

УДК 004.415.25

5.2.2. Математические, статистические и инструментальные методы экономики (физико-математические науки, экономические науки)

АНАЛИЗ ПРИКЛАДНОЙ ЦЕННОСТИ ДАННЫХ ПРИ ЧАСОВОМ ПРОГНОЗИРОВАНИИ СПРОСА (НА ПРИМЕРЕ УСЛУГИ ВЕЛОПРОКАТА)

Параскевов Александр Владимирович
старший преподаватель
SPIN-код: 2792-3483
e-mail paraskevov.a@kubsau.ru
ФГБОУ ВО Кубанский государственный аграрный университет имени И. Т. Трубилина, г. Краснодар, РФ

Сергеев Александр Эдуардович
доцент, кандидат физико-математических наук
e-mail galua1979@yandex.ru
ФГБОУ ВО Кубанский государственный аграрный университет имени И. Т. Трубилина, г. Краснодар, РФ

Цель исследования состоит в количественной оценке вклада календарных, погодных и операционных данных в точность часового прогноза спроса. Эксперимент выполнен на наборе Bike Sharing из репозитория UCI: 17 379 почасовых наблюдений системы Capital Bikeshare за 2011–2012 гг. Для всех вариантов применен ансамбль предельно рандомизированных деревьев с одинаковыми параметрами. Проверка проведена на четырех последовательных кварталах 2012 г.; обучающая выборка расширялась перед каждым кварталом. Календарная модель получила MAE 65,95 аренды в час. Добавление погоды снизило MAE до 50,62, или на 23,25%. Включение часовой истории спроса уменьшило MAE полного варианта до 30,79. Снижение относительно модели с календарем и погодой составило 39,17%, относительно недельного ориентира — 49,34%. После включения лагов погода уменьшила MAE на 0,81%; различие дневных ошибок статистически не подтверждено ($p = 0,147$)

Ключевые слова: АНАЛИЗ ДАННЫХ, ПРОГНОЗИРОВАНИЕ СПРОСА, СБОР ДАННЫХ, ВРЕМЕННОЙ РЯД, EXTRA TREES, ВЕЛОПРОКАТ, ХРОНОЛОГИЧЕСКАЯ ПРОВЕРКА

<http://dx.doi.org/10.21515/1990-4665-221-035>

UDC 004.415.25

5.2.2. Mathematical, statistical and instrumental methods of economics (physical and mathematical sciences, economic sciences)

ANALYSIS OF THE APPLICABLE VALUE OF DATA IN HOURLY DEMAND FORECASTING (USING THE EXAMPLE OF A BIKE-RENTAL SERVICE)

Paraskevov Alexander Vladimirovich
senior lecturer
RSCI SPIN-code: 2792-3483
e-mail paraskevov.a@kubsau.ru
Federal State Budgetary Educational Institution of Higher Education «Kuban State Agrarian University named after I.T. Trubilin», Krasnodar, Russia

Sergeev Alexander Eduardovich
Associate Professor, Candidate of Physics-mathematical Sciences
e-mail alvdmv@mail.ru
Federal State Budgetary Educational Institution of Higher Education «Kuban State Agrarian University named after I.T. Trubilin», Krasnodar, Russia

The goal of the study is to quantitatively assess the contribution of calendar, weather, and operational data to the accuracy of hourly demand forecasting. The experiment was conducted on the Bike Sharing dataset from the UCI repository: 17,379 hourly observations of the Capital Bikeshare system for 2011–2012. For all variants, an ensemble of extremely randomized trees with identical parameters was used. The validation was carried out on four consecutive quarters of 2012; the training set was expanded before each quarter. The calendar model achieved an MAE of 65.95 per hour. Adding weather reduced the MAE to 50.62, or by 23.25%. Including the hourly demand history reduced the MAE of the full variant to 30.79. The reduction relative to the calendar and weather model was 39.17%, and relative to the weekly benchmark — 49.34%. After the inclusion of lags, the weather reduced MAE by 0.81%; the difference in daily errors was not statistically significant ($p = 0.147$)

Keywords: DATA ANALYSIS, DEMAND FORECASTING, DATA COLLECTION, TIME SERIES, EXTRA TREES, BIKE RENTAL, CHRONOLOGICAL VERIFICATION

Введение

Организация сбора данных требует затрат на датчики, интеграцию источников, хранение и контроль качества. Точность прогноза сама по себе не определяет состав информационной системы. Итоговая ошибка отвечает на вопрос о качестве конкретной модели, но не указывает, какое поле обеспечило снижение ошибки и какое поле можно исключить. Для проектирования контура данных требуется эксперимент, в котором алгоритм и порядок проверки остаются постоянными, а состав входной информации меняется по группам.

Системы совместного использования велосипедов подходят для такого эксперимента по трем причинам. Каждая аренда имеет временную отметку, спрос меняется в течение суток, а внешняя среда включает календарь и погоду. Capital Bikeshare публикует историю поездок и описывает состав открытых записей: время начала и окончания, станции, длительность и тип пользователя [1]. Набор UCI Bike Sharing объединяет реальные журналы этой системы с погодными и календарными полями [2]. Его структура позволяет проверить ценность внутренних и внешних источников в одной задаче.

Исследование посвящено часовому прогнозу общего числа аренд. Объектом служит процесс формирования спроса на услуги городского велопроката. Предмет исследования — изменение ошибки прогноза при последовательном включении календарных, погодных и операционных признаков. Цель состоит в определении минимального состава данных, который дает измеримое повышение точности на горизонте один час.

Для достижения цели решены четыре задачи: сформированы группы признаков по источнику сбора; исключены поля с прямой утечкой целевой переменной; выполнена хронологическая проверка на четырех кварталах; рассчитаны абсолютные, относительные и статистические оценки прироста. Изложение задач используется как описание структуры

эксперимента, поскольку каждая операция связана с отдельным проверяемым результатом.

Рабочая гипотеза содержит два положения. Непрерывный журнал спроса должен давать больший вклад в часовой прогноз, чем внешние погодные данные, поскольку лаги отражают фактическое состояние сервиса непосредственно перед прогнозом. Погодные признаки должны снижать ошибку модели, основанной только на календаре, поскольку температура, влажность и осадки связаны с решением пользователя о поездке. Оба положения проверяются на одинаковом алгоритме.

Научная новизна состоит в оценке данных как ресурса информационной системы. Варианты сравниваются по происхождению признаков, а итог формулируется как порядок сбора. Такой подход отделяет вклад алгоритма от вклада информационного наполнения. Практический результат представлен перечнем обязательных полей, периодом хранения оперативной истории и границами использования погодного источника.

Исследования спроса и постановка эксперимента

Fanaee-T и Gama использовали журнал Capital Bikeshare для разметки событий и сопоставили прогнозные методы при наличии фоновой информации [3]. В работе подтверждено происхождение наблюдений; ряд пригоден для регрессии и обнаружения аномалий. Corcoran и соавторы исследовали пространственно-временные закономерности велопроката с учетом погоды и календарных событий [4]. Kim установил различия погодных эффектов между группами станций и часами суток [5]. Эти результаты обосновывают раздельное представление времени и погоды.

Cantelmo, Kucharski и Antoniou объединили историю поездок с температурой и осадками; проверка на данных Citi Bike включала свыше

17 млн поездок [6]. Pazdan, Kiec и D'Agostino сопоставили автоматические счетчики, поездки велопроката и погодные условия [7]. Ма и соавторы включили историю поездок, погоду, сведения о пользователях и землепользование в краткосрочный пространственно-временной прогноз [8]. Многоисточность повышает информационную емкость входа, однако стоимость подключения каждого источника остается отдельным управленческим вопросом.

Исследования 2024 г. расширили интерпретацию факторов спроса. Ngeni и соавторы применили объяснимое машинное обучение к станционным данным [9]. Cortez-Ordoñez, Vázquez и Sanchez-Espigares рассмотрели влияние погоды, сезонов, праздников и часовых профилей [10]. Работы подтверждают неоднородность влияния признаков. Для информационной системы требуется дополнительная оценка: насколько меняется ошибка после подключения целой группы полей при неизменном алгоритме и временной схеме проверки.

Sathishkumar, Park и Cho сравнили методы интеллектуального анализа для прогноза спроса в мегаполисе [11]. В отдельной работе Sathishkumar и Cho применили наборы Seoul Bike и Capital Bikeshare с погодными полями [12]. Основной результат таких сравнений связан с выбором регрессора. В настоящем исследовании алгоритм фиксирован. Переменной экспериментального плана служит доступность группы данных.

Схема эксперимента приведена на рисунке 1. Журнал преобразуется в три группы признаков. Затем формируются пять вариантов входа: недельный ориентир, календарь, календарь с погодой, календарь с историей и полный набор. Каждый вариант оценивается на одних тестовых часах. Такое построение исключает преимущество, возникающее из-за различного периода проверки.



Рисунок 1 – Схема оценки вклада групп данных

Схема экспериментального плана автора.

Данные и методы

Состав и подготовка набора

Набор Bike Sharing размещен в UCI Machine Learning Repository с DOI 10.24432/C5W894 [2]. Почасовой файл содержит 17 379 записей с 1 января 2011 г. по 31 декабря 2012 г. Полная календарная сетка периода включает 17,544 часов; отсутствие 165 временных отметок соответствует 0,94%. В записанных строках пропуски полей отсутствуют. Пропущенный час и нулевой спрос имеют разный смысл: минимальное зарегистрированное значение `snt` равно одной аренде. При подготовке создана полная часовая шкала, поэтому разрывы сохранились как пропуски.

Целевая переменная `snt` равна сумме `casual` и `registered`. Оба слагаемых исключены из входа, поскольку их использование дает целевой показатель до момента прогноза. Индекс `instant` также исключен. Дата и час преобразованы в циклические координаты. Нормированные погодные поля оставлены в шкале набора: `temp` делится на 41, `atemp` — на 50, `hum` —

на 100, windspeed — на 67 [2]. Такое решение сохраняет исходную воспроизводимую форму наблюдений.

Группировка признаков основана на способе получения. Календарные поля вычисляются из временной отметки и справочника праздников. Погодные поля поступают из внешнего источника. Операционные признаки рассчитываются из ранее зарегистрированного cnt. Таблица 1 связывает каждую группу с минимальным контуром сбора.

Таблица 1 – Группы признаков и источники формирования

Группа	Поля эксперимента	Требование к сбору
Календарная	час, день недели, месяц, сезон, праздник, рабочий день, циклические координаты, тренд	временная отметка, часовой пояс, справочник праздников
Погодная	класс погоды, температура, ощущаемая температура, влажность, скорость ветра	почасовая погода для территории обслуживания
Операционная	лаги 1, 24 и 168 часов; среднее и стандартное отклонение за 24 и 168 часов	счетчик аренд, непрерывная история минимум за семь суток

Формализация прогноза и признаков

Горизонт равен одному часу. В момент расчета доступны календарь прогнозируемого часа, погодные условия из текущего наблюдения или часового прогноза и фактический спрос до часа $t - 1$. Общая постановка имеет вид:

$$\hat{y}_t = f(X_t^{(c)}, X_t^{(w)}, X_t^{(o)}) \quad (1)$$

где $\hat{y}_t$ — прогноз числа аренд в час t ; f — регрессионная функция; $X^{(c)}t$ — календарный вектор; $X^{(w)}t$ — погодный вектор; $X^{(o)}t$ —

операционный вектор. Верхние индексы s , w и o обозначают источник признаков. Ни один компонент $X(o)t$ не содержит y_t .

Скользящее среднее операционной истории рассчитывается по формуле:

$$\bar{y}_{t,L} = \frac{1}{L} \sum_{i=1}^L y_{t-i}, \quad L \in \{24, 168\} \quad (2)$$

где $\bar{y}_{t,L}$ — среднее число аренд перед часом t ; L — длина окна; i — смещение в часах; y_{t-i} — ранее зарегистрированный спрос. Окно 24 отражает внутрисуточный уровень, окно 168 охватывает семь суток. Стандартное отклонение для тех же окон характеризует изменчивость. При разрывах использованы минимальные количества наблюдений 18 и 120 соответственно; отсутствующие лаги сохранены как пропуски, поддерживаемые реализацией алгоритма.

Алгоритм и хронологическая проверка

Для всех вариантов выбран алгоритм Extremely Randomized Trees, предложенный Geurts, Ernst и Wehenkel [13]. Ансамбль строит регрессионные деревья со случайным выбором порогов. Среднее прогнозов снижает зависимость результата от отдельного дерева. Единый алгоритм нужен для сопоставления информационных наборов: изменение ошибки связано с доступными полями и их взаимодействиями при заданной модели.

Ансамбль содержит 400 деревьев, максимальная глубина равна 24, минимальный лист включает одно наблюдение, доля рассматриваемых признаков равна 1,0. Параметры глубины, размера листа и доли признаков выбраны по периоду октябрь–декабрь 2011 г. из сетки {без ограничения; 24}, {1; 2; 4} и {0,7; 0,9; 1,0}. После выбора в тестовую процедуру вошел только 2012 г. Случайное состояние зафиксировано значением 42. Расчеты выполнены в Python с библиотекой scikit-learn [16].

$$\hat{y}_t = \frac{1}{M} \sum_{m=1}^M T_m(x_t) \quad (3)$$

где $M = 400$ — число деревьев; T_m — прогноз m -го дерева; x_t — вектор признаков часа t ; $\hat{y}_t$ — среднее ансамбля. Отрицательные прогнозы обнулялись, поскольку число аренд принимает значения от нуля.

Случайное перемешивание временного ряда не применялось. Bergmeir, Hyndman и Коо рассматривают ограничения стандартной перекрестной проверки для авторегрессионного прогноза [14]. В эксперименте использованы четыре последовательных тестовых блока: кварталы 2012 г. Перед каждым блоком обучение охватывало все доступные предшествующие наблюдения. В тесте каждого квартала лаги строились только по уже зарегистрированному спросу. Общий объем проверки составил 8 734 часа.

Таблица 2 – Варианты информационного обеспечения

Код	Доступные данные	Назначение сравнения
V0	спрос за тот же час предыдущей недели; при пропуске — лаг 24 часа	сезонный ориентир без обучения
V1	календарная группа	результат расчетной календарной схемы
V2	календарная и погодная группы	вклад внешнего погодного источника
V3	календарная и операционная группы	вклад журнала спроса без погоды
V4	все три группы	полный часовой прогноз

Метрики и статистическая проверка

Средняя абсолютная ошибка выражена в арендах за час:

$$MAE = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t| \quad (4)$$

где n — число тестовых часов; y_t — фактический спрос; $\hat{y}_t$ — прогноз; t — номер тестового наблюдения. MAE допускает прямое прикладное чтение: значение 30 означает среднее отклонение на 30 аренд в час.

Среднеквадратическая ошибка имеет вид:

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2} \quad (5)$$

где обозначения y , $\hat{y}$, t и n совпадают с формулой (4). Возведение отклонения в квадрат увеличивает вклад редких крупных ошибок.

Симметричная средняя абсолютная процентная ошибка рассчитана по формуле:

$$sMAPE = \frac{100\%}{n} \sum_{t=1}^n \frac{2|y_t - \hat{y}_t|}{|y_t| + |\hat{y}_t|} \quad (6)$$

где знаменатель содержит сумму модулей факта и прогноза. При одновременном нулевом значении слагаемое принято равным нулю. Метрика включена для относительного сопоставления часов с разным уровнем спроса. Выбор MAE, RMSE и sMAPE согласуется с задачей оценки точечных прогнозов; свойства метрик подробно рассмотрены Hyndman и Koehler [15].

Относительный прирост при переходе от варианта А к варианту В вычислен как:

$$G_{A \rightarrow B} = \frac{MAE_A - MAE_B}{MAE_A} \cdot 100\% \quad (7)$$

где MAE_A и MAE_B — средние абсолютные ошибки сравниваемых вариантов; $G_{A \rightarrow B}$ — снижение ошибки в процентах. Положительное

значение соответствует меньшей МАЕ варианта В. Для статистической проверки часовые абсолютные ошибки агрегированы по дням. Односторонний критерий знаковых рангов Вилкоксона проверяет уменьшение дневной МАЕ. Доверительный интервал разности построен бутстрэпом по 366 дням с 10 000 повторов.

Результаты

Структура часового спроса

Среднее число аренд за час равно 189,46, медиана — 142, стандартное отклонение — 181,39, максимум — 977. Среднее 2011 г. составляет 143,79, 2012 г. — 234,67. Разница между годами требует расширяющегося обучения: модель, обученная только на первом году, получает систематическое смещение в начале 2012 г. Причины роста нельзя установить по агрегированному набору, поскольку в нем отсутствуют изменения числа станций, тарифов и доступного парка.

Рабочие дни имеют три выраженных часа спроса. Среднее в 08:00 равно 477,0 аренды, в 17:00 — 525,3, в 18:00 — 492,2. Для выходных и праздников максимум расположен в дневном интервале: 372,7 аренды в 13:00. Различие профилей представлено на рисунке 2. Суточное среднее без признака рабочего дня скрывает периоды, для которых требуется запас пропускной способности.

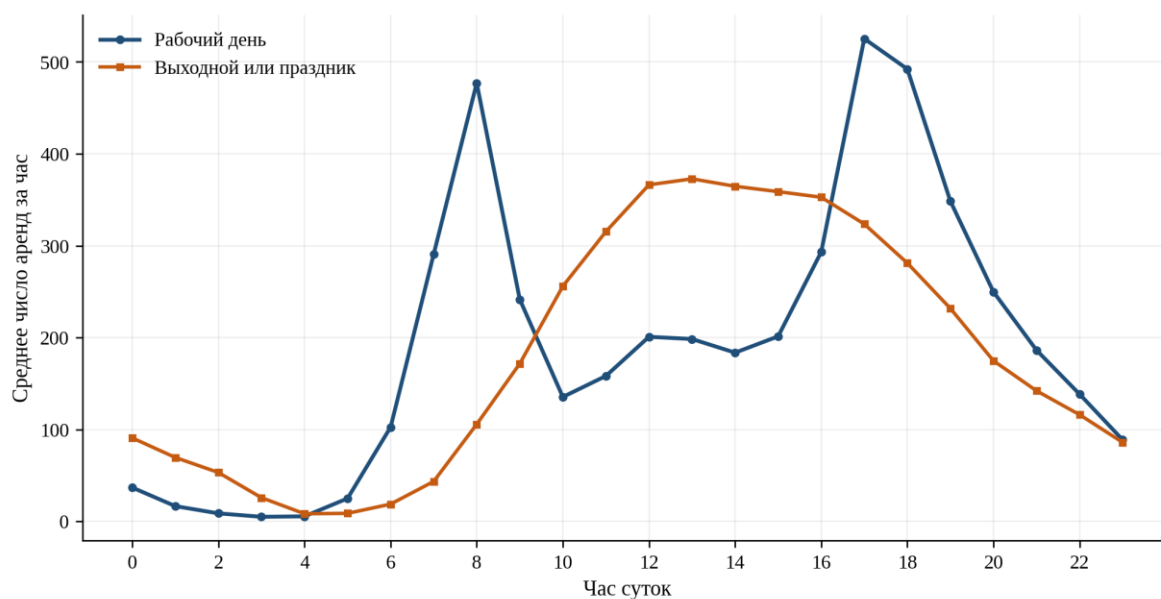


Рисунок 2 – Средний часовой профиль спроса по типу дня

Средний спрос при классе 1, включающем ясную погоду и малую облачность, равен 204,87 аренды. Для класса 2 с туманом или облачностью значение составляет 175,17, для класса 3 с дождем или снегом — 111,58. В классе 4 содержатся три часа, поэтому его среднее не используется для содержательного сравнения. Часовые профили трех представленных классов приведены на рисунке 3. Разница имеет описательный характер: класс погоды связан с сезоном, температурой и временем суток, поэтому средние нельзя трактовать как причинный эффект осадков.

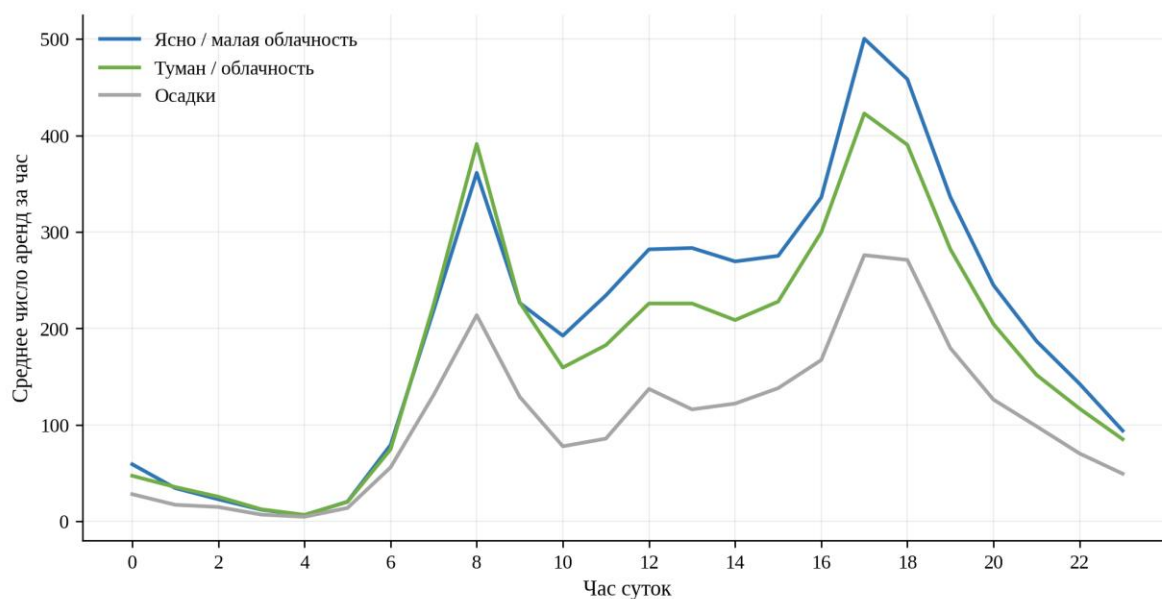


Рисунок 3 – Средний часовой спрос при разных погодных классах

Точность вариантов и вклад данных

Таблица 3 – Итоговые ошибки на 8 734 тестовых часах

Вариант	MAE	RMSE	sMAPE, %
V0. Недельный ориентир	60,77	104,33	35,18
V1. Календарь	65,95	103,70	36,57
V2. Календарь + погода	50,62	80,73	31,03
V3. Календарь + история	31,04	51,51	19,85
V4. Полный набор	30,79	50,46	19,65

Календарная модель V1 получила MAE 65,95. Ее ошибка выше недельного ориентира V0 с MAE 60,77. В первом квартале календарная модель достигает MAE 81,77. Причина связана с переходом к более высокому уровню спроса 2012 г.; деревья не экстраполируют уровень за пределы обучающих значений. Этот результат ограничивает распространенное утверждение о пользе аналитического алгоритма независимо от состава входа.

Погодная группа снижает MAE V1 с 65,95 до 50,62. Относительный прирост равен 23,25%. Средняя дневная разность ошибок составляет 15,56 аренды; 95%-й бутстрэп-интервал равен [13,07; 18,16], $p < 0,001$. Подключение внешнего источника оправдано для системы, в которой отсутствует оперативная история.

Операционные признаки меняют результат сильнее. Полный вариант V4 имеет MAE 30,79, RMSE 50,46 и sMAPE 19,65%. Снижение MAE относительно V2 равно 39,17%; средняя дневная разность равна 19,77 при 95%-м интервале [17,22; 22,35], $p < 0,001$. Относительно недельного ориентира MAE уменьшилась на 49,34%.

Вариант V3 без погоды получил MAE 31,04. Добавление погоды к календарю и истории снизило показатель на 0,81%. Интервал дневной разности [-0,14; 0,67] включает ноль, $p = 0,147$. Для горизонта один час недавняя история спроса содержит большую часть информации, связанной с текущей погодой. Этот вывод не переносится автоматически на прогноз за сутки, поскольку в таком режиме лаг $t - 1$ еще неизвестен.

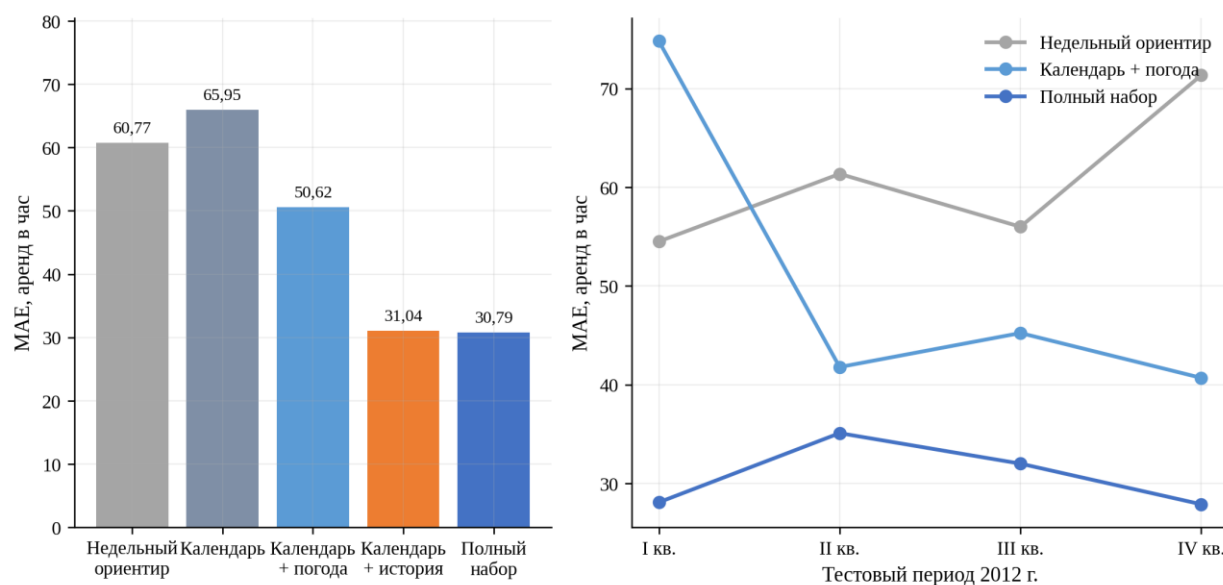


Рисунок 4 – MAE вариантов и устойчивость ошибки по кварталам

Таблица 4 – Квартальная MAE выбранных вариантов

Период	V0	V2	V4
I кв.	54,51	74,79	28,10
II кв.	61,35	41,80	35,10
III кв.	56,00	45,24	32,02
IV кв.	71,34	40,70	27,90

MAE полного варианта находится в диапазоне от 27,90 до 35,10 аренды в час. Недельный ориентир ухудшается до 71,34 в четвертом квартале. Полный вариант в том же периоде имеет MAE 27,90. Расширение обучающей выборки и оперативные лаги компенсируют изменение уровня ряда лучше фиксированной недельной копии.

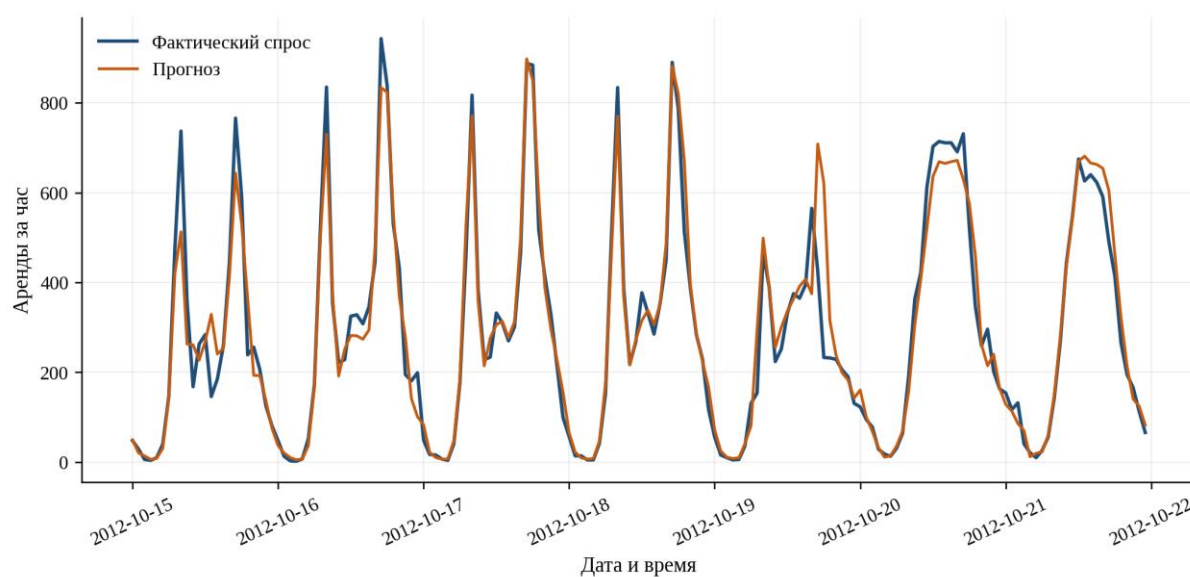


Рисунок 5 – Фактический и прогнозный спрос 15–21 октября 2012 г.

Групповая перестановка и ошибки при высоком спросе

Для проверки роли групп в полной модели все поля одной группы совместно перемешивались внутри тестового квартала. Средний прирост MAE равен 39,76 для календарной группы, 3,40 для погодной и 157,32 для операционной. Совместная перестановка сохраняет отношения между полями группы, но нарушает их соответствие целевому часу. Показатель отражает зависимость уже обученной полной модели и не является причинной оценкой.

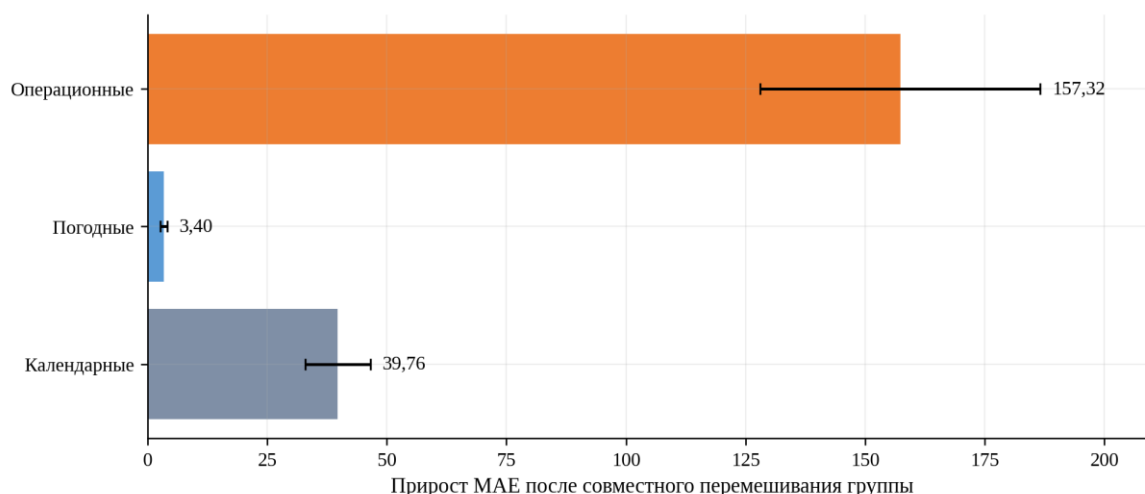


Рисунок 6 – Групповая перестановочная важность в полной модели

Порог верхних 10% фактического спроса равен 546 арендам в час. В этих 877 наблюдениях MAE полного варианта составляет 83,79, для остальных часов — 24,88. Абсолютная ошибка возрастает в часы пик. Для операционного применения одного среднего показателя недостаточно: контроль должен отдельно учитывать интервалы с высоким спросом, поскольку именно они формируют потребность в резерве велосипедов и обслуживающего персонала.

Практический порядок сбора и применения данных

Первый уровень информационной системы включает временную отметку, часовой пояс, число начатых аренд за час и признак полноты записи. Временная отметка требуется для календарных преобразований; счетчик формирует целевую переменную и лаги. Нулевая аренда записывается числом 0. Отсутствующая телеметрия хранится как пропуск с кодом причины. Такое разделение предотвращает искусственное снижение спроса при сбое канала.

Операционная история должна охватывать минимум 168 часов до момента прогноза. Этот срок обеспечивает лаг за неделю и недельные статистики. Хранение только суточного окна исключает сравнение с тем же часом предыдущей недели. Для повторного обучения нужен архив за

более длинный период; в эксперименте расширяющаяся выборка включала от одного года до 21 месяца. Минимальные семь суток относятся к онлайн-расчету признаков, а не к обучающему архиву.

Второй уровень включает рабочий день, праздник и погодные поля. Справочник праздников должен быть версионизируемым, поскольку перенос выходного меняет часовой профиль. Погодная запись связывается с часом по территории обслуживания и содержит источник, время получения и статус факта или прогноза. Для часового горизонта погода дает большой прирост при отсутствии истории и малый прирост после включения лагов. Поэтому ее интеграция следует за настройкой непрерывного журнала спроса.

Расчет выполняется каждый час после закрытия предыдущего интервала. На вход подаются календарь следующего часа, доступный погодный прогноз и лаги спроса. Выход содержит точечный прогноз, дату версии модели и признаки качества входа. При пропуске лага система использует остальные доступные поля и повышает флаг риска. Отдельный журнал хранит факт, прогноз и абсолютную ошибку; эти три поля позволяют вычислять дневную MAE без повторного обращения к обучающей выборке.

Контроль целесообразно разделить на обычные часы и верхний дециль спроса. В представленном наборе граница равна 546 арендам в час. Средняя ошибка верхнего дециля выше общей. Для планирования агрегированного резерва оператор использует прогноз часа и исторический квантиль ошибки. Расчет стационарного перераспределения по этому набору невозможен: часовой файл не содержит начальную станцию, конечную станцию, остаток велосипедов и свободные доки.

Таблица 5 – Минимальная схема данных для внедрения часового прогноза

Поле	Периодичность	Использование	Контроль
timestamp, timezone	каждый час	календарные признаки и соединение источников	уникальность часа, единый часовой пояс
rentals_count	каждый час	цель, лаги 1/24/168, окна 24/168	0 отделен от NULL; диапазон неотрицательный
workingday, holiday	по календарю	различение рабочих и нерабочих профилей	версия производственного календаря
weather_code, temp, atemp, hum, wind	каждый час	внешние условия прогнозируемого часа	источник, время получения, факт/прогноз
forecast, model_version	каждый запуск	контроль качества и воспроизводимость	связь с фактическим cnt

Ограничения исследования

Набор относится к Вашингтону 2011–2012 гг. Численные пороги спроса нельзя переносить на другую систему без повторного обучения. Методический результат допускает проверку на новом журнале: группы полей, горизонт и порядок хронологической оценки сохраняются, а метрики пересчитываются по локальным данным.

Агрегация выполнена по всей системе. Ошибка на уровне города не измеряет дефицит отдельной станции. Для решения задачи перераспределения нужны поездки с координатами или идентификаторами станций, остатки парка, вместимость доков и время перемещения обслуживающего транспорта. Эти поля входят в будущую станционную модель и расширяют стоимость сбора.

Погодные значения в UCI относятся к фактическим условиям часа. В производственном прогнозе используется прогноз погоды, содержащий собственную ошибку. Полученная оценка 0,81% после включения лагов может измениться при другом горизонте, погодном источнике или пространственной детализации.

Критерий Вилкоксона применен к дневным MAE, что снижает зависимость соседних часовых ошибок. Дни все еще образуют временную последовательность. Бутстрэп по дням оценивает разброс при выбранной схеме и не заменяет проверку на другом городе или году. Групповая

перестановка описывает зависимость полной модели от входа; причинные эффекты требуют отдельного дизайна.

Экономический эффект в денежных единицах не рассчитан. В наборе отсутствуют стоимость дефицита, цена перемещения велосипеда, тарифный доход и данные об отказах пользователей. Статья подтверждает сокращение прогнозной ошибки и определяет приоритет сбора. Перевод ошибки в затраты требует функции потерь оператора.

Заключение

Эксперимент на 8 734 тестовых часах количественно подтвердил полезность сбора данных для прогноза спроса. Календарь без оперативной истории получил MAE 65,95. Погодные поля снизили показатель до 50,62. Полный вход с лагами и скользящими статистиками достиг MAE 30,79, RMSE 50,46 и sMAPE 19,65%. Снижение MAE относительно недельного ориентира равно 49,34%.

Главный прикладной результат связан с последовательностью внедрения. Непрерывный часовой счетчик и история за 168 часов имеют первый приоритет. Погодная интеграция имеет второй приоритет для прогноза на один час при наличии лагов. Ее самостоятельная ценность сохраняется в календарной модели: MAE уменьшается на 23,25%. Различие между вариантом с историей и полным набором составляет 0,81% и не подтверждено на уровне дневных ошибок.

Результат задает проверяемое требование к информационной системе: сохранять каждый ожидаемый час, различать ноль и пропуск, хранить версию модели и сопоставлять прогноз с фактом. Для перехода от общесистемного планирования к стационарному перераспределению следует добавить идентификаторы станций, остатки и вместимость. Новая детализация должна пройти тот же групповой эксперимент, чтобы стоимость сбора связывалась с измеримым снижением ошибки.

Список литературы

1. Capital Bikeshare. System Data [Электронный ресурс]. URL: <https://capitalbikeshare.com/system-data> (дата обращения: 25.08.2026).
2. Fanaee-T H. Bike Sharing [Dataset]. UCI Machine Learning Repository, 2013. DOI: 10.24432/C5W894.
Fanaee-T H., Gama J. Event labeling combining ensemble detectors and background knowledge // Progress in Artificial Intelligence. 2014. Vol. 2. P. 113–127. DOI: 10.1007/s13748-013-0040-3.
3. Corcoran J., Li T., Rohde D., Charles-Edwards E., Mateo-Babiano D. Spatio-temporal patterns of a Public Bicycle Sharing Program: the effect of weather and calendar events // Journal of Transport Geography. 2014. Vol. 41. P. 292–305. DOI: 10.1016/j.jtrangeo.2014.09.003.
4. Kim K. Investigation on the effects of weather and calendar events on bike-sharing according to the trip patterns of bike rentals of stations // Journal of Transport Geography. 2018. Vol. 66. P. 309–320. DOI: 10.1016/j.jtrangeo.2018.01.001.
5. Cantelmo G., Kucharski R., Antoniou C. Low-Dimensional Model for Bike-Sharing Demand Forecasting that Explicitly Accounts for Weather Data // Transportation Research Record. 2020. Vol. 2674, no. 8. P. 132–144. DOI: 10.1177/0361198120932160.
6. Pazdan S., Kiec M., D’Agostino C. Impact of environment on bicycle travel demand—Assessment using bikeshare system data // Sustainable Cities and Society. 2021. Vol. 67. Art. 102724. DOI: 10.1016/j.scs.2021.102724.