Enhancing Hotel Reservation Show-Up Prediction

Group 7

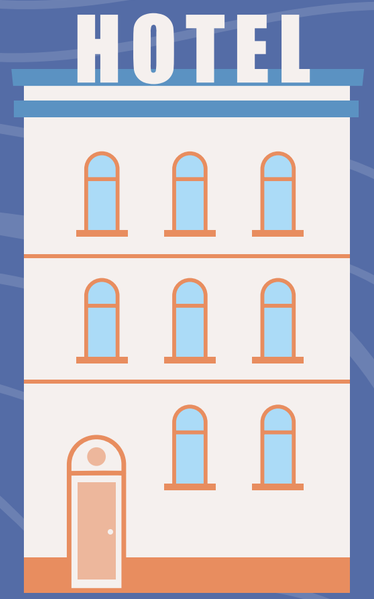
SC1015
Mini-Project

WHO ARE WE ?

CHENG LIN

ISAAC

JUN SHENG



WE ARE A GROUP OF DATA SCIENTISTS!

WHAT WE DO :

- ANALYSE VARIETY OF FACTORS
- DEVELOP PREDICTIVE MODELS
- FORECAST FUTURE DEMANDS

TABLE OF CONTENT

PRACTICAL MOTIVATION

REVENUE OPTIMISATION

Maximises profits and promotes long term growth

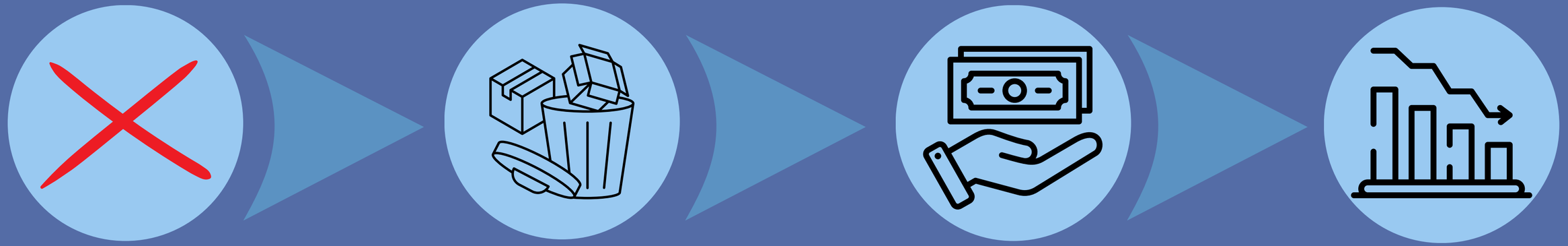
RESOURCE MANAGEMENT

Reducing wastage of resources, ensuring efficient resource allocation

CUSTOMER EXPERIENCE

Improving customer satisfaction and brand loyalty

PROBLEM FORMULATION



NO-SHOW

RESOURCE WASTAGE

OPPORTUNITY COST

LOWER PROFITS

HOW COULD WE PREDICT THE NUMBER OF NO-SHOWS IN ORDER TO MAINTAIN AND INCREASE PROFITS?

SAMPLE COLLECTION

```
RangeIndex: 119391 entries, 0 to 119390
Data columns (total 15 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   booking_id      119391 non-null  int64
 1   no_show         119390 non-null  float64
 2   branch          119390 non-null  object
 3   booking_month   119390 non-null  object
 4   arrival_month   119390 non-null  object
 5   arrival_day     119390 non-null  float64
 6   checkout_month  119390 non-null  object
 7   checkout_day    119390 non-null  float64
 8   country         119390 non-null  object
 9   first_time      119390 non-null  object
 10  room            97778 non-null   object
 11  price           94509 non-null   object
 12  platform        119390 non-null  object
 13  num_adults      119390 non-null  object
 14  num_children    119390 non-null  float64
```

	booking_id	no_show	branch	booking_month	arrival_month	arrival_day	checkout_month	checkout_day	country	first_time	room	price	platform	num_adults	num_children
0	94113	0.0	Changi	November	June	25.0	June	27.0	Singapore	Yes	Single	SGD$ 492.98	Website	1	0.0
1	86543	0.0	Orchard	August	November	28.0	November	29.0	Indonesia	Yes	King	SGD$ 1351.22	Website	2	0.0
2	75928	0.0	Changi	March	February	7.0	February	11.0	India	Yes	Single	None	Agent	1	0.0
3	66947	1.0	Orchard	September	October	1.0	October	3.0	China	Yes	Single	SGD$ 666.04	Website	1	0.0
4	106390	0.0	Orchard	March	June	20.0	June	24.0	Australia	Yes	Queen	USD$ 665.37	Website	1	0.0

DATA PREPARATION

```
# Drop rows with missing values
dataCleaned = data.dropna()
dataCleaned.info()
```

```
Index: 72897 entries, 0 to 119390
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   booking_id      72897 non-null  int64  
 1   no_show         72897 non-null  float64
 2   branch          72897 non-null  object 
 3   booking_month   72897 non-null  object 
 4   arrival_month   72897 non-null  object 
 5   arrival_day     72897 non-null  float64
 6   checkout_month  72897 non-null  object 
 7   checkout_day    72897 non-null  float64
 8   country         72897 non-null  object 
 9   first_time      72897 non-null  object 
 10  room            72897 non-null  object 
 11  price           72897 non-null  object 
 12  platform        72897 non-null  object 
 13  num_adults      72897 non-null  object 
 14  num_children    72897 non-null  float64
```

```
# Display unique values of 'price'
uniqueValues = dataCleaned['price'].unique()
print(uniqueValues)
```

```
['SGD$ 492.98' 'SGD$ 1351.22' 'SGD$ 666.04' ... 'USD$ 331.93'
 'USD$ 612.18' 'USD$ 1041.29']
```

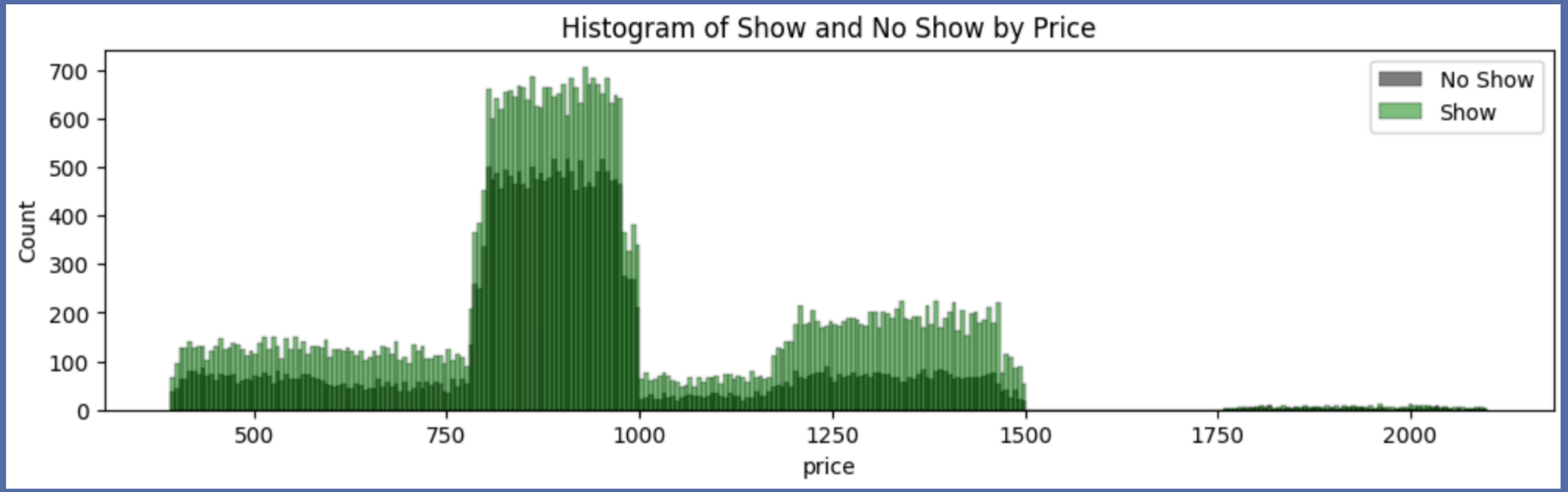
DATA PREPARATION

Errors	Variable	Remedy
Months not standardised	arrival_month	Convert all the months to lowercase and capitalised the first letter
Negative days	checkout_day	Remove rows with negative numbers
Price not in SGD	price	Convert USD to SGD and to floating point numbers
Numbers in string	num_adults	Map each string number to its respective integer number

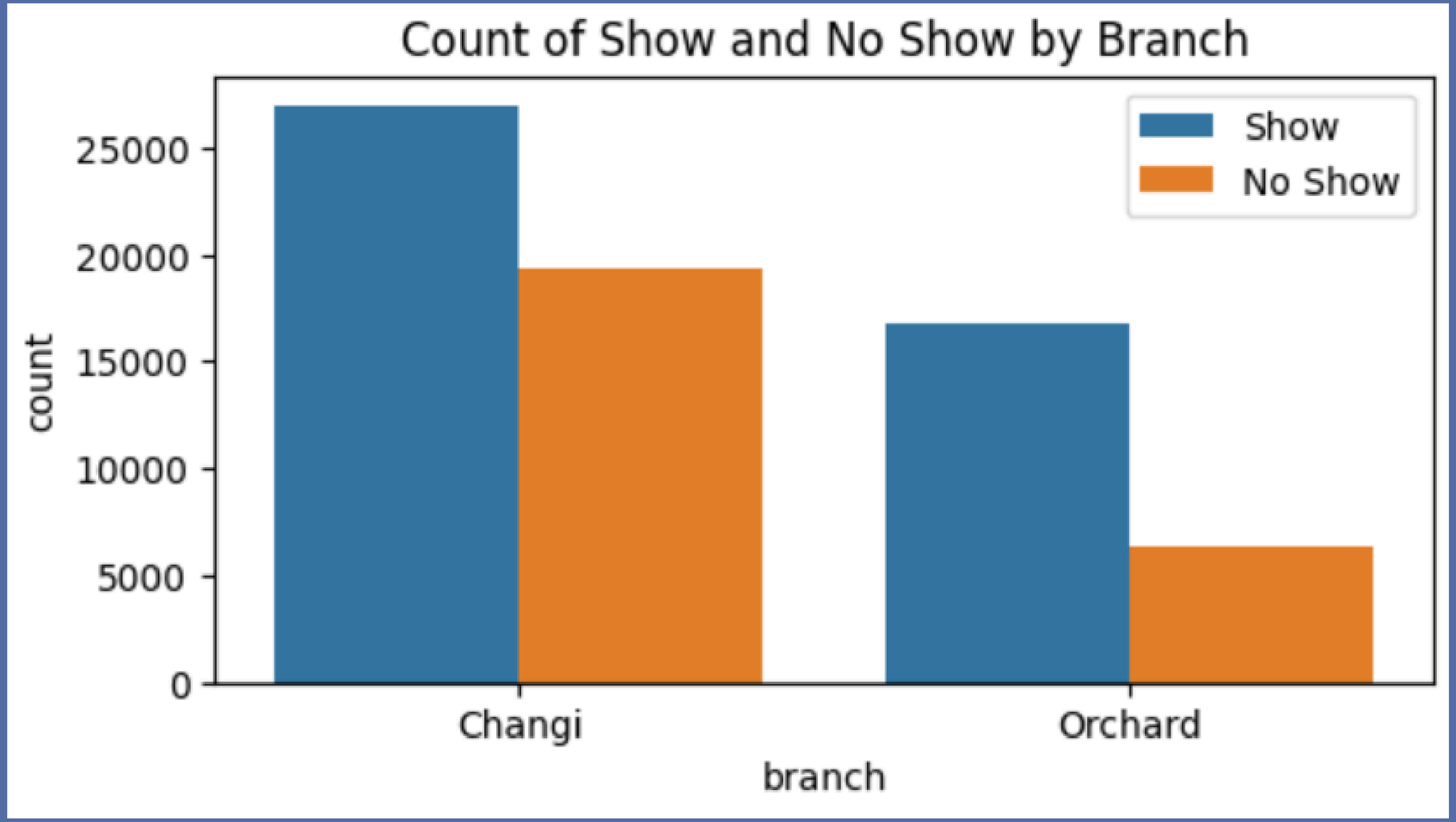
EXPLORATORY DATA ANALYSIS

	booking_id	no_show	branch	booking_month	arrival_month	arrival_day	checkout_month	checkout_day	country	first_time	room	price	platform	num_adults	num_children
0	94113	0.0	Changi	November	June	25.0	June	27.0	Singapore	Yes	Single	SGD$ 492.98	Website	1	0.0
1	86543	0.0	Orchard	August	November	28.0	November	29.0	Indonesia	Yes	King	SGD$ 1351.22	Website	2	0.0
2	75928	0.0	Changi	March	February	7.0	February	11.0	India	Yes	Single	None	Agent	1	0.0
3	66947	1.0	Orchard	September	October	1.0	October	3.0	China	Yes	Single	SGD$ 666.04	Website	1	0.0
4	106390	0.0	Orchard	March	June	20.0	June	24.0	Australia	Yes	Queen	USD$ 665.37	Website	1	0.0

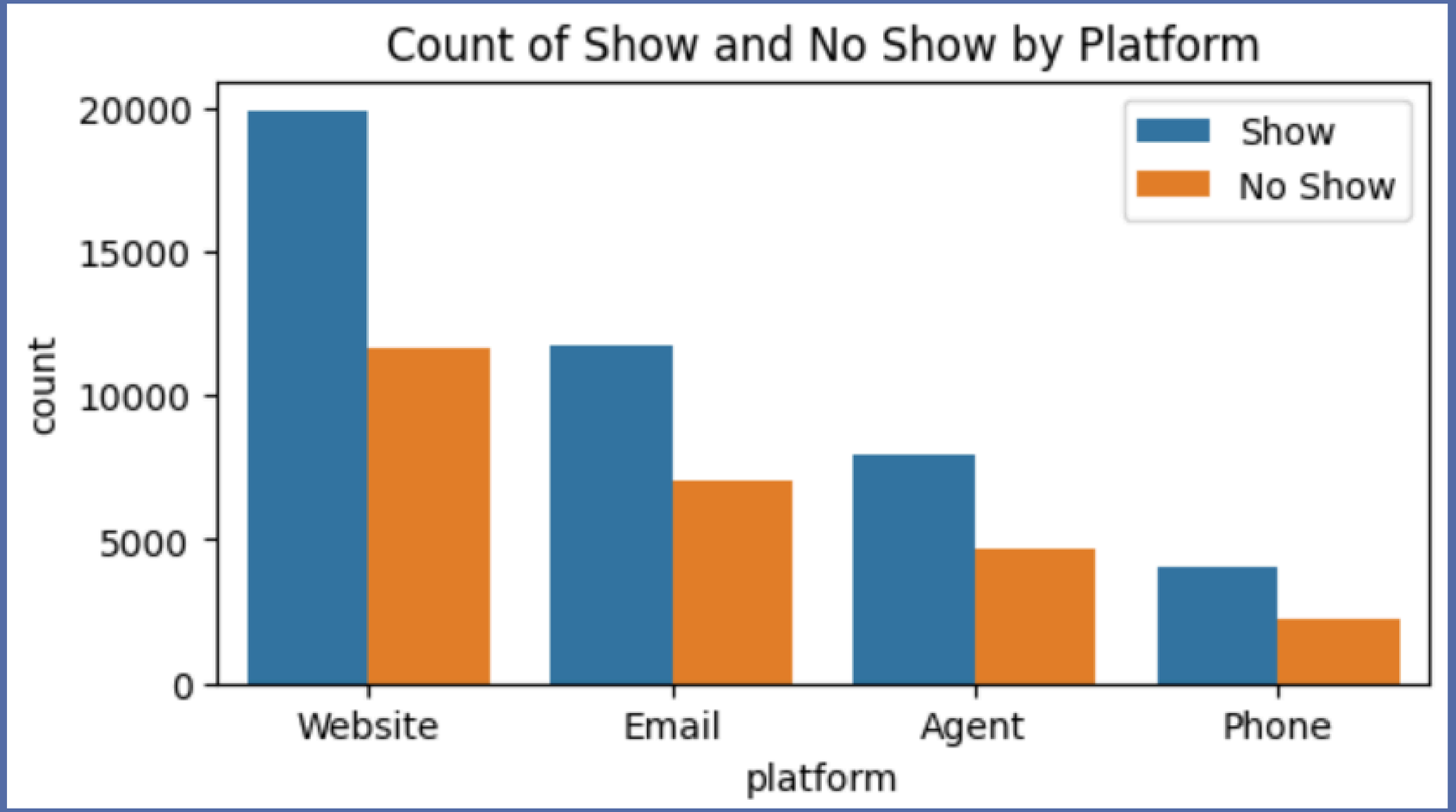
EXPLORATORY DATA ANALYSIS



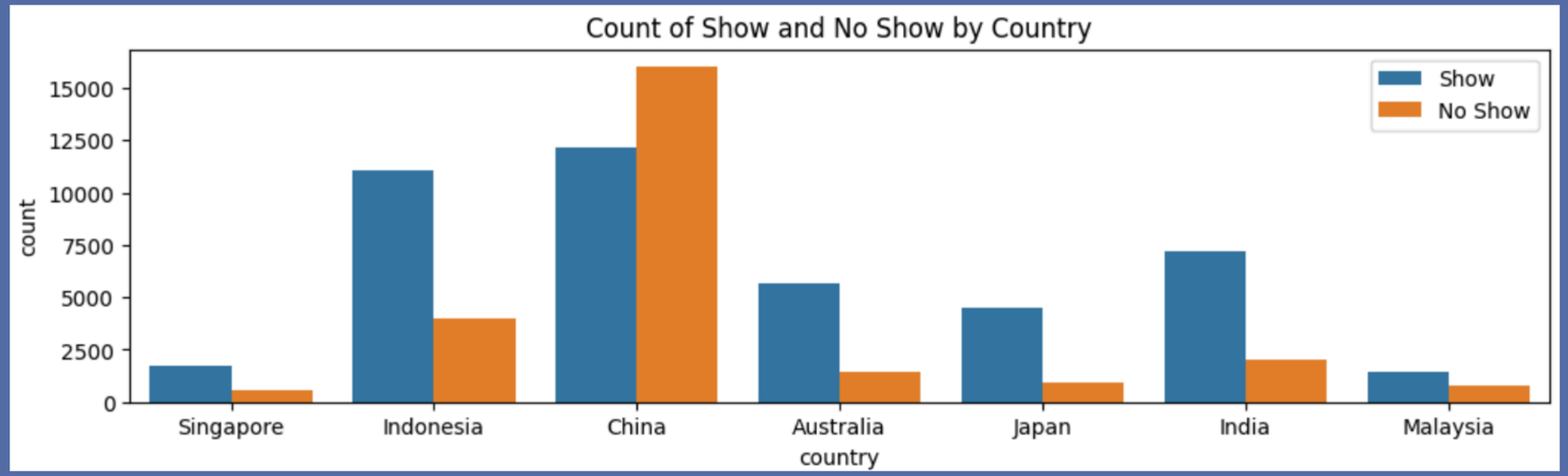
EXPLORATORY DATA ANALYSIS

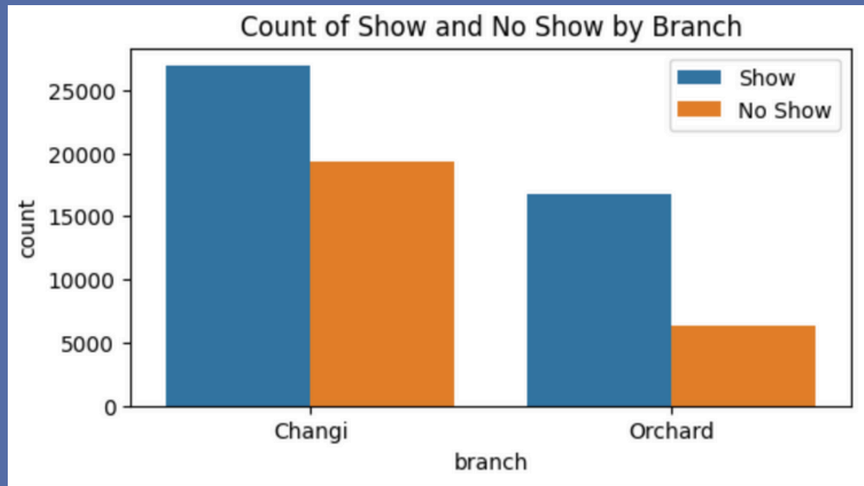


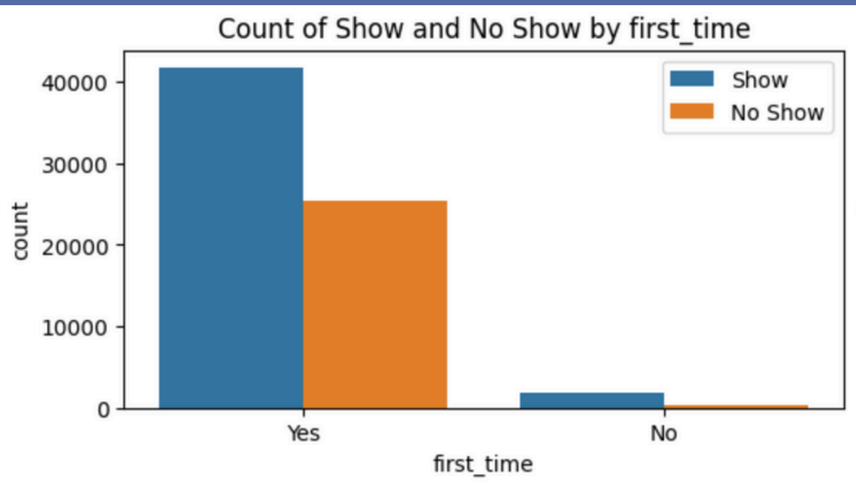
EXPLORATORY DATA ANALYSIS

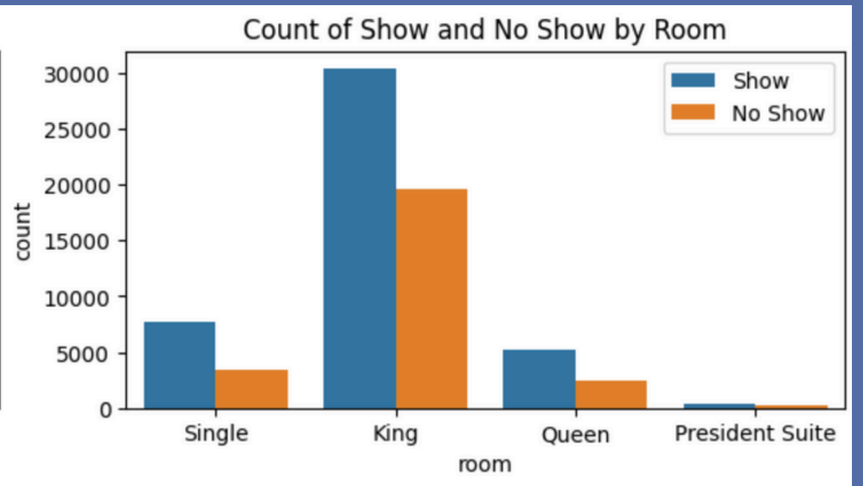


EXPLORATORY DATA ANALYSIS

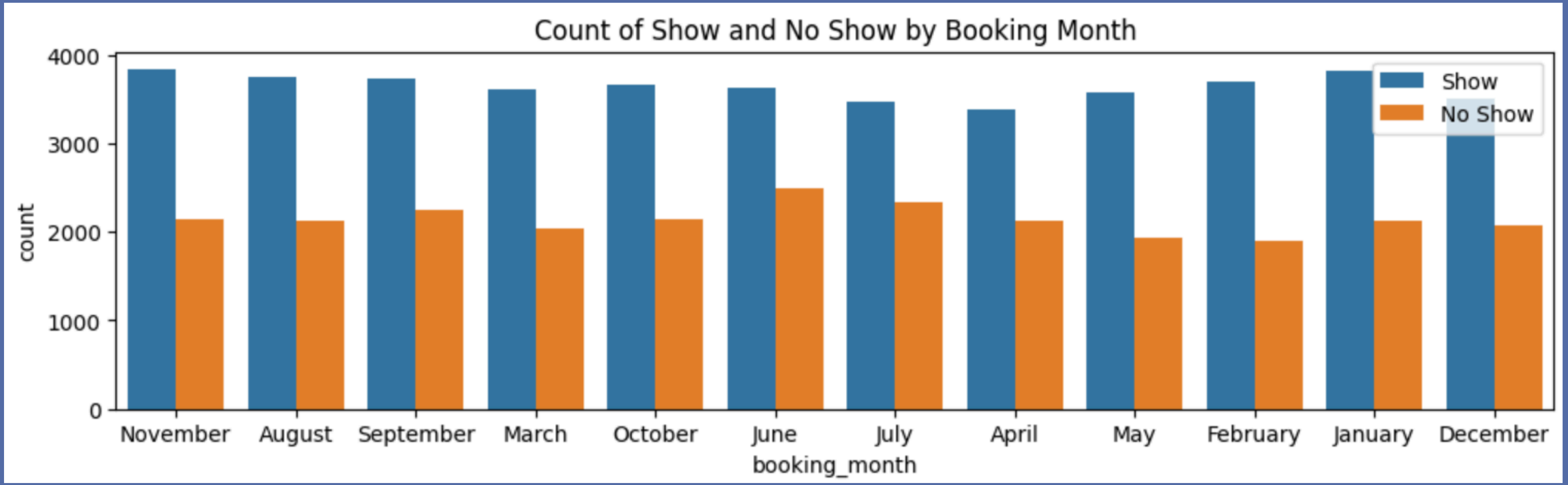








EXPLORATORY DATA ANALYSIS



CHI-SQUARE STATISTIC

CHI-SQUARE TEST IS A STATISTICAL METHOD USED TO DETERMINE IF THERE IS A SIGNIFICANT ASSOCIATION BETWEEN CATEGORICAL VARIABLES IN A DATASET

WHAT WE USED

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

$\chi 2$ = Chi-square statistic
O = Observed frequency
E = Expected frequency under null hypothesis (no association between variables)

OUR RESULTS

Chi-Square Test Results:

	Variable	Chi2
0	branch	1316.238151
1	booking_month	108.456938
2	arrival_month	318.578296
3	arrival_day	199.172407
4	checkout_month	301.388291
5	checkout_day	313.910383
6	country	8412.274998
7	first_time	491.668926
8	room	362.067572
9	platform	6.055378
10	num_adults	20.900278
11	num_children	42.410799

CRAMER'S V

ASSESS THE STRENGTH OF ASSOCIATION BETWEEN EACH CATEGORICAL PREDICTOR VARIABLE AND THE BINARY OUTCOME VARIABLE (NO-SHOW OR NOT).

WHAT WE USED

$$\text{Cramer's V} = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, c-1)}}$$

V = Cramer's V coefficient
$\chi 2$ = Chi-square statistic
n = Total number of observations
r = Number of rows in the contingency table
c = Number of columns in the contingency table

OUR RESULTS

Cramér's V Test Results:

	Variable	Cramers_V
0	branch	0.137708
1	booking_month	0.037486
2	arrival_month	0.066594
3	arrival_day	0.049388
4	checkout_month	0.064706
5	checkout_day	0.063981
6	country	0.348145
7	first_time	0.084111
8	room	0.071953
9	platform	0.006637
10	num_adults	0.016939
11	num_children	0.023838

VARIABLES WITH STRONG CORRELATIONS

COUNTRY

BRANCH

FIRST_TIME

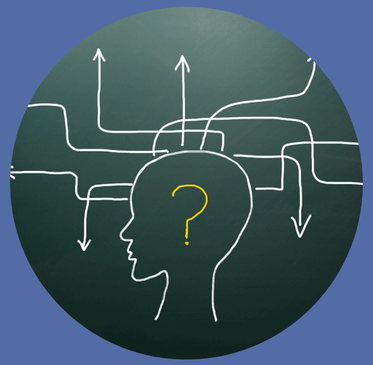
ROOM_TYPE

ARRIVAL_MONTH

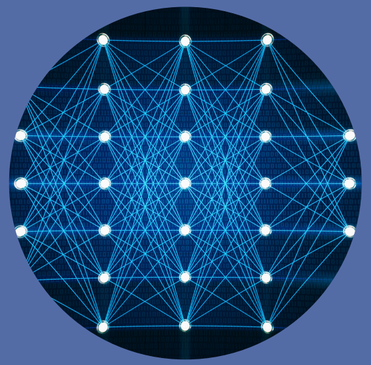
MACHINE LEARNING MODELS

Binary Logistic Regression

Random Forest



Decision Tree



Recurrent Neural Network

Convolutional Neural Network and LSTM Hybrid

METRICS USED TO EVALUATE THE MODELS

ACCURACY

ACCURACY =
TRUE POSITIVES + TRUE NEGATIVE
TOTAL PREDICTIONS

PRECISION

PRECISION =
TRUE POSITIVES
TRUE POSITIVES + FALSE POSITIVES)

RECALL

RECALL =
TRUE POSITIVE (TP)
TRUE POSITIVE (TP) + FALSE NEGATIVE (FN)

F1 SCORE

- SINGLE METRIC THAT BALANCES BOTH PRECISION AND RECALL
- HIGH F1 SCORES INDICATE BOTH HIGH PRECISION AND HIGH RECALL.

ROC-AUC

- REPRESENTS MODEL'S ABILITY TO DISCRIMINATE BETWEEN THE POSITIVE AND NEGATIVE CLASSES

BINARY LOGISTIC REGRESSION

STATISTICAL METHOD USED FOR BINARY CLASSIFICATION TASKS

WHAT WE DID

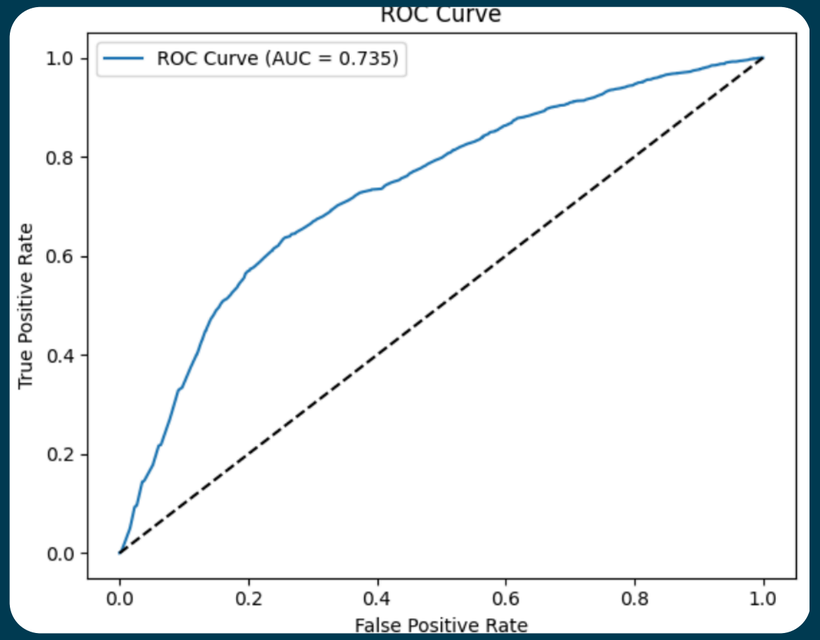
DATA PREPARATION :

- **One-hot encoding** used for categorical features.
- Split the dataset into features (input variables) and the target variable (no-show indicator)

MODEL TRAINING :

- Specify parameters such as **regularization strength (C)** to control model complexity and prevent overfitting

OUR RESULTS



ROC-AUC : 0.735 **F1 SCORE: 0.57**

ACCURACY: 0.72

DECISION TREES

BUILDING A TREE-LIKE MODEL THAT MAKES DECISIONS BASED ON THE VALUES OF INPUT FEATURES.

WHAT WE DID

DATA PREPARATION :

- **Label Encoder** used for categorical features.

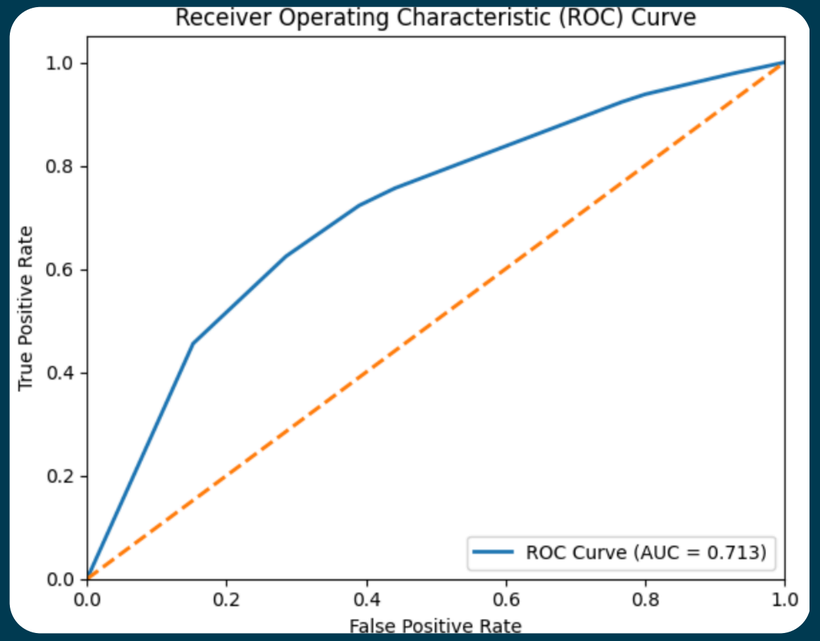
HYPERPARAMETER TUNING :

- **Grid search cross-validation** (GridSearchCV) to search for the best combination of hyperparameters

FEATURE IMPORTANCE :

- visualize the importance of **predictor variables**

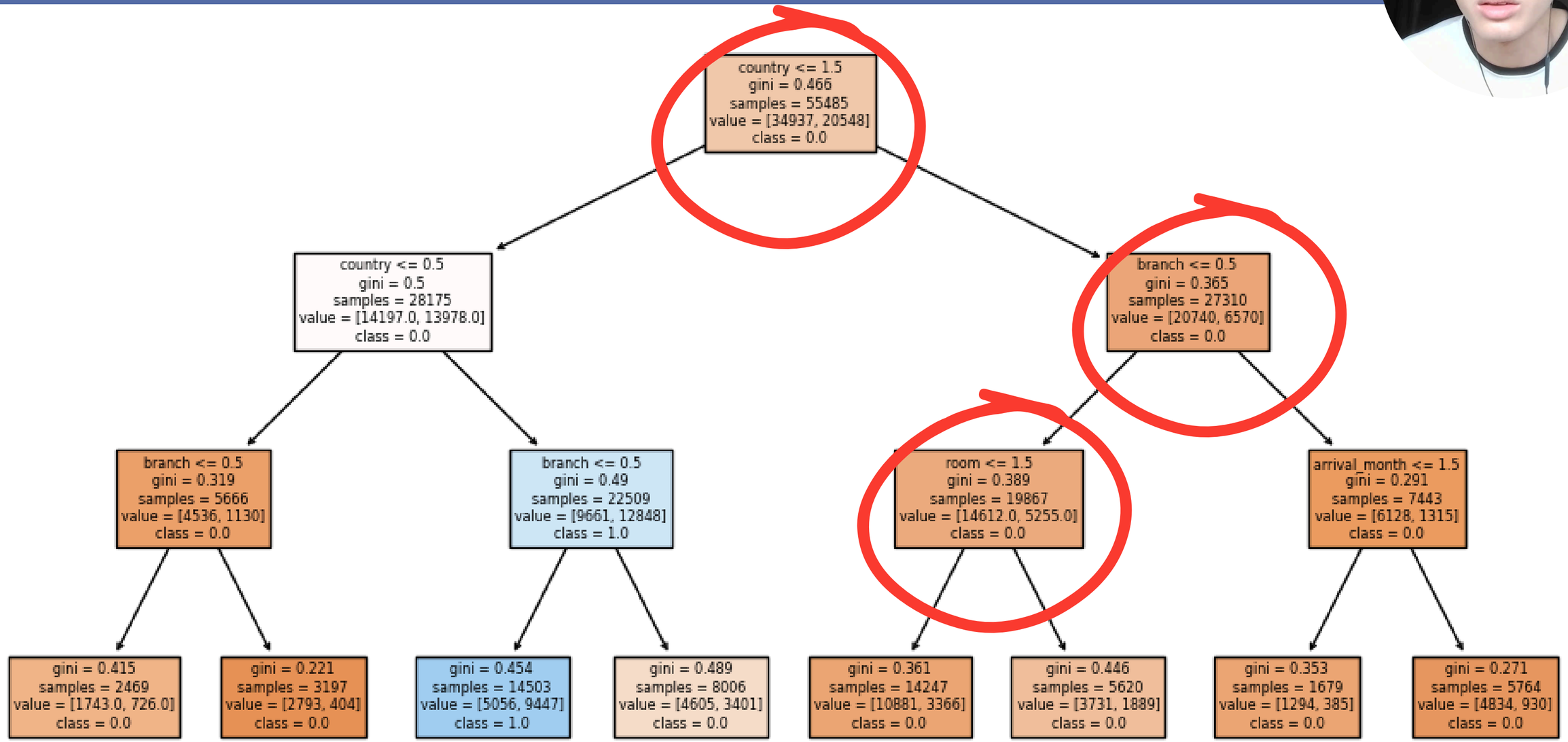
OUR RESULTS



ROC-AUC : 0.713 F1 SCORE: 0.53

ACCURACY: 0.70

DECISION TREES



RANDOM FOREST

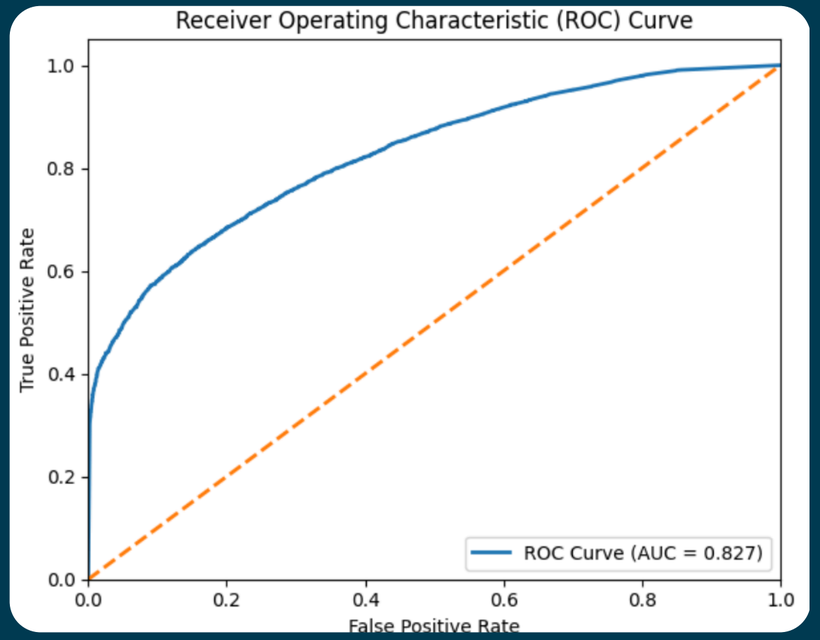
ENSEMBLE LEARNING METHOD THAT BUILDS MULTIPLE DECISION TREES DURING TRAINING TO IMPROVE ACCURACY AND GENERALIZATION OF PREDICTIONS

WHAT WE DID

CROSS VALIDATION:

- Split the available data into multiple subsets, called **"folds."**
- Process allows for more reliable estimates of the model's performance compared to a single train-test split

OUR RESULTS

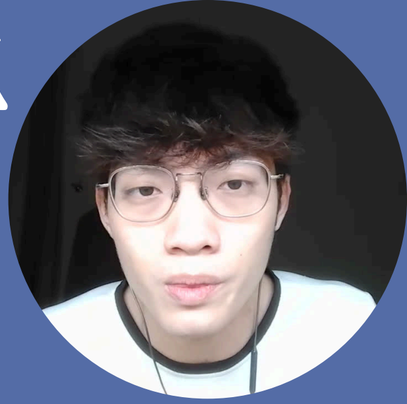


ROC-AUC : 0.827 F1 SCORE: 0.67

ACCURACY: 0.77

RECURRENT NEURAL NETWORK

LEVERAGING THE SEQUENTIAL NATURE OF RESERVATION DATA TO CAPTURE TEMPORAL DEPENDENCIES AND MAKE PREDICTIONS



WHAT WE DID

BINARY CROSS-ENTROPY :

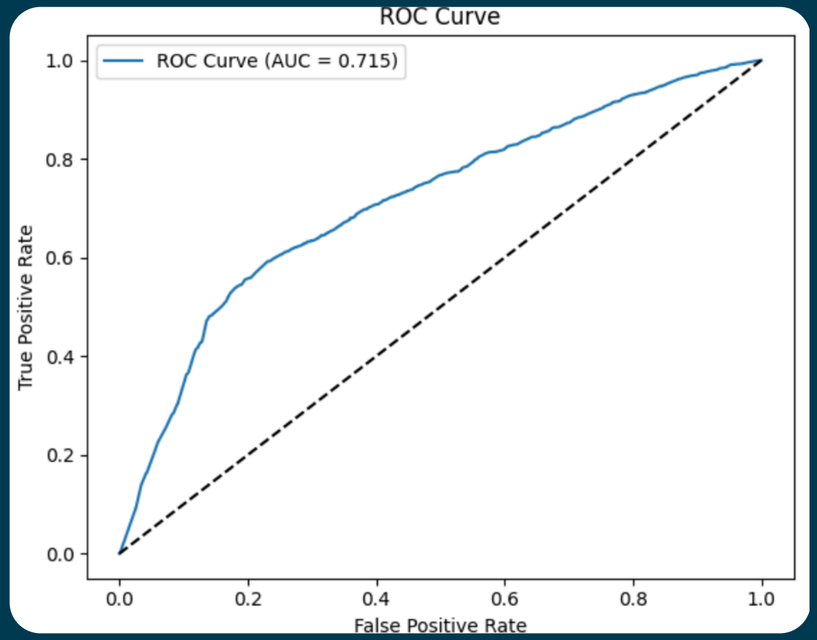
$$Probability(x) = \mu^{x}(1-\mu)^{1-x}$$

$$L = \prod_{i=1}^{N} p(x_i) = \prod_{i=1}^{N} \mu^{x_i}(1-\mu)^{1-x_i}$$

$$LossFunction = -\sum_{i=1}^{N}[y_i * log(y_{pred}) + (1-y_i) * log(1-y_{pred})]$$

$$BinaryCrossEntropy = -\frac{1}{N}\sum_{i=1}^{N}[y_i * log(y_{pred}) + (1-y_i) * log(1-y_{pred})]$$

OUR RESULTS



ROC-AUC : 0.715 **F1 SCORE: 0.53**

ACCURACY: 0.72

RECURRENT NEURAL NETWORK

BINARY CROSS-ENTROPY :

PROBABILITY MASS FUNCTION (PMF)

$$Probability(x) = \mu^x (1-\mu)^{1-x}$$

LIKELIHOOD (PRODUCT OF ALL X IN PMF)!

$$L = \prod_{i=1}^{N} p(x_i) = \prod_{i=1}^{N} \mu^{x_i} (1-\mu)^{1-x_i}$$

LOSS FUNCTION : EVALUATES HOW WELL ALGORITHM MODELS YOUR DATASET

$$LossFunction = -\sum_{i=1}^{N} [y_i * log(y_{pred}) + (1-y_i) * log(1-y_{pred})]$$

$$BinaryCrossEntropy = -\frac{1}{N} \sum_{i=1}^{N} [y_i * log(y_{pred}) + (1-y_i) * log(1-y_{pred})]$$

CONVOLUTION NEURAL NETWORK

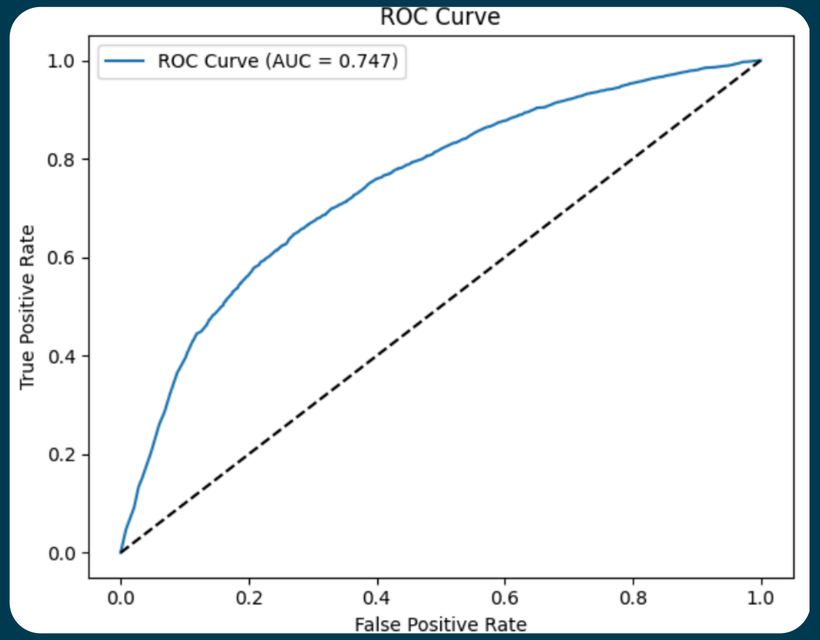
EXTRACTING MEANINGFUL FEATURES (E.G., RESERVATION INFORMATION) USING CONVOLUTIONAL LAYERS AND LEVERAGING THESE FEATURES TO MAKE PREDICTIONS

WHAT WE DID

COMBINATION OF CNN AND LSTM :

- Merging output representations of two types of layers to create a unified representation
- Captures both sequential and spatial information

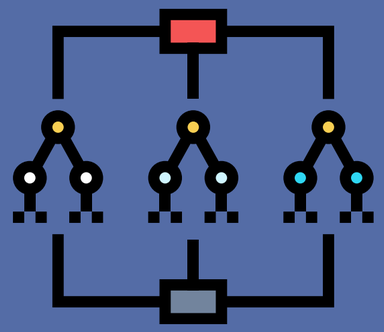
OUR RESULTS



ROC-AUC : 0.747 F1 SCORE: 0.57

ACCURACY: 0.70

INTELLIGENT DECISION



RANDOM FOREST MODEL

- HANDLING CATEGORICAL VARIABLES
- NON-LINEARITY
- ENSEMBLE LEARNING
- ROBUSTNESS TO OVERFITTING

HIGHEST ACCURACY OF 77%

THANK YOU