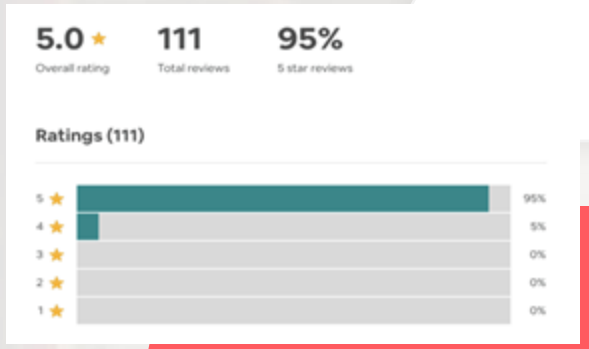
Introduction

Data Sources

Analysis and Results

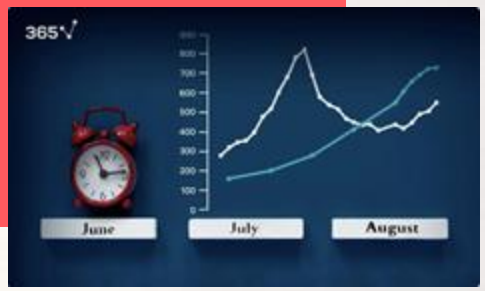
Business Impact, Recommendations and Implications



Data Sources

Three complementary datasets from Inside Airbnb were used to understand pricing, availability, and guest behavior.

1. **Listing Data**
   - Describes each Airbnb property and host in Washington, DC.
   - 6,257 listings with 79 property and host attributes.
   - Potential use: Market segmentation and identifying drivers of demand.
2. **Calendar Data**
   - Provides daily availability and pricing for every listing.
   - 2M+ entries across 7 dimensions.
   - Potential use: Estimate occupancy, derive price elasticity, and assess seasonality or demand shifts.



Data Sources

Review Data

Identifies guests through their **reviews**.

415,353 unique reviews across **6 dimensions**

Methodology

CRISP–DM Framework

- Standard process for developing data mining and predictive models.
- Steps: Business Understanding → Data Understanding → Preparation → Modeling → Evaluation → Deployment.

python™

- Core tool for data wrangling, visualization, and predictive modeling.
- Enabled statistical analysis and regression modeling to explore price–demand relationships.

RStudio®

- Used for trend and seasonality decomposition.
- Helped isolate underlying demand patterns and remove cyclical effects from the data.

ChatGPT

- Applied to augment qualitative insights from text data.
- Classified listings based on their description fields.
- Implement a Revenue Optimization Strategy

Analysis and Results

Modeling Objectives

- **Goal:** Estimate how price and listing attributes influence realized demand
- **Demand Proxy:** Length of stay and overall booking probability among historically completed Airbnb bookings
- **Key Considerations:**
  - Some of the available data reflects future supply, not guest behavior. Listing ranged [March 2025 - March 2026]
  - Restricted modeling to listings truly available for booking

Models Applied and Split Strategy

- **Four Functional Forms using Cross-Validation Lasso Regularization**
  - **Linear:** Stay Length = f(Price + Features)
  - **Exponential (Log-Linear):** ln(Stay Length + 1) = f(Price + Features)
  - **Constant Price Elasticity (Log-Log):** ln(Stay Length + 1) = f(ln(Price + 1), Features)
  - **Logistic Model:** ln(Overall Booking Probability/(1 - Overall Booking Probability + Epsilon) = f(Price, Features)
  - **Train, Validation, and Test Splits:** Time-based split:
    - ***Train*** ← Start Date (March 13 2025) - May 1 2025 Listings; ***Dims*** = (796,498 rows, 96 cols)
    - ***Validation*** ← May 2 2025 - July 1 2025 Listings; ***Dims*** = (570,945 rows, 96 cols)
    - ***Test*** ← July 2 2025 - Last date in which bookings occurred; ***Dims*** = (280,449 rows, 96 cols)

Predictors Used

- **Regressors Used Included**
  - **Pricing:** Nightly price
  - **Capacity:** Bedrooms, beds, accommodates, Room Type, Property Type
  - **Policies:** Minimum nights, maximum nights
  - **Forward Looking Availability:** Frequency of Availability over 30 days, 90 days, and 365 Days
  - **Reputation:** Overall rating, cleanliness, location, value, review count
  - **Temporal:** Weekend indicator, seasonal sine–cosine variables

Purpose: Capture pricing power, listing quality, supply pressure, and cyclical demand

Removal of Outlier Instances

Initial Technique (Effects of Third Pass Through

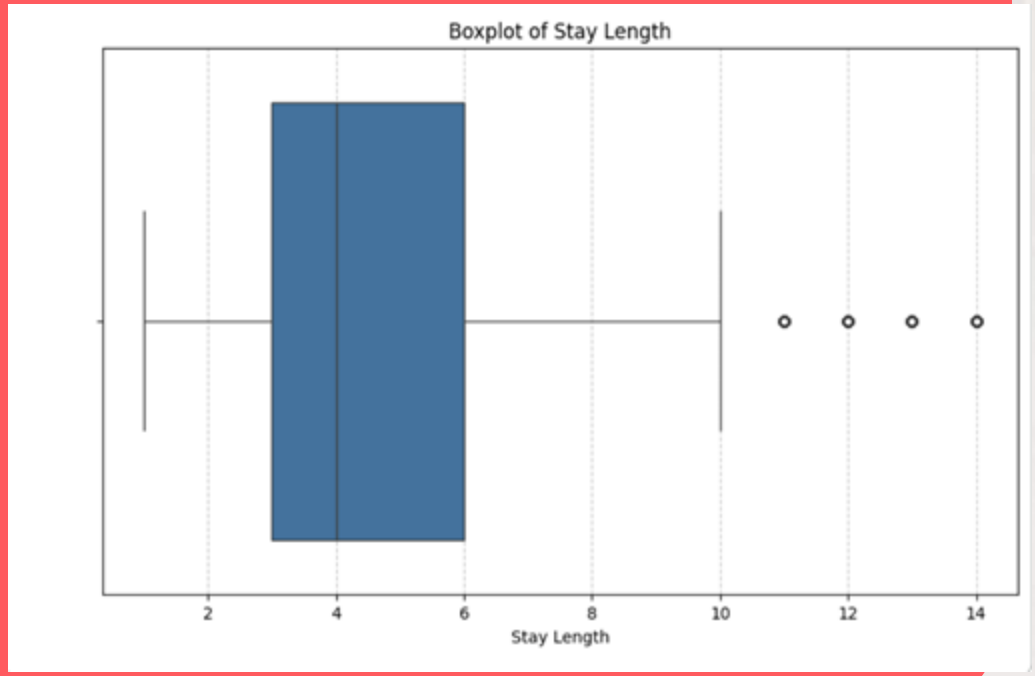
```
# Third pass of outlier removal
drop_rows_third_pass = booking.assign(
    stay_length_zscore = lambda x: np.abs((x.stay_length - x.stay_length.mean())/np.std(
        x.stay_length, ddof=1))
).loc[lambda x: np.abs(x.stay_length_zscore) >= 3].index

booking.drop(drop_rows_third_pass, axis=0, inplace=True)
```

```
booking.stay_length.describe()
```

	stay_length
count	1.982834e+06
mean	6.337001e+00
std	6.051852e+00
min	1.000000e+00
25%	3.000000e+00
50%	4.000000e+00
75%	7.000000e+00
max	3.800000e+01

dtype: float64

Ultimately Applied



Model Performance

Model Performance Comparison:

	RMSE	MAPE	WMAPE
Linear Model	1.889772	0.498545	0.360110
Exponential Model	1.915795	0.471634	0.359438
Constant Price Elasticity Model	1.915102	0.471546	0.359340
Practical Logistic Model	0.135964	0.279769	0.196993

- **RMSE**: Logistic's models RMSE is lowest by a wide margin, partially because it is predicting overall booking probability (a fraction between 0 and 1), while other models are predicting stay length (an integer value). Therefore, not directly comparable
- **MAPE**: Logistic model has smallest average percentage difference between predicted and actual values in its response variable, again by a very wide margin. For models predicting stay length, the CPE model has lowest MAPE closely followed by the Linear Model.
- **WMAPE**: Again, Logistic model has lowest WMAPE, followed by the Exponential and Linear Models (in that order).

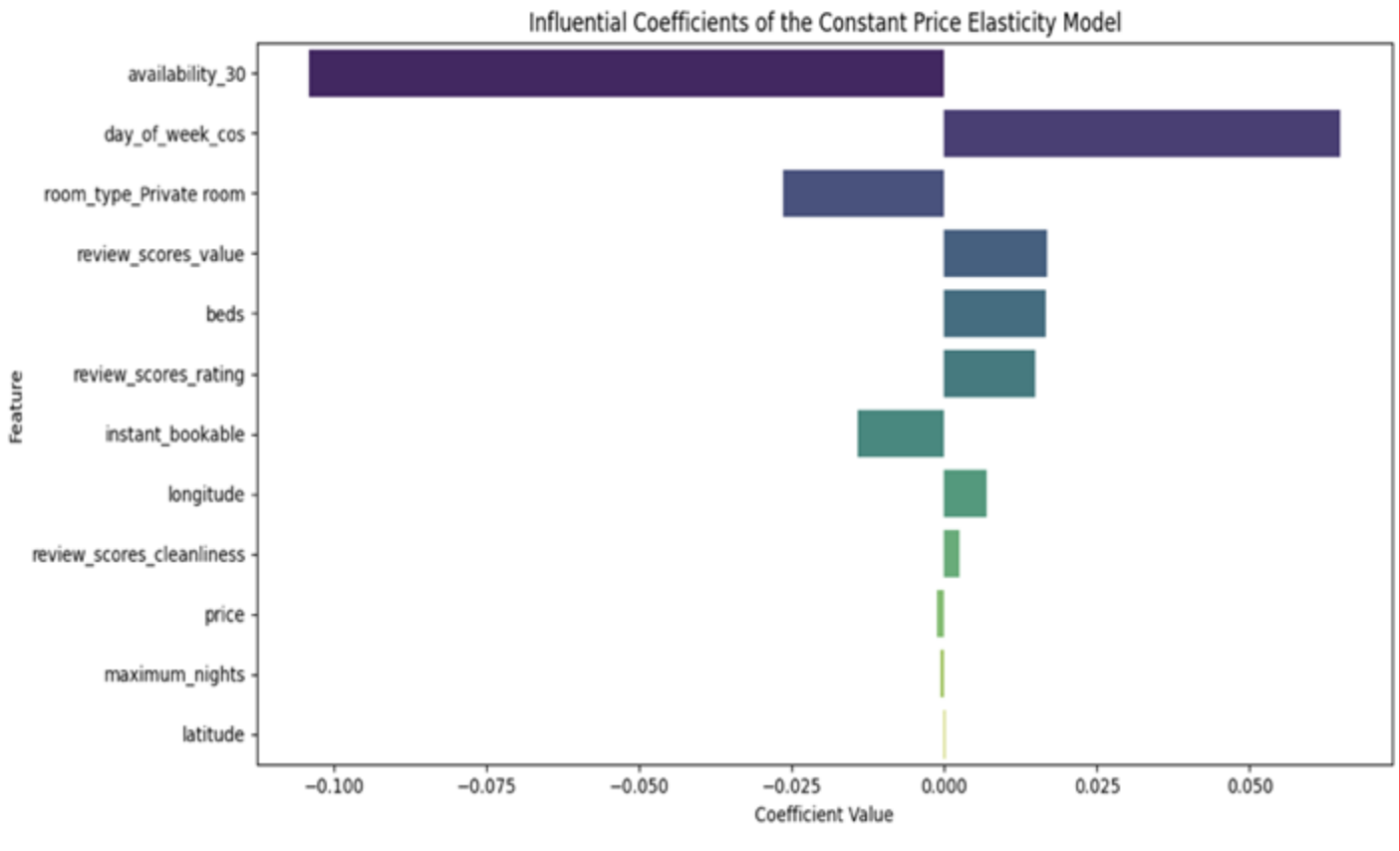
Conclusion

Given that the Logistic Model is predicting Overall Booking Probability and the others are predicting Stay Length, a direction comparison of RMSE is not valid. However, across the other metrics, our Logistic Model outperforms the other models drastically, demonstrating high accuracy in terms of percentage error metrics.

Among the models predicting Staty Length, all three have RMSE values that are high relative to the STD of Stay Length (which is ~2.95 days). This indicates that predicting Stay Length with high precision remains a challenge, but the models still capture some of the relationship. Would likely have to trim off more data (likely only to listings in which stay length of 2-7 days) was realized.

However, as is, if set on predicting Stay Length, the choice between Linear and CPE probably depends on whether the interpretability of price elasticity is important. Because it is, CPE model likely preferred.

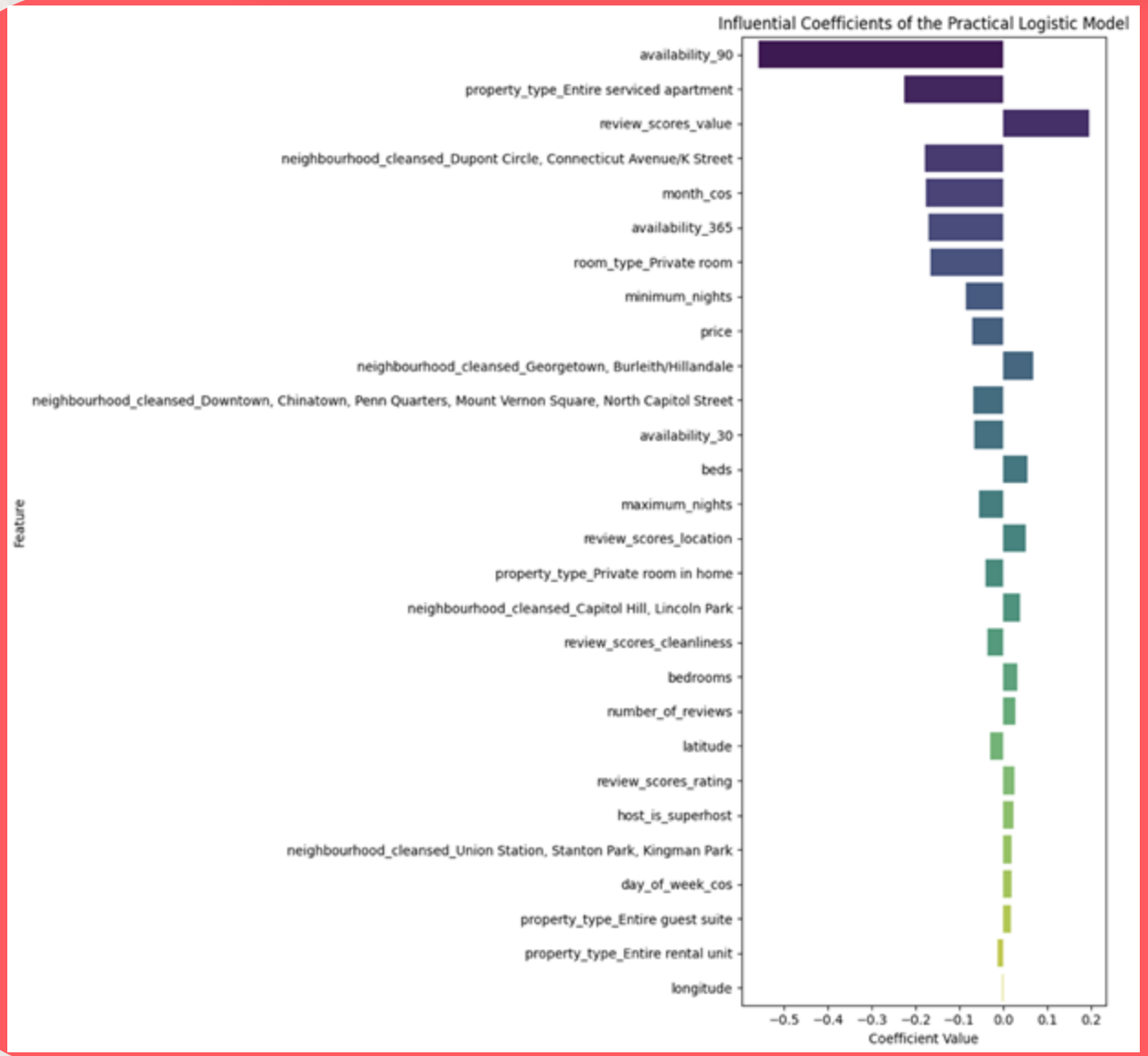
Influential Coefficients (Constant Price Elasticity Model)



Interpretation of Some Key Coefficients for CPE Model

- ***Price***: In the CPE model, the coefficient for price is the price elasticity of demand. The coefficient of ~-0.00119 tells us that a 1% increase in price is only associated with a ~0.00119% decrease in Stay Length.
- ***Availability_30***: This feature tells how many days a listing is available over the next 30 days. Because no other predictors, other than price, are log transformed in this model, and with its coefficient of ~-0.104, a one-unit increase for this feature causes, on average and holding all else constant, a 10.4% decrease in stay length. The same rule applies for all continuous predictors in the model.
- ***Room Type - Private Room***: This one hot encoded binary feature has a coefficient of ~-0.025. This means that, if the listing is for only a private room to be rented, the stay of length decreases by approximately (exp(0.025)-1)*100% =2.53%,, on average and all else constant. The same rule applies for all binary predictors in the data.

Influential Coefficients (Logistic Model)

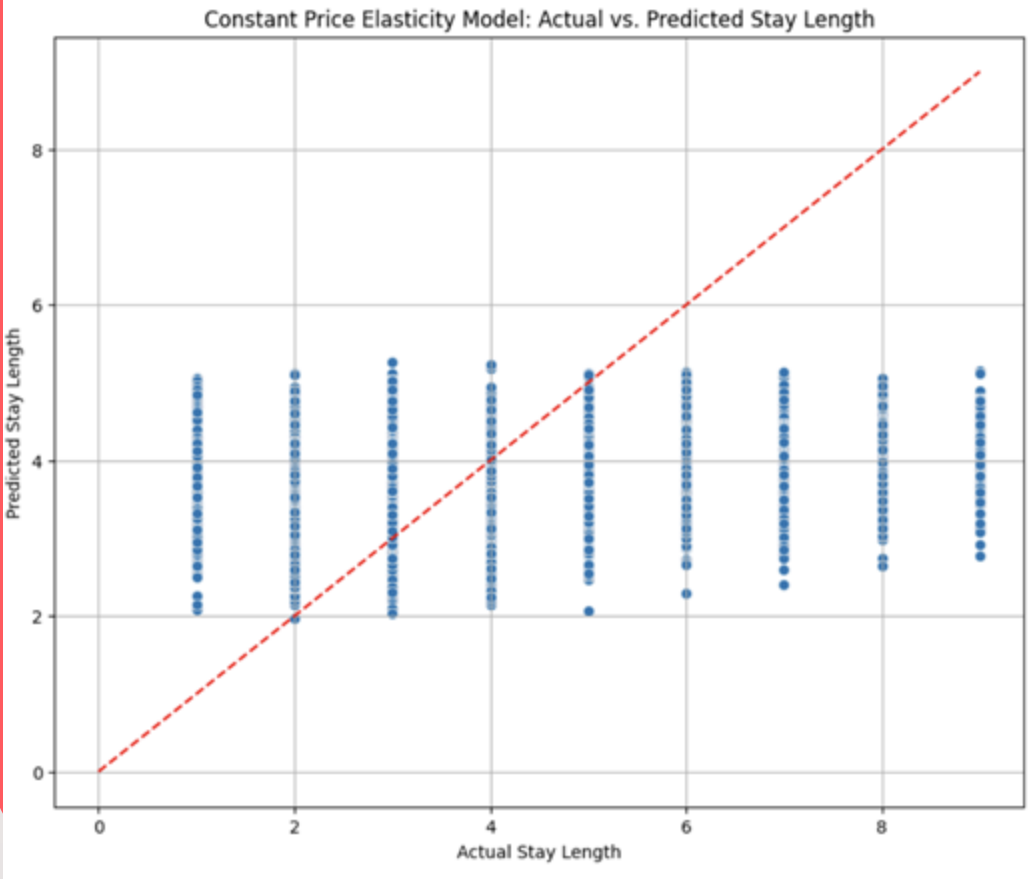


Interpretation of Some Key Coefficients for Logistic Model

- ***Price***: Price has a coefficient of ~-.07. Here, a one-unit increase in the scaled price decreases the odds of a booking occurring by exp(-0.703), or by about 6.8%, holding all else constant.
- ***Review Scores Value***: The scaled coefficient for Review Scores Value is ~0.197. Therefore, a one-unit increase in the scaled value of this variable increases the overall booking probability of a listing by about 21.8%, on average and holding all else constant.
- ***Property_type Entire Serviced Apartment***: For this binary feature, with a coefficient of ~-0.224, if the listing pertains to an "Entirely Serviced" Apartment, the likelihood of a booking occurring is 20.1% lower than the baseline property type, on average and holding all other factors constant.

Scatter Plot of Actual vs Predicted Values (Constant Price Elasticity Model)

```
Constant Price Elasticity Model (Stay Length):
  R-squared: 0.0330
  Mean Squared Error (MSE): 3.6676
```



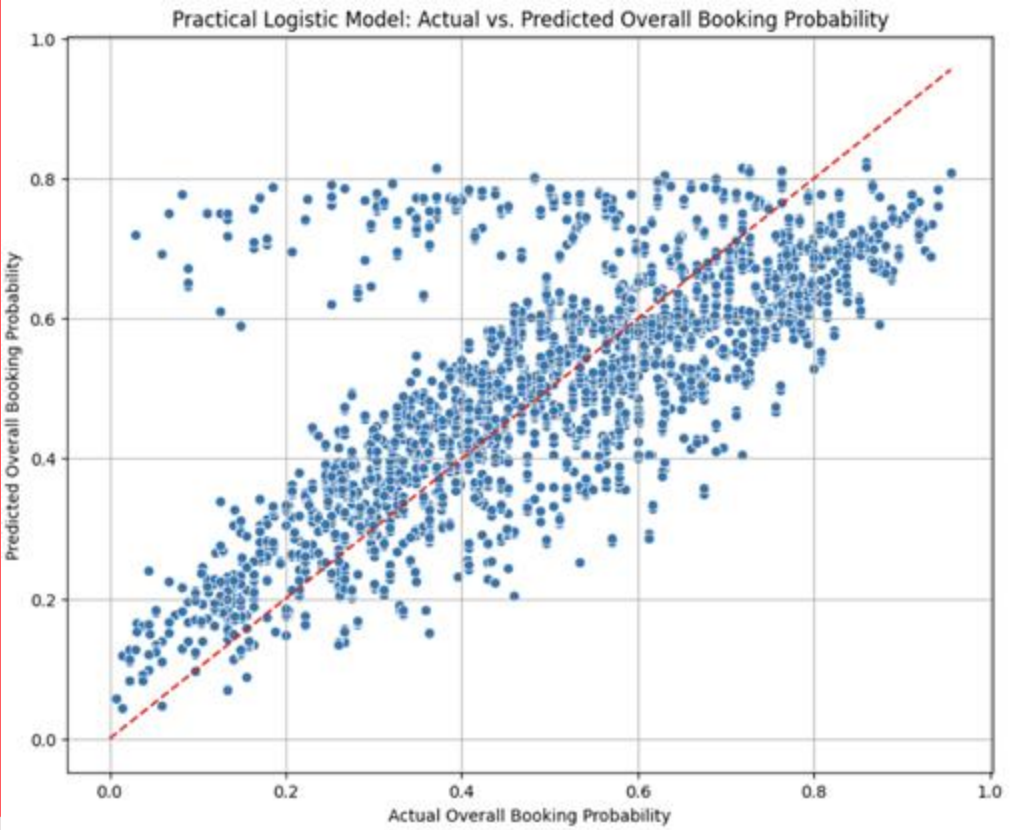
Why RMSE Appeared Strong but R-Squared is Poor

For the models predicting Stay Length, the RMSE values are around 1.89 to 1.91. In a vacuum, these seem strong and like they imply strong predictive models; however, the "goodness" of RMSE is dependent on the scale and variability in the target variable. Because Stay Length (in the validation data) has a STD of ~2.95 days, the RMSE values are very close to the STD. This means the models predicting stay length as is are performing no better than a very naive model that would predict the mean of Stay Length for every observation. If we just guessed the stay length for every booking, the average error would be roughly the standard deviation. So an RMSE of 1.89 (as presented in the CPE model) means the model is explaining some variance, but not a huge amount relative to just predicting the average.

The CPE Model explains only 3.30% of the variability in stay length (as shown by the R-Squared value). This is very poor and tells us that the features we've included (even after one-hot encoding and transformations) are not strongly capturing the drivers of stay length for these particular models.

Scatter Plot of Actual vs Predicted Values (Logistic Model)

```
Practical Logistic Model (Overall Booking Probability):
  AUC-ROC: 0.8415
```



The strong AUC score of the Logistic Model makes it a very valuable tool for predicting the probability of a booking occurring. Specifically, the AUC quantifies the logistic models overall ability to distinguish between a true positive class (a booking occurring) from the negative class (a booking not occurring) across all possible classification thresholds. Leveraging booking probabilities to derive actionable business insights was determined to be our strategy going forward.

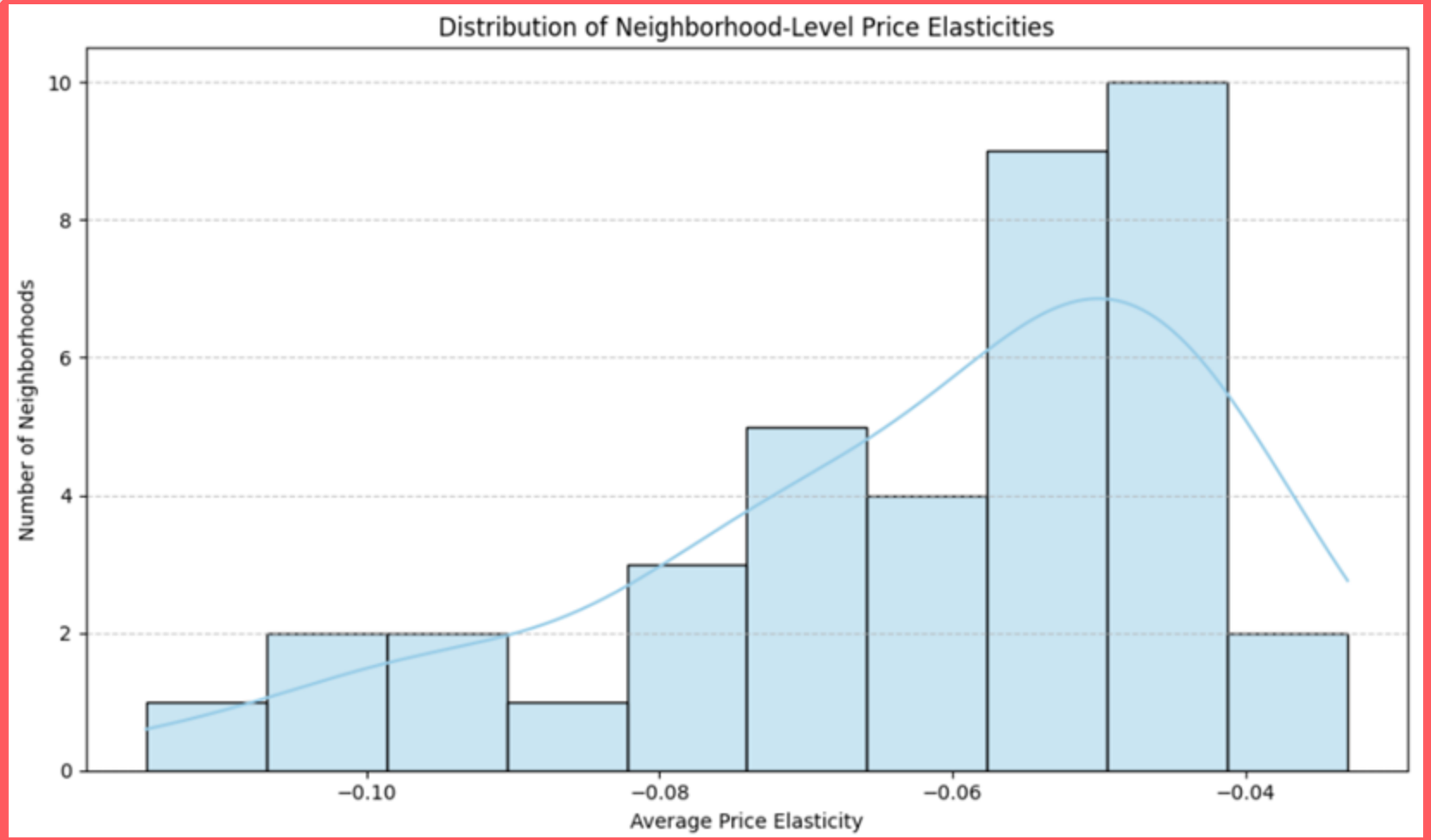
Price Elasticities per Neighborhood

```
# Elasticity = (beta_scaled_price / price_std_dev) * original_price * (1 - predicted_probability)
elasticities = (scaled_price_coef / price_std_dev) * original_price_val * (1 - logit_val_preds)
```

Neighborhood-Level Price Elasticities (Average):

neighbourhood_cleansed	price_elasticity
Southwest Employment Area, Southwest/Waterfront, Fort McNair, Buzzard Point	-0.115017
Cathedral Heights, McLean Gardens, Glover Park	-0.100308
Downtown, Chinatown, Penn Quarters, Mount Vernon Square, North Capitol Street	-0.100138
Douglas, Shipley Terrace	-0.094022
Dupont Circle, Connecticut Avenue/K Street	-0.092085
Georgetown, Burleith/Hillandale	-0.084919
Near Southeast, Navy Yard	-0.078200
Ivy City, Arboretum, Trinidad, Carver Langston	-0.076289
West End, Foggy Bottom, GWU	-0.075300
Sheridan, Barry Farm, Buena Vista	-0.072850
Colonial Village, Shepherd Park, North Portal Estates	-0.072315
Kalorama Heights, Adams Morgan, Lanier Heights	-0.070860
Brookland, Brentwood, Langdon	-0.070574
Historic Anacostia	-0.067347
Mayfair, Hillbrook, Mahaning Heights	-0.065471
Woodridge, Fort Lincoln, Gateway	-0.060019
Woodland/Fort Stanton, Garfield Heights, Knox Hill	-0.059910
Congress Heights, Bellevue, Washington Highlands	-0.0590
Howard University, Le Droit Park, Cardozo/Shaw	-0.0571
Spring Valley, Palisades, Wesley Heights, Foxhall Crescent, Foxhall Village, Georgetown Reservoir	-0.0557
Shaw, Logan Circle	-0.0555
Eastland Gardens, Kenilworth	-0.0521
Capitol Hill, Lincoln Park	-0.0519
Columbia Heights, Mt. Pleasant, Pleasant Plains, Park View	-0.0515
Twining, Fairlawn, Randle Highlands, Penn Branch, Fort Davis Park, Fort Dupont	-0.0514
Union Station, Stanton Park, Kingman Park	-0.0511
North Cleveland Park, Forest Hills, Van Ness	-0.0504
Cleveland Park, Woodley Park, Massachusetts Avenue Heights, Woodland-Normanstone Terrace	-0.0471
Deanwood, Burrville, Grant Park, Lincoln Heights, Fairmont Heights	-0.0458
Hawthorne, Barnaby Woods, Chevy Chase	-0.0449
Takoma, Brightwood, Manor Park	-0.0446
Fairfax Village, Naylor Gardens, Hillcrest, Summit Park	-0.0445
North Michigan Park, Michigan Park, University Heights	-0.0441
Capitol View, Marshall Heights, Benning Heights	-0.0433
Friendship Heights, American University Park, Tenleytown	-0.04323
Edgewood, Bloomingdale, Truxton Circle, Eckington	-0.0424
Lamont Riggs, Queens Chapel, Fort Totten, Pleasant Hill	-0.0412
Brightwood Park, Crestwood, Petworth	-0.0387
River Terrace, Benning, Greenway, Dupont Park	-0.0330

Price Elasticities per Neighborhood



Looking at the average neighborhood-price elasticies, none of their absolute values are above 1. Therefore, based on conventional definition, all the neighborhoods in Washington DC for Airbnb Listings in the snapshotted data exhibit inelastic demand with respect to price changes. This implies that while price does at least somewhat influence booking probability, the booking probability for Airbnb Listings across Washington DC for the time period on hand are not generally sensitive to pricing fluctuations.

Model Application: Applying an LLM to Optimize Revenue in Test Data

Due to time constraints, we decided to apply an LLM to calculate initial revenue from the listings in our test data dataset (using stay lengths, original prices, neighborhoods, and predicted probabilities) and simulate optimized revenue per neighborhood aft applying a general optimization strategy.

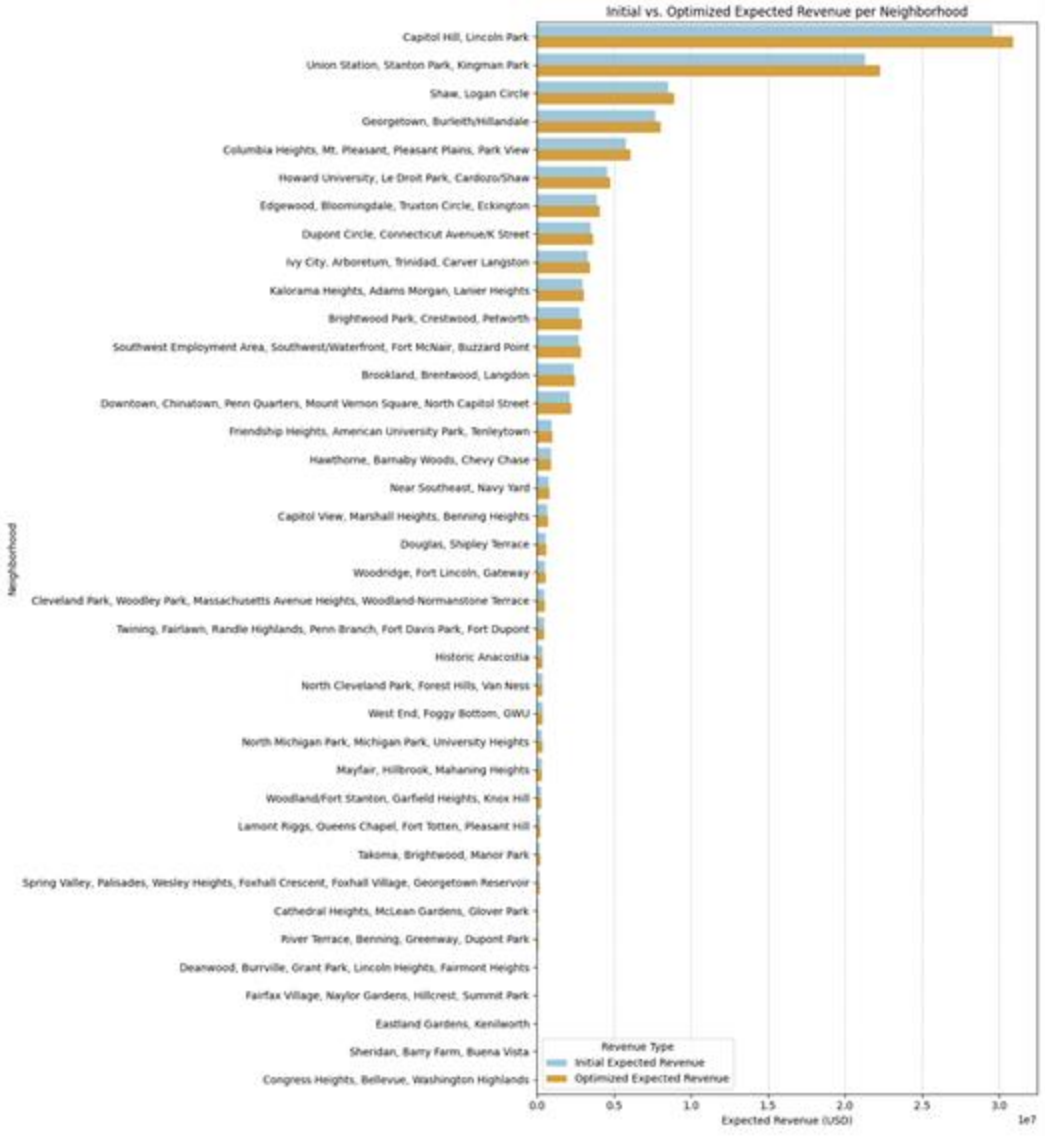
Revenue Comparison (Initial vs. Optimized with 5% Price Increase) per Neighborhood

neighbourhood_cleansed	Initial Expected Revenue	Optimized Expected Revenue	Revenue Change	Percentage Change
Eastland Gardens, Kenilworth	3.773176e+04	3.956753e+04	1.835771e+03	4.865319
Capitol View, Marshall Heights, Benning Heights	6.931428e+05	7.268136e+05	3.367089e+04	4.857714
Spring Valley, Palisades, Wesley Heights, Foxhall Crescent, Foxhall Village, Georgetown Reservoir	1.987621e+05	2.083625e+05	9.600367e+03	4.830079
Cleveland Park, Woodley Park, Massachusetts Avenue Heights, Woodland-Normanstone Terrace	5.008264e+05	5.250054e+05	2.417901e+04	4.827822
Sheridan, Barry Farm, Buena Vista	3.573498e+04	3.745980e+04	1.724820e+03	4.826698
Fairfax Village, Naylor Gardens, Hillcrest, Summit Park	3.895014e+04	4.082999e+04	1.879851e+03	4.826301
River Terrace, Benning, Greenway, Dupont Park	7.329865e+04	7.682579e+04	3.527136e+03	4.812007

Strategy for inelastic demand: Because each neghborhood exhibits inelastic demand, guests are relatively insensitive to price changes. Therefore, our revenue optimization strategy involves increasing prices.

- **Impact on total revenue:** Since the percentage decrease in booking probability is less than the percentage increase in price, the total expected revenue for the listings in each neighborhood is shown to increase. This is because the higher price per booking outweighs the smaller reduction in number of bookings.
- **Overall Revenue Increase:** A blanket 5% price increase across all neighborhoods, as demonstrated (only partially shown to left) resulted in a significant total revenue increase pof ~5.07M USD, corresponding to a 4.62% overall percentage increase in revenue.

Model Application: Applying an LLM to Optimize Revenue in Test Data



Neighborhood-Level Variation: While the overall trend is positive, the percentage change in revenue varies slightly across neighborhood. This variation reflects the differing price sensitivities (elasticites) of each neighborhood, even though all were identified as inelastic.

- **Highest Gains:** Neighborhood such as "Eastland Gardens, Kenilworth" saw a percentage increase of around 4.87%, indicating a very inelastic response to price changes.
- **Lower Gains:** Neighborhoods like "Georgetown, Burleith/Hillandale" had a slightly lower percentage increase of around 4.45% suggesting they are relatively less inelastic compared to others, but still benefit from a price increase.

Recommendations and Business Impact

Recommendations and Implications

- **Confirm Inelasticity**: The observed revenue increase from a price hike suggests that the initial assessment of inelastic demand for these neighborhoods is robust. For future optimization, a more granular analysis of elasticity within each neighborhood (e.g., segmenting by property type or other features) could refine price adjustments further.
- **Gradual Price Adjustments**: While a 5% increase was beneficial, a more sophisticated strategy would involve testing different price increases or decreases based on each neighborhood's precise elasticity, rather than a uniform percentage. Iterative A/B testing in live environments could fine-tune these adjustments.
- **Monitor Non-Price Factors**: Continuous monitoring and improvement of non-price factors (e.g., review scores, amenities, host responsiveness) remain crucial. Enhancing these aspects can potentially shift the entire demand curve upwards, allowing for higher prices without sacrificing booking probability, or even increasing booking probability at existing prices.
- **Competitive Landscape**: Always consider the competitive landscape. Even with inelastic demand, excessively high prices might drive customers to alternative markets or platforms if substitutes exist, a factor not explicitly modeled here.

Business Impact

- **Higher Revenue:** Price adjustments and policy tuning yield measurable gains without major reductions in demand
- **Better Occupancy:** Minimum-night optimization and availability monitoring help stabilize booking patterns
- **Improved Guest Experience:** Stronger reputation signals enhance trust and increase stay length
- **Smarter Host & Platform Decisions:** Data-driven nudges on pricing, policies, and quality drive consistent marketplace improvement

Thank You