Model Evaluation Report

Investment prediction project

Georgie Zaguirre

Table of Contents

1 Introduction

This report will outline the steps taken to develop a machine learning (ML) model that will make recommendations on which top 500 US are having the highest average growth to invest in. This report will also outline the model performance, insights and recommendations for future development.

2 Business problem statement

An American electronics conglomerate company are looking to buy new companies and rapid grow their company. Its leaders want to speed up the process of identifying which of the top 500 US companies are good acquisition targets by using ML algorithms.

This will be a binary classification problem where the algorithm will predict which companies are a good acquisition or not, by defining a good acquisition is a company with Average Annual Revenue Growth above 0.7.

3 Exploratory data analysis (EDA) and pre-processing

Initial data preparation and pre-processing of the Russell 3000 Index financial data were done, through first loading and merging various datasets and with total of 78 variables initially identified. The target variable Average Annual Revenue Growth was created by calculating the increase of revenue divided by the company's original revenue then divided it by the number of years separating the entries. Data were also cleaned, where missing data were filled and outliers removed.

The following analysis and insights were initially found:

3.1 Average Annual Revenue Growth distributions

a. The first result of the EDA showed that the target variable Average Annual Revenue Growth shows a positively skewed distribution:

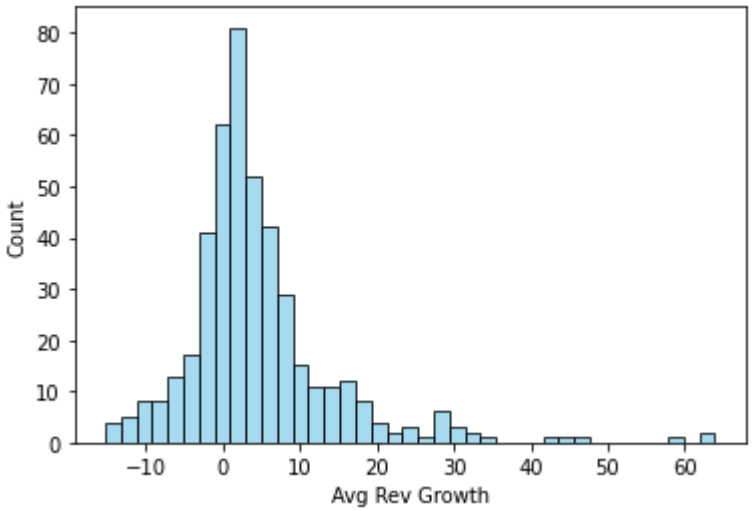


Figure 1: *Distribution of Average Revenue Growth*

b. Real Estate sector has the highest distribution Average revenue growth, followed by Health Care and Information Technology, while the Energy sector is at the bottom.

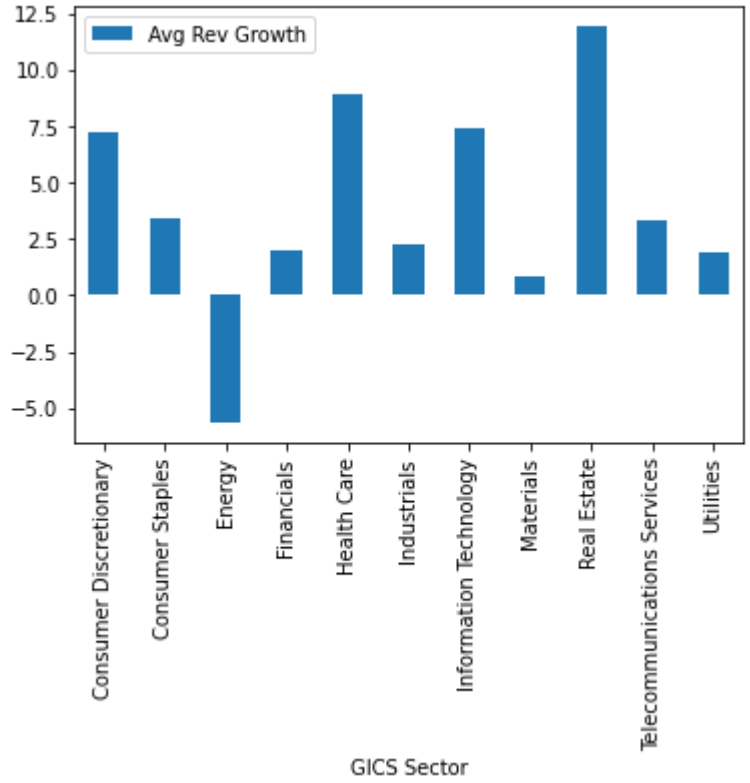


Figure 2: *Distribution of Average Revenue Growth by Sector*

c. Further analysis of the Average revenue growth by sector show Real Estate has a normal distribution but a few outliers do exist for Health Care and Information Technology.

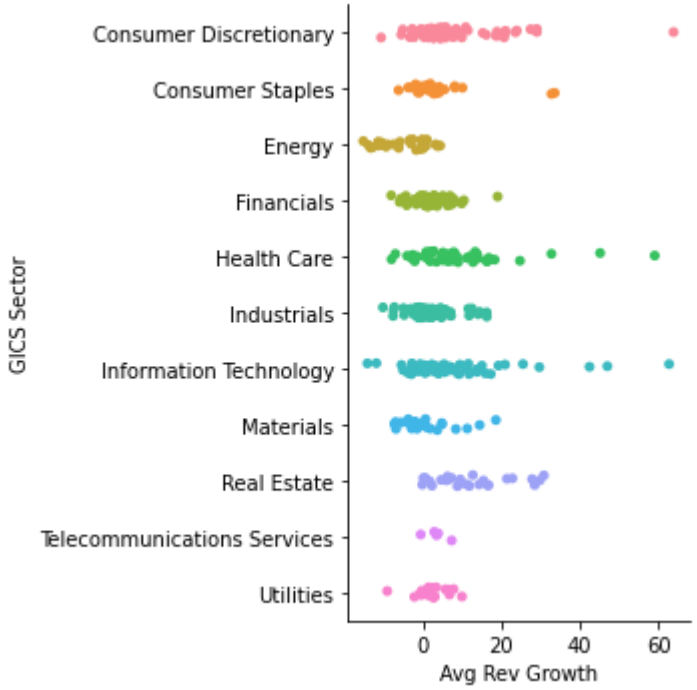


Figure 3: *Distribution of Average Revenue Growth by Sector using categorical distribution plots*

3.2 Correlation analysis

Correlation analysis have identified many highly correlated features where Pearson's correlation coefficient is either greater or equal to 0.85 or less than or equal to -0.85. For full list, see the Appendices.

3.2.1 Test hypothesis from correlation analysis

Correlation between Long-Term Investments and Total Liabilities & Equity were tested.

Null hypothesis: There IS a correlation between Long-Term Investments and Total Liabilities & Equity

Alternate hypothesis: There is NO correlation between Long-Term Investments and Total Liabilities & Equity

Results:
The 95% confidence interval show that correlation is between 0.92 and 0.99 and therefore accept the null hypothesis. This an example of hypothesis testing that can be done on any of the correlated features and by performing a test, will ensure that any domain-driven hypothesis about specific variables, causation or correlations can be tested with degree of confidence, rather the business 'gut-feel'.

4 Model development and evaluation

4.1 Model selection

The classifiers that that were initially evaluated and compared are Random Forest, Gradient boosting, Support Vector Machine (SVM), Logistic Regression and Naive Bayes. For evaluation, the metric Precision-Recall will be used to evaluate model performance as analysis of the classes show that the data set is mildly imbalanced (1= 67%, 0 = 33%). Since there is a higher cost for False positives, or incorrect predictions to invest companies that are actually making losses, the metric that are useful when comparing these models is the Precision (True negative), as opposed to a False negative means that the company is missing out on investing on growth investments but not necessarily losing money.

Base model results:

By looking at the metric Precision (True negative), Random Forest and SVM equally have the highest Precision and accuracy scores. However, Random Forest had slightly higher average accuracy score than SVM with average scores of 0.76 and 0.75 respectively, when cross validated 5 times. Since these are the best performing algorithms thus far, before feature selection and hyperparameter tuning, they have been shortlisted as the best performing base models and therefore no further evaluation will be done on the rest of the other algorithms. This analysis assumes that the other algorithms cannot be improved and can now be discarded.

	Random Forest	Gradient boosting	SVM	Logistic regression	Naïve Bayes
Model accuracy on test data	0.73	0.72	0.73	0.70	0.637037037
Mean Cross validated Score (5 folds)	0.76	0.75	0.75	0.75	0.69
Precision (True negative=0)	0.75	0.70	0.75	0.71	0.55
Precision (True positive=1)	0.73	0.73	0.73	0.70	0.65

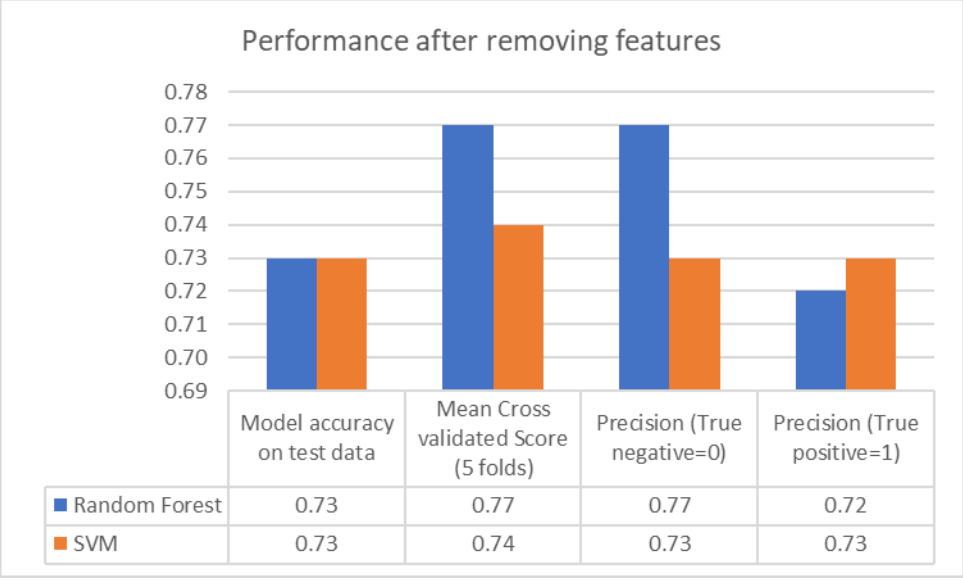
Figure 4: *Model accuracy score and precision comparison*

5 Feature selection

Feature selection is very important for the predictability of the models. Adding variables that are well known by domain experts can greatly inform and influence model accuracy.

This feature selection process was performed in two parts: Firstly, highly correlated features identified earlier were removed and independently, the second part was to add the features in the additional datasets 'prices.csv' and 'prices-split-adjusted.csv' to see if either or both will make a difference to the model performance.

5.1 Results of removing highly correlated features



	Model accuracy on test data	Mean Cross validated Score (5 folds)	Precision (True negative=0)	Precision (True positive=1)
Random Forest	0.73	0.77	0.77	0.72
SVM	0.73	0.74	0.73	0.73

Figure 5: *Model performance after removing correlated features*

Removing highly correlated features have improved the cross validated scores from 0.76 to 0.77 and precision scores from 0.75 to 0.77 of the Random Forest and is outperforming SVM.

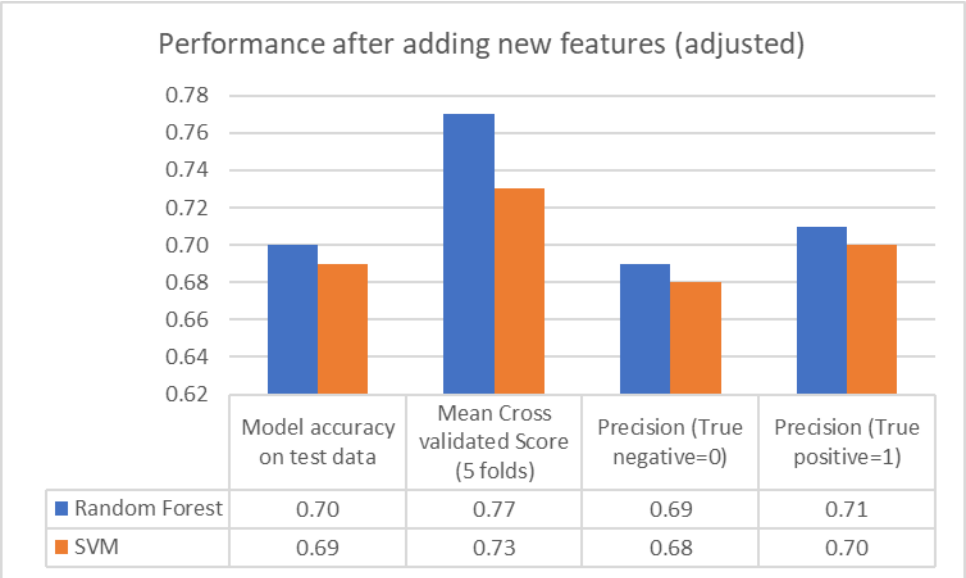
5.2 Results of adding new features

Stock's closing price is a common measure of how a share has performed during the day used by financial institutions, regulators and individual investors. For this reason, the feature 'close' and ‘volume’ were added, instead of other prices features.

	Model accuracy on test data	Mean Cross validated Score (5 folds)	Precision (True negative=0)	Precision (True positive=1)
Random Forest	0.70	0.77	0.68	0.71
SVM	0.68	0.74	0.65	0.73

Figure 6*: Model performance after adding new features prices and volumes*

The results above show that adding new features Closing prices or Volumes has not improved model performance.



	Model accuracy on test data	Mean Cross validated Score (5 folds)	Precision (True negative=0)	Precision (True positive=1)
Random Forest	0.70	0.77	0.69	0.71
SVM	0.69	0.73	0.68	0.70

Figure 7: *Model performance after adding new features prices and volumes (adjusted)*

The results above show that adding new adjusted features Closing prices or Volumes did not improved model performance either.

5.3 Conclusion on models and feature selection process

The final conclusion on the model selection is that Random Forest has highest performance with the highest Precision (0 - True negatives) at 0.77 and model accuracy. The improvement was achieved by removing some correlated features. The difference in percentage for the business can potentially mean differences in millions.

6 Model optimisation

6.1 Hyperparameter tuning

Hyperparameter tuning is the process of finding out the most optimised set of parameters for the selected model. This is important as again any minor improvements for the business can translate to differences in millions in gains or losses. Furthermore, a minor improvement can mean if the model will be productionised or not.

The following best parameters for the classifier Random Forest were found using a random search method:

n_estimators= 800,
max_depth=50,

Applying the above hyperparameters has improved the model accuracy score on test or unseen data by 1 % from 0.73 to 0.74 as shown below.

```
In [668]: randomForestClassifier.score(X_test, y_test)
Out[668]: 0.7407407407407407
```

6.2 Model Feature importance

Extracting the model feature importance can provide the business information on which features have had the most influence on its prediction and therefore help model explainability, track its outcomes and decisions. This is important for accountability, audibility and will help learn from mistakes and avoid errors.

For the Random Forest model, the top 10 important features were:

```
[   'Liabilities',
    'Accounts Receivable',
    'Non-Recurring Items',
    'Retained Earnings',
    'Common Stocks',
    'Other Equity',
    'Net Cash Flows-Financing',
    'Other Investing Activities',
    'Industry',
    'Other Assets']
[   0.06294031333457192,
    0.04055367659598279,
    0.03889673241193252,
    0.033400013395368716,
    0.030139628399118993,
    0.02744209123851058,
    0.025889878116595804,
    0.02578921813433785,
    0.025171055142253724,
    0.024446147245993862]
```

Figure 8: *Top 10 important features and weights*

Additionally, extracting the important features can help with tuning performance by selecting important features only to help avoid model overfitting. In the above scenario, the top 10 important features were fitted onto the Random Forest base model but the results showed that the performance was almost the same as the original dataset with 78 features. This shows that even starting with just a base model, even before feature selection and hyperparameters tuning, the Random Forest model can be just as effective.

7 Ethical and sustainability considerations

The ethical and sustainability issues that must be considered are model explainability issues, model drift and bias. Model outputs that cannot be explained or any changes to the data (data drift), degrades model predictability and risks of having unethically biased models and inaccurate predictions. Consequently, these issues can lead to models not being able to be productionised or unsustainable and cannot provide value to the business.

One of the key ways to address these issues is to analyse the bias and variance trade off so further improvements can be done on the models.

The Random Forest model seen here shows that further optimisation is needed to lower variance, at the cost of bias and to reduce number of features. The graph also shows the model benefits as more training data is added.

Figure 9: *Random Forest Learning curve*

8 Insights and recommendations

In conclusion, the selected model Random Forest performed the best with slight improvements after feature selection and hyperparameters tuning. It however requires further optimisation to lower the variance, it is recommended that an iterative process of feature selection and feature engineering with domain experts be done and to add more training data, test other algorithms and use other target variables.

While this model is not yet fit for production, the model evaluation process has however provided some valuable insights:

- Real Estate sector has the highest distribution Average Revenue Growth, followed by Health Care and Information Technology. The distribution of Average Revenue Growth of Real Estate sector showed it's less of a risky investment, with no outliers and with highest average growth.
- The model predicted 85 companies to have Average Revenue Growth of over 0.7, predicting Charter Communications having the highest growth, followed by Facebook and Procter & Gamble.

This report recommends that the company to research from a select portfolio of potential acquisitions from 85 companies the model has predicted to have the highest growth. See Appendices for the top 30 companies with the highest revenue growth.

Appendices

8.1 Correlated features (Pearson's correlation coefficient >= 0.85 or -0.85)

	FEATURE_1	FEATURE_2	CORRELATION
287	After Tax ROE	Pre-Tax ROE	0.994503
520	Cash Ratio	Quick Ratio	0.927476
569	Cash and Cash Equivalents	Long-Term Investments	0.947221
587	Cash and Cash Equivalents	Other Current Liabilities	0.989964
604	Cash and Cash Equivalents	Total Assets	0.922065
608	Cash and Cash Equivalents	Total Liabilities	0.925654
609	Cash and Cash Equivalents	Total Liabilities & Equity	0.922067
841	Cost of Revenue	Total Revenue	0.957399
905	Current Ratio	Quick Ratio	0.943008
1171	Earnings Before Interest and Tax	Earnings Before Tax	0.984904
1178	Earnings Before Interest and Tax	Income Tax	0.948434
1190	Earnings Before Interest and Tax	Net Cash Flow-Operating	0.942894
1193	Earnings Before Interest and Tax	Net Income	0.972390
1195	Earnings Before Interest and Tax	Net Income Applicable to Common Shareholders	0.969472
1196	Earnings Before Interest and Tax	Net Income-Cont. Operations	0.973145
1199	Earnings Before Interest and Tax	Operating Income	0.904820
1255	Earnings Before Tax	Income Tax	0.966300
1267	Earnings Before Tax	Net Cash Flow-Operating	0.910720
1270	Earnings Before Tax	Net Income	0.987491
1272	Earnings Before Tax	Net Income Applicable to Common Shareholders	0.987263
1273	Earnings Before Tax	Net Income-Cont. Operations	0.991957
1276	Earnings Before Tax	Operating Income	0.922193
1809	Income Tax	Net Income	0.925257
1811	Income Tax	Net Income Applicable to Common Shareholders	0.925568
1812	Income Tax	Net Income-Cont. Operations	0.955191
2358	Long-Term Investments	Other Current Liabilities	0.958393
2373	Long-Term Investments	Short-Term Debt / Current Portion of Long-Term...	0.912566
2375	Long-Term Investments	Total Assets	0.981104
2379	Long-Term Investments	Total Liabilities	0.985832
2380	Long-Term Investments	Total Liabilities & Equity	0.981105
2733	Net Cash Flow-Operating	Net Income	0.904001
2966	Net Income	Net Income Applicable to Common Shareholders	0.999524
2967	Net Income	Net Income-Cont. Operations	0.985630
2970	Net Income	Operating Income	0.923655
3121	Net Income Applicable to Common Shareholders	Net Income-Cont. Operations	0.985443
3124	Net Income Applicable to Common Shareholders	Operating Income	0.928589
3201	Net Income-Cont. Operations	Operating Income	0.907126
3759	Other Current Liabilities	Short-Term Debt / Current Portion of Long-Term...	0.911964
3761	Other Current Liabilities	Total Assets	0.927977
3765	Other Current Liabilities	Total Liabilities	0.932758
3766	Other Current Liabilities	Total Liabilities & Equity	0.927977
4292	Pre-Tax Margin	Profit Margin	0.907473
4916	Short-Term Debt / Current Portion of Long-Term...	Total Assets	0.911914
4920	Short-Term Debt / Current Portion of Long-Term...	Total Liabilities	0.913590
4921	Short-Term Debt / Current Portion of Long-Term...	Total Liabilities & Equity	0.911910
5074	Total Assets	Total Liabilities	0.997381
5075	Total Assets	Total Liabilities & Equity	1.000000
5149	Total Current Assets	Total Current Liabilities	0.907907
5383	Total Liabilities	Total Liabilities & Equity	0.997381

8.2 Top 30 companies with the highest revenue growth

Ticker Symbol	revGrowthAboveSeven	Security	SEC filings	GICS Sector	GICS Sub Industry	Address of Headquarters	Date first added	CIK	Avg Rev Growth
CHTR	1	Charter Communications	reports	Consumer Discretionary	Cable & Satellite	Stamford, Connecticut	8/9/16	1091667	63.911711
FB	1	Facebook	reports	Information Technology	Internet Software & Services	Menlo Park, California	23/12/13	1326801	62.773120
PG	1	Procter & Gamble	reports	Consumer Staples	Personal Products	Cincinnati, Ohio	NaN	80424	32.731780
MAA	1	Mid-America Apartments	reports	Real Estate	Residential REITs	Memphis, Tennessee	2/12/16	912595	29.780694
CRM	1	Salesforce.com	reports	Information Technology	Internet Software & Services	San Francisco, California	15/9/08	1108524	29.645818
UA	1	Under Armour	reports	Consumer Discretionary	Apparel, Accessories & Luxury Goods	Baltimore, Maryland	1/5/14	1336917	28.998415
KORS	1	Michael Kors Holdings	reports	Consumer Discretionary	Apparel, Accessories & Luxury Goods	New York, New York	13/11/13	1530721	28.994945
HCN	1	Welltower Inc.	reports	Real Estate	REITs	Toledo, Ohio	30/1/09	766704	28.450310
O	1	Realty Income Corporation	reports	Real Estate	Retail REITs	San Diego, California	7/4/15	726728	27.792258
LKQ	1	LKQ Corporation	reports	Consumer Discretionary	Distributors	Chicago, Illinois	23/5/16	1065696	18.613601
CELG	1	Celgene Corp.	reports	Health Care	Biotechnology	Summit, New Jersey	NaN	816284	18.229800
CTSH	1	Cognizant Technology Solutions	reports	Information Technology	IT Consulting & Other Services	Teaneck, New Jersey	NaN	1058290	17.251573
ABC	1	AmerisourceBergen Corp	reports	Health Care	Health Care Distributors	Chesterbrook, Pennsylvania	NaN	1140859	16.738028
CBG	1	CBRE Group	reports	Real Estate	Real Estate Services	Los Angeles, California	NaN	1138118	16.662746
BIIB	1	BIOGEN IDEC Inc.	reports	Health Care	Biotechnology	Weston, Massachusetts	NaN	875045	16.288480
LRCX	1	Lam Research	reports	Information Technology	Semiconductor Equipment	Fremont, California	29/6/12	707549	15.886568
AYI	1	Acuity Brands Inc	reports	Industrials	Electrical Components & Equipment	Atlanta, Georgia	3/5/16	1144215	14.386578
SRCL	1	Stericycle Inc	reports	Industrials	Industrial Conglomerates	Lake Forest, Illinois	19/11/08	861878	14.018236
BDX	1	Becton Dickinson	reports	Health Care	Health Care Equipment	Franklin Lakes, New Jersey	30/9/72	10795	13.747827
TSS	1	Total System Services	reports	Information Technology	Internet Software & Services	Columbus, Georgia	2/1/08	721683	13.743416
RHT	1	Red Hat Inc.	reports	Information Technology	Systems Software	Raleigh, North Carolina	27/7/09	1087423	13.610095
PDCO	1	Patterson Companies	reports	Health Care	Health Care Supplies	St. Paul, Minnesota	NaN	891024	12.024945
FBHS	1	Fortune Brands Home & Security	reports	Industrials	Building Products	Deerfield, Illinois	22/6/16	1519751	11.520671
AAP	1	Advance Auto Parts	reports	Consumer Discretionary	Automotive Retail	Roanoke, Virginia	9/7/15	1158449	11.384410
ADBE	1	Adobe Systems Inc	reports	Information Technology	Application Software	San Jose, California	5/5/97	796343	11.091760
SBUX	1	Starbucks Corp.	reports	Consumer Discretionary	Restaurants	Seattle, Washington	NaN	829224	10.844802
MHK	1	Mohawk Industries	reports	Consumer Discretionary	Home Furnishings	Amsterdam, New York	23/12/13	851968	9.863471
AMG	1	Affiliated Managers Group Inc	reports	Financials	Asset Management & Custody Banks	Beverly, Massachusetts	1/7/14	1004434	9.826433
KMX	1	Carmax Inc	reports	Consumer Discretionary	Specialty Stores	Richmond, Virginia	28/6/10	1170010	9.547858
VTR	1	Ventas Inc	reports	Real Estate	REITs	Chicago, Illinois	4/3/09	740260	9.226645