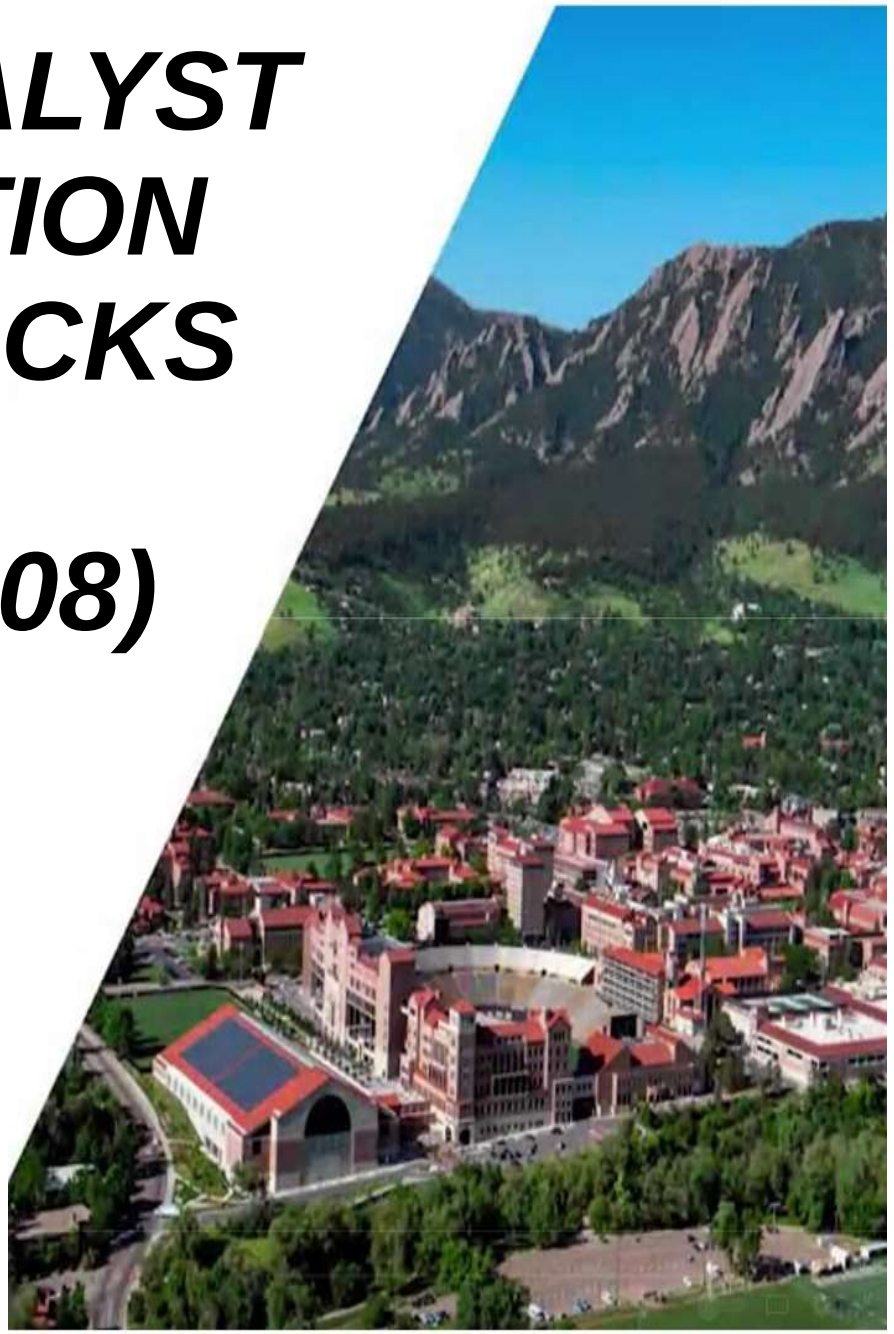
PREDICTING ANALYST RECOMMENDATION FOR S&P500 STOCKS

by XX (2025-05-08)



Master of Science in Data Science
UNIVERSITY OF COLORADO BOULDER

Table of contents

- Introduction and Key Objectives:
  - Problem Statement
  - Data Set
  - Tools & Techniques
- Data Source, Directories and Code Structure:
  - Description
- EDA, Cleaning, and Pre-Processing:
  - Tools & Techniques
- Models & Analysis:
  - Dimensionality Reduction through PCA
  - Predict Clusters with K-Means (Unsupervised)
  - Predict Clusters with Logistic Regression (Supervised)
- Conlussions and Final Comments
- Future Work
- References

Caution: Don’t Try This at Home

- Investments involve risk. You may loose some or all your money.
- This is an academic paper. It is not investment advise.

Introduction: Problem Statement

- Manage their financial portfolios with financial data.
- Financial Analysts issue recommendations, to "buy", "hold", or sell stocks.
- Can we derive a recommendation from the financial data?
- Gather financial data, clean it, PCA and then fit models to it.
- **Main objective: predict the mean analysts recommendation.**
- Data source: Yahoo Finance. Metrics, enginered features.
- Fit a K-Means Unsupervised Model on training and testing sets.
- Labels available (for Analyst Recommendations): fit a Supervised Logistic Regression Model, measure scores on both models and compare them.

Data Set: (yahoo.f)

sector	fullTimeEmployees	auditRisk	boardRisk	compensationRisk	shareHolderRightsRisk	overallRisk	priceHint	previousClose	open
Industrials	61500	1	7	6	4	4	2	142.08	140.62

- Main data source Yahoo Finance (YF)
- 179 financial features
- 500 constituents of the S&P500.
- Secondary data source: list of S&P500 tickers Wikipedia

- Code divided in two:
  - first section dedicated to sourcing data, done once.
  - second section dedicated to cleaning, EDA, and modelling
  - Each section is saved in its own .py file

EDA, Cleaning, Pre-processing

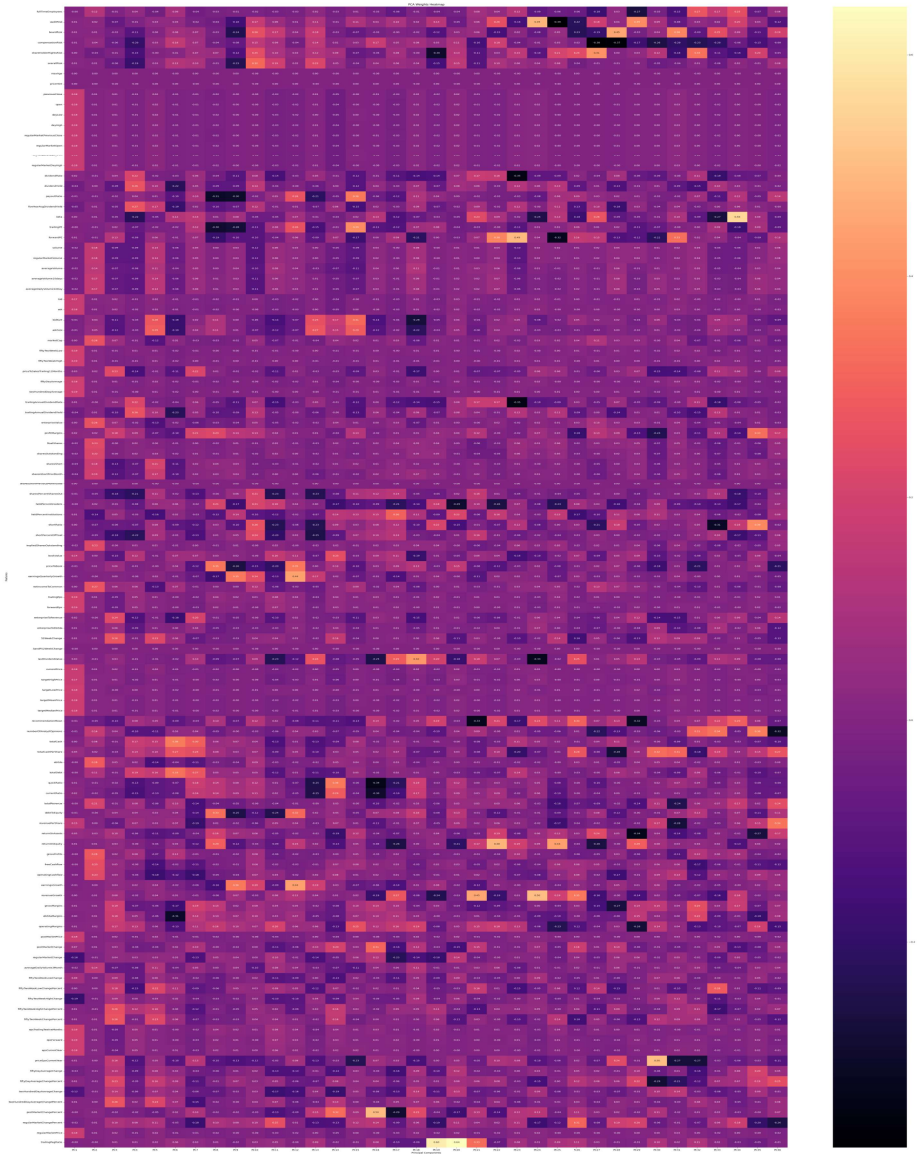
- Each stock has 179 features, with each feature describing a piece of information about the financials of the listing corporation.
- Condense features in a smaller number through PCA.
- The first reduction is in data type. 179 goes to 117.
- PCA needs numeric data: keep integers/ float data types.
- Compile all 500 stocks into one single data frame, and save the file for further cleaning and pre-processing.
- Not all stocks report the same information, and for many of them, the set is incomplete. 3M is a template.
- Cleaning is extensive. See details in Jupyter N and Code.
- Final after cleaning: 465 (from 502) / 111 features (from 179)
- no NaN or missing data, ready to pass it on to PCA algorithm.

PCA

- Scaling, splitting training and testing sets
- Create PCA object (cum explained var 95%).
- 111 to 36 Principal components.
- Principal Components hard to interprete.
- PCs are newly engineered variables
- No direct intuitive interpretation
- Linear combinations of the original features.
- Power of PCA: explain majority of original variance with the least amount of components.

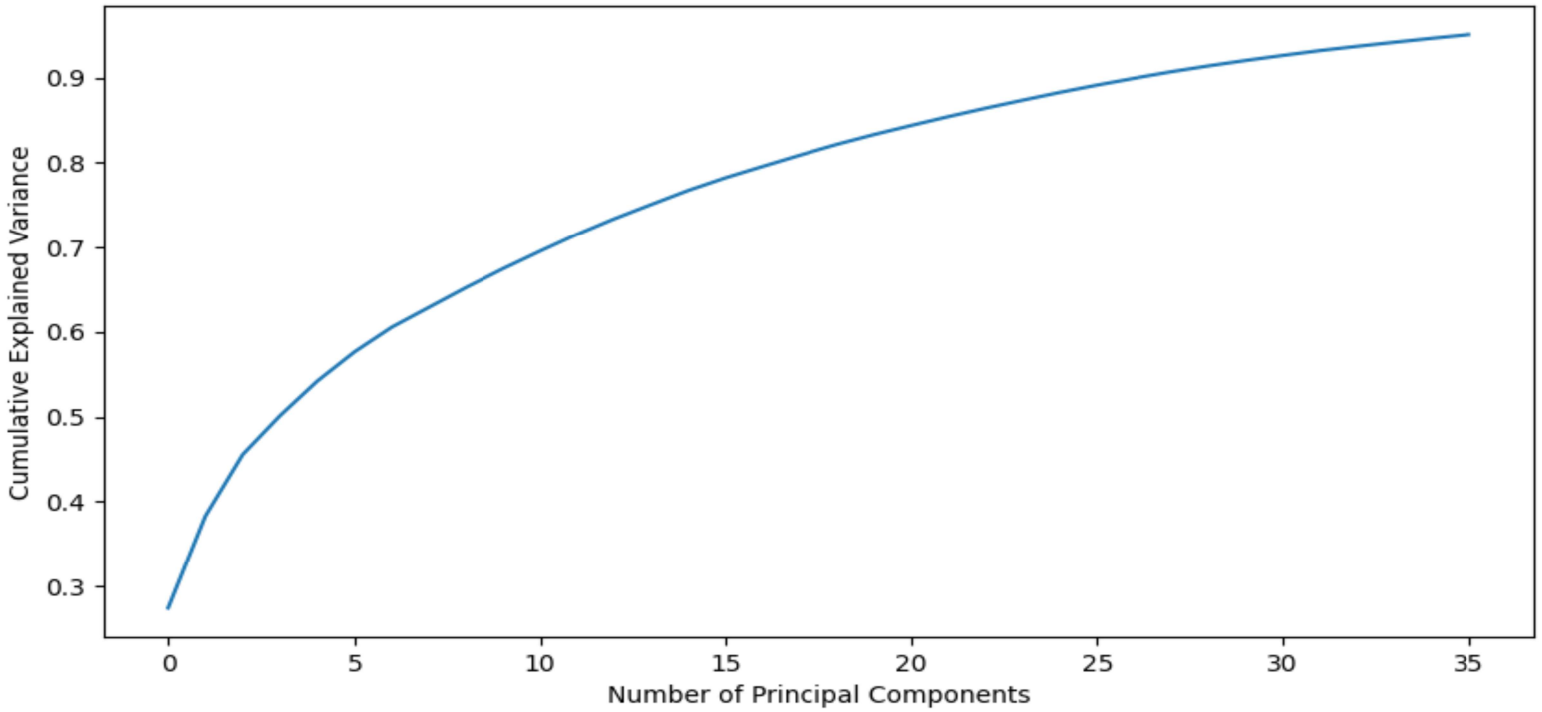
PCA (contd.)

- weight of each feature is the contribution of that feature to each principal component.
- y = features
- x = PC
- (x, y) = weight



PCA contd. Variance Explained

- First component largest variance explained.
- The proportion of variance each component explains represents the amount of information from the original data that is captured in a given component.



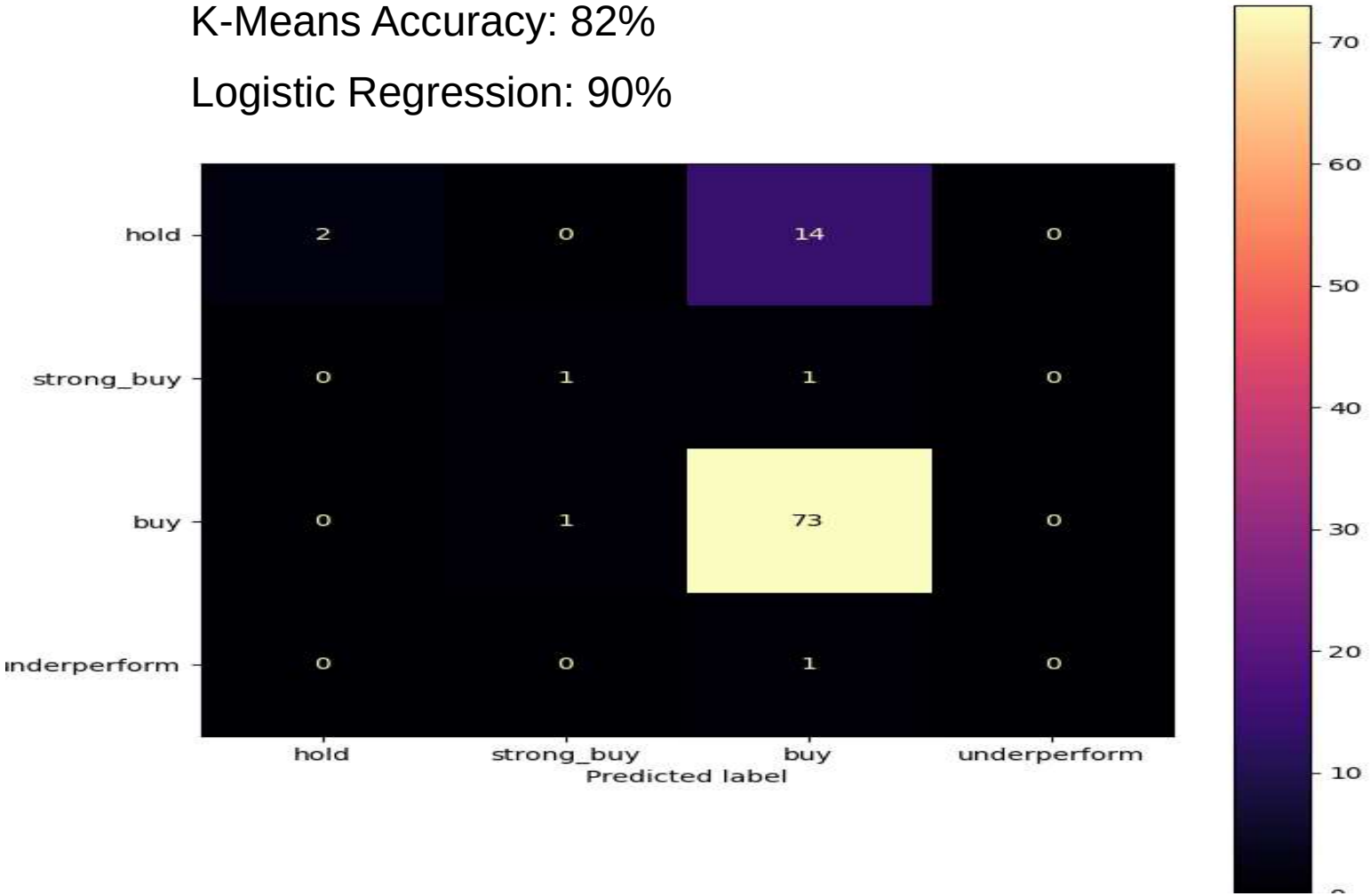
Fit K-Means (Unsupervised Model)

- Split the data: training and testing sets.
- Fit K-Means model, tune hyperparameters, measure accuracy.
-
- Compare results to supervised Logistic Regression.
-
- Not use the labels first. Fit K-Means to discover the 4 clusters [buy, strong buy, hold, underperform].
-
- Measure accuracy once the model is built, with labels.

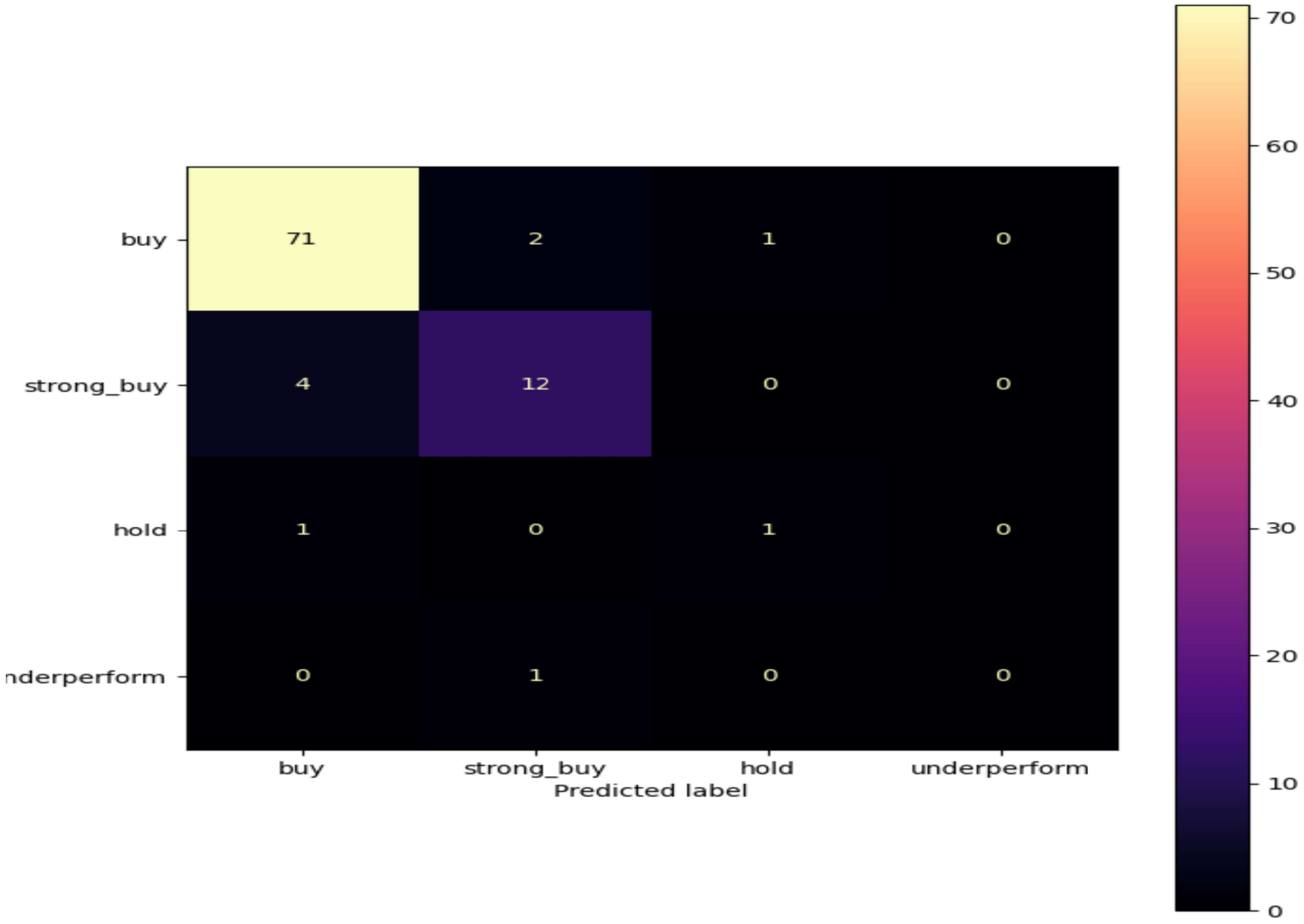
Results: Confusion M K-Means

K-Means Accuracy: 82%

Logistic Regression: 90%



Results: Confusion Logistic R



Conclussions and Final Comments

- K-Means finds the four clusters with 82% accuracy.
- The supervised LR method does even better, with almost 90% accuracy.
- Dimensionality reduction through PCA was an important step to be able to get the algorithms to deliver results in reasonable time.
- **Future Developments:**
  - SP500 10% of the total of 6000 securities in the US; small fraction of all securities worldwide.
  - Future extensions of this work would include all US stocks first with international markets added progressively.
-

Thank you !