Can we predict the date given the price, volume, opening price, lowest bid, ask price and closing bid of stock data from AT&T?

Candidate Number: 239106

December 2022

1 Introduction

1.1 Main Objectives

ATT are a public telecommunication company based in America. The dataset details the stock profile over a 20 year period. The main objective is to understand whether the date of a stock profile can be predicted from the price history and trading volume details. The objectives of analysis is to produce a model with the best accuracy when making predictions and to find the best data transformations to perform on the dataset to produce the best results. Another objective is to locate the best hyperparameters for the chosen models to produce the best results. The final objective is to find the most effective method with regards to analysing the results produced by the machine learning models.

1.2 Previous Findings

There is a substantial amount of previous research using regression on stock data. However, most of this research is concerned with trying to predict the price of a stock given other independent features. Bhuriya highlight a method using linear regression to predict the price [1]. The methodology used in this experiment will be similar of that to predict the date. Date is a independent feature in Bhuriya's dataset and price is dependant. When predicting the date, these features will be swapped. Although it is difficult to find research produced on date prediction with regards to stock data, there are regression methods published to predict the date in other industries. Sharma highlights the various regression methods that can be used such as linear and XGB (Extreme Gradient Boosting) regression to predict the date a customer is likely to purchase an item [2]. The results of Sharma's experiment found that XGB regression was more effective than linear regression. Linear regression is one of the methods which will be used in the proposed experiment so this is therefore a possible snag as it may not give the most accurate prediction.

2 Methodology

2.1 Technical Features

The AT&T dataset has 5387 instances and 29 features. Each instance in the dataset represents a different date with regards to the stock information. The date ranges from 03/01/2000 to 31/12/2020. Many of the features are redundant having minimal variation and many repeated values which is unlikely to improve regression models. Once these features have been removed the ones remaining are:

- DATE - The calendar date (Target Variable)
- PRC - Price
- VOL - Volume
- OPENPRC - Opening Price
- ASKHI - Ask or High Price
- BIDLO - Bid or Lowest Bid
- BID - Closing Bid

2.2 Methodology

2.2.1 Data Transformation

Out of the 29 features mentioned above, it appears that only the six refereed to can be used to predict the data. It was apparent that the BID column had 24 missing values; to resolve this, the instances were removed. This was decided as the data set is large enough that this should not effect the performance of the regression models. The DATE column also needed transforming; as it was in form of YYYYMMDD. The stock market is not open on the weekend therefore there were also many gaps in the dates; it was therefore decided that the dates would be swapped with integers beginning at zero making the data more continuous. Finally, the data will be standardized, normalized and transformed using principle component analysis(PCA) with n equal to 2 which is expected to effect the performance of the models.

2.2.2 Data Visualisation

Figure 1 to 6 highlights any correlation between the date and the predictor variables. It is clear that there is not a strong correlation between the variables other than in Figure 6 where there is a weak, slightly positive correlation.

2.2.3 Modelling

There are three regression models that have been chosen: Linear, Ridge and XGB Regression. Linear regression is a method which reduces an error function. The error function is the absolute distance between the data points and the line of best fit. Ridge regression is similar, however prevents overfitting by increasing bias and reducing variance using a function which penalises larger coefficients. L2 regularization is used which shrinks coefficients however they will never reach zero.
XGB regression [3] is a type of gradient boosting algorithm which is more efficient and scalable. XGB regression is robust to outliers and overfitting which makes it ideal for stock data which can have large fluctuations [4].
Linear regression does not have any hyperparameters. In ridge regression, the alpha value controls the effect of the penalty function. XGB regression has many hyperparameters however for the depth value which specifies the maximum depth of decision trees is going to be varied for the experiment. A tree with too many levels is prone to overfitting. When modelling, k-fold cross validation will be used to increase reliability and repeatability.

2.2.4 Analysis

R^2 and adjusted r^2 values will be used to measure of correlation between the predicted and expected value. A value of one means the predictor and target variables are perfectly correlated; a value of 0.5 means that 50% of the dependant variable is predicted by the framework. Adjusted r^2 is similar, however it is also effected by the number of independent variables.
Residual values will also be used to explore the difference between the expected values vs the predicted values.
SHAP (shapley additive explanation) values will be used to better understand the models. This is because they are easier to interpret than the coefficients as coefficients depend of the nature of the feature they are associated with.
Pearson Correlation Coefficient will be used to analyse the correlation between the predictor and target variables. Pearson Correlation Coefficient produces a value between zero and one to indicate correlation.

3 Results

3.1 Hyperparameter Tuning

Figures 7 and 8 highlights the r^2 values with different hyperperameters. For ridge regression, the lower the alpha value, the lower the r^2. However, due to a ridge regression model with a value of zero for alpha being equivalent to linear regression, it was decided to keep alpha at one. For XGB regression, a max tree depth of three is optimal for training. The default value for XGB regression is six suggesting that the model would be overfitted. All further experiments will be conducted with these hyperparameters.

3.2 R^2 and Adjusted R^2 Analysis

Figures 9 and 10 show the r^2 and the adjusted r^2 results. It is clear that XGB regression with the untransformed data provides the best results. Overall, the standardized data and the untransformed data both produce similar results for all models. The normalized data performs worse than the standardized and untransformed data over all regression models.

3.3 Principle Component Analysis (PCA)

In figures 9 and 10 the results of the r^2 and the adjusted r^2 using PCA are also displayed. For XGB regression, the loss of r^2 and adjusted r^2 using PCA is minimal. This means that if a similar experiment was to be conducted on a larger dataset,

one may want to use PCA to reduce the compute time required as the loss of accuracy may be worth the extra time required for training.

3.4 Residual Value Analysis

From the residual values shown in figures 11, 12 and 13, it is clear that all models are neither making consistent over or under predictions due to approximately equal variation on either side of zero. The figures further highlight that XGB is the better model and that the normalized data provides the worst results.

3.5 SHAP Value Analysis

Figures 14, 15 and 16 show the SHAP values using standardized data. The charts highlight how each model works differently with different SHAP values corresponding with each feature. For ridge and XGB regression ASKHI and BIDLO are the two most important features however with linear regression it is BID and PRICE. This is particularly fascinating as ridge and linear regression both produced very similar r^2 values however clearly got there using very different methods.

3.6 Pearson Correlation Coefficient

Figure 17 shows the results of Pearson's Correlation Coefficient. Contrasting with the SHAP values, it suggests that the volume of the stock has the highest effect on the predicted date. This was unexpected as the SHAP values ranked volume third or lower. From further research[5], it appears that the reason for this inconsistency is due to SHAP values focusing on how the independent variables effect the result of the predicted variable. Pearsons Correlation Coefficient simply produces a value based on how correlated two variables are.

3.7 Results Conclusion

The results explained above are fairly accurate with high best-case r^2 values, over 0.7 for the best models. The residual values highlight the level of accuracy with XGB regression always producing the tightest spread. Overall, the results are explainable however there are some outcomes such as the normalized data having dramatically worse results which were not expected. It appears the reason for this is simply due to the data not normalizing well and not being highly representative of the dataset.

4 Discussion

4.1 Key Findings

The results presented, show that it is possible to use regression to predict the date given the stock profile which was one of the objectives. Hyperparameter tuning is required to ensure that models are optimal as the default values were not optimal for the dataset. Throughout the experiment, it was clear that r^2 and adjusted r^2 were the correct analysis method allowing a clear comparison between models. There was minimal loss for r^2 vs adjusted r^2 with a mean difference of -0.002 suggesting that the features were correctly selected in the methodology. For PCA with the XGB model, there was minimal difference between the r^2 and adjusted r^2 suggesting that both the PCA features are equally key to the performance of the model. The residual values are a good indicator of accuracy highlighting the difference in performance between the models and different data formats. In conclusion, it is clear that the XGB model with untransformed data was the best.

4.2 Next Steps

For next steps in analysing the data, it could be interesting to test the models with stock data from the future. If the models still perform well, they could be used as an investing tool predict the date when a stock would hit a certain price or volume. The models could also be run the including stock profiles that predate 2000's which could result in greater accuracy. ATT went public in 1901 highlighting the volume of data that could be used to train the models. Training using other model types could also result in greater r^2 values such as using neural networks.

5 Appendix

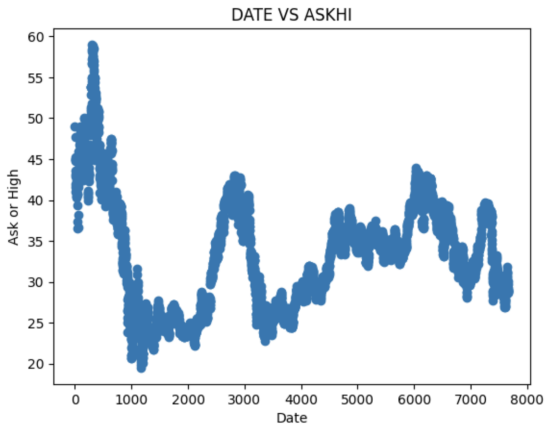

Figure 1: The relationship between the date and the asking/high price of the stock

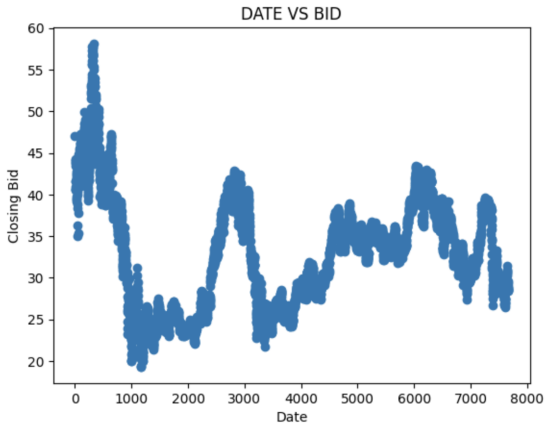

Figure 2: The relationship between the date and the closing bid of the stock

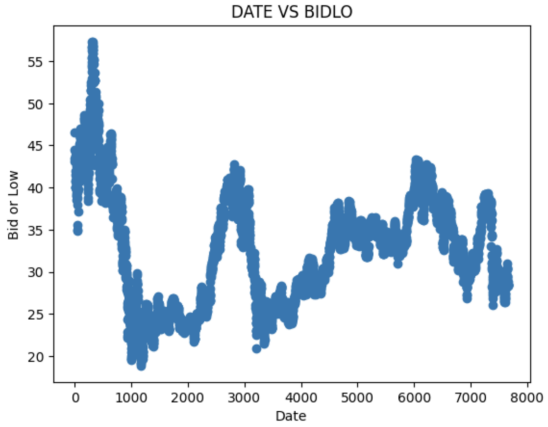

Figure 3: The relationship between the date and the closing/lowest bid of the stock

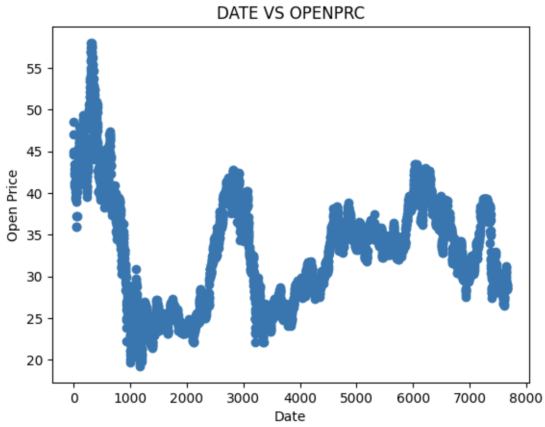

Figure 4: The relationship between the date and the opening price of the stock

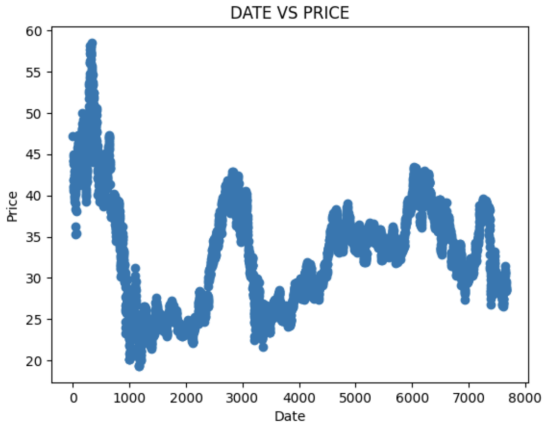

Figure 5: The relationship between the date and the price of the stock

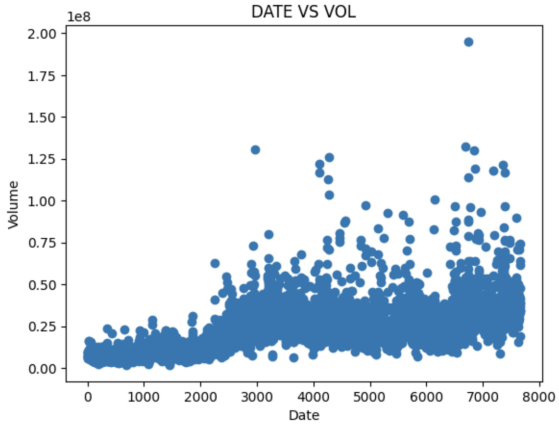

Figure 6: The relationship between the date and the volume of the stock

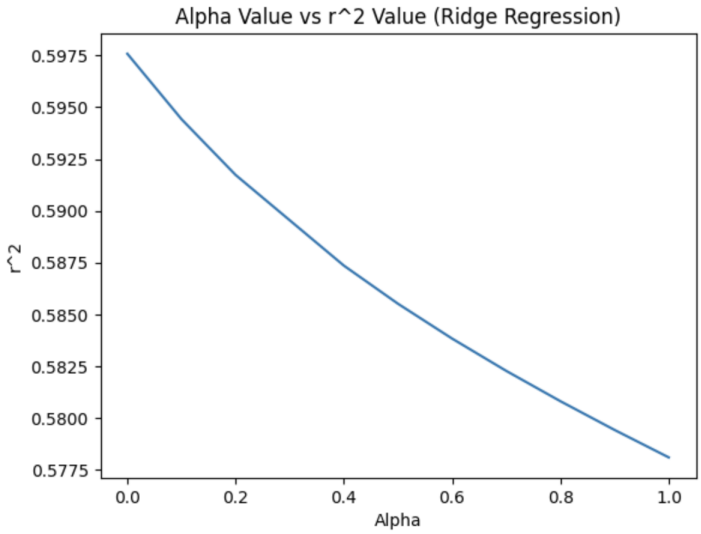

Figure 7: Hyperparameter tuning on ridge regression using standardized data

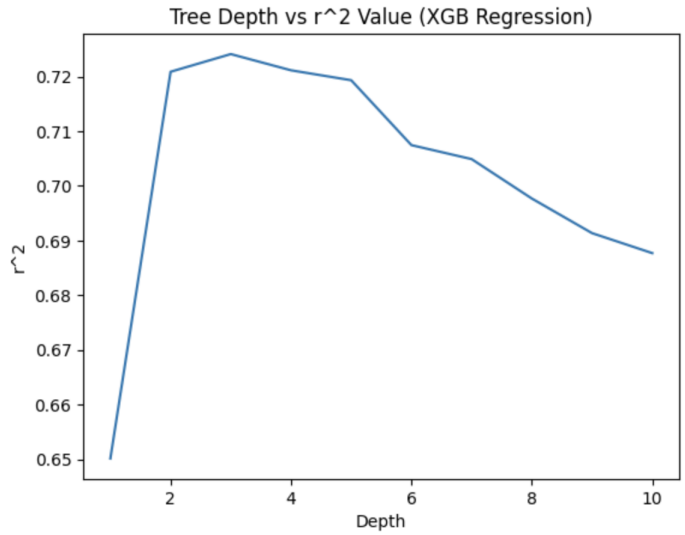

Figure 8: Hyperparameter tuning on XGB regression using standardized data

	Data Format			
Regression Model	Untransformed Data	Normalized Data	Standardized Data	PCA Standardized Data
Linear Regression	0.590	0.470	0.589	0.349
Ridge Regression	0.590	0.001	0.581	0.349
XGB Regression	0.738	0.480	0.724	0.712

Figure 9: The mean r^2 values of the different data formats using k-fold cross validation with each regression model rounded to 3 D.P

	Data Format			
Regression Model	Untransformed Data	Normalized Data	Standardized Data	PCA Standardized Data
Linear Regression	0.588	0.468	0.587	0.347
Ridge Regression	0.588	-0.005	0.579	0.347
XGB Regression	0.737	0.478	0.723	0.712

Figure 10: The mean adjusted r^2 values of the different data formats using k-fold cross validation with each regression model rounded to 3 D.P

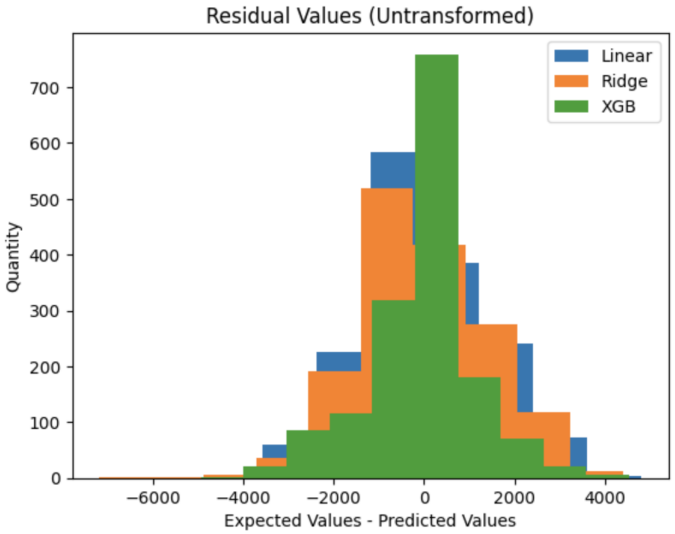

Figure 11: The quantity of residual values for a sample model using untransformed data

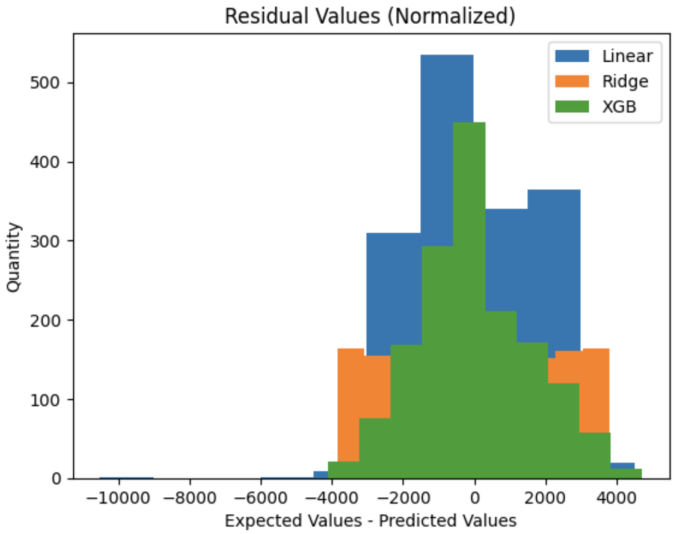

Figure 12: The quantity of residual values for a sample model using normalized data

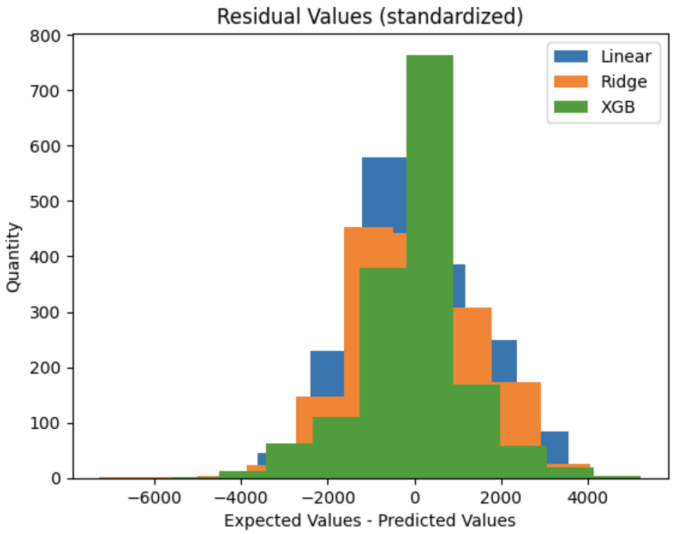

Figure 13: The quantity of residual values for a sample model using standardized data

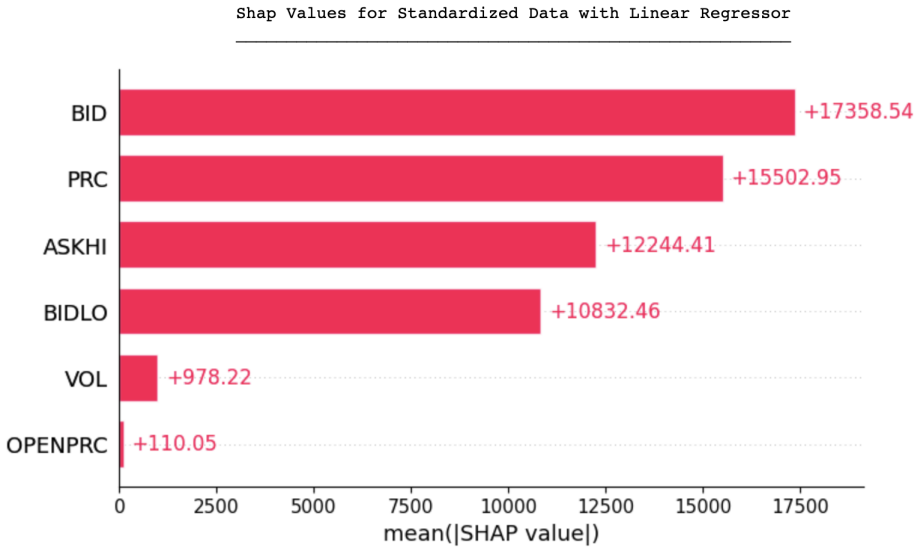

Figure 14: The SHAP values for a sample model using standardized data with a linear regressor

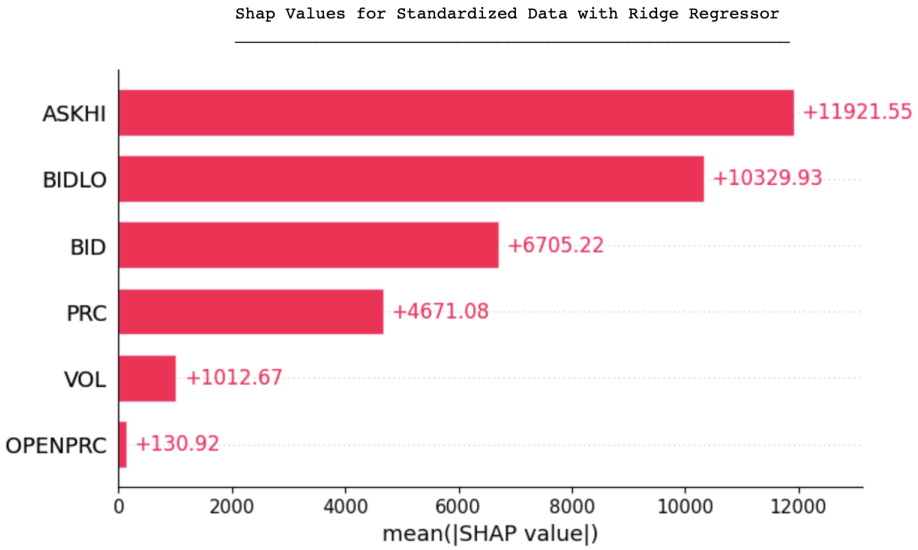

Figure 15: The SHAP values for a sample model using standardized data with a ridge regressor

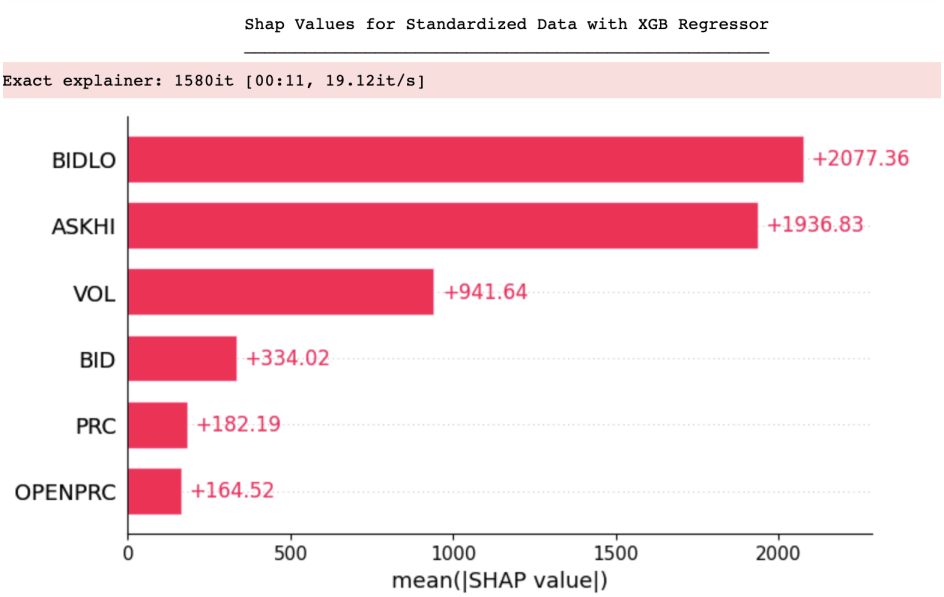

Figure 16: The SHAP values for a sample model using standardized data with a XGB regressor

	Independant Features					
Dependant Feature	PRC	VOL	OPENPRC	ASKHI	BIDLO	BID
Date	0.07	0.58	0.07	0.05	0.08	0.07

Figure 17: The Pearson Correlation Coefficients for the dataframe

References

[1] D. Bhuriya, G. Kaushal, A. Sharma, and U. Singh, "Stock market predication using a linear regression," in *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, vol. 2, 2017, pp. 510–513.

[2] A. Sharma, P. Randhawa, and H. F. Alharbi, "Statistical and machine learning approaches to predict the next purchase date: A review," in *2022 4th International Conference on Applied Automation and Industrial Diagnostics (ICAAID)*, vol. 1, 2022, pp. 1–7.

[3] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen *et al.*, "Xgboost: extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[4] X. Zhang, C. Yan, C. Gao, B. A. Malin, and Y. Chen, "Predicting missing values in medical data via xgboost regression," *Journal of healthcare informatics research*, vol. 4, no. 4, pp. 383–394, 2020.

[5] C. Molnar, *9.6 SHAP (SHapley Additive exPlanations)*. Christoph Molnar, 2022.