Unicorn Classifier

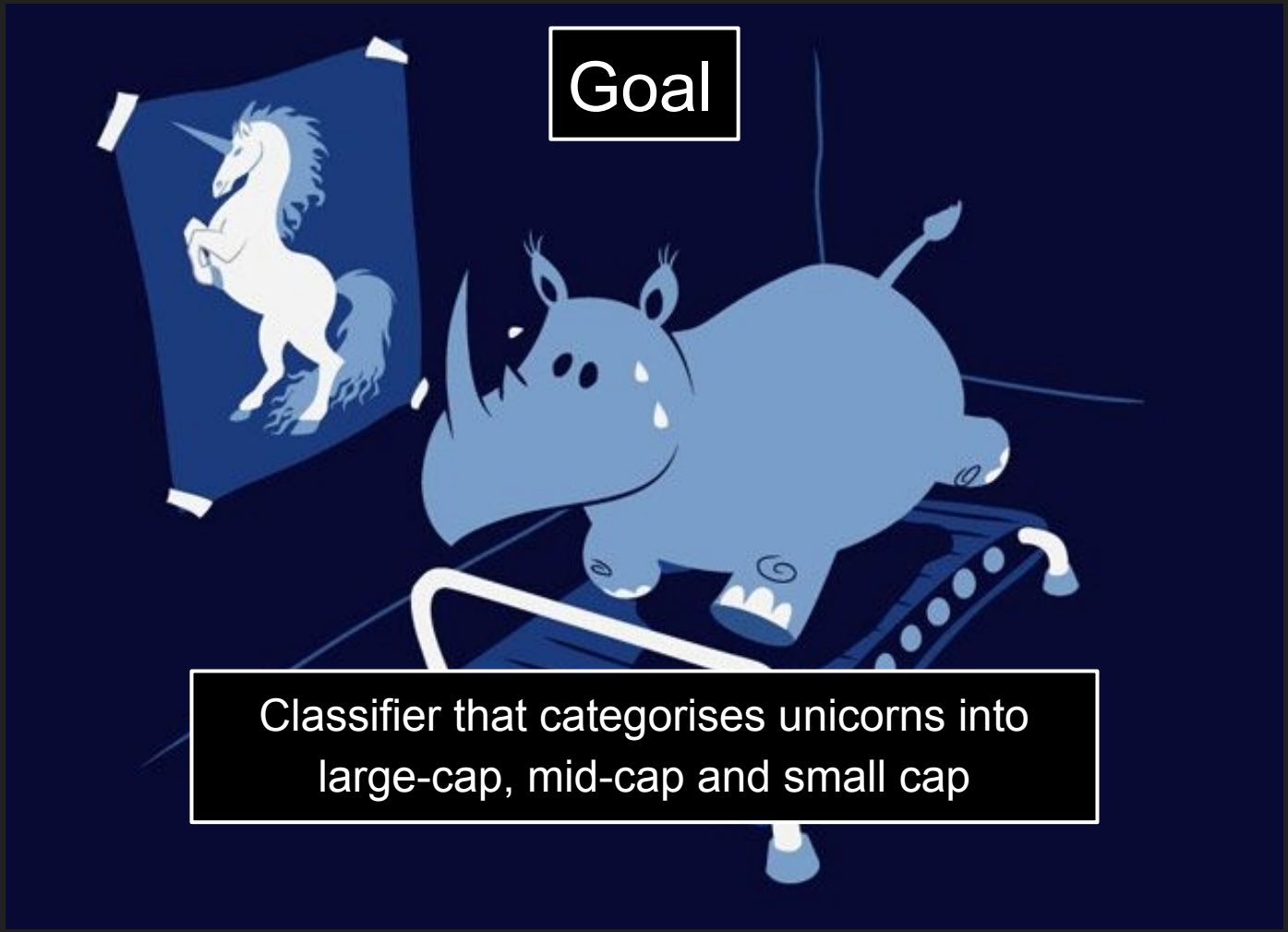
What Is An Unicorn?

I’m Fantastic?

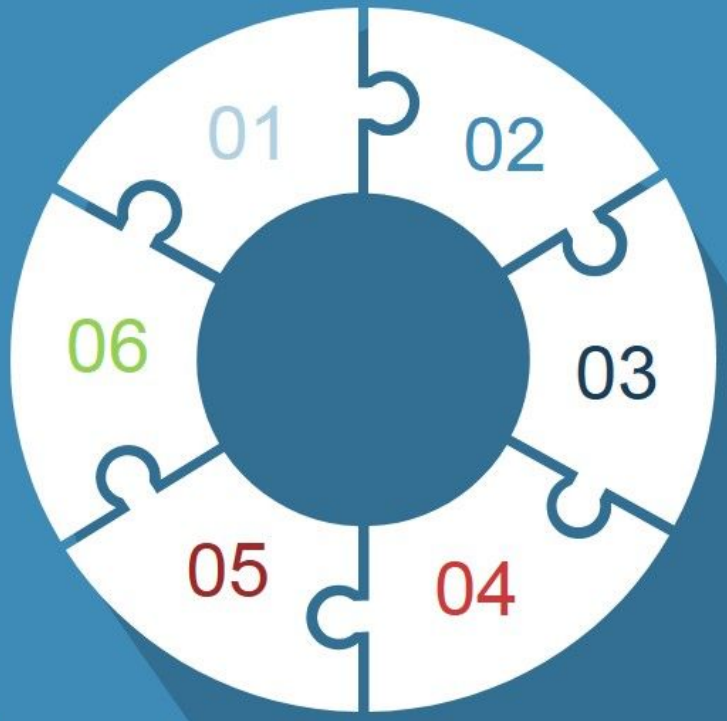
Average unicorn only worth half its valuation

How to know if a unicorn is really a unicorn?

Since market cap for public companies could represent market value, can we determine market cap for unicorns?


Goal
Classifier that categorises unicorns into large-cap, mid-cap and small cap

Capstone Approach



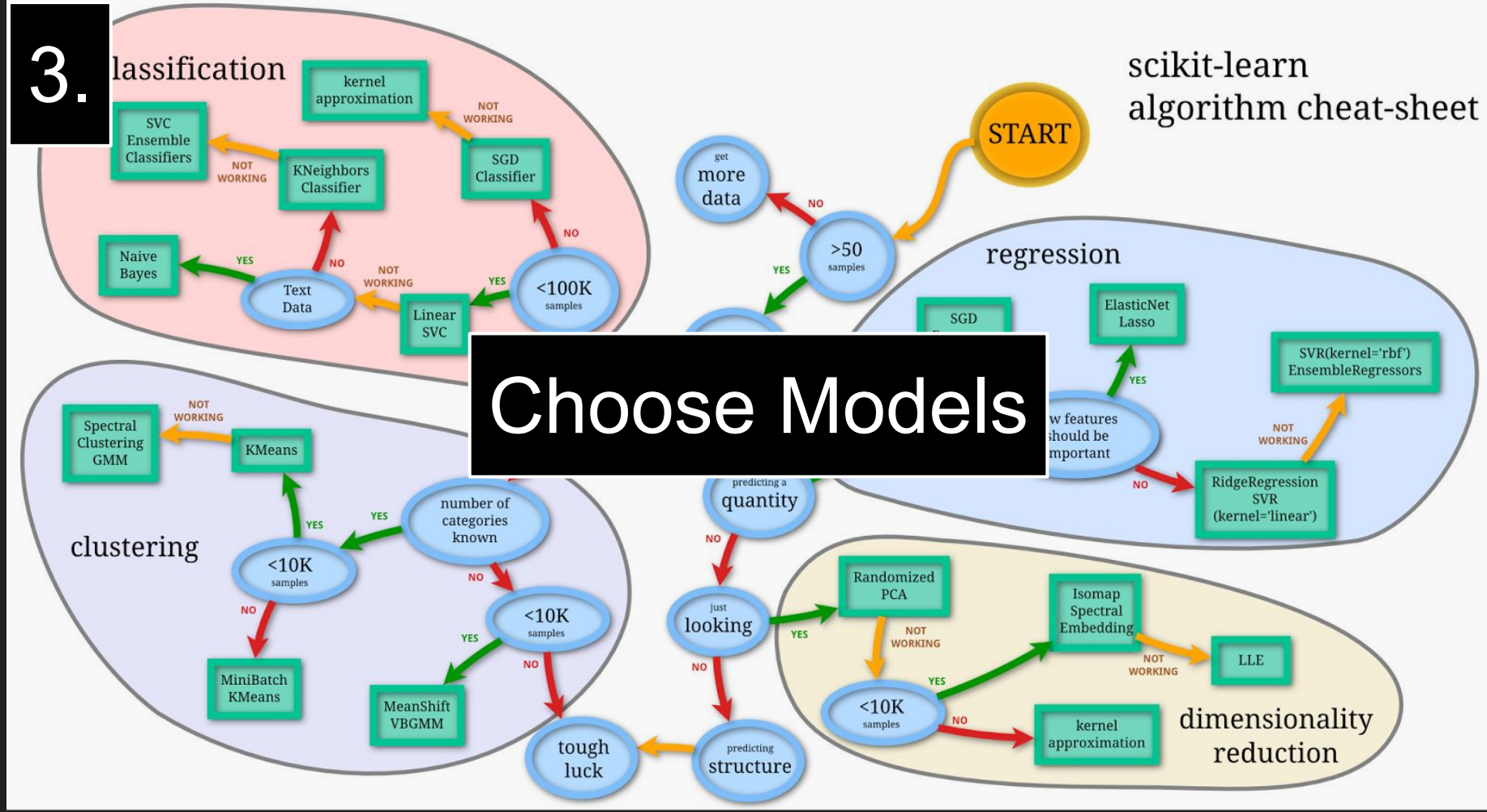
1.

Web Scraping

2.

EDA & Clean

3. Choose Models



4. Set Baseline

5.

Train Models

6.

Select and interpret best model

1.

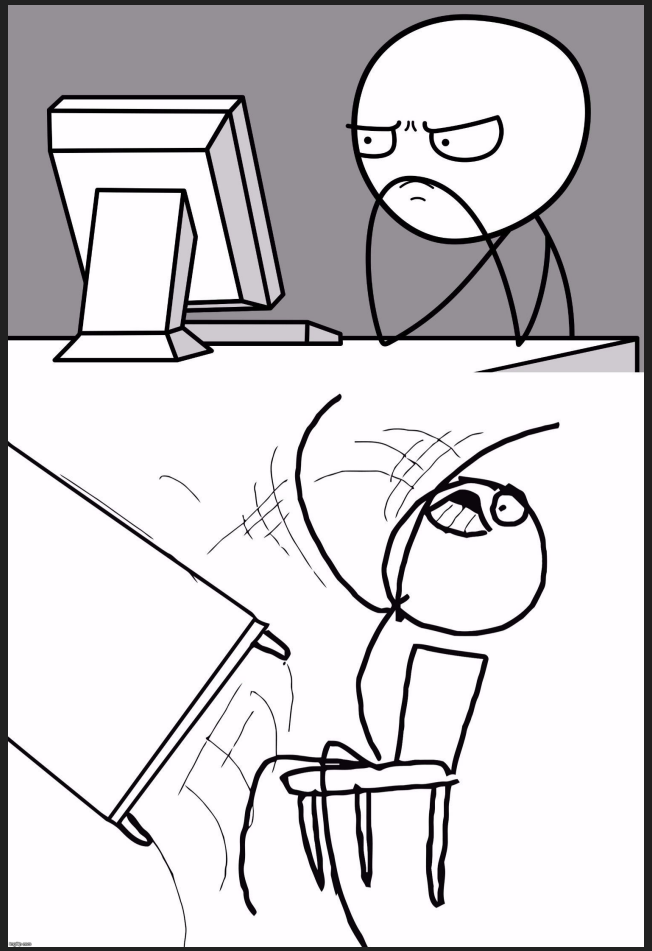
Web Scraping

- 2009 to 2018
- S&P 500 large-cap
- S&P 400 mid-cap
- S&P 600 small-cap

Dataset & Features

2 sets of data (S&P 500 and S&P 1000)
Total 25 million rows by 5 features (50 GB)

1. Security name (i.e. Apple, Google)
2. Account name (i.e. income, revenue, cash)
3. Value (i.e. numbers, text blocks)
4. Unit (i.e. USD, no. of shares)
5. Financial period (i.e. 2018Q4YTD)



2.

EDA & Clean

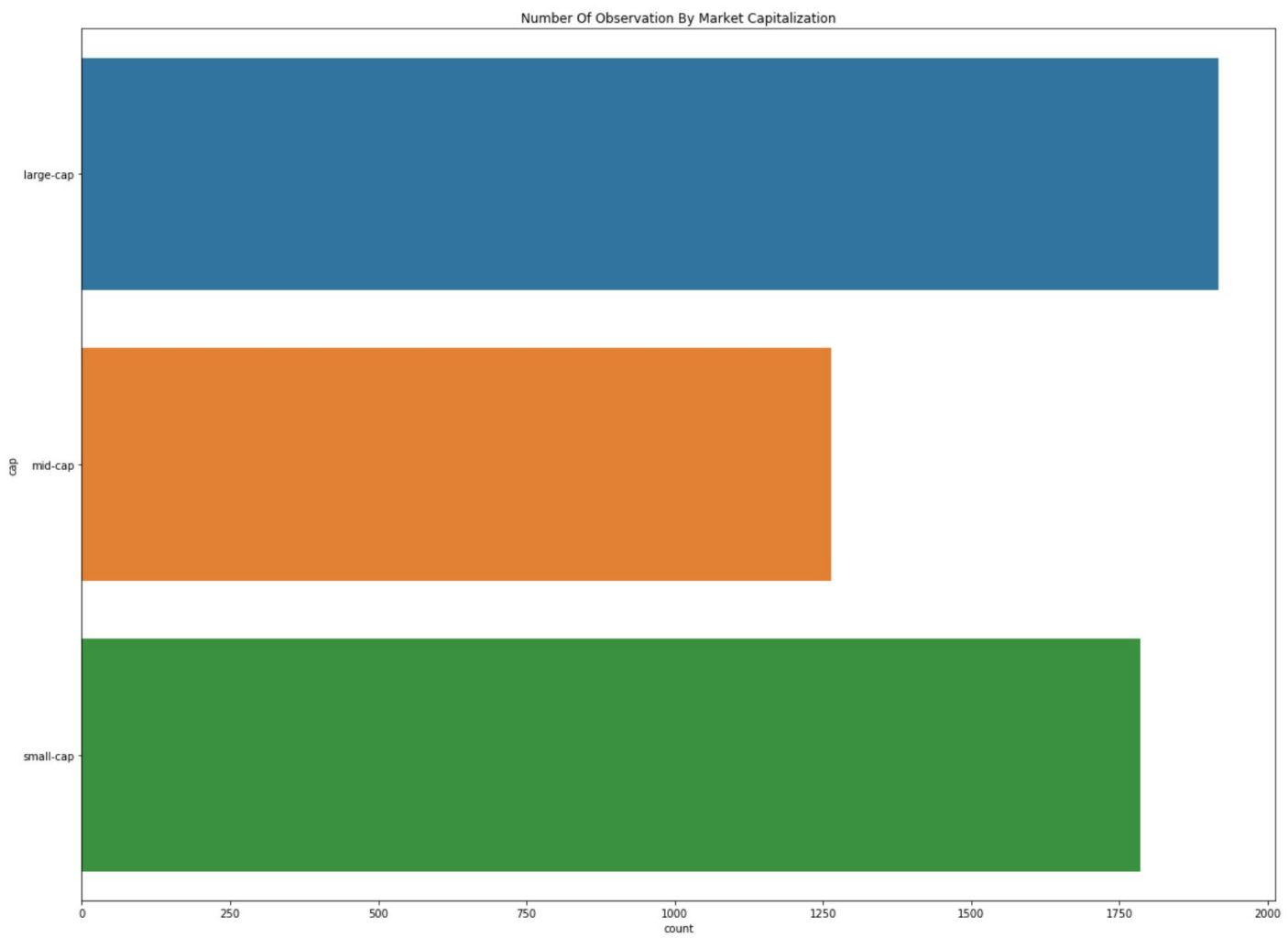
25 million rows by 5 features

- Security name (new row index)
- Account name (new feature)
- Value
- ~~Unit references~~
- Financial period (new row index)

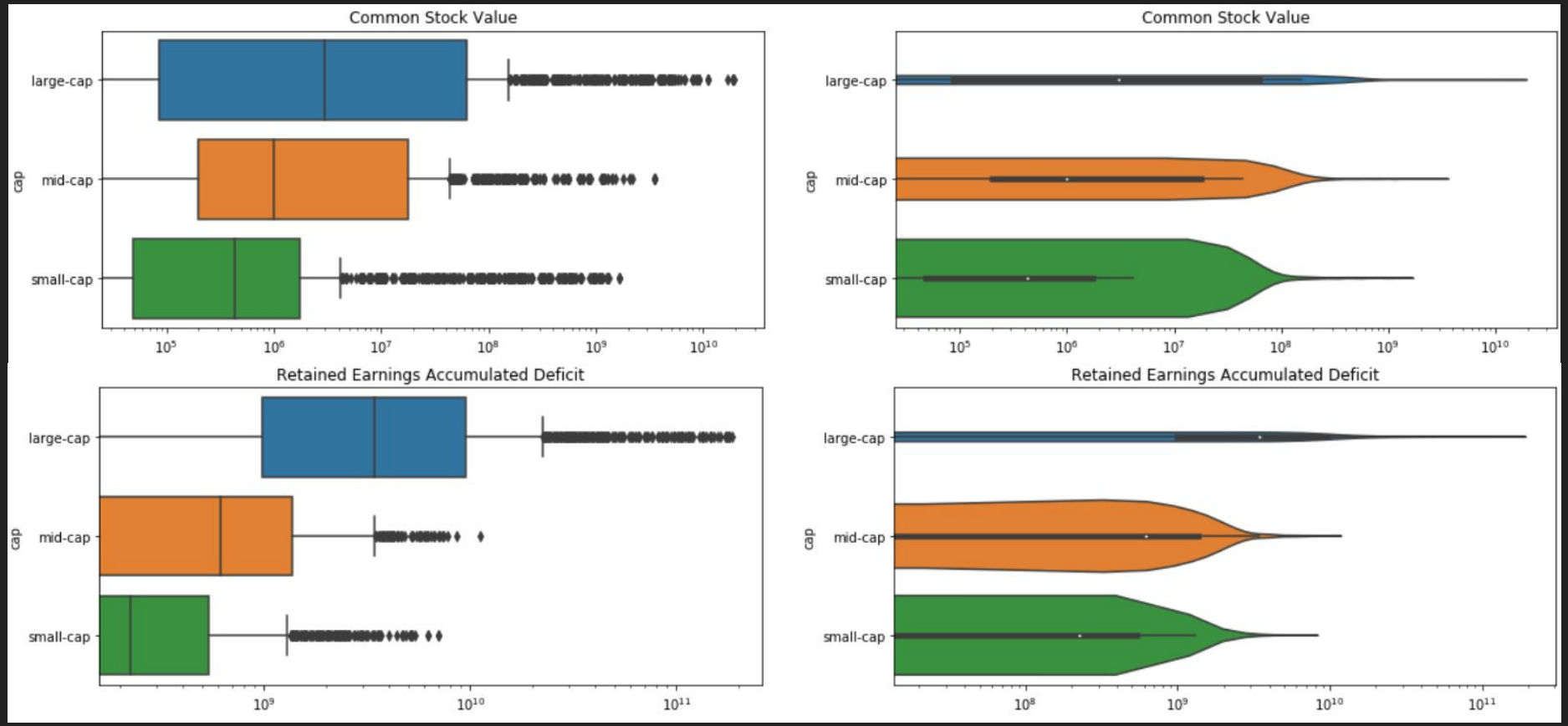
5,000* rows by ~~100,000~~ 50 features

- Cash
- Revenue
- Income

* Only Companies with December year end

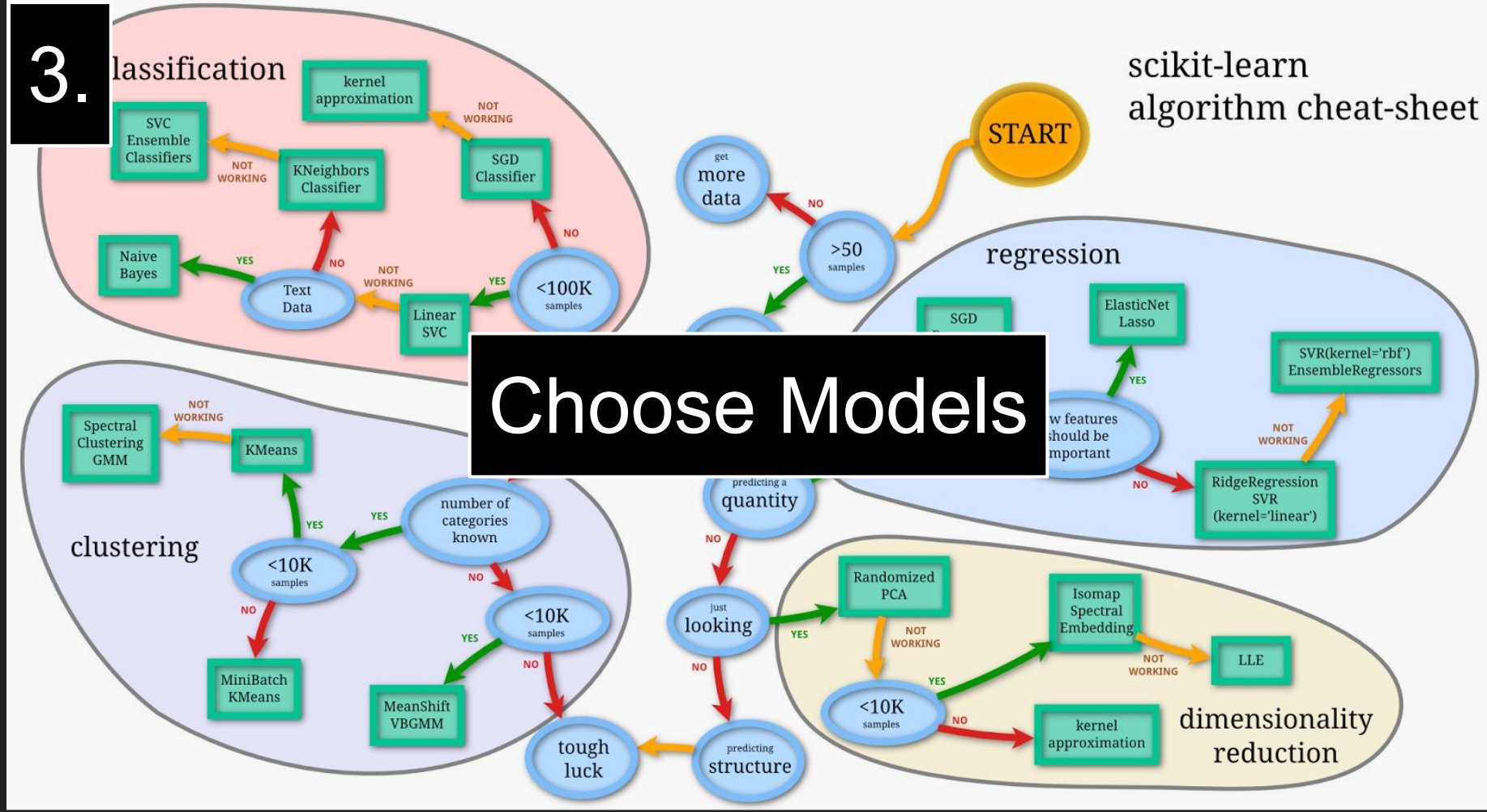

Number Of Observation By Market Capitalization
large-cap
mid-cap
small-cap
cap
0
250
500
750
1000
1250
1500
1750
2000
count

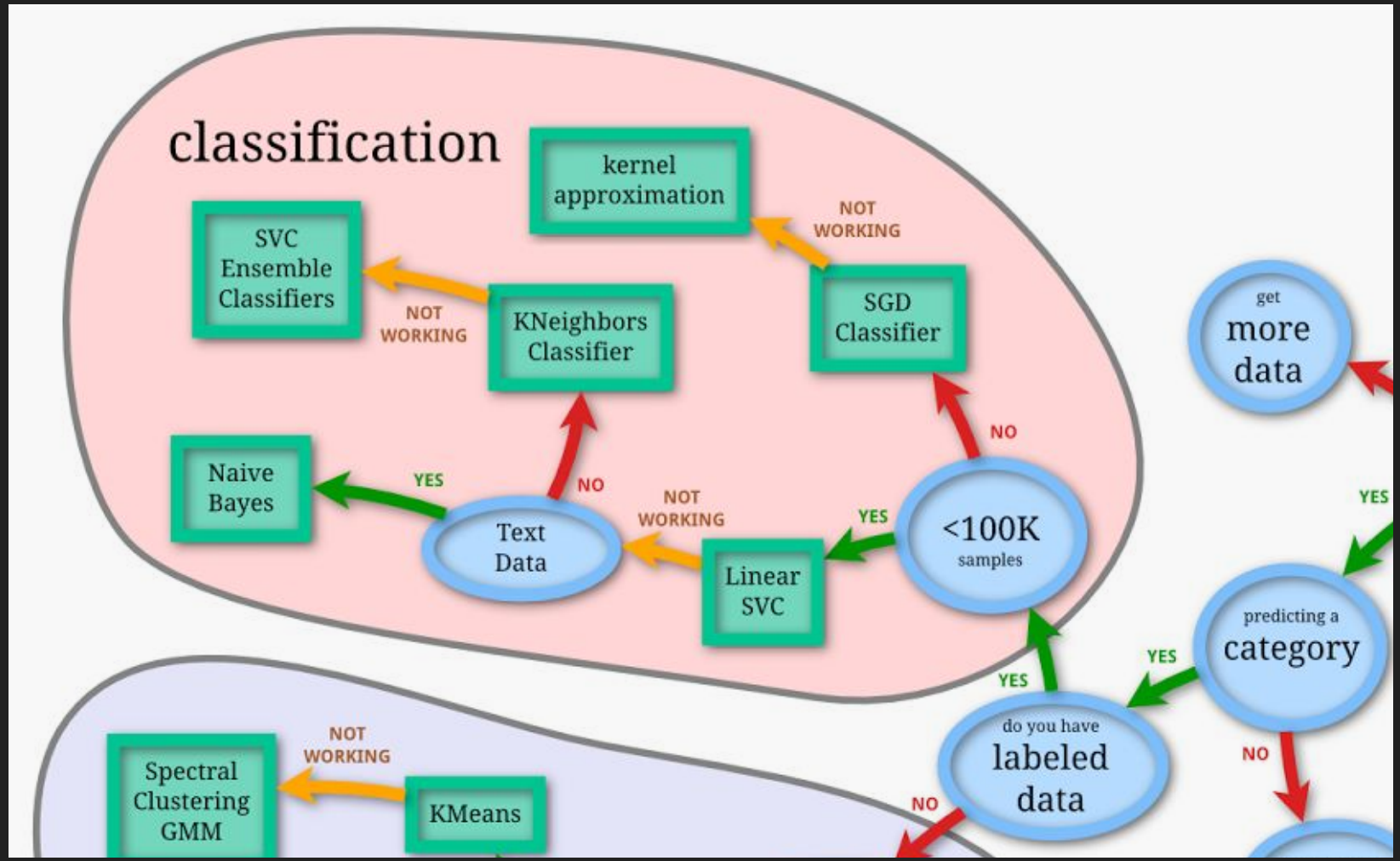
Values are correlated to market cap



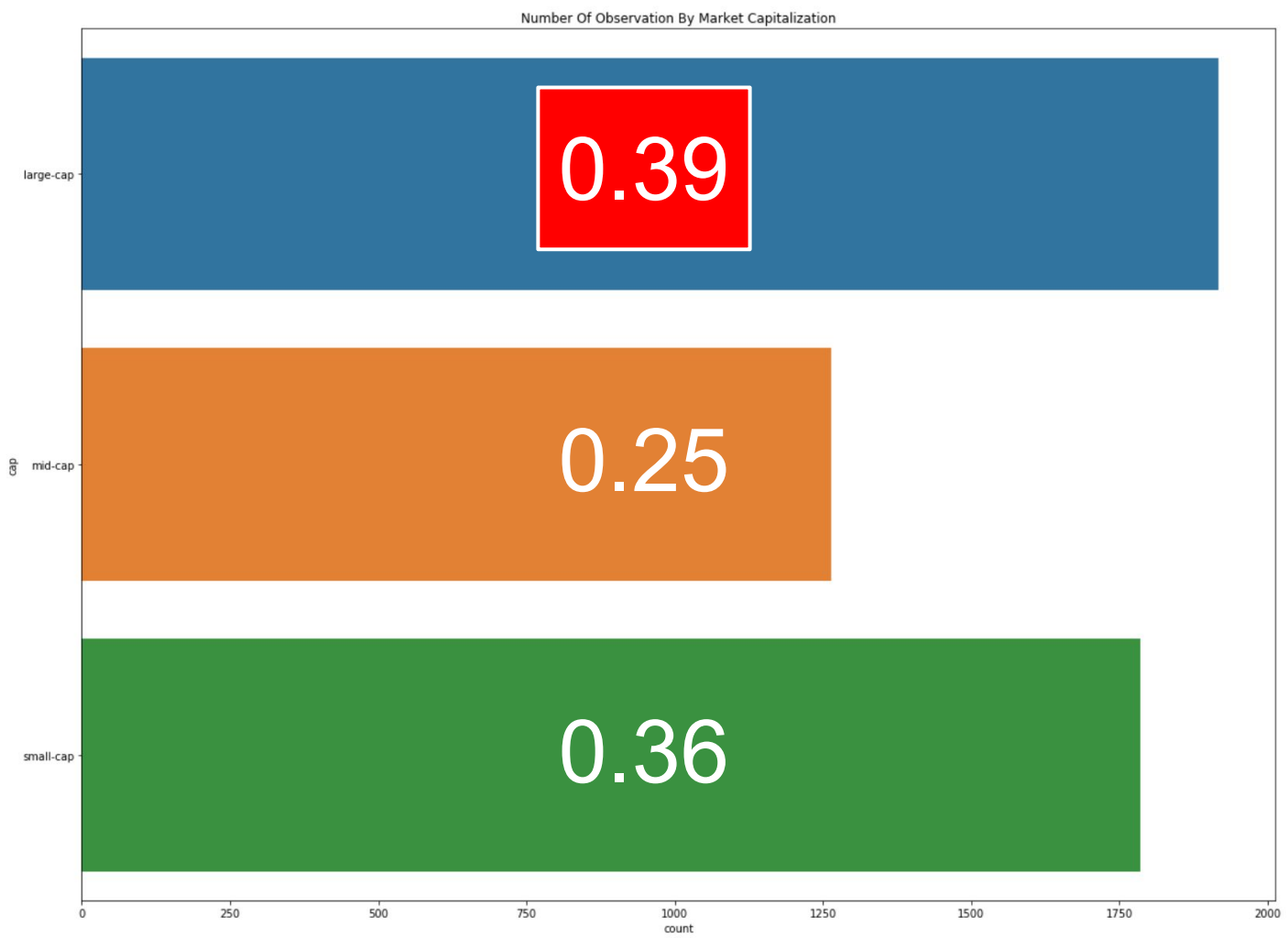
* on log scale

3. Choose Models




classification
kernel approximation
NOT WORKING
SVC Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
get more data
NO
Naive Bayes
YES
NO
NOT WORKING
YES
Text Data
Linear SVC
<100K samples
YES
predicting a category
YES
YES
do you have labeled data
NO
NOT WORKING
Spectral Clustering GMM
KMeans
NO

4.
Set Baseline
Classifier: Accuracy

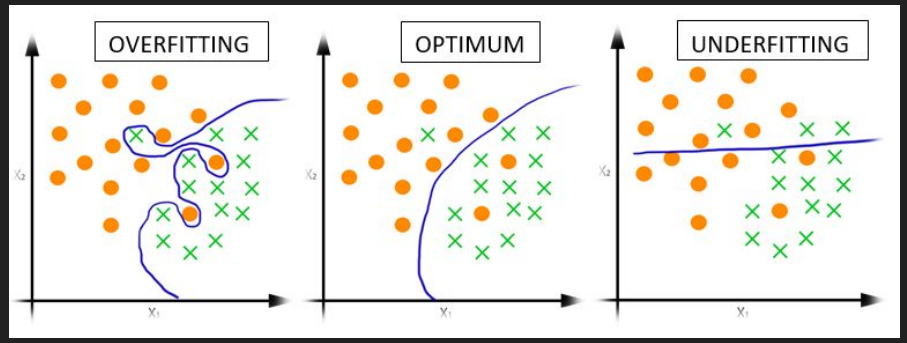

Number Of Observation By Market Capitalization
large-cap
mid-cap
small-cap
cap
0.39
0.25
0.36
0
250
500
750
1000
1250
1500
1750
2000
count

5.

Train Models

Modelling Approach

1. Convert text to numbers
2. Train test split @ 0.75, 0.25
3. Standardize features (μ = 0, σ = 1)
4. Train classification models
5. Repeat step 3 with dimensionality reduction (PCA), and train separately



Note: To reduce overfitting, steps 3 to 5 are assembled in pipelines for hyper parameter tuning with RandomizedSearchCV

Training Accuracies	Without PCA	With PCA
XGBoost	**0.95**	0.91
Gradient Boosting	0.94	0.89
VotingClassifier	0.91	0.89
K Nearest Neighbors	0.90	0.91
Bagging	0.89	0.90
Support Vector Machine	0.89	0.85
Adaptive Boosting (Adaboost)	0.88	0.91
LogisticRegression (L2 or no regularization)	0.86	0.85
LogisticRegression (Elastic Net)	0.86	0.83
LogisticRegression (L1 regularization)	0.85	0.85
Feedforward Neural Networks (Keras)	0.83	0.82
Gaussian Naive Bayes	0.82	0.73
Decision Tree	0.81	0.82
Bernoulli Naive Bayes	0.78	0.71
Random Forest	0.77	0.82
Extremely Randomized Trees (ExtraTrees)	0.75	0.77
Baseline	0.39	

6.

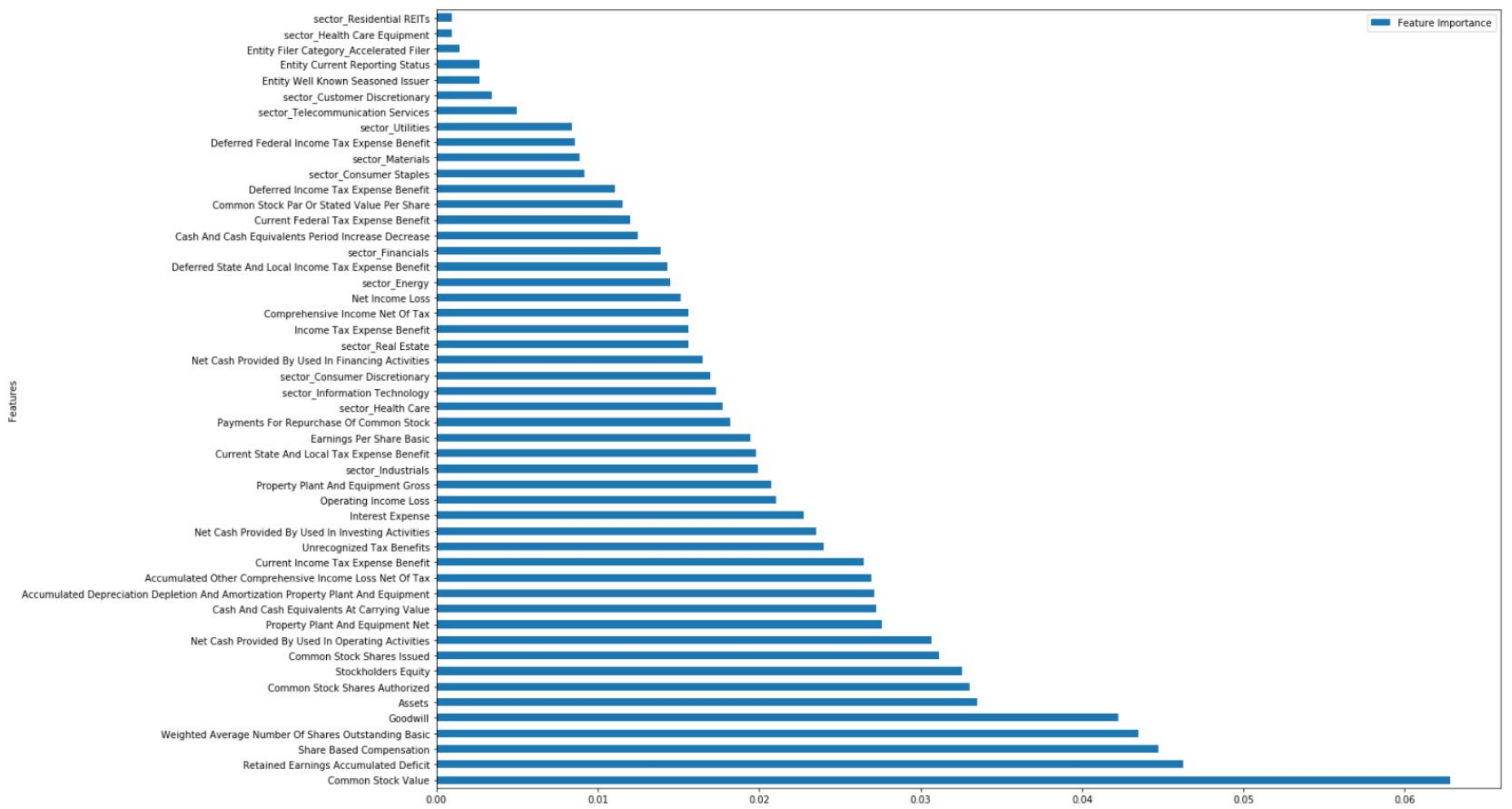
Select and interpret best model

Model Recommendation

- **XGBoost**
- Training accuracy: 0.95
- Test accuracy: 0.95

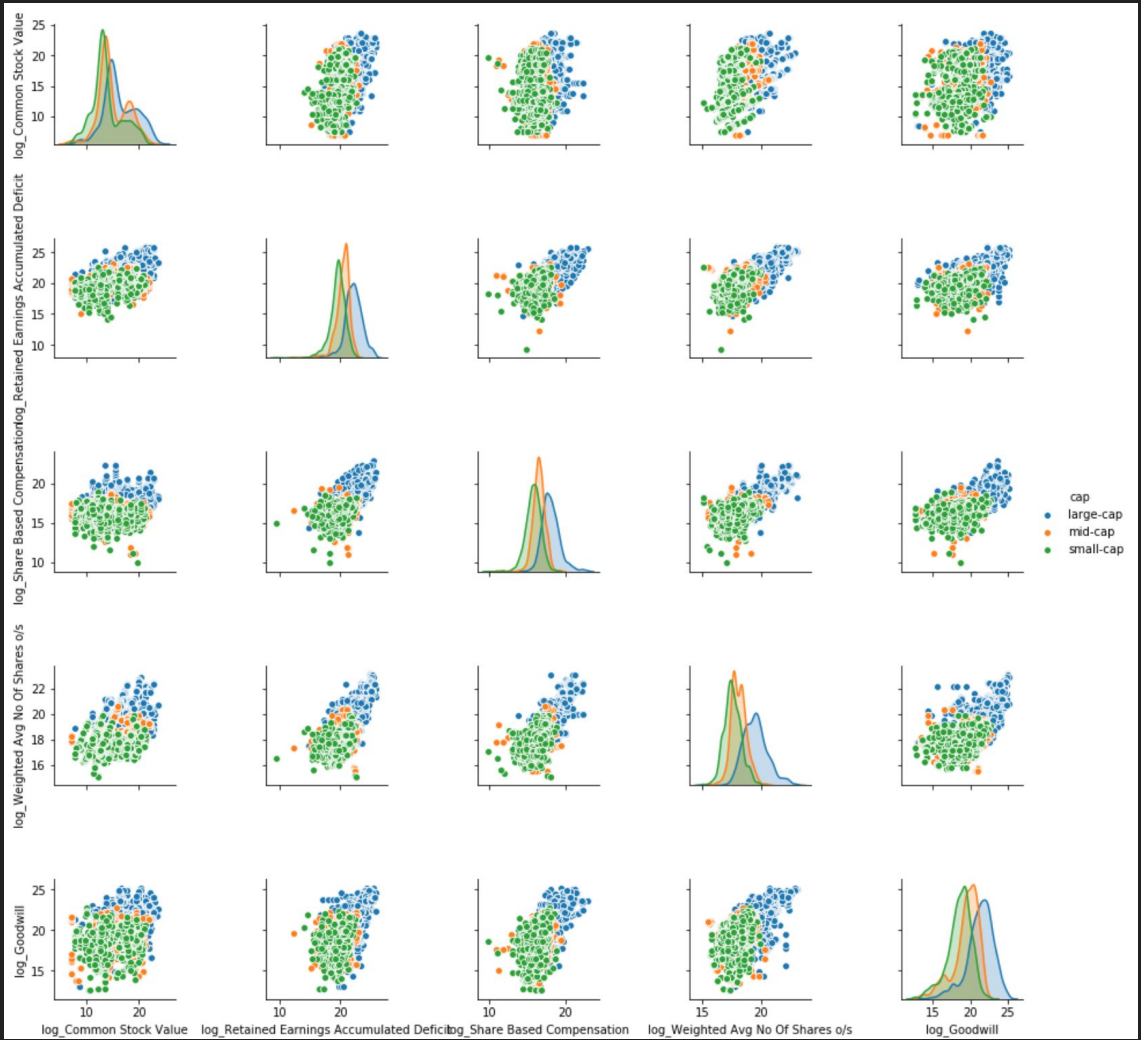
	Precision	Recall	F1 Score
Large-cap	1.00	0.99	0.99
Mid-cap	0.93	0.91	0.92
Small-cap	0.93	0.95	0.94

XGBoost Simplified


Feature Importance
Features
sector_Residential REITs
sector_Health Care Equipment
Entity Filer Category_Accelerated Filer
Entity Current Reporting Status
Entity Well Known Seasoned Issuer
sector_Customer Discretionary
sector_Telecommunication Services
sector_Utilities
Deferred Federal Income Tax Expense Benefit
sector_Materials
sector_Consumer Staples
Deferred Income Tax Expense Benefit
Common Stock Par Or Stated Value Per Share
Current Federal Tax Expense Benefit
Cash And Cash Equivalents Period Increase Decrease
sector_Financials
Deferred State And Local Income Tax Expense Benefit
sector_Energy
Net Income Loss
Comprehensive Income Net Of Tax
Income Tax Expense Benefit
sector_Real Estate
Net Cash Provided By Used In Financing Activities
sector_Consumer Discretionary
sector_Information Technology
sector_Health Care
Payments For Repurchase Of Common Stock
Earnings Per Share Basic
Current State And Local Tax Expense Benefit
sector_Industrials
Property Plant And Equipment Gross
Operating Income Loss
Interest Expense
Net Cash Provided By Used In Investing Activities
Unrecognized Tax Benefits
Current Income Tax Expense Benefit
Accumulated Other Comprehensive Income Loss Net Of Tax
Accumulated Depreciation Depletion And Amortization Property Plant And Equipment
Cash And Cash Equivalents At Carrying Value
Property Plant And Equipment Net
Net Cash Provided By Used In Operating Activities
Common Stock Shares Issued
Stockholders Equity
Common Stock Shares Authorized
Assets
Goodwill
Weighted Average Number Of Shares Outstanding Basic
Share Based Compensation
Retained Earnings Accumulated Deficit
Common Stock Value
0.00
0.01
0.02
0.03
0.04
0.05
0.06

All Observations

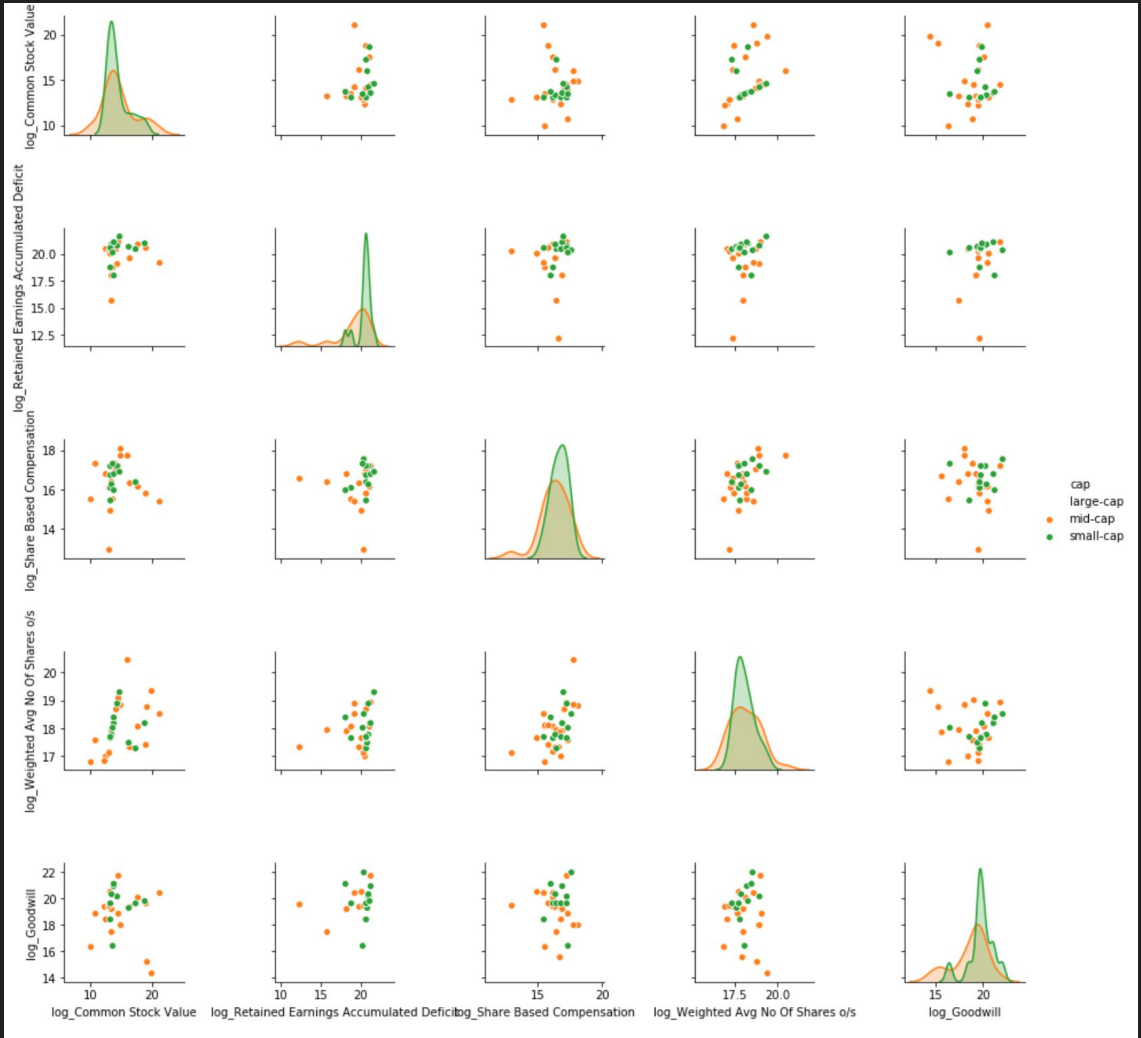
* on log scale



* Actual labels, not predictions

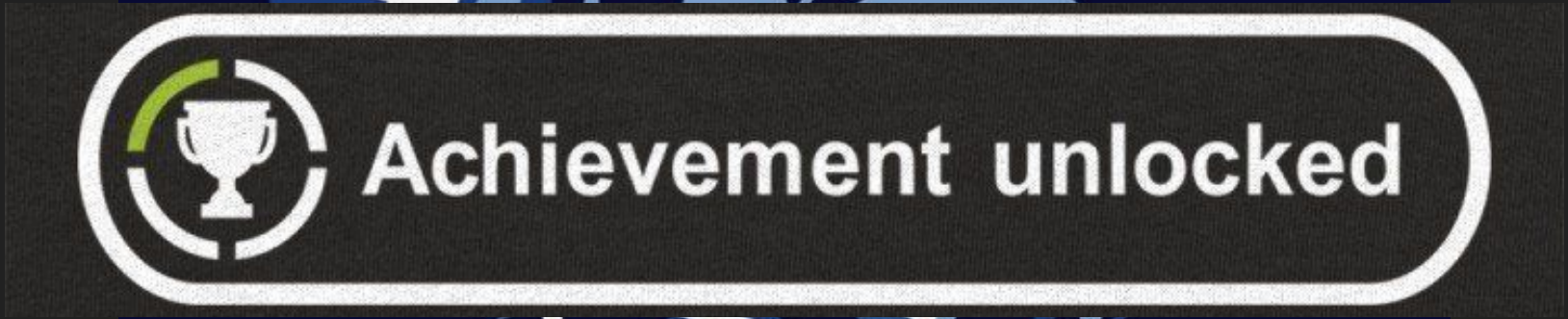
Wrong Predictions

* on log scale



* Actual labels, not predictions

Goal



Classifier that categorises unicorns into large-cap, mid-cap and small cap

VS

Differentiate Unicorn Size
Large-cap
Mid-cap
Small-cap

LEVEL UP

- Evaluate the classifier using unicorn’s actual financials
- Include non-December companies
- Expand data collection

https://github.com/junkaitoh/capstone_project