[Student Paper]

Financial Forecasting, Planning and Analysis Using Machine Learning

Priyal Jhunjhunwala*

Department of Computer Science and Engineering, University of Westminster, London, UK

Abstract

This paper studies multiple linear regression and time series regression: two regression methods, the task of which is to learn the relationship between various components of financial statements and the profits earned. Time series regression is used to understand and predict the behavior of dynamic systems such as the modeling and forecasting of economic, financial, biological, and engineering systems. For the machine learning methods that have been used to conduct this study, we have used a dataset that was created by using information provided by the luxury brands themselves to maintain accuracy and authenticity of data. This will help in providing and gaining accurate information. Three machine learning models; simple linear regression, multiple linear regression and times series regression were employed for conducting this study where we try to predict future values based on the historical data provided to us. The time series regression model then has various models which have been used after comparing and determining the best suited model according to the component that we have tried to predict. Our study adds to the existing literature by providing insights into which models are better suited for the type of predictions and the components when related to each other and time. Specifically, we find that all the brands in the fashion industry had their lowest financial performance in the year 2020 when compared with the data of the past 10 years but have also managed to recuperate and flourish since. The findings of the study have important implications for investors and fashion enthusiasts who want to explore different investment opportunities other than merchandise purchasing. These individuals can use the results given to make investments decisions which can be worth a large sum in monetary terms.

Keywords: Regression; Efficiency; Machine Learning

ARK: https://n2t.net/ark:/47543/JISCOM2023.v4i1.a32
Received August 20, 2023, accepted September 16, 2023, date of publication October 10, 2023

1. Introduction

The fashion industry has become one of the largest industries with millions of people all over the world taking an interest in their fashion choices and appearances. The existence of the industry has become of core importance in our day to day lives as people from all over the world depend on the creators, brands and even influencers to make decisions about what to wear and what to purchase. This industry not only provides the world with the most basic requirement being clothes but also provides a large amount of job opportunities to an array of people and provides us with a platform to express our creativity in ways where it could be found in any wardrobe in the world.

This industry is worth trillions of dollars with the large variety of merchandises that the labels and brands in this industry produce ranging from the smallest of jewellery to perfumes to statement clothing pieces. However, what many are not familiar with is that they can choose to invest their finances into this industry in ways other than simply just purchasing products. There are a lot of people who are interested in the fashion industry but not exactly enthusiastic about spending their finances on the products and then there are several people who are looking for investment opportunities but forget to consider that this industry is one of the largest industries with a net worth of 1.7 trillion USD as of 2022 which also makes it one of the best industries to invest into.

* Corresponding author.
E-mail address: priyal.jhunjhunwala@gmail.com

The dataset used for this study has been created using the financial information these luxury brands have announced over the years in order to maintain accuracy and authenticity so that the results that have been derived do not cause any harm to the people who use them to make their investment decisions. The dataset contains information on the various financial components such as revenue, cost of sales, gross margin, operating profit, profit after tax, eps and more. Each of these components have been predicted using their own historical data and then the relationship between the different components have been used to determine the dependent variable and then predict it.

In this study, three machine learning algorithms have been applied to the dataset, namely, simple linear regression, multiple linear regression, and time series regression to forecast future values. Time series regression was found to be the more suited model for such forecasting purposes where we are using historical data. The study explores and implements the named regression methods on the provided dataset and provides discussion and results accordingly.

2. Related works

For conducting this study numerous preexisting research papers were referred to. They have been listed below as follows:

Stock Trend Prediction Using Regression Analysis – A Data Mining Approach by S Abdulsalam Sulaiman Olaniyi, Adewole, Kayode S., Jimoh, R. G: This paper demonstrates and talks about using linear regression analysis with least squares method in order to forecast stock prices and moving averages to reduce error. The paper "Time series extrinsic regression Predicting numeric values from time series data" authored by Chang Wei Tan, Christoph Bergmeir, François Petitjean, and Geoffrey I. Webb studies the Time Series Extrinsic Regression method which aims to learn the relationship between time series and a continuous scalar variable. In the paper Fundamental Analysis and the Prediction of Earnings, the authors have tried to predict future earnings using statistical and contextual analysis.

3. Methods and Analysis

In this study three regression methods have been used, namely, Simple Linear Regression, Multiple Linear Regression, and Time Series Regression. Simple Linear Regression is a regression method where we have one independent variable and one dependent variable. The dataset is divided into training set and a test set where the training set is used to train the model. Here, we used this method along with the least squares method to reduce errors. Multiple Linear Regression is a regression method in which there are more than one independent variable and only one dependent variable. The dataset is divided into training set and test set where the training set is used to train the model.

Time Series Regression is a regression method that can help us understand and predict the behaviour of dynamic systems from observational or experimental data. It can be used for modelling and forecasting of economic, financial, biological, and engineering systems. This method has been used along with moving averages.

To conduct this study, we first import the necessary libraries of NumPy, pandas, matplotlib and seaborn. Then, we have tried to predict revenue based on historical data using simple linear regression using Sklearn libraries. We first perform train test split on the dataset and initialize the regression model. Then we perform linear regression to predict the values of the test set using least squares method to get the values. It is seen that even with least squares method, the error margin is big between the predicted values and the actual values.

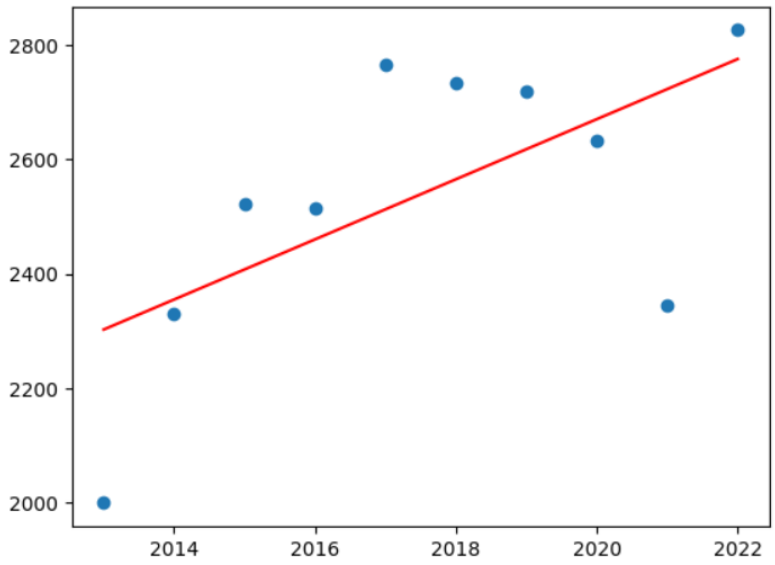

Fig 1

Next, we have tried performing the prediction of profit after taxation based on the other attributes of the dataset which affect the profit after tax of an organization over the years. This prediction has been done using multiple linear regression method where attributes such as revenue, cost of sales, Cost of sales, operating expenses, adjusting operating items, net finance (charge)/credit, and taxation collectively affect the profit of the organization. First, we drop the unnecessary attributes from the dataset, then perform one hot encoding method to make the processing of data easier and faster. After encoding, we perform train test split on the dataset and initialize the regression method. Upon applying regression to the test set to predict the values for certain years, we obtain the below given result.

```
[[313.92 341.1 ]
 [427.76 376.  ]
 [261.44 287.7 ]
 [299.08 397.  ]
 [311.82 332.3 ]]
```

Fig 2

In Fig 2, the values on the left-hand side are the predicted values and the values on the right-hand side are the actual values. The difference between the actual and predicted values is not that large for most cases.

After the simple and multiple linear regression methods, we have used Time Series regression method [7] considering that we are trying to forecast future values for our dataset based on historical data and time series. First, we use this regression method for forecasting revenue values for the next 5 years with the application of moving averages. The resultant graph can be seen below.

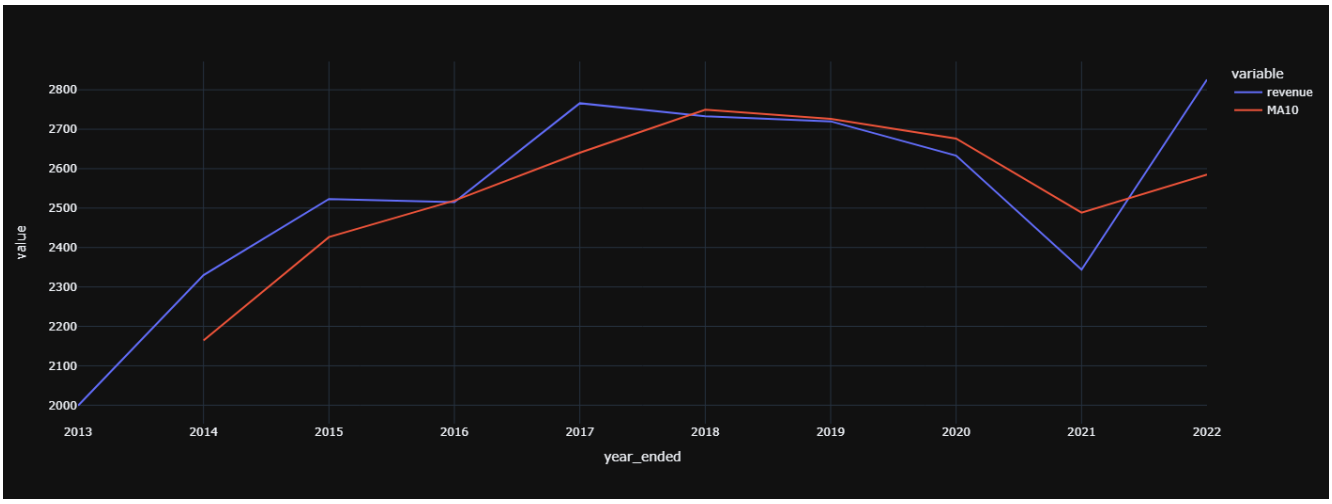

Fig 3

In the above figure Fig 3, we have first plotted the revenue and the moving averages over the years. We are going to use this data to perform forecasting values for the next 5 years. For this, the dataset is altered by dropping the unrelated columns and adding the column for moving averages using the below code.

```
# create a sequence of numbers
data['Series'] = np.arange(1,len(data)+1)
# drop unnecessary columns and re-arrange
data.drop(['adjustedoperating_profit','adjusteddiluted_eps','profit_after_tax','ope
rating_profit','dividendper_share'],axis=1, inplace=True)
data = data[['Series','year_ended','revenue','MA10']]
```

After this, we divide our dataset into training and test set manually. After this is done, we proceed to setup the time series regression model using the PyCaret library in Python. We compare the various models available under this type of regression method to find the best model for the provided data.

```
best = compare_models(sort = 'MAE')
```

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
xgboost	Extreme Gradient Boosting	137.2337	46190.5624	137.2337	nan	0.0530	0.0498	
lightgbm	Light Gradient Boosting Machine	285.8987	86176.9652	285.8987	nan	0.1126	0.1063	
dummy	Dummy Regressor	285.8987	86176.9636	285.8987	nan	0.1126	0.1063	

Fig 4

According to the results displayed in Fig 4, Extreme Gradient Boosting is found to be the best model and is applied to the dataset for forecasting. We first train this model using our training set and then create a series for the additional years we want to predict values for.

```
future_dates = pd.date_range(start = '2022', end = '2028', freq = 'BY')
future_df = pd.DataFrame()
future_df['year_ended'] = ['2022','2023','2024','2025','2026','2027']
future_df['Series'] = np.arange(145,(145+len(future_dates)))
future_df.head()
```

Then, we apply the model to these values to predict values.

```
predictions_future = predict_model(final_best, data=future_df)
```

```
predictions_future.head()
```

	year_ended	Series	MA10	prediction_label
0	2022	145	NaN	2520.593262
1	2023	146	2164.5	2527.901611
2	2024	147	2426.5	2527.901611
3	2025	148	2519.0	2566.026367
4	2026	149	2640.5	2825.999023

Fig 5

Fig 5 shows the predicted values under the column “prediction_label”. This data is then visualized to check for values.

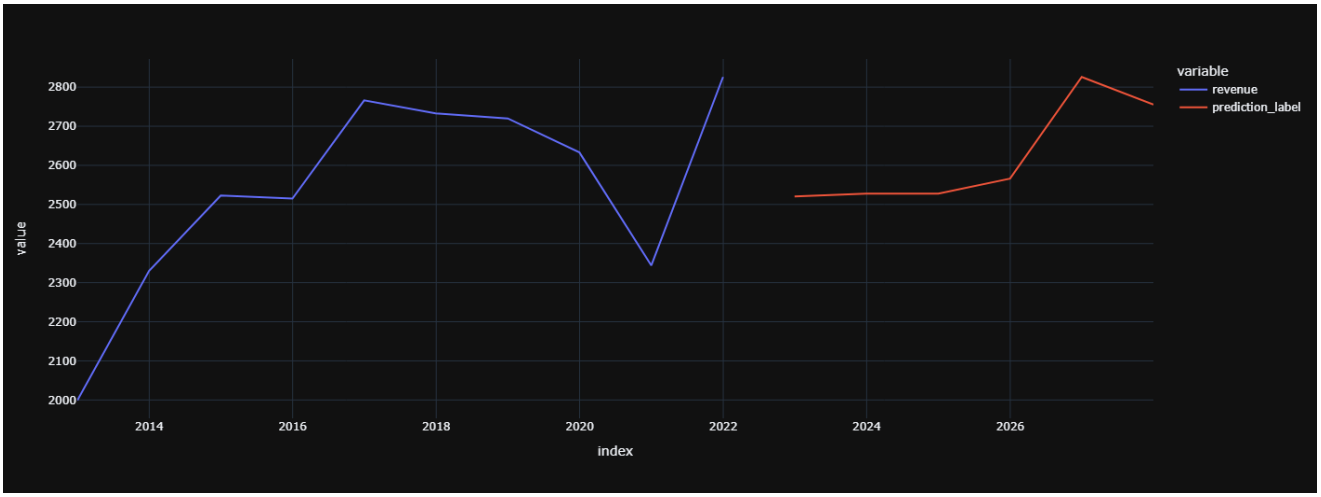

Fig 6

The above figure, Fig 6, shows us the actual values over the years and the forecasted values for the next 5 years.

The same methods that have been used to forecast revenue using time series regression have also been used to forecast profit after taxation.

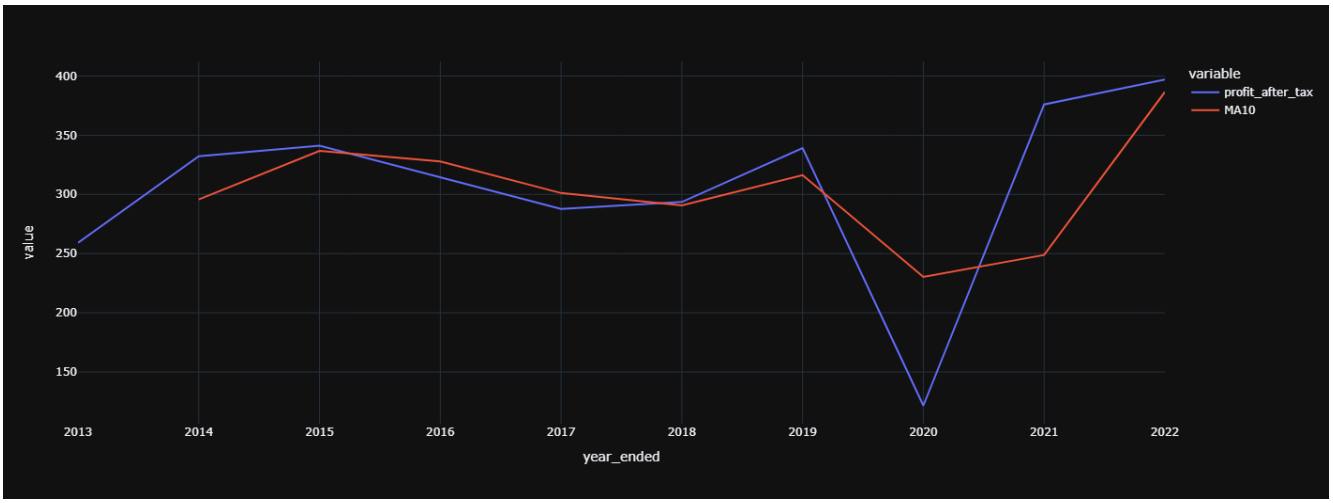

Fig 7

In the above figure Fig 7, we have first plotted the profit after tax and the moving averages over the years. We are going to use this data to perform forecasting values for the next 5 years. For this, the dataset is altered by dropping the unrelated columns and adding the column for moving averages using the below code.

```
# create a sequence of numbers
data['Series'] = np.arange(1,len(data)+1)
# drop unnecessary columns and re-arrange
data.drop(['revenue','adjustedoperating_profit','adjusteddiluted_eps','operating_pr
ofit','dividendper_share'],axis=1, inplace=True)
data = data[['Series','year_ended','profit_after_tax','MA10']]
```

After this, we divide our dataset into training and test set manually. After this is done, we proceed to setup the time series regression model using the PyCaret library in Python. We compare the various models available under this type of regression method to find the best model for the provided data.

```
best = compare_models(sort = 'MAE')
```

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lightgbm	Light Gradient Boosting Machine	19.0833	472.7042	19.0833	nan	0.0624	0.0652	
dummy	Dummy Regressor	19.0833	472.7030	19.0833	nan	0.0624	0.0652	
xgboost	Extreme Gradient Boosting	19.8223	505.8029	19.8223	nan	0.0634	0.0662	

Fig 8

According to the results displayed in Fig 8, Light Gradient Boosting Machine is found to be the best model and is applied to the dataset for forecasting. We first train this model using our training set and then create a series for the additional years we want to predict values for.

```
future_dates = pd.date_range(start = '2022', end = '2028', freq = 'BY')
future_df = pd.DataFrame()
future_df['year_ended'] = ['2022','2023','2024','2025','2026','2027']
future_df['Series'] = np.arange(145,(145+len(future_dates)))
future_df.head()
```

Then, we apply the model to these values to predict values.

```
predictions_future = predict_model(final_best, data=future_df)
predictions_future.head()
```

	year_ended	Series	MA10	prediction_label
0	2022	145	NaN	318.865263
1	2023	146	295.750000	318.865263
2	2024	147	336.700012	318.865263
3	2025	148	327.850006	318.865263
4	2026	149	301.149994	318.865263

Fig 9

Fig 9 shows the predicted values under the column “prediction_label”. This data is then visualized to check for values.

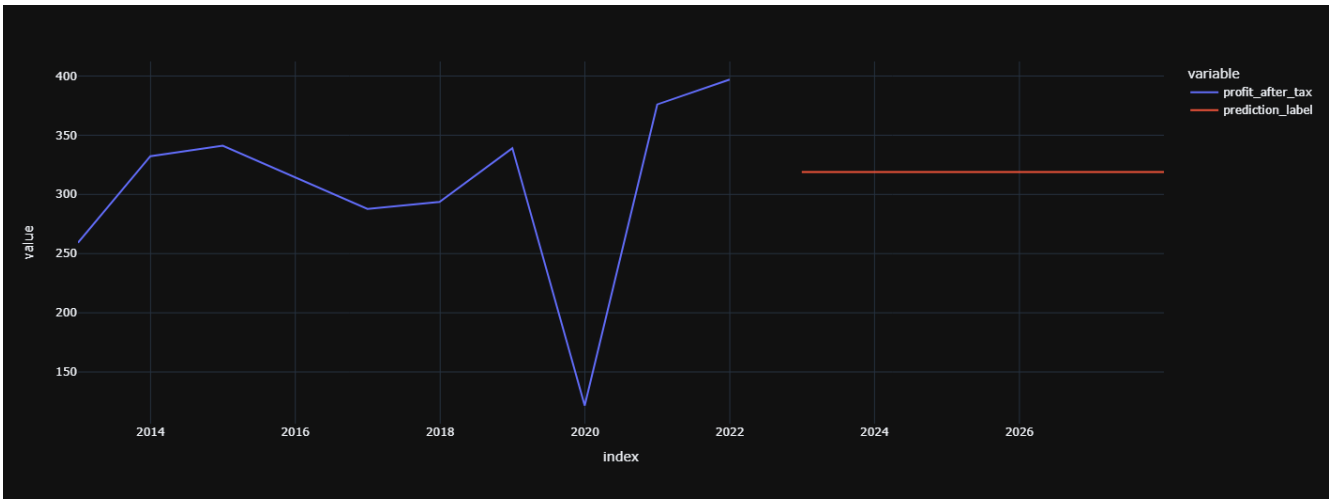

Fig 10

The above figure, Fig 6, shows us the actual values over the years and the forecasted values for the next 5 years.

4. Results and Discussion

The above methods and analyses, show us that simple linear regression method cannot be used to perform forecasting for revenue as it does not give accurate results and have a large error margin, whereas when the same is done using time series regression, the forecasting has been done more accurately. Through this we understand that time series regression methods work best when we want to forecast values based on historical data for a time series. The below given figures, Fig 11 and Fig 12 are for prediction using simple linear regression and time series regression methods respectively.

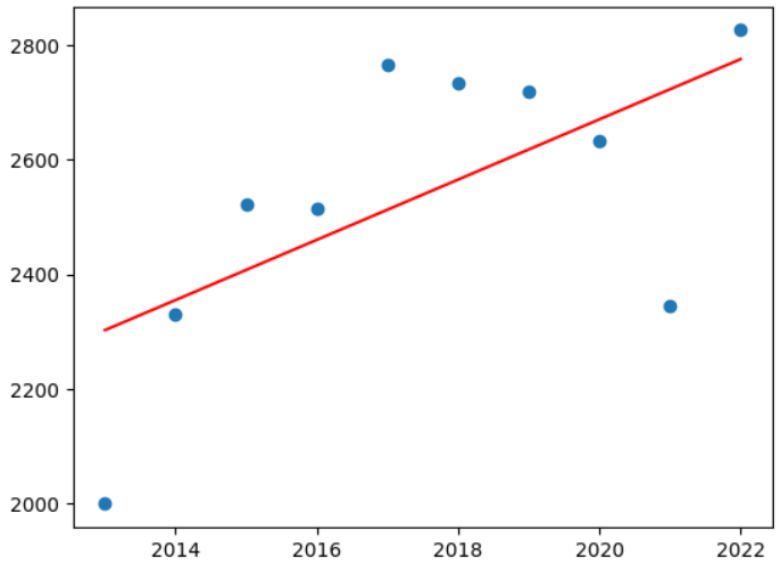

Fig 11

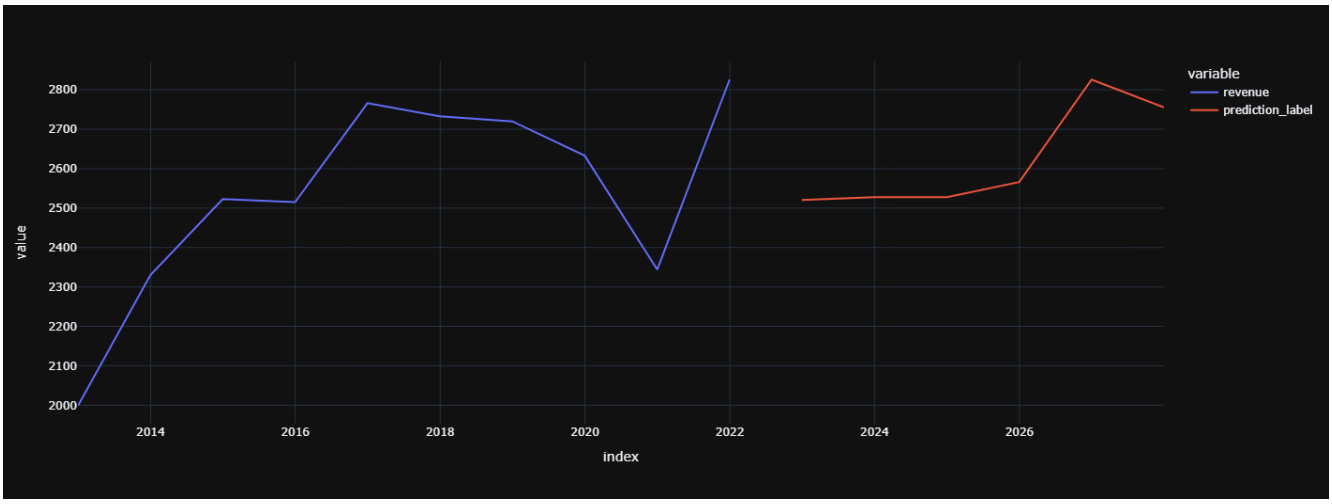

Fig 12

The above two figures are predicting the same variable with using the same parameters but different methods but it can be seen that time series regression method is better when we want to forecast future values.

In the below given figures, Fig 13 and 14, we can see the results that we have obtained for predicting the profit after tax for an organization using two different methods, multiple linear regression, and time series regression.

```
[[313.92 341.1 ]
 [427.76 376.  ]
 [261.44 287.7 ]
 [299.08 397.  ]
 [311.82 332.3 ]]
```

Fig 13

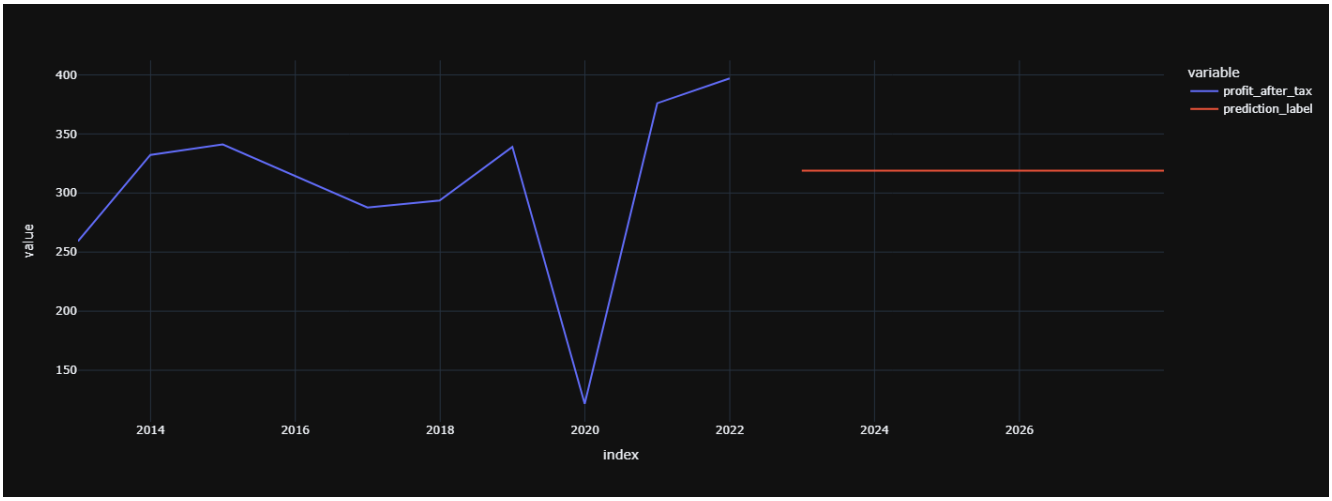

Fig 14

Here, we can see in the above figures that when we try to predict profit after tax using time series regression it does not give us proper results as this method does not consider the different attributes other than “year_ended” attribute that affect the profit after tax of an organization which is why using multiple linear regression gives us better results as this method is specifically designed to predict a variable based on the values of more than one variable and as seen in this case profit after tax is dependent on multiple other components of a financial statement, namely, revenue, cost of sales, operating expenses, adjusting operating items, net finance (charge)/credit, and taxation.

5. Conclusion and Future work

In conclusion, Time Series Regression is a regression method useful for forecasting values using historical data when there is a time series however multiple series regression works better when we want to predict a variable that is not only dependent on more than one independent variable but also has a time series.

The future scope of this paper includes trying to find an algorithm that lets us combine time series regression and multiple linear regression in order to obtain better results in predicting and forecasting variables having multiple independent variables.

References

[1] Stock Trend Prediction Using Regression Analysis – A Data Mining Approach by S Abdulsalam Sulaiman Olaniyi, Adewole, Kayode S., Jimoh, R. G

[2] Time series extrinsic regression Predicting numeric values from time series data by Chang Wei Tan, Christoph Bergmeir, François Petitjean, and Geoffrey I. Webb

[3] Fundamental Analysis and the Prediction of Earnings by Dyna Seng and Jason R. Hancock

[4] Fundamental Analysis, Future Earnings, and Stock Prices by Jeffery S. Abarbanell and Brian J. Bushee

[4] Time Series Regression - MATLAB & Simulink

[5] Linear Regression Using Least Squares | by Adarsh Menon | Towards Data Science

[6] Time Series Regression VII: Forecasting - MATLAB & Simulink Example

[7] Time Series Forecasting with PyCaret Regression Module | by Moez Ali | Towards Data Science