**Тестовое задание**

**Вопросы**

Ниже представлен список вопросов, на которые следует развёрнуто ответить, уделяя особое внимание нюансам задач класса «Marketing mix modeling».

1. Опишите последовательность, суть и нюансы основных этапов проекта по эконометрическому моделированию влияния медиакампаний (offline и Интернет) на целевой показатель (например, продажи). Какие этапы, на Ваш взгляд, являются наиболее ресурсоёмкими, а какие -  наукоёмкими? На каких этапах необходимо активное взаимодействие с клиентом и в чём оно выражается?
2. Каким образом можно сравнивать несколько моделей регрессии? Какие критерии и подходы при этом следует применять?
3. Статистика R^2 линейной регрессионной модели составляет 95%. Что это значит? Можно ли утверждать, что данная модель будет хорошим инструментом прогнозирования зависимой переменной?

**Практическое задание**

В прилагаемом файле «data\_for\_test.xlsx» представлены (на листе «Raw data») временные ряды предикторов и целевого показателя (суммарное количество телефонных звонков возможных покупателей в дилерские центры, реализующие автомобиль определённой марки). Описание факторов дано на листе «Data description». Допускается дополнение данных новыми факторами (в т.ч. производными от имеющихся).

**Основная задача: построить регрессионную модель предсказания зависимой величины с помощью предоставленных факторов. Подзадачи:**

1. провести разведочный анализ данных (описательные статистики, распределения, сводные таблицы, графики), среди графиков желателен график динамики целевой величины с отметками периодов рекламной активности; содержательно описать результаты анализа;
2. построить регрессионную модель, выполнить её валидацию на отдельной подвыборке;
3. оценить качество и адекватность модели, выполнить необходимые тесты выполнения предпосылок используемой модели (если тесты не выполняются, то объяснить, почему, и как с этим бороться);
4. количественно оценить изменение степени влияния рекламной активности на ТВ;
5. построить график динамики целевого показателя, дополненный графиком оценок целевого показателя, полученных с помощью модели;
6. дать содержательную интерпретацию полученным результатам и итоговой модели, сформулировать рекомендации по совершенствованию модели и привести дополнительные факторы, которые можно было бы использовать для улучшения объясняющей силы и прогностической способности модели.

**Результаты работы следует представить в виде текстового отчёта, включающего:**

1. описание задачи и факторов;
2. описание результатов разведочного анализа;
3. описание полученной модели, оценок её качества и адекватности, выявленных проблем;
4. интерпретацию выявленных закономерностей;
5. необходимые таблицы и графики.

**Отдельно следует приложить воспроизводимый программный код на предпочитаемом языке программирования, использованный для решения задачи. Также можно выполнить всё задание в виде единого файла Jupyter Notebook / R Markdown.**

**Срок выполнения задания:** 1 неделя с момента получения задания.

Выполнение ТЗ: Ответы на вопросы

1. Опишите последовательность, суть и нюансы основных этапов проекта по эконометрическому моделированию влияния медиакампаний (offline и Интернет) на целевой показатель (например, продажи). Какие этапы, на Ваш взгляд, являются наиболее ресурсоёмкими, а какие – наукоёмкими? На каких этапах необходимо активное взаимодействие с клиентом и в чём оно выражается?

**1. Этапы проекта по эконометрическому моделированию влияния медиакампаний**

**1.1. Этапы и их содержание**

**(1) Постановка задачи и сбор требований**

* Определяется цель анализа: объяснение драйверов продаж, оптимизация бюджетов, прогноз эффекта от медиа и т. д.
* Выбирается целевой показатель (sales, leads, awareness).
* Уточняются временные рамки, каналы коммуникации (TV, digital, radio, OOH, PR, promotions).
* Формируется гипотеза о том, какие факторы влияют на целевую метрику.

На этом этапе особенно важно плотное взаимодействие с клиентом: бизнес знает контекст, агентство – методологию. Здесь согласуются цели, источники данных и критерии успеха.

**(2) Сбор и подготовка данных**

* Интеграция источников: медиаспенд, GRP, digital impressions, данные CRM, макроэкономика, погода, промоактивности, цены, конкуренты.
* Гармонизация временных рядов: агрегация по неделям, устранение выбросов, проверка лагов.
* Кодирование категориальных переменных, создание индикаторов для событий.

Это самый **ресурсоёмкий** этап. Он требует много ручной проверки, выравнивания по календарям, устранения дубликатов и ошибок.

**(3) Обогащение данных и трансформации**

* Применяются **adstock-функции** (задержанный эффект рекламы) и **saturation-функции** (эффект убывающей отдачи).
* Могут добавляться лаги, логарифмирование, нормализация.
* Иногда строятся сезонные индексы.

Этап – **наукоёмкий**, требует понимания поведения аудитории и особенностей каналов.

**(4) Построение модели**

* Основной инструмент – **многомерная регрессия** (обычно OLS, иногда Bayesian или Ridge/Lasso для борьбы с мультиколлинеарностью).
* Проверяются гипотезы о значимости переменных, стабильности коэффициентов, мультиколлинеарности, автокорреляции остатков.
* Итерационный процесс: разные наборы переменных, формы функций adstock, периодичность лагов.

Этап – **наукоёмкий**, здесь концентрируется эконометрическая экспертиза.

**(5) Интерпретация и валидация**

* Рассчитываются **эластичности**, **ROI по каналам**, доля объяснённой вариации продаж.
* Сравниваются результаты по регионам или периодам.
* Проводится **out-of-sample validation** (проверка прогноза на новых данных).

Здесь важно совместное обсуждение с клиентом – интерпретация чисел требует бизнес-знания контекста.

**(6) Рекомендации и сценарное моделирование**

* Симулируются сценарии: «что если увеличить digital на 10%», «что если снизить ТВ».
* Разрабатывается оптимизационная модель (часто на базе solver’а или Python-пакетов вроде scipy.optimize).
* Клиент получает **инструмент для медиапланирования** и прогнозирования.

#### ****1.2. Ресурсоёмкие и наукоёмкие этапы****

| **Этап** | **Характеристика** | **Почему** |
| --- | --- | --- |
| Сбор и подготовка данных | Ресурсоёмкий | Много источников, несовпадения, ручная работа |
| Моделирование и трансформации | Наукоёмкий | Требует эконометрического мышления, статистических тестов, гипотез |
| Интерпретация и рекомендации | Совместный, бизнес-критичный | Нужно перевести математику в бизнес-решения |

#### ****1.3. Взаимодействие с клиентом****

* На старте – формулировка задач, уточнение данных.
* В середине – обсуждение предварительных моделей и логики влияния факторов.
* В конце – согласование интерпретаций и практических рекомендаций.

Ошибка – делать модель «в вакууме»: клиент должен подтверждать гипотезы, иначе модель потеряет доверие.

1. Каким образом можно сравнивать несколько моделей регрессии? Какие критерии и подходы при этом следует применять?

При сравнении нескольких моделей (например, разных наборов переменных или разных типов функций) применяют комбинацию **статистических и бизнес-критериев**:

#### ****2.1. Статистические критерии****

* R² (коэффициент детерминации)– **объяснённая доля вариации.**
* Adjusted R² **– корректирует R² на число предикторов.**
* AIC / BIC **– баланс качества и сложности модели (чем меньше, тем лучше).**
* RMSE / MAPE / MAE **– ошибки прогноза.**
* Out-of-sample performance – **тестирование на отложенной выборке.**

#### ****2.2. Эконометрические критерии****

* **Проверка** значимости коэффициентов **(p-values).**
* **Отсутствие** мультиколлинеарности **(VIF).**
* **Стационарность остатков (тест Дики-Фуллера).**
* **Нормальность и отсутствие автокорреляции.**

#### ****2.3. Бизнес-критерии****

* Реалистичность эластичностей (напр., ROI > 1000% — подозрительно).
* Согласованность с рыночной логикой (digital не должен быть отрицательным, если это не объясняется акциями конкурентов).
* Простота интерпретации: модель должна быть понятна заказчику.

Идеальная модель – это компромисс между статистическим качеством, простотой и бизнес-достоверностью.

1. Статистика R^2 линейной регрессионной модели составляет 95%. Что это значит? Можно ли утверждать, что данная модель будет хорошим инструментом прогнозирования зависимой переменной?

### ****Интерпретация R² = 95%****

**Что это значит:**

R² = 0.95 говорит, что **95% вариации зависимой переменной** объясняется включёнными факторами. На первый взгляд – высокая объясняющая способность.

**Стоит уточнить несколько важных моментов при разборе данной модели :**

* **Высокий R² не гарантирует хорошего прогноза.** Модель может быть **переобученной** или содержать **мультиколлинеарные переменные**, из-за чего коэффициенты нестабильны.
* В MMM (моделирование комплекса маркетинга) часто встречаются псевдо-корреляции (например, рост digital и рост продаж просто совпали по тренду).
* Если данные агрегированы по времени, то R² может быть искусственно высоким из-за тренда и сезонности.

**Что нужно проверить:**

* Остатки: случайны ли они (white noise)?
* Прогноз на новых данных — насколько точен?
* Насколько устойчива модель к удалению части выборки (cross-validation)?

**Вывод:**

R² = 95% — признак хорошего описания данных, **но не доказательство прогностической силы**. Для прогнозов нужно подтверждение на независимой выборке и логическая устойчивость модели.

Практическое задание

В прилагаемом файле «data\_for\_test.xlsx» представлены (на листе «Raw data») временные ряды предикторов и целевого показателя (суммарное количество телефонных звонков возможных покупателей в дилерские центры, реализующие автомобиль определённой марки). Описание факторов дано на листе «Data description». Допускается дополнение данных новыми факторами (в т.ч. производными от имеющихся). Основная задача: построить регрессионную модель предсказания зависимой величины с помощью предоставленных факторов. Подзадачи:

* + 1. провести разведочный анализ данных (описательные статистики, распределения, сводные таблицы, графики), среди графиков желателен график динамики целевой величины с отметками периодов рекламной активности; содержательно описать результаты анализа;
    2. построить регрессионную модель, выполнить её валидацию на отдельной подвыборке;
    3. оценить качество и адекватность модели, выполнить необходимые тесты выполнения предпосылок используемой модели (если тесты не выполняются, то объяснить, почему, и как с этим бороться);
    4. количественно оценить изменение степени влияния рекламной активности на ТВ;
    5. построить график динамики целевого показателя, дополненный графиком оценок целевого показателя, полученных с помощью модели;
    6. дать содержательную интерпретацию полученным результатам и итоговой модели, сформулировать рекомендации по совершенствованию модели и привести дополнительные факторы, которые можно было бы использовать для улучшения объясняющей силы и прогностической способности модели.

Результаты работы следует представить в виде текстового отчёта, включающего:

1. описание задачи и факторов;
2. описание результатов разведочного анализа;
3. описание полученной модели, оценок её качества и адекватности, выявленных проблем;
4. интерпретацию выявленных закономерностей;
5. необходимые таблицы и графики.
6. Отдельно следует приложить воспроизводимый программный код на предпочитаемом языке программирования, использованный для решения задачи. Также можно выполнить всё задание в виде единого файла Jupyter Notebook / R Markdown.

Краткий итог (что сделано)

Отчёт по заданию: Эконометрическое моделирование влияния медиаактивности на количество звонков (Marketing Mix Modeling)

* + - 1. **Описание задачи и факторов**

**Цель проекта:**  
 Построить и интерпретировать регрессионную модель зависимости количества звонков в дилерские центры (Y) от факторов маркетинговой активности (в частности – ТВ-рекламы), а также макроэкономических и сезонных переменных.

**Задачи:**

1. Провести разведочный анализ данных (EDA).
2. Построить модель, оценивающую влияние рекламных факторов.
3. Проверить её адекватность и устойчивость.
4. Качественно и количественно интерпретировать влияние ТВ.
5. Подготовить рекомендации по улучшению модели и данных.

**Описание факторов:**

| **Переменная** | **Описание** | **Тип** |
| --- | --- | --- |
| Y | Количество звонков в дилерские центры (целевой показатель) | Целевая |
| TV | Индекс или объем рекламной активности на ТВ | Медиафактор |
| OOH | Активность наружной рекламы | Медиафактор |
| Seasonal\_Sales | Сезонный индекс продаж | Контрольный |
| Usd\_rate | Курс доллара (макроэкономический фактор) | Контрольный |
| Date | Дата наблюдения | Временной |
| trend | Временной тренд, добавленный при анализе | Производный |

Для описания длительного эффекта рекламы применено **adstock-преобразование** (с коэффициентом затухания 0.7 для ТВ и 0.5 для OOH), отражающее эффект накопления и последующего убывания воздействия рекламных кампаний.

* + - 1. **Результаты разведочного анализа данных**

**Структура и временной диапазон данных:**

* Период наблюдений: январь 2013 – февраль 2015.
* Данные ежедневные, преобразованы в недельную агрегацию (109 недель), чтобы сгладить шум и лучше отразить длительные эффекты рекламы.

**Основные наблюдения:**

* Целевая переменная Y демонстрирует выраженную волатильность на дневном уровне и чёткие сезонные колебания на недельном.
* ТВ-активность (TV) имеет ярко выраженные периоды всплесков – весна и осень 2014 года, декабрь 2014.
* Корреляционный анализ (по агрегированным данным):
  1. Y и TV: +0.27
  2. Y и OOH: ≈ 0
  3. Y и Usd\_rate: +0.31

Также отмечены визуальные совпадения пиков звонков с периодами ТВ-активности.

**Вывод EDA:**

ТВ может оказывать краткосрочный положительный эффект на количество звонков, в то время как влияние наружной рекламы малозаметно. Валютный курс и сезонность также оказывают значимое воздействие.

* + - 1. **Описание модели, оценка качества и адекватности**

**Метод:**  
Линейная регрессия (OLS) для зависимой переменной ln(1 + Y) с предикторами:

* TV\_ad – adstock-преобразованная ТВ-активность (decay = 0.7)
* OOH\_ad – adstock-преобразованная OOH (decay = 0.5)
* Seasonal\_Sales, Usd\_rate, trend

**Разделение выборки:**

* Обучающая часть – 80% наблюдений.
* Тестовая часть – последние 20%.

**Качество модели:**

| **Метрика** | **Train** | **Test** |
| --- | --- | --- |
| R² | 0.546 | — |
| RMSE | — | 2043.4 |
| MAE | — | 1174.1 |

**Выдержка из результатов OLS:**

| **Переменная** | **Коэффициент** | **p-value** | **Интерпретация** |
| --- | --- | --- | --- |
| TV\_ad | +3.75e−06 | <0.001 | Положительное влияние ТВ |
| OOH\_ad | −2.81e−06 | 0.04 | Слабое отрицательное влияние |
| Usd\_rate | +0.0178 | <0.001 | Рост курса → больше звонков |
| trend | −0.0107 | <0.001 | Убывающий тренд |
| const | 1.319 | — | Базовый уровень |

**Диагностика:**

* **Durbin–Watson ≈ 1.82** → слабая положительная автокорреляция остатков.
* **VIF** < 2 для всех факторов → мультиколлинеарности нет.
* **Ljung–Box p > 0.05** → остатки не автокоррелированы.
* **ADF-тест остатков** → стационарность подтверждается.
* **Нормальность** остатков нарушена (возможен эффект выбросов).

**Вывод:**  
Модель адекватна по предпосылкам линейной регрессии, но умеренно объясняет вариацию целевой переменной. Прогноз на тестовой выборке ухудшается из-за резкого роста Usd\_rate и экстремальных значений TV.

* + - 1. **Интерпретация закономерностей**
* ТВ оказывает **статистически значимое положительное влияние**:  
  средняя **эластичность ≈ 0.077**, то есть рост ТВ-активности на 10% приводит к увеличению количества звонков примерно на 0.8%.
* Наружная реклама – незначимый и слегка отрицательный фактор, возможно из-за отсутствия данных о географическом охвате или временных лагов.
* Рост курса доллара коррелирует с увеличением активности звонков (возможен эффект «спешки перед ростом цен»).
* Негативный тренд указывает на постепенное снижение интереса к продукту или усиление конкуренции.
  + - 1. **Таблицы и графики**

### ****5.1. Динамика звонков и периодов ТВ-активности****

График временного ряда Y (синие линии) с красными точками, обозначающими недели с высокой ТВ-активностью. Видна синхронизация пиков звонков и ТВ.

### ****5.2. Наблюдаемые и прогнозные значения (Weekly)****

На графике факт и прогноз модели:

* В обучающем периоде модель описывает общие тренды.
* В тестовом периоде завышенные прогнозы – результат экстраполяции высоких значений Usd\_rate.

### ****5.3. Остатки модели****

Диаграмма остатков показывает отсутствие систематических сдвигов, но некоторая асимметрия присутствует.

* + - 1. **Рекомендации по совершенствованию модели**

1. **Данные:**
   * Добавить показатели digital-активности (поисковые запросы, клики, показы).
   * Учесть акции, скидки, стоимость автомобилей.
   * Добавить данные конкурентов (их медиаактивность).
2. **Моделирование:**
   * Использовать лог-лог форму для прямой оценки эластичностей.
   * Добавить функции насыщения (Hill, логистическую).
   * Применить Ridge/Lasso для регуляризации.
   * Разделить периоды до/после кризиса 2014 г.
3. **Валидация:**
   * Протестировать кросс-валидацию по временным блокам.
   * Рассмотреть ARIMAX-модель с временной структурой остатков.

**Код программы находится так же в папке под названием Test\_Task**