

Estimating Demand for Taxis at LaGuardia Airport

Adam Coviensky¹, Anuj Katiyal¹, Keerti Agrawal¹, Will Geary¹

Abstract

There is often a mismatch between the demand and supply of taxis at New York City's LaGuardia Airport (LGA). This can lead to long wait times for passengers in the taxi queues and for drivers in the taxi hold lots. Better predictions of taxi demand could improve demand-supply balance and, as a result, passenger and driver satisfaction. In this paper, we present several models to predict the hourly number of taxi pickups at LGA. These models are trained on a novel set of instrumental variables derived from publicly available datasets. We implemented and evaluated an ensemble of tree-based models, achieving a Mean-Absolute Error (MAE) of 56.9 and coefficient of determination (R^2) of 0.908. We also implemented a Long Short-Term Memory (LSTM) Recurrent Neural Network, which achieved an MAE of 48.1 and R^2 of 0.921. To enable NYC's Taxi and Limousine Commission (TLC) and The Port Authority of New York and New Jersey (PANYNJ) to interact with our model, we built an interactive web application that allows for stakeholders to generate hourly taxi pickup estimates.

Keywords

Regression, Time-Series Forecasting, LSTM, Transportation, Demand Estimation

¹Data Science Institute, Columbia University, Broadway W 120th St, New York, NY 10027

¹{ac4092, ak3979, ka2601, wcg2111}@columbia.edu

Contents

1	Introduction	1
2	Related Work and Contributions	2
3	Data and Exploration	2
3.1	Flight Data	3
3.2	LGA Yellow Taxi Data	4
3.3	Weather Data	4
3.4	Holiday Data	5
3.5	Taxi and Passenger Wait Times	5
4	Modeling Approaches and Results	5
4.1	Feature Selection and Importance	6
4.2	Machine Learning Models	6
4.3	Results and Discussion	8
4.4	Interactive WebApp	8
5	Limitations	8
6	Conclusion and Future Work	9
	Acknowledgments	9
	References	10

1. Introduction

LaGuardia Airport (LGA) is one of the three primary commercial airports in the New York's metropolitan area, serving approximately 30 million people annually. The airport is situated between Bowery Bay and Flushing Bay on the East River in northern Queens. It consists of four passenger terminals

and two main runways. LGA was opened as a commercial airport in 1939 and since 1947 it has been operated by The Port Authority of New York and New Jersey (PANYNJ), which oversees much of the regional transportation infrastructure within the geographical jurisdiction of the Port of New York and New Jersey.

Large-scale construction at LGA is causing traffic congestion and making transportation into and out of the airport difficult. Taxis report to a holding lot and wait to be dispatched to a taxi queue at one of the terminals. However, there is often a mismatch between the demand for taxis and the supply. As a result, when there is a shortage of taxis, passengers can end up waiting nearly an hour. Conversely, when there are too many taxis, drivers might wait in the hold lot for two or more hours.

In this paper, we implement and evaluate several models to estimate the number of taxi pickups occurring at LGA in a given hour. This serves several purposes, including:

1. **Communication with Drivers:** New York City's Taxi and Limousine Commission (TLC) is an agency of the New York City government that licenses and regulates the medallion taxis. The TLC currently alerts taxi drivers of taxi shortages at LGA after the shortage has already become apparent. Our model and interactive web application could help enable the TLC to employ a proactive communication strategy rather than a reactive one.
2. **Coordination at Airport:** The Port Authority coordinates the flow of taxis at LGA. Upon arrival at LGA,

taxis report to a hold lot and wait to be dispatched to a taxi queue at one of the airport terminals. Our model and interactive web application could help enable the PANYNJ to anticipate pickup demand and act accordingly.

With increasing competition from for-hire vehicle services such as Uber and Lyft, it is imperative for the medallion taxi industry to improve systemic efficiency. This would help drivers maximize their earnings, provide better customer service for passengers, and possibly reduce congestion.

In this paper, we focus on three main tasks: (i) Exploratory data analysis in order to gain useful insights from the data; (ii) developing a high performance regression model for estimating hourly taxi pickups; and (iii) developing an interactive web application allowing stakeholders to engage with our model. In task (i), we explore and analyze relevant datasets regarding taxi trips, flights, weather, holidays, date, time and seasonality. Task (ii) consists of two important sub-tasks: (a) Exploring the features that are relevant for analyzing the taxi demand at LGA and determining their predictive power; and (b) Exploring various supervised machine learning techniques to model this problem. This includes experimentation with different Machine Learning and Deep Learning models and parameter tuning using grid search and cross-validation. Finally, in task (iii), we built an interactive web application to provide an intuitive framework for technical and non-technical stakeholders to interact with our model and react accordingly.

The paper is organized as follows: Section 2 (**Related Work and Contributions**) describes previous works on taxi pickup prediction which can help in our estimates for taxi hourly demand estimates in New York City; the Section 3 (**Data and Exploration**) describes and analyzes the datasets and obtained features from these datasets which are used to build our models; Section 4 (**Modeling Approaches and Results**) discuss the experiments and provide insights on the features used for various regression models tried out for the task. The section also discusses the performance of the various models tried and present visualizations comparing the actual and predicted number of taxi pickups for the final models; Section 5 (**Limitations**) highlights the limitations of our approach and recommends improved data collection and sharing policies for consideration by the Port Authority and TLC; and finally Section 6 (**Conclusion and Future Work**) concludes the paper and lays the path for the future direction.

2. Related Work and Contributions

Over the last decade, there has been considerable work done on similar grounds of predicting taxi pickups in New York City. Tong et al [1] discuss a spatio-temporal model to predict taxi demands per unit time and region on large-scale online taxi platforms. They propose a high-dimensional linear regression model LinUOTD. Following their work, we have treated the time-related information such as hour of the day, the day of the week and the month as features in our regression

model. The paper further inspired us to pull more data and create higher dimensional features by also including interaction features for our model.

The research work by Kamga et al [2] analyzes the temporal and weather related changes in the taxi service supply and demand equilibrium. Their study finds that under all rainy conditions, regardless of severity, drivers get more fares in the city. Interestingly, even snowy conditions were found to have little to no effect on taxi fares over a small sample. Thus, they are less likely to come to the airports to wait for a fare creating a greater mismatch in supply and demand.

Anwar et al [3] present an app, ChangiNOW, to optimally queue taxis at each terminal in Singapore. Their statistical model is based on a two-queue system of passengers and taxis. They model the passengers at each terminal as a time-varying poisson distribution using historical data on the number of passengers arriving at each terminal every 15 minutes. They also estimate taxi arrival times based on taxi medallion numbers and real-time GPS coordinates. This illustrates what can be accomplished by TLC with more robust data collection processes. It also highlights the importance of data-sharing policies between LGA, Port Authority, and TLC.

Predicting New York City Taxi Demand [4] analysis focuses on predicting the number of taxi pickups given a one-hour time window and a location within New York City. This blog helped us develop intuition about how we can approach our problem as a regression problem. The important features used in the blog helped us understand the features that can play a major role in our analysis. They attempted regression techniques on their feature sets and published the RMSE and R^2 scores. Our analysis, results and metric followed very similar methodologies after framing the estimation as a regression problem. Other important blogs like [5] and [6] also tackled similar problems about analyzing taxi demands at JFK airport while answering many other issues related to the role of bias in the datasets and helped us better understand the demand estimation problem at hand.

The application of deep learning to multivariate time series data has gained attention in a wide variety of contexts, including financial markets, weather and traffic forecasting. Ma et al [7] demonstrate the use of long short-term memory (LSTM) neural network for traffic speed prediction and show that the LSTM neural network delivers superior performance and stability over other non-parametric approaches, which motivated us to frame our problem as a time series demand forecasting problem.

3. Data and Exploration

The flow chart in Figure 1 presents the datasets used to estimate the demand of taxis at the airport. We extract our target variable, the number of yellow taxi pickups occurring at LGA per hour, from taxi trip records released to the public by the NYC Taxi and Limousine Commission (TLC). Based on the literature survey and our intuitions about what could affect the demand for taxis, we collect and create several

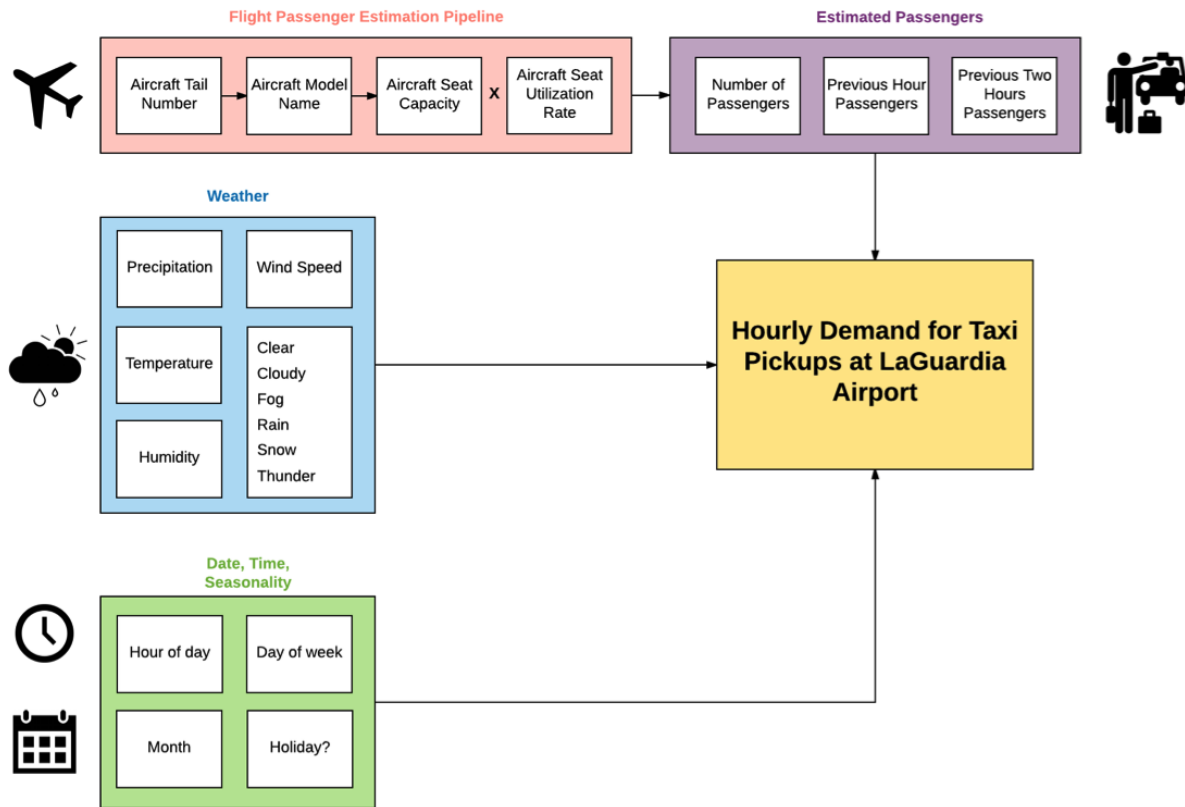


Figure 1. Feature Flow Chart for Machine Learning Models

additional datasets. We derive predictor variables from the datasets regarding flights arriving at LGA, seat capacities and seat utilization rates associated with those flights, weather data and holiday data.

3.1 Flight Data

We devised a novel approach to estimate the number of passengers arriving at LGA on an hourly basis. We start with data from the Bureau of Transportation Statistics (BTS), an agency housed within U.S. Department of Transportation (USDOT). The BTS publishes a database called Airline On-Time Performance Data. This table contains on-time arrival data for non-stop domestic flights by major airline carriers, and provides such additional items as departure and arrival delays, origin and destination airports, tail numbers, scheduled and actual departure and arrival times, cancelled or diverted flights, and other flight related information. Flight records were downloaded from this database on a monthly basis from January 2014 through July 2017, filtered to only include flights arriving at LaGuardia, and concatenated into a single table. Figure 2 below shows the average number of arrivals at LGA airport by hour since January 2014.

Next, we rely on the Federal Aviation Administration (FAA)’s N-Number Inquiry tool, which enables aircraft infor-

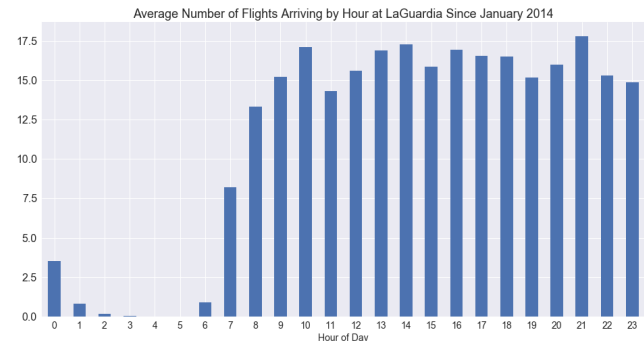


Figure 2. Average Flight Arrivals at LGA by Hour Since 2014

mation look-up by tail number. We wrote a Python script to automate querying the FAA N-Number database and built a dataset associating tail numbers (i.e. N14249) with aircraft model names (i.e. Boeing 747) for specific airline carriers. Then, we rely on SeatGuru, a consumer-focused flight comparison tool that provides seat capacity data. We wrote a python script to automate querying SeatGuru and built a dataset associating aircraft model names with passenger seat counts. Then, we use AviationDB’s Airline Traffic Query database,

which contains monthly seat utilization rates segmented by airline and airport. We associate monthly average seat utilization rates with airlines. Finally, we multiply incoming seat capacities by utilization rate to estimate incoming passengers.

Figure 3 shows the monthly average seat utilization for JetBlue, SouthWest, American, Delta and United flights arriving at LGA since 2014. Seat utilization for these major airlines hovers between 70% and 90%, with clear seasonal dips in January and February.

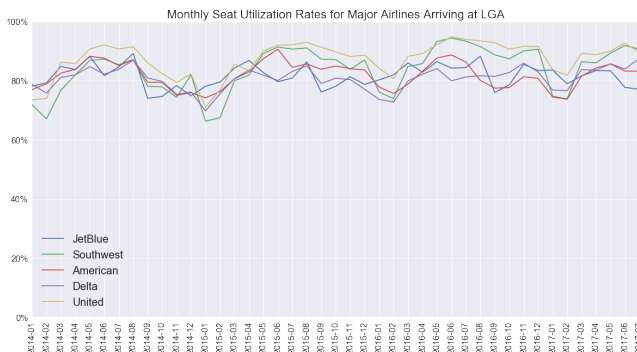


Figure 3. Monthly Average Seat Utilization Rates for Major Airlines Arriving at LGA Since 2014

3.2 LGA Yellow Taxi Data

The NYC Taxi and Limousine Commission (TLC) has released monthly taxi trip records to the public since 2009. We will use all months from 2014 to present day to build and test our model. The records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The yellow taxi data is collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab Livery Passenger Enhancement Programs (TPEP/LPEP). The trip data was not created by the TLC, and TLC makes no representations as to the accuracy of these datasets. Figure 4 shows a visualization highlighting the yellow taxi pickups in NYC during the month of January 2016, with highlighted region of LGA considered for our analysis.

The most important features in this dataset for our analysis are the pickup time and the pickup location. Aggregating by the total number of pickups per hour at LGA, we obtain our proxy for the taxi demand at the airport for any given hour. This serves as our target variable to train our machine learning models.

3.3 Weather Data

We obtained the weather data using the OpenWeatherMap History Bulk API. The API provides hourly NYC weather data that includes temperatures, pressure, wind-speed and weather categories from January 1st 2012 until October 2017. The obtained weather categories were analyzed and combined to obtain the major weather categories. This helped us reduce the number of weather main categories to the most important ones

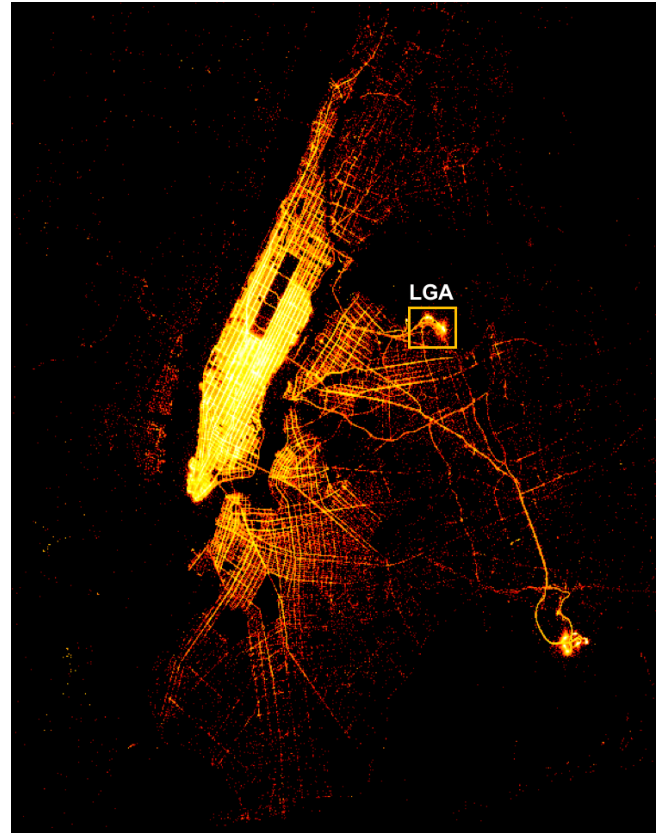


Figure 4. Visualization showing every yellow taxi pickup in NYC during the month of January 2016

namely clear, clouds, fog, rain, snow and thunderstorm, which further helped in reducing the number of one-hot encoded categories that get introduced in the future ML models. Another important benefit of doing this analysis was the removal of outliers, or removal of weather categories which were recorded only for a few hours in the last 4 years. Also, there were specific hours where multiple categories were recorded which was also considered while doing one-hot encoding for the weather categories. The few missing values for the weather types were filled by taking the weather conditions prevailing in the previous hours.

Further, an analysis of the different features related to the weather revealed that the main weather categories like clouds, rain, snow and thunderstorms do not contain enough mutual information with our target variable. Despite extreme conditions such as snow and thunderstorms affect the estimated number of incoming passengers and the number of taxi pickups drastically, they still don't act as important features in predicting taxi demands. The most important features related to weather affecting our taxi demands were temperature, humidity and wind speed at every hour.

The OpenWeather API did not provide us with precipitation data, which seems to play an important role in predicting taxi demands as understood from the literature review [4]. We thus obtain the historical data on hourly precipitation from

the Iowa State University Environmental Mesonet. The data obtained contains hourly totals of precipitation to a hundredth of an inch for the LGA weather station for the years 2014 to 2017. The precipitation data also did not act as an important feature in our ML models. We believe the flight data ended up being a major predictor for our target variable and reduced the feature importance for all weather related features.

3.4 Holiday Data

We have considered holidays as a feature in our analysis of the taxi demand at LGA. The federal holidays are obtained through Python Pandas API and it contains the US Federal Holidays. The hypothesis here is that during the holidays, people tend to travel more and hence it can lead to an increase in the taxi demand at the airport. Analyzing this hypothesis, we plot the average number of taxi pickups for every hour during the federal holidays and non-holiday days (shown below in Figure 5) and observe that there is a slight increase number of taxi pickups after noon on the days of federal holidays.

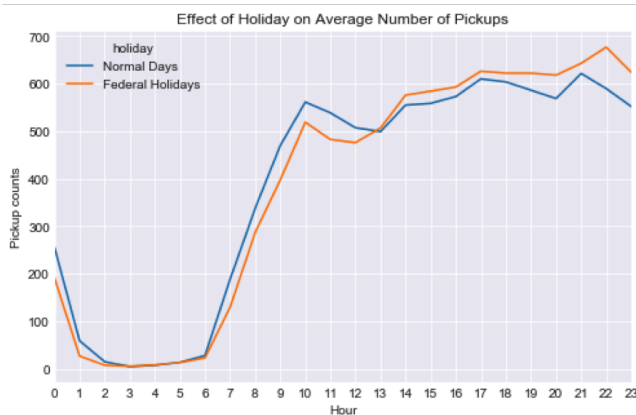


Figure 5. Effect of Federal Holidays on Number of Pickups

3.5 Taxi and Passenger Wait Times

We analyzed proprietary data provided by TLC including the number of taxis waiting in the hold lot and average passenger wait times on an hourly basis. Our goal was to understand the demand-supply gap in the recent months at LGA airport. For each hour, we aggregate the data from all the parking lots for the total waiting taxis and we also aggregate the total passenger wait times at the terminals and then normalize them to visualize the relationship. The Figure 6 below shows the number of taxis waiting in parking lots compared to the passengers wait time for every hour shown in Figure 7 below. Finally, we have estimated the number of taxis waiting at the end of each hour and used them as hourly biases, which were added to our model estimates and shown as the final results in our web application.

The figure 6 suggests that the taxis come to LGA for the morning flights, but then prefer not to come back later (because they can get better fares in Manhattan) resulting in longer wait times for passengers late afternoon. From

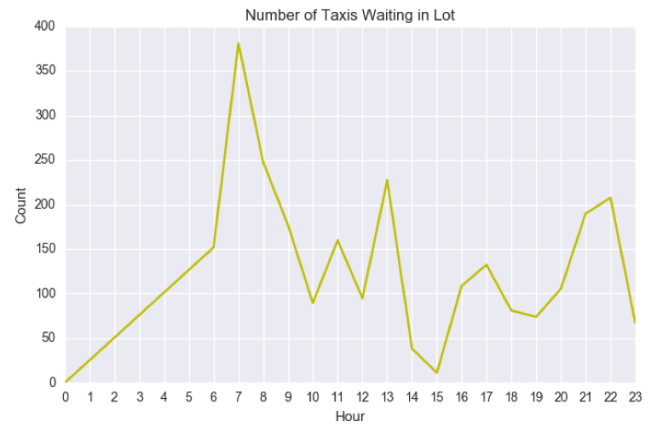


Figure 6. Average number of taxis waiting in hold lot per hour

the figure 7, it is evident that more taxis are needed at LGA after 1pm to match the demand and reduce wait times for the passengers. We also observe the hours when there are more than 100 taxis waiting in the holding lot but the passenger wait times still being over 20 minutes. This does indicate a civil engineering bottleneck occurring at LGA, limiting the efficient dispatching of taxis from the hold lot to terminal taxi queues.



Figure 7. Average passenger wait time on an hourly basis at LGA

4. Modeling Approaches and Results

In this section, we highlight the features which were used for our estimation of taxi demand at LGA and their importance while considering the mutual information between the features and target variable. After that, we talk about the machine learning and deep learning approaches attempted for estimation of taxi demands on an hourly basis. We further present the important results by comparing the models based on the evaluation metrics, specifically, R^2 and MAE . The final part of this section talks about the interactive webapp which was created as a deliverable for usage by New York City Taxi

and Limousine Commission.

4.1 Feature Selection and Importance

This section lists the features and how they were encoded before being used as input to our regression models.

1. *HourOfDay*: The hour of the day was extracted from the associated timestamp data. For the final ML models, the Hour of Day gave the best results by being encoded as continuous values.
2. *DayOfWeek*: The day of the week was extracted from the associated timestamp data. For the final ML models, the day of week gave the best results on being encoded as continuous values.
3. *Month*: The month data was extracted from the associated timestamp and gave the best results on being treated as continuous values.
4. *Weather*: The weather data was analyzed and divided into the Main Weather Categories. The main weather categories were one hot encoded with several hours having more than one weather type recorded adjusted accordingly.
5. *Holiday*: The Federal holiday data was obtained and used for the ML models. The data was treated as a boolean value, which is just an indicator of whether the aggregated hour belong to a holiday or not.
6. *Passengers*: The estimated number of flight passengers arriving at LGA per hour resulting from our flight passenger estimation pipeline.
7. *cancelled_departing_flights*: The average number of flights cancelled from LGA per hour. We expect taxi demand at the airport to increase due to flight cancellations.
8. *Avg_Delay_Arriving*: The average delay in minutes for flights arriving at LGA by hour. The average delays of arriving flights are highly correlated to the taxi demand at LGA.
9. *Prev_hour_Passengers*: The previous hour passengers simply refer to the count of the arriving passengers at the airport in the last hour. This feature plays a major role, which is indicative of the fact that the people tend to spend about an hour at the airport before taking a taxi (Figure 8).
10. *Prev_2hour_Passengers*: The previous two hour passengers simply refers to the count of the passengers arriving at the airport two hours prior. This feature also plays a major role as shown in Figure 8.

Figure 8 shows the importance of the top 10 features in our model using mutual information with the number of taxi pickups per hour. We see that the flight data is a critical part of our model. Also, as expected, the hour of day encoded as a continuous variable was found to be the most important feature, given the cyclical nature of the data.

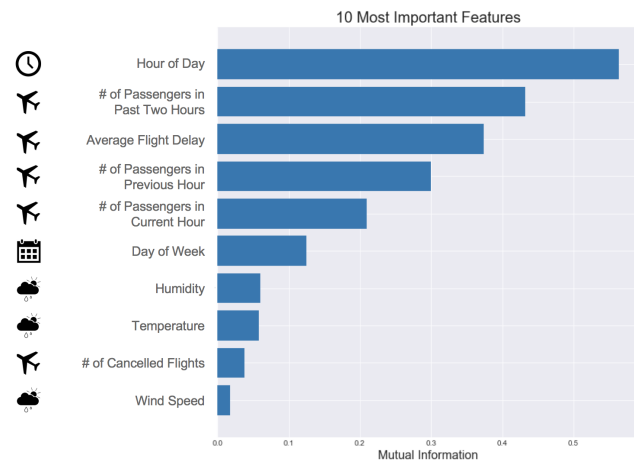


Figure 8. Ten features with highest mutual information with target variable

4.2 Machine Learning Models

Each model in this section was tuned using an exhaustive grid search with 5 fold cross-validation. We tuned the models to maximize the mean R^2 score amongst the cross-validation folds. Finally, we tested each model on a held out test set (15% of total data). Our baseline model is a linear regression model which predicts the number of taxi pickups from LGA with a R^2 of 0.698. To improve upon this baseline, we experiment with different regression models described below.

Tree Based Regressors The tree-based models tried for our regression analysis were a Random Forest Regressor (RF), Gradient Boosting Tree Regressor (GBR) and XGBoost Regressor (XGB). The best parameters for the models were obtained after extensive grid searches and resulted in GBR having a maximum depth of 10 nodes, maximum number of features considered at each split as 6 and 200 estimators. This GBR model yielded a test R^2 of 0.904, which is a significant improvement over regularized regression models. The best parameters for RF regressors resulted in deeper trees with maximum depth of 21 nodes and also 200 estimators. This time however, each estimator was warm started using the output of the previous tree, resulting in best test R^2 of 0.900. For the XGBoost regressor, we used extensive grid search and obtained for best tuned parameters having the maximum depth 7, column sample by level as 0.6, column sample by tree as 0.7, gamma as 1.1, learning rate as 0.1 and number of estimators set to 200. The XGB based regressor resulted in the test R^2 of 0.904.

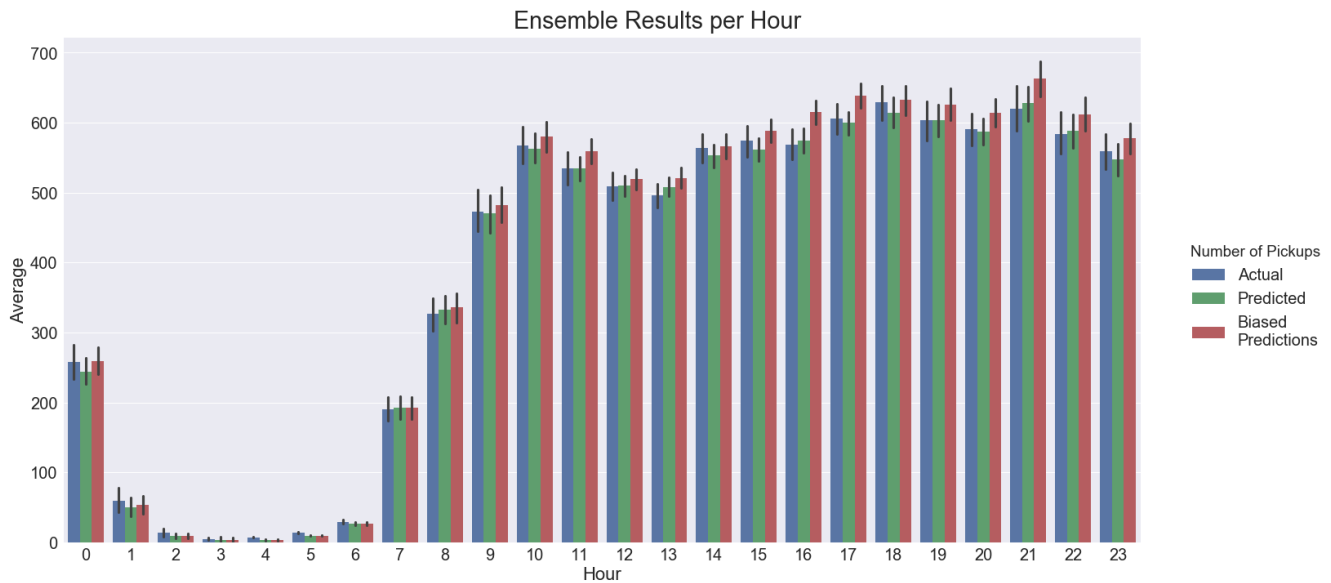


Figure 9. Ensemble model Predictions vs. Actual taxi estimates at LGA

Ensembles We created an ensemble of the above models, namely GBR and RF and XGB by stacking them and using their predictions in a final meta lasso regression model. The lasso meta regressor was trained using 5 fold cross-validation as shown in the figure below. Each of the initial models were trained on 4 folds of the data and then used for predictions on the final, validation fold. The lasso regressor was then trained using the predictions from the 3 models and the true values in the validation fold. This was repeated 5 times leaving out each of the folds separately. Finally, it yielded an R^2 of 0.908 on the test set. We also observed the coefficients it gave to each of the individual models: 0.19 for RF, 0.39 for the random forest and 0.50 for XGB. The final ensemble increased the R^2 by 0.004 over the XGB method alone.

Figure 9 shows the average actual number of taxi pickups by hour of day, compared to our model's original predictions and predictions adjusted to incorporate hourly biases based on passenger queue sizes. Considering that the NYC Taxi and Limousine Commission wanted the model to over-predict rather than under predict the taxi pickups at every hour of the day, after including biases based on Taxi and Passenger Wait times, we observe that our models work really well. The architecture for the ensemble model is shown in Figure 10.

Deep Learning Model Long Short-Term Memory, or LSTM, is a recurrent neural network comprised of internal gates which allow the model to be trained using back-propagation through time. We use the Python module Keras to build an LSTM for multivariate time series regression.

Preparing our dataset for the LSTM required framing the problem as a supervised learning problem. We frame the supervised learning problem as predicting the number of taxi pickups at the current hour (t) given the feature variables at the previous 23 time steps ($t-1, t-2, \dots, t-23$). 23 lagging timesteps were chosen to capture the ebb and flow of a single day. We

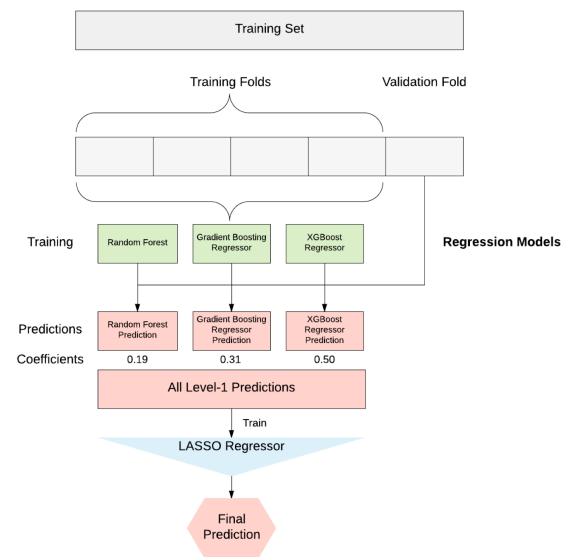


Figure 10. Architecture of our ensemble model

then normalize all features using Min-Max scaling. We experimented tuning hyper-parameters including the number of neurons, layers, epochs and batch sizes. Our best performing model has a simple architecture: a single LSTM layer with 100 neurons and final dense layer with linear activation, with MAE as our loss function and Adam as our optimization algorithm. Figure 11 shows the LSTM training and testing loss functions converging over 100 epochs.

The scatterplot in figure 12 shows the LSTM predictions vs. the actual values over six months of test data, from January 2017 through June 2017. The LSTM performed well, achieving R^2 of 0.921 and MAE of 48.1.

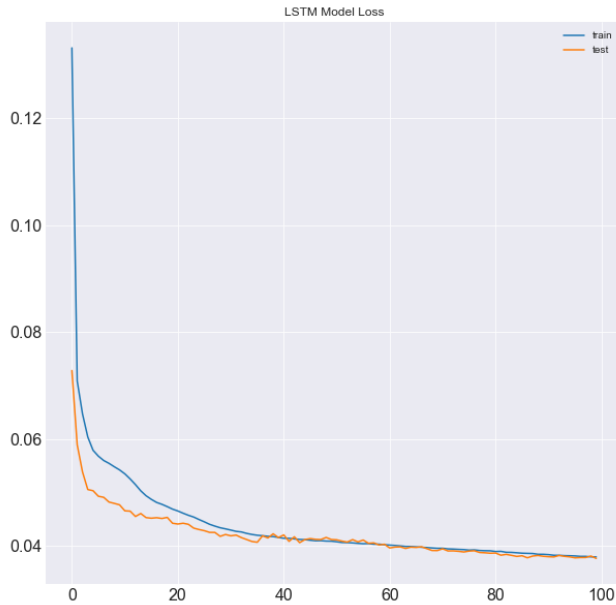


Figure 11. LSTM training and testing loss functions

4.3 Results and Discussion

Table 1 summarizes the best results for each model, obtained by training on 85% of the data points and testing on the remaining 15%. The hyperparameters used for each model were determined using extensive grid-search over the parameter values.

The table 1 compares the results from each of our models. The improvement column indicates the improvement compared to the test score of the baseline linear regression model. We see that the simple regression based models performed the worst. Adding regularization to the baseline linear regression model improved our test score by 4.3%. The Random Forest, Gradient Boosted Trees and XGBoost tree-based models perform fairly well on our data and exhibit a significant improvement, approximately 29%, over the baseline method. By stacking the regression models, the final model improves slightly compared to the individual tree-based models. Finally, the LSTM significantly improved results having achieved nearly 32% improvement over the baseline model. It will be the final model used for the predictions we present to the TLC at the conclusion of the project.

4.4 Interactive WebApp

The webapp created as a part of the project was one of the major contributions, as it will enable the end user to easily obtain predictions from the best trained regression models. Figure 13 shows the layout of the interactive webapp which is built using Dash in Python. The web application aims to provide an easy to use interface to input features related to the date, hour of the day, the current temperature and precipitation conditions along with an input area which can take in multiple weather categories. Based on the inputs provided, the webapp aims to predict the number of taxi pickups for the input date.

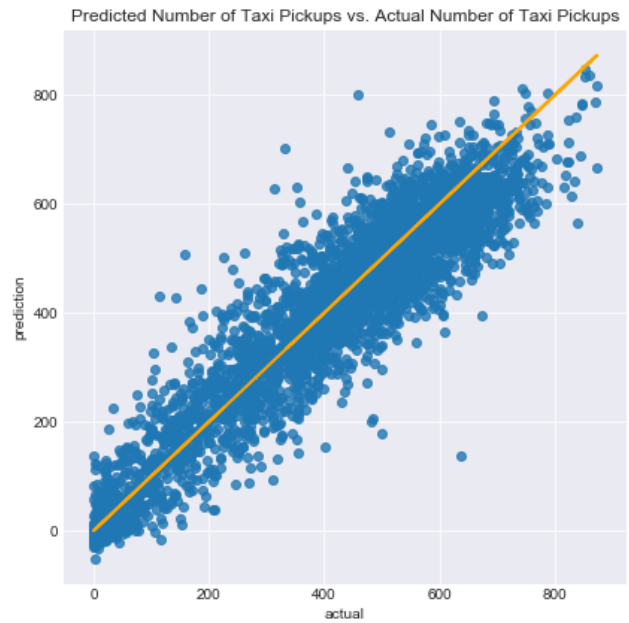


Figure 12. LSTM predictions vs. actuals over the test set

This prediction is based on our ensemble model with hourly biases, which performs nearly as well as the LSTM model and also provides a slight over estimation of demand at the airports, as requested by the NYC Taxi and Limousine Commission (TLC). The webapp also predicts the number of taxis needed at LGA for the next 6 hours. NYC TLC and Port Authority can utilize this web based application to observe the pick up trends for the near future and inform taxi drivers in a proactive way to decrease wait times for both the passengers in line and taxi drivers in the hold lots. The link for the demo of the app can be found here [8].

Estimating Taxi Demand at LaGuardia Airport

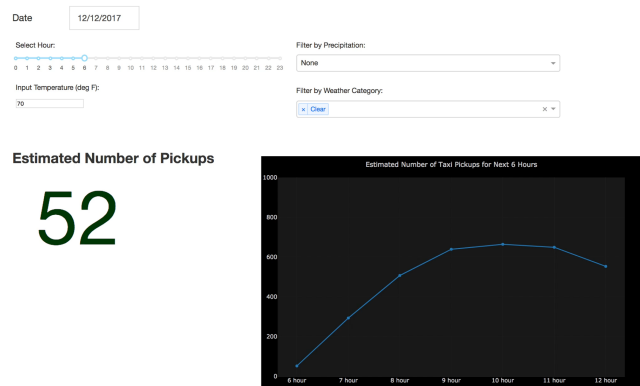


Figure 13. Taxi Demand Prediction WebApp

5. Limitations

Overall, our models for predicting taxi demand at the LGA airport perform very well. However, we do believe there are

Table 1. Table of Results

Model	MAE	Test Score (R2)	Improvement (%)
Baseline(Linear Regression)	117.34	0.697	-
Ridge Regression	109.74	0.727	4.3
KNN Regressor	84.57	0.809	16.1
SVM Regressor	83.25	0.813	16.6
Random Forest Regressor	59.30	0.900	29.1
Gradient Boosted Trees Regressor	58.04	0.904	29.7
XGBoost Regressor	58.83	0.904	29.7
Stacking Ensemble	56.9	0.908	30.2
Long Short Term Memory	48.1	0.921	32.1

multiple limitations we faced while working through this problem. We would like to highlight a few of the limitations here and make recommendations for future research on problems related to analyzing taxi demand at airports.

We arrive at our estimate for the number of incoming flight passengers per hour using several disparate datasets and a number of assumptions. For instance, we map incoming tail numbers from the US Bureau of Transportation Statistics to aircraft models from the Federal Aviation Administration. We then map aircraft models to passenger capacities according to SeatGuru. Finally, we estimate passenger arrivals by multiplying passenger capacities by seat utilization percentage from AviationDB. These seat utilization percentages are available as monthly averages, segmented by airline and airport. So, while we believe they provide the basis for reasonable assumptions, they are assumptions nonetheless. Given how well the features extracted from this flight passenger estimate pipeline performed in our models, we recommend that the NYC Taxi and Limousine Commission establish systemic data-sharing practices with LGA and Port Authority. LGA could provide exact incoming passenger counts each hour at each terminal. Using the exact passenger counts will allow TLC to improve the model and employ a data-driven taxi demand strategy.

Another major limitation here is to consider the target variable to be the number of taxi pickups as a proxy for the taxi demand at airport for every hour. Since taxi demand is affected by the number of passengers actually wanting a taxi and entering the queue, rather than only those passengers who ultimately take a taxi, the proxy used here is not a true reflection of taxi demand. With data on passenger queues, passenger wait times and taxi wait times, we could predict taxi demand more accurately. It would thus be very useful for Port Authority to collect data on the actual number of incoming passengers entering the taxi queue at each terminal for each hour.

The recent data provided by TLC for number of taxis waiting and passenger wait times are only available for the most recent 3 months. We went ahead and explored the data but cannot use this data in our models for taxi demand prediction for the last three years. Though, if such datasets continue to be accrued, can really help in improving the estimation of taxi demands at the airports on a terminal basis.

Following the work done by Anwar et al. at Changi airport in Singapore [3], we recommend TLC to collect data on real-time GPS locations as well as including medallion numbers or unique identifiers. This will allow TLC to accurately measure taxi wait times and how long it takes drivers to get to the airport allowing them to better estimate the supply of taxis over the next few hours.

6. Conclusion and Future Work

In this research work, we have (i) built a model to predict the number of pickups at LGA; (ii) analyzed the most predictive features in our regression models for estimation of taxi demands at LGA; and (iii) outlined new recommendations for improved data collection processes by LGA, TLC, and Port Authority. The models implemented along with the included biases perform very well and with the future implementation of strategic data collection and data sharing policies between the governmental agencies, our models have the potential to even more accurately predict the demand for taxis at LGA.

Furthermore, the developed models could be implemented at a terminal level, addressing the problem of supply-demand imbalance between the taxi passenger queues at each terminal at LGA. Finally, along with the models, two of the major contributions of this project are the creation of the pipeline to estimate the number of incoming passengers at LGA on an hourly basis using flight data, and the creation of the interactive web application to predict the future taxi demand at LGA given the input parameters. We hope that future research on predicting taxi demands can utilize our contributions from this project. Finally, we believe that the methodology can be replicated for the other airports in New York City metro area, JFK and EWR, or even elsewhere in the world.

Acknowledgment

We would like to thank the Data Science Institute at Columbia University for the opportunity to work on the Capstone project. We would also like to thank our mentor Eleni Drinea along with the TA Yogesh Garg for their involvement and guidance throughout the project. Finally, sincere thanks to the members of New York City Taxi and Limousine Commission who

have been involved in our discussions, and instrumental in completing this project.

References

- [1] Yongxin Tong, Yuqiang Chen, Zimu Zhou, Lei Chen, Jie Wang, Qiang Yang, Jieping Ye, and Weifeng Lv. The simpler the better: A unified approach to predicting original taxi demands based on large-scale online platforms. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17*, volume Part F129685, pages 1653 – 1662. ACM, 2017-08. 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2017); Conference Location: Halifax, NS, Canada; Conference Date: August 13-17, 2017.
- [2] Camille Kanga, M. Anil Yazici, and Abhishek Singhal. Hailing in the rain: Temporal and weather-related variations in taxi ridership and taxi demand-supply equilibrium. 01 2013.
- [3] Afian Anwar, Mikhail Volkov, and Daniela Rus. Changinow: A mobile application for efficient taxi allocation at airports. In *16th International IEEE Conference on Intelligent Transportation Systems, ITSC 2013, The Hague, The Netherlands, October 6-9, 2013*, pages 694–701, 2013.
- [4] Shuo Zhang Yunrou Gong, Bin Fang and Jingyu Zhang. Predict new york city taxi demand. <https://nycdatascience.com/blog/student-works/predict-new-york-city-taxi-demand/>, 2016.
- [5] Chris Whong. Should i stay or should i go? nyc taxis at the airport. <http://bit.ly/2pfamIC>, 2014.
- [6] Todd W. Schneider. Analyzing 1.1 billion nyc taxi and uber trips, with a vengeance. <http://bit.ly/2BAHav9>, 2015.
- [7] Xiaolei Ma, Zhimin Tao, Yinhai Wang, Haiyang Yu, and Yunpeng Wang. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C: Emerging Technologies*, 54(Supplement C):187 – 197, 2015.
- [8] Predict taxi demand webapp. <https://vimeo.com/247898437>.