

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО
ОБРАЗОВАНИЯ
МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

На правах рукописи

Легенчук Иван Георгиевич

Главный исполнительный директор, Iron Capital Ltd.

**Прогнозирование выручки компании с помощью регрессии опорных
векторов новостных данных**

Развёрнутая аннотация доклада

на

**XXV Ясинской (Апрельской) международной научной конференции по
проблемам развития экономики и общества**

Москва

2025

Айн Рэнд, замечательная американская писательница советского происхождения, в своём романе «Атлант расправил плечи», вышедшем в свет в 1957 году, указывала, что деньги являются материальным выражением того принципа, что люди могут взаимодействовать при помощи торговли и платить ценностью за ценность. Производя товары и услуги, человек убеждён, что обменивает их на продукт чужого труда с помощью денег – «эти кусочки бумаги, как и золото, содержат энергию людей, производящих ценности». За восемь лет до этого, Беджамин Грэхем в своей книге «Разумный инвестор» предлагал инвестору сопоставлять цену акций компании, которые ему предлагают приобрести, с их справедливой стоимостью, и приобретать лишь те акции, различие между рыночной ценой и стоимостью которых, наибольшее, подчёркивая, при этом, что значение для оценки имеет лишь будущая прибыль компании.

Действительно, метод дисконтирования денежных потоков, Фишер, 1930, позволяет оценить компанию на основе оценки её будущих денежных потоков, так или иначе, создаваемых персоналом, контролирующими акционерами и основателями компании. В качестве одного из основных недостатков метода является невозможность достаточно точно спрогнозировать рост на горизонте 3-5 лет, Chan et al., 2003. Важность роста обусловлена необходимостью спрогнозировать выручку компании, которая, в свою очередь является аргументом функции денежного потока, Damodaran, 2002:

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

Таким образом, предметом научных интересов автора является прогнозирование выручки (SALES) или её относительного изменения - роста. С практической же точки зрения, при фиксации прочих составляющих формулы денежного потока, в том числе, EBITDA рентабельности (*EBITDAmargin*), амортизации (*DA*), ставки налога на прибыль (*Tax*) и изменения оборотного капитала (ΔWC), прогнозирование выручки компании, осуществляемое с помощью методов машинного обучения в реальном времени, позволит определять справедливую стоимость компании и использовать полученные данные для управления портфелем акций.

В целях выбора метода, автором был проведён анализ существующих методов машинного обучения, применяемых для прогнозирования временных рядов, выработки продукции и выручки. Shi et al., 2012 применили интегрированную модель авторегрессии скользящего среднего ARIMA, искусственную нейронную сеть (ANN) и метод опорных векторов (SVM) для прогнозирования мощности и скорости ветра. Aries et al., 2013 показали, что SVM превзошёл ANN в предсказании рыночной волатильности на основе

данных твиттера. Из всех методов машинного обучения, глубокая нейронная сеть и случайный лес обнаружили наименьшую среднюю абсолютную процентную ошибку при прогнозировании выручки на рынке модной одежды, Loureiro, 2018. Lu et al., 2019 доказал, что SVM превзошёл ANN в предсказании потребления газа. Dedi, 2020, предсказывая спрос на электроэнергию, убедился в превосходстве сети долгой краткосрочной памяти (LSTM) над SVM. Наконец, Ма, 2021 показал, что двухканальная свёрточная нейронная сеть (CNN) лучше предсказывает розничные продажи, чем SVM.

Таким образом, автором был сделан выбор в пользу регрессии опорных векторов – метода, учитывающего нелинейные закономерности в данных, а качество работы которого сравнимо с нейронной сетью долгой краткосрочной памяти и свёрточной нейронной сетью.

В качестве входных данных, автором были использованы данные по квартальной выручке тридцати пяти компаний, торгующихся на Московской бирже и новости на русском языке с 2004 по 2011 год. Новости были отфильтрованы по ключевым словам в отношении безработицы. Таким образом, в датасет за каждую неделю из исследуемого периода попали две новости о безработице, о числе первичных и повторных обращений по безработице в США. По данным квартальной выручки компаний был вычислен прирост относительно год к году, а новости о безработице векторизованы с помощью Yandex GPT. Новостные векторы внутри одного квартала суммированы. Данные были разделены на обучающий, тестовый и валидационный датасет. На обучающем проведено обучение SVR алгоритма с помощью библиотеки sklearn. На тестовом датасете были вычислены ошибки – средняя абсолютная процентная ошибка (MAPE), коэффициент детерминации (R^2) и среднеквадратичная ошибка (RMSE).

На основании полученных данных по компании Магнит - $R^2 = 0.87$, MAPE = 0.24, RMSE = 0.05 при среднем значении прироста выручки 0.25 в тестовом периоде – автором был сделан вывод о допустимости прогнозирования выручки компании с помощью регрессии опорных векторов новостей о безработице в США.

Полученные предсказания выручки, с учётом предпосылки о её отклонении в пределах одного стандартного отклонения в течение прогнозного периода (исторические приросты выручки Магнита подчиняются нормальному нормальному закону), позволили рассчитать справедливую стоимость компании по методу дисконтирования денежных потоков в момент публикации новостей по безработице в США, Damodaran, 2002. Полученные расчёты были положены в основу торгового алгоритма, применение которого

к рыночным данным в период с июля 2019 года по январь 2022 года позволил получить доходность в размере 41.3% годовых.

Библиография.

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., & Jas'ic, 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. Фишер, Ирвинг. "Теория процентных ставок". Нью-Йорк 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.