NYC Taxi Fare Prediction Using Temporal and Spatial Data

Simone Bevilacqua
University of Illinois at Chicago
Graduate student
sbevi@uic.edu

Niccolò Brembilla
University of Illinois at Chicago
Graduate student
nbrem@uic.edu

Brian Rosca
University of Illinois at Chicago
Graduate student
brosca2@uic.edu

Tommaso Tognoli
University of Illinois at Chicago
Graduate Student
ttogn@uic.edu

***Index Terms*—Taxi Fare Prediction, Data Preprocessing, Regression Analysis, Predicting Modeling, Neural Networks, Backpropagation**

I. Introduction

Accurately predicting taxi fares in urban environments presents a significant challenge in transportation planning and pricing. Our project focused on developing and comparing multiple machine-learning approaches to predict New York City taxi fares using the 2019 yellow taxi ride dataset. We implemented traditional regression models, ensemble methods, and neural networks, incorporating temporal, spatial, and environmental data to create a comprehensive prediction system.

The complexity of urban taxi pricing stems from numerous variables influencing fare calculations, including time of day, day type (weekday, weekend, or holiday), weather conditions, and demand fluctuations. Traditional fare estimation methods often struggle to account for these dynamic factors, leading to uncertainty for riders and inefficiencies in urban transportation systems.

Our research leveraged the geographical specificity of New York City to create a highly contextualized prediction model. We could integrate precise, location-specific data sources. We preprocessed and merged weather data from NYC's weather stations to match exact weather conditions with each ride's timestamp, providing a granular environmental context. Similarly, we incorporated a comprehensive holiday dataset that identified holidays and distinguished between weekdays and weekends, allowing our models to capture temporal patterns in fare variations. This preprocessing approach ensured that each ride in our dataset was enriched with accurate, time-specific environmental and temporal conditions that directly influenced taxi fares.

Our research aimed to address these challenges by developing multiple predictive models that could accurately estimate taxi fares while accounting for these various influences. Through careful preprocessing and feature engineering, we created a robust dataset that formed the foundation for our analysis, with each ride containing precise contextual information about its operating conditions.

The project employed three main approaches to fare prediction:

- Traditional regression models that established a baseline for prediction accuracy
- Ensemble methods that combined multiple predictors to improve performance
- A deep neural network architecture featuring multiple hidden layers with batch normalization and dropout for enhanced generalization

This multi-model approach allowed us to compare different methodologies and identify the most effective techniques for fare prediction. We could evaluate the trade-offs between model complexity and prediction accuracy by implementing both traditional machine learning methods and modern neural network architectures. Our work builds upon previous research in transportation pricing prediction while addressing limitations identified in earlier studies. Unlike some prior work that relied on limited timeframes or simplified feature sets, our analysis incorporated a full year of data and a comprehensive set of external factors, aiming to create more robust and generalizable models for real-world applications. The impact of our preprocessing methodology on model performance and the comparative analysis of different modeling approaches will be discussed in detail in subsequent sections.

II. Problem Statement

The challenge of predicting taxi fares in New York City presented several complex data preprocessing considerations that significantly impacted our model development approach. While the base prediction task appeared straightforward, the underlying data complexity required careful consideration of multiple preprocessing strategies to create a reliable prediction model.

Our primary technical challenge centered on determining the optimal approach to handle various fare components and outliers in the dataset. Each preprocessing iteration revealed different aspects of the data that required attention:

Initially, we focused on basic data cleaning by removing invalid entries, such as trips with zero distance or negative fares. However, this approach revealed a more nuanced challenge regarding treating total fare amounts versus individual fare components. Our second preprocessing iteration addressed this by subtracting tip amounts from total fares, acknowledging that tips represent a discretionary (and, thus, uncontrollable) component that could skew our prediction models.

The identification and handling of outliers proved particularly challenging. While extreme values in fare amounts could represent legitimate airport trips or special circumstances, they could also indicate data entry errors or unusual situations that might not generalize well for our models. Our third preprocessing iteration addressed this by implementing a fare ceiling of $350 and utilizing the Isolation Forest algorithm to detect and remove statistical outliers.

The temporal and spatial dimensions of our dataset introduced additional complexity. We needed to integrate weather conditions and holiday information effectively while maintaining data integrity. This required careful consideration of how to handle missing values and edge cases, particularly when merging multiple data sources. For example, weather data needed to be precisely matched with ride timestamps, and holiday designations needed to account for both official holidays and weekend variations in fare patterns.

These preprocessing challenges directly influenced our modeling approach, as each cleaning decision impacted the balance between maintaining valuable data patterns and removing noise that could compromise model performance. Our iterative preprocessing approach aimed to systematically address these challenges while preserving the essential characteristics of NYC taxi fare patterns.

III. Data Analysis & Preprocessing Methodology

Our approach to preparing the NYC taxi fare dataset evolved through three distinct iterations, each addressing specific data quality and model performance concerns. This iterative refinement process was crucial in developing a robust foundation for our prediction models.

A. Initial Data Assessment

The raw dataset contained several components that required careful consideration:

- Temporal features, including pickup and dropoff times
- Spatial information through pickup and dropoff location IDs
- Fare components including base fare, tips, taxes, and surcharges
- Trip metrics such as distance and duration

B. Preprocessing Evolution

Our preprocessing strategy evolved through three major iterations:

1) First Iteration - Basic Cleaning: The initial preprocessing phase focused on fundamental data quality issues:

- Removed trips with zero or negative distances
- Eliminated negative fare amounts
- Converted time information into useful features (hour of day, day type)
- Integrated weather and holiday data
- Applied Isolation Forest for basic outlier detection

2) Second Iteration - Fare Component Analysis: The second iteration refined our treatment of fare components:

- Separated tip amounts from total fares
- Maintained only essential fare components
- Enhanced location zone processing
- Refined the categorical encoding of service zones

3) Third Iteration - Advanced Outlier Treatment: The final preprocessing phase introduced more sophisticated data cleaning:

- Implemented a $350 fare ceiling to remove extreme outliers
- Enhanced weather data integration
- Refined holiday and weekend categorization
- Generated visualization of temporal patterns
- Further refined zone-based features

C. Feature Engineering

Through these iterations, we developed several derived features:

- Time-based features: trip duration, times of the day
- Day type indicators: weekday (1), weekend (2), holiday (3)
- Weather conditions mapped to exact ride times
- Service zone classifications: Airports (1), Boro Zone (2), Yellow Zone (3)

IV. Regression Models

Our analysis employed multiple regression approaches to predict NYC taxi fares, with particular attention to how different preprocessing strategies affected model performance. We implemented and evaluated several algorithmic approaches, ranging from simple linear models to more sophisticated ensemble methods.

A. Model Implementation

Our implementation included several key approaches. We utilized linear regression with both standard and one-hot encoded versions as baseline models. Principal Component Analysis served as our dimensionality reduction technique with varying component numbers. K-Nearest Neighbors provided a non-parametric approach, while our tree-based regressors were implemented with three distinct criteria: Squared Error, Friedman MSE, and Poisson.

We also analyzed the correlation between different features (reported in the picture above) to understand what models could work best. Since there is a robust correlation between some particular features and our predicted value (i.e.,

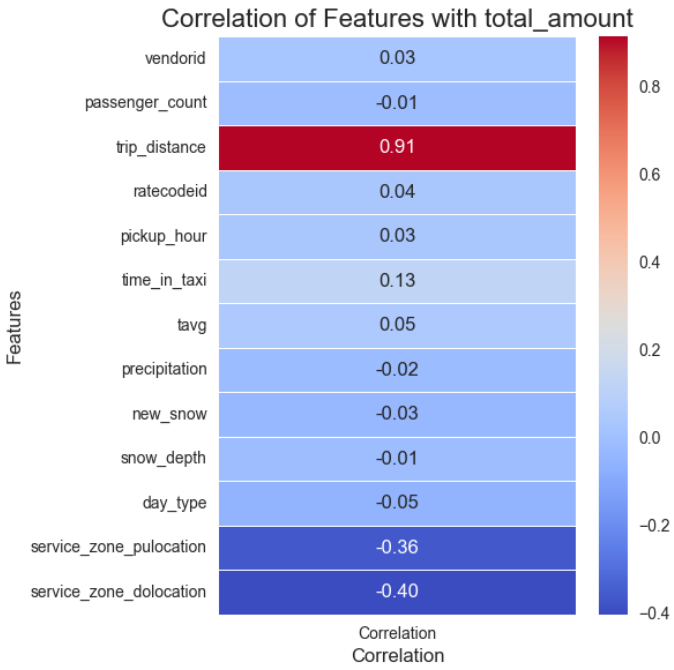

Fig. 1. Performance metrics across different numbers of PCA components showing R² Score, MSE, and MAE trends.

"trip_distance" seems to be the most crucial feature for the prediction), we thought that even simple models could capture them and give us acceptable results.

B. Linear Regressors

We started by fitting a simple Linear Regressor. We then tried to get better results by using one-hot encoding on categorical features (i.e. "vendorid") to see if the model would perform better. We noticed that there was no particular improvement between the two, and this could be explained by the fact that all the categorical features had a very low correlation to the prediction, so they would not improve our results by a lot. In general, the two models were acceptable with R2 around 0.85, but with an elevated MAE (around 1.86).

C. PCA (Principal Component Analysis)

Because the correlation matrix was so skewed towards certain features, we tried to apply PCA to the dataset before fitting the model.

As shown in Figure 2, our PCA analysis revealed several critical insights through the three key performance metrics. In fact, a sharp improvement was observed when increasing from 2 to 4 components, reaching approximately 0.85 and maintaining this level through 9 components. The MSE dropped dramatically from about 60 with 2 components to approximately 10 with 4 components, while the MAE followed a similar pattern, stabilizing around four.0 after 4 components. These results suggest that 4-5 components capture the essential variance in our data, with minimal gains from additional components. In general, when fitting the model on the new reduced dataset, the predictions did not deteriorate too much. In particular, as little as 4 components were enough to achieve similar results as the basic Linear Regressor, and this can be explained by the fact that only few features actually influence the output.

D. Comparative Model Performance

Figure 3 presents a comprehensive comparison of our models across three key performance metrics. The tree-based regressors consistently demonstrated superior performance across all metrics. Most notably, in terms of Mean Squared Error, linear models and their one-hot encoded variants showed similar MSE values of approximately 9.97, while the PCA implementation resulted in a slightly higher MSE at 10.68. KNN demonstrated improved performance with an MSE of 4.62, though the tree-based regressors excelled with MSE values around 2.60, representing a significant improvement over simpler models.

The progression in Mean Absolute Error performance closely mirrored the MSE results. Linear models and PCA showed comparable MAE values around 1.86-1.96, while KNN achieved an improved MAE of 0.82. The tree-based regressors again demonstrated superior performance with consistent MAE values of 0.70.

In terms of R² scores, while all models showed strong performance with values above 0.80, the tree-based regressors achieved the highest values at 0.96, compared to approximately 0.85 for linear models and PCA implementations.

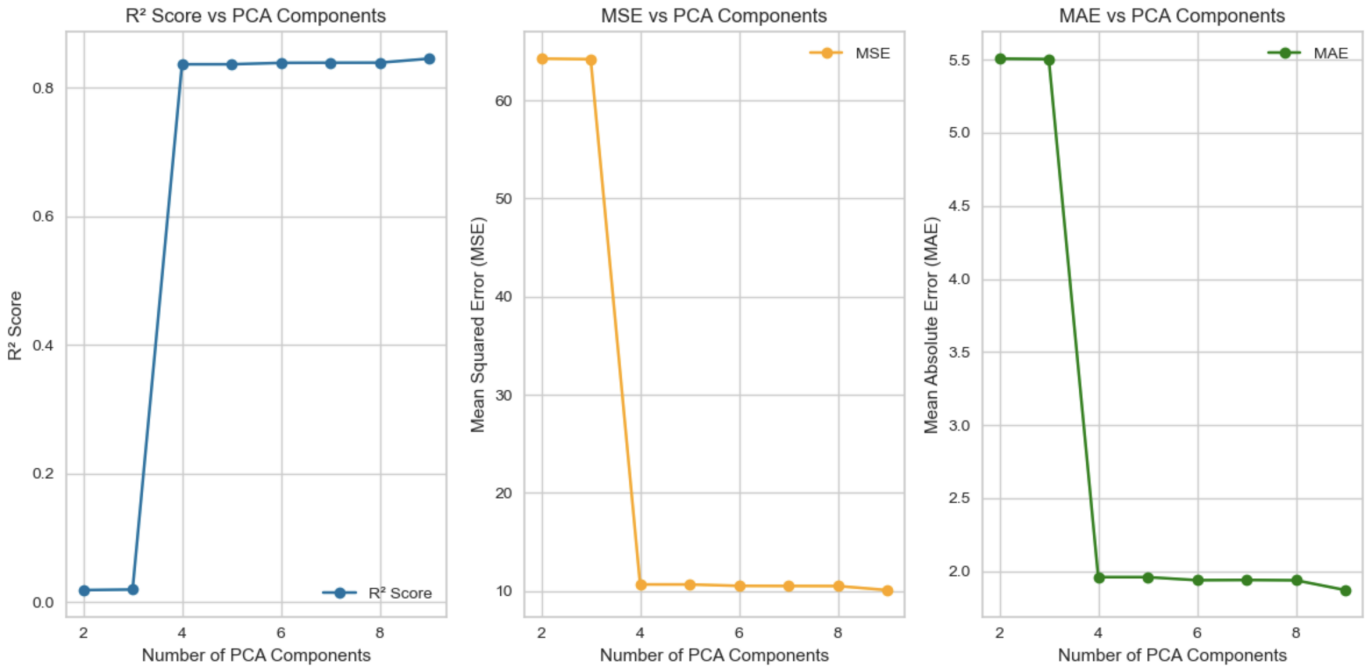

Fig. 2. Performance metrics across different numbers of PCA components showing R² Score, MSE, and MAE trends.

Fig. 3. Performance comparison of different regression models across MSE, MAE, and R² metrics.

E. Key Findings

The results illustrated in Figures 2 and 3 demonstrate that tree-based regressors consistently outperformed other approaches across all metrics. This superiority can be attributed to their ability to capture non-linear relationships in the preprocessed data, particularly the complex interactions between temporal, spatial, and weather-related features. The preprocessing steps, especially the careful handling of outliers and categorical variables, appeared to particularly benefit the tree-based models, as evidenced by their superior performance metrics.

V. Ensemble Models

Our analysis extended to ensemble methods, which combine multiple base learners to improve prediction accuracy and reduce overfitting. We implemented four distinct ensemble approaches, each offering unique advantages in handling the complexities of taxi fare prediction.

A. Random Forest

The Random Forest model demonstrated exceptional performance with an MSE of 1.25 and an R² score of 0.98, showing particularly strong prediction accuracy in the common fare ranges (10-40 USD). The model's high R² score indicates its robust ability to capture the complex relationships between our preprocessed features and fare amounts.

B. AdaBoost

The AdaBoost implementation showed different characteristics, with an MSE of 8.87 and an R² score of 0.86. While the model demonstrated good overall performance, it showed increased scatter in predictions, particularly for higher fare values, suggesting some difficulty in capturing extreme cases.

C. Gradient Boosting

The Gradient Boosting model achieved excellent results with an MSE of 2.20 and an R² score of 0.97. The scatter plot reveals a tighter clustering of predictions around the ideal fit line compared to AdaBoost, particularly in the mid-range fare values. This model demonstrated strong capability in

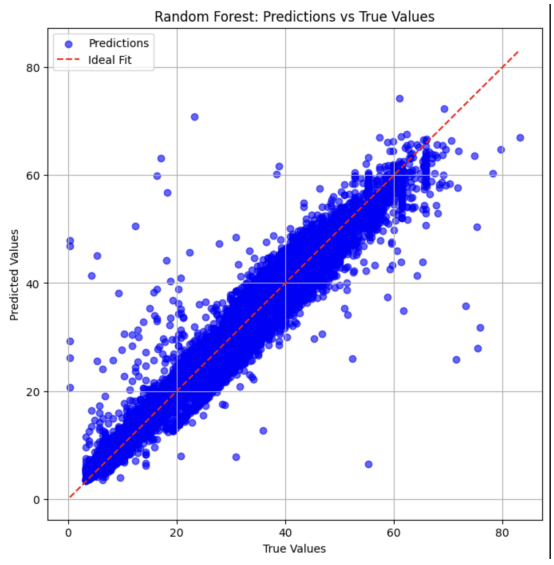

Fig. 4. Random Forest predictions versus true values, demonstrating strong correlation (R^2 = 0.98) with minimal scatter, particularly in the lower and middle fare ranges.

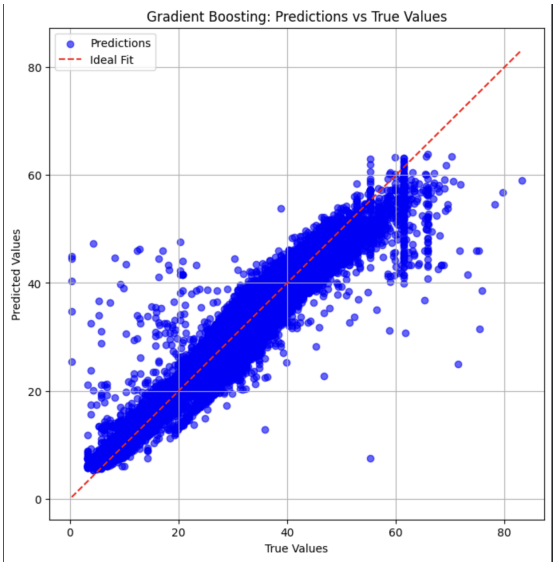

Fig. 6. Gradient Boosting predictions versus true values, showing improved performance over AdaBoost with tighter clustering around the ideal fit line (R^2 = 0.97).

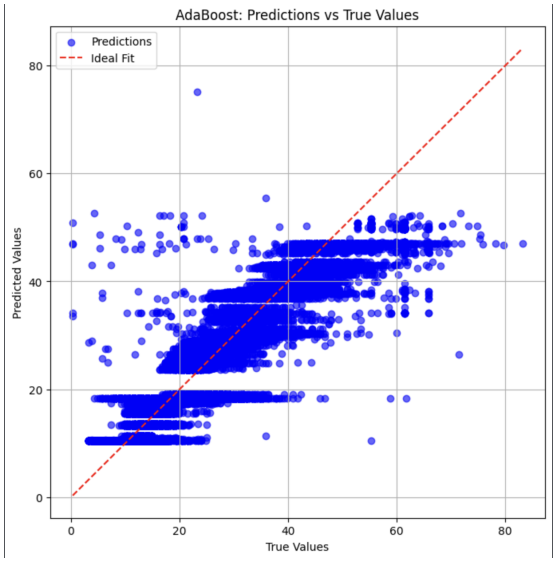

Fig. 5. AdaBoost predictions versus true values, showing increased scatter compared to Random Forest but maintaining reasonable prediction accuracy (R^2 = 0.86).

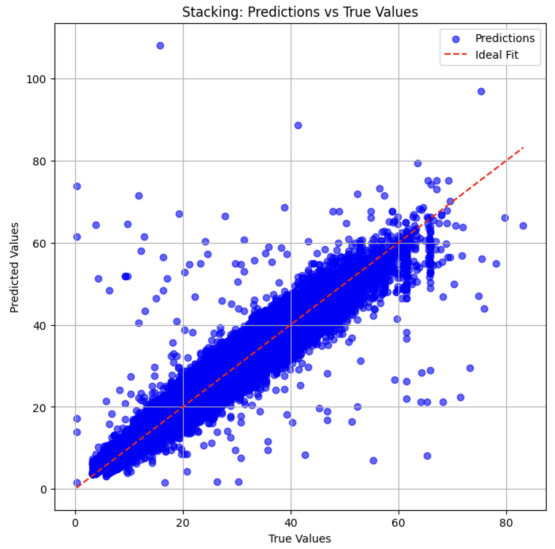

Fig. 7. Stacking ensemble predictions versus true values, showing consistent performance across the fare range with some increased scatter at higher values (R^2 = 0.96).

handling the various feature interactions introduced through our preprocessing steps.

D. Stacking

The Stacking ensemble approach used as base models a Decision Tree Regressor and an Elastic Net, achieving an MSE of 2.44 and an R^2 score of 0.96. This method demonstrated robust performance by leveraging the predictive capabilities of multiple models, though with slightly higher variance in predictions compared to the Gradient Boosting approach.

E. Comparative Model Performance

A comprehensive comparison of our ensemble methods reveals distinct performance characteristics across different metrics. To visualize these differences effectively, we compiled the key performance indicators for each model.

The comparative analysis shown in Figure 8 highlights several key findings.

First, the Random Forest model exhibited the lowest error metrics, with an MSE of 1.25 and MAE of 0.54, significantly outperforming other approaches in prediction accuracy. This superior performance can be attributed to its ability to han-

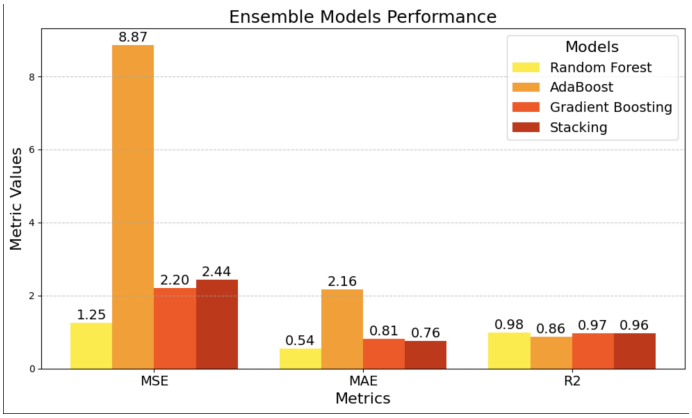

Fig. 8. Comparative performance metrics across ensemble models showing MSE, MAE, and R^2 scores. Random Forest demonstrates superior performance in error metrics while maintaining competitive R^2 scores.

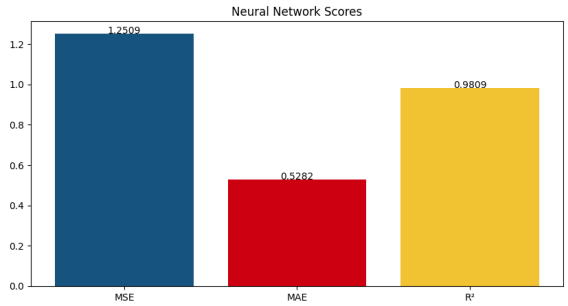

Fig. 9. Performance metrics of the FFNN showing MSE, MAE, and R^2 scores. The Neural Network demonstrates equal performance to the best ensemble model.

dle the complex feature interactions introduced through our preprocessing steps.

Second, while AdaBoost showed the highest error rates (MSE: 8.87, MAE: 2.16), it maintained a respectable R^2 score of 0.86, suggesting that despite higher absolute errors, it still captured the general trends in the data effectively.

Finally, both Gradient Boosting and Stacking demonstrated strong middle-ground performance, with MSE values of 2.20 and 2.44 respectively, and R^2 scores above 0.96. These models proved particularly effective at balancing prediction accuracy with model complexity.

This comprehensive evaluation suggests that while each ensemble method offered unique advantages, the Random Forest model provided the most consistent and accurate predictions across our evaluation metrics. Its superior performance validates our preprocessing approach, particularly in handling the temporal and spatial features of our taxi fare dataset.

VI. Feed-Forward Neural Network (FFNN)

Among the models we evaluated, the Feed-Forward Neural Network (FFNN) emerged as the most effective for taxi fare prediction, surpassing even the ensemble models in our study. The FFNN, with an architecture of 13 input nodes followed by layers of 512, 256, 128, 64, 32, and a single output node, achieved an MAE of 0.5282, an MSE of 1.2509, and an R^2 score of 0.9809. These metrics confirm the FFNN's superior ability to capture the intricate patterns in the dataset, making it particularly effective for accurate predictions.

The model was trained on a reduced dataset of 1.2 million rows, demonstrating that this architecture is well-suited for medium-sized datasets. Key design choices such as a combination of linear and ReLU activation functions, dropout layers (rate=0.4) to mitigate overfitting, and batch normalization after each layer played a critical role in enhancing model stability and generalization.

To optimize the training process, we employed the AdamW optimizer with a weight decay of $1 \cdot 10^{-5}$, complemented by a ReduceLROnPlateau scheduler with an initial learning rate of $1 \cdot 10^{-3}$. This configuration allowed the model to adaptively refine its learning rate during 50 training epochs. Despite the model's exceptional performance, the training process required significant computational time, ranging from 30 minutes to 1 hour, highlighting the trade-off between computational resources and prediction accuracy.

VII. Results and Analysis

The results of our regression and ensemble model evaluations reveal several key findings regarding the performance of various approaches to NYC taxi fare prediction. Across all models, preprocessing techniques, including feature encoding, dimensionality reduction, and outlier management, played a critical role in enhancing predictive accuracy.

Among regression models, tree-based regressors demonstrated the best performance, achieving R^2 scores as high as 0.96 and significantly outperforming linear and KNN models in terms of error metrics. Notably, the PCA-based regression models achieved stable performance with R^2 values around 0.85, suggesting they effectively captured essential variance with minimal components. However, tree-based methods excelled in leveraging the complex interactions among temporal, spatial, and weather features.

The ensemble methods highlighted the advantages of advanced algorithms in predictive modeling. The Random Forest model achieved the highest overall performance with an MSE of 1.25 and an R^2 score of 0.98. This performance can be attributed to its ability to handle non-linear relationships and interactions among features while minimizing overfitting through its ensemble nature. Similarly, the Neural Network model performed as effectively as Random Forest, benefiting from its capacity to model complex, non-linear relationships in high-dimensional data, capturing subtle interactions that other models struggled with.

Gradient Boosting closely follows with an R^2 of 0.97 and an MSE of 2.20, demonstrating strong robustness due to its iterative nature, which optimizes the ensemble by correcting errors sequentially. Stacking also performed strongly, combin-

ing diverse base models to balance prediction accuracy and model complexity.

In contrast, AdaBoost and linear regression models exhibited notable limitations. AdaBoost, with an R^2 of 0.86 and an MSE of 8.87, struggled with high-variance features, particularly outliers and extreme fare values, as it assigns higher weights to mispredicted data points, amplifying their influence. Similarly, linear regression models, including their one-hot encoded variants, were limited by their inherent assumption of linearity. This assumption led to suboptimal performance in capturing the non-linear and complex feature interactions in the dataset, yielding relatively high error metrics.

Overall, the analysis confirms that ensemble models like Random Forest and Gradient Boosting, as well as Neural Networks, are highly effective for this task. Their success arises from their ability to capture non-linear relationships, integrate complex feature interactions, and adapt to diverse fare ranges. By contrast, the weaker performance of AdaBoost and linear regression highlights the importance of model selection and the challenges of handling high-dimensional, non-linear data in predictive modeling.

While the Random Forest and Neural Network models demonstrated comparable performance, with R^2 scores of 0.98 and exceptional predictive metrics, their practical applicability diverges due to differences in computational efficiency and scalability. Random Forest offers a significant advantage in terms of training time, completing the process relatively quickly while still capturing non-linear relationships and complex feature interactions effectively. This efficiency makes it a more practical choice for applications requiring rapid deployment or iterative model updates.

In contrast, the Neural Network, despite its superior capacity to model intricate patterns, requires substantially more computational resources and time to train. The FFNN's architecture, designed for high-dimensional data, shows promise for future expansions, such as incorporating additional features like seasonality or dynamic temporal patterns. However, such enhancements would likely increase the model's training complexity and time requirements, potentially limiting its immediate usability in time-sensitive contexts.

Therefore, while the Neural Network's flexibility positions it as a robust solution for scenarios involving evolving feature sets, the Random Forest emerges as the more pragmatic choice for current applications, balancing accuracy with computational efficiency.

VIII. Future Works

The scope of this study highlights several promising directions for future research aimed at refining taxi fare prediction models and broadening their applicability. While our analysis demonstrated the effectiveness of ensemble methods and neural networks, there remains significant potential for optimization and exploration in key areas.

Future work could focus on integrating real-time data streams, such as live traffic updates and ride demand surges, into prediction models. This enhancement would allow for more context-aware and dynamic fare estimates, particularly under varying traffic conditions or during high-demand periods.

A critical avenue for future research lies in analyzing the profound impact of the COVID-19 pandemic on urban transportation patterns. Our current study utilized data from 2019, representing a pre-pandemic baseline of peak urban mobility. Extending this analysis to include data from 2020 through 2023 would provide valuable insights into:

- The immediate impact of lockdown measures on taxi usage and pricing dynamics
- Changes in ride-sharing preferences during various pandemic phases
- The evolution of urban transportation patterns during the recovery period
- The emergence of new pricing models adapted to post-pandemic realities
- Comparative analysis between pre-pandemic, pandemic, and post-pandemic transportation behaviors

This temporal expansion would not only capture the unprecedented disruption caused by COVID-19 but also help understand the resilience and adaptability of urban transportation systems during global crises. Furthermore, analyzing how different cities worldwide responded to and recovered from the pandemic would provide insights into cultural and regional variations in urban mobility patterns.

Exploring more diverse and comprehensive datasets remains another critical step. While this study focused on a single year of New York City taxi data, future research could analyze data spanning multiple years or incorporate information from different cities to better understand regional fare patterns and urban transportation dynamics. This cross-cultural analysis could reveal how different societies utilize and value taxi services, potentially leading to more adaptable and culturally aware pricing models.

Additionally, future studies could focus on deploying these models in real-world scenarios, emphasizing scalability and computational efficiency. Designing lightweight models for integration into mobile apps or ride-hailing platforms (e.g., Uber and Lyft) would bridge the gap between academic research and practical application.

By pursuing these research directions, particularly the analysis of pandemic-related changes, the potential of predictive modeling for taxi fares can be further realized, driving improvements in both fare accuracy and broader urban transportation systems. Understanding how global events like pandemics affect transportation patterns could also help develop more resilient and adaptive pricing models for future crisis scenarios.

IX. Novelty Statement

Our approach to taxi fare prediction significantly improves upon previous work in several key areas. For instance, the Cab Price Predictor with 96% Accuracy [3] uses basic models and does not include any data preprocessing or additional contextual features. While this model demonstrates high accuracy, it

is based on only one month of data and lacks the integration of other relevant data, such as weather conditions or holiday information, that could influence taxi fare predictions. Our model, by contrast, incorporates comprehensive data from a full year and includes advanced preprocessing techniques, such as outlier detection and feature engineering, along with more sophisticated features that improve the model's ability to make accurate predictions under diverse conditions.

Similarly, the Uber & Lyft Price Analysis (Traditional vs DNN) [1] uses preprocessing and deep learning models, but it is limited to only a single month of data. This short time frame restricts the model's ability to capture long-term trends or account for seasonal variations in ride-sharing prices. In contrast, our approach utilizes multi-year data, which helps the model better understand fluctuations in taxi operations, fare structures, and urban mobility patterns, offering more generalized predictions.

The Predict Trip Duration project [4] uses the same dataset we used but focuses exclusively on predicting trip duration rather than fare prediction. While trip duration is certainly an important factor in fare determination, our model goes beyond this by incorporating multiple contextual features—such as weather, taxi availability, and holiday data—into the fare prediction process, making it a more comprehensive and practical tool for real-world applications.

Finally, the NYC Taxi 2019 Dataset Analysis and Clustering project [2] applies clustering techniques (specifically KMeans) to the same dataset but does not focus on fare prediction. The clustering approach is useful for understanding the general structure of the dataset, but it falls short when it comes to generating accurate predictions for taxi fares.

By utilizing a more robust dataset, integrating additional data sources, and applying advanced predictive models, our work offers significant improvements over these prior studies in terms of accuracy, generalizability, and practical applicability.

X. Contributions

Our group was composed of four graduate students, and the contributions are detailed below:

- Simone Bevilacqua: worked on the ensemble models and the training of neural network.
- Niccolò Brembilla: worked on the preprocessing and the training neural network.
- Brian Rosca: worked on the preprocessing and the writing of the proposal/final report.
- Tommaso Tognoli: worked on the regression models and the preparation of the presentation.

References

[1] Khaled Ashraf. *Uber-Lyft Price Analysis: Traditional vs. DNN*. URL: https://www.kaggle.com/code/khaledashrafm3wad/uber-lyft-price-analysis-traditional-vs-dnn/notebook. (accessed: 05.12.2024).

[2] Azamatjon Khasanzoda. *NYC Taxi 2019 Dataset: Analysis and Clustering*. URL: https://www.kaggle.com/code/azamatjonkhasanzoda/nyc-taxi-2019-dataset-analysis-and-clustering. (accessed: 05.12.2024).

[3] Prince Raj. *Cab price predictor with 96% accuracy*. URL: https://www.kaggle.com/code/idiotprofessor/cab-price-predictor-with-96-accuracy. (accessed: 05.12.2024).

[4] Ismael Solace. *Predict trip duration*. URL: https://www.kaggle.com/code/ismaeldwikat/predict-trip-duration. (accessed: 05.12.2024).