UK Metal and Mining Corporate Earning Analysis

Lauren Strutt
Computer Science Department
University of Bristol
Bristol, United Kingdom
ub20693@bristol.ac.uk

Yen Nguyen
Computer Science Department
University of Bristol
Bristol, United Kingdom
xc18328@bristol.ac.uk

Parth Patel
Computer Science Department
University of Bristol
Bristol, United Kingdom
ar20791@bristol.ac.uk

Sanad Sha'ban
Computer Science Department
University of Bristol
Bristol, United Kingdom
py20044@bristol.ac.uk

Victoria Smith García
Engineering Mathematics Department
University of Bristol
Bristol, United Kingdom
lj20127@bristol.ac.uk

Manuel Rodríguez De Guzmán Martínez
Engineering Mathematics Department
University of Bristol
Bristol, United Kingdom
fj20166@bristol.ac.uk

***Abstract*—This report attempts to determine the most significant factors affecting the earnings of companies within the FTSE 350 Metal and Mining. Our analysis was performed using financial and economic data from Bloomberg, which included past data on unemployment, inflation, economic events and the stock market prices of companies within the industry. In order to achieve this, multiple regression, elastic-net regression, XGBoost Regressor and Random Forest Regression were used.**

***Index Terms*—XGBoost, Random Forest, Elastic-Net Regression, Hypothesis Testing, Statistical Regression Analysis**

I. Introduction

The Metal and Mining Industry is an integral part of the UK economy, employing over 230,000 people and directly contributing £10.7 billion to the country's GDP. [1].

Despite comprising of 11,000 companies, the UK Metal and Mining sector is highly concentrated, with Rio Tinto (RIO), Anglo American (AAL) and Antofagasta (ANTO) accounting for the vast majority of the market capitalisation [2], followed by smaller companies like Glencore, BHP and EVRAZ.

Between 2006 and 2015, RIO increased in market value by approximately 30%, while AAL decreased in value by 60%. This differing behaviour occurred despite both companies experiencing the same commodity prices and economic conditions. Hence, all mining companies must "have a clear plan for differential value creation, beyond relying on commodity prices" [3]. This prompts investigation into the important variables that drive the earnings of the main UK metal and mining companies. In this report, we tackle this, focusing on the companies within the FTSE 350 Metal and Mining Index; a weighted index of the 350 UK Metal and Mining companies with the largest market capitalisation [2].

To perform our data analysis, we sought to measure feature importance using Elastic Net Regression, Random Forest, XGBoost and Multiple Regression.

A. Project Goals

This report will provide valuable insights for industry leaders, investors, and policymakers, offering a solid foundation for making informed decisions and understanding the dynamics of the FTSE 350 Metal and Mining Index. Through a rigorous analysis of historical data and relevant economic indicators, we aim to contribute to the development of robust investment strategies and an enhanced understanding of the factors that drive the performance of the UK's Metals and Mining sector. This is of particular importance, since past commodity and equity prices have exceeded expectations, making investments in commodities more attractive as long-term investments [4].

Additionally, our findings will enable effective feature selection for other data scientists. This is paramount since ineffective feature selection can otherwise lead to incomplete information and redundant and noisy features [5]. Feature selection is key in financial time series forecasting; removing redundant features decreases computation time during training and increases model accuracy by reducing overfitting, because there is less noise due to the removal of irrelevant data [6].

On balance, by understanding the most important features which affect the earnings of these companies, we can better predict the companies which will benefit from certain changes in financial or economic conditions and other data scientists can use the findings from our project to train predictive financial models on a reduced subset of features, which are shown to be most important and hence increasing the efficiency of these predictive models.

In order to fulfil these aims, we consider the relative feature importance of a range of variables based on economic and financial using findings from elastic net regression, multiple regression, XGBoost and Random Forest.

B. Process

Throughout our project, we have adhered to Cross-Industry Standard Process for Data Mining (CRISP-DM) [7]. By following CRISP-DM, it provides a systematic and structured approach to the project, ensuring that all relevant aspects are considered. Also, using a standard methodology as CRISP-DM, ensures that results are consistent, reliable, and reproducible.

Fig. 1. CRISP DM [7].

C. Related Work

There is limited previous work on data analysis in determining the most important variables driving the earnings of the main UK Metal and Mining companies. The most similar study, [3], analyses the most important variables affecting the value of mining companies between 2006-2015. Their study focused on 4 international companies, which included Anglo American and Rio Tinto. From the results of their Pearson correlation coefficient analysis and hypothesis testing, MacDiarmid et al. concluded revenue, commodity price and EBITDA multiple (1) were the most important determinants of company value. Their study, however, focused on correlation and did not use modeling to perform feature importance, as this study aims to do. In addition, they have a broader aim in investigating the important factors affecting not just UK metal and mining companies, but metal and mining companies internationally.

$$\text{EBITDA multiple} = \frac{\text{Enterprise Value}}{\text{EBITDA}} \quad (1)$$

Feature selection methods have been analysed in previous studies like the one published by Chih-Fong Tsai, where methods including t-test, correlation matrix, stepwise regression, principle component analysis (PCA) and factor analysis (FA) are compared to solve bankruptcy predictions [8]. In this paper, the predicted model used a Multi-layer perceptron (MLP) Neural Network. The paper analysed five different data sets and they used average accuracy and a number of Type I and II errors in order to evaluate and compare the performance of these methods. Using both these loss measures, their results concluded that the t-test feature selection outperformed the other methods outlined. However, it is important to consider that the paper only focuses on bankruptcy prediction using MLP. Therefore, the difference in data sets between this study and the one outlined in this report, may give rise to different results.

In the realm of stock price direction prediction, the use of ensemble methods has shown promising results. A study reported by Michel Ballings [9] investigates the performance of ensemble methods, such as Random Forest, AdaBoost, and Kernel Factory, and compares them to single classifier models, including Neural Networks, Logistic Regression, Support Vector Machines, and K-Nearest Neighbors. By examining data from 5,767 European companies, their study aimed to predict stock price direction one year ahead using the area under the receiver operating characteristic curve (AUC) as a performance measure. They found that Random Forest was the top-performing algorithm, followed by Support Vector Machines, Kernel Factory, and AdaBoost. This indicates that ensemble methods should be considered in future studies of stock price direction prediction, as they have demonstrated superior performance compared to single classifier models. Thus, incorporating ensemble methods like Random Forest may be helpful to improve prediction accuracy and better understand the factors influencing corporate earnings in the UK Metal and Mining industry.

Other studies, such as the one posed in [10] analyses the most significant indicators affecting stock market trends using feature selection algorithms. In this paper, He Yuqinq et. al. employed Principal Component Analysis (PCA), Genetic Algorithms, and Sequential Forward Search (SFS) to analyse a series of technical indicators commonly used in stock market research. The study uses data from the Shanghai Stock Exchange Composite Index and compares the results obtained from the three algorithms with a previous study. It highlights the strengths and weaknesses of various approaches, offering guidance for future studies in financial market analysis. The paper concludes that PCA is the most reliable and accurate method for this study, while Genetic Algorithm is suggested as a better option for situations with large dimensions due to its ability to leverage randomness. The results may differ in other contexts, such as FTSE 350 stock movements.

II. Data

The raw data we used to explore the variables most important in determining the earnings of the main UK metal and mining companies were accessed via Bloomberg. It was in a tabular form in CSV files, which were categorized into two main directories; economic and financial. The financial data comprised market indices related to the UK metal and mining industry and the UK government. Industry data covered six companies; Rio Tinto, Glencore, Anglo American, Antofagasta, Evraz and BHP. The economic data included major economic indicators like inflation and unemployment, as well as qualitative data, like economic events. In each data set, the data were provided from a wide range of dates in different frequencies. The data within the financial category was provided from January 2000 to the end of January 2022,

TABLE I
EXPLANATION OF VARIABLES CONSIDERED IN DETERMINING EBITDA (EARNINGS BEFORE INTEREST, TAXES, DEPRECIATION, AND AMORTIZATION).

Feature	Description
F3METL_PX_LAST	The last price traded off the FTSE 350 Industrial Metals index.
PX_VOLUME	The number of shares traded per day; reflecting market activity and the interest of investors.
PX_LAST	The last price of a company's stock at the end of a trading day.
PE_RATIO	The ratio of a company's stock price to its earnings per share; commonly used as a valuation metric for companies because it compares value to earnings.
F3METL_PX_VOLUME	Number of shares traded of the FTSE 350 Industrial Metals index, which reflects the confidence of investors in this industry.
F3METL_EBITDA	EBITDA of the FTSE 350 Industrial Metals index.
F3METL_IDX_GEN_EARN	The total earnings of companies in the FTSE 350 Metals index.
SPX_VOLUME	The number of shares traded in Standard and Poor's 500 index, which includes 500 of the largest companies listed on stock exchanges in the USA.
GBPUSD_PX_LAST	The last price of the British pound sterling against the US dollar; important indicator of UK economy.
GUKG10_PX_LAST	The last price of 10-year UK government bonds; considered since an increase lowers interest rates, increasing demand for commodities.
Inflation	The monthly inflation rate in Retail Price Index (RPI).
Unemployment	The percentage of the population without a paid job who are able to work.
BCOMIN_PX_LAST	The last price of the Bloomberg Commodity Index, which tracks the prices of various commodities.
SPX_LAST	The last price of the S&P 500 index.

and this data was provided for every weekday. However, this was not consistent across companies, since for some companies like Glencore, data was made available from a later date. On the other hand, the economic data set date range varied drastically where data was provided from dates ranging from 1991 to 2022.

A. Data Quality and Understanding

Data was well-labeled and mostly had consistent frequencies and formats throughout. Whilst most of the data provided was clean, some columns had missing values. The missing data in most stock indices were assumed to be Missing Completely At Random (MCAR), as there were no discernible patterns or consecutive occurrences [11]. The units for numeric columns were not specified, but are generally consistent and easy to infer. One of the minor inconsistencies we observed was that some columns (e.g.: RIO TINTO EBITDA) start off as values with four decimal places, but the values are rounded to integers later.

We found that the value that was most indicative of the earnings of individual companies was EBITDA, which stands for earnings before interest, taxes, depreciation and amortization [12]. This value had a half-yearly frequency. Setting EBITDA as our target variable, we considered fourteen features to be the potential factors that impact earnings. These features are discussed in table I.

B. Data Exploration

To get an initial understanding of the relationship between the earnings and the chosen features, we plotted several line graphs displaying each feature along with EBITDA over time. Note that two different scales were plotted to account for the different ranges between the features and EBITDA, so that the plots are insightful. The relationship between certain features and EBITDA over some periods was clear from the graphs. For instance, in the middle graph of Figure 2, there seems to be a positive linear relationship between EBITDA and RIO_PX_LAST, which is especially clear in the 2010 to 2022 period.

To investigate this further, we computed the Pearson's correlation coefficient (PCC) between each company's EBITDA and the features. PCC is a measure of the linear correlation of two variables, and is given by:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{2}$$

where $\bar{x}, \bar{y}$ are the sample means for $\boldsymbol{x}, \boldsymbol{y}$. The values of PCC are shown in Figure 3. The high positive coefficient of 0.75 between RIO_PX_LAST and RIO_EBITDA further supports the claim above.

Finally, the graphs allowed for the visual detection of unusual occurrences and outliers, which happen when some variable takes an extremely unusual value [13]. The plots in Figure 2 show that RIO's EBITDA had an unusual dip around the year 2013. As this is the target variable, we first visually checked that this dip was not justified by any of the features. As it did not seem to correspond to an unusual dip or rise in any of the features, this prompted further investigation. We found that this was due to a natural disaster that affected one of RIO's mining sites [14], which verified that this is an outlier that needs to be dealt with before analysis.

Moreover, referring to the bottom graph of Figure 2, there is an extreme spike in PE Ratio in 2017. The spike is over a period of approximately 6 months. This is 100 times greater than the usual value that PE Ratio takes.

III. DATA PREPROCESSING

Having gathered an understanding of the data set and features available, ensuring the data was in a format appropriate to be fed into the models was important. This meant ensuring that data sets were cleaned, making sure the frequency of the data was consistent with the frequency of the target variable and merging similar data sets together.

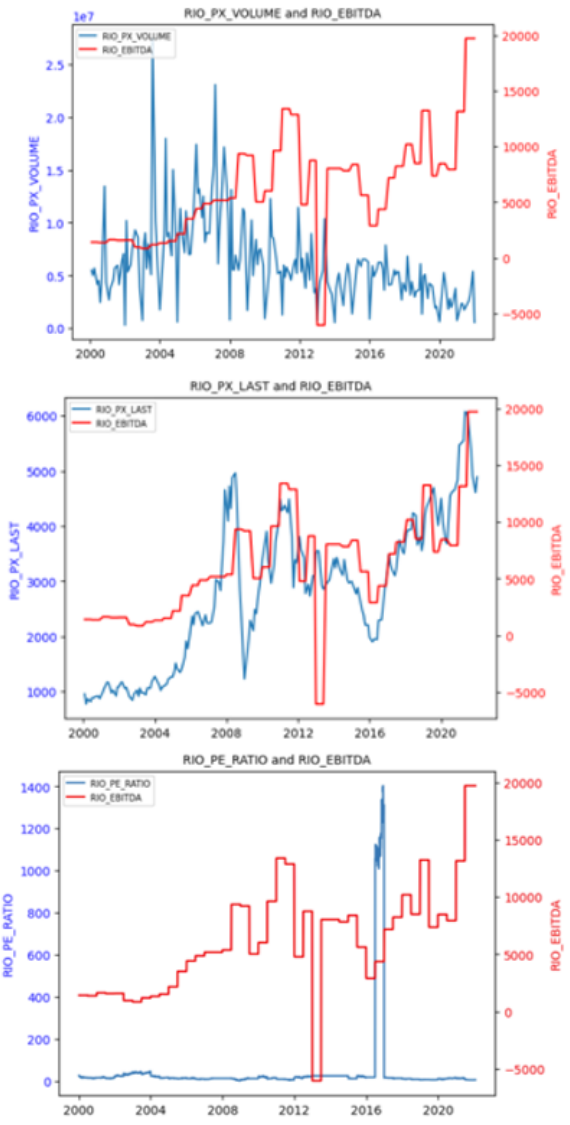

Fig. 2. Graph of RIO_PX_VOLUME(Top), RIO_PX_LAST(Middle) and RIO_PE_RATIO(Bottom) over time alongside RIO_EBITDA

A. Data Aggregation

Despite the data sets containing missing data, these were not replaced with imputed values. The frequency of the data should be consistent with the frequency of the target variable, which was in most cases half-yearly, though with minor inconsistencies. The data for the other features where EBITDA was constant was aggregated to match the frequency of EBITDA. We decided that imputing missing values to then aggregate the data would introduce unjustifiable errors. This was further supported by the observation that the missing values were sporadic throughout the data set and no consecutive pattern was found. Thus, the features were directly aggregated into the desired frequency.

For each of the stock indices, the open, high, low, and close price (OHLC) is provided, which drastically increased the total number of features and dimensionality of our data. Furthermore, for each index, all of these values are highly correlated with each other, and thus the decision to use just the PX_LAST price was made as this is widely used in stock price modelling [15], [16], [17]. This simplified the models and reduced the amount of multicollinearity present within the models. As a result, the total number of features 32 was reduced to 14.

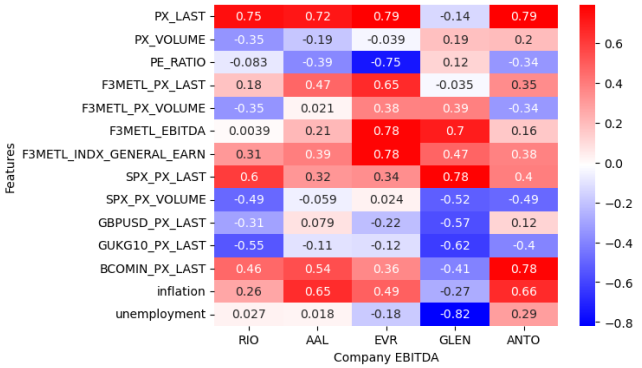

Fig. 3. A heat map illustrating the Pearson correlation matrix between company EBITDA and corresponding features

For the volume of stock traded for the index, the median was an appropriate method. The median is a method that is less sensitive to outliers and thus more suited in this situation.

A cumulative multiplicative total was used to aggregate inflation rates. This feature was provided at the end of each month, where it represented the percentage increase in RPI over that month. In other words, if the value at the end of some months was r, then the RPI increased by $r\%$ during that month. To aggregate n consecutive values $r_1, r_2, ..., r_n$ into one value r, the following formula was used:

$$r = 100 \times ((\prod_{i=1}^{n}(1 + \frac{r_i}{100})) - 1), \tag{3}$$

which is the percentage increase in RPI over the n month period.

B. Outlier Detection

Many models are sensitive to outliers, and thus dealing with these in an appropriate manner is required to enhance the validity of the results.

Having mentioned the outlier in PE_ratio in Data Exploration, this outlier needs to be dealt with to reduce the chance of obtaining inflated results. This outlier can be explained due to a variety of factors, such as the decline in China's economic growth [18], one of RIO's main markets. This was dealt with by using the median over the whole data range.

Another outlier noticed was within RIO_EBITDA as discussed in the Data Exploration section. This was dealt with by using the mean. Despite this method's sensitivity to outliers, the outlier of concern was not that extreme in comparison to the data and hence we considered it appropriate to use the mean.

C. Data Scaling

Features being on different scales can result in giving certain features inflated weighting due to it having a larger variance. This diminishes the interpretability of the results for feature importance. Scaling also allows for robustness to outliers.

All the features are therefore required to be in a consistent scale before modelling and thus a standard scaling method was used. This is done by removing the mean and scaling to unit variance, as outlined in Equation 4, where x is the sample value, μ is the sample mean and σ in the sample standard deviation. This was implemented using the Python function StandardScaler from the sci-kit learn library [19].

$$z = \frac{(x - \mu)}{\sigma} \tag{4}$$

IV. Machine Learning Models

Since the aim of this project is feature importance analysis, the interpretability of the results was one of the main considerations when choosing an appropriate model. The three obstacles we faced with our data was its high dimensionality, its limited sample size due to the frequency of the target variable (EBITDA), and multicollinearity. High multicollinearity occurs when there is significant inter-correlation between the features [20]. This is an issue as many models make the assumption that features are independent. Using models like multiple regression in the presence of multicollinearity leads to issues with interpretability, as the sign and magnitude of some feature's coefficient can be affected by another correlated feature [21]. There are several ways to gauge the degree of multicollinearity within the features. We used the variance inflation factor (VIF), which can be calculated for each feature i as follows:

$$VIF_i = \frac{1}{1 - R_i^2}, \tag{5}$$

where R_i^2 is the R-squared value resulting from regressing feature i on the other features:

$$R_i^2 = 1 - \frac{RSS}{TSS}, \tag{6}$$

where RSS = Residual Sum of Squares, and TSS = Total Sum of Squares [20].The higher the VIF, the greater the effect of multicollinearity. In particular, a VIF greater than 5 indicates high multicollinearity [20]. The VIF values for the features of Rio Tinto are presented in Table II.

TABLE II
VARIANCE INFLATION FACTOR OF THE PREDICTORS OF RIO_EBITDA

Feature	VIF
BCOMIN_PX_LAST	13.655087
GUKG10_PX_LAST	10.520676
F3METL_EBITDA	9.347436
SPX_PX_LAST	8.398919
RIO_PX_LAST	7.945560
GBPUSD_PX_LAST	7.333783
SPX_PX_VOLUME	7.039644
F3METL_PX_LAST	5.429277
F3METL_INDX_GENERAL_EARN	4.724539
RIO_PX_VOLUME	4.596148
F3METL_PX_VOLUME	3.430198
Unemployment	2.207091
Inflation	1.732675
RIO_PE_RATIO	0.639271

We found several methods in the literature to mitigate the effect of multicollinearity. However, the issue was that some of them would hinder interpretability. Methods such as PCA and combining correlated features [22] were deemed inappropriate, as we aimed to investigate the effect of each feature on earnings.

In light of the aforementioned challenges, we opted for four appropriate models. The key idea was that if different models emphasised the same feature, that would imply that the feature has a significant relationship to company earnings. The first two were tree-based ensemble methods, namely XGBoost and Random Forest. Tree-based models are known to be interpretable even in the presence of multicollinearity, as they do not make the independence assumption [23]. Moreover, ensemble tree-based methods have a high capability in handling mixed data, outliers, collinearities, and can therefore cope effectively with more complex data [24].

The third model was Elastic Net Regression, which is a regularised linear regression model. Due to the regularisation term within their loss functions, regularised regression models act as embedded feature selection methods, by pushing the coefficients of insignificant features to zero. Two common regularized linear regression models are Ridge, which uses an L1-norm penalty term, and Lasso which uses an L2-norm penalty term. As it combines the two norms, Elastic Net was found to outperform both Ridge and Lasso in high-dimensional settings where multicollinearity is present [25]. Moreover, it has a grouping effect in correlated features [26], meaning that highly correlated features tend to be selected together or left out together. This is desirable in our case, because if a significant feature is highly correlated with another feature, then both could affect earnings.

We intended to carry out statistical significance analysis for the Elastic Net Regression model, however, we found that this is a particularly involved task for the regularized version [27]. Therefore, we carried out hypothesis testing on the unregularized multiple regression model which forms our 4^{th} model. Despite this model being sensitive to multicollinearity,

inference about feature importance could still be made with caution.

Finally, all of the models considered in this project have a history of being successful within financial studies such as [28], [29]. Results show that no method excels for all problems and every method gives different insights into the data [30].

A. Models

1) Elastic Net Regression: EN was proposed in [26], and is implemented within Scikit-Learn python library which we used. The model aims to solve the following optimization problem:

$$\underset{\boldsymbol{\omega}}{\operatorname{argmin}}\Big\{\frac{1}{2n}\|\boldsymbol{X}\boldsymbol{\omega}-\boldsymbol{y}\|_2^2+\alpha\rho\|\boldsymbol{\omega}\|_1+\frac{\alpha(1-\rho)}{2}\|\boldsymbol{\omega}\|_2^2\Big\}$$

where $\boldsymbol{X}$ is the feature matrix, $\boldsymbol{y}$ is the target variable (EBITDA), n is sample size, $\boldsymbol{\omega}$ is the vector of coefficients, and ρ and α are hyperparameters. Writing $\boldsymbol{\omega}^*$ as the minimiser of the objective function above, the fitted model $\hat{f}$ would then be:

$$\hat{f}(\boldsymbol{x})=(\boldsymbol{\omega}^*)^T\cdot\boldsymbol{x}$$

The hyperparameters were tuned by combining grid search and Leave-One-Out-Cross-Validation(LOOCV) as in [26]. Grid search involves testing combinations of ρ and α and choosing the combination that minimises the cross-validation error measured by Mean-Squared-Error. LOOCV was found to be appropriate due to the small sample size [31]. This is where all but one data point is used to train the model, and then the model is tested on this left-out data point. This is repeated for all data points, and then the results are averaged out to return a final value for a specific hyper-parameter set. The hyper-parameter set that performs the best is returned. Finally, feature importance in this model is made based on the magnitude of the estimated coefficients, with higher magnitude implying more significance [32].

2) XGBoost: XGBoost, (eXtreme gradient boosting) is an ensemble method that is known for its speed, its excellent performance [33], and is a non-linear modelling technique. The base learners are decision trees, which are added to the ensemble iteratively such that the errors induced by previous learners are corrected by future learners. Recently created learners put a greater emphasis on data points that previous learners performed weakly on, and thus to improve any weak points within the ensemble. One benefit of XGboost is that it has the ability to handle missing data which was a prominent issue with our data set. [34]

XGBoost is commonly used for predictive analysis, however, it can return a feature importance score which makes it an appropriate model for this use case. The feature importance score from the model reflects how much a single feature contributes to the model, and a higher score indicates that the feature is more important. This intuitive interpretability provides another benefit of choosing XGBoost.

Within the grid search, the following hyperparameters were used: the max depth of a tree, the number of trees in the ensemble, the learning rate, the regularization terms, and the γ term which determines the threshold for further splitting of a leaf node in a tree to occur. Despite the model having more hyperparameters, it was necessary to tune those parameters that seemed appropriate to our small data set size. XGBoost was implemented through the XGBoost open source Python library.

3) Random Forests: Random forest is also an ensemble method that uses decision trees as its base learners. However, the trees are constructed independently as opposed to iteratively in XGBoost. Two concepts that random forest covers are Bagging and random feature selection at nodes. Each individual base learner is independently trained on a random subset of data [35]. When splitting at a node, a random subset of features are considered for the split, which helps trees to be less correlated with each other, allowing to deal with multicollinearity, noise in the data and providing diversity within the ensemble which reduces overfitting.

Feature importance score is also an output from the Random Forest model like XGBoost, and thus the ease of interpretability makes this model appropriate for the use case.

Similar to XGBoost, the hyperparameters that were used within the grid search to find the optimal hyperparameter setting includes the number of trees and the maximum depth of each tree. This was implemented using scikit-learn.

4) Multiple Regression and Hypothesis Testing: This involves fitting a simple regression model:

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\omega}+\boldsymbol{\epsilon}$$

where $\boldsymbol{y}$ is the output vector, $\boldsymbol{X}$ is the feature matrix and $\boldsymbol{\omega}$ is the fitted coefficients, and $\boldsymbol{\epsilon}$ is vector of residuals [36]. The coefficients $\boldsymbol{\omega}$ are obtained by minimising the Ordinary Least Squares (OLS) function:

$$\underset{\boldsymbol{\omega}}{\operatorname{argmin}}\sum_{i=1}^{n}(y_i-(\boldsymbol{\omega}^T\boldsymbol{x}_i))^2$$

where x_i is the i^{th} input data and y_i is i^{th} data point of the target variable.

When all the features are scaled feature importance is typically inferred by the magnitude of the coefficients, or by applying hypothesis testing. The Statsmodels library in Python conducts a t-test hypothesis test on the regression model fitted. Following from [37], the null ($\boldsymbol{H}_0$) and alternate ($\boldsymbol{H}_1$) hypotheses are as follows:
For feature i,

$$\boldsymbol{H}_0:\omega_i=0$$

$$\boldsymbol{H}_1:\omega_i\neq 0$$

where ω_i is the true coefficient of the i^{th} feature in the multiple regression model. The t-test verifies whether the feature being considered has a significant effect on the multiple regression model or not. A multiple regression model is fitted under the null hypothesis, assuming that the i^{th} feature has no effect on the model. The t-test statistic is then used to determine how extreme an observation is deemed to be under this null

hypothesis, as an extreme result observed would suggest that $\boldsymbol{H}_0$ is not appropriate and representative of the situation.

The t-test statistic is defined as the following:

$$t = \frac{\hat{\omega}_i - 0}{S(\hat{\omega}_i)}, \tag{7}$$

where $S(\hat{\omega}_i)$ is the standard error of the i^{th} estimated coefficient [36].

One can calculate the p-value which is mathematically formulated as $\mathbb{P}(T > |t|)$. This is automatically returned by the Statsmodel library, allowing for easy analysis and evaluation of the results. If the p-value falls below a certain user-set threshold, which is known as the significance level α, this represents that the observation is extreme under the null hypothesis and thus the interpretation is that the feature is interpreted as important for this model. However, with the presence of multicollinearity, the coefficients can be inflated and so may the p-values. However, a relationship in the following direction holds. If a feature has a significant p-value, this implies that this feature is a significant feature from a linear model perspective [38]. Note, that the other direction doesn't necessarily hold due to multicollinearity discussed above.

TABLE III
TABLE OF RESULTING P-VALUES WHEN USING MULTIPLE REGRESSION ON RIO TINTO, ANGLO AMERICAN AND ANTOFAGASTA, WHERE P≤0.1 ARE COLOUR CODED IN GREEN

Feature	p-value		
	RIO	AAL	ANTO
PX_LAST	0.524	0.056	0.469
PX_VOLUME	0.145	0.048	0.943
PE_RATIO	0.448	0.056	0.521
F3METL_PX_LAST	0.15	0.146	0.02
F3METL_PX_VOLUME	0.995	0.046	0.571
F3METL_EBITDA	0.099	0.178	0.087
F3METL_INDX_GENERAL_EARN	0.102	0.424	0.120
BCOMIN_PX_LAST	0.220	0.341	0.733
SPX_PX_LAST	0.024	0.112	0.066
SPX_PX_VOLUME	0.158	0.637	0.845
GBPUSD_PX_LAST	0.684	0.750	0.011
GUKG10_PX_LAST	0.822	0.545	0.142
inflation	0.024	0.420	0.015
unemployment	0.051	0.111	0.182

V. EVALUATION OF FIT OF MODELS AND LAG-SHIFT

After running the data set on the models mentioned in the previous section, the fit of the models was visualized. One interesting thing to note from the top graph of Figure 4 is that there seems to be a lag between the fitted line from the model and true EBITDA. The intuitive reasoning behind that was that the effect of the features did not take place immediately, but affected the earnings in the next period.

To test this further, the features used within the models were all shifted by one to account for this observation. The last value of EBITDA was removed to account for this shift. The model before the lag achieved an R^2 score of 0.614 and the R^2 after the shift was 0.755, indicating that implementing that lag resulted in a model that fits the data better. This

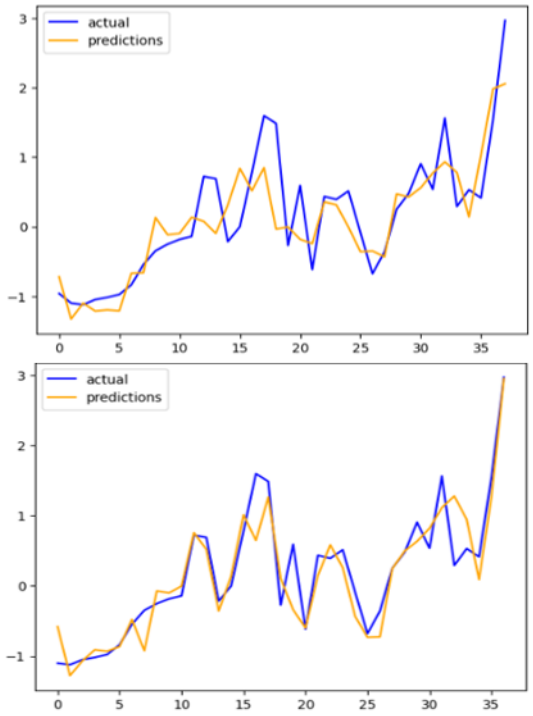

Fig. 4. Two graphs showing the fit of multiple regression model before(Top) and after(Bottom) shifting predictor features by lag of 1

provides a more appropriate fit, and was investigated, verified and implemented in other models.

VI. ANALYSIS OF RESULTS

Overall, as expected, all models had different feature importance results. This is due to different models having different methodologies and assumptions as previously mentioned. To analyse the results we have accumulated, we have to identify which features are deemed important by multiple models and examine the correlation between features and EBITDA to identify if there is a strong association between a feature and the company's earnings. We consider features with a p-value of $p < 0.1$ for the Multiple Regression model to be statistically significant as in [39].

For the 3 biggest companies in the UK Metal and Mining industry, the companies' PX_LAST value was deemed to be an important feature by Elastic Net Regression, XGBoost and Random Forest, ranking relatively highly. Something to note is that the p-value results are inconsistent with this hypothesis about PX_LAST. This can be explained due to high collinearity for PX_LAST as in II, and thus resulting in inflated p-values. A company's stock price can represent investors' confidence in the company, which can result in higher investment from investors. This extra investment can be invested into the company which can result in increased earnings.

From Figure 5, 6 and 7, it can be seen that SPX_LAST is highly ranked. These are all huge companies and thus

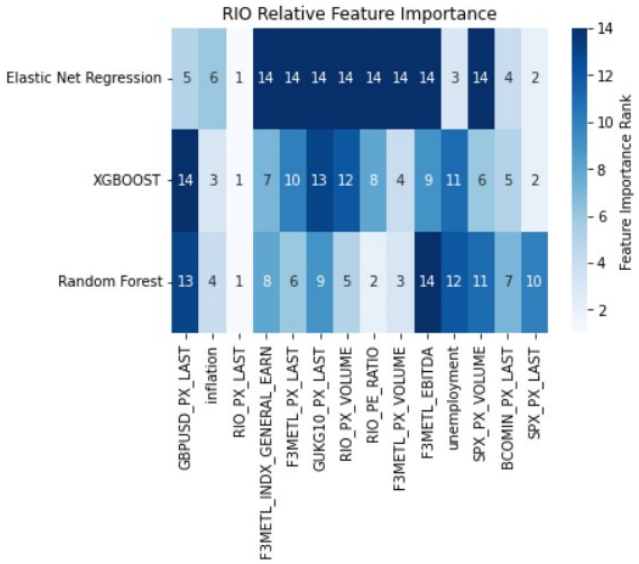

Fig. 5. RIO Relative feature importance results

would trade internationally especially RIO. This is most likely due to the United States being their second largest revenue-generating region, generating 15.9% of its revenue share [40], and being listed in the New York Stock Exchange [41]. Similar to Rio Tinto, Anglo American is a leading global company with many mining operations and projects in the U.S.A. and SPX was highly ranked by the EN model. Antofagasta has significant ties with the United States through its subsidiary company 'Twin Metals Minnesota', which is headquartered in Minnesota, and looking to open new operational copper and nickel mines around the state of Minnesota [42].

Inflation can be interpreted to be an important feature from the results collected from our models. Inflation was more important for RIO and ANTO as indicated by their higher relative ranking and also being statistically significant according to table III. Inflation is an important macroeconomic factor that affects the operations costs of a company. Bigger companies will have more complex operations and thus are more susceptible to inflation rates.

EVR and GLEN seem to be affected by the F3METL index as indicated by F3METL_EBITDA and F3METL_LAST appearing frequently. GLEN outputs the 3 major metals that F3METL index tracks; Aluminium, Copper and Zinc. Thus, if the index is performing well, this indicates there will be a high demand for them and thus will result in greater earnings.

It can be seen that bigger companies are affected by factors such as SPX_LAST, PX_LAST, and inflation whereas the "relatively" smaller companies are affected by the F3METL index. This suggests that bigger companies are less affected by the F3METL index, as it can be thought that they control the movements of the market as opposed to smaller companies that would more likely follow the movements of the market.

VII. Summary and Discussion

The aim of this project was to determine the features most important in driving the earnings of the UK metal and mining companies. There have been a wide variety of models applied to similar areas; results show that no method excels for all problems and every method gives different insights into the data [30], hence we opted to use a combination of models.

Although our project successfully identified the relative feature importance for four different models, it is important to acknowledge the limitations of our findings. One limitation concerned the frequency of EBITDA. Matching the frequency of the features with the frequency of EBITDA meant we were unable to detect short-term volatility. For example, Rio Tinto said the relaxation of covid-19 restrictions would induce high volatility in the short term as a result of shortages of labour and disruptions to supply chains [43], which our analysis would not be able to determine. Moreover, the frequency of EBITDA imposed a very small sample size, which causes overfitting, affecting the reliability of our models and analysis. Using target variables that are less-obvious than EBITDA but more frequent could result in more reliable results. Our methodology would still apply regardless of the target variable used to indicate earnings. Moreover, the features considered were influenced by what was provided in the dataset as opposed to what seemed most appropriate. Our methodology can adapt easily to all the limitations mentioned above, and thus can be used by these companies who have access to more frequent earnings, as well as a wider range of features. This versatility was one of our considerations when deciding our modelling methodology. Another limitation of our methodology was using the Pearson correlation measure, since this does not account for time lags (which we saw profound evidence for in Figure 4) and it can only be used to identify linearly correlated variables. A more appropriate technique for measuring correlation in future work is to use cross-correlation, which accounts for lags and has been used in [44] and [45].

Another potential limitation is the reliability of the data given; since in the recent past, Glencore, Anglo American and Rio Tinto all faced fines by misleading investors [46], [47], [48]. Therefore, in analysing our findings, it should be clear that the data on companies may not be completely accurate, since companies wish to claim to have higher earnings to inflate their share price. However, the data must be accurate in order for us as data scientists to make accurate results. Thus, in this project, we aim to be completely transparent about where our data comes from to allow the reader to make an informed decision about the potential limitations of the data.

References

[1] "A foundation industry." https://www.ukmetalscouncil.org/industry-history, 2023.
[2] M. Garside, "Market capitalization of leading mining companies in the united kingdom (uk) in march 2023," *Statista*, 2023.
[3] J. MacDiarmid, T. Tholana, and C. Musingwini, "Analysis of key value drivers for major mining companies for the period 2006–2015," *Resources Policy*, vol. 56, pp. 16–30, 2018. Special issue on Mineral Economics research at Wits: Papers in honor of Richard Minnitt.

[4] S. Baurens, "Valuation of metals and mining companies," *Immobilien-wirtschaft Aktuell*, 2010.

[5] D. P. Williams, V. Myers, and M. S. Silvious, "Mine classification with imbalanced data," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 3, pp. 528–532, 2009.

[6] R. G. Pajares, J. M. Benítez, and G. S. Palmero, "Feature selection for time series forecasting: A case study," in *2008 Eighth International Conference on Hybrid Intelligent Systems*, pp. 555–560, 2008.

[7] R. Wirth and J. Hipp, "Crisp-dm: Towards a standard process model for data mining," *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 01 2000.

[8] C.-F. Tsai, "Feature selection in bankruptcy prediction," *Knowledge-Based Systems*, vol. 22, no. 2, pp. 120–127, 2009.

[9] M. Ballings, D. Van den Poel, N. Hespeels, and R. Gryp, "Evaluating multiple classifiers for stock price direction prediction," *Expert Systems with Applications*, vol. 42, no. 20, pp. 7046–7056, 2015.

[10] Y. He, K. Fataliyev, and L. Wang, "Feature selection for stock market analysis," in *Neural Information Processing* (M. Lee, A. Hirose, Z.-G. Hou, and R. M. Kil, eds.), (Berlin, Heidelberg), pp. 737–744, Springer Berlin Heidelberg, 2013.

[11] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: a review," May 2010.

[12] J. Grant and L. Parker, *EBITDA!* Elsevier science, 2002.

[13] C. C. Aggarwal, *Outlier Analysis*. Springer Nature, Jan 2013.

[14] D. Petley, "The 2013 bingham canyon mine failure: Insights into the giant manefay landslide," May 2017.

[15] S. Mehtab and J. Sen, "Stock price prediction using convolutional neural networks on a multivariate time series," Aug. 2021.

[16] G. D. et. al., "Artificial neural network models for forecasting stock price index in the bombay stock exchange," 2021.

[17] Y. Du, "Application and analysis of forecasting stock price index based on combination of arima model and bp neural network," 2017.

[18] "China gdp growth rate 1961-2023."

[19] scikit learn, "sklearn.preprocessing.standardscaler specification," 2022.

[20] N. Shrestha, "Detecting multicollinearity in regression analysis," *American Journal of Applied Mathematics and Statistics*, vol. 8, no. 2, pp. 39–42, 2020.

[21] R. K. Paul, "Multicollinearity : Causes , effects and remedies," 2008.

[22] J. I. Daoud, "Multicollinearity and regression analysis," in *Journal of Physics: Conference Series*, vol. 949, p. 012009, IOP Publishing, 2017.

[23] S. Chowdhury, Y. Lin, B. Liaw, and L. Kerby, "Evaluation of tree based regression over multiple linear regression for non-normally distributed data in battery performance," 2021.

[24] D.-S. Cao, J.-H. Huang, Y.-Z. Liang, Q.-S. Xu, and L.-X. Zhang, "Tree-based ensemble methods and their applications in analytical chemistry," *TrAC Trends in Analytical Chemistry*, 2012.

[25] T. Sirimongkolkasem and R. Drikvandi, "On regularisation methods for analysis of high dimensional data," *Annals of Data Science*, 12 2019.

[26] H. Zou and T. Hastie, "Regularization and Variable Selection Via the Elastic Net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, pp. 301–320, 03 2005.

[27] S. Zhao, D. Witten, and A. Shojaie, "In Defense of the Indefensible: A Very Naïve Approach to High-Dimensional Inference," *Statistical Science*, vol. 36, no. 4, pp. 562 – 577, 2021.

[28] S. Du, D. Hao, and X. Li, "Research on stock forecasting based on random forest," in *2022 IEEE 2nd International Conference on Data Science and Computer Application (ICDSCA)*, pp. 301–305, 2022.

[29] M. A. Muslim and Y. Dasril, "Company bankruptcy prediction framework based on the most influential features using xgboost and stacking ensemble learning," *International Journal of Electrical and Computer Engineering*, vol. 11, pp. 5549–5557, 2021.

[30] P. Schmude, "Feature selection in multiple linear regression problems with fewer samples than features," in *Bioinformatics and Biomedical Engineering: 5th International Work-Conference, IWBBIO 2017, Granada, Spain, April 26–28, 2017, Proceedings, Part I 5*, pp. 85–95, Springer, 2017.

[31] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[32] R. Suchting, E. T. Hébert, P. Ma, D. E. Kendzor, and M. S. Businelle, "Using Elastic Net Penalized Cox Proportional Hazards Regression to Identify Predictors of Imminent Smoking Lapse," *Nicotine & Tobacco Research*, vol. 21, pp. 173–179, 09 2017.

[33] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[34] Z. E. Aydin and Z. K. Ozturk, "Performance analysis of xgboost classifier with missing data," *Manchester Journal of Artificial Intelastic-clligelasticncelastic and Applielasticd Scielasticnces (MJAIAS)*, vol. 2, no. 02, p. 2021, 2021.

[35] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[36] R. J. Freund, W. J. Wilson, and P. Sa, *Regression analysis*. Elsevier, 2006.

[37] G. Kończak, "On testing the significance of the coefficients in the multiple regression analysis," *Acta Universitatis Lodziensis. Folia Oeconomica*, no. 269, 2012.

[38] F. E. Harrell, "Regression modeling strategies," *Bios*, vol. 330, no. 2018, p. 14, 2017.

[39] M. Xu, "A study on the correlation between financial status of listed companies and chinese stock market: Base on multiple linear regression analysis," in *2021 2nd International Conference on Big Data Economy and Information Management (BDEIM)*, pp. 340–343, 2021.

[40] S. R. Department, "Rio tinto - statistics & facts," *Statista*, 2023.

[41] "Rio tinto adr each rep 1 ord rio." https://www.nyse.com/quote/XNYS:RIO.

[42] "The metals we'll mine are the metals you use." https://www.twin-metals.com/.

[43] "Ftse 350 review: Is discipline holding back mining returns?." https://www.investorschronicle.co.uk/ideas/2023/02/02/ftse-350-review-is-discipline-holding-back-mining-returns/, 2023. Accessed: May 7, 2023.

[44] L. Kullmann, J. Kertész, and K. Kaski, "Time-dependent cross-correlations between different stock returns: A directed network of influence," *Phys. Rev. E*, vol. 66, p. 026125, Aug 2002.

[45] B. Podobnik, D. Wang, D. Horvatic, I. Grosse, and H. E. Stanley, "Time-lag cross-correlations in collective phenomena," *Europhysics Letters*, vol. 90, p. 68001, jun 2010.

[46] O. Gill, "Anglo american faces attack over 'worthless' coal mines." https://www.telegraph.co.uk/business/2021/06/05/anglo-american-faces-attack-worthless-coal-mines/, Jun 2021.

[47] R. Pat Sweet, "Fca mines rio tinto for record £27m fine over disclosure failings." https://www.accountancydaily.co/fca-mines-rio-tinto-record-ps27m-fine-over-disclosure-failings#:~:text=FTSE%20100%20mining%20giant%20Rio%20Tinto%20has%20been,the%20company%20and%20its%20former%20CEO%20and%20CFO.

[48] "Glencore entered guilty pleas to foreign bribery and market manipulation schemes." https://www.justice.gov/opa/pr/glencore-entered-guilty-pleas-foreign-bribery-and-market-manipulation-schemes#:~:text=Under%20the%20terms%20of%20the%20plea%20agreement%2C%20which,forfeiture%20and%20disgorgement%20in%20the%20amount%20of%20%24272%2C185%2C792, Jan 2023.

VIII. Appendices

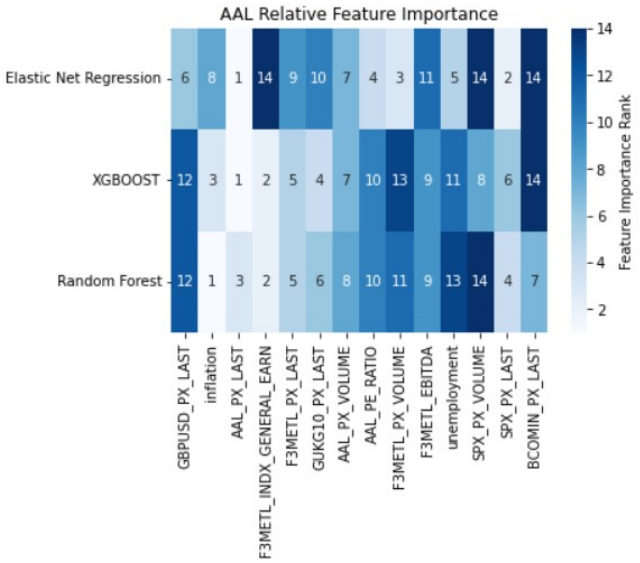

Fig. 6. AAL Relative feature importance

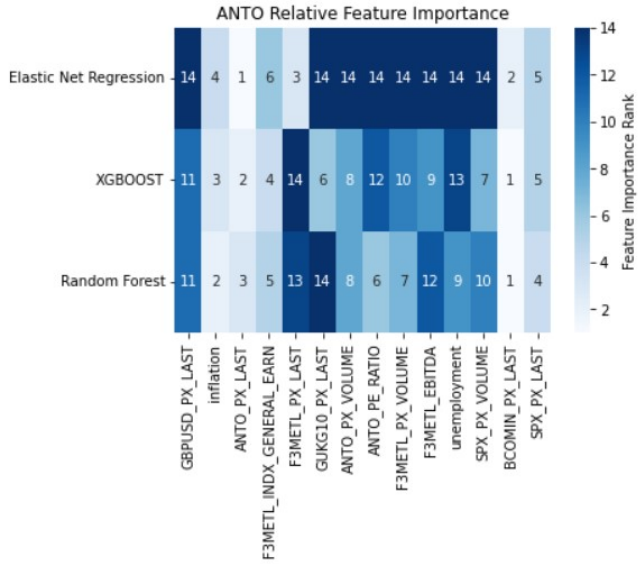

Fig. 7. ANTO Relative feature importance

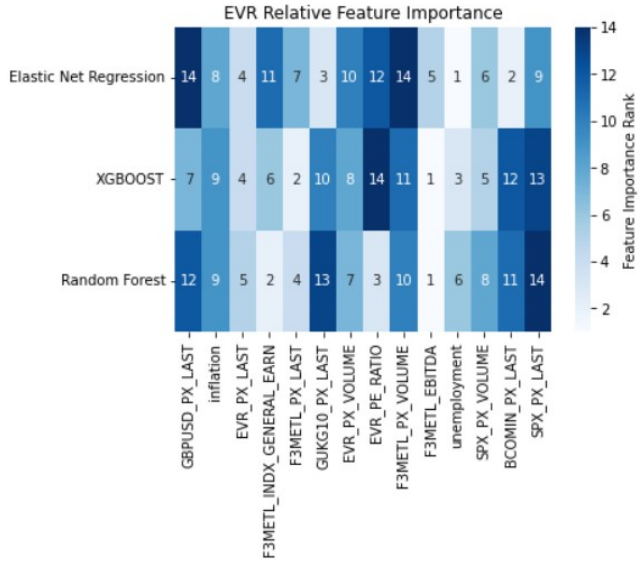

Fig. 8. EVR Relative feature importance

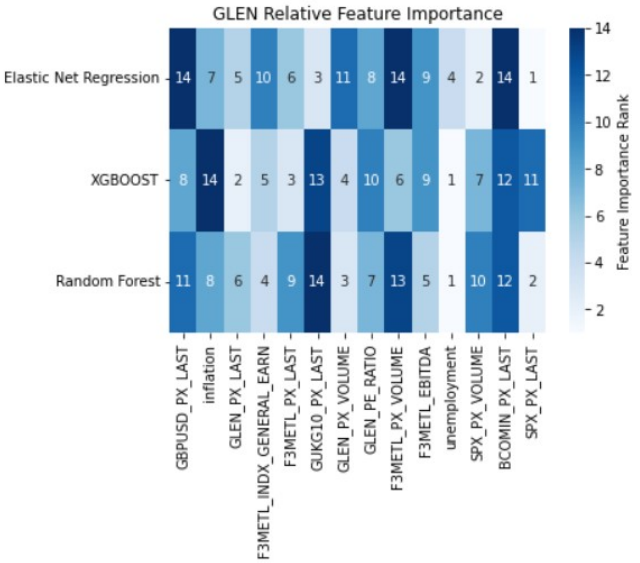

Fig. 9. GLEN Relative feature importance

Individual Reflection

Project 16: UK Metal and Mining Corporate Earning Analysis

Manuel Rodriguez de Guzman Martinez
Computer Science Department
University of Bristol
Bristol, United Kingdom
fj20166@bristol.ac.uk

***Abstract*— This individual reflective discussion describes my role, achievements, and learnings from the group project. It will describe the dynamics of the group, the adversities faced, and the work input to complete the project.**

I. My Role In the group

During this project, I played an active role in my group as a team member. I participated in group discussions and attended all the meetings the group had. Furthermore, I met with some members of the group individually to get the work done. I had some tasks which were individual, and others which were collaborating with another team member. I brought forward innovative ideas and previous work done and ensured that our project remained on track.

In the beginning, we decided to individually explore the dataset. After the first week, individual exploratory analysis had been done. In the following two weeks, we cleaned imputed and aggregated the datasets. We decided as a group on which datasets to work with. Then, we set tasks to be completed individually. I focused on previous studies, researching previous feature selection methods, feature analysis and predictive models in the stocks and mining sector. I gathered all the information in a Word document, updating it each time progress was made. I shared my advancements with the group in our regular weekly meetings. After the models were done, I focussed on the explanation and analysis of the models: why outliers happened, the interpretability of the results and improvements.

In the presentation, we worked through the CRISP-DM standard which suited perfectly as there are six components to it and we are 6 group members. We divided the task, and it was agreed I was doing the sections on further improvements, challenges, deployment, and evaluation.

For the final report, each member wrote the part they worked on. I wrote the introduction and literature review. Furthermore, I wrote collaboratively with Lauren the section on Data and Data Preprocessing. After all the sections were imputed, we as a team went through it. Proofreading, adding editing, and correcting each other's work. I also worked on referencing, helping others to correctly do so. Once each member was happy with the final report, we finished it.

II. Achievements

A. Effective Teamwork Work Balance and Technical Ability

One of my key achievements during this project was my ability to effectively communicate with my group members. This allowed us to collaborate well and achieve our goals efficiently. I learned that effective communication is crucial when working on a group project. By establishing good communication, through meetings in teams and in person, having a WhatsApp group chat and having a drive to share our work, we could work collectively and individually well.

By having a clear workflow, we distributed the work that needed to be done efficiently. As people had different preferences for each task, the work was distributed according to their preferences to maximise the individual contribution to the team. We managed to successfully deliver a report in the analysis of features in the dataset. We develop and demonstrate models used with justifications, and completed the analysis, and visualisation for our results.

III. Challenges Faced

One major challenge the group faced came from the easter break. As we had an international group (including me) easter break meant going back home which temporarily slowed our progress. However, we managed to overcome this challenge by setting (although I had a -7-hour difference) more realistic targets. By doing so, we continued working efficiently. Also, challenges faced by our data were high dimensionality, its limited sample size due to the frequency of the target variable (EBITDA), and multicollinearity. We managed to overcome the problems and finally produce the required results.

IV. Outcomes of the course

A. Evolution and Linking to Course

I developed an understanding of how to be able to work with an initial dataset, explore it (visualisations and understanding), pre-process it (aggregate, outliers, scaling etc..), fit models to the dataset and finally conclude with results. This process has helped me increase my understanding of how a real-life problem can be dealt with in practical data science. This project related to the ideas taught in the unit in several ways. Firstly, I practised data acquisition and ensured ethical considerations and privacy on the dataset. Secondly, I managed and wrangled data, which was necessary to fit the desired model to produce results. Finally, created visualisations, which were helpful to present our results to a non-technical audience and to explore the initial dataset. Also, presented our findings and overall work as a report which helps a non-technical reader understand how we came up with our findings. In addition, it gave me hands-on experience in conducting research and analysing data, which reinforced my knowledge of research methods and techniques.

V. Conclusion

Overall, I am extremely happy with the dynamics of my group. We had clear communication, great distribution of tasks and regular deadlines. These three things were key to handling the project and completing it. This project taught me the value of teamwork and the importance of effective communication in achieving shared goals. I also learned valuable research skills and gained experience in working on a real-world project. I feel that this project has prepared me well for future professional endeavours.