Стратегии инвестирования с использованием моделей машинного обучения

К. М. Баязитов

Выпускная квалификационная работа
09.04.01 — Информатика и вычислительная техника
Научный руководитель: А. В. Ильницкая

20 июня 2024 г.

Слайд об исследованиях

Цель исследования —

Повышение качества моделей прогнозирования временных рядов на примере динамики курса акций.

Предположение —

Внешние факторы, влияющие на курс акций, заложены в ответы опытных инвесторов.

Решение —

Предлагается использовать в модели помимо данных временного ряда также агрегированные ответы опытных инвесторов.

Постановка задачи прогнозирования

$y_1, y_2, ..., y_T$ - временной ряд, $y_i \in \mathbb{R}$.

Требуется получить модель временного ряда:

$$\hat{y}_{t+k}(\mathbf{w}) = f_{t,k}(y_{t-M+1}, ..., y_t; \mathbf{w})$$

$$k = 1, ..., K,$$

где
M - размер входного окна,
K - горизонт прогнозирования,
$\mathbf{w}$ - вектор параметров модели.
Функция потерь $\mathcal{L}$, используемая при обучении модели:

$$\mathcal{L}(\mathbf{w}, \mathbf{Y}) = \sum_{t=M}^{T-K} \sum_{k=1}^{K} (f_{t,k}(y_{t-M+1}, ..., y_t; \mathbf{w}) - y_{t+k})^2,$$

Оптимизационная задача:

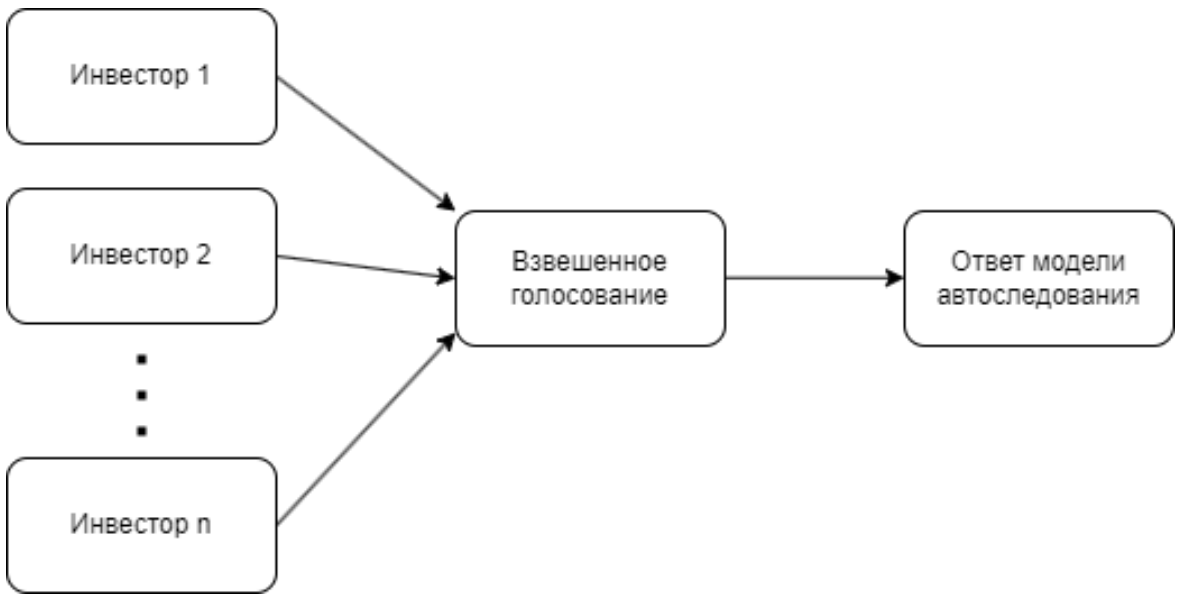
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{W}} \mathcal{L}(\mathbf{w}, \mathbf{Y}).$$

Модель автоследования

Автоследование — способ инвестирования, при котором все желающие могут подключиться к стратегии более опытного инвестора (он же автор стратегии) и автоматически повторять все его сделки на своем счете.

$$\text{Ответ инвестора} = \frac{\text{Сумма сделки}}{\text{Объем портфеля}}$$

Путем усреднения ответов инвесторов о продаже или покупке акций составляется временной ряд $a_1, ..., a_N, a_i \in [-1, 1]$.



Экспериментальные данные

Эксперимент проводится для данных динамики курса акций YNDX.
Задается временной ряд

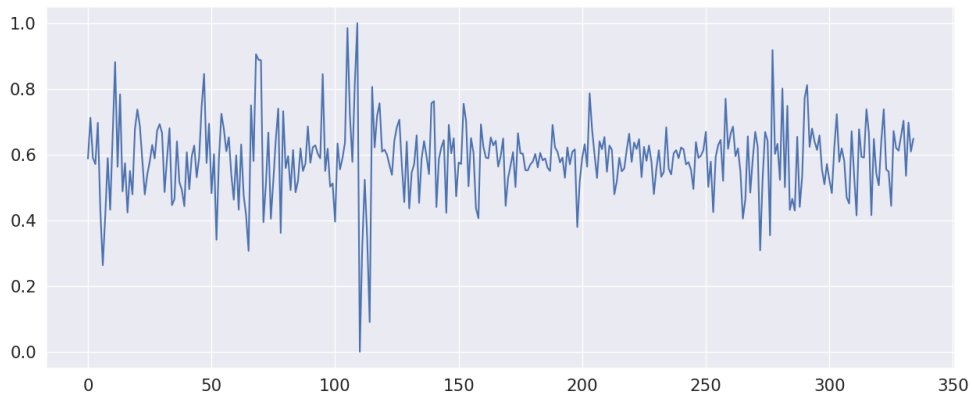
$$x_1, x_2, x_3, ..., x_N, \quad x_i \in \mathbb{R}^5$$

$$x_i = \begin{bmatrix} c_i & o_i & h_i & l_i & a_i \end{bmatrix}^T,$$

где
c_i - цена закрытия,
o_i - цена открытия,
h_i - максимальная цена,
l_i - минимальная цена,
a_i - ответ модели автоследования ($a_i = 0$ в базовом варианте обучения модели)

Стационарность

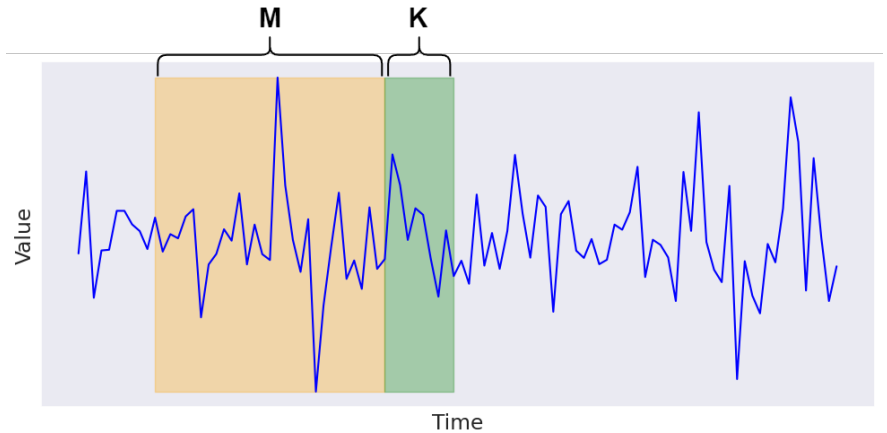
Ряд приводится к стационарному виду следующими преобразованиями:
1) Дифференцирование: $y'_t = y_t - y_{t-1}$
2) Сезонное дифференцирование $y''_t = y'_t - y'_{t-s},\ s = 5$



Для проверки ряда на стационарность используется критерий KPSS:
Для исходного ряда $p - value < 0.01$
Для полученного ряда $p - value > 0.01$

Составление выборки

Методом скользящего окна составляется выборка $\mathfrak{D} = (\mathbf{X}, \mathbf{Y})$:



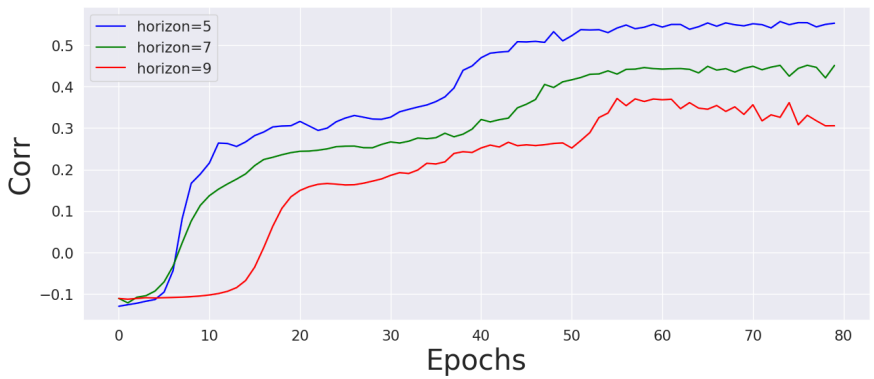
$$\mathbf{X} = \begin{pmatrix} x_1 & x_2 & \dots & x_M \\ x_2 & x_3 & \dots & x_{M+1} \\ x_3 & x_4 & \dots & x_{M+2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N-K-M+1} & x_{N-K-M+2} & \dots & x_{N-K} \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} c_{M+1} & c_{M+2} & \dots & c_{M+K} \\ c_{M+2} & c_{M+3} & \dots & c_{M+K+1} \\ c_{M+3} & c_{M+4} & \dots & c_{M+K+2} \\ \vdots & \vdots & \vdots & \vdots \\ c_{N-K+1} & c_{N-K+2} & \dots & c_N \end{pmatrix},$$

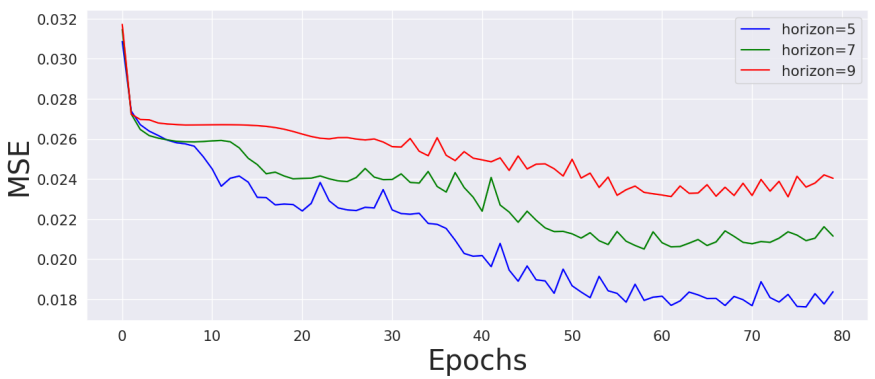
где M - размер окна, K - горизонт прогнозирования.

В соотшении 80/20 выборка делится на обучающую и тестовую части.

Горизонт прогнозирования

В качестве тестируемой модели используется Seq2Seq архитекутра на основе LSTM.
На графиках показаны метрики корреляции Пирсона и среднеквадратичной ошибки в зависимости от горизонта прогнозирования.





С увеличением горизонта прогнозирования качество модели ухудшается.

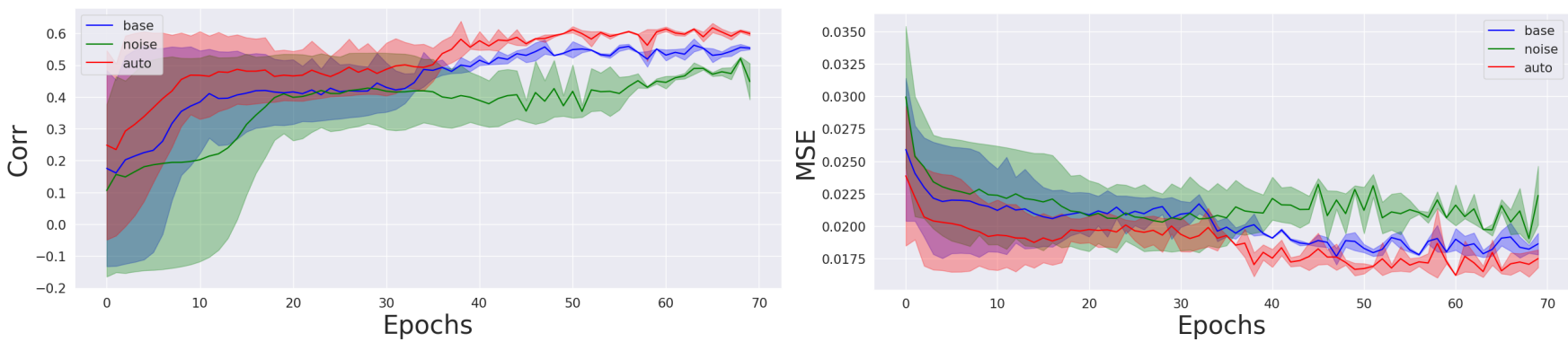
Анализ предложенного метода

В качестве тестируемой модели используется Seq2Seq архитекутра на основе LSTM.

Проводится сравнение тестируемой модели с моделями, где в качестве дополнительных данных используются:

1) Ответы модели автоследования
2) Нормальный шум $\mathcal{N}(0, 1)$

На графиках показаны метрики корреляции Пирсона и среднеквадратичной ошибки.



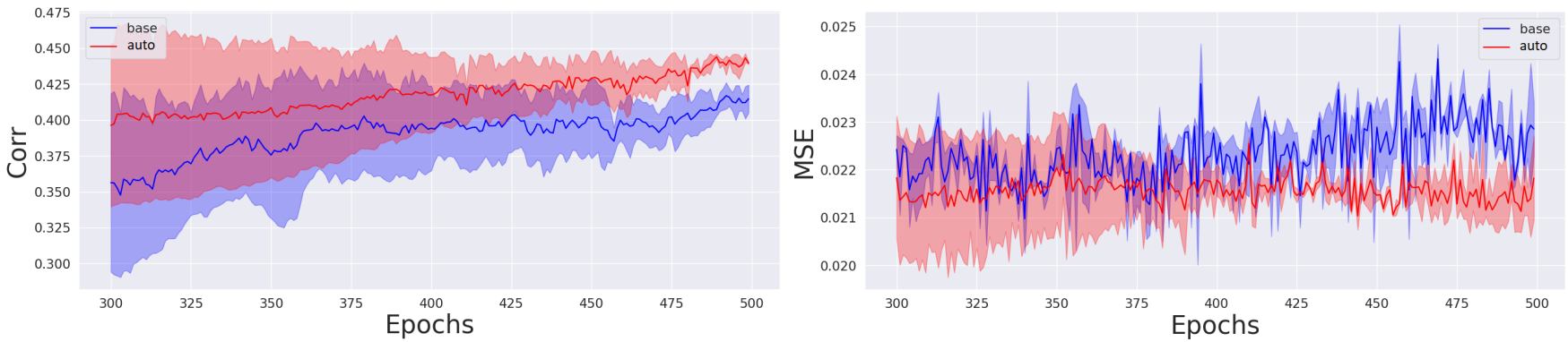
Модель, использующая ответы модели автоследования, показывает лучшее значение метрик.

Анализ предложенного метода

В качестве тестируемой модели используется Seq2Seq архитекутра на основе модели Transformer.
Проводится сравнение тестируемой модели с моделями, где в качестве дополнительных данных используются:

1) Ответы модели автоследования

На графиках показаны метрики корреляции Пирсона и среднеквадратичной ошибки.



Модель, использующая ответы модели автоследования, показывает лучшее значение метрик.

Сравнение результатов

Модель	**Дополнительные данные**	**Корреляция Пирсона**	**MSE**
ARIMA (3, 0, 3)		0,340	0,0183
Seq2Seq LSTM	—	$0{,}553 \pm 0{,}004$	$0{,}0186 \pm 0{,}0008$
Seq2Seq LSTM	Автоследование	$0{,}599 \pm 0{,}006$	$0{,}0175 \pm 0.0007$
Seq2Seq Transformer	—	$0{,}415 \pm 0{,}010$	$0{,}0229 \pm 0{,}0006$
Seq2Seq Transformer	Автоследование	$0{,}440 \pm 0{,}001$	$0{,}0218 \pm 0{,}0009$

Выводы

1. Предложен метод повышения качества модели при использовании дополнительных данных.
2. Предложен метод агрегации знаний опытных инвесторов.
3. Проведен вычислительный эксперимент на реальных данных динамики курса акций YNDX.
4. Проведен анализ выбора горизонта прогнозирования.