Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ
КАФЕДРА СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

Обработка датасета

Продажи в вендинговых автоматах

Студент ИУ5-65Б _____ А. А. Ларин
(группа) (подпись, дата) (И.О. Фамилия)

Руководитель НИР _____ Ю. Е. Гапанюк
(подпись, дата) (И.О. Фамилия)

2025 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ

Заведующий кафедрой ИУ5
(индекс)
В.И. Терехов
(И.О. Фамилия)
(подпись)
(дата)

З А Д А Н И Е
на выполнение научно-исследовательской работы

по теме Обработка датасета продаж в вендинговых аппаратах

Студент группы ИУ5-65Б
Ларин Алексей Андреевич

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)
ИССЛЕДОВАТЕЛЬСКАЯ

Источник тематики (кафедра, предприятие, НИР) КАФЕДРА

График выполнения НИР:
25% к _____ нед., 50% к _____ нед., 75% к _____ нед., 100% к _____ нед

Техническое задание: Выбрать датасет для задачи классификации или регрессии. Провести анализ данных, выбрать признаки и построить не менее пяти моделей машинного обучения, включая две ансамблевые. Настроить гиперпараметры моделей, оценить их качество с использованием не менее трех метрик.

Оформление научно-исследовательской работы:

Расчетно-пояснительная записка на _14_ листах формата А4.

Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

Дата выдачи задания «_07_» _февраля___ 2025 г.

Руководитель НИР _____ Ю. Е. Гапанюк
(подпись, дата) (И.О. Фамилия)

Студент _____ А.А. Ларин
(подпись, дата) (И.О. Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В настоящем исследовании рассматривается задача прогнозирования выбора категории товара при покупке через вендинговые аппараты, а именно – будет ли куплен товар из категории "Food" , или же будет выбрана другая категория. Эта задача представляет собой интересную проблему бинарной классификации , позволяющей выявить ключевые факторы, влияющие на поведение покупателей в точках продажи без участия персонала.

Актуальность исследования обусловлена стремлением компаний, эксплуатирующих вендинговые машины, к более эффективному управлению ассортиментом, оптимизации загрузки товаров и повышению рентабельности оборудования. Понимание закономерностей выбора продуктов позволяет улучшать маркетинговые стратегии, учитывать локальные особенности и адаптировать предложения под предпочтения целевой аудитории.

Применение методов машинного обучения в данном контексте открывает новые возможности для автоматизации анализа данных, прогнозирования спроса и принятия обоснованных решений в сфере управления торговыми точками.

Целью работы является построение, обучение и сравнительный анализ нескольких моделей машинного обучения, включая ансамблевые методы, для решения задачи бинарной классификации: куплен ли товар из категории "Food" или.

В ходе исследования проводится предварительная обработка данных, подбор и сравнение различных моделей, настройка их гиперпараметров, оценка качества с использованием подходящих метрик, а также формирование выводов о применимости и эффективности рассмотренных подходов.

ОСНОВНАЯ ЧАСТЬ

Постановка задачи

Имеется открытый набор данных о продажах товарах в вендиговых аппаратах, содержащий информацию о купленных товарах и локации. Необходимо спрогнозировать купили товар из категории food или из другой.

Для решения задачи требуется:

1. Выполнить разведочный анализ данных;
2. Обработать пропуски и закодировать категориальные переменные;
3. Провести масштабирование признаков;
4. Сформировать обучающую и тестовую выборки;
5. Построить не менее пяти моделей (включая две ансамблевые);
6. Оценить их качество по нескольким метрикам (точность, полнота, F1-мера);
7. Настроить гиперпараметры моделей;
8. Сравнить результаты и обосновать выбор финальной модели.

Загрузка и просмотр данных

Этот датасет содержит 9350 записей данных о покупках со следующими признаками:

1. Status — статус транзакции.
2. Device ID — уникальный идентификатор устройства.
3. Location — местоположение вендинговой машины.
4. Machine — тип или модель вендинговой машины.
5. Product — наименование проданного товара.
6. Category — категория товара (например, Food, Water, Carbonated и т.д.).
7. Transaction — уникальный номер транзакции.
8. TransDate — дата совершения транзакции.
9. Type — тип транзакции.
10. RCoil — идентификатор катушки товара.
11. RPrice — цена товара за единицу.
12. RQty — количество товара в транзакции.
13. MCoil — второй идентификатор катушки.
14. MPrice — вторая цена (возможно, промо-цена).
15. MQty — второе количество товара (например, при акции).
16. LineTotal — сумма по строке транзакции (RPrice × RQty).
17. TransTotal — общая сумма всей транзакции.
18. Prcd Date — дата обработки транзакции системой.

Цель – изучить распределение данных, взаимосвязи между признаками и построить модель прогнозирования итоговой оценки.

Разведочный анализ данных

Рассмотрим распределение числовых признаков:

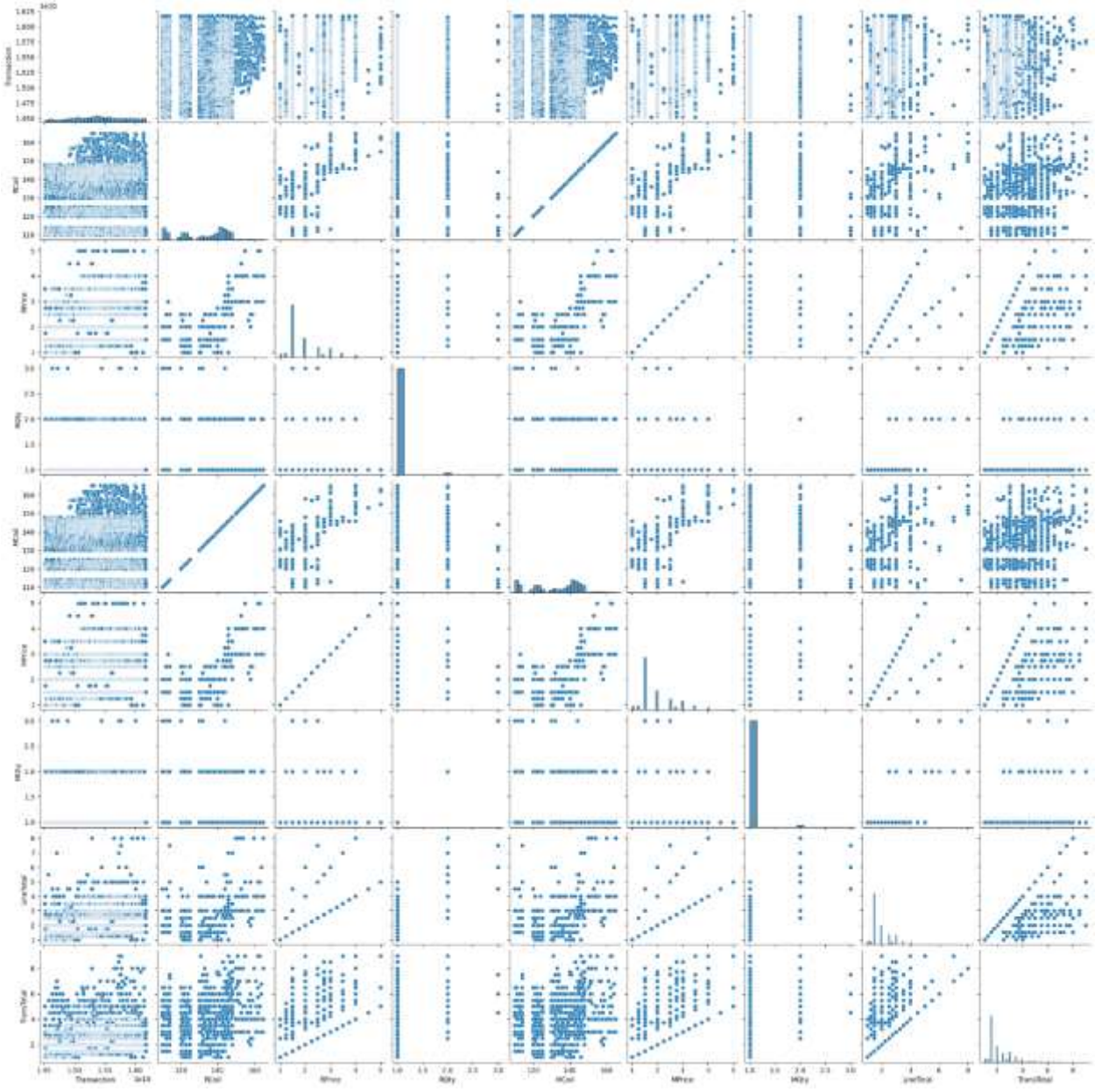

Рисунок 1 – Распределение числовых признаков

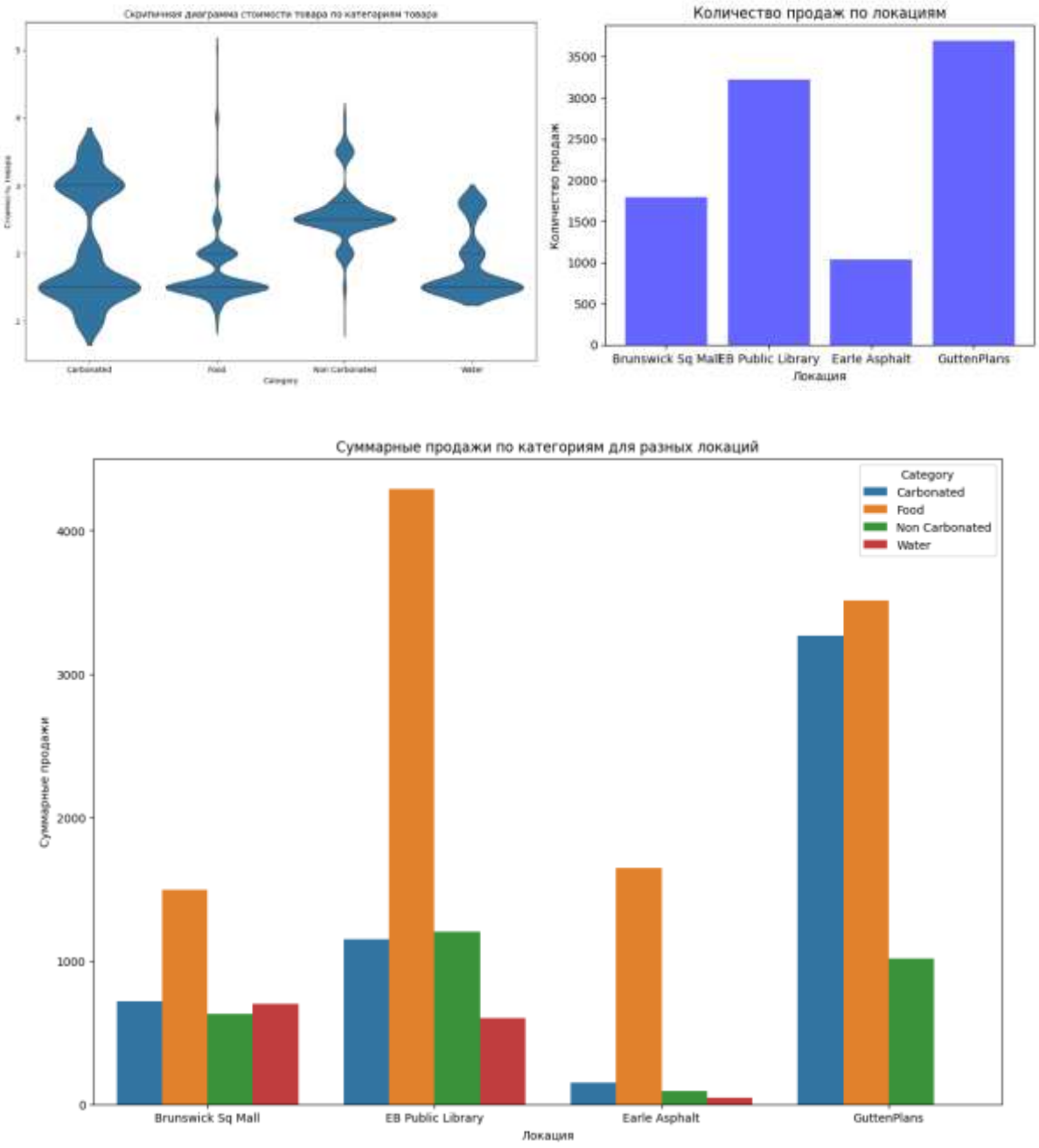

Рис 2. – Анализ категорий и локаций

В большинстве локаций больше всего доминирует еда, однако по сумме продаж остальные примерно равны категории “food”.

Категория “food”в большинстве своём стоит 1 и 2 денежные ед, в отличии от других категорий где существенную часть занимают более дорогие товары – это может являться одной из причин большего количества покупок чем во всех других категориях.

В будние дни покупают больше чем в выходные, а по будним дням количество покупок равномерное.

Рассмотрим корреляционную матрицу признаков:

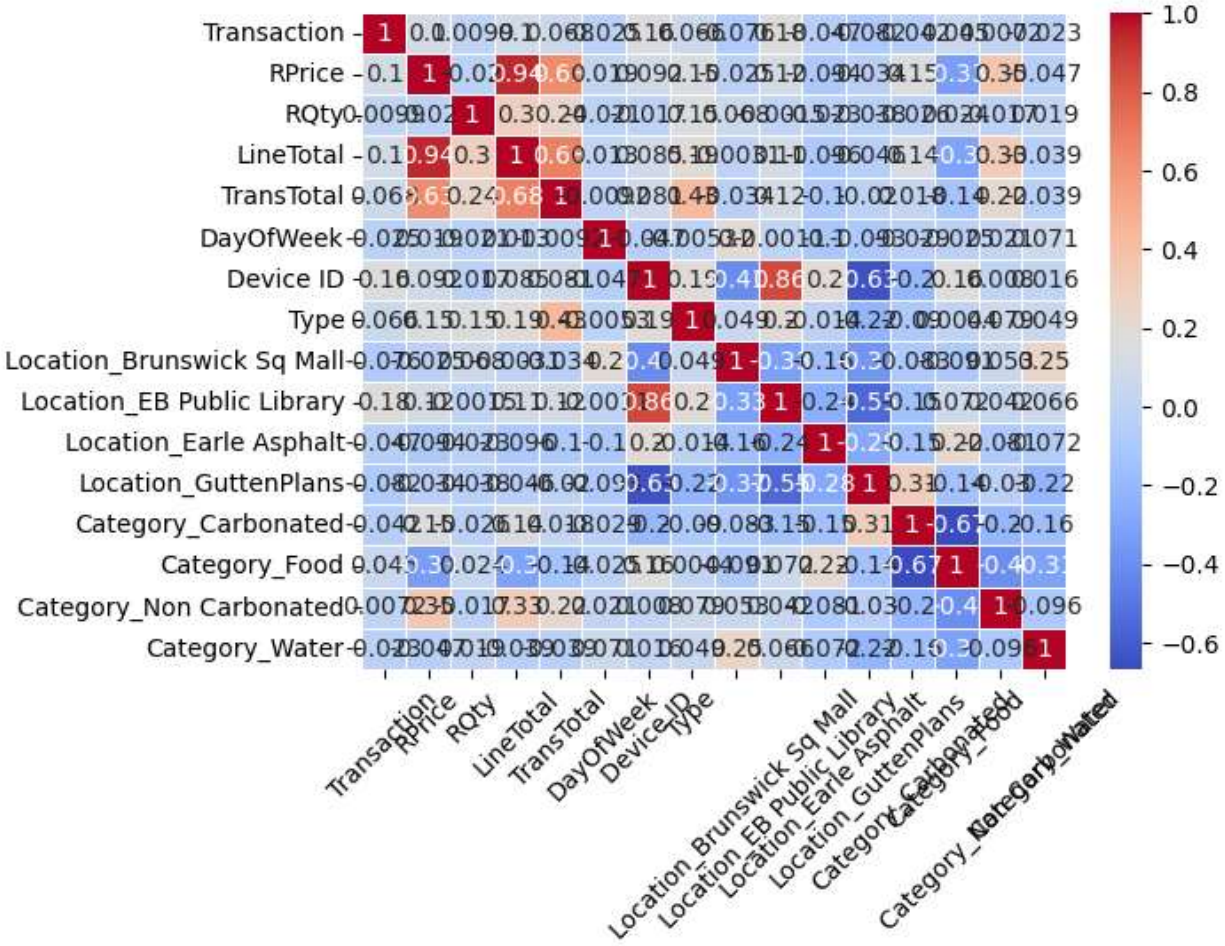

Рисунок 3 – Корреляционная матрица

Полученная корреляционная матрица свидетельствует о :

1. наличии зависимости между устройствами и локациями потому что на каждой лакаци расположены конкретные машины.
2. Цены и количества товаров в заказе – это свидетельствует что дорогие товары покупают несколько.
3. Категория и локации незначительно зависят друг от друга – это связано с тем что по графикам есть главная категория, кроме GuttenPlans, где доминируют сразу 2 категории товаров.

Обработка и преобразование признаков

Все категориальные признаки были закодированы с помощью OneHotEncoder.

Были удалены строки со status = ‘unlinked’ – этот статус означает что товар не был оплачен или при оплате произошла ошибка. При любом из случаев товар не был выдан, а значит при анализе такие покупки не должны учитываться.

Удалили признаки:

1. Status – по причине связи со статусом оплаты. В этом же этапе ранее мы удалили неудавшиеся покупки и теперь эта колонка имеет единственное значение.
2. Machine – по причине связи с Device id. При этом device id прикреплён к конкретным машинам, а machine нет.
3. RCoil – по причине большого количества уникальных значений и малой связанности с самой категорией продукта
4. MCoil, Rpice – по причине дублирования значение. Это было обнаружено при анализе датасета.
5. Product – по причине большого количества данных
6. Product Date – по причине большого количества данных и малой связанности с самой категории продукта

Построение и сравнение моделей

Были построены и обучены следующие модели:

1. Логистическая регрессия
2. Дерево решений
3. Метод опорных векторов
4. Метод k ближайших соседей
5. Случайный лес
6. Градиентный бустинг

Все модели были обучены на одной и той же выборке (80% обучающая, 20% тестовая), и поначалу использовались базовые параметры без настройки.

Для оценки моделей будем рассматривать следующие метрику F1 на графике и accuracy, precision, recall при выводе результата обучения модели

В результате почти все модели показали accuracy и F1-score выше 0.99. Худший результат показал методы логической регрессии и метод опорных векторов. Методы Decision Tree, Random Forest, Gradient Boosting были отброшены из-за переобучения

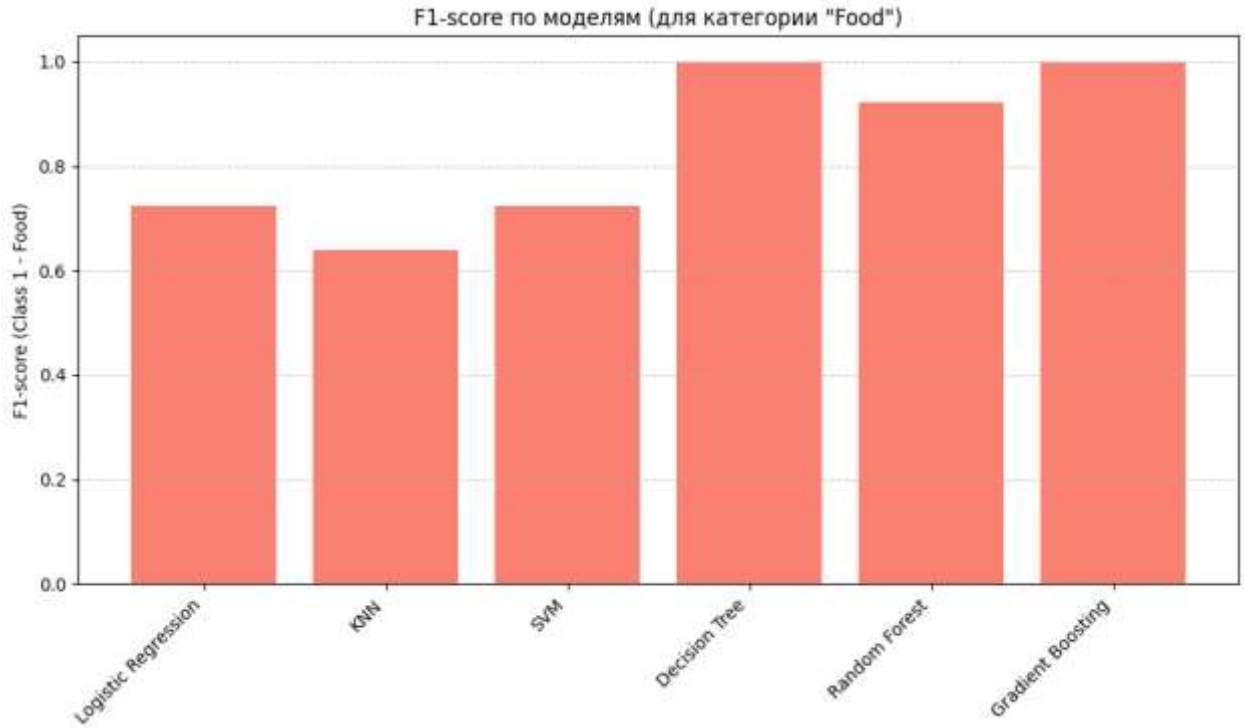

Рисунок 4 – Сравнение f-store baseline-моделей

Подбор гиперпараметов

Попробуем улучшить показатели метода K ближайших соседей путем подбора гиперпараметров с использованием GridSearchCV.

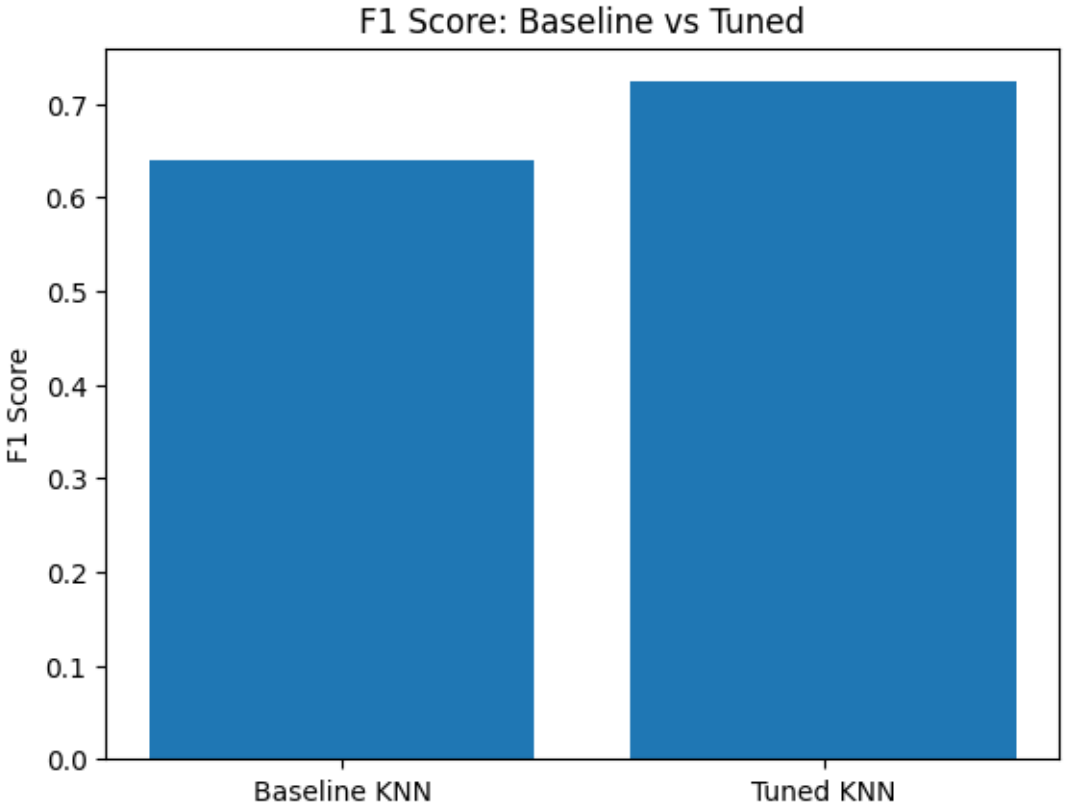

Рисунок 5 – Сравнение f1 для baseline-модели и улучшенной модели

В результате подбора гиперпараметров удалось поднять точность метода K ближайших соседей.

ЗАКЛЮЧЕНИЕ

В ходе выполнения научно-исследовательской работы была решена задача прогнозирования, будет ли куплен товар из категории "Food" , на основе данных о транзакциях вендинговых автоматов. Для достижения цели был проведён комплексный анализ данных, включавший разведочный анализ, обработку признаков и построение нескольких моделей машинного обучения для решения задачи бинарной классификации .

Разведочный анализ данных позволил выявить ключевые особенности распределений по таким признакам, как локация, тип машины, категория товара, цена, количество и день недели. Анализ корреляций показал отсутствие сильных линейных зависимостей между числовыми признаками, что указывает на необходимость использования нелинейных методов и кодирования категориальных переменных.

Построение и сравнение моделей машинного обучения показало, что большинство алгоритмов (дерево решений, случайный лес и градиентный бустинг) демонстрируют переобучение (ROC ≈ 1) на предварительно подготовленных данных. Метод k-ближайших соседей (KN) показал более скромные результаты в baseline-версии без настройки параметров.

Оптимизация гиперпараметров модели KNN с помощью GridSearchCV позволила незначитьльно улучшить её эффективность.

Таким образом, в рамках исследования были успешно реализованы и протестированы несколько моделей прогнозирования категории товара, что открывает возможность их дальнейшего применения для анализа спроса, оптимизации ассортимента и повышения эффективности управления вендинговыми аппаратами.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Student Performance & Learning Style // Kaggle URL: https://www.kaggle.com/datasets/adilshamim8/student-performance-and-learning-style/data (дата обращения: 29.05.2025).

2. Supervised learning // scikit-learn URL: https://scikit-learn.org/stable/supervised_learning.html (дата обращения: 29.05.2025).

3. Sokolova, M., Lapalme, G. (2009). A Systematic Analysis of Performance Measures for Classification Tasks. Information Processing & Management.

4. Statistical data visualization // Seaborn URL: https://seaborn.pydata.org/ (дата обращения: 29.05.2025).

5. Breiman, L. (2001). Random Forests. Machine Learning.

6. KNN Hyperparameters: A Friendly Guide to Optimization // ProgrammingR URL: https://www.programmingr.com/knn-hyperparameters-a-friendly-guide-to-optimization/ (дата обращения: 29.05.2025).