1. **Бизнес-анализ:**

* *Определение бизнес-целей:*

Оптимизировать поставки продуктов в каждый магазин

* *Оценка текущей ситуации:*

Нет данных, вся информация в предоставленном дата сете:

* *Определение целей аналитики:*

предсказать продажи каждого продукта в конкретном магазине, выявить факторы высоких продаж

* *Подготовка плана проекта*

1. **Анализ данных:**

Данные предоставлены файлом csv. 8523 записи

|  |  |
| --- | --- |
| Variable | Description |
| Item\_Identifier | Id продукта |
| Item\_Weight | Вес продукта |
| Item\_Fat\_Content | Содержание жира в продукте |
| Item\_Visibility | %полок, отведенный под наш продукт в конкретном магазине |
| Item\_Type | Категория продукта |
| Item\_MRP | Максимальная цена продажи продукта |
| Outlet\_Identifier | Идентификатор магазина |
| Outlet\_Establishment\_Year | Год открытия магазина |
| Outlet\_Size | Площадь магазина |
| Outlet\_Location\_Type | Тип города, в котором расположен магазин |
| Outlet\_Type | Признак является ли магазин продуктовым или супермаркетом |
| Item\_Outlet\_Sales | Продажи продукта в конкретном магазине. Именно ее и надо предсказываь |

Построены зависимости переменных, оценена корреляция

1. **Подготовка данных:**

Данные поделены на train и test в соотношении 70:30

Обработаны пропущенные значения заменой на моду для категориальных признаков. Исправлены ошибки в написании.

Заполнен вес для всех продуктов.

Outlet\_Size заменена на численные значения. Создано несколько фич по средним значениям максимальной цены продажи продукта в каждом магазине.

Год открытия магазина не значим.

Сгруппировала id продукта по группам и применила LabelEncoder и к id продукта и к новому признаку по id продукта. Для оставшихся категориальных фич -> get\_dummies.

Нормализовать данные.

1. **Моделирование:**

Построены LinearRegression(), SVR(), RandomForestRegressor() с применением GridSearchCV. Для лучших вариантов этих 3х моделей применен StratifiedKFold и линейная регрессия в качестве алгоритма второго уровня. С помощью feature\_importances\_ определены 3 наиболее влияющих признака: Outlet\_Type\_Supermarket Type3, Outlet\_Type\_Grocery Store, Item\_MRP.

1. **Оценка решения:**

Качество моделей оценено с помощью R^2 ~ 0,6. Ну так, надо бы ещё посмотреть на данные. Возможно определить среднее значение таргета по продукту в трейновой выборке. А также попробовать применить более сложные модели, которые ещё не изучены.. Больше внимания в анализ и подготовку данных.

1. **Внедрение**

Работа залита на github с комментариями, что и является итоговым документом.