

ПРОГНОЗИРОВАНИЕ ПРОДАЖ НА ОСНОВЕ ДИАГНОСТИЧЕСКИХ ДЕРЕВЬЕВ С РАЗБИЕНИЕМ РЯДА ПО ФОРМУЛЕ СТЕРДЖЕССА

Предлагается подход к прогнозированию продаж на основе решающих деревьев с формированием признаков путем разбиения ряда по формуле Стерджесса, расчета характеристик интервалов и отбора признаков с использованием различающей матрицы.

Введение

Точность прогнозирования спроса является ключевым фактором управления запасами. Традиционные методы (ARIMA) не учитывают нелинейность, сезонность и выбросы. Алгоритмы решающих деревьев (Random Forest, XGBoost) эффективны, но требуют обоснованного отбора признаков. Цель – разработка методики формирования признаков на основе разбиения ряда по формуле Стерджесса и различающей матрицы.

I. Методика формирования признаковового пространства

Исходные данные – открытый датасет компании «1С» (ежедневные продажи за 34 месяца, 60 магазинов, 22170 товаров) [1]. Методика включает агрегацию до помесечных данных, разбиение ряда по формуле Стерджесса на 6 интервалов ($k = 1 + \log_2(34)$), расчет для каждого интервала среднего, стандартного отклонения, выбросов (3σ) и прогноза AR(1), формирование 50 признаков (лаги lag_1, lag_2, lag_3, lag_6, lag_12 и сезонные признаки), построение различающей матрицы с дискретизацией на 5 классов и выделение редукта жадным алгоритмом, а также обучение моделей Decision Tree, Random Forest и XGBoost с временной кросс-валидацией.

II. Результаты экспериментального исследования

Применение различающей матрицы и жадного алгоритма позволило выделить ядро критически важных признаков, входящих во все редукты: lag_1 (продажи за предыдущий месяц), interval_5_mean (среднее значение за последний интервал), interval_5_autoregress_pred (авторегрессионный прогноз), is_december (признак декабря). Минимальный редукт составил 8 признаков: lag_1, interval_5_mean, is_december, interval_5_autoregress_pred, lag_12, interval_5_std, interval_4_std, lag_6.

Сравнительный анализ качества моделей для трех вариантов признакового пространства представлен в таблице 1.

Таблица 1 – Сравнение качества моделей

Модель	Признаки	RMSE	MAE	R^2
DT	50	1.24	0.58	0.63
DT	15	1.18	0.55	0.66
DT	8	1.21	0.56	0.64
RF	50	0.98	0.42	0.78
RF	15	0.95	0.40	0.79
RF	8	0.97	0.41	0.78
XGB	50	0.96	0.39	0.80
XGB	15	0.93	0.38	0.81
XGB	8	0.92	0.37	0.82

Наилучшие результаты показывает XGBoost с редуктом из 8 признаков ($RMSE = 0.92$, $MAE = 0.37$, $R^2 = 0.82$). Сокращение признаков с 50 до 8 не снизило точность, а для XGBoost даже улучшило ее за счет устранения шумовых признаков.

III. Выводы

В результате проведенного исследования разработана методика прогнозирования продаж на основе диагностических деревьев с разбиением ряда по формуле Стерджесса и отбором признаков с использованием различающей матрицы. Эксперимент на датасете «1С» подтвердил эффективность: количество признаков сокращено с 50 до 8 при сохранении точности ($RMSE = 0.92$ для XGBoost). Практическая значимость работы заключается в возможности интеграции разработанного программного модуля в учетные системы на платформе «1С» для автоматизации прогнозирования спроса в торговых организациях Республики Беларусь.

Список литературы

1. Kaggle Inc. Predict Future Sales: Competition Dataset [Электронный ресурс] / Kaggle Inc. – Режим доступа: <https://www.kaggle.com/competitions/competitive-data-science-predict-future-sales/data>. – Дата доступа: 11.03.2026.
2. scikit-learn Developers. Machine Learning in Python: Decision Trees and Ensemble Methods [Электронный ресурс] / scikit-learn Developers. – 2025. – Режим доступа: <https://scikit-learn.org/stable/modules/tree.html>. – Дата доступа: 28.02.2026.

Куприянчик Анна Александровна, магистрант кафедры информационных технологий автоматизированных систем БГУИР, hannakuprianchik@gmail.com.