МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«ТЮМЕНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

ИНСТИТУТ МАТЕМАТИКИ И КОМПЬЮТЕРНЫХ НАУК
Кафедра программной и системной инженерии

РЕКОМЕНДОВАНО К ЗАЩИТЕ В ГЭК
Заведующий кафедрой
д.т.н., профессор
_____ А.Г. Ивашко
_____ 2023 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
магистерская диссертация

СИСТЕМА ПРОГНОЗИРОВАНИЯ ФИНАНСОВЫХ ПОКАЗАТЕЛЕЙ РЕГИОНА"

09.04.03 Прикладная информатика
Магистерская программа «Информационные системы анализа данных»

Выполнил работу
студент 2 курса
очной формы обучения *(Подпись)* Волковинский Александр Юрьевич

Научный руководитель
д.т.н, профессор *(Подпись)* Ивашко Александр Григорьевич

Рецензент
к.ф.-м.н, доцент *(Подпись)* Семихин Дмитрий Витальевич

г. Тюмень

2023

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Прогнозирование финансовых показателей является одной из ключевых задач в финансовой индустрии, это основа для финансового планирования и бюджетирования, т.е. составления стратегических, текущих и оперативных планов и бюджетов.

Макроэкономические показатели представляют собой сводные, усредненные по экономике в целом показатели объемов производства и потребления, доходов и расходов, структуры, эффективности, уровня благосостояния, экспорта и импорта, темпов экономического роста и др. По своей сути это статистические данные, которые отражают общие тенденции в экономических условиях конкретной страны, региона или сектора экономики. Они оказывают влияние на принятие органами государственной власти решений в политической, социальной, экономической сферах, являются ориентиром для предпринимателей, отдельных граждан, зарубежных партнеров в их хозяйственной деятельности. [14]

В рамках прогнозирования показателей финансового плана применяемые методы разделяют на три группы:

- методы экспертных оценок;
- методы экстраполяции;
- методы экономико-математического моделирования.

Первый метод основан на анализе мнений компетентных специалистов по вопросам динамики финансовых процессов. Проводится в форме специальных процедур анкетирования и интервьюирования. Эксперты должны обладать высокой квалификацией и владеть профессиональными знаниями и навыками в сфере управления финансами.

Второй метод – это распространение на будущее тенденций, которые сложились в ретроспективе. Целесообразность применения метода экстраполяции определяется степенью стабильности или инерционности динамики развития экономической системы. Меньше всего используются финансовые показатели

микроэкономики, которые являются менее инерционными. В свою очередь более стабильными считаются динамика развития финансовых индикаторов на уровне макроэкономики. Обычно данный метод используется в совокупности с другими.

Методы экономико-математического моделирования основываются на построении моделей, которые описывают динамику финансовых показателей относительно воздействующих на финансовые процессы факторов.

Выделяют следующие виды экономико-математического моделирования: корреляционное моделирование; оптимизационное моделирование; многофакторное экономико-математическое моделирование.

Сущность первого вида заключается в определении корреляционной зависимости между двумя исследуемыми показателями в динамике и последующем прогнозировании одного из них относительно изменения другого, принятого за базу.

Традиционные методы прогнозирования, такие как статистические модели и экспертные оценки, имеют свои ограничения и не всегда могут обеспечить высокую точность прогнозов. Одной из проблем традиционных методов прогнозирования является их ограниченность в обработке больших объемов данных. Кроме того, они не всегда учитывают сложные взаимосвязи между различными факторами, которые могут влиять на изменение финансовых показателей. Это может привести к неточным прогнозам и потере денежных средств. Еще одной проблемой является нестабильность финансовых рынков, которая может привести к неожиданным изменениям.

В связи с этим, в последние годы все большее внимание уделяется применению алгоритмов машинного обучения для прогнозирования финансовых показателей.

Целью данной работы является создание модели машинного обучения на основе макроэкономических показателей и взаимодействию с банками физических лиц для прогнозирования таких финансовых активов как:

- наличная валюта;
- долговые ценные бумаги;

Для достижения поставленной цели нужно решить следующие **задачи**:

- найти и изучить первичные источники для сбора данных;
- организовать сбор данных для разных форматов документов - web-страницы, pdf, xlsx;
- провести первичный анализ данных на наличие пропусков и формата показателей;
- провести анализ полученных входных параметров;
- разработать модель машинного обучения для прогнозирования целевых показателей на основе имеющихся данных;
- провести оценку качества обученных моделей на тестовых данных.

Для успешной подготовки и защиты выпускной квалификационной работы обучающимся использовались средства и методы физической культуры и спорта с целью поддержания должного уровня физической подготовленности, обеспечивающую высокую умственную и физической работоспособность. В режим рабочего дня включались различные формы организации занятий физической культурой (физкультпаузы, физкультминутки, занятия избранным видом спорта) с целью профилактики утомления, появления хронических заболеваний и нормализации деятельности различных систем организма.

В рамках подготовки к защите выпускной квалификационной работы автором созданы и поддерживались безопасные условия жизнедеятельности, учитывающие возможность возникновении чрезвычайных ситуаций.

ГЛАВА 1. СБОР ДАННЫХ

1.1 Макроэкономические показатели

Система национальных счетов (СНС) – это совокупность взаимосвязанных показателей и классификаций, применяемых для отражения информации о всех фазах экономического процесса и функционировании экономики страны в определенный период. Использование СНС позволяет анализировать ВВП страны и другие макроэкономические показатели в разрезе всех фаз общественного воспроизводства: производства, распределения, обмена и потребления. [20]

Основные макроэкономические показатели РФ представлены в СНС в и включают в себя:

- ВВП - показатель, отражающий рыночную стоимость всех конечных товаров и услуг (то есть предназначенных для непосредственного употребления, использования или применения), произведённых за год во всех отраслях экономики на территории конкретного государства для потребления, экспорта и накопления;
- ВРП - показатель, измеряющий валовую добавленную стоимость, исчисляемый путём исключения из суммарной валовой продукции объёмов её промежуточного потребления.;
- инфляция - устойчивое повышение общего уровня цен на товары и услуги; процесс обесценивания денег, падение их покупательной способности вследствие чрезмерного выпуска (эмиссии) или сокращения товарной массы в обращении при неизменном количестве выпущенных денег;
- безработица - наличие в стране людей, составляющих часть экономически активного населения, которые трудоспособны, но не могут найти работу;
- ключевая ставка - основной инструмент денежно-кредитной политики, ведущейся Банком России;

- коэффициент Джини - статистический показатель степени расслоения общества данной страны или региона по какому-либо изучаемому признаку. Используется для оценки экономического неравенства;
- индекс потребительской уверенности - индикатор, разработанный для измерения потребительской уверенности, определенной как степень оптимизма относительно состояния экономики, который население выражает через своё потребление и сбережение;
- цена на нефть – стоимость нефти марки Brent;
- средняя цена на 1 квадратный метр общей площади жилья

Макроэкономические показатели являются важным инструментом для принятия решений в области экономики и финансов, и их анализ может помочь в прогнозировании изменений финансовых активов на основе макроэкономических факторов.

1.2 Методы сбора данных

Данные по отдельности собирались с разных ресурсов и в разных форматах: pdf-файлы и excel-таблицы.

В связи с этим написать универсальный парсер для сбора данных не представляется возможным. Для решения проблемы было написано несколько парсеров. Далее собранные данные сохранялись в xlsx-файлы с использованием библиотеки openpyxl – пакет Python для чтения и записи в Excel. Далее полученные xlsx-файлы были объедены в единый csv-файл, где данные сгруппированы по годам и кварталам.

1.2.1 Парсер PDF-файлов

Парсер разработан с использованием библиотек tabula и pandas. Tabula - пакет Python, который может читать таблицы из PDF-файлов и конвертировать полученные данные в DataFrame пакета pandas. Pandas – пакет Python для обработки и анализа данных.

```python
import tabula

from IPython.display import display
import pandas as pd

rub = False
name = '2012_4'
country = 'РУБ' if rub else 'ИН'
# Read pdf into list of DataFrame
dfs = tabula.read_pdf(f'Docs/Наличная валюта и вклады {country}/{name}.pdf', pages='all')
```

Рисунок 1 – чтение PDF с помощью Tabula

15	Белгородская область	18 258 838 0	0	877 435 5 485 744	8 299 423
16	Брянская область	4 051 599 0	0	148 998 80 910	3 080 397
17	Владимирская область	9 420 281 450	0	476 749 248 952	5 703 896
18	Воронежская область	33 484 763 0	0	1 844 239 4 897 369	20 873 063
19	Ивановская область	5 079 320 24	0	206 524 26 542	4 114 827
20	Калужская область	7 875 708 711	0	761 866 488 676	4 633 581
21	Костромская область	4 844 205 59	0	182 612 2 149 062	2 005 610
22	Курская область	6 124 090 0	0	267 337 2 468 343	2 307 523
23	Липецкая область	7 893 945 0	0	3 291 104 575 033	3 015 169

Рисунок 2 – полученный DataFrame из PDF

```python
result = pd.concat([dfs[0], dfs[1]], ignore_index=True)
#result = dfs[0].append(dfs[1])
#display(result.head(50))
#result.to_excel('Docs/Наличная валюта и вклады ИН/temp.xlsx')
result.to_excel(f'Docs/Наличная валюта и вклады {country}/{name}.xlsx')
```

Рисунок 3 – сохранение данных в Excel

1.2.2 Парсер Excel-файлов

Парсер разработан с использованием библиотеки pandas. Tabula - пакет Python, который может читать таблицы из PDF-файлов и конвертировать полученные данные в DataFrame пакета pandas. Pandas – пакет Python для обработки и анализа данных.

```python
# загрузка существующего файла
workbook_original = load_workbook(filename='usd.xlsx')
workbook_result = load_workbook(filename='result.xlsx')

# получение нужного листа
sheet = workbook_original.active

new_sheet = workbook_result.active
c=0
i = 4
startRow = 1
for index in range(i, doc_len):
    if sheet[f'H{i}'] != None:
        new_sheet[f'D{startRow}'] = sheet[f'H{i}']
    i += 1
    startRow += 1

# сохранение изменений
workbook_result.save('sr2.xlsx')
```

Рисунок 4 – пример сохранения данных из таблицы Excel

1.3 Исходный вид входных данных и полученный набор признаков

Большинство исходных данных официальность статистики имели схожий вид: в качестве строк принимались субъекты РФ, а в колонках значения по годам (кварталам) или транспонированная таблица (годы в строках, значения в столбцах).

Валовый региональный продукт на душу населения – данные взяты с Росстата [2] в разделе «Национальные счета» в excel-формате. Единица измерения «Рубль», пропуски отсутствуют. Для соответствующего года для итоговых данных брался год в исходной таблице.

Валовой региональный продукт на душу населения по субъектам Российской Федерации в 1998-2015гг. (рублей)								
	1998г.	1999г.	2000г.	2001г.	2002г.	2003г.	2004г.	2005г.
Валовой региональный продукт по субъектам Российской Федерации (валовая добавленная стоимость в текущих основных ценах)-всего	**15 371,1**	**26 200,6**	**39 532,3**	**49 474,8**	**60 611,4**	**74 840,5**	**97 691,9**	**125 658,7**
Центральный федеральный округ	**16 564,4**	**31 118,7**	**48 205,0**	**58 851,5**	**75 739,2**	**94 244,6**	**121 487,7**	**164 887,9**
Белгородская область	12 242,8	21 398,0	27 969,5	33 126,7	41 327,4	50 271,4	75 629,4	95 911,2
Брянская область	7 659,1	11 752,4	17 413,5	21 511,9	27 020,0	31 953,4	37 719,1	49 923,4
Владимирская область	9 350,2	15 457,1	21 073,3	27 170,0	32 923,6	40 809,4	49 353,4	58 261,0
Воронежская область	9 082,1	14 808,3	20 365,1	24 905,4	34 789,6	42 237,5	49 530,0	56 534,5
Ивановская область	6 804,5	9 765,2	14 240,0	18 947,2	23 396,9	29 192,4	35 732,7	40 039,1
Калужская область	9 330,4	14 891,4	22 438,0	30 201,9	35 708,4	47 136,5	56 325,6	69 192,2
Костромская область	10 971,7	17 450,4	21 984,7	29 668,3	35 109,5	40 741,1	52 661,0	63 304,4
Курская область	11 909,9	17 093,8	23 677,7	28 946,1	36 545,7	46 131,2	63 512,1	72 995,3
Липецкая область	13 216,6	25 079,9	39 050,9	41 308,6	58 065,7	79 661,2	117 959,2	121 376,2
Московская область	12 329,5	19 753,4	26 687,7	35 569,3	47 323,5	62 023,3	79 833,2	104 738,3
Орловская область	10 641,3	17 800,8	25 168,4	31 676,1	41 322,8	49 342,3	54 740,1	64 180,4
Рязанская область	10 000,6	15 691,0	22 070,3	29 645,8	37 164,2	48 977,6	58 094,5	70 665,8
Смоленская область	10 358,8	18 562,0	25 798,1	33 575,4	39 983,1	47 084,3	54 178,9	63 687,0

Рисунок 5 – Исходная таблица с данными по ВРП субъектов

Потребительская уверенность – данные взяты с Росстата [2] в разделе «Уровень жизни» в excel-формате. Предоставлены разрезе кварталов по годам. Пропуски отсутствуют.

Индекс потребительской уверенности

баланс %

	Всего	в том числе по полу		в том числ	
		мужчины	женщины	до 30 лет	30-
1998 год					
IV квартал	-58	-57	-59	-51	
1999 год					
I квартал	-51	-49	-52	-47	
II квартал	-48	-47	-50	-44	
III квартал	-47	-45	-49	-40	
IV квартал	-35	-35	-35	-31	
2000 год					
I квартал	-25	-23	-26	-20	
II квартал	-17	-16	-18	-11	
III квартал	-18	-17	-20	-16	
IV квартал	-17	-17	-17	-12	
2001 год					
I квартал	-17	-16	-17	-14	
II квартал	-14	-12	-17	-11	
III квартал	-10	-8	-11	-4	
IV квартал	-10	-7	-12	-5	
2002 год					

Рисунок 6 - Исходная таблица с данными потребительской уверенности

Средняя цена 1 квадратный метр общей площади – данные взяты с портала ЕМИСС [11] в разделе «Средняя цена 1 кв. м общей площади квартир на рынке жилья» в excel-формате. Предоставлены разрезе кварталов и регионов РФ. Единица измерения «Рубль», пропуски отсутствуют.

		2013			
		I квартал	II квартал	III квартал	IV кв
Российская Федерация	Все типы квартир	48 794,73	49 330,65	49 959,49	
Центральный федеральный округ	Все типы квартир	57 030,89	57 624,52	58 968,35	
Белгородская область	Все типы квартир	51 219,75	52 380,54	52 686,18	
Брянская область	Все типы квартир	30 879,37	30 655,6	30 965,43	
Владимирская область	Все типы квартир	38 799,81	38 485,04	38 572,06	
Воронежская область	Все типы квартир	39 591,2	39 771,16	41 574,97	
Ивановская область	Все типы квартир	34 380,19	35 186,61	35 089,68	
Калужская область	Все типы квартир	49 207,61	49 381,18	49 925	
Костромская область	Все типы квартир	32 383,6	32 953,72	33 241,9	
Курская область	Все типы квартир	32 351,27	32 464,35	32 526,3	
Липецкая область	Все типы квартир	41 639,28	42 768,3	42 108,58	
Московская область	Все типы квартир	73 463,82	73 876,02	74 891,97	
Орловская область	Все типы квартир	36 324,91	36 697,8	36 815,73	
Рязанская область	Все типы квартир	36 055,13	35 722,9	35 781,76	

Рисунок 7 - Средняя цена 1 кв. м общей площади квартир на рынке жилья (рубль) за 2013

Индекс Джинни – данные взяты с web-сайта [12] в разделе «Статистика по России» в csv-формате. Предоставлены в разрезе годов и регионов РФ. Пропуски отсутствуют.

Коэффициент Джини рассчитывается как соотношение доходов самых богатых и самых бедных слоев население. Чем больше значение коэффициента — тем выше неравенство в обществе.

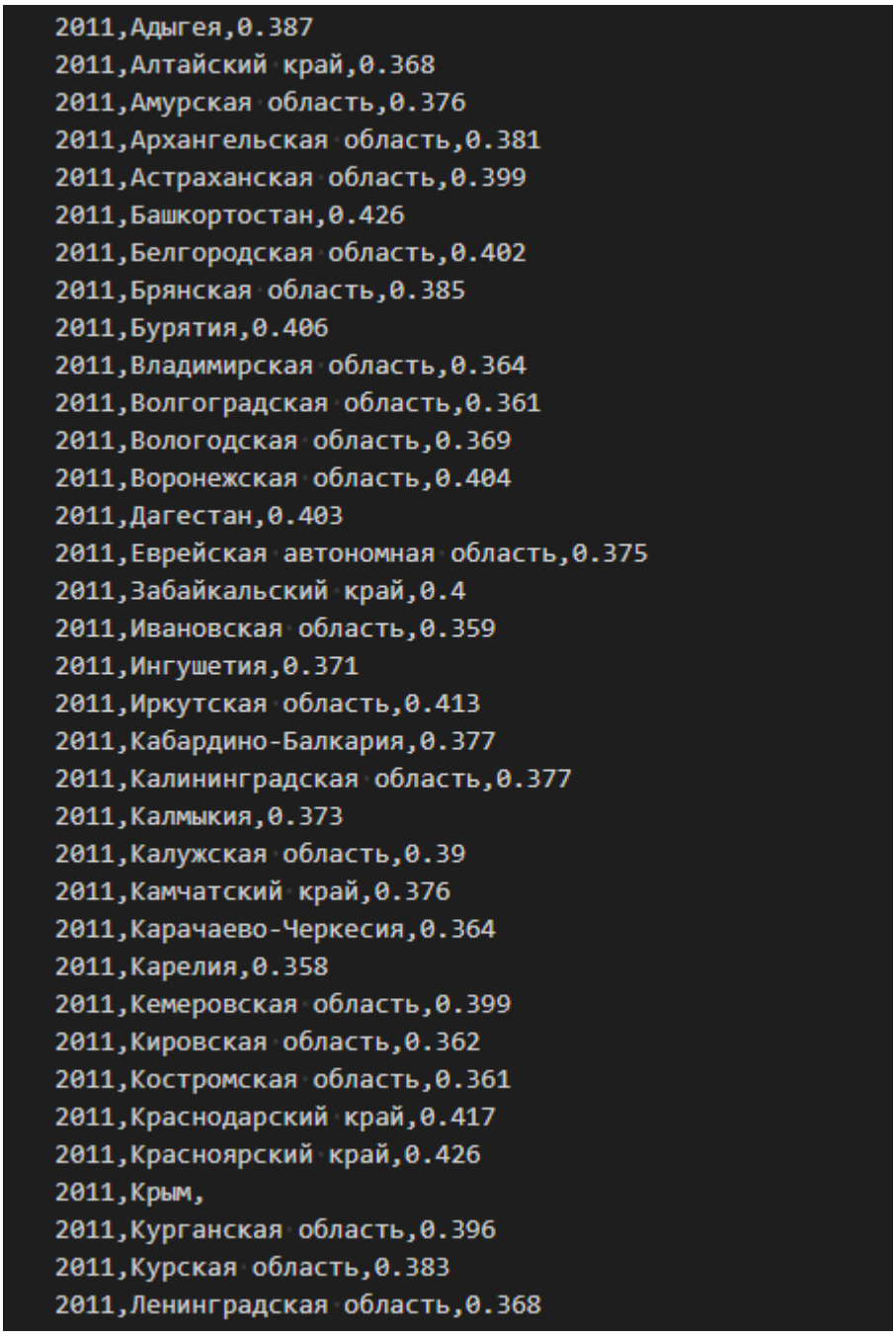


Рисунок 8 – csv-файл с данными индекса Джини по годам и регионам

Уровень безработицы – данные взяты с Росстата [2] в разделе «Статистика/Официальная статистика/Рынок труда, занятость и заработная плата/Трудовые ресурсы, занятость и безработица» в excel-формате. Предоставлены разрезе регионов по годам. Единица измерения процент (количество безработных к количеству трудоспособных), пропуски отсутствуют.

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Российская Федерация	**10,6**	**9,0**	**7,9**	**8,2**	**7,8**	**7,1**	**7,1**	**6,0**	**6,2**	**8,3**	**7,3**	**6,5**
Центральный федеральный округ	**7,8**	**6,0**	**5,1**	**5,1**	**4,7**	**4,3**	**4,0**	**3,1**	**3,6**	**5,8**	**4,6**	**4,1**
Белгородская область	6,1	6,4	8,1	8,3	6,2	5,9	5,6	4,3	3,9	4,7	5,2	4,3
Брянская область	13,4	9,8	8,6	7,4	8,9	6,8	6,8	6,4	6,5	10,7	8,0	7,1
Владимирская область	12,0	10,2	10,4	10,2	9,1	9,0	10,8	6,7	5,7	8,7	6,1	5,7
Воронежская область	10,1	9,6	8,9	8,3	8,7	7,6	5,4	5,2	5,2	8,6	7,5	6,4
Ивановская область	10,4	5,9	6,9	6,5	4,7	6,9	4,2	4,3	5,2	10,8	7,6	6,6
Калужская область	8,8	6,1	6,6	6,1	6,3	5,6	5,6	5,0	4,6	6,1	6,7	5,6
Костромская область	8,8	5,9	5,0	6,1	5,9	4,9	4,9	3,3	4,9	8,2	6,0	5,1
Курская область	10,5	10,5	7,1	8,5	7,5	7,2	7,2	4,9	6,3	8,8	8,2	6,3
Липецкая область	8,7	6,4	4,9	4,5	4,3	8,2	4,9	2,7	5,0	5,6	4,5	4,9
Московская область	7,8	5,6	4,4	4,4	3,8	3,2	3,0	2,0	2,7	4,8	3,3	3,7
Орловская область	8,6	7,9	6,7	8,0	6,2	6,3	5,9	5,5	6,0	9,8	8,9	6,3
Рязанская область	9,5	10,9	8,1	8,4	5,8	5,2	4,9	3,7	5,4	9,0	8,4	7,2
Смоленская область	12,3	9,9	11,1	10,8	9,1	7,7	8,1	6,7	7,0	7,8	7,4	7,6
Тамбовская область	8,3	12,0	9,5	9,1	9,5	8,6	8,7	9,2	9,1	9,1	7,8	6,6
Тверская область	9,1	7,5	4,7	6,6	5,4	5,8	4,5	4,1	5,1	7,7	6,6	6,0
Тульская область	10,0	5,3	6,1	5,3	4,7	5,0	2,8	2,6	3,4	6,0	5,8	5,3
Ярославская область	7,4	6,7	3,8	5,7	4,6	3,9	2,9	3,4	5,8	7,9	7,5	5,1
г. Москва	3,9	2,1	1,4	1,3	1,6	0,8	1,6	0,8	0,9	2,8	1,8	1,4

Рисунок 9 – Исходная таблица с данными по уровню безработицы в процентах

Среднедушевые доходы – данные взяты с Росстата [2] в разделе «Статистика/Официальная статистика/Население/Уровень жизни» в excel-формате. Предоставлены разрезе регионов по годам и кварталам. Единица измерения рубли, пропуски отсутствуют.

	2016 год				2017 год			
	I квартал	II квартал	III квартал	IV квартал	I квартал	II квартал	III квартал	IV квартал
Российская Федерация	**26646**	**30234**	**30540**	**36150**	**27763**	**31307**	**31325**	**36619**
Центральный федеральный округ	**35612**	**39888**	**39961**	**45556**	**36968**	**41769**	**40646**	**47758**
Белгородская область	24586	31261	30677	32775	28024	30005	27970	34763
Брянская область	19749	24404	24487	27178	22525	25038	25834	26177
Владимирская область	20672	22371	22527	23649	21913	23883	23069	24507
Воронежская область	26665	29480	28978	32061	27531	29387	29375	31173
Ивановская область	20461	22912	23883	27142	22324	22809	23773	29696
Калужская область	26010	29823	29035	30606	27882	28384	26515	31290
Костромская область	21492	22325	22464	26182	22597	23753	23603	25602
Курская область	21856	24852	25842	29000	23406	25208	25607	29361
Липецкая область	24386	27362	27543	32451	25616	28571	27935	32950
Московская область	40222	40693	39346	46025	38707	41479	41008	48218
Орловская область	19231	23371	22968	26157	21750	23610	23167	26427
Рязанская область	20295	22936	22389	29434	20685	22911	22741	29984
Смоленская область	21087	23892	22604	26606	23259	25295	24197	25592
Тамбовская область	21009	25386	26532	29659	22004	24254	26766	30437
Тверская область	20953	24519	23809	26678	21565	22998	23135	28707
Тульская область	23893	27061	26828	29229	24703	26941	27120	29310
Ярославская область	24125	27096	26487	29312	25160	27637	26279	29023
г. Москва	53279	61262	62549	71434	55971	66527	63648	76263
Северо-Западный федеральный округ	**29650**	**33432**	**31850**	**36571**	**30654**	**34631**	**32692**	**38750**
Республика Карелия	24415	26350	26020	27971	25053	26753	26670	30335
Республика Коми	29969	32857	29910	33699	29019	31761	30057	36831
Архангельская область	28051	31639	31422	34019	29563	32617	31874	33892

Рисунок 10 – Исходная таблица с данными по среднедушевым доходам

Долговые бумаги – данные взяты с Росстата [13] в excel-формате. Предоставлены разрезе регионов по годам и кварталам. Единица измерения миллион рублей, пропуски присутствуют.

Выпущенные на внутреннем рынке долговые ценные бумаги по номинальной стоимости									
(по состоянию на дату)									
Млн руб.									
	01.01.2013	01.02.2013	01.03.2013	01.04.2013	01.05.2013	01.06.2013	01.07.2013	01.08.2013	01.09.2013
Выпущенные долговые ценные бумаги - Итого									
Итого	8 516 503	8 394 063	8 520 761	8 613 444	8 772 739	8 808 909	8 913 700	8 998 623	9 096 183
краткосрочные	13 000	13 000	15 000	15 000	15 000	15 000	15 000	15 000	15 000
долгосрочные	8 503 503	8 381 063	8 505 761	8 598 444	8 757 739	8 793 909	8 898 700	8 983 623	9 081 183
остаточный срок до погашения менее 1 года									
остаточный срок до погашения более 1 года									
Центральный банк	-	-	-	-	-	-	-	-	-
краткосрочные	-	-	-	-	-	-	-	-	-
долгосрочные	-	-	-	-	-	-	-	-	-
остаточный срок до погашения менее 1 года									
остаточный срок до погашения более 1 года									
Кредитные организации	1 136 389	1 116 889	1 200 371	1 219 871	1 242 371	1 240 425	1 233 855	1 253 510	1 265 993
краткосрочные	-	-	-	-	-	-	-	-	-
долгосрочные	1 136 389	1 116 889	1 200 371	1 219 871	1 242 371	1 240 425	1 233 855	1 253 510	1 265 993
остаточный срок до погашения менее 1 года									
остаточный срок до погашения более 1 года									
Страховщики	15 000	15 000	15 000	15 000	18 000	18 000	18 000	18 000	18 000
краткосрочные	-	-	-	-	-	-	-	-	-
долгосрочные	15 000	15 000	15 000	15 000	18 000	18 000	18 000	18 000	18 000
остаточный срок до погашения менее 1 года									
остаточный срок до погашения более 1 года									
Другие финансовые организации	1 045 141	1 048 031	1 084 698	1 137 851	1 150 222	1 154 307	1 174 287	1 188 062	1 223 262
краткосрочные	-	-	-	-	-	-	-	-	-
долгосрочные	1 045 141	1 048 031	1 084 698	1 137 851	1 150 222	1 154 307	1 174 287	1 188 062	1 223 262
остаточный срок до погашения менее 1 года									
остаточный срок до погашения более 1 года									
Органы государственного управления	4 414 257	4 335 634	4 282 403	4 233 274	4 369 000	4 407 631	4 415 712	4 450 124	4 510 771
краткосрочные	-	-	-	-	-	-	-	-	-
долгосрочные	4 414 257	4 335 634	4 282 403	4 233 274	4 369 000	4 407 631	4 415 712	4 450 124	4 510 771
остаточный срок до погашения менее 1 года									
остаточный срок до погашения более 1 года									
Нефинансовые организации	1 835 416	1 808 209	1 867 989	1 937 148	1 922 846	1 918 246	2 001 546	2 018 627	2 007 857
краткосрочные	-	-	2 000	2 000	2 000	2 000	2 000	2 000	2 000
долгосрочные	1 835 416	1 808 209	1 865 989	1 935 148	1 920 846	1 916 246	1 999 546	2 016 627	2 005 857
остаточный срок до погашения менее 1 года									
остаточный срок до погашения более 1 года									
Нерезиденты	70 300	70 300	70 300	70 300	70 300	70 300	70 300	70 300	70 300
краткосрочные	13 000	13 000	13 000	13 000	13 000	13 000	13 000	13 000	13 000
долгосрочные	57 300	57 300	57 300	57 300	57 300	57 300	57 300	57 300	57 300

Рисунок 11 – Исходная таблица с данными по выпущенным долговым бумагам

Таблица 37.1

Средства клиентов в иностранной валюте по кредитным организациям, зарегистрированным в данном регионе, на 1.05.12г.

тыс. руб.

	Всего	в том числе				
		Средства бюджетов на расчетных счетах	Средства государственных и других внебюджетных фондов на расчетных счетах	Средства организаций на расчетных и прочих счетах	Депозиты и прочие привлеченные средства юридических лиц (кроме кредитных организаций)	Вклады физических лиц
1	2	3	4	5	6	7
ЦЕНТРАЛЬНЫЙ ФЕДЕРАЛЬНЫЙ ОКРУГ	5 791 385 559	29 609 776	277	777 032 977	2 706 427 132	1 969 154 525
Белгородская область	500 931	0	0	45 400	0	455 518
Брянская область	30	0	0	30	0	0
Владимирская область	429 066	0	0	19 535	0	409 475
Воронежская область	62 437	0	0	296	0	62 138
Ивановская область	626 112	0	0	17 063	0	609 048
Калужская область	773 430	0	0	12 168	103	759 434
Костромская область	6 065 758	0	0	345 827	535 409	5 176 855
Курская область	610 849	0	0	186 888	0	347 696
Липецкая область	1 508 343	0	0	901 887	117 694	488 761
Московская область	11 112 035	0	0	263 079	3 693 098	7 087 248
Орловская область	138 301	0	0	51 970	0	86 331
Рязанская область	579 067	0	0	78 550	1 119	490 239
Смоленская область	2 861 516	0	0	113 519	332 832	2 414 786
Тамбовская область	24 878	0	0	1 537	0	23 179
Тверская область	442 087	0	0	26 767	120 428	294 829
Тульская область	595 499	0	0	42 765	0	529 791
Ярославская область	310 614	0	0	52 252	58 725	199 500
г.Москва	5 764 744 606	29 609 776	277	774 873 444	2 701 567 724	1 949 719 697
СЕВЕРО-ЗАПАДНЫЙ ФЕДЕРАЛЬНЫЙ ОКРУГ	162 373 451	0	0	29 233 182	57 980 362	71 275 518
Республика Карелия	19 507	0	0	12 165	0	6 938
Республика Коми	560 551	0	0	3 592	4 417	552 292
Архангельская область	24 293	0	0	478	18 186	5 552
Вологодская область	8 523 720	0	0	2 683 421	2 963 471	2 874 843
Калининградская область	6 055 155	0	0	2 508 645	1 288 983	2 147 157
Ленинградская область	659 501	0	0	95 326	0	559 702
Мурманская область	600 469	0	0	321 448	2 936	236 944
Новгородская область	109 089	0	0	20 973	0	86 213
Псковская область	27 166	0	0	5 858	0	21 304
г.Санкт-Петербург	145 794 000	0	0	23 581 276	53 702 369	64 784 573
ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ ОКРУГ	10 474 573	0	0	2 595 155	1 707 363	5 834 552

Рисунок 12 – Исходный PDF-файл с данными по наличной иностранной валюте и вкладам за 2 квартал 2012

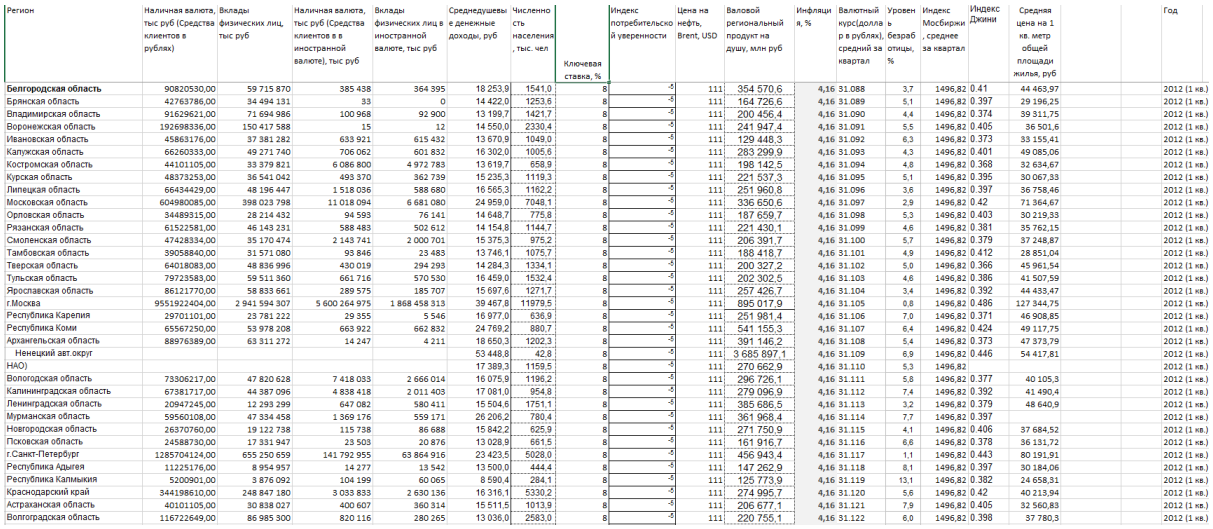
Таблица 37.2

Средства клиентов в рублях (по головным офисам кредитных организаций и филиалам, расположенным на территории региона)* на 1.05.14г.

тыс. руб.

	Всего	из них				
		Средства бюджетов на расчетных счетах	Средства государственных и других внебюджетных фондов на расчетных счетах	Средства организаций на расчетных и прочих счетах	Депозиты и прочие привлеченные средства юридических лиц (кроме кредитных организаций)	Вклады физических лиц
1	2	3	4	5	6	7
ЦЕНТРАЛЬНЫЙ ФЕДЕРАЛЬНЫЙ ОКРУГ	15 159 879 451	2 012 240	94 819	3 408 372 425	5 166 917 569	6 436 491 073
Белгородская область	110 780 813	3 416	6 795	14 479 837	12 010 280	84 135 683
Брянская область	55 922 123	2 274	0	5 013 012	4 348 090	46 423 627
Владимирская область	118 269 369	22 966	1 442	10 528 385	9 062 331	98 394 882
Воронежская область	318 448 437	58 019	0	50 251 880	50 713 042	214 014 711
Ивановская область	64 875 669	19 678	0	6 217 465	3 370 802	55 108 971
Калужская область	102 582 860	22 734	0	13 096 187	15 615 493	73 719 515
Костромская область	56 168 488	11 667	0	3 509 919	3 466 628	49 129 299
Курская область	62 234 000	8 149	0	6 506 684	4 635 223	51 011 389
Липецкая область	84 118 178	7 388	0	9 219 879	8 080 206	66 538 022
Московская область	843 381 134	555 116	14 102	180 771 622	96 260 448	561 317 372
Орловская область	47 783 956	659	0	4 767 472	2 294 152	40 342 822
Рязанская область	88 612 893	12 491	0	10 481 324	6 087 457	71 734 720
Смоленская область	60 388 316	20 440	0	6 896 220	4 983 409	48 294 933
Тамбовская область	58 098 608	3 947	0	7 136 476	3 867 436	46 637 340
Тверская область	89 513 729	16 903	0	13 919 865	3 161 427	72 240 784
Тульская область	107 624 532	14 402	1 393	11 542 386	11 820 919	84 078 173
Ярославская область	123 414 959	15 170	0	18 842 207	20 080 104	83 911 858
г.Москва	12 767 661 387	1 216 821	71 087	3 035 191 605	4 907 060 122	4 689 456 972
СЕВЕРО-ЗАПАДНЫЙ ФЕДЕРАЛЬНЫЙ ОКРУГ	2 537 693 648	3 116 467	715	542 792 243	514 437 884	1 454 731 893
Республика Карелия	37 513 726	9 176	0	3 586 527	1 747 113	31 917 885
Республика Коми	79 574 959	29 859	0	7 226 459	4 964 493	67 038 482
Архангельская область	147 442 871	7 087	0	23 493 980	36 189 614	87 588 546
Вологодская область	100 942 591	17 671	0	13 210 518	15 187 696	72 378 824
Калининградская область	77 858 092	2 536	0	14 064 392	7 877 137	55 589 593
Ленинградская область	24 059 717	29 087	0	3 718 669	4 016 771	16 233 661
Мурманская область	78 629 600	15 657	0	10 213 479	4 504 680	63 683 278
Новгородская область	35 097 052	2 019	715	5 199 254	3 033 142	26 596 937
Псковская область	35 035 246	2 124	0	4 487 832	6 067 790	24 345 126
г.Санкт-Петербург	1 921 539 794	3 001 251	0	457 591 133	430 849 448	1 009 359 561
ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ ОКРУГ	1 106 703 672	193 605	25 621	193 772 483	82 225 580	823 989 373
Республика Адыгея (Адыгея)	16 758 551	1 351	0	1 534 693	806 206	14 376 033
Республика Калмыкия	7 155 649	393	0	980 910	258 960	5 893 739
Краснодарский край	496 730 795	56 674	25 621	86 679 863	38 899 922	369 314 720

Рисунок 13 – Исходный PDF-файл с данными по наличной национальной валюте и вкладам за 2 квартал 2014



Регион	Наличная валюта, тыс руб (Средства клиентов в рублях)	Вклады физических лиц, тыс руб	Наличная валюта, тыс руб (Средства клиентов в в иностранной валюте), тыс руб	Вклады физических лиц в иностранной валюте, тыс руб	Среднедушевые денежные доходы, руб	Численность населения, тыс. чел	Ключевая ставка, %	Индекс потребительской уверенности	Цена на нефть, Brent, USD	Валовой региональный продукт на душу, млн руб	Инфляция, %	Валютный курс(доллар в рублях), средний за квартал	Уровень безработицы, %	Индекс Мосбиржи, среднее за квартал	Индекс Джини	Средняя цена на 1 кв. метр общей площади жилья, руб		Год
Белгородская область	90820530,00	59 715 870	385 438	364 395	18 253,9	1541,0	8	-5	111	354 570,6	4,16	31.088	3,7	1496,82	0.41	44 463,97		2012 (1 кв.)
Брянская область	42763786,00	34 494 131	33	0	14 422,0	1253,6	8	-5	111	164 726,6	4,16	31.089	5,1	1496,82	0.397	29 196,25		2012 (1 кв.)
Владимирская область	91629621,00	71 694 986	100 968	92 900	13 199,7	1421,7	8	-5	111	200 456,4	4,16	31.090	4,4	1496,82	0.374	39 311,75		2012 (1 кв.)
Воронежская область	192698336,00	150 417 588	15	12	14 550,0	2330,4	8	-5	111	241 947,4	4,16	31.091	5,5	1496,82	0.405	36 501,6		2012 (1 кв.)
Ивановская область	45863176,00	37 381 282	633 921	615 432	13 670,9	1049,0	8	-5	111	129 448,3	4,16	31.092	6,3	1496,82	0.373	33 155,41		2012 (1 кв.)
Калужская область	66260333,00	49 271 740	706 062	601 832	16 302,0	1005,6	8	-5	111	283 299,9	4,16	31.093	4,3	1496,82	0.401	49 085,06		2012 (1 кв.)
Костромская область	44101105,00	33 379 821	6 086 800	4 972 783	13 619,7	658,9	8	-5	111	198 142,5	4,16	31.094	4,8	1496,82	0.368	32 634,67		2012 (1 кв.)
Курская область	48373253,00	36 541 042	493 370	362 739	15 235,3	1119,3	8	-5	111	221 537,3	4,16	31.095	5,1	1496,82	0.395	30 067,33		2012 (1 кв.)
Липецкая область	66434429,00	48 196 447	1 518 036	588 680	16 565,3	1162,2	8	-5	111	251 960,8	4,16	31.096	3,6	1496,82	0.397	36 758,46		2012 (1 кв.)
Московская область	604980085,00	398 023 798	11 018 094	6 681 080	24 959,0	7048,1	8	-5	111	336 650,6	4,16	31.097	2,9	1496,82	0.42	71 364,67		2012 (1 кв.)
Орловская область	34489315,00	28 214 432	94 593	76 141	14 648,7	775,8	8	-5	111	187 659,7	4,16	31.098	5,3	1496,82	0.403	30 219,33		2012 (1 кв.)
Рязанская область	61522581,00	46 143 231	588 483	502 612	14 154,8	1144,7	8	-5	111	221 430,1	4,16	31.099	4,6	1496,82	0.381	35 762,15		2012 (1 кв.)
Смоленская область	47428334,00	35 170 474	2 143 741	2 000 701	15 375,3	975,2	8	-5	111	206 391,7	4,16	31.100	5,7	1496,82	0.379	37 248,87		2012 (1 кв.)
Тамбовская область	39058840,00	31 571 080	93 846	23 483	13 746,1	1075,7	8	-5	111	188 418,7	4,16	31.101	4,9	1496,82	0.412	28 851,04		2012 (1 кв.)
Тверская область	64018083,00	48 836 996	430 019	294 293	14 284,3	1334,1	8	-5	111	200 327,2	4,16	31.102	5,0	1496,82	0.366	45 961,54		2012 (1 кв.)
Тульская область	79723583,00	59 511 360	661 716	570 530	16 459,0	1532,4	8	-5	111	202 302,5	4,16	31.103	4,6	1496,82	0.386	41 507,59		2012 (1 кв.)
Ярославская область	86121770,00	58 833 661	289 575	185 707	15 697,6	1271,7	8	-5	111	257 426,7	4,16	31.104	3,4	1496,82	0.392	44 433,47		2012 (1 кв.)
г.Москва	9551922404,00	2 941 594 307	5 600 264 975	1 868 458 313	39 467,8	11979,5	8	-5	111	895 017,9	4,16	31.105	0,8	1496,82	0.486	127 344,75		2012 (1 кв.)
Республика Карелия	29701101,00	23 781 222	29 355	5 546	16 977,0	636,9	8	-5	111	251 981,4	4,16	31.106	7,0	1496,82	0.371	46 908,85		2012 (1 кв.)
Республика Коми	65567250,00	53 978 208	663 922	662 832	24 769,2	880,7	8	-5	111	541 155,3	4,16	31.107	6,4	1496,82	0.424	49 117,75		2012 (1 кв.)
Архангельская область	88976389,00	63 311 272	14 247	4 211	18 650,3	1202,3	8	-5	111	391 146,2	4,16	31.108	5,4	1496,82	0.373	47 373,79		2012 (1 кв.)
Ненецкий авт.округ					53 448,8	42,8	8	-5	111	3 685 897,1	4,16	31.109	6,9	1496,82	0.446	54 417,81		2012 (1 кв.)
НАО)					17 389,3	1159,5	8	-5	111	270 662,9	4,16	31.110	5,3	1496,82				2012 (1 кв.)
Вологодская область	73306217,00	47 820 628	7 418 033	2 666 014	16 075,9	1196,2	8	-5	111	296 726,1	4,16	31.111	5,8	1496,82	0.377	40 105,3		2012 (1 кв.)
Калининградская область	67381717,00	44 387 096	4 838 418	2 011 403	17 081,0	954,8	8	-5	111	279 096,9	4,16	31.112	7,4	1496,82	0.392	41 490,4		2012 (1 кв.)
Ленинградская область	20947245,00	12 293 299	647 082	580 411	15 504,6	1751,1	8	-5	111	385 686,5	4,16	31.113	3,2	1496,82	0.379	48 640,9		2012 (1 кв.)
Мурманская область	59560108,00	47 334 458	1 369 176	559 171	26 206,2	780,4	8	-5	111	361 968,4	4,16	31.114	7,7	1496,82	0.397			2012 (1 кв.)
Новгородская область	26370760,00	19 122 738	115 738	86 688	15 842,2	625,9	8	-5	111	271 750,9	4,16	31.115	4,1	1496,82	0.406	37 684,52		2012 (1 кв.)
Псковская область	24588730,00	17 331 947	23 503	20 876	13 028,9	661,5	8	-5	111	161 916,7	4,16	31.116	6,6	1496,82	0.378	36 131,72		2012 (1 кв.)
г.Санкт-Петербург	1285704124,00	655 250 659	141 792 955	63 864 916	23 423,5	5028,0	8	-5	111	456 943,4	4,16	31.117	1,1	1496,82	0.443	80 191,91		2012 (1 кв.)
Республика Адыгея	11225176,00	8 954 957	14 277	13 542	13 500,0	444,4	8	-5	111	147 262,9	4,16	31.118	8,1	1496,82	0.397	30 184,06		2012 (1 кв.)
Республика Калмыкия	5200901,00	3 876 092	104 199	60 065	8 590,4	284,1	8	-5	111	125 773,9	4,16	31.119	13,1	1496,82	0.382	24 658,31		2012 (1 кв.)
Краснодарский край	344198610,00	248 847 180	3 033 833	2 630 136	16 316,1	5330,2	8	-5	111	274 995,7	4,16	31.120	5,6	1496,82	0.42	40 213,94		2012 (1 кв.)
Астраханская область	40101105,00	30 838 027	400 607	360 314	15 511,5	1013,9	8	-5	111	206 677,1	4,16	31.121	7,9	1496,82	0.405	32 560,83		2012 (1 кв.)
Волгоградская область	116722649,00	86 985 300	820 116	280 265	13 036,0	2583,0	8	-5	111	220 755,1	4,16	31.122	6,0	1496,82	0.398	37 780,3		2012 (1 кв.)

Рисунок 14 – Полученный набор данных для предсказания показателя наличная валюта

Кредитные организации	Страховщики	Другие финансовые организации	Органы государственного управления	Нефинансовые организации	Нерезиденты	Среднедушевые денежные доходы, руб	Численность населения, тыс. чел	Ключевая ставка, %	Индекс потребительской уверенности	Цена на нефть, Brent, USD	ВВП ,млрд руб	Инфляция, %	Валютный курс(доллар в рублях), средний за месяц	Уровень безработицы, %	Индекс Мосбиржи, среднее за месяц	Индекс Джини	Средняя цена на 1 кв. метр общей площади жилья, руб	Год
1 136 389	15 000	1 045 141	4 414 257	1 835 416	70 300	**21 864,6**	143,3	5,5	-7	115,77	16640	**7,07**	30.2465	0,05	1622,13	41.9	48 794,73	**01.01.2013**
1 116 889	15 000	1 048 031	4 335 634	1 808 209	70 300	**21 864,6**	143,3	5,5	-7	115,77	16640	**7,07**	30.1631	0,06	1534,41	41.9	48 794,73	**01.02.2013**
1 200 371	15 000	1 084 698	4 282 403	1 867 989	70 300	**21 864,6**	143,3	5,5	-7	115,77	16640	**7,07**	30.8003	0,06	1460,04	41.9	48 794,73	**01.03.2013**
1 219 871	15 000	1 137 851	4 233 274	1 937 148	70 300	**23 579,1**	143,3	5,5	-6	101,9	17507	**7,24**	31.3502	0,06	1407,21	41.9	49 330,65	**01.04.2013**
1 242 371	18 000	1 150 222	4 369 000	1 922 846	70 300	**23 579,1**	143,3	5,5	-6	101,9	17507	**7,24**	31.3059	0,06	1331,43	41.9	49 330,65	**01.05.2013**
1 240 425	18 000	1 154 307	4 407 631	1 918 246	70 300	**23 579,1**	143,3	5,5	-6	101,9	17507	**7,24**	32.3068	0,05	1275,44	41.9	49 330,65	**01.06.2013**
1 233 855	18 000	1 174 287	4 415 712	2 001 546	70 300	**24 228,6**	143,3	5,5	-7	108	19003	**7,02**	32.7408	0,05	1313,38	41.9	49 959,49	**01.07.2013**
1 253 510	18 000	1 188 062	4 450 124	2 018 627	70 300	**24 228,6**	143,3	5,5	-7	108	19003	**7,02**	33.0249	0,05	1290,96	41.9	49 959,49	**01.08.2013**
1 265 993	18 000	1 223 262	4 510 771	2 007 857	70 300	**24 228,6**	143,3	5,5	-7	108	19003	**7,02**	32.6017	0,05	1422,49	41.9	49 959,49	**01.09.2013**
1 306 554	18 000	1 264 231	4 569 162	2 041 163	87 800	**25 928,2**	143,3	5,5	-11	111	20104	**6,48**	32.0992	0,05	1480,42	41.9	50 208,31	**01.10.2013**
1 372 134	18 000	1 304 152	4 695 398	2 036 763	102 800	**25 928,2**	143,3	5,5	-11	111	20104	**6,48**	32.6940	0,06	1402,93	41.9	50 208,31	**01.11.2013**
1 355 316	18 000	1 311 834	4 794 914	2 126 878	102 800	**25 928,2**	143,3	5,5	-11	111	20104	**6,48**	32.8807	0,05	1442,73	41.9	50 208,31	**01.12.2013**

Рисунок 15 – Полученный набор данных для предсказания показателя долговые ценные бумаги (кредитные организации, страховщики, другие финансовые организации, органы государственного управления, нефинансовые организации, нерезиденты)

ГЛАВА 2 ОБАБОТКА И АНАЛИЗ ДАННЫХ

После сбора и объедение данных в общий файл необходимо было провести анализ и обработку данных. Для этого использовалась библиотека Pandas.

Pandas – библиотека, написанная на Python для анализа и обработки данных. Работа с библиотекой строится поверх NumPy. Представляет спец. структуры данных и операции для обработки числовых таблиц и временных рядов. Применяется не только для сбора и очистки данных, но и для анализа и моделирования данных.

Обработка данных — это важный этап подготовки данных для машинного обучения. Она включает в себя различные методы, такие как очистка данных, преобразование данных, масштабирование и выбор признаков. Очистка данных включает в себя удаление выбросов, заполнение пропущенных значений и удаление дубликатов. Преобразование данных может включать в себя преобразование категориальных признаков в числовые, а также создание новых признаков на основе существующих. Масштабирование данных может быть полезным для нормализации данных и улучшения производительности модели. Выбор признаков включает в себя выбор наиболее значимых признаков для обучения модели. Все эти методы помогают улучшить качество модели машинного обучения.

2.1 Обработка исходного набора данных

На первом этапе на основе сформированного DataFrame была определена информация о кол-ве и проценте пропусков (Таблица 2).

Показатель	**% пропусков**	**Тип данных**
Наличная валюта, тыс руб (Средства клиентов в рублях)	0.058824	float64
Вклады физических лиц, тыс руб	0.057353	float64

Среднедушевые денежные доходы, руб	0.008824	float64
Численность населения, тыс. чел	0.000000	float64
Ключевая ставка, %	0.000000	float64
Индекс потребительской уверенности	0.000000	int64
Цена на нефть, Brent, USD	0.000000	float64
Валовой региональный продукт на душу, млн руб	0.005882	float64
Инфляция, %	0.000000	float64
Валютный курс (доллар в рублях), средний за квартал	0.000000	float64
Уровень безработицы, %	0.000000	float64
Индекс Мосбиржи, среднее за квартал	0.000000	float64
Индекс Джини	0.017647	float64
Средняя цена на 1 кв. метр общей площади жилья, руб	0.051471	float64
Регион	0	string

Таблица 1 – Процент пропусков в данных

Как видно выше, процент пропусков в показателях незначительный, поэтому пропуски заменены средним значением по показателю (Исключая значения по Москве и Санкт-Петербургу).

Низкий процент пропусков обусловлен наличием официальной статистики по годам и регионам. Значения интересующих показателей либо имеются в полном объеме, либо имеется менее 50% от общего количества данных. В случае малого объема данных показатели не рассматривались далее в построении модели, т.к.

эффективно заменить такие пропуски не предоставляется возможным. Это такие показатели как, например, инфляционные ожидания населения и инфляционные ожидания предприятий.

Из всех входных данных, только регион имеет категориальный тип данных. Категориальные данные - это данные с ограниченным числом уникальных значений или категорий (например, пол или религия). Категориальные поля могут быть текстовыми или числовыми, в которых категории закодированы числовыми кодами (например, 0 = Женский, а 1 = Мужской). Также эти данные называются качественными данными. Для работы с такими признаками надо произвести кодирование категориальных признаков - процедуру, которая представляет собой некоторое преобразование категориальных данных в численное представление по некоторым правилам, в зависимости от выбранной стратегии преобразования. [15]

Непрерывные данные — это данные, измеренные на интервальной шкале, для которых существует и порядок значений (можно сравнить больше или меньше), и расстояния между двумя значениями. Например, 1 литр и 3 литра. 3 больше, чем 1, а разница между значениями 2 литра (3-1).

Для корректной работы простым линейным моделям в качестве входных признаков мы должны передавать только численные данные, следовательно преобразование обязательно. Иначе построение модели невозможно, поскольку sklearn выдаст ошибку ValueError для столбца с регионом.

Существует два наиболее популярных кодировщика для категориальных данных – Label Encoder и One-Hot Encoder.

Label Encoder – преобразование представляет собой однозначное соответствие число – уникальное значение. Первое значение кодируется нулем, второе единицей и так далее. Такое кодирование создаст значимый недостаток – создание несуществующей зависимости в данных. Например, в результате кодирования получим Брянская область = 1, а Омская область = 33. Из этого следует, что Омская область превосходит по какому-то признаку Брянскую в 33 раза, что является неверным. Такой вывод можем сделать исходя из определения непрерывных

данных, которые являются входными для модели, мы не можем корректно интерпретировать что такое 1 и 33 для регионов.

One-Hot Encoder - тип кодирования, который основывается на создании бинарных признаков (столбцов), которые показывают принадлежность к уникальному значению. Например, для строки Омской области единица будет только в соответствующем столбце, во всех остальных будет 0. Главный недостаток One-Hot Encoder'a - существенное увеличение объема данных, так как большое количество уникальных значения признака приведет к большому количеству новых столбцов. Например, для 10 уникальных значений признака будет добавлено 10 новых столбцов.

Label Encoder и One-Hot Encoder представлены в библиотеке sklearn, для кодирования региона выбран способ One-Hot. При кодировании региона в датасет добавлено столбцов, равное количеству уникальных регионов в наборе данных.

2.2 Анализ взаимосвязей входных признаков

Для полученного набора данных следует определить связь между показателями в наборе данных. Для этого используется корреляционная матрица.

Корреляционная матрица — это таблица, которая показывает связь между парами переменных в наборе данных. Она используется для изучения того, какие переменные коррелируют друг с другом и насколько сильно. Каждый элемент матрицы представляет собой коэффициент корреляции между двумя переменными. Коэффициент корреляции может быть положительным, отрицательным или равным нулю, что указывает на отсутствие корреляции между переменными. Корреляционная матрица может быть использована для определения наиболее важных переменных в наборе данных и для поиска скрытых связей между переменными.

Для формирования и отображения корреляционной матрицы был использован пакет Python – searbon. Seaborn – библиотека для Python, имеющая возможность визуализации данных и тесно связана с объектами DataFrame библиотеки Pandas.

```
import seaborn as sns

sns.set(rc={'figure.figsize':(12,7)})
# корреляционная матрица
corr = df.corr()
sns.heatmap(corr, cmap="YlGnBu", annot=True)
```

Рисунок 16 – Построение корреляционной матрицы

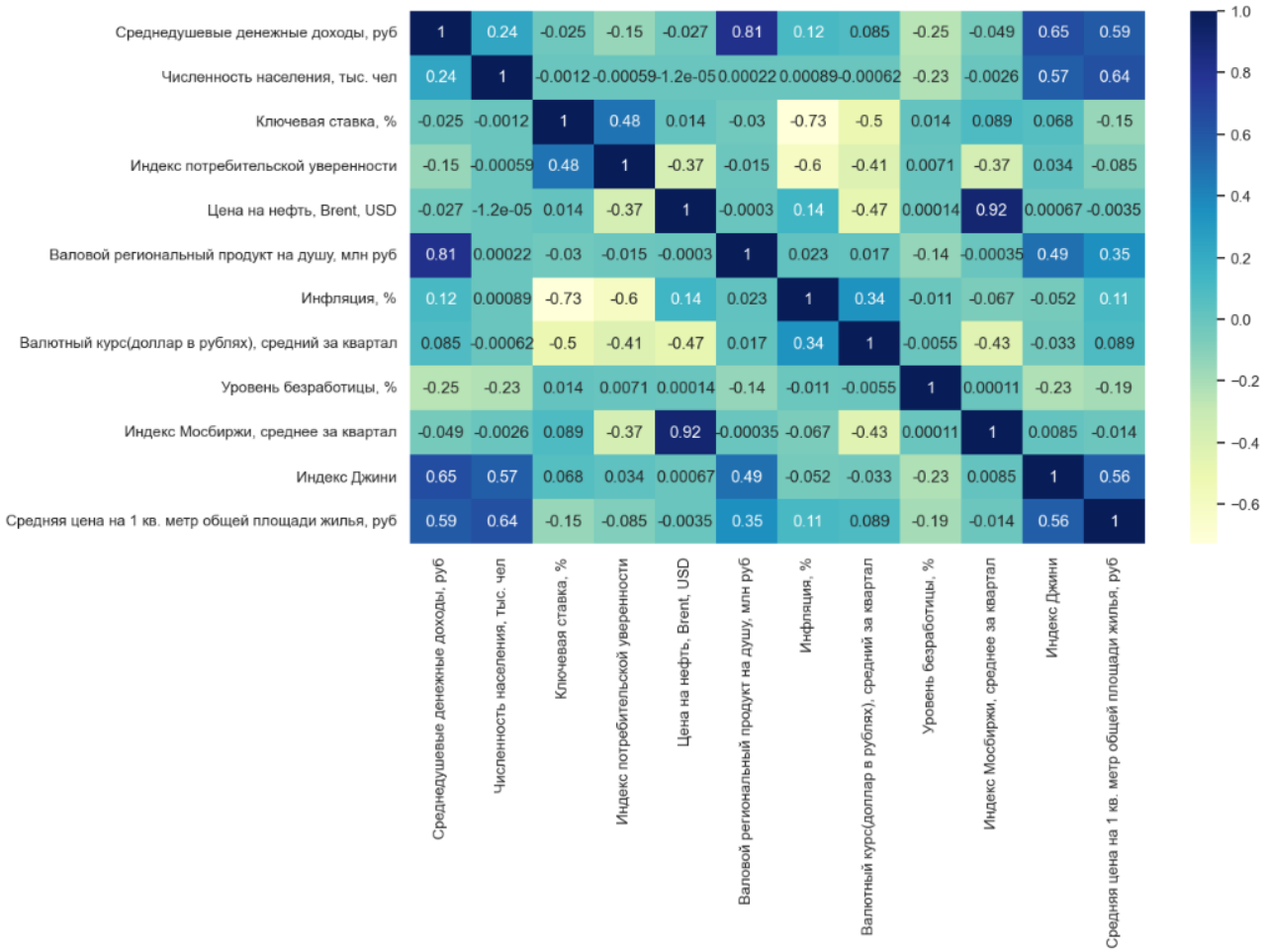


Рисунок 17 – Корреляционная матрица входных параметров

Простой коэффициент корреляции (Пирсона) вычисляется по формуле:

$$r = \frac{\sum_{i=1}^{n}(x_n - \bar{x})(y_n - \bar{y})}{n\sigma_x\sigma_y},$$

Где n— число статистических наблюдений, x и y — случайные переменные. Значения коэффициента корреляции всегда расположены в диапазоне от -1 до 1 и интерпретируются следующим образом:

- если коэффициент корреляции близок к 1, то между переменными наблюдается положительная корреляция. Иными словами, отмечается высокая степень связи между переменными. В данном случае, если значения переменной x будут возрастать, то и переменная y также будет увеличиваться [16];
- если коэффициент корреляции близок к -1, это означает, что между переменными имеет место сильная отрицательная корреляция. Если значение x будет возрастать, то y будет уменьшаться, и наоборот [16];
- промежуточные значения, близкие к 0, будут указывать на слабую корреляцию между переменными и, соответственно, низкую зависимость. Иными словами, поведение переменной x не будет совсем (или почти совсем) влиять на поведение y (и наоборот) [16].

Очевидно, что если корреляция между переменными высокая, то, зная поведение входной переменной, проще предсказать поведение выходной, и полученное предсказание будет точнее (говорят, что входная переменная хорошо «объясняет» выходную). Однако, чем выше корреляция наблюдается между переменными, тем очевиднее связь между ними, например, взаимозависимость между ростом и весом людей.

На полученной матрице отметим признаки с коэффициентом корреляции более 0.5:

Среднедушевые доходы и ВРП на душу – 0.81;

Среднедушевые доходы и Индекс Джини – 0.65;

Среднедушевые доходы и Средняя цена за квадратный метр – 0.59;

Численность населения и Индекс Джини – 0.57;

Численность населения и Средняя цена за квадратный метр – 0.64;

Цена на нефть и Индекс Мосбиржи – 0.92

Индекс Джини и Средняя цена за квадратный метр – 0.56.

Отдельно отметим связь ключевой ставки и инфляции – коэффициент корреляции составляет -0.73, что говорит об отрицательной корреляции.

Выделив признаки с высокой корреляцией, получим 4 сильно связанных признака: индекс Джини, Цена на нефть, индекс Мосбиржи, среднедушевые доходы.

Сильная взаимосвязь входных признаков в линейных классификаторах может привести к проблемам мультиколлениарности и переобучения, а также косвенно говорить о «проклятии размерности».

Мультиколлинеарность - тесная корреляционная взаимосвязь между отбираемыми для анализа факторами, совместно воздействующими на общий результат, которая затрудняет оценивание регрессионных параметров.

Для дальнейшего анализа этих признаков, рассмотрим их значение и способ расчёта.

Индекс Мосбиржи - взвешенный по капитализации композитный индекс, рассчитываемый на основе цен наиболее ликвидных российских акций крупнейших и динамично развивающихся российских эмитентов, представленных на Московской бирже. Российский рынок ценных бумаг традиционно в большей степени состоит из акций компаний — экспортеров нефти, газа, металлов, удобрений и т. д. Именно доходы от экспорта сырья являются ключевым источником роста для российской экономики. С помощью множества регуляторных и рыночных механизмов этот доход перераспределяется в иные сектора экономики, к которым относятся другие компании из индекса: банки, розничные магазины, поставщики интернет-услуг, телекоммуникационные компании, компании из сферы жилой недвижимости и другие. [17] Из этого следует, что индекс уже учитывает цену на нефть, а значит, можем удалить этот входной признак.

Стоимость квадратного метра напрямую зависит от региона и средних доходов этого региона, о чем говорит связь признака со среднедушевыми доходами. Явный пример Москва и областные центры России. Также уберем этот признак.

Рассмотрим Индекс Джини, численность населения и среднедушевые доходы. Геометрически индекс Джини это площадь под кривой Лоренца между долей доходов и долей населения в процентах, следовательно индекс уже учитывает численность населения и доходы населения, являясь некой обобщенной экономической характеристикой между этими признаки.

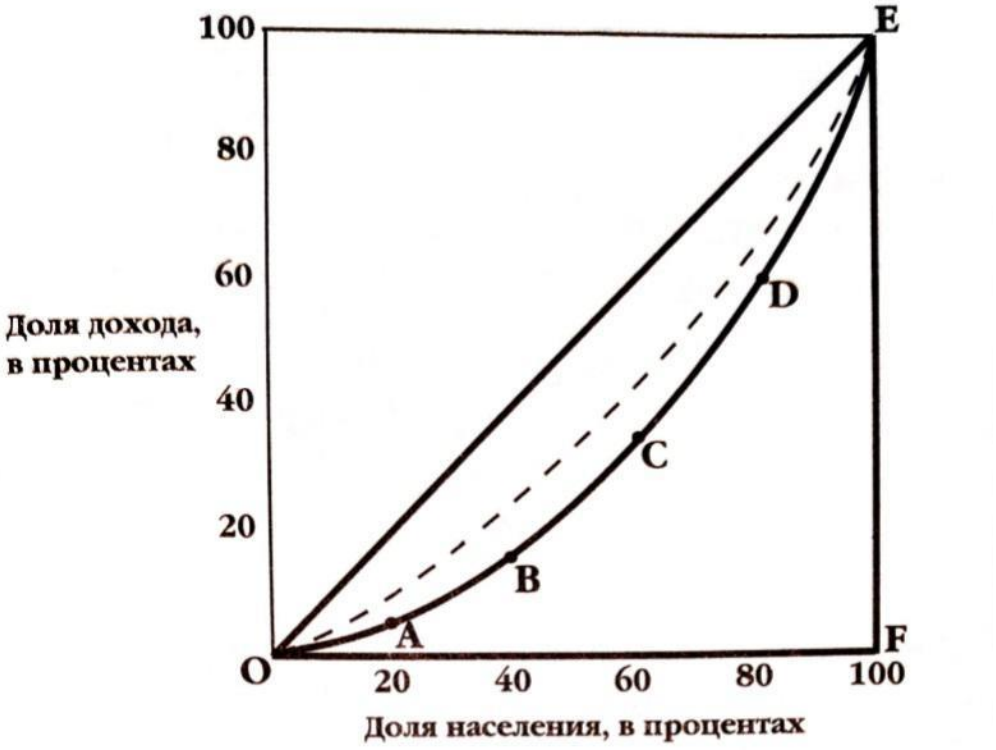


Рисунок 18 - кривая Лоренца

Исходя из того, что признаки уже учитываются, уберем их из входных параметров для моделей.

Проведя исключение признаков, составим новую корреляционную матрицу.

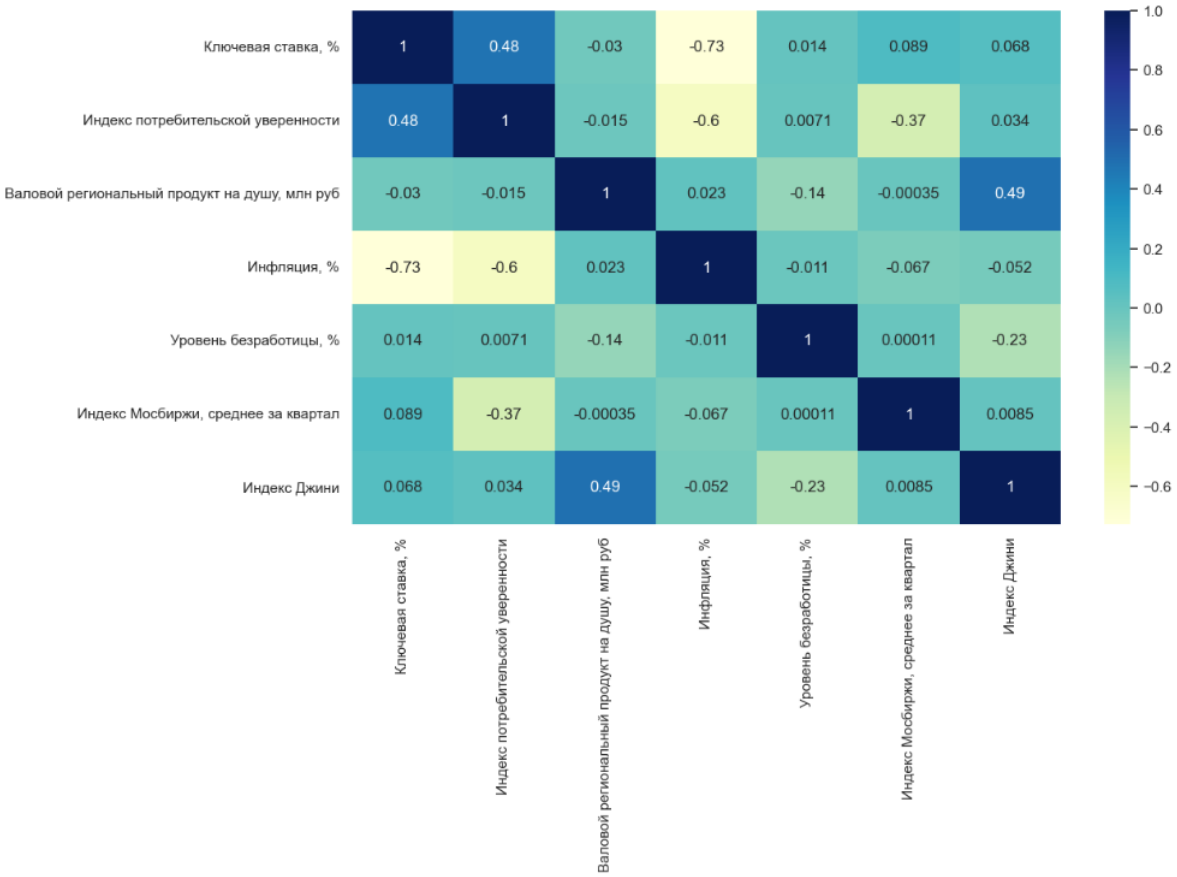


Рисунок 19 – полученные входные признаки для построения модели

На полученной матрице убраны признаки, корреляция которых более 0.5. Мультиколлениарности нет, поэтому можем использовать такой набор входных признаков.

ГЛАВА 3 ВЫБОР МЕТРИКИ ОЦЕНКИ КАЧЕСТВА МОДЕЛИ

Наиболее типичными мерами качества в задачах регрессии являются:

- Средняя квадратичная ошибка (MSE);
- Средняя абсолютная ошибка;
- Коэффициент детерминации;
- Средняя абсолютная процентная ошибка.

Средняя квадратичная ошибка (MSE) - MSE применяется в ситуациях, когда необходимо выделить большие ошибки и выбрать модель, которая исключает большие ошибки. Большие ошибки становятся весомее из-за того, что значение ошибки прогноза возводится в квадрат. Модель, которая дает нам меньшее значение MSE, можно сказать, что что у этой модели меньше больших ошибок.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|a(x_i) - y_i|$$

Средняя абсолютная ошибка - Среднеквадратичный функционал сильнее штрафует за большие отклонения по сравнению со среднеабсолютным, и поэтому более чувствителен к выбросам. При использовании любого из этих двух функционалов может быть полезно проанализировать, какие объекты вносят наибольший вклад в общую ошибку — не исключено, что на этих объектах была допущена ошибка при вычислении признаков или целевой величины.

Среднеквадратичная ошибка подходит для сравнения двух моделей или для контроля качества во время обучения, но не позволяет сделать выводов о том, на сколько хорошо данная модель решает задачу. Например, MSE = 10 является очень плохим показателем, если целевая переменная принимает значения от 0 до 1, и очень хорошим, если целевая переменная лежит в интервале (10000, 100000). В таких ситуациях вместо среднеквадратичной ошибки полезно использовать коэффициент детерминации — R2.

Коэффициент детерминации измеряет долю дисперсии, объясненную моделью, в общей дисперсии целевой переменной. Фактически, данная мера качества — это нормированная среднеквадратичная ошибка. Если она близка к единице, то модель хорошо объясняет данные, если же она близка к нулю, то прогнозы сопоставимы по качеству с константным предсказанием.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(a(x_i) - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Средняя абсолютная процентная ошибка - коэффициент, не имеющий размерности, с простой интерпретацией. Его можно измерять в долях или процентах. Если у вас получилось, например, что MAPE=11.4%, то это говорит о том, что ошибка составила 11,4% от фактических значений. Основная проблема данной ошибки — нестабильность.

$$MAPE = 100\% \times \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - a(x_i)|}{|y_i|}$$

MAPE является полезной метрикой для оценки точности модели в задачах, где важна точность прогнозирования в процентном соотношении. Например, задачи прогнозирования продаж или доходов.

Однако, MAPE также имеет свои недостатки. Она не может быть использована в случаях, когда значения равны нулю или пропущены, так как такие значения приводят к делению на ноль (видно по формуле MAPE). Кроме того, она может быть чувствительна к выбросам в данных, что может привести к искажению результатов.

На основе анализа полученного набора данных принято решение использовать метрики Коэффициент детерминации (R2) и среднюю абсолютную процентную ошибку (MAPE), т.к целевые признаки имеют непрерывный тип данных.

Непрерывные данные — данные, которые могут принимать любые значения в некотором интервале. Над непрерывными значениями можно проводить арифметические операции, и они имеют смысл. Примерами непрерывных данных являются: рост, вес, количество товара, прибыль и т.д.

ГЛАВА 4 ПОДБОР ГИПЕРПАРАМЕТРОВ ДЛЯ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

Подбор параметров – одна из важных задач для построения модели машинного обучения. Изменение параметров модели может принципиально повлиять на ее качество. Например, модель может переобучиться. Перебор этих параметров вручную может занять колоссальное количество времени. Для решения этой проблемы существует функция GridSearchCV из библиотеки sklearn.model_selection. GridSearchCV – это очень мощный инструмент для автоматического подбирания параметров для моделей машинного обучения. GridSearchCV находит наилучшие параметры, путем обычного перебора: он создает модель для каждой возможной комбинации параметров. Если у нас есть M гиперпараметров и для каждого задано N возможных значений, то число вариантов равно M x N и для каждого нужно обучить модель и определить ее точность. Если мы используем перекрестную проверку (cross-validation), то это число надо умножить на число частей, на которые мы разбиваем набор данных. [18]

Параметры GridSearchCV:

- estimator — модель которую хотим обучать (алгоритм);
- param_grid — передаем какие параметры хотим подбирать, GridSearchCV на всех параметрах попробует сделать обучение;
- CV — сколько разрезов кросс-валидации мы ходим сделать;
- scoring — выбор метрики ошибки (для разных задач можно выбрать разные функции ошибки).

Данный метод хоть и работает не очень быстро, но, при этом, экономит достаточно времени по сравнению с ручным перебором тех же параметров, чем дает явное преимущество в использовании при построении моделей.

ГЛАВА 5 ВЫБОР И ПОСТРОЕНИЕ МОДЕЛИ

Целевые признаки имеют непрерывный тип данных, поэтому для анализа будем рассматривать регрессивные алгоритмы. Регрессивные алгоритмы осуществляют прогнозирование одной или нескольких непрерывных числовых переменных, например прибыли или убытков, на основе других атрибутов в наборе данных. В качестве моделей возьмем модели случайного леса и градиентного бустинга.

При подгонке моделей машинного обучения к наборам данных разбиваем набор данных на два набора:

1. Набор для обучения: используется для обучения модели (80% исходного набора данных).

2. Тестовый набор: используется для получения объективной оценки производительности модели (20% исходного набора данных).

Разделение произведено с помощью train_test_split библиотеки sklearn.

```
# разделение на обучающий и тестовый наборы данных
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Рисунок 20 – Разделение выборки

5.1. Построение модели случайного леса (RandomForestRegressor)

В качестве одной из моделей для обучения был выбран случайный лес. Случайный лес — это алгоритм машинного обучения, который использует множество деревьев решений для принятия решений. Каждое дерево в случайном лесу обучается на случайной подвыборке данных и использует только некоторые признаки для принятия решений. Это позволяет уменьшить переобучение и повысить точность прогнозирования.

Кроме того, случайный лес имеет ряд преимуществ перед другими алгоритмами машинного обучения. Во-первых, он может обрабатывать большие объемы данных и многомерные признаки. Во-вторых, он может работать с

данными, содержащими пропущенные значения и выбросы. В-третьих, он может использоваться для решения задач классификации и регрессии.

На основе анализа данных можно предположить, что модель случайного леса хорошо подходит, т. к. данные имеют сложную структуру и нелинейные зависимости между признаками. Также эта модель хорошо показывается себя на большом объеме данных, т. к. способен его обрабатывать и не склонен к переобучению.

5.1.1. Подбор параметров модели

Параметры, используемые при построении модели случайного леса:

- n_estimators - количество деревьев в лесу;
- max_depth -Максимальная глубина дерева. Если None, то узлы расширяются до тех пор, пока все листья не станут чистыми или пока все листья не будут содержать выборок меньше, чем min_samples_leaf - Минимальное количество выборок, необходимое для конечного узла. Точка разделения на любой глубине будет рассматриваться только в том случае, если она оставляет по крайней мере min_samples_leaf обучающие выборки в каждой из левой и правой ветвей. Это может иметь эффект сглаживания модели, особенно в регрессии.;
- min_samples_split - Минимальное количество выборок, необходимое для разделения внутреннего узла.

```
clf = RandomForestRegressor()
parametrs = { 'n_estimators': range (90, 200, 10),
              'max_depth': range (8, 20, 2),
              'min_samples_leaf': range (1,8),
              'min_samples_split': range (2,10,2) }

grid = GridSearchCV(clf, parametrs, cv=5)
grid.fit(X_train, y_train)
display(grid.best_params_)
```

Рисунок 21 – задание параметров для GridSearchCV

Через «GridSearch» библиотеки sklearn были подобраны такие гиперпараметры модели, с которым оценка качества достигает максимума:

```
{'max_depth': 18,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'n_estimators': 130}
```

Рисунок 22 – Подобранные параметры через GridSearchCV

5.1.2. Реализация модели машинного обучения для прогнозирования наличной валюты

Для реализации модели использовался sklearn — библиотека машинного обучения для языка программирования Python. Она предоставляет широкий спектр алгоритмов машинного обучения, включая методы классификации, регрессии, кластеризации, а также методы выбора признаков и предобработки данных.

Подготовленная выборка была разделена на входные и выходные данные. Далее определен размер тестовой выборки - 0.2. Для этого использовался модуль sklearn.model_selection, метод train_test_split.

Через класс RandomForestRegressor была проинициализирована модель случайного леса. Кол-во деревьев равно 130. По результатам обучения получили оценку качества модели (R2 score) 0.98.

```
R2 score: 0.98
```

Рисунок 23 – результат R2 score

По результатам обучения получили MAPE = 0.25, что значит ошибка составила 25% от фактических значений. Ниже представлены графики соотношения предсказанных значений и фактических на примере целевых показателей (синий график – фактическое значение, оранжевый – предсказанное моделью).

Как видим из графика, модель почти справилась с «выбросами» данных в виде Москвы и Санкт-Петербурга

Через класс RandomForestRegressor была проинициализирована модель случайного леса.

Рисунок 24 – Предсказание показателя «Наличная валюта» через случайный лес

Показатели:

- R2 – 0.95
- MAPE – 0.25

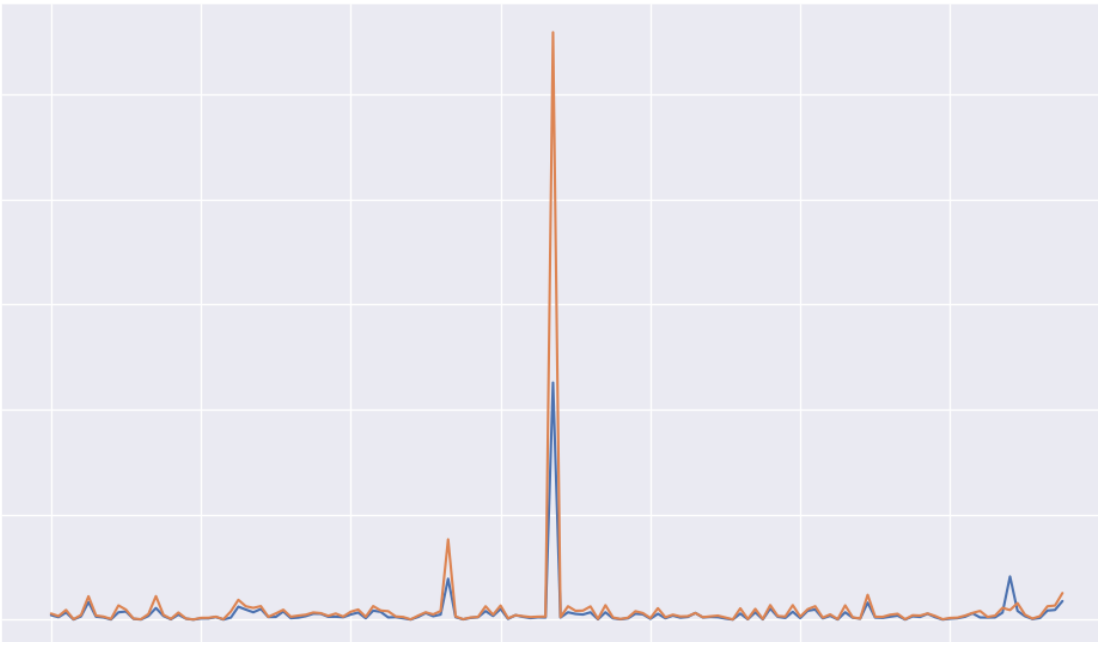


Рисунок 25 – Предсказание показателя «Вклады физических лиц» через случайный лес

- R2 – 0.94
- MAPE – 0.23

5.1.3. Реализация модели машинного обучения для прогнозирования показателей долговых ценных бумаг

Показатели рассчитаны аналогично пункту 4.1.1 через GridSearchCV.

Подготовленная выборка была разделена на входные и выходные данные. Далее определен размер тестовой выборки - 0.2. Для этого использовался модуль sklearn.model_selection, метод train_test_split.

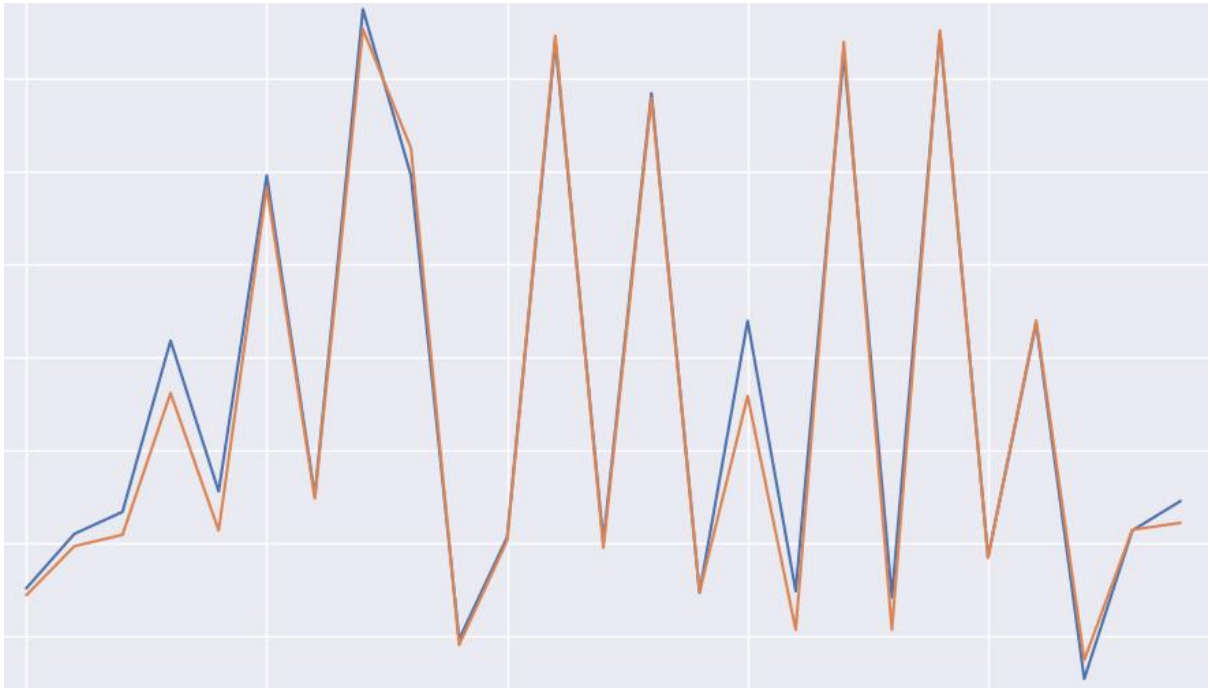


Рисунок 26 – Предсказание показателя «Кредитные организации»

Показатели:

- R2 – 0.96
- MAPE – 0.28

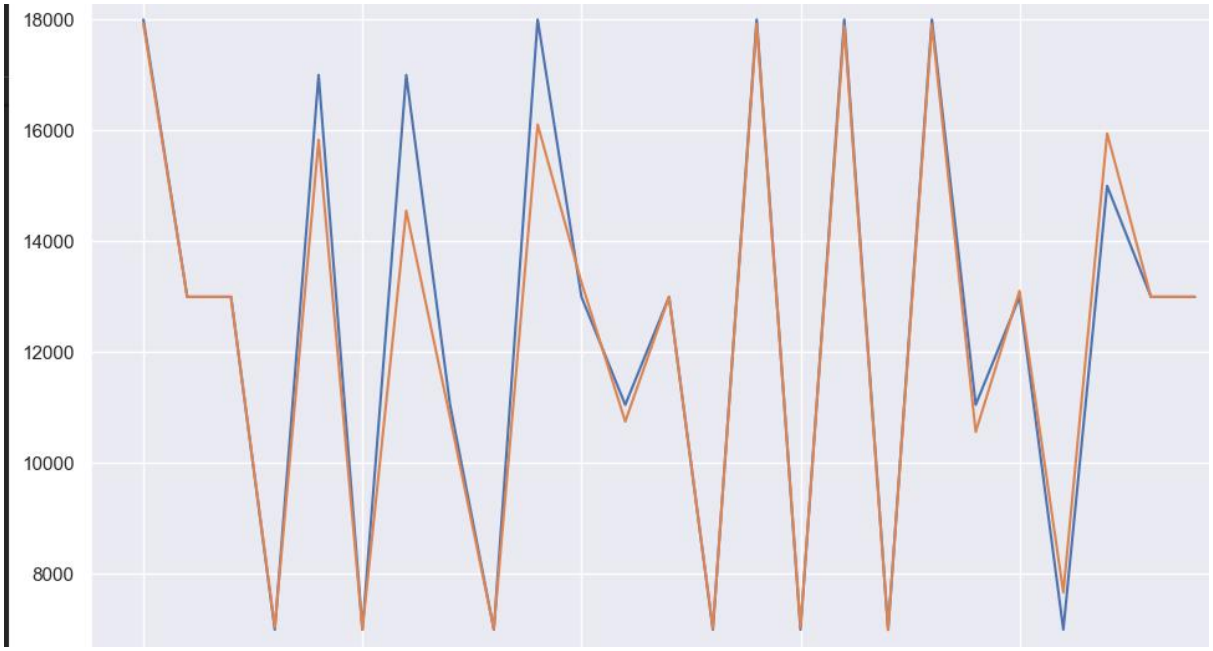


Рисунок 27 – Предсказание показателя «Страховщики»

Показатели:

- R2 – 0.94
- MAPE – 0.22

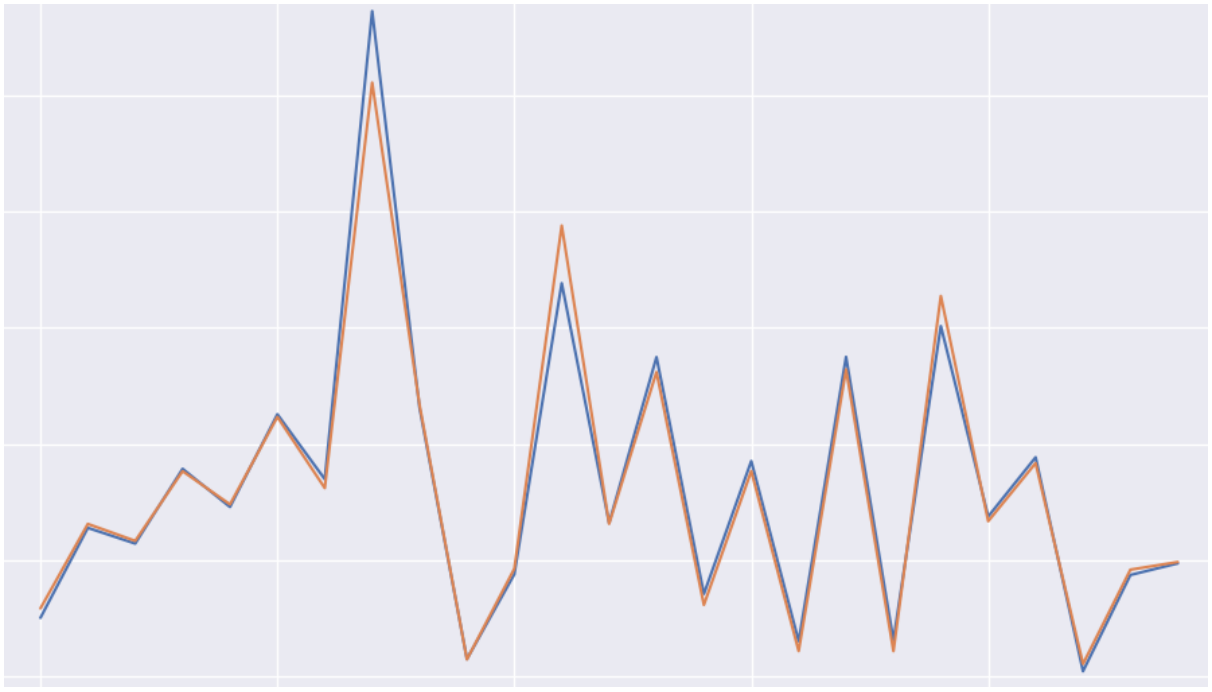


Рисунок 28 – Предсказание показателя «Другие финансовые организации»

Показатели:

- R2 – 0.95
- MAPE – 0.26

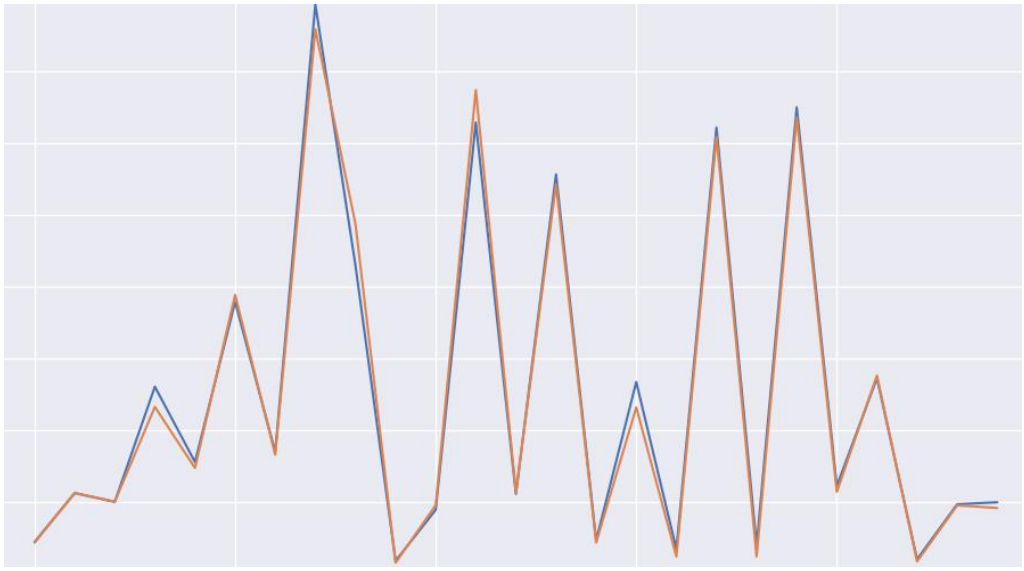


Рисунок 29 – Предсказание показателя «Органы государственного управления»

Показатели:

- R2 – 0.96
- MAPE – 0.31

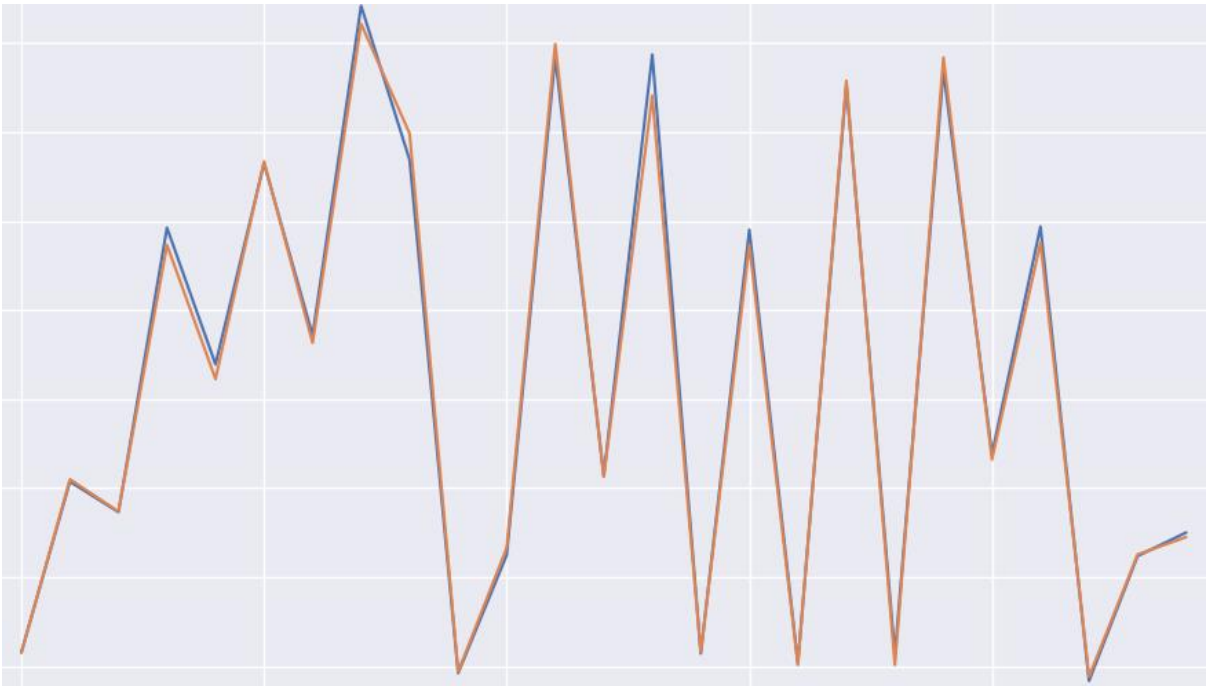


Рисунок 30 – Предсказание показателя «Нефинансовые организации»

Показатели:

- R2 – 0.96
- MAPE – 0.21

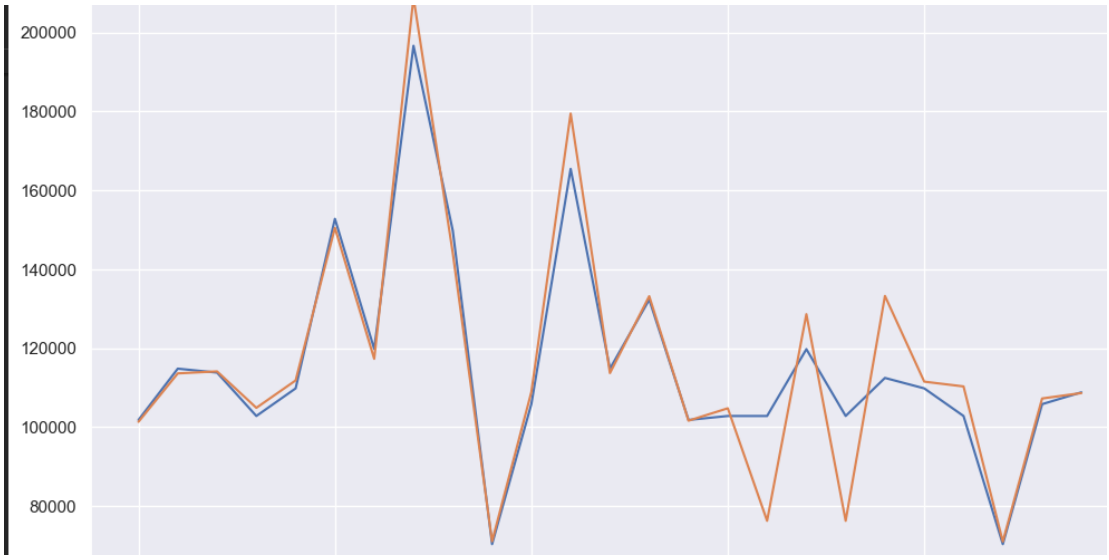


Рисунок 31 – Предсказание показателя «Нерезиденты»

Показатели:

- R2 – 0.84
- MAPE – 0.26

5.2. Построение моделм градиентного бустинга (GradientBoostingRegressor)

Градиентный бустинг — это техника машинного обучения для задач классификации и регрессии, которая строит модель предсказания в форме ансамбля слабых предсказывающих моделей, обычно деревьев решений. [19]

Эта функция используется для создания модели регрессии с повышением градиента. Она принимает обучающие данные и метки в качестве параметров.

Строит аддитивную модель поэтапно вперед; позволяет оптимизировать произвольные дифференцируемые функции потерь. На каждом этапе дерево регрессии аппроксимируется отрицательным градиентом заданной функции потерь.

Бустинг — это техника построения ансамблей, в которой предсказатели построены не независимо, а последовательно

Это техника использует идею о том, что следующая модель будет учится на ошибках предыдущей. Они имеют неравную вероятность появления в последующих моделях, и чаще появятся те, что дают наибольшую ошибку. Предсказатели могут быть выбраны из широкого ассортимента моделей, например, деревья решений, регрессия, классификаторы и т.д. Из-за того, что предсказатели обучаются на ошибках, совершенных предыдущими, требуется меньше времени для того, чтобы добраться до реального ответа. Но мы должны выбирать критерий остановки с осторожностью, иначе это может привести к переобучению. [19]

5.2.1. Подбор параметров модели для прогнозирования наличных средств

Параметры, используемые при построении модели бустинга:

- learning_rate- скорость обучения;
- max_depth -Максимальная глубина отдельных регрессионных оценок

- loss – функция потерь (['absolute_error', 'squared_error'])
- subsample - Доля выборок, которые будут использоваться для подбора отдельных базовых учащихся. Если значение меньше 1,0, это приводит к стохастическому повышению градиента.

```
clf = GradientBoostingRegressor()
parametrs = { 'loss': ['absolute_error', 'squared_error'],
              'learning_rate': [0.1, 0.3, 0.5],
              'subsample': [0.8, 0.9, 1.0],
              'max_depth': range (3, 8, 1)
}

grid = GridSearchCV(clf, parametrs, cv=5)
grid.fit(X_train, y_train)
display(grid.best_params_)
```

Рисунок 32 – задание параметров для GridSearchCV

Через «GridSearch» библиотеки sklearn были подобраны такие гиперпараметры модели, с которым оценка качества достигает максимума:

```
{'learning_rate': 0.1,
 'loss': 'squared_error',
 'max_depth': 6,
 'subsample': 0.8}
```

Рисунок 33 – Подобранные параметры через GridSearchCV

5.2.2. Реализация модели машинного обучения для прогнозирования наличных средств

Подготовленная выборка была разделена на входные и выходные данные. Далее определен размер тестовой выборки - 0.2. Для этого использовался модуль sklearn.model_selection, метод train_test_split.

Через класс GradientBoostingRegressor была проинициализирована бустинга. По результатам обучения получили оценку качества модели (R2 score) 0.96.

По результатам обучения получили MAPE = 0.21, что значит ошибка составила 21% от фактических значений. Ниже представлены графики соотношения

предсказанных значений и фактических на примере целевых показателей (синий график – фактическое значение, оранжевый – предсказанное моделью).

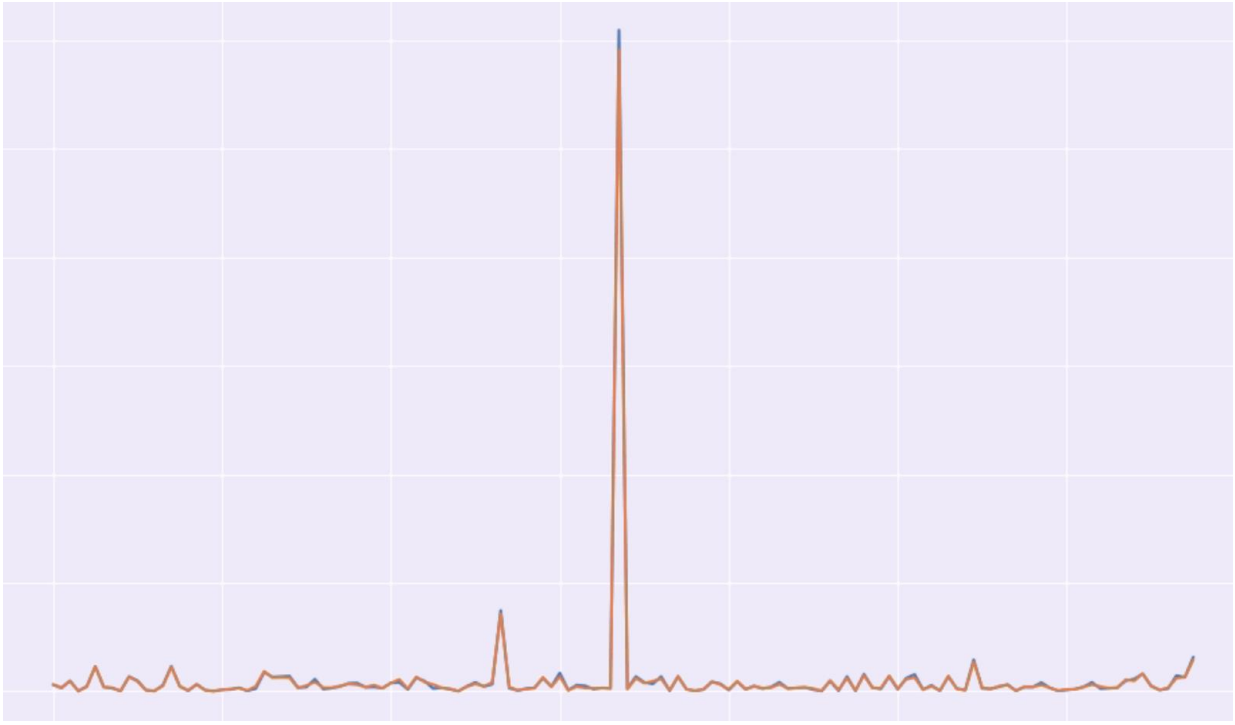


Рисунок 34 – Предсказание показателя «Наличная валюта, тыс руб (Средства клиентов в рублях)» GradientBoostingRegressor

Показатели:

- R2 – 0.9
- MAPE – 0.21

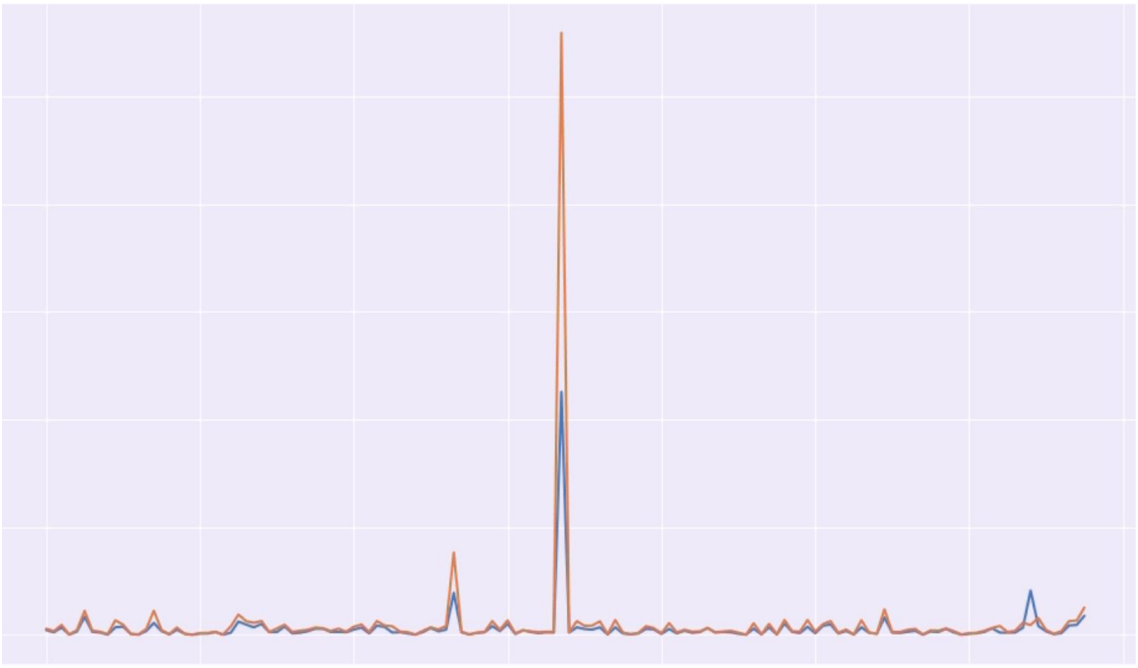


Рисунок 35– Предсказание показателя «Вклады физических лиц, тыс руб» GradientBoostingRegressor

Показатели:

- R2 – 0.93
- MAPE – 0.27

5.2.3. Подбор параметров модели для прогнозирования показателей долговых ценных бумаг

Параметры для построения модели бустинга аналогичны пункту 4.2.1

```
clf = GradientBoostingRegressor()
parametrs = { 'loss': ['absolute_error', 'squared_error'],
              'learning_rate': [0.1, 0.3, 0.5],
              'subsample': [0.8, 0.9, 1.0],
              'max_depth': range (3, 8, 1)
}

grid = GridSearchCV(clf, parametrs, cv=5)
grid.fit(X_train, y_train)
display(grid.best_params_)
```

Рисунок 36 – задание параметров для GridSearchCV

Через «GridSearch» библиотеки sklearn были подобраны такие гиперпараметры модели, с которым оценка качества достигает максимума:

```
{'learning_rate': 0.5,
 'loss': 'absolute_error',
 'max_depth': 3,
 'subsample': 0.8}
```

Рисунок 37 – Подобранные параметры через GridSearchCV

5.2.4. Реализация модели машинного обучения прогнозирования показателей долговых ценных бумаг

```
# разделение на обучающий и тестовый наборы данных
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# обучение модели
params = {'learning_rate': 0.5,
 'loss': 'absolute_error',
 'max_depth': 3,
 'subsample': 0.8
}
gbr = GradientBoostingRegressor(**params)
gbr.fit(X_train, y_train)


# важность признаков
importances = gbr.feature_importances_
for i, importance in enumerate(importances):
    print(f'Признак {i}: {importance:.4f}')
y_pred = gbr.predict(X_test)
```

Рисунок 38 – полученная модель бустинга

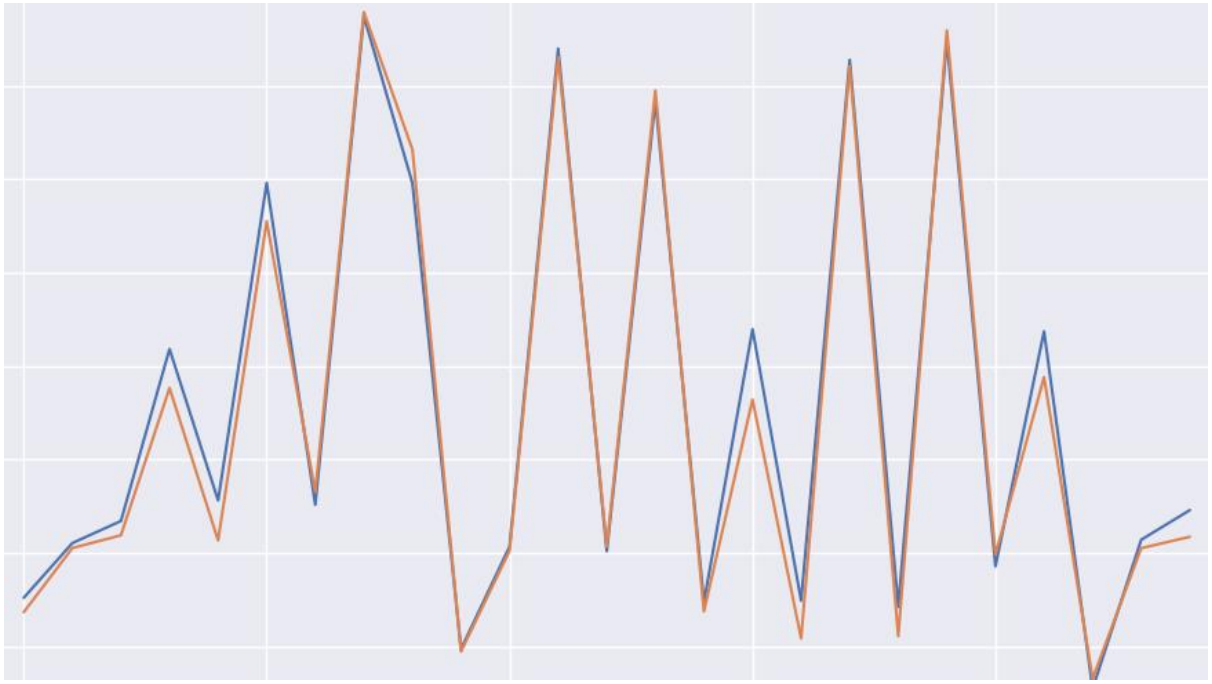


Рисунок 39 – Предсказание показателя «Кредитные организации»

Показатели:

- R2 – 0.95
- MAPE – 0.3

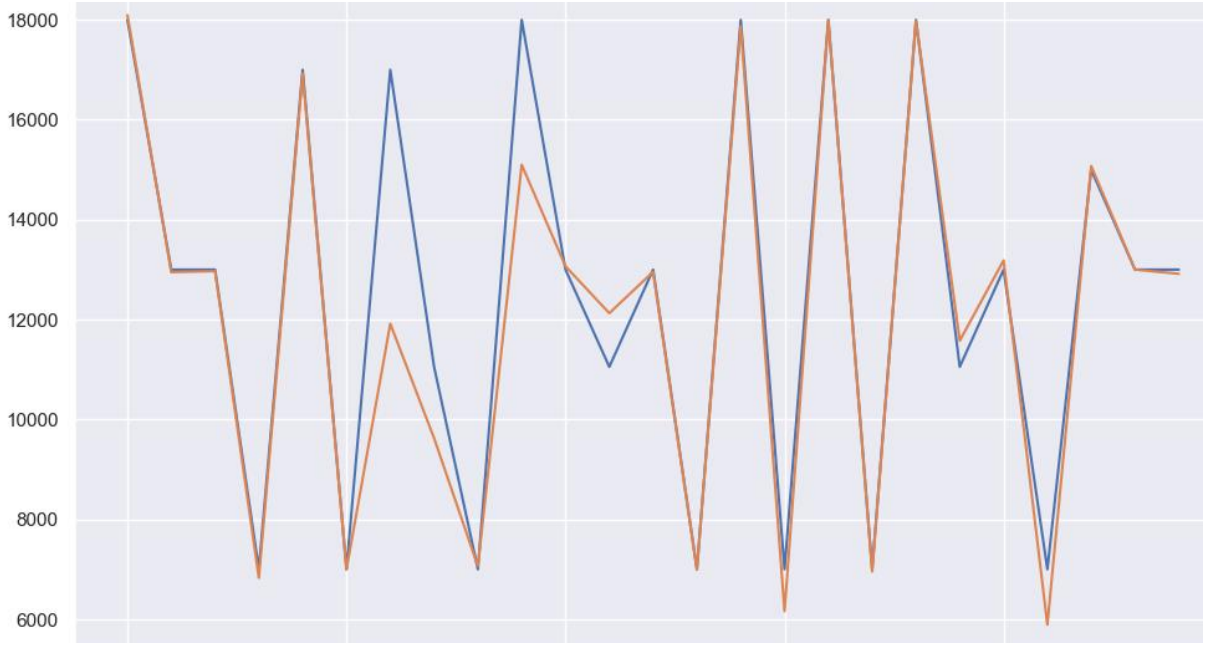


Рисунок 40 – Предсказание показателя «Страховщики»

Показатели:

- R2 – 0.90

- MAPE – 0.23

Рисунок 41 – Предсказание показателя «Другие финансовые организации»

Показатели:

- R2 – 0.95
- MAPE – 0.31

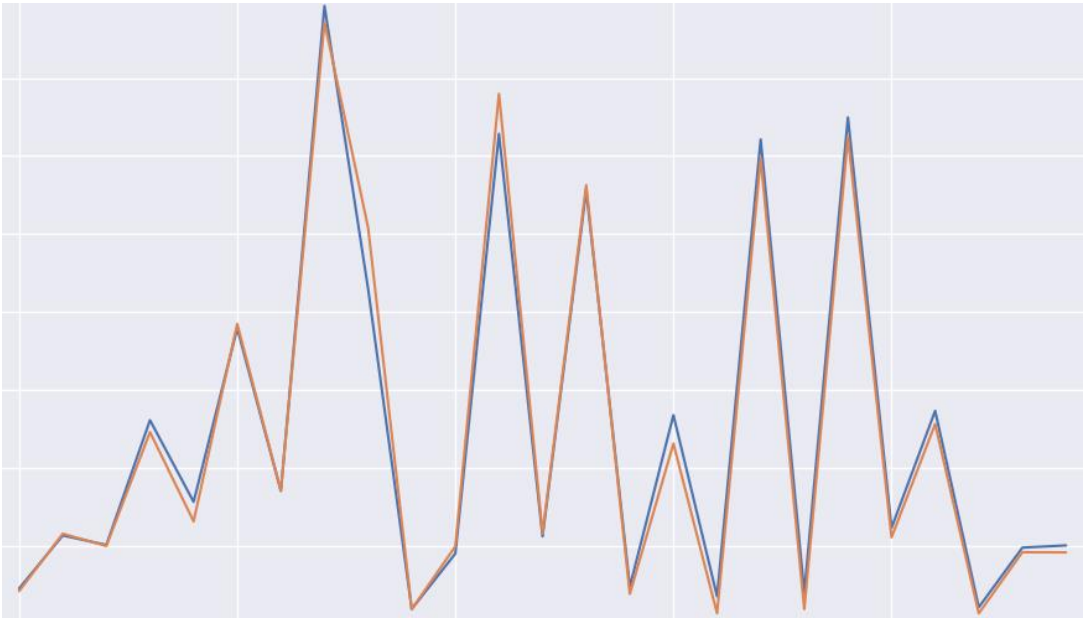


Рисунок 42 – Предсказание показателя «Органы государственного управления»

Показатели:

- R2 – 0.98
- MAPE – 0.35

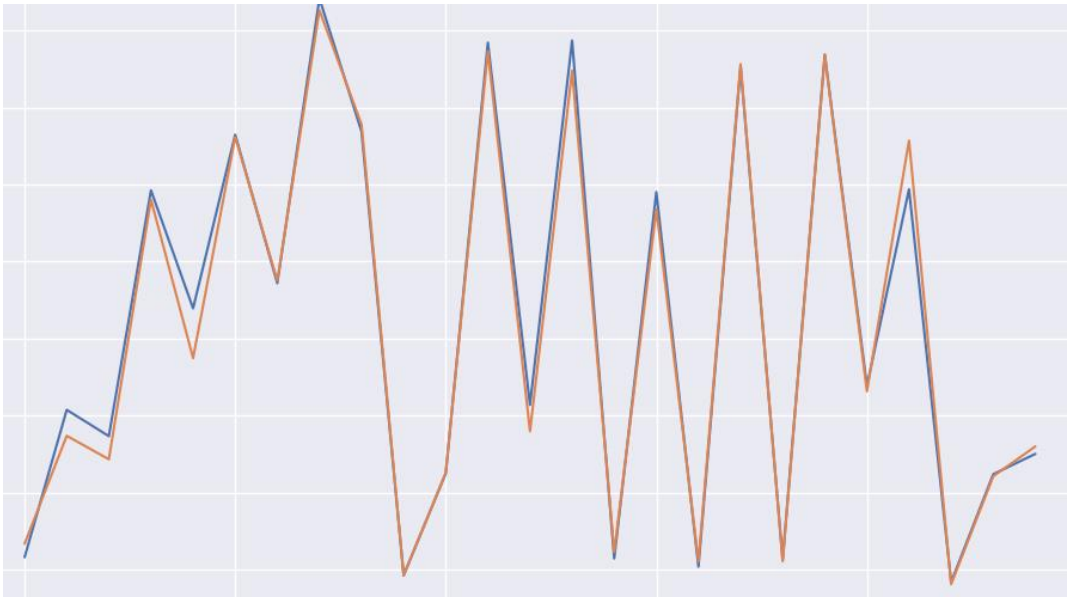


Рисунок 43 – Предсказание показателя «Нефинансовые организации»

Показатели:

- R2 – 0.97
- MAPE – 0.4

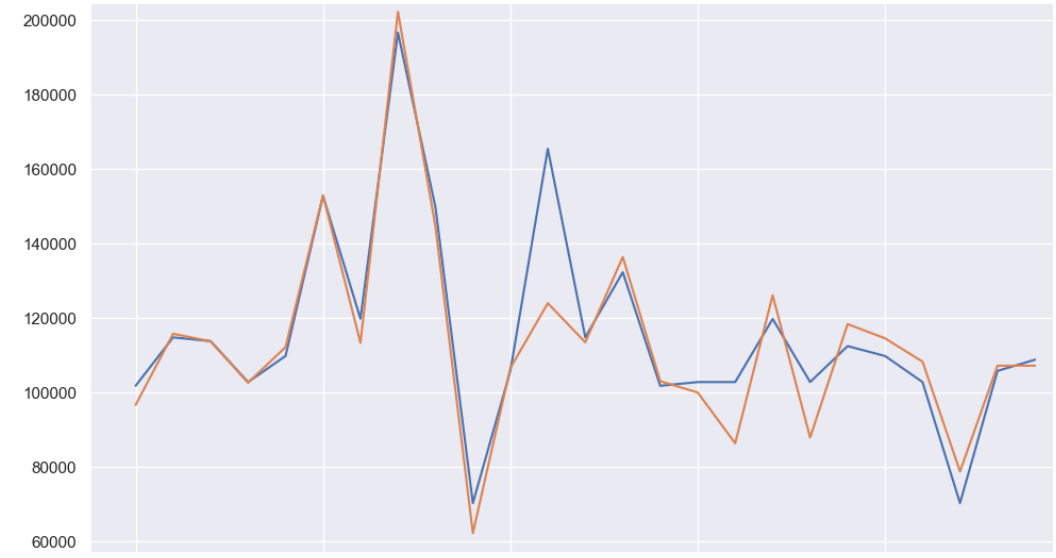


Рисунок 44 – Предсказание показателя «Нерезиденты»

Показатели:

- R2 – 0.85
- MAPE – 0.34

5.3. Сравнение моделей

Обе модели, и случайный лес и градиентный бустинг, показывают схожие метрики на подобранных параметрах через GridSearchCV. Модель случайного леса дает при этом лучшую метрику качества MAPE. Сравним все полученные значения по метрике MAPE:

Признак	Случайный лес	Градиентный бустинг
Наличная валюта, тыс руб (Средства клиентов в рублях)	0.25	0.31
Вклады физических лиц, тыс руб	0.23	0.27
Кредитные организации	0.28	0.3
Страховщики	0.22	0.23
Другие финансовые организации	0.26	0.331 ы
Органы государственного управления	0.31	0.35
Нефинансовые организации	0.21	0.24
Нерезиденты	0.26	0.33

Таблица 2 – сравнение полученных метрик

Исходя из полученных метрик, можно сделать вывод что для решения данной задачи больше подходит модель случайного леса. В среднем модель предсказывает значение с ошибкой в 25% от фактического значения.

ЗАКЛЮЧЕНИЕ

В ходе проделанной работы была разобрана предметная область, макроэкономические показатели, рассмотрены модели прогнозирования целевых признаков, возможные метрики качества модели.

Результатом проделанной работы стали обученные модели – случайный лес и градиентный бустинг. Проведен анализ и их сравнение на основе сформированной выборки. Предварительно были найдены и изучены первичные источники и разработаны парсеры для сбора данных. Проведен первичный анализ на наличие пропусков и формата показателей, а также корреляционных связей. Полученная модель случайного леса позволяет предсказать показатели со средней ошибкой 25%.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. ЕМИСС – единая межведомственная информационно-статистическая система [Электронный ресурс]. Режим доступа: https://russia.duck.consulting/maps/104/2017.
2. Федеральная служба государственной статистики [Электронный ресурс]. Режим доступа: https://rosstat.gov.ru/statistic.
3. InvestFunds – независимый источник данных для частного инвестора в России [Электронный ресурс]. Режим доступа: https://myfin.by/.
4. Банкирша.com, Инфляция в России [Электронный ресурс]. Режим доступа: https://bankirsha.com/uroven-inflyacii-v-rossiyskoy-federacii-po-godam.html.
5. Индекс Мосбиржи [Электронный ресурс]. Режим доступа: https://www.moex.com/ru/index/IMOEX/archive/?from=2013-03-01&till=2023-05-01&sort=TRADEDATE&order=desc.
6. Банк России [Электронный ресурс]. Режим доступа: https://www.cbr.ru/hd_base/KeyRate/?UniDbQuery.Posted=True&UniDbQuery.From=17.09.2013&UniDbQuery.To=12.05.2023.
7. Macrotrends – Цена на нефть [Электронный ресурс]. Режим доступа: https://www.macrotrends.net/2516/crude-oil-prices-70-year-historical-chart.
8. Документация scikit-learn [Электронный ресурс]. Режим доступа: https://scikit-learn.org/stable/;
9. Документация TensorFlow [Электронный ресурс]. Режим доступа: https://www.tensorflow.org/api_docs;
10. Макроэкономические показатели [Электронный ресурс]. Режим доступа: https://www.banki.ru/;
11. Средняя цена 1 кв. м. общей площади квартир на рынке [Электронный ресурс]. Режим доступа: https://www.fedstat.ru/indicator/31452.

12. Индекс Джинни [Электронный ресурс]. Режим доступа: https://russia.duck.consulting/maps/104/2016.
13. Статистические показатели банковского сектора Российской Федерации [Электронный ресурс]. Режим доступа: https://www.cbr.ru/statistics/bank_sector/review/.
14. Макроэкономические показатели [Электронный ресурс]. Режим доступа: https://www.banki.ru/wikibank/makroekonomicheskie_pokazateli/.
15. Типы полей [Электронный ресурс]. Режим доступа: https://www.ibm.com/docs/ru/spss-statistics/saas?topic=tab-field-variable-types
16. Коэффициент корреляции [Электронный ресурс]. Режим доступа: https://wiki.loginom.ru/articles/correlation-coefficient.html
17. Индекс Мосбиржи [Электронный ресурс]. Режим доступа: https://www.tinkoff.ru/invest/research/etf/imoex/
18. Подбор параметров с помощью GridSearchCV [Электронный ресурс]. Режим доступа: https://vc.ru/ml/147132-kak-avtomaticheski-podobrat-parametry-dlya-modeli-mashinnogo-obucheniya-ispolzuem-gridsearchcv
19. Градиентный бустинг [Электронный ресурс]. Режим доступа: https://neurohive.io/ru/osnovy-data-science/gradientyj-busting/
20. Система национальных счетов [Электронный ресурс]. Режим доступа: https://www.banki.ru/wikibank/sistema_natsionalnyih_schetov/
21.