DBA3803 / DSC3216 Group Project

TheDailyProphet (Team 8)

E0318558	CHIAM XIU ZHEN
E0004365	GAN CHIN BOON
E0191659	LIM YUCHENG EAGAN
E0319030	NAI JING WEN, BERNESSA
E0200936	ZHOU KAI JING

Contents

1 Introduction / Data Preprocessing

2 Model 1 : Regression

3 Model 2 : Time Series

4 Model 3 : Clustering

5 Further Analysis / Conclusion

Section 1

Introduction &
Data Preprocessing

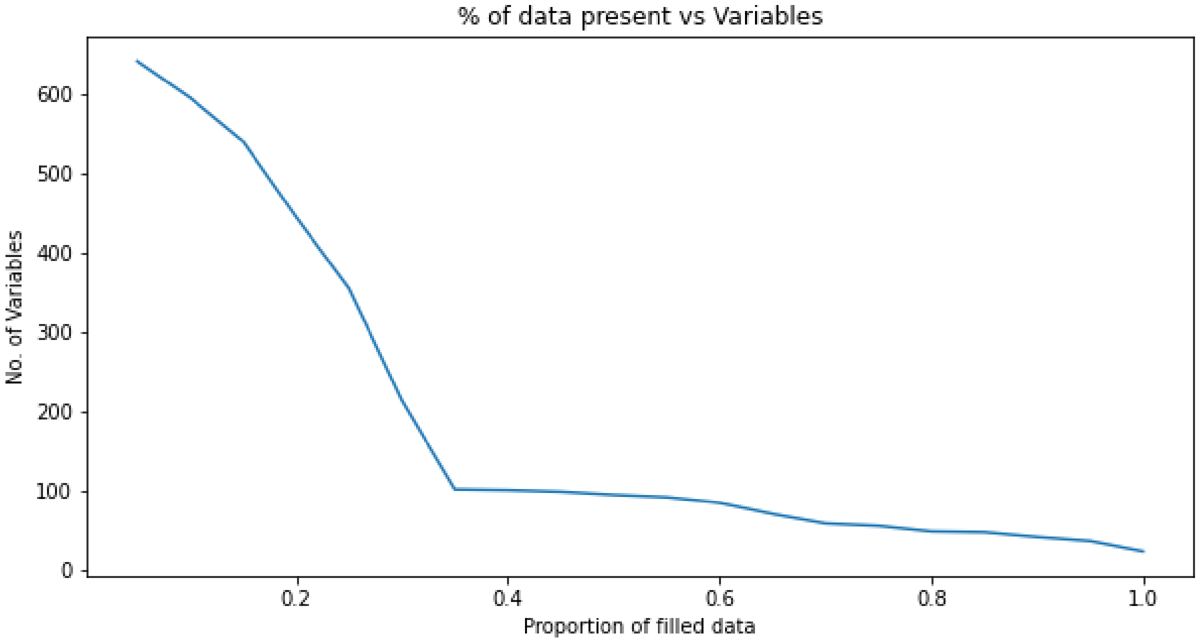
Introduction

- **Objectives**
  - Understand market trends
  - Recognize different market segments
  - Forecast company revenues
  - Predict stock prices

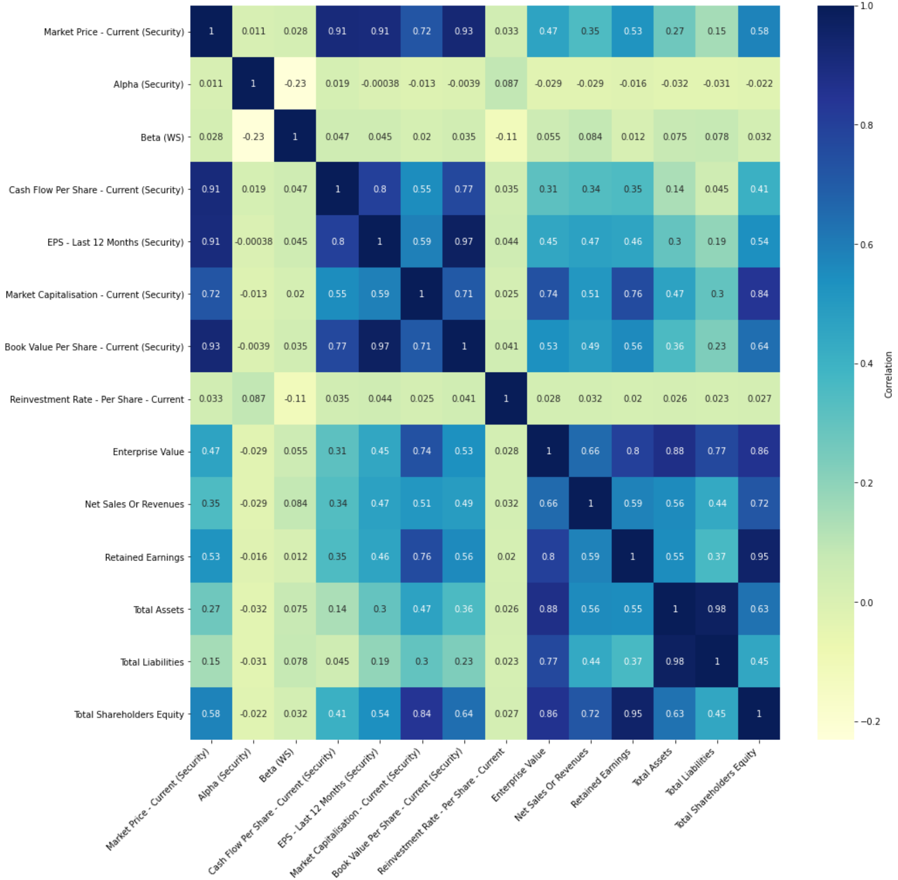
- **Model Overview**
  - Regression
  - Time Series
  - Clustering

Data Preprocessing

- Long vs Wide Data



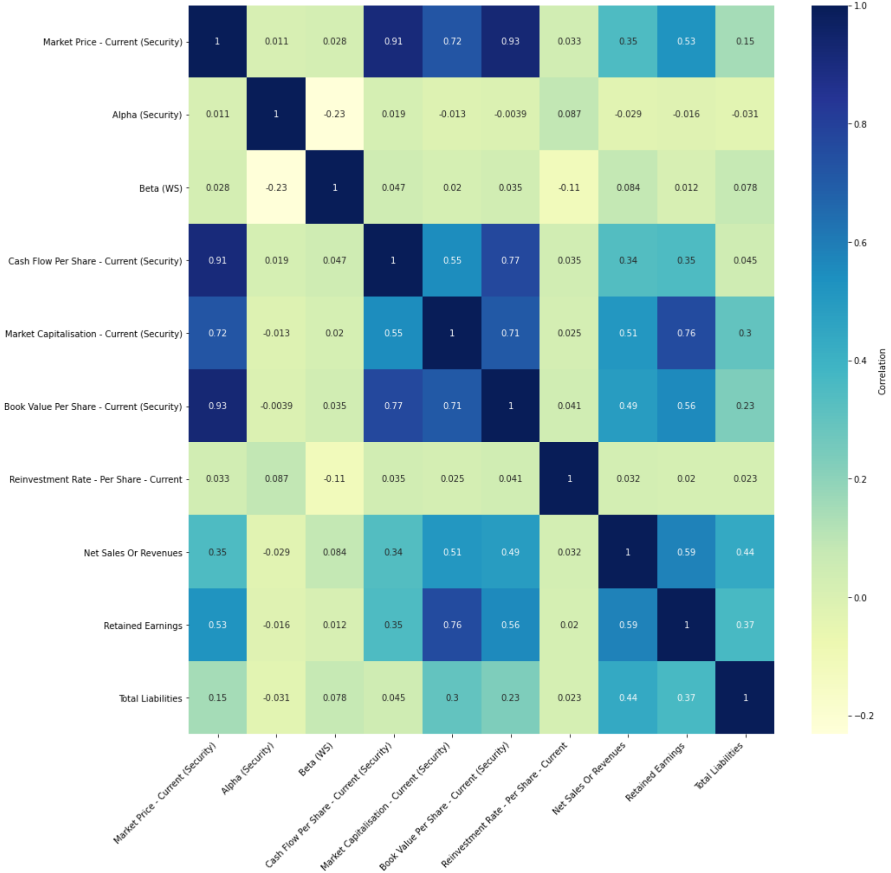
Data Preprocessing



	Market Price - Current (Security)	Alpha (Security)	Beta (WS)	Cash Flow Per Share - Current (Security)	EPS - Last 12 Months (Security)	Market Capitalisation - Current (Security)	Book Value Per Share - Current (Security)	Reinvestment Rate - Per Share - Current	Enterprise Value	Net Sales Or Revenues	Retained Earnings	Total Assets	Total Liabilities	Total Shareholders Equity
Market Price - Current (Security)	1	0.011	0.028	0.91	0.91	0.72	0.93	0.033	0.47	0.35	0.53	0.27	0.15	0.58
Alpha (Security)	0.011	1	-0.23	0.019	-0.00038	-0.013	-0.0039	0.087	-0.029	-0.029	-0.016	-0.032	-0.031	-0.022
Beta (WS)	0.028	-0.23	1	0.047	0.045	0.02	0.035	-0.11	0.055	0.084	0.012	0.075	0.078	0.032
Cash Flow Per Share - Current (Security)	0.91	0.019	0.047	1	0.8	0.55	0.77	0.035	0.31	0.34	0.35	0.14	0.045	0.41
EPS - Last 12 Months (Security)	0.91	-0.00038	0.045	0.8	1	0.59	0.97	0.044	0.45	0.47	0.46	0.3	0.19	0.54
Market Capitalisation - Current (Security)	0.72	-0.013	0.02	0.55	0.59	1	0.71	0.025	0.74	0.51	0.76	0.47	0.3	0.84
Book Value Per Share - Current (Security)	0.93	-0.0039	0.035	0.77	0.97	0.71	1	0.041	0.53	0.49	0.56	0.36	0.23	0.64
Reinvestment Rate - Per Share - Current	0.033	0.087	-0.11	0.035	0.044	0.025	0.041	1	0.028	0.032	0.02	0.026	0.023	0.027
Enterprise Value	0.47	-0.029	0.055	0.31	0.45	0.74	0.53	0.028	1	0.66	0.8	0.88	0.77	0.86
Net Sales Or Revenues	0.35	-0.029	0.084	0.34	0.47	0.51	0.49	0.032	0.66	1	0.59	0.56	0.44	0.72
Retained Earnings	0.53	-0.016	0.012	0.35	0.46	0.76	0.56	0.02	0.8	0.59	1	0.55	0.37	0.95
Total Assets	0.27	-0.032	0.075	0.14	0.3	0.47	0.36	0.026	0.88	0.56	0.55	1	0.98	0.63
Total Liabilities	0.15	-0.031	0.078	0.045	0.19	0.3	0.23	0.023	0.77	0.44	0.37	0.98	1	0.45
Total Shareholders Equity	0.58	-0.022	0.032	0.41	0.54	0.84	0.64	0.027	0.86	0.72	0.95	0.63	0.45	1

- **Manual Feature Selection**
  - Key Financial Statement items selected
  - Excludes Financial Ratios
- **Correlation Matrix Heatmap**
  - Used to quickly identify heavily correlated features
  - Highly correlated features removed as part of feature selection

Data Preprocessing



	Market Price - Current (Security)	Alpha (Security)	Beta (WS)	Cash Flow Per Share - Current (Security)	Market Capitalisation - Current (Security)	Book Value Per Share - Current (Security)	Reinvestment Rate - Per Share - Current	Net Sales Or Revenues	Retained Earnings	Total Liabilities
Market Price - Current (Security)	1	0.011	0.028	0.91	0.72	0.93	0.033	0.35	0.53	0.15
Alpha (Security)	0.011	1	-0.23	0.019	-0.013	-0.0039	0.087	-0.029	-0.016	-0.031
Beta (WS)	0.028	-0.23	1	0.047	0.02	0.035	-0.11	0.084	0.012	0.078
Cash Flow Per Share - Current (Security)	0.91	0.019	0.047	1	0.55	0.77	0.035	0.34	0.35	0.045
Market Capitalisation - Current (Security)	0.72	-0.013	0.02	0.55	1	0.71	0.025	0.51	0.76	0.3
Book Value Per Share - Current (Security)	0.93	-0.0039	0.035	0.77	0.71	1	0.041	0.49	0.56	0.23
Reinvestment Rate - Per Share - Current	0.033	0.087	-0.11	0.035	0.025	0.041	1	0.032	0.02	0.023
Net Sales Or Revenues	0.35	-0.029	0.084	0.34	0.51	0.49	0.032	1	0.59	0.44
Retained Earnings	0.53	-0.016	0.012	0.35	0.76	0.56	0.02	0.59	1	0.37
Total Liabilities	0.15	-0.031	0.078	0.045	0.3	0.23	0.023	0.44	0.37	1

- **Correlation-based Feature Selection**
  - Further selection with correlation threshold of 0.8
  - Graph on the left shows post-feature selection correlation matrix
  - Same feature selection process for time series and clustering datasets

Data Preprocessing

- **Time Series Preprocessing**
  - Time series and clustering datasets were separately prepared
  - Time Series data focuses on largest date range for analysis (to best capture market trends) → United Engineers Limited (UEL)
  - UEL has >30 data points → Last 30 years of revenue data used for forecasting

- **Clustering Preprocessing**
  - Clustering data focuses on the year with the most companies with complete info → 2014
- **Feature Scaling & Cross Validation**
  - Different features’ values varies a lot (by many magnitudes) → Feature Scaling
  - Train-test split (30% test-set size) → For out-of-sample cross validation

Section 2

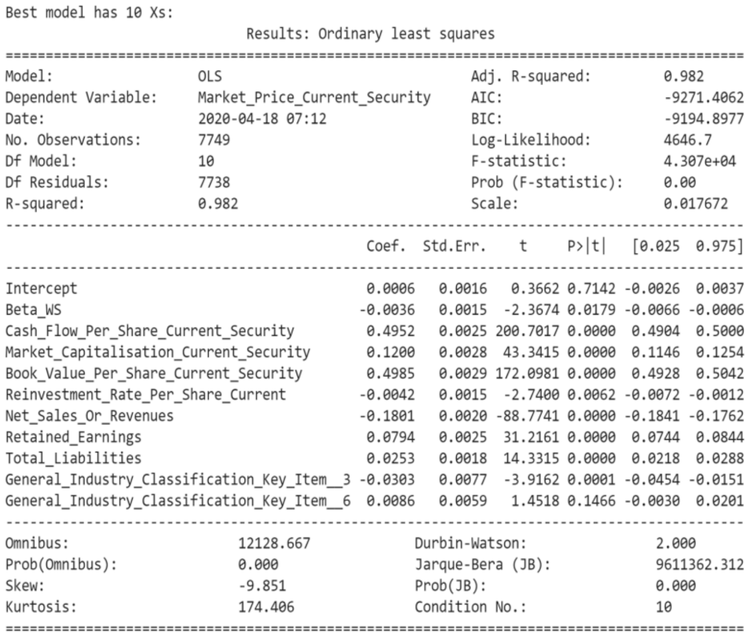
Model 1 : Regression

Regression Models

- **Base Regression Models**
  - OLS Linear Regression model
  - SK-Learn Linear Regression model
- **Regularised Regression Models**
  - Lasso (With L_1 Regularization)
  - Ridge (With L_2 Regularization)
  - Elastic Net Regression
- **Other Regression Models**
  - Random Forest Regressor

	RMSE
OLS (Variable Selection)	**0.141528**
Linear Regression	**0.141515**
Lasso Regression	**0.141515**
Ridge Regression	**0.141509**
Elastic Net Regression	**0.141515**
Random Forest Regression	**0.007314**

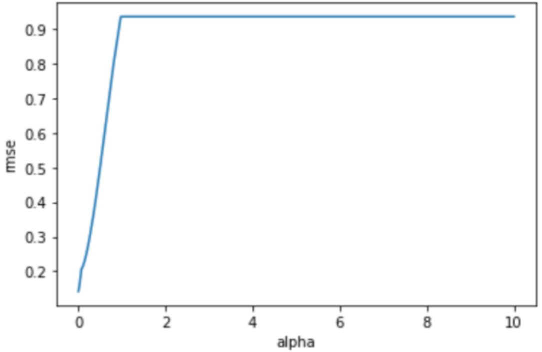
Regression Models



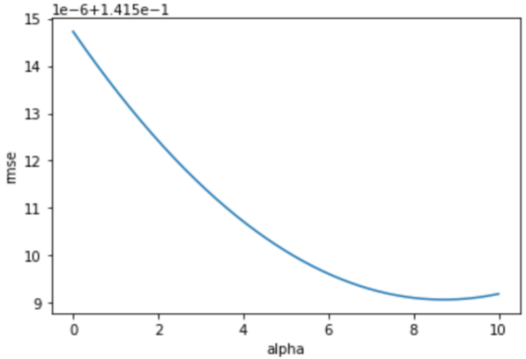
```
Best model has 10 Xs:
                              Results: Ordinary least squares
=============================================================================================
Model:                  OLS                              Adj. R-squared:         0.982
Dependent Variable:     Market_Price_Current_Security    AIC:                   -9271.4062
Date:                   2020-04-18 07:12                 BIC:                   -9194.8977
No. Observations:       7749                             Log-Likelihood:         4646.7
Df Model:               10                               F-statistic:            4.307e+04
Df Residuals:           7738                             Prob (F-statistic):     0.00
R-squared:              0.982                            Scale:                  0.017672
---------------------------------------------------------------------------------------------
                                              Coef.  Std.Err.    t     P>|t|   [0.025  0.975]
---------------------------------------------------------------------------------------------
Intercept                                     0.0006   0.0016   0.3662 0.7142 -0.0026  0.0037
Beta_WS                                      -0.0036   0.0015  -2.3674 0.0179 -0.0066 -0.0006
Cash_Flow_Per_Share_Current_Security          0.4952   0.0025 200.7017 0.0000  0.4904  0.5000
Market_Capitalisation_Current_Security        0.1200   0.0028  43.3415 0.0000  0.1146  0.1254
Book_Value_Per_Share_Current_Security         0.4985   0.0029 172.0981 0.0000  0.4928  0.5042
Reinvestment_Rate_Per_Share_Current          -0.0042   0.0015  -2.7400 0.0062 -0.0072 -0.0012
Net_Sales_Or_Revenues                        -0.1801   0.0020 -88.7741 0.0000 -0.1841 -0.1762
Retained_Earnings                             0.0794   0.0025  31.2161 0.0000  0.0744  0.0844
Total_Liabilities                             0.0253   0.0018  14.3315 0.0000  0.0218  0.0288
General_Industry_Classification_Key_Item__3 -0.0303   0.0077  -3.9162 0.0001 -0.0454 -0.0151
General_Industry_Classification_Key_Item__6  0.0086   0.0059   1.4518 0.1466 -0.0030  0.0201
---------------------------------------------------------------------------------------------
Omnibus:                     12128.667          Durbin-Watson:                 2.000
Prob(Omnibus):               0.000              Jarque-Bera (JB):              9611362.312
Skew:                        -9.851             Prob(JB):                      0.000
Kurtosis:                    174.406            Condition No.:                 10
=============================================================================================
```

- **OLS Regression model**
  - Based on Adjusted R², 10 variables were chosen
- **OLS vs SK-Learn Regression model**
  - RMSE of 0.141528 vs 0.141515
  - Feature selection does not necessarily improve model performance
    - Model Parsimony reduces prediction variance but increases prediction bias → Could still lead to overall increase in model error
    - Random train-test split

Regression Models



```
Best alpha = 0.0
Best RMSE (Lasso Regression) = 0.14151472303855223
```



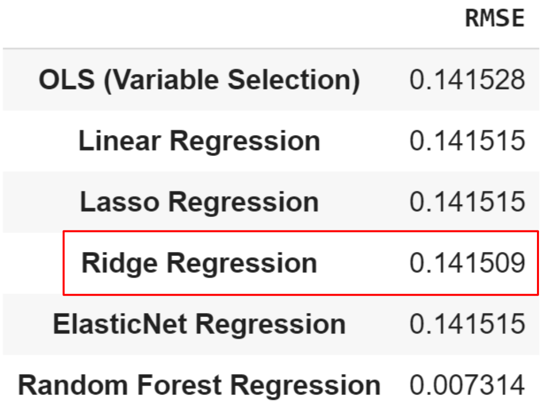
```
Best alpha = 8.72
Best RMSE (Ridge Regression) = 0.1415090718784026
```

- **Regularised Regression model**
  - Prevents overfitting by introducing penalty term
- **Lasso Regression model**
  - RMSE: 0.141515
- **Ridge Regression model**
  - RMSE: 0.141509
  - It is the only model that is better than previous Linear Regression Model
  - Alpha parameter of 8.72
- **Elastic Net Regression model**
  - RMSE: 0.141515

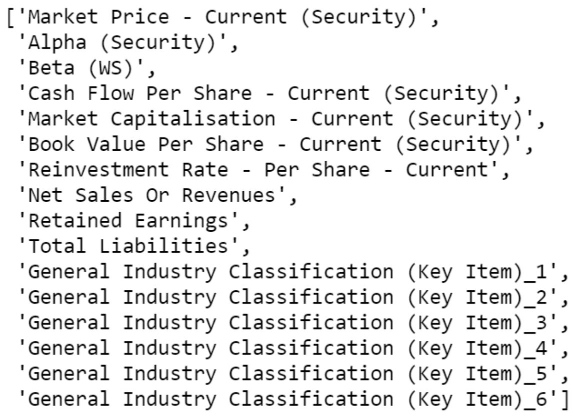
Regression Models

- Random Forest Regression Model
  - RMSE: 0.007314
  - Improves 20 times better than other models
  - Could be a case of overfitting
- In Summary
  - Ridge regression is the best based on our evaluation term of RMSE
  - All other RMSE also only differ slightly

    Could be due to manual variable selection that resulted in a parsimonious model



	RMSE
OLS (Variable Selection)	0.141528
Linear Regression	0.141515
Lasso Regression	0.141515
Ridge Regression	0.141509
ElasticNet Regression	0.141515
Random Forest Regression	0.007314



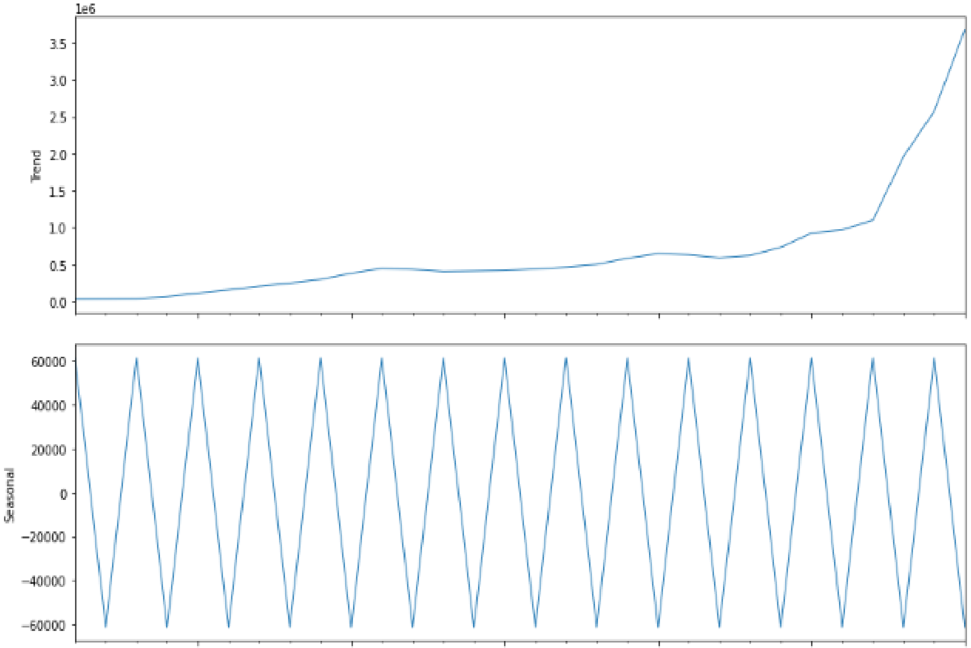
```
['Market Price - Current (Security)',
 'Alpha (Security)',
 'Beta (WS)',
 'Cash Flow Per Share - Current (Security)',
 'Market Capitalisation - Current (Security)',
 'Book Value Per Share - Current (Security)',
 'Reinvestment Rate - Per Share - Current',
 'Net Sales Or Revenues',
 'Retained Earnings',
 'Total Liabilities',
 'General Industry Classification (Key Item)_1',
 'General Industry Classification (Key Item)_2',
 'General Industry Classification (Key Item)_3',
 'General Industry Classification (Key Item)_4',
 'General Industry Classification (Key Item)_5',
 'General Industry Classification (Key Item)_6']
```

Section 3

Model 2 :
Time Series

Methodologies

Time Series Decomposition Plot



Rationale: Providing accurate valuation of the companies

Forecast Variable: Revenue

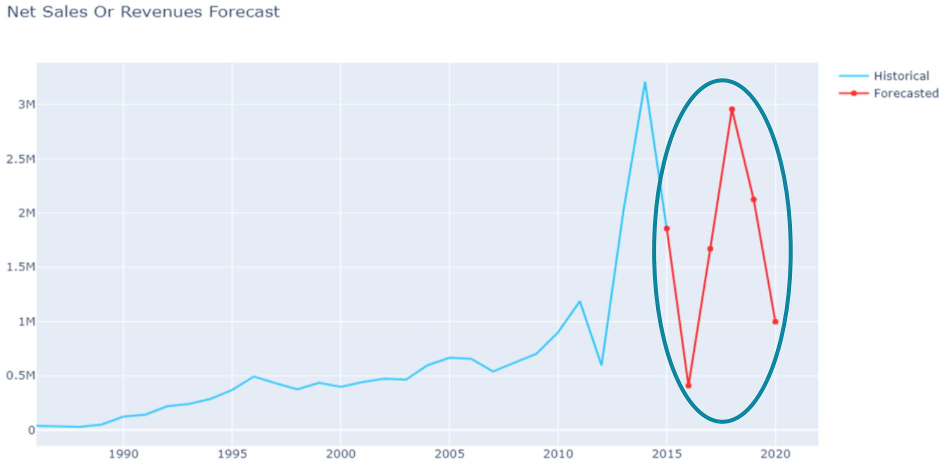
Aim: Revenue projection for the next 5 years

Decomposition Plot

→ Additive trend

Methodologies

SARIMAX Model (Auto ARIMA with Backward Stepwise Selection)



BSS → Drops least significant variables (using p-values)

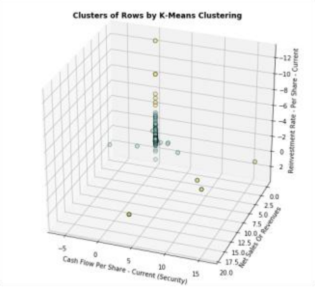
Model Evaluation Criterion: AIC

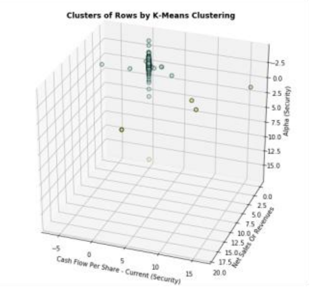
Forecasts	Y1	Y2	Y3	Y4	Y5
Forecasted Revenue ($)	406,000	1,660,000	2,950,000	2,130,000	1,010,000
Residual	192,000				

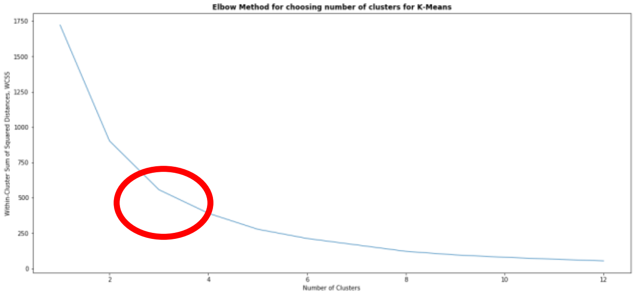
Section 4

Model 3 : Clustering

Clustering Analysis



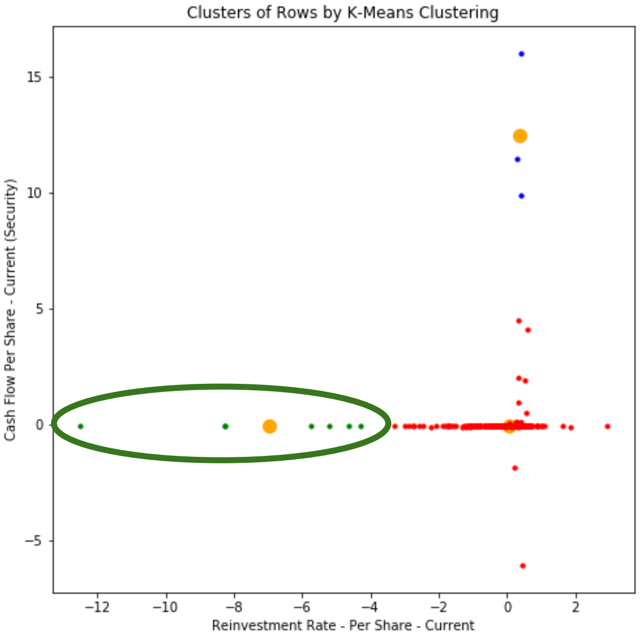




- **K-Means Model**
  - 'k-means++' algorithm
  - Improve initialization of cluster centers
  - Identify more distinct clusters
  - Improve overall clustering

- **Law of Parsimony**
  - Elbow graph method at inflection point
  - Within cluster sum of squares (WCSS)
  - Optimal number of clusters is 3

Cash Flow & Reinvestment Rate



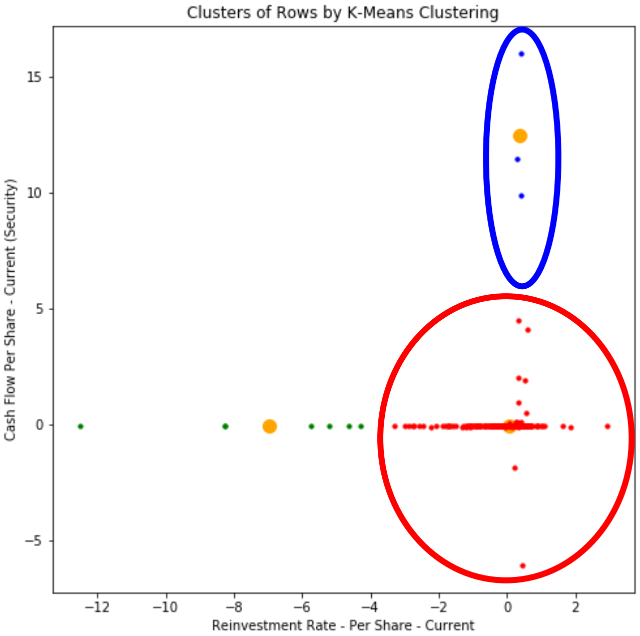
Green Cluster : Startups / SMEs

- Negative reinvestment rates
- Near-zero cash flow per share
- Typical for startups and SMEs

Negative Reinvestment Rates

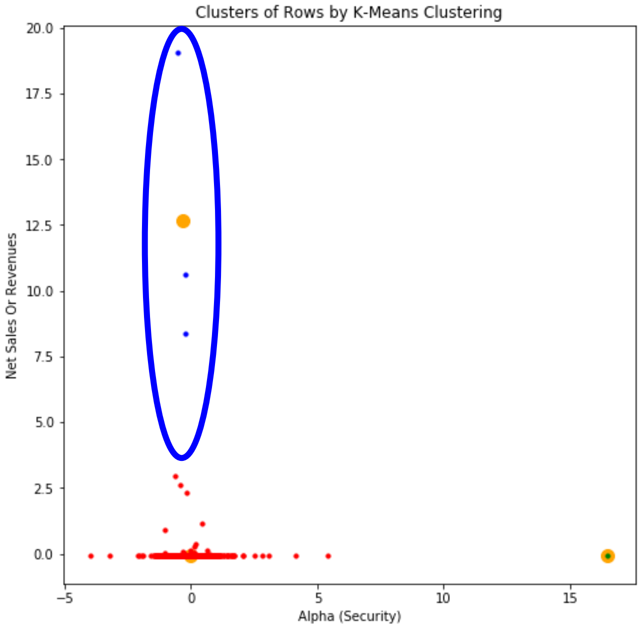
- Temporary lumpy capital expenditures
- Volatile working capital

Cash Flow & Reinvestment Rate



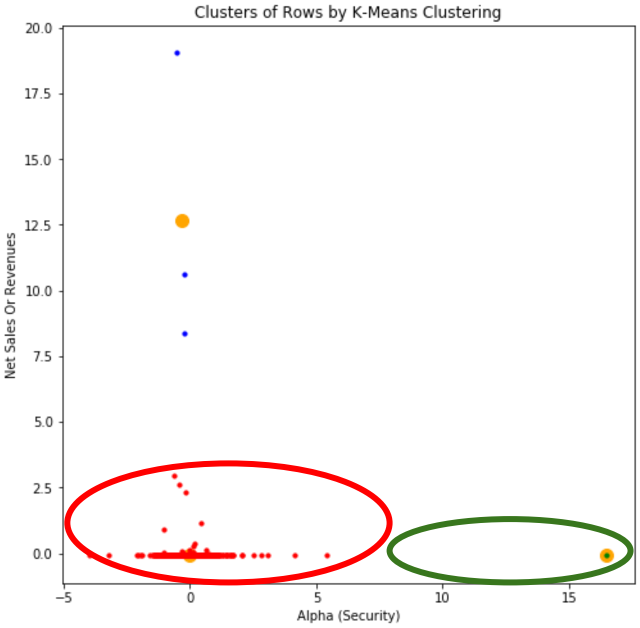
- **Blue Cluster : Large MNCs**
  - Positive stable cash streams
  - Typically do not reinvest much
  - Past the growth stage
- **Red Cluster : Growth / Expansion**
  - Near-zero cash flow per share
  - Reinvestment rate depends on growth strategy

Revenue & Alpha



- **Alpha**
  - Performance against a benchmark
  - Plotted against the company's revenue
- **Blue Cluster**
  - Big cash cow companies
  - Large revenues from stable businesses
  - Near-zero alpha due to accurate valuation and well-established branding

Revenue & Alpha



- **Green Cluster**
  - Anomalous amount of investment returns
  - Possibility of insider trading
- **Red Cluster**
  - Growing startups
  - Do not have very high revenues
  - High alphas to compensate higher risk

Section 5

Further Analysis & Conclusion

Insights

Start-Ups / SMEs

Growing Companies

MNCs

Overall market trend indicates a linearly increase

Revenue Forecasting: SARIMAX

Stock Price Prediction: Ridge Regression

Conclusion

Model does not account for external factors

Global Pandemics

Non-linear growth of technology firms

Conclusion

Important to account for qualitative factors

Company Transparency & Corporate Governance

Financial News

Conclusion

Important to account for qualitative factors

Other Techniques

Natural Language Processing

Thank You