INDIANA UNIVERSITY
BLOOMINGTON

INDIANA UNIVERSITY BICENTENNIAL

Crossroads Classic Analytics Challenge

-Team Zenith

Workflow

- Introduction and Problem Statement
- Data Description and Preprocessing
- Exploratory Data Analysis
- Model Building and Hyper Parameters Tuning
- Summary

Introduction and Problem Statement

- It is a struggle for NFL teams face on game days without having an understanding of which of the fans will be attending the game.

- Thus, given the data of attendees for the last 2 seasons for the 8 matches in a season, the ask to develop and train a model to accurately predict the number of actual attendees for Colts games based on a variety of factors.

Data Description & Pre-Processing

Data Description

```
RangeIndex: 1158228 entries, 0 to 1158227
Data columns (total 34 columns):
 #   Column           Non-Null Count    Dtype
---  ------           --------------    -----
 0   acct_id          1158228 non-null  object
 1   acct_type_desc   1074743 non-null  object
 2   event_name       1158228 non-null  object
 3   event_date       1158228 non-null  object
 4   plan_event_name  906065 non-null   object
 5   comp_name        1084292 non-null  object
 6   section_name     1158228 non-null  int64
 7   row_name         1158228 non-null  object
 8   SeatNum          1158228 non-null  int64
 9   price_code       1158228 non-null  object
 10  PC1              1158228 non-null  object
 11  Price            1158228 non-null  float64
 12  paid             1018942 non-null  object
 13  add_datetime     1084292 non-null  object
 14  class_name       1158228 non-null  object
 15  status           1158228 non-null  object
 16  Sales_Source     5578 non-null     float64
 17  isHost           1158228 non-null  int64
 18  SeatType         1158228 non-null  object
 19  TicketClass      1158228 non-null  object
 20  Start Year       126519 non-null   float64
 21  LastYear         126519 non-null   float64
 22  Term             126411 non-null   float64
 23  TicketType       1158228 non-null  object
 24  SeatUniqueID     1158228 non-null  object
 25  Season           1158228 non-null  int64
 26  ClubExpYear      126519 non-null   float64
 27  Tenure           1074743 non-null  float64
 28  UniqueID         1158228 non-null  object
 29  isAttended       1158228 non-null  object
 30  Resold           135466 non-null   object
 31  ResalePrice      134496 non-null   float64
 32  ResaleDate       135466 non-null   object
 33  isSTM            348509 non-null   float64
dtypes: float64(9), int64(4), object(21)
memory usage: 300.4+ MB
```

Training Data
(1158228, 34)

Test Data
(128688, 33)

```
RangeIndex: 128688 entries, 0 to 128687
Data columns (total 33 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   acct_id          128688 non-null  object
 1   acct_type_desc   120755 non-null  object
 2   event_name       128688 non-null  object
 3   event_date       128688 non-null  object
 4   plan_event_name  101679 non-null  object
 5   comp_name        121980 non-null  object
 6   section_name     128688 non-null  int64
 7   row_name         128688 non-null  object
 8   SeatNum          128688 non-null  int64
 9   price_code       128688 non-null  object
 10  PC1              128688 non-null  object
 11  Price            128688 non-null  int64
 12  paid             113829 non-null  object
 13  add_datetime     121980 non-null  object
 14  class_name       128688 non-null  object
 15  status           128688 non-null  object
 16  Sales_Source     634 non-null     float64
 17  isHost           128688 non-null  int64
 18  SeatType         128688 non-null  object
 19  TicketClass      128688 non-null  object
 20  Start Year       14072 non-null   float64
 21  LastYear         14072 non-null   float64
 22  Term             14060 non-null   float64
 23  TicketType       128688 non-null  object
 24  SeatUniqueID     128688 non-null  object
 25  Season           128688 non-null  int64
 26  ClubExpYear      14072 non-null   float64
 27  Tenure           120755 non-null  float64
 28  UniqueID         128688 non-null  object
 29  Resold           10582 non-null   object
 30  ResalePrice      10429 non-null   float64
 31  ResaleDate       10582 non-null   object
 32  isSTM            38303 non-null   float64
dtypes: float64(8), int64(5), object(20)
memory usage: 32.4+ MB
```

Data Description

section_name	Term	SeatType	acct_type_desc	PC1
SeatNum	Season	UniqueID	Resold	plan_event_name
Price	ClubExpYear	class_name	SeatUniqueID	add_datetime
Sales_Source	Tenure	acct_id	TicketClass	ResaleDate
isHost	ResalePrice	comp_name	price_code	event_date
Start Year	isSTM	TicketType	status	section_name
LastYear	event_name	row_name	paid–	isAttended (Target)

Number of numerical columns: 13
Number of categorical columns: 21

Data Cleaning

- Dropped repeated and extra information:
  - Acct_id, Sales_Source, SeatNum, UniqueID
  - SeatUniqueID, Price_code,
  - ClubExpYear, LastYear, Start Year,

- Comp_name - reduced category to Not complimentary and others
- Paid - Y, P or N

- Fill NaN with 0

- Label encoder - categorical columns with cardinality <10.
- Count encoder - categorical columns with cardinality >= 10.

Feature Engineering

- Extracted
  - Months from Event date
  - Weekday or Weekend from the Event Date (new column as:“month” and “day”)
- Days between event_date - resale_date or
  - event_date - add_datetime (new column as: “dates”)
- Difference between Price and ResalePrice (new column as: “price_diff”)

	event_date	month	day
0	2021-10-17	10	1
1	2021-11-14	11	1
2	2021-09-19	9	1
3	2021-12-18	12	1
4	2021-11-04	11	0
5	2021-08-15	8	1

Synthesized Columns Snapshot

ResaleDate > event_date : 14571 rows

	month	day	event_date	ResaleDate	add_datetime	Price	dates	ResalePrice	price_diff
0	10	1	2021-10-17	NaT	2021-03-31 16:08:52	111.0	199.0	NaN	NaN
1	10	1	2021-10-17	NaT	2021-03-31 16:08:19	111.0	199.0	NaN	NaN
2	10	1	2021-10-17	NaT	2021-03-31 16:10:20	111.0	199.0	NaN	NaN
3	10	1	2021-10-17	NaT	2021-03-31 16:08:21	111.0	199.0	NaN	NaN
4	10	1	2021-10-17	NaT	2021-04-20 16:03:18	0.0	179.0	NaN	NaN
5	10	1	2021-10-17	NaT	2021-03-31 16:09:01	111.0	199.0	NaN	NaN
6	10	1	2021-10-17	2021-11-30 00:31:19	2021-03-31 16:08:07	111.0	-45.0	272.0	-161.0
7	10	1	2021-10-17	2021-09-12 09:17:22	2021-03-31 16:08:07	111.0	34.0	473.6	-362.6
8	10	1	2021-10-17	2021-11-04 16:02:58	2021-03-31 16:08:07	111.0	-19.0	72.0	39.0
9	10	1	2021-10-17	NaT	2021-03-31 16:08:07	111.0	199.0	NaN	NaN

Extending Beyond

2D virtual Tour of seating arrangements of the Stadium.

Source: https://www.colts.com/game-day/stadium-maps

Extending Beyond

- Fetched the stadium's seating information from the internet and mapped the section_name to Section categories name:
  - Terrace - (500 and 600 levels)
  - Loge - (300 and 400 levels)
  - Street - (100 and 200 levels)

(new column as:"Section_category")

Source: https://blog.ticketiq.com/blog/lucas-oil-stadium-seating-chart

Extending Beyond

- section_name to a Section type namely:
  - Center
  - Corner
  - End

(new column as:“Section_type”)

Source:https://blog.ticketiq.com/blog/lucas-oil-stadium-seating-chart

Revised Data Description

Price	Term	SeatType	acct_type_desc	PC1
ResalePrice	Season	month	Resold	plan_event_name
price_diff	Tenure	class_name	dates	Section_category
section_name	isSTM	day	TicketClass	section_type
isHost	event_name	comp	price_code	TicketType
row_name	paid	status	section_name	isAttended

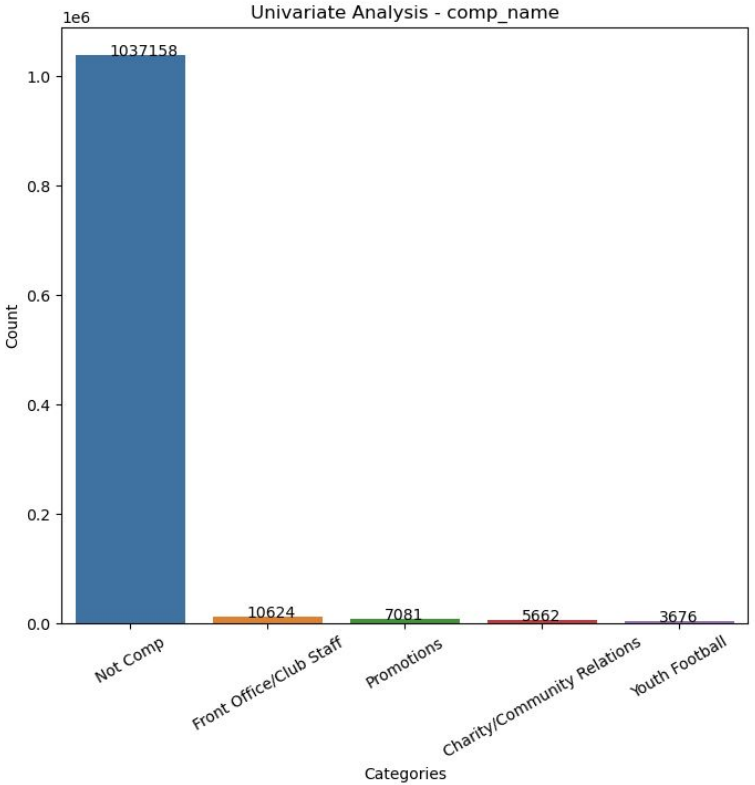
Number of numerical columns: 3
Number of categorical columns: 26

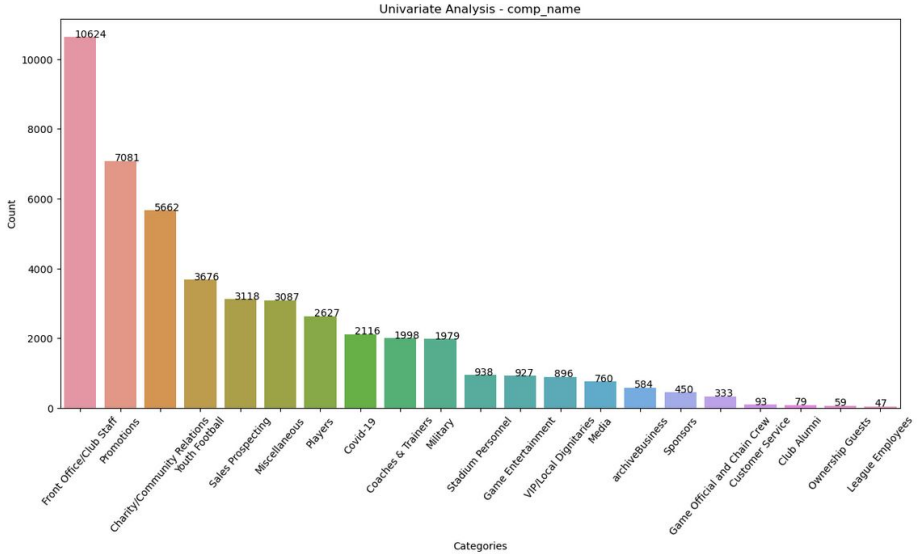
Workflow

- Introduction and Problem Statement
- Data Description and Preprocessing
- Exploratory Data Analysis
- Model Building and Hyper Parameters Tuning
- Summary

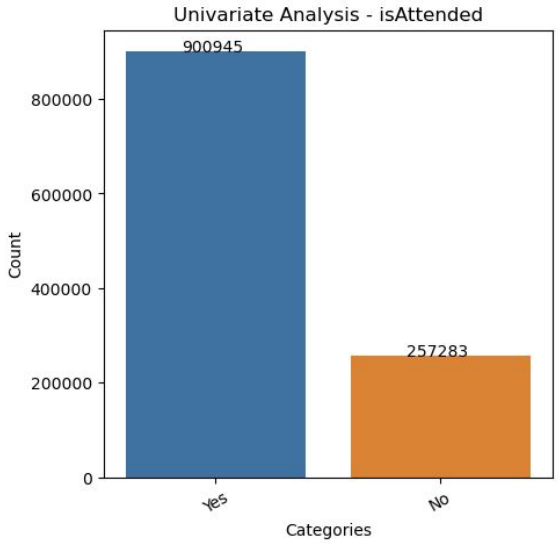
Exploratory Data Analysis

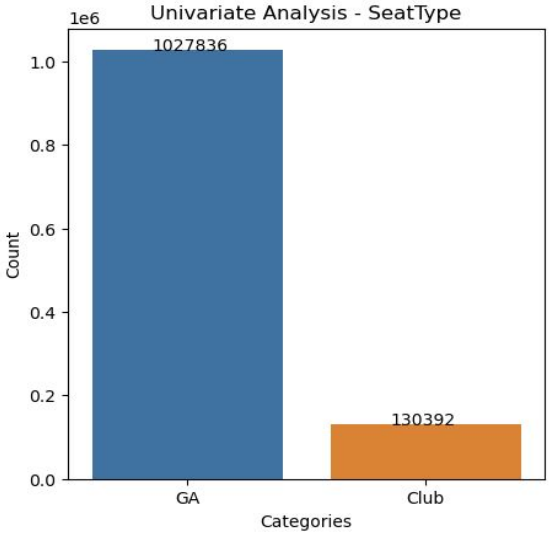
Univariate Analysis - comp_name

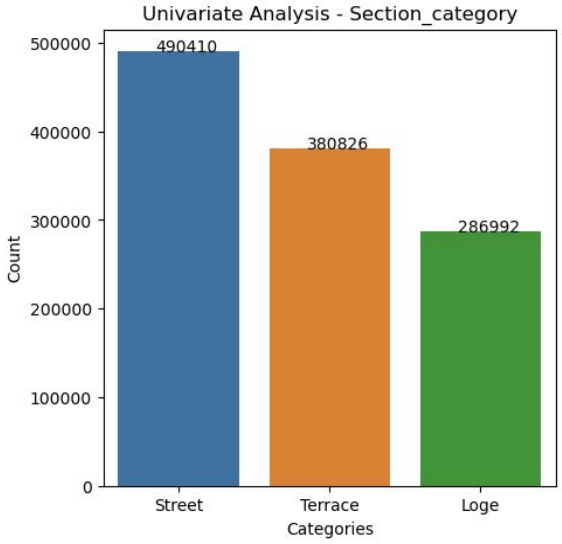




Univariate Analysis

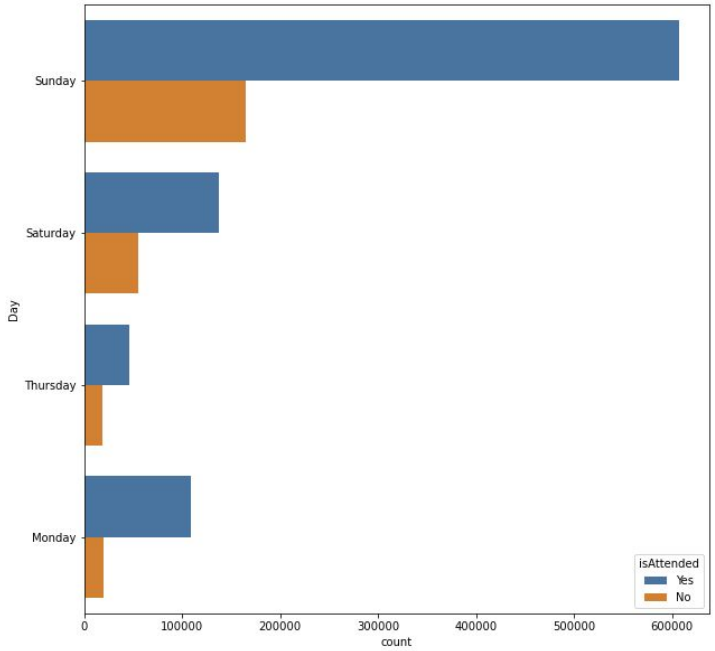






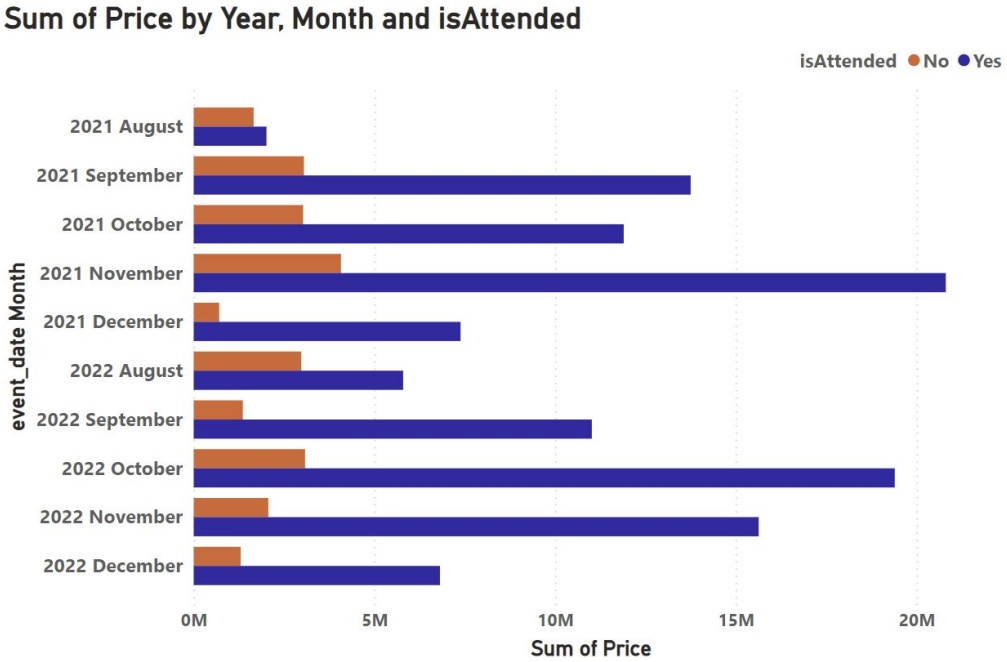
isAttended **SeatType** **Section_category**

Bivariate Analysis



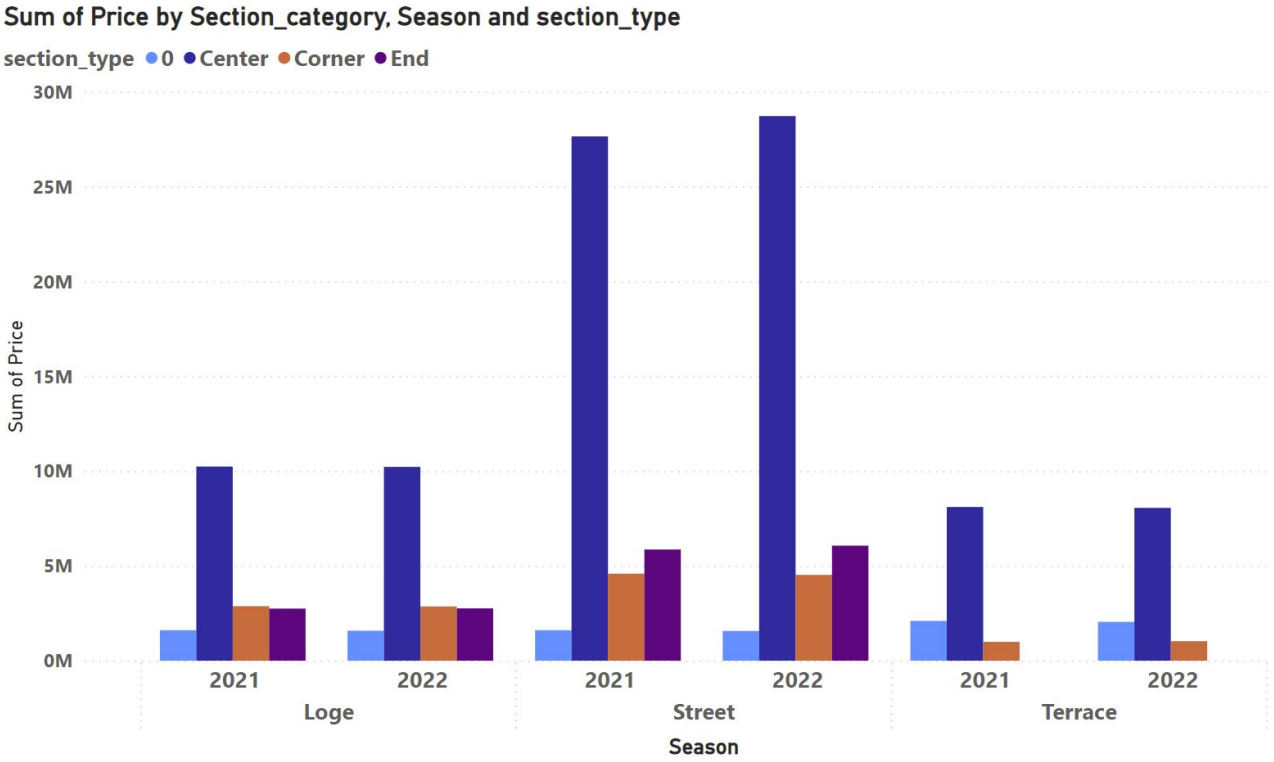
isAttended column against Day

Bar Plots



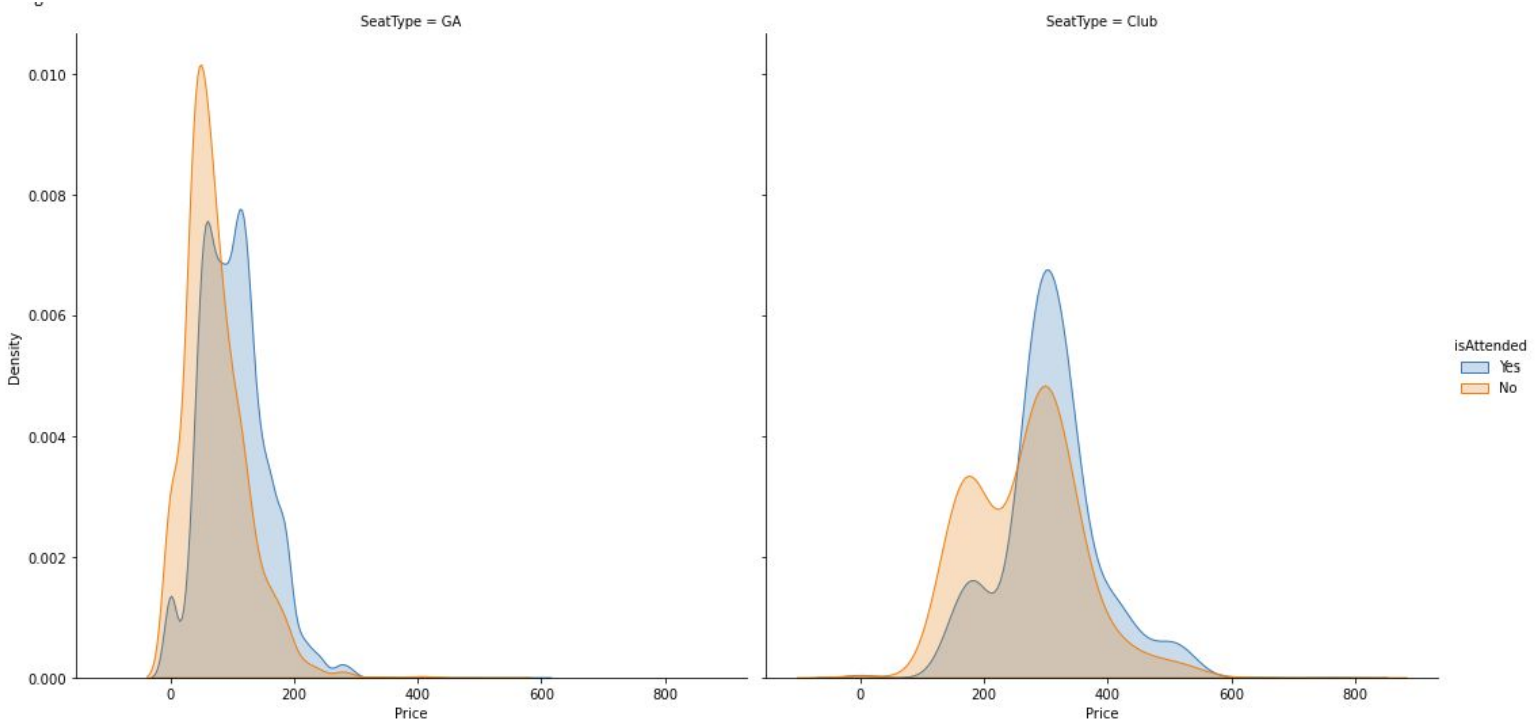
Explains the Prices vs Event_date (Month-wise) with isAttended as Legend

Bar Plots

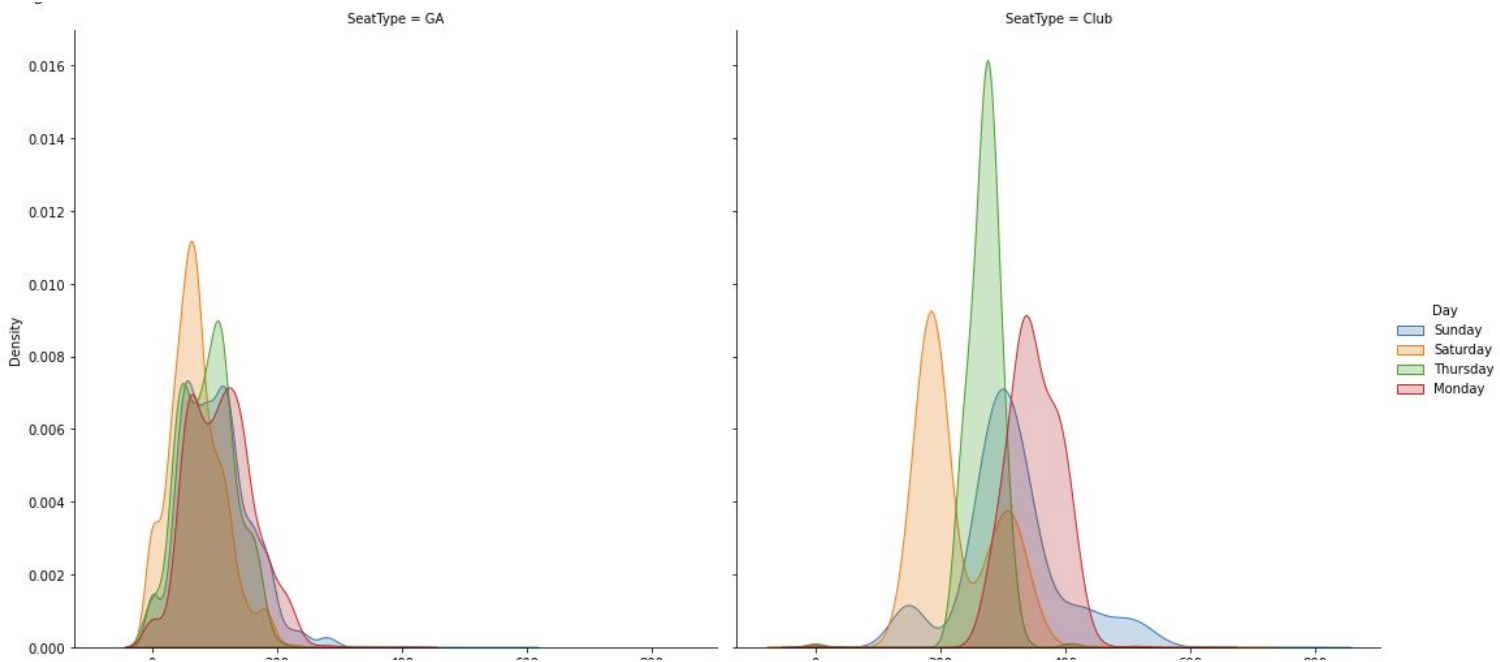


Barplots for Seasons vs Section _Category vs SectionType against the Price

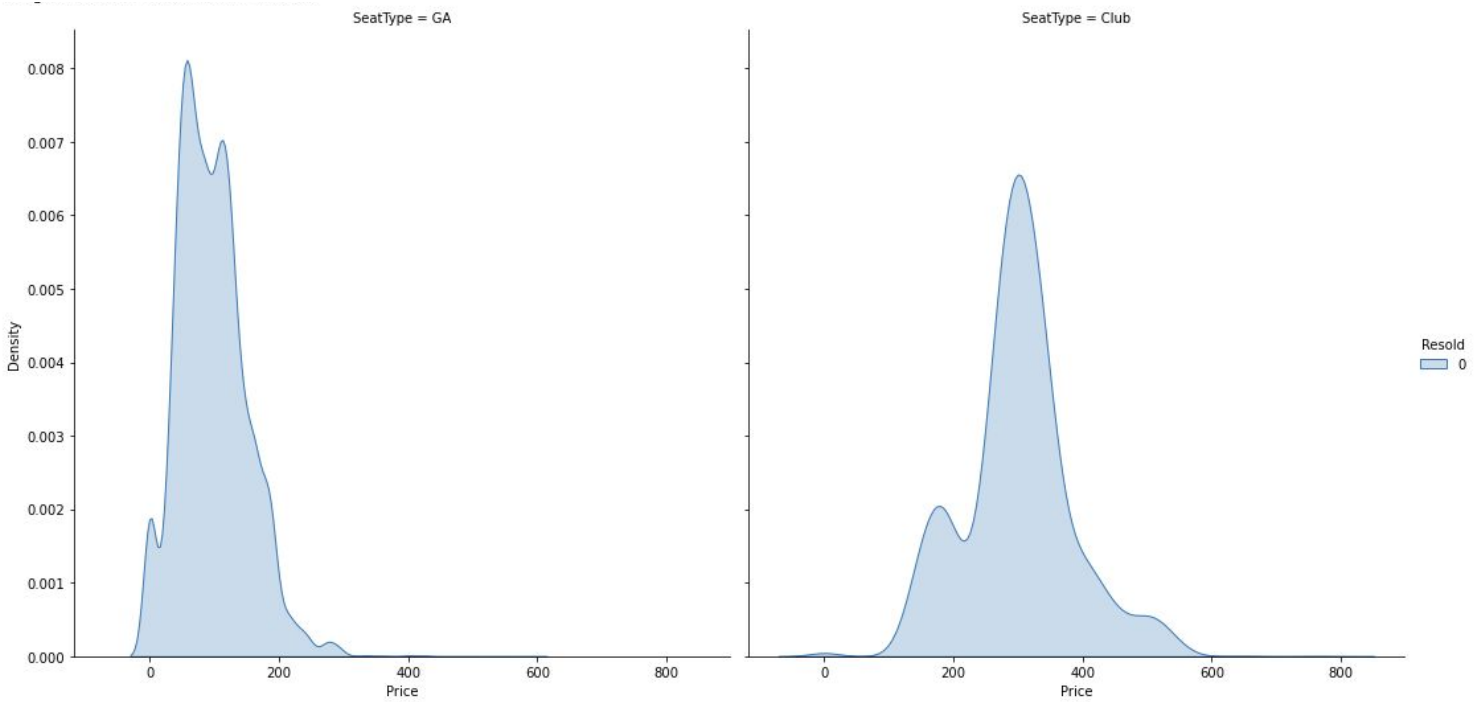
Density plot - Price vs SeatType



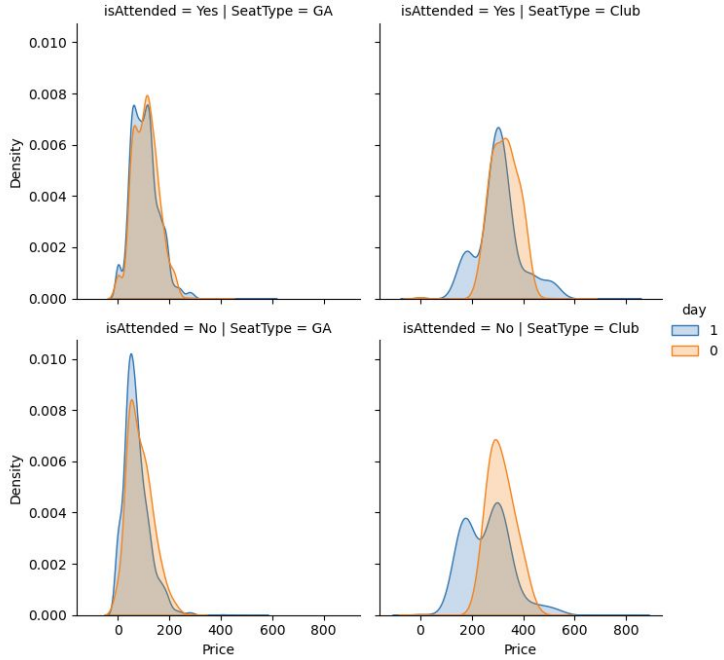
Density plot - Price vs SeatType for Day



Density plot - Price vs SeatType for Resold Price



Density plot - Day vs SeatType vs isAttended for Price



Workflow

- Introduction and Problem Statement
- Data Description and Preprocessing
- Exploratory Data Analysis
- Model Building and Hyper Parameters Tuning
- Summary

Model Building - Baseline Models

Default Parameters

Implementation and Comparison

- Dropped repetitive information: No
- Fabricated new features: No
- Hyper parameter values: Default
- Model Selection: Random Split

Model	Validation Accuracy
Decision Tree	80.72%
Random Forest	83.47%
XGBoost	84.63%
CatBoost	84.42%

Model Building - Optimized Models

Hyperparameter Tuning

Enhanced Strategies

- Feature Engineering: Yes
- Dropped repetitive columns: Yes
- Hyper parameter values: Tuned parameters as per the models
- Model Selection: Stratified 5 fold Cross Validation

Hyperparameter Tuning and Optimization

- RandomSearch, GridSearch, and BayesianSearch CV for Hyperparameter tuning and Optimization
- Best Parameters:

CatBoost:
max_depth= 12,
iterations =1000,
learning_rate= 0.1,
early_stopping_rounds = 8

Random Forest:
n_estimators=100,
max_depth=250,
max_features =auto

XGBoost:
gamma = 1,
reg_lambda =1,
max_depth = 10

Model Evaluation

Model	Cross Validation Accuracy	Test Data Accuracy	F1 Score*
Random Forest	74.4%	80.2%	0.91
XGBoost	72.9%	78%	0.90
CatBoost	73.6%	79.4%	0.89

* F1 score = 2*precision*recall/(precision + recall)

Summary

- Engineered 6 new features: day, **month**, **price_diff**, section_category, section_type, **dates**
- Encoded data using label encoding and frequency encoding
- Tuned hyperparameters using GridSearchCV
- Best Model: Random Forest with parameters max_depth = 100, n_estimators = 250 gave 0.8021 test accuracy and **0.91 F1 score**
- Chances of improvement
  a. Better hyperparameter tuning
  b. Transfer Deep Learning

Thank You

Team Zenith

Shachi Chaugule
Rajesh Sharma
Ashish Patidar