FEDERAL STATE AUTONOMOUS EDUCATIONAL

INSTITUTION OF HIGHER EDUCATION

MOSCOW INSTITUTE OF PHYSICS AND TECHNOLOGY

(NATIONAL RESEARCH UNIVERSITY)

On the rights of the manuscript

Ivan Legenchuk

Chief executive officer, Iron Capital Ltd.

Forecasting a company's revenue using regression of news data vectors

Detailed abstract of the report on

XXV Yasin (April) International Academic Conference on Economic and Social Development

Moscow

2025

Ayn Rand, a remarkable American writer of Soviet origin, in her novel Atlas Shrugged, published in 1957, pointed out that money is a material expression of the principle that people can interact through trade and pay value for value. By producing goods and services, a person is convinced that he will exchange them for the product of someone else's labor with the help of money – "these pieces of paper, like gold, contain the energy of people who produce values." Eight years earlier, Benjamin Graham, in his book The Intelligent Investor, suggested that an investor compare the price of shares of a company that he was offered to purchase with their fair value, and purchase only those shares whose difference between the market price and the value of which was greatest, while emphasizing that only the future value mattered.

Indeed, the method of discounting cash flows, Fisher, 1930, allows you to evaluate a company based on an assessment of its future cash flows, one way or another, created by the staff, controlling shareholders and the founders of the company. One of the main disadvantages of the method is the inability to accurately predict growth over a 3-5-year horizon, Chan et al., 2003. The importance of growth is due to the need to predict the company's revenue, which, in turn, is an argument for the cash flow function, Damodaran, 2002:

$$CF = (SALES\ x\ EBITDAmargin - DA)\ x\ (1 - Tax) - \Delta WC$$

Thus, the subject of the author's research interests is forecasting revenue (SALES) or its relative change - growth. From a practical point of view, when fixing other components of the cash flow formula, including EBITDA margin (EBITDAmargin), depreciation (DA), income tax rate (Tax) and changes in working capital (ΔWC), forecasting the company's revenue using real-time machine learning methods will allow determining the fair value of the company and use the data obtained to manage the stock portfolio.

In order to choose a method, the author analyzed existing machine learning methods used to predict time series, product production and revenue. Shi et al., 2012 applied the ARIMA integrated moving average autoregression model, an artificial neural network (ANN), and the support vector machine (SVM) method to predict wind power and speed. Aries et al., 2013 showed that SVM outperformed ANN in predicting market volatility based on twitter data. Of all the machine learning methods, deep neural network and random forest found the smallest average absolute percentage error in predicting revenue in the fashion market, Loureiro, 2018. Lu et al., 2019 proved

that SVM surpassed ANN in predicting gas consumption. Dedi, 2020, predicting the demand for electricity, was convinced of the superiority of the long-term short-term memory (LSTM) network over SVM. Finally, Ma, 2021 showed that a two-channel convolutional neural network (CNN) is better at predicting retail sales than SVM.

Thus, the author made a choice in favor of regression of support vectors, a method that takes into account nonlinear patterns in the data, and the quality of which is comparable to a neural network of long–term short-term memory and a convolutional neural network.

As input data, the author used quarterly revenue data from thirty-five companies traded on the Moscow Stock Exchange and news in Russian from 2004 to 2011. The news was filtered by keywords regarding unemployment. Thus, for each week from the study period, two news items about unemployment and the number of initial and repeated applications for unemployment in the United States were included in the dataset. According to the companies' quarterly revenue data, year-on-year growth was calculated, and news about unemployment was vectorized using Yandex GPT. The news vectors within one quarter are summarized. The data was divided into a training, test, and validation dataset. The SVR algorithm was trained using the sklearn library. Errors were calculated on the test dataset – the average absolute percentage error (MAPE), the coefficient of determination (R2) and the RMS error (RMSE).

Based on the data obtained for company Magnit - R2 = 0.87, MAPE = 0.24, RMSE = 0.05 with an average revenue increase of 0.25 in the test period – the author concluded that it is acceptable to predict the company's revenue using regression of the reference vectors of news about unemployment in the United States.

The revenue predictions obtained, taking into account the premise that it will deviate within one standard deviation during the forecast period (Magnit's historical revenue increases are subject to the normal law), allowed us to calculate the company's fair value using the cash flow discounting method at the time of the publication of unemployment news in the United States, Damodaran, 2002. The calculations obtained were used as the basis for a trading algorithm, the application of which to market data in the period from July 2019 to January 2022 allowed to obtain a yield of 41.3% per annum.

1. Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). Time series analysis: Forecasting and control (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall;
2. Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159–175;
3. Haykin, S. (1994). Neural networks: A comprehensive foundation (2nd ed.). New Jersey: Upper Saddle River;
4. Haykin, S., Ukrainec, A.M. (1996). A Modular Neural Network for Enhancement of Cross-polar Radar Targets. Neural Networks, Vol. 9, No. 1, pp. 143-168;
5. Agrawal, D., & Schorling, C. (1996). Market share forecasting: An empirical comparison of artificial neural networks and multinomial logit model. Journal of Retailing, 72(4), 383–407;
6. Ainscough, T. L., & Aronson, J. E. (1999). An empirical investigation and comparison of neural networks and regression for scanner data analysis. Journal of Retailing and Consumer Services, 6(4), 205–217;
7. Stock, J. H., & Watson, M. W. (1999). A comparison of linear and non-linear university models for forecasting economic time series. In R. F. Engle & H. White (Eds.), Cointegration, causality, and forecasting: A Festschrift in honour of Clive W.J. Granger (pp. 1–44). Oxford: Oxford University Press;
8. Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159–175;
9. Wold, S., Sjostrom, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 58(2), 109–130;
10. Hornquist, M., Hertz, J., & Wahde, M. (2003). Effective dimensionality for principal component analysis of time series expression data. Biosystems, 71(3), 311–317;
11. Poh, H. L., Yao, J., & Jasˇic, T. (1998). Neural networks for the analysis and forecasting of advertising and promotion impact. International Journal of Intelligent Systems in Accounting, Finance & Management, 7, 253–258;
12. Mastorocostas, P. A., Theocharis, J. B., & Petridis, V. S. (2001). A constrained orthogonal least-squares method for generating TSK fuzzy models: Application to short-term load forecasting. Fuzzy Sets and Systems, 118, 215–233;
13. Goldberg, D. E. (1989). Genetic algorithms in search, optimization and machine learning. Reading, MA: Addison-Wesley;
14. Michalewicz, Z. (1996). Genetic algorithms + data structures = evolution programs (3rd ed.). Berlin: Springer-Verlag;

15. Doganis, P. et al(2006). Time series sales forecasting for short shelf-life food products based on artificial neural networks and evolutionary computing, Journal of Food Engineering 75 (2006) 196–204;
16. Lorek KS, Willinger GL (2011) Multi-step-ahead quarterly cash-fow prediction models. Account Horizons 25:71–86;
17. Brown LD, Rozef MS (1979) Univariate time-series models of quarterly accounting earnings per share: a proposed model. J Account Res 17:179–189
18. LOUIS K. C. CHAN, JASON KARCESKI, and JOSEF LAKONISHOK (2003) The level and persistence of growth rates.
19. Lorek KS (2014) A critical assessment of the time-series literature in accounting pertaining to quarterly accounting numbers. Adv Account 30:315–321;
20. Baginski SP, Branson BC, Lorek KS, Willinger GL (2003) A time-series approach to measuring the decline in quarterly earnings persistence. Adv Account 20:23–42;
21. Kwon SS, Yin J (2015) A comparison of earnings persistence in high-tech and non-high-tech firms. Rev Quant Finan Acc 44:645–668;
22. Lorek KS, Willinger GL (2011) Multi-step-ahead quarterly cash-fow prediction models. Account Horizons 25:71–86;
23. Vapnik V (1995) The nature of statistical learning theory. Springer, New York
24. Fisher, Irving. "The theory of interest rates." New York 43 (1930)
25. Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput 14:199–222;
26. Fischer, J.A. et al A machine learning approach to univariate time series forecasting of quarterly earnings. Review of Quantitative Finance and Accounting (2020) 55:1163–1179;
27. Meulstee, P., Pechenizkiy M., Food Sales Prediction: “If Only It Knew What We Know”, IEEE, 2009;
28. G. Widmer and M. Kubat. Learning in the presence of concept drift and hidden contexts. Mach. Learn., 23(1):69–101, 1996;
29. C. Schaffer. Technical note: Selecting a classification method by cross-validation. Machine Learning, 13(1):135– 143, 1993;
30. A. Tsymbal, M. Pechenizkiy, P. Cunningham, and S. Puuronen. Handling local concept drift with dynamic integration of classifiers: Domain of antibiotic resistance in nosocomial infections. In Proceedings of CBMS 2006, International Symposium on Computer-Based Medical Systems, pages 679–684, Los Alamitos, CA, USA, 2006. IEEE Computer Society;

31. A. Tsymbal, M. Pechenizkiy, P. Cunningham, and S. Puuronen. Dynamic integration of classifiers for handling concept drift. Information Fusion, 9(1):56–68, 2008;
32. Liu, X. & Ichise, R., Food Sales Prediction with Meteorological Data — A Case Study of a Japanese Chain Supermarket, Data Mining and Big Data pp 93-104;
33. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9(8), 1735–1780 (1997);
34. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Networks 5(2), 157–166 (1994);
35. Hinton, G.E., Salakhutdinov, R.R. Reducing the dimensionality of data with neural networks. Science 313(5786), 504–507 (2006);
36. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. 11, 3371–3408 (2010);
37. Amiri, H., Resnik, P., Boyd-Graber, J., Daum´e III, H.: Learning text pair similarity with context-sensitive autoencoders. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, pp. 1882–1892, August 2016;
38. Cox, D.R.: The regression analysis of binary sequences. J. Roy. Stat. Soc. Ser. B 20(2), 215–242 (1958);
39. De Bie T., Cristianini N., Rosipal R. Eigenproblems in pattern recognition. In: Handbook of Geometric Computing. Ed. E.B. Corrochano. Berlin, Springer, 2005, pp. 129–167. doi: 10.1007/3-540-28247-5_5.