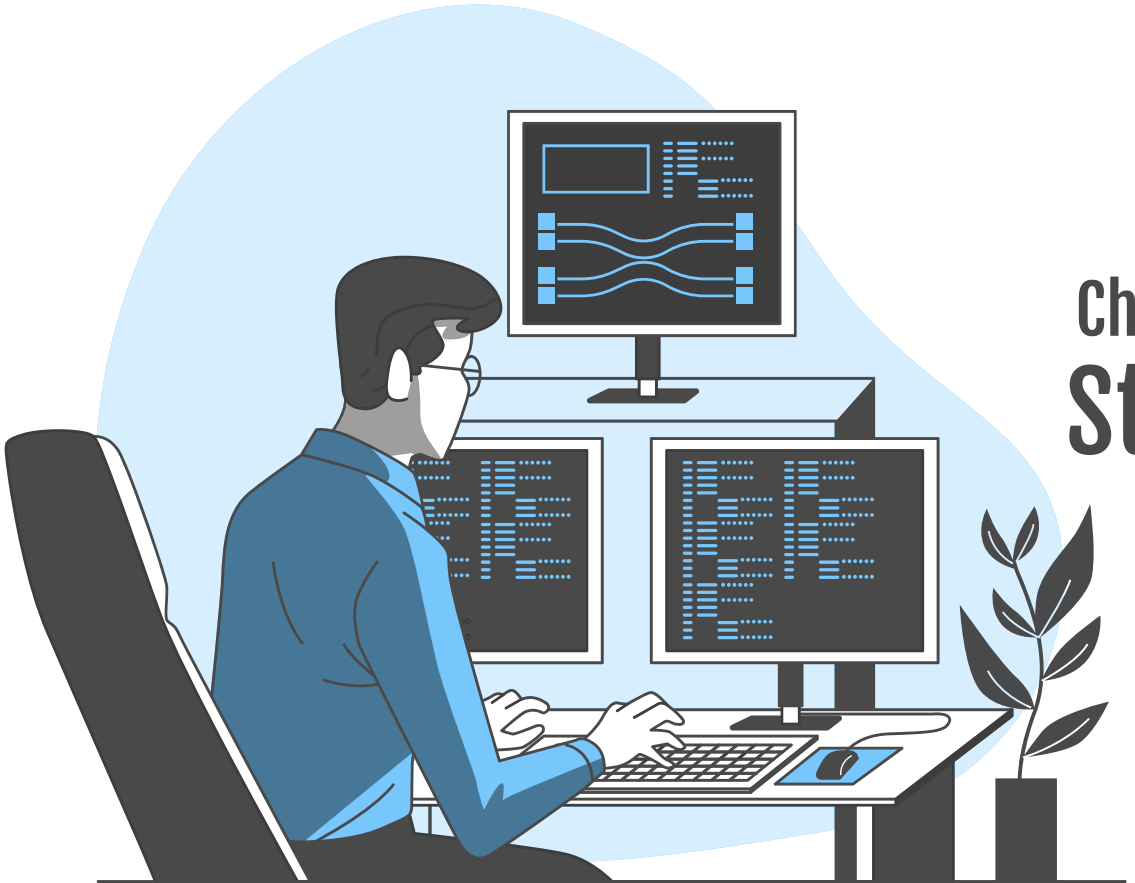


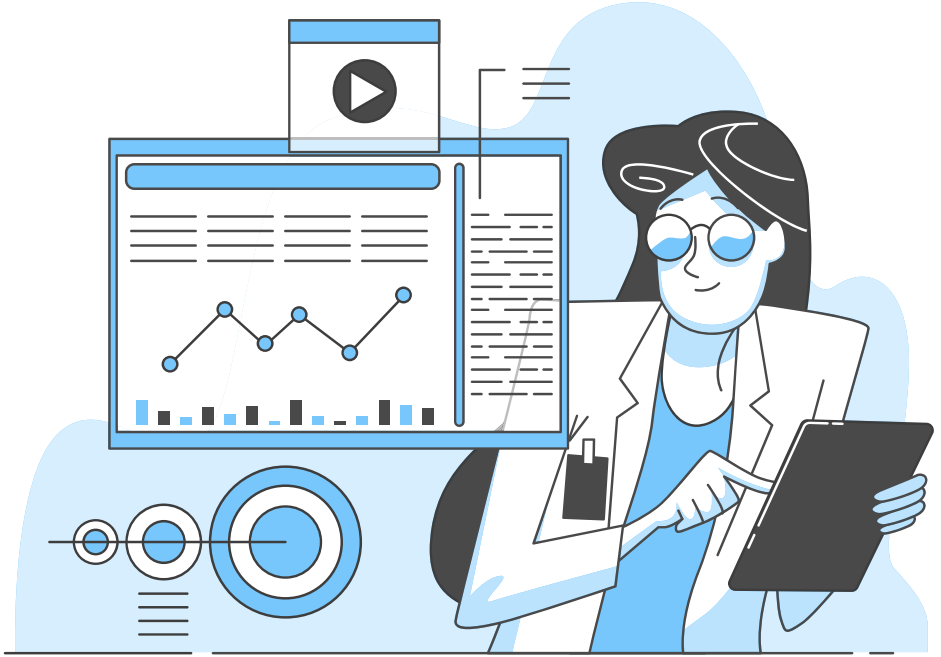
Choosing Outperforming Stocks with Machine Learning

Logan Chalifour

12/21/2021

Agenda

1. Introduction
2. Data Cleaning and EDA
3. Unsupervised Learning
4. Supervised Learning
5. Conclusion



1
Introduction

Introduction

- **Stock picking is hard!**
  - Over the past year, the S&P 500 has returned ~26% to investors
  - Choosing stocks to outperform the rest of the market is a difficult process
- **Past**
  - Investors would make stock predictions through a bottom up approach
  - Financial ratios were analyzed manually to make judgements about a company
  - Equity research, high fees, lots of bias, and room for mistakes
- **Future**
  - Automated trading, high frequency trading, and rob-advisors
  - Machine learning has allowed us to analyze large amounts of data in little time
  - Smarter decisions with your money
- **Question:**
  - Can we construct a supervised learning algorithm that predicts outperforming stocks using financial ratios?
  - If so, what financial ratios are most influential in choosing stocks that outperform the greater market?

2
Data Cleaning & EDA

Our Data Sources

01

Wikipedia

General identifying information about S&P 500 companies

02

FactSet

Financial ratios commonly discussed when valuing a stock

Data Cleaning

Merge Data

Combine data sources on the ticker symbol

...

Data Types

Change data types to the appropriate type for modeling

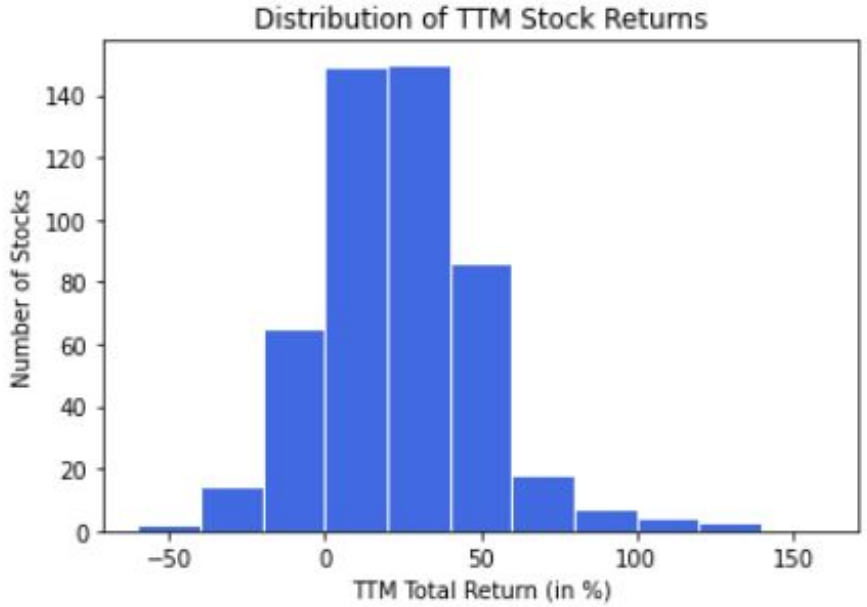
...

Null Values

Find null values and fill with column averages

...

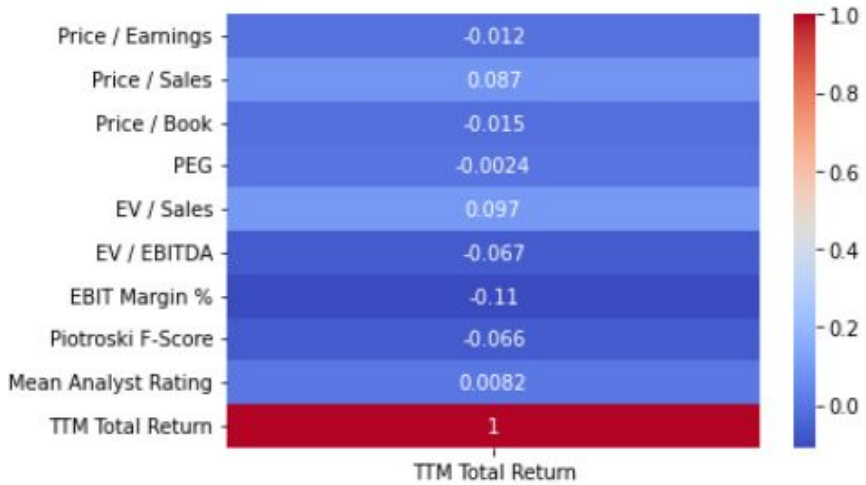
Target Variable of Interest



NOTES:

- Unimodal
- Centered around ~20%
- Few outliers at high and low end

Features Heatmap

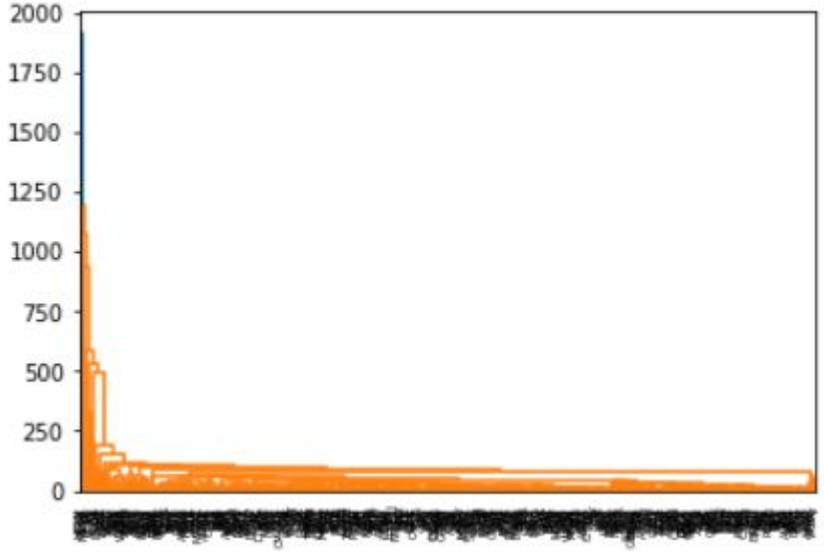


NOTES:

- Mostly minimal correlations
- Shows how hard stock picking is
- Even a slight given by these variables will be helpful to us

3
Unsupervised Learning

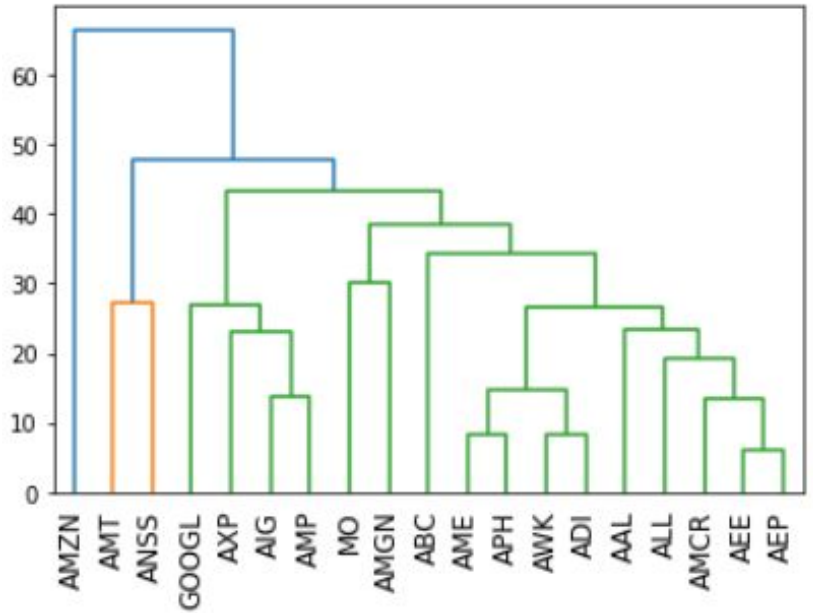
Hierarchical Clustering



NOTES:

- Too many stocks to picture visibly on one
- Let's construct a smaller one

Condensed Dendrogram



NOTES:

- Much clearer dendrogram for 20 stocks
- Three distinct clusters grouped by likeness
- Similar financial profiles due to their ratios

4

Supervised Learning

Preping Our Model

Binary Target

Outperform index (1) or underperform index (0) over TTM

...

Split X and Y Sets

Separate feature variables from the target variable

...

Test-Train Split

80-20 split for training and testing data respectively

...

S&P 500 Stock Performances

Outperform

> 26%

225

Underperform

< 26%

274

*One stock dropped for missing values

Our Models

BEST MODEL!

01

Random Forest

Accuracy:

.63

02

Gaussian NB

Accuracy:

.57

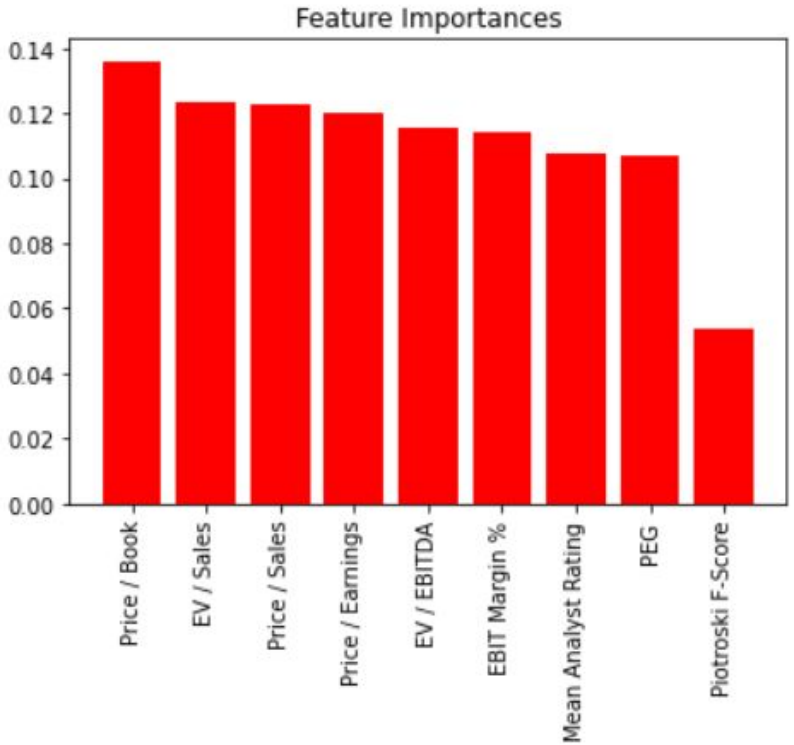
03

SVC

Accuracy:

.54

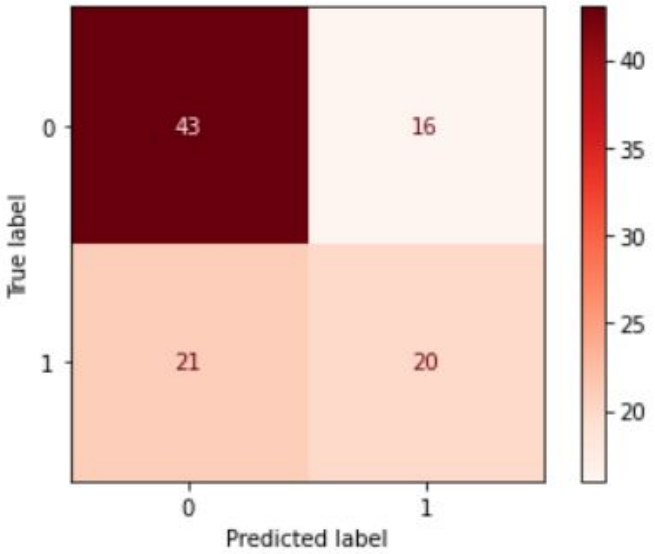
Feature Importances



NOTES:

- Price / Book, EV / Sales, and Price / Sales were most important
- Piotroski F-Score was clearly the least important feature

Confusion Matrix and Evaluation



Sensitivity: 0.49

Specificity: 0.73

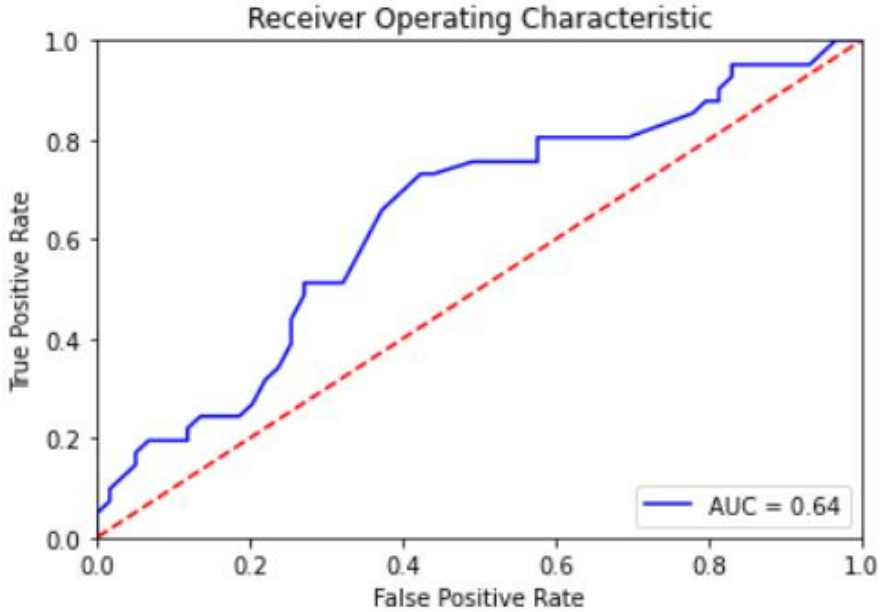
Precision: 0.56

F1 Score: 0.52

NOTES:

- Model excels at identifying underperforming stocks with a specificity of .73
- Worse job predicting outperforming stocks with a sensitivity of .49

Receiver Operating Curve



NOTES:

- Happy to see TP rate consistently above the null model at each respective FP rate
- AUC of .64 is traditionally not that helpful, however, in this case, it is because of the nature of the problem we are trying to solve

5

Conclusion

Conclusion

- **Random forest successful**
  - Very useful in avoiding underperforming stocks
  - Hitting on nearly half of overperforming stocks
- **Favorable evaluation metrics given the problem**
  - .63 accuracy well above null model
  - .73 specificity
  - .49 sensitivity
- **Financial metrics impacting stock performance**
  - Price / Book, EV / Sales, and Price / Sales were the most important features
  - Piotroski F-Score was the least important feature
- **Future implementation**
  - Gaussian Naive Bayes and SVCs also showed promise, but not as much as the random forest model
  - Further research in deep learning through the use of multilayer perceptrons may uncover even better-performing models