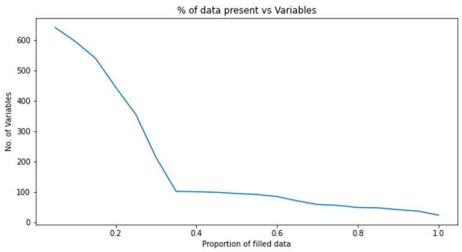
TheDailyProphet E0318558 E0004365 E0191659 E0319030 E0200936

1 Introduction

In this group project, we will attempt to synthesize, break down, analyze, and predict the accounting data from Worldscope that has been provided. Methodologies covered here include Regression, Time Series, and Clustering Models. The aim of this project is to help us understand market trends, different types of company segments, forecast company revenues, as well as predict the stock prices of these companies based on market trends.

2 Data Preprocessing & Visualization

Before further analysis, data preprocessing was performed on the dataset with initial steps to remove empty features, since they do not value-add to our analysis. This leaves us with a dataset of shape (53277, 1247). Before further variable selection, we can see from the graph that there is a tradeoff between long and wide data. Hence, if the features selected were more complete, there would be less features to choose from.



For the purposes of this project, market price is our regression model's target variable, since the ability to accurately predict companies' stock market prices is tangible and invaluable to investors. For our predictors, key items from the companies' financial statements (FS) were manually selected, and financial ratios were excluded since they are derived from multiple FS items, leaving us less choice of features and potentially leading to high correlation with other FS items. Univariate, bivariate and correlation analysis were also performed on our dataset with visualizations to help us understand the patterns and trends within each variable and with each other. After further feature selection using the correlation matrix and a correlation threshold of 0.8, our final feature set for the regression model is as shown above.

```
'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'
```

For our time series dataset, it was prepared separately by identifying a company with the largest number of complete data points for us to perform our analysis. The larger date range allows us to better capture any market trends that might be present over the three decades of data from United Engineers Limited (UEL). For our clustering dataset, it focuses on the Year 2014 since it gives us the greatest number of companies with complete information.

Finally, feature scaling was performed on our dataset to normalize the data before fitting in our regression, time series and clustering models since some features vary greatly in the range of their raw data values (by many magnitudes), which might affect our model performance. A train-test split, with 30% test-set size, was also done on our data to allow out-of-sample cross validation for evaluation of model performance across different models.

3 Model 1: Regression Analysis

	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

For our regression analysis, a total of 6 models were built and evaluated using RMSE as our evaluation metric. Our base linear regression models include SK-Learn's *LinearRegression()* model, as well as statsmodels' *ols()* model with feature selection. Based on the cross-validated predictive performance of these 2 base models, it tells us that the feature selection in *ols()* model does not necessarily bring about an improvement in the model's out-of-sample predictive performance despite a better model fit based on AIC / BIC. Even though model parsimony improves model error by reducing variance, the corresponding tradeoff from increased bias might lead to an overall reduction in predictive power. The difference between the RMSE of the 2 models could also be a result of random selection during the train-test split, with different *random_states* leading to different *ols()* and *LinearRegression()* results.

Building on our base linear regression models, we introduced regularization in the form of Lasso (L_1 Regularization), Ridge (L_2 Regularization) and Elastic Net Regression Models. Regularization aims to prevent overfitting by introducing a penalty term based on the model's feature weights to the model's loss function. This adjusts feature weights towards 0, improving model parsimony while achieving feature selection. From our results, Ridge Regression is the only regularization model that improves on the predictive performance of our previous linear regression model, at an alpha-level of 8.72.

We also performed regression with *RandomForestRegressor()*, which resulted in a model predictive error which is roughly 20 times better than the other models. However, having an extremely low amount of predictive error is indicative of a case of overfitting, which is even more likely in Random Forest models considering that they have a greater tendency to overfit data compared to other models in general.

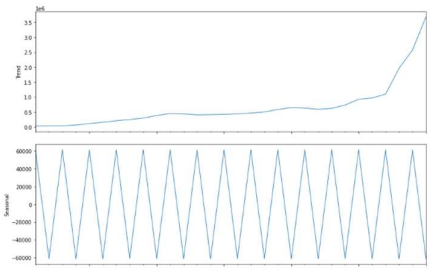
Overall, since Random Forest regression very likely constitutes a case of overfitting, Ridge Regression is our best regression model in terms of predictive power, with an RMSE of 0.142. However, the rest of our linear regression model also exhibits similar levels of predictive error, and this could be due to the fact that manual variable selection has allowed us to narrow down our feature set to one that is parsimonious, yet is well related to and sufficiently explains our target variable.

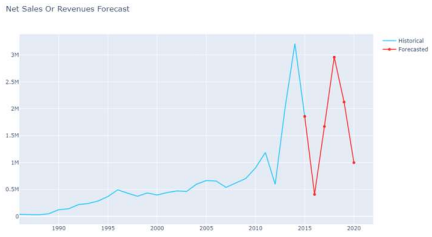
4 Model 2: Time Series Analysis

For our time series analysis, a different set of exogenous variables, with the target forecast variable Revenue, was chosen because revenue projection is important for investors in producing an accurate valuation of companies. Our aim is to provide revenue projection for the next 5 years, which is typical for valuation projections since further projections beyond the

horizon are prone to inaccuracies brought about by unpredictable changes in market conditions like the subprime mortgage crisis in 2008.

We observe from an initial plot of historical revenues for United Engineers Limited that except for the spike in 2014, their revenue has been increasing linearly over time. Hence, time series decomposition was performed with an additive trend. From our decomposition plots on the left, we can see that decomposition on a 2-year basis (freq=2) provides the best patterns of seasonality amongst (1, 2, 5 and 10 years) and hence, the rest of our forecast will be conducted with 2-years seasonality.

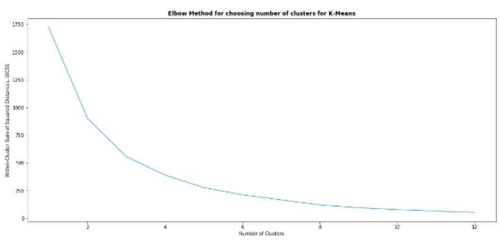


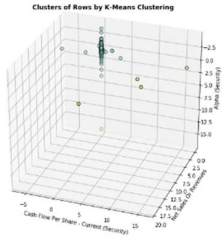


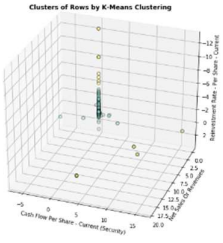
Our choice of model is pmdarima's *auto_arima()*, which helps determine the best (p,d,q) and seasonal order within a given range. Forecasts were done for each exogenous variable independently before being lagged by 1, 2 and 3 periods to prepare our exogenous dataset for revenue forecasting. Our SARIMAX model was then run together with a Backward Stepwise Selection to drop the least significant variables (highest p-values) based on the AIC criterion to maintain model parsimony. The forecasted revenues based on the above model are 406000, 1660000, 2950000, 2130000 and 1010000 for the next 5 periods (as seen in the graph above), with an error of roughly 192000.

5 Model 3: Clustering Analysis

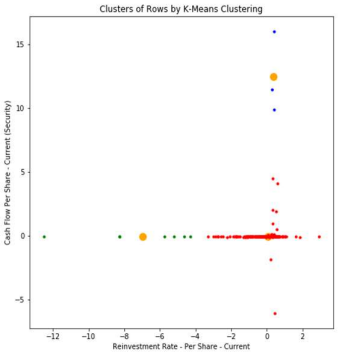
For our clustering analysis, K-Means models were used with the 'k-means++' algorithm to improve the initialization of cluster centers, allowing us to identify more distinct clusters and improve overall clustering. From our elbow graph, based on the within cluster sum of squares (WCSS) and the inflection point, the optimal number of clusters is 3 for all our analyses, and using 2 separate 3D graphs with different triplets of features, we are able to best visualize these clusters below.

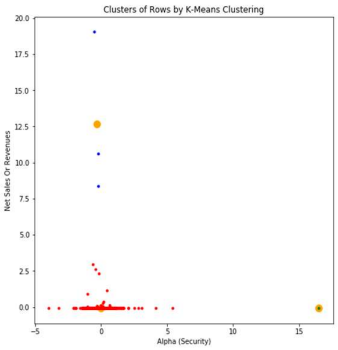






Based on the analysis of reinvestment rate (RR) and cash flow per share (CFPS) on the left, firms in the green cluster with negative RR and near-zero CFPS are likely to be startups and SMEs since negative RR reflects temporary lumpy capital expenditures or volatile working





capital for most firms[1], and is also very typical for startups and SMEs. On the other hand, large MNCs in the mature stage of business tend to have positive stable cash streams and typically do not reinvest as much as other companies since they are past the growth stage, as indicated by the blue cluster. The remaining companies in red are in the growth and expansion stages of business as reflected by their near-zero CFPS and their RR depends on the growth strategy of each individual company.

Our next analysis (shown on the left) is done on the security's Alpha, which represents its performance against a benchmark[2], against the company's Revenue. Aside from the green cluster, which is likely a security with an anomalous amount of investment returns with the possibility of insider trading, our blue cluster is indicative of big cash cow[3] companies with a higher possibility of large revenues from their stable businesses and having near-zero Alpha due to their well-established brands and accurate valuations. On the other hand, companies in the red cluster are either growing or are startups and these companies tend not to have very high revenues but have high Alphas to compensate for the higher risk their investors take.

6 Further Analysis & Conclusion

Through this project, we have managed to distill some insights which we hope can provide budding investors with a basic understanding of companies and the market. In summary, we have identified that there are generally 3 types of companies in the market: Startups / SMEs, growing companies and MNCs. Also, the overall market trend is likely to increase linearly over time (assuming it follows UEL's revenue trend), and the best tools identified for revenue forecasting and stock price prediction are SARIMAX and Ridge Regression respectively, with relatively reasonable predictive errors from their RMSE scores.

With that in mind, the performance of these models is constrained by our feature set, and are unable to account for external factors such as COVID-19, which resulted in global stock market crashes comparable to the 2008 Financial Crisis, and other factors such as the explosive non-linear growth of technology firm securities. When analyzing companies and the market, there are also other important qualitative factors that should be accounted for, such as financial news, company transparency and corporate governance practices, which all require additional techniques not covered here, like Natural Language Processing.

[1] https://pages.stern.nyu.edu/~adamodar/New_Home_Page/valquestions/growth.htm

[2] https://www.cnbc.com/id/45777498

[3] https://www.bcg.com/about/our-history/growth-share-matrix.aspx