Московский финансово-промышленный университет «Синергия»

ИТОГОВЫЙ ПРОЕКТ

по курсу «Инженер машинного обучения»

«АВТОМАТИЗАЦИЯ ОБРАБОТКИ ДАННЫХ»

Слушатель: Егоров А. А.
Группа: группа ИМО-92501/2МОп

Москва 2025

Сводный отчёт по продажам (мульти-источник)

Единицы измерения: суммы — руб.; количество — шт.

Данные: SQL/CSV/Excel/API.

Классификация: целевая метка = чек ≥ 80-й перцентиль по сумме чека.

Метрики: ROC-AUC и F1 для классификации; RMSE/MAE/R^2 для регрессии.

Идентификаторы трактуются как категориальные ключи; графики id↔id не строятся.

Сравнение источников (проверка с данными из БД)

Таблица

Источник	Строк, шт.	Сумма, руб.	Средний чек, руб.
api	200	13689.81	68.45
db	2000	169506.52	84.75
file	2200	169506.52	84.75

Числовые сводки (basic_stats)

Числовые сводки (basic_stats)

Колонка	count	mean	std	min	25%	50%	75%	max
amount	4200	83.98	76.02	0.41	30.13	61.62	113.9	609.4

Источник api: Числовые сводки (basic_stats)

Колонка	count	mean	std	min	25%	50%	75%	max
amount	200	68.45	56.78	2.06	25.25	53.08	95.37	250.2

Источник db: Числовые сводки (basic_stats)

Колонка	count	mean	std	min	25%	50%	75%	max
amount	2000	84.75	76.78	0.41	30.72	62.55	114.1	609.4

Источник file: Числовые сводки (basic_stats)

Колонка	count	mean	std	min	25%	50%	75%	max
amount	2000	84.75	76.78	0.41	30.72	62.55	114.1	609.4

ML-результаты

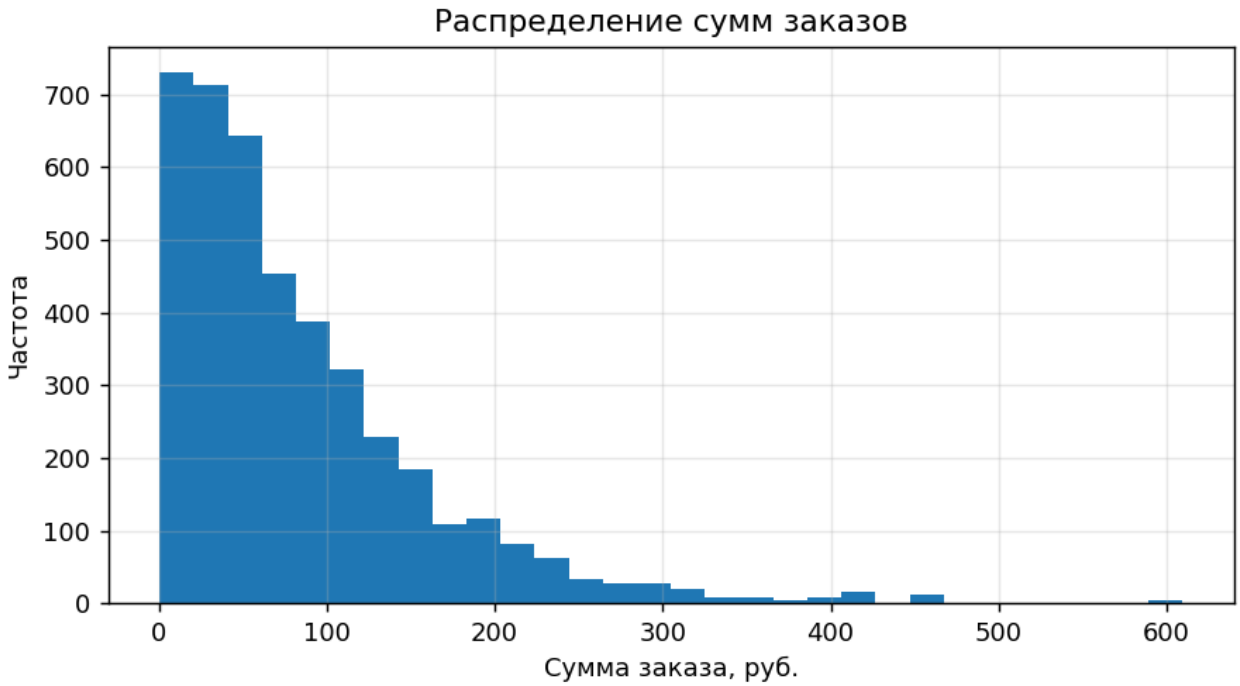
Коэффициенты LogReg (топ-20 по |коэф.|)

Признак	Коэффициент	|Коэф.|
dow	0.054581	0.054581
month	0.046309	0.046309
is_weekend	0.013135	0.013135
_amount_prev_mean	0.011403	0.011403
_cnt_prev	-0.006865	0.006865

Важность признаков (RandomForest, топ-20)

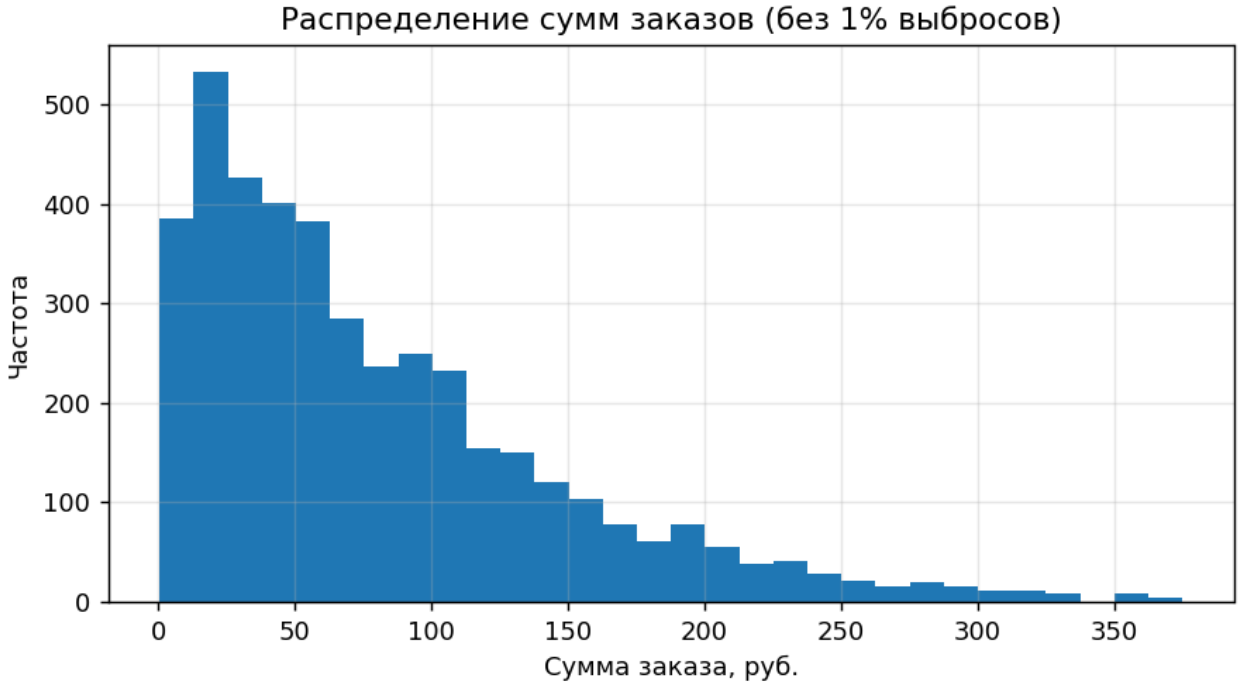
Признак	Важность
_amount_prev_mean	0.465811
month	0.202643
_cnt_prev	0.192667
dow	0.124672
is_weekend	0.014206

Рисунок 1. amount_hist.png



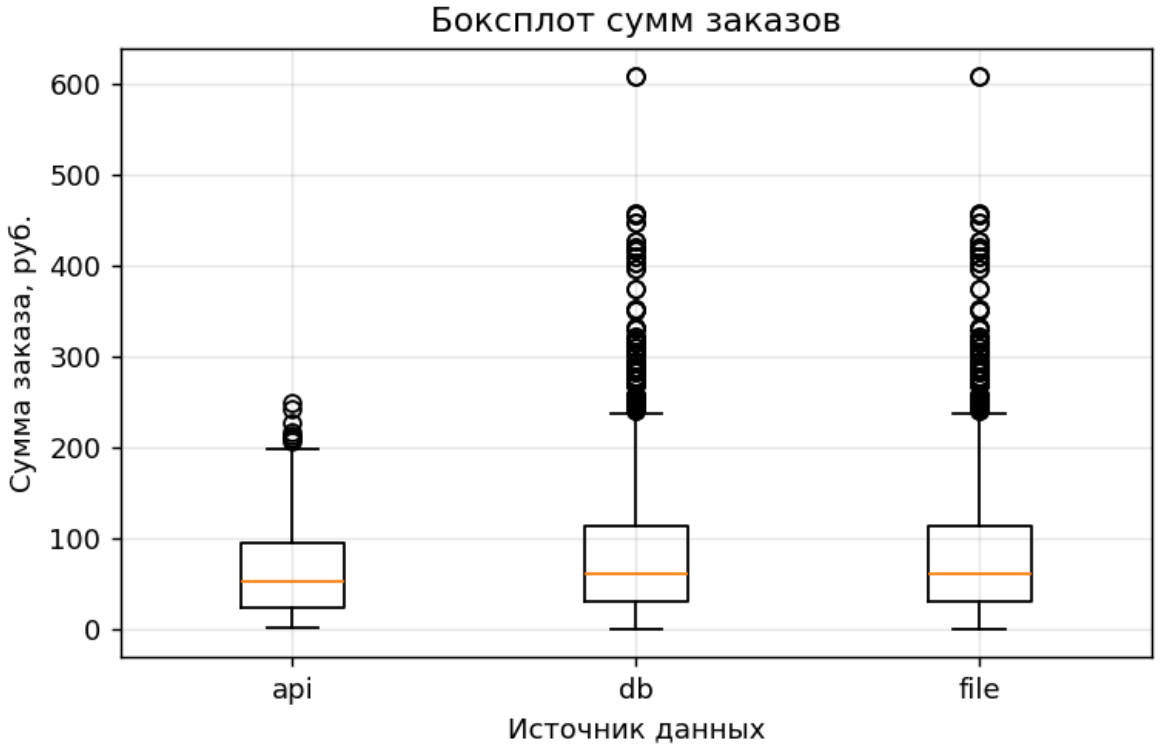
Гистограмма выручка; ось X — Выручка, руб.; ось Y — частота.

Рисунок 2. amount_hist_trim.png



Гистограмма выручка (<= Q0.99); X — Выручка, руб.; Y — частота.

Рисунок 3. amount_box.png

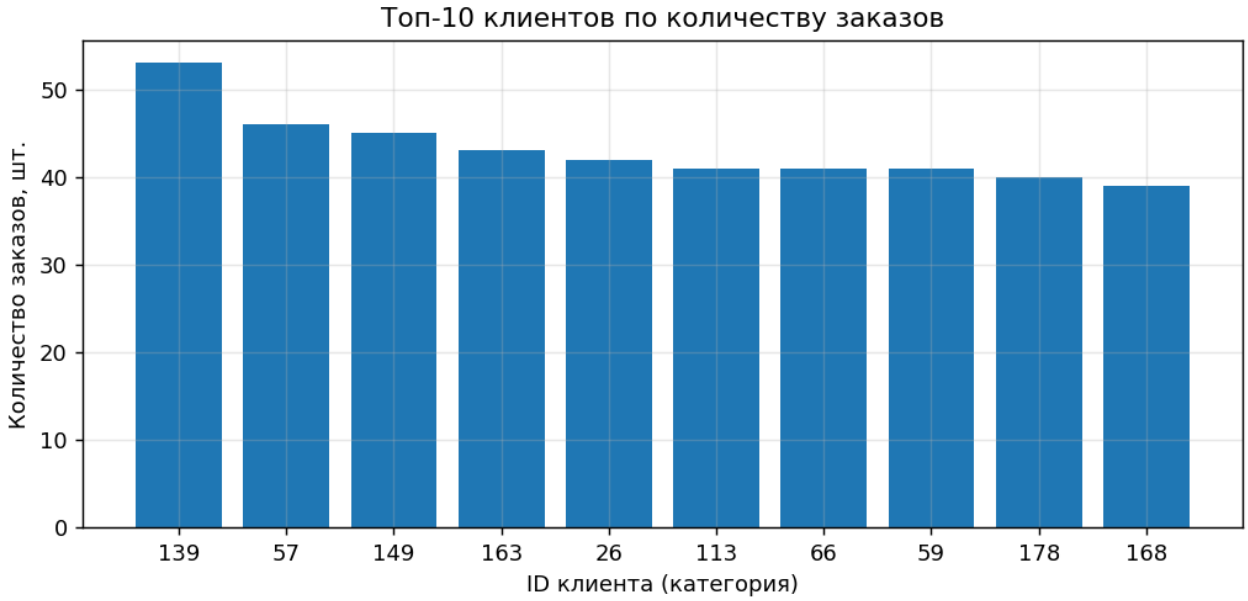


Боксплот по источникам; X — источник; Y — Выручка, руб..

Рисунок 4. orders_over_time.png

Динамика количества заказов; X — дата; Y — Заказы, шт.

Рисунок 5. top_customers.png

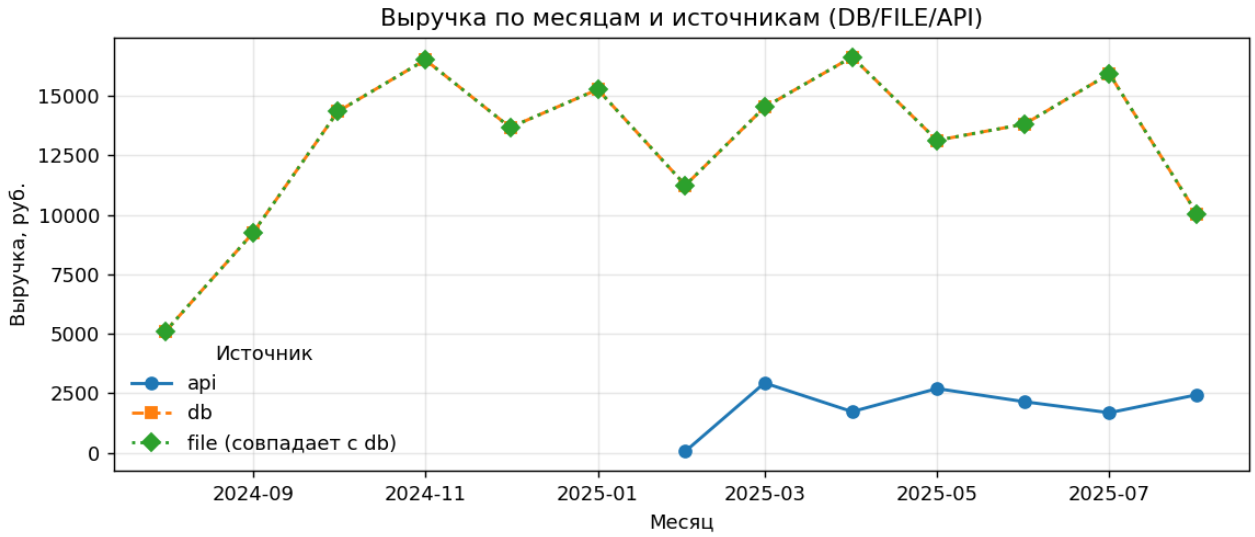


Рейтинг клиентов по числу заказов; X — ID клиента (категория); Y — Заказы, шт.

Рисунок 6. monthly_revenue.png

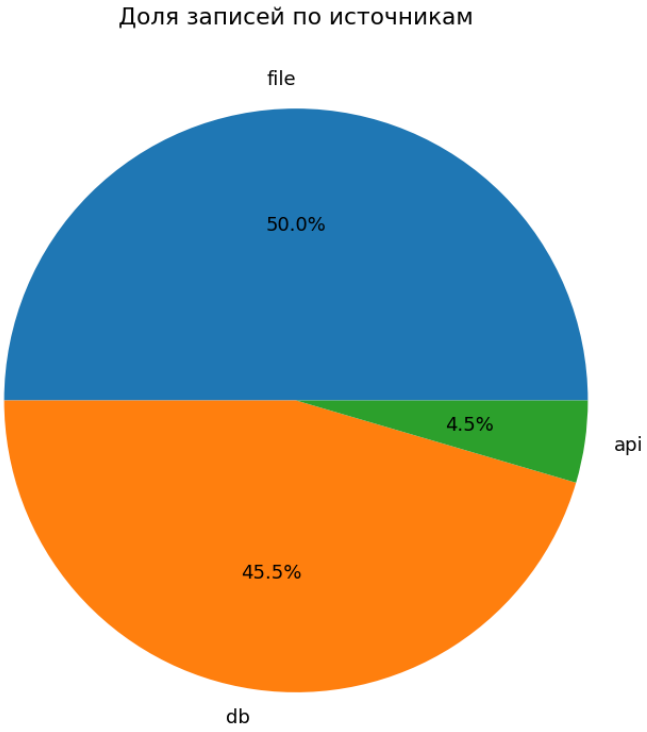
Выручка по месяцам; X — месяц; Y — Выручка, руб..

Рисунок 7. monthly_revenue_by_source.png



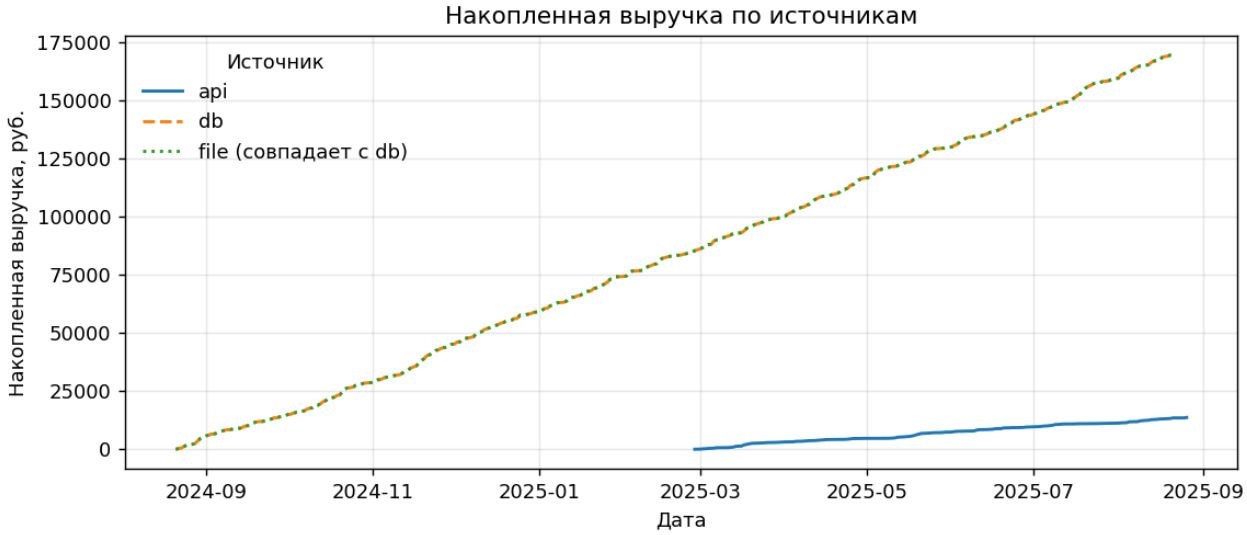
Выручка по месяцам с разбивкой по источникам; X — месяц; Y — Выручка, руб.. Совпадающие кривые помечены.

Рисунок 8. source_share_pie.png



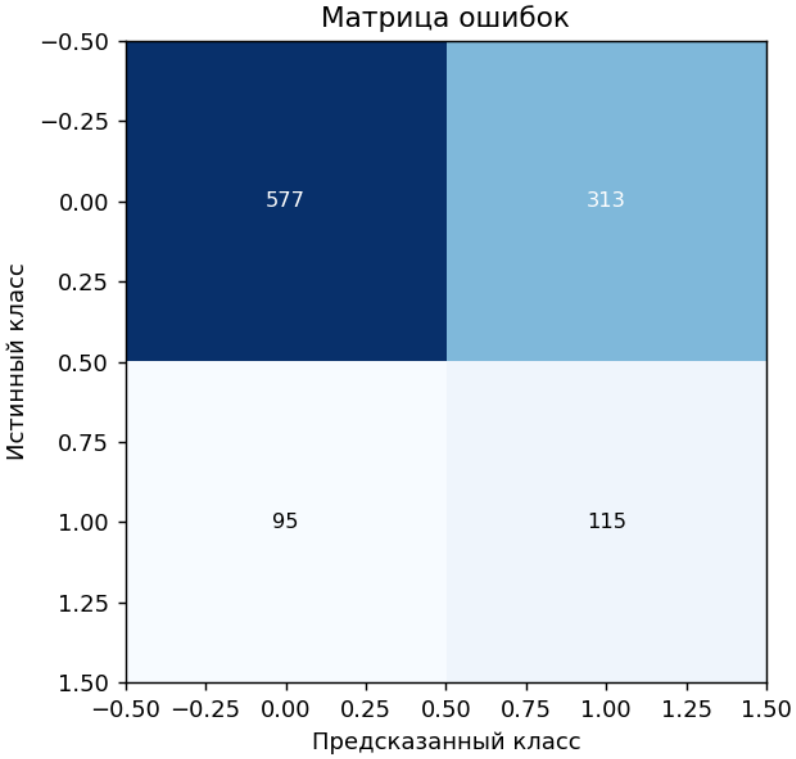
Доли источников в объединённом наборе.

Рисунок 9. cumulative_revenue_by_source.png



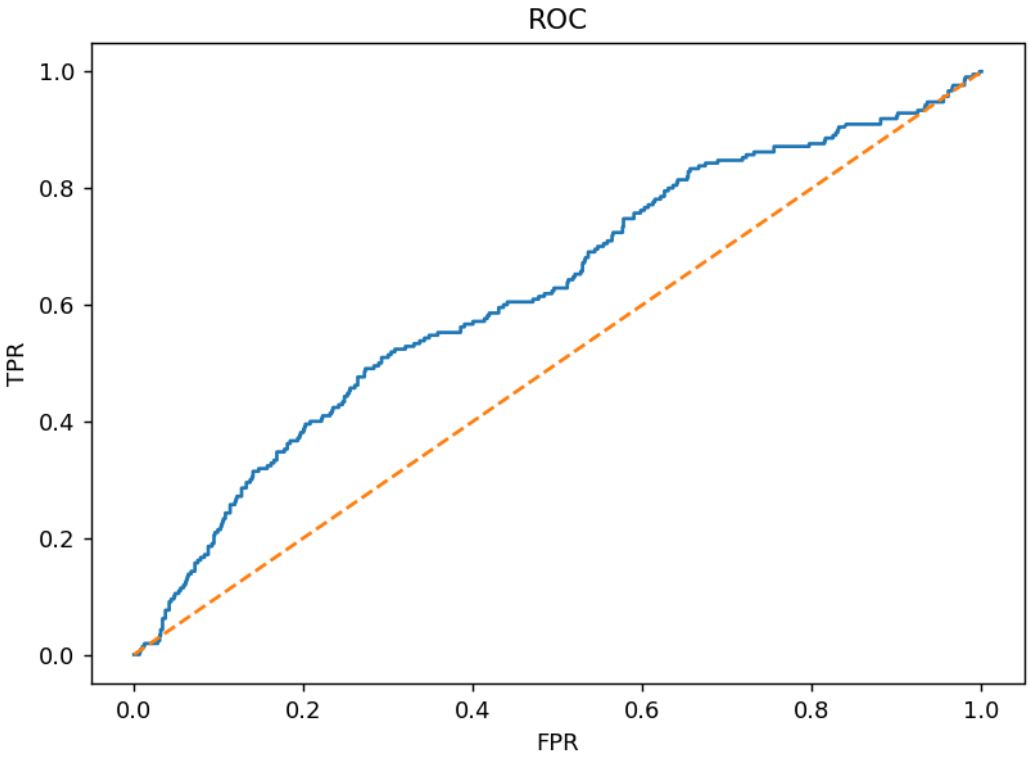
Накопленные суммы по источникам; X — дата; Y — Выручка (накопленная), руб.. Совпадающие кривые помечены.

Рисунок 10. clf_confusion.png



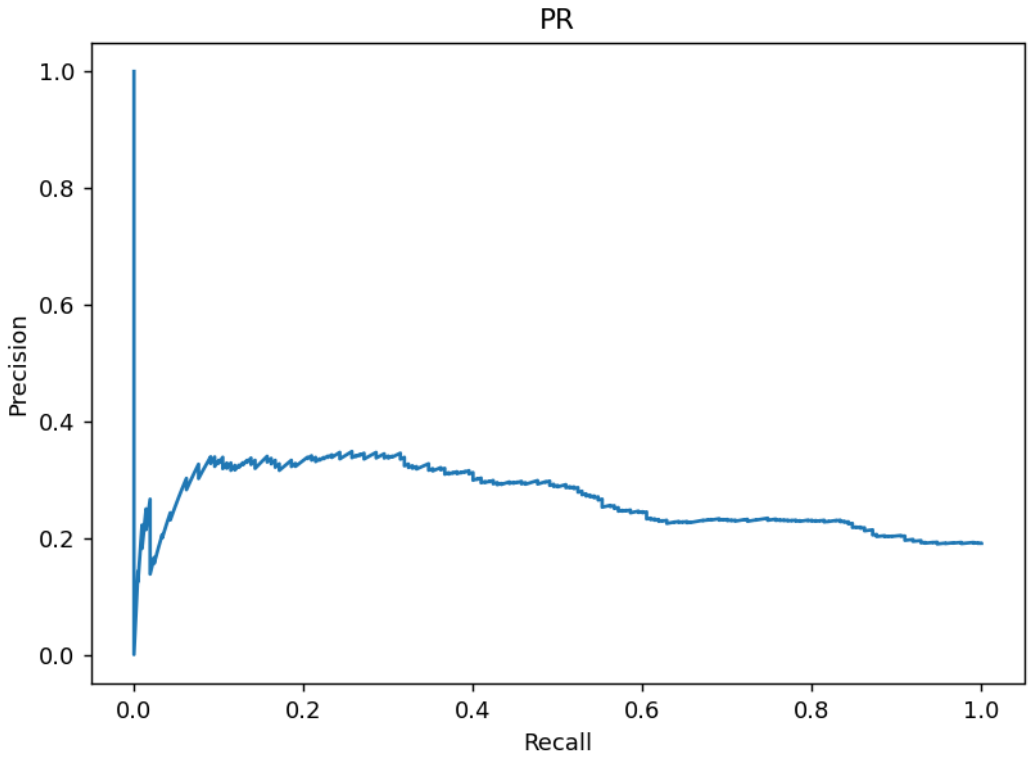
Матрица ошибок: диагональ — верные ответы.

Рисунок 11. clf_roc.png



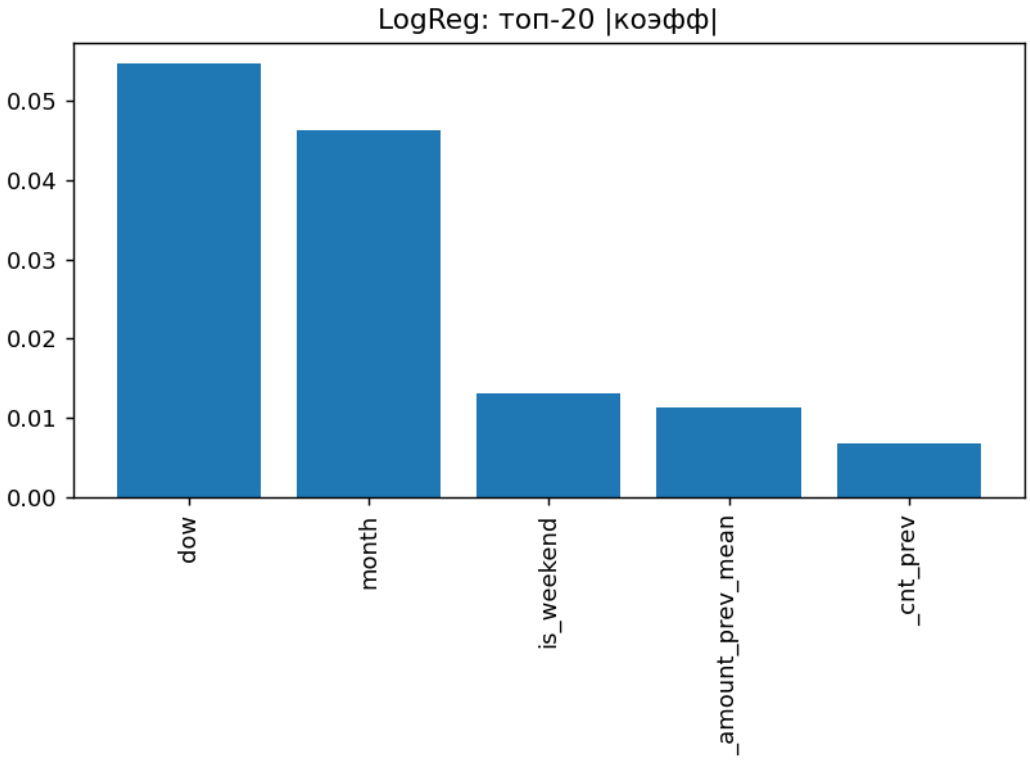
ROC-кривая; ROC-AUC — качество ранжирования.

Рисунок 12. clf_pr.png



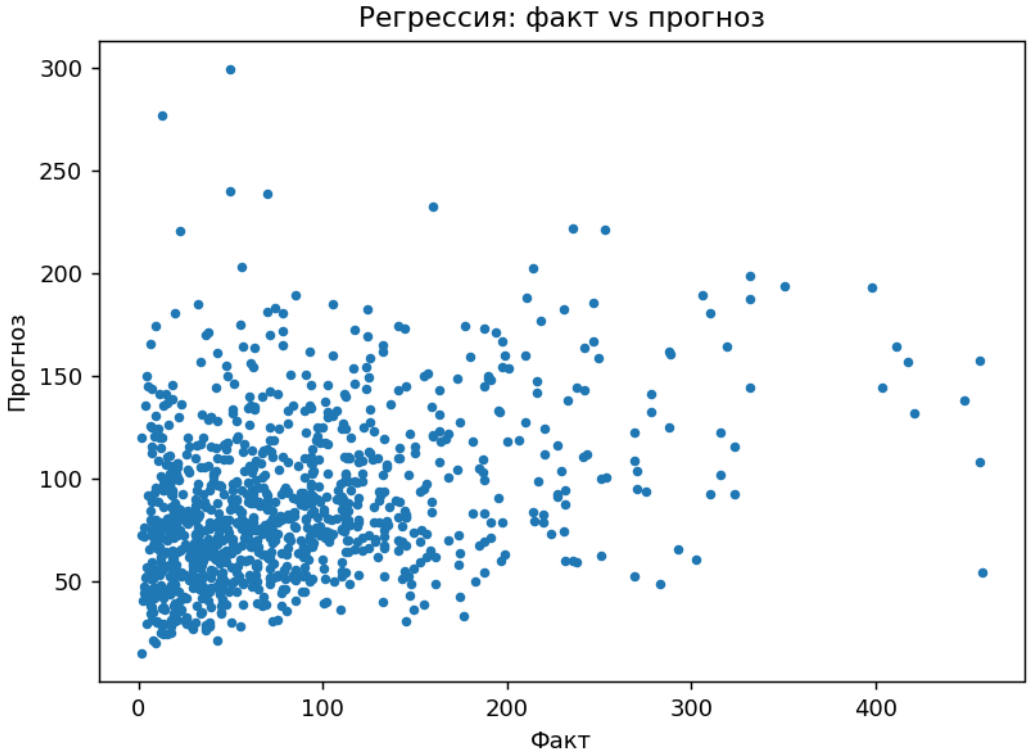
Precision–Recall кривая.

Рисунок 13. clf_top_coef.png



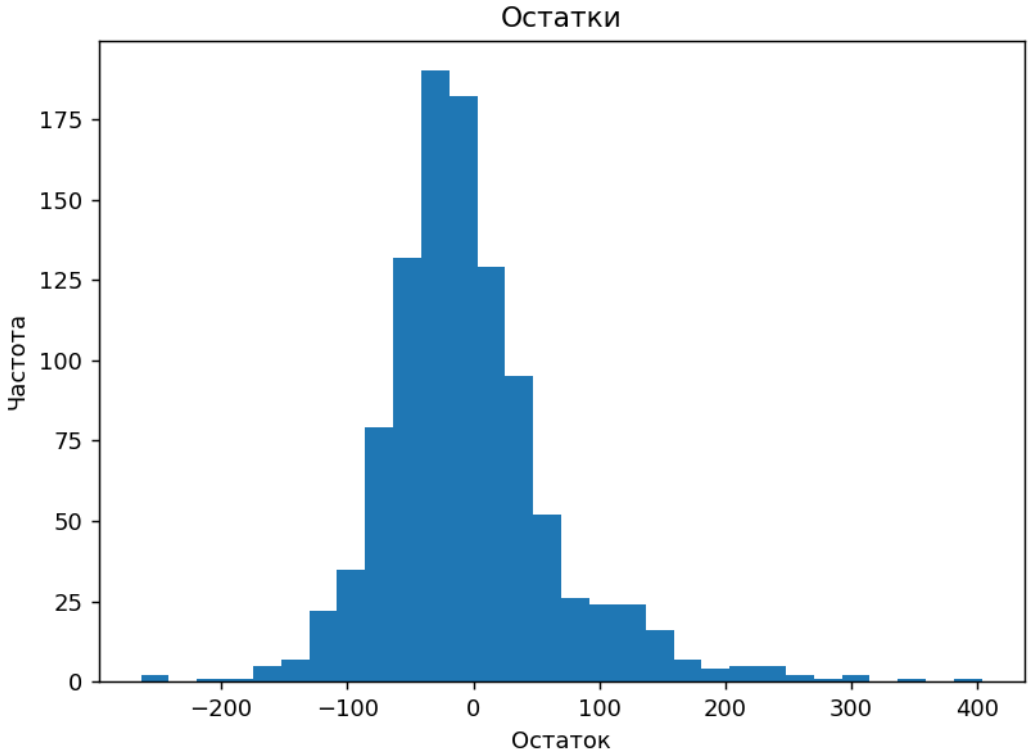
Наиболее влияющие признаки по |коэф|.

Рисунок 14. reg_scatter.png



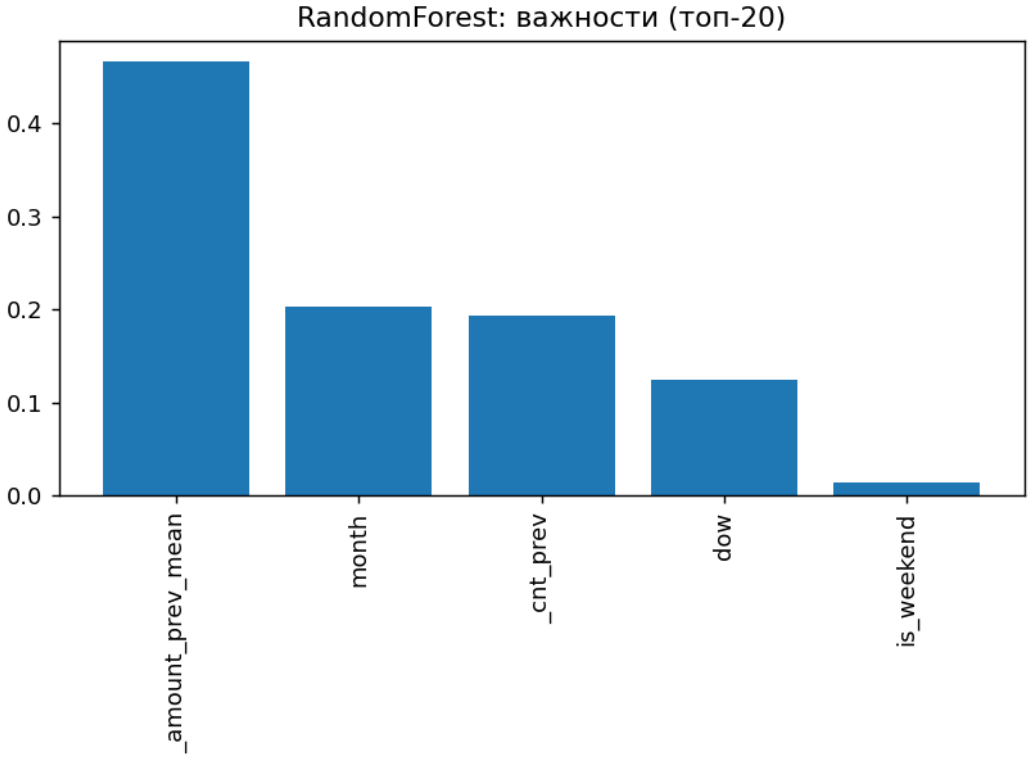
Скаттер факт–прогноз.

Рисунок 15. reg_residuals.png



Распределение остатков.

Рисунок 16. reg_feature_importance.png



Важности признаков RF.