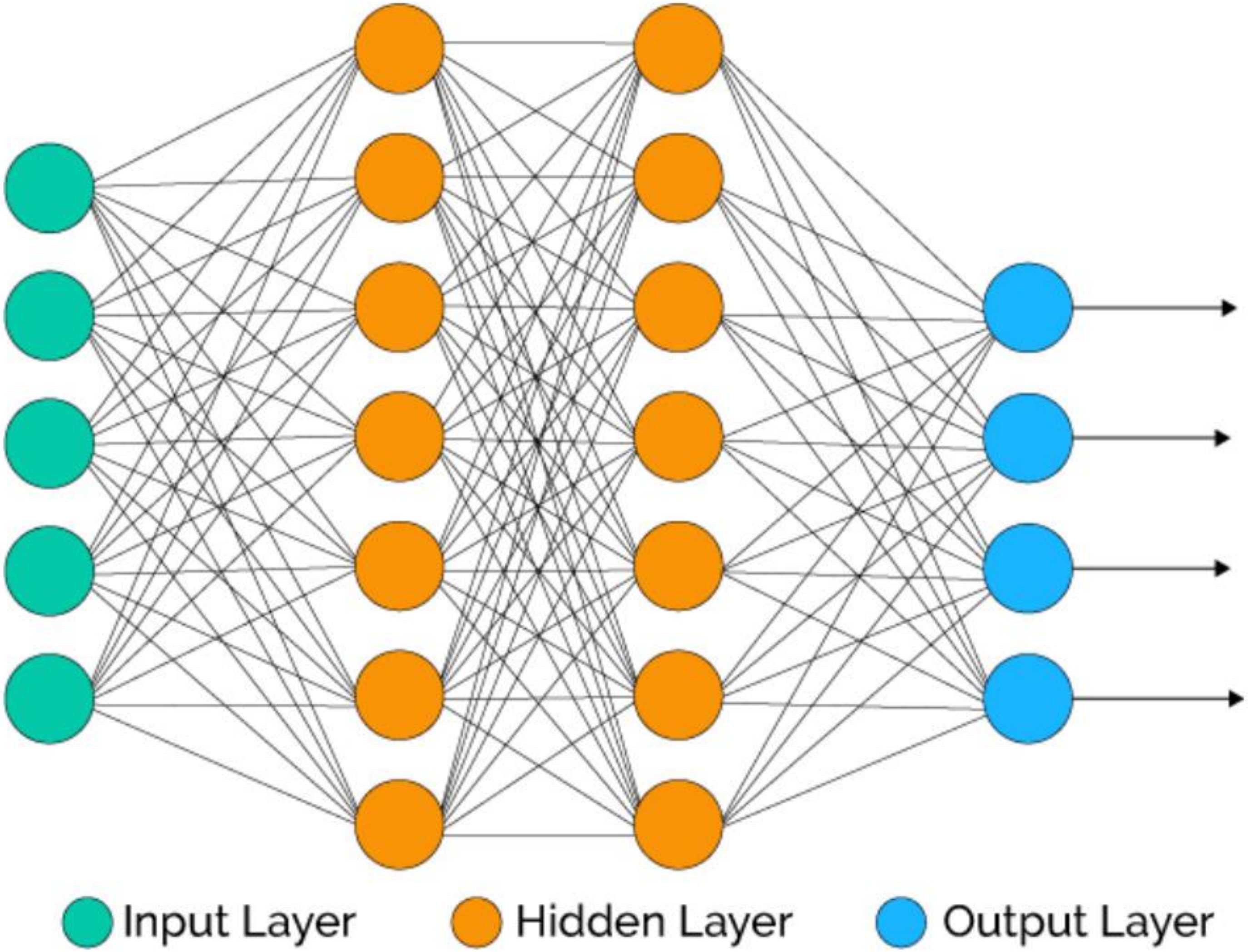
3. Neural networks

- Artificial neural networks are models based upon non-linear parameterisation of the input parameters
- Extremely powerful
- Models hard to interpret
- Interactions are automatically fitted and not possible to control

Deep learning



***Source**: ‘Deep Learning with R’ by François Chollet with J. J. Allaire*
[*https://www.manning.com/books/deep-learning-with-r*](https://www.manning.com/books/deep-learning-with-r)



Source: *http://blogs.rstudio.com/tensorflow/posts/2018-01-11-keras-customer-churn/*

Deep learning with keras

- TensorFlow developed by the Google Brain team
- Keras runs on top of TensorFlow as a high level interface specifically for neural networks
- Depth relates to number of layers in the model
- Each layer is a simple data transformation specified by weights; the optimal weights are 'learnt'
- For more information see: https://keras.rstudio.com/

Technique summary

	Logistic regression with ridge regression/ lasso	Decision trees with gradient boosting in xgboost	Neural networks with tensorflow/keras
Quick model summary	Standard logistic regression with all predictor variables included in the model. Predictor variables are 'shrank' rather then dropped.	Multiple decision trees with trees fitted to the errors from previous trees.	Regressions fed into other regressions. Outputs from the previous regression are transformed to create interactions.
Interactions	Only if explicitly entered	Automatically fitted. Can be controlled with depth of trees.	Automatically fitted and not possible to control
Number of parameters to 'tune'	Very few	Medium amount	Lots
Interpretability	Easiest, though there may be a lot of parameters	Generally not very easy though depends on parameters used.	Hardest

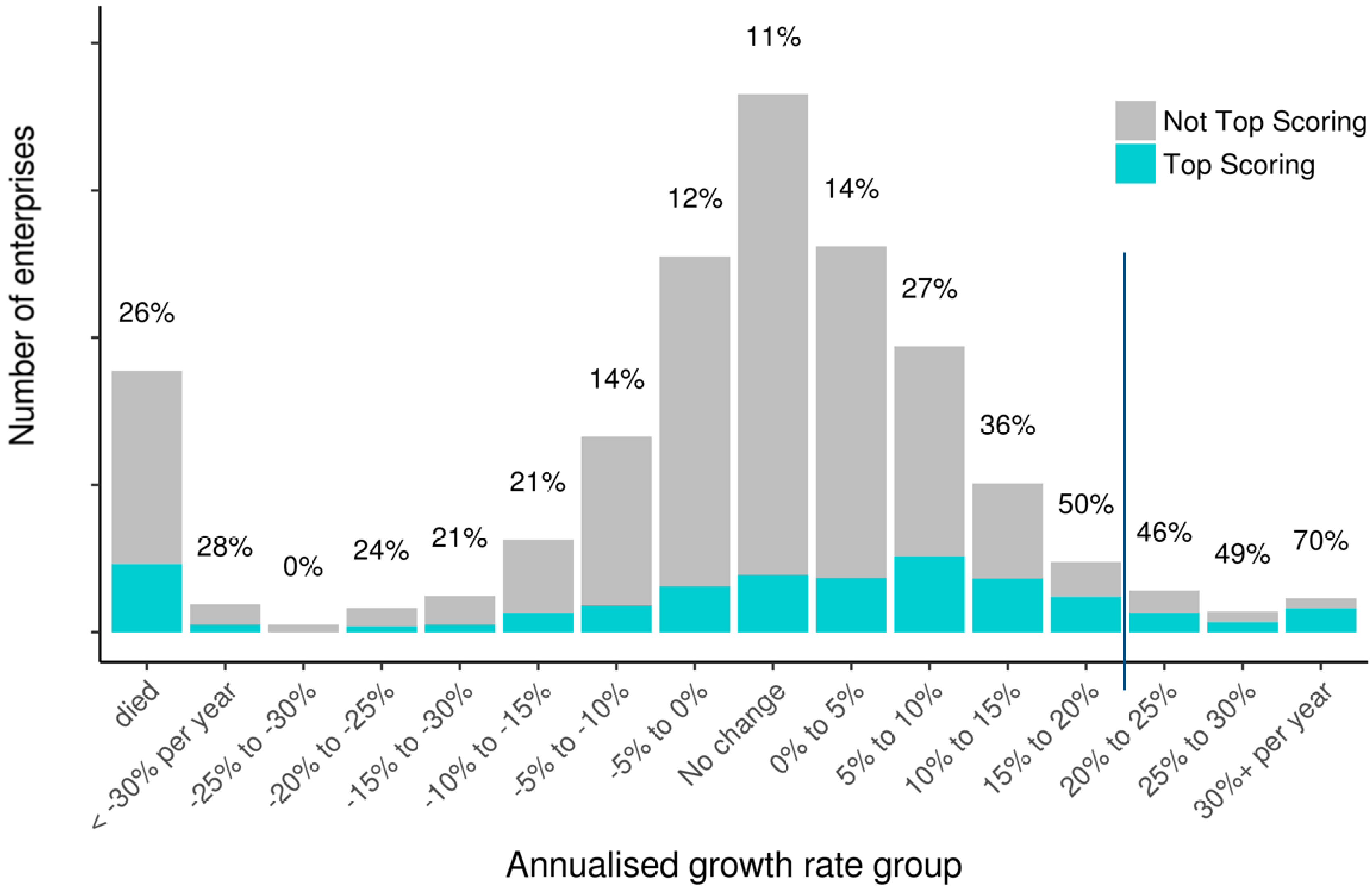
Indicative results

E.g. Using the model created for a sector where 5% will go on to be high growth, if the top 20% scoring enterprises were contacted:

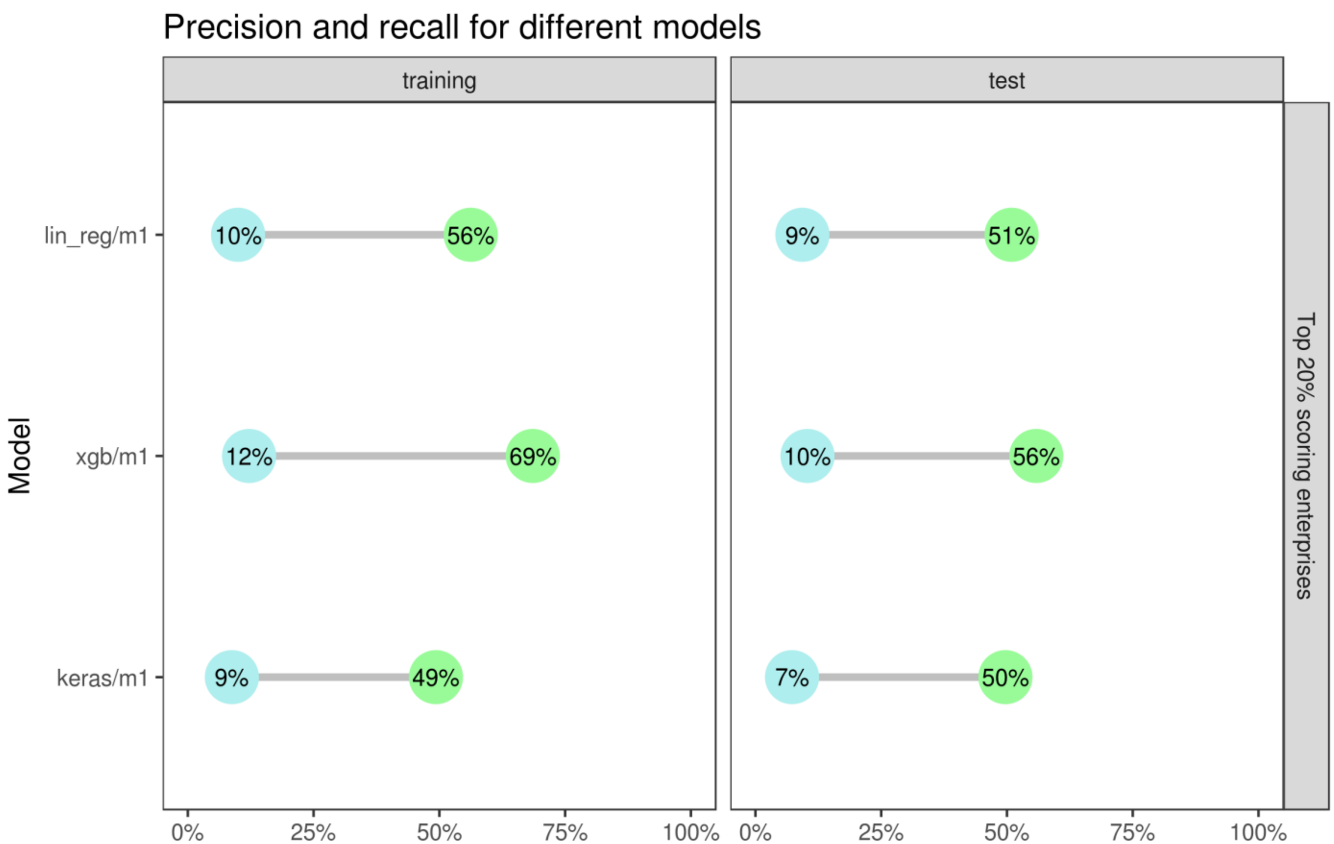
- ~20% of those contacted would go on to be high growth
- ~50% of the high growth enterprises would be within the ones contacted

Example of results presentation – for illustrative purposes only

Labels show the percentage of growth rate group which are high scoring enterprises



Comparing techniques



Recall
Proportion of true positives predicted to be positive

Precision
Proportion of those predicted to be positive that are truly positive

Model robustness

- Similar results from:
  - 3 different sectors
  - Out of time sample (applied the model created with 2013-2016 data to 2010-2013 data)
  - Changing the definition of high growth from 20% per year to 15%

Any questions?

cathy.atkinson@beis.gov.uk