Improving NYC Yellow Taxi Earnings through Predictive Modeling of Tips and Fares

Matty J. Maloni
Department of Electrical and Computer Engineering
University of Florida,
Gainesville, FL
mmaloni1@ufl.edu

***Abstract* – Maximizing taxi earnings depends on both fares and tips, with the latter shaped by factors beyond driver control. This study analyzes 2023 NYC Yellow Taxi data to identify determinants of fares and tips using Linear and Lasso Regression with cross-validation. Key predictors include trip distance, time, passenger count, and surcharges. Results show fares are highly structured, driven mainly by distance and fees, while tips are less predictable but influenced by timing, payment method, and pickup location. These insights highlight how predictive modeling can guide drivers in optimizing schedules and trip strategies to improve earnings.**

I. Introduction

A. *Industry Context*

In 2024, NYC yellow taxis provided just 40.7 million rides, a steep decline from 146.1 million trips in 2015 [1]. In less than a decade, ridership has contracted by more than 70%, signaling a shift in the city's transportation. It is no secret that the increased use of ridesharing apps like Uber and Lyft has significantly contributed to this decline. In 2015, High Volume For Hire Vehicles (FHVs) (like Uber and Lyft) completed 41.6 million trips in NYC. In 2024, they completed 239.5 million. This dramatic reversal (Fig. 1) illustrates how ridesharing has fundamentally reshaped the industry, reducing the customer base available to traditional taxi drivers.

B. *Impact on Drivers*

The decline in ridership has created severe financial pressure for NYC taxi drivers. At the industry's peak in the early 2010s, a single medallion—the license required to operate a yellow cab—sold for over $1.2 million[2]. With the rise of ridesharing, medallion values have collapsed to a fraction of that price, leaving many owners burdened by debt. Meanwhile, average driver earnings have stagnated at roughly $46,000 per year (before expenses such as insurance and fuel)[3]. Because fares are strictly regulated by the Taxi and Limousine Commission, drivers cannot raise prices to offset losses. As a result, maximizing tips, optimizing schedules, and making strategic trip choices have become essential for sustaining livelihoods in an increasingly competitive market.

C. *Research Goals*

To address this problem, this study applies predictive modeling to the New York City Yellow Taxi dataset to provide interpretable insights that can help drivers and companies identify profitable strategies.

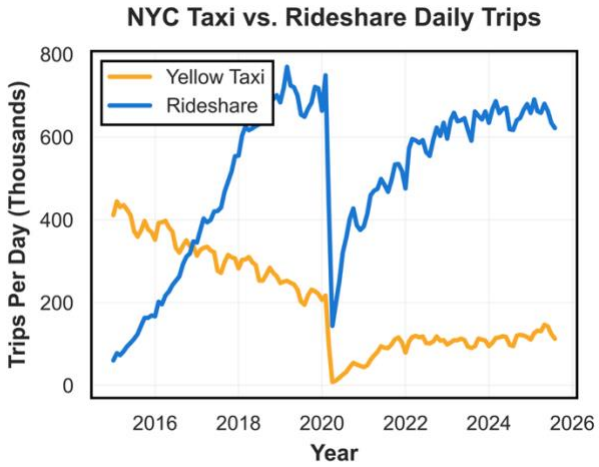

Figure 1: NYC Yellow Taxi vs High Volume FHVs

II. Data

A. *Dataset Introduction*

This study uses the publicly available NYC Yellow Taxi and Limousine Commission trip record dataset, which contains detailed information about individual yellow taxi rides. Each record includes trip times, locations, passenger counts, distances, and payment information. Crucially, the dataset reports both fare amount and tip amount, allowing us to study the structured component of driver income (fares) and the variable component (tips). The subset analyzed for this project covers trips from 2023, with approximately 10,000 rides sampled for analysis. It includes the following key variables: trip characteristics (pickup and dropoff times, passenger count, trip distance, and rate codes); financial variables (fare amount, tip amount, tolls, and surcharges including MTA tax, congestion surcharge, and airport fee); and categorical features (vendor ID, payment type, pickup and dropoff zones).

B. *Data Preprocessing*

Before modeling, several steps were performed to clean and prepare the raw dataset. Pickup and dropoff timestamps were converted into standard datetime objects to enable

consistent time-based calculations. Invalid records, such as trips with negative fares, zero distance, or missing passenger counts, were also removed to maintain data integrity.

Several columns were dropped entirely: mta_tax and improvement_surcharge were dropped because they were constant across all trips. The text-based pickup and dropoff location fields were discarded, with analysis relying instead on the precise ID-based location codes (PULocationID and DOLocationID). Store_and_fwd_flag was removed since it relates only to server communication and not to passenger or trip behavior.

Missing financial fields, including fare_amount, extra, tolls_amount, congestion_surcharge, and airport_fee, were imputed with a value of 0, reflecting the fact that these surcharges are only applied under certain conditions. In addition, passenger_count was imputed using the median value to maintain realistic distributions of ride occupancy, while ratecodeid was imputed with the most frequent category to preserve consistency across trips.

C. *Feature Engineering*

From the cleaned dataset, several new variables were derived to capture temporal patterns, fare structures, and passenger behavior more effectively. Trip duration was computed in minutes from the pickup and dropoff timestamps (after preprocessing ensured valid times) and retained as a continuous variable. After these derivations, the raw pickup and dropoff datetime fields were dropped to avoid redundancy.

Two categorical features were extracted from the pickup timestamp: day of week (Monday through Sunday) and time of day, grouped into six bins (Early Morning, Morning, Afternoon, Evening, Night, and Late Night). Six bins were used instead of the recommended 4 because using only the Morning, Afternoon, Evening, and Night bins resulted in a ~35% data loss during the non-included hours. These variables allowed the models to account for differences in travel and tipping behavior across time periods.

A new financial feature, pre-tip total amount, was created by summing all fare components except the tip (fare_amount + extra + tolls_amount + congestion_surcharge + airport_fee). This variable isolates the structured component of trip revenue from the variable component represented by passenger tipping. In addition, other features exhibiting high collinearity were dropped to improve model stability and reduce overfitting.

C. *Feature Transformation*

After feature engineering, the dataset contained a mix of continuous, categorical, and high-cardinality variables. To ensure compatibility with regression models and to make coefficients interpretable, additional transformations were applied.

1) Standard Scaling. Continuous features such as trip_distance, total_time, and pre_tip_total_amount were standardized to zero mean and unit variance. This step is especially important for Lasso Regression, since its penalty term is scale-sensitive and would otherwise overweight features with larger numerical ranges. Scaling also ensures that regression coefficients can be more directly compared across variables.

2) One-Hot Encoding (OHE). Categorical variables with a limited number of categories were converted into binary indicator vectors. These included vendorid, payment_type, day_of_week, time_of_day, and surcharge-related flags (congestion_surcharge and airport_fee). OHE prevents the model from assuming spurious numerical relationships between categories (treating vendor 2 as "larger" than vendor 1) and allows it to assign separate coefficients to each group.

3) Target Encoding for High-Cardinality Locations. Pickup and dropoff location IDs (PULocationID and DOLocationID) have hundreds of unique values, making one-hot encoding impractical. Instead, a target encoding approach was used: each location ID was replaced with the mean fare amount observed for that location within the training folds. This transformation condenses hundreds of categories into meaningful continuous predictors while avoiding high-dimensional sparsity. To prevent data leakage, target encoding was recomputed inside each cross-validation split, ensuring that encodings reflected only the training data available to the model at that stage.

III. Exploratory Data Analysis

A. *Correlation Analysis*

Pearson's correlation matrix (Fig. 2) shows strong dependencies between fare_amount, total_amount, and pre_tip_total_amount (all > 0.98), indicating redundancy that could cause data leakage. Moderate correlations also appear between trip_distance and both fare_amount and pre_tip_total_amount (≈0.89), reflecting the natural link between distance and fare. In contrast, tip_amount exhibits weaker associations with other features, suggesting tipping behavior is less structurally determined.

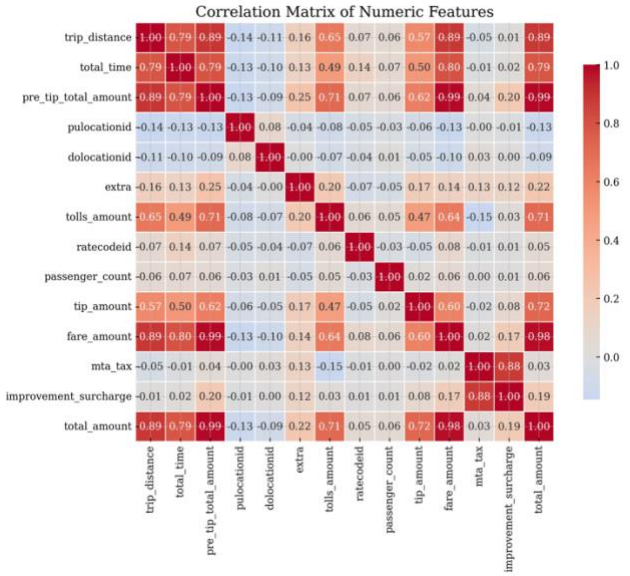

Figure 2: Pearson correlation matrix of numeric variables in the taxi dataset. Strong dependencies are observed between fare_amount, total_amount, and pre_tip_total_amount, while tipping behavior shows weaker associations with structural trip variables.

B. *Geographic Variation in Tips*

To identify which pickup locations yield the highest tips, the average tip amount per trip across all pickup zones was calculated, normalizing by ride counts to avoid bias from high-volume areas. Fig. 3 shows the top 10 pickup locations by average tip amount (minimum of 30 trips). The results indicate that specific zones—most notably location IDs 132, 138, and 70 (JFK, Kip's Bay, and East Elmhurst, respectively)—consistently generate the highest average tips, often associated with longer trips to or from major hubs such as airports and central business districts.

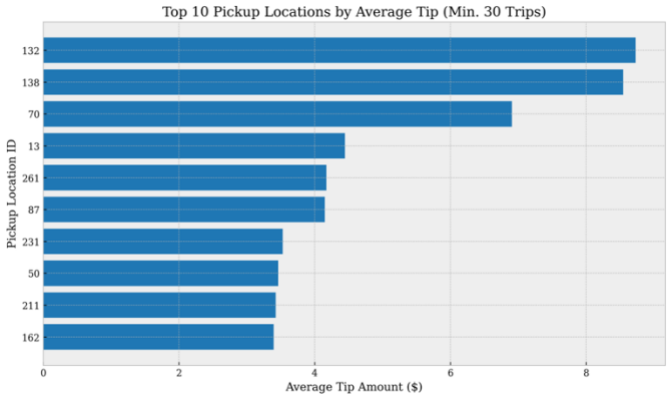

Figure 3: Top 10 pickup locations by normalized average tip amount. Zones with fewer than 30 trips were excluded to reduce noise.

C. *Temporal Variation in Tips*

To analyze the interaction between time and tipping behavior, we examined average tip amounts across both day of week and time of day. Fig. 5 presents a heatmap of mean tips normalized per trip. The results show that Friday and Saturday evenings yield the highest average tips, consistent with leisure and nightlife travel patterns. In contrast, weekday mornings, dominated by routine commutes, are associated with relatively lower tips.

These findings reinforce that tipping behavior is not solely tied to fare amount or trip distance, but also reflects passenger context and trip purpose, providing actionable insights for drivers in scheduling their shifts.

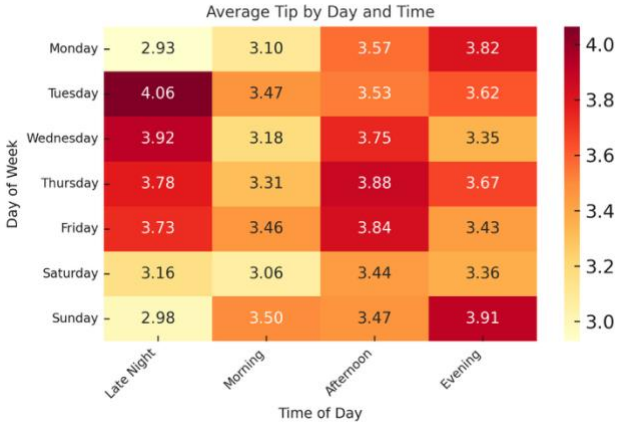

Figure 4: Average tip amount by day of week and time of day. Higher tips are concentrated in evening and weekend periods, while weekday mornings show the lowest tipping behavior.

IV. Modeling

A. *Pipeline Setup*

All modeling was implemented within scikit-learn pipelines, which combined preprocessing, encoding, and model estimation into a single workflow. Before model training, the dataset was divided into training (80%) and testing (20%) sets using a fixed random seed to ensure reproducibility. The training split was used for cross-validation and hyperparameter tuning, while the held-out test split was reserved strictly for final performance evaluation.

To prevent data leakage, different feature sets were used depending on the prediction task: when predicting tips, the variables tip_amount and total_amount were excluded, while for fare prediction, fare_amount, pre_tip_total_amount, and total_amount were removed. This ensured that each model relied only on features realistically available to drivers before a trip's outcome is known.

Each pipeline began with a custom LocationTargetEncoder, which applied target encoding to the high-cardinality pickup and dropoff location IDs. The encoded dataset was then passed into a ColumnTransformer that directed features through column-specific preprocessing: continuous variables were imputed (using either the median, mode, or a constant zero) and standardized, categorical variables were one-hot encoded after imputation, and fee-like binary flags were treated as categorical indicators. Finally, the transformed dataset was passed to the estimator, either Multiple Linear Regression as the baseline or Lasso Regression as the regularized model.

B. *Model Fitting and Cross-Validation*

Once the preprocessing and feature transformation pipelines were established, we trained and evaluated both Multiple Linear Regression and Lasso Regression models. Model selection was carried out using the training split only, with the held-out test split reserved for final evaluation.

For the tip prediction task, both Linear and Lasso pipelines were fit, and hyperparameter tuning was conducted on the Lasso model using 5-fold cross-validation (CV). Candidate values of the regularization parameter α were searched over the range [10^-5, 10^-2]. The best model was selected based on mean CV R^2 performance, and the resulting estimator was then evaluated once on the test set.

For the fare prediction task, we followed the same procedure, this time training on features constructed without fare_amount, pre_tip_total_amount, and total_amount to prevent leakage. Separate pipelines were built for Linear and Lasso Regression, and tuning for Lasso again relied on a grid search across values α [10^-5, 1] with 5-fold CV.

V. Results

A. *Model Evaluation Stratgey*

Model performance was measured using the coefficient of determination (R^2) as the primary metric, with 95% confidence intervals (CI) estimated from 5-fold cross-validation on the training set. The held-out test set was used once for final evaluation. Hyperparameters for Lasso regression were tuned within cross-validation using grid search over α values.

B. *Predicting Tips*

Both multiple Linear Regression and Lasso Regression pipelines were trained. The baseline Linear Regression model achieved a cross-validated R^2 of 0.5607 with a 95% CI of [0.5141, 0.6074] and a test R^2 of 0.5518**.** The Lasso model, tuned over values of α ranging from 10^{-5} to 10^{-2}, resulted in a cross-validated R^2 of 0.5607 with a 95% CI of [0.5141, 0.6073] and a test R^2 of 0.5518**.** While Lasso regularization did not increase predictive accuracy compared to standard Linear Regression, it identified features with little or no contribution to tip prediction.

The models revealed that tips are most strongly influenced by trip distance, pre-tip fare, and additional fees. Each extra mile raised expected tips by about $0.56, while every added dollar in pre-tip fare contributed roughly $1.65. Time and day patterns were also important: afternoons and Thursdays yielded higher tips, whereas evenings and Sundays were less favorable. Passenger count and many minor surcharges showed little or a negative influence.

Lasso highlighted large coefficients for certain rate codes and vendor IDs, but since these reflect administrative categories outside driver control, they were excluded from recommendations. It also eliminated weak predictors such as Saturday pickups, Vendor ID 1, Payment Type 3, and certain surcharge categories, sharpening the focus on actionable strategies. While Lasso did not improve R^2 compared to Linear Regression, it clarified the drivers of tipping behavior. The most effective strategies are to prioritize longer rides, afternoon and Thursday shifts, and higher-fare passengers (Fig. 5).

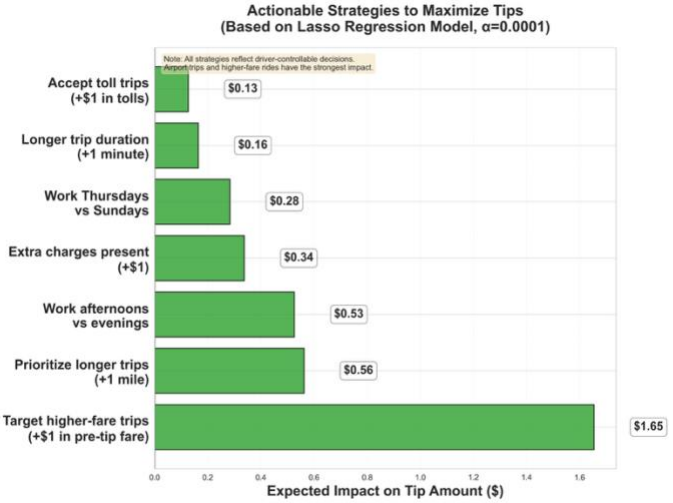

Figure 5: Expected increase in tip amount by controllable driver strategies, as identified by the Lasso regression model.

C. *Predicting Fare*

Both multiple Linear Regression and Lasso Regression pipelines were trained. The baseline Linear Regression model achieved a cross-validated R^2 of 0.927 with a 95% CI of [0.899, 0.946] and a test R^2 of X. The Lasso model, tuned over α values from 10^{-5} to 10^{-2}, selected $\alpha = 10^{-5}$, yielding a cross-validated R^2 of 0.922 and a test R^2 of 0.927 with a 95% CI of [0.899, 0.946]. Because the confidence intervals overlapped, Lasso did not significantly improve accuracy compared to Linear Regression.

Still, Lasso improved interpretability by removing weak predictors such as Vendor ID 2, Payment Type 4, and Monday pickups. The dominant drivers of fare were trip distance (+$8.76 per mile) and trip duration (+$5.32 per minute), which directly reflect the metered pricing system. Location also mattered: pickups in high-fare areas increased fares by $3.69, drop-offs by $1.71. Tips were positively associated with fare (+$1.51 per dollar), while tolls (+$0.68) and extra charges (+$0.35) added smaller amounts. Passenger count and time of day had little effect.

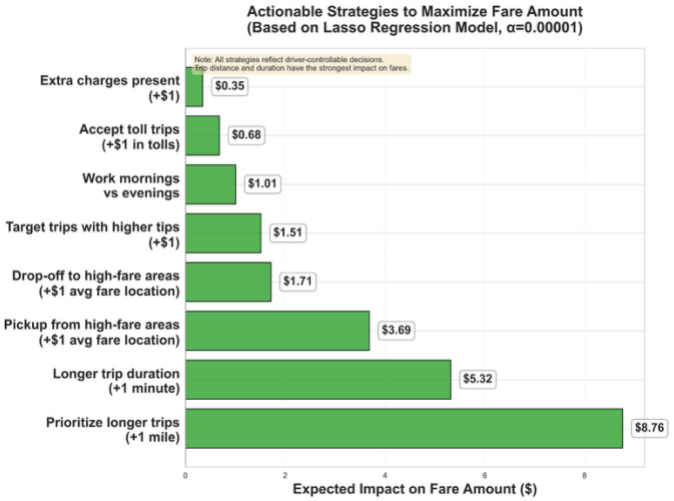

Figure 6: Expected impact of actionable strategies on fare amount, estimated using a Lasso regression model (α = 0.00001).

CONCLUSION

This study shows that while fares are largely determined by fixed factors such as distance and surcharges, tips are influenced by timing, location, and payment methods. Linear and Lasso Regression models highlighted actionable strategies, including prioritizing longer trips, working during high-tipping periods, and targeting profitable pickup zones. Although predictive accuracy for tips remains limited, these insights demonstrate how data-driven approaches can help drivers and companies optimize trip choices and scheduling to improve earnings in a challenging industry.

REFERENCES

[1] [New York City Taxi and Limousine Commission, "Aggregated Reports," NYC.gov, 2024. [Online]. Available: https://www.nyc.gov/site/tlc/about/aggregated-reports.page

[2] National Credit Union Administration, "Timeline of the NYC Taxi Medallion Crisis," NCUA, 2022. [Online]. Available: https://ncua.gov/news/responding-collapse-new-york-city-taxi-medallion-market/timeline-nyc-taxi-medallion-crisis

[3] Salary.com, "Taxi Driver Salary in New York, NY," Salary.com, Jan. 2025. [Online]. Available: https://www.salary.com/research/salary/benchmark/taxi-driver-salary/new-york-ny