MGMT 683 Final Project
Airbnb Property Churn Analysis

Group 18

Nagarjuna Chidarala
Sai Teja Devalla
Seonkyu Kim
Chaitanya Sanaboina

Content

Data Description

- Airbnb data in Washington
- Panel data consisting of property data over periods

Period	Property_ID
5	A
5	B
5	C
5	D
6	A
6	B
6	E
7	B
7	E

Variable Groups

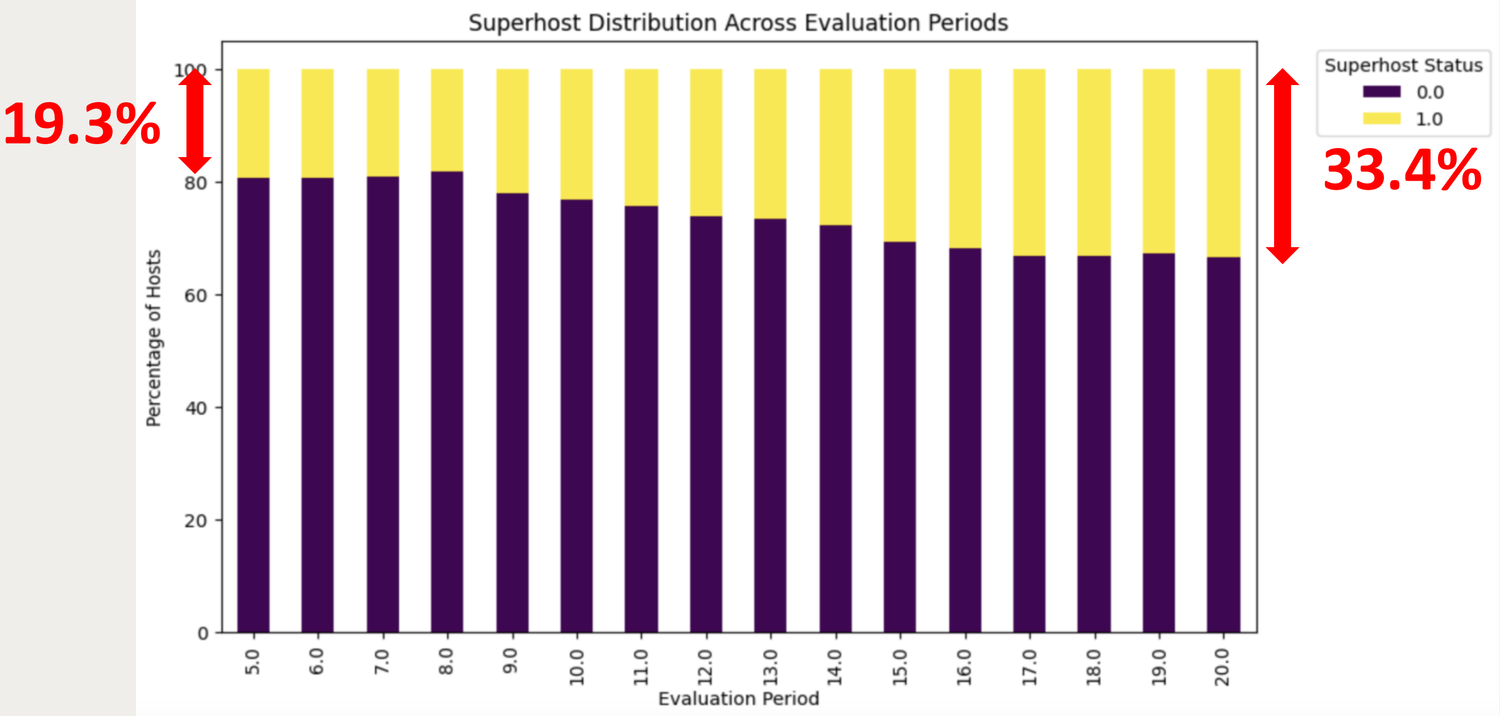
- Superhost Status
- Reviews
- Bookings
- Revenues
- Tract & Zip level demographics
- Location

And many more..

Data Details

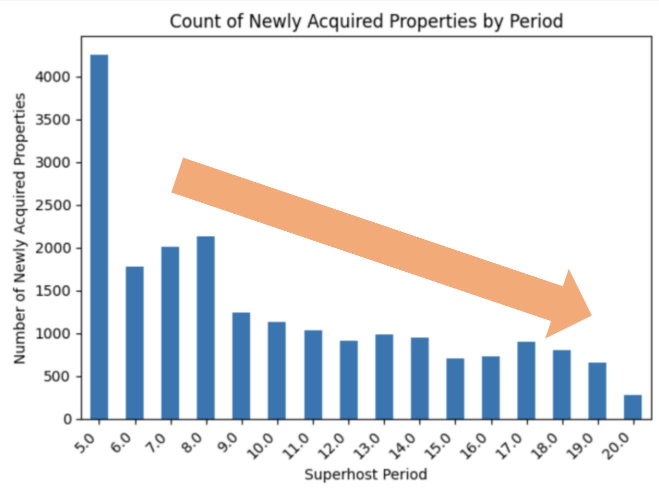
- 16 periods
- 133,741 records
- 20,526 unique properties
- 12,073 unique hosts

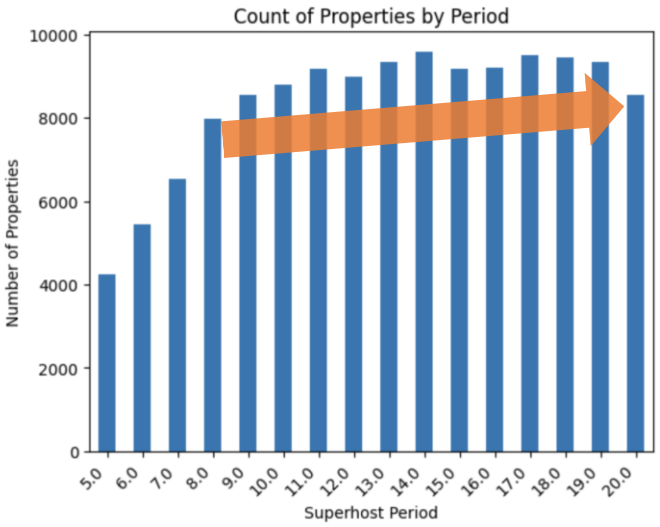
The proportion of superhosts increased.

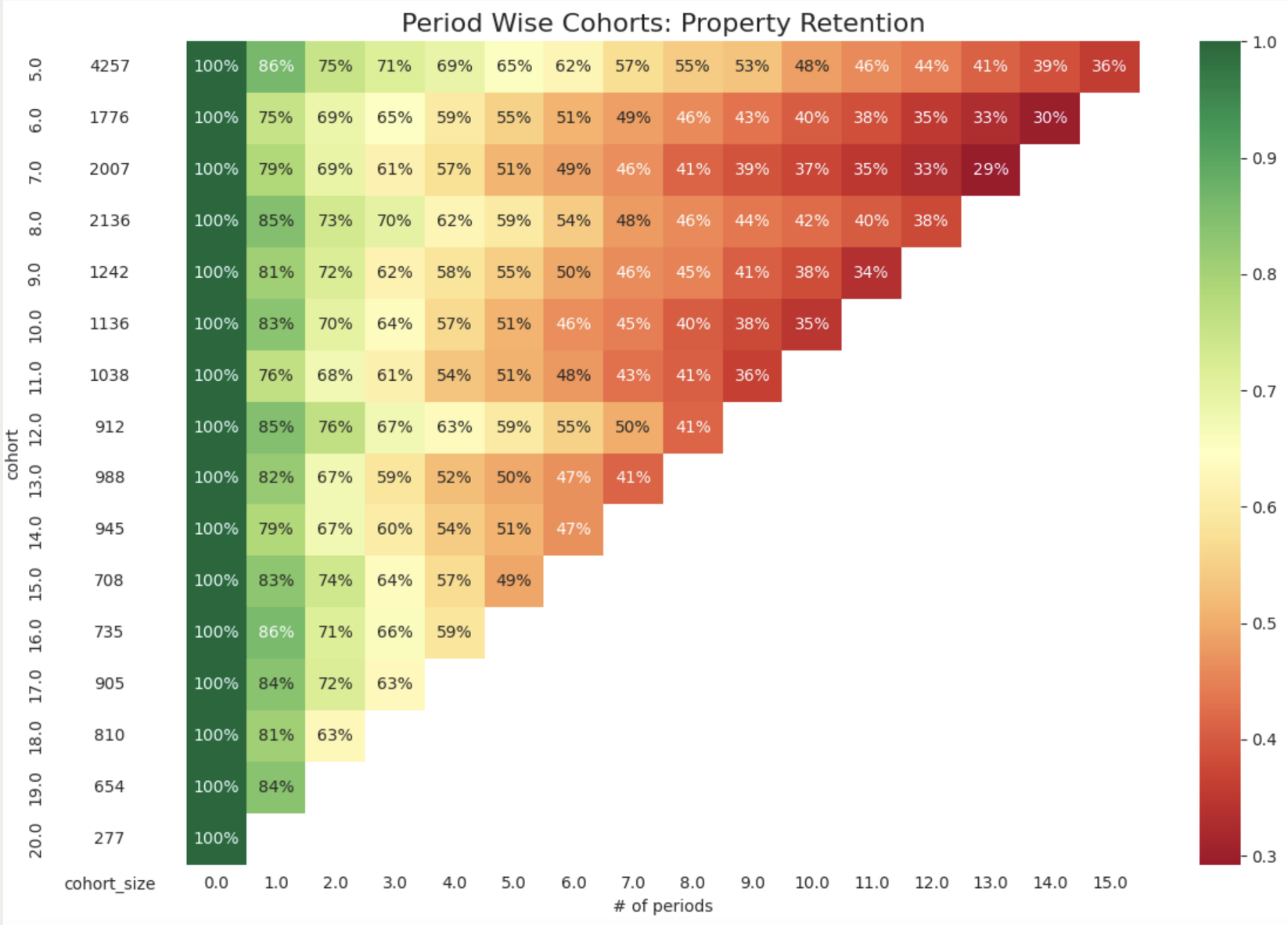


- Host at least 10 trips
- Maintain 90% response rate for guest requests
- Complete all confirmed reservations without cancellation
- Receive 5-star review at least 80% of the time

Only about 30% of properties survived up to 3 years.







Data Analytics Objectives

- Figure out the **reasons for property churn**.
- Identify the **properties prone to churn**.
- Help Airbnb define **targeted marketing and promotional activities** that will help retain the properties on the platform.

Train(80%), Validation(20%): Period 5~19 / Test: Period 20 Prediction

Defining Churn

Period	Property_ID
5	A
5	B
5	C
5	D
6	A
6	B
6	E
7	B
7	E

Period	Property_ID	Churn	Description
5	A	0	
5	B	0	
5	C	1	Didn't survive in the next period
5	D	1	Didn't survive in the next period
6	A	1	Didn't survive in the next period
6	B	0	
6	E	0	
7	B		Last period, to be predicted
7	E		Last period, to be predicted

Feature Engineering & Variable Selection

- New column "months_with_bnb"

  = Difference between "created_date" & "Scraped Date" in months

- Variable Selection (Drop variables)
  1) Repeated variables
  2) Columns that can be feature-engineered by existing columns
  3) Variables that will not add much value to the churn (our intuition)

→ 79 columns concerned + 1 new column added = 80 variables

- Explore 80 variables with boxplot → **30 variables** selected for model

Data Preprocessing

1) Replaced missing values of **occupancy rate** & **revenue** with **0** (No days with "booked_days = 0")

2) Replaced missing values of other variables with **medians**

```
selected_rows_revenue_nan = prd_not20_model_vars[prd_
print(selected_rows_revenue_nan)
```

```
        revenue  booked_days  booked_days_avePrice
25983       NaN          NaN                   NaN
40557       NaN          NaN                   NaN
27093       NaN          NaN                   NaN
27054       NaN          NaN                   NaN
54216       NaN          NaN                   NaN
...         ...          ...                   ...
118986      NaN          NaN                   NaN
120232      NaN          NaN                   NaN
81443       NaN          NaN                   NaN
106517      NaN          NaN                   NaN
108907      NaN          NaN                   NaN

[57209 rows x 3 columns]
```

```
prd_not20_model_vars['booked_days'].value_counts()
```

```
booked_days
1.0      2346
2.0      2176
3.0      2175
4.0      2121
7.0      1841
         ...
151.0       1
128.0       1
140.0       1
137.0       1
```

Logistic Regression (Backward-Elimination for p-value > 0.1)

- Selected 30 Input variables
- Target variable: Churn
- Period 5 ~ 19
- Threshold = 0.5

Confusion Matrix		Predicted Value	
		0	1
True Value	0	22,514	17
	1	2,496	13

Accuracy = 0.8996

Sensitivity = 0.0051

Specificity = 0.9992

- Selected 30 Input variables
- Target variable: Churn
- Period 5 ~ 19
- Threshold = 0.1

Confusion Matrix		Predicted Value	
		0	1
True Value	0	15,266	7,265
	1	1,036	1,473

Accuracy = 0.6685

Sensitivity = 0.5871

Specificity = 0.6776

Coefficient Interpretation

- 1% increase in **rating_ave_pastYear** decreases property churn by $(e^{0.620 \times 0.01} - 1) \times 100\% = 0.620\%$

- 1% increase in **hostResponseAverage_pastYear** decreases property churn by $(e^{0.011 \times 0.01} - 1) \times 100\% = 0.011\%$

- 1% increase in **months_with_bnb** decreases property churn by $(e^{0.015 \times 0.01} - 1) \times 100\% = 0.015\%$

- 1% increase in **Max Guests** increases property churn by $(e^{0.012 \times 0.01} - 1) \times 100\% = 0.012\%$

More Sophisticated Models

1) Decision Tree

- No threshold tuning
- Classification at 0.5

Confusion Matrix		Predicted Value 0	Predicted Value 1
True Value	0	20,641	1,890
	1	1,796	713

Accuracy = 0.8528

Sensitivity = 0.2842

Specificity = 0.9161

2) Random Forest

- Threshold = 0.5

Confusion Matrix		Predicted Value 0	Predicted Value 1
True Value	0	22,503	28
	1	2,231	278

Accuracy = 0.9098

Sensitivity = 0.1108 / Specificity = 0.9988

- Threshold = 0.12

Confusion Matrix		Predicted Value 0	Predicted Value 1
True Value	0	16,970	5,561
	1	663	1,846

Accuracy = 0.7514

Sensitivity = 0.7358 / Specificity = 0.7532

3) Gradient Boosting

- Threshold = 0.5

Confusion Matrix		Predicted Value 0	Predicted Value 1
True Value	0	22,514	17
	1	2,408	101

Accuracy = 0.9032

Sensitivity = 0.0403 / Specificity = 0.9992

- Threshold = 0.095

Confusion Matrix		Predicted Value 0	Predicted Value 1
True Value	0	15,623	6,908
	1	760	1,749

Accuracy = 0.6938

Sensitivity = 0.6971 / Specificity = 0.6934

Model Comparison

Model/Metric	Threshold	Accuracy	Sensitivity	Specificity
Logistic Regression	0.1	0.67	0.58	0.67
Decision Tree	0.5	0.85	0.28	0.91
Random Forest	0.12	0.75	0.73	0.75
Gradient Boosting	0.095	0.69	0.69	0.69

Random Forest Model works better than other models.

Random Forest (Validation Set of Period 5~19)

Confusion Matrix		Predicted Value	
		0	1
True Value	0	16,970	5,561
	1	663	1,846

7,407
- 3,807: revenue > 0
  - 996 Actual Churn
  - 2,811 Did not Churn
- 3,600: revenue = 0

Random Forest (Predict Period 20)

8,545
- 3,703: revenue > 0
  - 1,780: Churn Probability > 0.12 → **Target for Marketing & Promotion**
  - 1,923: Churn Probability <= 0.12
- 4,842: revenue = 0

Revenue Analysis

- Logistic Regression (Period 5 ~ 19)
- Input variables: 30 variables
- Target variable: revenue_label
  (1: revenue$=$0, 0: revenue$\neq$0)

Confusion Matrix		Predicted Value	
		0	1
True Value	0	53,240	14,747
	1	10,227	46,982

Accuracy = 0.8005

Sensitivity = 0.8212

Specificity = 0.7831

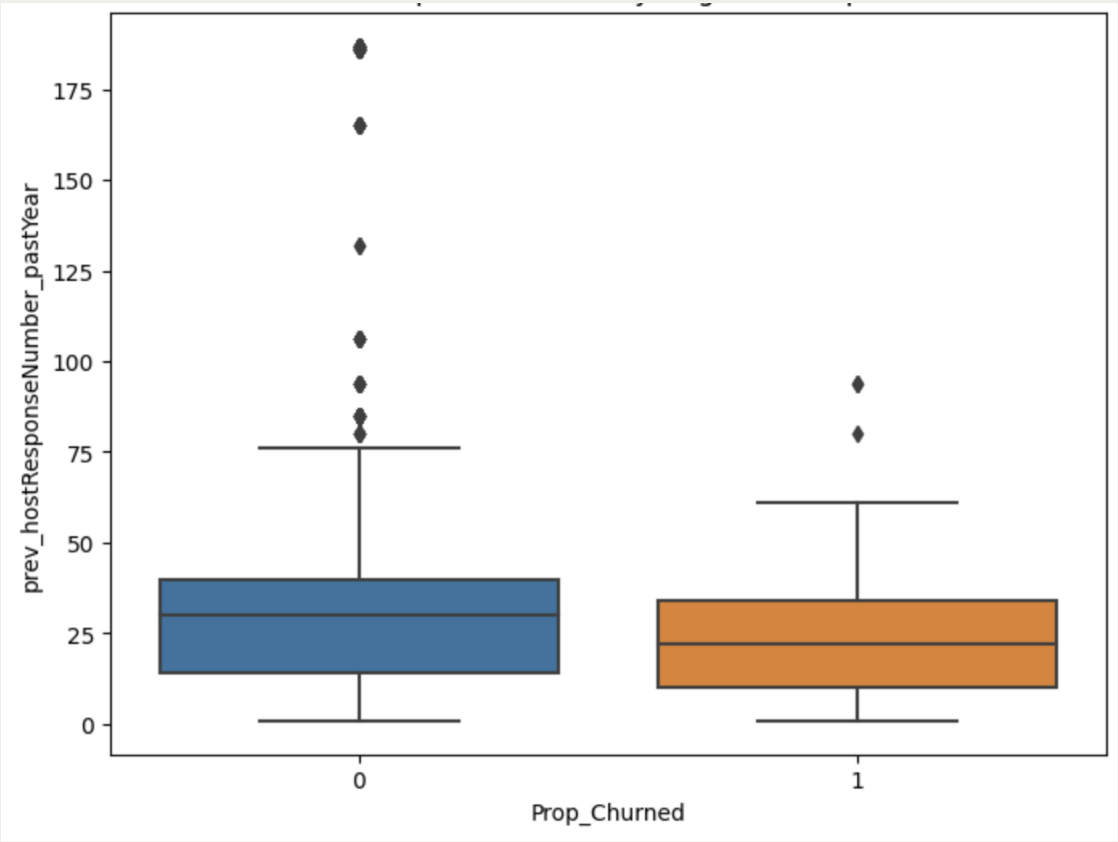
Coefficient Interpretation (revenue_label)

- The "revenue = 0" probability for a **churned property** is $(e^{0.055} - 1) \times 100\% = 5.65\%$ higher than that for a non-churned property.

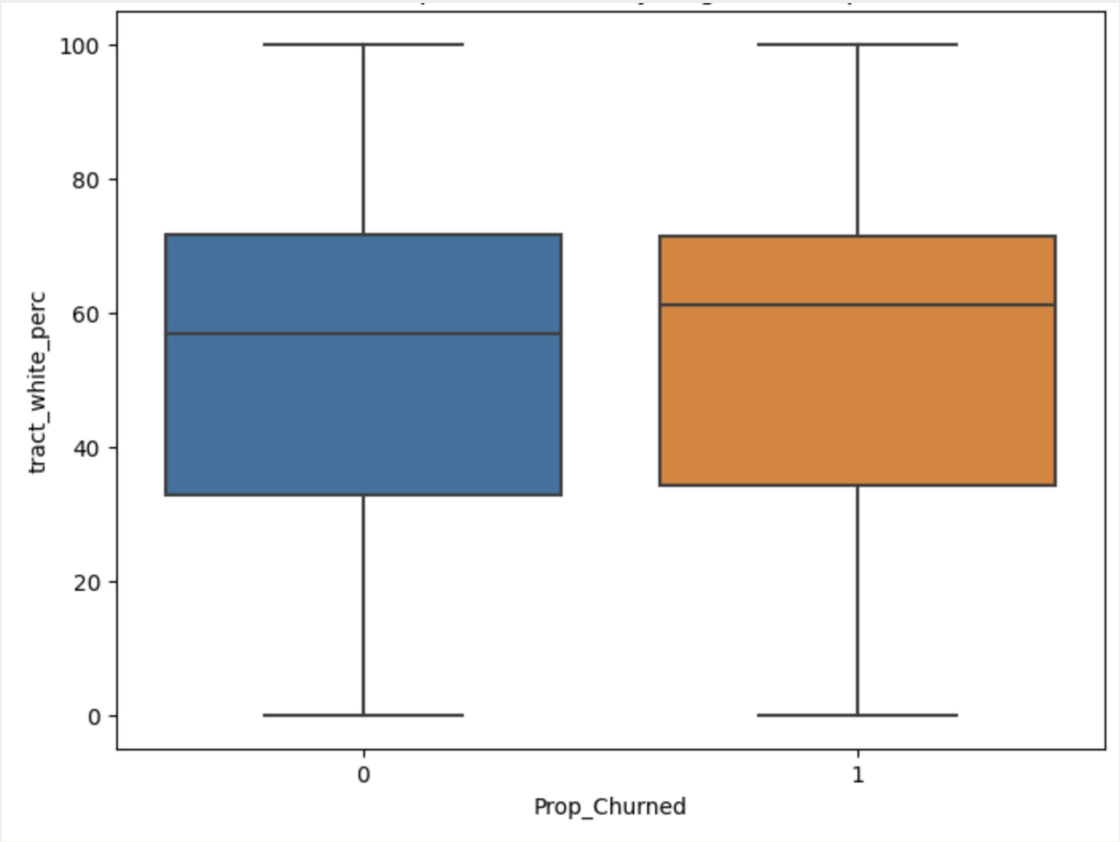
- 1% increase in **months_with_bnb** increases "revenue = 0" probability by $(e^{0.028\times0.01} - 1) \times 100\% = 0.028\%$

- 1% increase in **prev_Number of Reviews** decreases "revenue = 0" probability by $(e^{0.026\times0.01} - 1) \times 100\% = 0.026\%$

Exhibit 1. Variable Selection Boxplot



Selected Variable



Non-Selected Variable

Exhibit 2. Logistic Regression Result (Backward Elimination for p-value>1.0, threshold=0.5)

Accuracy: 0.8996405750798722
Confusion Matrix:
[[22514 17]
[2496 13]]
Sensitivity: 0.0051813471502590676
Specificity: 0.9992454839998225

Logit Regression Results

Dep. Variable:	Prop_Churned	No. Observations:	100156
Model:	Logit	Df Residuals:	100131
Method:	MLE	Df Model:	24
Date:	Fri, 08 Dec 2023	Pseudo R-squ.:	0.04981
Time:	18:44:49	Log-Likelihood:	-29787.
converged:	True	LL-Null:	-31348.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	2.8859	0.176	16.367	0.000	2.540	3.231
superhost_period_all	0.0267	0.003	8.092	0.000	0.020	0.033
rating_ave_pastYear	-0.6159	0.033	-18.477	0.000	-0.681	-0.551
numReviews_pastYear	-0.0003	6.76e-05	-4.261	0.000	-0.000	-0.000
numReservedDays_pastYear	-0.0002	2.01e-05	-9.054	0.000	-0.000	-0.000
numReserv_pastYear	0.0003	3.29e-05	8.126	0.000	0.000	0.000
prev_numReservedDays_pastYear	0.0001	2.07e-05	6.101	0.000	8.59e-05	0.000
hostResponseNumber_pastYear	-0.0031	0.001	-4.231	0.000	-0.004	-0.002
hostResponseAverage_pastYear	-0.0113	0.001	-15.868	0.000	-0.013	-0.010
prev_hostResponseNumber_pastYear	0.0046	0.001	6.077	0.000	0.003	0.006
available_days	-0.0040	0.000	-23.061	0.000	-0.004	-0.004
booked_days	-0.0093	0.001	-8.037	0.000	-0.012	-0.007
booked_days_avePrice	0.0015	0.000	13.896	0.000	0.001	0.002
Number of Photos	-0.0148	0.001	-11.829	0.000	-0.017	-0.012
Nightly Rate	-0.0003	6.6e-05	-4.494	0.000	-0.000	-0.000
Rating Overall	-0.0014	0.001	-1.894	0.058	-0.003	4.89e-05
revenue	-4.919e-05	8.09e-06	-6.084	0.000	-6.5e-05	-3.33e-05
prev_occupancy_rate	0.1514	0.066	2.302	0.021	0.023	0.280
tract_asian_perc	0.0071	0.004	1.907	0.056	-0.000	0.014
zip_white_nothispanic_percent	-0.0016	0.001	-2.272	0.023	-0.003	-0.000
Nightly Rate_tractQuartile	0.0485	0.012	4.020	0.000	0.025	0.072
tract_superhosts	0.0127	0.003	4.365	0.000	0.007	0.018
tract_superhosts_ratio	-0.7430	0.137	-5.439	0.000	-1.011	-0.475
tract_prev_superhosts	-0.0125	0.003	-4.255	0.000	-0.018	-0.007
months_with_bnb	-0.0162	0.001	-22.850	0.000	-0.018	-0.015

Exhibit 3. Optimal Threshold and other metrics of Logistic Regression

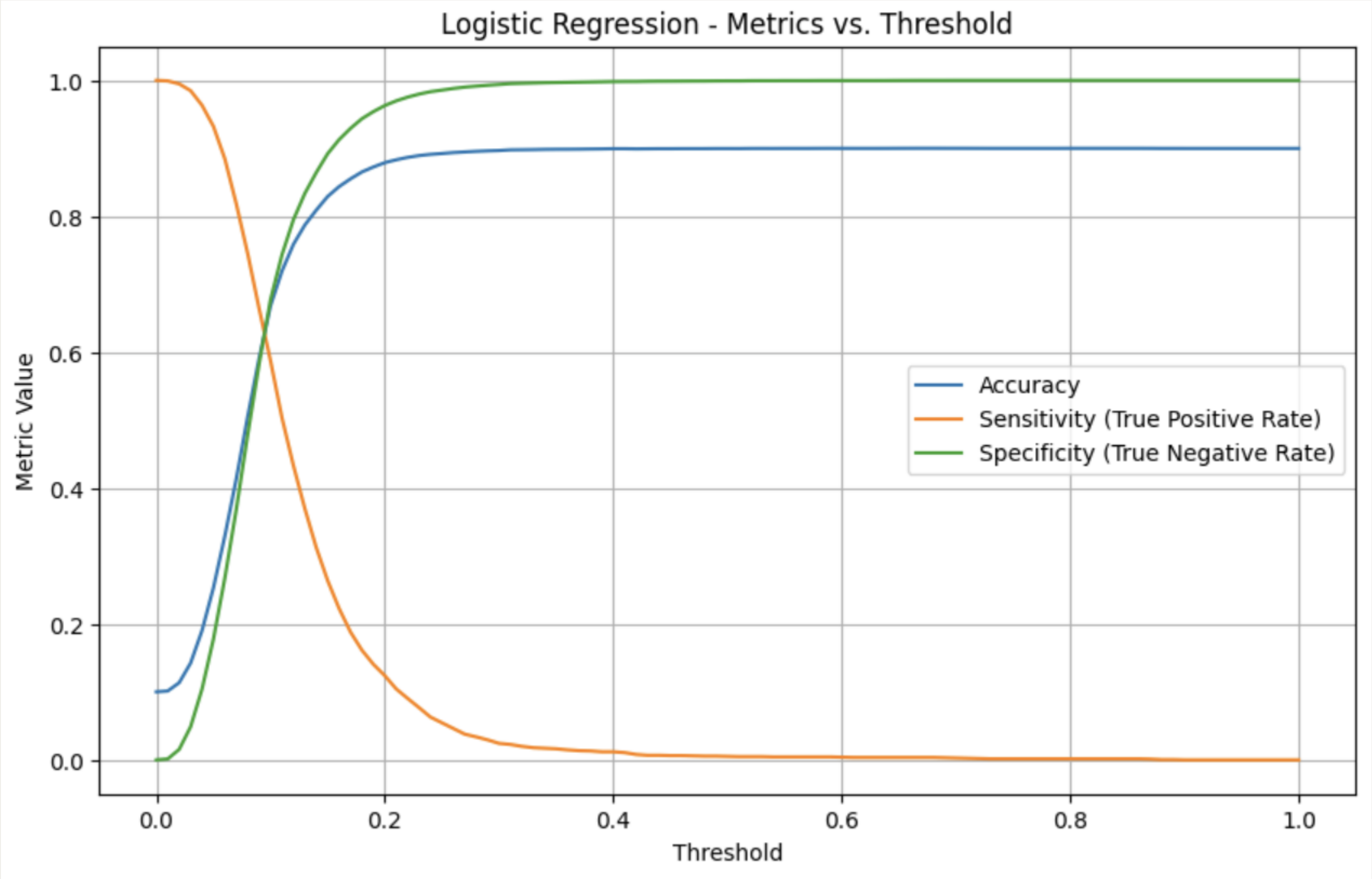


Exhibit 4. Optimal Threshold and other metrics of Random Forest

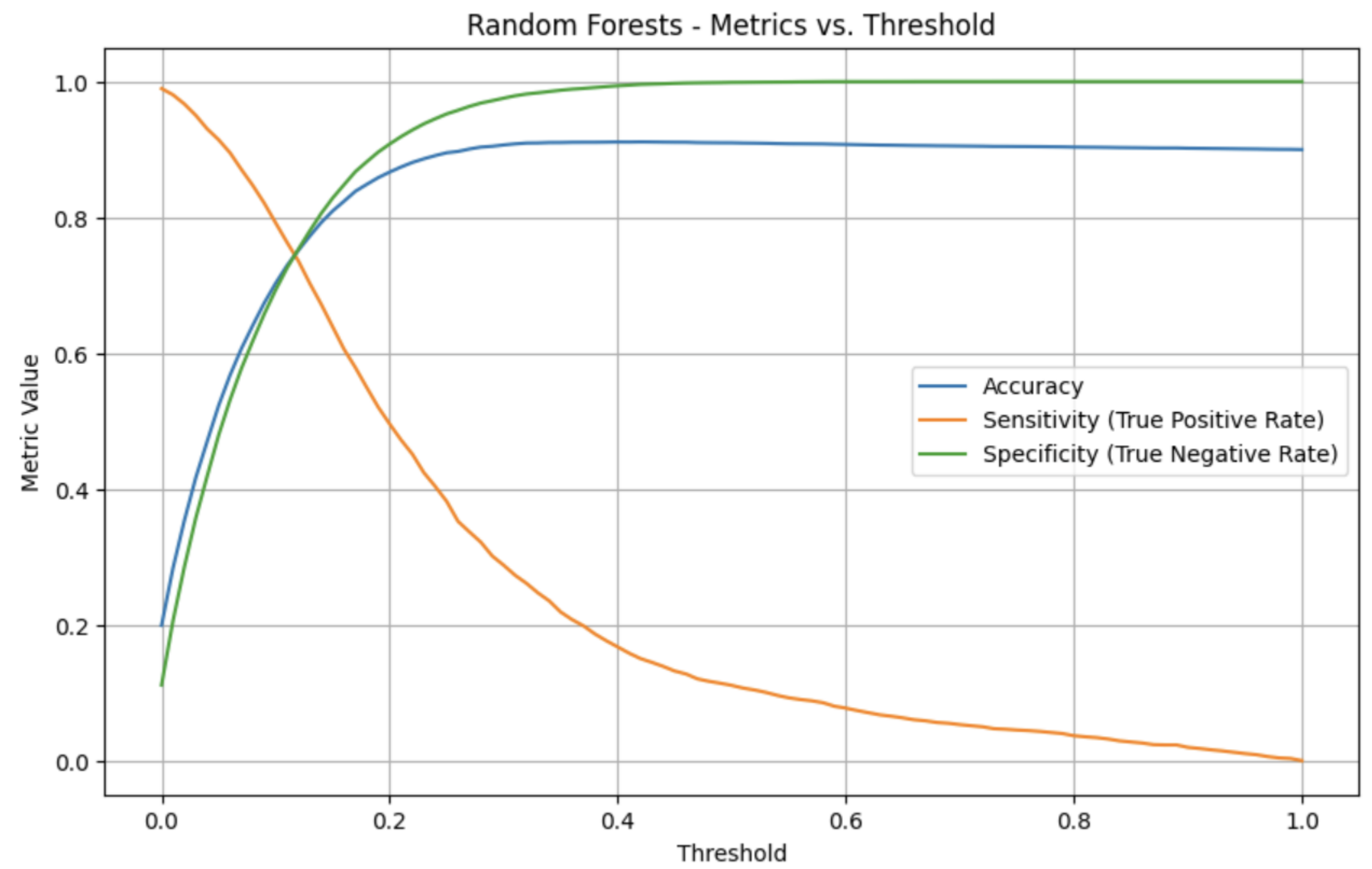


Exhibit 5. Feature Importance (Random Forest)

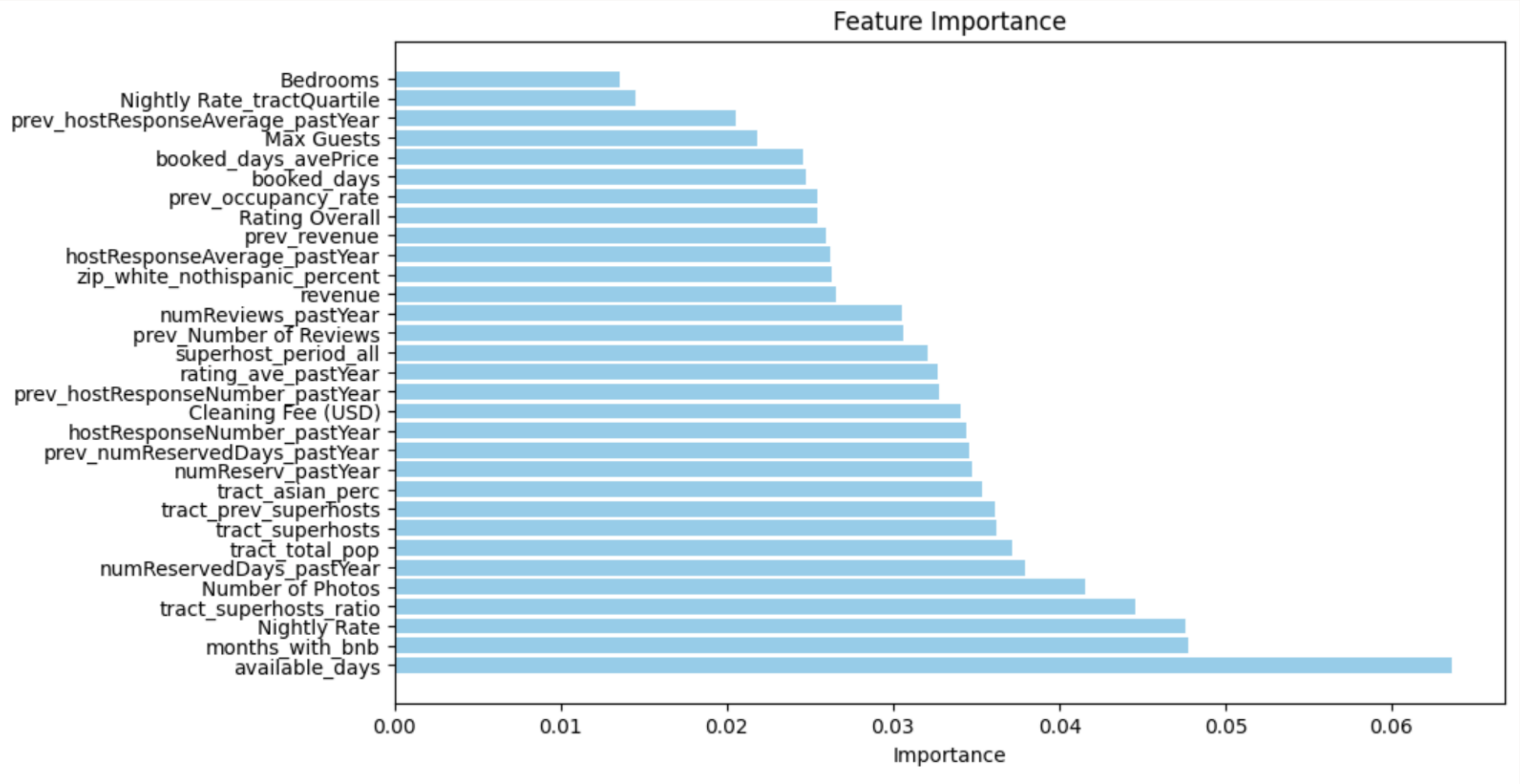


Exhibit 6. Optimal Threshold and other metrics of Gradient Boosting

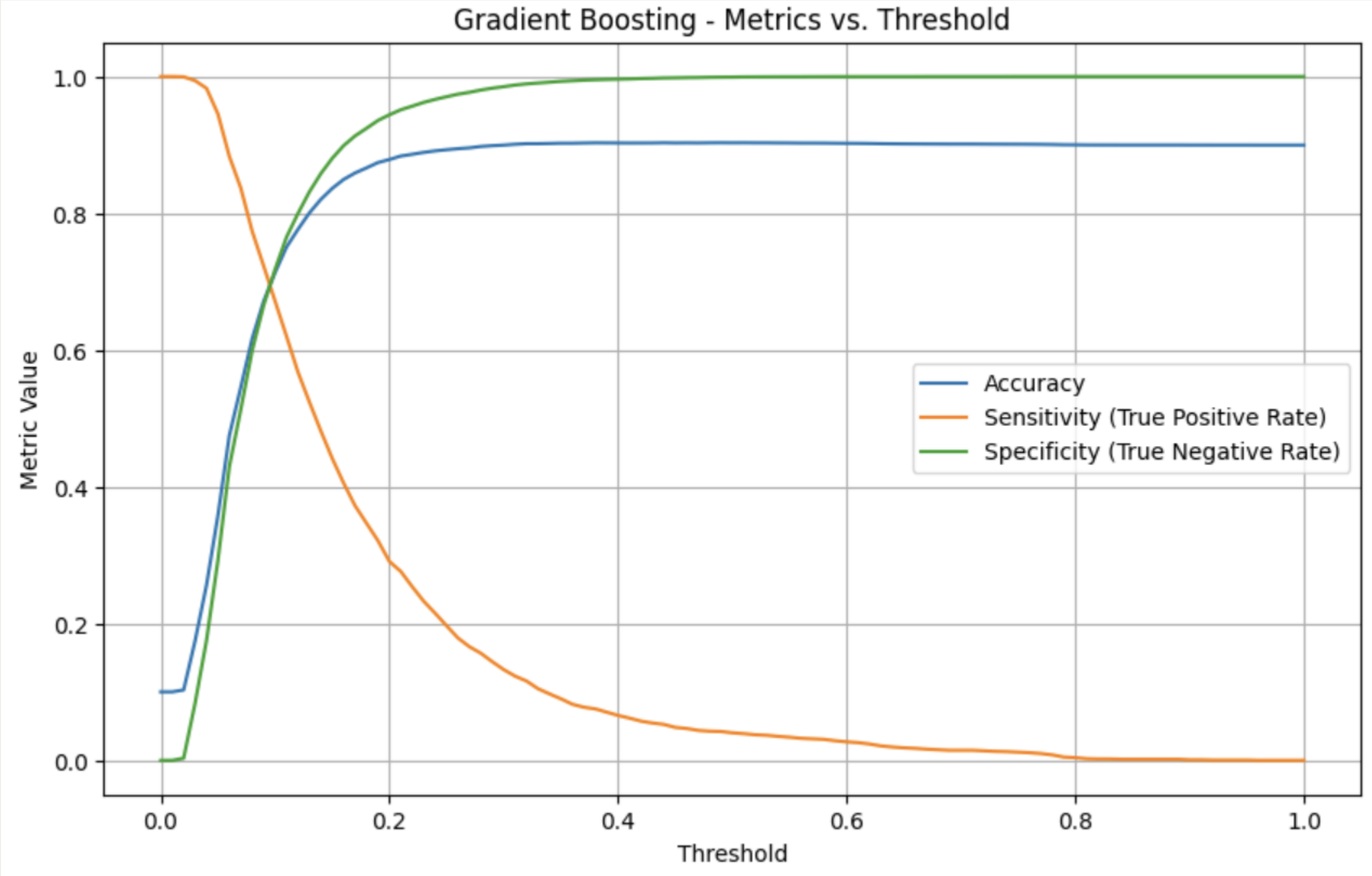


Exhibit 7. Feature Importance (Gradient Boosting)

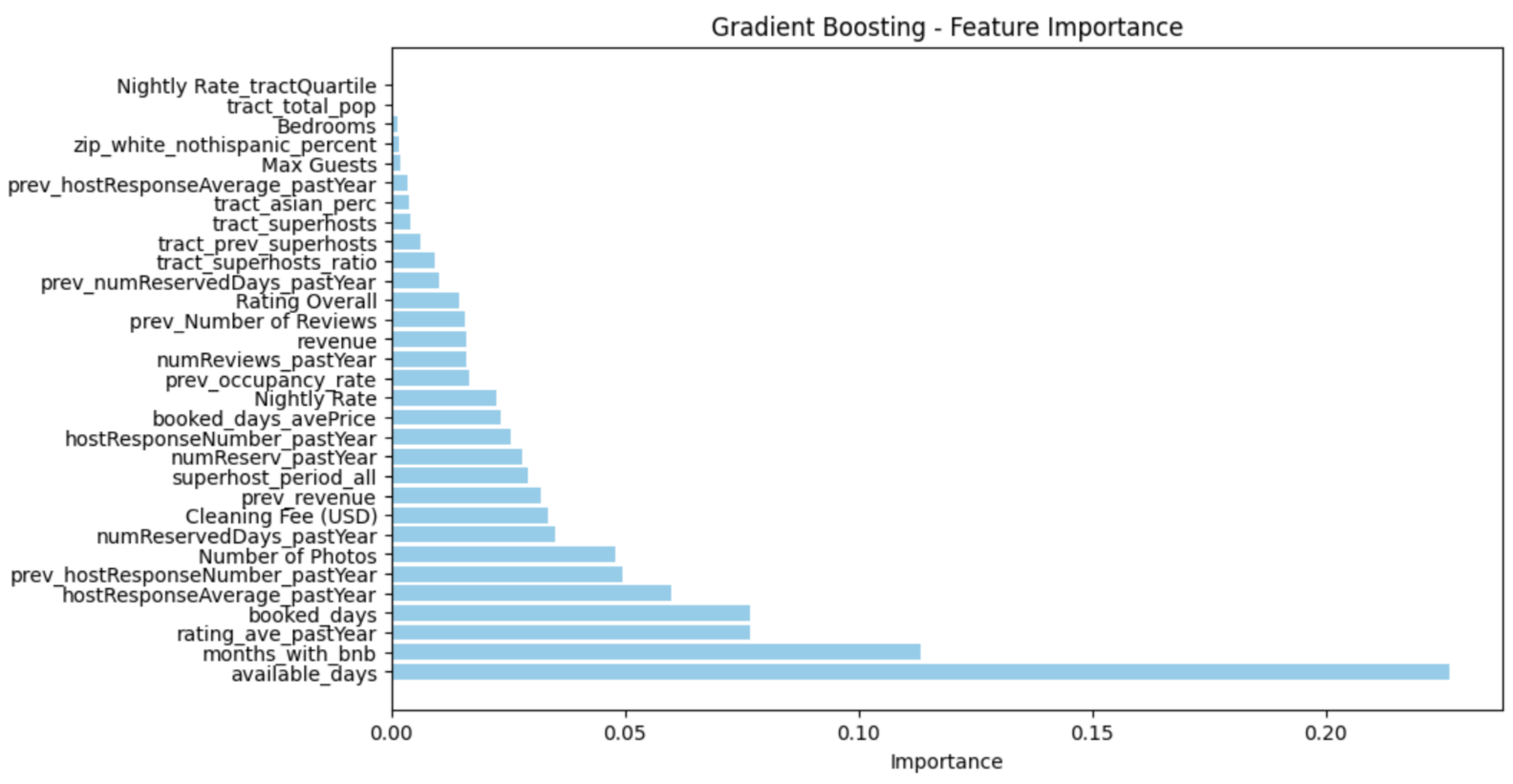


Exhibit 8.
Logistic Regression Result (Revenue Analysis)

Logit Regression Results

Dep. Variable:	revenue_label	No. Observations:	125196
Model:	Logit	Df Residuals:	125165
Method:	MLE	Df Model:	30
Date:	Thu, 07 Dec 2023	Pseudo R-squ.:	0.3595
Time:	19:05:15	Log-Likelihood:	-55282.
converged:	True	LL-Null:	-86315.
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.5556	0.145	3.835	0.000	0.272	0.840
rating_ave_pastYear	0.2198	0.027	8.260	0.000	0.168	0.272
numReviews_pastYear	-0.0007	6.31e-05	-11.313	0.000	-0.001	-0.001
numReservedDays_pastYear	0.0002	1.55e-05	9.897	0.000	0.000	0.000
numReserv_pastYear	-0.0005	2.8e-05	-16.982	0.000	-0.001	-0.000
prev_numReservedDays_pastYear	-1.916e-05	1.91e-05	-1.005	0.315	-5.65e-05	1.82e-05
hostResponseNumber_pastYear	-0.0032	0.001	-6.203	0.000	-0.004	-0.002
hostResponseAverage_pastYear	-0.0002	0.001	-0.153	0.878	-0.002	0.002
prev_hostResponseNumber_pastYear	0.0014	0.001	2.615	0.009	0.000	0.003
prev_hostResponseAverage_pastYear	0.0001	0.001	0.118	0.906	-0.002	0.002
available_days	-0.0011	0.000	-8.624	0.000	-0.001	-0.001
booked_days	0.0081	0.001	14.857	0.000	0.007	0.009
booked_days_avePrice	-0.0104	0.000	-62.753	0.000	-0.011	-0.010
Bedrooms	0.2126	0.014	15.744	0.000	0.186	0.239
Max Guests	-0.0431	0.005	-8.140	0.000	-0.053	-0.033
Cleaning Fee (USD)	-0.0003	0.000	-1.507	0.132	-0.001	9.02e-05
Number of Photos	-0.0366	0.001	-41.458	0.000	-0.038	-0.035
Nightly Rate	0.0051	0.000	51.332	0.000	0.005	0.005
prev_Number of Reviews	-0.0260	0.000	-68.990	0.000	-0.027	-0.025
Rating Overall	-0.0091	0.001	-14.476	0.000	-0.010	-0.008
prev_revenue	-0.0005	6.52e-06	-76.550	0.000	-0.001	-0.000
prev_occupancy_rate	1.5534	0.063	24.674	0.000	1.430	1.677
tract_total_pop	-1.622e-05	5.44e-06	-2.979	0.003	-2.69e-05	-5.55e-06
tract_asian_perc	0.0242	0.003	9.124	0.000	0.019	0.029
zip_white_nothispanic_percent	0.0052	0.000	11.112	0.000	0.004	0.006
Nightly Rate_tractQuartile	-0.0003	0.009	-0.032	0.974	-0.017	0.017
tract_superhosts	-0.0141	0.002	-7.307	0.000	-0.018	-0.010
tract_superhosts_ratio	-0.0125	0.090	-0.140	0.889	-0.188	0.163
tract_prev_superhosts	0.0183	0.002	9.649	0.000	0.015	0.022
months_with_bnb	0.0281	0.000	58.112	0.000	0.027	0.029
Prop_Churned	0.0554	0.024	2.303	0.021	0.008	0.103

References

- Is Airbnb broken? - https://finshots.in/archive/is-airbnb-broken/

Team Composition

Nagarjuna Chidarala

Sai Teja Devalla

Seonkyu Kim

Chaitanya Sanaboina


Thank you