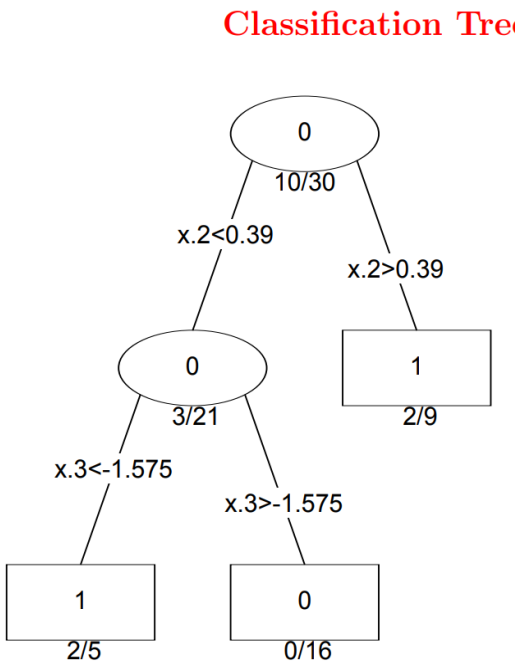
Big Data for Public Policy

7. Ensembles and Explanations

Elliott Ash & Malka Guillot

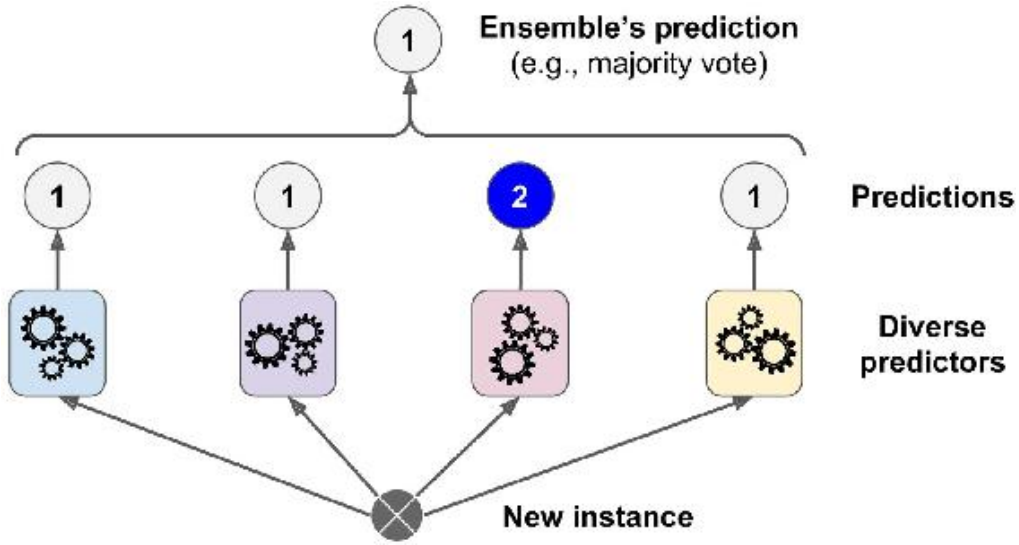
Outline

Decision Trees



- Decision trees learn a series of binary splits in the data based on hard thresholds.
  - if yes, go right; if no, go left.
- Can have additional splits as you move through the tree.
- fast and interpretable, but performance is often poor.

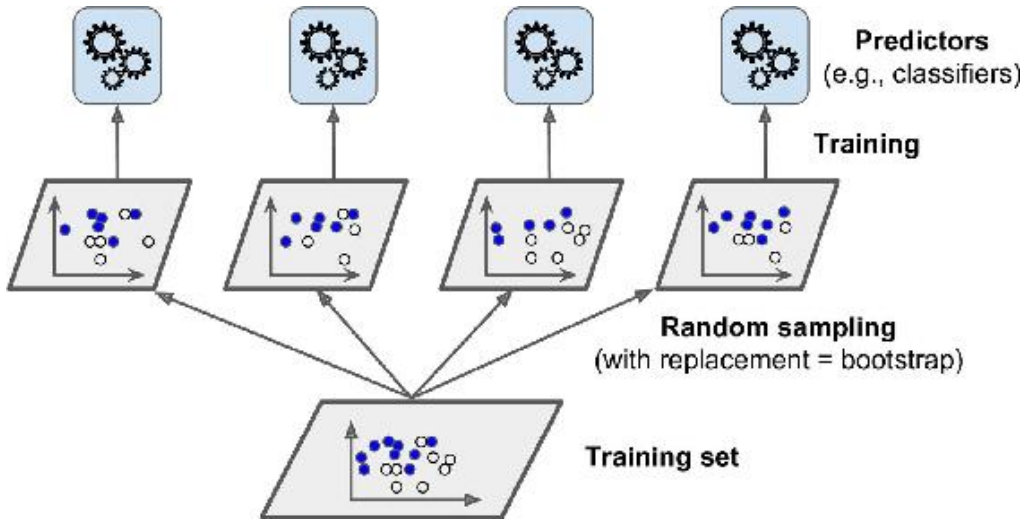
Voting Classifier



- voting classifiers generally out-perform the best classifier in the ensemble.
  - more diverse algorithms will make different types of errors, and improve your ensemble's robustness.

Bootstrapping

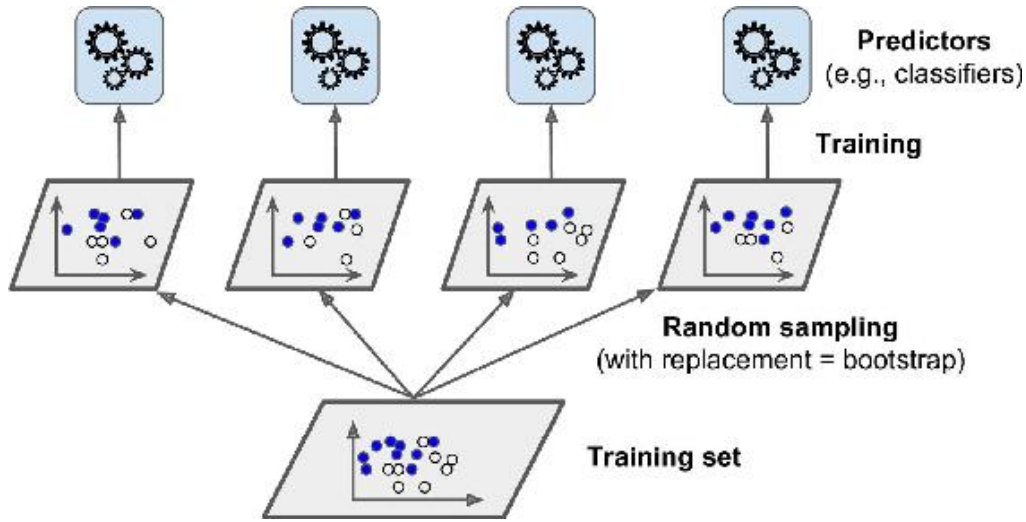
- Rather than use the same data on different classifiers, one can use different subsets of the data on the same classifier:



  - This is called bootstrapping.
  - can also use different subsets of features across subclassifiers.

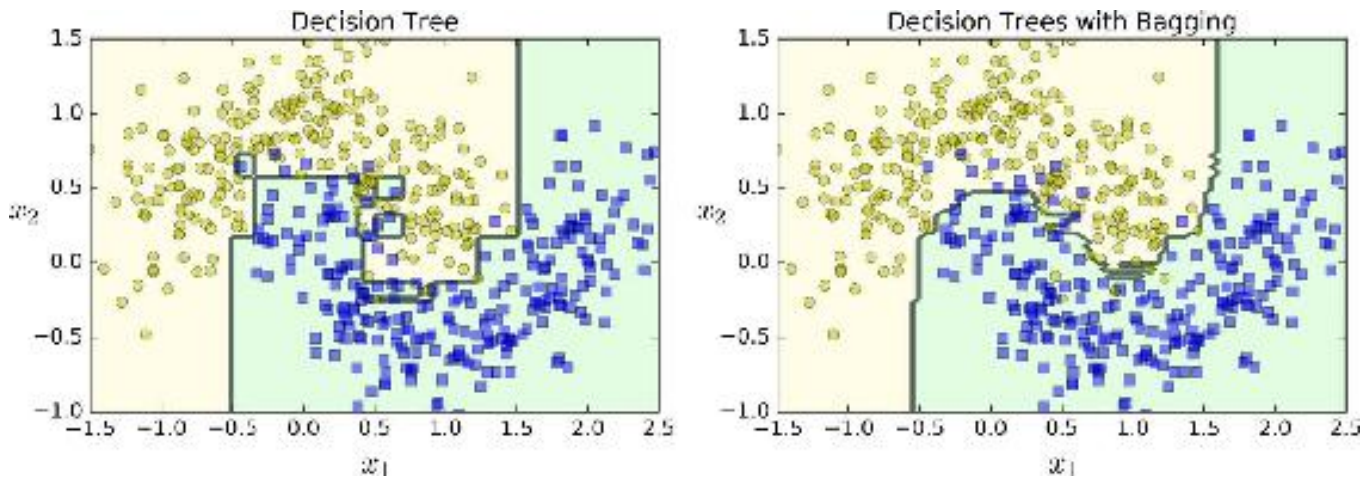
Bootstrapping

- Rather than use the same data on different classifiers, one can use different subsets of the data on the same classifier:



  - This is called bootstrapping.
  - can also use different subsets of features across subclassifiers.
- The ensemble predicts by aggregating the predictions:
  - for classification, use the most frequent prediction
  - for regression, use the average output

Bootstrapping Benefits



- While the individual predictors have a higher bias than a predictor trained on all the data, aggregation reduces both bias and variance.
  - Generally, the ensemble has a similar bias but lower variance than a single predictor trained on all the data.
- Predictors can be trained in parallel using separate CPU cores.

Random Forests

Random Forests are optimized ensembles of bootstrapped decision trees:

Random Forests

Random Forests are optimized ensembles of bootstrapped decision trees:

1. Each voting tree gets its own sample of data.

Random Forests

Random Forests are optimized ensembles of bootstrapped decision trees:

1. Each voting tree gets its own sample of data.
2. At each tree split, a random sample of features is drawn, only those features are considered for splitting.

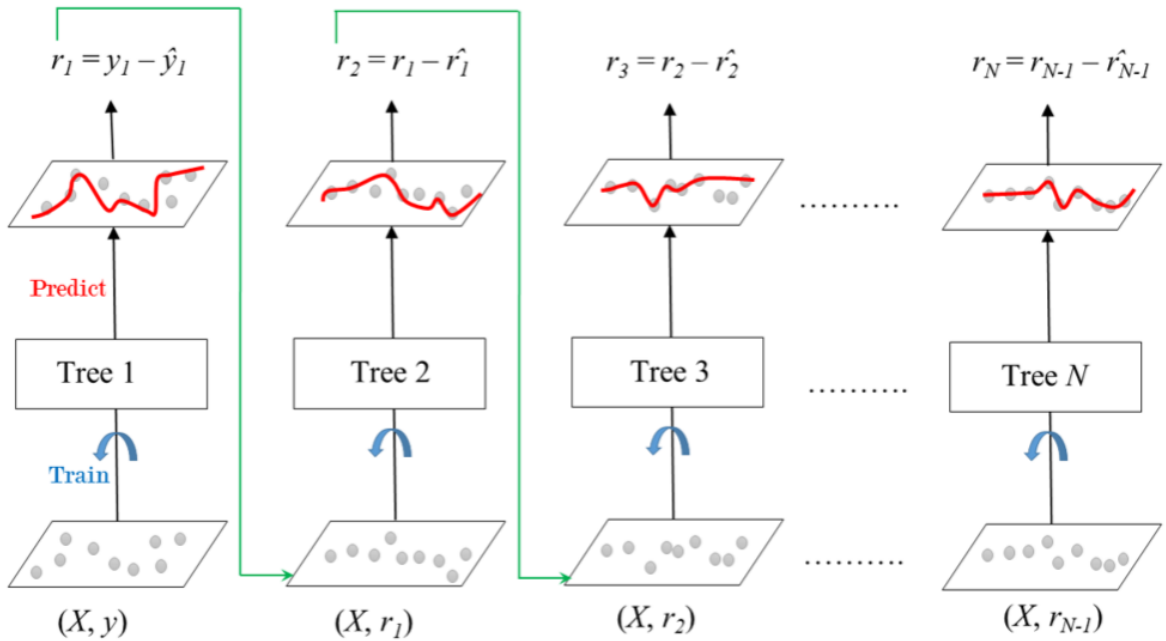
Random Forests

Random Forests are optimized ensembles of bootstrapped decision trees:

1. Each voting tree gets its own sample of data.
2. At each tree split, a random sample of features is drawn, only those features are considered for splitting.
3. For each tree, error rate is computed using data outside its bootstrap sample.

Gradient Boosting Machines

- Gradient boosting refers to an additive ensemble of trees:



- Adds additional layers of trees to fit the residuals of the first layers

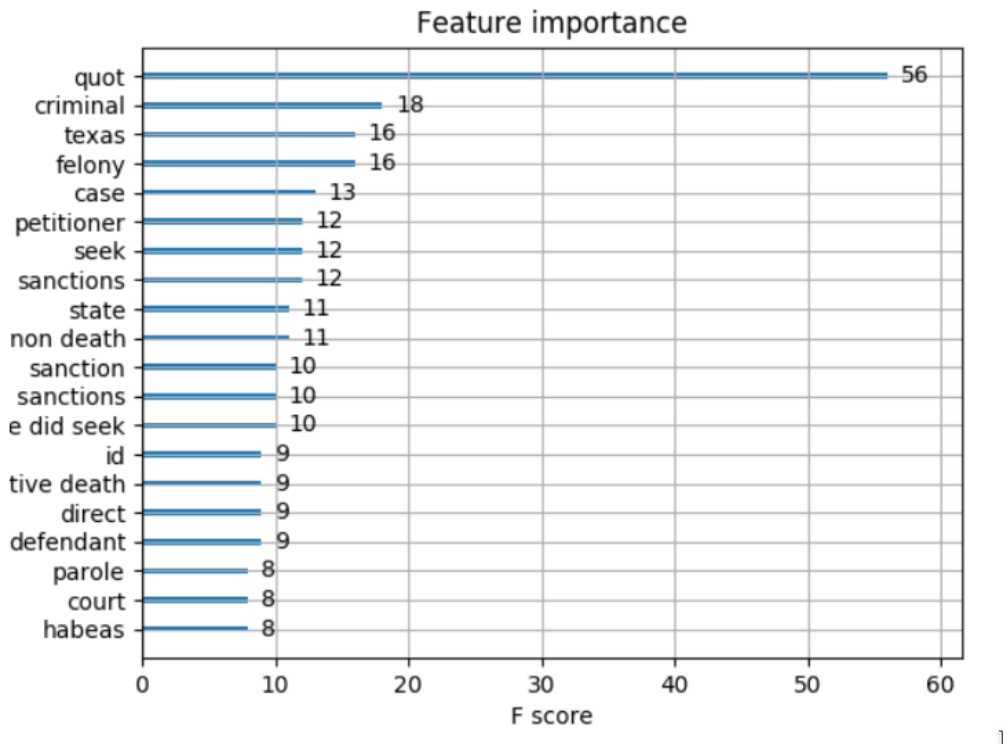
XGBoost

- GBC's get state-of-the-art performance in classification tasks with structured data (e.g. Geron 2019)
- XGBoost (available in Python or R) is a good starting point for any machine learning task:
  - easy to use
  - actively developed
  - efficient / parallelizable
  - provides model explanations
  - takes sparse matrices as input

Feature Importance

```
from xgboost import plot_importance
plot_importance(xgb_reg, max_num_features=20)
```

```
<IPython.core.display.Javascript object>
```



- Random forests and boosted trees provide a metric of feature importance that summarizes how well each feature contributes to predictive accuracy.

Outline

https://xkcd.com/

Why interpretability?

- In machine learning, helps with debugging.

Why interpretability?

- In machine learning, helps with debugging.
- In research, helps with measurement validity.

Why interpretability?

- In machine learning, helps with debugging.
- In research, helps with measurement validity.
- In applications, can be helpful to users.

Why interpretability?

- In machine learning, helps with debugging.
- In research, helps with measurement validity.
- In applications, can be helpful to users.
- In decision systems, can help subjects feel fairly treated.

Some features of “good” explanations

- A good explanation answers a “why” question.

Some features of "good" explanations

- A good explanation answers a "why" question.
- **Contrastive**: they explain not just why a certain prediction was made, but why it was made instead of other predictions.

Some features of "good" explanations

- A good explanation answers a "why" question.
- **Contrastive**: they explain not just why a certain prediction was made, but why it was made instead of other predictions.
- **Selective**: explanations should be short.

Some features of "good" explanations

- A good explanation answers a "why" question.
- **Contrastive**: they explain not just why a certain prediction was made, but why it was made instead of other predictions.
- **Selective**: explanations should be short.
- **Social**: explanations should be targeted to the relevant audience.

Some features of "good" explanations

- A good explanation answers a "why" question.
- **Contrastive**: they explain not just why a certain prediction was made, but why it was made instead of other predictions.
- **Selective**: explanations should be short.
- **Social**: explanations should be targeted to the relevant audience.
- **Outlier-focused**: if one of the input features is abnormal, that should be the focus of the explanation.

Interpretable Models

Algorithm	Linear	Monotone	Interaction
Linear regression	X	X	
Logistic regression		X	
Decision trees		~	X
k-nearest neighbors			

- **Linearity**: association between features and target is modelled linearly.
  - in addition, L1 penalty can enforce sparsity.

Interpretable Models

Algorithm	Linear	Monotone	Interaction
Linear regression	X	X	
Logistic regression		X	
Decision trees		~	X
k-nearest neighbors			

- **Linearity**: association between features and target is modelled linearly.
  - in addition, L1 penalty can enforce sparsity.
- **Monotonicity**: the relationship between a feature and the target outcome always goes in the same direction over the entire range of the feature.

Interpretable Models

Algorithm	Linear	Monotone	Interaction
Linear regression	X	X	
Logistic regression		X	
Decision trees		~	X
k-nearest neighbors			

- **Linearity**: association between features and target is modelled linearly.
  - in addition, L1 penalty can enforce sparsity.
- **Monotonicity**: the relationship between a feature and the target outcome always goes in the same direction over the entire range of the feature.
- **No interactions:** allowing interactions between features improves predictive performance but hurts interpretability.

Feature Importance

- What features are most important for prediction?

Feature Importance

- What features are most important for prediction?

Permutation feature importance algorithm (Fisher, Rudin, and Dominici 2018):

- Estimate any model, compute mean squared error.

Feature Importance

- What features are most important for prediction?

Permutation feature importance algorithm (Fisher, Rudin, and Dominici 2018):

- Estimate any model, compute mean squared error.
- For each feature j:
  - generate new dataset where feature j is permuted (scrambled)
  - generate predictions and estimate new error.
  - feature importance of j is (proportional or absolute) increase in error (ratio or difference).

Feature Importance Plot

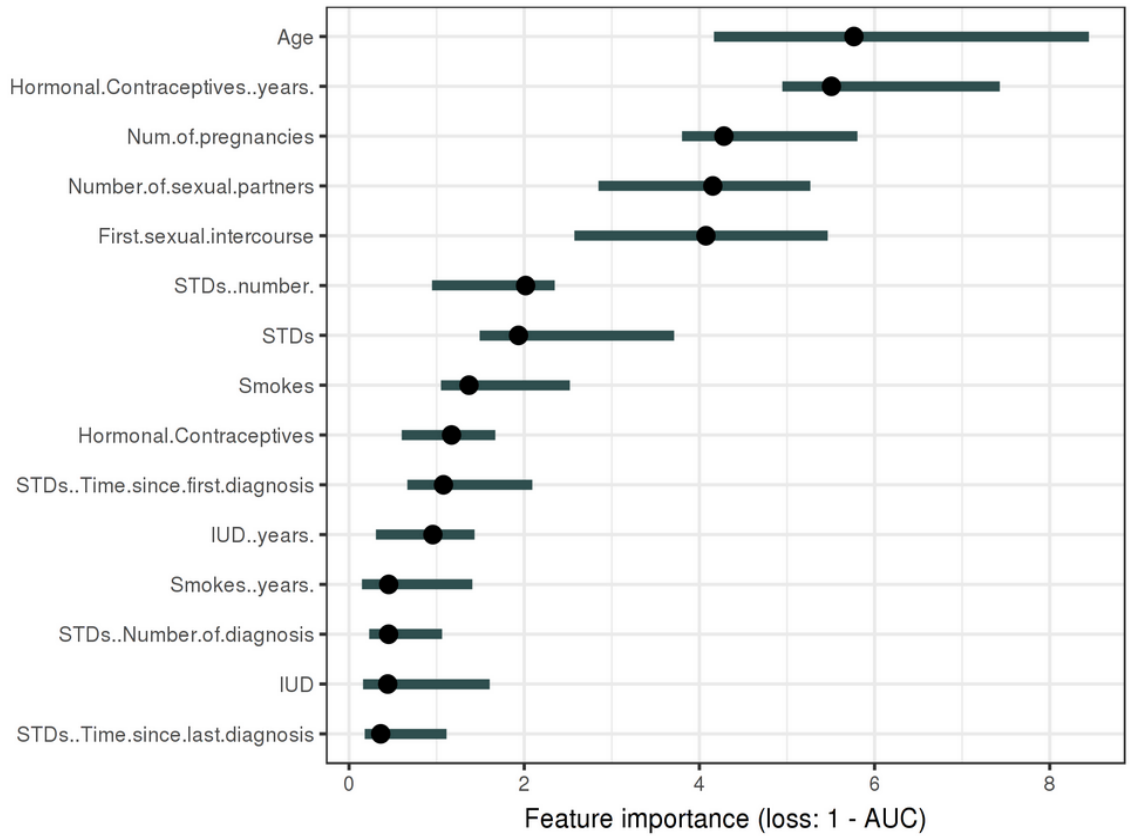


FIGURE 5.29: The importance of each of the features for predicting cervical cancer with a random forest. The most important feature was Age. Permuting Age resulted in an increase in 1-AUC by a factor of 5.76

Global Surrogate Model

1. Get predictions $\hat{y}$ of the black box model from the data X.

Global Surrogate Model

1. Get predictions $\hat{y}$ of the black box model from the data X.
2. Train an interpretable model (lasso, decision tree, etc) on X with $\hat{y}$ as the label.
   - This is the surrogate model!

Global Surrogate Model

1. Get predictions $\hat{y}$ of the black box model from the data X.
2. Train an interpretable model (lasso, decision tree, etc) on X with $\hat{y}$ as the label.
   - This is the surrogate model!
3. Validate that the surrogate model replicates the predictions of the black box model
   - e.g., compute R^2 or $F1$ between black box $\hat{y}$ and surrogate $\hat{\hat{y}}$.
   - doesn't need to be in held out test set.

Local Surrogate (LIME)

- LIME = local interpretable model-agnostic explanations.
- Isolates the features which are most important at a particular data point.

Local Surrogate (LIME)

- LIME = local interpretable model-agnostic explanations.
- Isolates the features which are most important at a particular data point.

1. Select data point to explain

Local Surrogate (LIME)

- LIME = local interpretable model-agnostic explanations.
- Isolates the features which are most important at a particular data point.

1. Select data point to explain
2. Perturb dataset (locally) and get black box predictions for the new points.

Local Surrogate (LIME)

- LIME = local interpretable model-agnostic explanations.
- Isolates the features which are most important at a particular data point.

1. Select data point to explain
2. Perturb dataset (locally) and get black box predictions for the new points.
3. Train an interpretable surrogate model on the perturbed dataset (weighted by proximity to initial data point).
   - This is the "local" surrogate model.
   - use lasso with high L1 penalty to get a sparse explanation.

LIME for Text

1. Generate new texts by randomly *removing* words from the original document.

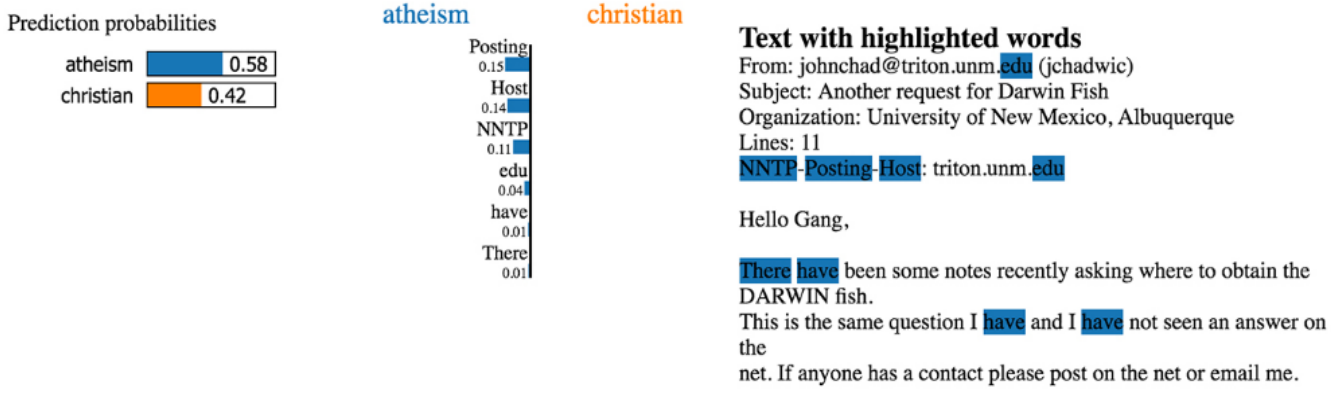
LIME for Text

1. Generate new texts by randomly *removing* words from the original document.
2. Form predictions $\hat{y}$ from black box model for these perturbed documents.

LIME for Text

1. Generate new texts by randomly *removing* words from the original document.
2. Form predictions $\hat{y}$ from black box model for these perturbed documents.
3. Train lasso on dataset of binary features for each word, equaling one if word appears, to predict $\hat{y}$.
   - weight by proximity to initial data point (one minus the proportion of words dropped)

```
exp = explainer.explain_instance(test_example,
classifier.predict_proba, num_features=6)
```



Practical Advice for Research and Applications

1. for gradient boosting, use the contained feature importance.
2. for regression, examine coefficients
3. look at highest and lowest ranked documents for $\hat{y}$
4. report a few example documents with LIME highlighting

Outline

Cross-Domain (Transfer) Learning

- A recent but now widespread approach to machine learning is **transfer learning**:
  - train a model in a big labeled dataset
  - apply in a smaller (mostly) unlabeled dataset

Cross-Domain (Transfer) Learning

- A recent but now widespread approach to machine learning is **transfer learning**:
  - train a model in a big labeled dataset
  - apply in a smaller (mostly) unlabeled dataset
- In NLP:
  - transfer learning is intuitive because NLP tasks share common knowledge about language.
  - labeled data is scarce/expensive, so learn tasks on tons of unlabeled data.

Cross-Domain (Transfer) Learning

- A recent but now widespread approach to machine learning is **transfer learning**:
  - train a model in a big labeled dataset
  - apply in a smaller (mostly) unlabeled dataset
- In NLP:
  - transfer learning is intuitive because NLP tasks share common knowledge about language.
  - labeled data is scarce/expensive, so learn tasks on tons of unlabeled data.
  - BERT and GPT-2 are the big success stories in transfer learning.

Osnabruegge, Ash, and Morelli 2020

This paper takes the idea of transfer learning to the political science context.

- Learn to predict political topics from text in a labeled corpus (party manifestos from Comparative Manifesto Project)

Osnabruegge, Ash, and Morelli 2020

This paper takes the idea of transfer learning to the political science context.

- Learn to predict political topics from text in a labeled corpus (party manifestos from Comparative Manifesto Project)
- Apply model to classify topics in unlabeled corpus (parliamentary speeches).

Osnabruegge, Ash, and Morelli 2020

This paper takes the idea of transfer learning to the political science context.

- Learn to predict political topics from text in a labeled corpus (party manifestos from Comparative Manifesto Project)
- Apply model to classify topics in unlabeled corpus (parliamentary speeches).
- Use for empirical analysis of electoral institutions and speech content.

Comparative Manifesto Project Corpus

- 115,410 annotated English-language political statements
  - hundreds of political party platforms from English-speaking countries.

Comparative Manifesto Project Corpus

- 115,410 annotated English-language political statements
  - hundreds of political party platforms from English-speaking countries.
- Each statement gets a CMP code $\boldsymbol{y}_i$, e.g. "decentralization", "education"
  - n_y =44 topics
    - some topics are somewhat esoteric, such as "marxist analysis"

Comparative Manifesto Project Corpus

- 115,410 annotated English-language political statements
  - hundreds of political party platforms from English-speaking countries.
- Each statement gets a CMP code $\boldsymbol{y}_i$, e.g. "decentralization", "education"
  - n_y =44 topics
    - some topics are somewhat esoteric, such as "marxist analysis"
  - also: 8 broader "topic domains" (external relations, freedom and democracy, political system, economy, welfare and quality of life, fabric of society, social groups, and no topic)

Featurizing the Statements

- Standard featurization steps:
  - remove capitalization, punctuation, stopwords
  - construct n-grams up to length 3
  - remove n-grams appearing in less than 10 statements or more than 40 percent of statements

Featurizing the Statements

- Standard featurization steps:
  - remove capitalization, punctuation, stopwords
  - construct n-grams up to length 3
  - remove n-grams appearing in less than 10 statements or more than 40 percent of statements
- $n_x = 19,734$ features
  - compute tf-idf-weighted n-gram frequencies

Prediction Model

- Models:
  - regularized logistic
  - gradient boosting
  - bidirectional transformer

Prediction Model

- Models:
  - regularized logistic
  - gradient boosting
  - bidirectional transformer

Predict the CMP code in a held-out sample of manifesto corpus statements:

- 44-topic specification:
  - test-sample accuracy $= 0.54$
  - training-sample accuracy $= .70$
  - choosing randomly would be correct 2% of the time; choosing most-frequent category (other topic) would be correct 15% of the time.

Prediction Model

- Models:
  - regularized logistic
  - gradient boosting
  - bidirectional transformer

Predict the CMP code in a held-out sample of manifesto corpus statements:

- 44-topic specification:
  - test-sample accuracy = 0.54
  - training-sample accuracy = .70
  - choosing randomly would be correct 2% of the time; choosing most-frequent category (other topic) would be correct 15% of the time.
- 8-topic specification:
  - test-sample accuracy = 0.64
  - training-sample accuracy = 0.76
  - choosing randomly == 0.125 accuracy, choosing most-frequent category would be correct 30% of the time.

	Economy	External Relations	Fabric of society	Freedom & Democracy	Political system	Social groups	Welfare & quality of life	No topic / Other	Total true
Economy	5270	93	131	40	301	254	1108	0	**7197**
External Relations	175	1207	137	83	85	49	209	1	**1946**
Fabric of society	269	107	1785	90	204	115	618	1	**3189**
Freedom and Democracy	102	60	135	631	219	35	177	0	**1359**
Political system	608	71	186	137	1255	65	542	1	**2865**
Social groups	493	51	185	29	111	1230	818	0	**2917**
Welfare and quality of life	1033	66	316	58	267	293	7138	0	**9171**
No topic / Other	58	6	37	9	34	7	55	3	**209**
Total predicted	**8008**	**1661**	**2912**	**1077**	**2476**	**2048**	**10665**	**6**	
Total predicted / Total true	**1.11**	**0.85**	**0.91**	**0.79**	**0.86**	**0.70**	**1.16**	**0.03**	

New Zealand Parliamentary Speeches

- All parliament speeches, 1987-2002.
  - $437\mathrm{K}$ speeches in total, removed procedural remarks, short speeches, and foreign-language speeches, to get 290K for analysis.
  - speeches linked to speaker and parliamentary debate type.

New Zealand Parliamentary Speeches

- All parliament speeches, 1987-2002.
  - $437\mathrm{K}$ speeches in total, removed procedural remarks, short speeches, and foreign-language speeches, to get 290K for analysis.
  - speeches linked to speaker and parliamentary debate type.
- Apply featurization pipeline and classifier to get predicted topic probabilities for each speech.

(a) economy

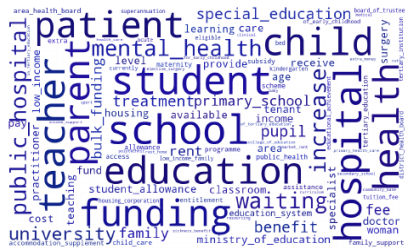
(b) external relations

(e) political system

(f) social groups

(c) fabric of society

(d) freedom and democracy



(g) welfare and quality of life

(h) no topic

Validation with Target-Corpus Annotation

Table 3: Human Coding vs. Predicted Manifesto Topics

	Top 1	Top 3	Top 5	N
8 topics				
Welfare and quality of life	62	91	98	796
Political system	57	90	98	1,069
External relations	56	84	91	94
Fabric of society	55	87	97	433
Economy	54	85	95	721
Social groups	37	71	88	325
Freedom and democracy	37	71	88	545
no topic	1	2	12	192
Total	51	82	92	4,175

- We sampled 4,175 NZ speeches and a manifesto coder annotated them.
- In 44-topic spec, overall top-1 accuracy is 41% and top-3 accuracy is 65%

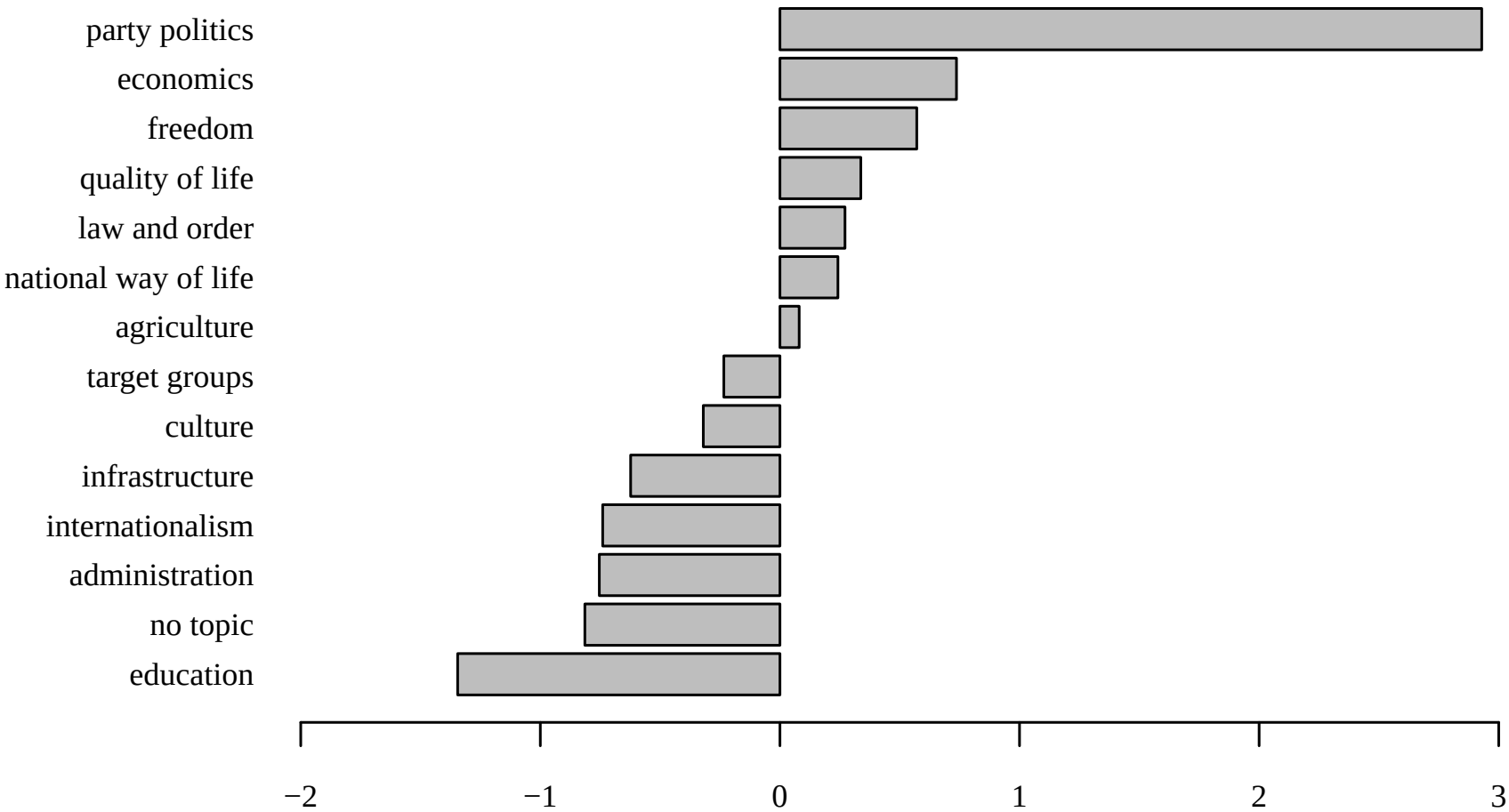
Experiment: Electoral Reform in New Zealand

- A 1993 reform in New Zealand moved from majoritarian to proportional representation:
  - **Majoritarian (first past the post)**: two parties, single party controls parliament.
  - **Proportional representation**: many minority parties, coalition governments.

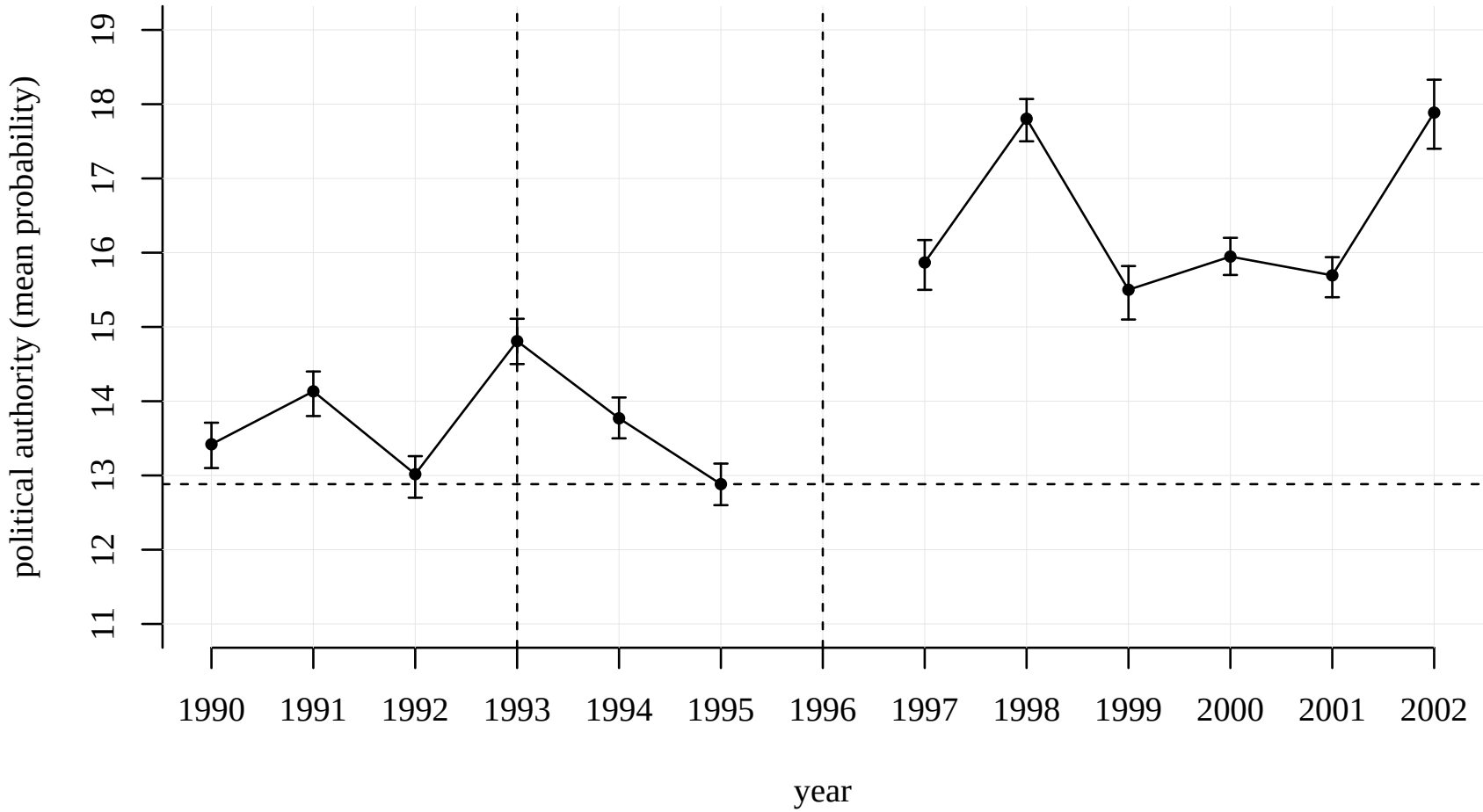
Experiment: Electoral Reform in New Zealand

- A 1993 reform in New Zealand moved from majoritarian to proportional representation:
  - **Majoritarian (first past the post)**: two parties, single party controls parliament.
  - **Proportional representation**: many minority parties, coalition governments.
- How did it affect speech topics in the New Zealand Parliament?

Change in Parliament Attention due to Reform



Change in Party Politics Topic



Example "Party Politics" Speech

Osnabruegge, Ash, Morelli (2019)

"I have seen seven Opposition leaders in my time, but I have never seen a leader as relentlessly negative as Helen Clark. . . . How could anybody be so negative, day in, day out? It could get into the Guinness Book of Records. She does not have a positive word to say about anything. It is all negative, negative, negative."

- Parliamentarian Richard Prebble, 15 Feb 1999

Outline

Overview

- Apply ML to budget factors to analyze corruption
- Produce empirical results in political economy / development
  - Evaluate the dynamic (and spillover) effects of audits
- **Policy upshot:** Our method/results could provide inputs to policy decisions about corruption

- In Brazil, local governments have a central role in the provision of a variety of public goods (e.g., primary education, culture, health care, housing, transportation and municipal infrastructure)

- In Brazil, local governments have a central role in the provision of a variety of public goods (e.g., primary education, culture, health care, housing, transportation and municipal infrastructure)
- In May 2003 the Brazilian government introduced an innovative anticorruption program.
- The program is based on the auditing of public spending of **randomly selected** municipalities

- In Brazil, local governments have a central role in the provision of a variety of public goods (e.g., primary education, culture, health care, housing, transportation and municipal infrastructure)
- In May 2003 the Brazilian government introduced an innovative anticorruption program.
- The program is based on the auditing of public spending of **randomly selected** municipalities

- A report containing a list of all irregularities and malpractices founded is sent to competent authorities for prosecution and made publicly available on the CGU website in about 3 months (Internet and other media)

Corruption Audit Data

- Brollo et al. (2013) provide corruption audit data from 1481 Brazilian municipalities, disclosed from July 2003 to March 2010 (the first 29 lotteries)
  - We use the measure of **narrow corruption**: intentional misconduct, including illegal procurement, fraud, favoritism, and over-invoicing

Local Public Finance Data

- The annual municipality budget has detailed information about the categories of *expenditure*, *revenue*, as well as *assets* and *liabilities* positions
- We collected data for 2001 through 2012, for all Brazilian municipalities

Local Public Finance Data

- The annual municipality budget has detailed information about the categories of *expenditure*, *revenue*, as well as *assets* and *liabilities* positions
- We collected data for 2001 through 2012, for all Brazilian municipalities
- In total we have 797 variables (Revenue 250, Expenditure 334, Active 100, Passive 79)
- Pre-processing:
  - Standardize budget features
  - Impute missing values with mean of observed value for each variable – add dummy variables to indicate missing

Model Training: Gradient Boosted Classifier

- Shuffle dataset into 80% training set and 20% test set
- Tuned hyperparameters in the training set using five-fold cross-validation
  - (e.g., max depth of trees and learning rate)
- Take tuned model to get performance metrics in the test set

Model Performance in Test Set

	Gradient Boosting				
	Standard	+ Pop.	Missing dummies	+ Pop. & missing dummies	OLS
	(1)	(2)	(3)	(4)	(5)
Brollo et al (2013)					
Accuracy	0.750	0.761	0.740	0.764	0.594
AUC-ROC	0.814	0.824	0.793	0.834	0.562
F1	0.665	0.685	0.663	0.687	0.413
Avis et al (2018)					
Accuracy	0.869	0.869	0.866	0.869	0.760
AUC-ROC	0.918	0.923	0.921	0.917	0.791
F1	0.719	0.717	0.709	0.719	0.464

Notes: The table provides prediction performance for the four different specifications used with the gradient boosting model and OLS.

- Test-set accuracy of $\sim 75\%$ [$\sim 86\%$], much better than guessing (58%) [75%] and predictions from OLS (59%) [76%]

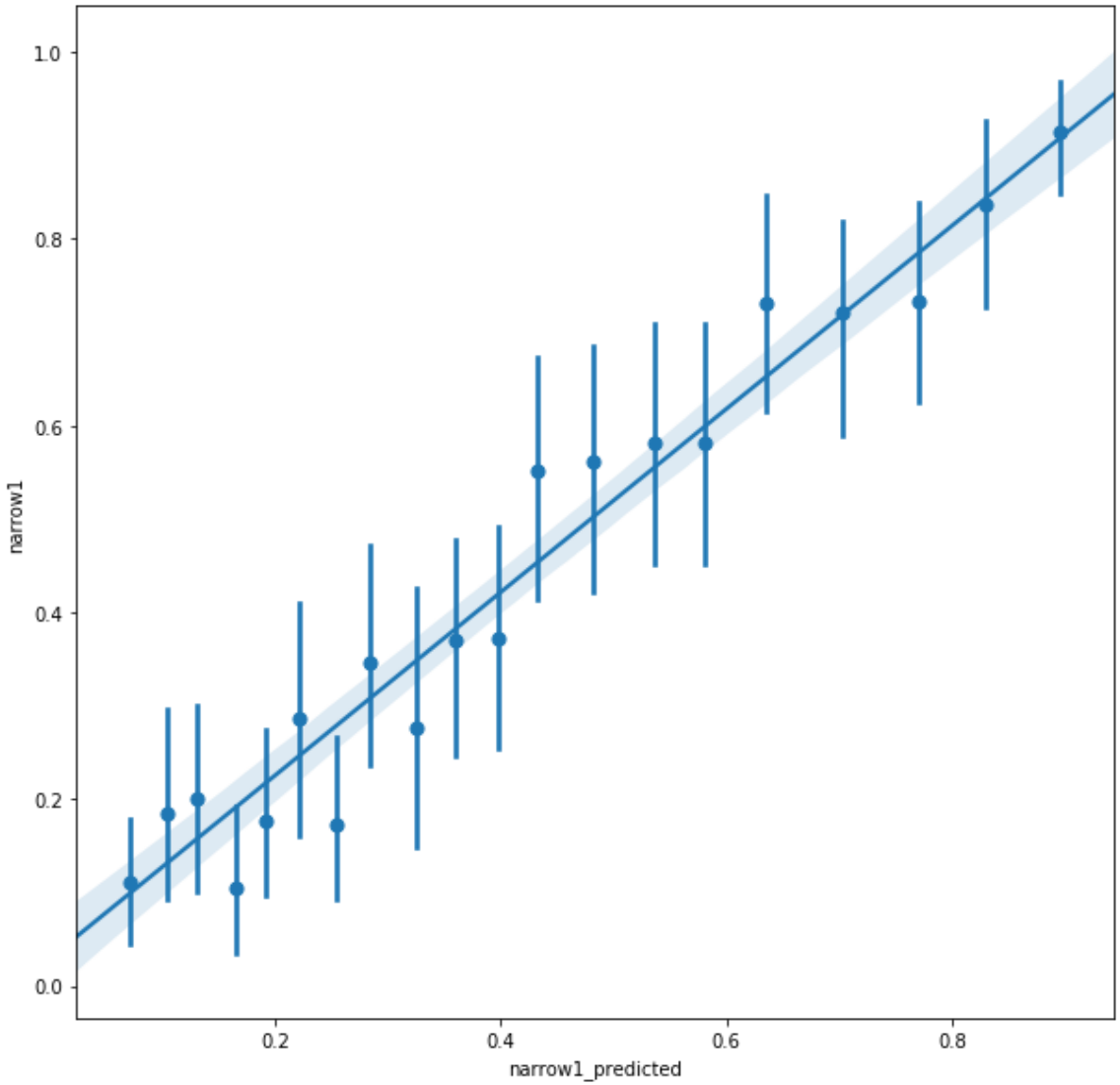
Confusion Matrix

Table: Confusion Matrices for Binary Prediction Tasks

Corruption

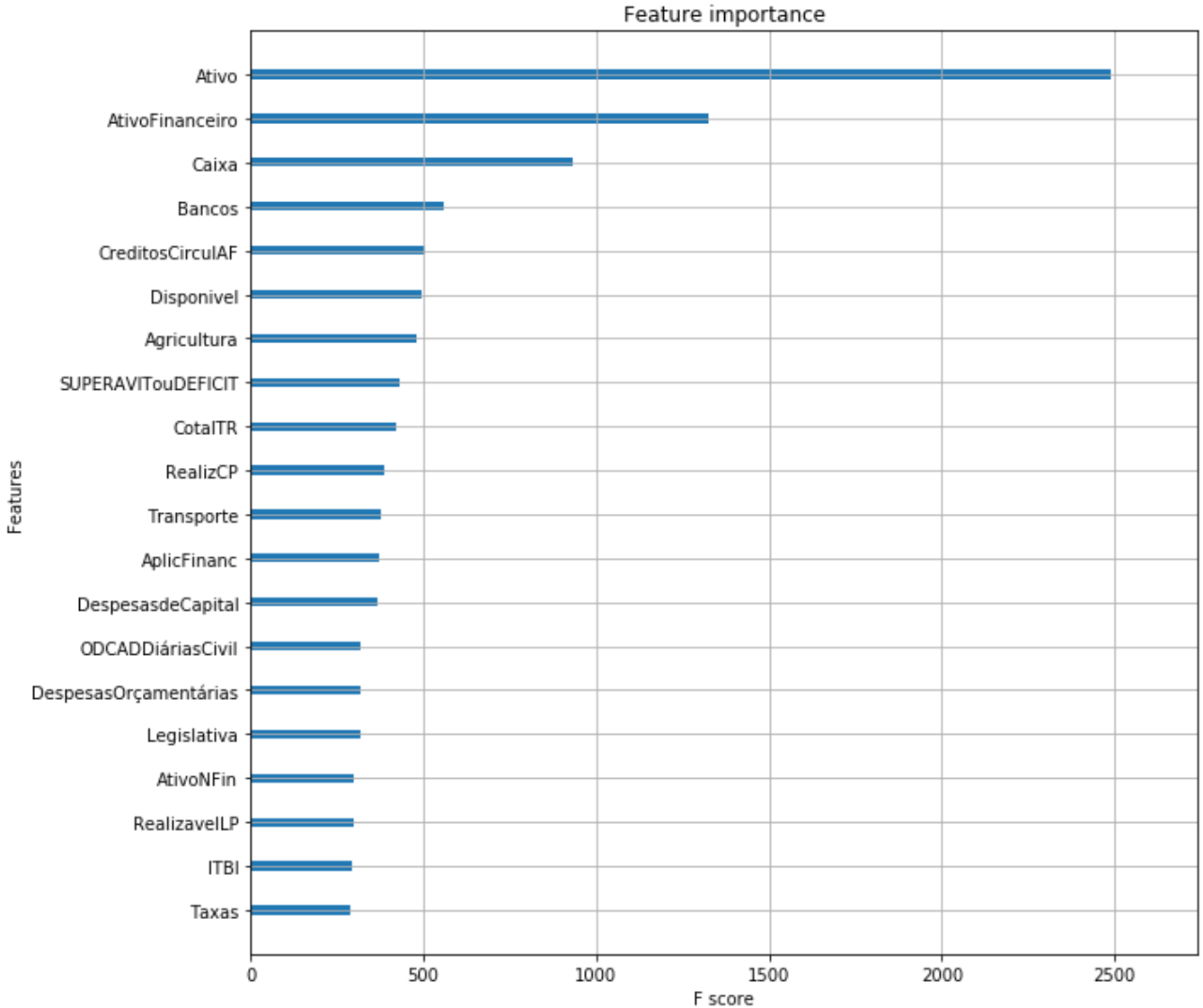
	Prediction	
Truth	Not Corrupt	Corrupt
Not Corrupt	614	100
Corrupt	185	313

Binscatter: True Corrupt Rate vs Predicted Prob.



Notes. Binscatter diagram of average true corruption (vertical axis) against binned predicted corruption (horizontal axis).

Gradient Boosting Feature Importance



Most Important Budget Features

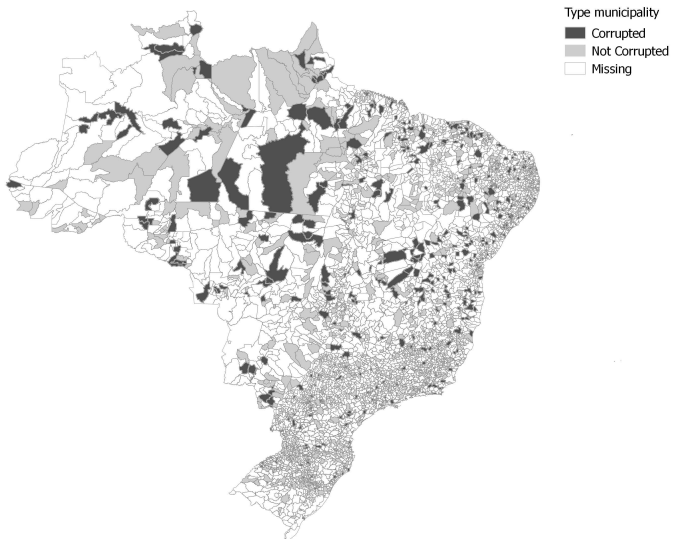
- **Assets** (Ativo):
  - Financial assets (Ativo financeiro): liquid assets (Disponível), cash (Caixa) bank deposits (Bancos), financial investments (Aplic Financ), outstanding loan credit (Créditos circulação, Out Valor Realiz),
  - Non-financial assets (Ativo não Financeiro): short term (Realiz CP) and long term (Realiz LP)
- **Revenue**:
  - Taxes (Taxas): property taxes (CotaITR, IPTU, ITBI).

Most Important Budget Features

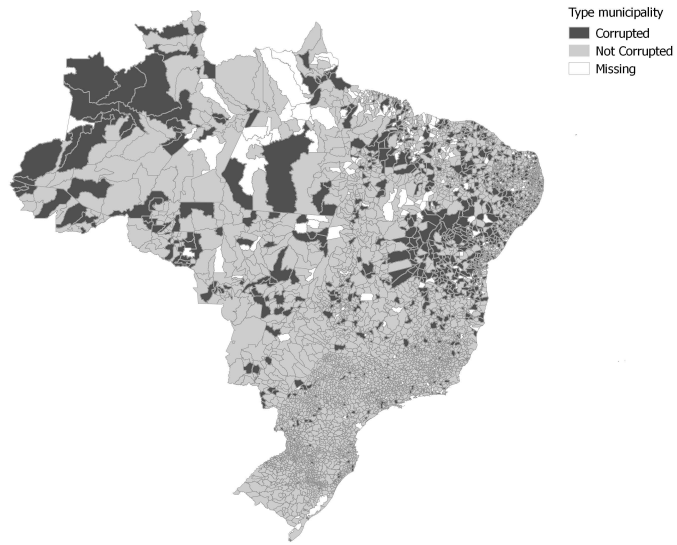
- **Expenditures** (Despesas Orçamentárias):
  - Investment expenditures (Despesadecapital)
  - Current expenditures "other" categories: civil servant per diems (ODCAD Diárias Civil), payments for the previous fiscal year (ODCAD Desp Exerc Anteriores), and payments for services provided by individuals business (ODCAD Out Serviç Terceiros PF)
  - Expenditure by function: agricultural spending (agricoltura), transportation (Transporte), communication (Comunicações)
- **Budget deficit** (Supeavit ou Deficit)

Applying to Full Dataset

- Take model trained on audited municipality-terms and predict probability of corruption in all municipalities and all years
- We regressed predicted corruption in pre-audit years on having an audit, and there was no difference in any specification



(a) Actual Corruption



(b) Predicted Corruption

Effect of Revenue Shocks on Corruption

Effect of Revenue Shocks on Corruption

- Brollo et al. (2013) test whether a windfall of public revenues (federal transfers) can lead to an increase in rent-seeking by the public administration (as measured by a subsequent increase in corruption)
- Allocation of transfers relies on exogenous population thresholds

$$FPM_i^k = \frac{FPM_k \lambda_i}{\sum_{i \in k} \lambda_i}$$

- Theoretical transfers $\neq$ Actual Transfers

Estimating Equations

First stage – instrument actual transfer with predicted transfer based on population:

$$\tau_i = g(P_i) + \alpha_\tau \hat{\tau}_i + \delta_t + \gamma_s + u_i \tag{1}$$

Second stage – estimate coefficient on instrumented transfer:

$$y_i = g(P_i) + \beta_y \tau_i + \delta_t + \gamma_s + \epsilon_i \tag{2}$$

- polynomial $g(\cdot)$ in population P_i, time fixed effects δ_t, state fixed effects γ_s

2SLS Estimates

	Predicted Corruption			
	standard	+ population	missing dummies	+ population & missing dummies
	(1)	(2)	(3)	(4)
Sample A: audited cities				
Actual transfers	0.00747	0.00696	0.00659	0.00773
	(0.00375)**	(0.00389)*	(0.00380)*	(0.00386)**
Observations	1115	1115	1115	1115
Adjusted R^2	0.254	0.241	0.246	0.249
Sample B: all cities				
Actual transfers	0.00351	0.00681	0.00412	0.00535
	(0.00117)***	(0.00120)***	(0.00119)***	(0.00118)***
Observations	5809	5809	5809	5809
Adjusted R^2	0.515	0.471	0.508	0.512
Sample C: excluding cities audited				
Actual transfers	0.00215	0.00620	0.00313	0.00421
	(0.00113)*	(0.00115)***	(0.00115)***	(0.00112)***
Observations	4693	4693	4693	4693
Adjusted R^2	0.596	0.549	0.589	0.594

Effects of FPM transfers on (predicted) corruption measures. Each cell reports the estimated coefficient of actual FPM transfers (instrumented with theoretical FPM transfers) in a regression where the dependent variable corresponds to each column heading. The regression controls for a third-order polynomial in normalized population size, term dummies, and macro-region dummies as in equation (7). Robust standard errors clustered at the municipal level are in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

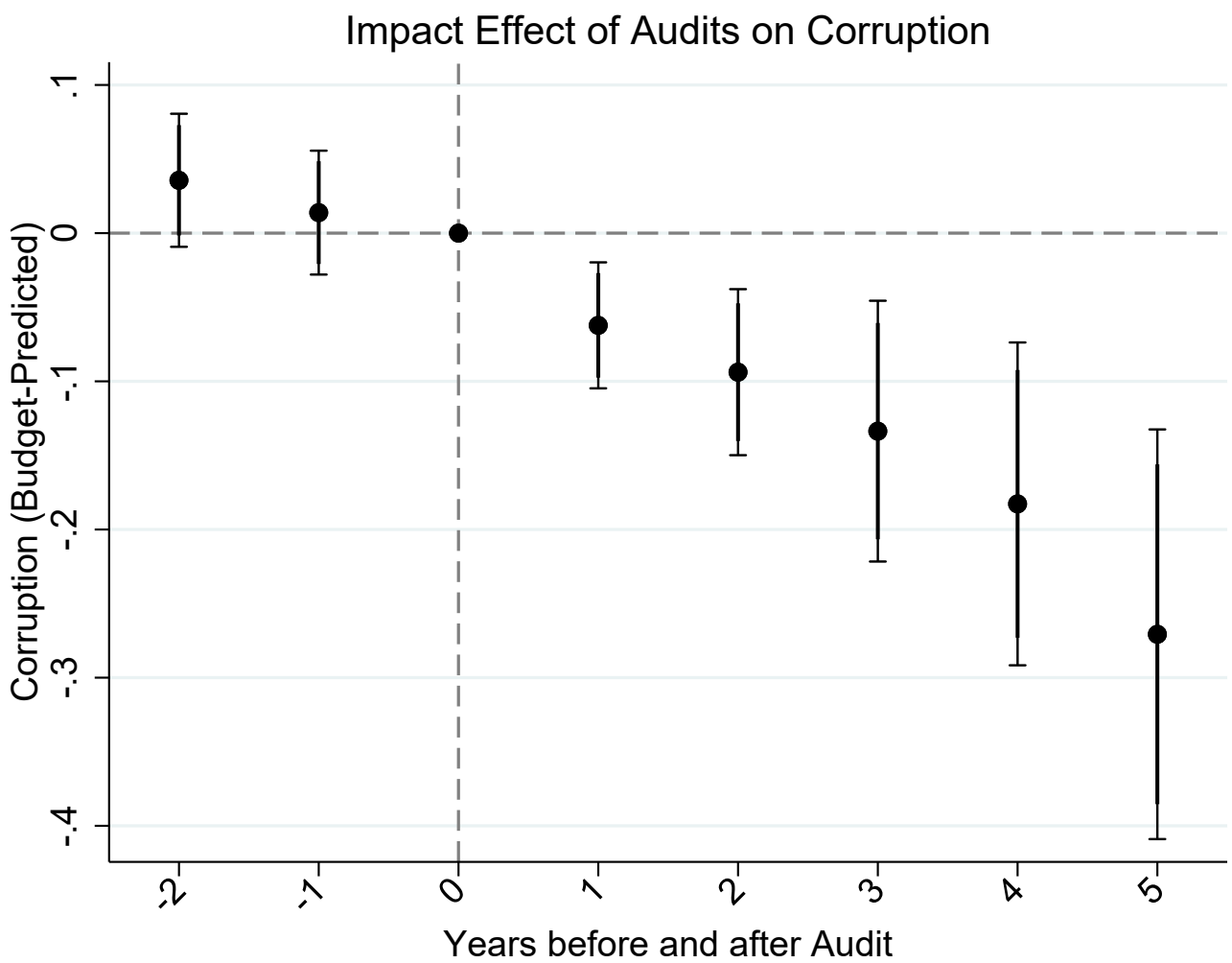
Event Study: Effect of Audits on Corruption

Event Study: Effect of Audits on Corruption

$$y_{ist} = \sum_{k=-2, k\neq 0}^{5} \beta_k D_{ist}^k + \delta_i + \lambda_i \cdot t + \gamma_{st} + X'_{ist}\phi + \epsilon_{ist} \qquad (3)$$

- D_{ist}^K is a dummy variable for k years before and after an audit
- δ_i, municipality FE
- $\lambda_i \cdot t$ municipality trend
- γ_{st} state-year FE
- Sample restricted to non-trained observations and only audited

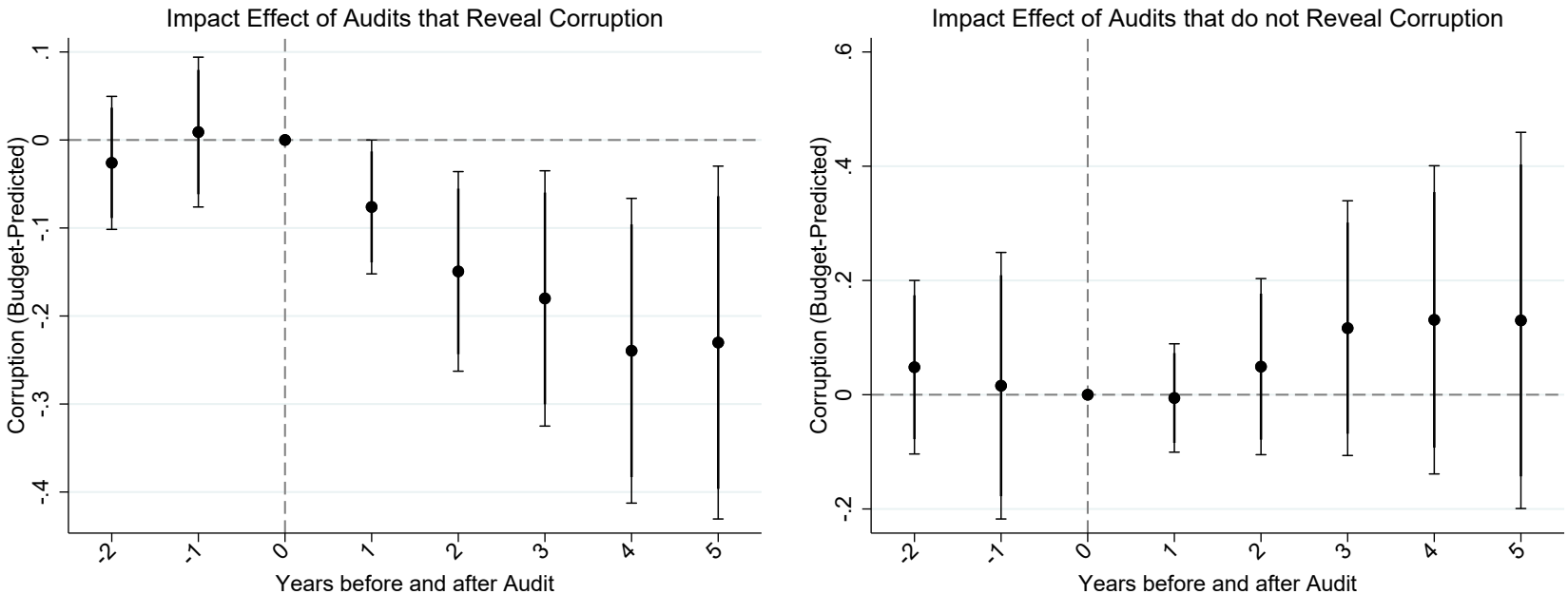
Figure: Event Study: Effect of Audits on Fiscal Corruption



Error spikes give 95% (horizontal bars) and 90% (bold lines) confidence intervals, with standard error clustered by state.

Event Study: By Audit Outcome

Figure: Event Study: Effect of Audits on Fiscal Corruption



Error spikes give 95% (horizontal bars) and 90% (bold lines) confidence intervals, with standard error clustered by state.