УДК 004.8; 656.1

ПРОГНОЗИРОВАНИЕ ВРЕМЕНИ ДВИЖЕНИЯ ТРАНСПОРТНЫХ СРЕДСТВ НА СЕГМЕНТНОМ УРОВНЕ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Даукараев Дамир Арманулы
Магистрант, образовательная программа 7М06138 – «Информационные системы», кафедра Информационные технологии,
Казахский университет технологии и бизнеса имени К. Кулажанова,
г.Астана, Республика Казахстан

***Научный руководитель**: Кантуреева М.А., Доктор PhD*

В статье рассматривается задача прогнозирования времени движения транспортных средств на сегментном уровне с использованием методов машинного обучения. Актуальность исследования обусловлена необходимостью повышения точности оценки времени движения в интеллектуальных транспортных системах, что является важным фактором для оптимизации управления транспортными потоками и повышения качества обслуживания общественного транспорта. В качестве исходных данных использован открытый реальный датасет мониторинга движения автобусов города Астаны, сформированный на основе GPS-данных и приведённый к формату GTFS.

В рамках исследования выполнена предобработка сегментно-ориентированных транспортных данных с целью предотвращения утечки целевой информации и формирования информативных временных и маршрутных признаков. Для решения задачи регрессии реализованы и сравнительно оценены модели Random Forest, градиентного бустинга XGBoost и рекуррентные нейронные сети LSTM, включая вариант с эмбеддингами категориальных признаков. Оценка качества моделей проводилась с использованием временного разбиения выборки и стандартных метрик регрессионного анализа.

Экспериментальные результаты показали, что модель XGBoost обеспечивает наивысшую объясняющую способность, достигая коэффициента детерминации выше 0,84, тогда как модель LSTM с эмбеддингами демонстрирует минимальные значения средней абсолютной и среднеквадратичной ошибок, обеспечивая наибольшую точность прогнозирования. Полученные результаты подтверждают эффективность как табличных, так и последовательных моделей для сегментного прогнозирования времени движения и указывают на перспективность их

комбинированного использования в интеллектуальных транспортных системах.

Ключевые слова: интеллектуальные транспортные системы, прогнозирование времени движения, сегментный уровень, машинное обучение, XGBoost, LSTM, транспортные данные.

Введение

В условиях интенсивной урбанизации и роста транспортных потоков задача точного прогнозирования времени движения транспортных средств приобретает особую актуальность. Надёжная оценка времени прохождения участков маршрута является ключевым элементом интеллектуальных транспортных систем, поскольку напрямую влияет на эффективность управления дорожным движением, оптимизацию маршрутов, снижение задержек и повышение качества транспортного обслуживания пассажиров [1]. Неточность прогнозов времени движения приводит к неэффективному использованию транспортной инфраструктуры, увеличению времени в пути и росту эксплуатационных издержек.

Традиционные методы оценки времени движения, основанные на детерминированных и статистических моделях, зачастую не способны в полной мере учитывать сложную динамику транспортных процессов, влияние временных факторов, а также нелинейные зависимости между параметрами движения [2]. В связи с этим в последние годы всё большее внимание уделяется применению методов машинного обучения, которые позволяют выявлять скрытые закономерности в больших объёмах транспортных данных и обеспечивать более высокую точность прогнозирования времени движения [3].

Особый интерес представляет прогнозирование времени движения на сегментном уровне, при котором маршрут рассматривается как последовательность отдельных сегментов между контрольными точками или остановками. Такой подход позволяет учитывать локальные особенности движения, повышает гибкость прогнозных моделей и делает возможной их интеграцию в системы оперативного управления транспортом [4]. Вместе с тем сегментный уровень моделирования предъявляет повышенные требования к качеству данных, корректности экспериментального дизайна и предотвращению утечки целевой информации.

Несмотря на наличие значительного числа исследований в области прогнозирования времени движения транспортных средств, остаются актуальными вопросы сравнительной оценки различных моделей машинного обучения на сегментно-ориентированных данных, а также анализа влияния временных и маршрутных признаков на точность прогнозов [5]. В этой связи проведение экспериментальных исследований с использованием реальных

данных мониторинга транспортных поездок представляет существенный научный и практический интерес.

Целью данной работы является разработка и экспериментальная оценка моделей машинного обучения для прогнозирования времени движения транспортных средств на сегментном уровне. Для достижения поставленной цели в работе решаются следующие задачи: анализ и предобработка сегментных транспортных данных, формирование информативных признаков, построение и обучение регрессионных моделей, а также сравнительная оценка их точности с использованием корректных протоколов валидации. Полученные результаты могут быть использованы при разработке и внедрении интеллектуальных транспортных систем.

Материал и методы исследования

В последние годы задача прогнозирования времени движения транспортных средств активно исследуется в рамках интеллектуальных транспортных систем и транспортной аналитики. Развитие технологий сбора транспортных данных и рост вычислительных возможностей привели к широкому применению методов машинного обучения для решения задач прогнозирования временных характеристик движения [7]. В отличие от классических детерминированных и статистических моделей, методы машинного обучения позволяют учитывать сложные нелинейные зависимости между параметрами транспортного потока и обеспечивают более высокую точность прогнозирования.

Результаты современных исследований подтверждают эффективность применения как классических алгоритмов машинного обучения, так и глубоких нейронных сетей для прогнозирования времени движения. В частности, в работе [8] проведено сравнительное исследование алгоритмов k-ближайших соседей, рекуррентных нейронных сетей (LSTM) и архитектур Transformer, показавшее преимущество глубоких моделей в условиях сложной и нестабильной дорожной обстановки. В исследовании [9] предложен двухэтапный метод отбора признаков, позволяющий существенно сократить размерность входных данных без потери точности прогнозирования.

Отдельное направление исследований связано с интеграцией дополнительных факторов, влияющих на движение транспортных средств. В работе [10] предложена OD-ориентированная модель прогнозирования времени движения с учётом внешних факторов, таких как погодные условия и поведенческие характеристики водителей, что позволило повысить качество прогнозов по сравнению с базовыми моделями.

Данные

В рамках настоящего исследования использован открытый реальный датасет мониторинга движения общественного транспорта, сформированный на основе необработанных GPS-данных и приведённый к формату GTFS (General

Transit Feed Specification) [6]. Данный набор данных содержит сегментно-ориентированную информацию о движении транспортных средств, включая временные характеристики, идентификаторы маршрутов и параметры прохождения сегментов.

Каждая запись датасета соответствует прохождению одного сегмента маршрута в рамках отдельной транспортной поездки, что позволяет формализовать задачу прогнозирования времени движения на сегментном уровне. Общий объём выборки составляет несколько сотен тысяч сегментов, что обеспечивает статистическую надёжность эксперимента и позволяет рассматривать задачу в условиях, приближённых к реальной эксплуатации интеллектуальных транспортных систем.

Ключевыми атрибутами датасета являются идентификатор поездки, направление движения, номер сегмента маршрута, точки начала и окончания сегмента, а также временные параметры, характеризующие движение транспортного средства. В качестве целевой переменной в основном эксперименте используется время движения по сегменту, выраженное в секундах.

Предобработка данных

На этапе предобработки данных особое внимание уделено предотвращению утечки целевой информации. Из набора признаков были исключены параметры, напрямую связанные с фактическим временем прохождения сегмента и формируемые после его завершения. Такой подход позволил обеспечить корректность постановки задачи прогнозирования и объективность экспериментальной оценки моделей машинного обучения.

Временные атрибуты были преобразованы путём извлечения агрегированных признаков, отражающих циклическую природу транспортных процессов, в частности часа начала движения и дня недели. Категориальные признаки, включая сегмент маршрута, направление движения и идентификаторы поездок, были закодированы с использованием методов, совместимых с применяемыми моделями машинного обучения. Пропущенные значения, при их наличии, обрабатывались с использованием стандартных методов очистки и фильтрации данных.

Методы машинного обучения

Задача прогнозирования времени движения транспортных средств формализована как задача регрессии, в которой по набору временных, маршрутных и сегментных признаков требуется оценить значение времени прохождения отдельного сегмента маршрута. Такой подход является стандартным для задач прогнозирования временных характеристик движения в интеллектуальных транспортных системах и широко используется в современных исследованиях [11].

В рамках настоящего исследования были реализованы и сравнительно оценены несколько моделей машинного обучения, представляющих различные классы алгоритмов. В качестве базовой линейной модели использована линейная регрессия, позволяющая оценить вклад отдельных признаков и служащая интерпретируемым ориентиром для сравнения с более сложными методами. Несмотря на ограниченные возможности в учёте нелинейных зависимостей, линейные модели часто применяются в задачах прогнозирования времени движения в качестве базового уровня качества [12].

В качестве нелинейной ансамблевой модели применён метод случайного леса, основанный на построении множества решающих деревьев и их агрегировании. Данный алгоритм устойчив к шуму, способен учитывать сложные взаимодействия между признаками и широко используется для прогнозирования транспортных характеристик при работе с разнородными табличными данными [13].

Для повышения точности прогнозирования использованы методы градиентного бустинга, которые последовательно обучают ансамбль слабых моделей с целью минимизации ошибки предсказания. Алгоритмы градиентного бустинга хорошо зарекомендовали себя в задачах прогнозирования времени движения и, согласно ряду исследований, демонстрируют более высокую точность по сравнению с классическими моделями и случайным лесом при работе с сегментно-ориентированными транспортными данными [14].

Обучение всех моделей проводилось с фиксированными значениями параметров воспроизводимости, что обеспечило стабильность полученных результатов и возможность повторения эксперимента. Для корректной оценки качества прогнозирования использовалось временное разбиение данных, исключающее использование информации из будущих периодов при обучении моделей.

Метрики оценки качества моделей

Для количественной оценки точности прогнозирования времени движения транспортных средств использованы стандартные метрики регрессионного анализа, широко применяемые в задачах прогнозирования транспортных характеристик. Выбор данных метрик обусловлен их интерпретируемостью и возможностью комплексной оценки качества моделей.

Средняя абсолютная ошибка (MAE)

Средняя абсолютная ошибка характеризует среднее по выборке абсолютное отклонение прогнозируемых значений от фактических и позволяет оценить точность модели в тех же единицах измерения, что и целевая переменная.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \widehat{y_i}|$$

где,

y_i — фактическое значение времени движения для i-го наблюдения,

$\hat{y}_i$ — прогнозируемое значение,
N — количество наблюдений.

Среднеквадратичная ошибка (RMSE)

Среднеквадратичная ошибка является более чувствительной к большим отклонениям и позволяет оценить наличие крупных ошибок прогнозирования, что особенно важно в задачах транспортного моделирования.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

Чем меньше значение RMSE, тем выше точность модели прогнозирования.

Коэффициент детерминации (R^2)

Коэффициент детерминации показывает долю дисперсии целевой переменной, объясняемую моделью, и используется для оценки качества аппроксимации.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

где,

$\hat{y}_i$ — среднее значение фактических наблюдений.

Значение коэффициента $\mathbf{R^2}$ лежит в диапазоне от 0 до 1, при этом более высокие значения свидетельствуют о лучшем качестве модели.

Использование совокупности метрик MAE, RMSE и $\mathbf{R^2}$ позволяет получить комплексную оценку качества моделей машинного обучения. MAE отражает среднюю величину ошибки в интерпретируемых единицах, RMSE позволяет выявить модели с крупными отклонениями прогнозов, а коэффициент детерминации характеризует общую объясняющую способность модели. Такой подход обеспечивает объективное сравнение различных моделей при прогнозировании времени движения транспортных средств на сегментном уровне.

Результаты и обсуждение

В рамках экспериментального исследования была проведена сравнительная оценка моделей машинного обучения для прогнозирования времени движения транспортных средств на сегментном уровне с использованием реальных данных движения автобусов города Астаны. Для всех моделей применялось временное разбиение выборки, исключающее утечку информации из будущих периодов и обеспечивающее корректную оценку качества прогнозирования в условиях, приближённых к реальной эксплуатации интеллектуальных транспортных систем.

Результаты сравнительной оценки моделей машинного обучения для прогнозирования времени движения представлены в таблице 1. Для оценки качества прогнозирования использовались метрики средней абсолютной ошибки, среднеквадратичной ошибки и коэффициента детерминации.

В качестве базового уровня качества была рассмотрена модель случайного леса (Random Forest). Полученные результаты показали, что данная модель обеспечивает среднюю абсолютную ошибку 40,95 секунды и коэффициент детерминации 0,804. Значение MAE менее одной минуты свидетельствует о высокой точности прогнозирования времени движения на сегментном уровне и подтверждает применимость ансамблевых методов для анализа транспортных данных. В то же время относительно высокое значение RMSE указывает на чувствительность модели к отдельным выбросам и сложным сценариям движения.

На следующем этапе была исследована модель градиентного бустинга XGBoost, продемонстрировавшая улучшение качества прогнозирования по всем используемым метрикам. Значение MAE снизилось до 30,12 секунды, RMSE — до 72,02 секунды, а коэффициент детерминации увеличился до 0,847. Это означает, что модель XGBoost объясняет около 85 % вариации фактического времени движения, что является высоким показателем для задач сегментного прогнозирования. Полученные результаты подтверждают эффективность методов градиентного бустинга при работе с табличными транспортными данными, содержащими нелинейные зависимости и взаимодействия признаков.

Для оценки влияния временной структуры маршрута на качество прогнозирования была реализована последовательная модель на основе рекуррентной нейронной сети LSTM. Базовая версия модели без использования эмбеддингов категориальных признаков показала ограниченную объясняющую способность ($R^2 = 0,407$), несмотря на умеренное снижение средней абсолютной ошибки по сравнению с моделью случайного леса. Это указывает на то, что без специализированных механизмов обработки категориальных признаков последовательные модели не в полной мере используют доступную информацию.

Существенное улучшение результатов было достигнуто при использовании LSTM с эмбеддингами категориальных признаков сегмента маршрута. Лучшая модель, выбранная по критерию ранней остановки, обеспечила значение MAE 28,35 секунды и RMSE 61,53 секунды, что является наилучшим результатом среди всех рассмотренных моделей по данным метрикам. При этом коэффициент детерминации составил 0,646, что ниже значения, полученного для XGBoost, но значительно превосходит показатели базовой версии LSTM.

Таблица 1. Сравнительная оценка моделей прогнозирования времени движения

Модель	MAE, сек	RMSE, сек	MAE, мин	RMSE, мин	R^2
Random Forest	40,95	81,60	0,68	1,36	0,804
XGBoost	30,12	72,02	0,50	1,20	0,847
LSTM (без эмбеддингов)	37,74	79,78	0,63	1,33	0,407
LSTM + Embeddings (K=12)	28,35	61,53	0,47	1,03	0,646

Сравнительный анализ моделей машинного обучения представлен также в виде графиков (рисунки 1–3). Как видно из рисунка 1, последовательная модель LSTM с эмбеддингами демонстрирует наименьшее значение MAE. Согласно рисунку 2, та же модель обеспечивает минимальное значение RMSE. В то же время, как показано на рисунке 3, модель XGBoost характеризуется наибольшим значением коэффициента детерминации, что указывает на её более высокую объясняющую способность.

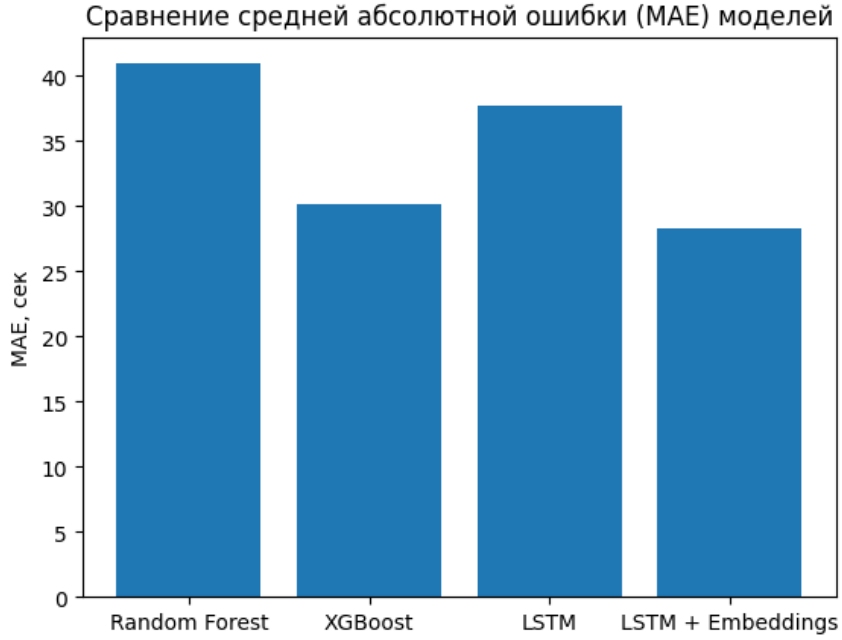

Рисунок 1. Сравнение средней абсолютной ошибки (MAE) моделей прогнозирования

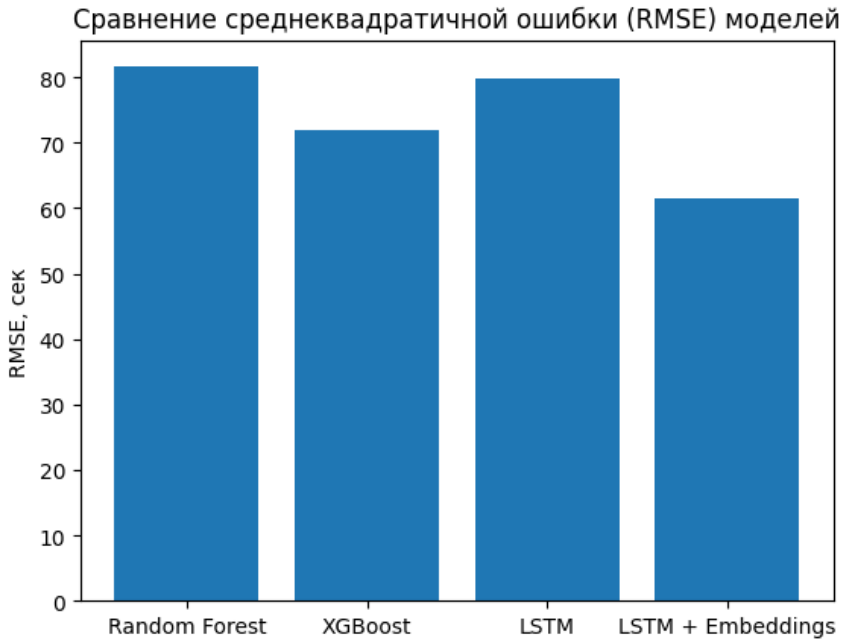

Рисунок 2. Сравнение среднеквадратичной ошибки (RMSE) моделей прогнозирования

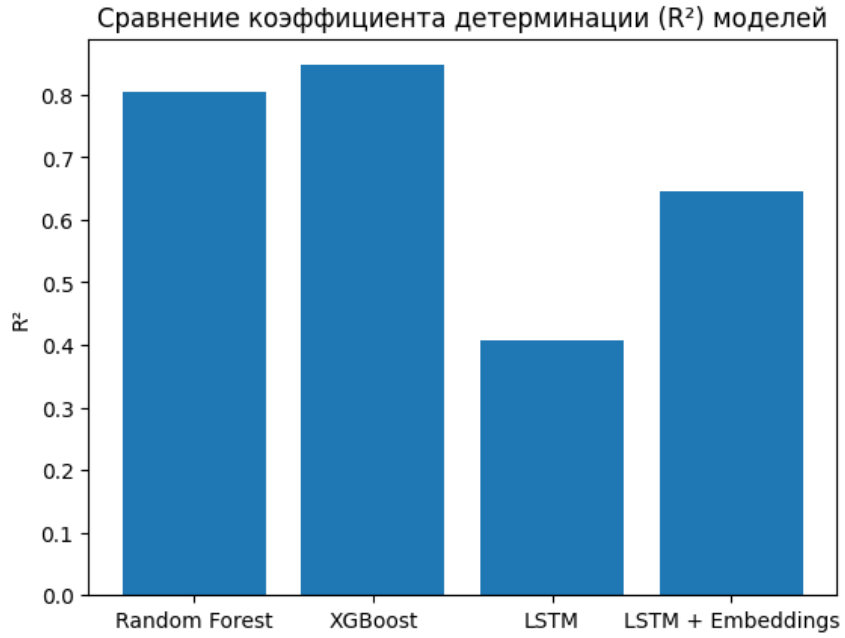

Рисунок 3. Сравнение коэффициента детерминации (R^2)

Сравнительный анализ результатов показывает, что различные классы моделей обладают различными преимуществами. Модель XGBoost демонстрирует наивысшую объясняющую способность и наиболее полно описывает общую вариацию времени движения, тогда как последовательная

модель LSTM с эмбеддингами обеспечивает минимальные средние ошибки прогнозирования, что особенно важно для практических задач оценки времени прибытия и оперативного управления маршрутами общественного транспорта. Это указывает на целесообразность выбора модели в зависимости от конкретной прикладной задачи, а также подтверждает перспективность комбинированных и гибридных подходов к прогнозированию времени движения в интеллектуальных транспортных системах.

Выводы

В данной работе была рассмотрена задача прогнозирования времени движения транспортных средств на сегментном уровне с использованием методов машинного обучения и последовательного моделирования. Исследование проведено на основе реальных данных движения автобусов города Астаны, представленных в сегментно-ориентированном формате, что обеспечило практическую значимость и воспроизводимость полученных результатов.

В ходе экспериментов выполнена сравнительная оценка нескольких классов моделей, включая ансамблевые методы машинного обучения и рекуррентные нейронные сети. Показано, что модель случайного леса обеспечивает устойчивый базовый уровень качества прогнозирования с ошибкой менее одной минуты, что подтверждает применимость ансамблевых методов для анализа транспортных данных. Модель градиентного бустинга XGBoost продемонстрировала наивысшую объясняющую способность, обеспечив максимальное значение коэффициента детерминации и наиболее полное описание вариации времени движения транспортных средств.

Последовательная модель LSTM без использования эмбеддингов категориальных признаков показала ограниченную эффективность, что указывает на недостаточность прямого применения рекуррентных нейронных сетей к сегментно-ориентированным транспортным данным. Вместе с тем введение эмбеддингов категориальных признаков позволило существенно повысить качество последовательного моделирования. Модель LSTM с эмбеддингами обеспечила наименьшие значения средней абсолютной и среднеквадратичной ошибок среди всех рассмотренных моделей, что особенно важно для практических задач оценки времени прибытия и оперативного управления маршрутами общественного транспорта.

Полученные результаты показывают, что выбор модели прогнозирования должен определяться конкретной прикладной задачей. Методы градиентного бустинга целесообразно использовать в задачах анализа и объяснения транспортных процессов, тогда как последовательные модели с эмбеддингами являются более предпочтительными при минимизации средней ошибки прогнозирования. Это подтверждает перспективность комбинированных и

гибридных подходов, сочетающих преимущества табличных и последовательных моделей.

В качестве направлений дальнейших исследований можно выделить разработку гибридных архитектур, объединяющих градиентный бустинг и нейронные последовательные модели, использование механизмов внимания и графовых представлений маршрутов, а также расширение набора признаков за счёт включения внешних факторов, таких как погодные условия и загруженность дорожной сети. Реализация данных направлений может способствовать дальнейшему повышению точности прогнозирования времени движения в интеллектуальных транспортных системах.

Список использованной литературы

1. Hassan, M., Al Nafees, A., Shraban, S. S., et al. (2025). Application of machine learning in intelligent transport systems: A comprehensive review and bibliometric analysis. *Discover Civil Engineering*, 2, Article 98. https://doi.org/10.1007/s44290-025-00256-2

2. Kandiri, A., Ghiasi, R., Nogal, M., & Teixeira, R. (2024). Travel time prediction for an intelligent transportation system based on a data-driven feature selection method considering temporal correlation. *Transportation Engineering*, 18, Article 100272. https://doi.org/10.1016/j.treng.2024.100272

3. Jang, J. (2024). Travel time prediction using machine learning algorithms: Focusing on k-NN, LSTM, and Transformer. *The Open Transportation Journal*, 18, Article e26671212356139. https://doi.org/10.2174/0126671212356139241101070347

4. Lei, J., Chen, Y., Han, Q., Zeng, L., & He, G. (2025). Effective bus travel time prediction system of multiple routes: Introducing PMLNet based on MDARNN. *Applied Sciences*, 15(14), Article 8104. https://doi.org/10.3390/app15148104

5. Shi, C., Zou, W., Wang, Y., Zhu, Z., Chen, T., Zhang, Y., & Wang, N. (2025). Enhancing travel time prediction for intelligent transportation systems: A high-resolution origin–destination-based approach with multi-dimensional features. *Sustainability*, 17(5), Article 2111. https://doi.org/10.3390/su17052111

6. Mansurova, A., Mussina, A., Aubakirov, S., Nugumanova, A., & Yedilkhan, D. (2025). From raw GPS to GTFS: A real-world open dataset for bus travel time prediction. *Data*, 10(8), Article 119. https://doi.org/10.3390/data10080119

7. Chen, M.-Y., Chiang, H.-S., & Yang, K.-J. (2022). Constructing cooperative intelligent transport systems for travel time prediction with deep learning approaches. *IEEE Transactions on Intelligent Transportation Systems*, 23(9), 16590–16599. https://doi.org/10.1109/TITS.2022.3148269

8. Abdollahi, M., Khaleghi, T., & Yang, K. (2020). An integrated feature learning approach using deep learning for travel time prediction. *Expert Systems with Applications*, 139, Article 112864. https://doi.org/10.1016/j.eswa.2019.112864

9. Chughtai, J.-U.-R., Haq, I. U., Shafiq, O., & Muneeb, M. (2022). Travel time prediction using hybridized deep feature space and machine learning based heterogeneous ensemble. *IEEE Access*, 10, 98127–98139. https://doi.org/10.1109/ACCESS.2022.3206384

10. Shanthappa, N. K., Mulangi, R. H., & Manjunath, H. M. (2024). Origin–destination demand prediction of public transit using graph convolutional neural network. *Case Studies on Transport Policy*, 17, Article 101230. https://doi.org/10.1016/j.cstp.2024.101230

11. Jiang, X., & Adeli, H. (2005). Dynamic wavelet neural network model for traffic flow forecasting. *Journal of Transportation Engineering*, 131(10), 771–782. https://doi.org/10.1061/(ASCE)0733-947X(2005)131:10(771)

12. Gao, Y., Zhou, C., Rong, J., Wang, Y., & Liu, S. (2022). Short-term traffic speed forecasting using a deep learning method based on multitemporal traffic flow volume. *IEEE Access*, 10, 82384–82395. https://doi.org/10.1109/ACCESS.2022.3195353

13. Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random forest algorithm overview. *Babylonian Journal of Machine Learning*, 2024, 69–79. https://doi.org/10.58496/BJML/2024/007

14. Szczepanek, R. (2022). Daily streamflow forecasting in mountainous catchment using XGBoost, LightGBM and CatBoost. *Hydrology*, 9(12), Article 226. https://doi.org/10.3390/hydrology9120226

КӨЛІК ҚҰРАЛДАРЫНЫҢ ҚОЗҒАЛЫС УАҚЫТЫН СЕГМЕНТТІК ДЕҢГЕЙДЕ МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІН ҚОЛДАНА ОТЫРЫП БОЛЖАУ

Даукараев Дамир Арманұлы

Ғылыми жетекші: *Кантуреева М.А., PhD докторы*

Бұл мақалада машиналық оқыту әдістерін қолдану арқылы көлік құралдарының қозғалыс уақытын сегменттік деңгейде болжау мәселесі қарастырылады. Зерттеудің өзектілігі интеллектуалды көлік жүйелерінде қозғалыс уақытын дәл бағалау қажеттілігімен байланысты, себебі бұл көлік ағындарын тиімді басқару және қоғамдық көлік қызметінің сапасын арттыру үшін маңызды болып табылады. Зерттеу барысында Астана қаласындағы автобустар қозғалысының GPS деректері негізінде қалыптастырылған және

GTFS форматына келтірілген ашық нақты деректер жиынтығы пайдаланылды.

Деректерді алдын ала өңдеу кезеңінде мақсатты айнымалы бойынша ақпараттың ағып кетуін болдырмау және уақыттық әрі маршруттық белгілерді қалыптастыру жүзеге асырылды. Регрессия есебін шешу үшін Random Forest, градиенттік бустинг XGBoost және LSTM рекурренттік нейрондық желілері, сондай-ақ категориялық маршруттық белгілерге арналған эмбеддингтері бар LSTM моделі қолданылып, салыстырмалы түрде бағаланды. Модельдердің сапасы уақытқа тәуелді деректерді бөлу тәсілі және стандартты регрессиялық метрикалар арқылы бағаланды.

Эксперименттік нәтижелер XGBoost моделінің ең жоғары түсіндіру қабілетін қамтамасыз ететінін және детерминация коэффициентінің 0,84-тен жоғары мәнге жеткенін көрсетті, ал эмбеддингтері бар LSTM моделі MAE және RMSE метрикалары бойынша ең төмен қателіктерді қамтамасыз етті. Алынған нәтижелер қозғалыс уақытын сегменттік деңгейде болжау үшін табличалық және тізбектік модельдерді қолданудың тиімділігін және интеллектуалды көлік жүйелерінде оларды біріктіріп пайдаланудың перспективалылығын дәлелдейді.

Кілт сөздері: интеллектуалды көлік жүйелері, қозғалыс уақытын болжау, сегменттік деңгей, машиналық оқыту, XGBoost, LSTM, көлік деректері.

TRAVEL TIME PREDICTION OF VEHICLES AT THE SEGMENT LEVEL USING MACHINE LEARNING METHODS

Daukaraev Damir Armanuly

Scientific Supervisor: *Kantureyeva M.A., PhD*

This paper addresses the problem of segment-level travel time prediction using machine learning methods. The relevance of the study is driven by the need to improve travel time estimation accuracy in intelligent transportation systems, which is essential for efficient traffic flow management and improving the quality of public transport services. The study is based on an open real-world dataset of bus movements in the city of Astana, constructed from raw GPS data and transformed into the GTFS format.

Data preprocessing was performed to prevent target information leakage and to construct informative temporal and route-related features. Several regression models were implemented and comparatively evaluated, including Random Forest, gradient boosting (XGBoost), and recurrent neural networks (LSTM), as well as an LSTM

model with embeddings for categorical route features. Model evaluation was conducted using a time-aware train–test split and standard regression metrics.

Experimental results demonstrate that the XGBoost model achieves the highest explanatory power, with a coefficient of determination exceeding 0.84, while the LSTM model with embeddings provides the lowest prediction errors, achieving the best performance in terms of MAE and RMSE. The findings confirm the effectiveness of both tabular and sequential modeling approaches for segment-level travel time prediction and highlight the potential of their combined use in intelligent transportation systems.

Keywords: intelligent transportation systems, travel time prediction, segment-level modeling, machine learning, XGBoost, LSTM, transportation data.

REFERENCES

1. Hassan, M., Al Nafees, A., Shraban, S. S., et al. (2025). Application of machine learning in intelligent transport systems: A comprehensive review and bibliometric analysis. *Discover Civil Engineering*, 2, Article 98. https://doi.org/10.1007/s44290-025-00256-2

2. Kandiri, A., Ghiasi, R., Nogal, M., & Teixeira, R. (2024). Travel time prediction for an intelligent transportation system based on a data-driven feature selection method considering temporal correlation. *Transportation Engineering*, 18, Article 100272. https://doi.org/10.1016/j.treng.2024.100272

3. Jang, J. (2024). Travel time prediction using machine learning algorithms: Focusing on k-NN, LSTM, and Transformer. *The Open Transportation Journal*, 18, Article e26671212356139. https://doi.org/10.2174/0126671212356139241101070347

4. Lei, J., Chen, Y., Han, Q., Zeng, L., & He, G. (2025). Effective bus travel time prediction system of multiple routes: Introducing PMLNet based on MDARNN. *Applied Sciences*, 15(14), Article 8104. https://doi.org/10.3390/app15148104

5. Shi, C., Zou, W., Wang, Y., Zhu, Z., Chen, T., Zhang, Y., & Wang, N. (2025). Enhancing travel time prediction for intelligent transportation systems: A high-resolution origin–destination-based approach with multi-dimensional features. *Sustainability*, 17(5), Article 2111. https://doi.org/10.3390/su17052111

6. Mansurova, A., Mussina, A., Aubakirov, S., Nugumanova, A., & Yedilkhan, D. (2025). From raw GPS to GTFS: A real-world open dataset for bus travel time prediction. *Data*, 10(8), Article 119. https://doi.org/10.3390/data10080119

7. Chen, M.-Y., Chiang, H.-S., & Yang, K.-J. (2022). Constructing cooperative intelligent transport systems for travel time prediction with deep learning approaches. *IEEE Transactions on Intelligent Transportation Systems*, 23(9), 16590–16599. https://doi.org/10.1109/TITS.2022.3148269

8. Abdollahi, M., Khaleghi, T., & Yang, K. (2020). An integrated feature learning approach using deep learning for travel time prediction. *Expert Systems with Applications*, 139, Article 112864. https://doi.org/10.1016/j.eswa.2019.112864

9. Chughtai, J.-U.-R., Haq, I. U., Shafiq, O., & Muneeb, M. (2022). Travel time prediction using hybridized deep feature space and machine learning based heterogeneous ensemble. *IEEE Access*, 10, 98127–98139. https://doi.org/10.1109/ACCESS.2022.3206384

10. Shanthappa, N. K., Mulangi, R. H., & Manjunath, H. M. (2024). Origin–destination demand prediction of public transit using graph convolutional neural network. *Case Studies on Transport Policy*, 17, Article 101230. https://doi.org/10.1016/j.cstp.2024.101230

11. Jiang, X., & Adeli, H. (2005). Dynamic wavelet neural network model for traffic flow forecasting. *Journal of Transportation Engineering*, 131(10), 771–782. https://doi.org/10.1061/(ASCE)0733-947X(2005)131:10(771)

12. Gao, Y., Zhou, C., Rong, J., Wang, Y., & Liu, S. (2022). Short-term traffic speed forecasting using a deep learning method based on multitemporal traffic flow volume. *IEEE Access*, 10, 82384–82395. https://doi.org/10.1109/ACCESS.2022.3195353

13. Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random forest algorithm overview. *Babylonian Journal of Machine Learning*, 2024, 69–79. https://doi.org/10.58496/BJML/2024/007

14. Szczepanek, R. (2022). Daily streamflow forecasting in mountainous catchment using XGBoost, LightGBM and CatBoost. *Hydrology*, 9(12), Article 226. https://doi.org/10.3390/hydrology9120226