

АНАЛИЗ КНИЖНЫХ МАГАЗИНОВ

ЦЕЛИ РАБОТЫ

- Собрать данные из 3 книжных магазинов
- Создать модель, наиболее точно предсказывающую цену книги по собранным выше признакам
- Создать модель, наиболее точно предсказывающую магазин, в котором была совершена покупка по собранным выше признакам

СБОР ДАННЫХ

Сбор данных производился по следующему алгоритму:

1. В качестве магазинов были выбраны: Книжный лабиринт, Читай-город, Молодая Гвардия.
2. В каждый из них приходил один из 3 экспертов. С помощью рандомайзера он выбирал случайный отдел, случайный стеллаж и случайную книгу на нем.
3. Были собраны следующие признаки:
 - категория книги,
 - название книги,
 - название магазина
 - цена
 - количество страниц,
 - визуальная привлекательность (1 - 5)
 - качество обложки (1 - 5)
 - качество переплёта (1 - 5)
 - качество бумаги (1 - 5)
 - качество печати (1 - 5)
 - формат книги (0-3)
 - размер книги (0-3)
 - номер эксперта(1-3)

АНАЛИЗ И ПРЕДОБРАБОТКА СОБРАННЫХ ДАННЫХ

Для числовых признаков были рассчитаны средние(в том числе и для каждого представленного магазина), отклонения и квантили различных уровней. С данными можно ознакомиться в файле с вычислениями.

МАТРИЦА КОРЕЛЛЯЦИЙ ПРИЗНАКОВ

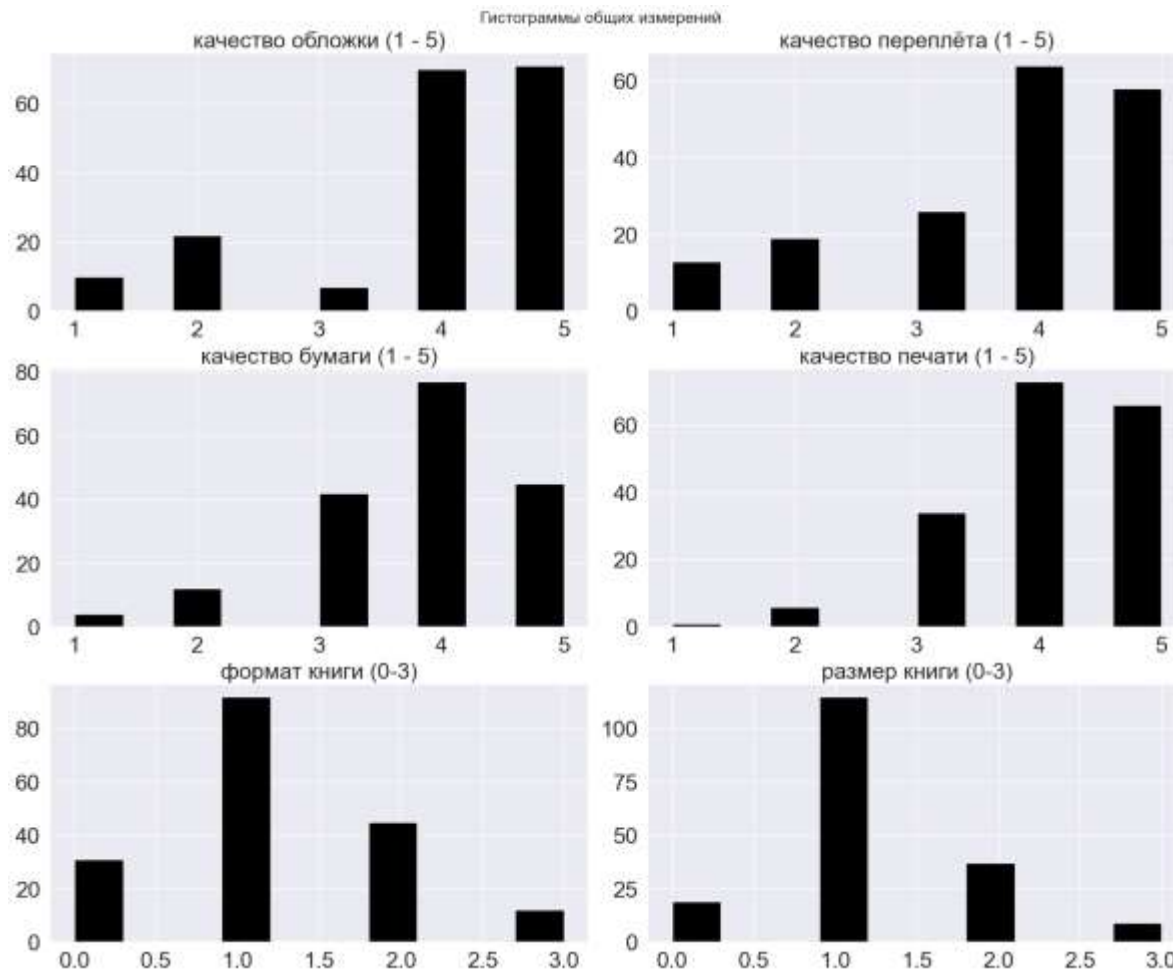
	цена	количество страниц	визуальная привлекательность (1-5)	качество обложки (1-5)	качество переплетения (1-5)	качество бумаги (1-5)	качество печати (1-5)	формат книги (0-3)	размер книги (0-3)	экспорт	классический дизайн	модерн дизайн	новый город
цена	1.00	0.18	0.06	0.24	0.20	0.38	0.34	0.48	0.45	-0.01	0.64	-0.28	-0.38
количество страниц	0.18	1.00	0.06	-0.02	-0.12	-0.17	-0.10	0.18	0.39	-0.10	-0.07	0.04	0.03
визуальная привлекательность (1-5)	0.06	0.06	1.00	0.34	0.26	0.21	0.15	0.06	0.06	-0.18	-0.06	-0.09	0.15
качество обложки (1-5)	0.24	-0.02	0.34	1.00	0.78	0.48	0.29	0.40	0.35	0.01	0.19	-0.03	-0.08
качество переплетения (1-5)	0.20	-0.12	0.26	0.78	1.00	0.51	0.35	0.34	0.25	0.17	0.14	0.02	-0.16
качество бумаги (1-5)	0.38	-0.17	0.21	0.48	0.51	1.00	0.54	0.38	0.31	0.06	-0.13	0.04	-0.18
качество печати (1-5)	0.34	-0.10	0.15	0.29	0.35	0.54	1.00	0.18	0.19	0.25	0.23	-0.18	-0.38
формат книги (0-3)	0.48	0.18	0.06	0.40	0.34	0.38	0.18	1.00	0.77	0.03	0.15	-0.02	-0.13
размер книги (0-3)	0.45	0.39	0.06	0.35	0.25	0.31	0.19	0.77	1.00	0.02	-0.12	-0.02	-0.10
экспорт	-0.01	-0.10	-0.18	0.01	0.17	0.06	0.25	0.03	0.02	1.00	-0.00	0.48	-0.48
классический дизайн	0.64	-0.07	-0.06	0.19	0.14	0.15	0.23	0.15	0.12	-0.00	1.00	-0.50	-0.50
модерн дизайн	-0.28	0.04	-0.09	-0.03	0.02	0.04	-0.02	-0.02	-0.02	0.48	-0.50	1.00	-0.50
новый город	-0.38	0.03	-0.10	-0.08	-0.16	-0.18	-0.38	-0.13	-0.10	-0.48	-0.50	-0.50	1.00

АНАЛИЗ ОСНОВНЫХ КОРРЕЛЯЦИЙ

Из данной таблицы можно сделать следующие выводы:

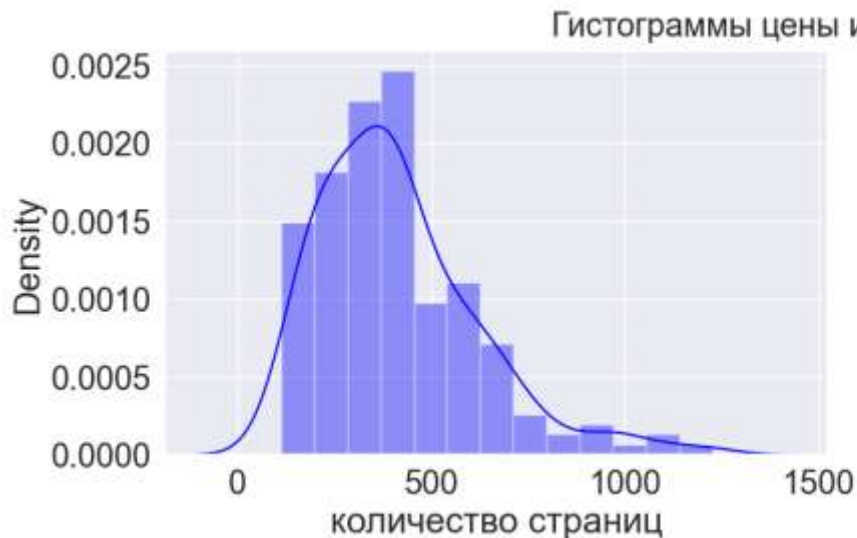
- Цены на книги конкретно в магазине "Книжный Лабиринт" сильно больше, чем в остальных магазинах.
- Субъективные параметры(качество бумаги, качество печати и т.д.) достаточно сильно зависят друг от друга.
- Размер книги достаточно сильно зависит от формата книги(что впринципе логично).
- Есть связь между магазинами равна -0.5, т.к. это бинаризованный признак, и у одной книги одновременно может быть только один магазин.

Гистограммы признаков

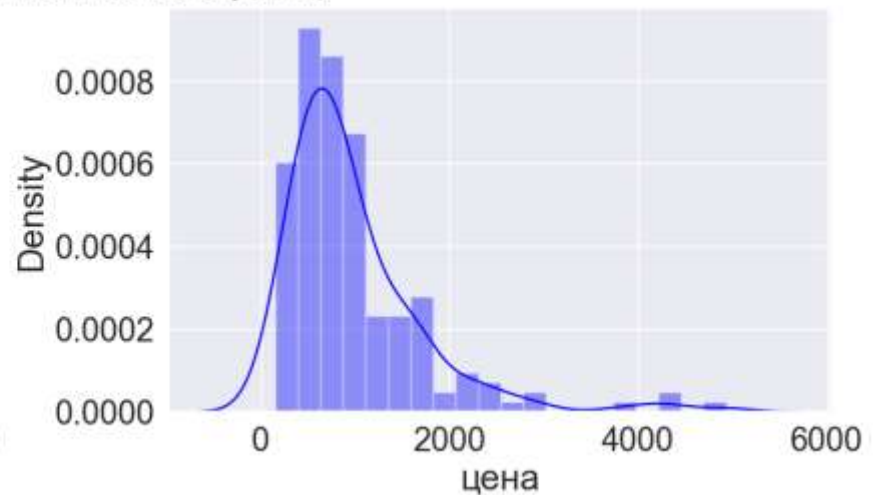


ПРОВЕРКА НА НОРМАЛЬНОСТЬ

Гистограммы цены и количества страниц имеют распределения, похожие на нормальные. Проверим это с помощью статистических тестов Андерсона-Дарлинга и Шапиро-Уилка:



A-D: statistic = 9.64, critical value = 1,069, alpha = 0.01 –отвергаем
S-U: statistic=0.77,
pvalue=2.69e-15 - отвергаем



A-D: statistic = 3.21, critical value = 1,069, alpha = 0.01 –отвергаем
S-U: 0.91, pvalue=1.83e-08 - отвергаем

ПРЕДСКАЗЫВАНИЕ ЦЕНЫ КНИГИ

Бинаризуем признаки “магазин”, “категория книги” и построим 3 регрессии: Lasso regression, RandomForestRegressor, XGBRegressor. Для каждой из них определим оптимальные параметры.

Получим следующие результаты:

Название	Lasso regression	RandomForest Regressor	XGBRegressor
Результаты с изначальными гиперпараметрами	351.44	350.72	357.50
Результаты с оптимальными гиперпараметрами	331.29	328.17	302.81

Наилучший результат показала модель XGBRegressor с rmse 302.81.

ПРЕДСКАЗАНИЕ МАГАЗИНА

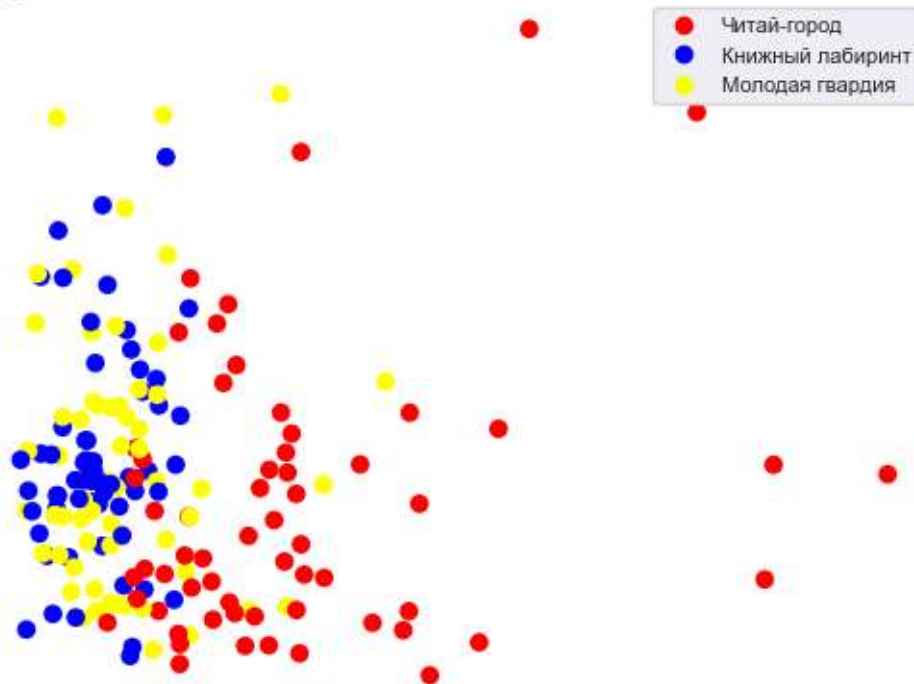
В данном случае использовано 4 классификатора:
RandomForestClassifier, XGBClassifier, GaussianNB, VotingClassifier(из
всех 3 данных классификаторов). Метрика – accuracy.

Название	RandomForest Classifier	XGBClassifier	GaussianNB	VotingClassifier
Результаты с изначальными гиперпараметрам и	0.64	0.64	0.58	0.64
Результаты с оптимальными гиперпараметрам и	0.70	0.76	0.58	0.76(все веса помимо XGBClassifier занулились)

ВИЗУАЛЬНОЕ ПРЕДСТАВЛЕНИЕ ДАННЫХ

Было выполнено понижение размерности до 2 с помощью PCA.

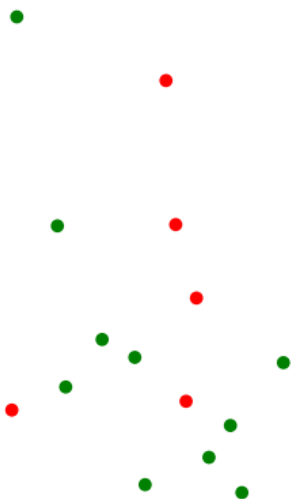
Распределение всех книг по магазинам



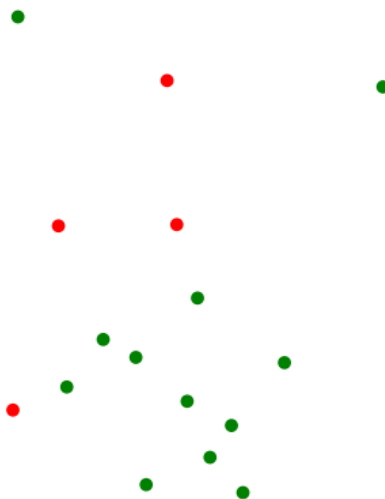
ВИЗУАЛЬНОЕ ПРЕДСТАВЛЕНИЕ КАЖДОГО КЛАССИФИКАТОРА

Зеленым цветом показано правильное предсказание, красным – неверное. Данные, используемые на тренировке, не рассматриваются.

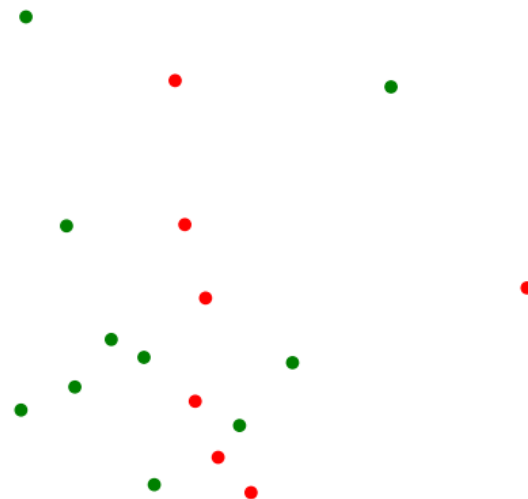
RF



XGB



BAY



ИТОГИ

- Была получена модель, предсказывающая цену книги со среднеквадратичной ошибкой 302
- Была получена модель, предсказывающая магазин с accuracy = 0.76

Спасибо за внимание!

Со кодом вы можете ознакомиться по ссылке:

https://github.com/Snackkie/Book_analysis/blob/main/book_analysis.ipynb

С самими данными вы можете ознакомиться по ссылке:

https://github.com/Snackkie/Book_analysis/blob/main/book_reviews.xlsx