# Отчёт по итоговому заданию к курсу

# «Аналитика для бизнеса»: Повышение точности подготовки Паспорта проекта

**Тема:** оценка предварительной экономической целесообразности покупки списанной автомобильной и спецтехники для последующей переработки (лом + запчасти). Ограничение – цена выкупа не всегда известна на момент оценки проекта (лота).

*Уточнение*: кейс реальный, но все данные по нему (в т.ч.: поставщики, тип, год выпуска, состояние, и т.д.) для служебного пользования, поэтому для итоговой работы был использован фейковый, сгенерированный, датасет.

**Датасет**: синтетический, n = 500 партий спецтехники.

Ссылки: [GitHub](https://github.com/kazacoder/mgimo_data_analisys), [Colab](https://colab.research.google.com/drive/1wj3_h16WAlo3LqeKXBDDUOEZdwbOpuEY?usp=sharing)

Ключевые цифры (по датасету)  
- Всего партий: 500.  
- Доля партий, убыточных даже без учёта цены покупки (break\_even < 0): 8.2%.  
- Средний break-even (максимально допустимая цена покупки, чтобы не уйти в убыток): 76 383 ₽.  
- Средняя ожидаемая выручка от переработки (scrap + parts): 207 332 ₽.  
- Средняя стоимость переработки: 130 949 ₽.  
- Для 40% партий (200) в датасете известна реальная предложенная продавцом цена; среди них 77.0% оказались потенциально прибыльны.  
- Средний относительный маржин (%) — 63.24%.  
  
1. Формулировка проблемы  
Проблема: заранее (до подтверждения цены продавцом) определить, какие предложения о покупке списанной автомобильной и спецтехники целесообразно оценивать дальше, а какие сразу отсечь как экономически невыгодные. На практике это нужно для экономии времени экспертов и ресурсов на осмотр и оценку только тех лотов, которые с высокой вероятностью окажутся прибыльными.  
  
2. Ключевые SMART-вопросы  
1. Какая доля приходящих лотов может быть надёжно отсеяна как явно убыточные до получения цены продавца (с порогом уверенности ≥ 95%)?  
2. Какие характеристики партии (type, age\_years, condition, weight\_kg, supplier) наилучшим образом предсказывают положительную маржу при покупке по цене, не превышающей break-even?  
3. Каково ожидаемое изменение экономического результата (средняя прибыль на тонну и суммарная прибыль), если внедрить автоматический предфильтр, снижающий количество проверяемых лотов экспертами на 40%?  
4. При каком уровне неопределённости по цене (разброс относительно break-even) проект становится слишком рискованным для подтверждения без дополнительного осмотра?  
  
3. Какие данные и из каких источников нужно собрать  
Обязательные данные (доступны до получения цены от продавца):  
- type, age\_years, weight\_kg, condition, supplier\_id, photos, typical\_content.  
Желательные данные (после согласия на оценку): лабораторные/инспекционные замеры, true\_offer\_price, результаты переработки.  
Внешние данные: рыночные цены на лом и запчасти, стоимость переработки.

Как собирать: сочетание ручного ввода (приём заявок), интеграция с CRM/1С (для поставщиков), стандартизированная форма загрузки фотографий, и справочники по типам техники (для типовой массы/состав).  
  
4. Очистка и подготовка данных  
1. Стандартизация форматов и единиц (приведение всех дат/чисел к единым типам, единицы — кг, ₽, годы.).  
2. Дедупликация и валидация (масса, возраст).

3. Нормализация наименований поставщиков и типов: использование справочников и нечеткий поиск (для устранения дубликатов по названиям поставщиков).  
4. Обработка пропусков: заполнение из справочников или маркировка для ручной проверки.  
5. Создание признаков:

* est\_metal\_kg — ориентировочная масса металла = weight\_kg \* metal\_fraction\_by\_type \* condition\_multiplier.
* est\_parts\_value — ориентировочная стоимость годных запчастей (по типу и condition).
* expected\_margin\_rough = est\_total\_revenue - processing\_cost (без учёта purchase\_price)

## **5. Агрегирующие показатели (KPIs)**

Estimated total revenue (R):

R = est\_scrap\_revenue + est\_parts\_value (в dataset — est\_total\_revenue).

Использование: базовая оценка верхней границы дохода от лота.

Processing cost (C):

C = f(weight\_kg, type, condition, proc\_mult) — в dataset processing\_cost.

Использование: оценка внутренних затрат.

Break-even purchase price (B):

B = R - C — максимальная цена покупки без убытка. (в датасете: break\_even\_purchase\_price)

Использование: ключевой фильтр — если продавец предлагает цену > B, проект убыточен.

Relative margin (M%):

M% = (R - C) / C \* 100% (dataset: relative\_margin\_pct).

Использование: ранжирование лотов по привлекательности (вне зависимости от абсолютных сумм).

Flag unprofitable without purchase (F):

F = 1 if B < 0 else 0 — случаи, когда даже при покупке за ₽0 проект убыточен. (dataset: flag\_unprofitable\_no\_purchase).

Probability of profit at offered price (P):

Для партий с известными true\_offer\_price: P = I[R - C - true\_offer\_price > 0]. Для неизвестных — можно моделировать P через модель классификации.

**Примеры интерпретации (гипотетические)**

* Если B ≤ 0: автоматически отклоняем лот (экономически нецелесообразно — 8.2% в нашем датасете).
* Если M% > 50% и B высок (> медианы), лот — приоритетный для оценки.
* Если M% низок (< 10%) и B близок к 0 — требовать дополнительного осмотра/доп. данных; высокая вероятность риска.

**Как агрегаты ответят на SMART-вопросы**

* Чтобы измерить «долю лотов, которые можно безопасно отсеять» — считать F + те с B < threshold\_small (например, < 10 000 ₽). Precision/recall можно оценить на валидационной выборке (где есть цены).
* Чтобы определить критерии для модели — использовать relative\_margin\_pct, type, age\_years, condition как предикторы и is\_profitable (при известной цене) как цель.

## **6. Визуализации**

6.1 Scatter: Processing cost vs Estimated total revenue (с линией break-even y=x)

Почему: мгновенно показывает, какие лоты оказываются ниже линии (R < C) — убыточные без учёта цены покупки, и какие находятся существенно выше.

Интерпретация: точки ниже линии — flag\_unprofitable\_no\_purchase. Точки чуть выше линии — низкая маржа; следует требовать осторожности.

6.2 Boxplot: Estimated total revenue по типам техники

Почему: позволяет увидеть распределение выручки по типам (Truck/Excavator/...). Рекомендуется для приоритизации типов и корректировки цены выкупа для разных типов.

Интерпретация: если медиана по типу Excavator выше, чем у Loader, — закупки Excavator более перспективны.

6.3 Гистограмма: распределение break-even purchase price

Почему: показывает плотность допустимых цен покупки; удобно выбирать пороги для «автоматической отсечки».

Интерпретация: если большая масса партий имеет B < 0 или B < 20k ₽, то имеет смысл отсеить их заранее.

6.4 Bar chart: % партий с B < 0 по типам техники

Почему: позволяет определить типы техники, которые чаще всего убыточны даже без покупки.

Интерпретация: типы с высокой долей убыточных — «красная зона» закупок (например, Forklift может быть чаще убыточен из-за малой массы и невысокой цены запчастей).

6.5 Heatmap корреляций (numeric)

Почему: проверка линейных связей между факторами (weight, age, cost, revenue). Помогает увидеть, какие признаки важны для предсказания B и margin.

Интерпретация: сильная корреляция est\_total\_revenue с est\_scrap\_revenue указывает, что основной драйвер прибыли — металлолом/запчасти.

Как ответы меняются в зависимости от результатов визуализаций:

Если scatter показывает большой кластер ниже y=x - стратегически нужно сократить приём определённых типов или усилить предварительную оценку.

Если boxplot показывает большие медианные различия между типами - можно вводить дифференцированные правила отбора по типу.

Если гистограмма B смещена влево (многие низкие/отрицательные значения) – текущий набор лотов низкодоходен, и метод закупок нужно пересмотреть.

7. Поиск взаимосвязей и методы

**Переменные**

Целевые: is\_profitable\_if\_offer (для тех лотов, где есть цена), break\_even\_purchase\_price, relative\_margin\_pct.

Предикторы: type, age\_years, condition, weight\_kg, supplier, est\_metal\_kg, est\_parts\_value.

**Как искать связи**

Корреляционный анализ: Pearson/Spearman между числовыми признаками (weight, age, est\_metal\_kg, est\_total\_revenue, processing\_cost, break\_even). Помогает быстро выделить сильные линейные связи.

— В датасете, как правило, est\_total\_revenue коррелирует сильно с est\_scrap\_revenue и est\_parts\_value. processing\_cost коррелирует с weight\_kg.

Групповые сравнения (ANOVA / Kruskal-Wallis): проверить, значимо ли различается break\_even и relative\_margin между типами техники и condition.

**Классификация/регрессия:**

Классификация (Binary): цель — is\_profitable\_if\_offer (при известных предложениях)

Регрессия: предсказать break\_even\_purchase\_price как непрерывную величину (R — C). Методы: Linear Regression, Random Forest Regressor, GB Regressor.

Кросс-валидация и валидация по поставщикам: использовать K-fold CV; рекомендовано также проводить validation by supplier (leave-one-supplier-out), чтобы оценить обобщаемость на новых продавцах.

**Как интерпретации могут меняться**

Если age\_years и condition оказываются слабо связанными с profit, это значит, что нужно полагаться больше на weight и тип (или на дополнительно извлекаемые признаки из фото).

Если supplier даёт сильный эффект, то стоит ввести supplier-level правила (репутация, контракты).

8. Управленческое действие

Действие: внедрить автоматический предварительный фильтр лотов который до получения цены продавца отсекает:

* Лоты с break\_even\_purchase\_price < 0 (автоматический отказ). На исследуемом датасете это ~8.2% партий — экономия ресурсов на ненужных осмотрах.
* Лоты с break\_even\_purchase\_price ниже низкого порога (например, < 10 000 ₽) — высокий риск, отклонять/автоматически предлагать альтернативные сценарии (перепродажа/разборка на месте).
* Для остальных — применять классификатор вероятности прибыльности (обученный на исторических данных):
  + Если P(profit) > 0.9 - «зелёный», можно переходить к формальному согласованию и быстрой проверке.
  + Если 0.6 < P ≤ 0.9 - «жёлтый», требуется быстрая доп.фотография / удалённый осмотр оператором.
  + Если P ≤ 0.6 - «красный», отклонение или полная ручная экспертиза.

**Ожидаемый эффект (примерная калькуляция по датасету):**

* Отклоняя автоматически 8.2% явно убыточных лотов, сокращаем нагрузку на экспертов и немедленно экономим время.
* При внедрении фильтра нацеленного на 40% сокращение ручных проверок (сортировка предложений), по conservative estimate, более 70% из проверенных «зелёных» лотов будут прибыльны (в датасете известно, что при имеющихся предложениях 77% прибыльны).
* Экономический эффект (пример): если компания осматривает 2000 лотов в год, сокращение ручных проверок на 40% сэкономит ~800 осмотров; при стоимости осмотра 3 000 ₽ это ~2.4 млн ₽ в год прямой экономии, плюс избегание покупки убыточных лотов (8.2% \* 2000 ≈ 164 лота, если бы их покупали — значительная потенциальная экономия).

Изображение выглядит как текст, снимок экрана, линия, диаграмма

Содержимое, созданное искусственным интеллектом, может быть неверным.  
Изображение выглядит как текст, диаграмма, линия, План

Содержимое, созданное искусственным интеллектом, может быть неверным. Изображение выглядит как текст, диаграмма, снимок экрана, График

Содержимое, созданное искусственным интеллектом, может быть неверным.



