

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

**ПРОГНОЗИРОВАНИЕ СОБЫТИЯ ПРИОБРЕТЕНИЯ
НЕДВИЖИМОСТИ ИПОТЕЧНЫМИ КЛИЕНТАМИ
СТРОИТЕЛЬНОЙ КОМПАНИИ И ИХ КЛАСТЕРИЗАЦИЯ**

по программе профессиональной переподготовки:

«Анализ данных на языке Python»

Выполнил: Лиситчук Антон Олегович
Руководитель: к.т.н. Семендяев Родион Юрьевич

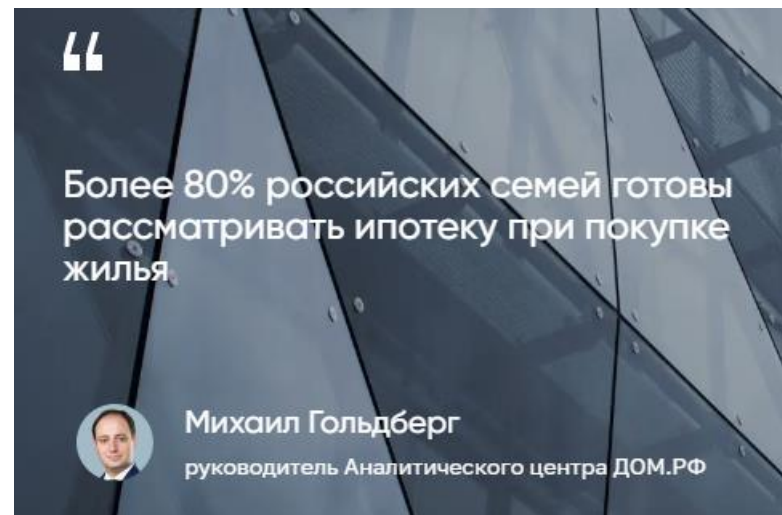
Санкт-Петербург, 2023

Актуальность

В настоящее время бóльшая часть квартир на первичном рынке приобретается с использованием ипотеки. В крупных строительных компаниях существуют специальные отделы, которые помогают клиентам подать заявку в банк для получения ипотечного кредита.

Для прогнозирования объема продаж важно понимать, какие клиенты с наибольшей вероятностью станут покупателями и от каких характеристик это зависит.

Важным моментом является понимание портрета покупателя, целевой аудитории



Исходные данные

- В наборе представлены данные о клиентах, подавших заявку на ипотеку
- Событие покупки недвижимости является результирующим признаком
- Категориальные признаки закодированы в числовые значения
- Период подачи заявок: с июня 2021 года по март 2023 года включительно

	Дата	Тип занятости	Тип дохода	Тип ипотеки	Стоимость, млн	Доля ПВ	Срок кредита, лет	Образование	СП	Пол	Возраст	Регион покупателя	Категория работы	Имущество	Купил
0	2021-06-01	1	1	1	13.51	NaN	10.00	0 0	1.00	34.00	2	0	0	0	
1	2021-06-01	1	1	1	13.51	NaN	10.00	0 0	1.00	34.00	2	0	0	0	
2	2021-06-03	1	1	1	NaN	NaN	10.00	5 2	1.00	35.00	1	3	0	0	
3	2021-06-07	1	1	1	6.66	40.00	10.00	4 1	1.00	44.00	1	0	1	0	
4	2021-06-07	1	1	1	6.66	40.00	10.00	4 1	1.00	44.00	1	5	1	0	
...
11549	2023-03-16	1	5	2	23.23	35.42	20.00	4 2	2.00	49.00	1	4	1	0	
11550	2023-01-19	1	1	3	24.73	83.00	30.00	5 2	1.00	37.00	1	5	1	0	
11551	2022-08-03	1	1	2	16.50	69.69	20.00	6 2	2.00	50.00	1	4	0	0	
11552	2022-09-26	1	1	2	15.05	21.00	20.00	0 1	1.00	40.00	3	0	0	0	
11553	2022-12-12	1	1	1	32.80	15.23	30.00	4 2	2.00	26.00	1	2	1	0	

11554 rows × 15 columns

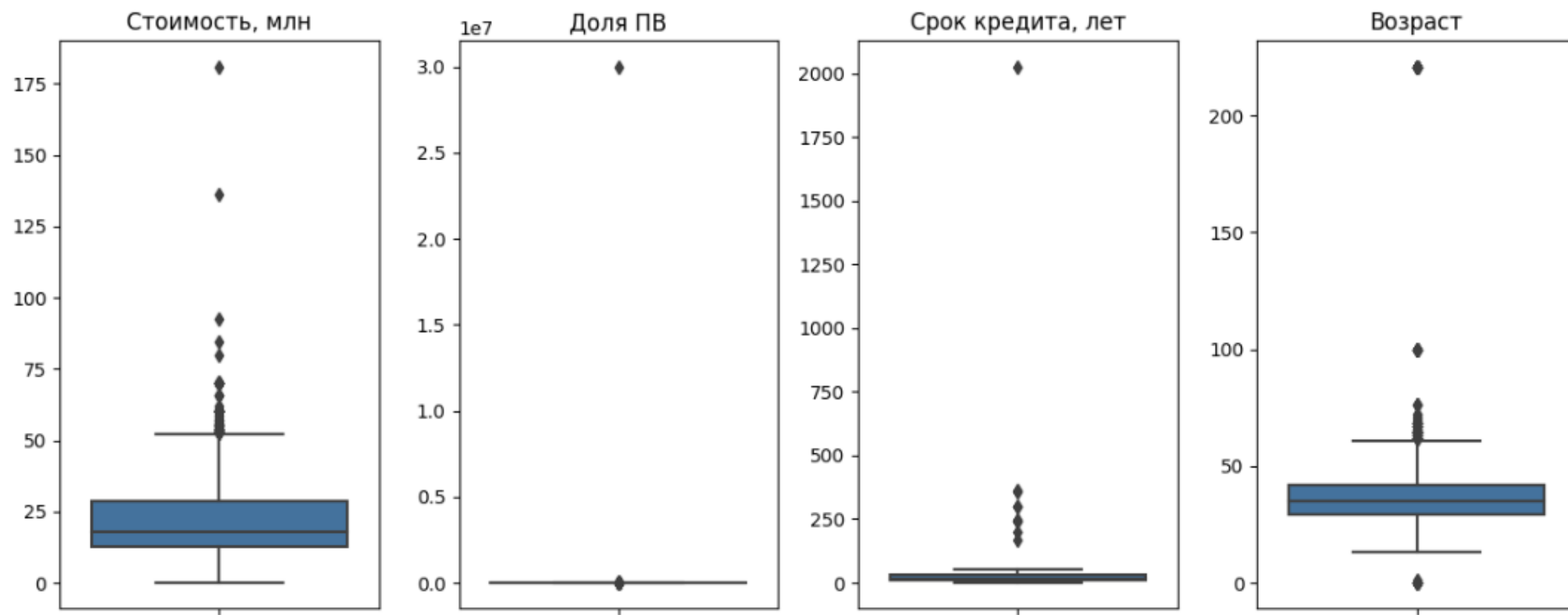
Цель работы

Спрогнозировать поведение клиентов ипотечного кредитования на основе их анализа

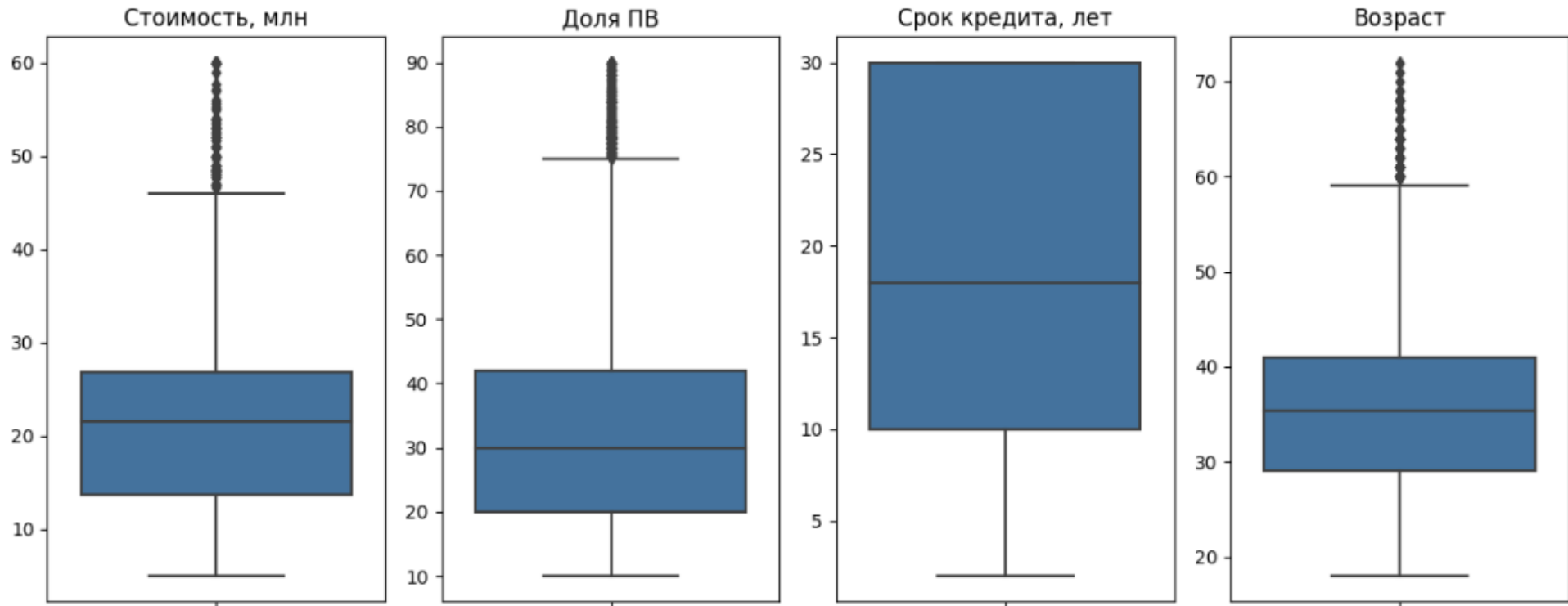
Задачи:

1. Провести сбор и подготовку данных
2. Провести анализ и предварительную обработку данных
3. Выбрать и обучить модель классификатор, позволяющую прогнозировать событие приобретения недвижимости
4. Оценить качество модели
5. Провести анализ клиентов
6. Провести кластеризацию покупателей и дать описание каждому кластеру

Данные до обработки



Данные после обработки



Стоимость от 5 до 60 млн руб.
Доля первоначального взноса от 10% до 90%

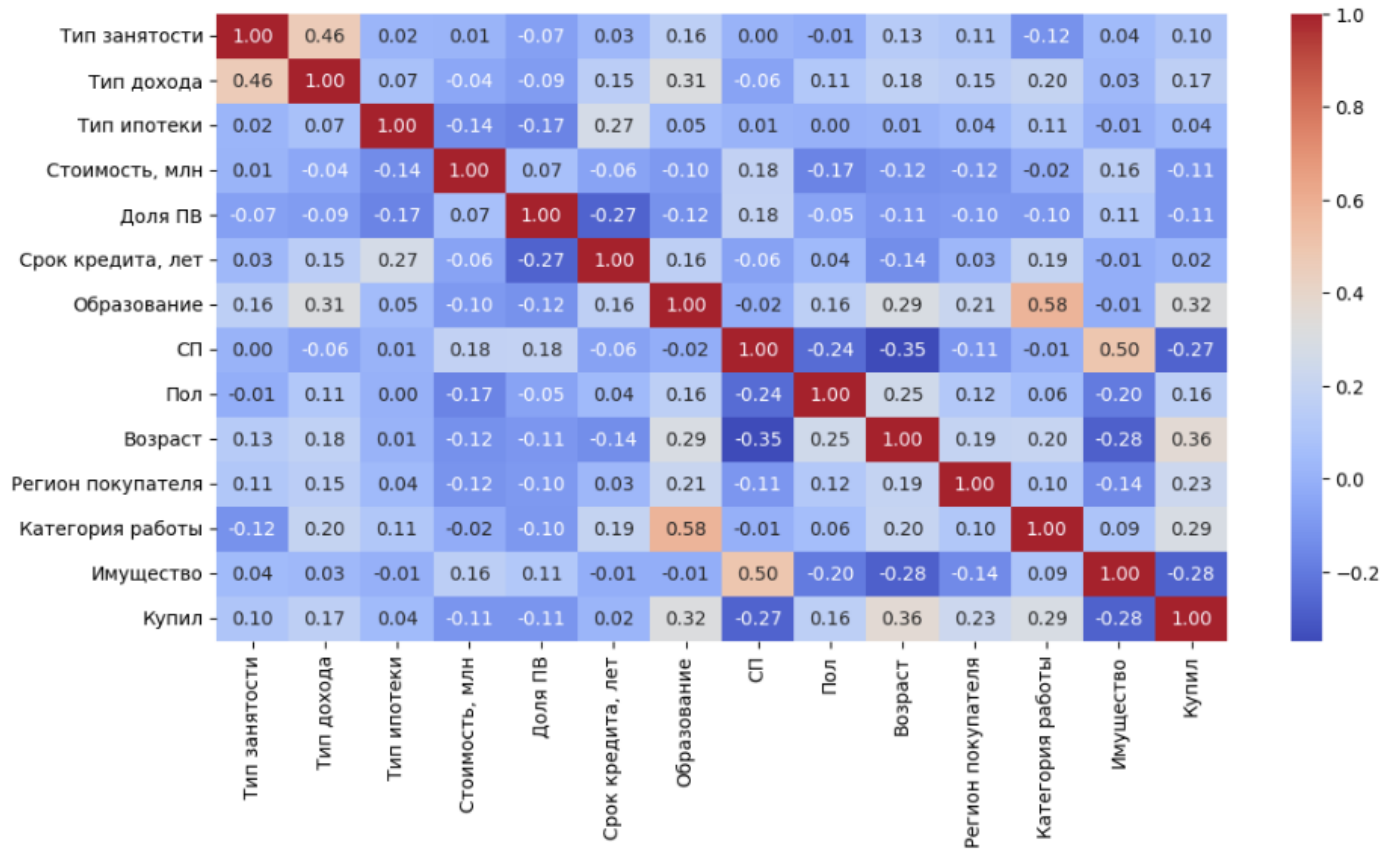
Срок кредита от 2 до 30 лет
Возраст от 18 до 75 лет

Корреляционная матрица

Результативный признак распределен относительно равномерно: 44%/56%

Наибольшую корреляцию с результативным признаком имеют следующие признаки:

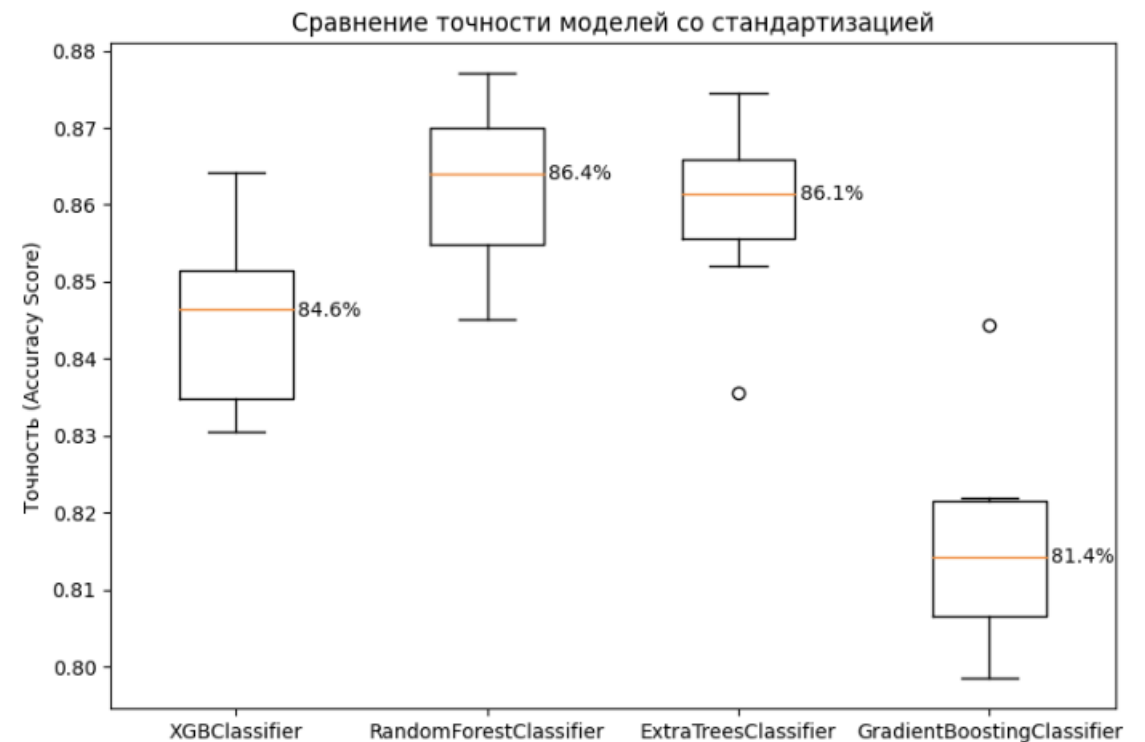
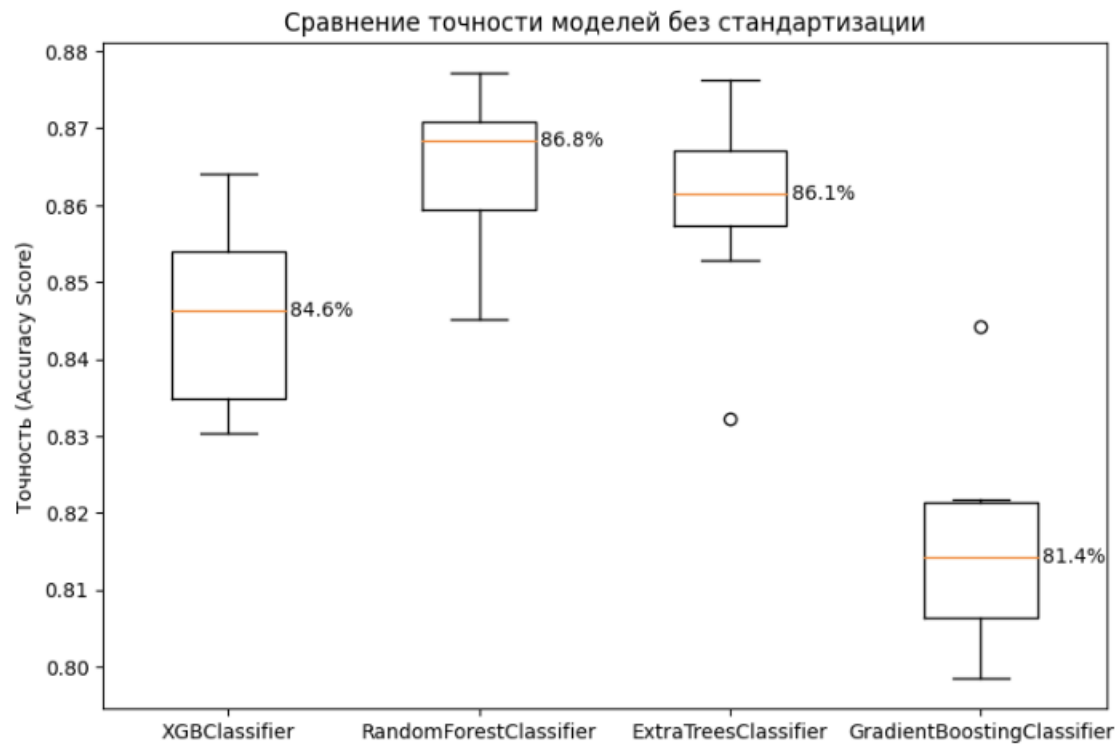
- Возраст (0,36)
- Образование (0,32)
- Категория работы (0,29)
- Регион покупателя (0,23)
- Семейное положение (-0,27)
- Имущество (-0,28)



Метод Спирмена (method='spearman')

Выбор моделей классификаторов

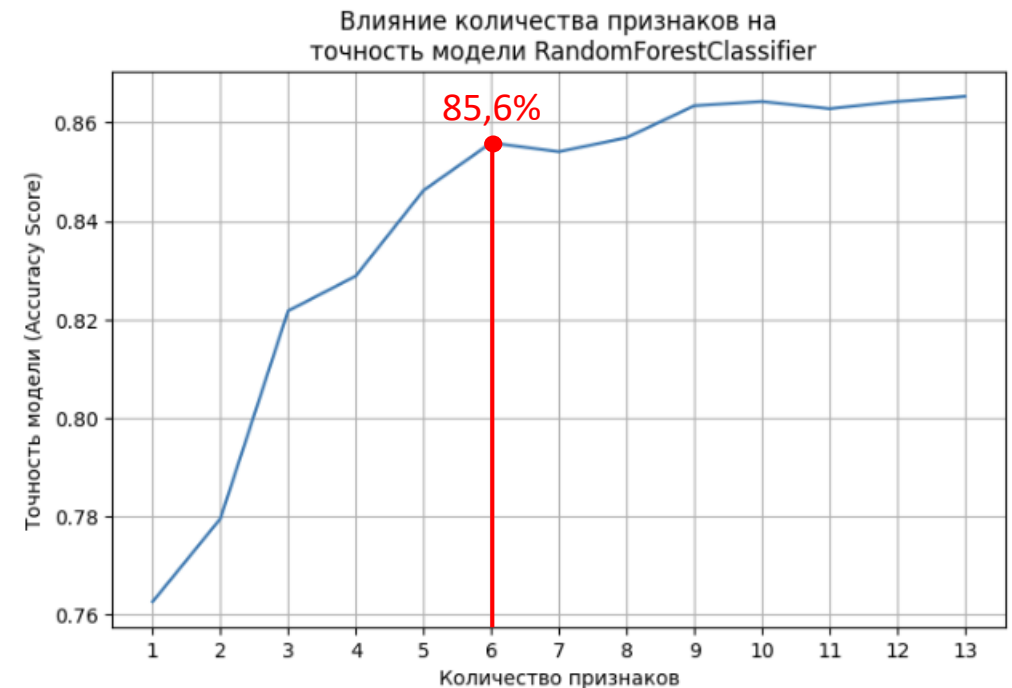
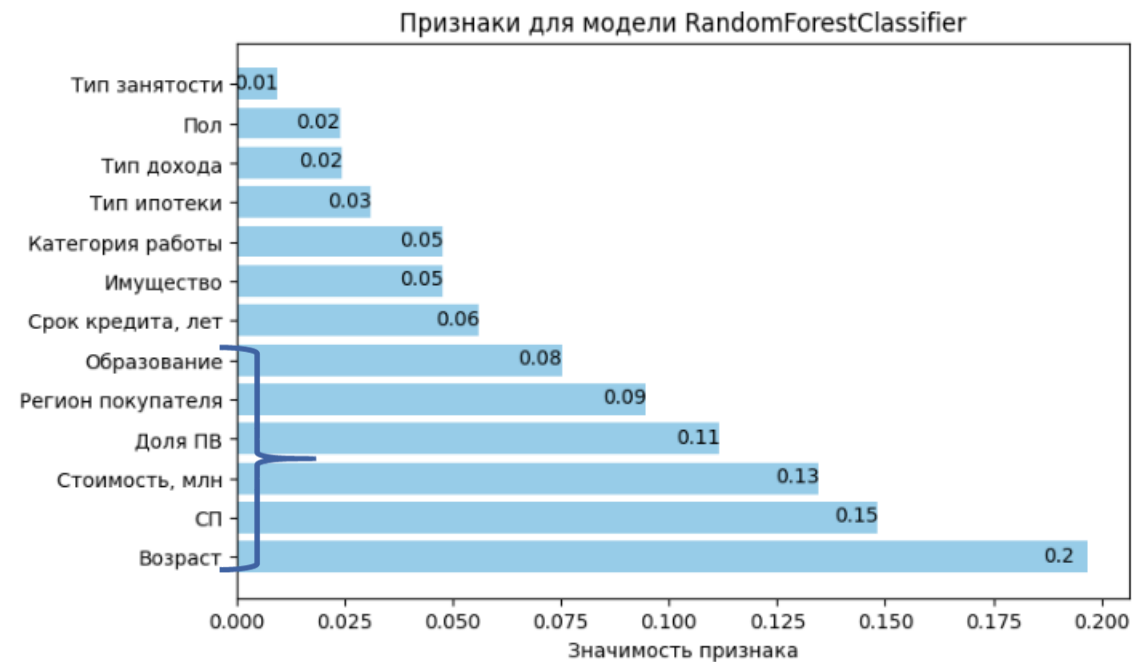
Лучшую точность (accuracy score) показывает модель «RandomForestClassifier» - ансамблевый метод классификации, который строит несколько деревьев решений на разных подвыборках данных и усредняет их прогнозы. Точность не увеличивается при использовании стандартизации.



Оценка значимости и отбор признаков

Наибольшее влияние на результат прогнозирования оказывает «Возраст», наименьшее – «Тип занятости»

Можно оставить 6 признаков вместо 13, практически без ухудшения точности модели (accuracy score).



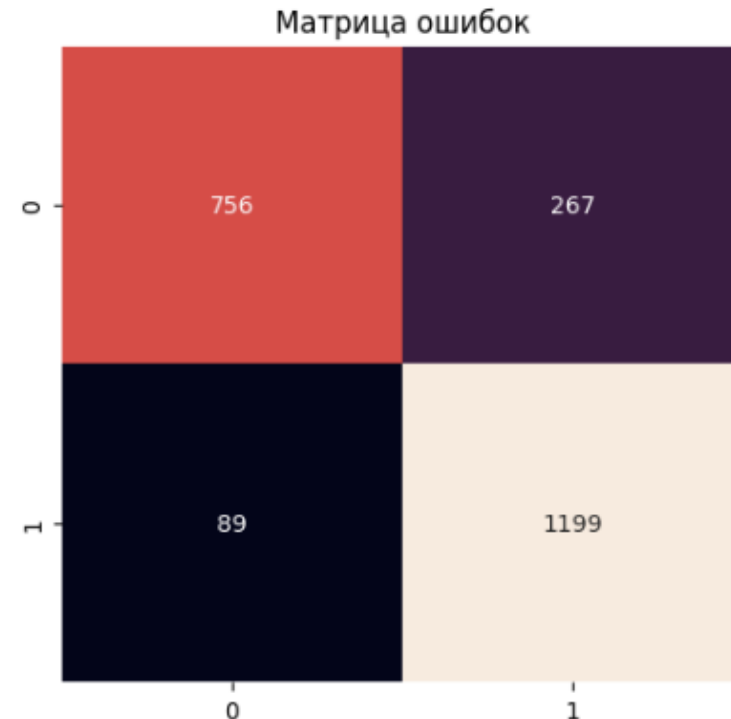
Оценка качества модели

Модель: RandomForestClassifier
Размер тестовой выборки: 20%

Точность (accuracy score): 85%
Точность для 0 класса (precision): 90%
Точность для 1 класса (precision): 88%

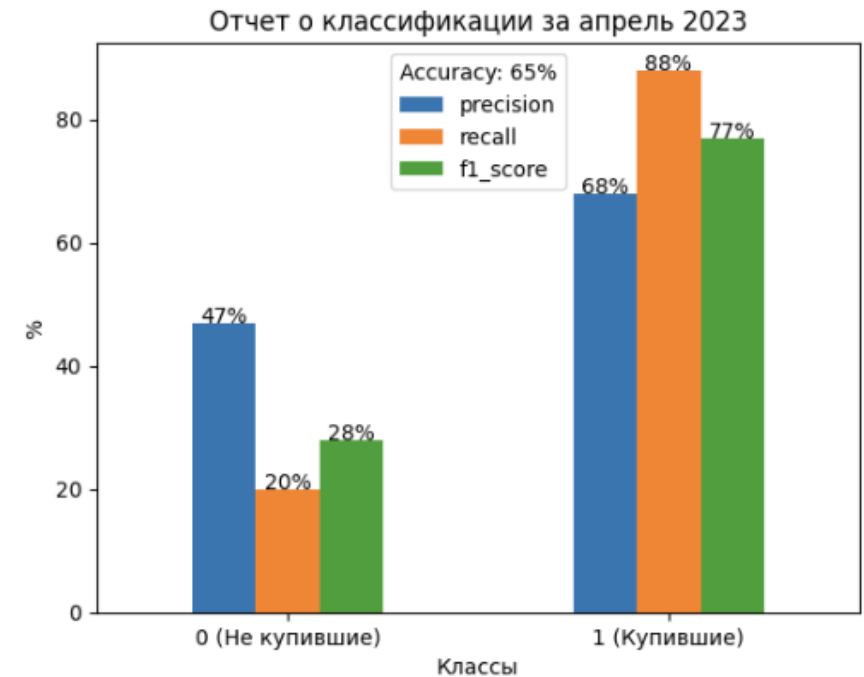
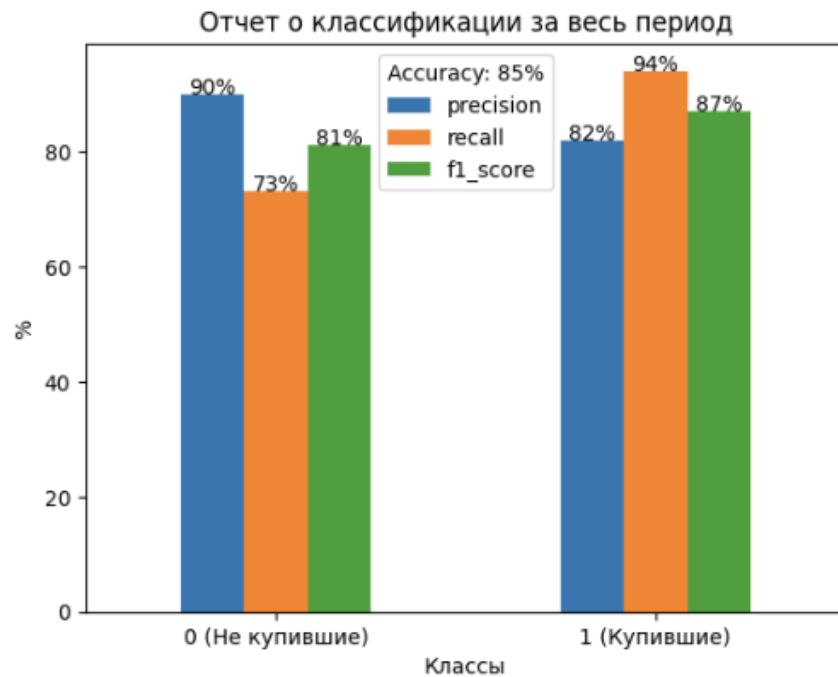
Вероятность пропуска наблюдения 0 класса: 27%
Вероятность пропуска наблюдения 1 класса: 6%

Модель лучше предсказывает 1 класс – тех, кто совершил покупку.



Прогнозирование на новых наблюдениях

В ранее обученную модель были переданы данные, которые поступили за апрель 2023 года.



Из полученных показателей можно сделать вывод о том, что между датой подачи заявки на ипотеку и покупкой квартиры существует временной лаг. На момент получения данных не все клиенты еще успели совершить покупку.

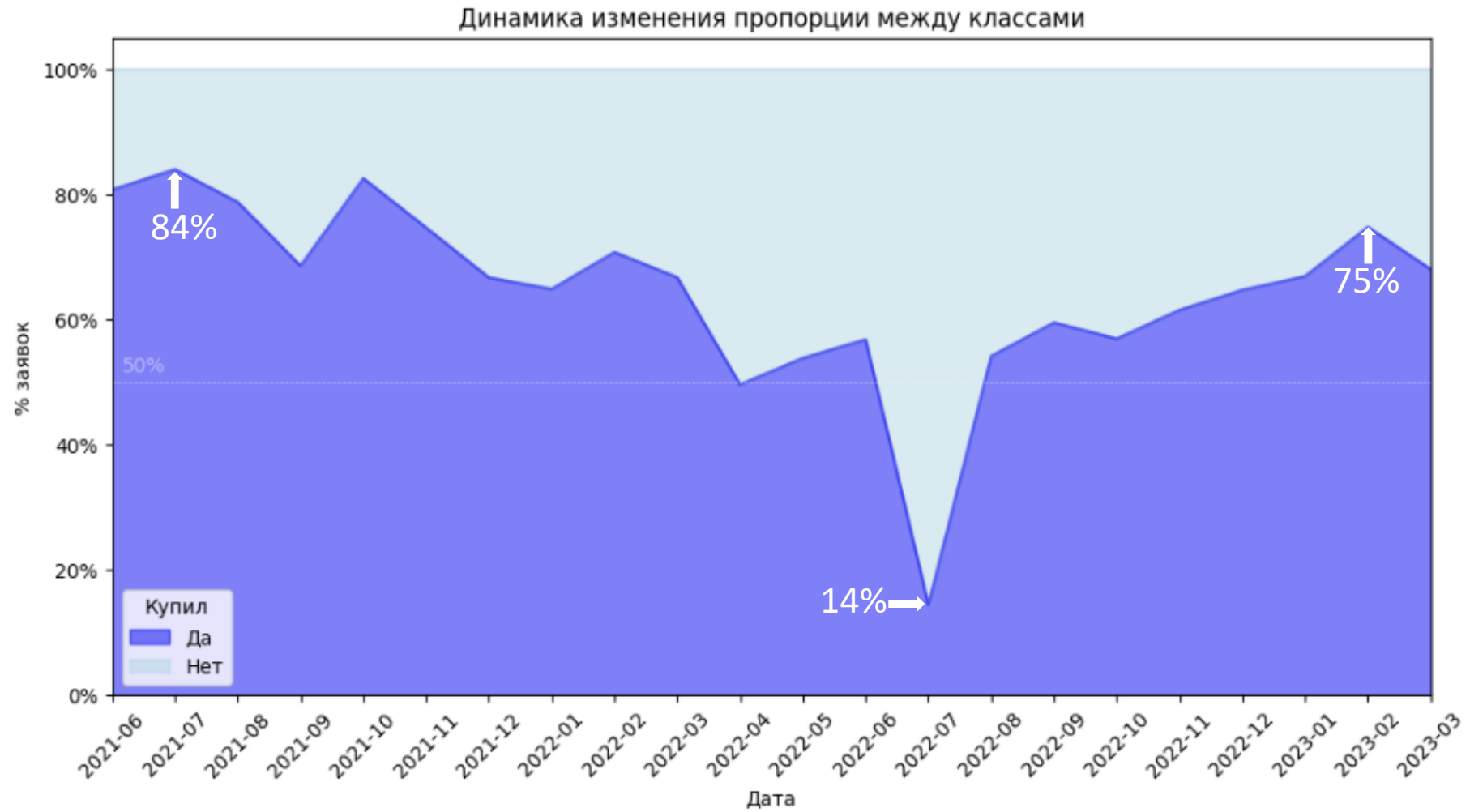
Анализ клиентов

Типичные наблюдения наиболее значимых признаков, влияющих на событие покупки квартиры, в разбивке набора данных на купивших и не купивших клиентов

	Средний возраст	СП	Средняя стоимость, млн	Средняя доля ПВ, %	Регион покупателя	Образование	Кол-во	Доля, %
Купил								
Нет	31.40	Гражданский брак	23.41	35.68	Мск	Не указано	5015	43.40
Да	38.54	Женат (замужем)	20.23	30.72	Мск	Высшее	6539	56.60

- Количество купивших преобладает
- Чаще покупают более взрослые люди с высшим образованием, состоящие в браке
- Чаще не покупают более молодые люди, не указывающие информацию об образовании и находящиеся в гражданском браке

Анализ клиентов

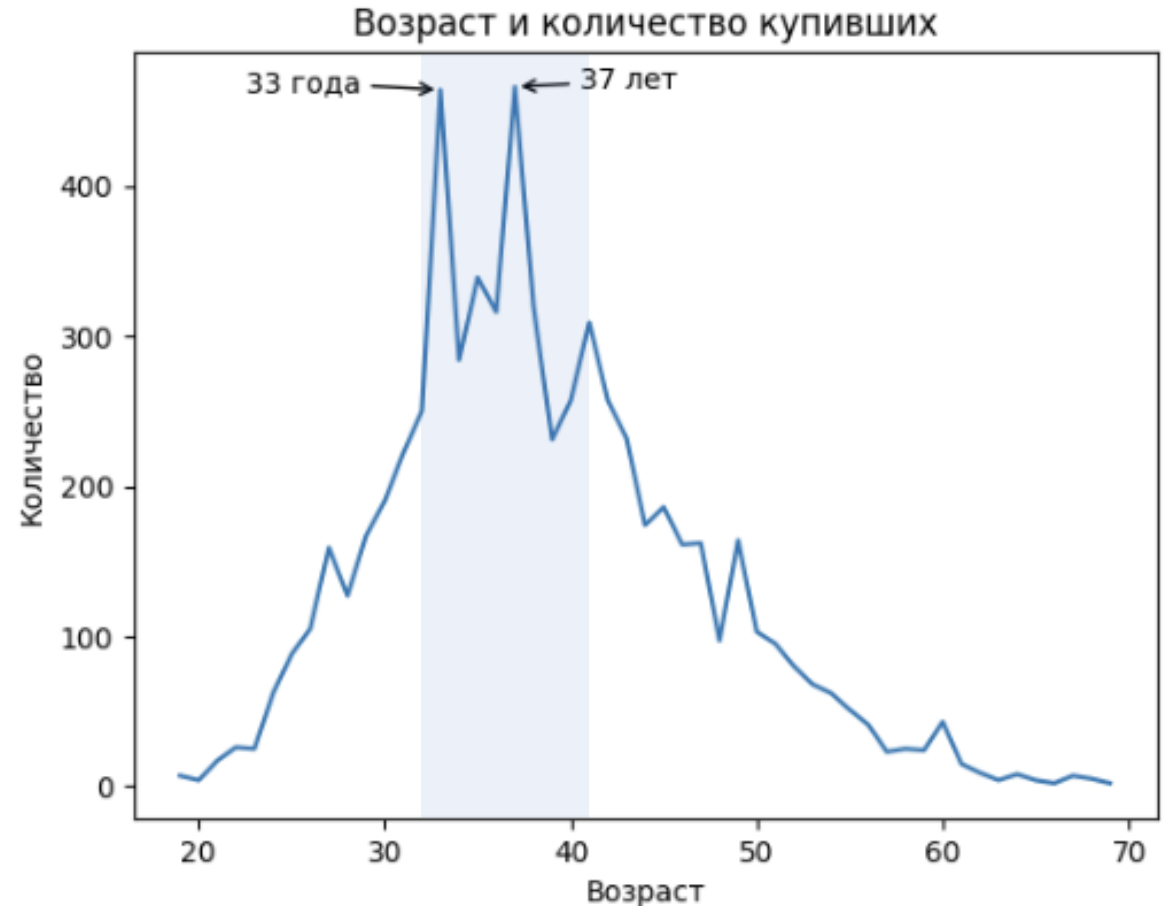


Анализ покупателей

Возраст покупателей является наиболее важным признаком, определяющим событие покупки

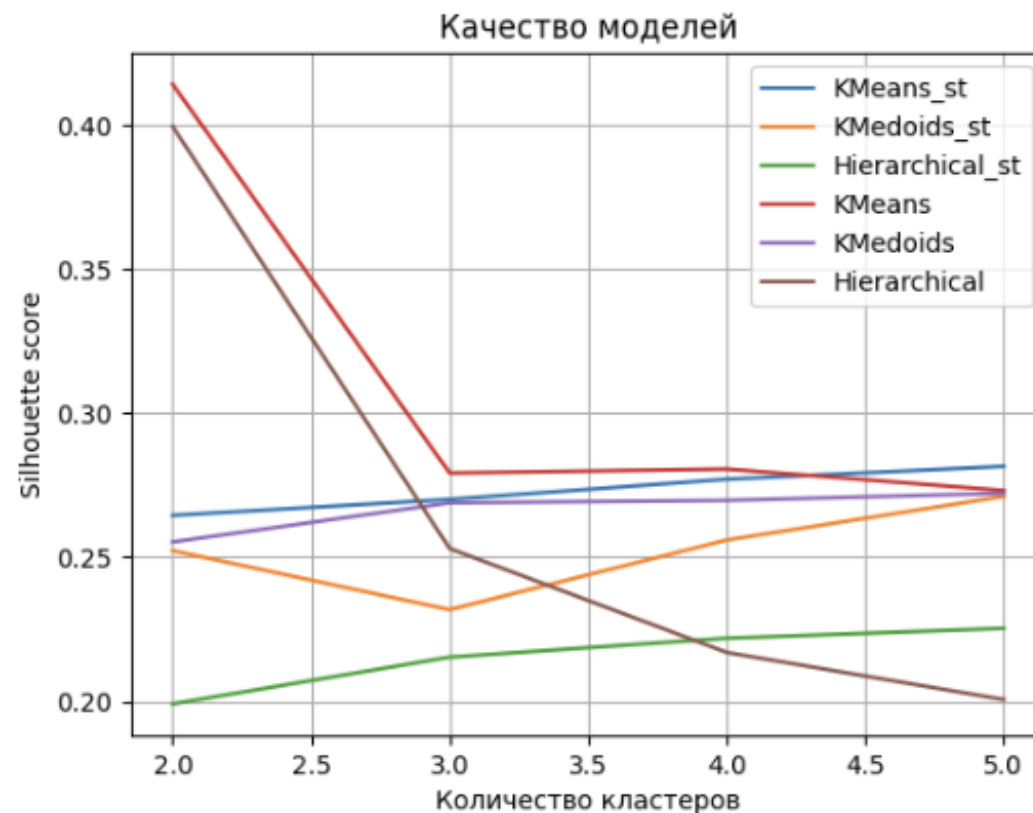
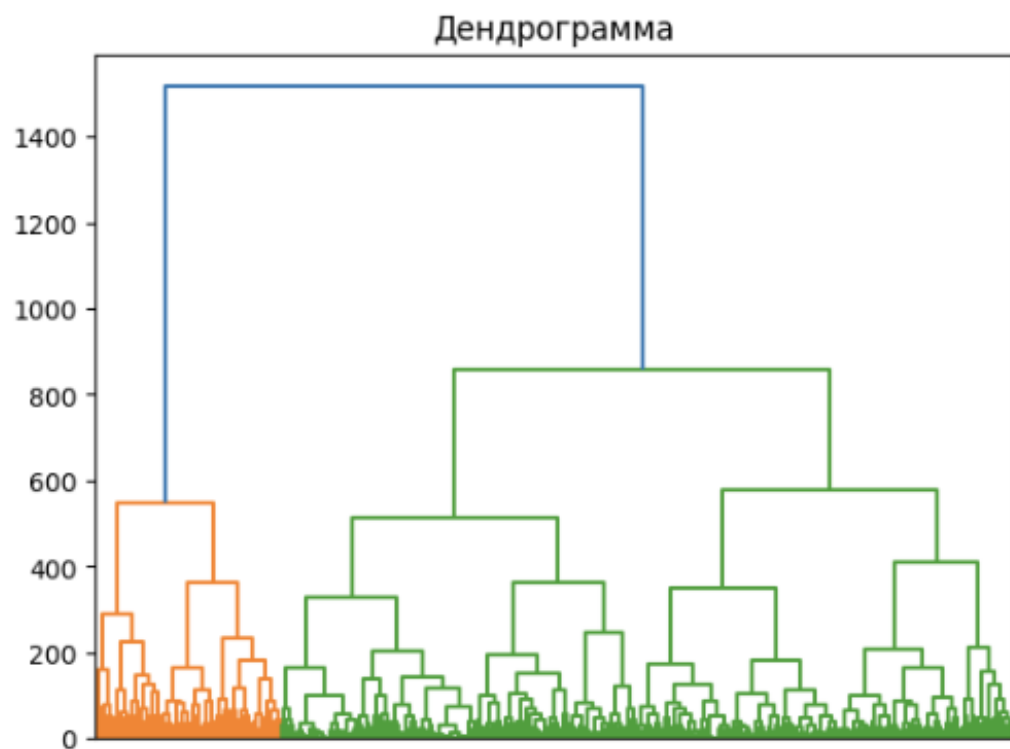
Наиболее частыми покупателями являются люди в возрасте от 32 лет до 41 года (50% наблюдений)

Пики покупательской активности приходятся на 33 года и 37 лет (по 7% наблюдений)



Кластеризация покупателей

Наиболее высокий показатель силуэта у метода K-средних без стандартизации с двумя кластерами, но как видно на дендрограмме, в этом случае один из кластеров будет значительно меньше. Допустимо сделать кластеризацию на 3 кластера.



Кластеризация покупателей

Кластеризация методом К-средних (Kmeans) с разделением на 3 кластера без стандартизации данных

Кластер	Стоимость, млн	Доля ПВ	Срок кредита, лет	Возраст	Кол-во	Доля, %
1	20.42	28.33	12.75	41.56	2685.00	41.06
2	19.49	20.43	27.34	35.01	2656.00	40.62
3	21.44	58.89	18.56	39.61	1198.00	18.32

Средняя стоимость покупки во всех кластерах примерно одинаковая

- Кластер 1 - Наиболее взрослые заемщики с небольшим сроком кредита. Средняя доля первоначального взноса 28%
- Кластер 2 - Наиболее молодые заемщики с большим сроком кредита. Средняя доля первоначального взноса 20%
- Кластер 3 - Заемщики с большим первоначальным взносом, в среднем 59%

Распределение по кластерам



Заключение

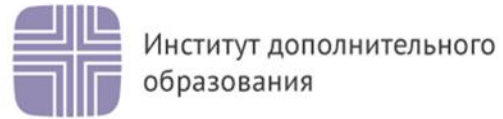
1. В данной работе была выбрана и обучена модель классификатор, позволяющая прогнозировать событие приобретения недвижимости клиентом с точностью 85%. Модель лучше предсказывает тех клиентов, которые совершают покупку, пропуская лишь 6% таких случаев. Это позволяет использовать ее для прогнозирования объема продаж.

2. Между датой подачи заявки и покупкой может наблюдаться существенный временной лаг, что снижает точность прогнозирования нулевого класса на новых наблюдениях.

3. Изначальный набор данных, состоящий из 13 признаков, является избыточным. Его можно сократить до 6, практически без потери качества прогноза.

4. Наиболее значимым признаком, влияющим на событие покупки, является возраст. Наиболее частыми покупателями являются клиенты в возрасте от 32 лет до 41 года.

5. Кластеризация покупателей на три группы показала, что группа с большим первоначальным взносом составляет лишь 18% от общего числа покупателей. В остальных случаях доля первоначального взноса в среднем менее 30%, что говорит о важности ипотечного кредитования.



ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

**ПРОГНОЗИРОВАНИЕ СОБЫТИЯ ПРИОБРЕТЕНИЯ
НЕДВИЖИМОСТИ ИПОТЕЧНЫМИ КЛИЕНТАМИ
СТРОИТЕЛЬНОЙ КОМПАНИИ И ИХ КЛАСТЕРИЗАЦИЯ**

по программе профессиональной переподготовки:

«Анализ данных на языке Python»

Выполнил: Лиситчук Антон Олегович
Руководитель: к.т.н. Семендяев Родион Юрьевич

Санкт-Петербург, 2023