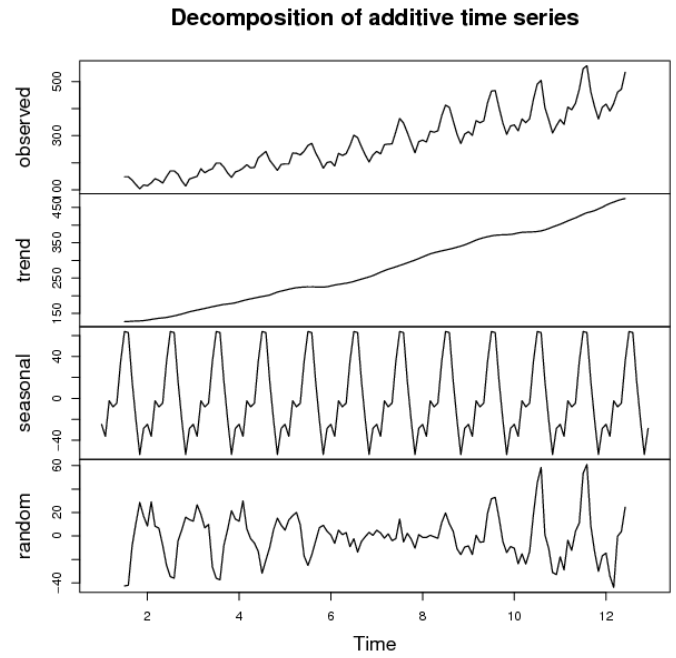
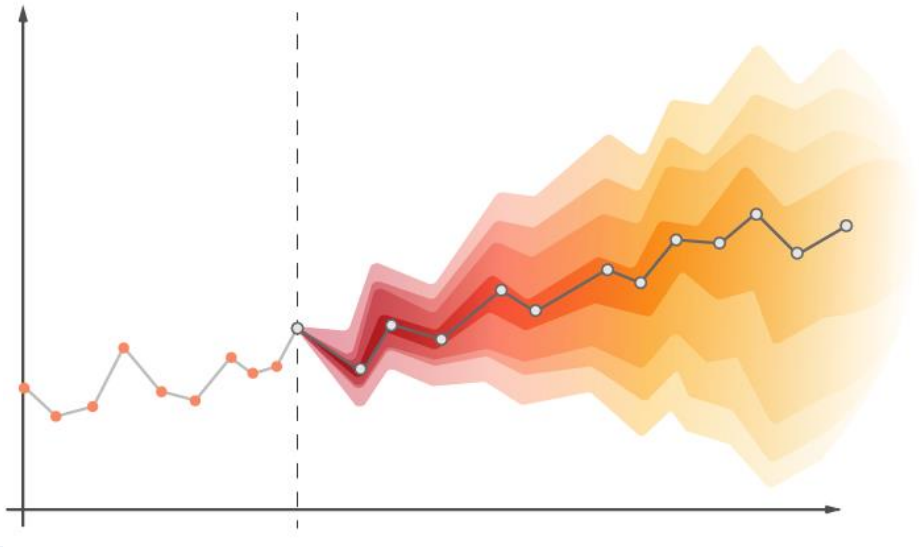
Анализ эффективности нейронных сетей для среднесрочного прогнозирования временных рядов разных частот

Хорунженко Аркадий Сергеевич
Группа 22712
Научный руководитель: **к.э.н. Макушев Василий Леонидович**
Рецензент: на данный момент отсутствует
Новосибирский Государственный университет

Актуальность

- Прогнозирование позволяет автоматизировать и оптимизировать процессы принятия решений на основе анализа изменения показателей во времени.
- Нейронные сети, как один из методов искусственного интеллекта, позволяют строить сложные модели, учитывающие множество факторов, что улучшает точность прогнозов и позволяет решать более сложные задачи.





Область применения

- Экономика и финансы: принятие решений в области инвестиций
- Медицина: прогнозирование уровня заболеваемости
- Энергетика: планирование и управление энергосистемами
- Логистика: управление трафиком, прогнозирование пробок

Цели и задачи

- Цель - исследовать эффективность нейронных сетей для прогнозирования временных рядов с учётом различной разряженности экономических данных, а также в сравнительном анализе с традиционными методами прогнозирования.

Цели и задачи

- Задачи:

1. Провести обзор литературы и проанализировать существующие методы прогнозирования временных рядов.
2. Изучить информацию о статистических методах и методах машинного обучения.
3. Подготовить данные для построения прогноза.
4. Подобрать оптимальные параметры для различных архитектур нейронных сетей и методов обучения на основе анализа результатов экспериментов с использованием различных наборов данных.
5. Разработать и апробировать модель для прогнозирования различных временных рядов, оценить ее эффективность по сравнению с традиционными методами, а также сравнить эффективность разработанных моделей на основе различных критериев качества прогнозирования.
6. Проанализировать применимость различных архитектур и методов обучения нейронных сетей для прогнозирования финансовых временных рядов, и оценить их преимущества и недостатки.
7. Сделать выводы о эффективности нейронных сетей для построения прогнозов динамики финансовых временных рядов, а также о возможных направлениях дальнейших исследований в области прогнозирования временных рядов.

Статистические модели, применявшиеся в исследовании

$$X_t = c + \sum_{i=1}^{p} a_i X_{t-i} + \varepsilon_t$$

Авторегрессионная модель
AR(p)

$$X_t = \sum_{i=1}^{q} b_i \varepsilon_{t-i}$$

Модель скользящего среднего
MA(q)

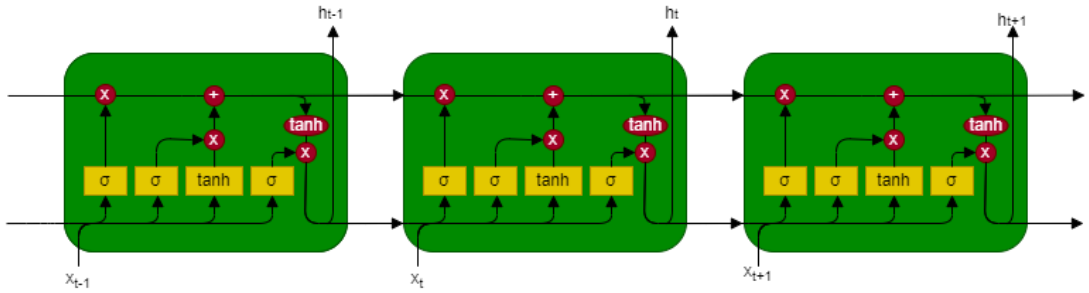
$$X_t = c + \sum_{i=1}^{p} a_i X_{t-i} + \sum_{i=1}^{q} b_i \varepsilon_{t-i} + \varepsilon_t$$

Модель авторегрессии — скользящего среднего
ARMA(p,q)

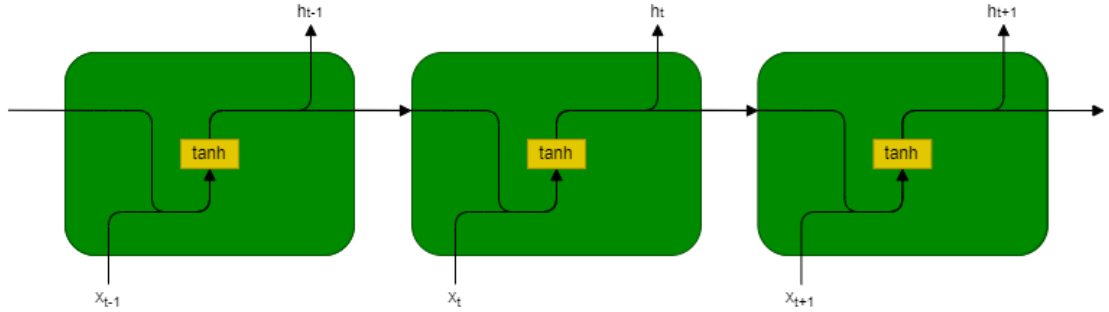
$$\Delta^d X_t = c + \sum_{i=1}^{p} a_i X_{t-i} + \sum_{i=1}^{q} b_i \varepsilon_{t-i} + \varepsilon_t$$

Интегрированная модель авторегрессии — скользящего среднего
ARIMA(p, d, q)

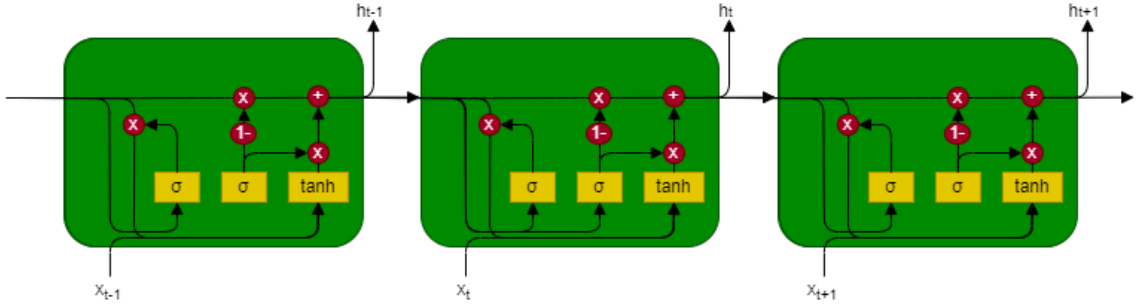
Архитектуры нейронных сетей



LSTM



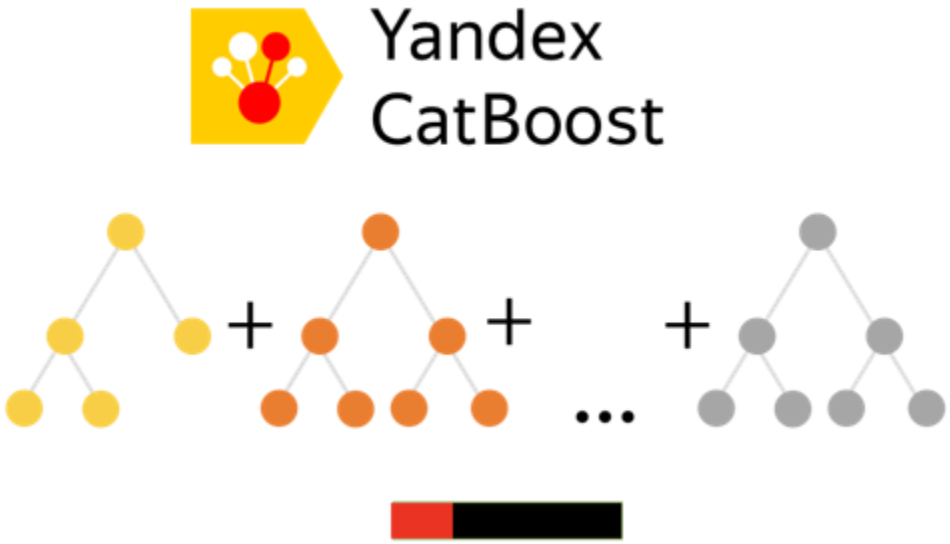
RNN



GRU

Градиентный бустинг CatBoost

По результатам исследования было решено использовать только градиентный бустинг catboost

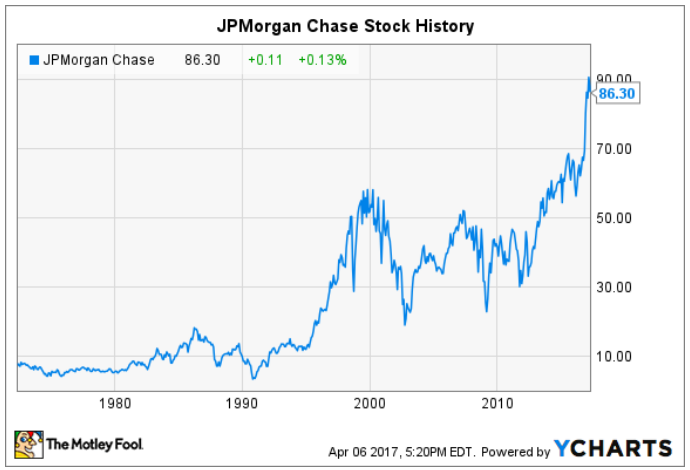


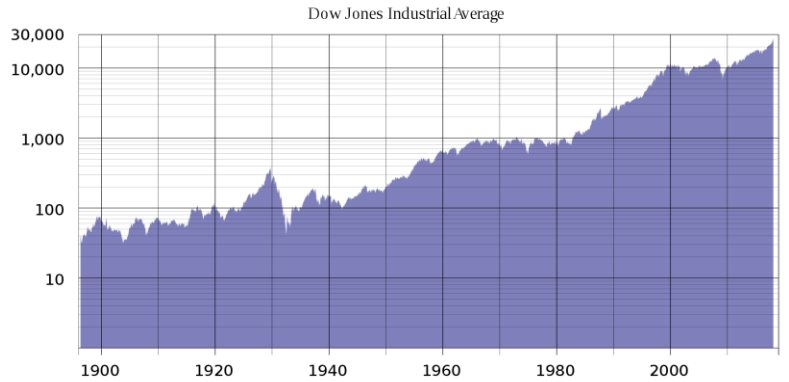
CatBoost — это библиотека градиентного бустинга, которая использует небрежные (oblivious) деревья решений, чтобы вырастить сбалансированное дерево. Одни и те же функции используются для создания левых и правых разделений (split) на каждом уровне дерева.

Ход работы

На данном этапе исследования взяты следующие наборы данных за период 2010 - 2015 г.г.:

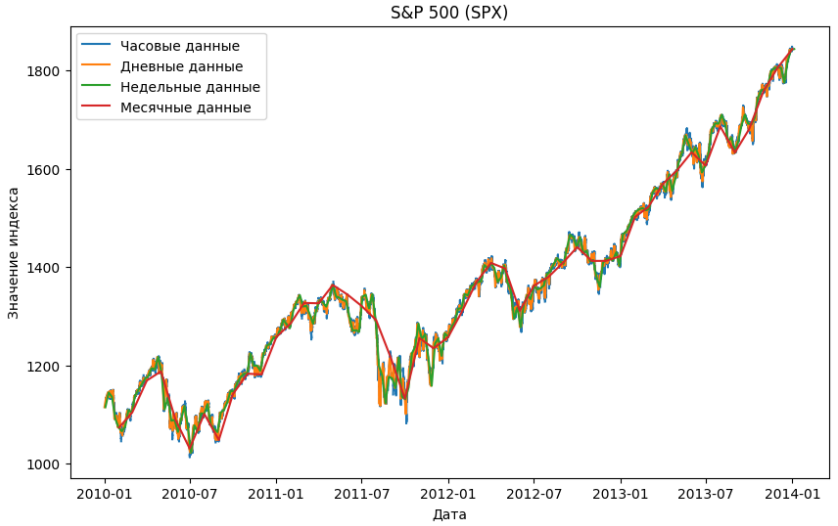
- Фондовый индекс S&P500 (SPX)
- Фондовый индекс NASDAQ Composite (IXIC)
- Промышленный индекс Dow Jones (DJI)
- Акции JPMorgan Chase & Co (JPM)

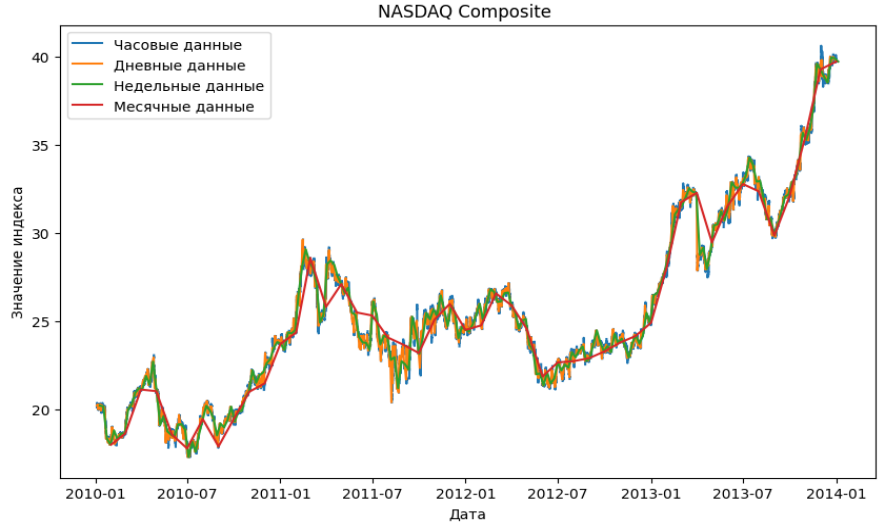


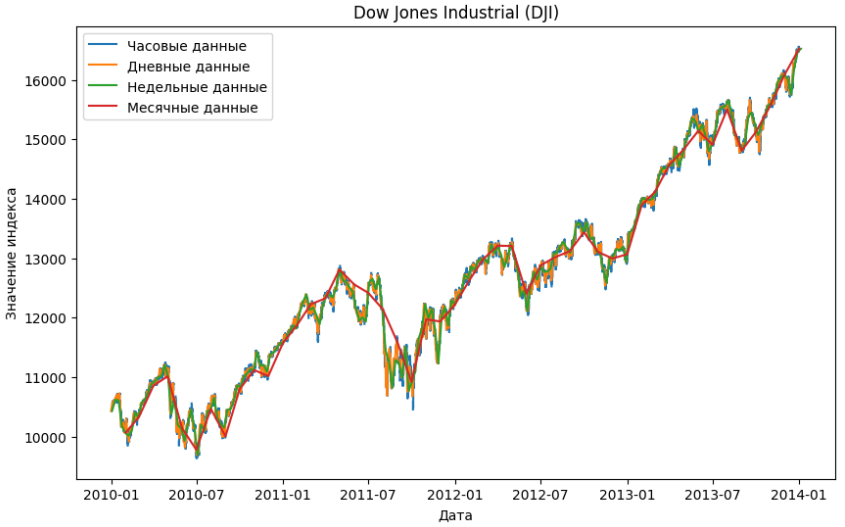


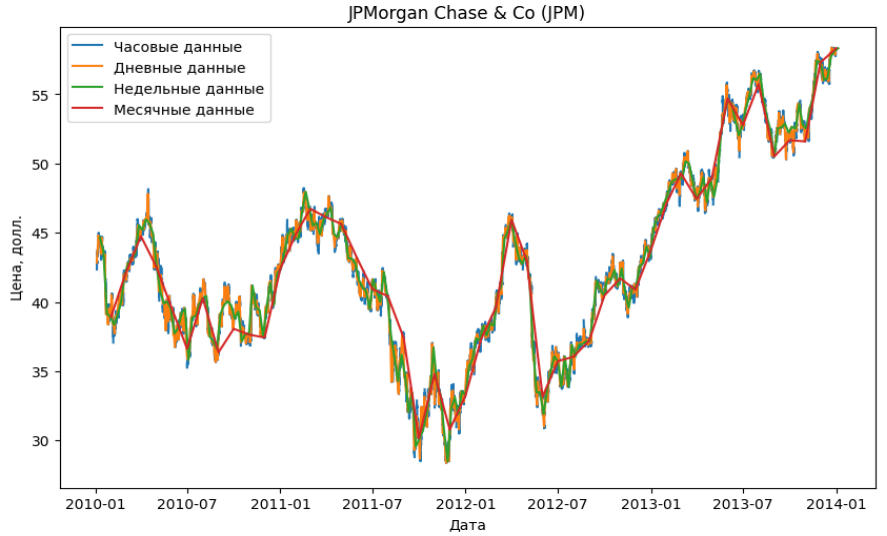
Ход работы

Исходные наборы данных были получены с минутным шагом, а затем преобразованы в данные с разной частотой: часовые, дневные, недельные и месячные.



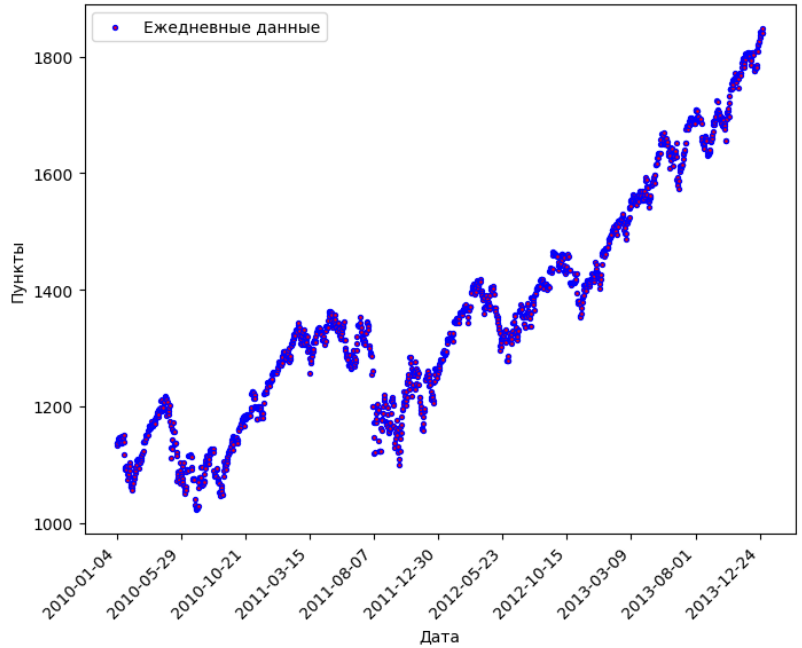




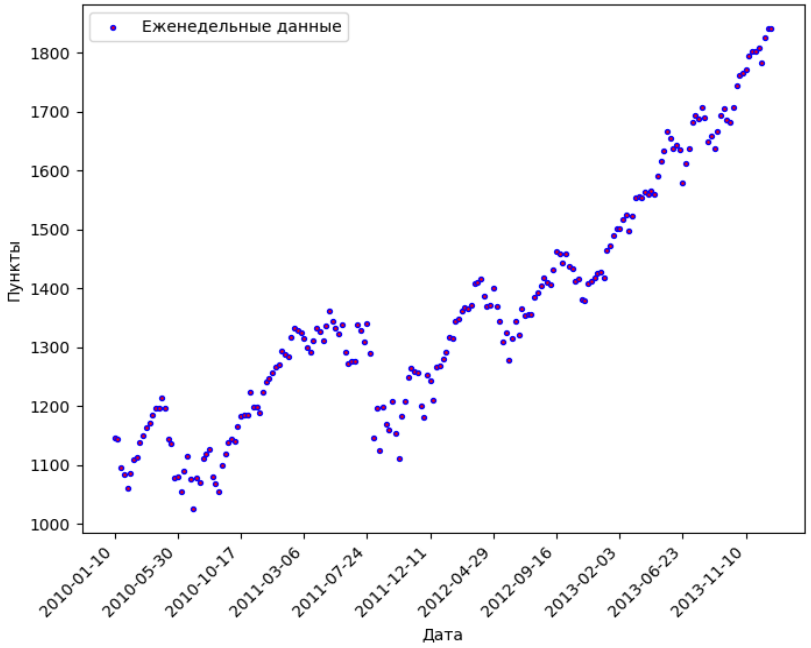


Ряды исходных данных

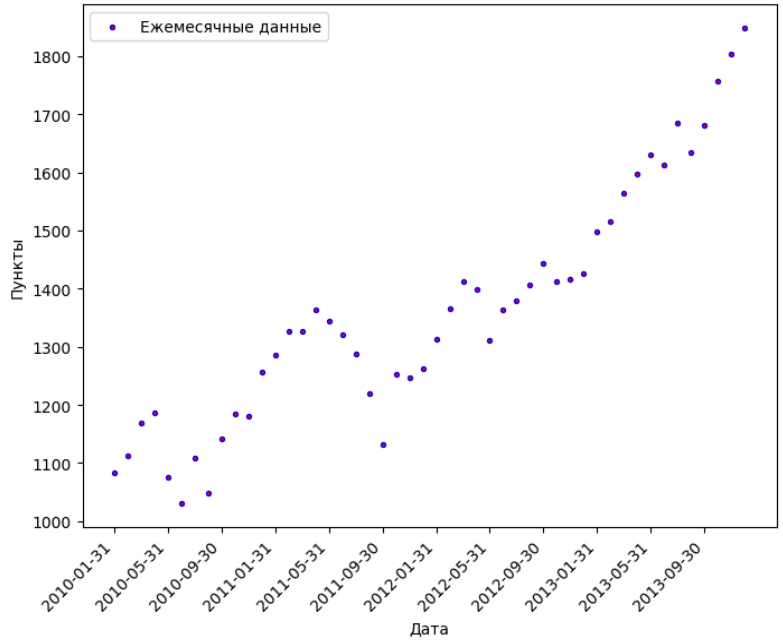
Пример данных S&P500 разной часоты



Дневные данные



Недельные данные



Месячные данные

Ход работы

Для каждого набора данных был проведён статистический анализ. Пример для S&P500 дневных данных:

Сравнение показателей среднего арифметического (mean) и медианы (median) свидетельствует о правосторонней асимметрии (т.к. mean > median);

Значение коэффициента вариации свидетельствует об однородности исходных данных (CV = 0.146 < 0.33);

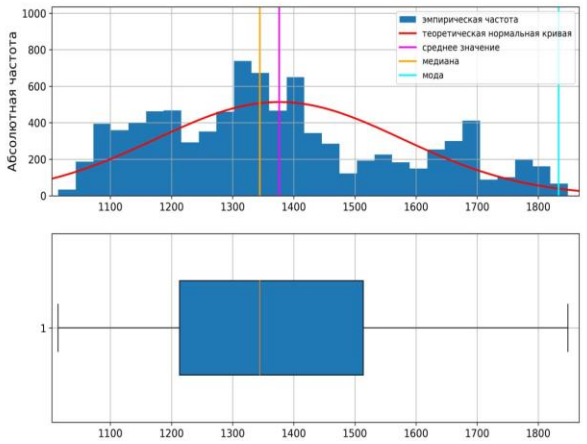
Значение показателя асимметрии skew (As) свидетельствует о значительной правосторонней асимметрии (As = 0.57, |As| > 0.5, As > 0);

Значение показателя эксцесса (Es) свидетельствует о плосковершинном распределении (Es = -0.502);

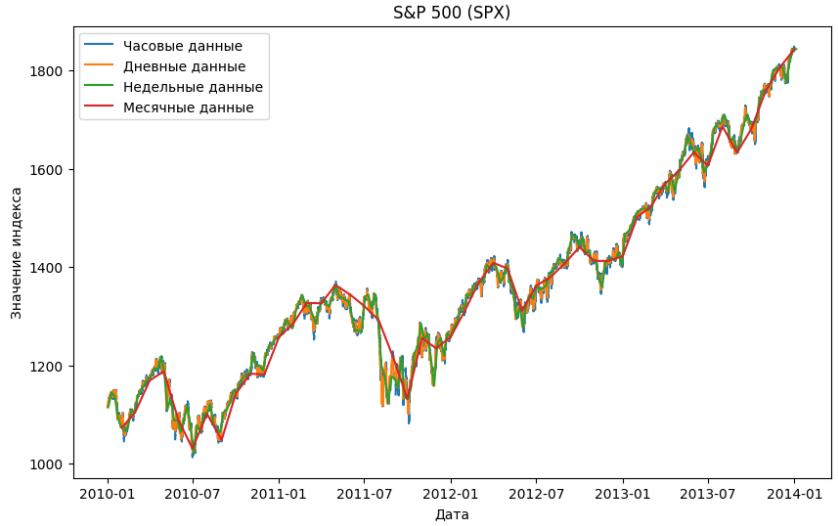
Коробчатая диаграмма показывает отсутствие аномальных значений (выбросов) для всей совокупности.

Вероятностные графики свидетельствует о том, что скорее всего закон распределения отличается от нормального.

characteristic	evaluation	conf.int.low	conf.int.high	abs.err.	rel.err.(%)	note
count	1458					
mean	1357.927	1347.739	1368.115	5.194	0.382	
median	1329.188	1322.327	1336.32	6.509	0.489	distribution is positive skewed
mode	1136.52					
variance	39329.034	36622.369	42348.819	728.315	1.852	
standard deviation	198.315	191.37	205.788	3.67251	1.852	
mean absolute deviation	131.945					
min	1022.58					
5%	1084.704					
25% (Q1)	1197.578					
50% (median)	1329.188					
75% (Q3)	1461.348					
95%	1744.561					
max	1848.36					
range = max - min	825.78					
IQR = Q3 - Q1	263.77					
CV = std/mean	0.146	0.142	0.1507	0.0028	1.892	CV <= 0.33 (homogeneous population)
QCD = (Q3-Q1)/(Q3+Q1)	0.099					
skew (As)	0.57			0.064	11.24	distribution is highly positive skewed
kurtosis (Es)	-0.503			0.128	-25.467	platykurtic distribution



коробчатая диаграмма и гистограмма распределения



Ход работы

Исходные наборы данных по очевидным причинам имеют пропуски в выходные и праздничные дни. Для решения этой проблемы были рассмотрены следующие варианты заполнения пропусков

1. Сдвиг даты
2. Заполнение пропусков значениями последнего рабочего дня
3. Линейная интерполяции (между последним и следующим рабочим днём)

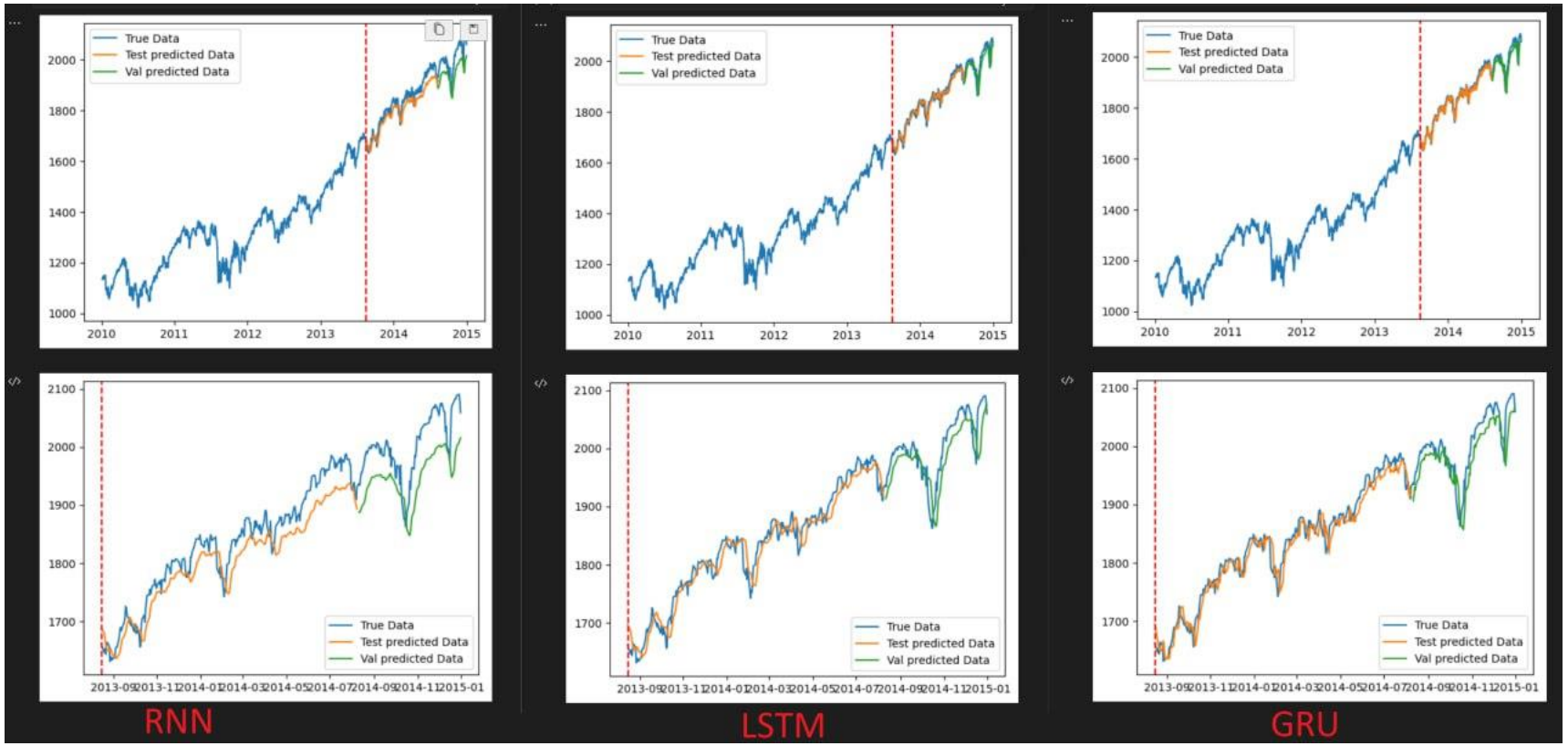
По результатам проведения пробного прогнозирования и анализа данных, было принято решение использовать линейное заполнение пропусков (вариант 3). Этот метод показал наилучшие результаты и обеспечил наиболее точные прогнозы по сравнению с другими рассмотренными вариантами. Кроме того, линейная интерполяция является более естественным способом заполнения пропусков, так как она учитывает тенденции изменения данных до и после пропущенных значений, создавая плавный переход между известными точками.

Ход работы

Результаты моделирования для датасета S&P500(2010-2015)
с применением различных моделей нейронных сетей
на малом окне прогноза

Весь датасет с участком предсказания и теста

Предсказание: Тест и валидация



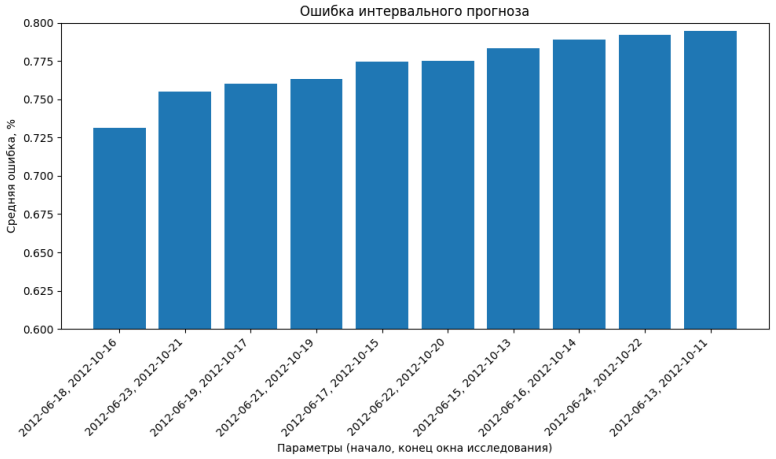
Ход работы

По результатам анализа предварительных результатов было принято решение проводить прогноз со следующими параметрами:

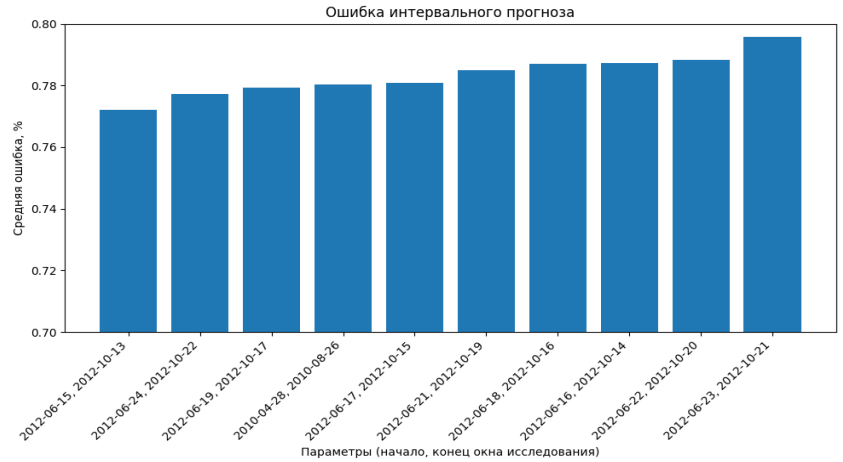
- Дневные данные: обучающая выборка – 200 (для ML 160 обучение и 40 тестирование), прогноз 90 дней.
- Недельные данные: обучающая выборка – 29 (для ML 23 обучение и 6 тестирование), прогноз 13 недель.
- Месячные данные: обучающая выборка – 7 (для ML 6 обучение и 1 тестирование), прогноз 3 месяцев.

Ход работы

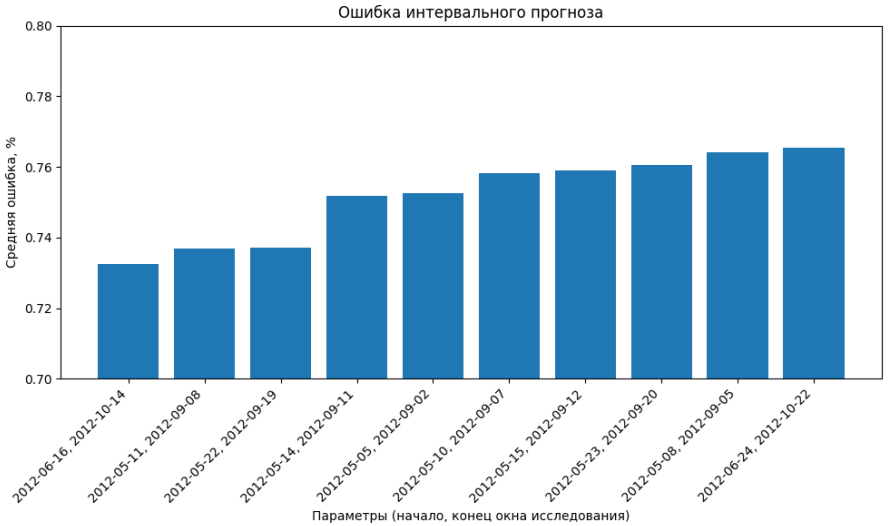
Лучшие модели машинного обучения для дневного набора данных



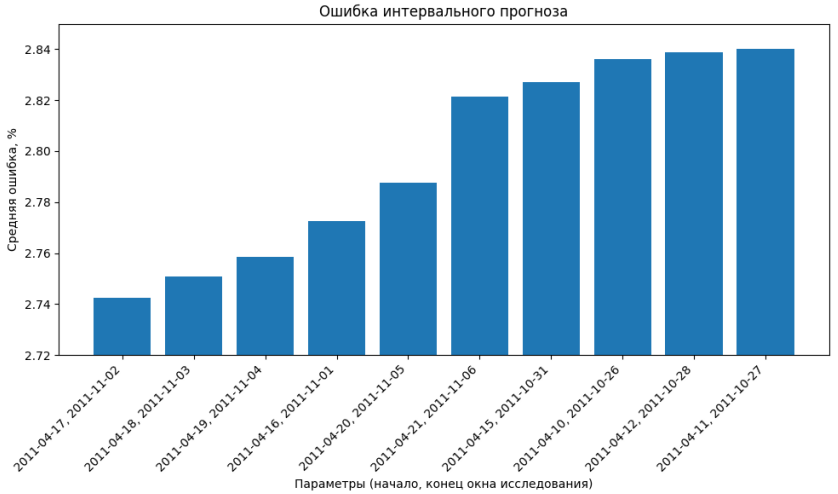
GRU



RNN

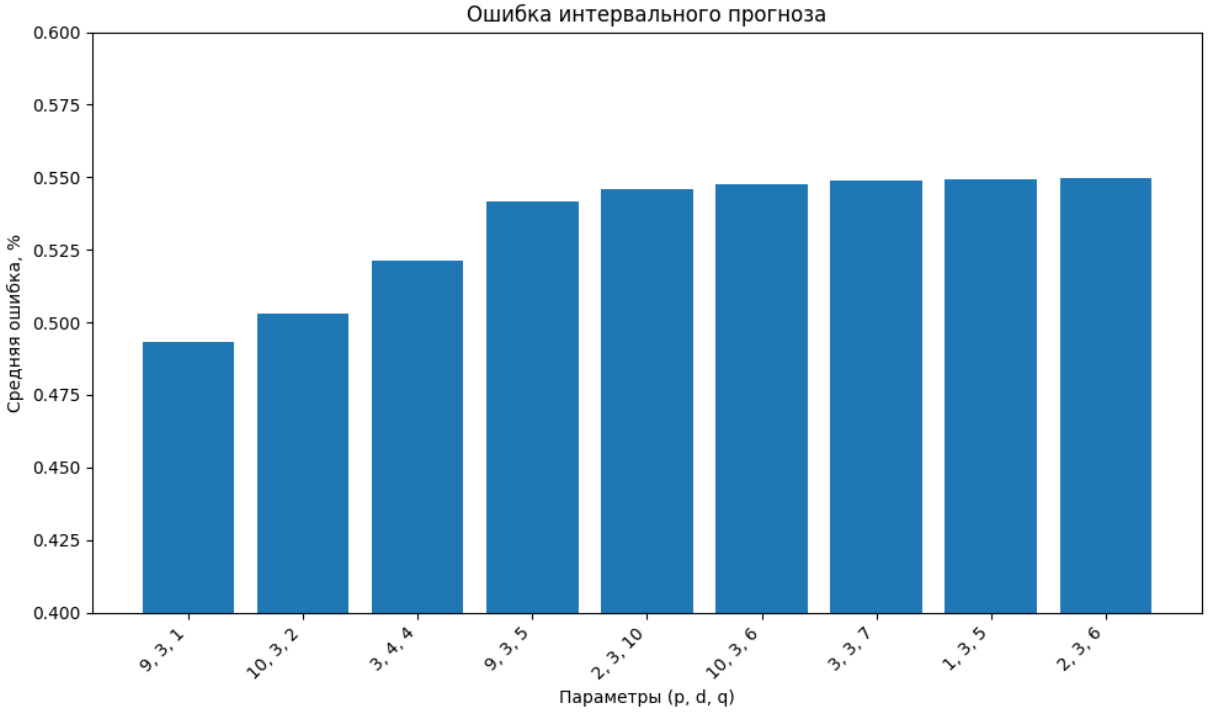


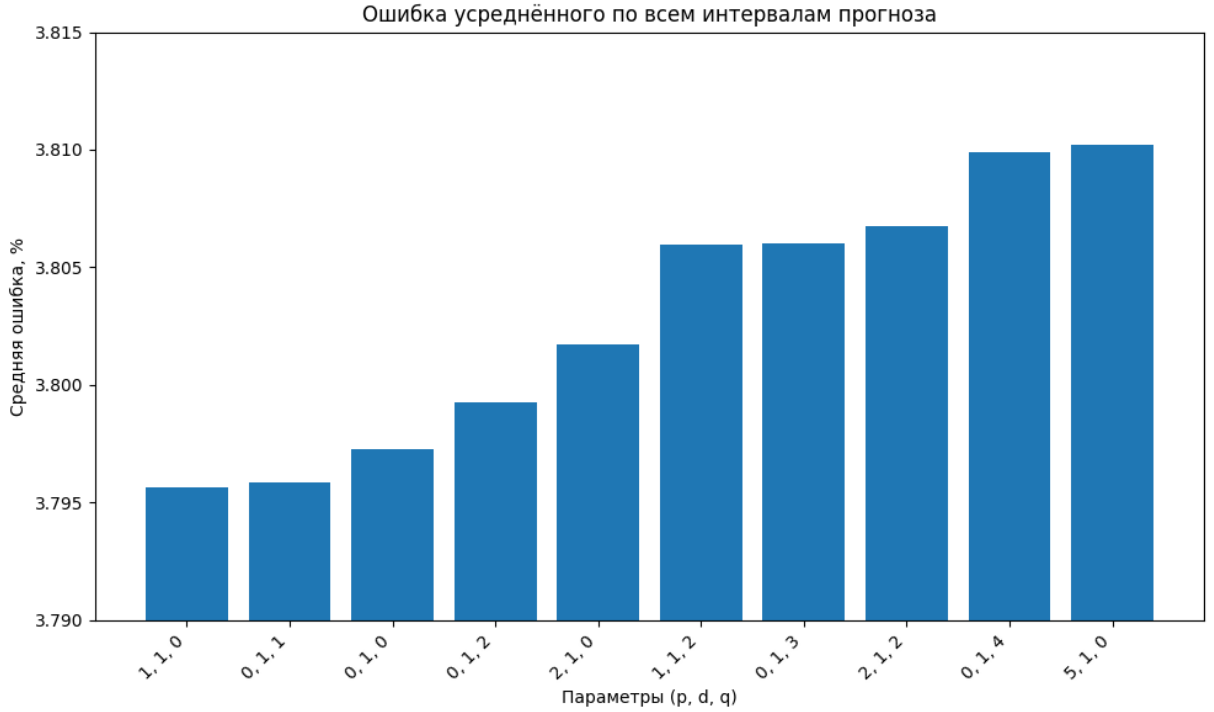
LSTM



CatBoost

Лучшие ARIMA модели

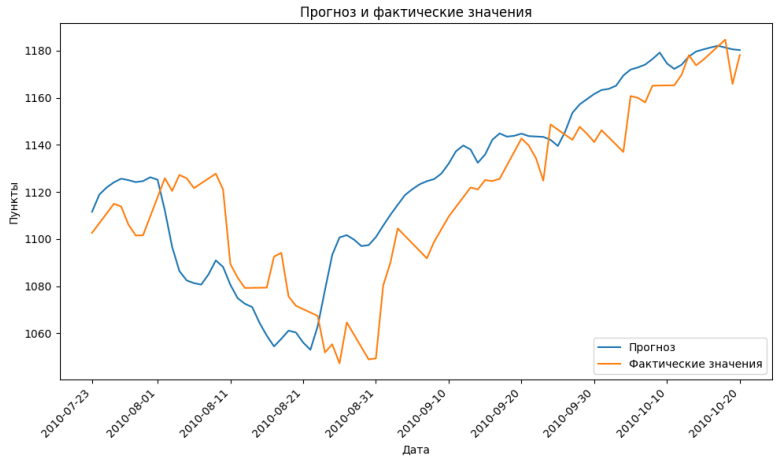




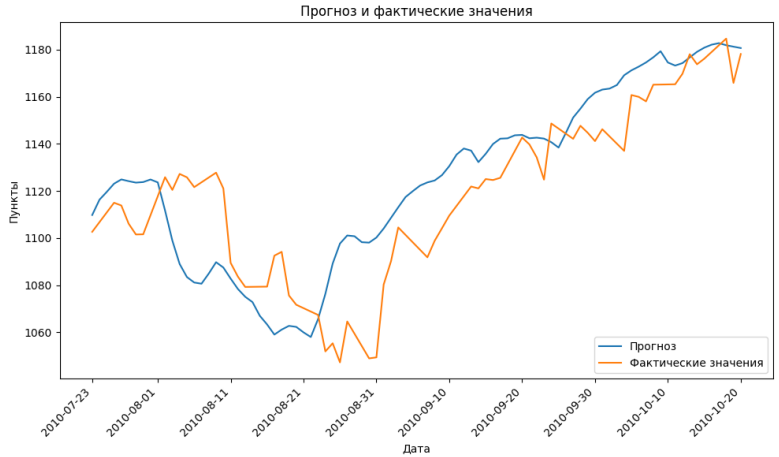
Прогноз и фактические значения

Прогноз S&P500 для дневного датасета (первые 290 дней)

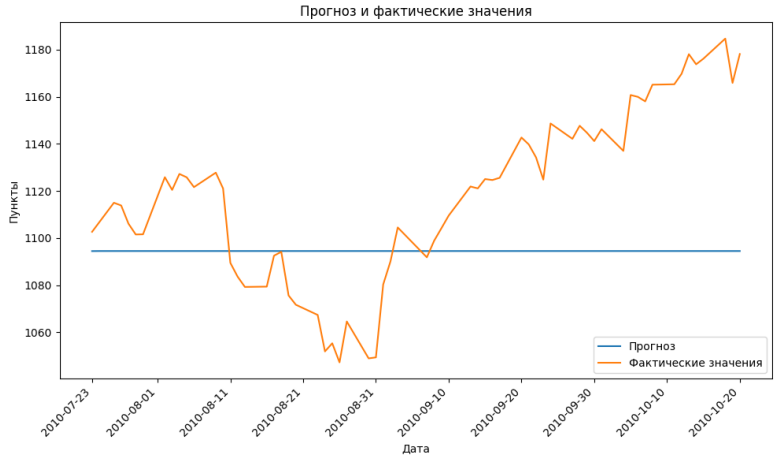
LSTM



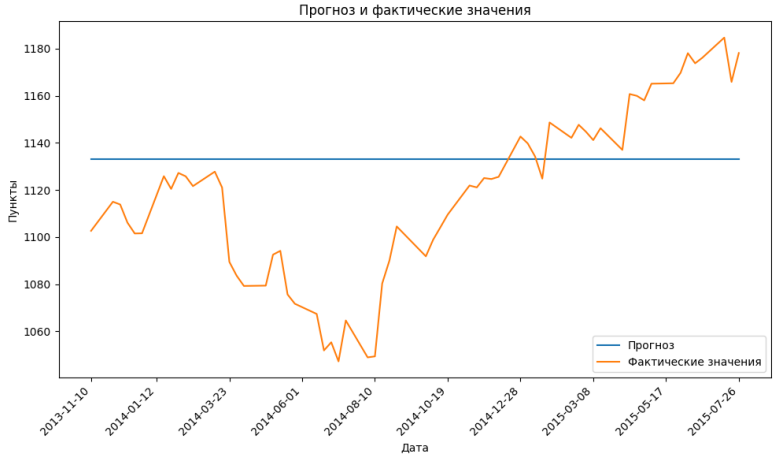
RNN



GRU

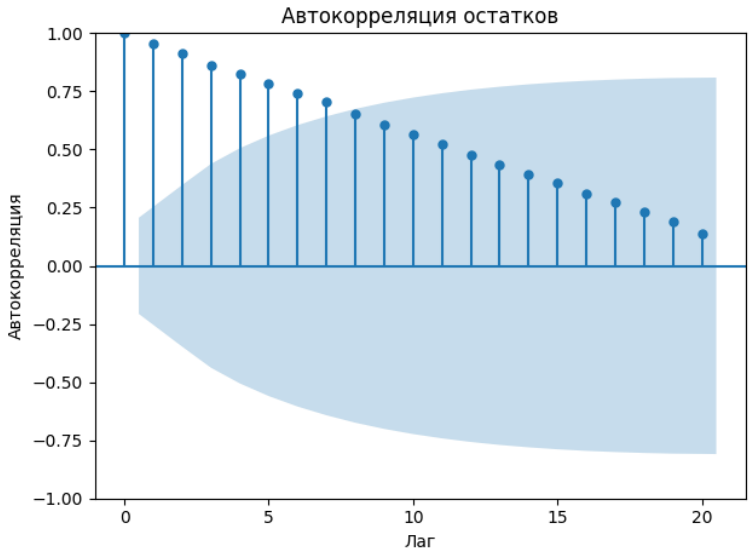


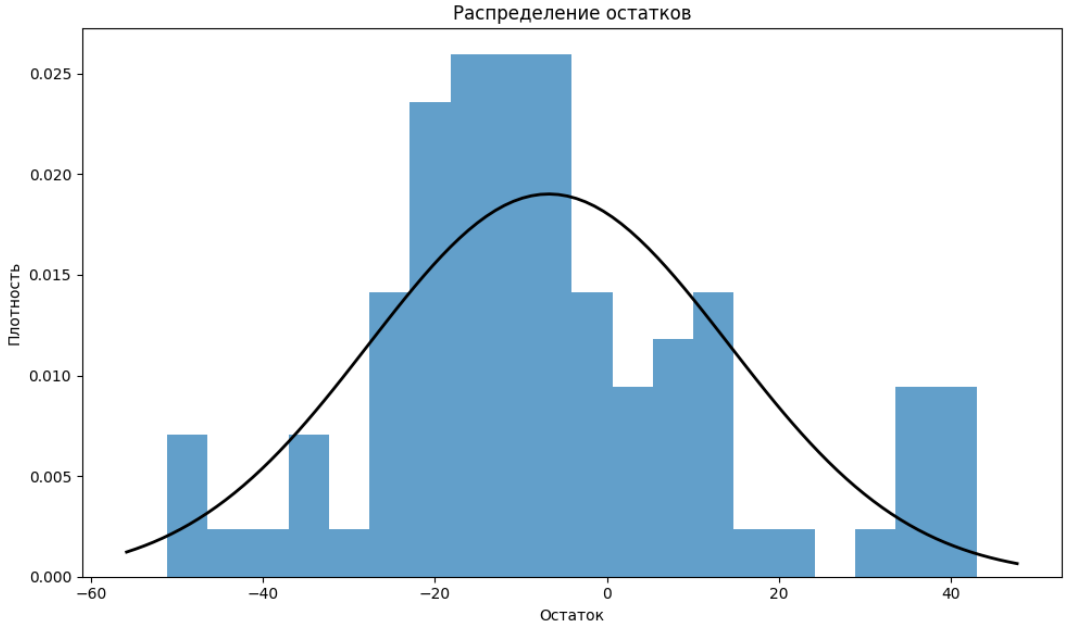
ARIMA(1,1,0)

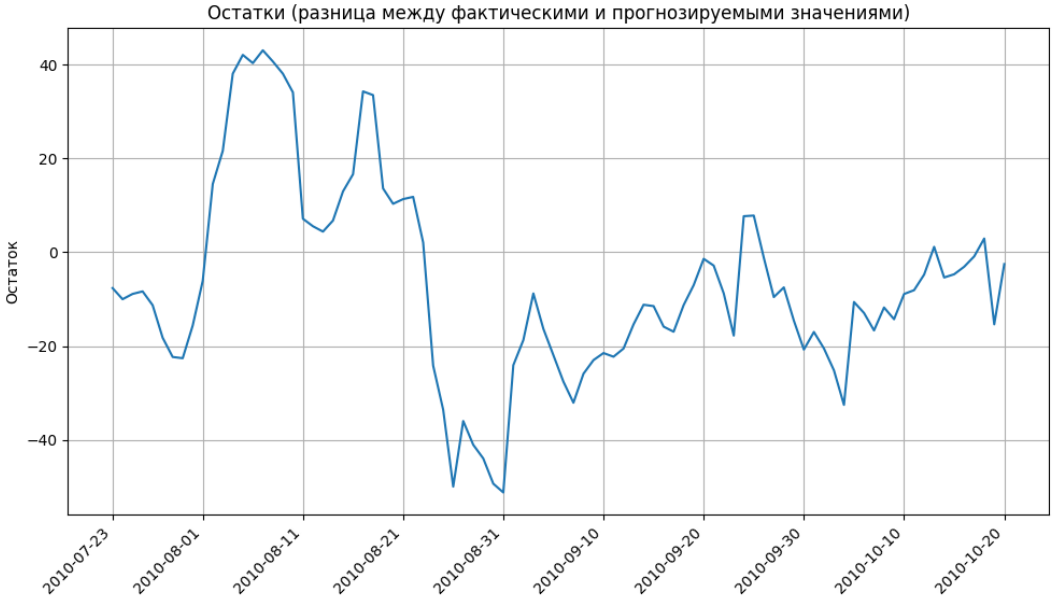


CatBoost

Проверка адекватности модели – исследование остатков







Результаты дневного прогноза S&P500

Название модели	Общее количество данных	Количество данных для обучения	Количество данных для теста	Количество данных для прогнозирования	Минимальное абсолютное отклонение, ед.	Максимальное абсолютное отклонение, ед.	Минимальное абсолютное отклонение, %	Максимальное абсолютное отклонение, %	MAPE
ARIMA(1,1,0)	290	200	-	90	0.339	90.211	0.03	7.61	3.13
GRU	290	160	40	90	0.249	52.617	0.02	5.02	1.65
LSTM	290	160	40	90	0.092	51.009	0.01	4.86	1.45
RNN	290	160	40	90	0.212	54.365	0.02	5.19	1.7
CatBoost	290	160	40	90	1.167	85.891	0.1	8.2	2.87

Результаты прогноза

Модель\Набор данных	SAND500	NASDAQ	JPM	DJI
ARIMA(1,1,0)	3.795654	6.636714	7.552283	3.45601
ARIMA(0,1,2)	3.799254	6.645998	7.551882	3.456715
ARIMA(10,0,5)	4.416458	6.521338	7.774164	3.943847
Catboost	7.592884	11.13878	13.02852	6.568545
GRU	1.713403	2.894893	3.216144	1.512084
LSTM	1.847384	3.439067	3.333534	1.61693
RNN	1.606113	1.597926	3.094778	1.455146

Средняя ошибка дневного прогноза

Модель\Набор данных	SAND500	NASDAQ	JPM	DJI
ARIMA(0,1,1)	3.857547	6.895388	7.66842	3.521682
ARIMA(1,0,0)	4.218864	6.610514	7.7427	3.823635
ARIMA(2,1,0)	3.888299	7.061621	7.893707	3.520209
Catboost	7.66917	11.19656	12.99161	6.602439
GRU	6.486913	8.883824	9.012148	5.80613
LSTM	3.712315	5.987191	7.770421	3.330939
RNN	5.626548	8.544423	10.15468	4.862223

Средняя ошибка недельного прогноза

Результаты прогноза

Модель\Набор данных	SAND500	NASDAQ	JPM	DJI
ARIMA(0,1,0)	4.555761	7.661597	9.676473	4.175079
ARIMA(1,0,0)	5.619049	8.057557	9.516905	4.931365
Catboost	6.568675	9.473433	11.03782	5.577207
GRU	5.998952	8.162465	9.445526	5.901631
LSTM	7.620905	10.09989	9.929543	6.733517
RNN	4.977977	7.918587	9.797744	4.548647

Средняя ошибка месячного прогноза

Модель\Набор данных	SAND500	NASDAQ	JPM	DJI
ARIMA(0,1,1)	3.857547	6.895388	7.66842	3.521682
ARIMA(1,0,0)	4.218864	6.610514	7.7427	3.823635
ARIMA(2,1,0)	3.888299	7.061621	7.893707	3.520209
Catboost	7.66917	11.19656	12.99161	6.602439
GRU	6.486913	8.883824	9.012148	5.80613
LSTM	3.712315	5.987191	7.770421	3.330939
RNN	5.626548	8.544423	10.15468	4.862223

Средняя ошибка недельного прогноза

Результаты

1. Реализованы и обучены модели нейронных сетей с архитектурами RNN, LTSM, GRU
2. Реализованы и обучены традиционные модели прогнозирования такие как AR, MA, ARMA, ARIMA
3. Реализован код для работы с моделью градиентного бустинга catboost
4. Реализованный код оформлен в виде jupyter notebook’ов и в виде библиотеки и будет выложен в ближайшие месяцы после экспертной валидации
5. Сформирована рекомендация о применимости нейронных сетей

Дальнейшие задачи

1. Анализ влияния добавления различных параметров на качество прогноза для определения их положительного или отрицательного влияния на результат.
2. Дополнение исследования результатами прогноза на часовых и минутных данных, которые на данный момент не были включены в анализ в связи с высокой вычислительной сложностью.
3. Анализ таких моделей как SARIMA, SARIMAX и их продолжений для сравнения их эффективности с моделями глубокого обучения.
4. Подбор оптимальных параметров для моделей машинного обучения (ML) и глубокого обучения (DL) с целью улучшения качества прогнозов.
5. Анализ эффективности реализованных моделей на датасетах часовой и минутной частоты для оценки их применимости к прогнозированию временных рядов с высокой гранулярностью.

Выводы по результатам работы

- Для всех датасетов и для всех временных периодов и для всех индексов оптимальным порядком интегрирования для авторегрессионной модели является d=1.
- Значение порядка авторегрессии для абсолютного большинства оптимальных моделей не превышает 1.
- Рассмотренные модели показали хорошее качество прогнозирования на непрерывных стационарных рядах при условии отсутствия серьезных политический и экономических потрясений, которые могут повлиять на деятельность рынка в целом.
- Модели глубокого обучения показали лучшие результаты на данных дневной частоты. При снижении частоты данных до недельной, преимущество нейронных сетей над моделью ARIMA уменьшается. На месячных данных модель ARIMA превзошла нейронные сети, указывая на эффективность классических статистических моделей при низкой частоте данных. Модель CatBoost продемонстрировала худшие результаты на всех частотах данных и не рекомендуется для среднесрочного прогнозирования временных рядов без категориальных признаков. Выбор оптимальной модели зависит от частоты данных: нейронные сети лучше подходят для высокой частоты, а классические статистические модели.