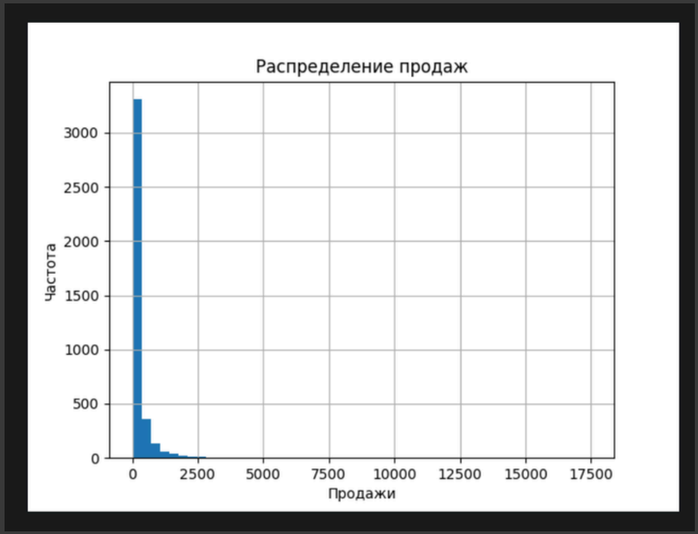
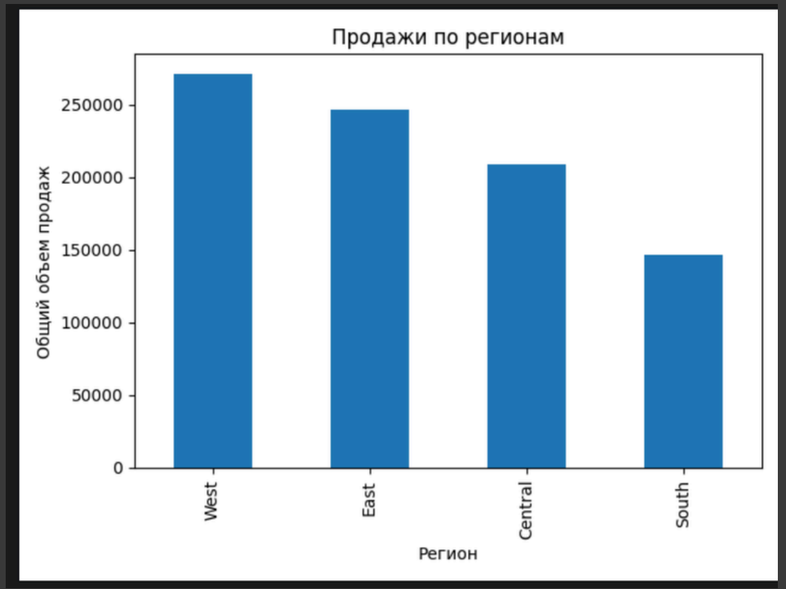
Постановка бизнес-задачи

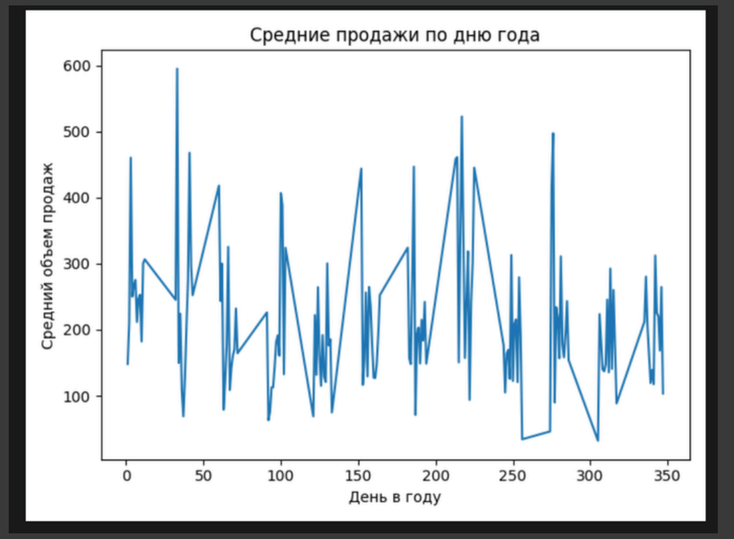
Целью проекта является анализ исторических данных о продажах и построение модели, способной предсказывать объём продаж на основе временных и категориальных признаков. Полученные результаты могут быть использованы для улучшения планирования, управления запасами и оценки влияния различных факторов на продажи.

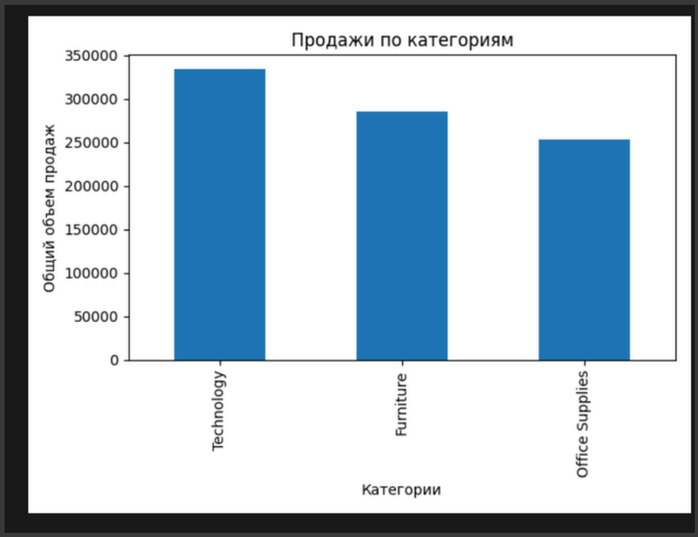
Course_Final/Final.py at main · elwaykabusiness-bot/Course_Final

Ключевые инсайты EDA

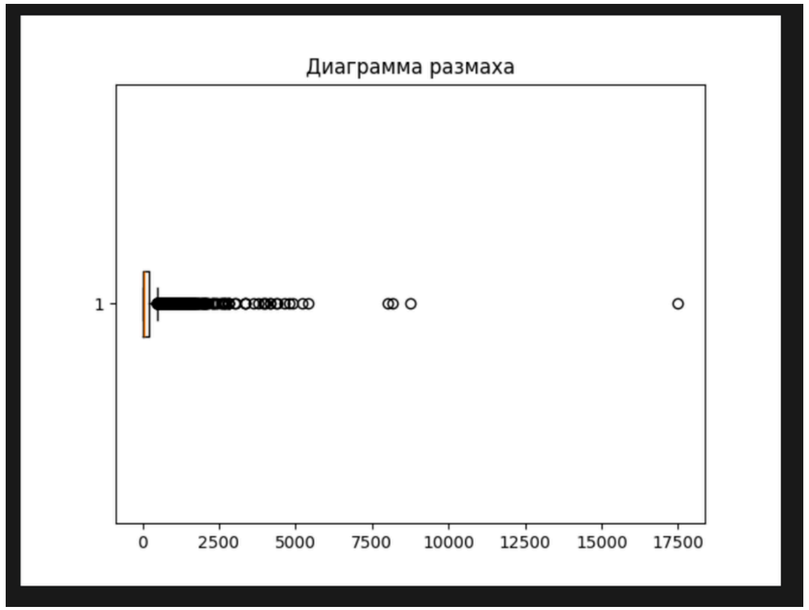








Ключевые инсайты EDA

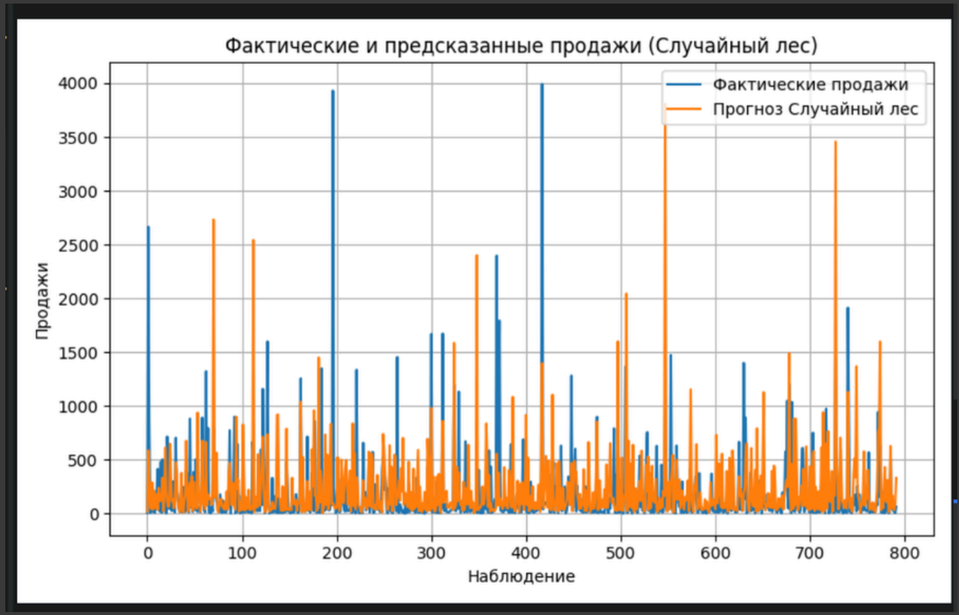


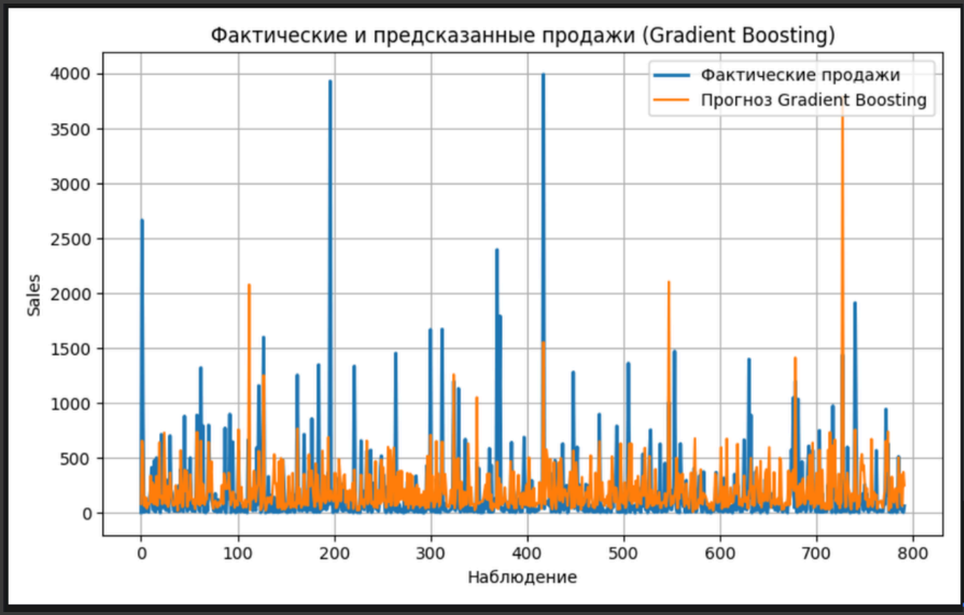
В результате EDA были получены ключевые инсайты: продажи имеют выраженную сезонность, различаются по регионам и категориям, а также в данных присутствуют выбросы, связанные с редкими крупными заказами.

Выбор и качество модели

Для решения задачи я выбрал четыре различные модели, относящиеся к разным классам алгоритмов.
Линейную регрессию как базовую интерпретируемую модель, случайный лес как ансамблевый метод на основе bagging, градиентный бустинг как ансамблевый метод на основе boosting и многослойный персептрон как нейросеть.
Все модели обучались и сравнивались на одинаковом разбиении данных, что обеспечивает корректность сравнения.

Выбор и качество модели





Выбор и качество модели

Для оценки качества моделей использовались метрики MAE, RMSE и коэффициент детерминации R^2. Оценка проводилась отдельно на обучающей и тестовой выборках, что позволило выявить переобучение и сравнить обобщающую способность моделей.

На обучающей выборке наилучшие результаты показал градиентный бустинг с минимальными значениями MAE = 180.79 и RMSE = 398.25, а также наибольшим значением R^2 = 0.535, что указывает на хорошую способность модели аппроксимировать обучающие данные. Линейная регрессия и случайный лес продемонстрировали сопоставимые, но более слабые результаты, с R^2 около 0.23–0.25. Многослойный персептрон показал умеренное качество на обучающей выборке, однако уступил градиентному бустингу по всем метрикам.

На тестовой выборке лучшие значения R^2 показали линейная регрессия (R^2 = 0.280) и случайный лес (R^2 = 0.279), при практически одинаковых значениях MAE и RMSE. Градиентный бустинг продемонстрировал заметное снижение качества на тестовых данных (R^2 = 0.170), что свидетельствует о переобучении модели. Многослойный персептрон показал наихудшие результаты на тестовой выборке, с наибольшими значениями ошибок и минимальным R^2 = 0.112.

Выбор и качество модели

```
Обучающие метрики

Линейная регрессия
MAE  = 198.54
RMSE = 511.82
R2   = 0.232

Случайный лес
MAE  = 188.97
RMSE = 505.93
R2   = 0.250

Градиентный бустинг
MAE  = 180.79
RMSE = 398.25
R2   = 0.535

Многослойный перспетрон
MAE  = 188.89
RMSE = 474.12
R2   = 0.341
```

```
Тестовая метрики

Линейная регрессия
MAE  = 195.95
RMSE = 406.46
R2   = 0.280

Случайный лес
MAE  = 194.01
RMSE = 406.77
R2   = 0.279

Градиентный бустинг
MAE  = 197.49
RMSE = 436.33
R2   = 0.170

Многослойный перспетрон
MAE  = 205.87
RMSE = 451.44
R2   = 0.112
```

Таким образом, несмотря на более высокое качество сложных моделей на обучающей выборке, наилучшую обобщающую способность на тестовых данных продемонстрировала линейная регрессия, что делает её наиболее устойчивой моделью в рамках текущего набора признаков.

Интерпретация и практическая ценность

Анализ важности признаков показал, что наибольшее влияние на объем продаж оказывают товарные характеристики. Самым значимым признаком является подкатегория Machines, за которой следуют категория Office Supplies и подкатегория Copiers. Это указывает на то, что тип и назначение товара вносят основной вклад в формирование объема продаж.

Значимое влияние также оказывают другие подкатегории товаров, такие как Furnishings, Accessories, Phones и Storage, а также категория Furniture, что подтверждает важность структуры ассортимента при прогнозировании продаж.

Временные признаки month и year имеют меньший вклад по сравнению с товарными характеристиками, однако наличие месяца среди наиболее значимых признаков указывает на умеренную сезонность продаж.

```
Интерпретация
cat__Sub-Category_Machines       0.231203
cat__Category_Office Supplies    0.186366
cat__Sub-Category_Copiers        0.132544
cat__Sub-Category_Furnishings    0.103553
num__month                       0.055654
cat__Category_Furniture          0.054727
cat__Sub-Category_Accessories    0.050086
cat__Sub-Category_Phones         0.040434
cat__Sub-Category_Storage        0.024143
num__year                        0.013531
dtype: float64
```

Интерпретация и практическая ценность

Таким образом, модель в первую очередь опирается на товарную и категориальную принадлежность, а временные факторы играют вспомогательную роль. Полученные результаты позволяют использовать модель для прогнозирования продаж по ассортименту, анализа вклада отдельных товарных групп и поддержки управленческих решений в области планирования запасов и управления ассортиментом.

Выводы

В работе был реализован полный цикл анализа данных и машинного обучения для прогнозирования продаж. Проведено сравнение нескольких моделей, показавшее, что более сложные алгоритмы склонны к переобучению, а наибольшее влияние на продажи оказывают товарные характеристики. Качество прогнозирования умеренное, что указывает на необходимость расширения набора признаков и дальнейшего улучшения модели.

Дальнейшие шаги и улучшения

Расширение набора признаков

Текущая модель использует в основном категориальные и простые временные признаки. Для повышения качества прогнозирования можно добавить:

- информацию о скидках и промоакциях,
- количество проданных единиц, прибыль и маржинальность,
- характеристики клиентов (частота покупок, средний чек),
- более детальные временные признаки (неделя, день недели, праздники).

Работа с выбросами и распределением целевой переменной

Возможные улучшения:

- логарифмирование целевой переменной,
- ограничение экстремальных значений,
- обучение моделей отдельно для разных сегментов товаров.

Использование специализированных моделей для временных рядов поскольку данные имеют временную природу.