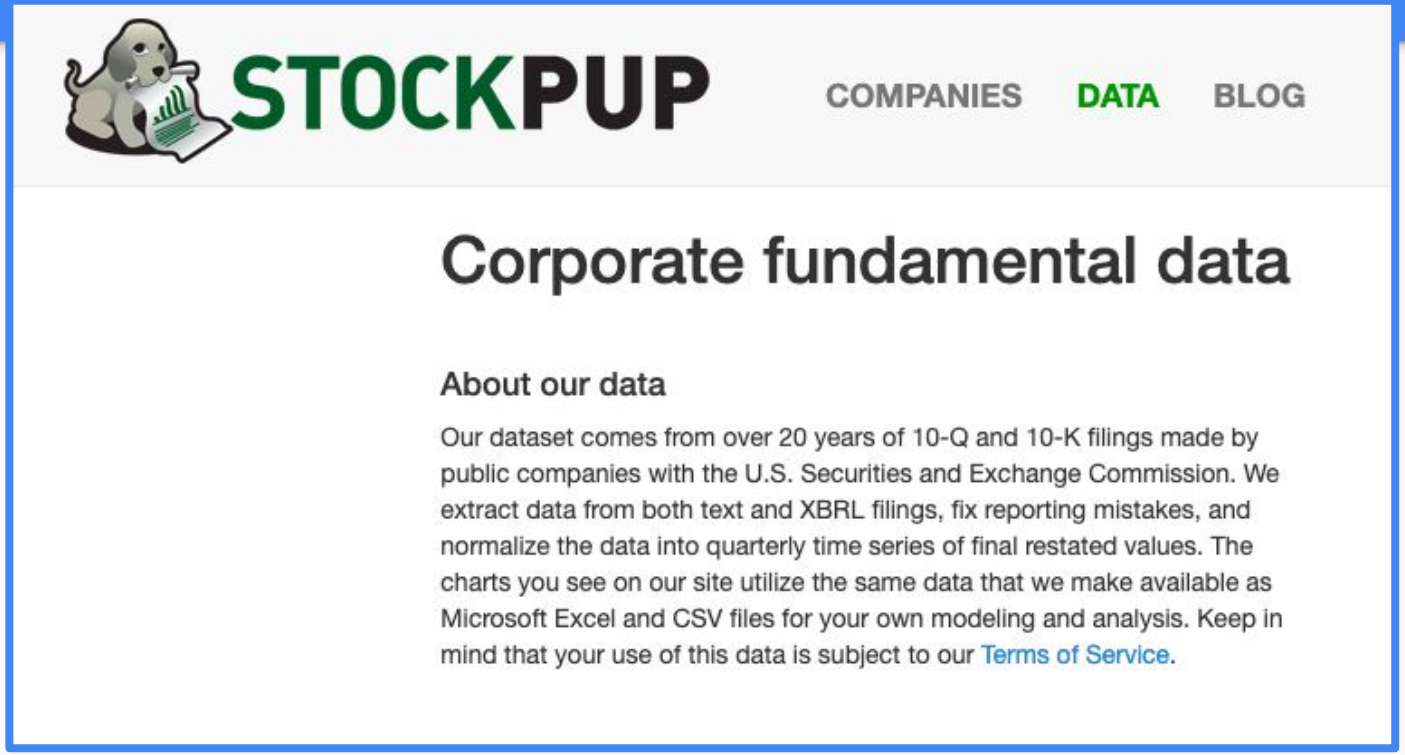
Stocks and Fundamental Data

Predicting if a stock will go up 5% or more in the next quarter

The Data


STOCKPUP
COMPANIES
DATA
BLOG
Corporate fundamental data
About our data
Our dataset comes from over 20 years of 10-Q and 10-K filings made by public companies with the U.S. Securities and Exchange Commission. We extract data from both text and XBRL filings, fix reporting mistakes, and normalize the data into quarterly time series of final restated values. The charts you see on our site utilize the same data that we make available as Microsoft Excel and CSV files for your own modeling and analysis. Keep in mind that your use of this data is subject to our Terms of Service.

The Data

- 756 Stocks with respective fundamental data
- Cleaned Data:
  - Filled NaNs with Backfill, then removed any further NaNs
  - Left with <2,000 rows

Altering the Data

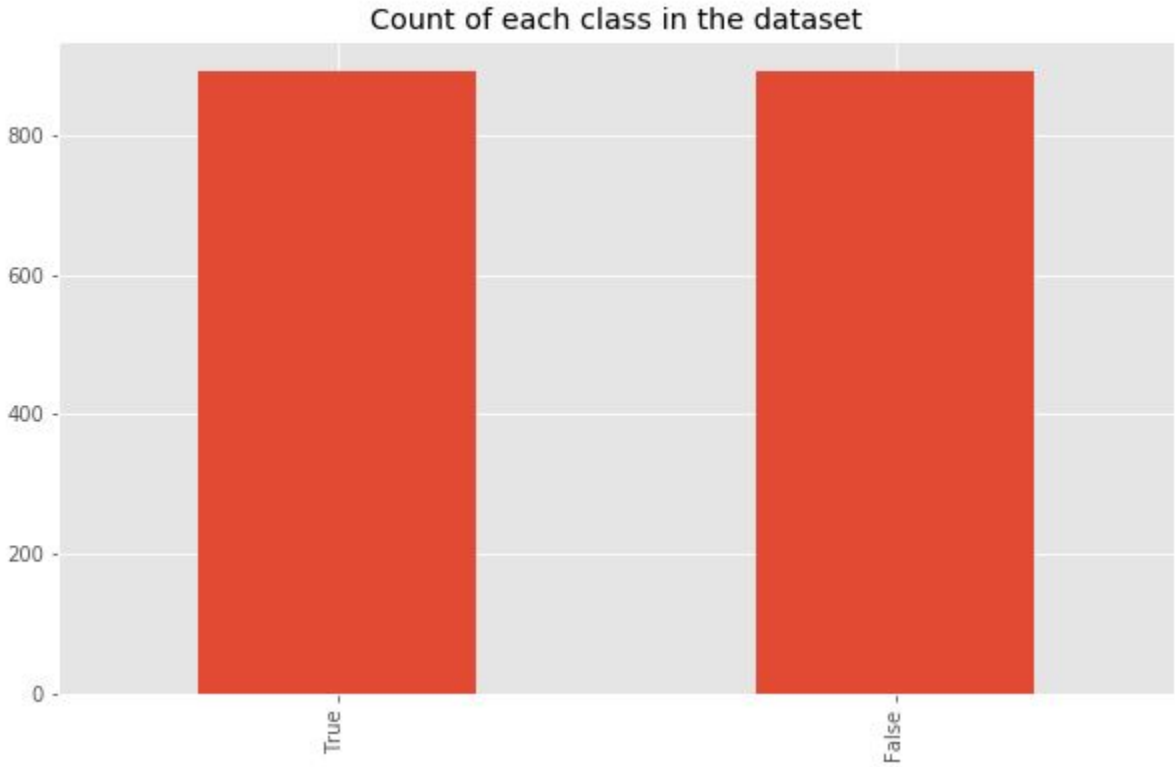
(Feature Engineering)

- New columns derived from original features:
  - Values were percent improvement from past quarter to current quarter.
- Classification column (True or False):
  - True if future quarter price increased > 5% compared to current quarter
  - False if otherwise
- Balanced the Classes:
  - Removed any excess classes of True or False

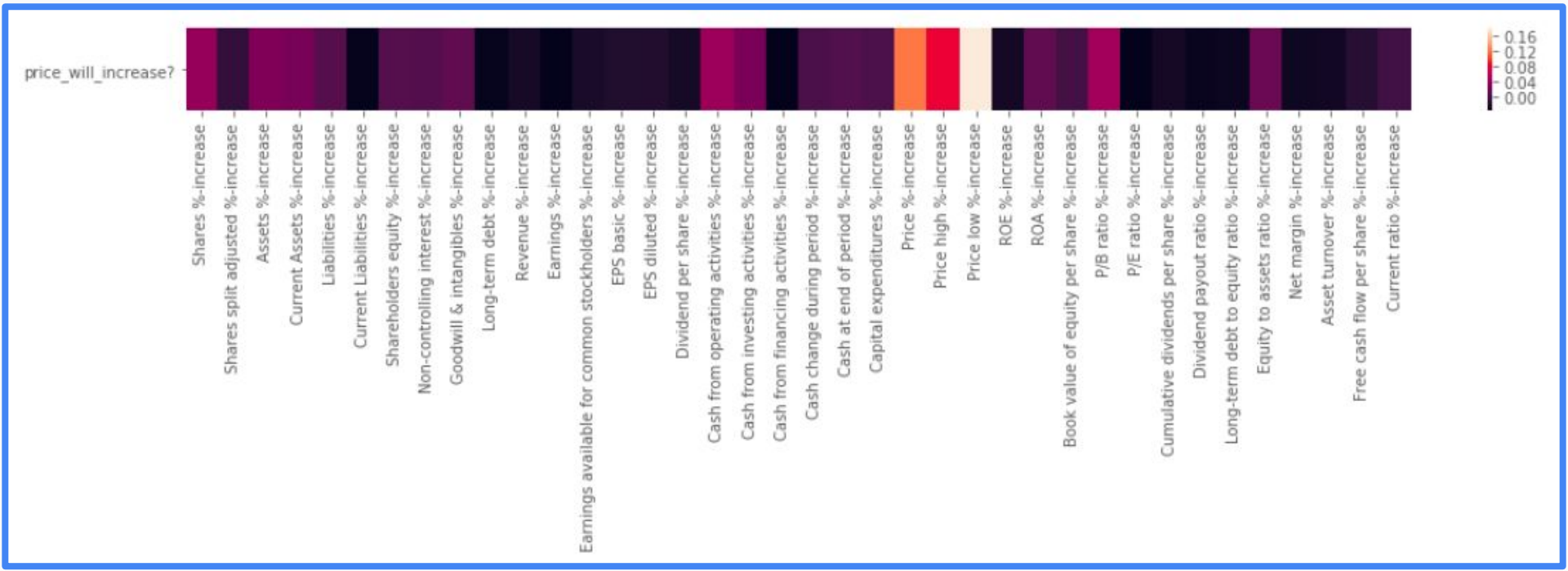
The Final DataSet Preview

	price_will_increase?	Shares %-increase	Shares split adjusted %-increase	Assets %-increase	Current Assets %-increase	Liabilities %-increase	Current Liabilities %-increase	Shareholders equity %-increase	Non-controlling interest %-increase	Goodwill & intangibles %-increase	Long-term debt %-increase	Revenue %-increase	Earnings %-increase
0	True	0.39	0.39	1.08	-7.24	1.13	-10.98	0.97	0.00	0.0	4.61	9.65	179.71
7	True	0.00	0.00	1.79	8.34	1.98	0.11	1.33	0.00	0.0	3.40	1.59	89.22
31	True	0.84	0.84	-0.79	-6.19	-1.16	-2.70	-0.06	0.65	0.0	-2.60	7.94	36.54
35	True	0.08	0.08	1.92	10.52	2.16	-4.36	1.29	6.12	0.0	4.78	-0.87	16.07
39	True	0.03	0.03	1.64	5.33	1.97	6.04	0.77	17.94	0.0	0.43	-9.07	-36.69

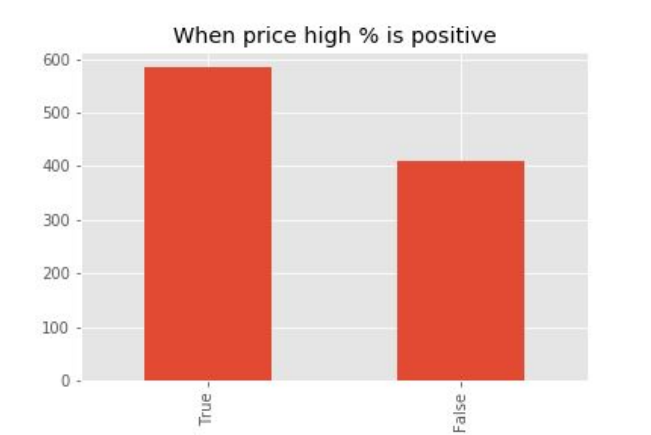
Class Balance

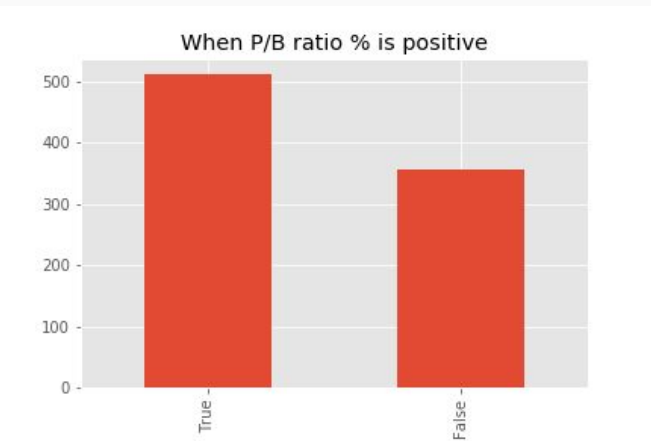


Features Correlation



Count of True/False when specific features are positive

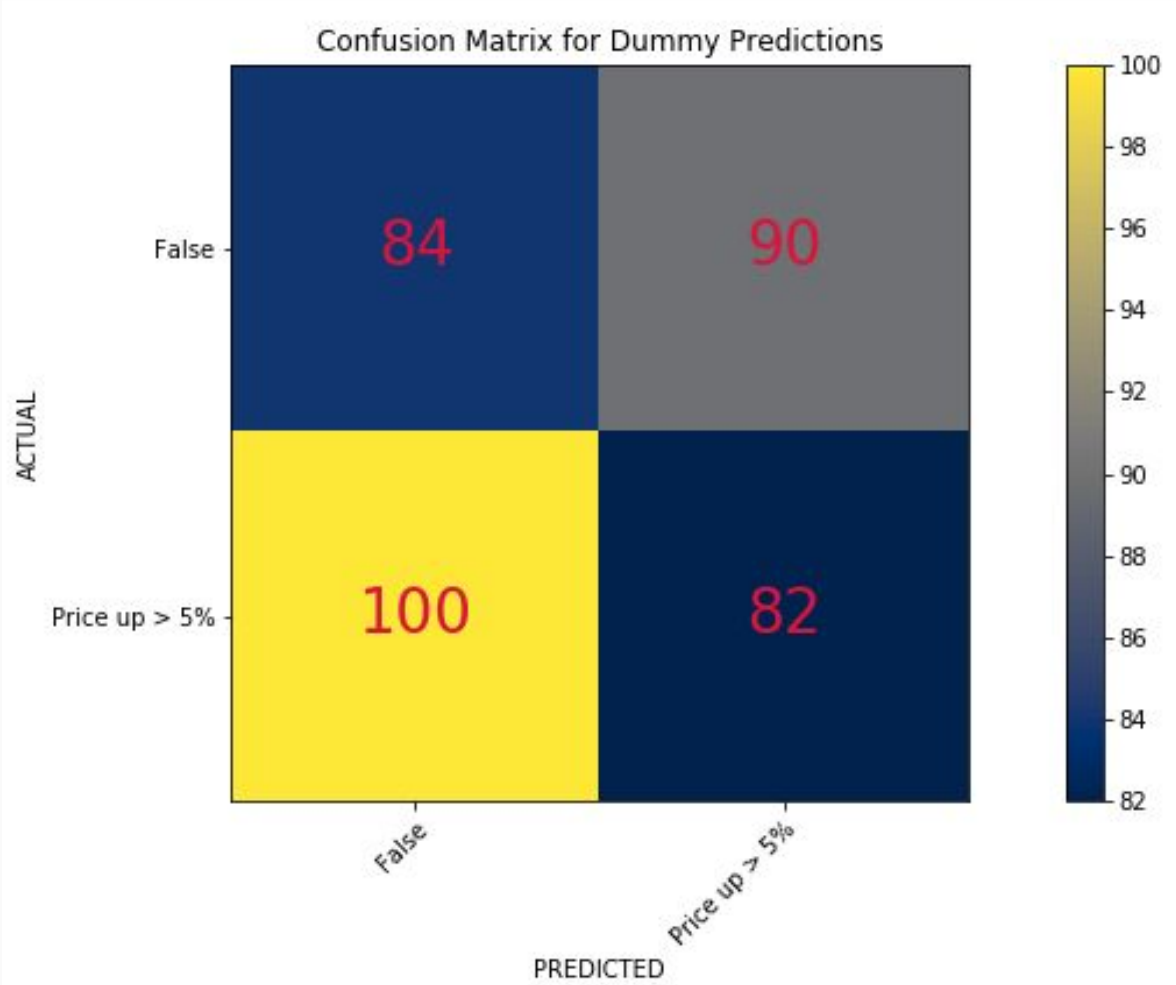




Modeling

Baseline Model: Dummy Classifier

- Accuracy:
  - 46%
- F1 Score:
  - 46%



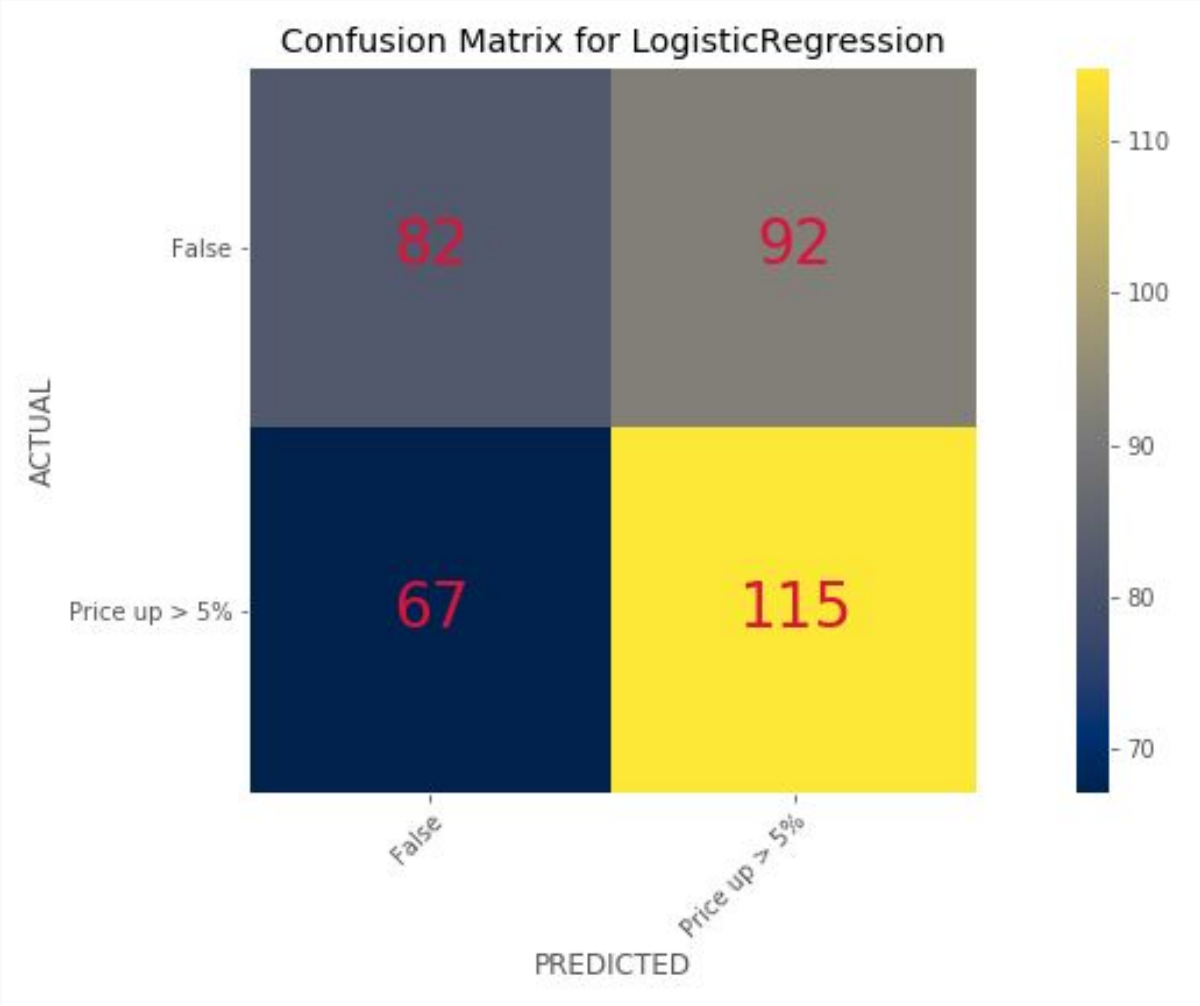
Models Used

Optimized using GridSearchCV

- Logistic Regression
- K Nearest Neighbors
- Decision Tree
- Random Forest
- XGBoost

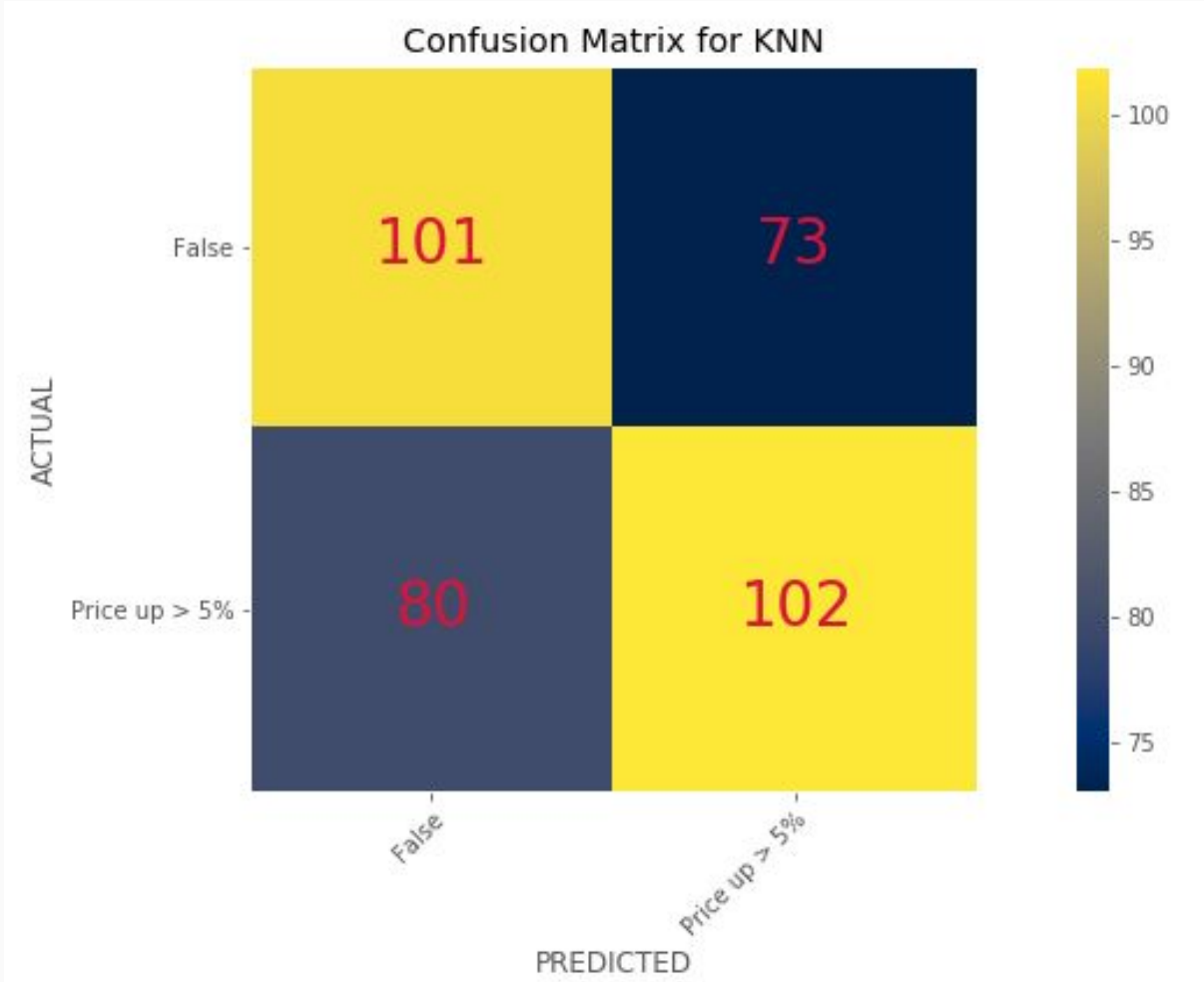
Logistic Regression

- Accuracy:
  - 56%
- F1 Score:
  - 60%



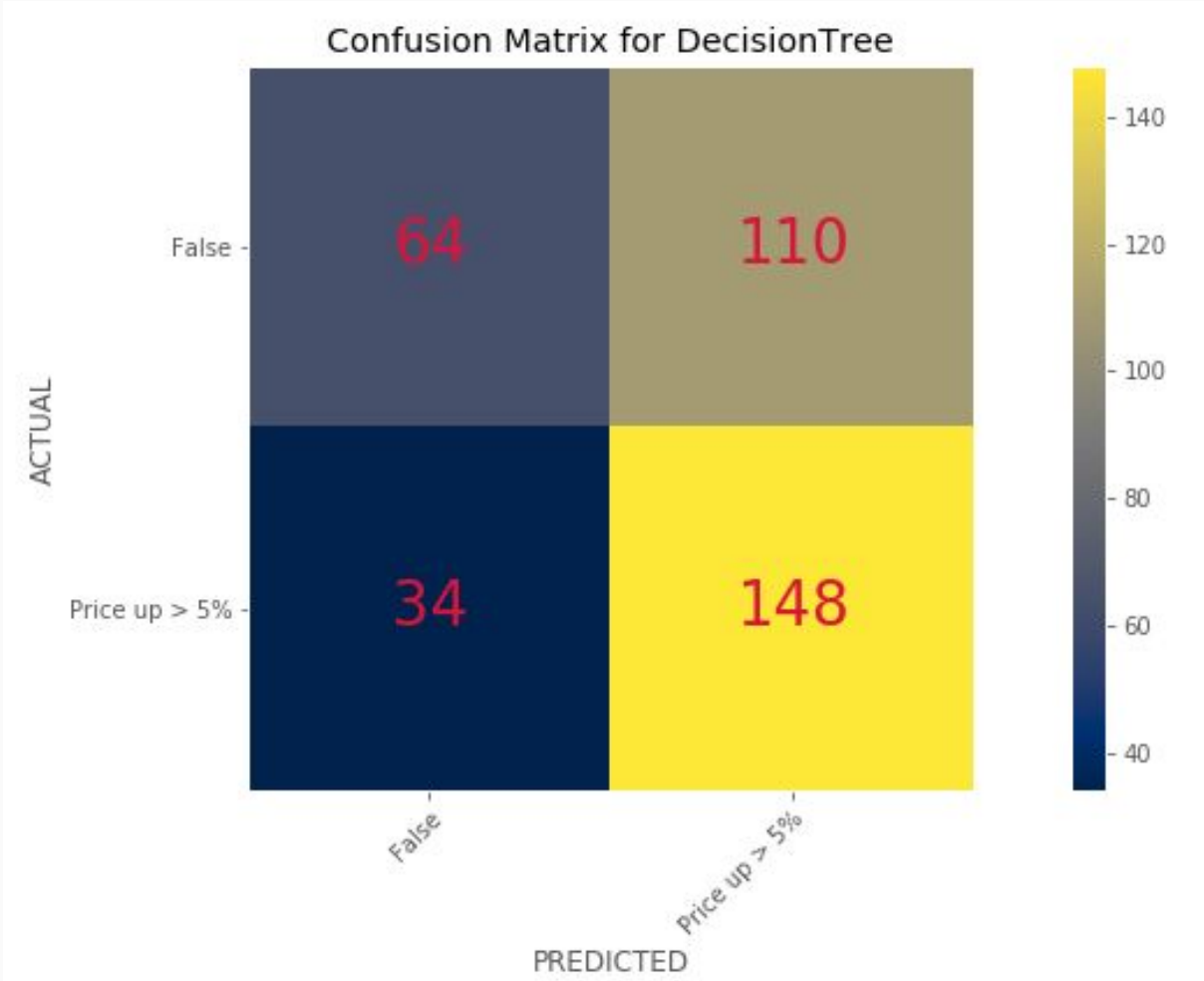
K Nearest Neighbors

- Accuracy:
  - 57%
- F1 Score:
  - 57%



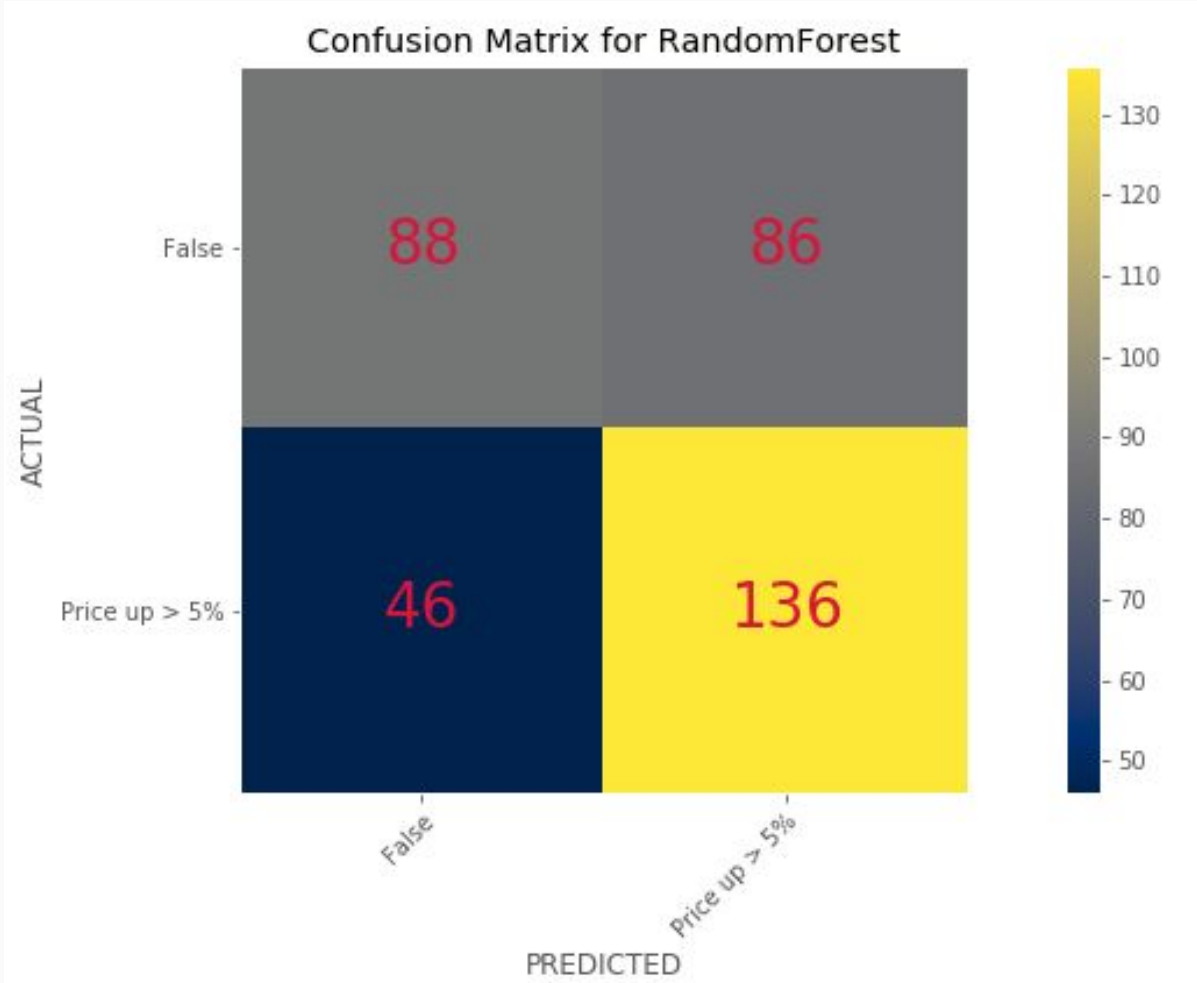
Decision Tree

- Accuracy:
  - 59%
- F1 Score:
  - 67%



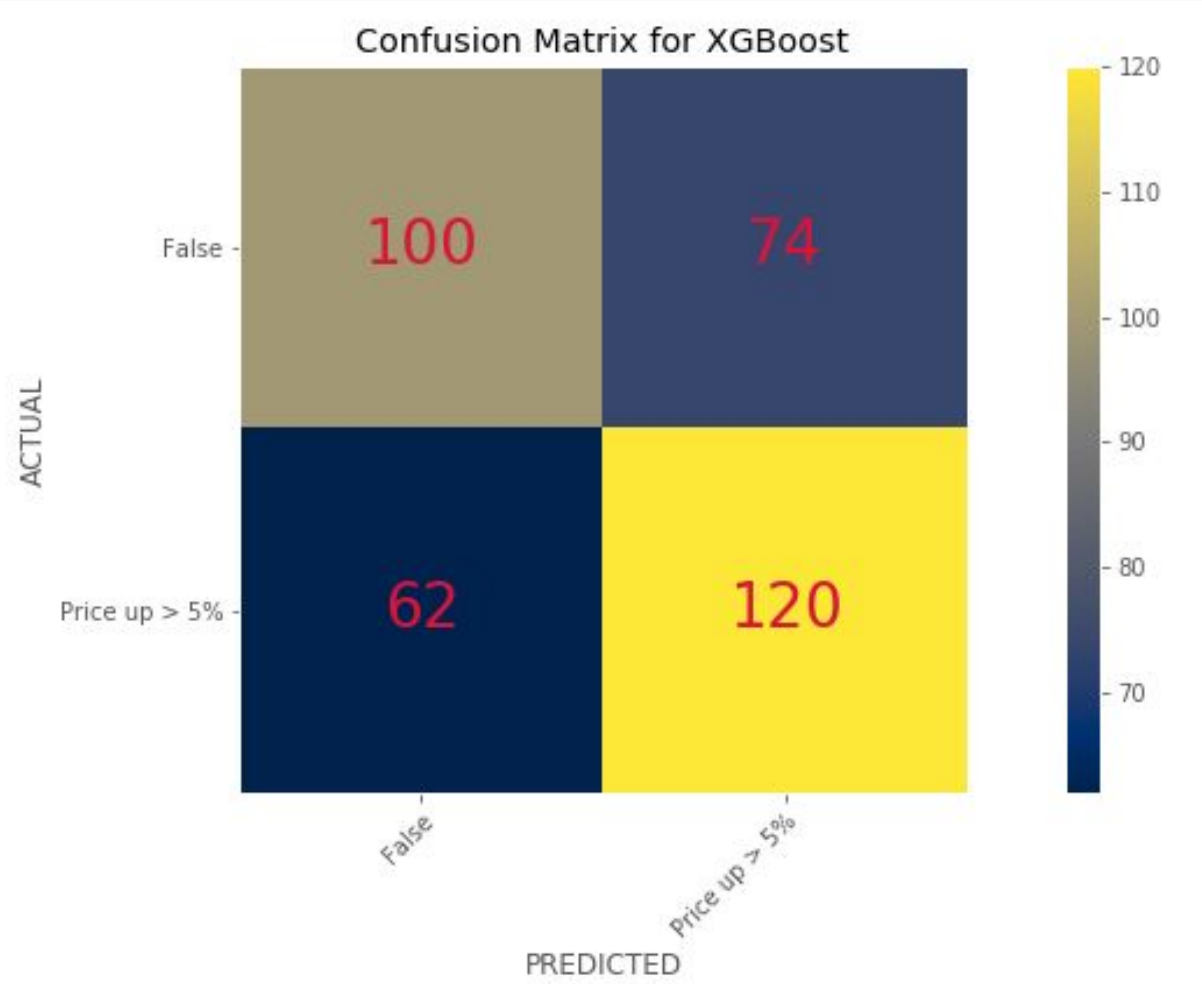
Random Forest

- Accuracy:
  - 58%-62%
- F1 Score:
  - 62%-67%



XGBoost

- Accuracy:
  - 61%
- F1 Score:
  - 63%



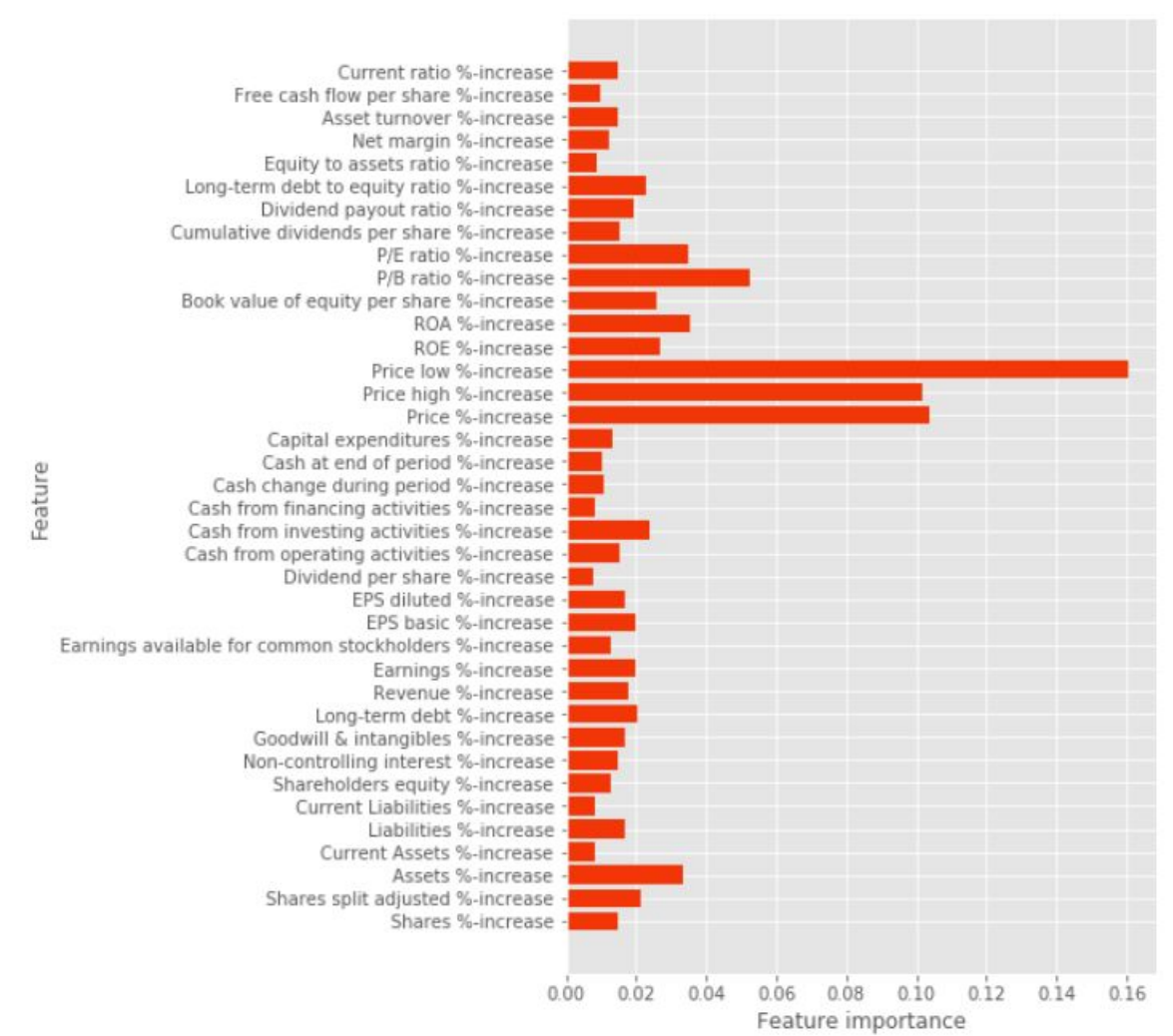
Best Model

- For the Risk-Averse:
  - **XGBoost**
    - Most consistent
- For the Risk-Takers:
  - **Random Forest**
    - More opportunities

Which Features were important?

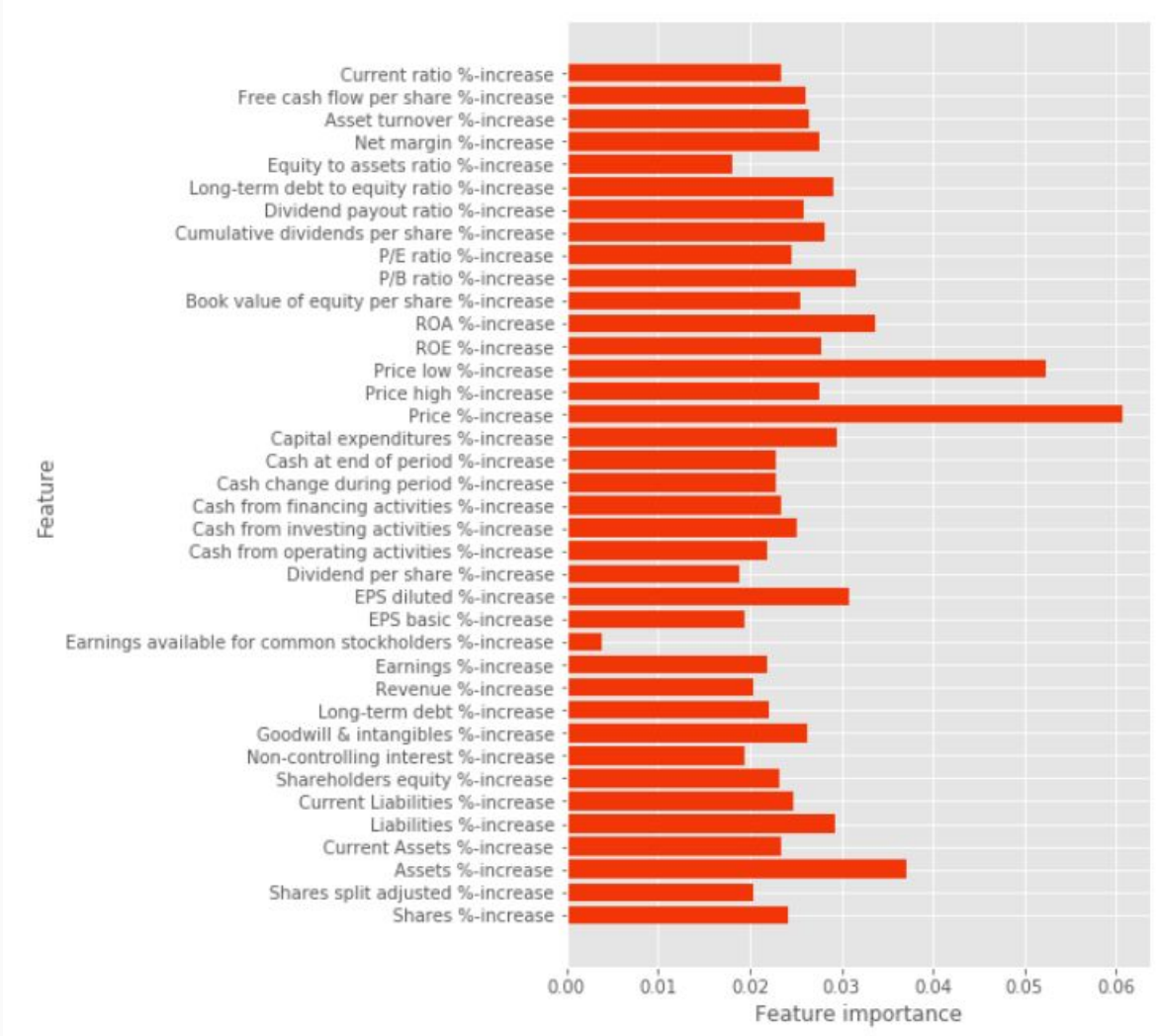
Random Forest

- Price Low
- Price High
- Price
  - Emphasis on Price Low



XGBoost

- Price Low
- Price
  - Both are almost equally important



Potential Improvements

- Add feature interactions
- Eliminate unimportant features
- Look for other sources
- Potentially run more models (SVM, Naive Bayes).

Summary

- Best models:
  - XGBoost
  - Random Forest
    - Depending on your risk tolerance
- Look for Price High/Low changes