SUPERHOST CHURN PREDICTION

TEAM 12

MEET OUR TEAM

Roshan Raj Singh

Shourya Chouhan

Shubhankar Sharma

Sohan Sahoo

IDEATION

- Revenue prediction and saturation of Superhosts in a Neighborhood
- Churn Prediction
- Optimization of the current dynamics
- Impact on Non Superhost properties if there is an increase in Number of Superhosts

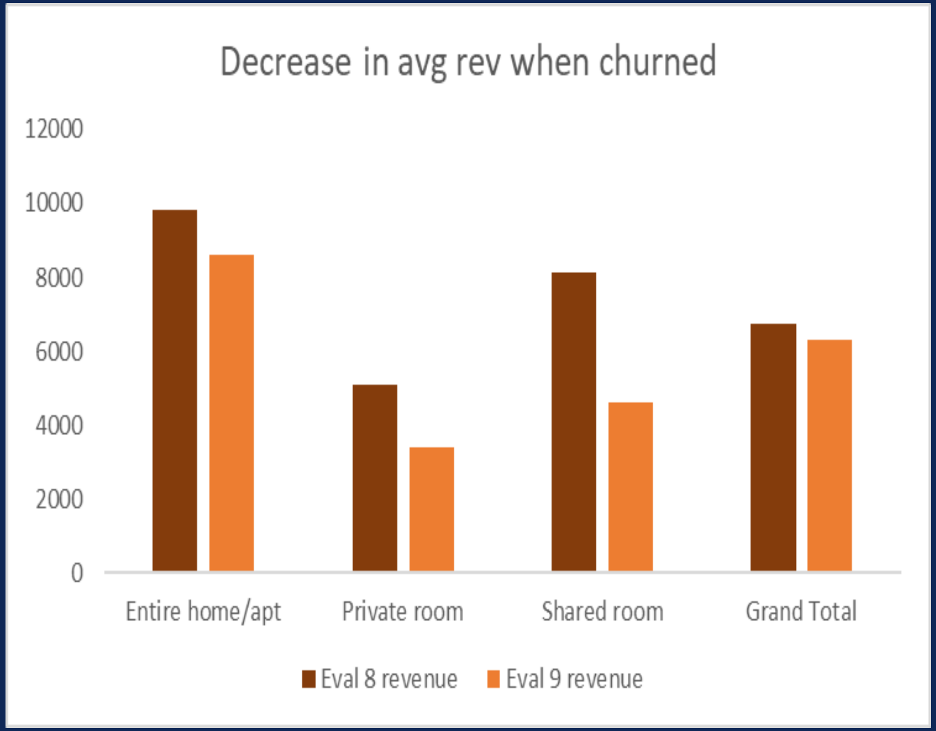
PROBLEM DEFINITION

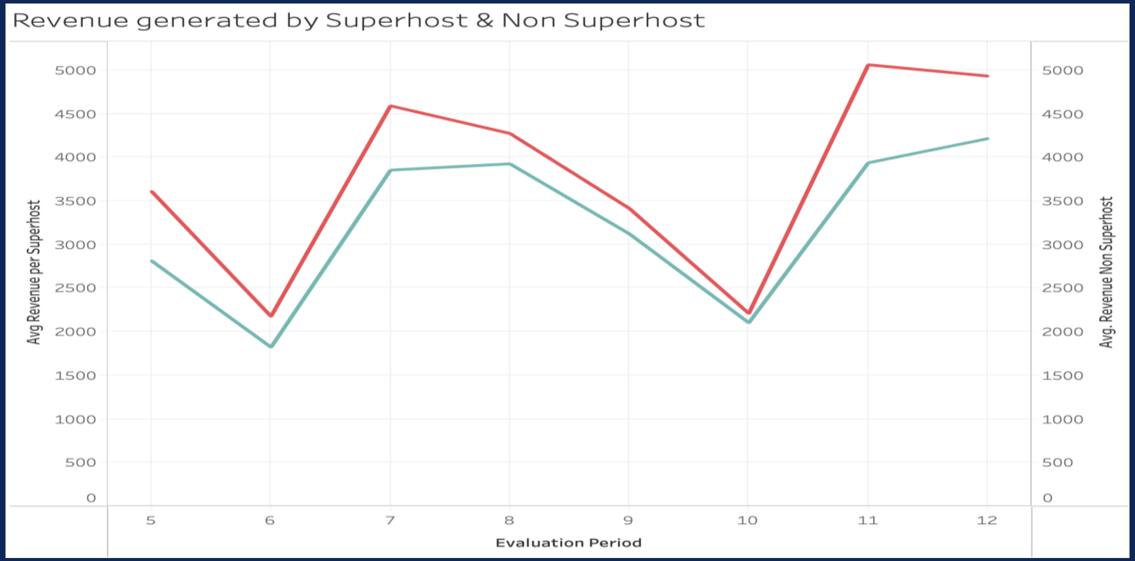
To predict whether a Superhost in the current evaluation period is going to churn in the next evaluation period, i.e. whether a Superhost is going to lose their Superhost status in the next evaluation period.

BUSINESS PROBLEM

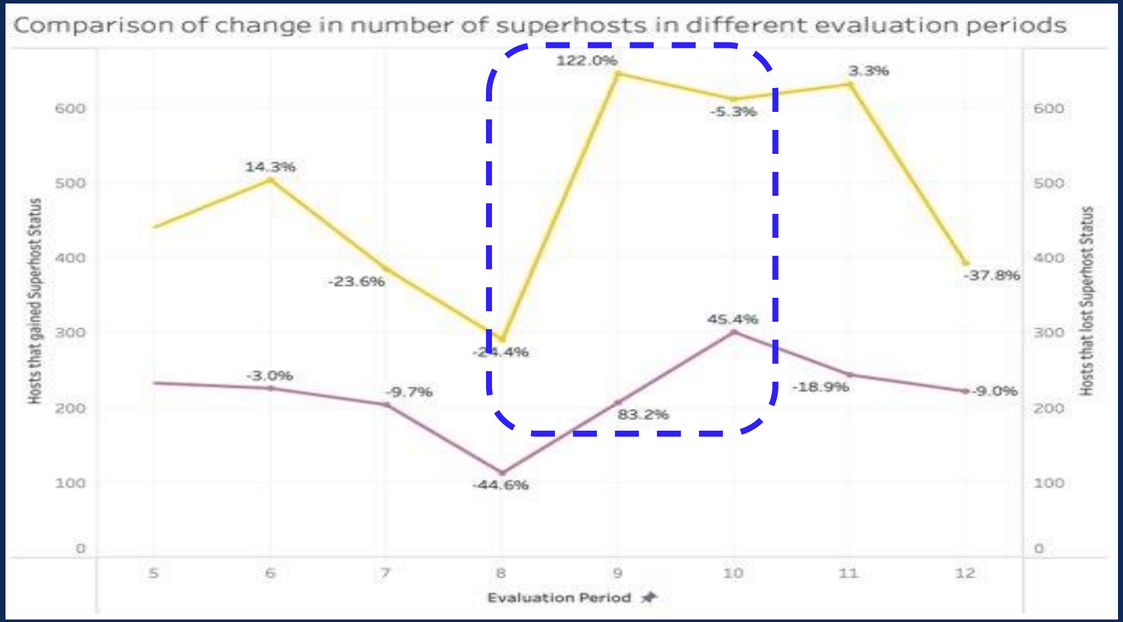
Superhost churn business impact

Average Loss of Revenue	12%
Average Loss of Reviews	21%
Average Loss of Occupancy	18%





Avg revenue of Superhost properties is higher than Non Superhost properties



Rate of net change in Superhost has decreased

PROCESS FLOW

Sample	Explore	Modify	Model	Assess
Choosing the right cities as well as evaluation period for enough and valuable datapoints	Understanding the dataset and exploring dependency of target with predictor variables	Cleaning which involved dealing with missing values and outliers, normalising the data, making it ready for the model	Ran different models and tuning parameters to figure the best fit for the dataset	Assess what factors are causing the superhosts to lose their superhost status

CHALLENGES FACED

Over 30% of the dataset contained missing values

During the data preprocessing phase, diverse columns required aggregation using distinct aggregation methods.

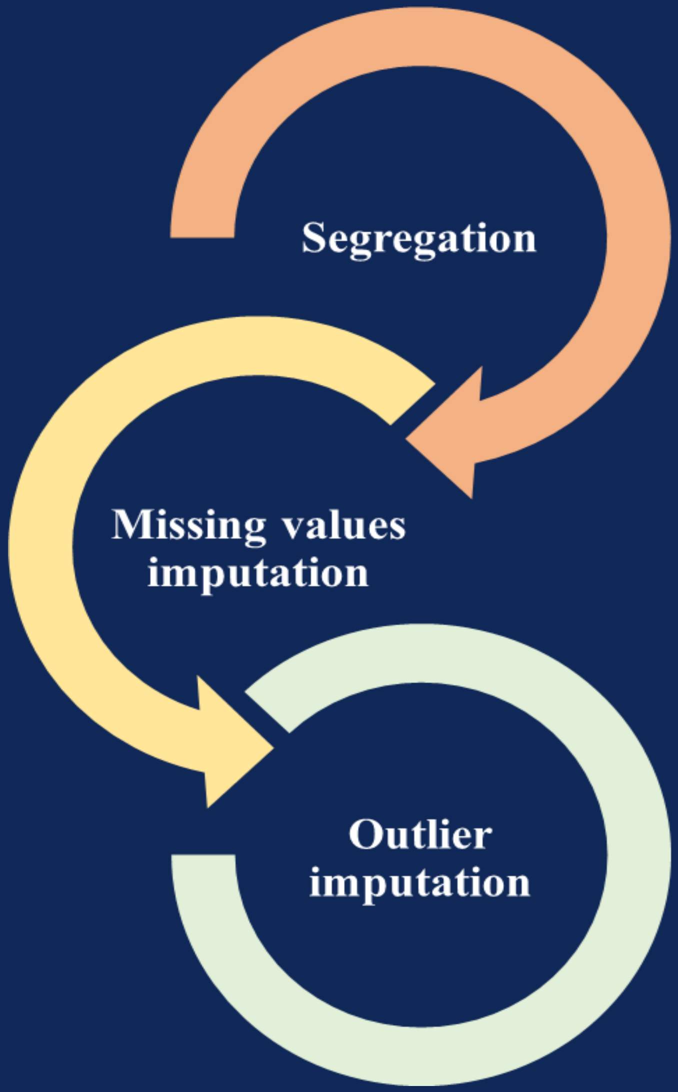
Encountering a substantial imbalance in our dataset, with Class 0 comprising 7185 instances and Class 1 only 1423 instances

To rectify this imbalance, we implemented the Synthetic Minority Over-sampling Technique (SMOTE)

DATA PREPARATION PROCESS

- Established the criteria for selecting Superhost IDs, mapping each ID to its churn status in the upcoming evaluation period
- Determined the optimal evaluation period required for training the model to accurately predict Superhost churn in the subsequent assessment period
- Procedure was executed across eight diverse cities, ensuring an ample and varied dataset for robust model training

DATA PREPROCESSING

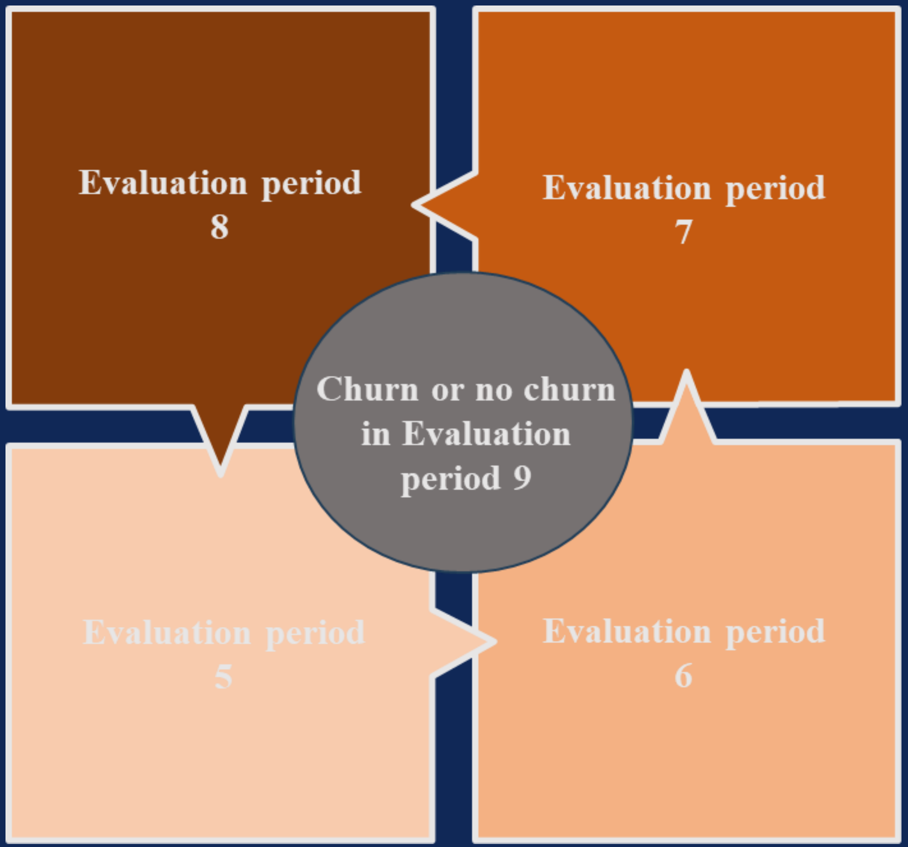


- Identified of numerical and categorical columns
- Converted specified columns to string or numeric type

- Imputed missing values for interval columns using the group median based on 'Airbnb Property ID'
- Imputed missing values for binary columns using the group mode based on 'Airbnb Property ID'
- Impute missing values for categorical columns with the mode value within each property type based on 'Airbnb Property ID'

- Applied the lower fence imputation function to impute missing values in numerical columns

DATA AGGREGATION



Sum

Prev_revenue

Prev_numreviews

Average

Prev_ratings

Prev_occupancy

MODEL BUILDING PROCESS

- Data Preprocessing
- Variable Selection
- Dataset Partitioning
- Class Imbalance Handling
- Model Exploration
- Hyperparameter Tuning
- Threshold Optimization

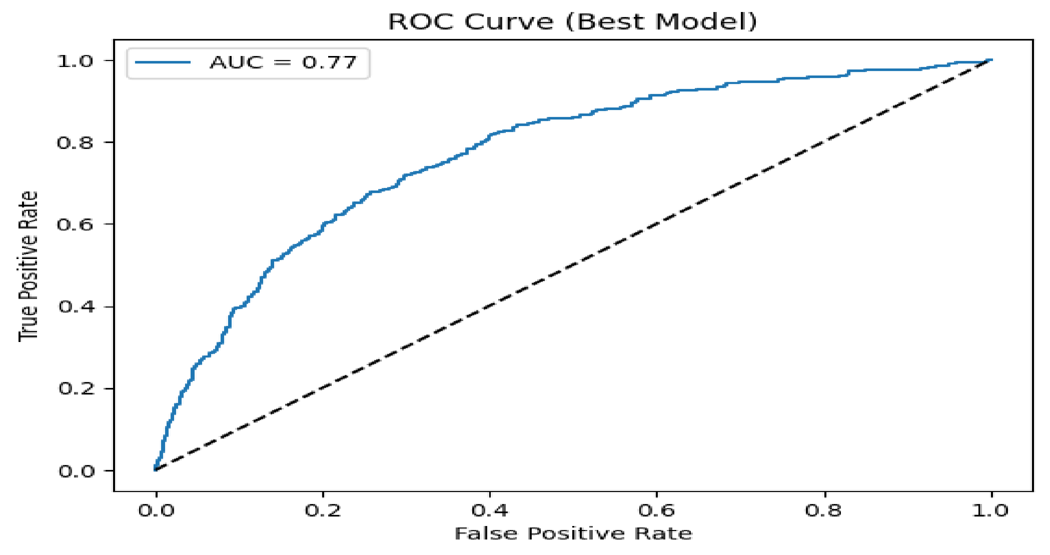
MODEL COMPARISION AND SELECTION

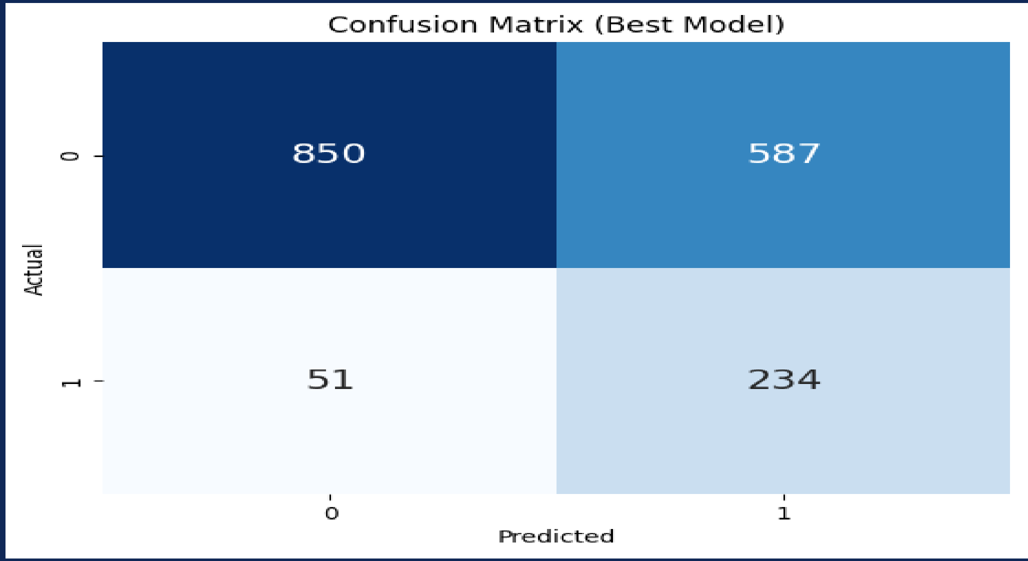
- Logistic Regression
- Decision Tree
- Random Forest
- Gradient Boosting
- Neural Network

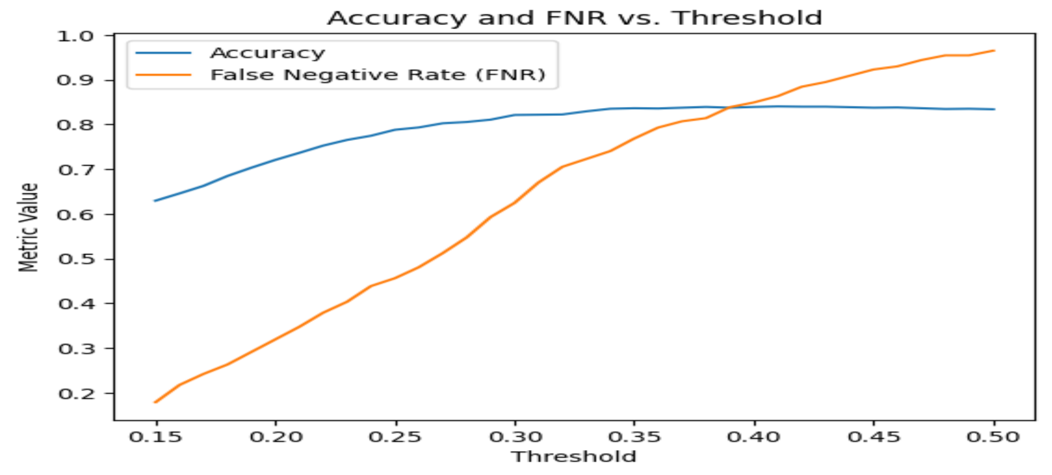
Without Threshold Optimisation					
Metrics for Comparision	Logistic Regression	Decision Tree	Random Forest	Gradient Boosting	Neural Network
Accuracy	0.83	0.8	0.85	0.83	0.75
ROC	0.77	0.64	0.79	0.64	0.63
Sensitivity	0.04	0.41	0.18	0.33	0.45
Interpretability of Result	High	High	Low	Low	Very Low

With Threshold Optimisation					
Metrics for Comparision	Logistic Regression	Decision Tree	Random Forest	Gradient Boosting	Neural Network
Accuracy	0.84	0.8	0.68	0.83	0.8235
ROC	0.77	0.64	0.8	0.79	0.71
Sensitivity	0.82	0.41	0.8	0.77	0.72
Interpretability of Result	High	High	Low	Low	Very Low
Result	Choosen because of High interpretability and High Sensitivity	No Improveme	Lower Sensitivity than Logistic Regression	Not Choosen because of lower Precision and ROC than Random Forest	Low Sensitivity, not choosen

BEST MODEL – LOGISTIC REGRESSION



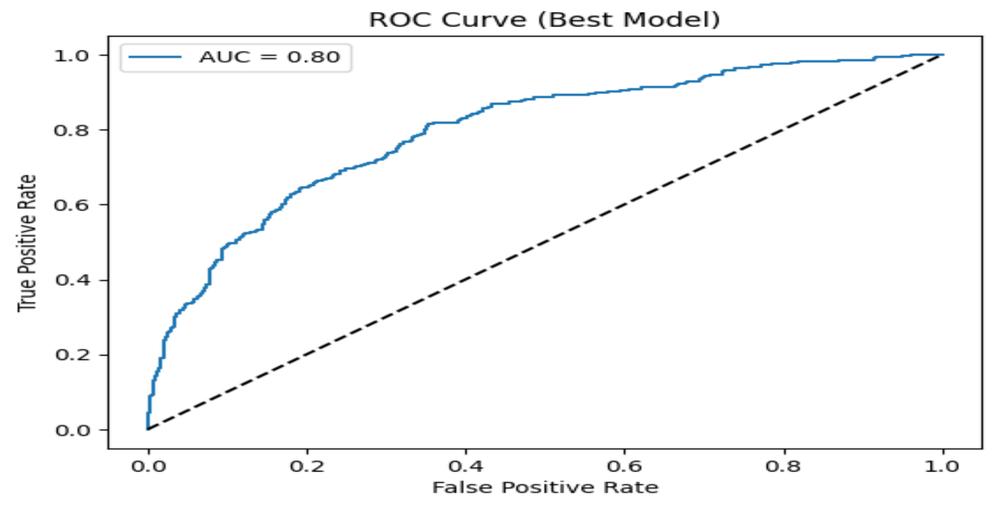


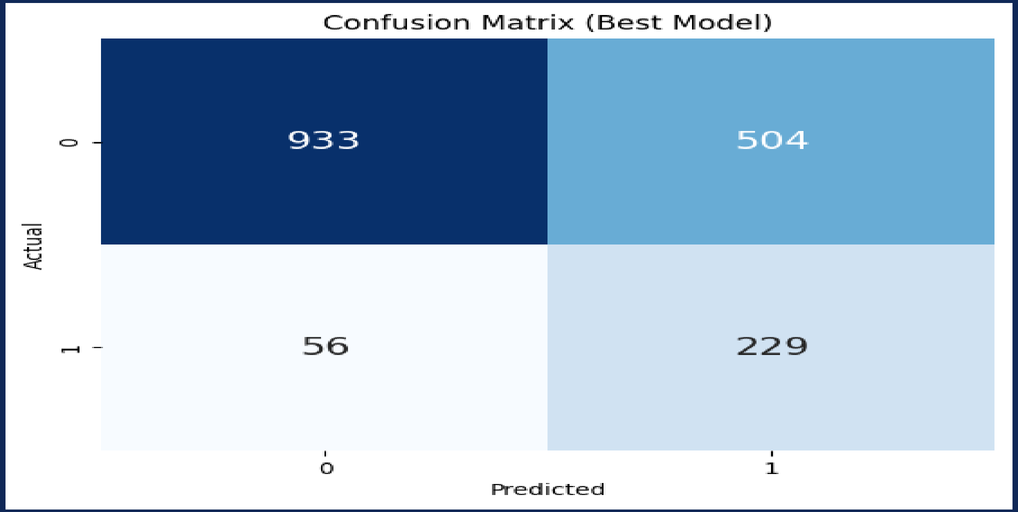


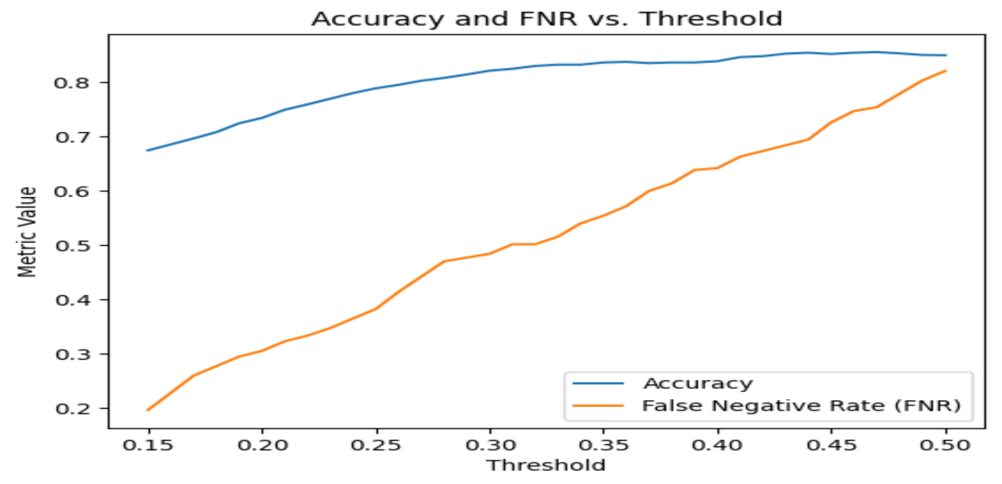
Cost vs Accuracy

- 234 out of 285 actual churned hosts were correctly identified
- Took a hit on accuracy to decrease the false negative rate
- Allows to correctly capture those hosts with a high propensity to churn than random forest
- High Sensitivity, easy Interpretability

RANDOM FOREST RESULTS







Cost vs Accuracy

- 229 out of 285 actual churned hosts were correctly identified
- Took a hit on accuracy to decrease the false negative rate
- Allows to correctly capture those hosts with a high propensity to churn

PREDICTED PROBABILITY OF CHURN

Top 1% of Airbnb Host IDs with Max Probabilities:

	Airbnb Host ID	churn_prob
441	1314045	0.970245
3596	105117303	0.863864
1362	7706697	0.855675
998	4204562	0.855645
2018	19962052	0.851194
3152	62328018	0.812814
637	2092314	0.772158
875	3407346	0.766331
958	3931950	0.761759
2326	26228378	0.740260
2534	33163646	0.736711
1358	7678072	0.716573
2983	50573299	0.711565
3167	63383282	0.710371
3463	93148942	0.701796
770	2689908	0.692237
1846	15613411	0.671532
3516	97513787	0.657037
1951	18053985	0.656661
2164	22652051	0.647604
3509	97022024	0.641047
1749	13161042	0.637497

Prediction

- Evaluation period 10 taken into consideration
- Aggregation of data for evaluation period 6, 7, 8 and 9 as suggested previously in data modeling stage
- Predicted churn for different Host IDs with probabilities

Suggested actions that can be taken by Airbnb

- Analyze internal and external factors contributing to churn specific to these host property's location and property type
- Airbnb can proactively push the superhosts to meet the requirements
- Better planning to nullify the churn by converting non-superhosts to superhosts

CONCLUSION

Provide operational auditing services to Airbnb superhosts to help them retain their superhost status

Understand which superhosts are going to lose their superhost status in the next evaluation period

Identify the factors that lead to the change in superhost status and how each factor affects the churn rate

Help superhosts by advising them on the said factors to improve on the service and retain their superhost status for the next evaluation period.

EXTRA SLIDES FOR REFERENCE

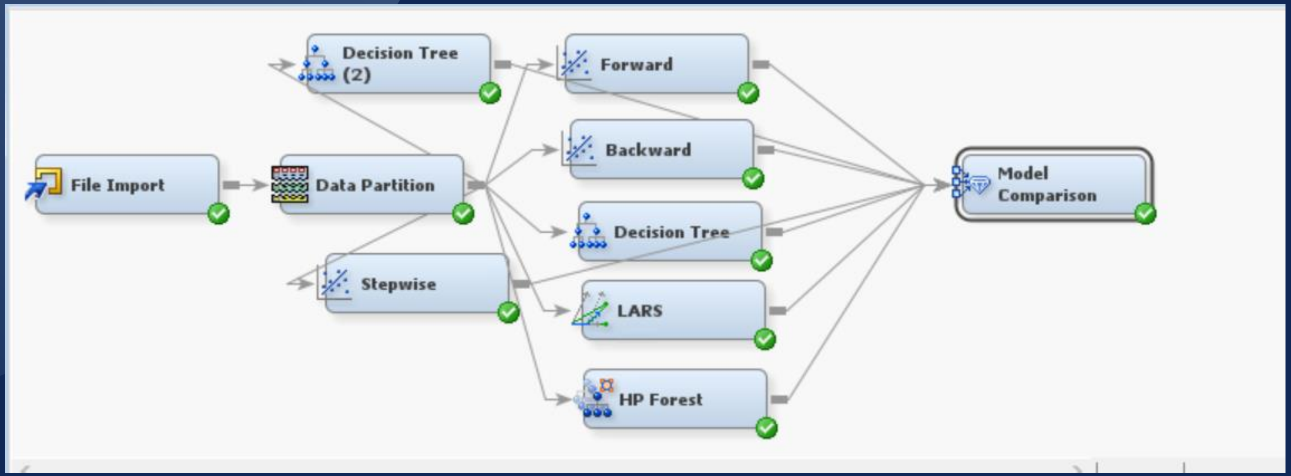
TOOLS USED

- Excel
- SAS Enterprise Miner
- Tableau
- Jupyter And Colab

DATA AND ITS SUMMARY

	Superhost	Non Superhost
Average Revenue	$4252.83	$3977.14
Proportion of Hosts	7185	1424
Average Number of Reviews	39.66	26.06

BACKWARD REGRESSION




```
The selected model, based on the misclassification rate for the validation data, is the model trained in Step 6. It consists of the following effects:

Intercept  Bedrooms  Instantbook_Enabled  Max_Guests  Minimum_Stay  Nightly_Rate  Number_of_Photos  Number_of_Reviews  Rating_Overall  booked_days_period_city  hostRespo
numReserv_pastYear  numReservedDays_pastYear  numReviews_pastYear  num_5_star_Rev_pastYear  occupancy_rate  prev_Nightly_Rate  prev_Rating_Overall  prev_available_days  
prev_hostResponseNumber_pastYear  prev_numCancel_pastYear  prev_numReserv_pastYear  prev_numReviews_pastYear  prev_num_5_star_Rev_pastYear  prev_occupancy_rate  prev_pro
prop_5_StarReviews_pastYear  revenue  revenue_period_city  tract_asian_perc  tract_black_perc  tract_housing_units  tract_total_pop  tract_white_perc  zip_asian_nothispa
zip_black_nothispanic_percent  zip_hispanic_or_latino_anyrace  zip_hispanic_or_latino_anyrace_p  zip_total_population  zip_white_nothispanic  zip_white_nothispanic_perce
```

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

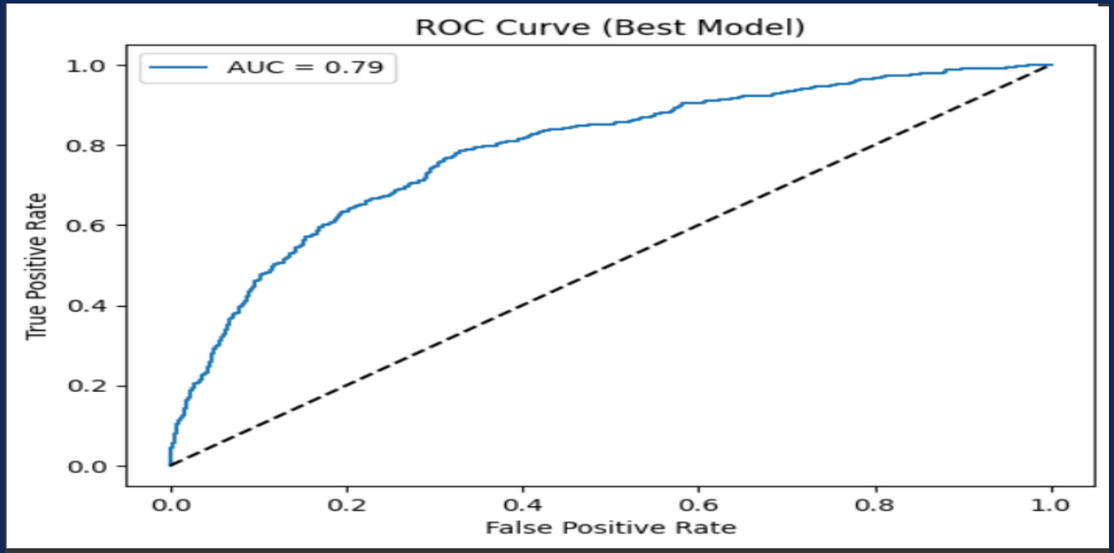
-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
4628.275	3981.262	647.0129	49	<.0001

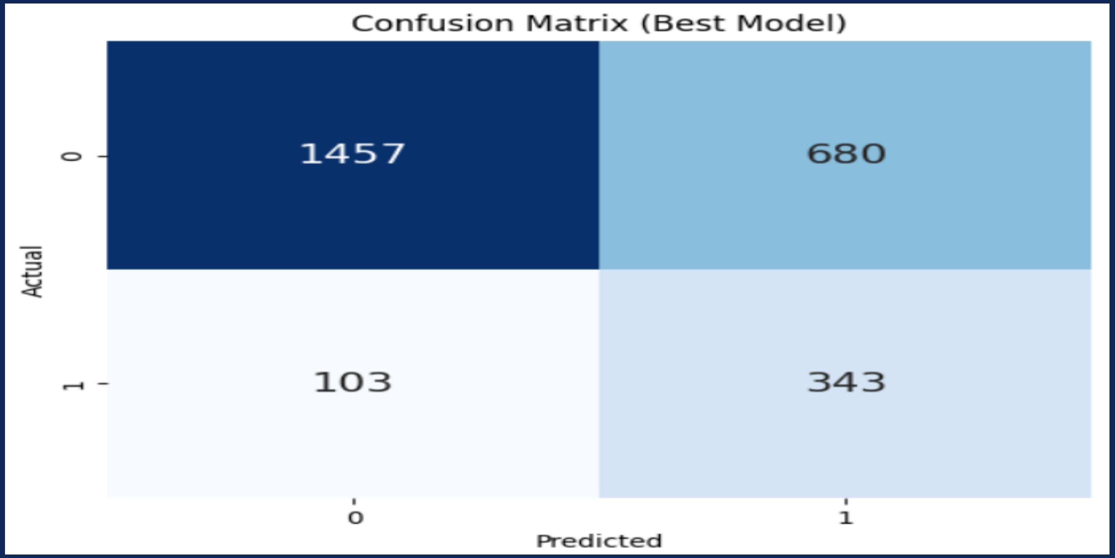
LOGISTIC REGRESSION RESULTS

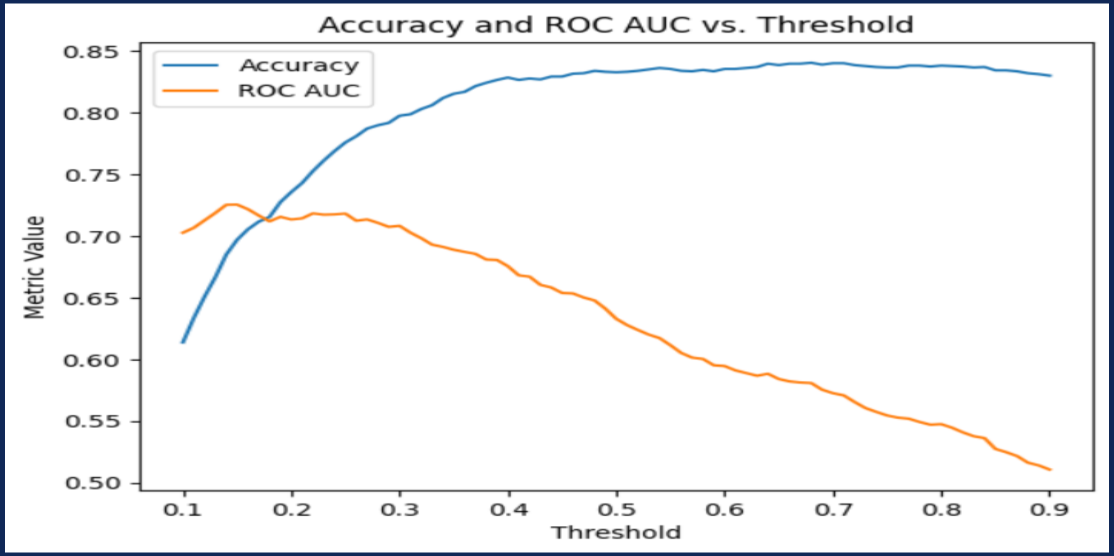
	coef	std err	z	P>|z|	[0.025	0.975]
const	9.6045	3.329	2.885	0.004	3.080	16.129
rating_ave_pastYear	-0.2981	1.164	-0.256	0.798	-2.579	1.982
numReviews_pastYear	0.3542	0.041	8.720	0.000	0.275	0.434
numCancel_pastYear	0.6129	0.408	1.501	0.133	-0.188	1.413
num_5_star_Rev_pastYear	-0.4411	0.048	-9.284	0.000	-0.534	-0.348
prop_5_StarReviews_pastYear	-2.1415	1.762	-1.216	0.224	-5.594	1.311
prev_rating_ave_pastYear	-0.8132	0.837	-0.971	0.332	-2.454	0.828
prev_numReviews_pastYear	-0.0781	0.032	-2.410	0.016	-0.142	-0.015
prev_numCancel_pastYear	0.1000	0.181	0.551	0.582	-0.256	0.456
prev_num_5_star_Rev_pastYear	0.1027	0.037	2.779	0.005	0.030	0.175
prev_prop_5_StarReviews_pastYear	0.8737	1.273	0.686	0.492	-1.621	3.368
numReservedDays_pastYear	-0.0020	0.001	-1.485	0.137	-0.005	0.001
numReserv_pastYear	0.0085	0.007	1.256	0.209	-0.005	0.022
prev_numReservedDays_pastYear	0.0004	0.001	0.322	0.747	-0.002	0.003
prev_numReserv_pastYear	-0.0055	0.007	-0.751	0.453	-0.020	0.009
hostResponseNumber_pastYear	-0.0064	0.015	-0.436	0.663	-0.035	0.022
hostResponseAverage_pastYear	-0.0420	0.015	-2.750	0.006	-0.072	-0.012
prev_hostResponseNumber_pastYear	0.0012	0.009	0.125	0.900	-0.017	0.020
prev_hostResponseAverage_pastYear	0.0118	0.013	0.898	0.369	-0.014	0.037
prev_available_days	8.667e-05	5.99e-05	1.447	0.148	-3.07e-05	0.000
prev_available_days_aveListedPrice	0.0005	0.001	0.402	0.688	-0.002	0.003
prev_booked_days	-0.0139	0.004	-3.619	0.000	-0.021	-0.006
prev_booked_days_avePrice	0.0006	0.001	0.684	0.494	-0.001	0.002
Bedrooms	0.0233	0.070	0.331	0.741	-0.115	0.161
Bathrooms	-0.0706	0.098	-0.721	0.471	-0.263	0.121
Max Guests	0.0020	0.027	0.074	0.941	-0.050	0.054
Minimum Stay	0.0006	0.001	0.608	0.543	-0.001	0.003
Number of Photos	-0.0099	0.003	-3.036	0.002	-0.016	-0.004
Instantbook Enabled	0.0801	0.086	0.935	0.350	-0.088	0.248
Nightly Rate	-0.0005	0.001	-0.903	0.366	-0.001	0.001
prev_Nightly Rate	0.0004	0.001	0.553	0.580	-0.001	0.002
Number of Reviews	-0.0075	0.002	-3.895	0.000	-0.011	-0.004
prev_Number of Reviews	0.0003	0.000	0.604	0.546	-0.001	0.001
Rating Overall	-0.0553	0.011	-4.861	0.000	-0.078	-0.033
prev_Rating Overall	0.0493	0.013	3.761	0.000	0.024	0.075
revenue	-2.807e-06	1.07e-05	-0.262	0.793	-2.38e-05	1.82e-05
occupancy_rate	0.0322	0.236	0.137	0.891	-0.430	0.494
prev_revenue	9.625e-06	1.73e-05	0.558	0.577	-2.42e-05	4.35e-05
prev_occupancy_rate	1.7715	0.368	4.811	0.000	1.050	2.493

- Key variables like 'rating_avg_pastyear', 'numcancel_pastyear', 'prop_5_starreviews', 'prev_bookeddays', 'no_of_photos' were identified based on p value

GRADIENT BOOSTING RESULTS



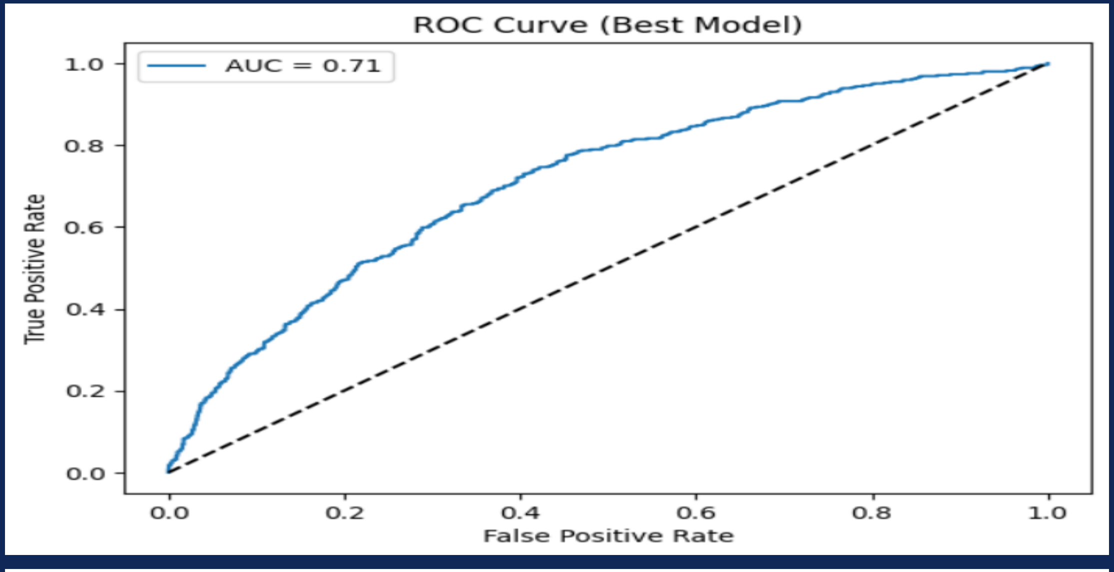


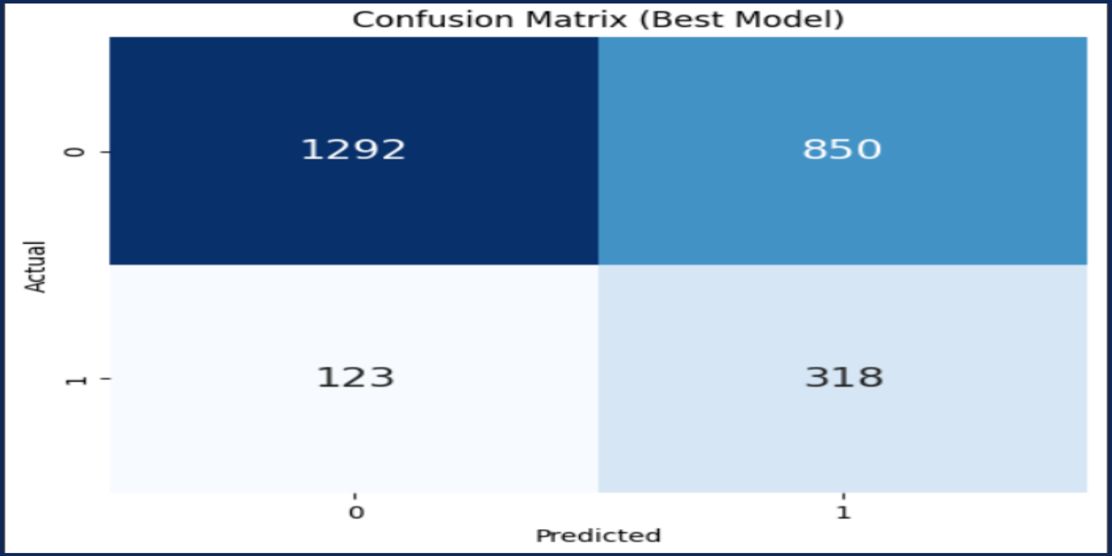


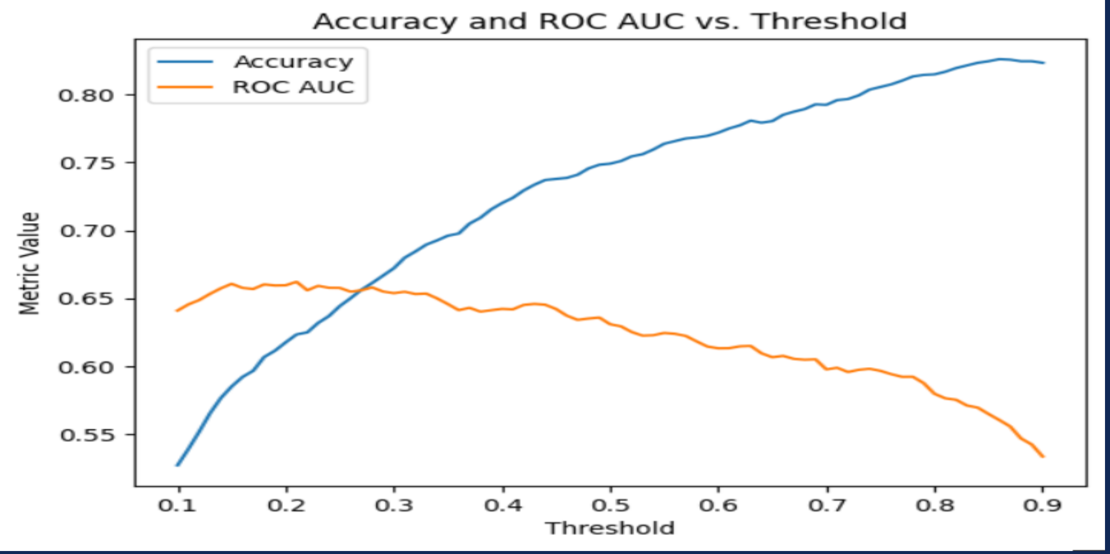
Cost vs Accuracy

- Train Test Split of 70:30
- 343 out of 441 actual churned hosts were correctly identified
- Took a hit on precision to decrease the overall accuracy of predicting churned superhosts
- Allows to correctly capture those hosts with a high propensity to churn

NEURAL NETWORK RESULTS







Cost vs Accuracy

- Train Test Split of 70:30
- 318 out of 441 actual churned hosts were correctly identified
- Took a hit on precision to decrease the overall accuracy of predicting churned superhosts
- Allows to correctly capture those hosts with a high propensity to churn

THANK YOU