Big Data-Driven Trip Duration Prediction in Urban Transportation Systems

Elias Dritsas, Maria Trigka and Phivos Mylonas
Department of Informatics and Computer Engineering
University of West Attica, Egaleo Park Campus, 12243 Athens, Greece
Email:{idritsas,mtrigka,mylonasf}@uniwa.gr

***Abstract*—This paper presents a machine learning (ML)-based approach to trip duration prediction using a large-scale mobility dataset from New York City (NYC) yellow taxi, incorporating both raw and engineered features to model spatiotemporal and fare-related factors that influence trip duration. Six regression models including Linear Regression (LR), Support Vector Regressor (SVR), Random Forest (RF), Gradient Boosting (GB), XGBoost, and Multi-Layer Perceptron (MLP), were trained and evaluated using standard metrics namely, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). The RF model achieved the best performance, with an MAE of 53.83 sec, an RMSE of 143.38 sec, and an R^2 score of 0.929. These results demonstrate the suitability of ensemble tree-based models for predictive analytics in intelligent transport systems and outline future directions for enhancing performance using contextual and real-time data streams.**

***Index Terms*—Big Data, Trip Duration Prediction, Models, Machine Learning, Urban Mobility**

I. Introduction

Accurate prediction of taxi trip duration is a key enabler for efficient urban mobility systems. As cities grow increasingly complex and data-rich, transportation analytics must leverage scalable, data-driven approaches to provide timely and accurate travel-time estimates. The widespread deployment of global positioning system (GPS)-enabled services and digital fare systems has enabled the collection of massive volumes of trip data, allowing for predictive modelling at an unprecedented scale. In this context, ML techniques, particularly those applied to large-scale spatiotemporal datasets, offer a compelling path forward for enhancing the responsiveness and efficiency of urban transport operations [1].

Recent studies have shown that combining raw mobility data with engineered features such as time-of-day or fare-per-mile can significantly improve the quality of trip duration predictions. However, identifying suitable models that can handle data heterogeneity, non-linearity, and scale is still an open challenge. Different algorithms offer trade-offs between accuracy, interpretability, and deployment cost, especially when used in city-scale applications involving millions of trips per month [2].

Motivated by these challenges, this work explores the effectiveness of various supervised ML models in predicting taxi trip duration from real-world NYC yellow taxi data. The goal is to evaluate how well different regression algorithms perform under consistent preprocessing conditions and to identify models that strike the best balance between predictive power and practical deployment feasibility. The main contributions of this study are as follows:

- A feature-rich dataset of taxi trips by combining raw spatial, temporal, and fare-based attributes.
- A systematic comparison of six ML regressors, LR, SVR, RF, GB, XGBoost, and MLP, using three standard evaluation metrics (MAE, RMSE, and R^2).
- Empirical evidence that RF consistently outperforms all other models in terms of accuracy, generalization, and training efficiency, while maintaining robustness and relatively low computational complexity due to parallelizable training and minimal preprocessing requirements.

The rest of the paper is organised as follows. In Section II, we describe the dataset and the adopted methodology. Next, in Section III, we discuss the acquired research results. Finally, conclusions and future directions are outlined in Section IV.

II. Material and Methods

This section presents the dataset, preprocessing pipeline, and ML methodology employed for trip duration prediction. We describe the construction of a feature-rich dataset derived from real-world urban mobility records. Furthermore, we detail the regression models and evaluation metrics used to assess predictive performance under consistent experimental conditions.

A. Data Preprocessing and Description

The dataset used in this study is sourced from the publicly available records of the NYC taxi in January 2015. It includes detailed logs of individual yellow taxi trips, each represented by spatio-temporal information (pickup_datetime, dropoff_datetime, and coordinates), trip context (passenger_count, trip_distance), and fare-related data (fare_amount, tip_amount, tolls_amount, and total_amount). To ensure data quality, a systematic cleaning process was applied to eliminate entries with missing or corrupt values, implausible distances, durations, or geographic coordinates, and zero or negative fares. After filtering, over 1 million valid trip records were retained for analysis.

A total of 11 features were used to support regression-based prediction of trip_duration. These included both raw variables directly extracted from the original records,

such as trip_distance, fare_amount, passenger_count, and total_amount, and engineered attributes derived during preprocessing. The target variable trip_duration was computed as the difference (dropoff_datetime - pickup_datetime). Additional features, such as pickup_hour, day_of_week, and is_weekend, were extracted from timestamps to capture temporal patterns, while log_trip_distance and fare_per_mile were introduced to mitigate skewness and model spatial-economic behavior. Features not subject to explicit filtering, either because they were created from clean inputs or inherently valid, were retained without further modification.

Overall, data preprocessing served a dual purpose. First, to enforce strict quality control via rule-based filtering (to remove outliers), and second, to enhance the feature space through transformation and enrichment. This two-phase approach enabled the construction of a robust, ML–ready dataset tailored to modeling urban trip duration. Table I summarizes the selected features, the applied filtering rules, their descriptions, and key statistical properties in the cleaned dataset.

TABLE I
SUMMARY OF APPLIED FILTERING RULES, FEATURE TYPES, AND STATISTICS FOR SELECTED FEATURES.

Feature	Filtering Rule	Type	Description	Statistics / Distribution
trip_distance	> 0	Numeric (continuous)	Total distance of the trip in miles	2.78 ± 3.33, [0.01–99.9]
fare_amount	> 0	Numeric (continuous)	Fare charged to the passenger (excluding tip)	11.80 ± 9.47, [0.01–360.0]
passenger_count	$1 \leq x \leq 6$	Nominal (categorical)	Number of passengers in the taxi	Mode: 1, 70.6%
pickup_hour	derived	Numeric (cyclical hour)	Hour of the day when the trip started from pickup_datetime	Mode: 19:00 (7.1%), Evening hours dominant, Range: [0–23]
tip_amount	none	Numeric (continuous)	Additional tip paid by the passenger	1.19 ± 1.60, [0.0–100.0]
tolls_amount	none	Numeric (continuous)	Toll charges incurred during the trip	0.38 ± 1.41, [0.0–27.5]
total_amount	none	Numeric (continuous)	Sum of fare, tip, and tolls	13.38 ± 10.22, [0.01–365.4]
day_of_week	none, derived	Nominal (categorical)	Day of the week (0=Monday, ..., 6=Sunday) from pickup_datetime	0: 10.65%, 1: 11.76%, 2: 13.97%, 3: 16.39%, 4: 16.27%, 5: 19.05%, 6: 11.91%
is_weekend	none, derived	Nominal (categorical)	Binary, indicates whether the trip occurred on a weekend if day_of_week ≥ 5	0: 69.04%, 1: 30.96%
log_trip_distance	none, derived	Numeric (continuous)	Log-transformed trip distance $\log(\texttt{trip\_distance} + 10^{-5})$ to avoid log(0)	0.98 ± 0.58, [-4.61–4.60]
fare_per_mile	none, derived	Numeric (continuous)	Fare normalized by distance	8.06 ± 29.91, [0.0–998.1]
trip_duration	$60 \leq x \leq 7200$	Numeric (continuous)	Trip duration (sec), used as the prediction target	740.52 ± 545.21, [60.0–7150.0]

B. Correlation and Mutual Information Analysis

To assess feature relevance, Pearson correlation coefficients (PCC) were computed for continuous variables, while mutual information (MI) was calculated for all features. PCC captures linear dependence, whereas MI detects both linear and nonlinear associations and supports categorical inputs [3].

As shown in Table II, fare_amount, total_amount, and log_trip_distance yield the highest PCC values ($\rho > 0.791$), confirming their strong linear association with trip duration. Tip_amount and tolls_amount show moderate correlation, while fare_per_mile exhibits negligible linear dependence ($\rho = -0.064$) but a relatively high MI score (0.548), revealing a nonlinear contribution.

Discrete and temporal features such as pickup_hour, day_of_week, is_weekend, and passenger_count were excluded from PCC but yield non-zero MI values. Notably, pickup_hour (MI = 0.096) captures time-of-day effects not visible through linear correlation. Overall, MI values ranged from 0.009 to 1.404, offering a unified scale for evaluating both dominant and subtle relationships.

The ranking in Table II highlights a consistent pattern: spatial and fare-related attributes dominate in both linear and nonlinear relevance metrics, confirming their interpretable impact on trip duration. Temporal and normalized features, though ranked lower, may encode context-specific or interaction effects not captured by global statistics. Given the diversity of models considered, all features were retained to preserve potential nonlinear or conditional contributions across learners.

TABLE II
PEARSON CORRELATION AND MUTUAL INFORMATION SCORES WITH RESPECT TO *trip_duration*

Feature	PCC (ρ)	MI
fare_amount	0.859	1.404
total_amount	0.834	1.164
log_trip_distance	0.827	0.696
trip_distance	0.791	0.697
fare_per_mile	-0.064	0.548
tip_amount	0.400	0.403
tolls_amount	0.337	0.302
pickup_hour	-	0.096
day_of_week	-	0.035
is_weekend	-	0.026
passenger_count	-	0.009

C. Machine Learning Models and Evaluation Metrics

This study employs a suite of regression algorithms to model the relationship between trip attributes and taxi ride duration. The models encompass linear, kernel-based, ensemble, and neural network methods, each defined by unique loss functions, learning procedures, and representational assumptions.

Let $D_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{n}$ be the training dataset, where $x_i \in \mathbb{R}^p$ represents the feature vector of trip-related attributes (e.g., distance, fare, time) and $y_i \in \mathbb{R}$ denotes the corresponding trip duration in seconds. Similarly, let $D_{\text{test}} = \{(x_j, y_j)\}_{j=1}^{m}$ denote the test set, and $\bar{y} = \frac{1}{m}\sum_{j=1}^{m} y_j$ be the empirical mean duration. The objective is to learn a predictive function $f : \mathbb{R}^p \to \mathbb{R}$ such that $f(x_j) \approx y_j$, i.e., the estimated duration closely approximates the true value. The following ML models were trained to approximate this mapping function f.

LR [4] models trip duration as a linear combination of input features: $\hat{y} = x^\top \beta + \beta_0$. Parameters are estimated by minimising the residual sum of squares: $L(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^\top \beta - \beta_0)^2$.

SVR [5] is a kernel-based method minimizing prediction error within an ε-insensitive margin: $\min_{w,b,\xi,\xi^*} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$ s.t. $\begin{cases} y_i - \langle w, \phi(x_i)\rangle - b \leq \varepsilon + \xi_i \\ \langle w, \phi(x_i)\rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$
where $\phi(\cdot)$ is a kernel mapping and C controls error tolerance.

RF [6] constructs an ensemble of T decision trees trained on bootstrapped samples, with random feature subsets. The final prediction is the mean of all tree outputs: $\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f^{(t)}(x)$

GB [7] builds an additive model in a stage-wise manner: $f_M(x) = \sum_{m=1}^{M} \rho_m h_m(x)$ where $h_m(x)$ is the m-th weak learner trained to approximate the negative gradient of the loss function.

XGBoost [8] is a highly optimized gradient boosting algorithm using second-order Taylor approximation of the loss function and regularization: $\mathcal{L}^{(t)} = \sum_i \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2\right] + \Omega(f_t)$ where g_i and h_i are first and second-order gradients.

MLP [9] is a feed-forward neural network composed of an input layer, one or more hidden layers, and an output layer. Each neuron computes a non-linear activation over weighted inputs. The network is trained using backpropagation and stochastic gradient descent to minimize the mean squared error.

To evaluate the performance of the trained regression models, three commonly adopted metrics were used: MAE, RMSE, and the Coefficient of Determination (R^2). These metrics provide complementary insights into the prediction quality, error sensitivity, and variance explanation of the models [10].

MAE $= \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$ quantifies the average magnitude of the prediction errors without considering their direction. It is robust to outliers compared to RMSE and retains the same unit as the target variable. It provides an interpretable measure of the expected absolute deviation between predicted and actual values.

RMSE $= \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$ penalizes larger errors more than MAE due to the squaring of residuals, making it more sensitive to outliers. It is particularly suitable when large deviations are undesirable, and the goal is to optimize for both accuracy and consistency. Like MAE, it is expressed in the same unit as the target variable.

$\boldsymbol{R^2} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$ represents the proportion of variance in the dependent variable that is predictable from the independent variables. A score of 1 indicates perfect prediction, while a score of 0 indicates that the model performs no better than a mean-based baseline. Negative values can occur if the model performs worse than the baseline. This metric is scale-independent, facilitating the comparison of model performance across different datasets.

Together, these metrics provide a robust basis for comparing model performance from multiple perspectives: average prediction accuracy (MAE), sensitivity to large errors (RMSE), and explanatory power (R^2).

D. Model Training and Optimization

All models were trained using a standardised experimental pipeline to ensure consistency and comparability across the selected regression algorithms. The dataset of NYC Yellow Taxi trips was split into 80% training and 20% testing subsets. Stratified sampling was not applied, as the target variable—trip duration—is continuous and not class-based. This random split preserved the natural distribution of durations and ensured adequate representation of all trip types across both subsets.

Model training was implemented using `scikit-learn` (v1.4.1) and `XGBoost` (v1.7.6) in Python 3.10 [11]. Experiments were conducted on a workstation with an Intel Core i7 CPU and 32 GB RAM. For each model, a grid search with 3-fold cross-validation was performed on the training set to identify the optimal hyperparameters based on minimum validation RMSE. After hyperparameter tuning, the best configuration was selected and retrained on the full training set, then evaluated on the hold-out test set using MAE, RMSE, and R^2 metrics.

The hyperparameter space explored and the selected configurations are summarized in Table III.

TABLE III
OPTIMAL HYPERPARAMETERS FOR THE ML MODELS.

Model	**Optimal Hyperparameters**
LR	Applied without regularization
SVR	kernel = rbf, C = 1.0, epsilon = 0.2
RF	n_estimators=100, max_depth=15, min_samples_split=2
GB	n_estimators=100, learning_rate=0.1, max_depth=5
XGBoost	n_estimators=100, learning_rate=0.1, max_depth=5, verbosity=0
MLP	hidden_layer_sizes=(100,), max_iter=300, activation='relu', solver='adam'

This experimental setup ensured methodological rigor across all models, enabling a fair and meaningful comparative analysis of their performance in predicting taxi trip duration. No feature selection or dimensionality reduction was applied, as all cleaned and engineered features were retained to preserve information and support generalization.

III. RESULTS AND DISCUSSION

This section presents the experimental results and analyses the predictive performance of the six ML models. We report their accuracy in estimating trip duration using standard regression metrics. The discussion highlights differences in model behavior, generalization capacity, and suitability for real-world deployment.

A. Performance Evaluation

The six regression models were evaluated on a hold-out test set using three standard regression metrics: MAE, RMSE, and R^2. The evaluation results are reported in Table IV.

Among all models, the RF achieved the best predictive accuracy, with an MAE of 53.83 seconds, RMSE of 143.38 seconds, and an R^2 score of 0.929, indicating that it explains over 92% of the variance in trip duration. This result highlights the model's ability to effectively capture non-linear dependencies and complex interactions among the features in the dataset.

The XGBoost and GB models also performed competitively, with MAE values of 114.35 and 116.70 seconds, respectively, and R^2 scores above 0.71. These models demonstrate strong generalization capability and are well-suited for structured regression tasks involving tabular features.

On the other hand, the LR and SVR models showed lower performance, with MAE values exceeding 130 seconds and R^2 scores below 0.65. These results reflect the limited

capacity of these models to account for the nonlinear and context-dependent nature of trip durations in a dynamic urban environment.

The MLP Regressor achieved intermediate performance, with an MAE of 127.94 seconds and an R^2 score of 0.657. While capable of modeling non-linear patterns, it was more sensitive to initialization and required careful tuning.

The observed performance differences suggest that ensemble-based models, particularly Random Forest and XGBoost, are better equipped to handle the heterogeneous and interaction-rich nature of the taxi trip data. The relatively low errors and high R^2 scores achieved by these models indicate their suitability for real-world deployment in applications such as travel-time estimation, taxi fleet management, and route optimization.

TABLE IV
EVALUATION RESULTS OF THE ML MODELS.

Model	**MAE (s)**	**RMSE (s)**	$\boldsymbol{R^2}$
LR	131.23	189.67	0.646
SVR	157.88	216.45	0.584
RF	53.83	143.38	0.929
GB	116.70	171.12	0.715
XGBoost	114.35	168.40	0.722
MLP	127.94	186.02	0.657

B. Interpretability and Deployment Considerations

The RF model clearly outperformed all other approaches, offering the highest accuracy and generalization ability with minimal tuning. Its ensemble architecture effectively captured complex relationships between spatial, temporal, and economic features, making it a strong candidate for practical deployment in mobility analytics and intelligent transportation systems. The model's ability to handle feature heterogeneity and its built-in resistance to overfitting contributed to its dominance across all evaluation metrics.

Other ensemble methods, such as XGBoost and GB, also performed competitively, though with higher training complexity and slightly lower predictive power. These models remain viable for deployment scenarios that can accommodate additional optimization overhead and require fine-grained control over model regularization. LR and SVR underperformed due to their limited flexibility in modelling nonlinear interactions. While computationally efficient and interpretable, their accuracy is insufficient for operational scenarios requiring real-time travel-time predictions at the city scale. The MLP regressor, despite its theoretical capacity for non-linear learning, showed marginal improvement over linear models. Its sensitivity to hyperparameter tuning and lower stability make it less suitable for mid-scale, tabular mobility datasets without extensive optimization.

Overall, ensemble tree-based models, especially RF, offer the most favourable balance between accuracy, robustness, and ease of deployment for big data regression tasks in urban mobility domains.

IV. CONCLUSION

This study investigated the application of supervised ML models to predict taxi trip duration using a large-scale, real-world mobility dataset from New York City. After a rigorous preprocessing and feature engineering pipeline, six regression models, namely LR, SVR, RF, XGBoost, GB and MLP were trained and evaluated using standard metrics: MAE, RMSE, and R^2. The experimental results demonstrated that ensemble-based methods, particularly RF, significantly outperformed linear, kernel-based, and neural models in terms of both accuracy and robustness.

The findings confirm that tree-based ensembles are well-suited for modeling structured urban transportation data, offering a favorable balance between predictive performance and deployment feasibility. Simpler models, although more interpretable and computationally efficient, failed to capture the non-linear dynamics of urban mobility patterns. The study also highlighted practical trade-offs between model complexity, training stability, and generalization capacity.

Future research directions include integrating external data sources, such as weather conditions, public events, and real-time traffic feeds, to further enhance prediction accuracy. Additionally, exploring model distillation, interpretable surrogate modeling, and reinforcement learning for adaptive routing and dynamic pricing represents promising avenues for operational deployment in smart transportation systems.

REFERENCES

[1] Y. Wang, F. Currim, and S. Ram, "Deep learning of spatiotemporal patterns for urban mobility prediction using big data," *Information Systems Research*, vol. 33, no. 2, pp. 579–598, 2022.

[2] M. Abdollahi, T. Khaleghi, and K. Yang, "An integrated feature learning approach using deep learning for travel time prediction," *Expert Systems with Applications*, vol. 139, p. 112864, 2020.

[3] H. Gong, Y. Li, J. Zhang, B. Zhang, and X. Wang, "A new filter feature selection algorithm for classification task by ensembling pearson correlation coefficient and mutual information," *Engineering Applications of Artificial Intelligence*, vol. 131, p. 107865, 2024.

[4] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, "Linear regression," in *An introduction to statistical learning: With applications in python*. Springer, 2023, pp. 69–134.

[5] F. Zhang and L. J. O'Donnell, "Support vector regression," in *Machine learning*. Elsevier, 2020, pp. 123–140.

[6] S. J. Rigatti, "Random forest," *Journal of Insurance Medicine*, vol. 47, no. 1, pp. 31–39, 2017.

[7] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, pp. 1937–1967, 2021.

[8] J. Chen, F. Zhao, Y. Sun, and Y. Yin, "Improved xgboost model based on genetic algorithm," *International Journal of Computer Applications in Technology*, vol. 62, no. 3, pp. 240–245, 2020.

[9] R. Kruse, S. Mostaghim, C. Borgelt, C. Braune, and M. Steinbrecher, "Multi-layer perceptrons," in *Computational intelligence: a methodological introduction*. Springer, 2022, pp. 53–124.

[10] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation," *Peerj computer science*, vol. 7, p. e623, 2021.

[11] J. Brownlee, *XGBoost With python: Gradient boosted trees with XGBoost and scikit-learn*. Machine Learning Mastery, 2016.