WIE 3007 DATA MINING & WAREHOUSING

Leveraging Data Mining for Enhanced Hotel Booking Strategies

Prepared by

- Kalkazbek Alikhan S2043519
- Cher Jia Wen U2005415/1
- Thayananth Kumaresan U2105590
- Ahmed Ishag Mohamed S2124564
- Visalini a/p Vijayan U2102811

Lecturer: Prof Dr Teh Ying Wah

Table of Contents

1 INTRODUCTION

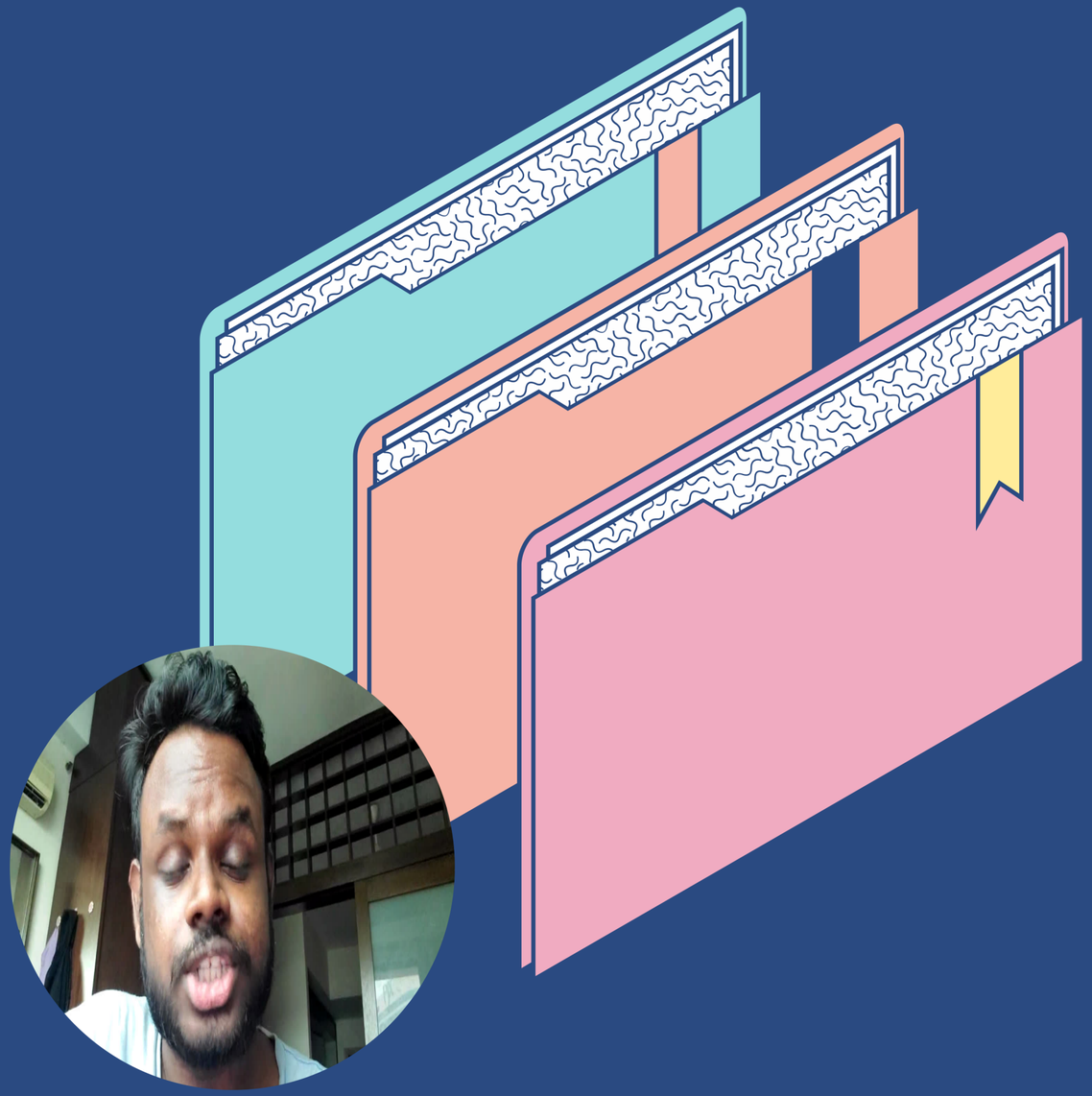
2 DATASET SELECTION

3 DATA UNDERSTANDING

4 SAS SEMMA METHODOLOGY

5 CONCLUSION

6 REFERENCES



1 Introduction

Background

OVERVIEW - HOTEL BOOKINGS

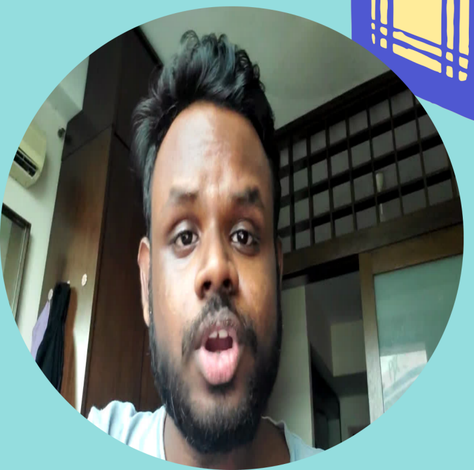
- Central to the hospitality industry.
- Complex interplay of booking patterns and customer preferences.
- Dataset includes diverse variables: demographics, booking details.

Objectives

- To uncover recurring patterns in booking behavior for adapting to seasonal trends and optimizing resource allocation.

- To identify factors influencing cancellations by analyzing the dataset for development of strategies to mitigate cancellations and enhance revenue management.

- To develop predictive models to forecast the likelihood of booking cancellations so that stakeholders can implement preventive measures and improve overall booking success rates.



2
Dataset Selection

Dataset

DATASET SOURCE & INFORMATION

kaggle

Hotel Reservations Dataset

Can you predict if customer is going to cancel the reservation ?

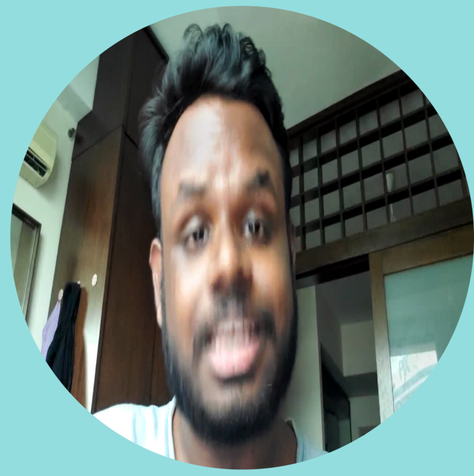
Data Card | Code (169) | Discussion (19)

About Dataset

Context

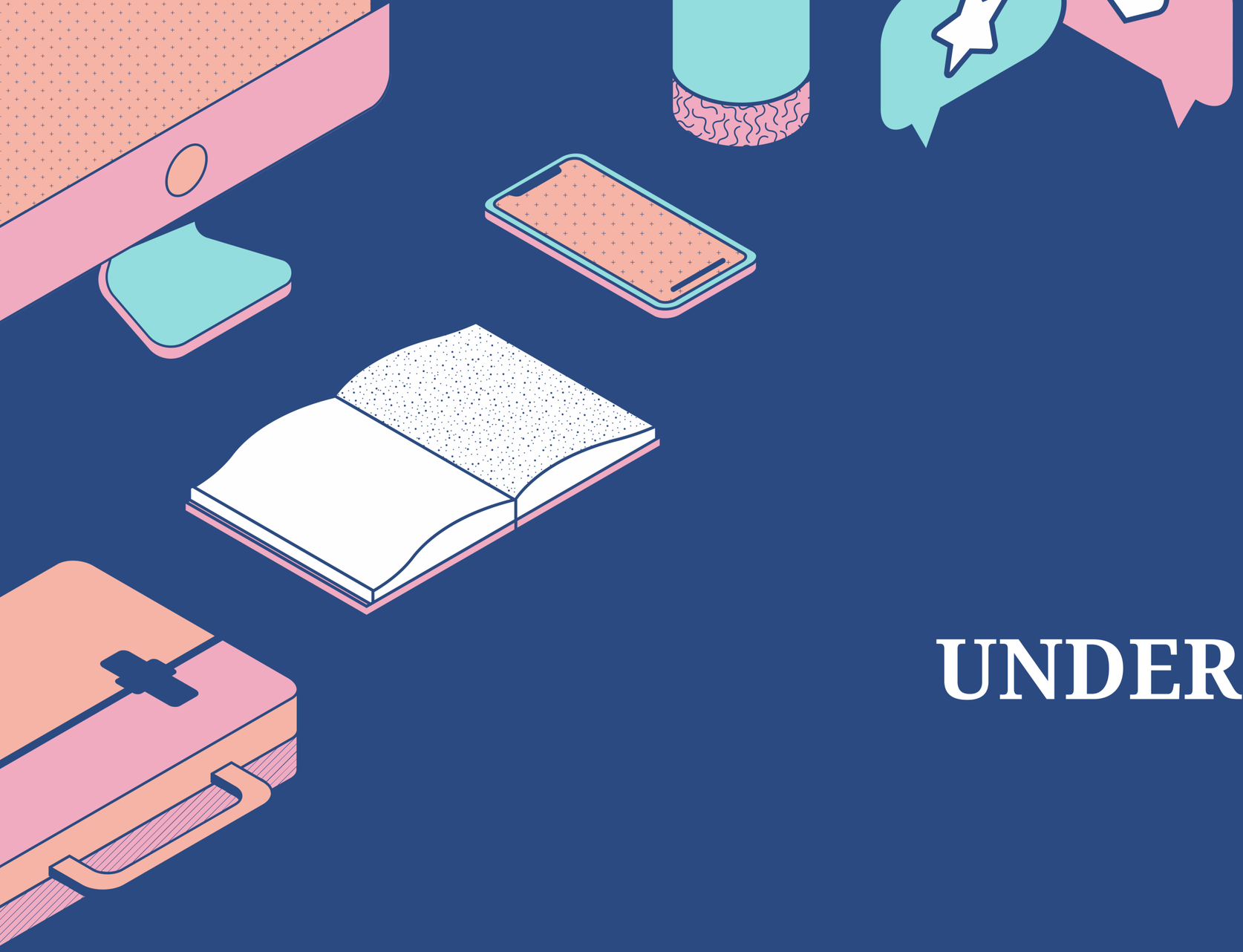
The online hotel reservation channels have dramatically changed booking possibilities and customers' behavior. A significant number of hotel reservations are called-off due to cancellations or no-shows. The typical reasons for cancellations include change of plans, scheduling conflicts, etc. This is often made easier by the option to do so free of charge or preferably at a low cost which is beneficial to hotel guests but it is a less desirable and possibly revenue-diminishing factor for hotels to deal with.

Features	*no_of_adults, no_of_children, no_of_weekend_nights, no_of_week_nights, type_of_meal_plan, required_car_parking_space, room_type_reserved, lead_time, arrival_year, arrival_month, arrival_date, market_segment_type, repeated_guest, no_of_previous_cancellations, no_of_previous_bookings_not_canceled, avg_price_per_room, no_of_special_requests.*
Target Variables	*booking_status (Flag indicating if the booking was canceled or not).*



JUSTIFICATION

1. **Relevance to Hospitality Industry:**
   - Focuses on hotel bookings, crucial for industry insights.
   - Directly applicable to current hospitality challenges.
2. **Real-world Application of Data Mining:**
   - Aligns with practical use of data mining in hospitality.
   - Aims to extract patterns and insights using SAS SEMMA methodology.
3. **Potential for Predictive Analysis:**
   - 'Booking_status' allows for predictive modeling.
   - Understanding cancellation patterns can optimize booking strategies.



3

DATA UNDERSTANDING

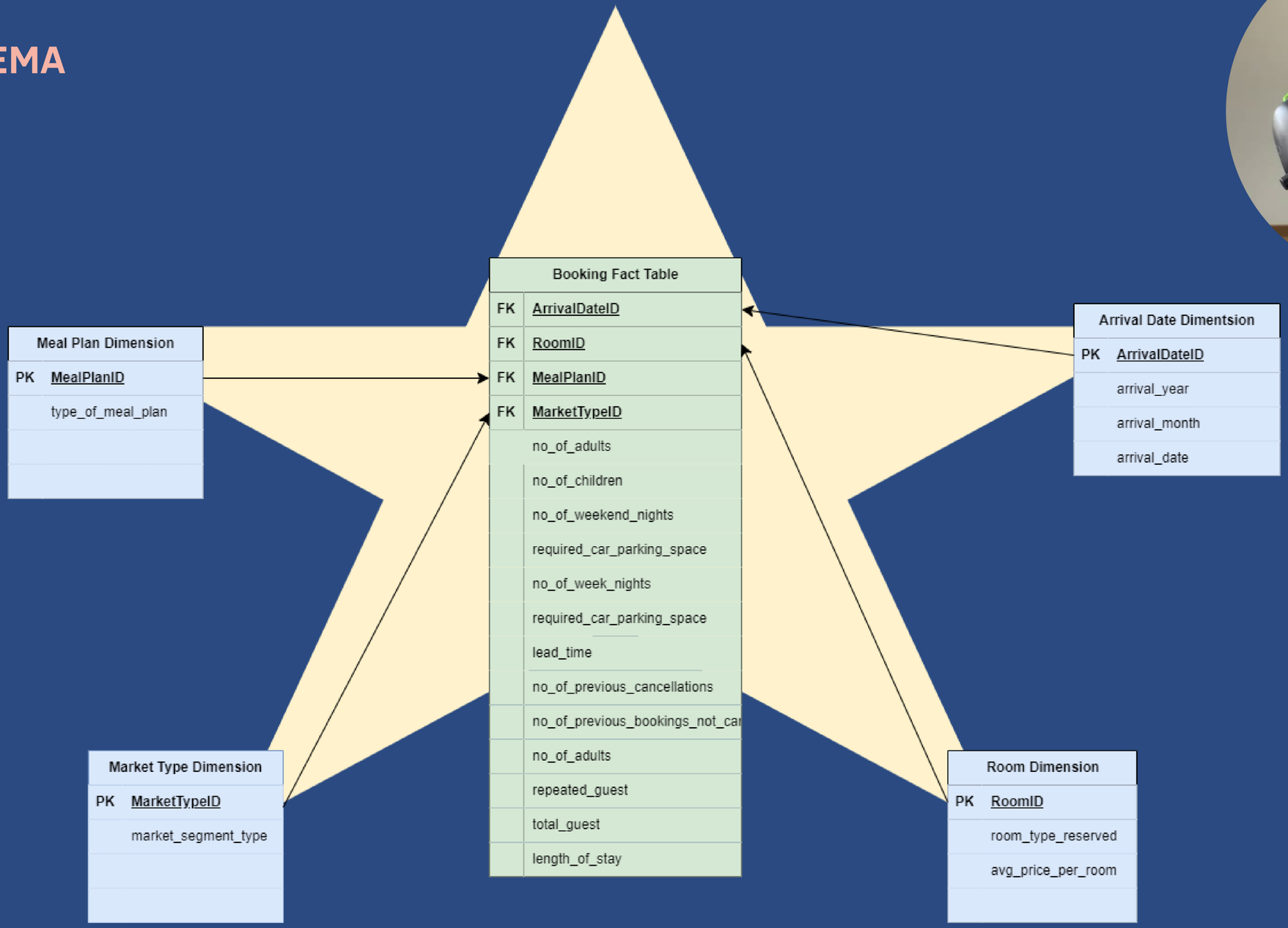
Structure & Features

- Number of Entries (Rows): 36275
- Number of Features (Columns): 19

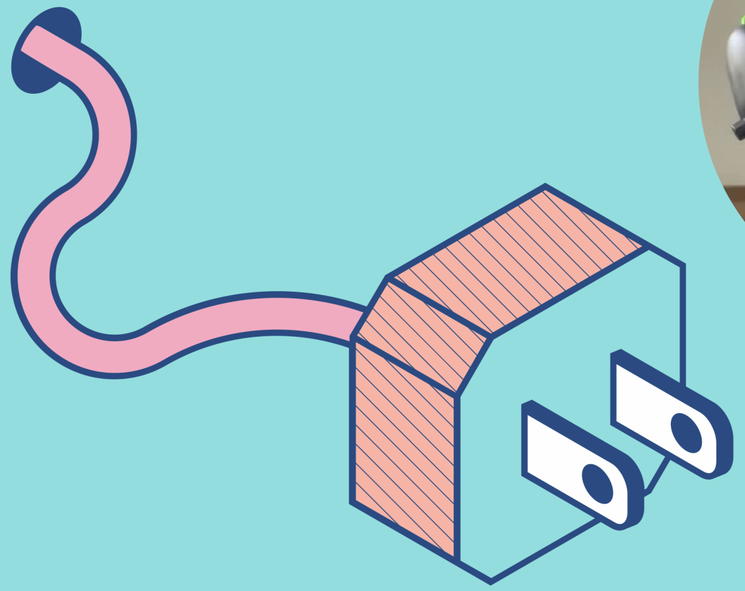
DATA DICTIONARY

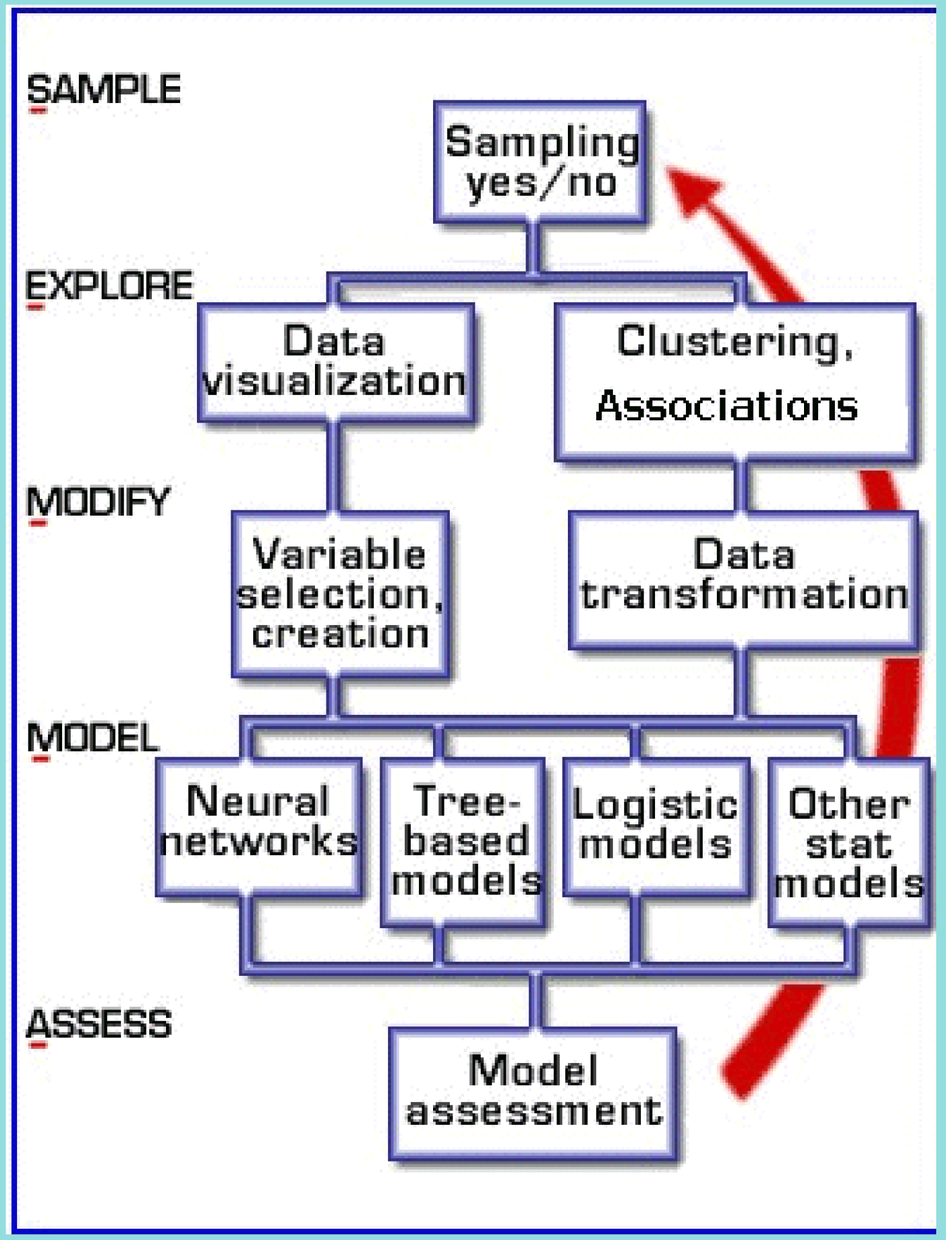
Field Name	Description	Data Type
Booking_ID	Unique identifier of each booking	String or Integer
no_of_adults	Number of adults	Integer
no_of_children	Number of children	Integer
no_of_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel	Integer
no_of_week_nights	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel	Integer
type_of_meal_plan	Type of meal plan booked by the customer	String
required_car_parking_space	Does the customer require a car parking space? (0 - No, 1 - Yes)	Integer (0 or 1)
room_type_reserved	Type of room reserved by the customer; values are ciphered (encoded) by INN Hotels	String
lead_time	Number of days between the date of booking and the arrival date	Integer
arrival_year	Year of arrival date	Integer
arrival_month	Month of arrival date	Integer
arrival_date	Date of the month	Integer
market_segment_type	Market segment designation	String
repeated_guest	Is the customer a repeated guest? (0 - No, 1 - Yes)	Integer (0 or 1)
no_of_previous_cancellations	Number of previous bookings that were canceled by the customer prior to the current booking	Integer
no_of_previous_bookings_not_canceled	Number of previous bookings not canceled by the customer prior to the current booking	Integer
avg_price_per_room	Average price per day of the reservation; prices of the rooms are dynamic (in euros)	Float or Decimal
no_of_special_requests	Total number of special requests made by the customer (e.g., high floor, view from the room)	Integer
booking_status	Flag indicating if the booking was canceled or not	String or Integer

STAR SCHEMA

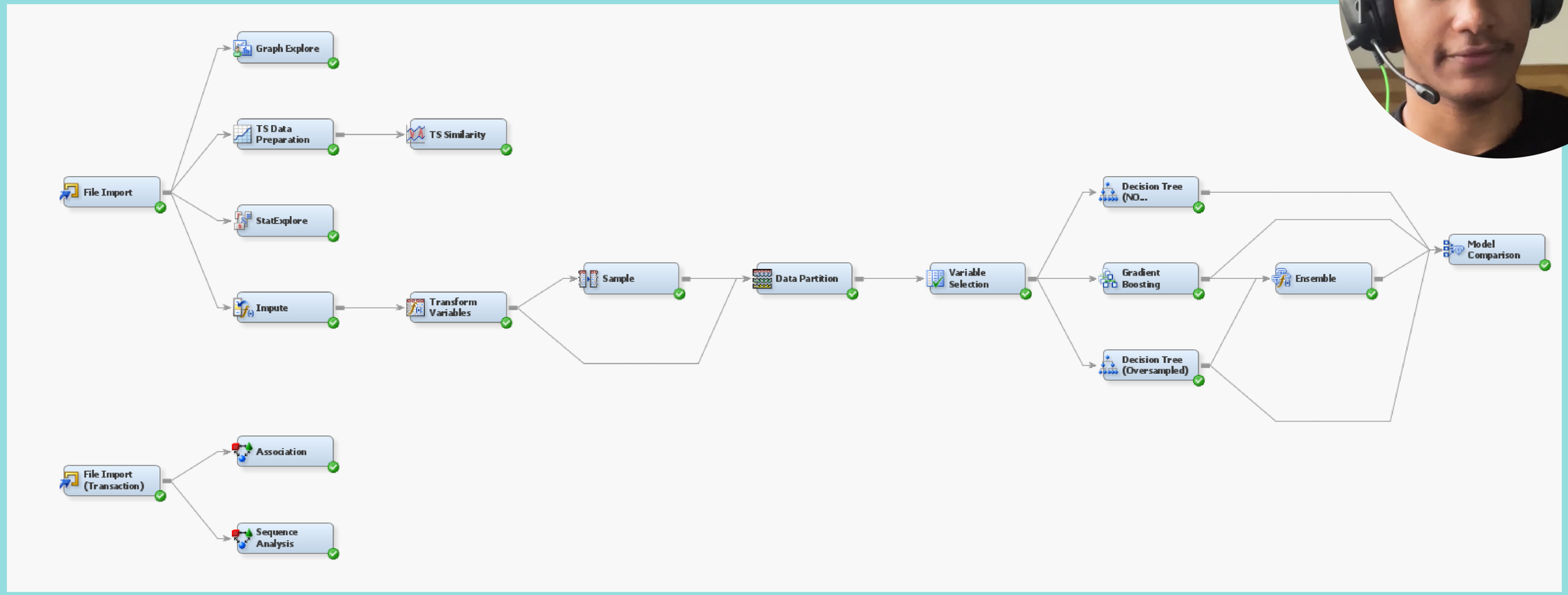

Meal Plan Dimension
PK MealPlanID
type_of_meal_plan
Booking Fact Table
FK ArrivalDateID
FK RoomID
FK MealPlanID
FK MarketTypeID
no_of_adults
no_of_children
no_of_weekend_nights
required_car_parking_space
no_of_week_nights
required_car_parking_space
lead_time
no_of_previous_cancellations
no_of_previous_bookings_not_car
no_of_adults
repeated_guest
total_guest
length_of_stay
Arrival Date Dimentsion
PK ArrivalDateID
arrival_year
arrival_month
arrival_date
Market Type Dimension
PK MarketTypeID
market_segment_type
Room Dimension
PK RoomID
room_type_reserved
avg_price_per_room

4
SAS SEMMA METHODOLOGY





FINAL MODEL



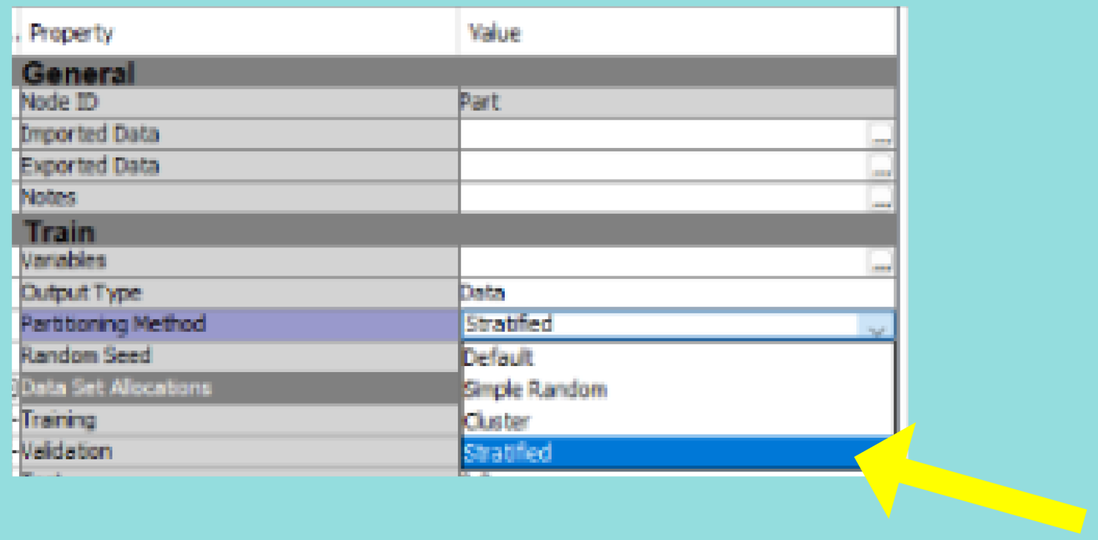
This documentation outlines a comprehensive **predictive modelling analysis** conducted on a dataset related to hotel bookings. The objective of this analysis is to develop accurate and robust models for predicting booking outcomes. The process encompasses various stages, including **data exploration, preprocessing, feature engineering,** and **the evaluation** of multiple machine learning algorithms.

Stage 1 - Sample

NODES

- Splits data into subsets for training, testing and validation stages of model development.
- Get good indication of how well the model can generalize to larger datasets.

CONFIGURATION



- Partitioning Method: **Stratified**
- Dataset Allocations:
  - 70% Training
  - 30% Validation

RESULTS

Output

Partition Summary

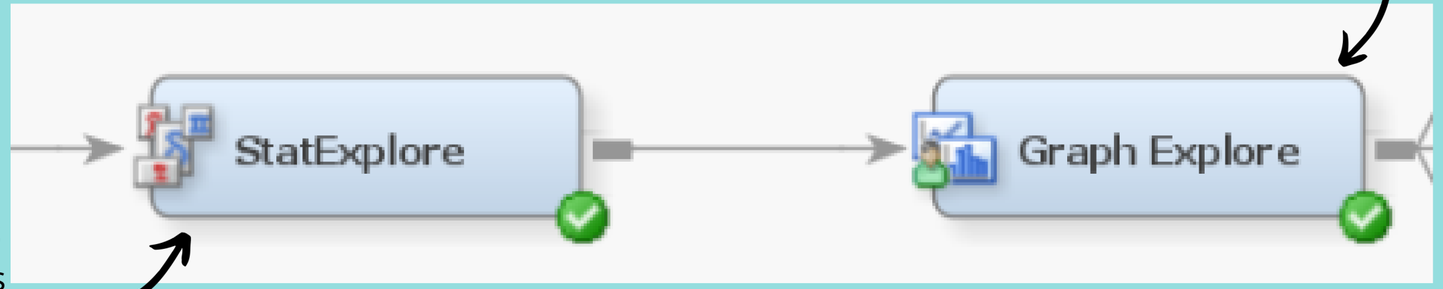
Type	Data Set	Number of Observations
DATA	EMWS3.Trans_TRAIN	30000
TRAIN	EMWS3.Part_TRAIN	20999
VALIDATE	EMWS3.Part_VALIDATE	9001

Stage 2 - Explore

EXPLORATORY DATA ANALYSIS

Explore large volumes of data graphically to uncover patterns and trends and to reveal extreme values in the database.

- Examine variable distributions & statistics
- Summarize the statistics



Variable Summary

Role	Measurement Level	Frequency Count
INPUT	BINARY	1
INPUT	INTERVAL	13
INPUT	NOMINAL	4
REJECTED	INTERVAL	1
TARGET	BINARY	1

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
arrival_date	INPUT	15.58393	8.723193	30000	0	1	16	31	0.030832	-1.15523
arrival_month	INPUT	7.428633	3.064031	30000	0	1	8	12	-0.34745	-0.93175
arrival_year	INPUT	2017.822	0.382631	30000	0	2017	2018	2018	-1.68251	-442.61
avg_price_per_room	INPUT	103.4134	35.05049	29914	86	0	99.45	375.5	0.626596	2.572624
lead_time	INPUT	85.5006	86.07222	30000	0	0	57	443	1.292343	1.184088
length_of_stay	INPUT	3.008267	1.773669	30000	0	0	3	24	2.190381	11.80667
no_of_children	INPUT	0.105033	0.403741	30000	0	0	0	10	4.867538	41.30415
no_of_previous_bookings_not_canc	INPUT	0.155133	1.771696	30000	0	0	0	58	19.32383	463.6431
no_of_previous_cancellations	INPUT	0.0227	0.35316	30000	0	0	0	13	25.54338	767.5436
no_of_special_requests	INPUT	0.6207	0.787349	30000	0	0	0	5	1.142626	0.862264
no_of_week_nights	INPUT	2.202133	1.402191	30000	0	0	2	17	1.560763	7.466199
no_of_weekend_nights	INPUT	0.806133	0.868547	30000	0	0	1	7	0.737165	0.257155
total_guests	INPUT	1.950867	0.650385	30000	0	1	2	12	0.898389	5.26545

Abnormally high skewness and kurtosis

no_of_children	INPUT	0.105033	0.403741	30000	0	0	0	10	4.867538	41.30415
no_of_previous_bookings_not_canc	INPUT	0.155133	1.771696	30000	0	0	0	58	19.32383	463.6431
no_of_previous_cancellations	INPUT	0.0227	0.35316	30000	0	0	0	13	25.54338	767.5436

Variable	Role	Mean	Standard Deviation	Non Missing	Missing
arrival_date	INPUT	15.58393	8.723193	30000	0
arrival_month	INPUT	7.428633	3.064031	30000	0
arrival_year	INPUT	2017.822	0.382631	30000	0
avg_price_per_room	INPUT	103.4134	35.05049	29914	86

Presence of 86 missing values

Chi-Square statistics

Input	Chi-Square	Df	Prob
lead_time	5502.5163	4	<.0001
no_of_special_requests	995.0204	4	<.0001
arrival_year	934.2196	1	<.0001
market_segment_type	715.4852	4	<.0001
arrival_month	545.4652	4	<.0001
avg_price_per_room	534.2270	5	<.0001
repeated_guest	339.8776	1	<.0001
required_car_parking_space	222.7899	1	<.0001
type_of_meal_plan	219.1374	3	<.0001
no_of_week_nights	184.8766	4	<.0001
length_of_stay	154.2180	4	<.0001
no_of_weekend_nights	121.9469	4	<.0001
no_of_previous_bookings_not_canc	51.2550	4	<.0001
room_type_reserved	43.9391	6	<.0001
total_guests	38.2518	2	<.0001
no_of_previous_cancellations	27.1623	3	<.0001
arrival_date	10.8054	4	0.0288
no_of_children	0.4336	2	0.8051

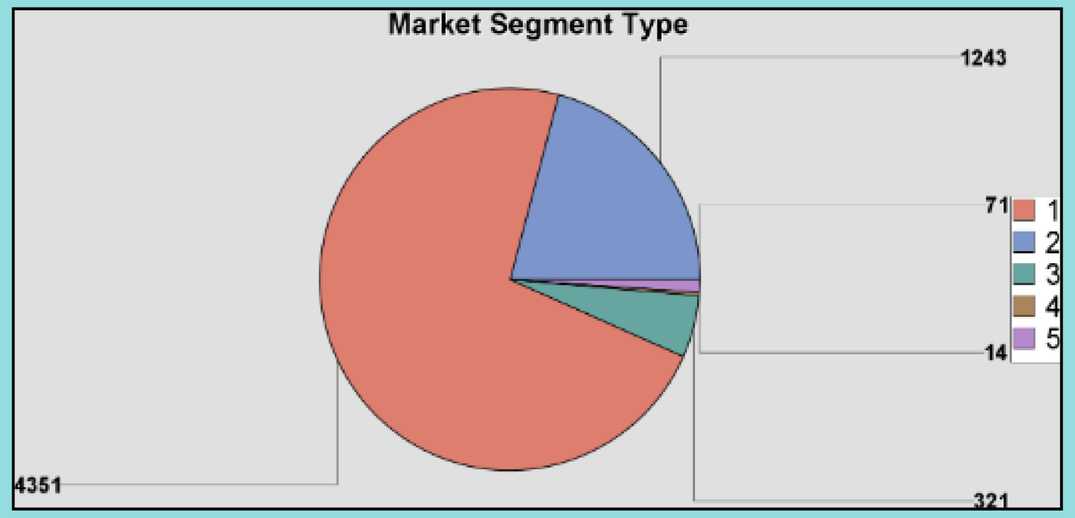
Highly Significant Predictors

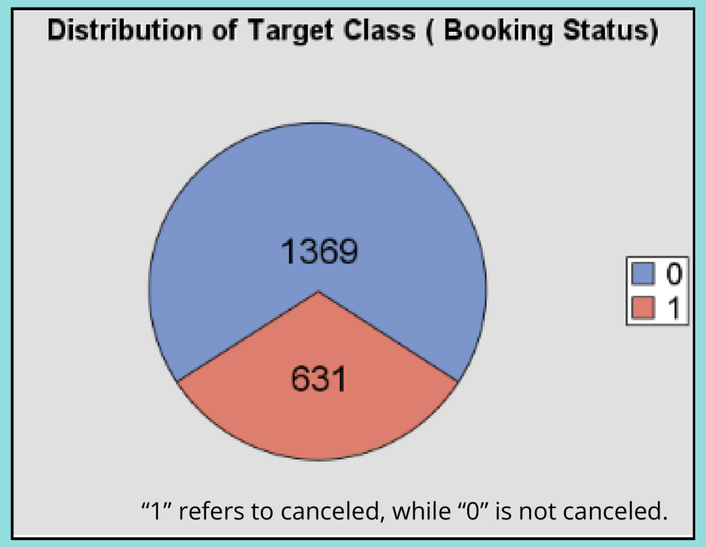
Moderately Significant Predictors

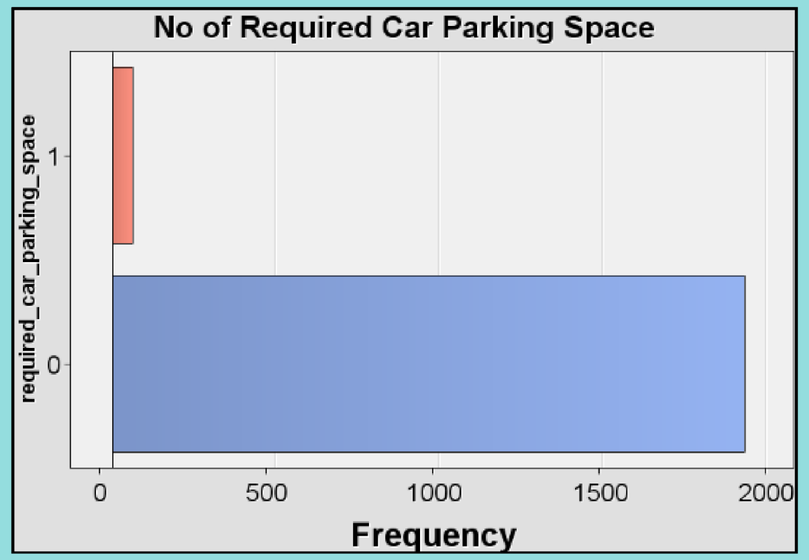
Less Significant Predictors

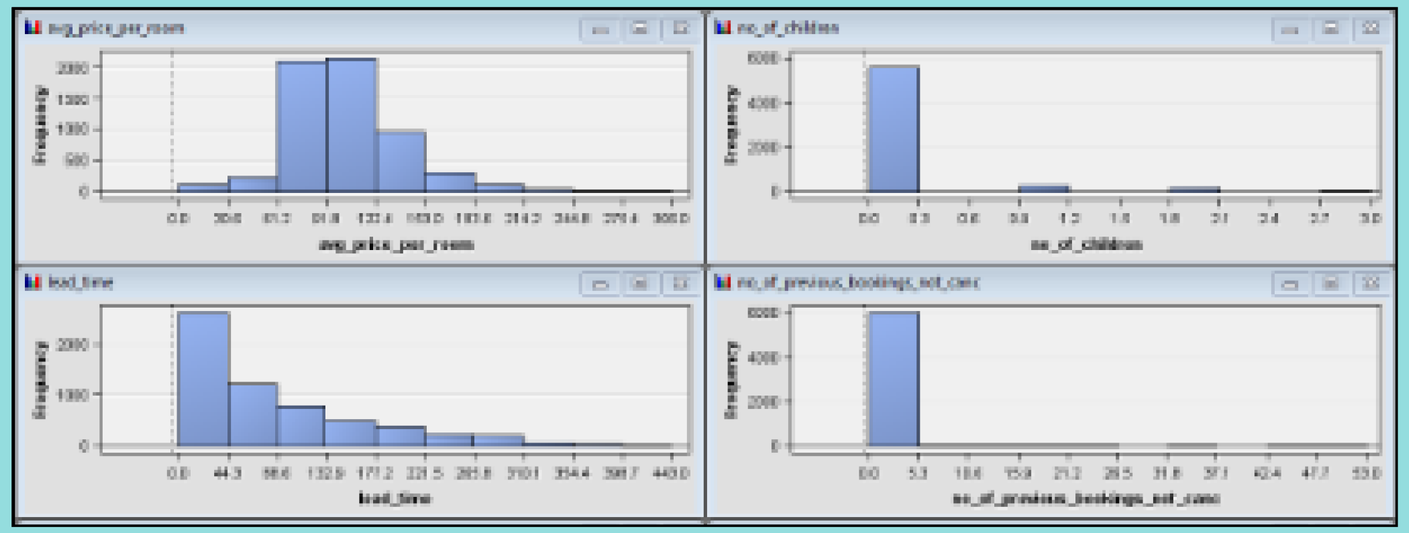
Insignificant Predictors

Sample graphs made with Graph Explore Node

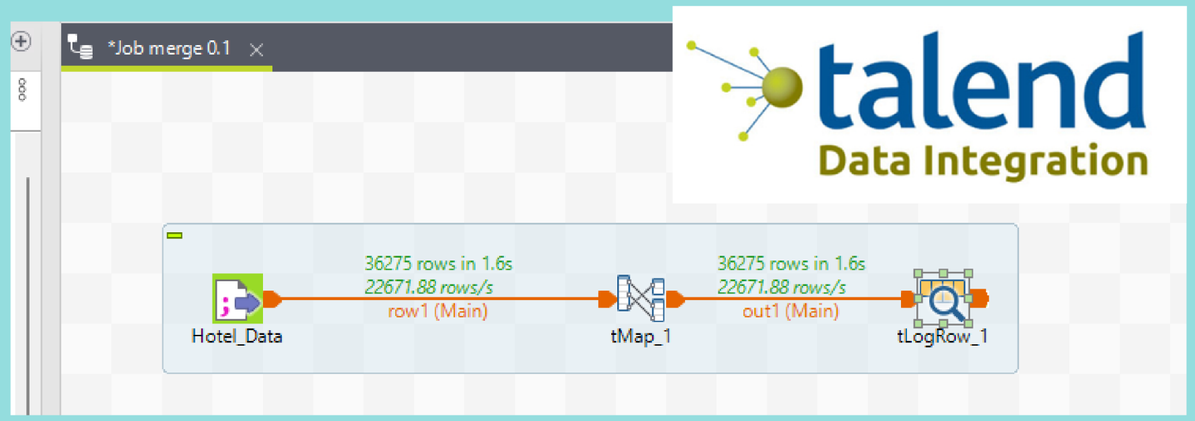


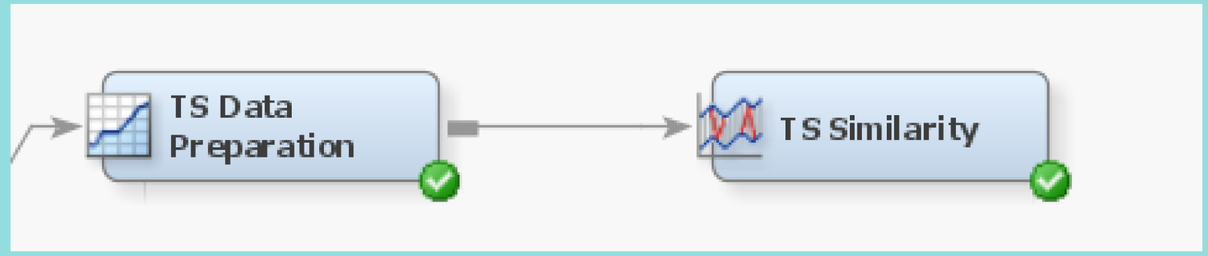






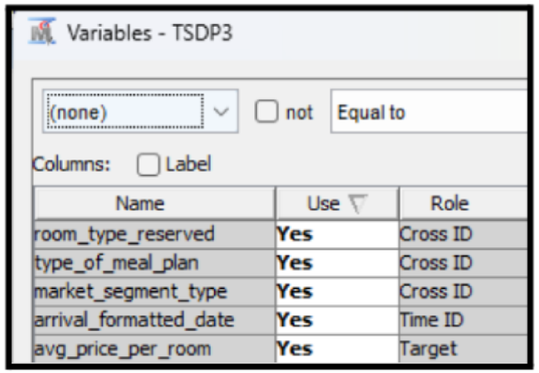
TIME SERIES SIMILARITY & CLUSTERING

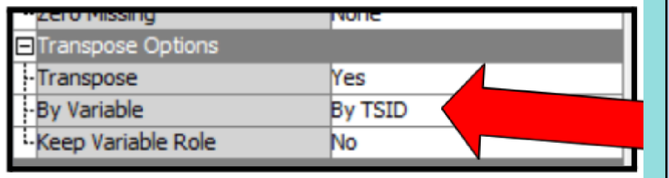




Configuration

In Time Series Similarity, the **TS Data Preparation** was configured to set Transpose by variable **"By TSID"** as shown in the picture below and then the variables were selected for t analysis are shown in the right picture.





Results:

TSID Map Summary Table

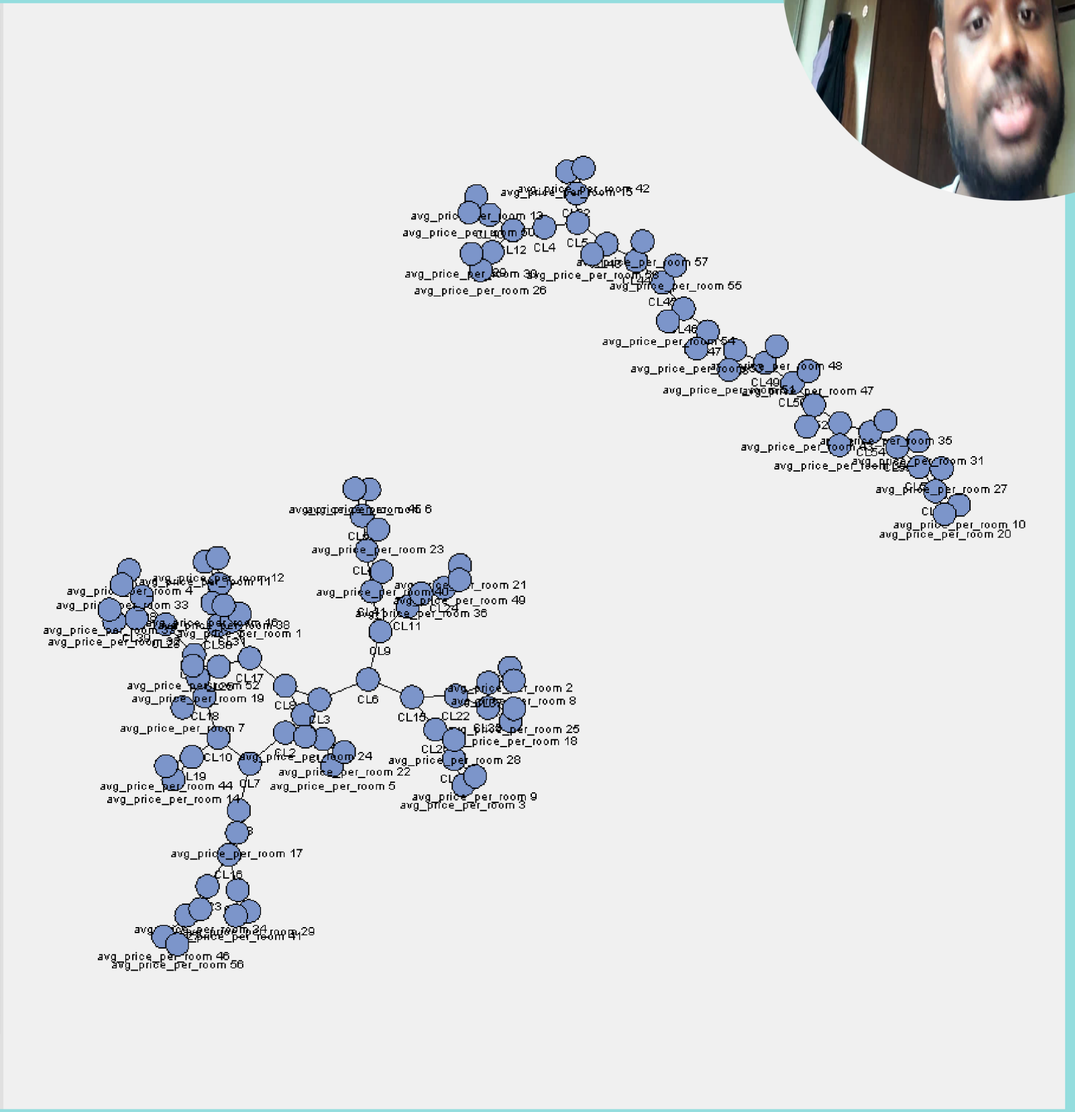
Name	Level	Count	Percent
booking_status	0	55	63.21839
booking_status	1	32	36.78161
market_segment_type	1	28	32.18391
market_segment_type	3	12	13.7931
market_segment_type	4	4	4.597701
market_segment_type	5	16	18.3908
market_segment_type	2	27	31.03448
room_type_reserved	7	9	10.34483
room_type_reserved	6	12	13.7931
room_type_reserved	5	12	13.7931
room_type_reserved	4	18	20.68966
room_type_reserved	3	3	3.448276
room_type_reserved	2	10	11.49425
room_type_reserved	1	23	26.43678
type_of_meal_plan	0	18	20.68966
type_of_meal_plan	3	3	3.448276
type_of_meal_plan	2	21	24.13793
type_of_meal_plan	1	45	51.72414
TSID		87	100

Total of 87 time series were created

TIME SERIES SIMILARITY

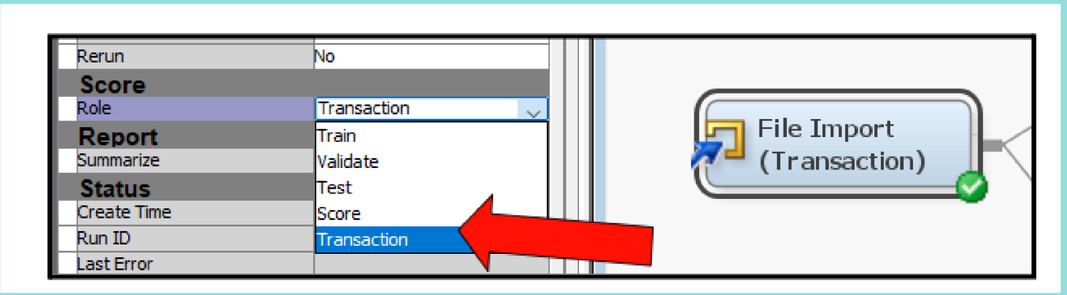
Time Series Similarity Index

Time Series Cluster Plot



ASSOCIATION ANALYSIS

Configuration of Data :



Coding to prepare data :

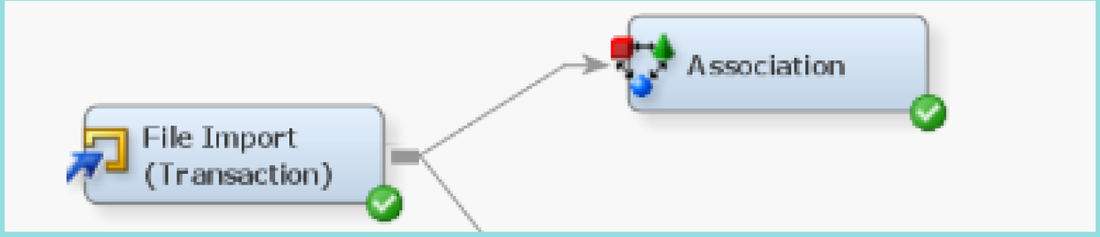
+ Code + Markdown

Preparing Data for Association and Sequence Analysis

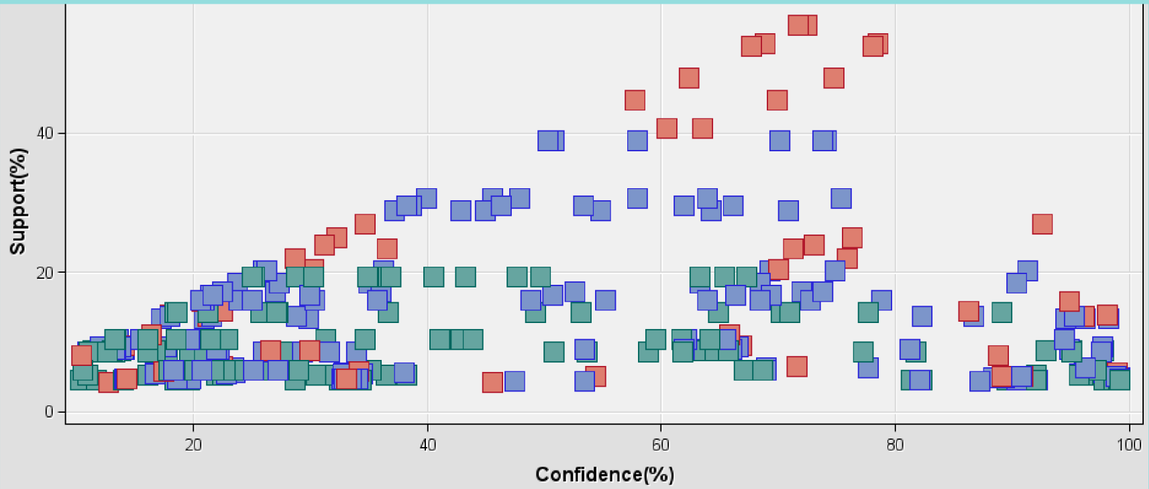
```
# Step 1: Change 'Not Selected' to 'No Meal Plan' in 'type_of_meal_plan'
hotel_data['type_of_meal_plan'] = hotel_data['type_of_meal_plan'].replace('Not Selected', 'No Meal Plan')

# Step 2: Change 0 to 'No Car Parking' and 1 to 'Yes Car Parking' in 'required_car_parking_space'
hotel_data['required_car_parking_space'] = hotel_data['required_car_parking_space'].map({0: 'No Car Parking', 1: 'Yes Car Parking'})

# Step 3: Change 0 to 'No Repeated Guest' and 1 to 'Repeated Guest' in 'repeated_guest'
hotel_data['repeated_guest'] = hotel_data['repeated_guest'].map({0: 'No Repeated Guest', 1: 'Repeated Guest'})
```

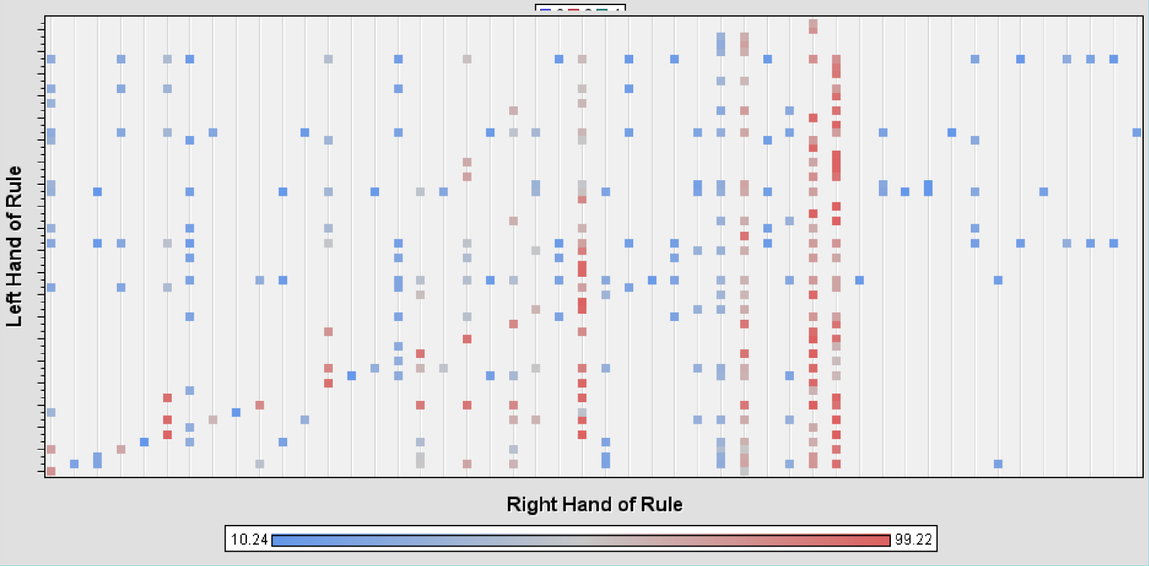
```
import pandas as pd

selected_columns = ['Booking_ID', 'room_type_reserved', 'type_of_meal_plan', 'required_car_parking_space', 'market_segment_type', 'repeated_guest', 'booking_status']

# Create a new DataFrame with selected columns
transaction_data = hotel_data[selected_columns]  # Copy all columns

# Convert the DataFrame to the desired format
transformed_data = pd.DataFrame(columns=['Booking_ID', 'Transaction_Details'])

for index, row in transaction_data.iterrows():
    booking_id = row['Booking_ID']
    transaction_details = row.drop('Booking_ID').astype(str).values
    transformed_data = pd.concat([transformed_data, pd.DataFrame({'Booking_ID': [booking_id] * len(row.drop('Booking_ID')), 'Transaction_Details': transaction_details})])
```

```
transformed_data.to_csv('Dataset\AssociatianAnalysisPreppedData.csv', index=False)
```



Results:





Rules Table

Relations	Expected Confidence(%)	Confidence(%)	Support(%)	Lift	Rule
2	32.76	45.57	4.15	1.39	Meal Plan 2 ==> Canceled
4	32.76	38.19	5.25	1.17	Room_Type 4 & Online & Meal Plan 1 ==> Canceled
3	32.76	38.05	5.48	1.16	Room_Type 4 & Online ==> Canceled
2	32.76	36.51	23.36	1.11	Online ==> Canceled
3	32.76	36.03	17.24	1.10	Online & Meal Plan 1 ==> Canceled
3	32.76	35.71	15.99	1.09	Room_Type 1 & Online ==> Canceled
4	32.76	34.61	10.25	1.06	Room_Type 1 & Online & Meal Plan 1 ==> Canceled
3	32.76	34.46	5.46	1.05	Room_Type 4 & Meal Plan 1 ==> Canceled
4	32.76	34.28	4.58	1.05	Room_Type 1 & Online & Not Selected ==> Canceled
2	32.76	34.16	5.70	1.04	Room_Type 4 ==> Canceled
3	32.76	33.98	4.62	1.04	Online & Not Selected ==> Canceled
3	32.76	33.45	4.65	1.02	Room_Type 1 & Not Selected ==> Canceled
2	32.76	33.12	4.68	1.01	Not Selected ==> Canceled

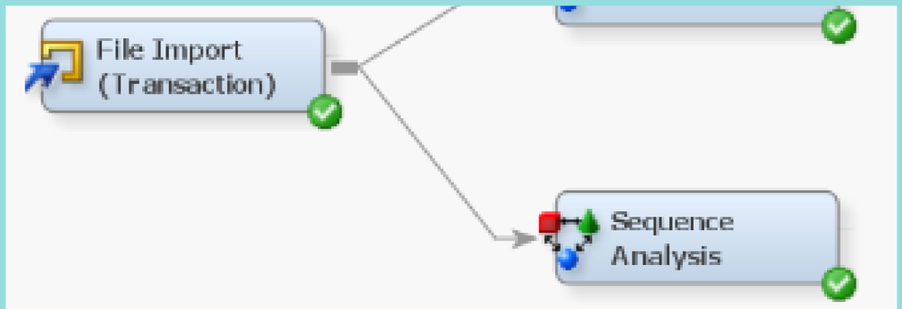
Cancel Booking:
Those who selected Meal Plan 2, Room Type 4, Online will cancel their booking with a confidence level of 32.76% while the highest confidence level is when customers select Meal Plan 2 at 45.57%

Rules Table

Relations	Expected Confidence(%)	Confidence(%)	Support(%)	Lift	Transaction Count	Rule	Right Hand of Rule
4	67.24	90.62	4.53	1.35	1642.0	Room_Type 1 & Meal Plan 1 & Corporate ==> Not_Canceled	Not_Canceled
3	67.24	90.34	4.57	1.34	1656.0	Room_Type 1 & Corporate ==> Not_Canceled	Not_Canceled
3	67.24	89.33	4.92	1.33	1783.0	Meal Plan 1 & Corporate ==> Not_Canceled	Not_Canceled
2	67.24	89.09	4.95	1.33	1797.0	Corporate ==> Not_Canceled	Not_Canceled
3	67.24	72.76	16.02	1.08	5812.0	Offline & Meal Plan 1 ==> Not_Canceled	Not_Canceled
4	67.24	71.01	14.29	1.06	5182.0	Room_Type 1 & Offline & Meal Plan 1 ==> Not_Canceled	Not_Canceled
3	67.24	70.13	38.97	1.04	14136	Room_Type 1 & Meal Plan 1 ==> Not_Canceled	Not_Canceled
2	67.24	70.05	20.33	1.04	7375.0	Offline ==> Not_Canceled	Not_Canceled
2	67.24	68.82	52.81	1.02	19156	Meal Plan 1 ==> Not_Canceled	Not_Canceled
3	67.24	68.40	18.38	1.02	6667.0	Room_Type 1 & Offline ==> Not_Canceled	Not_Canceled
2	67.24	67.75	52.54	1.01	19058	Room_Type 1 ==> Not_Canceled	Not_Canceled

Not Cancel booking:
Those who choose Room Type 1, Meal Plan 1, Corporate, Offline group will not cancel their booking with expected confidence level of 67.24% while the highest confidence level is when customer choose Room Type 1, Meal Plan 1 and Corporate at 90.62%.

SEQUENCE ANALYSIS



Sequence analysis is performed, by choosing the same Association node but was renamed to Sequence Analysis

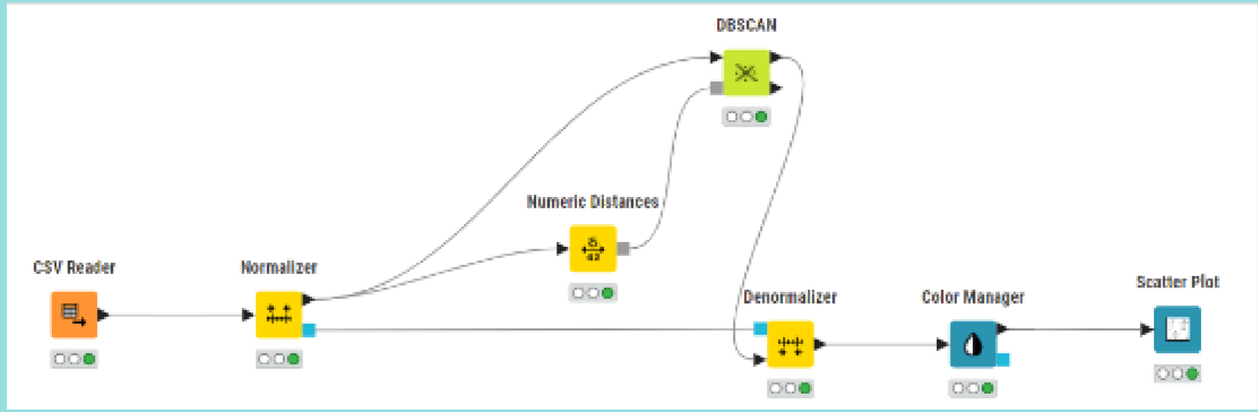
Configuration:

- Export Rule by ID is set to “Yes”

Rule Description

Map	Rule
RULE1	Room_Type 1 & Meal Plan 2 ==> Offline
RULE2	Offline ==> Room_Type 1 & Meal Plan 2
RULE3	Room_Type 1 & Offline ==> Meal Plan 2
RULE4	Meal Plan 2 ==> Room_Type 1 & Offline
RULE5	Meal Plan 2 ==> Offline
RULE6	Offline ==> Meal Plan 2
RULE7	Room_Type 1 & Online ==> Not Selected & Canceled
RULE8	Not Selected & Canceled ==> Room_Type 1 & Online
RULE9	Room_Type 1 & Online & Not_Canceled ==> Not Selected
RULE10	Not Selected ==> Room_Type 1 & Online & Not_Canceled
RULE11	Room_Type 1 & Online ==> Not Selected
RULE12	Not Selected ==> Room_Type 1 & Online
RULE13	Room_Type 1 & Not_Canceled & Meal Plan 1 ==> Corporate
RULE14	Corporate ==> Room_Type 1 & Not_Canceled & Meal Plan 1
RULE15	Not_Canceled & Not Selected ==> Room_Type 1 & Online
RULE16	Room_Type 1 & Online ==> Not_Canceled & Not Selected
RULE17	Not Selected ==> Room_Type 1 & Online & Canceled
RULE18	Room_Type 1 & Online & Canceled ==> Not Selected
RULE19	Room_Type 4 & Canceled ==> Online & Meal Plan 1
RULE20	Online & Meal Plan 1 ==> Room_Type 4 & Canceled
RULE21	Room_Type 4 ==> Online & Meal Plan 1 & Canceled
RULE22	Online & Meal Plan 1 & Canceled ==> Room_Type 4
RULE23	Room_Type 4 ==> Online & Meal Plan 1
RULE24	Online & Meal Plan 1 ==> Room_Type 4
RULE25	Room_Type 1 & Corporate ==> Not_Canceled & Meal Plan 1
RULE26	Corporate ==> Not_Canceled & Meal Plan 1
RULE27	Room_Type 4 ==> Online & Not_Canceled & Meal Plan 1
RULE28	Online & Not_Canceled & Meal Plan 1 ==> Room_Type 4
RULE29	Not_Canceled & Corporate ==> Room_Type 1 & Meal Plan 1
RULE30	Corporate ==> Room_Type 1 & Meal Plan 1
RULE31	Room_Type 4 & Not_Canceled ==> Online & Meal Plan 1
RULE32	Online & Meal Plan 1 ==> Room_Type 4 & Not_Canceled

DBSCAN CLUSTERING



Open for Innovation
KNIME

Knime is used as the tool to perform clustering.

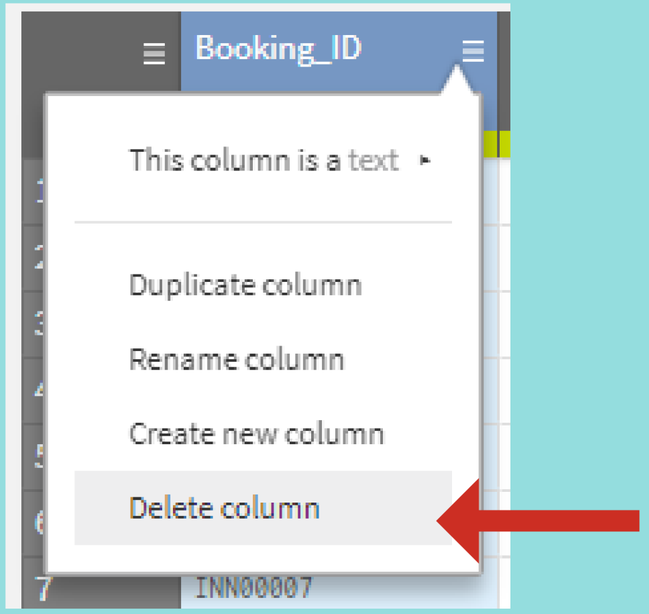
Results:

2 successfull clusters were formed !

Stage 3 - Modify

In this phase, the data is parsed and cleaned, being then passed onto the modelling stage, and explored if the data requires refinement and transformation. The process utilized Talend DataPrep tool.

TASK 1 : REMOVE UNWANTED VARIABLE



Booking ID

- identified as irrelevant variable for the prediction problem
- was removed from the data using Talend DataPrep

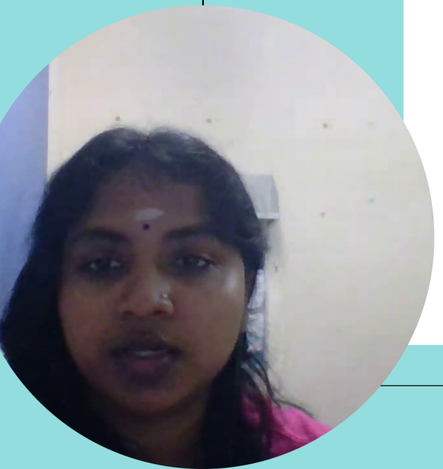
TASK 2 : ENCODE CATEGORICAL VARIABLES

Variables below were in text which **needed to be converted to numerical format.**

Before encoding:

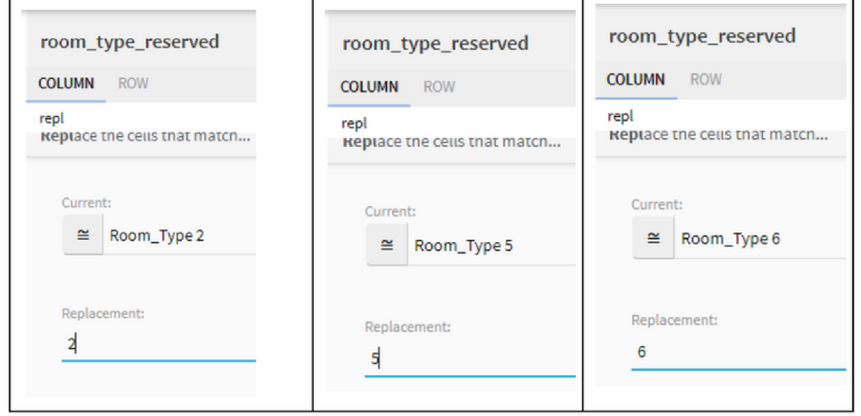
The encoding of variables are as below :

- **room_type_reserved**
  - Room_Type 1 → 1
  - Room_Type 2 → 2
  - Room_Type 3 → 3
  - Room_Type 4 → 4
  - Room_Type 5 → 5
  - Room_Type 6 → 6
- **type_of_meal_plan**
  - Not_selected → 0
  - Meal Plan 1 → 1
  - Meal Plan 2 → 2
  - Meal Plan 3 → 3
- **Market_segment_type**
  - Online → 1
  - Offiline → 2
  - Corporate → 3
  - Aviation → 4
  - Complemenatry → 5
- **booking_status**
  - Cancelled →1
  - Not_cancelled → 0

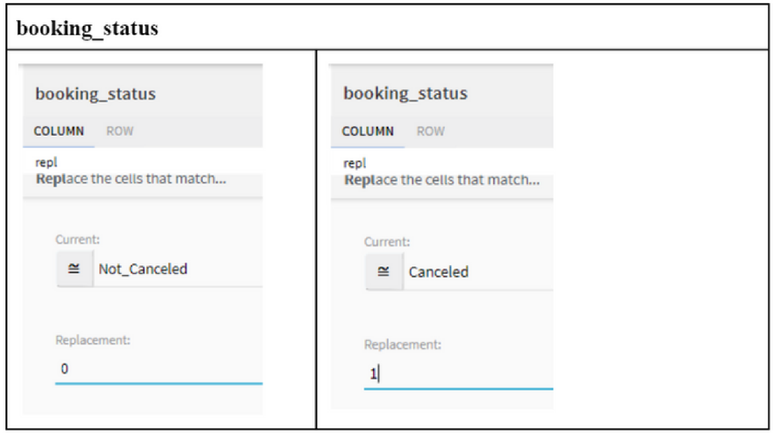


After encoding:

room_type_reserved



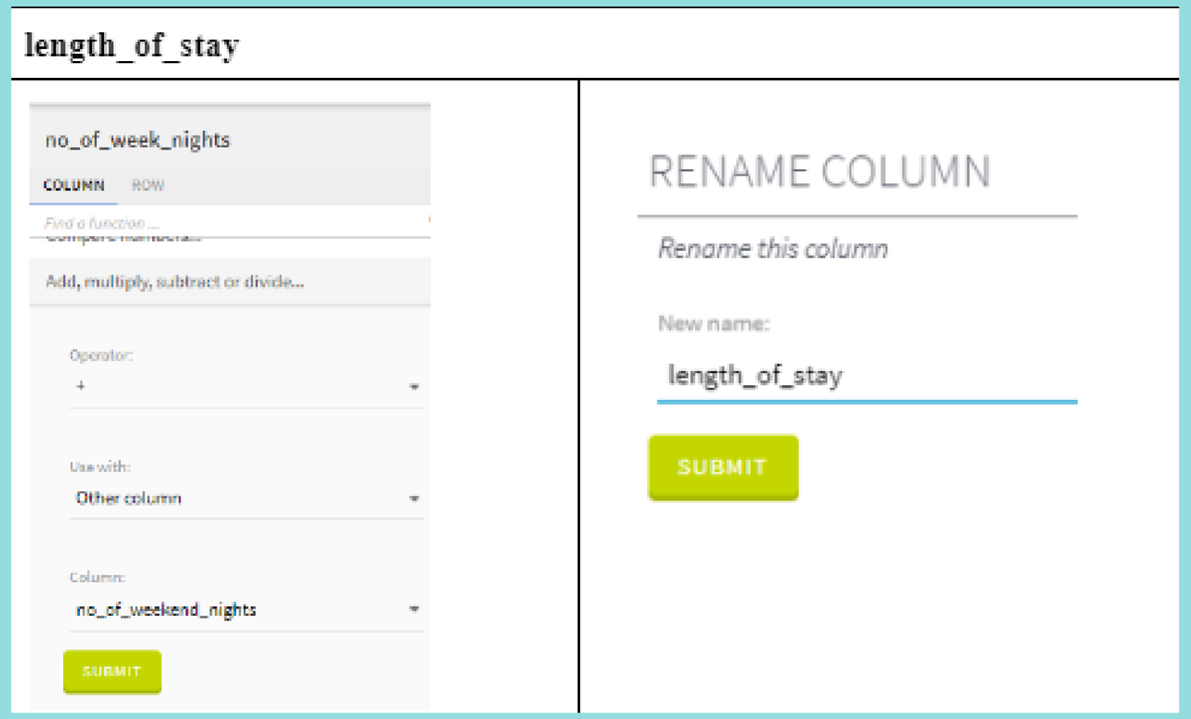
booking_status



TASK 3 : FEATURE ENGINEERING

- process of extracting features from raw data to support training a downstream statistical model.

length_of_stay = no_of_weekend_nights + no_of_weekday_nights
total_guests = no_of_adults + no_of_child



TASK 4 : VARIABLE PREPARATION

- data was loaded into SAS Enterprise Miner
- prepare the variables by assigning roles and levels to the variables.

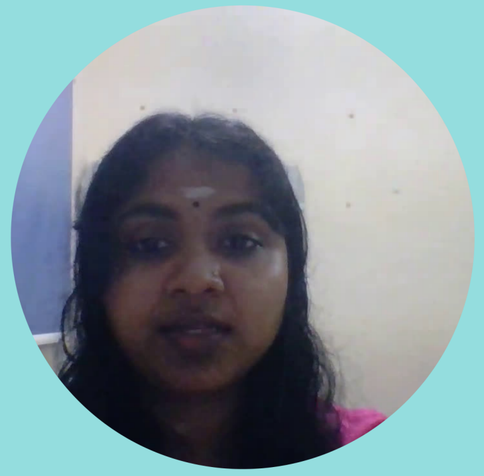
Role

- booking_status : Input → Target

Level

- booking_status : Interval → Binary
- market_segment_type: Interval → Nominal
- repeated_guest : Interval → Binary
- required_car_parking_space : Interval → Binary
- room_type_reserved : Interval → Nominal
- type_of_meal_plan : Interval → Nominal

Name	Role	Level
arrival_date	Input	Interval
arrival_month	Input	Interval
arrival_year	Input	Interval
avg_price_per_r	Input	Interval
booking_status	Input	Interval
lead_time	Input	Interval
length_of_stay	Input	Interval
market_segment	Input	Interval
no_of_children	Input	Interval
no_of_previous_	Input	Interval
no_of_previous_	Input	Interval
no_of_special_r	Input	Interval
no_of_week_nig	Input	Interval
no_of_weekend	Input	Interval
repeated_guest	Input	Interval
required_car_pa	Input	Interval
room_type_rese	Input	Interval
total_guests	Input	Interval
type_of_meal_p	Input	Interval



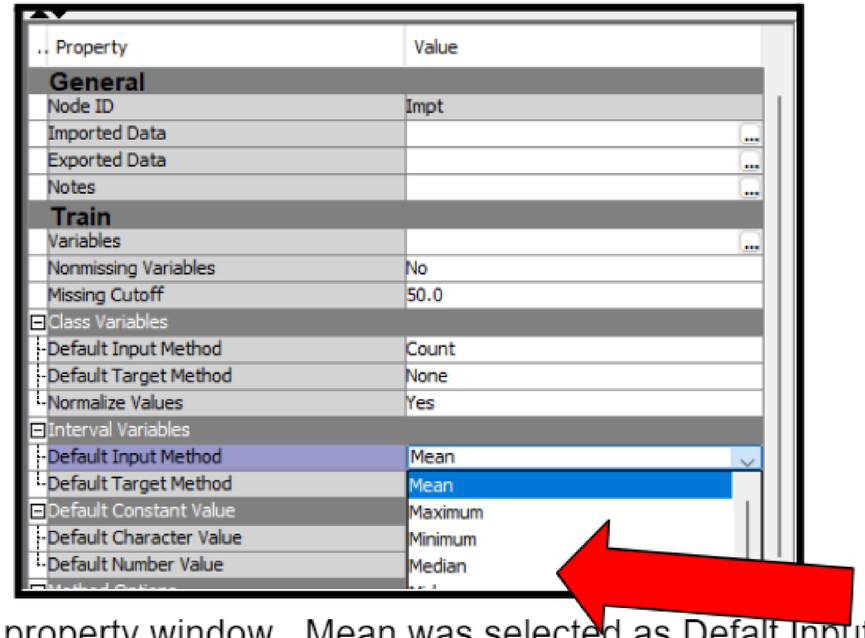
Name	Role	Level
arrival_date	Input	Interval
arrival_month	Input	Interval
arrival_year	Input	Interval
avg_price_per_room	Input	Interval
booking_status	Target	Binary
lead_time	Input	Interval
length_of_stay	Input	Interval
market_segment_type	Input	Nominal
no_of_children	Input	Interval
no_of_previous_bookings_not_canc	Input	Interval
no_of_previous_cancellations	Input	Interval
no_of_special_requests	Input	Interval
no_of_week_nights	Input	Interval
no_of_weekend_nights	Input	Interval
repeated_guest	Input	Nominal
required_car_parking_space	Input	Binary
room_type_reserved	Input	Binary
total_guests	Input	Interval
type_of_meal_plan	Input	Nominal

TASK 5 : IMPUTE MISSING VALUES

- There were some missing values found in our dataset as indicated at Explore stage. Thus this task was crucial to address this problem.

Node Name :	Impute	**Configuration**

The Impute Node is more versatile and provides multiple imputation methods to handle missing values. It allows users to choose from a variety of imputation techniques, such as regression imputation, k-nearest neighbor imputation, and more.



In the property window , Mean was selected as Defalt Input Method. The mean is the most common measure of a variable's central tendency; it is an unbiased estimate of the population mean. The mean is the preferred statistic to use to replace missing values if the variable values are at least roughly symmetric

TASK 5 : IMPUTE MISSING VALUES

- After imputing missing values:

Result

Imputation Summary
Number Of Observations

Variable Name	Impute Method	Imputed Variable	Indicator Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
REP_avg_price_per_room	MEAN	IMP_REP_avg_price_per_room	M_REP_avg_price_per_room	105.011	INPUT	INTERVAL	Replacement: avg_price_per_room	541

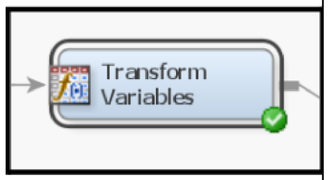
The variable "REP_avg_price_per_room" underwent imputation using the mean imputation method.These findings summarize the imputation process for the specified variable

- **Imputed Variable:**
  - The imputed values were stored in a new variable called "IMP_REP_avg_price_per_room."
- **Imputed Value:**
  - The mean imputed value for "REP_avg_price_per_room" is 105.011.

TASK 6 : DATA TRANSFORMATION

- As indicated in the Explore stage some of the variables were suffering from skewness.

Node Name :	Transform Variables	**Configuration**



The Transform Variables node allows users to create new variables from existing variables. The original variables used in the calculations will remain unchanged, however new column(s) will be added to the dataset.

As highlighted in explore stage on the high skewness and kurtosis problem , 3 variables are selected for the transformation which are again :

- no_of_children
- no_of_previous_bookings_not_canc
- no_of_previous_cancellations

Name	Method	Number of Bins	Role	Level
M_REP_avg_price_per_room	**Default**	4	Input	Binary
required_car_parking_space	**Default**	4	Input	Binary
booking_status	**Default**	4	Target	Binary
no_of_special_requests	**Default**	4	Input	Interval
arrival_month	**Default**	4	Input	Interval
length_of_stay	**Default**	4	Input	Interval
IMP_REP_avg_price_per_room	**Default**	4	Input	Interval
arrival_date	**Default**	4	Input	Interval
avg_price_per_room	**Default**	4	Rejected	Interval
no_of_week_nights	**Default**	4	Input	Interval
lead_time	**Default**	4	Input	Interval
VAR1	**Default**	4	Rejected	Interval
no_of_previous_cancellations	Default	4	Input	Interval
no_of_children	Default	4	Input	Interval
total_guests	Dummy Indica	4	Input	Interval
no_of_weekend_nights	Equalize	4	Input	Interval
arrival_year	Exponential	4	Input	Interval
no_of_previous_bookings_not_car	Group Rare Le	4	Input	Interval
market_segment_type	Inverse	4	Input	Nominal
type_of_meal_plan	Log		Input	Nominal
repeated_guest	Log 10		Input	Nominal
room_type_reserved	**Default**	4	Input	Nominal

For the method Log was selected as transformation method for these variables

Results

Transformations Statistics

Source	Method	Variable Name	Formula	Skewness
Output	Computed	LOG_no_of_children	log(no_of_children + 1)	3.629558
Output	Computed	LOG_no_of_previous_bookings_...	log(no_of_previous_bookings_not_canc + 1)	8.916943
Output	Computed	LOG_no_of_previous_cancellatio...	log(no_of_previous_cancellations + 1)	14.35091

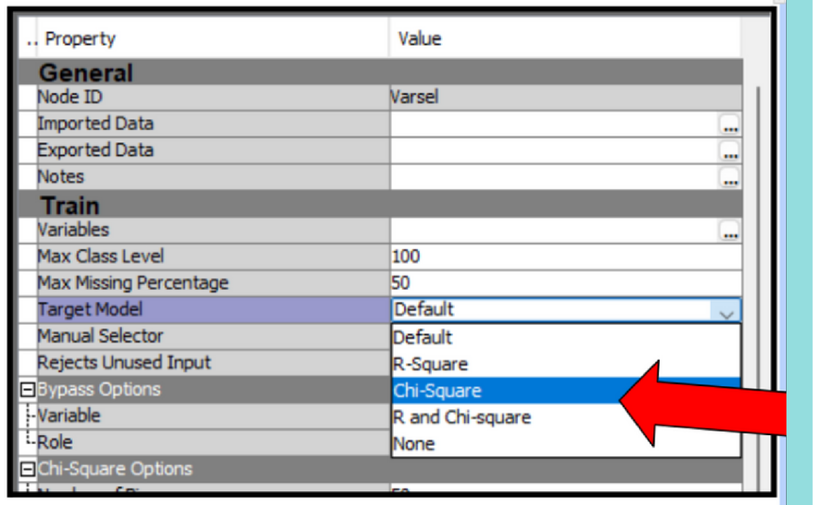
The above table indicates the newly created variables as the output of running the node. As we can see the transformed the variables has reduced skewness

TASK 7 : VARIABLE SELECTION

- To ensure highly predictive variables are selected for the model.

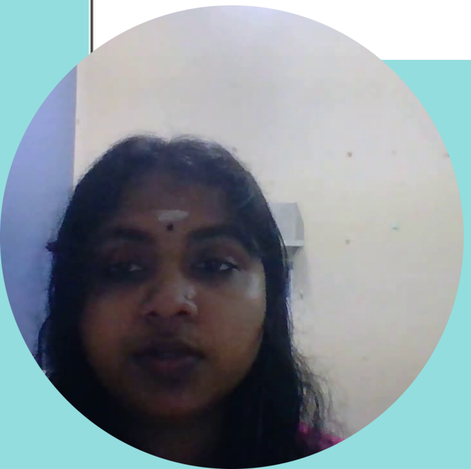
Node Name :	Variable selection	**Configuration**

The Variable Selection node quickly identifies input variables which are useful for predicting the target variable or variables. The input status is assigned to these variables.

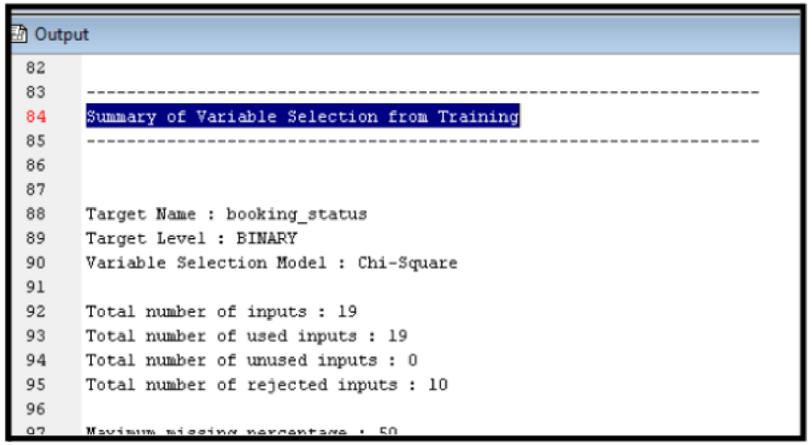


Property	Value
General	
Node ID	Varsel
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Max Class Level	100
Max Missing Percentage	50
Target Model	Default
Manual Selector	Default
Rejects Unused Input	R-Square
Bypass Options	Chi-Square
Variable	R and Chi-square
Role	None
Chi-Square Options	

Chi Square is chosen since it is classification problem.



Result

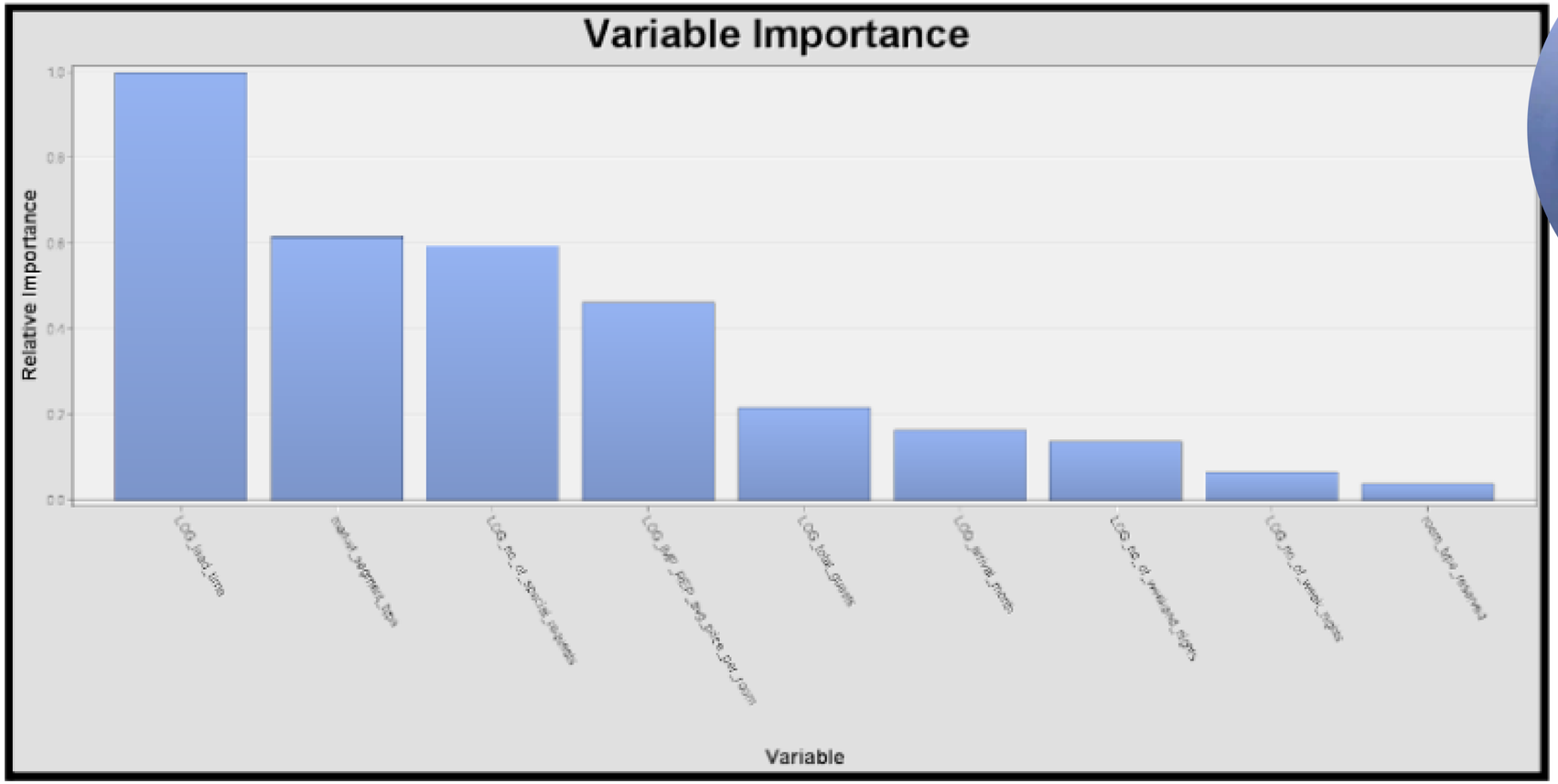


```
Output
82
83  ------------------------------------------------------------------
84  Summary of Variable Selection from Training
85  ------------------------------------------------------------------
86
87
88  Target Name : booking_status
89  Target Level : BINARY
90  Variable Selection Model : Chi-Square
91
92  Total number of inputs : 19
93  Total number of used inputs : 19
94  Total number of unused inputs : 0
95  Total number of rejected inputs : 10
96
97  Maximum missing percentage : 50
```

The summary indicates that Chi-Square was used for variable selection with the target variable being "booking_status" (a binary outcome). Here are key points from the summary:

- **Target Variable:**
  - **Name:** booking_status
  - **Type:** Binary (two possible outcomes)
- **Variable Selection Method:**
  - **Model**: Chi-Square
- **Number of Inputs:**
  - **Total Inputs**: 19
  - **Used Inputs**: 19
  - **Unused Inputs**: 0

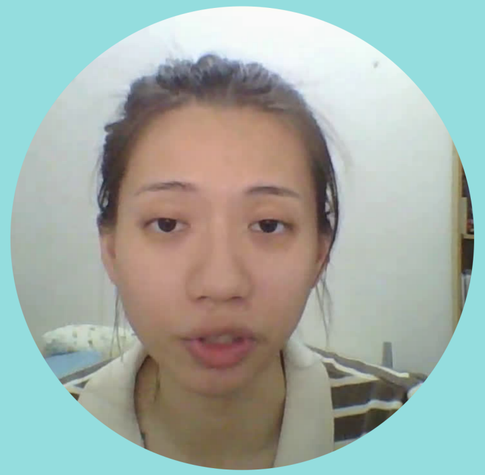
TASK 7 : VARIABLE SELECTION



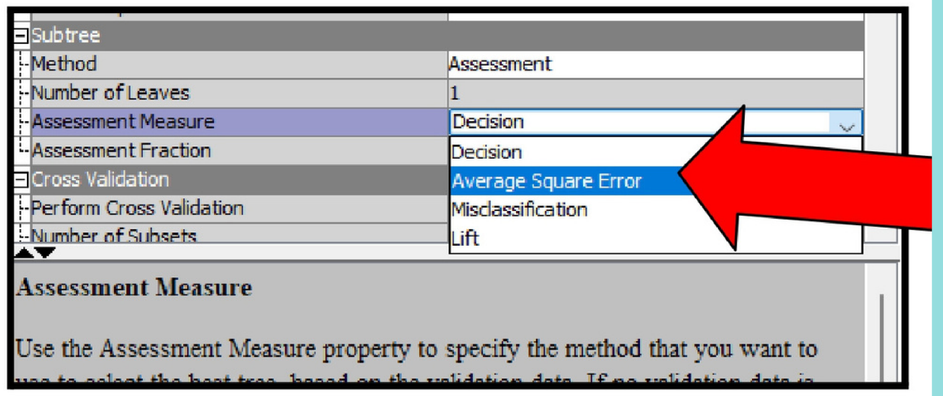
The graph above indicates the order on relative importance of the tha variables/ As we can see LOG_lead_time has the highest importance while room_type_reserved has the least importance.

Stage 4 - Model

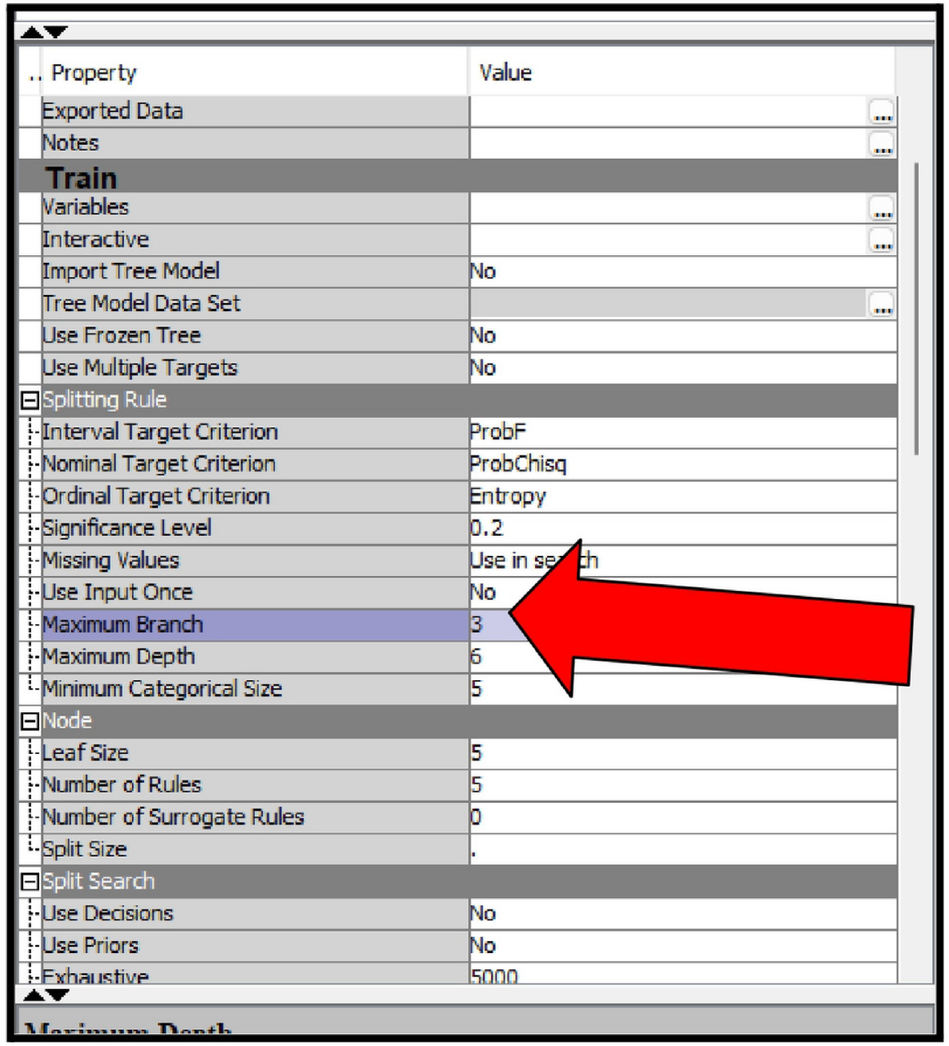
DECISION TREE



Configuration:



For the assessment measure Average Square Error was selected.



DECISION TREE

Result:

Fit Statistics

Target=booking_status Target Label=' '

Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	20999.00	9001.00
MISC	Misclassification Rate	0.16	0.16
MAX	Maximum Absolute Error	0.99	0.99
SSE	Sum of Squared Errors	4725.49	2016.10
ASE	Average Squared Error	0.11	0.11
RASE	Root Average Squared Error	0.34	0.33
DIV	Divisor for ASE	41998.00	18002.00
DFT	Total Degrees of Freedom	20999.00	.

- Misclassification rate :16%
- Maximum absolute error: 0.99
- Sum of squared errors (SSE):
  - 4,725.49 for training; 2,016.10 for validation
- Average squared error (ASE): 0.11
- Root average squared error (RASE):
  - 0.34 for training; 0.33 for validation

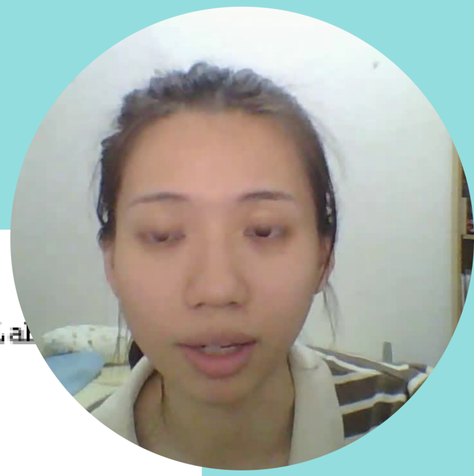
Event Classification Table

Data Role=TRAIN Target=booking_status Target Lab

False Negative	True Negative	False Positive	True Positive
1995	12797	1332	4875

Data Role=VALIDATE Target=booking_status Target Label=' '

False Negative	True Negative	False Positive	True Positive
851	5479	577	2094

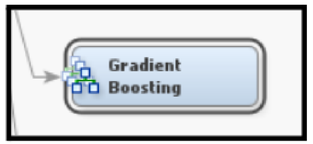


Accuracy

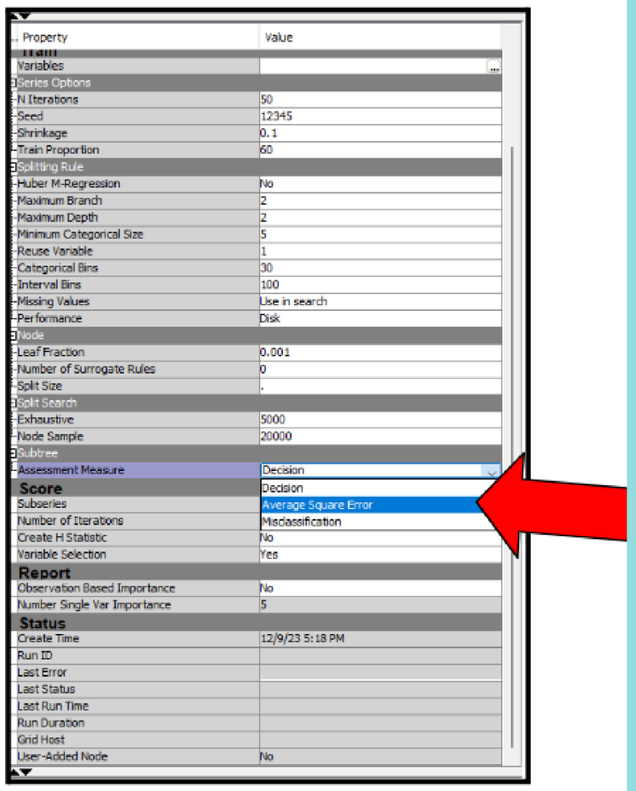
- correctly predicting instances where bookings were not cancelled (Target 0).
  - 85.26% in training, 85.43% in validation
- correctly identifying instances where bookings were cancelled (Target 1)
  - 81.42% in training and 81.13% in validation.

GRADIENT BOOSTING

Node Name :	Gradient Boosting	Configuration

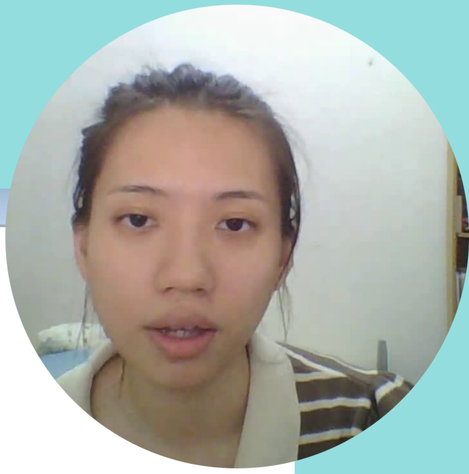


This node uses a partitioning algorithm described in "Greedy Function Approximation: A Gradient Boosting Machine," and "Stochastic Gradient Boosting" by Jerome Friedman. A partitioning algorithm is one that searches for an optimal partition of the data defined in terms of the values of a single variable.



Property	Value
Train	
Variables	
Series Options	
N Iterations	50
Seed	12345
Shrinkage	0.1
Train Proportion	60
Splitting Rule	
Huber M-Regression	No
Maximum Branch	2
Maximum Depth	2
Minimum Categorical Size	5
Reuse Variable	1
Categorical Bins	30
Interval Bins	100
Missing Values	Use in search
Performance	Disk
Node	
Leaf Fraction	0.001
Number of Surrogate Rules	0
Split Size	.
Split Search	
Exhaustive	5000
Node Sample	20000
Subtree	
Assessment Measure	Decision
Score	Decision
Subseries	Average Square Error
Number of Iterations	Misclassification
Create H Statistic	No
Variable Selection	Yes
Report	
Observation Based Importance	No
Number Single Var Importance	5
Status	
Create Time	12/9/23 5:18 PM
Run ID	
Last Error	
Last Status	
Last Run Time	
Run Duration	
Grid Host	
User-Added Node	No

Result:



Output

Fit Statistics

Target=booking_status Target Label=' '

Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	20999.00	9001.00
SUMW	Sum of Case Weights Times Freq	41998.00	18002.00
MISC	Misclassification Rate	0.18	0.18
MAX	Maximum Absolute Error	0.94	0.92
SSE	Sum of Squared Errors	5416.54	2304.31
ASE	Average Squared Error	0.13	0.13
RASE	Root Average Squared Error	0.36	0.36
DIV	Divisor for ASE	41998.00	18002.00
DFT	Total Degrees of Freedom	20999.00	.

- Misclassification Rate: 18%.
- Maximum Absolute Error: 0.94 for training; 0.92 for validation.
- Sum of Squared Errors (SSE): 5,416.54 for training; 2,304.31 for validation
- Average Squared Error (ASE): 0.13
- Root Average Squared Error (RASE): 0.36
- Effective Sample Size: is 41,998.00 for training; 18,002.00 for validation.

GRADIENT BOOSTING

Classification Table

Data Role=TRAIN Target Variable=booking_status Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	83.6635	90.8698	12839	61.1410
1	0	16.3365	36.4920	2507	11.9387
0	1	22.8197	9.1302	1290	6.1431
1	1	77.1803	63.5080	4363	20.7772

Data Role=VALIDATE Target Variable=booking_status Target Label=' '

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	83.9082	91.1823	5522	61.3487
1	0	16.0918	35.9593	1059	11.7654
0	1	22.0661	8.8177	534	5.9327
1	1	77.9339	64.0407	1886	20.9532

Training Dataset:

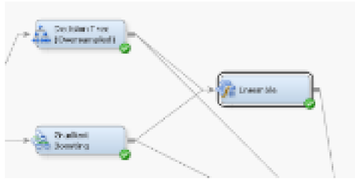
- 83.66% accuracy in correctly predicting non-cancelled bookings (Target 0)
- 77.18% accuracy in predicting cancelled bookings (Target 1).
- The model struggled to differentiate between non-cancelled (Outcome 0) and cancelled bookings (Outcome 1) in both training and validation datasets, resulting in lower accuracies for Target 1 in the non-cancelled category

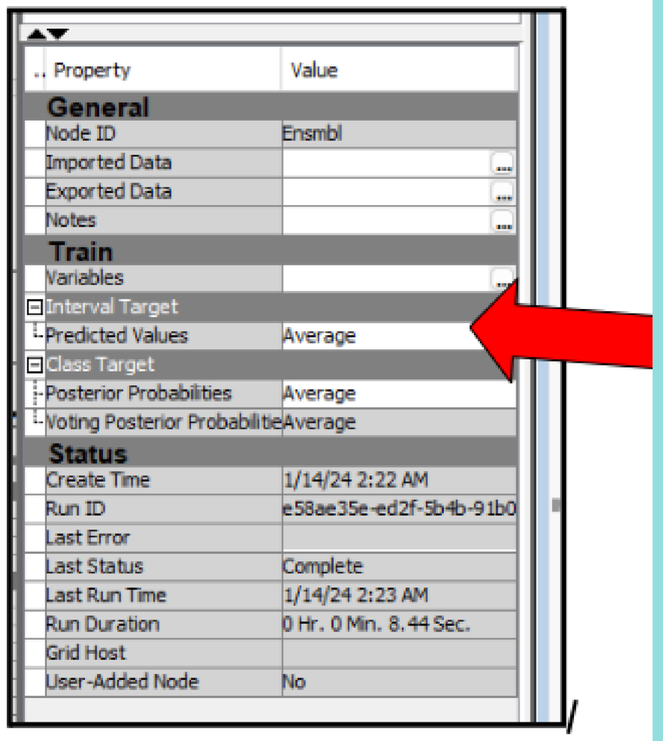
Validation Dataset:

- 83.91% accuracy for correctly predicting non-cancelled bookings (Target 0)
- 77.93% accuracy for predicting cancelled bookings (Target 1).
- The misclassification rates remained consistent at 16.09% for non-cancelled bookings and 22.07% for cancelled bookings in the validation dataset.

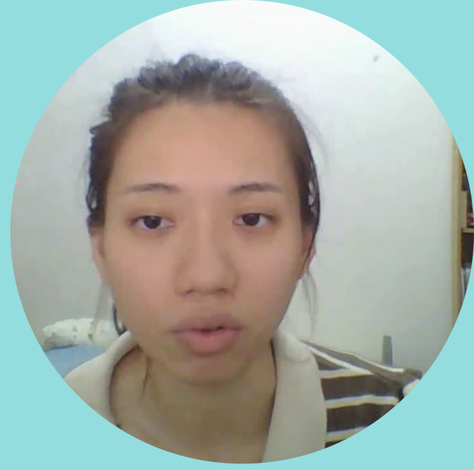
ENSEMBLE

Node Name:	Ensemble	Configuration





Property	Value
General	
Node ID	Ensmbl
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Interval Target	
Predicted Values	Average
Class Target	
Posterior Probabilities	Average
Voting Posterior Probabilitie	Average
Status	
Create Time	1/14/24 2:22 AM
Run ID	e58ae35e-ed2f-5b4b-91b0
Last Error	
Last Status	Complete
Last Run Time	1/14/24 2:23 AM
Run Duration	0 Hr. 0 Min. 8.44 Sec.
Grid Host	
User-Added Node	No



Result:

Fit Statistics

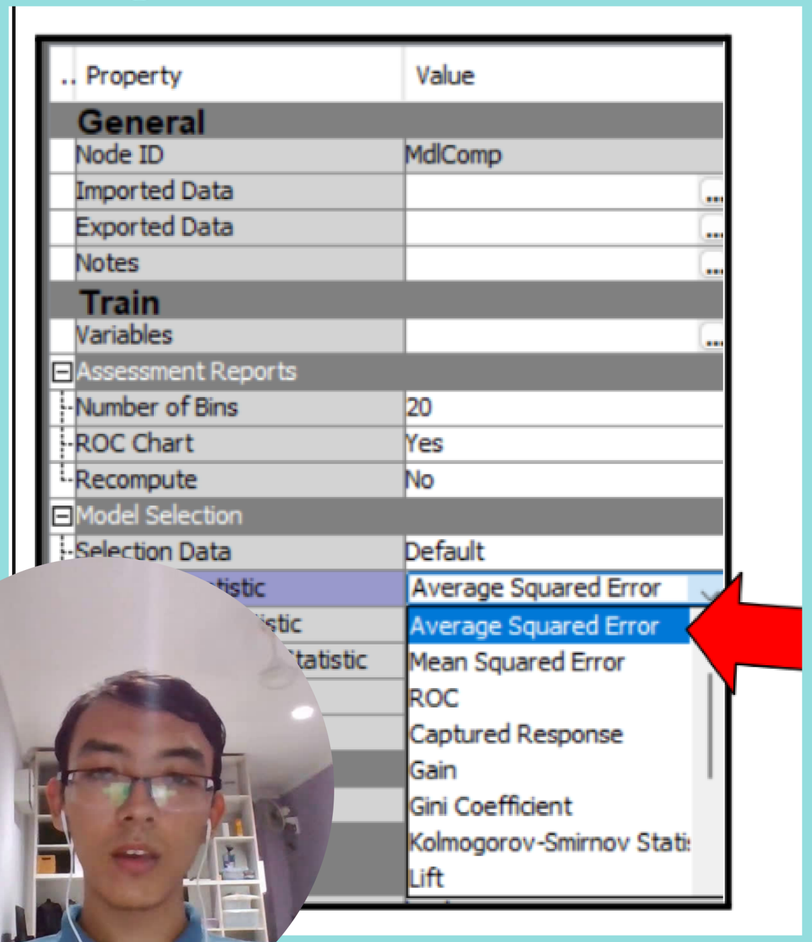
Target=booking_status Target Label=' '

Fit Statistics	Statistics Label	Train	Validation
ASE	Average Squared Error	0.13	0.13
DIV	Divisor for ASE	41998.00	18002.00
MAX	Maximum Absolute Error	0.89	0.88
NOBS	Sum of Frequencies	20999.00	9001.00
RASE	Root Average Squared Error	0.36	0.36
SSE	Sum of Squared Errors	5401.23	2293.98
DISF	Frequency of Classified Cases	20999.00	9001.00
MISC	Misclassification Rate	0.16	0.16
WRONG	Number of Wrong Classifications	3394.00	1446.00

Stage 5 - Asses

- the model is evaluated for how useful and reliable it is for the studied topic.

Configuration



Fit Statistics

Model Selection based on Valid: Average Squared Error (_VASE_)

Selected Model	Model Node	Model Description	Valid: Average Squared Error	Train: Average Squared Error	Train: Misclassification Rate	Valid: Misclassification Rate
Y	Tree	Decision Tree	0.11199	0.11252	0.15777	0.15743
	Boost	Gradient Boosting	0.12800	0.12897	0.18082	0.17698
	Reg	Regression	0.22014	0.22013	0.32716	0.32719

Report

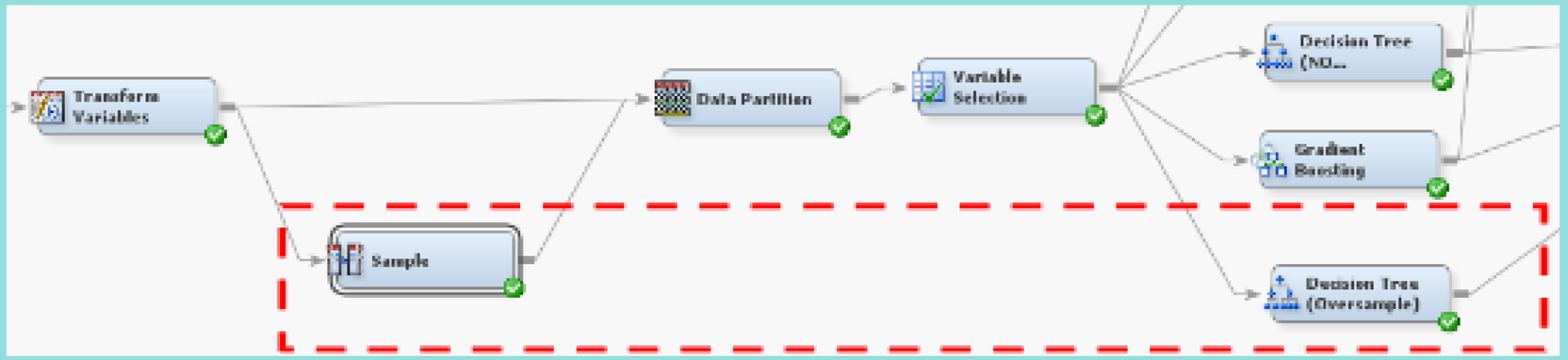
Selected Model

Target	booking_status
Model Node	Tree
Model Description	Decision Tree
Selection Criteria	Valid: Average Squared Error

Decision tree was chosen as the **final predictive model**

OVERFITTING & CLASS IMBALANCE

- To address the overfitting & class imbalance, a Sample node is added and configured to oversampling.



Configuration:

Train	
Variables	...
Output Type	Data
Sample Method	Stratify
Random Seed	12345
Size	
Type	Computed
Observations	.
Percentage	10.0
Alpha	0.01
PValue	0.01
Cluster Method	Random
Stratified	
Criterion	Equal
Ignore Small Strata	No
Minimum Strata Size	5
Level Based Options	
Level Selection	Event
Level Proportion	100.0
Sample Proportion	50.0
Oversampling	
Adjust Frequency	Yes
Based on Count	Yes
Exclude Missing Levels	No

Configuration for Oversampling in Sample Node

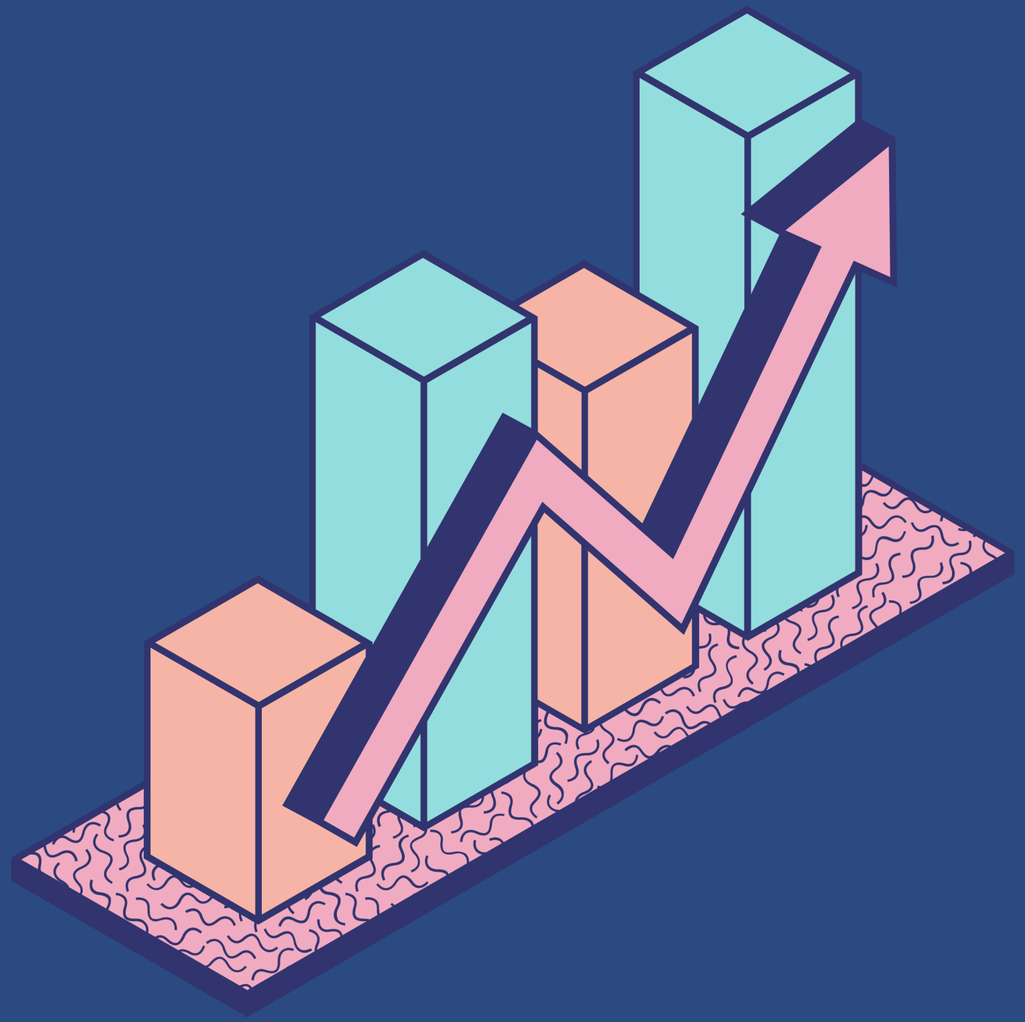
Result:

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Train: Average Squared Error
Y	Tree	Tree	Decision Tree (Oversampled)	booking_st...	0.114707
	Tree2	Tree2	Decision Tree (NO Oversample)	booking_st...	0.116244
	Neural	Neural	Neural Network	booking_st...	0.129993
	Boost	Boost	Gradient Boosting	booking_st...	0.12863
	Ensmbl	Ensmbl	Ensemble	booking_st...	0.128607
	Reg	Reg	Regression	booking_st...	0.220126

The oversampled data has performed better than the data without oversampled and it has proven that we successfully addressed the overfitting and class imbalance.

5
Conclusion



Key Findings

BOOKING PATTERNS AND CUSTOMER PREFERENCES

- Observe distinct patterns in booking behavior, shedding light on the factors influencing successful reservations.
- Understanding customer preferences to tailor services to meet guest expectations.

FACTORS INFLUENCING CANCELLATIONS

- Insights gained from the dataset offer a foundation for proactive measures and improved revenue management

Recommendations & Future Work

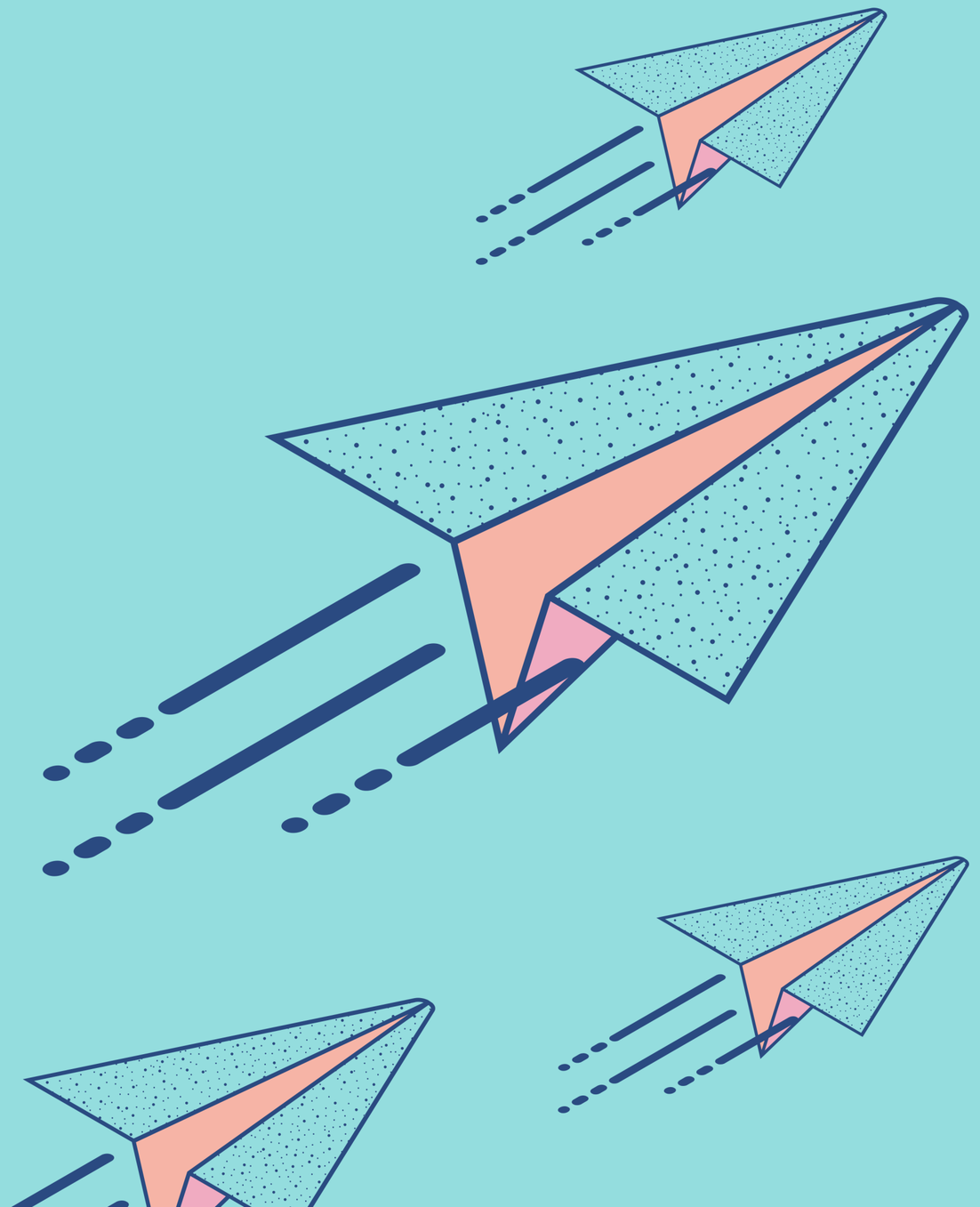
DYNAMIC PRICING STRATEGIES

Investigate the implementation of dynamic pricing strategies based on historical booking patterns and customer behavior.

PERSONALIZATION IN MARKETING

Explore avenues for further personalization in marketing campaigns to target specific customer segments effectively.

REAL-TIME MONITORING

Implement real-time monitoring systems to promptly identify potential cancellations a
preventive actions.

6
References