

Сравнительный анализ трех задач

Кузнецов Г.И.

Введение

Проведем сравнение нескольких выполненных работ по нескольким ключевым позициям.

Для начала кратко вспомним основные концепции и задачи, поставленные в предыдущих работах. После чего сравним их согласно плану.

План сравнения:

- 1.Понимание бизнес-целей (Business Understanding)
- 2.Начальное изучение данных (Data Understanding)
- 3.Подготовка данных (Data Preparation)
- 4.Моделирование (Modeling)
- 5.Оценка (Evaluation)
- 6.Внедрение (Deployment)

Работа 1: Создание предиктивной модели рейтинга мобильных приложений

1. Провести анализ данных

2. Ответить на вопросы:

- как информация о приложении влияет на рейтинг пользователей?
- чем отличается статистика приложений для разных групп?

Unnamed: 0		id	track_name	size_bytes	currency	price	rating_count_tot	rating_count_ver	user_rating	user_rating_ver
0	1	281656475	PAC-MAN Premium	100788224	USD	3.99	21292	26	4.0	4.5
1	2	281796108	Evernote - stay organized	158578688	USD	0.00	161065	26	4.0	3.5
2	3	281940292	WeatherBug - Local Weather, Radar, Maps, Alerts	100524032	USD	0.00	188583	2822	3.5	4.5
3	4	282614216	eBay: Best App to Buy, Sell, Save! Online Shop...	128512000	USD	0.00	262241	649	4.0	4.5
4	5	282935706	Bible	92774400	USD	0.00	985920	5320	4.5	5.0

часть датасета

Работа 2: Построение предиктивной модели оценки надежности заемщика

1. Провести разведочный анализ данных по данным скоринга.
2. Определить достаточность данных для определения кредитного скоринга, социального скоринга.

Data columns (total 62 columns):

#	Column	Dtype
0	id	int64
1	rn	int64
2	pre_since_opened	int64
3	pre_since_confirmed	int64
4	pre_pterm	int64
5	pre_fterm	int64
6	pre_till_pclose	int64
7	pre_till_fclose	int64
8	pre_loans_credit_limit	int64
9	pre_loans_next_pay_summ	int64
10	pre_loans_outstanding	int64
11	pre_loans_total_overdue	int64
12	pre_loans_max_overdue_sum	int64
13	pre_loans_credit_cost_rate	int64
14	pre_loans5	int64
15	pre_loans530	int64
16	pre_loans3060	int64
17	pre_loans6090	int64
18	pre_loans90	int64
19	is_zero_loans5	int64
...		
60	fclose_flag	int64
61	flag	int64

число признаков

Работа 3: Разработка многофакторной модели для оценки стоимости недвижимости

1. Собрать данные по стоимости недвижимости в выбранном регионе.
2. Выбрать признаки, влияющие на стоимость разных типов недвижимости в регионе.
3. Вдвинуть гипотезы о том, какие признаки наиболее и наименее влияют на стоимость разных типов недвижимости, проверить гипотезы.
4. Смоделировать модель прогнозирования стоимости разных видов недвижимости в выбранном регионе на основе разных методов обучения, сравнить эффективность моделей.
5. Дать рекомендации покупателям и продавцам недвижимости.

	price	district	underground	street	house_number	floor
0	62598031.0	Дорогомилово	Москва-Сити	NaN	NaN	11
1	76250000.0	Пресненский	Москва-Сити	Краснопресненская набережная	14Ак1	25
2	28172548.0	Даниловский	Павелецкая	NaN	NaN	6
3	103000000.0	Красносельский	Сухареvская	переулок Даев	31С2	5
4	89850000.0	Пресненский	Улица 1905 года	NaN	NaN	2

часть датасета

floors_count	total_meters	rooms_count	residential_complex	url
18	45.8	2	Бадаевский ЖК	https://www.cian.ru/sale/flat/313877981/
61	61.0	2	Capital Towers	https://www.cian.ru/sale/flat/315795006/
27	40.5	2	Эра ЖК	https://www.cian.ru/sale/flat/301260611/
5	152.0	3	NaN	https://www.cian.ru/sale/flat/317646469/
22	99.7	2	Лайф Тайм ЖК	https://www.cian.ru/sale/flat/316363605/

Понимание бизнес-целей (Business Understanding)

Перед началом любого анализа данных важно четко определить цели проекта.

В этих трех задачах мы работали с разными бизнес-потребностями: от предсказания рейтинга мобильных приложений до оценки кредитоспособности заемщиков и прогнозирования стоимости недвижимости.

Каждый проект требовал уникального подхода к постановке задачи и интерпретации результатов.

Задача	Цель	Особенности
Рейтинг приложений	Анализ влияния данных на рейтинг, сравнение групп приложений	Простой датасет, мультиклассовая/регрессия
Скоринг заемщиков	Оценка надежности клиента (кредитный/социальный скоринг)	Сильный дисбаланс (97% vs 3%), сложный формат данных (.pq)
Оценка недвижимости	Прогнозирование цены на основе гибридных подходов	Самостоятельный сбор данных, проверка гипотез

Начальное изучение данных (Data Understanding)

На этом этапе мы провели первичный анализ данных, оценили их структуру, качество и потенциальные проблемы.

В зависимости от задачи, данные были представлены в разных форматах — от простых CSV-файлов до сложных паркет-таблиц.

Особое внимание пришлось уделить проблеме дисбаланса классов в задаче кредитного скоринга и самостоятельному сбору актуальных данных по недвижимости.

Задача	Источник данных	Проблемы
Рейтинг приложений	2 CSV-файла	Простые, чистые данные
Скоринг заемщиков	Паркет-файлы (.pq)	Дисбаланс, сложность загрузки
Недвижимость	Парсинг актуальных данных (2025)	Отсутствие готового датасета, feature engineering

Подготовка данных (Data Preparation)

Качество данных напрямую влияет на результат моделирования.

В каждом проекте мы выполнили очистку данных, обработку пропусков и преобразование признаков.

Для задачи с дисбалансом классов (3% положительных примеров) был применен метод undersampling. В случае с недвижимостью — проведен тщательный feature engineering на основе собранных данных.

Задача	Методы обработки	Фичи
Рейтинг приложений	Работа с категориальными признаками, в остальном чисто	Категориальные + числовые
Скоринг заемщиков	Undersampling (только 1 + случайные 0), обработка .pq	Социальные + кредитные признаки
Недвижимость	Парсинг, очистка, отбор признаков	Геоданные, параметры объектов

Моделирование (Modeling)

Для решения поставленных задач мы тестировали различные алгоритмы машинного обучения.

В проектах с рейтингом приложений и стоимостью недвижимости сравнивались четыре модели, тогда как в задаче кредитного скоринга из-за особенностей данных использовался только Random Forest. Выбор оптимальной модели осуществлялся на основе анализа метрик.

Задача	Модели	Лучшая модель
Рейтинг приложений	Линейная, Decision Tree, Random Forest, Gradient Boosting	Gradient Boosting
Скоринг заемщиков	Только Random Forest (из-за сложности данных)	Random Forest
Недвижимость	Линейная, Decision Tree, Random Forest, Gradient Boosting	Gradient Boosting

Оценка (Evaluation)

Оценка качества моделей проводилась с учетом специфики каждой задачи. Для регрессии использовались метрики MSE и R^2 , для классификации — precision, recall и F1-мера.

Анализ важности признаков позволил выявить ключевые факторы, влияющие на целевую переменную в каждом случае.

Задача	Метрики	Выводы
Рейтинг приложений	RMSE, R^2 , MAE	Важность описания, скриншотов
Скоринг заемщиков	F1, ROC-AUC, Classification report (precision, recall, accuracy)	Важность отдельных платежей, типа кредита, социальных признаков
Недвижимость	RMSE, MAE, R^2	Сильное влияние площади, инфраструктуры (район, близость к метро)

Внедрение (Deployment)

Результаты каждого проекта имеют практическую ценность для бизнеса.

Модель оценки рейтинга может помочь разработчикам улучшать свои приложения, кредитный скоринг — снизить риски банка, а прогнозирование цен на недвижимость — поддержать принятие решений покупателями и продавцами.

Успешное внедрение требует не только технической реализации, но и понятной интерпретации результатов для конечных пользователей.

Задача	Где применимо	Рекомендации
Рейтинг приложений	Рекомендации разработчикам	Улучшение описаний и категорий
Скоринг заемщиков	Банковский скоринг	Учет социальных данных
Недвижимость	Риелторские сервисы	Динамическое ценообразование

Выводы (по сравнительному анализу)

1. Разнообразие задач — разные подходы

Несмотря на общую структуру CRISP-DM, каждый проект потребовал уникальных решений: от борьбы с дисбалансом данных в скоринге до самостоятельного сбора актуальной информации по недвижимости.

2. Качество данных — основа успеха

Наиболее сложным этапом оказалась подготовка данных, особенно при работе с «сырыми» источниками (парсинг, паркет-файлы). Глубокий EDA и feature engineering критически важны для построения качественных моделей.

3. Выбор модели зависит от контекста

Gradient Boosting показал лучшие результаты в задачах регрессии. Простые модели (линейная регрессия) могут быть полезны для интерпретируемости.

4. Практическая ценность

Все модели решают конкретные бизнес-задачи: от снижения рисков банка до рекомендаций по ценообразованию. Важно не только построить модель, но и обеспечить её внедрение в рабочие процессы.

Спасибо за внимание