Стратегии инвестирования с использованием моделей машинного обучения

К. М. Баязитов

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

20 июня 2024 г.

Слайд об исследованиях

Цель исследования —

Повышение качества моделей прогнозирования временных рядов на примере курса акций.

Предположение —

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

Решение

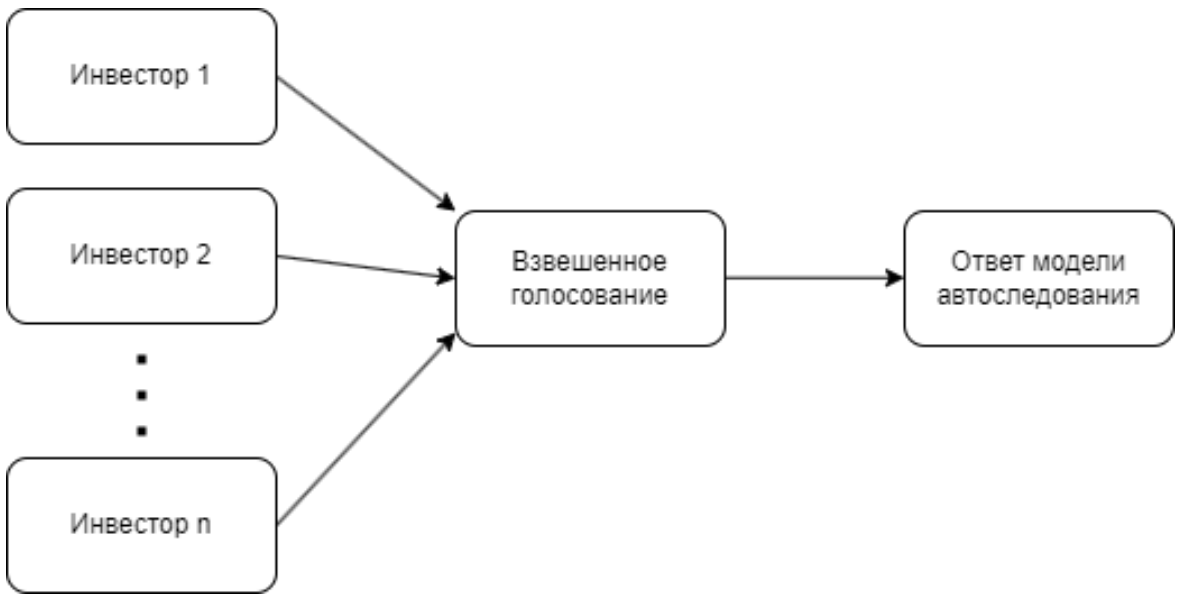
Предлагается при обучении моделей использовать помимо данных временного ряда также агрегированные ответы опытных инвесторов.

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 $a_0, ..., a_N, a_i \in [-1, 1]$.



Постановка задачи прогнозирования

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

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 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:

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

Экспериментальные данные

Эксперимент проводится для данных курса акций YNDX.
Задается временной ряд

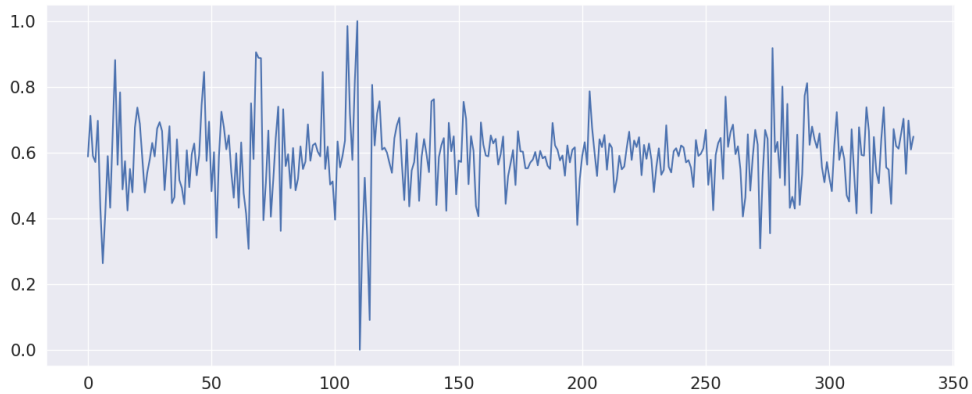
$$x_0, x_1, x_2, ...x_N, \quad x_i \in \mathbb{R}^5$$

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

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

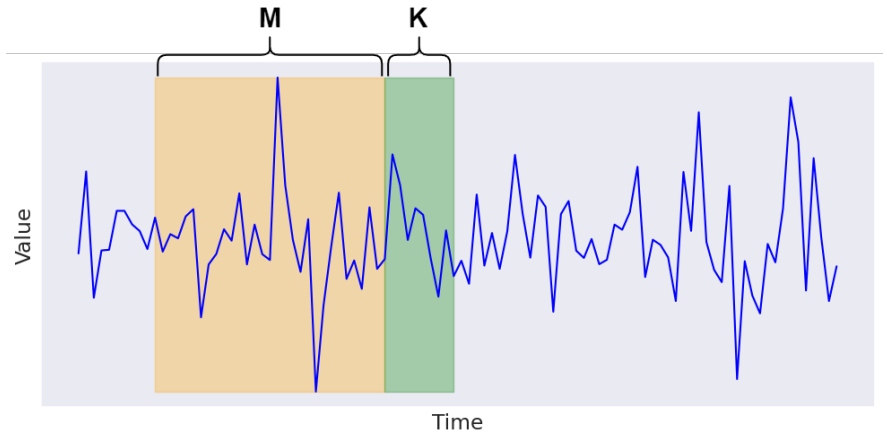
Ряд приводится к стационарному виду следующими преобразованиями:
1) Дифференцирование: $y'_t = y_t - y_{t-1}$
2) Сезонное дифференцирование $y''_t = y'_t - y'_{t-s},\ s = 5$



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 $\mathfrak{D} = (\mathbf{X}, \mathbf{Y})$:



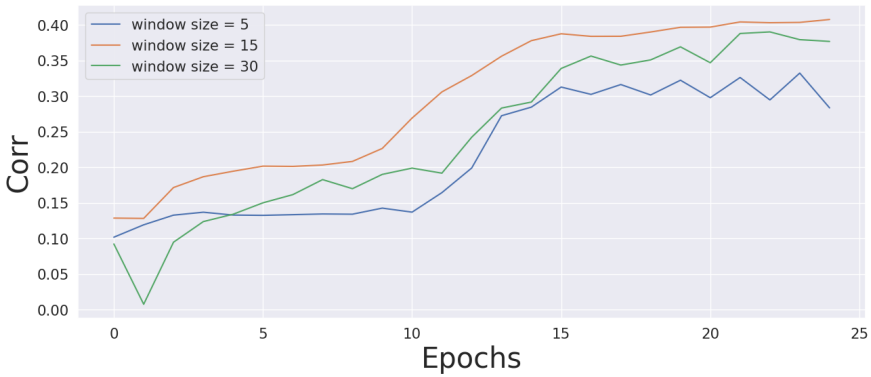
$$\mathbf{X} = \begin{pmatrix} x_0 & x_1 & \dots & x_M \\ x_1 & x_2 & \dots & x_{M+1} \\ x_2 & x_3 & \dots & x_{M+2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{N-K-M} & x_{N-K-M+1} & \dots & x_{N-K} \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} c_{M+1} & c_{M+2} & \dots & c_{M+K} \\ c_{M+2} & c_{M+3} & \dots & c_{M+K+1} \\ c_{M+3} & c_{M+4} & \dots & c_{M+K+2} \\ \vdots & \vdots & \vdots & \vdots \\ c_{N-K+1} & c_{N-K+2} & \dots & c_N \end{pmatrix},$$

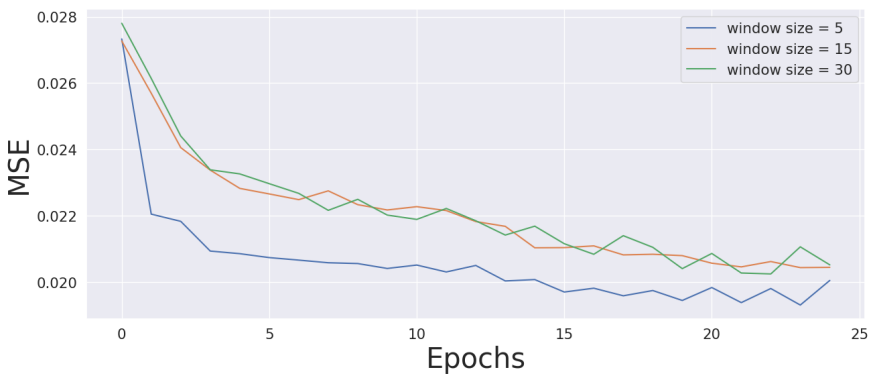
где M - размер окна, K - горизонт прогнозирования.

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

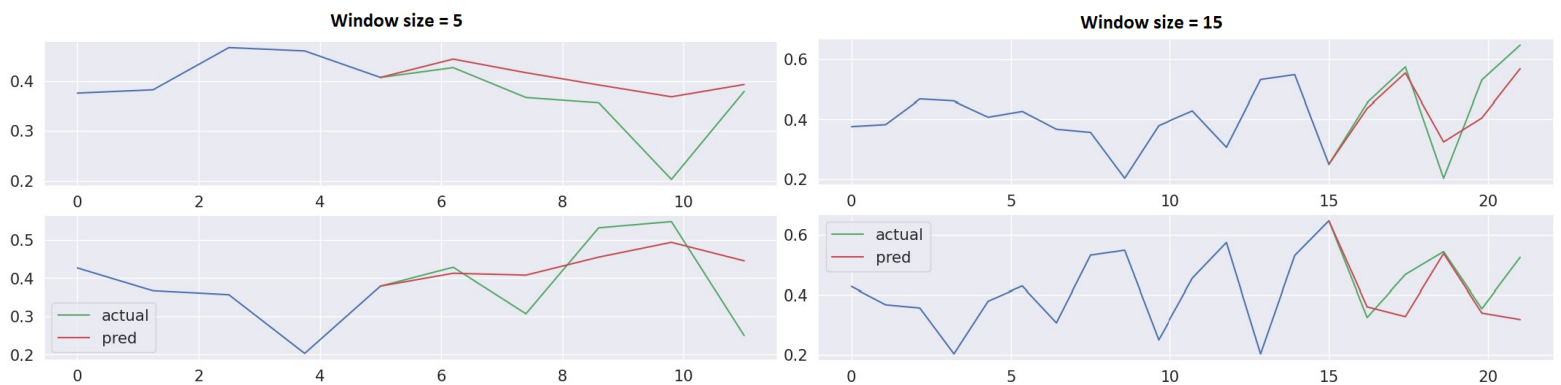
Выбор размера окна

В качестве базовой модели используется Seq2Seq архитекутра на основе LSTM. На графиках показаны метрики корреляции Пирсона и среднеквадратичной ошибки в зависимости от размера входного окна.



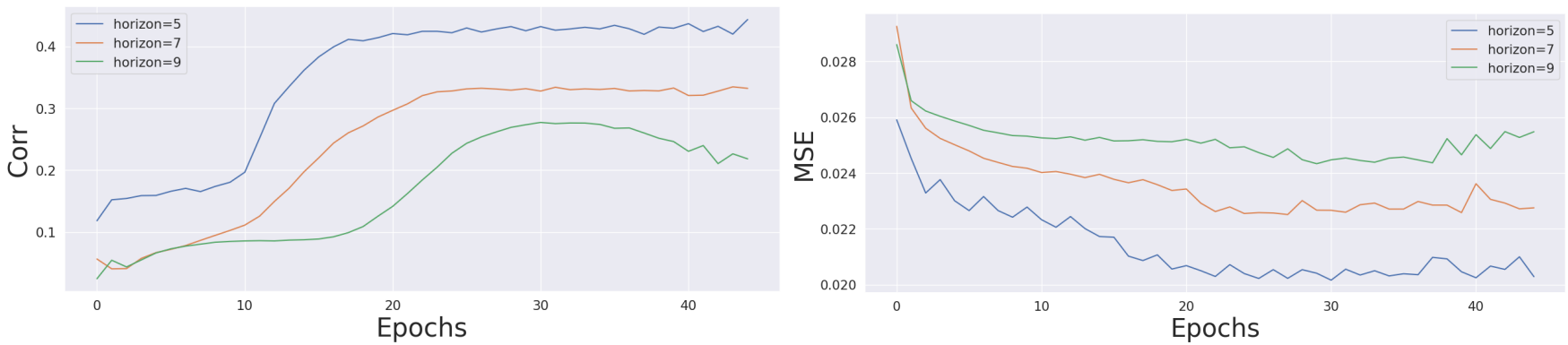


Визуализация прогнозов моделей:



Выбор горизонта прогнозирования

На графиках показаны метрики корреляции Пирсона и среднеквадратичной ошибки в зависимости от горизонта прогнозирования.



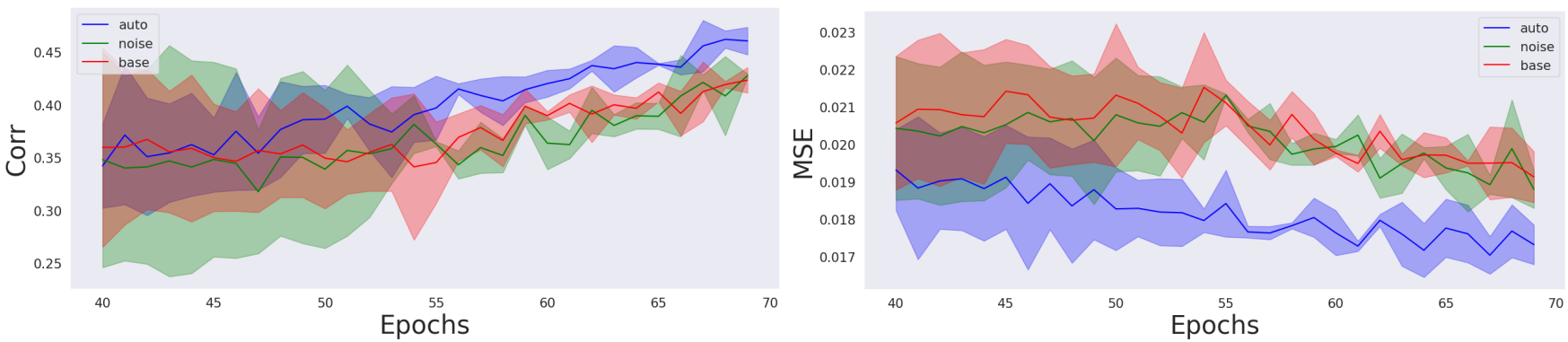
С увеличением горизонта прогнозирования качество модели ухудшается.

Анализ предложенного метода

Проводится сравнение базовой 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.

ARIMA

Порядок модели выбирается на основе критерия AIC.

Модель обучается на первых 80 % данных. Зафиксированные параметры используются для оценки качества прогнозирования на N шагов на оставшихся 20 % данных.

Сравнение результатов

Модель	Корреляция Пирсона	MSE
Seq2Seq	0.424 (+0 %)	0.019 (-0 %)
Seq2Seq + Автоследование	0.461 (+8.7 %)	0.017 (-10.5 %)
ARIMA	-	-

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 курса акций YNDX.
4. Проведен анализ выбора размера окна и горизонта прогнозирования.