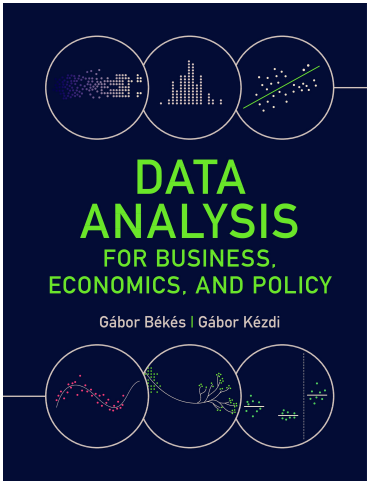
6. Firm exit case study

Gabor Bekes

Data Analysis 3: Prediction

2021

Slideshow for the Békés-Kézdi Data Analysis textbook



- Cambridge University Press, 2021 April
- Available in paperback, hardcover and e-book
- **gabors-data-analysis.com**
  - Download all data and code `https://gabors-data-analysis.com/data-and-code/`
- This slideshow is for **Chapter 17** case study - with some extensions

Firm exit case study: Case study: background

- Banks and business partners are often interested in the stability of their customers.
- Predicting which firms will be around to do business with is an important part of many prediction projects.
- Working with financial and non-financial information, your task may be to predict which firms are more likely to default than others.

Firm exit case study: business case

- Suppose we work for a consultancy, whose aim is to advise banks on client selection or purchasing managers on supplier selection.
- "We do business with a firm, is this firm going to be around in the near future?" - they may ask.

- Our aim is to predict corporate default - exit from the market.
- That is it.
  - Not more specific.
  - We have to figure out and decide on target, features, etc.

Firm exit case study: **bisnode-firms** dataset

- Firm data
- Many different type of variables
  - Financial
  - Management
  - Ownership
  - Status (HQ)
- Dataset is a panel data
  - We created earlier
  - Rows are identified by company id (comp-id) and year.
- We'll focus on a cross-section of 2012.

Firm exit case study: Label (target) engineering

- Defining our target.
- In the data, there is no "exit" - we have to define it!
- A firm is operational in year t, but is not in business in $t + 2$.
- The target is hence a binary variable called exit,
  - 1 if the firm exited within 2 years
  - 0 otherwise.
- This definition is broad
  - Defaults / forced exit
  - Orderly closure
  - Acquisitions

Firm exit case study: Sample design

- Look at a cross section
  - Year=2012
  - status_alive=1
  - Keep if established in 2012
- We do not care about all firms. Not very small and very large
  - Below 10 million euros
  - Above 1000 euros
- Hardest call: keep when important variables are not missing
  - Balance sheet like liquid assets
  - Ownership like foreign
  - Industry classification
- End with 19K observation, 20% default rate

Firm exit case study: Features - overview

- Key predictors
  - size: sales, sales growth
  - management: foreign, female, young, number of managers
  - region, industry, firm age
  - other financial variables from the balance sheet and P&L.
- For financial variables, we use ratios (to sales or size of balance sheet).
- Here it will turn out be important to look at functional form carefully, especially regarding financial variables.
- Mix domain knowledge and statistics.

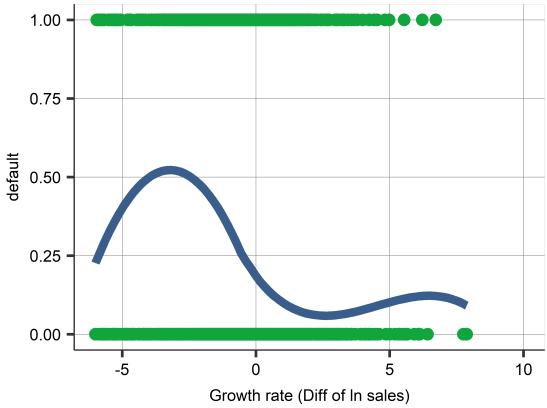
Firm exit case study: Feature engineering

- Growth rates
  - 1 year growth rate of sales. Log difference.
  - Could use longer time period. Lose observations
  - Should depend on client needs. Maybe: I am interested in 3y+ firms.
- Ownership, management info sometimes weak
  - Could drop additional firms
  - Again, depends on business
- Sometimes simplify (unless big data)
  - ceo_young = ceo_age_mod<40 & ceo_age_mod>15
  - Industry categories - too many, need merge
  - Foreign ownership - above a threshold
- Functional form - logs, polynomials
  - Look at relationships in scatterplot, loess and decide

Firm exit case study: Feature engineering

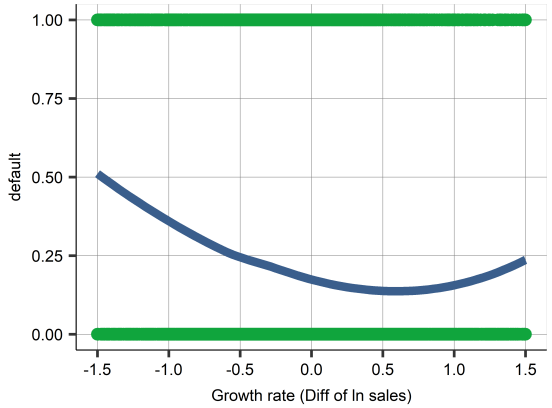
- May need to make cleaning steps.
- Create binary variables (flags) when implementing changes to values.
- When financial values are negative: replace with zero and add a flag to capture imputation.
- Make changes
  - for values that may have additional information in non-linear way
  - Value is exactly 0

Firm exit case study: Firm sales growth



- Annual growth in sales (difference in log sales) vs default
- Weird shape...

Firm exit case study: Firm sales growth

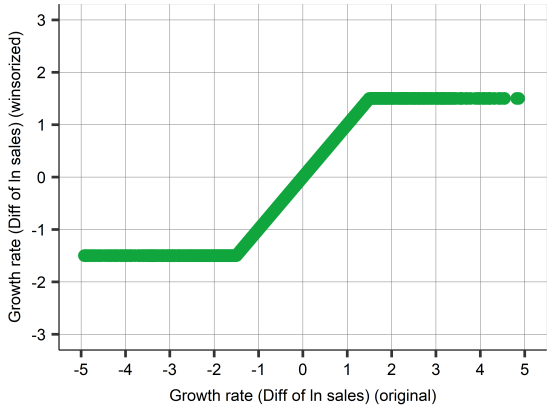


- Annual growth in sales vs default
- Weird shape...
- ... because of extremes, really
- few firms below, say -1.5 and above 1.5
- The rest looks ok

Firm exit case study: Winsorizing

- When edge of a distribution is weird
  - Not just a u-shaped polynomial
  - Domain knowledge helps!
- Winsorizing is a process to keep observations with extreme values in sample
- for each variable, we
  - identify a threshold value, and replace values outside that threshold with the threshold value itself
  - and add a flag variable.
- Two ways to do it:
  - an automatic approach, where the lowest and highest 1 percent or 5 percent is replaced and flagged.
  - Pick thresholds by domain knowledge as well as by looking at lowess. Preferred.

Firm exit case study: Firm sales growth



- The winsorized value simply equals original value in a range and flat below/after.

Case study: firm exit: Model features 1

- **Firm**: Age of firm, squared age, a dummy if newly established, industry categories, location regions for its headquarters, and dummy if located in a big city.
- **Financial 1**: Winsorized financial variables: fixed, liquid (incl current), intangible assets, current liabilities, inventories, equity shares, subscribed capital, sales revenues, income before tax, extra income, material, personal and extra expenditure.
- **Financial 2**: Flags (extreme, low, high, zero - when applicable) and polynomials: Quadratic terms are created for profit and loss, extra profit and loss, income before tax, and share equity.
- **Growth**: Sales growth is captured by a winsorized growth variable, its quadratic term and flags for extreme low and high values.

Firm exit case study: Model features 2

- **HR**: For the CEO: female dummy, winsorized age and flags, flag for missing information, foreign management dummy; and labor cost, and flag for missing labor cost information.
- **Data Quality**: Variables related to the data quality of the financial information flag for a problem, and the length of the year that the balance sheet covers.
- **Interactions**: Interactions with sales growth, firm size, and industry.

Firm exit case study: Models

Models (number of predictors)

- Logit M1: handpicked few variables ($p = 11$)
- Logit M2: handpicked few variables + Firm ($p = 18$)
- Logit M3: Firm, Financial 1, Growth ($p = 35$)
- Logit M4: M3 + Financial 2 + HR + Data Quality ($p = 79$)
- Logit M5: M4 + interactions ($p = 153$)
- Logit LASSO: M5 + LASSO ($p = 142$)

  - Number of coefficients = N of predictors +1 (constant)

Firm exit case study: Data

- $N = 19,036$
- $N = 15,229$ in work set (80%)
  - Cross validation 5x training + test sets
  - Used for cross-validation
- $N = 3,807$ in holdout set (20%)
  - Used only for diagnostics of selected model.

Firm exit case study: Comparing model fit

	Variables	Coefficients	CV RMSE
Logit M1	4	12	0.374
Logit M2	9	19	0.366
Logit M3	22	36	0.364
Logit M4	30	80	0.362
Logit M5	30	154	0.363
Logit LASSO	30	143	0.362

- *5-fold cross-validated on work set, average RMSE*

Firm exit case study: Comparing model fit

	Variables	Coefficients	CV RMSE
Logit M1	4	12	0.374
Logit M2	9	19	0.366
Logit M3	22	36	0.364
Logit M4	30	80	**0.362**
Logit M5	30	154	0.363
Logit LASSO	30	143	0.362

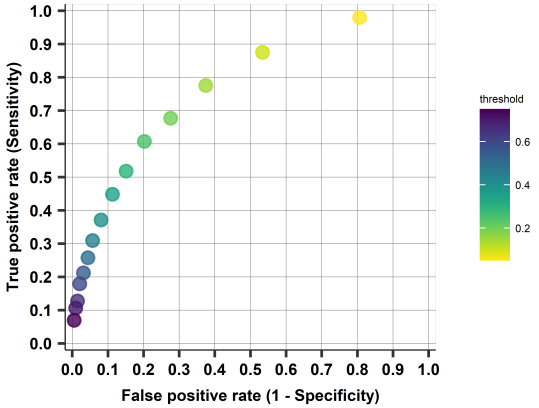
- *5-fold cross-validated on work set, average RMSE*

Will use Logit M4 model as benchmark

Classification

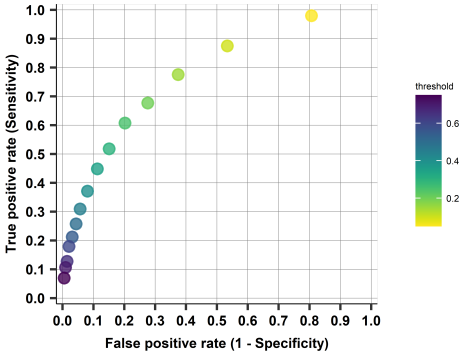
- Picked a model on RMSE/Brier score
- For classification, we will need a threshold

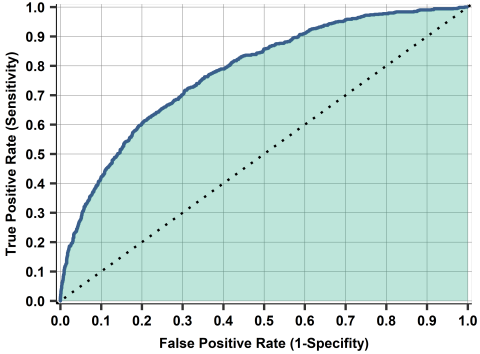
Firm exit case study: ROC curve



- ROC curve shows trade-off for various values of the threshold
  - Go through values of the ROC curve for selected threshold values,
  - between 0.05 and 0.75, by steps of 0.05

Firm exit case study: ROC curves





Firm exit case study: AUC

Model	RMSE	AUC
Logit M1	0.374	0.738
Logit M2	0.366	0.771
Logit M3	0.364	0.777
Logit M4	**0.362**	**0.782**
Logit M5	0.363	0.777
Logit LASSO	**0.362**	0.768

- Can calculate the AUC for all our models
- Model selection by RMSE or AUC
- Here: same (could be different if close)

Firm exit case study: Comparing two thresholds

- Take the Logit M4 model, predict probabilities and use that to classify on the holdout set
- Two thresholds: 50% and 20%
- Predict exit if probability > threshold

Firm exit case study: Comparing two thresholds

- Predict exit if probability > threshold

	Threshold: 0.5			Threshold: 0.2		
	Actual stay	Actual exit	Total	Actual stay	Actual exit	Total
Predicted stay	75%	15%	90%	57%	7%	64%
Predicted exit	4%	6%	10%	22%	14%	36%
Total	79%	21%	100%	79%	21%	100%

Firm exit case study: Threshold choice consequences

- Having a higher threshold leads to
  - fewer predicted exits:
    - 10% when the threshold is 50% (36% for threshold 20%).
  - fewer false positives (4% versus 22%)
  - more false negatives (15% versus 7%).
- The 50% threshold leads to a higher accuracy rate than the 20% threshold
  - 50% threshold: 75% + 6% = 81%
  - 20% threshold: 57% + 14% = 71%
  - even though the 20% threshold is very close to the actual proportion of exiting firms.

Summary

First option: no loss fn

- On the work set, do 5 fold CV and loop over models
  - Do Probability predictions
  - Calculate average RMSE on test for each fold
  - Draw ROC Curve and calculate AUC for each fold
- Pick best model based on avg RMSE
- Take best model and estimate RMSE on holdout–>best guess for live data performance
- Output: probability ranking - most likely to least likely.
- Show ROC curve and confusion table with logit on holdout 4 at $t = 0.5$ and $t = 0.2$ - to illustrate trade-off.

Firm exit case study: The loss function

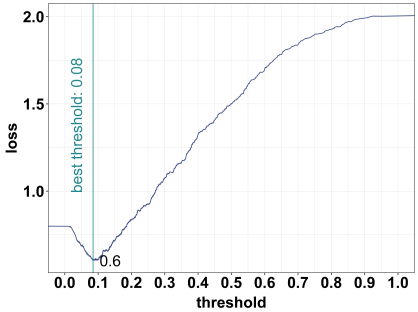
- Loss function = FN, FP
  - What matters is FN/FP
- FN=10
  - If the model predicts staying in business and the firm exits the market (a false negative), the bank loses all 10 thousand euros.
- FP=1
  - If predict exit and the bank denies the loan but the firm stays in business in fact (a false positive), the bank loses the profit opportunity of 1 thousand euros.
- With correct decisions, there is no loss.

Firm exit case study: Finding the threshold

- Find threshold by formula or algo

- Formula: the optimal classification threshold is $1/11 = 0.091$
- Algo: search thru possible cutoffs

Firm exit case study: Finding the threshold



- Consider all thresholds $T = 0.01$, 0.02—1
- Calculate the expected loss for all thresholds
- Pick when loss function has the minimum
- *Done in CV, this is fold Nr.5.*

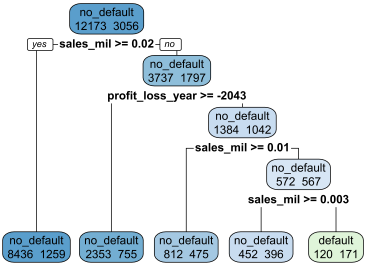
Firm exit case study

- Model selection process
  - Predict probabilities
  - Use predicted probabilities and loss function to pick optimal threshold
  - Use that threshold to calculate expected loss
  - Pick model with smallest expected loss (in 5-fold CV)
- We run the threshold selection algorithm on the work set, with 5-fold cross-validation.
  - Best is model Logit M4
  - the optimal classification threshold by algo is 0.082. Close to formula (0.091)
  - The average expected loss of 0.64.

Firm exit case study:Summary of process with loss function

- On the work set, do 5 fold CV and loop over models
  - Do Probability predictions
  - Calculate average RMSE on each test folds
  - Draw ROC Curve and find optimal threshold with loss function (1,10)
  - show: threshold search - loss plots and ROC curve for fold 5
- Summarize: for each model: average of optimal thresholds, threshold for fold 5, average expected loss, expected loss for fold Nr.5.
- Pick best model based on average expected loss
- Take best model, re-estimate it on work set + find optimal threshold and estimate expected loss on holdout set
- Confusion table on holdout with optimal threshold–>what to expect in live data.

Firm exit case study: CART



- CART
- a small tree we built for illustration purposes
- with only three variables:
  - firm size (sales),
  - binary variable for having a foreign management
  - Binary if the firm is new.
- Terminal nodes with share of exit predictions

Firm exit case study: Random Forest

- The model outperforms the logit models, with a cross validated RMSE of 0.358 and AUC of 0.808.
- We used predicted probabilities to find the optimal thresholds, and used this to make the classification.
- The expected loss: 0.587
  - smaller than for the best logit (0.642)
- For the random forest we re-estimate the model on work set, and do prediction on holdout set.
  - Holdout RMSE RF is 0.358 (vs 0.366 best logit)
  - Holdout AUC is 0.808 vs 0.784 for best logit.

Firm exit case study: Random Forest

- Note that finding the optimal threshold is rather important.
- If used a 0.5 threshold, the expected loss jumped to -1.540 vs -0.587 for the best threshold model.
- This is 2.6 times the loss from the optimal threshold.
- The default option in random forest (and many ML models) for classification is majority voting
- Majority voting is threshold=50%

Firm exit case study: Random Forest

- Note that finding the optimal threshold is rather important.
- Used a 0.5 threshold, the expected loss jumped to -1.540 vs -0.587 for the best threshold model.
- This is 2.6 times the loss from the optimal threshold.
- The default option in random forest (and many ML models) for classification is majority voting
- Majority voting is threshold=50% - NO!!!!!
  - Don't use it!!!!!!
    - Unless loss function: FN=FP

Repetition for sake of argument

If you don't have a loss function, you can't classify.

Firm exit case study: Random Forest

- No loss function
  - Predict probabilities
- Loss function
  - Predict probabilities
  - Take these probabilities and classify by threshold selected
- Alternative: use threshold and change the classification rule
  - Can be done in caret/ranger

Firm exit case study: Comparing two thresholds

- Predict exit if probability>10.9%
- Expected loss: $(1.33 \times \underline{10} + 45.4 \times \underline{1})/100 = 0.587$

	actual stay	actual exit
predicted stay	33.6%	1.3%
predicted exit	45.4%	19.7%

Firm exit case study: Summary of process with RF

- Run probability forest on work set with 5-CV
- Get average (ie over the folds) RMSE and AUC
- Now use loss function (1,10) and search for best thresholds and expected loss over folds
- Show ROC, loss on fold 5
- Optimal Threshold, average expected loss is calculated
- Take model to holdout and estimate RMSE, AUC and expected loss–>what you expect in live data
- *+1 Show expected loss with classification RF and default majority voting to compare*

Firm exit case study: Summary of model for **model selection**

Model	Preds	Coeffs	RMSE	AUC	threshold	exp. loss
Logit M1	11	12	0.374	0.736	0.089	0.722
Logit M4	36	79	0.362	0.784	0.082	0.619
Logit LASSO	36	143	0.362	0.768	0.106	0.642
RF probability	36	n.a.	0.354	0.808	0.098	0.587

- RMSE, AZC, Threshold, Loss: all 5-fold CV results (averages).

Firm exit case study: Business application

- Consider this setup

- For each firm we review, we get 1000 euros in revenues,
- Loss function: loans to bad companies $= -10,000$ euros,
- missed loans to good ones $= -1,000$ euros.

Firm exit case study: Business application

- Simplest model 1 classifies with expected loss 0.722 euro per firm, the Random Forest model has 0.587 euro.
- Building a better model yields 135 euros higher profit per firm

$$(0.722 - 0.587) \times 1000 = 135$$

- If we do 1000 deals, it is $135,000$ euros in profit.
- If a regulator asks for an interpretable model, we shall compare with the logit M4 model and have $103,000$ euros in expected profit.
- Why does it matter?

Firm exit case study: Business application

- Random Forest gets us 135K profit, best logit is 103K compared to some simple model.
- We can take this and compare to development costs
- Profit for good analysis.