
Predicting Financial Statement Auditor
Supervised Learning: Classification
Appendix
Findings
Future Work
Background & Purpose
Methodologies
Jason Dunleavy, Data Scientist

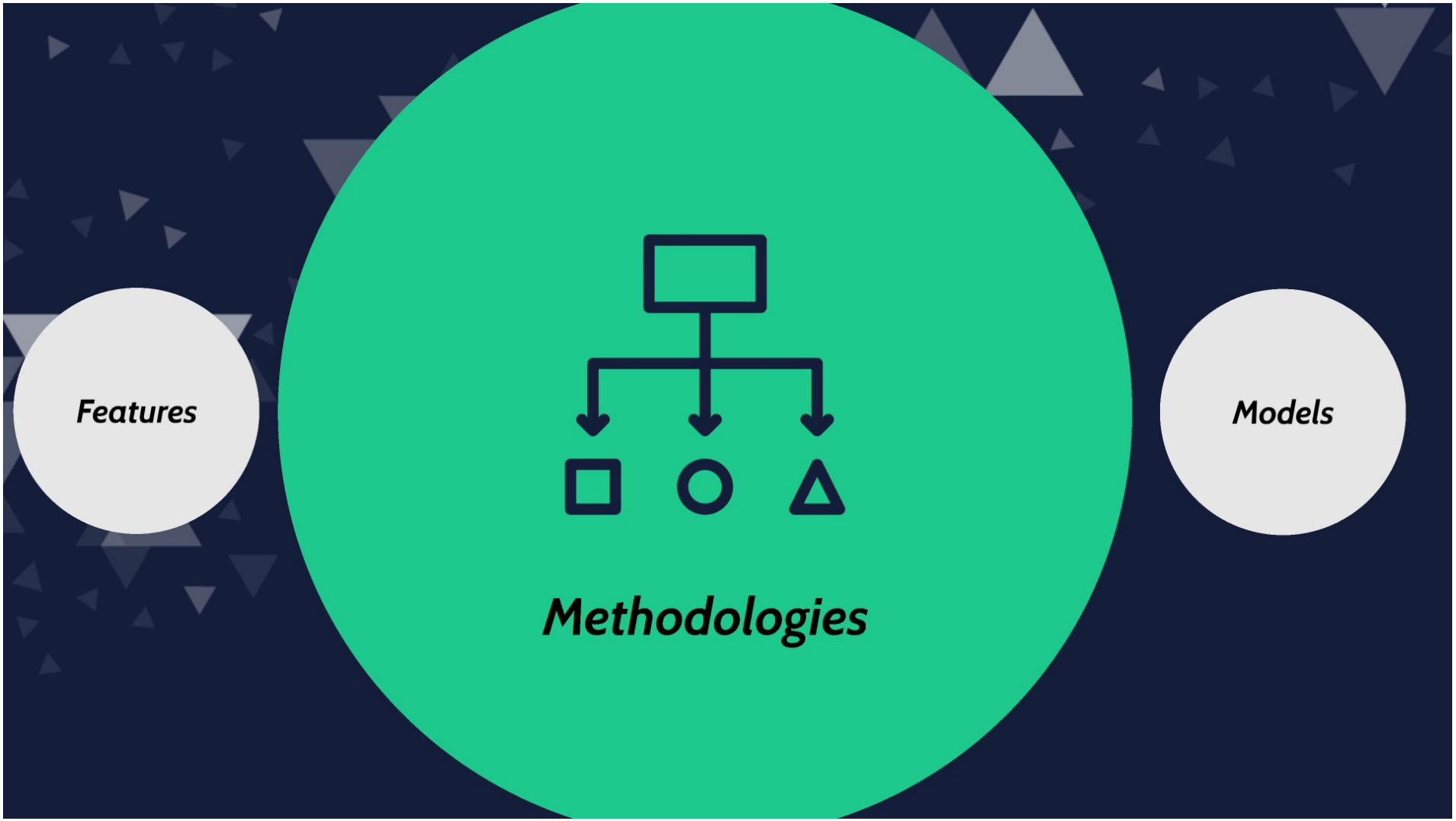

Introduction
Background &
Purpose
Data

Big Four Effect

- Need for quality audits
  - Fewer restatements
  - More accurate forecasts
  - Lower cost of capital
- Why effect exists?
  - Better audit methodologies/tools/technologies
  - Hiring effect
  - Self-selection effect
- Target (Binary)
  - Big Four (PwC, EY, Deloitte, KPMG)
  - Other

Data

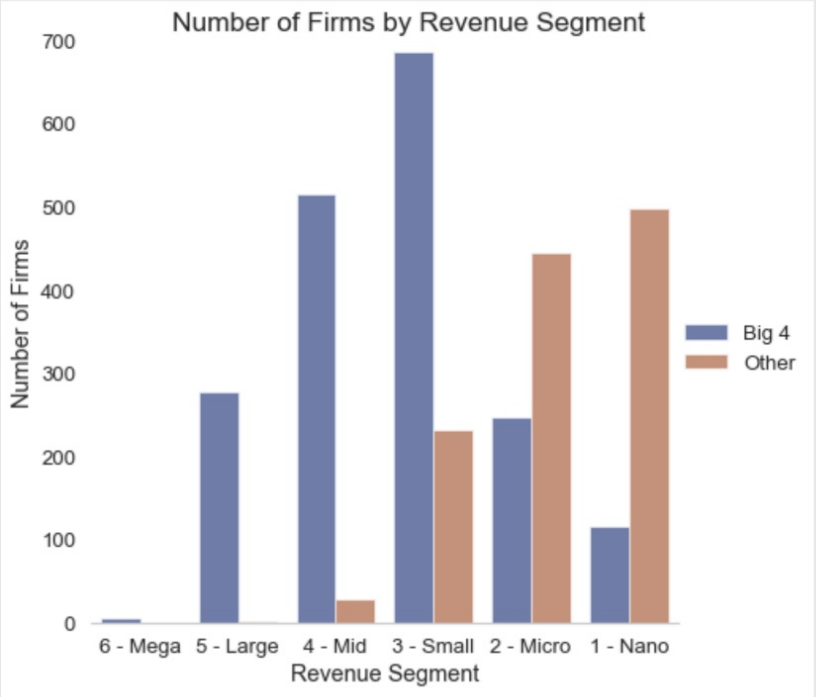
- Data sources:
  - PCAOB: Public Company Auditor Search
  - Quandl: US Company Fundamentals
- Created/queried PostgreSQL database

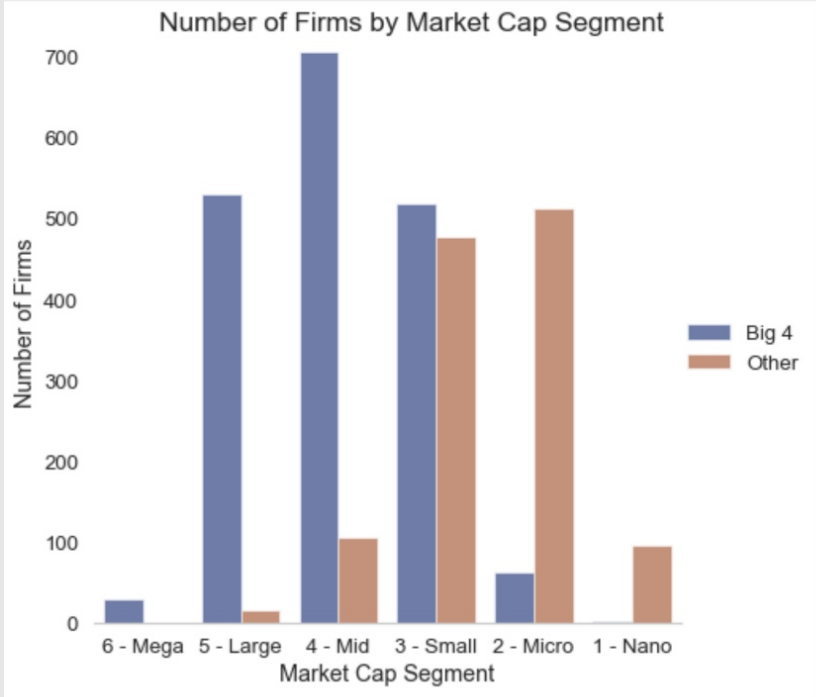

Features
Methodologies
Models

Features

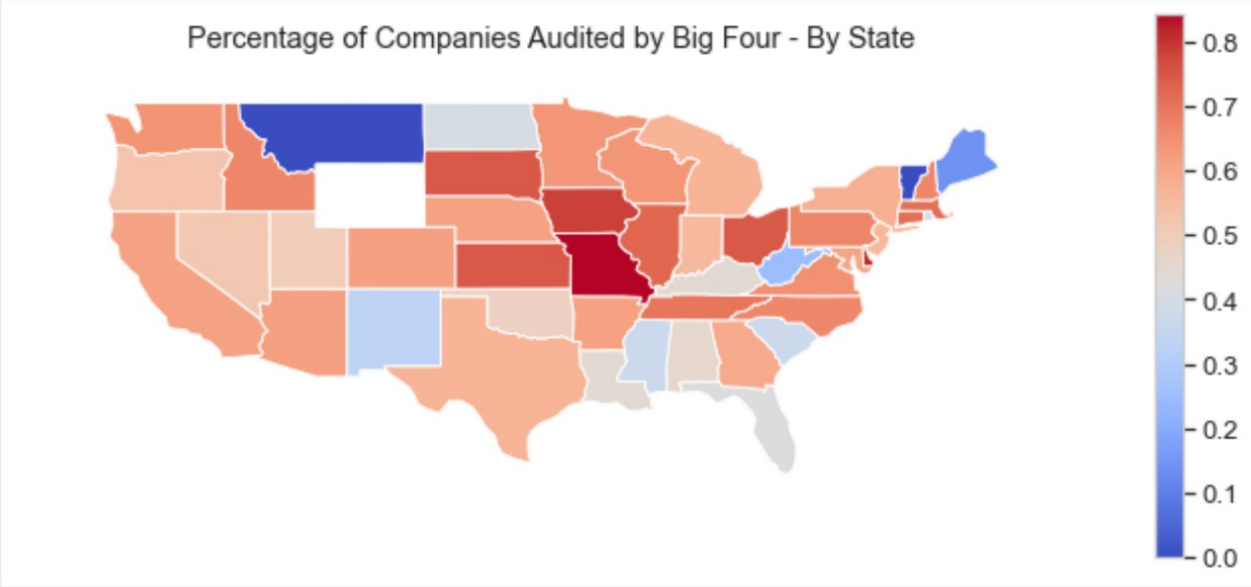
- Cash Flows (Operating, Investing, Etc)
  - Feature engineering: Inflows vs. Outflows
- Balance Sheet Metrics
  - Assets, Cash, Liabilities, Debt, Equity
- Income Statement Metrics
  - Revenue
  - EBITDA
- Enterprise value
- Market value
- Location (State)
- Exchange
- Industry Sector

EDA


Number of Firms by Revenue Segment
Number of Firms
700
600
500
400
300
200
100
0
6 - Mega
5 - Large
4 - Mid
3 - Small
2 - Micro
1 - Nano
Revenue Segment
Big 4
Other

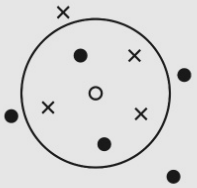

Number of Firms by Market Cap Segment
Number of Firms
700
600
500
400
300
200
100
0
6 - Mega
5 - Large
4 - Mid
3 - Small
2 - Micro
1 - Nano
Market Cap Segment
Big 4
Other

EDA

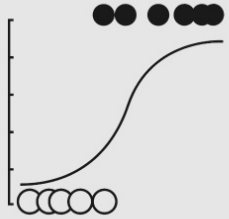


* For additional EDA visualizations see appendix

Models

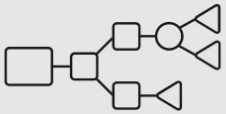


KNN



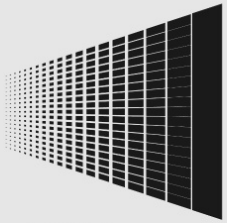
Logistic Regression

Naive Bayes

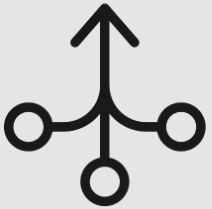


Decision Tree

Random Forest



XGBoost

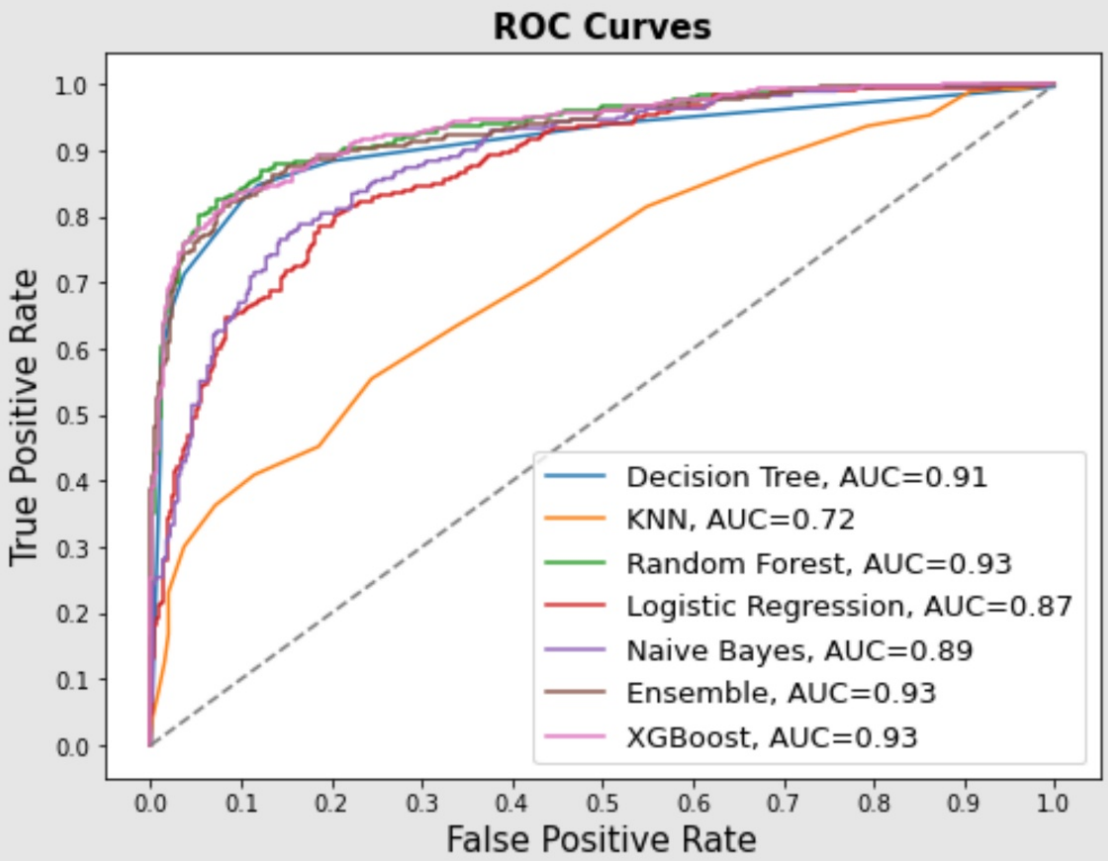


Ensemble

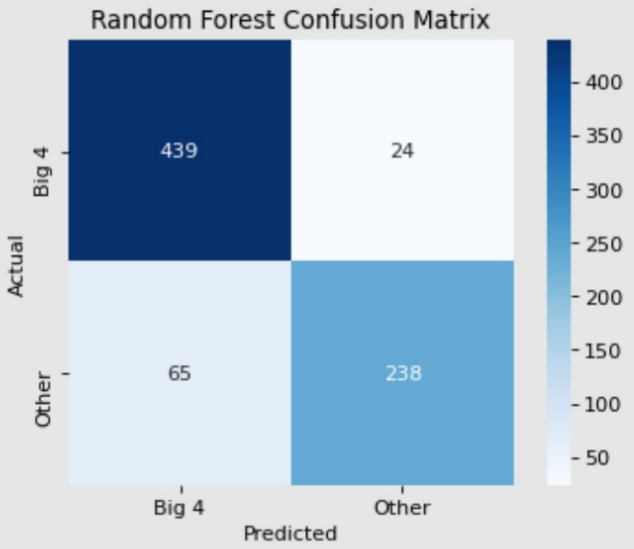

Metrics
Findings
Model Insight

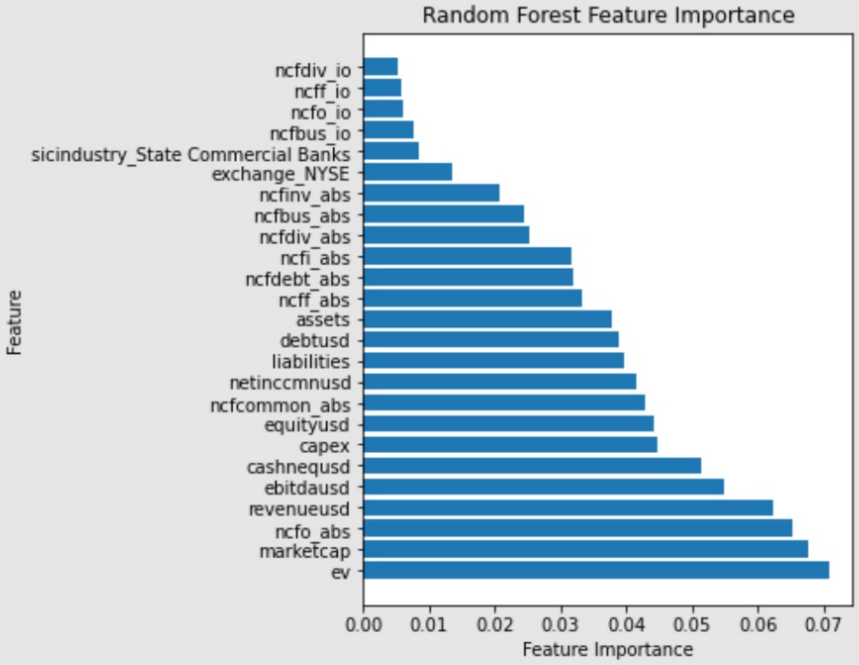
Metrics

	Accuracy	F1 Score
KNN	.64	.67
Naive Bayes	.79	.79
Logistic Regression	.80	.78
Decision Tree	.87	.86
Random Forest	**.88**	**.88**
XGBoost	.88	.87
Ensemble	.87	.87

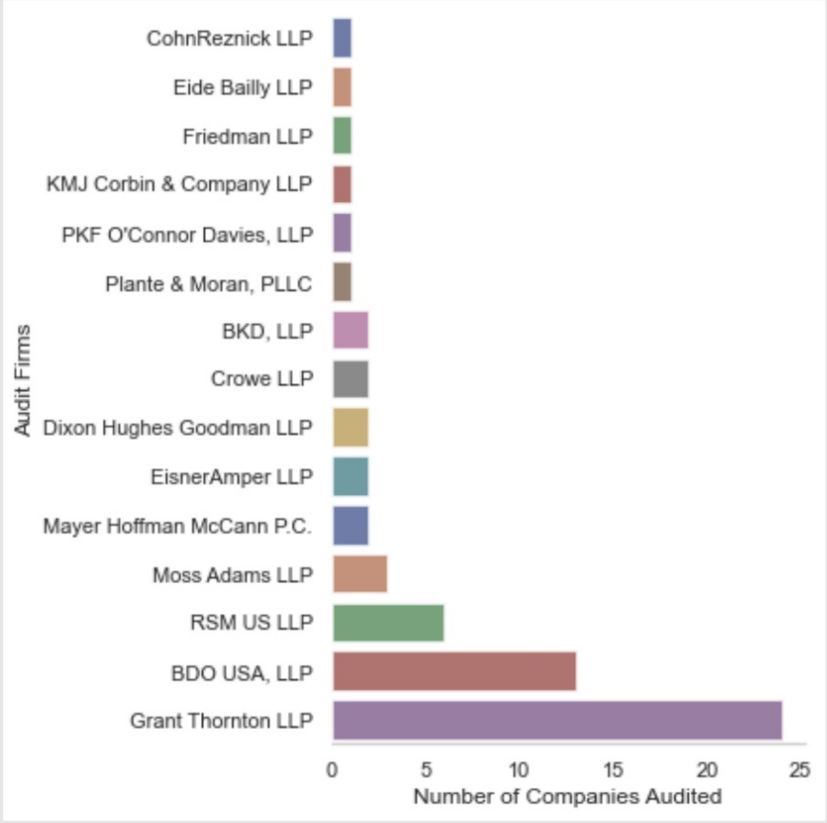

ROC Curves
True Positive Rate
False Positive Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Decision Tree, AUC=0.91
KNN, AUC=0.72
Random Forest, AUC=0.93
Logistic Regression, AUC=0.87
Naive Bayes, AUC=0.89
Ensemble, AUC=0.93
XGBoost, AUC=0.93

Random Forest

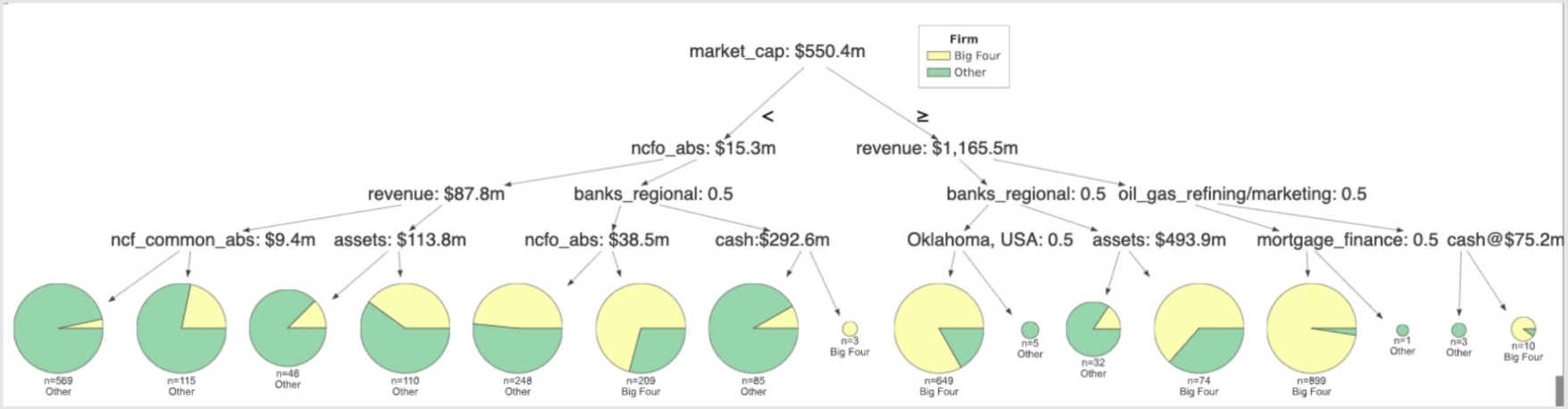




False Positives (n=65)



Decision Tree

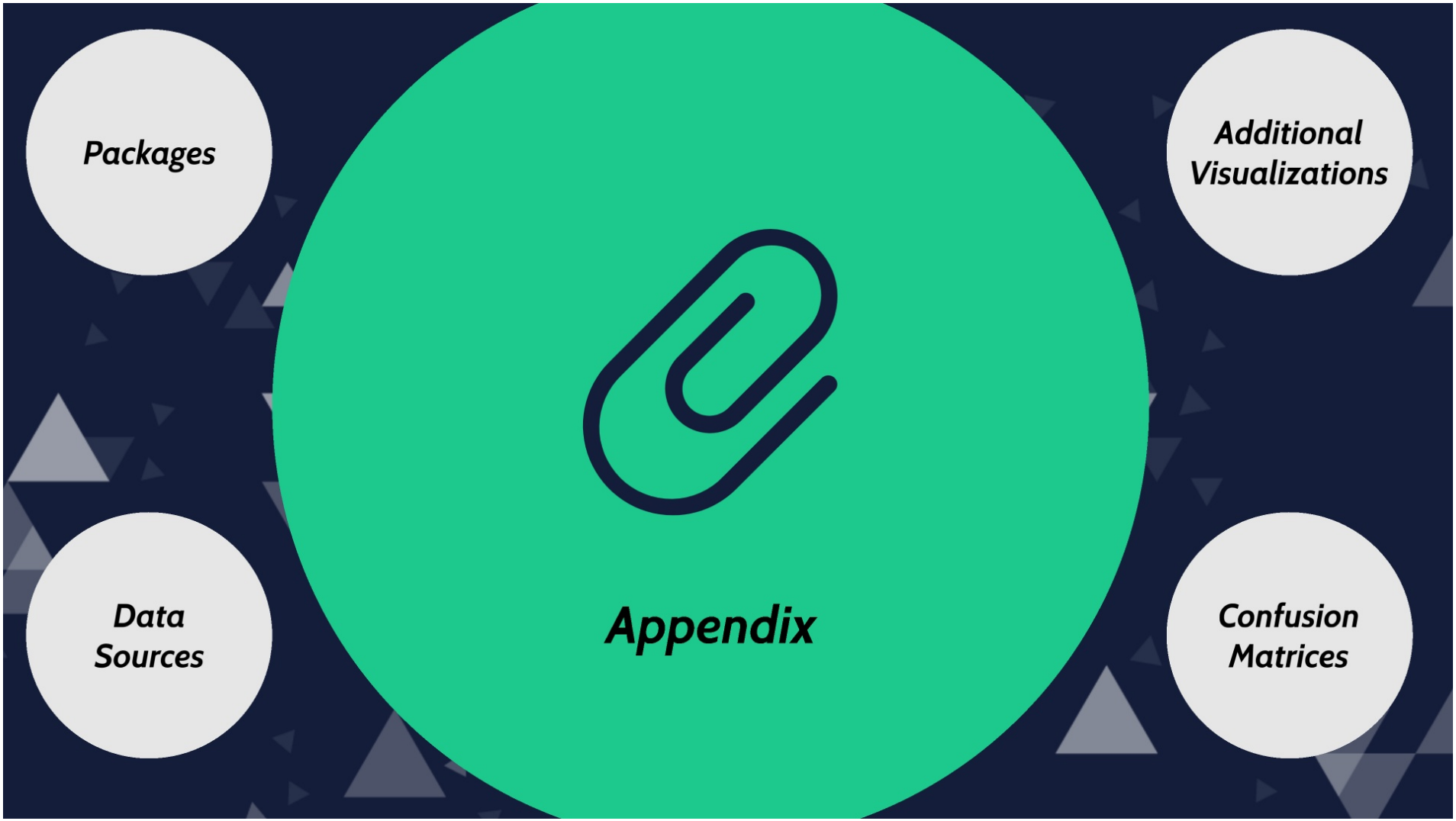



Future Work
Future Work
Q&A

Future Work

- Multiclass classification
  - Identify specific firm
- Time-series analyses
  - When do companies upgrade?
- Create internal tool for audit firms to predict probably of winning a bid

Q&A


Packages
Additional Visualizations
Appendix
Data Sources
Confusion Matrices

Packages

NumPy

pandas

matplotlib

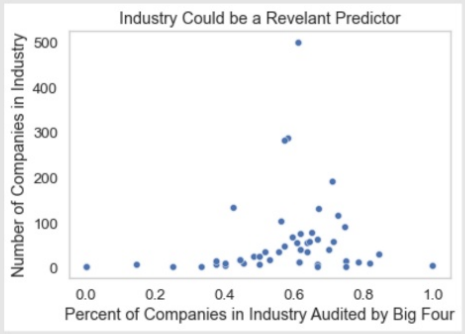
seaborn

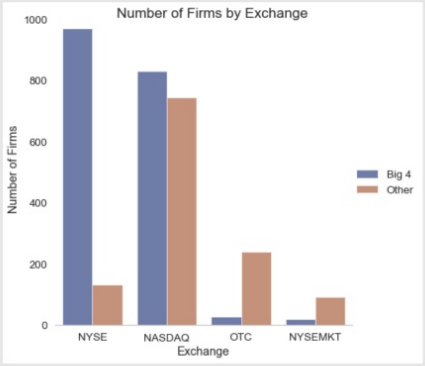
Data Sources

PCAOB: Auditor Search
https://pcaobus.org/resources/auditorsearch

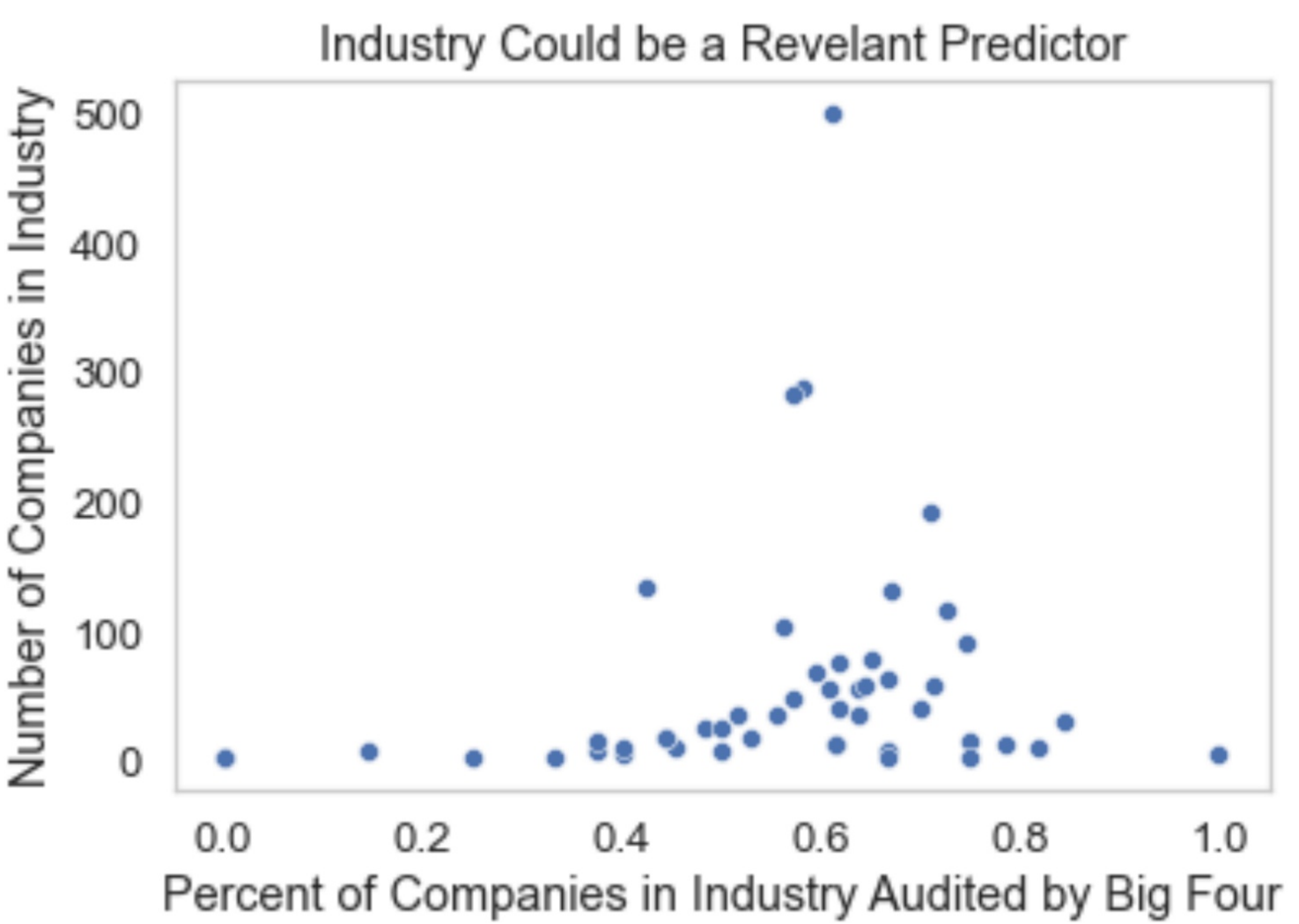
Quandl: US Company Fundamentals
https://www.quandl.com/databases/SF1/data

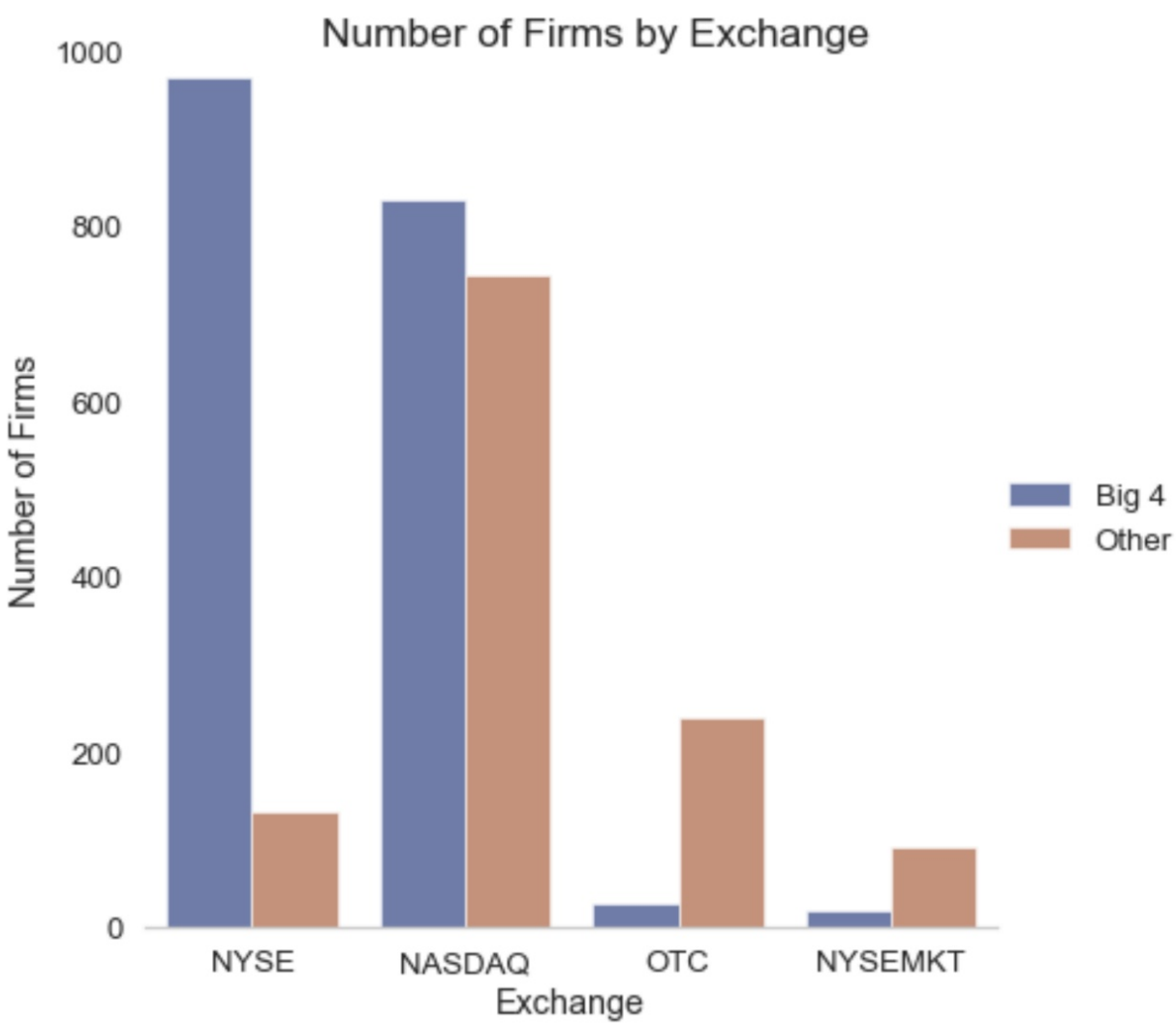
Additional Visualizations



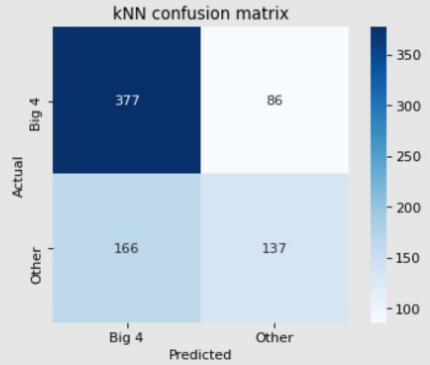


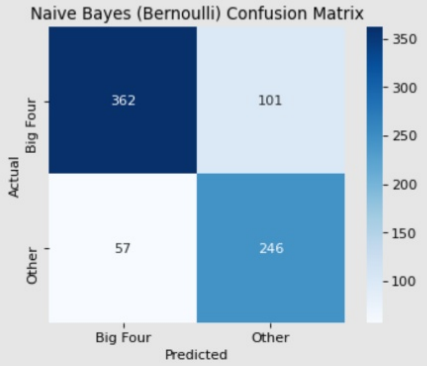

Imbalanced Data
Number of Companies
1750
1500
1250
1000
750
500
250
0
Big 4
Other
Auditor

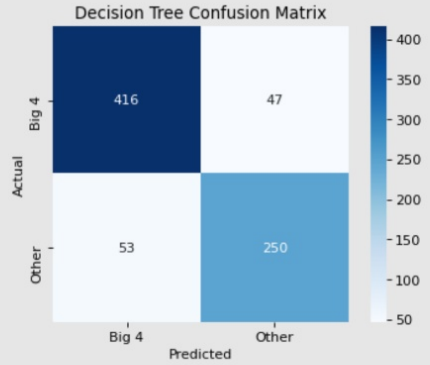

Industry Could be a Revelant Predictor
Number of Companies in Industry
500
400
300
200
100
0
0.0
0.2
0.4
0.6
0.8
1.0
Percent of Companies in Industry Audited by Big Four

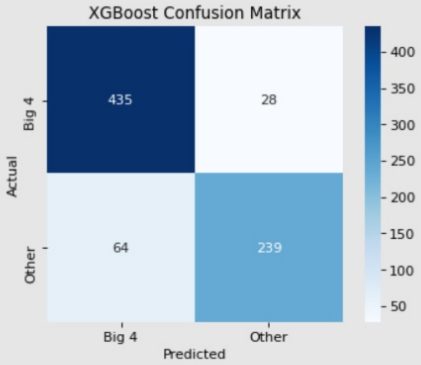

Number of Firms by Exchange
1000
800
600
400
200
0
Number of Firms
NYSE
NASDAQ
OTC
NYSEMKT
Exchange
Big 4
Other

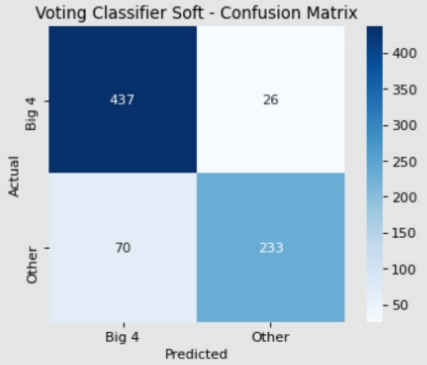
Confusion Matrices












Predicting Financial Statement Auditor
Supervised Learning: Classification
Appendix
Findings
Future Work
Background & Purpose
Methodologies
Jason Dunleavy, Data Scientist