James Lo
Aditya Bhattacharya
Springboard Data Science
7 August 2023

Capstone 2: Project Proposal

The Problem

The 99 DAPT SAO IH Hotel Booking problem is a Kaggle challenge on predicting hotel booking cancellations with machine learning. Using the Kaggle challenge, the following scenario has been created:

> *For Sunset Pier Hotels and Resorts, booking cancellations have always been a pain point. Cancellations negatively impact the hotel's revenue and bottom line where sometimes the hotel may not be able to book the room again in time. Due to cancellations, the hotel estimates a 14% loss in revenues. While two seasons ago the institution implemented risk mitigation in the form of requiring a booking deposit, management of the hotel suspect that this has dissuaded customers from booking the hotel altogether. Management wants to implement a data-driven approach to their risk management strategy to mitigate cancellations.*

The Client

Sunset Pier Hotels and Resorts.

The Dataset

The datasets are provided on Kaggle. A training set and validation set are included.

Columns include type of hotel, binary variable indicating reservation cancellation (0 = not cancelled, 1 = cancelled), number of days between reservation date and hotel check-in date (when the reservation was made), date of arrival at the hotel (for when the reservation is), number of weekend days in the reservation, number of weekdays in the reservation, number of adults, number of children, number of babies, type of meal included in the reservation, country of the customer, customer marketing segmentation, sales channel through which the reservation was made, has the customer already stayed at the hotel? (0 = no, 1 = yes), how many reservations the customer has canceled in the past, how many bookings the customer has made and not canceled in the past, desired room type, type of room assigned, number of changes to the booking between booking date and entry/cancellation, type of deposit made at time of booking, id of the agent who made the reservation (na if the reservation was not made by an agent), id of the company that made the reservation (na if the reservation is not corporate), how many days did it take to confirm the reservation, booking type, average daily rate, average price for each night in the reservation, number of parking spaces required in the reservation, number of special requests in the reservation (double bed, floor, room with a view), and date the reservation was last updated.

The Approach

First the data will be cleaned. Next all variables will be examined and features may or may not be implemented to provide insight on the effect of the variables on cancellation. Afterwards, several binary classification methods will be explored and implemented including (but not limited to) random forest, logistic regression, and support vector machines. Evaluation will be determined by using the Mean F1-Score. Time will be spent fine tuning the various models and trying new approaches until a 95% accuracy is achieved. Model selection will be purely based on highest F1-score.

Deliverables

The following will be delivered in a GitHub repo:

- Jupyter notebooks containing details and implementation of the predictive models utilizing the data science method.
- Project Report
- Slide Deck

Problem Statement Worksheet (Hypothesis Formation)

How can Sunset Pier Hotels and Resorts implement data backed risk mitigation strategies for the next hotel season that (a) reduce their loss in revenues due to cancellation to sub 10% and (b) do not dissuade the clients from booking, therefore increasing overall revenues by 5%?

1 Context

For Sunset Pier Hotels and Resorts, booking cancellations have always been a pain point. Cancellations negatively impact the hotel's revenue and bottom line where sometimes the hotel may not be able to book the room again in time. Due to cancellations, the hotel estimates a 14% loss in revenues. While two seasons ago the institution implemented risk mitigation in the form of requiring a booking deposit, management of the hotel suspect that this has dissuaded customers from booking altogether. Management wants to implement a data-driven approach to their risk management strategy to mitigate cancellations.

2 Criteria for success

- Implementation of a predictive model which can determine whether a client will cancel or not with at least a 70% accuracy rate which will translate into a reduction in loss of revenue.
- Suggestions for a data backed risk mitigation strategy that minimize potential to dissuade clients.

3 Scope of solution space

- Determine the effect of the many variables including (but not limited to) lead time, length of stay, number of guests, party composition (presence of children and/or babies), customer's track record (as a past customer), deposit type, average daily rate, etc. on cancellation rate.
- Thoroughly examine whether customer market segmentation has an effect on cancellations.
- Analyze the cancellation rates by agent and by company.
- Analyze cancellation rates by customer booking type.
- Suggest deposit size as a function of predicted cancellation rate.

4 Constraints within solution space

- Data sources are restricted to the provided CSV files.
- While the problem and solution are general, the training data is constrained to only 2 businesses.

5 Stakeholders to provide key insights

- Chief Data Scientist
- Chief Financial Officer

6 What key data sources are required?

- tb_hotel_traintest.cvs - the provided training set
- tb_hotel_feat_valid_2.csv - the provided validation set
- tb_hotel_sample_valid.csv - a sample submission file specifying format