Thomas Dinh

DSE 6211

Project 1 Analytic Plan

September 7, 2023

Analytic Plan: Leveraging Neural Networks for classification on hotel cancellation

1. Define the Business Need:

- ABC Hotels is looking to identify Bookings that have a high risk of cancellation.

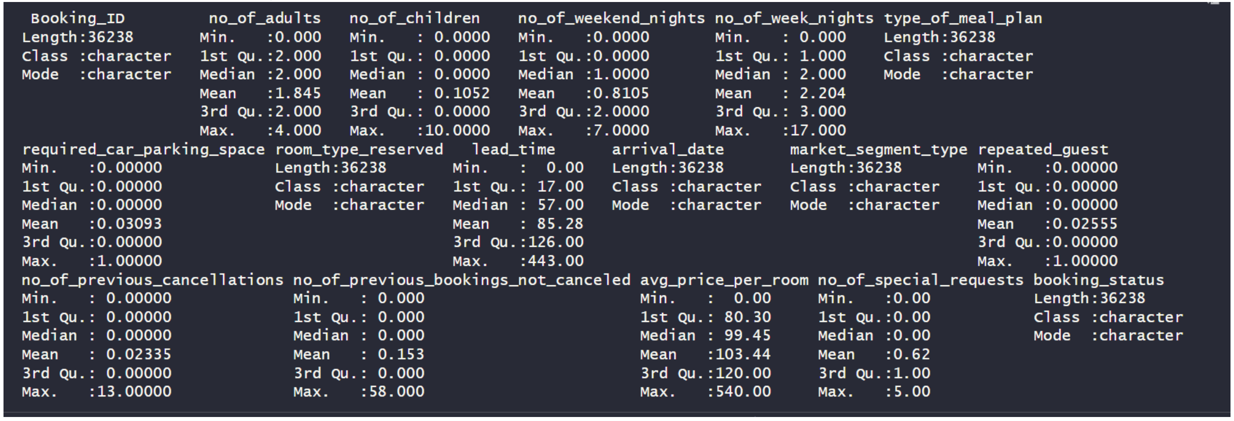
2. Problem Statement:

- Since this is a supervised classification task, we find the risk of cancellation between 0 and 1. (0 is No for cancellation, 1 is Yes for cancellation)
  - Booking Status is our dependent value we are looking for, while certain columns will be used to as the independent variables to predict for booking status

3. Data Understanding:

- The data sets dimensions are (36238, 17)
- 36348 data points and 17 variables
- No missing values
- Column Names:
  - "Booking_ID", "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_date", "market_segment_type", "repeated_guest", "no_of_previous_cancellations", "no_of_previous_bookings_not_canceled", "avg_price_per_room", "no_of_special_requests", "booking_status"

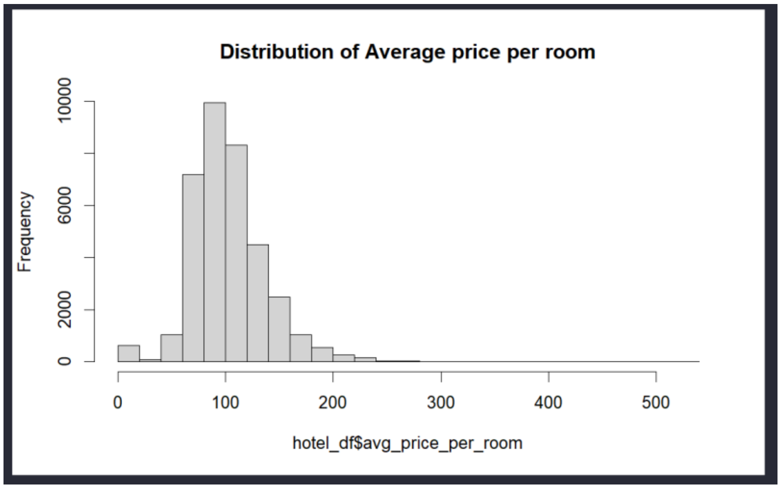
Hotel Data Summary:

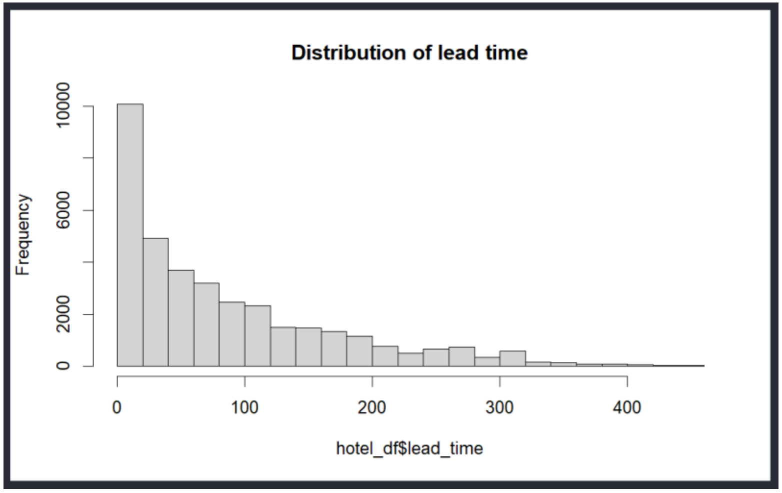


```
  Booking_ID         no_of_adults   no_of_children    no_of_weekend_nights no_of_week_nights type_of_meal_plan
 Length:36238       Min.   :0.000   Min.   : 0.0000   Min.   :0.0000       Min.   : 0.000    Length:36238
 Class :character   1st Qu.:2.000   1st Qu.: 0.0000   1st Qu.:0.0000       1st Qu.: 1.000    Class :character
 Mode  :character   Median :2.000   Median : 0.0000   Median :1.0000       Median : 2.000    Mode  :character
                    Mean   :1.845   Mean   : 0.1052   Mean   :0.8105       Mean   : 2.204
                    3rd Qu.:2.000   3rd Qu.: 0.0000   3rd Qu.:2.0000       3rd Qu.: 3.000
                    Max.   :4.000   Max.   :10.0000   Max.   :7.0000       Max.   :17.000
 required_car_parking_space room_type_reserved   lead_time      arrival_date       market_segment_type repeated_guest
 Min.   :0.00000            Length:36238       Min.   :  0.00   Length:36238       Length:36238        Min.   :0.00000
 1st Qu.:0.00000            Class :character   1st Qu.: 17.00   Class :character   Class :character    1st Qu.:0.00000
 Median :0.00000            Mode  :character   Median : 57.00   Mode  :character   Mode  :character    Median :0.00000
 Mean   :0.03093                               Mean   : 85.28                                          Mean   :0.02555
 3rd Qu.:0.00000                               3rd Qu.:126.00                                          3rd Qu.:0.00000
 Max.   :1.00000                               Max.   :443.00                                          Max.   :1.00000
 no_of_previous_cancellations no_of_previous_bookings_not_canceled avg_price_per_room no_of_special_requests booking_status
 Min.   : 0.00000             Min.   : 0.000                       Min.   :  0.00     Min.   :0.00           Length:36238
 1st Qu.: 0.00000             1st Qu.: 0.000                       1st Qu.: 80.30     1st Qu.:0.00           Class :character
 Median : 0.00000             Median : 0.000                       Median : 99.45     Median :0.00           Mode  :character
 Mean   : 0.02335             Mean   : 0.153                       Mean   :103.44     Mean   :0.62
 3rd Qu.: 0.00000             3rd Qu.: 0.000                       3rd Qu.:120.00     3rd Qu.:1.00
 Max.   :13.00000             Max.   :58.000                       Max.   :540.00     Max.   :5.00
```

4. Data Processing:

- The columns with the outliers we will take care of are Lead Time and Average price per room





- 2 way I was thinking about processing the data for outliers is standardizing or using box plots
- Standardization of columns:
    - Required room, # of adults, # of children, # of weekend nights, # of weeknights, lead time, repeated guests, # of previous bookings, # of previous cancellations, # of previous not canceled, average price, # of special requests
- One-hot encoding of columns:
    - Type of meal plan, room type, market segment type, booking segment
- Excluded columns:
    - Date and booking_id
    - booking_id doesn't carry any meaningful information related to the problem, it may not provide any valuable predictive power to the model.
    - Date is a maybe, because I will have to see more graphs and research to see how to use it efficiently

5. Model Selection:

- Feedforward Neural Networks
    - They are versatile and can be adapted to various classification tasks by adjusting their architecture and hyperparameters.
- Activation Functions
    - ReLu
        - Promotes faster training convergence, mitigates vanishing gradient problem, widely used in hidden layers.
    - Sigmoid
        - Output values in the range (0, 1), suitable for binary classification where you want to estimate class probabilities.
- Consideration for model hyperparameters and optimization techniques will adjust accordingly

6. Model Development:

- Split the data 80%/20%
- Multiple types of cross validations but one I will be starting off with is K-fold cross validation
- Combination of Oversampling and Undersampling:
    - I will apply both oversampling and undersampling techniques to balance the dataset. For example, you can use SMOTE to oversample the minority class and random undersampling for the majority class.

7. Evaluation Metrics:

- Accuracy
- Precision
- Recall
- F1-score
- ROC AUC

8. Conclusion:

- Summarize the project's objectives, methods, and expected outcomes.