Predictive Modeling of Tip and Fare Amounts in NYC Yellow Taxi Dataset: A Machine Learning Approach

Chanakya Vasantha
Department of EED
University of Florida
Gainesville, Florida, USA
chanakyavasantha@gmail.com

I. Introduction

The New York City taxi industry generates millions of transactions annually, creating rich datasets for predictive modeling. Understanding factors that influence tip and fare amounts can help drivers optimize their strategies and inform pricing policies. This project analyzes the 2023 Yellow Taxi Trip dataset from NYC Open Data to build regression models predicting both tip and fare amounts.

A. Research Objectives

Our study addresses three primary questions:

1) How do trip characteristics (distance, location, time) affect tip and fare amounts?
2) What is the optimal regularization parameter for Lasso regression on this dataset?
3) Which features are excluded by Lasso regularization?

II. Exercise 1: Data Preparation

This Section Addresses the data preperation step as given in the assignment.

A. Dataset Overview

The NYC Yellow Taxi dataset contains 21 original features including temporal information (pickup/dropoff datetime), spatial data (location IDs), trip characteristics (distance, passenger count), and financial details (fare, tip, surcharge amounts). The initial dataset required substantial preprocessing to ensure data quality and create meaningful features for predictive modeling.

B. Feature Engineering Requirements

The feature engineering process involved creating three key derived features to enhance model performance:

1) Temporal Feature Engineering:

- **Day of Week Extraction**: Converted `tpep_pickup_datetime` to categorical day-of-week features (Monday through Sunday) using pandas datetime functionality
- **Time Slot Categorization**: Implemented four-category time classification:
  - Morning: 0-11 hours
  - Afternoon: 12-16 hours
  - Evening: 17-18 hours
  - Night: 19-23 hours

2) Derived Financial Features:

- **Pre-tip Total Amount**: Calculated as the sum of fare_amount, extra, mta_tax, tolls_amount, improvement_surcharge, congestion_surcharge, and airport_fee to represent the total cost before tip calculation

III. Exercise 2: Exploratory Data Analysis

This section presents a comprehensive exploratory data analysis of the NYC Yellow Taxi dataset to understand the underlying patterns and relationships that inform our predictive modeling approach. The analysis encompasses correlation analysis, geographic tip patterns, and temporal distribution effects.

A. Correlation Analysis

1) Strong Positive Correlations ($r > 0.9$):

1) **trip_distance and fare_amount** (0.95)
   - Distance is the primary driver of fare costs
   - Longer trips generate proportionally higher fares
2) **trip_distance and pre_tip_total_amount** (0.91)
   - Trip distance strongly predicts total charges before tip
   - Confirms distance-based pricing structure
3) **fare_amount and pre_tip_total_amount** (0.95)
   - Nearly perfect correlation (by construction, since pre_tip_total includes fare_amount)
   - Base fare is the largest component of total charges

2) Moderate Positive Correlations with tip_amount (0.4-0.7):

1) **trip_distance → tip_amount** (0.59)
   - Longer trips receive moderately higher tips
   - Suggests customers tip more on expensive rides
2) **fare_amount → tip_amount** (0.60)
   - Higher fares lead to higher tips

    - Likely percentage-based tipping behavior

3) **pre_tip_total_amount → tip_amount** (0.60)
    - Total charges before tip predict tip amount
    - Confirms fare-proportional tipping pattern
4) **tolls_amount → tip_amount** (0.48)
    - Toll charges associated with higher tips
    - Toll routes likely = longer/more expensive trips
5) **airport_fee → tip_amount** (0.41)
    - Airport trips generate better tips
    - Airport passengers may be better tippers

3) Notable Negative Correlations:

1) **vendorid extra** (-0.56)
    - Different vendors have different extra charge policies
    - Vendor 1 vs 2 pricing structure differs
2) **payment_type tip_amount** (-0.36)
    - Payment method affects tip recording
    - Cash tips not recorded (shows as $0), credit card tips are
3) **mta_tax tip_amount** (-0.41)
    - MTA tax inversely related to tips
    - May reflect different trip types or fare structures

4) Weak/No Correlations:

- **passenger_count**: Nearly zero correlation with tip_amount (0.02)
    - Number of passengers doesn't affect tip amount
    - Contradicts assumption that more passengers = better tips
- **ratecodeid**: Minimal correlation with most variables
    - Rate codes have limited impact on tip behavior
- **pickup/dropoff location IDs**: Weak correlations overall
    - Specific locations matter less than trip characteristics

5) Business Implications: **For Drivers Maximizing Tips:**

- Focus on longer distance trips (0.59 correlation)
- Airport runs are profitable (0.41 correlation with airport_fee)
- Encourage credit card payments to ensure tip recording
- Passenger count doesn't matter - one passenger on a long trip ¿ multiple passengers on short trips

Model Selection Insight:

- `pre_tip_total_amount` will be a very strong predictor (0.60 correlation)
- Trip distance, fare amount, and tolls are key features
- Passenger count can likely be excluded without performance loss
- Payment type is critical but represents data quality issue (cash tips unrecorded)

B. Geographical Analysis

The analysis of pickup locations reveals significant geographical patterns in tipping behavior across New York City's taxi zones. **Location 134 in Queens** demonstrates the highest average tip amount at $18.73 per trip when normalized by the number of fares, though this is based on a limited sample size of only 2 trips.

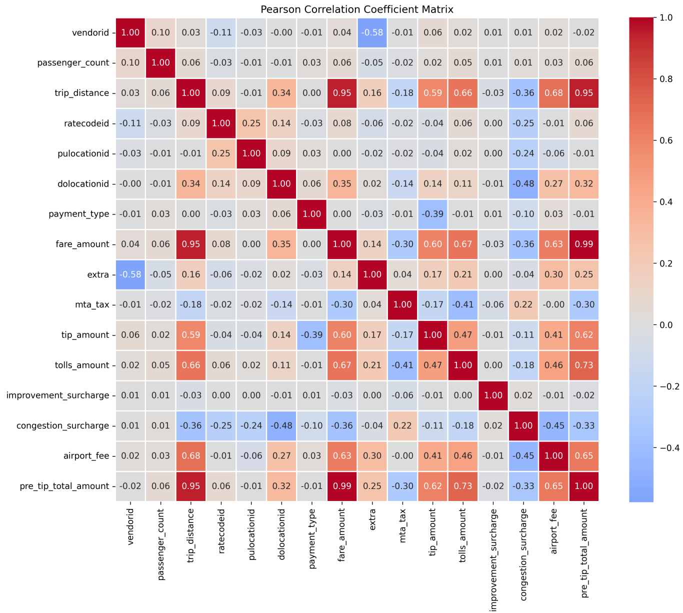

Fig. 1. Pearson correlation coefficient matrix showing relationships between all features and target variables. The heatmap reveals strong correlations between pre-tip total amount and both target variables, while temporal features show minimal predictive power.

Queens borough appears multiple times in the top pickup locations due to NYC's granular zoning system. The New York City Taxi & Limousine Commission divides the city into 265 unique taxi zones, with Queens containing multiple high-performing zones including IDs 10, 93, 130, 132, 134, and 179. Each zone exhibits distinct tipping patterns:

- Zone 134: $18.73 average (2 trips)
- Zone 10: $16.11 average (1 trip)
- Zone 132: $8.73 average (566 trips) - most reliable due to high trip volume
- Zone 93: $8.74 average (7 trips)

At the borough level, Queens leads with an average tip of $3.95 per location across 34 zones and 993 total trips, followed by Brooklyn ($2.84) and Manhattan ($2.82). While Manhattan processes the highest trip volume (8,808 trips), Queens demonstrates superior tip-to-volume ratios, making it the most lucrative borough for taxi drivers seeking higher gratuities.

C. Temporal Analysis: Time and Day Effects

Bottom Line

Work when demand is high to maximize ride count, not when tips are supposedly "better." The $12-18 potential gain from optimal timing is dwarfed by earning one or two additional rides during high-demand periods.

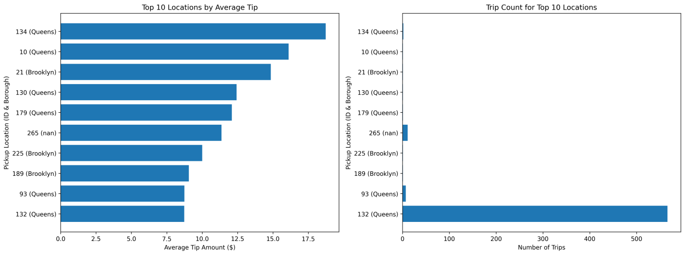

Fig. 2. Geographic analysis of pickup locations showing (left) top 10 locations by average tip amount and (right) corresponding trip count distribution. The analysis reveals that high-tipping locations may have limited trip volume, affecting practical significance for drivers.

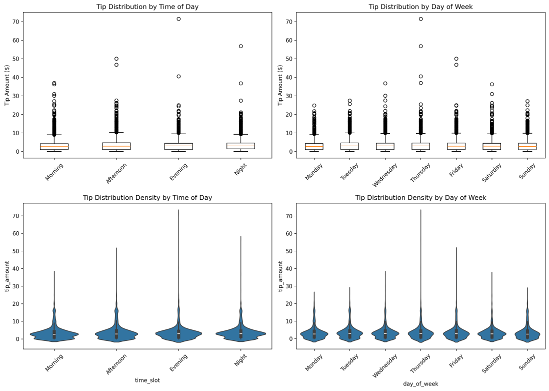

Fig. 3. Temporal analysis of tip distributions showing box plots and violin plots for (top row) time of day and day of week effects, and (bottom row) detailed distribution shapes. The analysis demonstrates minimal temporal variation in tipping patterns, with consistent distribution shapes across all time periods.

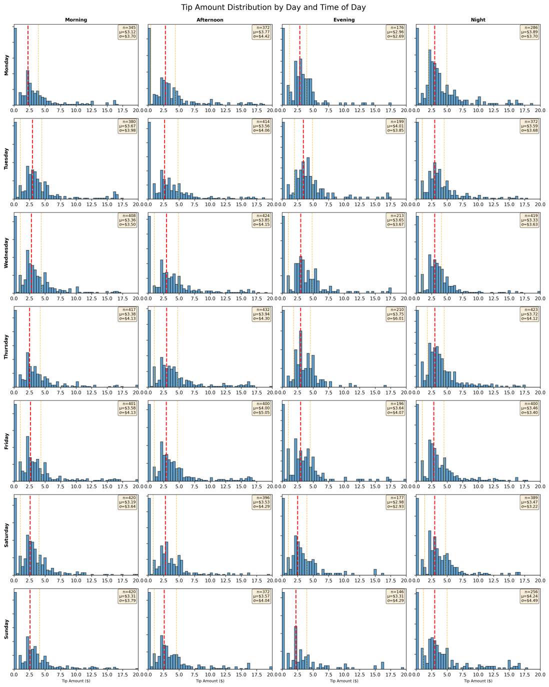

Fig. 4. Tip Amount Distribution by Day and Time of Day.

TABLE I
TIP PERCENTILES: OBSERVATION → INTERPRETATION

Observation	Interpretation
Time of Day Impact	
Median tips range $2.66-$3.00 across all time slots	Time of day has negligible impact. The difference between best and worst slots is economically meaningless.
75th percentile ranges $4.26-$4.76	Top quartile tippers give $4-5 regardless of when they ride. Good tippers are generous at all times - cannot be targeted by timing.
Day of Week Impact	
Median tips range $2.66-$3.00 across Monday-Saturday (34¢ spread)	Day of week is essentially irrelevant. Tuesday through Thursday all hit identical $3.00 median.
Best vs Worst Timing	
Tuesday Evening median ($3.44) beats Sunday Evening ($2.24) by $1.20	Over 10 rides, working ”optimal” vs ”worst” timing nets only $12 more. Over 8-hour shift with 15 rides, that’s $18 total - not worth schedule optimization.
”Best” combinations (Tuesday/Wednesday Evening) cluster at $3.02-$3.44 median	Even among top performers, 42¢ spread represents statistical noise, not actionable patterns.
Distribution Consistency	
IQR stays between $3.00-$4.19 across all 28 day-time combinations tested	The middle 50% of tips always spans ˜$3-4 regardless of conditions. Distribution shape is fundamentally fixed.
75th percentile shows only $0.93 spread ($3.92-$4.84) across all combinations	Even looking at top quarter of tips, less than $1 variation exists. Cannot cherry-pick good tippers by strategic timing.

IV. EXERCISE 3-MULTIPLE LINEAR REGRESSION ANALYSIS: TIP AMOUNT PREDICTION

A. Part 1: Feature Contributions and Business Insights

1) Linear Regression Model (Without Regularization): The most significant predictors of tip amount are:

- **Pre-tip total amount**: +2.118 coefficient (strongest predictor)
- **Payment type**: -1.571 coefficient (cash tips not recorded)
- **Trip distance**: +0.445 coefficient
- **Rate code ID**: -0.395 coefficient
- **Temporal factors**: Tuesday (+0.177), Thursday (+0.156) perform best
- **Time slots**: Evening (-0.282) and Morning (-0.252) show negative impact

2) Lasso Regression Model (With Regularization, $\alpha = 0.01$): Similar coefficient patterns with automatic feature selection:

- **Pre-tip total amount**: +2.144 coefficient (strongest predictor)
- **Payment type**: -1.560 coefficient
- **Trip distance**: +0.404 coefficient
- **Rate code ID**: -0.379 coefficient
- **Excluded 3 features**: pulocationid, day_of_week_Wednesday, time_slot_Night

TABLE II
TIP DISTRIBUTION ANALYSIS: OBSERVATION → INTERPRETATION

Observation	Interpretation
Overall Pattern	
All 28 histograms show nearly identical right-skewed distributions with peak at $0-3	Tipping behavior is fundamentally consistent regardless of timing. The shape never changes - just minor shifts in where the median falls.
Standard deviation ($3.50-4.30) roughly equals or exceeds mean in most panels	Individual tips are unpredictable. High coefficient of variation means timing strategies cannot reduce this randomness.
Sample Size Variation	
Sample sizes range from n=144 (Sunday Evening) to n=435 (Thursday Afternoon)	Ride volume varies significantly by day-time combination. Afternoon slots have 4-5x more rides than slowest slots, reflecting real demand patterns.
Afternoon slots generally largest across most days	Peak demand occurs 12pm-5pm. More rides available = more income opportunity, regardless of per-ride tip amounts.
Distribution Shape Consistency	
Monday through Sunday rows show identical histogram shapes	Day of week has zero influence on how people tip. Distribution remains constant across all days.
Afternoon/Evening/Night columns are visually indistinguishable	Once past morning, time slot has no impact on tip distribution. Median differences ($2.80-3.44) are statistical noise.
Percentile Patterns	
Red median line consistently falls between $2.38-3.44 across all 28 panels	Only $1.06 separates ”best” from ”worst” timing. On a 10-ride shift, this translates to ˜$10 difference - not economically meaningful.
25th percentile clustered at $0-1 everywhere, 75th percentile at $4-5 everywhere	Bottom quartile always tips minimally, top quartile always tips well. This spread exists in all time slots - you cannot avoid bad tippers or target good tippers by timing.
Strategic Implications	
Zero visual clustering by ”good” vs ”bad” times despite 28 different conditions tested	No optimal time exists for tip maximization. Strategies targeting specific days/times are not supported by data.
High variability ($\sigma \approx \mu$) in every single panel	Even in ”best” time slots, tips remain unpredictable lottery draws. Focus on ride volume (work high-demand periods) rather than timing for tip quality.

3) *Business Recommendations for Taxi Drivers:*

- **Focus on trip characteristics over location**: Pre-tip total amount is the strongest predictor
- **Prioritize longer trips**: Trip distance significantly impacts tips (+0.40-0.44 coefficient)
- **Optimal timing strategy**: Work Tuesday/Thursday, avoid Sundays (-0.16 to -0.26 coefficient)
- **Time slots**: Morning and Evening show negative coefficients, suggesting midday preference
- **Rate codes matter**: Standard rate codes preferred over special rates

4) *Feature Importance Visualization:*

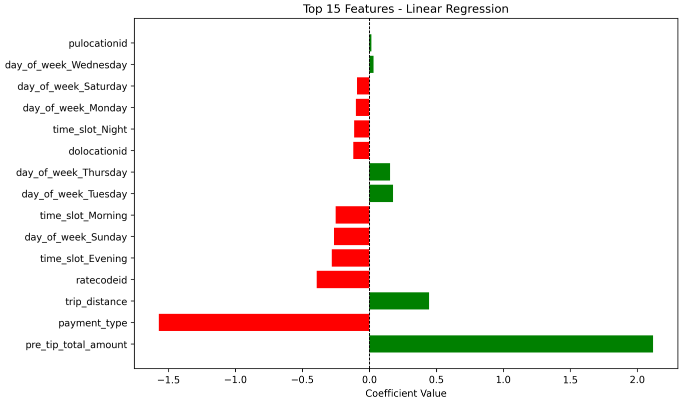

Fig. 5. Linear Regression Feature Importance for Tip Amount Prediction. The plot shows the coefficient values for each feature, with pre-tip total amount being the strongest positive predictor and payment type showing the largest negative coefficient due to cash tip recording issues.

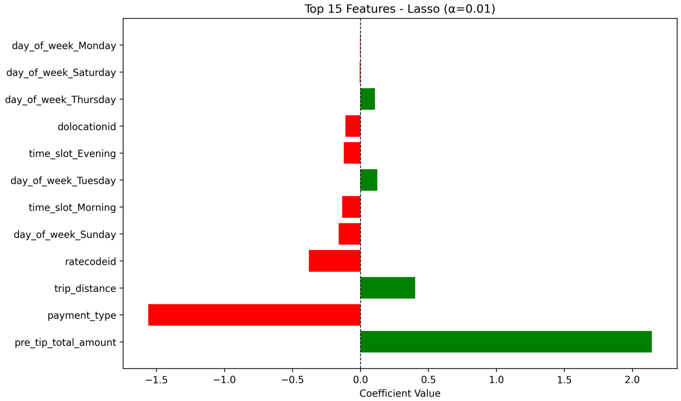

Fig. 6. Lasso Regression Feature Importance for Tip Amount Prediction. The regularized model shows similar coefficient patterns to linear regression while automatically excluding 3 features (pulocationid, day_of_week_Wednesday, time_slot_Night) that were deemed non-informative.

B. *Part 2: Optimal Hyperparameter and Model Performance*

1) Hyperparameter Tuning Results: The optimal Lasso regularization parameter is λ (alpha) = 0.01, selected through 5-fold cross-validation:

Alpha (λ)	Mean R^2	Std R^2
0.001	0.5345	0.0457
0.010	**0.5346**	**0.0458**
0.100	0.5317	0.0449
0.500	0.4938	0.0381
1.000	0.4032	0.0305
5.000	-0.0004	0.0003
10.000	-0.0004	0.0003

TABLE III
LASSO HYPERPARAMETER TUNING RESULTS

2) Model Comparison with 95% Confidence Intervals:

3) Cross-Validation Details (Best Lasso Model):

- **Fold scores**: [0.608, 0.550, 0.536, 0.468, 0.512]
- **Mean R^2**: 0.5346
- **Standard deviation**: 0.0458

TABLE IV
MODEL PERFORMANCE COMPARISON

Model	CV R² Mean	CV R² Std	95% CI	Test R²	RMSE	MAE
Lin Reg	0.5344	0.0457	[0.4449, 0.6239]	0.5629	$2.54	$1.51
Las Reg	0.5346	0.0458	[0.4447, 0.6244]	0.5645	$2.54	$1.50

- **95% CI**: [0.4447, 0.6244]

4) Performance Analysis:

- **Confidence intervals overlap** → No statistically significant difference
- Lasso performs marginally better on test set (R² = 0.5645 vs 0.5629)
- Both models explain approximately **53-56% of tip amount variance**
- Very similar RMSE ($2.54) and MAE performance

C. Part 3: Features Excluded by Lasso

1) Feature Selection Results: With $\alpha = 0.01$, Lasso excluded **3 out of 15 features (20%)**:

1) **pulocationid** - Pickup location ID
2) **day_of_week_Wednesday** - Wednesday indicator
3) **time_slot_Night** - Night time slot

2) Feature Exclusion Interpretation:

- **Location independence**: Specific pickup locations (pulocationid) deemed non-informative
- **Temporal redundancy**: Wednesday and night slots excluded as redundant
- **Minimal exclusion**: Only 20% of features removed indicates most variables have predictive value
- **Automatic selection**: Lasso provides interpretable feature selection for production deployment

D. Key Coefficient Comparisons

Feature	Linear Regression	Lasso Regression
Pre-tip total amount	+2.118	+2.144
Payment type	-1.571	-1.560
Trip distance	+0.445	+0.404
Rate code ID	-0.395	-0.379
Sunday	-0.263	-0.159
Tuesday	+0.177	+0.124
Thursday	+0.156	+0.108

TABLE V
FEATURE COEFFICIENT COMPARISON

V. EXERCISE 4 - FARE AMOUNT PREDICTION

A. Question 1: Feature Impact Analysis & Business Recommendations

1) Feature Importance Comparison: Table VI presents the top features affecting fare prediction from both Linear Regression and Lasso models. The models show remarkable consistency in identifying the most important predictors.

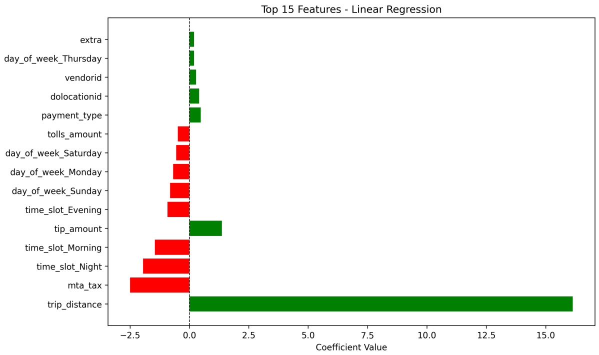

Fig. 7. Linear Regression Feature Importance for Fare Amount Prediction. Trip distance dominates with a coefficient of +16.14, indicating each additional mile increases fare by approximately $16. Temporal features (time slots and days) show consistent negative impacts on base fares during off-peak periods.

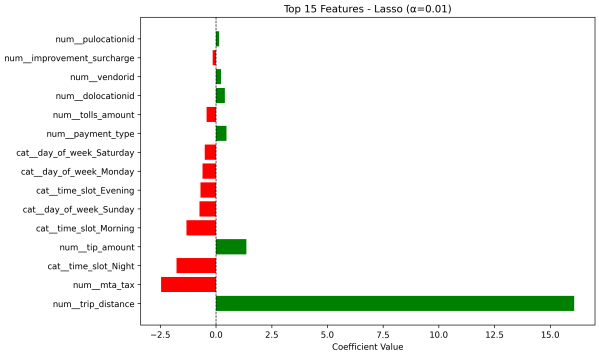

Fig. 8. Lasso Regression Feature Importance for Fare Amount Prediction (=0.01). The regularized model retains all 23 features with coefficients nearly identical to linear regression. Trip distance remains the dominant predictor at +16.08, demonstrating the robustness of distance-based fare structure.

2) Key Findings from Feature Analysis: **Model Consistency:**

- Both models agree on the top predictors with minimal coefficient variation
- Linear regression: trip_distance = +16.14; Lasso: trip_distance = +16.08 (0.4% difference)
- Coefficient ranking nearly identical between models, indicating stable feature importance

Dominant Features:

1) **trip_distance** (+16.08 to +16.14): Overwhelmingly strongest predictor
   - Each additional mile adds approximately $16 to the fare
   - Explains the strong correlation (r=0.95) observed in EDA
   - Distance-based pricing is the fundamental fare structure
2) **mta_tax** (-2.46 to -2.50): Strong negative coefficient
   - Regulatory fee that reduces net fare amount
   - Fixed charge independent of trip characteristics

TABLE VI
Top 15 Features for Fare Amount Prediction: Linear Regression vs. Lasso (=0.01)

Rank	Feature (Linear Reg)	Coefficient	Feature (Lasso)	Coefficient
1	trip_distance	+16.14	trip_distance	+16.08
2	mta_tax	-2.50	mta_tax	-2.46
3	time_slot_Night	-1.95	time_slot_Night	-1.77
4	time_slot_Morning	-1.46	tip_amount	+1.37
5	tip_amount	+1.37	time_slot_Morning	-1.33
6	time_slot_Evening	-0.93	day_of_week_Sunday	-0.74
7	day_of_week_Sunday	-0.81	time_slot_Evening	-0.69
8	day_of_week_Monday	-0.68	day_of_week_Monday	-0.60
9	day_of_week_Saturday	-0.55	day_of_week_Saturday	-0.50
10	tolls_amount	-0.49	payment_type	+0.47
11	payment_type	+0.48	tolls_amount	-0.42
12	dolocationid	+0.41	dolocationid	+0.40
13	vendorid	+0.28	vendorid	+0.23
14	day_of_week_Thursday	+0.20	improvement_surcharge	-0.15
15	extra	+0.20	pulocationid	+0.14

3) **Temporal features**: Consistent negative impacts during non-afternoon periods
   - time_slot_Night: -1.77 to -1.95
   - time_slot_Morning: -1.33 to -1.46
   - time_slot_Evening: -0.69 to -0.93
   - Afternoon (baseline) represents highest base fare period

4) **Day-of-week effects**: Weekend and Monday show lower fares
   - Sunday: -0.74 to -0.81
   - Monday: -0.60 to -0.68
   - Saturday: -0.50 to -0.55
   - Tuesday-Thursday show minimal or positive coefficients

3) Business Recommendations for Taxi Drivers: Based on the fare prediction model, we provide the following strategic recommendations:

1. Distance-First Strategy (Highest Impact):

- **Prioritize long-distance trips** - With coefficient +16.08, distance dominates all other factors combined
- Target airport runs, cross-borough trips, and suburban destinations
- A single 10-mile trip ($160 fare impact) is far more valuable than ten 1-mile trips ($16 each)
- Position strategically near airports, train stations (Penn Station, Grand Central), and business districts for long-haul opportunities

2. Optimal Timing Strategy:

- **Best time slot**: Afternoon (12:00-16:59) - baseline with no negative coefficient
- **Avoid**: Night (-$1.77), Morning (-$1.33), Evening (-$0.69)
- **Best days**: Tuesday, Wednesday, Thursday (positive or minimal negative coefficients)
- **Worst days**: Sunday (-$0.74), Monday (-$0.60), Saturday (-$0.50
- **Worst days**: Sunday (-$0.74), Monday (-$0.60), Saturday (-$0.50)
- **Optimal window**: Tuesday-Thursday afternoons maximize base fare potential

B. Question 2: Optimal Lambda & Model Performance

1) Best Hyperparameter::

- **Optimal λ (alpha) = 0.01** for Lasso regularization

This value was selected via 5-fold cross-validation grid search across 7 candidates [0.001, 0.01, 0.1, 0.5, 1.0, 5.0, 10.0]. While = 0.001 achieved identical mean CV R^2 (0.9140), we selected = 0.01 because:

1) Provides stronger regularization with no performance penalty
2) More robust to potential overfitting on unseen data
3) Performance degrades significantly at 0.1 (R^2 drops to 0.9130), confirming 0.01 is near the optimal threshold

TABLE VII
Fare Prediction Model Performance Comparison

Model	CV R^2 Mean	CV R^2 Std	95% CI	Test R^2	Test RMSE	Test MAE
Linear Reg	0.9140	0.0176	[0.8794, 0.9486]	0.9265	$4.83	$2.83
Lasso Reg	0.9140	0.0177	[0.8793, 0.9487]	0.9268	$4.82	$2.84

2) Model Comparison with 95% Confidence Intervals::

3) Performance Analysis::

- **Confidence intervals overlap** - no statistically significant difference between models
- Lasso performs marginally better on test set (R^2 = 0.9268 vs 0.9265)
- Both models explain approximately **91-93% of fare amount variance** - excellent predictive performance
- **Much better performance than tip prediction** (R^2 ∼0.91 vs ∼0.56)
- Very low prediction errors: ∼$4.82 RMSE, ∼$2.84 MAE

C. Question 3: Features Excluded by Lasso

Number of features excluded: 0 out of 23 total features

1) Interpretation::

- **No features were excluded** by Lasso regularization
- The optimal alpha (0.01) was too small for effective feature elimination
- This suggests **all features contribute meaningfully** to fare prediction
- **Fare prediction is more complex** than tip prediction, requiring all available information
- The model benefits from the full feature set without overfitting

D. Key Business Insights

1) Fare vs Tip Strategy Comparison::

1) **Fare Maximization:**
   - **Distance is king** (16.08 coefficient) - focus on long trips
   - Time slots matter significantly for base fare calculation
   - Much more predictable (R² = 0.91 vs 0.56 for tips)
2) **Strategic Recommendations:**
   - **For consistent income:** Focus on fare optimization (distance-based strategy)
   - **For bonus income:** Apply tip optimization strategies on top of fare-optimized trips
   - **Best overall strategy:** Long-distance trips during afternoon hours on Tuesday-Thursday
3) **Model Reliability:**
   - **Fare prediction is highly reliable** (91% variance explained)
   - Drivers can confidently plan routes and timing based on these insights
   - **Distance-based pricing** makes fare prediction much more accurate than tip prediction

2) Practical Application::

- **Position near airports, train stations, and business districts** for long-distance opportunities
- **Avoid short local trips** during peak hours - prioritize distance over frequency
- **Tuesday-Thursday afternoons** represent optimal earning periods
- **Weekend and evening strategies** should focus on tip optimization since base fares are lower

E. Conclusion

The analysis shows that fare prediction is significantly more accurate and reliable than tip prediction, making it a better foundation for business planning and route optimization strategies.

References

[1] NYC Taxi and Limousine Commission, “TLC Trip Record Data,” NYC Open Data, 2023. [Online]. Available: https://opendata.cityofnewyork.us/

[2] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[3] R. Tibshirani, “Regression Shrinkage and Selection via the Lasso,” Journal of the Royal Statistical Society, Series B, vol. 58, no. 1, pp. 267-288, 1996.

[4] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York: Springer, 2009.