

Predicting Financial Statement Auditor

Supervised Learning: Classification



Jason Dunleavy, Data Scientist



Background & Purpose

Introduction

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

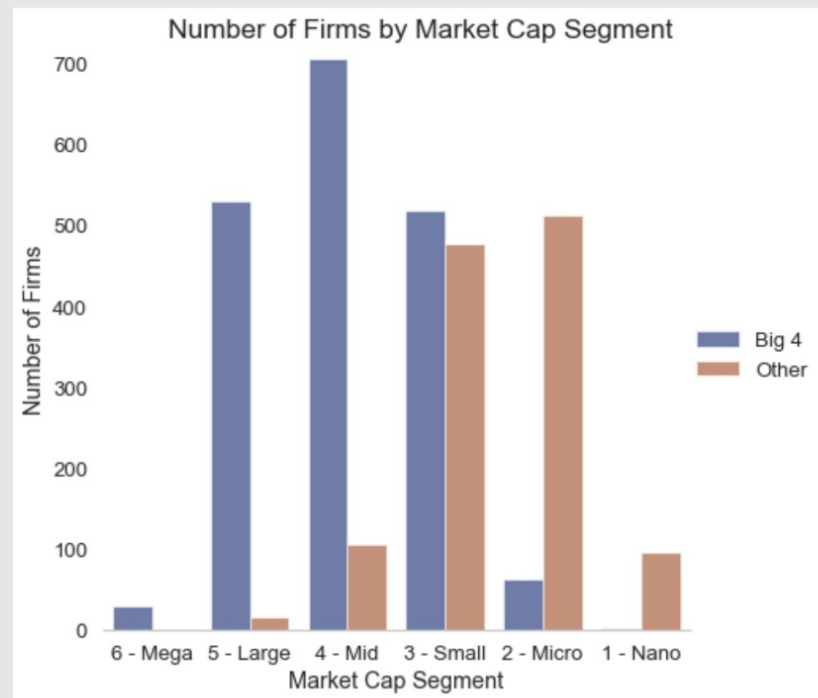
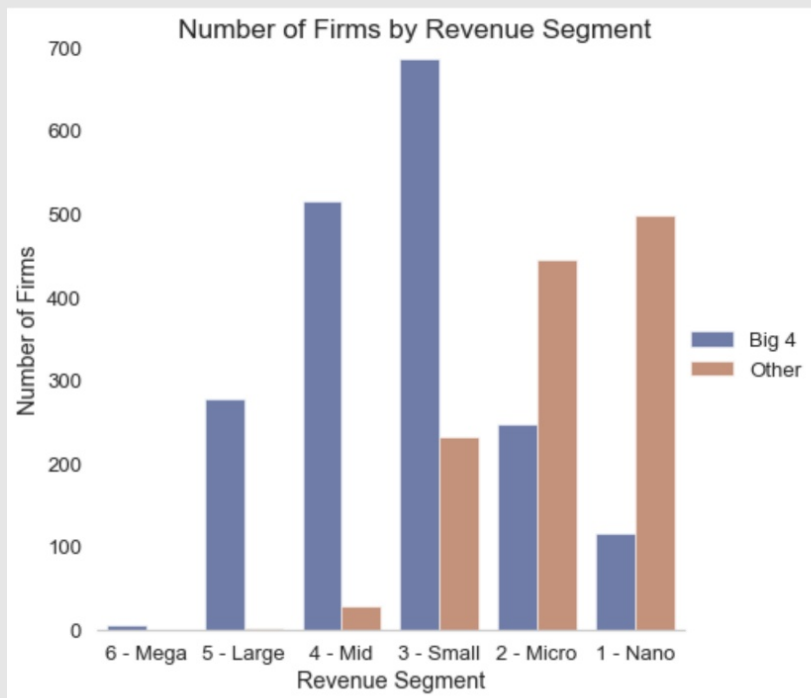




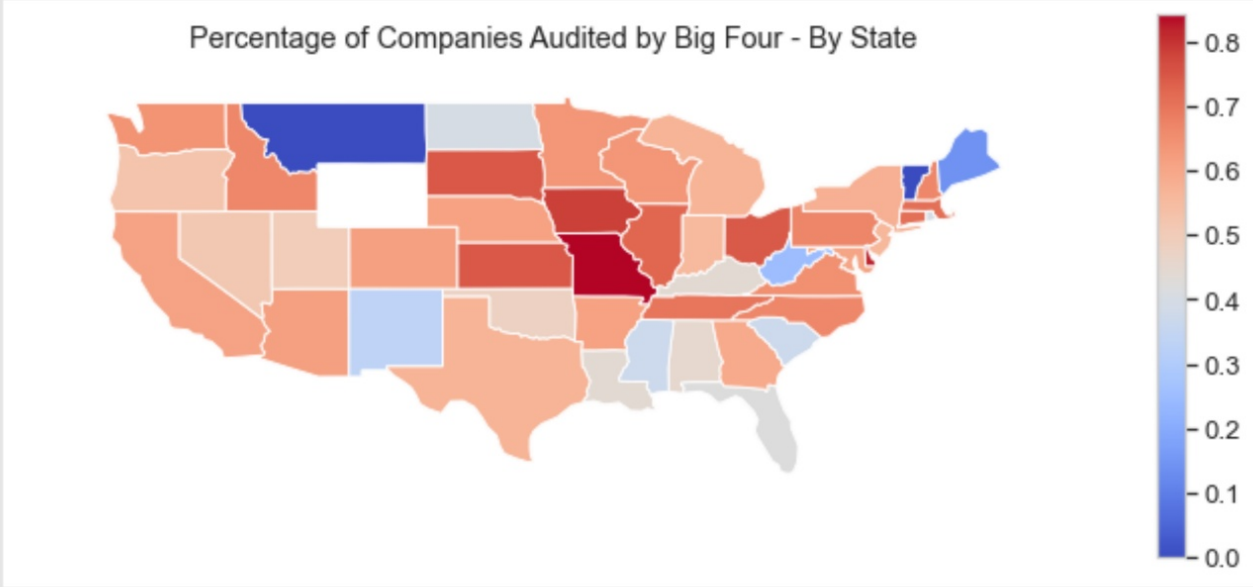
Features

- 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

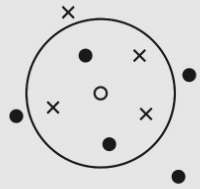


EDA

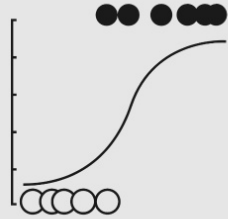


* For additional EDA visualizations see appendix

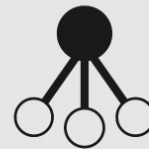
Models



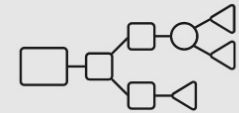
KNN



Logistic Regression



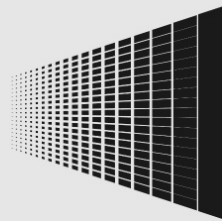
Naive Bayes



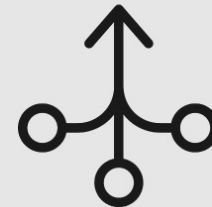
Decision Tree



Random Forest



XGBoost

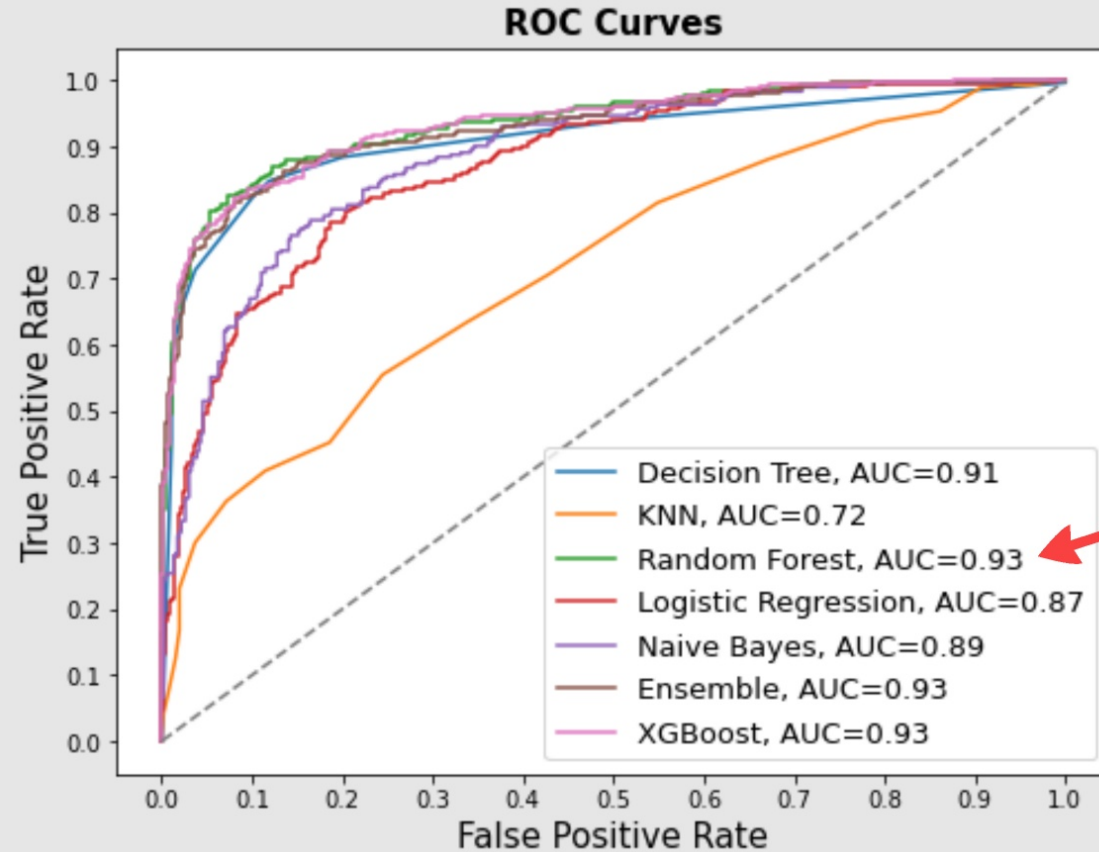


Ensemble

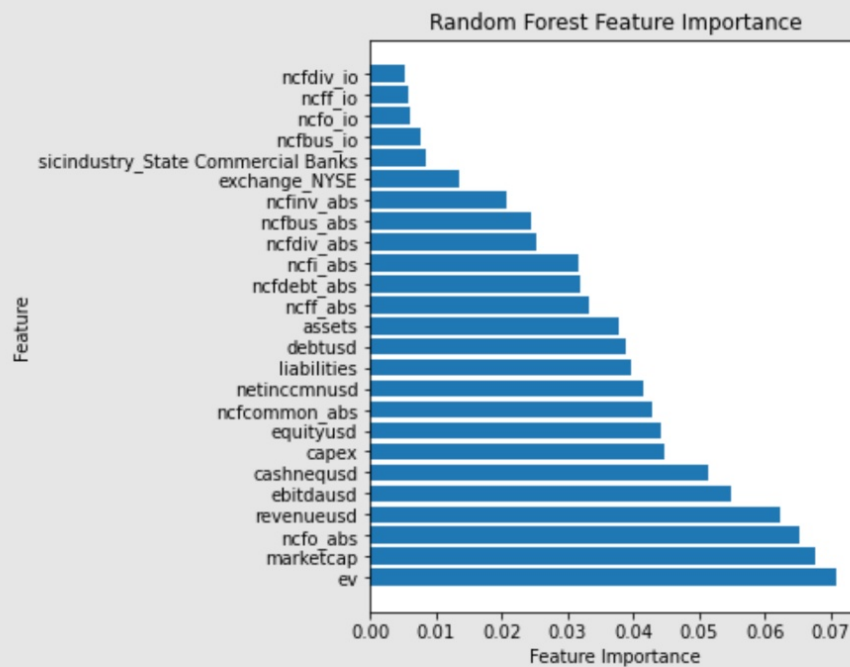
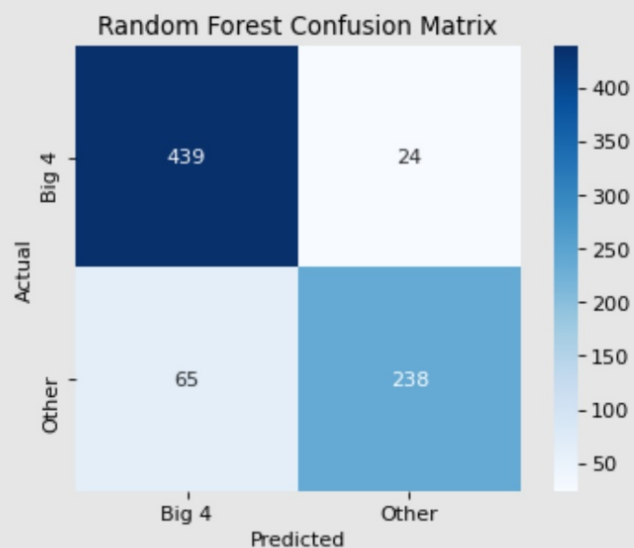


Metrics

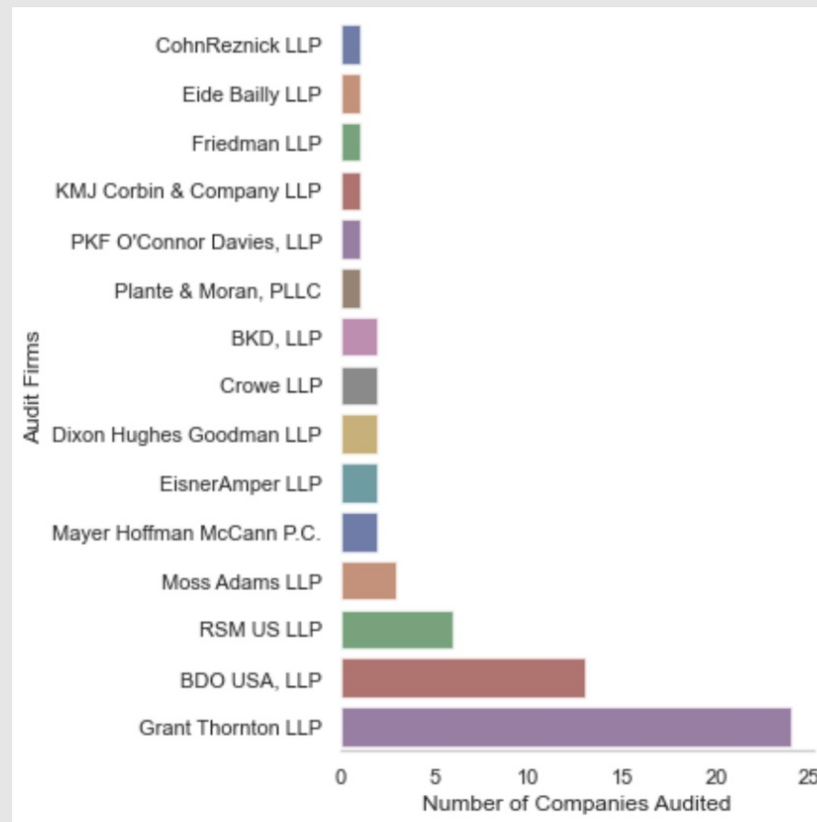
	Accuracy	F1 Score
<i>KNN</i>	.64	.67
<i>Naive Bayes</i>	.79	.79
<i>Logistic Regression</i>	.80	.78
<i>Decision Tree</i>	.87	.86
<i>Random Forest</i>	.88	.88
<i>XGBoost</i>	.88	.87
<i>Ensemble</i>	.87	.87



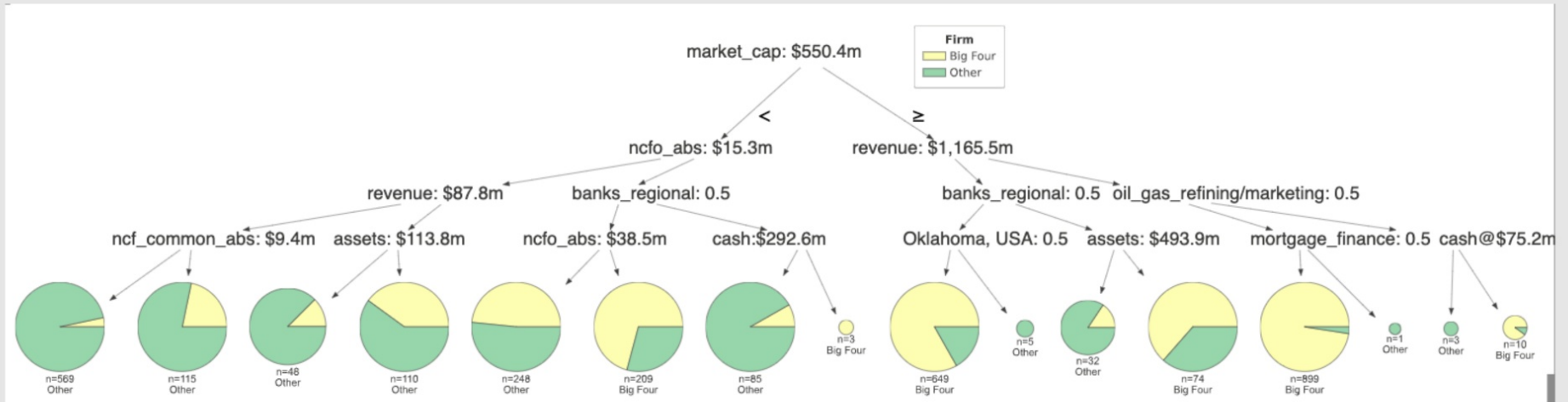
Random Forest



False Positives (n=65)



Decision Tree

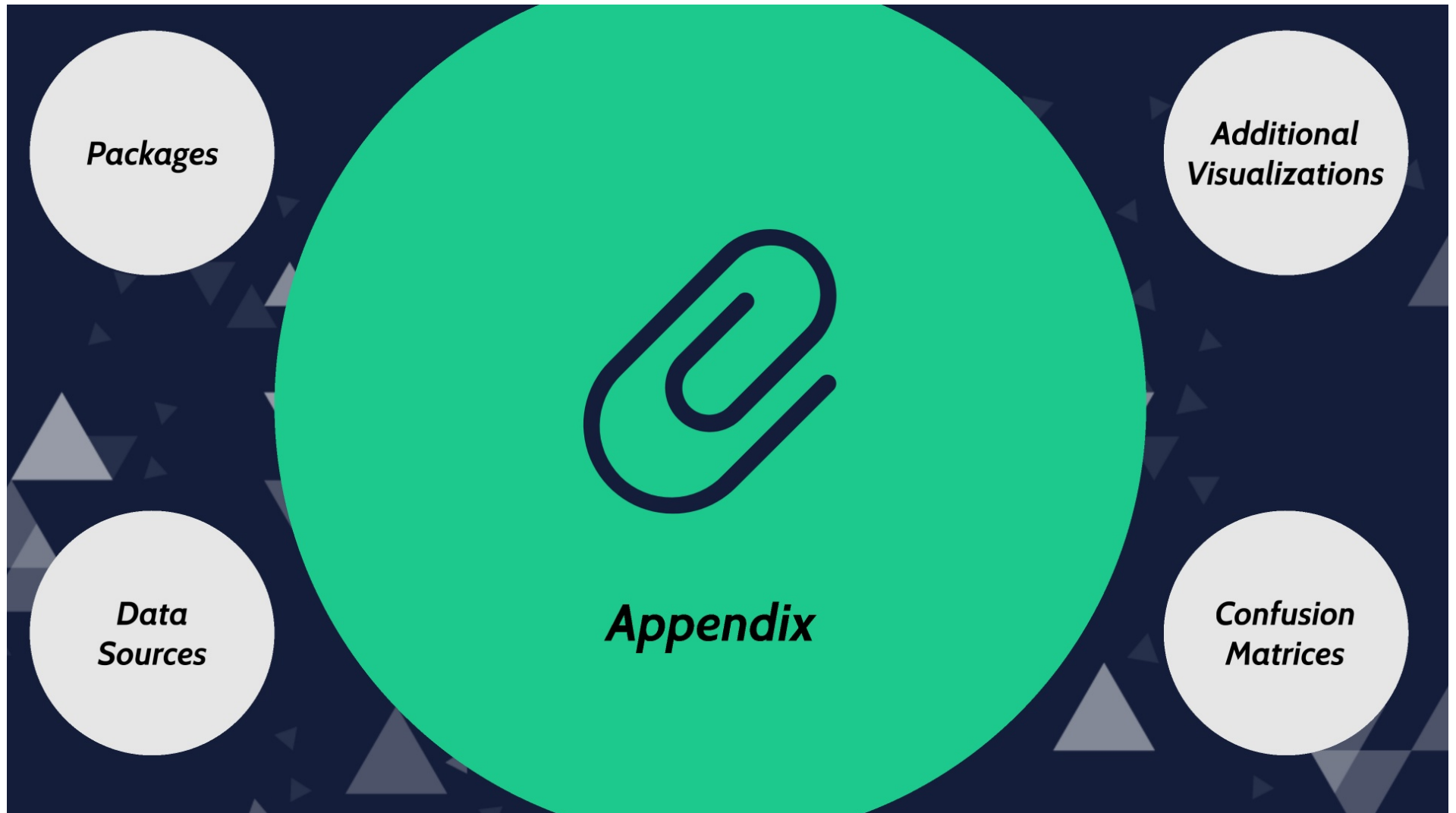




Future Work

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





Packages/Tools



Data Sources

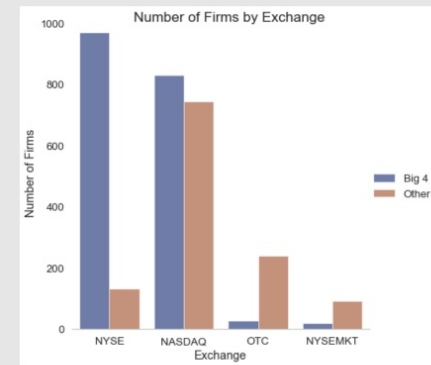
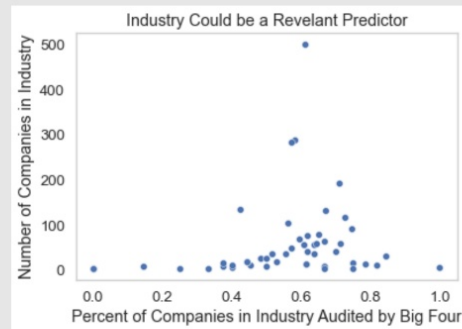
PCAOB: Auditor Search

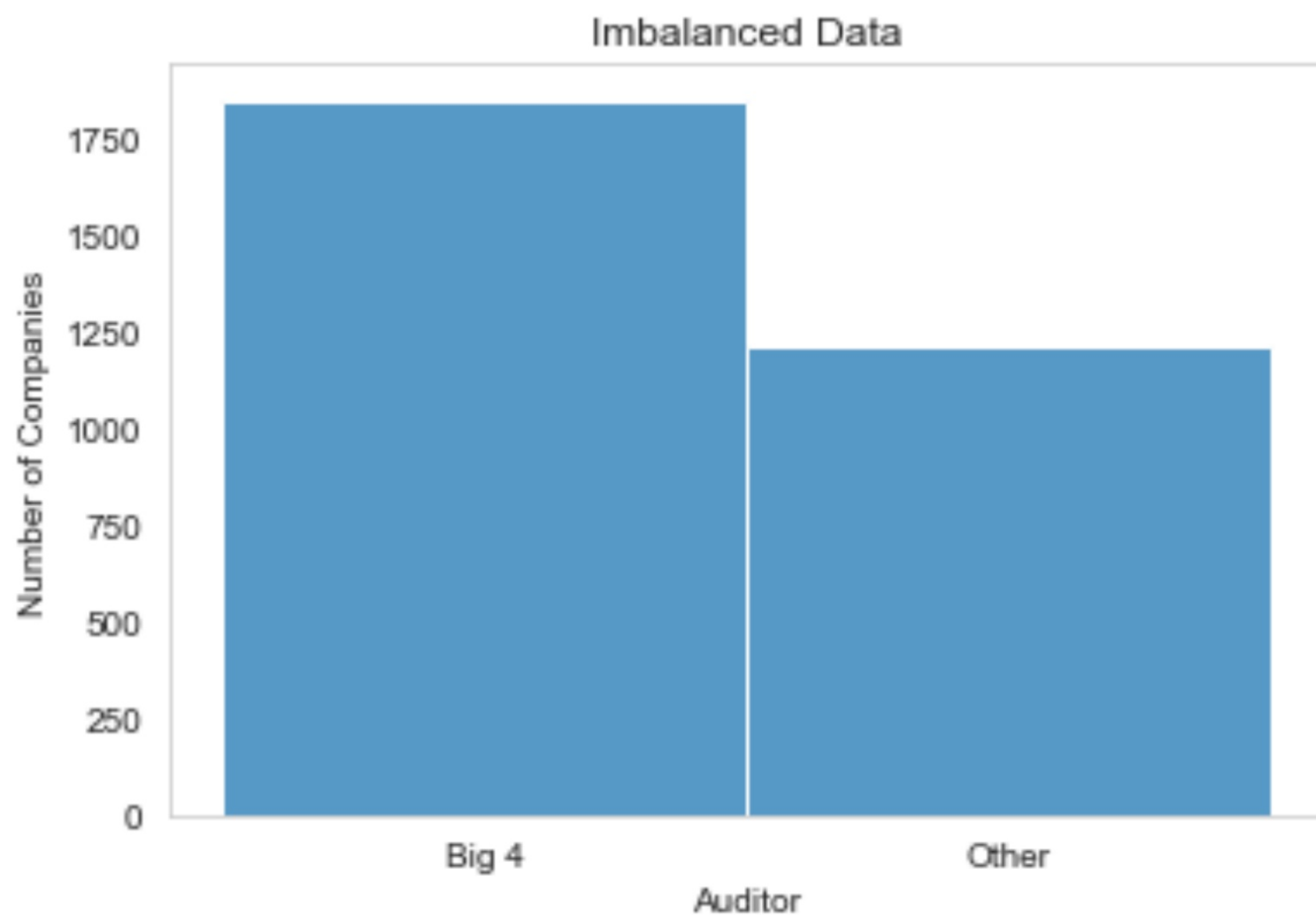
<https://pcaobus.org/resources/auditorsearch>

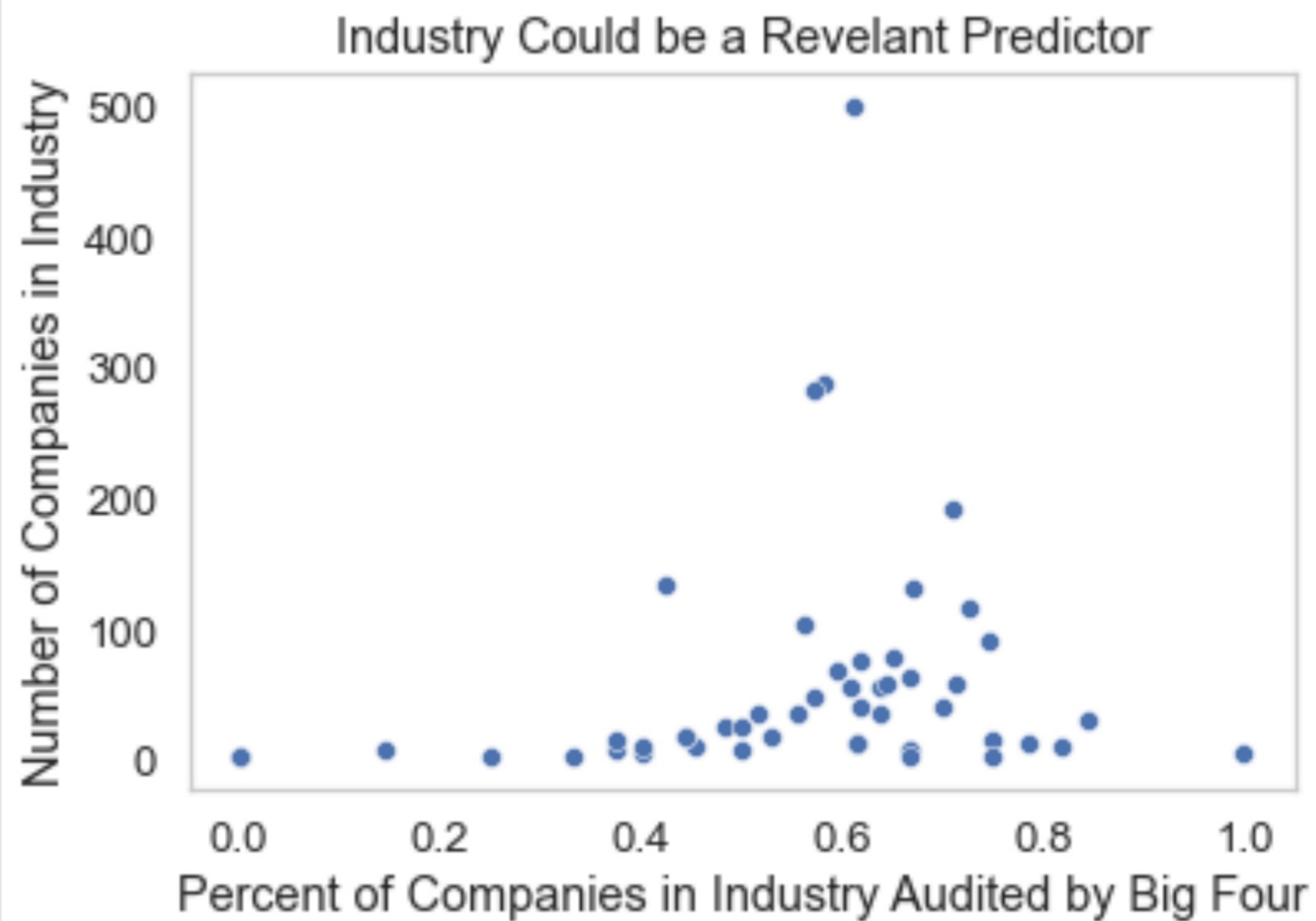
Quandl: US Company Fundamentals

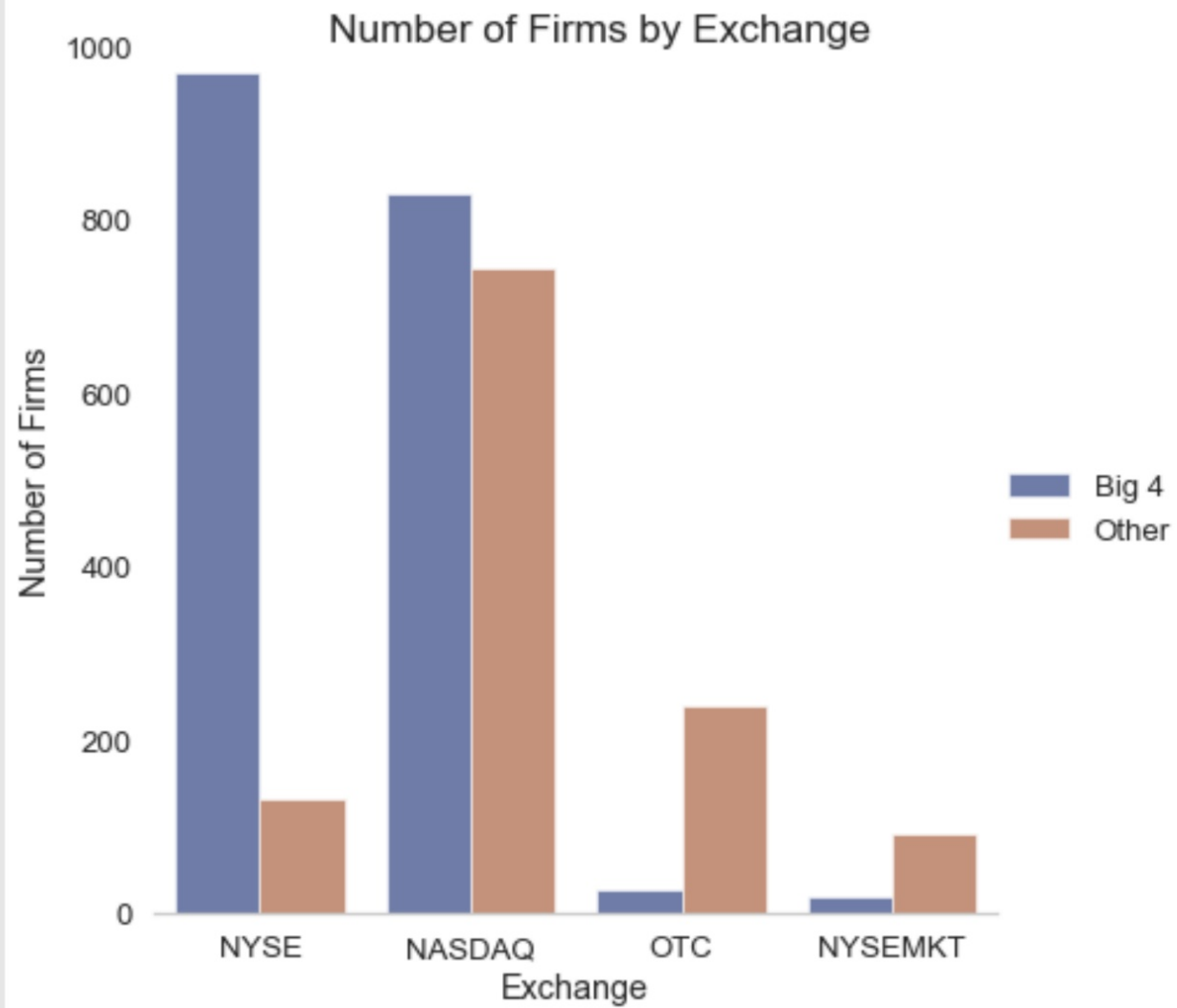
<https://www.quandl.com/databases/SFI/data>

Additional Visualizations

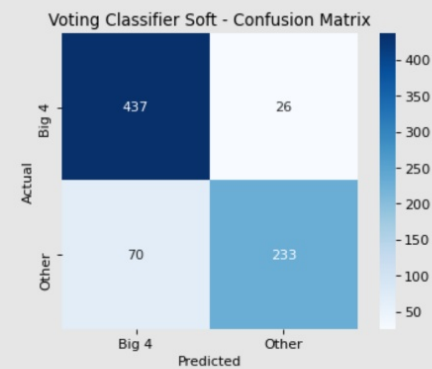
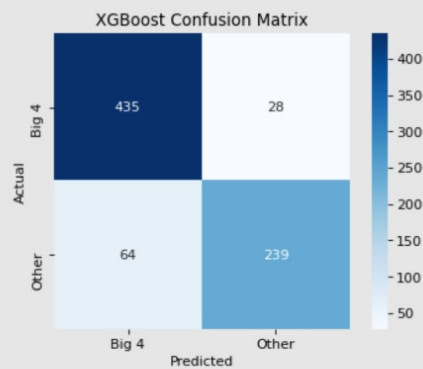
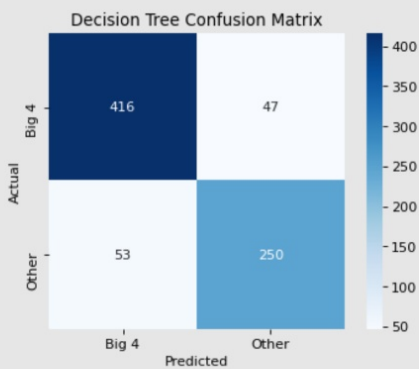
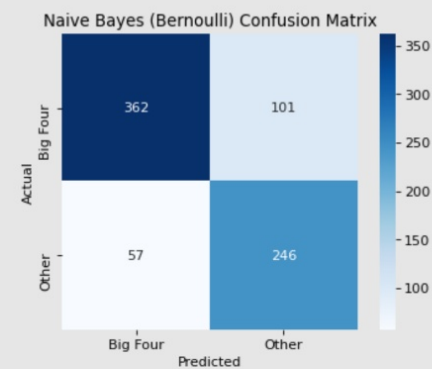
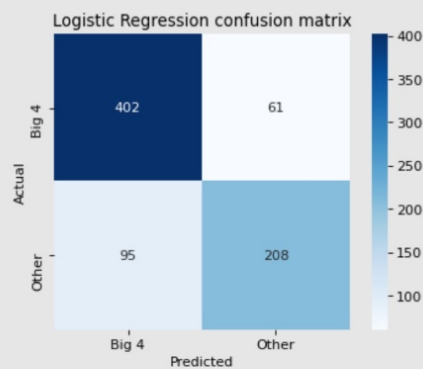
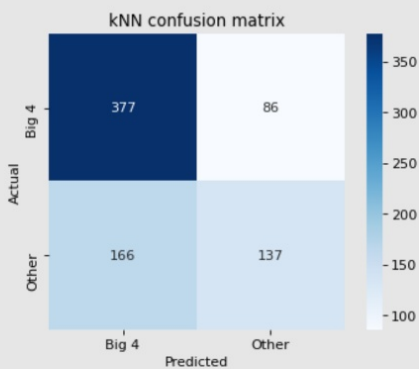








Confusion Matrices



Predicting Financial Statement Auditor

Supervised Learning: Classification



Jason Dunleavy, Data Scientist