Forecasting Airport Ride-Share Demand: A Predictive Model Using Flight Activity and Temporal Trends

Niraj Patel
pniraj745@gmail.com
https://doi.org/10.34293/ICETCS2024.ch005

Abstract- *The growth of for-hire vehicle (FHV) services, such as Uber and Lyft, has revolutionized urban mobility, with airports emerging as critical demand hubs due to fluctuations in passenger volumes and flight schedules. This study presents a predictive model for airport ride-share demand, leveraging the NYC Taxi and Limousine Commission's high-volume trip data and the Federal Aviation Administration's hourly flight data. By integrating flight arrival and departure statistics with temporal factors, we analyze how flight volume influences FHV demand at major New York City airports. Using Elastic Net regression, we develop and evaluate models for airport pick-ups and drop-offs, demonstrating a substantial predictive capability for high-demand periods. Our results indicate that future flight departures correlate strongly with airport drop-offs, while past arrivals primarily drive pick-up demand. The model explains over 81% of the variability in ride volumes at the airports, providing actionable insights for ride-share companies to optimize surge pricing, enhance fleet allocation, and improve customer satisfaction. Additionally, our findings can assist airport authorities in better managing traffic flows and inform travelers seeking to minimize transportation costs by avoiding peak periods. This research establishes a foundation for dynamic demand forecasting in transportation, with potential applications in real-time demand management and cost optimization strategies.*

INTRODUCTION

High-volume for-hire vehicle services such as Uber and Lyft dominate the private transportation industry around the world. One main competitive advantage these companies have over traditional taxi services is the algorithms they use to model demand using the vast amounts of data they have access toand thus are able to successfully employ multiple pricingstrategies (surge pricing, fees, etc.) to maximize profits. Airport surge pricing is a well-known strategy used by these services as airports are areas of anomalously high demand driven by external factors relating to flight and passenger volume. In this paper, we attempt to model this external factor of demand using TLC New York datasets [1] paired with hourly flight data from ASPM [2].We aim to identify the differences in hourly volume distribution in airport TLC zones (compared to other TLC zones across NYC) using hypothesis tests and form a more robust model for demand for trips from and to airport TLC zones by including and measuring the effect of the ASPM flight data using a linear regression (ElasticNet) based model.

We use the FHVHV dataset as it is most relevant to our research, as surge pricing is most prevalent in high-volume services, and thus, airport volume modeling is most valuable for such trips. We study flight volume data from the ASPM dataset, particularly departures and arrival numbers by hour and facility. We aggregate hourly FHV counts by facility (using location IDs to identify airport trips), and include temporal variables such as day to account for seasonal trends. Our target audience is both high-volume FHV services (such as Uber, Lyft, etc.) who may use this model to inform surge pricing and their drivers who may use this to better allocate supply, as well as customers (and customer-facing agencies) who may seek to minimize transportation costs by avoiding high volume times. We select 01/2021-07/2021 as our dataset to analyze and train our model, being the earliest available complete dataset for five. We select 02/2022-04/2022 as our testing dataset to evaluate if our models generalize well to recent (and future) periods.

DATA PROCESSING AND ANALYSIS

A. Dataset

In this analysis, we rely primarily on two datasets. Firstly, the NYC Taxi & Limousine Commission (NYCTLC) For-Hire Vehicle High Volume (FHVHV) dataset [1], which records trip details for for-hire vehicle rides commissioned through a ”High Volume” service (such as Uber, Lyft etc). Secondly, we use the Federal Aviation Administration (FAA) Aviation System Performance Metrics (ASPM) dataset [2], which contains metrics about flight departures and arrivals at airports across the USA at an hourly level.

B. High-Level Pre-processing (Data Extraction and Aggregation)

1. First, we begin preprocessing by removing irrelevant data from the TLC dataset. As we plan to analyze ride volume by the hour, we are most interested in counting ”valid rides,” and thus, we must exclude any invalid rides. For the purposes of our analysis, this entails culling any records with a negative fare, as these records do not represent actual rides but rather charge-backs of preceding rides. Apart from this, we do not exclude any outlying data from our analysis, as outlying column values (apart from date time values, which will automatically be excluded on our subsequent join with the flight dataset) will not affect our analysis and are not a sufficient indicator of an “invalid ride” (e.g. an outlying column value does not sufficiently invalidate a whole record from being counted and does not adversely affect our analysis). We must also filter our data to only contain rides to or from airports using the relevant TLC zones (1, 132, 138), generating two separate datasets for airport pickups and dropoffs, respectively.
2. Next, we must obtain the relevant statistics for our analysis. This was done by first grouping records by requesting ”airport” ”hour” and ”date”, then aggregating the ride counts for each group. We also generate columns representing the subsequent and preceding 5 hours and dates corresponding to each ride count to allow temporally shifted analysis (5 hours was chosen as most international flights recommend a 3-hour early arrival, and we assume a maximal expected travel time of 2 hours). These new columns will allow us to iteratively join with the flight dataset and obtain departure and arrival data spanning the preceding 5 hours and following 5 hours as features for our analysis.

C. Feature Selection

Now that we have aggregated and generated several relevant features to predict ride counts, we must choose which are relevant to the two problems presented.

1. ***Airport Dropoffs Features:*** We begin by examining the relevant categorical features, such as day and facility, using ANOVA. We do not consider hour, date, and month as these would require a much larger number of indicator variables to model and may encourage over fitting due to over parameterization. Furthermore, they may serve to associate the data with a given timestamp and thus reduce the generalisability of the model to other years. The hour is specifically not considered as daily time information is already somewhat encoded within flight data.

Table 1: One Way ANOVA - Day

<table>
<tr><th></th><th>Sum_sq</th><th>df</th><th>F</th><th>PR(>F)</th></tr>
<tr><td>C(Day)</td><td>6.272285e+06</td><td>6.0</td><td rowspan="2">38.286814</td><td rowspan="2">2.718924e-46</td></tr>
<tr><td>Residual</td><td>3.035925e+08</td><td>11119.0</td></tr>
</table>

Table 2: One Way ANOVA - Facility

<table>
<tr><th></th><th>Sum_sq</th><th>df</th><th>F</th><th>PR(>F)</th></tr>
<tr><td>C(Facility)</td><td>6.338480e+07</td><td>2.0</td><td rowspan="2">1430.195648</td><td rowspan="2">0.0</td></tr>
<tr><td>Residual</td><td>2.464800e+08</td><td>11123.0</td></tr>
</table>

For airport dropoffs, we intuitively expect departures in the next few hours to have the largest explanatory impact on ride counts (as people will be dropped off at the airport for their future departures).We can use Pearson’s correlation coefficient to ascertain which other features may be correlated to ride count.

As expected, future departure features are most heavily correlated to our response. However, future arrivals also seem to be positively correlated to counts. Furthermore, past arrivals seem to have a somewhat notable negative correlation to counts. In rationalizing these results, it is possible that these correlations are merely the result of being heavily collinear with future departures, particularly for the future arrivals columns (as these are obviously temporally linked).

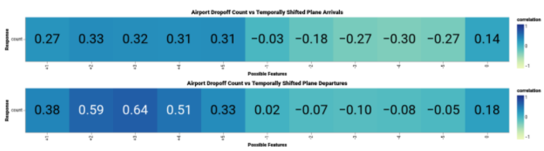


Figure 1: Correlation between Time-Shifted Arrival/Departure Features and Dropoff Count

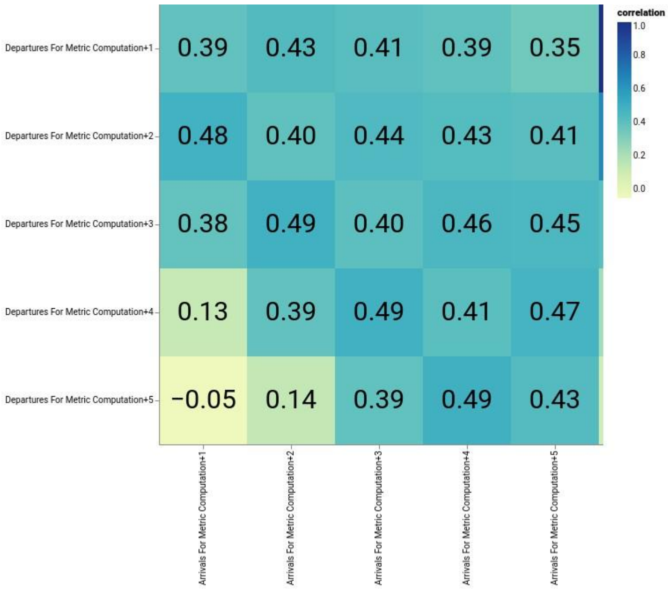


Figure 2: Correlation between Pairs of Time-Shifted Arrivals vsDepartures

This correlation matrix 2 demonstrates the correlation between pairs of arrivals/departures by temporal difference. Note that this mostly generalizes to past arrivals/departures as the feature values are identical (just shifted).

On the other hand, it is plausible that past arrivals may actually negatively impact ride volumes. Past arrivals may indicate a lower available supply of vehicles (as they may be busy picking people up), resulting in a lesser drop-off volume.

Hence, as per the above, we choose departures up to 5 hours into the future and arrivals up to 5 hours in the past as our relevant flight data columns. Furthermore, we also include facility and day indicator variables (dummy variables) as per the ANOVA analysis III.

2. ***Airport Pickup Features:*** Conversely, for airport pickups, we intuitively expect past arrivals to have the greatest correlation with ride counts. Once again, we use Pearson's correlation coefficient.

Table 3: Two Way ANOVA with Interaction - Day & Facility

	Sum_sq	df	F	PR(>F)
C(Facility)	6.385527e+07	2.0	1492.334127	0.000000e+00
C(Day)	6.742758e+06	6.0	52.527365	3.904188e-64
C(Facility):C(Day)	2.152066e+06	12.0	8.382503	5.128350e-16
Residual	2.375851e+08	11105.0		

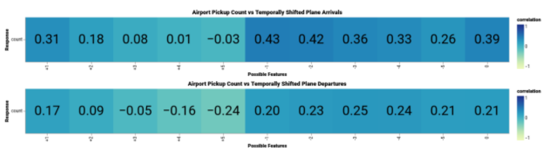


Figure 3: Correlation between Time-Shifted Arrival/Departure Features and Pickup Count

Once again, as expected, past arrivals have the heaviest correlation with our response. In this case, we also observe a non-trivial (albeit lesser) correlation with certain future departure variables, as well as a similarly high correlation with past departures. Again, the response's correlation with past departures can be explained by a high correlation between future departures and arrivals (of which arrivals are clearly and intuitively more explanatory). In this case, we have also chosen to include future departures, although the intuition for such is less clear than that of the former case. It could possibly be due to drivers being more likely to accept rides when there is a high demand for dropoffs in the future (i.e. more departures), as to return to their origin airport. We also include day and facility columns. The ANOVA analysis here, for the sake of brevity, follows similar methods and results as per III.

Models

A. Elastic Net Least Squares Regression (OLS)

We begin with a simple regression model, using Elastic Netas our model. We standardise our data to the standard normal distribution in order to work well with the regularisation implemented by Elastic Net. We use the below pipeline 4 to process and fit our data, as well as choose our hyper parameters using a cross-validated (k=5) grid search on the training data. We create dummy indicator variables for day and facility in order to include these in our OLS model.

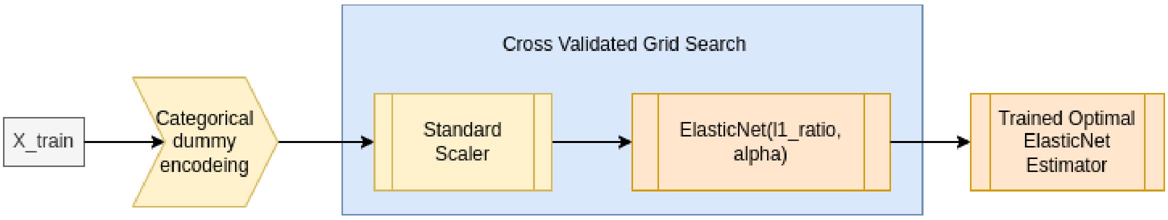


Figure 4: Elastic Net Pipeline

Clearly, our model does not properly distinguish between the three airports and tends to underestimate JFK and LGA volumes, especially at extreme data points. This is likely because the model does not account for the vast differences in distribution between the airports and the differences in the impact of other factors based on the airport facility. Therefore, next, we will explore a model with interaction to remedy this.

B. Elastic Net Least Squares Regression with Interaction

As demonstrated by the test evaluation of our model above and ANOVA analysis III, there is evidence of interaction between the categorical variables as well as between continuous variables and categorical variables. Thus, we will generate interaction terms for our second model. Interaction terms will allow us to model the effect of relevant variables on one another. For example, as above, there are indicators of interaction between day and facility as well as facility and the other numerical features, meaning that we can evaluate and model the effect of the facility on the impact of the other factors in the model. This means that each facility may have a different level/type of impact from arrivals/departures at different points in time, and busy days at each airport may differ. We introduce interaction for all pairs of features in order to capture any such relationships and allow elastic regression to minimize overfitting that may result from over parameterization.

An interesting result of our grid search is the choice of 1 as the l1 ratio for both datasets. This is likely because generating interaction creates a large number of features. As such, l1 regression is optimal. It ensures that the coefficients are sparse and minimizes the number of relevant features to the model, thus preventing any fitting to noise from the vast number of features (which, due to interaction, may be highly correlated with one another).

C. Evaluation and Comparison

In addition to the above plots 57, we also evaluate our models using Mean Absolute Error (MAE) and the Coefficient of Determination *R*2 We selected these metrics because they are both quite easily interpretable in the context of our problem. Below are metric outputs from our two models for both datasets, with information on training and testing performance.

As we can see, the interaction model clearly performs much better than the base model. However, it is evident that while the base model generalizes well with little overfitting (has much more similar training and test *R*2), the interaction model experiences a significant dropoff in performance when testing. This is likely because of the introduction of a large number of interaction terms. While a large of these have 0 coefficients (153/253 and 145/253 have non-zero coefficients) thanks to LASSO, there is still a much larger number of factors contributing to noise, and thus, it is likely that there is a significantly larger amount of overfitting in this model compared to the base model.

Table 4: CV Grid Search Avg. R^2 Scores across Folds for l1 Ratio and α

			l1 ratio								l1 ratio				
			.1	**.5**	.7	.9	1				**.1**	.5	.7	.9	1
Pickups	α	**.1**	.728	**.729**	.728	.728	.727	*Dropoffs*	α	**.1**	**.681**	.681	.680	.679	.678
		1	.641	.690	.713	.727	.727			1	.547	.620	.655	.679	.679
		10	.175	.286	.388	.576	.710			10	.023	.118	.215	.425	.619

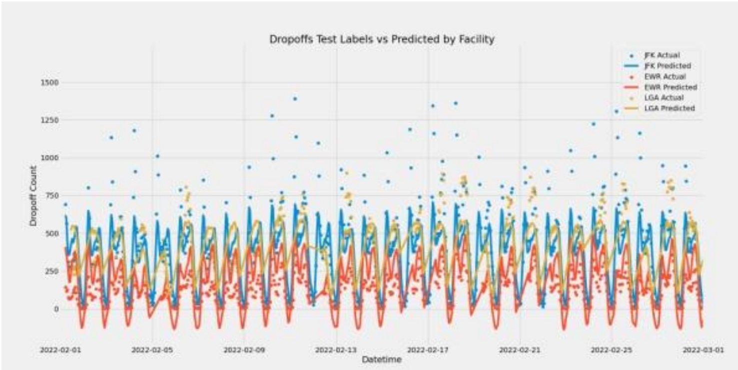


Figure 5: Dropoffs Performance Visualisation on a Subset of Test Data by Facility

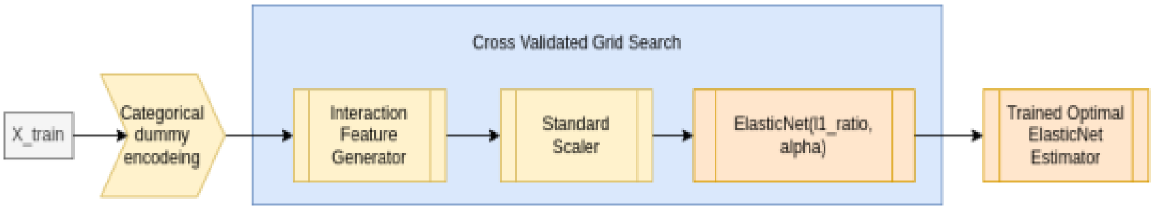


Figure 6: Elastic Net with Interaction Pipeline

On the other hand, as demonstrated by our testing in a 2022 dataset, the interaction model still handily outperformed the base model on both metrics, meaning that despite a larger impact of overfitting, the interaction model generalizes sufficiently well to perform a better predictive model for our problem.

More specifically, as per the $R2$, our interaction model is able to explain 81.6% and 59.1% of variance in the test set's response (dropoff and pickup counts, respectively) through linear relationships with features in the model compared to the non-interaction model's 70.0% and49.4%.Furthermore, the MAE of the interaction model was 87.4 for the dropoff dataset, meaning that it, on average, 'mis-predicts' dropoff counts by 87.4 units. This is contrasted by the much higher errors from not only the base model (116.3, 113.9 for dropoffs and pickups), but also compared to pickups with interaction. This not only demonstrates that the interaction model is less prone to error than the base model, even when extrapolating, but that the model for the pickup dataset is significantly less robust, more prone to overfitting and significantly less accurate a text rapolating. This is further reinforced by the large disparity in training and test $R2$, which is only amplified in the interaction model, which is a strong indicator of overfitting. This can likely be attributed to the test distribution deviating significantly from the test distribution in the given time periods (as the model performs well otherwise, for example, on the dropoff dataset).

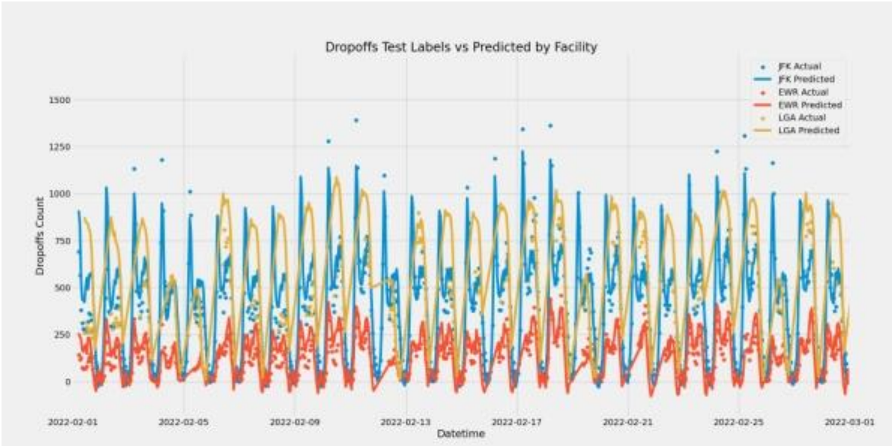
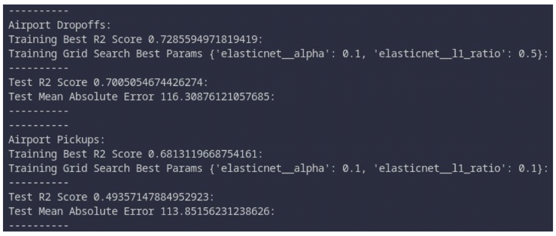
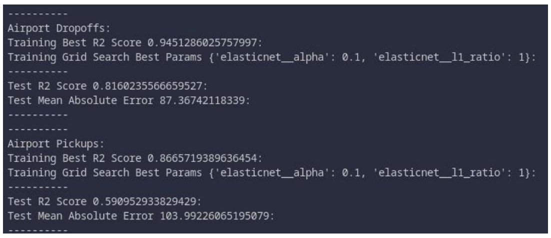


Figure 7: Dropoffs Performance Visualisation on a Subset of Test Data by Facility with Interaction



```
----------
Airport Dropoffs:
Training Best R2 Score 0.7285594971819419:
Training Grid Search Best Params {'elasticnet__alpha': 0.1, 'elasticnet__l1_ratio': 0.5}:
----------
Test R2 Score 0.7005054674426274:
Test Mean Absolute Error 116.30876121057685:
----------
----------
Airport Pickups:
Training Best R2 Score 0.6813119668754161:
Training Grid Search Best Params {'elasticnet__alpha': 0.1, 'elasticnet__l1_ratio': 0.1}:
----------
Test R2 Score 0.49357147884952923:
Test Mean Absolute Error 113.85156231238626:
----------
```

Base



```
----------
Airport Dropoffs:
Training Best R2 Score 0.9451286025757997:
Training Grid Search Best Params {'elasticnet__alpha': 0.1, 'elasticnet__l1_ratio': 1}:
----------
Test R2 Score 0.8160235566659527:
Test Mean Absolute Error 87.36742118339:
----------
----------
Airport Pickups:
Training Best R2 Score 0.8665719389636454:
Training Grid Search Best Params {'elasticnet__alpha': 0.1, 'elasticnet__l1_ratio': 1}:
----------
Test R2 Score 0.590952933829429:
Test Mean Absolute Error 103.99226065195079:
----------
```

(b) Interaction

Figure 8: Model Test & Train Output Metrics

RECOMMENDATIONS

We address the implications of this paper on each of the possible stakeholders identified in the introduction. Firstly, for ride-sharing services whose data is present in the FHVHV dataset, we have demonstrated a strong relationship between flight arrivals and departures and ride counts, indicating a possible driver of demand for airport pickup/dropoff rides. Therefore, we recommend that ride-sharing services incorporate temporally shifted real-time flight data (including flight schedules and history) in their pricing and allocation algorithms in order to maximize profit. This could simply be done by utilizing public or private flight APIs to capture flight data and include it as a factor in any surge pricing models used by such services. Secondly, for ride-sharing service drivers, this model may be used as part of either the ride-sharing service and/or an external application to predict optimal standby locations to maximize ride potential. For example, drivers may be notified when indicators of high future dropoff/pickup volume (e.g. when a lot of flights are departing or leaving an airport), allowing them to plan their shifts to take on airport-faring passengers. This may require additional data to model the origin of airport passengers in order to utilize the dropoff model to inform optimal ride-seeking locations. On the other hand, the pickup model could be used to inform drivers when and which airport to stand by for maximal ride probability. Finally, for customers and customer-facing businesses (e.g., travel agencies), an additional factor in getting the cheapest ticket may beavoiding surge pricing, for example, by avoiding high ride volume times for flights to minimize additional transportation costs. For example, a travel agency may use this model to advise passengers to travel during the night, as the lower number of plane arrivals/departures leads to a lower volume of rides, thus relaxing surge pricing and increasing ride availability on landing/departure.

References

[1] New York City Taxi and Limousine Commission, Tlc trip record data.https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[2] US Federal Aviation Authority, Aviation system performance metrics.