
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

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


Introduction
Background &
Purpose
Data

Data

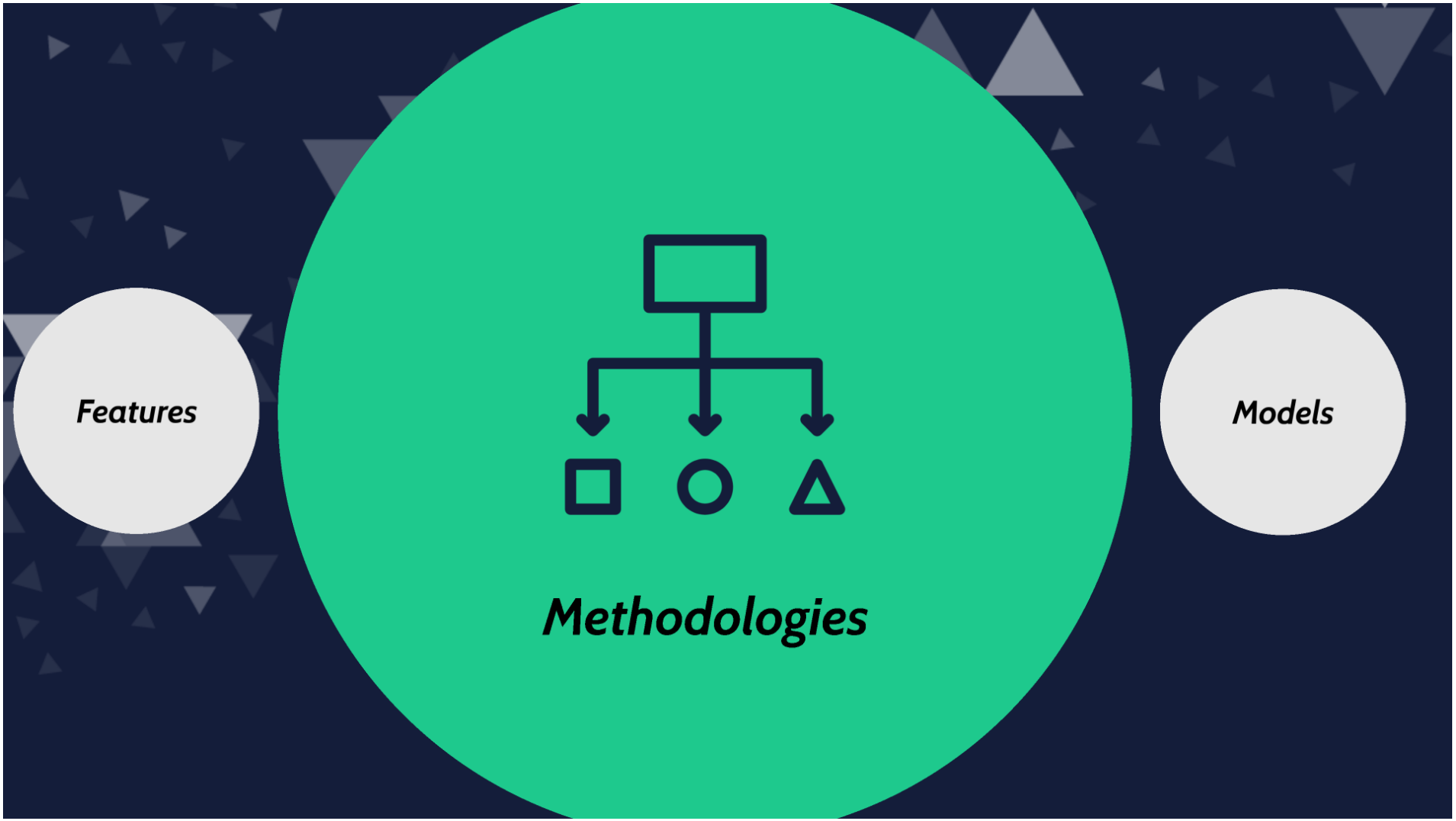
- Data sources:
  - PCAOB: Public Company Auditor Search
  - Quandl: US Company Fundamentals
- Created/queried PostgreSQL database

Data

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


Introduction
Background &
Purpose
Data


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


Features
Methodologies
Models

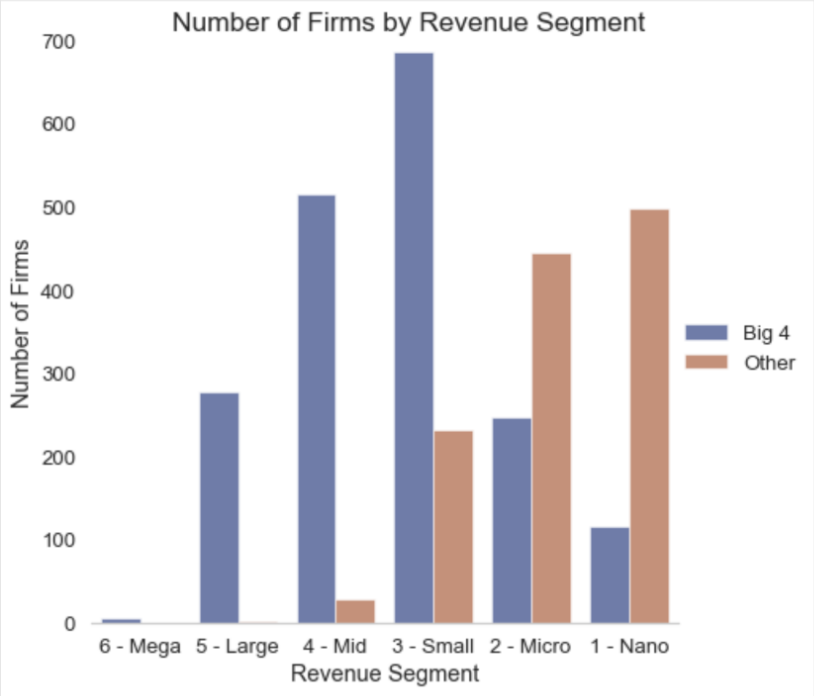
Features

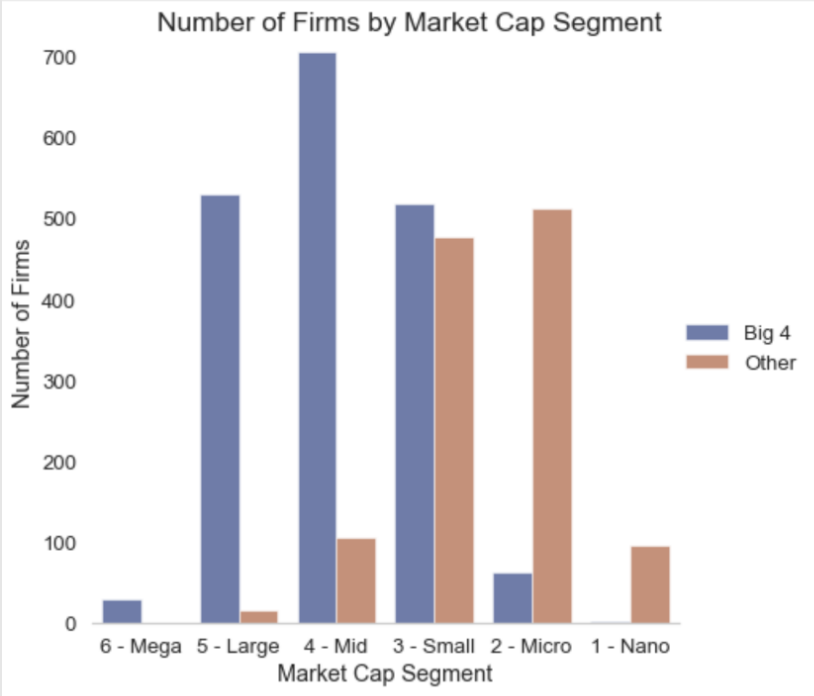
- Cash Flows (Operating, Investing, Etc)
  - Feature engineering: Inflows vs. Outflows
- Balance Sheet Metrics
  - Assets, Cash, Liabilities, Debt, Equity
- Income Statement Metrics
  - Revenue
  - EBITDA
- Enterprise value
- Market value
- Location (State)
- Exchange
- Industry Sector

Features

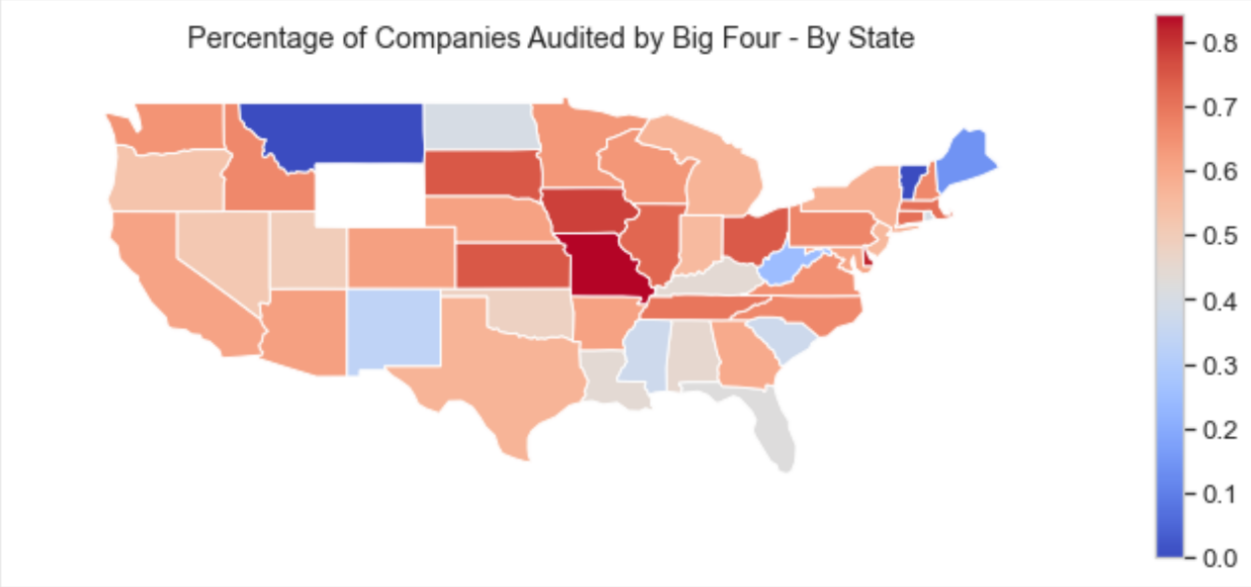
- Cash Flows (Operating, Investing, Etc)
  - Feature engineering: Inflows vs. Outflows
- Balance Sheet Metrics
  - Assets, Cash, Liabilities, Debt, Equity
- Income Statement Metrics
  - Revenue
  - EBITDA
- Enterprise value
- Market value
- Location (State)
- Exchange
- Industry Sector

EDA

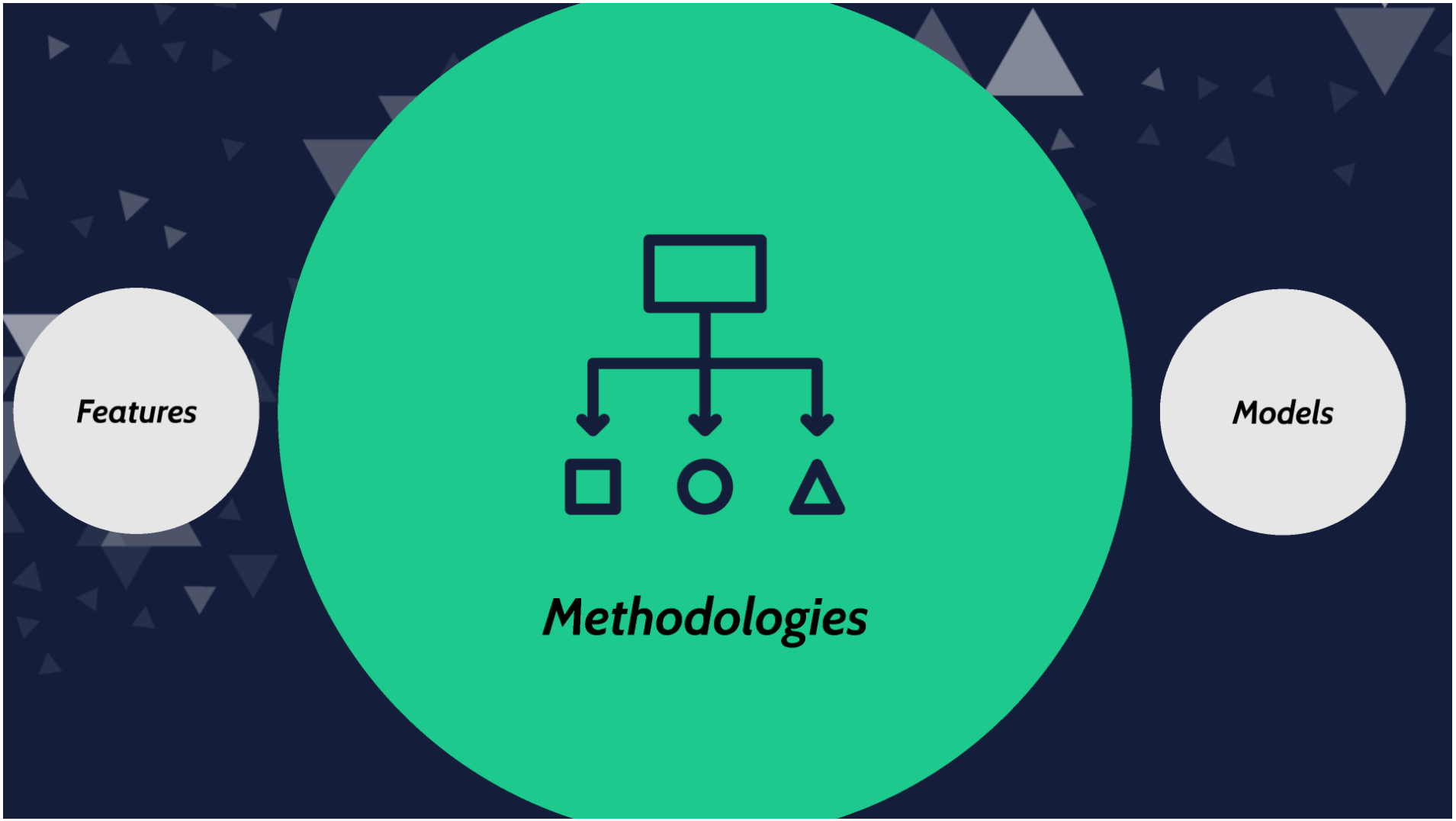

Number of Firms by Revenue Segment
Number of Firms
700
600
500
400
300
200
100
0
6 - Mega
5 - Large
4 - Mid
3 - Small
2 - Micro
1 - Nano
Revenue Segment
Big 4
Other


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

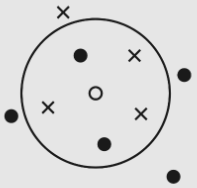
EDA



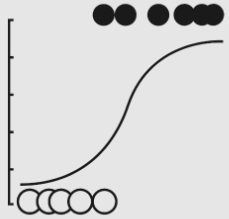
* For additional EDA visualizations see appendix


Features
Methodologies
Models

Models

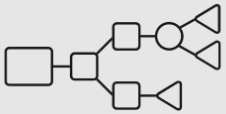


KNN



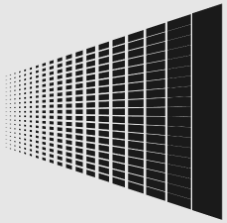
Logistic Regression

Naive Bayes

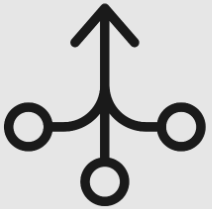


Decision Tree

Random Forest

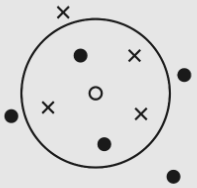


XGBoost

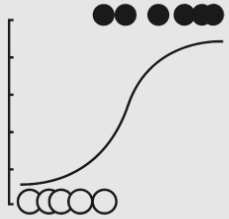


Ensemble

Models

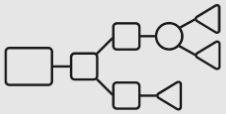


KNN



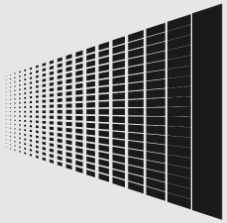
Logistic Regression

Naive Bayes

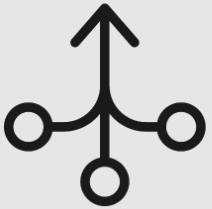


Decision Tree

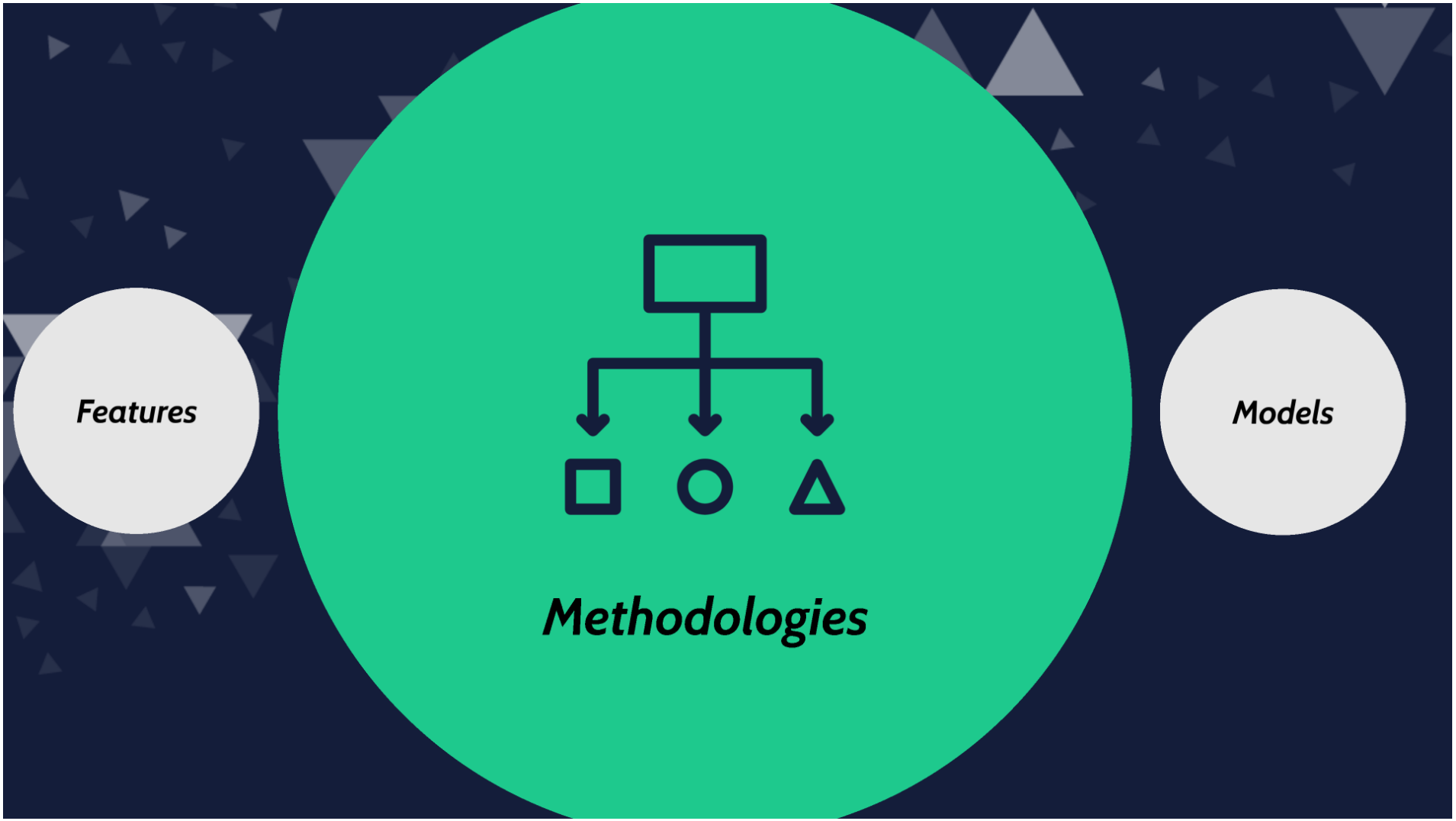
Random Forest



XGBoost



Ensemble


Features
Methodologies
Models


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

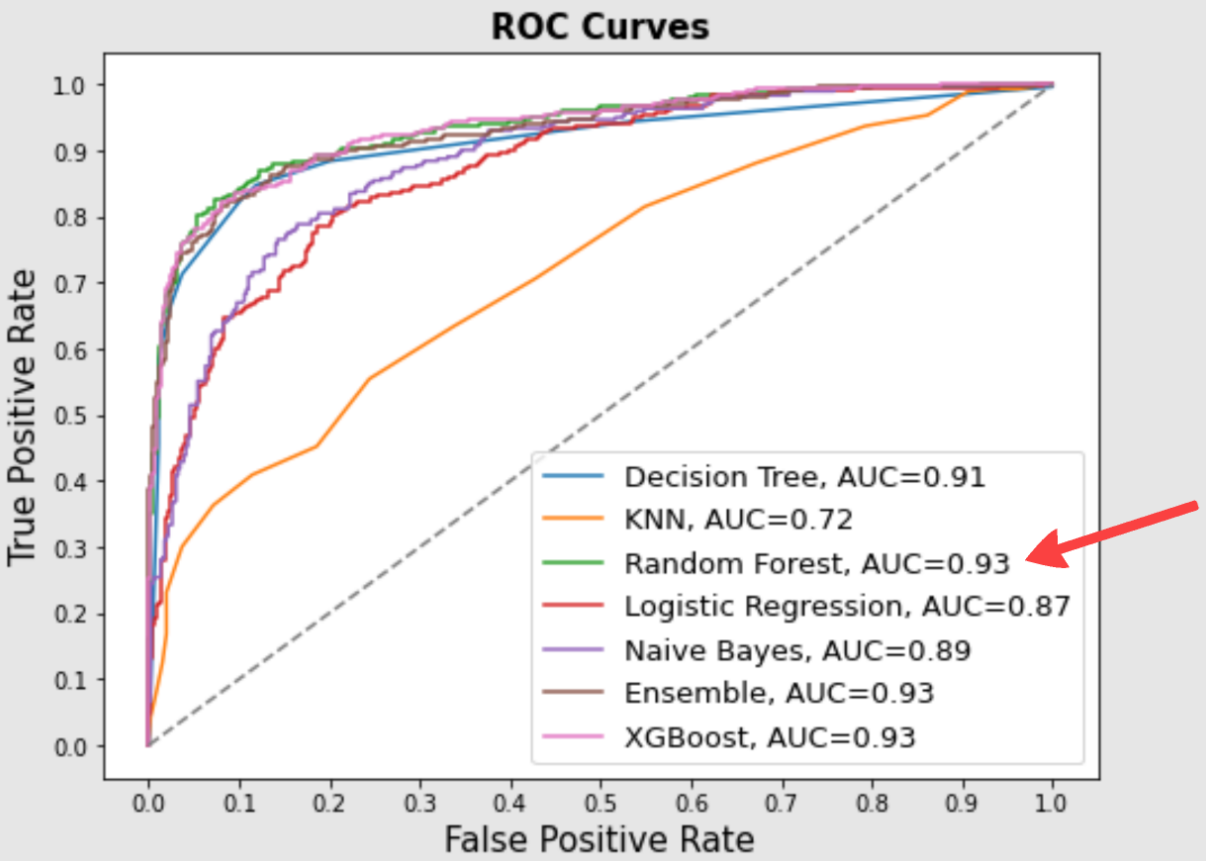

Metrics
Findings
Model Insight

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

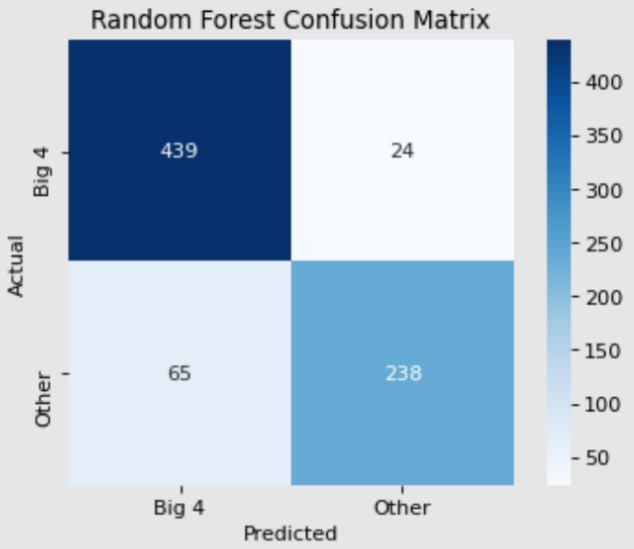
Metrics

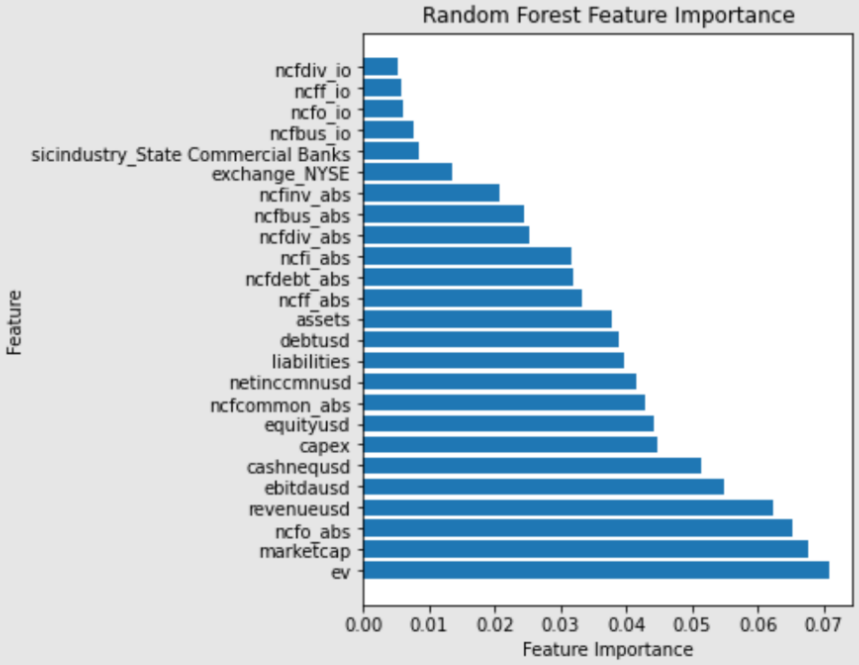
	Accuracy	F1 Score
KNN	.64	.67
Naive Bayes	.79	.79
Logistic Regression	.80	.78
Decision Tree	.87	.86
Random Forest	.88	**.88**
XGBoost	.88	.87
Ensemble	.87	.87


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

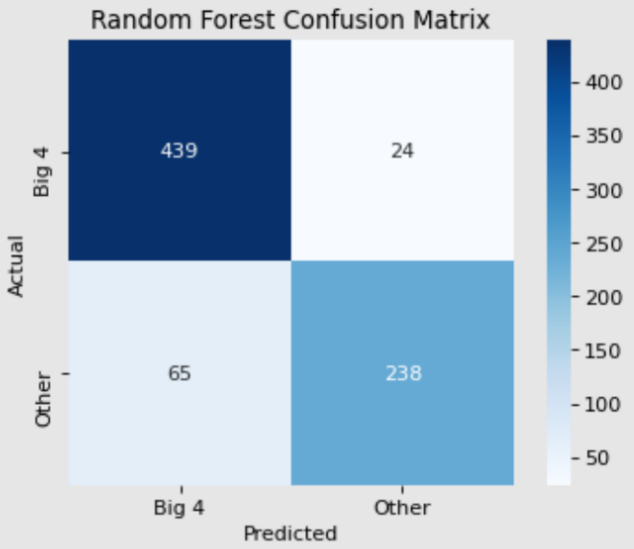

Metrics
Findings
Model Insight

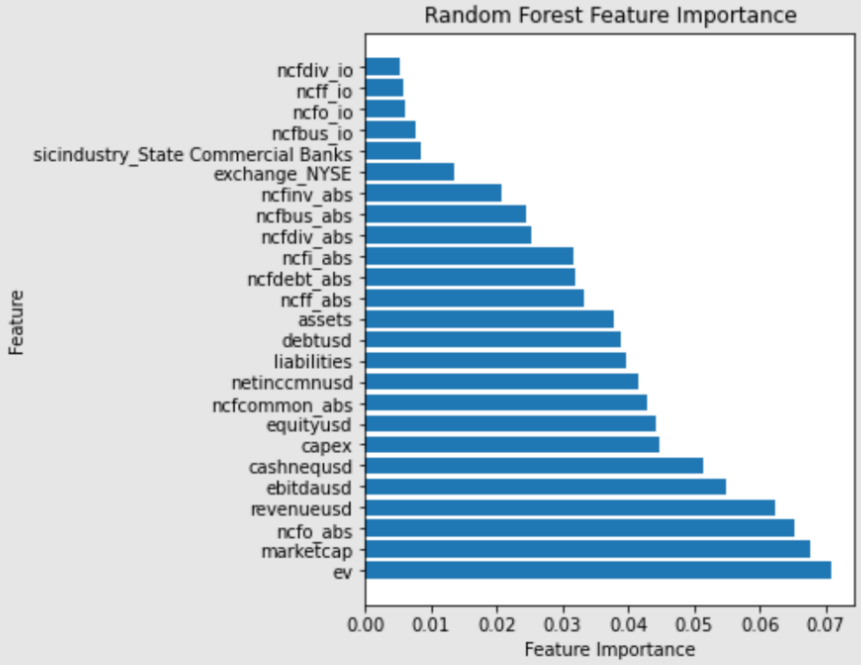
Random Forest





Random Forest





False Positives (n=65)