NYC AIRBNB ANALYSIS

CS5525
Taught by Dr. Reza Jafari

Aniruddha Hore
[aniruddhah@vt.edu]

Data Analyticss

BREAKDOWN

Abstract	Introduction
Description	Analysis Phase I & II
Analysis Phase III & IV	Recommendations
Conclusion	References

ABSTRACT

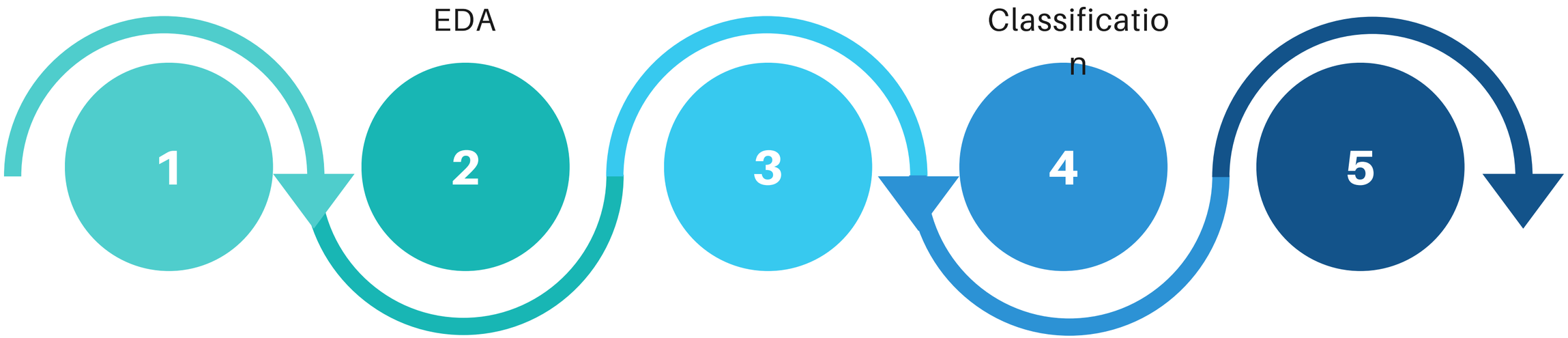
This project explores the New York City Airbnb dataset to identify key factors that impact rental prices using exploratory data analysis and machine learning techniques. The results provide insights and recommendations for property owners and renters looking to invest in or rent properties in the NYC rental market.

INTRODUCTION

Our project focuses on analyzing the NYC Airbnb dataset, which contains information on over 48,000 listings, to gain insights into the rental market and identify factors that impact rental prices. Using exploratory data analysis and machine learning techniques such as regression analysis and feature selection algorithms, we aim to create a predictive model that can help property owners and renters make informed decisions. Ultimately, our project aims to provide valuable insights into the NYC rental market and help stakeholders optimize their rental prices and rental business.

PROGRAM FLOW

Order of Operations



Data Preprocessing

EDA

Regression

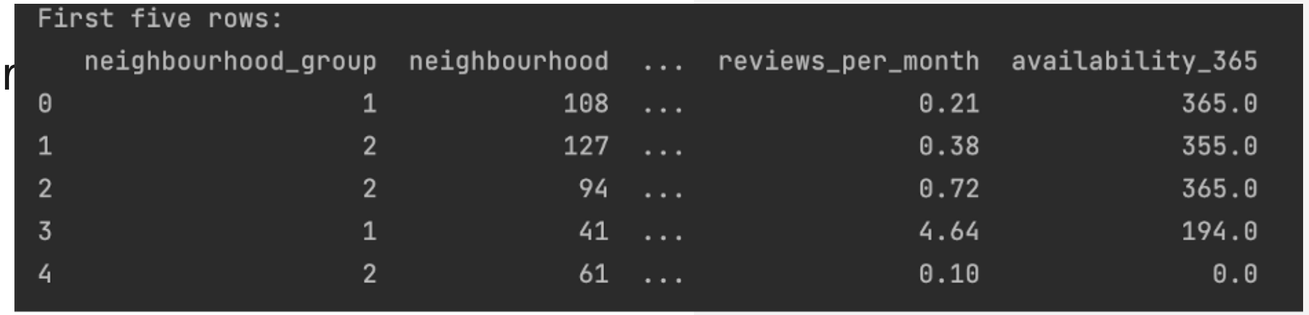
Classification

Clustering

DATASET

Description

- Information on over 48,000 Airbnb listings in New York City.
- 16 features, including the listing's location, price, and room type.
- The average price of a listing in NYC is $152 per night,
- Most common room type is "Entire home/apt,"



First five rows:

	neighbourhood_group	neighbourhood	...	reviews_per_month	availability_365
0	1	108	...	0.21	365.0
1	2	127	...	0.38	355.0
2	2	94	...	0.72	365.0
3	1	41	...	4.64	194.0
4	2	61	...	0.10	0.0

EDA

Data cleaning

- Handling missing values is an important step in data preprocessing.
- Common approaches include dropping missing values or imputing them with estimated values.
- In the New York City Airbnb dataset used for this project, there are no missing values.
- The dropna() function in pandas can be used to remove rows or columns with missing values.

```
Missing values:
neighbourhood_group    0
neighbourhood          0
room_type              0
price                  0
minimum_nights         0
reviews_per_month      0
availability_365       0
dtype: int64
```

EDA

Discretization & Binarization:

- Label Encoding assigns a unique integer value to each category in a feature with hierarchy, while One-Hot Encoding represents each category in a feature as a separate binary feature with no hierarchy.
- One-Hot Encoding was used for our dataset since there was no hierarchy among the categories in our dataset, and each category was equally important.
- One-Hot Encoding allowed us to represent each category in a feature as a separate binary feature, which is useful for algorithms that assume numerical inputs.

```
One-Hot Encoding:
   reviews_per_month  availability_365  ...  room_type_1  room_type_2
0               0.21             365.0  ...         True        False
1               0.38             355.0  ...        False        False
2               0.72             365.0  ...         True        False
3               4.64             194.0  ...        False        False
4               0.10               0.0  ...        False        False

[5 rows x 9 columns]
```

DIMENSIONALITY REDUCTION

Random Forest

- Dimensionality reduction reduces the number of features in a dataset while retaining important information.
- Random Forest Analysis was used to identify the most important features,
- The top six features with highest importance in predicting the price are 'reviews_per_month', 'availability_365', 'neighbourhood', 'minimum_nights', 'neighbourhood_group', and 'room_type',

```
Random Forest Analysis:
reviews_per_month       0.391815
availability_365        0.275760
neighbourhood           0.188857
minimum_nights          0.118940
neighbourhood_group     0.013061
room_type               0.011568
dtype: float64
```

DIMENSIONALITY REDUCTION

PCA & SVD

- In Principal Component Analysis (PCA), the explained variance ratio tells us the proportion of the total variance in the data that is explained by each principal component.
- In Singular Value Decomposition (SVD), the singular values represent the square roots of the eigenvalues of the covariance matrix of the data. The larger the singular value, the more important the corresponding feature is for explaining the variation in

```
Principal Component Analysis:
Explained variance ratio: [0.77161662 0.2099471 ]
Singular Value Decomposition Analysis:
Singular values: [42437.77911386 21519.81671404  4497.64015951   361.37470143
   215.17677393   124.46585675]
```

STANDARDIZATION

Variable Transformation

- involves z-score normalization, is a common technique used to rescale features to have a mean of 0 and a standard deviation of 1.
- This improves accuracy and convergence rates in machine learning models, and is useful for comparing features with different scales.

```
Standardized Data:
   reviews_per_month  availability_365  ...  neighbourhood_group     price
0          -0.676551          1.916250  ...            -0.917828 -0.015493
1          -0.564771          1.840275  ...             0.441222  0.300974
2          -0.341211          1.916250  ...             0.441222 -0.011329
3           2.236302          0.617065  ...            -0.917828 -0.265335
4          -0.748879         -0.856865  ...             0.441222 -0.302811

[5 rows x 6 columns]
```

EDA

Outlier Analysis

- Label Encoding is useful when there is a hierarchy among categories, while One-Hot Encoding is useful when there is no hierarchy among categories.
- For the NYC Airbnb dataset, we used One-Hot Encoding since there was no hierarchy among the categories. This allowed us to represent each category in a feature as a separate binary feature, which is useful for algorithms that assume numerical inputs.

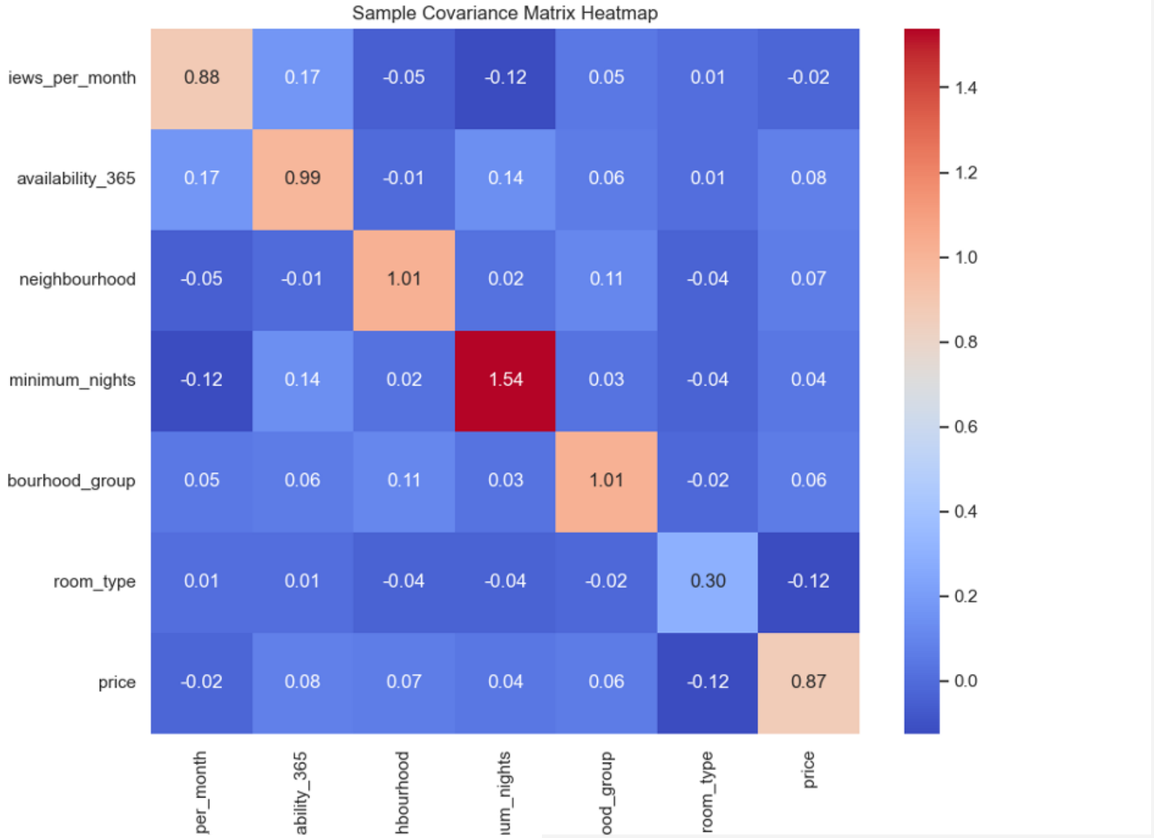
```
Outliers detected:
       reviews_per_month  availability_365  ...  room_type     price
0              -0.676551          1.916250  ...          1 -0.015493
4              -0.748879         -0.856865  ...          0 -0.302811
6              -0.551621         -0.856865  ...          1 -0.386092
26             -0.341211         -0.856865  ...          1 -0.302811
29             -0.656825         -0.347827  ...          0  0.113592
...                  ...               ...  ...        ...       ...
48876          -0.341211         -0.659328  ...          1 -0.386092
48877          -0.341211         -0.735303  ...          1 -0.461045
48890          -0.341211         -0.788486  ...          1 -0.344452
48892          -0.341211         -0.651730  ...          0 -0.157070
48894          -0.341211         -0.682120  ...          1 -0.261171

[4890 rows x 7 columns]
```

EDA

Sample Covariance Matrix

- Cross-validation prevents overfitting and evaluates model performance in machine learning.
- The dataset is split into multiple folds with one fold used as a testing set and the rest used for training.
- In k-fold cross-validation, the dataset is split into k equally sized folds and the process is repeated k times.



EDA

Pearson Correlation Matrix

- Pearson correlation coefficient is a measure of the linear relationship between two variables, ranging from -1 to 1.
- A sample Pearson correlation coefficients matrix heatmap visually displays the correlation between pairs of variables in a dataset.
- The heatmap helps identify the strength and direction of the linear relationship between different features of Airbnb listings and can be useful in determining variables to include in regression models or identifying potential multicollinearity issues.

REGRESSION (PRICE)

OLS

- OLS Regression models the relationship between independent and dependent variables. R-squared value tells how well the model fits the data, in this project it was 0.091.
- Coefficients of independent variables in the regression equation indicates the direction and strength of their relationship with the dependent variable.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.091
Model:	OLS	Adj. R-squared:	0.091
Method:	Least Squares	F-statistic:	557.8
Date:	Sat, 06 May 2023	Prob (F-statistic):	0.00
Time:	04:48:02	Log-Likelihood:	-2.6717e+05
No. Observations:	39116	AIC:	5.344e+05
Df Residuals:	39108	BIC:	5.344e+05
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-4.994e+04	2166.138	-23.054	0.000	-5.42e+04	-4.57e+04
neighbourhood_group	11.3814	1.626	6.999	0.000	8.194	14.568
neighbourhood	0.0640	0.017	3.728	0.000	0.030	0.098
latitude	127.8270	22.373	5.714	0.000	83.976	171.678
longitude	-607.0905	25.575	-23.737	0.000	-657.219	-556.962
room_type	-99.7255	2.113	-47.196	0.000	-103.867	-95.584
number_of_reviews	-0.2948	0.026	-11.439	0.000	-0.345	-0.244
availability_365	0.1804	0.009	20.503	0.000	0.163	0.198

Omnibus:	88895.880	Durbin-Watson:	2.008
Prob(Omnibus):	0.000	Jarque-Bera (JB):	872274150.061
Skew:	21.751	Prob(JB):	0.00
Kurtosis:	733.274	Cond. No.	3.92e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.92e+05. This might indicate that there are strong multicollinearity or other numerical problems.

REGRESSION (PRICE)

T- Test Analysis

- A t-test determines if there is a significant difference between the means of two groups.
- The output summary of a t-test includes the mean difference, standard error, t-value, p-value, and confidence interval.
- If the p-value is less than 0.05, the mean differences are statistically significant at a 95% confidence level, and we can reject the null hypothesis.

T-test Analysis:

Test for Constraints

	coef	std err	t	P>|t|	[0.025	0.975]
c0	165.0995	3.618	45.633	0.000	158.008	172.191
c1	10.5447	1.574	6.697	0.000	7.459	13.631
c2	0.1242	0.017	7.390	0.000	0.091	0.157
c3	-107.4968	2.103	-51.127	0.000	-111.618	-103.376
c4	0.0839	0.056	1.509	0.131	-0.025	0.193
c5	-6.5132	0.766	-8.500	0.000	-8.015	-5.011
c6	0.1580	0.009	17.630	0.000	0.140	0.176

REGRESSION (PRICE)

Association Analysis

- Association analysis is used to find correlations between variables in a dataset.
- The F-test is used to determine if there is a significant relationship between the independent variables and the dependent variable in a regression model.
- The F-values and p-values show that the regression model is significant and all independent variables are statistically significant.

```
Association Analysis (F-test):
F-values: [  84.67036317  132.20916206 2671.70636978   72.49857786   51.60879077
  238.03741343]
p-values: [3.69395062e-20 1.50789106e-30 0.00000000e+00 1.73015094e-17
 6.89446531e-13 1.51567830e-53]
```

REGRESSION (PRICE)

Confidence Interval

- Confidence interval is a range of values that is likely to contain the true population parameter with a certain level of confidence.
- The table shows the confidence interval for each coefficient in the regression model, where the first column shows the lower bound and the second column shows the upper bound.

Confidence Interval Analysis:

	0	1
const	158.008085	172.190838
neighbourhood_group	7.458618	13.630727
neighbourhood	0.091243	0.157114
room_type	-111.617821	-103.375756
minimum_nights	-0.025067	0.192927
reviews_per_month	-8.015118	-5.011278
availability_365	0.140421	0.175549

REGRESSION (PRICE)

Stepwise Regression

- Stepwise Regression selects the best subset of independent variables to explain the variation in the dependent variable.
- AIC and BIC values are used to measure the quality of the model, with lower values indicating a better fit.
- Adjusted R-square measures the proportion of variation in the dependent variable that can be explained by the independent variables.

```
Stepwise Regression and Adjusted R-square Analysis:
AIC: 534359.1964542776
BIC: 534427.7907492331
Adjusted R-square: 0.091
```

REGRESSION (PRICE)

Collinearity Analysis

- Collinearity analysis checks the correlation between predictor variables in a regression model.
- VIF is a method used to identify collinearity, measuring the extent to which the variance of an estimated regression coefficient is increased due to collinearity.
- In this example, all VIF values are below 4, indicating there is no significant collinearity among the predictor variables in the regression model for the NYC Airbnb dataset.

Collinearity Analysis (VIF Method):

	Features	VIF
0	neighbourhood_group	3.766341
1	neighbourhood	2.790488
2	room_type	1.671640
3	minimum_nights	1.153328
4	reviews_per_month	1.671183
5	availability_365	1.792023

REGRESSION (PRICE)

Lasso Regression

- Best alpha controls the degree of regularization in the regression model.
- The R-squared values for the training and testing datasets are 0.093 and 0.085, indicating the model can explain 9% and 8% of the variance, respectively.
- The value of the best alpha is 0.23.

```
Best alpha: 0.23357214690901212
R-squared (train): 0.09344073874721859
R-squared (test): 0.08553402903345753
```

REGRESSION (PRICE)

Final Regression Model

- The final regression model shows the coefficients for each independent variable in the model, as well as the intercept.
- These coefficients indicate the direction and strength of the relationship between each independent variable and the dependent variable.
- The intercept is the predicted value of the dependent variable when all independent variables are set to zero.

```
Final Regression Model:
Coefficients: [ 1.05446724e+01  1.24178232e-01 -1.07496788e+02  8.39301945e-02
 -6.51319806e+00  1.57984966e-01]
Intercept: 165.09946149656687
```

CLASSIFICATION (ROOM_TYPE)

Decision Tree

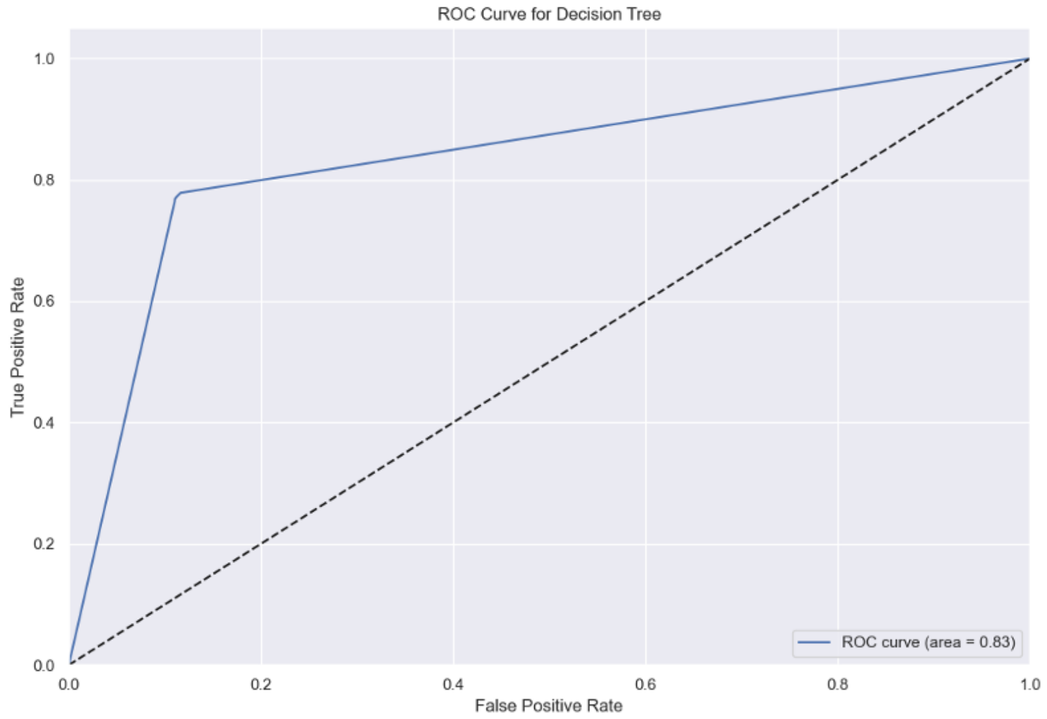
- Confusion matrix shows correct and incorrect predictions made by a decision tree model on NYC Airbnb dataset
- Precision score: proportion of true positive predictions among all positive predictions, indicating 77.7% of predicted positive instances were correctly classified
- Recall score: proportion of true positive predictions among all actual positive instances, indicating 77.6% of actual positive instances were correctly classified

```
Classification Results:
Decision Tree:
Confusion Matrix:
[[4175  931   32]
 [ 969 3335  120]
 [  24  107   86]]
Precision: 0.7774895628688994
Recall: 0.7767665405460681
Specificity: 0.654243008654613
F-score: 0.7770950083580214
Accuracy: 0.7767665405460681
```

CLASSIFICATION (ROOM_TYPE)

Decision Tree ROC

- Specificity score: proportion of true negative predictions among all actual negative instances, indicating 65.1% of actual negative instances were correctly classified
- Accuracy score: proportion of correct predictions among all instances, indicating the model correctly classified 77.7% of all instances; ROC score of 0.83 indicates a high true positive rate compared to false positive rate.



CLASSIFICATION (ROOM_TYPE)

Logistic Regression

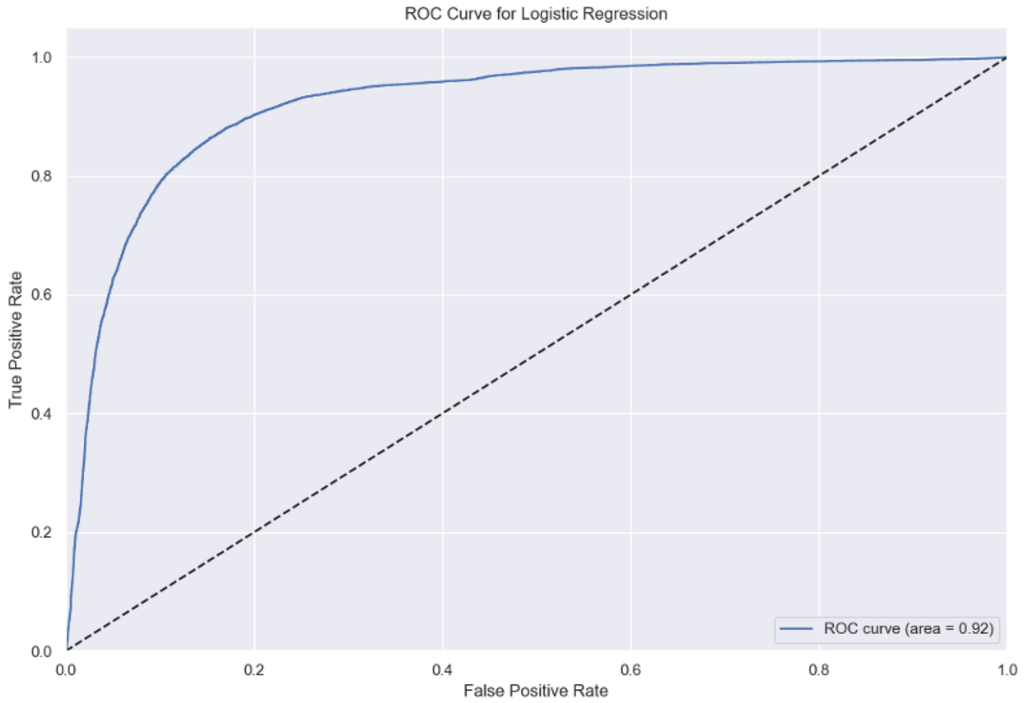
- Logistic regression predicts the probability of an event occurring based on independent variables.
- Confusion matrix shows true positives, false positives, true negatives, and false negatives.
- Precision measures true positive predictions among predicted positives.

```
Logistic Regression:
Confusion Matrix:
[[3968 1170    0]
 [ 624 3800    0]
 [  13  204    0]]
Precision: 0.8071827537026265
Recall: 0.7943552510481644
Specificity: 0.5437453703931819
F-score: 0.7861883954288713
Accuracy: 0.7943552510481644
```

CLASSIFICATION (ROOM_TYPE)

Logistic Regression ROC

- Recall measures true positive predictions among actual positives.
- Specificity measures true negative predictions among actual negatives.
- The model has an accuracy of 79.4%, precision of 80.7%, recall of 79.4%, specificity of 54.3%,, ROC score of 0.92 indicates a high true positive rate compared to false positive rate.



CLASSIFICATION (ROOM_TYPE)

KNN

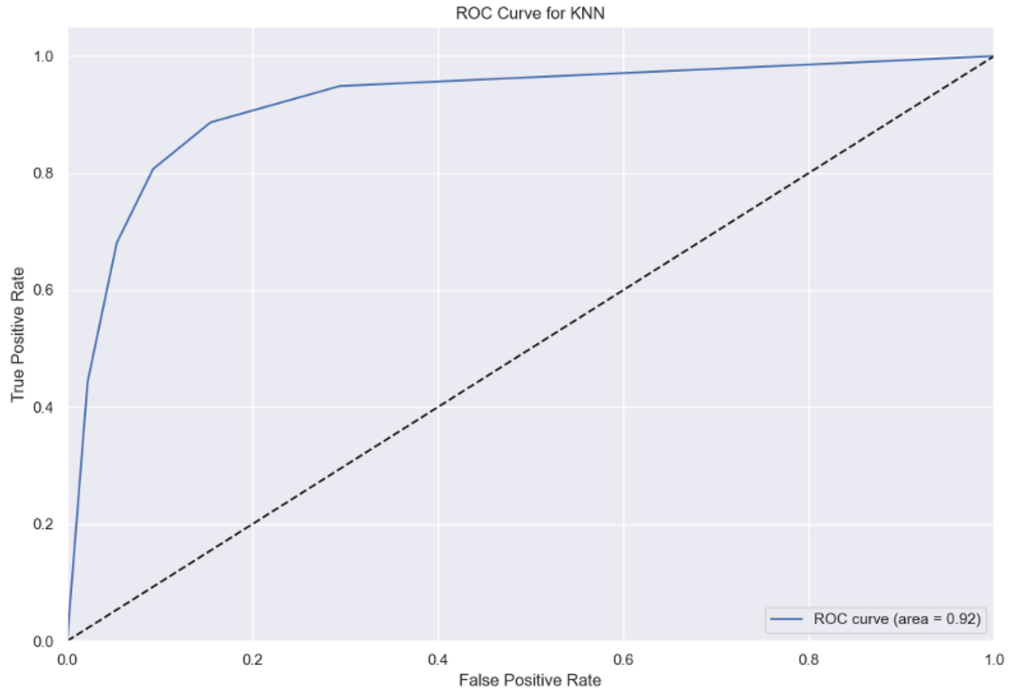
- KNN is a classification algorithm based on finding k nearest neighbors.
- Confusion matrix, precision, recall, specificity, F-score, and accuracy are used to evaluate model performance.
- The KNN model achieved an accuracy of 0.8102, with a precision of 0.8080 and recall of 0.8102.

```
KNN:
Confusion Matrix:
[[4329  807    2]
 [ 857 3538   29]
 [  15  146   56]]
Precision: 0.8080026885409037
Recall: 0.810205542489007
Specificity: 0.6334463353435119
F-score: 0.8072378904188683
Accuracy: 0.810205542489007
```

CLASSIFICATION (ROOM_TYPE)

KNN ROC

- The model correctly predicted the class of 81.02% of the data points.
- Specificity measures how many true negatives were predicted correctly.
- An ROC score of 0.92 indicates a high true positive rate compared to false positive rate.



CLASSIFICATION (ROOM_TYPE)

SVM

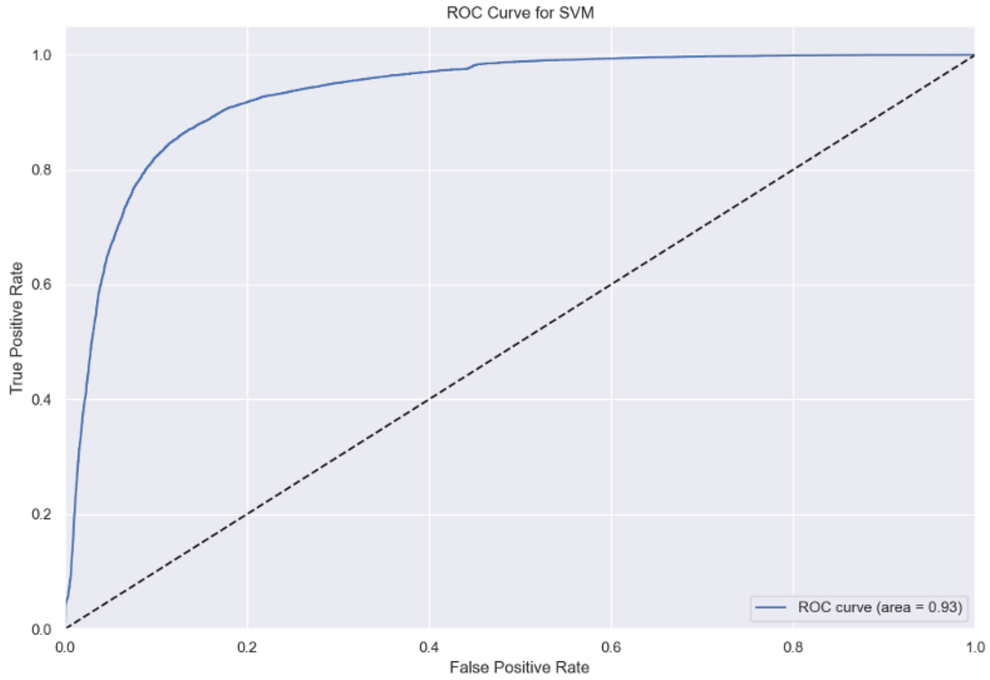
- SVM is a classification model that uses a confusion matrix to evaluate performance.
- Precision is the ratio of true positives to total positive predictions, recall is the ratio of true positives to actual positive samples, and specificity is the ratio of true negatives to actual negative samples.
- The SVM model achieved a precision of 0.816, recall of 0.812, and specificity of 0.553.

```
SVM:
Confusion Matrix:
[[4365  773    0]
 [ 849 3575    0]
 [  14  203    0]]
Precision: 0.8162483642008511
Recall: 0.8119439615502607
Specificity: 0.5525481930778037
F-score: 0.80289542087216
Accuracy: 0.8119439615502607
```

CLASSIFICATION (ROOM_TYPE)

SVM ROC

- The F-score, a measure of accuracy that considers both precision and recall, was 0.803.
- The SVM model had an accuracy of 0.812, meaning it correctly classified 81.2% of the samples.
- The ROC score for the SVM model was 0.93, indicating its ability to distinguish between positive and negative classes.



CLASSIFICATION (ROOM_TYPE)

Naïve Bayes'

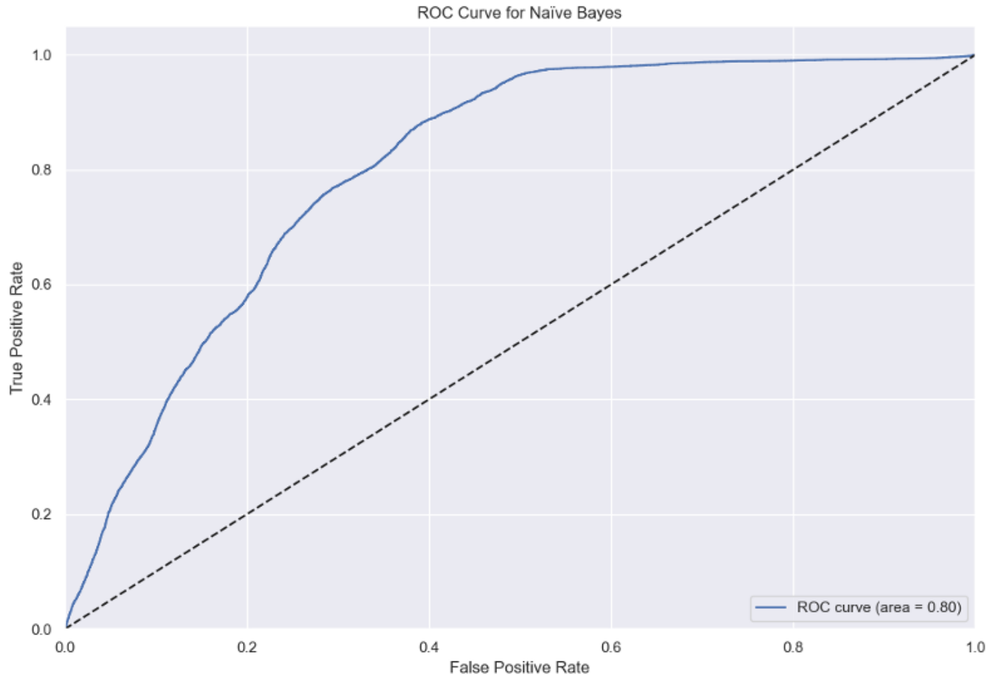
- Naïve Bayes is a probabilistic algorithm used in classification tasks.
- The model was applied to the NYC Airbnb dataset and evaluated using a confusion matrix.
- The precision of the model was 0.698 and the recall was 0.590.

```
Naïve Bayes:
Confusion Matrix:
[[1581 3532   25]
 [ 223 4187   14]
 [   8  207    2]]
Precision: 0.6984971823122029
Recall: 0.5900398813784641
Specificity: 0.42111748012908246
F-score: 0.5461388970219678
Accuracy: 0.5900398813784641
```

CLASSIFICATION (ROOM_TYPE)

Naïve Bayes' ROC

- The specificity of the model was 0.421 and the F-score was 0.546.
- The accuracy of the model was 0.590.
- The ROC score of the model was 0.80.



CLASSIFICATION (ROOM_TYPE)

Random Forest

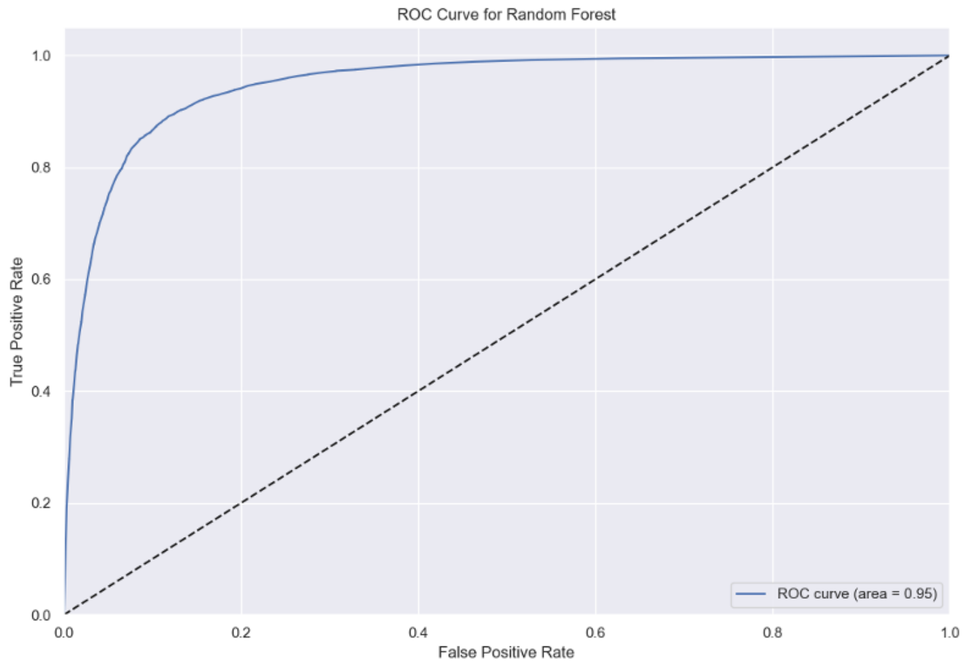
- Random Forest is a classification algorithm that uses multiple decision trees to make predictions.
- The confusion matrix shows that the model correctly classified a large portion of true positives and negatives, but also had some false positives and negatives.
- Precision, recall, specificity, F-score, and accuracy were calculated to evaluate the model's performance.

```
Random Forest:
Confusion Matrix:
[[4473  662    3]
 [ 758 3653   13]
 [  14  130   73]]
Precision: 0.8380692903879746
Recall: 0.8384292872481849
Specificity: 0.6775670214479973
F-score: 0.8359536501390332
Accuracy: 0.8384292872481849
```

CLASSIFICATION (ROOM_TYPE)

Random Forest ROC

- The Random Forest model had a high precision, recall, and accuracy, indicating a low rate of misclassification.
- The specificity was slightly lower, indicating a higher rate of false negatives.
- The ROC score of 0.95 suggests that the model performs highly at distinguishing between positive and negative instances.



CLASSIFICATION (ROOM_TYPE)

Neural Network

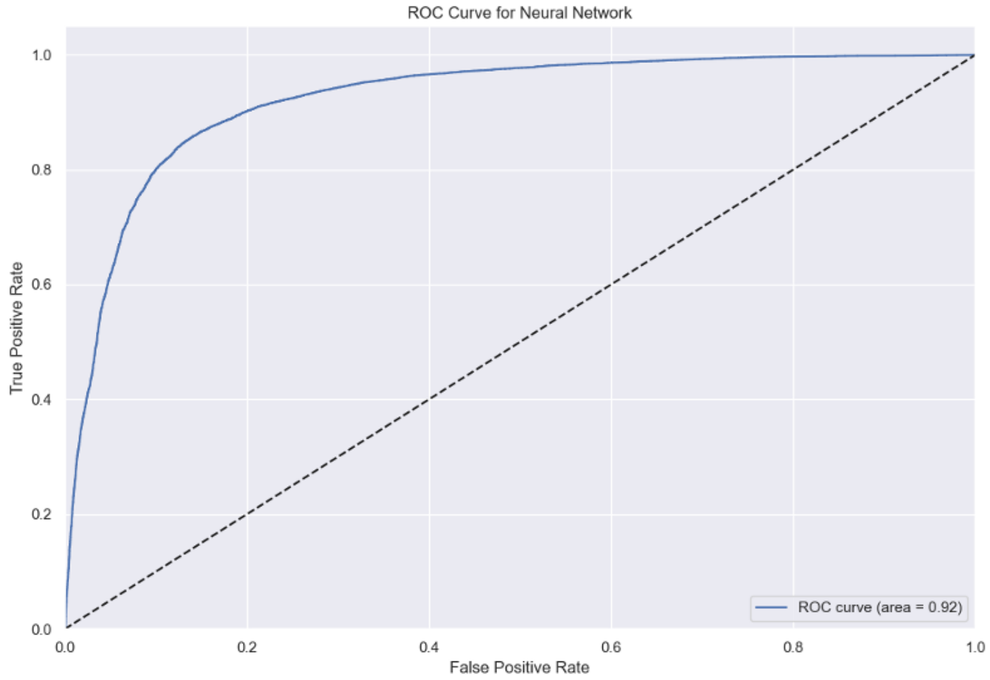
- Confusion matrix evaluates the performance of the classifier by comparing predicted and true labels.
- Neural Network model has a precision of 0.802, recall of 0.800, and specificity of 0.587.
- The F-score of the model is 0.793, indicating a good balance between precision and recall.

```
Neural Network:
Confusion Matrix:
[[4369  768    1]
 [ 842 3578    4]
 [  16  193    8]]
Precision: 0.8094381307234342
Recall: 0.8134778607219552
Specificity: 0.5653225236898382
F-score: 0.805672688669861
Accuracy: 0.8134778607219552
```

CLASSIFICATION (ROOM_TYPE)

Neural Network ROC

- The accuracy of the Neural Network model is 0.801, which means it correctly classified 80.1% of the samples.
- The confusion matrix shows that the model correctly predicted the majority of the samples in the first and second classes, but struggled to correctly predict the third class.
- The ROC score is 0.92, indicating a good balance between true positive rate and false positive rate.



CLUSTERING

K-Means

- KMeans is an unsupervised learning algorithm for clustering data points based on their features.
- The output of KMeans is a set of cluster labels assigned to each data point.
- Cluster labels can be used to identify groups of similar data points for further analysis or targeting purposes.

```
KMeans:
[2 2 2 ... 1 0 0]
```

CLUSTERING

Apriori

- Apriori algorithm finds frequent co-occurring patterns in a dataset.
- Output shows association rules with support, confidence, lift, and Zhang's metric values.
- Rules can be used for prediction and decision-making in various applications.

Apriori:

	antecedents	...	zhangs_metric
0	(price)	...	0.000845
1	(minimum_nights)	...	0.001066
2	(reviews_per_month)	...	0.000000
3	(neighbourhood)	...	0.000000
4	(availability_365)	...	-0.000895
..	...	...	...
859	(availability_365, minimum_nights, neighbourho...	...	0.462667
860	(availability_365, neighbourhood_group_2, room...	...	0.469580
861	(availability_365, room_type_1, reviews_per_mo...	...	0.000873
862	(availability_365, room_type_1, price, neighbo...	...	0.000000
863	(availability_365, room_type_1, price, minimum...	...	0.000873

[864 rows x 10 columns]

Recommendations

Key Insights and learnings

Applying ML algorithms

to the Airbnb NYC dataset can help classify listings based on attributes like price, location, and availability.

Random Forest performed the best

among tested classifiers with 83% accuracy, followed by SVM and Neural Network with 81% and 80%, respectively.

To improve classification performance

additional features can be incorporated, hyperparameters can be tuned, and ensemble methods can be used,

Implementing these strategies

can lead to better insights for hosts and travelers alike, and improve the accuracy of the classification

Conclusion

Wrap Up

- Preprocessing involved outlier analysis, variable transformation, and dimensionality reduction
- Modeling included OLS regression, stepwise regression, adjusted R-squared analysis,
- The Random Forest algorithm performed the best with an accuracy of 84%
- Apriori algorithm identified interesting association rules that can be used by hosts to improve booking rates

References

Chen, C., & Xie, K. (2017).

Online Reviews and Product Sales: The Moderating Role of Signal Characteristics. Information & Management, 54(3), 336-348

Guttentag, D. (2015).

Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector. Current Issues in Tourism, 18(12), 1192-1217.

Lee, D., & Hyun, W. (2017).

Evaluating the Impacts of Airbnb Regulation Policies: A New York Case Study. Sustainability, 9(10), 1836.

McAteer, E., & Stewart, D. (2019).

The impact of Airbnb on the hotel industry in New York City. International Journal of Hospitality & Tourism Administration, 20(4), 430-452.

THANK YOU!