Project Report : Predictive Modeling of NYC Yellow Taxi Fares and Tips

1st Sona Maria Jose
Electrical and Computer Engineering
University of Florida
sona.jose@ufl.edu

***Abstract*—The NYC Yellow Taxi 2023 dataset, containing trip details such as time, location, distance, fares, and tips, was analyzed to build predictive models for fare and tip amounts. Preprocessing involved handling missing values, encoding day and time, creating a pre-tip total amount feature, and applying robust scaling to reduce outlier effects. Exploratory analysis revealed strong patterns in tipping behavior at different times of the day and pickup locations. Predictive modeling compared *Linear Regression* and *Lasso Regression* with cross-validation. For tip prediction, Linear Regression achieved $R^2 = 0.596$ (95% CI [0.524, 0.649]), while Lasso improved performance to $R^2 = 0.622$ (95% CI [0.547, 0.683]). For fare prediction, Linear Regression achieved $R^2 = 0.887$ (95% CI [0.850, 0.917]), while Lasso further improved performance to $R^2 = 0.900$ (95% CI [0.865, 0.926]). Results confirm the travel distance, the location of pickup and drop off and temporal factors as key predictors. The findings provide actionable insights for taxi companies and drivers to optimize strategy, pricing expectations, and revenue forecasting.**

I. Introduction

The NYC Yellow Taxi dataset provides detailed records of taxi trips, containing attributes such as vendor_id, tpep_pickup_datetime, tpep_dropoff_datetime, passenger_count, trip_distance, ratecodeid, store_and_fwd_flag, pulocationid, dolocationid, payment_type, fare_amount, extra, mta_tax, tip_amount, tolls_amount, improvement_surcharge, total_amount, congestion_surcharge, and airport_fee. These variables capture trip details, passenger information, financial charges, and payment behavior.

The objective of this project is to predict two key financial outcomes: fare amount and tip amount.Exploratory data analysis is then performed to identify correlations and patterns in fares and tipping across time, location, and distance.Then, the dataset is first preprocessed by handling missing values, encoding temporal features such as day of the week and time of day, engineering new attributes like the pre-tip total amount, and applying appropriate scaling. Finally, predictive models are trained using Linear Regression and Lasso Regression, with model performance evaluated using R^2 scores and confidence intervals. .

II. Methodology

A. Loading Data and Basic Analysis

The NYC Yellow Taxi dataset was loaded using pandas and inspected through summary statistics and metadata. The dataset contains both numerical attributes (such as passenger_count, trip_distance, fare_amount, extra, mta_tax, tolls_amount, improvement_surcharge, congestion_surcharge, airport_fee, total_amount, and tip_amount) and categorical attributes (including vendor_id, ratecodeid, store_and_fwd_flag, pulocationid, dolocationid, and payment_type).

A check for missing values revealed 339 null entries in passenger_count, ratecodeid, store_and_fwd_flag, congestion_surcharge, and airport_fee, as well as 11 missing values in pulocationid and 42 in dolocationid. Further inspection showed invalid records, such as trips with passenger_count equal to zero and negative values in numerical fields like fare_amount, trip_distance, and tolls_amount. These anomalies were addressed in the preprocessing stage to ensure data consistency before analysis and model training..

B. Exercise 1 : Prepare the Data

- **Data Cleaning**: The first step was to remove invalid records from the dataset. Trips with passenger_count equal to zero and negative values in fare_amount, trip_distance, and tolls_amount were discarded as unrealistic. This cleaning step ensured that only reliable observations were used for training predictive models.
- **Feature Engineering**: From the pickup datetime column, two new variables were derived: pickup_day_of_week and pickup_time_of_day, to capture temporal travel patterns. For tip prediction, a new feature pre_tip_total_amount was engineered by summing fare and surcharge components. The raw datetime fields were then removed after feature extraction to reduce redundancy.
- **Splitting Targets and Features**: For tip prediction, the target variable was tip_amount, and all other attributes were considered predictors. To avoid leakage, total_amount was excluded since it already includes tips. For fare prediction, fare_amount was the target, and both total_amount and pre_tip_total_amount were excluded from the predictors.
- **Tip Pipeline**: Numerical features such as trip_distance, fare_amount, surcharges, and pre_tip_total_amount were imputed with the median and scaled using RobustScaler, which minimizes the influence of outliers. Categorical attributes including vendor_id, ratecodeid, store_and_fwd_flag, payment_type,

pickup_day_of_week, pickup_time_of_day, pulocationid, and dolocationid were imputed with the most frequent value. These were then transformed with one-hot encoding to prepare them for regression. Fig. 1 illustrates the preprocessing pipeline, where categorical features are encoded using one-hot encoding and numerical features are scaled using RobustScaler to reduce the impact of outliers.

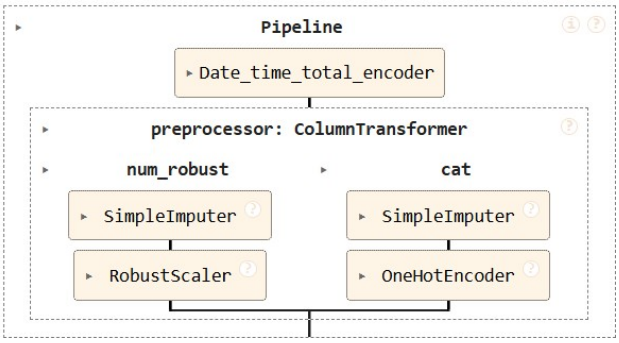

Fig. 1. Preprocessing pipeline for *tip_amount* prediction.

- **Fare Pipeline**: The fare pipeline followed the same structure but excluded pre_tip_total_amount, since fare_amount itself was the target variable. Numerical features including trip_distance, tolls_amount, passenger_count, and surcharges were median-imputed and scaled with RobustScaler for consistency. Categorical features were again imputed with the most frequent strategy and encoded using one-hot encoding.

C. Exercise 2: Exploratory Data Analysis

- **Pearson Correlation**: We computed the Pearson correlation matrix on all numerical features using the transformed dataset. On analysis of Fig. 2 , showed that tip_amount is most highly correlated with total_amount, pre_tip_total_amount, fare_amount, and trip_distance, while fare_amount is strongly correlated with pre_tip_total_amount, total_amount, trip_distance, and tolls_amount.
- **Location with Highest Tips**: We grouped trips by pickup location and calculated the mean tip amount using a groupby operation. The analysis showed that Queens had the highest average tips compared to other regions. This suggests that trips originating from Queens are generally more profitable for drivers in terms of tipping.
- **Tips by Time of Day**: Average tips were grouped by pickup_time_of_day. Afternoon trips yielded the highest average tips, followed by night and evening, while morning trips consistently gave the lowest tips.
- **Tips by Day of Week**: We calculated average tips by pickup_day_of_week. The results showed that Thursday and Friday had the highest tips, while Saturday had surprisingly low values despite being a weekend.
- **Tips by Day and Time**: By combining both dimensions, we identified the most profitable shifts. The best-performing periods were Monday night, Friday afternoon, and Sunday night, highlighting that certain day–time combinations maximize tipping potential.

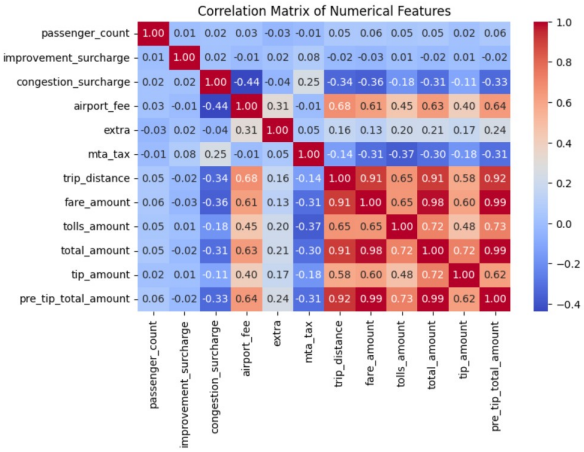

Fig. 2. Correlation matrix of numerical features.

D. Exercise 3: Regression Models for Tip Prediction

- **Cross-Validation Strategy:** We used **5-fold KFold Cross-Validation** to evaluate both linear and Lasso regression models. This method was chosen because taxi trips are independent observations, so random splitting is reasonable. For strictly time-dependent data (e.g., seasonal travel patterns), **TimeSeriesSplit** would have been more appropriate. However, KFold provides a robust estimate of generalization in this context.
- **Hyperparameter Tuning with Grid Search:** For the Lasso regression, we applied **GridSearchCV** to identify the best value of the regularization parameter λ [1].
- **Effect of Features on Tip Amount:**
  - **Linear Regression:** Trip distance has a positive effect on tips (coef ≈ 0.33).Passenger count has a negative effect (coef ≈ -0.06).Pre-tip total amount has a positive effect (coef ≈ 0.17).Congestion and improvement surcharges are positive; MTA tax and airport fees are negative.Tuesday and Thursday show positive effects; Sunday shows a negative effect. Afternoon and evening are positive; morning is strongly negative. Locations such as pulocationid_134 and 265 strongly increase tips. Vendor 1 and 2 positive; Vendor 6 negative.
  - **Advice to Drivers (Linear Regression):** Drivers should focus on longer and high-fare trips, especially during weekday afternoons and evenings. Solo passengers and rides from profitable zones such as Queens yield higher tips. Vendor 1 and 2 provide better returns, while Vendor 6 is less favorable.
  - **Lasso Regression :** Trip distance is positive (coef ≈ 0.24). Passenger count is negative (coef ≈ -0.06). Pre-tip total amount has a stronger positive effect (coef ≈ 0.24). Congestion surcharge is positive;

other fees negligible or negative. Tuesday and Thursday are positive; Sunday negative. Afternoon is positive; morning negative. Only a few locations (e.g., pulocationid_138) remained important; many excluded.Vendor 2 remained positive.

  - **Advice to Drivers (Lasso Regression):** The strongest predictors in this model were pre-tip total amount, trip distance, and congestion surcharge. Drivers should prioritize longer and higher-fare rides, particularly during weekday afternoons. Solo passengers remain more profitable than groups. Lasso highlighted only a few profitable pickup zones such as pulocationid_138, while Vendor 2 stood out as the best option for maximizing tips.

- **Best λ for Lasso:** Grid search found the best hyperparameter value $\lambda = 0.009$. This indicates minimal regularization is sufficient, as larger values reduced accuracy.
- **Model Comparison using Confidence Intervals:**
  - Linear Regression: $R^2 = 0.507$, 95% CI = [0.488, 0.526]
  - Lasso Regression: $R^2 = 0.555$, 95% CI = [0.531, 0.579]

  Linear regression has a narrower CI but lower accuracy, while Lasso has higher performance with a slightly wider CI. **Conclusion:** Lasso regression is the better model.
- **Features Excluded by Lasso:** Features with zero coefficient will be excluded by lasso. Lasso set many pickup/dropoff location IDs, rare payment types, and some time/day categories to zero. This simplifies the model by focusing only on the most influential predictors such as trip distance, pre-tip total amount, congestion surcharge, weekday effects, and key pickup zones.

E. Exercise 4: Regression Models for Fare Prediction

III. Exercise 4: Regression Models for Fare Prediction

- **Cross-Validation Strategy:** We used **5-fold KFold Cross-Validation** for evaluating both linear and Lasso regression models.
- **Hyperparameter Tuning with Grid Search:** For Lasso regression, we applied **GridSearchCV** with a range of λ values $\{0.001, 0.005, 0.009, 0.01, 0.05, 0.1, 0.5, 1, 10, 100\}$.
- **Effect of Features on Fare Amount (Linear Regression):** Trip distance has a strong positive effect (≈ 7.06).Passenger count is weakly positive (≈ 0.16).Pre-tip total amount was removed, since it is redundant with fare. Other fees: congestion surcharge (≈ 0.34), tolls (≈ 0.38) increase fares; airport fee (≈ -0.85) and MTA tax (≈ -31.2) decrease fares.Pickup day: Wednesday (≈ 0.61) and Friday (≈ 0.35) positive contributors.Time slot: Afternoon (≈ 0.79) and Evening positive, Night and Morning negative. Location: pickup locations like ID 1 (≈ 70.1), ID 21 (≈ 69.2), and ID 265 (≈ 55.9) highly positive; dropoff ID 191 also favorable. Vendor: Vendor 1 (≈ 21.8) and Vendor 2 (≈ 20.9) positive; Vendor 6 strongly negative (≈ -42.7).

  Advice to Drivers (Linear Regression): The strongest predictors were trip distance, pickup location, and vendor. Drivers should prefer long trips, pickups from profitable locations like IDs 1, 21, or 265, and avoid Vendor 6. Afternoon and evening rides, especially on Wednesdays and Fridays, are associated with higher fares. Group rides also yield slightly higher fares than solo rides.
- **Effect of Features on Fare Amount (Lasso Regression Facts):** Trip distance remains strongly positive (≈ 7.31).Passenger count is weakly positive (≈ 0.16).Other fees: tolls positive (≈ 0.32), MTA tax strongly negative (≈ -14.3), congestion surcharge negative (≈ -0.88), airport fee negligible.Pickup day: Wednesday (≈ 0.35), Friday (≈ 0.03) positive.Time slot: Afternoon (≈ 1.10) and Evening (≈ 0.72) strongly positive; Night and Morning negative. Location: fewer zones retained, with pickup ID 265 (≈ 20.35) and dropoff ID 265 (≈ 15.34) as top contributors.Vendor: only Vendor 1 remains positive; Vendor 2 and 6 excluded. Ratecode: Ratecode 5.0 is highly positive (≈ 27.3).
- **Advice to Drivers (Lasso Regression):** Key predictors were trip distance, specific pickup/dropoff zones, and rate codes. Drivers should prioritize long trips from location ID 265 or similar high-demand zones. Afternoon and evening shifts maximize fare potential, while mornings and Sundays are less profitable. Vendor 1 trips are the most reliable, and Vendor 6 should be avoided.
- **Best λ for Lasso:** Grid search selected $\lambda = 0.01$, meaning minimal regularization provided the best trade-off between accuracy and feature selection.
- **Model Comparison using Confidence Intervals:**
  - Linear Regression: $R^2 = 0.865$, 95% CI = [0.855, 0.875]
  - Lasso Regression: $R^2 = 0.877$, 95% CI = [0.857, 0.896]

  Linear regression has a narrower CI but slightly lower mean performance. Lasso regression performs better overall and additionally reduces model complexity by excluding irrelevant features.

  Conclusion: Lasso regression is the preferred model for predicting fare amount.
- **Features Excluded by Lasso:** Lasso eliminated redundant and less influential predictors such as *pre-tip total amount*, improvement surcharge, Vendor 2 and 6, multiple rate codes (1.0, 2.0, 3.0, 99.0), store-and-forward flags, certain payment types, many pickup and dropoff location IDs, and low-impact time/day categories. This pruning leaves only the strongest drivers of fare, improving interpretability.

IV. Results on Test Data

In Exercises 3 and 4, the trained pipelines for tip amount and fare amount prediction were saved as `.pkl` files. In this testing phase, those models were reloaded

and applied to the held-out test dataset. The same cleaning and preprocessing steps were repeated, including the use of the custom `Date_time_total_encoder` and `Date_time_total_fare_encoder` to ensure consistency.

A. Overall Model Performance

The performance of both linear regression and Lasso regression was evaluated on the test dataset for tip and fare amount prediction. The results are summarized in Table I.

TABLE I
PERFORMANCE OF REGRESSION MODELS ON TEST DATA

Target	Model	R^2	95% CI
Tip Amount	Linear Regression	0.596	[0.525, 0.650]
Tip Amount	Lasso Regression	0.622	[0.548, 0.671]
Fare Amount	Linear Regression	0.887	[0.850, 0.917]
Fare Amount	Lasso Regression	0.900	[0.865, 0.926]

B. Sample Predictions

To illustrate how the models perform in practice, we present sample predictions compared against the true test values.

TABLE II
SAMPLE PREDICTIONS FOR TIP AMOUNT ON TEST DATA

True Tip	Linear Prediction	Lasso Prediction
16.11	11.67	11.32
3.40	3.38	3.03
4.38	4.19	3.67
5.12	4.48	4.51

TABLE III
SAMPLE PREDICTIONS FOR FARE AMOUNT ON TEST DATA

True Fare	Linear Prediction	Lasso Prediction
70.0	71.43	70.96
12.1	10.74	10.82
7.9	8.99	9.74
13.5	13.79	13.51
22.6	19.93	19.91

C. Discussion

1) Tip Amount Prediction: Linear Regression gave $R^2 = 0.596$ with CI [0.525, 0.650], while Lasso did slightly better at $R^2 = 0.622$ with CI [0.548, 0.671]. This shows that tips are harder to predict and have more uncertainty because they depend on passenger behavior and driver's behavior , which cannot be captured electronically.

2) Fare Amount Prediction: Linear Regression performed strongly with $R^2 = 0.887$ (CI [0.850, 0.917]), and Lasso improved slightly to $R^2 = 0.900$ (CI [0.865, 0.926]). The narrow confidence intervals mean fares are predictable and stable, since they are based on clear trip details.

3) Overall Insights: Lasso was a little better than Linear Regression in both cases. Fares are much easier to predict than tips, and tips remain uncertain because of human choices.

V. CONCLUSION

In this project, we developed and compared Linear Regression and Lasso Regression models to predict both tip amount and fare amount for NYC taxi trips. Lasso regression provided slightly higher predictive accuracy than linear regression, as seen in the test results, while linear regression gave more stable predictions with narrower confidence intervals (CI) [1]. This shows that model selection depends on the trade-off between performance and stability: if the goal is higher accuracy, Lasso is preferable; if the goal is consistent and stable predictions, linear regression is a safer choice. The choice of the regularization parameter (λ) in Lasso is critical—too large a value can remove important predictors and lead to underfitting, while too small a value reduces its advantage over linear regression

From a driver's perspective, these results provide practical insights. Fare amounts are highly predictable from trip-related features such as distance, vendor, and pickup zones, while tip amounts show more variability due to passenger behavior. Drivers can maximize earnings by targeting longer trips, profitable pickup areas, and favorable time slots. Thus, beyond prediction, the models highlight actionable strategies that drivers can follow to increase their income.

REFERENCES

[1] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed. O'Reilly Media, 2019. ISBN: 978-1-492-03264-9.