# Star Hotels Business Case - Analysis

SARAVANAN RANGARAJAN (SR)

# Contents



# Business Problem Overview

A significant number of hotel bookings 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. Such losses are particularly high on last-minute cancellations.

The new technologies involving online booking channels have dramatically changed customers' booking possibilities and behavior. This adds a further dimension to the challenge of how hotels handle cancellations, which are no longer limited to traditional booking and guest characteristics.

The cancellation of bookings impact a hotel on various fronts:

1. Loss of resources (revenue) when the hotel cannot resell the room.
2. Additional costs of distribution channels by increasing commissions or paying for publicity to help sell these rooms.
3. Lowering prices last minute, so the hotel can resell a room, resulting in reducing the profit margin.
4. Human resources to make arrangements for the guests.

# Objective

The increasing number of cancellations calls for a Machine Learning based solution that can help in predicting which booking is likely to be canceled. Star Hotels Group has a chain of hotels in Portugal, they are facing problems with the high number of booking cancellations and have reached out to your firm for data-driven solutions. You as a data scientist have to analyze the data provided to find which factors have a high influence on booking cancellations, build a predictive model that can predict which booking is going to be canceled in advance, and help in formulating profitable policies for cancellations and refunds.

# Solution Approach

Per the requirement is to perform Model building and perform Logistic Regression and provide actionable insights.

Following are the steps to perform

1. Understand Dataset
2. Sanity Check
3. Missing values treatment
4. Outliers Detection and Treatment
5. EDA
6. Test & Train Model
7. Perform Logistic  Regression
8. Performance of model Validation
9. Test for MultiCollinearity
10. Check VIF
11. Optimize threshold
12. Decision Tree analysis
13. Try to reduce overfitting
14. Recommendations

# Data Overview

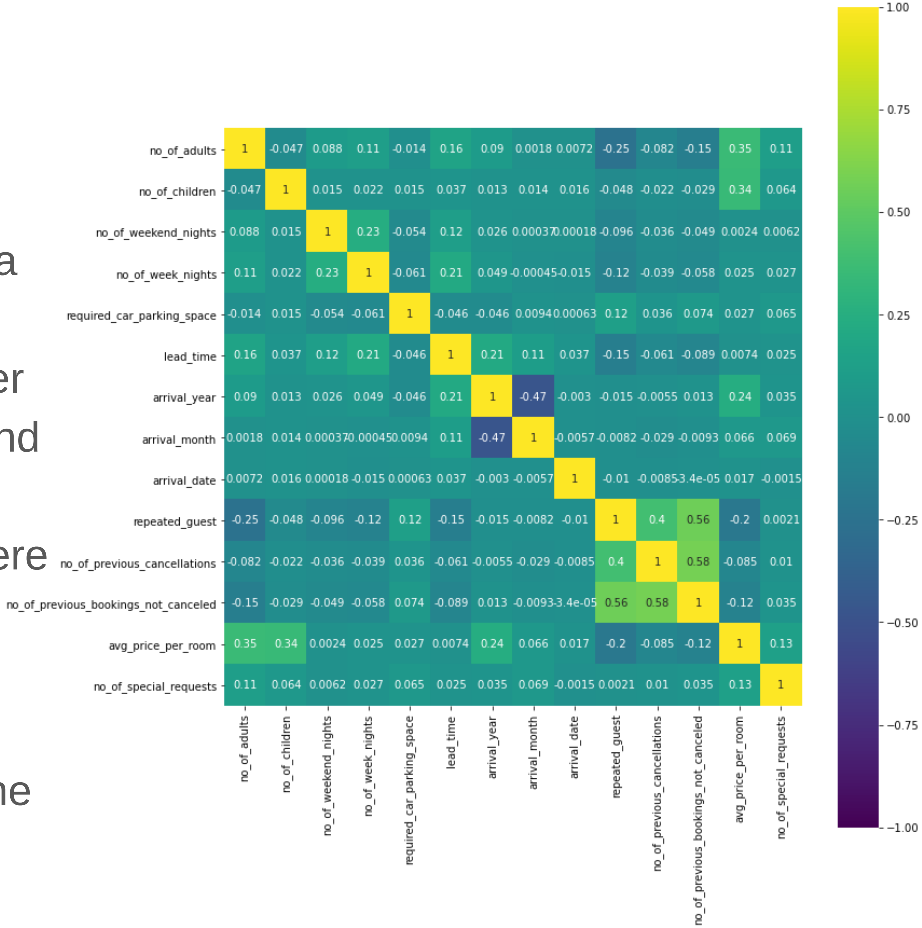
- no_of_adults: Number of adults
- no_of_children: Number of Children
- no_of_weekend_nights: Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
- no_of_week_nights: Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel
- type_of_meal_plan: Type of meal plan booked by the customer:
  - Not Selected – No meal plan selected
  - Meal Plan 1 – Breakfast
  - Meal Plan 2 – Half board (breakfast and one other meal)
  - Meal Plan 3 – Full board (breakfast, lunch, and dinner)
- required_car_parking_space: Does the customer require a car parking space? (0 - No, 1- Yes)
- room_type_reserved: Type of room reserved by the customer. The values are ciphered (encoded) by Star Hotels.
- lead_time: Number of days between the date of booking and the arrival date
- arrival_year: Year of arrival date
- arrival_month: Month of arrival date
- arrival_date: Date of the month
- market_segment_type: Market segment designation.
- repeated_guest: Is the customer a repeated guest? (0 - No, 1- Yes)
- no_of_previous_cancellations: Number of previous bookings that were canceled by the customer prior to the current booking
- no_of_previous_bookings_not_canceled: Number of previous bookings not canceled by the customer prior to the current booking
- avg_price_per_room: Average price per day of the reservation; prices of the rooms are dynamic. (in euros)
- no_of_special_requests: Total number of special requests made by the customer (e.g. high floor, view from the room, etc)
- booking_status: Flag indicating if the booking was canceled or not.

# Data Overview – Contd.

- It has 14 Numerical values & 4 Categorical values ( Room type, type of meal plan, market segment type, booking status)
- Null checks have been done to check the data. No Null values exist.
- Duplicate checks have been done. Duplicate data exist and it has been removed for model.
- Check the distinct values in the Categorical columns and check the data spread for general information.
- Missing Values has been replaced by Mean values for average price per room.
- One hot Encoding was used for booking status columns.
- Univariate & Bi variate analysis have been done for understanding of data.
- Outlier Treatment was done.
- Initial analysis show Lead time and No of special request may have impact on Cancellation.

# Correlation Analysis

- Highest positive correlation between the number of previous bookings canceled and previous bookings not canceled by a customer and repeated guest.
- Negative correlation between the number of special requests from the customer and the booking status, indicating if a customer had some special requests there is chances of cancellation that may decrease
- Positive correlation between booking status and lead time, indicating higher the lead time higher are the chances of cancellation



# Model Evaluation Criteria

- Predictions can be
  - Predicting a customer will not cancel their booking but, the customer will cancel their booking.
  - Predicting a customer will cancel their booking but, the customer will not cancel their booking.
- In this Business scenario, reducing both the types of losses is important so we would go by checking F1 score (Reduce False Positive and False Negative)

# Model Performance (Logistic Regression)

- Logistic Regression (sklearn)

Training set performance:
Accuracy: 0.7928732006844948
Precision: 0.7320415879017014
Recall: 0.6134046134046134
F1: 0.667492593590089

Test set performance:
Accuracy: 0.790808737179989
Precision: 0.7355689939527212
Recall: 0.6101231190150479
F1: 0.666999002991027

•Coefficients of required_car_parking_space, arrival_month,repeated_guest, no_of_special_requests, room_type_reserved_Room_Type(s), market_segment_type_Offline are negative an increase in mentioned variables will lead to a decrease in chances of a customer canceling their booking.

•Coefficients of other variables are positive an increase in these will lead to a increase in the chances of a customer canceling their booking.

# Logistic Regression (Statsmodel)

| Training | Logistic Regression sklearn | Logistic Regression- 0.31 Threshold | Logistic Regression- 0.41 Threshold |
|---|---|---|---|
| Accuracy | 0.79 | 0.77 | 0.79 |
| Recall | 0.61 | 0.80 | 0.71 |
| Precision | 0.73 | 0.63 | 0.68 |
| F1 | 0.67 | 0.71 | 0.69 |

| Testing | Logistic Regression sklearn | Logistic Regression- 0.31 Threshold | Logistic Regression- 0.41 Threshold |
|---|---|---|---|
| Accuracy | 0.79 | 0.77 | 0.79 |
| Recall | 0.61 | 0.80 | 0.71 |
| Precision | 0.74 | 0.63 | 0.69 |
| F1 | 0.67 | 0.71 | 0.70 |

# Logistic Regression - Keypoints

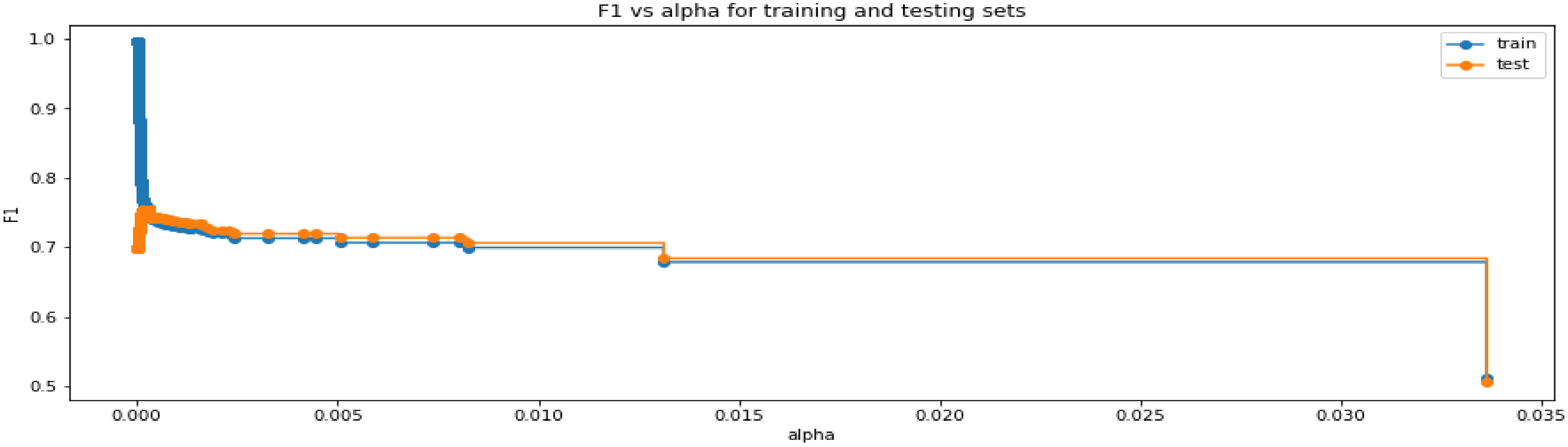
- We have been able to build a predictive model that can be used by the hotel to predict which bookings are likely to be cancelled with an F1 score of 0.69 on the training set.
- The model with default threshold the model will give a low recall but good precision score
- The model with a 0.31 threshold the model will give a high recall but low precision score
- The model with a 0.41 threshold the model will give a balance recall and precision score

# Decision Tress (Pre – Pruning)

- Decision Tree Classifier was used to predict key variables for Business case
- Default model was overfitting
- Decision Tree Grid Search Technique for Hyperparameter tuning approach was used to reduce over fitting
- Following parameters were used
- "max_depth": [5, 10, 15, None],
- "criterion": ["entropy", "gini"],
- "splitter": ["best", "random"],
- "min_impurity_decrease": [0.00001, 0.0001, 0.01],
- F1 Score was used to Train the model and fit.
- Obviously tree has become simpler and it is legible
- Performance has improved.
- Important features as per graph are Lead time, No. of Special Requests, Market Segment Type Online, Average Price per room, arrival month.

# Decision Trees (Post- Pruning)

- Cost Complexity pruning method was performed for post pruning
- CCP Alpha and Impurities were calculated
- In DecisionTreeClassifier, this pruning technique is parameterized by the cost complexity parameter, ccp_alpha. Greater values of ccp_alpha increase the number of nodes pruned. Here we only show the effect of ccp_alpha on regularizing the trees and how to choose a ccp_alpha based on validation scores.



# Decision Trees - Summary

| Train | Tree - Default | Tree - Pre_Prune | Tree - Post_Prune |
| --- | --- | --- | --- |
| Accuracy | 1.00 | 0.83 | 0.82 |
| Recall | 0.99 | 0.78 | 0.80 |
| Precision | 1.00 | 0.74 | 0.72 |
| F1 | 1.00 | 0.76 | 0.76 |

| Test | Tree - Default | Tree - Pre_Prune | Tree - Post_Prune |
| --- | --- | --- | --- |
| Accuracy | 0.79 | 0.82 | 0.82 |
| Recall | 0.69 | 0.77 | 0.80 |
| Precision | 0.69 | 0.73 | 0.72 |
| F1 | 0.69 | 0.75 | 0.75 |

# Decision Trees – Key Points

•After post pruning we can see that it has given better Recall, Precision and F1 Score
•Key variables remain almost same as pre prune.
•Decision Tree model performs better
•Key variables are Lead Time, No of Special Requests, Market Segment Type and Average price per room.

# Business Recommendations

- No of special requests and lead time are pivotal parameters in cancellations.
- Bookings made in less than 150 days prior to arrival has less chances compared to booking made more than 150 days before.
- Any special requests are made, Hotel should have to be more cautious on those reservations
- Also, Holiday months like Christmas, Thanksgiving months have less cancellation, where as summer holidays see more cancellations so Hotel might run some campaigns during that period
- Hotel also might charge cancellation fee for More lead time or special request reservations

# Questions

- Questions give different Perspective, most of the time result in idea.

greatlearning

*Power Ahead*

# Happy Learning !