A prediction model that identify how many R&D cost is good for IBM's stock price

Le Zhou – MS of Data Science
Seidenberg School Of Computer Science and information Systems, Pace University

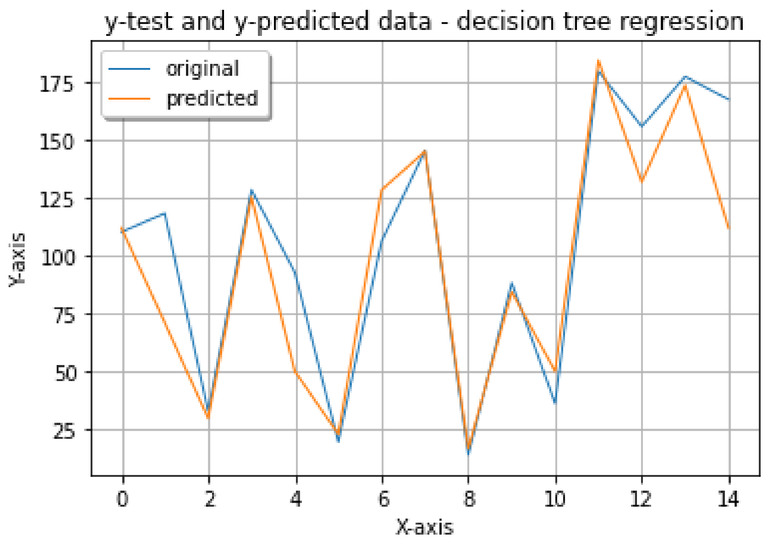
Introduction

Accurate prediction is hard in chaotic commercial system. However, the logic of analysis financial report give a new idea to revisit the process that people evaluate a company by accounting experience. Accounting is the process of recording financial transactions pertaining to a business and historic standard which is designed by exports. We could build a relation to quickly analysis their accounting's change. It is hard to find an explainable mathematics model between R&D and company's developing potential, but it is reasonable for the relation and exploring its application.

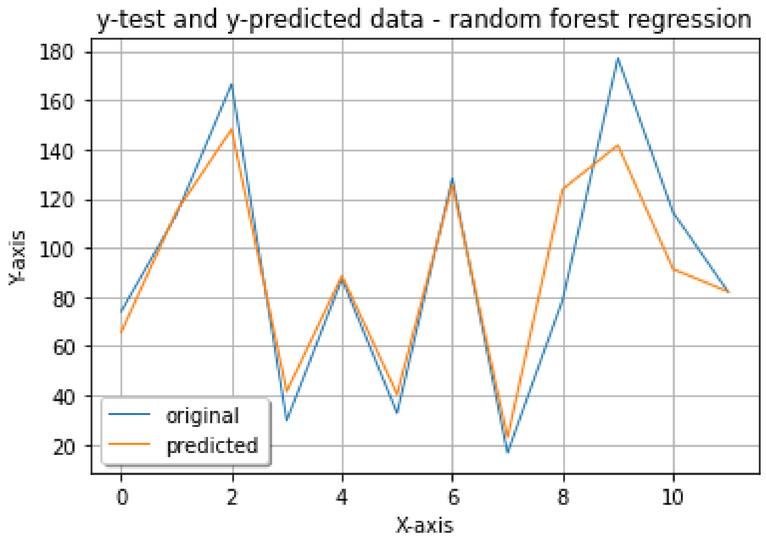
	Date	R&D	Nasdaq_value	stockprice
0	1994–06–30	1.091000e+09	705.960022	14.041587
1	1994–09–30	1.053000e+09	764.289978	16.640774
3	1995–03–31	9.130000e+08	817.210022	19.628345
4	1995–06–30	9.740000e+08	933.450012	22.944551
9	1996–09–30	1.115000e+09	1226.920044	29.756214
...	...	...	...	...
104	2020–06–30	1.582000e+09	10058.769531	115.458893
105	2020–09–30	1.515000e+09	11167.509766	116.319313
106	2020–12–31	1.611000e+09	12888.280273	120.344170
107	2021–03–31	1.630000e+09	13246.870117	127.399620
108	2021–06–30	1.657000e+09	14503.950195	140.143402

Methodology

Decision tree regression



Random Forest regression

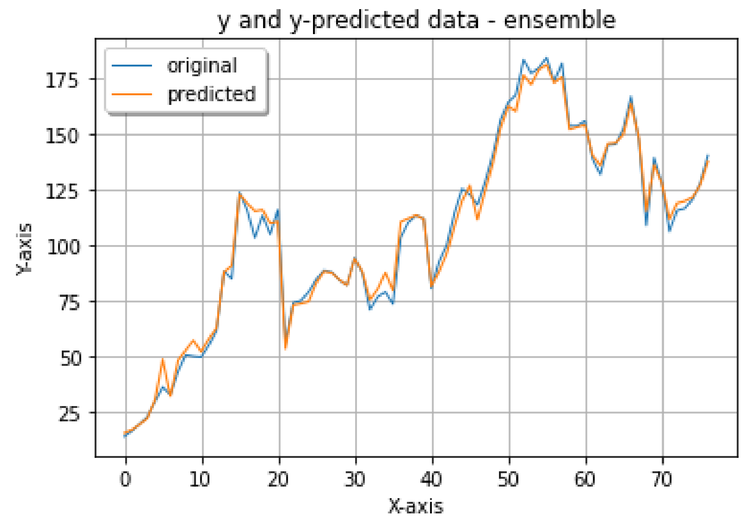


Ensemble methods – voting regression

Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model

Results

	Decision tree	Random forest	Voting regression
R-squared	0.81	0.95	0.99
MSE	562.89	371.75	16.45
RMSE	15.42	185.88	3.05



The result will indicate the relation between R&D and IBM's stock price, which remind users to have a better budget in R&D for next year

Conclusion

1. Time series data is a way to load massive accounting data in regression model.
2. Decision tree regression and random forest regression can treat non-linear issues and less feature.
3. Random forest also is a suitable addition by voting model. It often adapt in same conditions because random forest is a bagging of decision tree.
4. The model reach the intended result that accuracy is 99%. R&D could build an accurate relationship with IBM's stock price, which is a workable method to predict the company's value then guide R&D management

Acknowledgements

I would like to express my very great appreciation to Dr Christelle Scharff. Advice given by Dr. Christelle Scharff has been a great help in this capstone project

Reference

[1] Tang, L., Pan, H., & Yao, Y. (2018). K-nearest neighbor regression with principal component analysis for financial time series prediction. *Proceedings of the 2018 International Conference on Computing and Artificial Intelligence - ICCAI 2018*. https://doi.org/10.1145/3194452.3194467
[2] Sen, J., Mehtab, S., & Dutta, A. (2021). Stock price prediction using machine learning and LSTM-based deep learning models. https://doi.org/10.36227/techrxiv.15103602.v1
[3] Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., & Smith, N. A. (2009). Predicting risk from financial reports with regression. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics on - NAACL '09*. https://doi.org/10.3115/1620754.1620794
[4] Wang, H. (2020). Stock price prediction based on machine learning approaches. *Proceedings of the 3rd International Conference on Data Science and Information Technology*. https://doi.org/10.1145/3414274.3414275
[5] Sen, J., Mehtab, S., & Dutta, A. (2021). Stock price prediction using machine learning and LSTM-based deep learning models. https://doi.org/10.36227/techrxiv.15103602.v1