Министерство цифрового развития

Федеральное государственное бюджетное образовательное учреждение высшего образования
«Сибирский государственный университет телекоммуникаций и информатики»
(СибГУТИ)

Кафедра прикладной математики и кибернетики

Отчёт

по лабораторной работе № 4 «Классификация методом дерева решений»

Выполнил:

студент группы ИП-312

Прозоренко К.В

Работу проверил: старший преподаватель кафедры ПМиК
Дементьева К.И.

Новосибирск 2025 г.

Цель: Сформировать комплексное понимание различных методов регрессионного анализа и выработать навыки осознанного выбора моделей в зависимости от характеристик данных и решаемой задачи.

Ход выполнения:

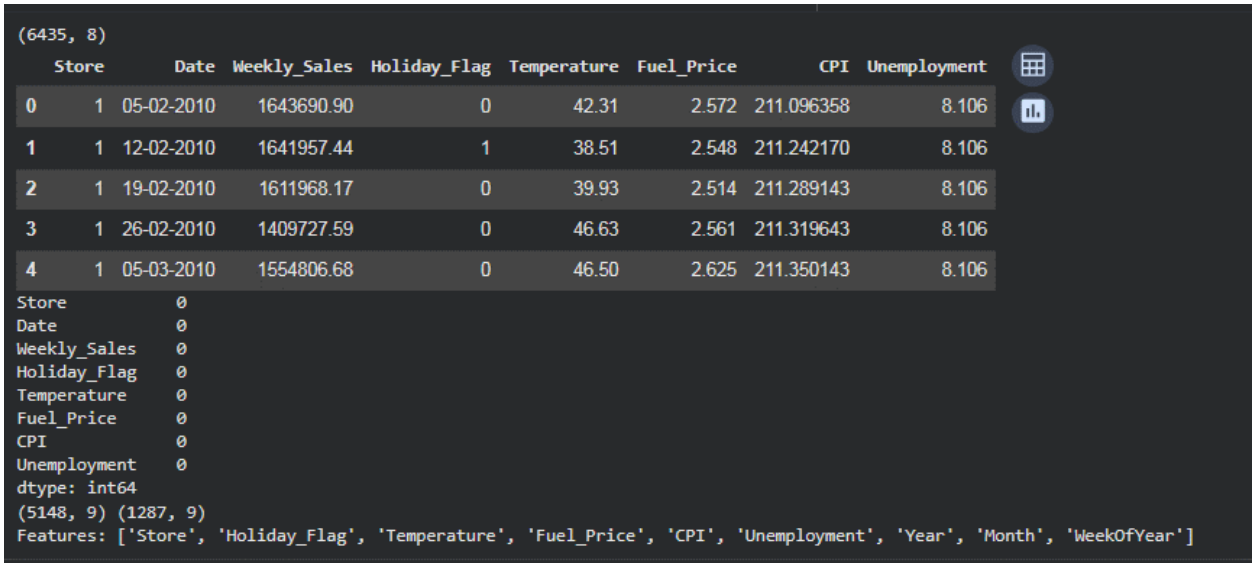
Датасет: Walmart_Sales.csv (6435 строк, 8 исходных столбцов), целевая переменная - Weekly_Sales, признаки включают Store, Date, Holiday_Flag, Temperature, Fuel_Price, CPI, Unemployment.

После преобразования Date извлечены признаки Year, Month, WeekOfYear, что позволяет учитывать сезонность без сложных временных моделей.

Предобработка данных

- Пропущенные значения: в выбранной версии датасета пропусков не обнаружено (все суммы NaN = 0), поэтому imputation не потребовался.
- Кодирование категориальных признаков: Store (и другие категориальные при наличии) закодированы методом Label Encoding по условию задания.
- Разделение данных: выполнен split на обучающую и тестовую выборки (80/20) с фиксированным random_state для воспроизводимости.
- Масштабирование: для Lasso/ElasticNet применён StandardScaler, т.к. регуляризация чувствительна к масштабу признаков (штраф зависит от величины коэффициентов).

Что важно отметить как ограничение: Label Encoding для Store превращает магазин в “числовую ось”, и линейная модель может интерпретировать это как упорядоченность, хотя по смыслу это просто идентификатор.



(6435, 8)

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
0	1	05-02-2010	1643690.90	0	42.31	2.572	211.096358	8.106
1	1	12-02-2010	1641957.44	1	38.51	2.548	211.242170	8.106
2	1	19-02-2010	1611968.17	0	39.93	2.514	211.289143	8.106
3	1	26-02-2010	1409727.59	0	46.63	2.561	211.319643	8.106
4	1	05-03-2010	1554806.68	0	46.50	2.625	211.350143	8.106

```
Store           0
Date            0
Weekly_Sales    0
Holiday_Flag    0
Temperature     0
Fuel_Price      0
CPI             0
Unemployment    0
dtype: int64
(5148, 9) (1287, 9)
Features: ['Store', 'Holiday_Flag', 'Temperature', 'Fuel_Price', 'CPI', 'Unemployment', 'Year', 'Month', 'WeekOfYear']
```

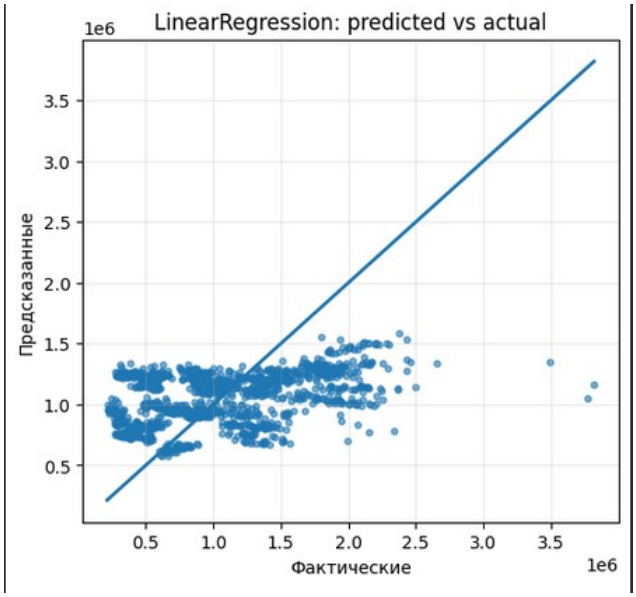
Обучение моделей и качество

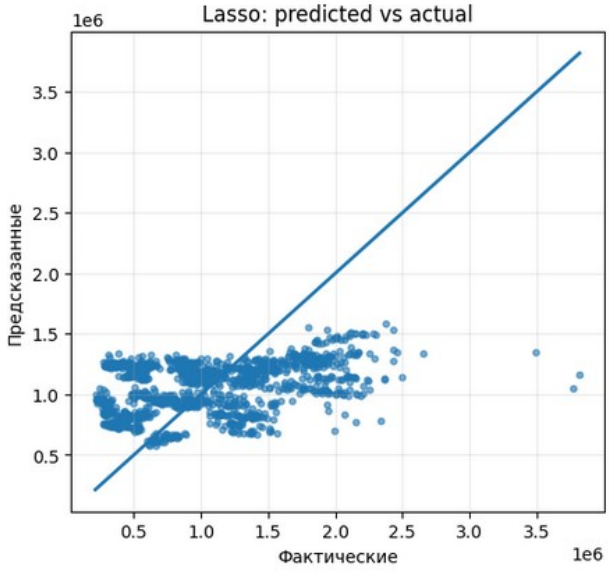
Обучены базовые модели: LinearRegression, Lasso, ElasticNet.

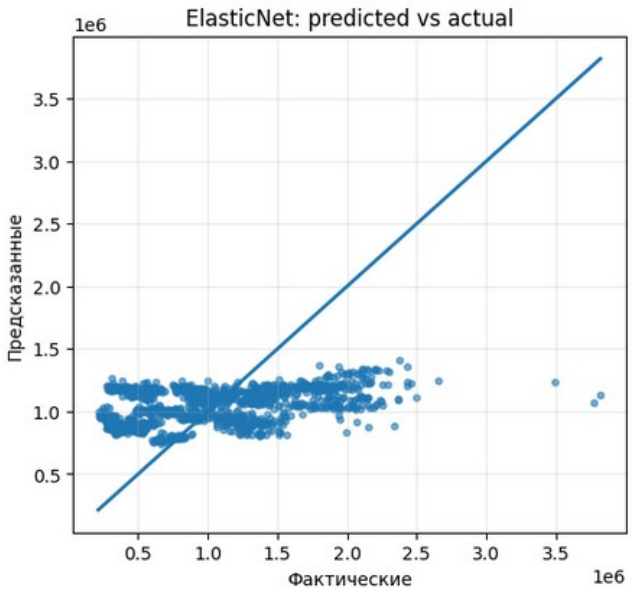
Метрики на тесте: MSE, RMSE, MAE, R^2.

Model	MSE	RMSE	MAE	R2
LinearRegression	2.741099e+11	523555.075800	433079.082388	0.149135
Lasso	2.741113e+11	523556.401001	433080.277580	0.149131
ElasticNet	2.818143e+11	530861.859842	443202.441878	0.125220

Графики predicted vs actual:
В ваших графиках заметен существенный разброс, то есть линейные признаки объясняют ограниченную долю вариации продаж.







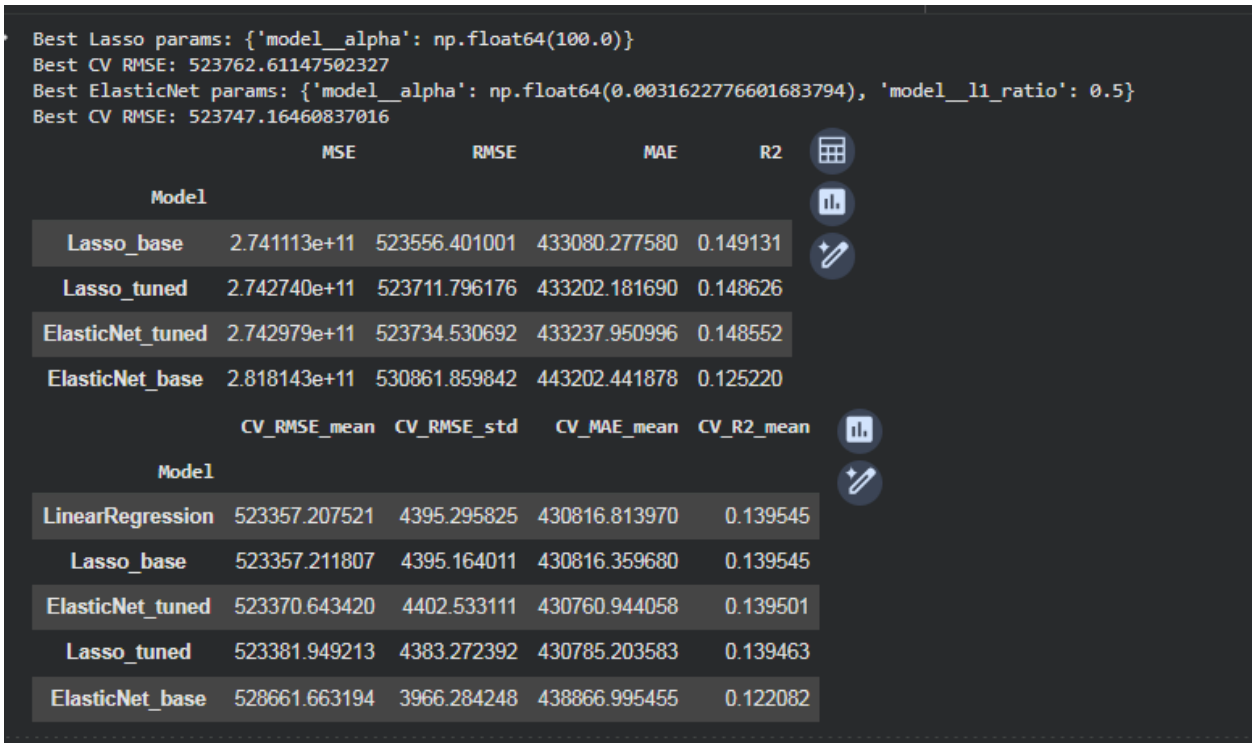
Сравнение коэффициентов:

- LinearRegression показывает “базовый” вклад признаков.
- Lasso добавляет L1-регуляризацию, которая может уменьшать часть коэффициентов вплоть до нуля (эффект отбора признаков).
- ElasticNet комбинирует L1 и L2, что часто полезно при коррелированных признаках; параметр l1_ratio задаёт долю L1 в смеси.

Коэффициенты LinearRegression и Lasso (base) почти совпадают, что указывает на слабое влияние регуляризации при выбранном alpha; в ElasticNet (base) веса сжаты сильнее, что привело к недообучению и ухудшению метрик.

	LinearRegression	Lasso	ElasticNet
Month	-198485.914781	-197256.482566	-1.767979
WeekOfYear	198184.864324	196957.521571	551.864766
Store	-195967.554186	-195965.705611	-127076.623021
CPI	-90341.076777	-90340.797549	-48092.318655
Unemployment	-45677.246482	-45679.694061	-32328.526120
Holiday_Flag	19808.164334	19804.433327	13303.685209
Temperature	-15629.884440	-15620.713326	-15422.319697
Year	-5669.604452	-5658.065622	-2692.834416
Fuel_Price	-126.309600	-125.226188	2065.164020

Подбор гиперпараметров и кросс-валидация



```
Best Lasso params: {'model__alpha': np.float64(100.0)}
Best CV RMSE: 523762.61147502327
Best ElasticNet params: {'model__alpha': np.float64(0.0031622776601683794), 'model__l1_ratio': 0.5}
Best CV RMSE: 523747.16460837016
```

Model	MSE	RMSE	MAE	R2
Lasso_base	2.741113e+11	523556.401001	433080.277580	0.149131
Lasso_tuned	2.742740e+11	523711.796176	433202.181690	0.148626
ElasticNet_tuned	2.742979e+11	523734.530692	433237.950996	0.148552
ElasticNet_base	2.818143e+11	530861.859842	443202.441878	0.125220

Model	CV_RMSE_mean	CV_RMSE_std	CV_MAE_mean	CV_R2_mean
LinearRegression	523357.207521	4395.295825	430816.813970	0.139545
Lasso_base	523357.211807	4395.164011	430816.359680	0.139545
ElasticNet_tuned	523370.643420	4402.533111	430760.944058	0.139501
Lasso_tuned	523381.949213	4383.272392	430785.203583	0.139463
ElasticNet_base	528661.663194	3966.284248	438866.995455	0.122082

По моим результатам тюнинг Lasso качество не улучшил (на тесте RMSE стало немного хуже, R^2 чуть снизился) - это нормальная ситуация, когда регуляризация не даёт выигрыша на данных, где линейная модель уже близка к оптимуму, а лишний штраф только добавляет смещение.

Для ElasticNet тюнинг заметно улучшил качество относительно “базового” ElasticNet, потому что базовые параметры давали слишком сильную регуляризацию, а подбор нашёл более подходящую.

Вывод: На текущих признаках линейные модели дают близкое качество; Lasso не дал прироста после тюнинга, а ElasticNet существенно улучшился после подбора гиперпараметров, что указывает на чувствительность ElasticNet к настройкам регуляризации и ограниченную объясняющую способность линейной зависимости для Weekly_Sales.

Ссылка на Google Collab: https://colab.research.google.com/drive/12g4zdglMl-aBSS6cBxhEWqkXLlvA4jL4?usp=sharing