MACHINE LEARNING ON A TRAVEL BUSINESS

Melisa Bardhi

February 22nd, 2021

TABLE OF CONTENTS

I. Data Source
II. Predicting Cancellations
III. Classifying Transactions
IV. Sentiment Analysis

DATA

Travel business data with 4238 observations and 21 variables

Key categorical data:

- Stars
- hotel_id
- AvgUserRating
- Type
- hasSpecialRequest
- hasFreeCancellation

Key non-categorical data:

- roomCount
- numberOfBookedNights
- numberOfReviews
- Total booking price

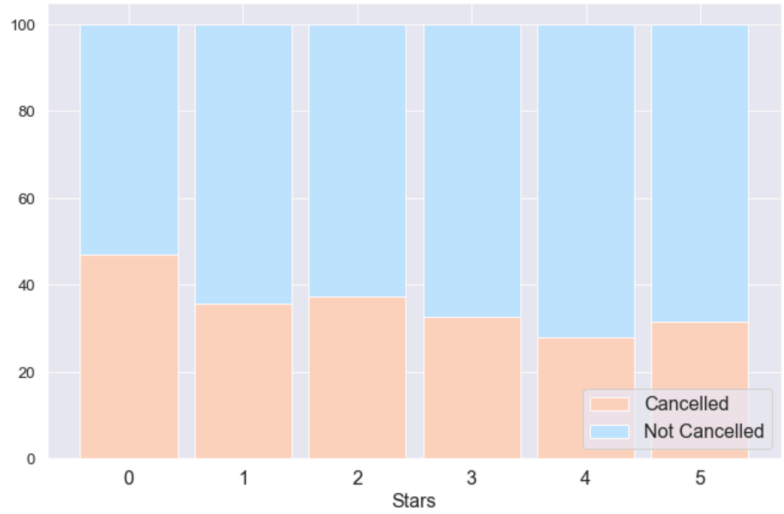
GOALS

1) Predict Venue Cancellations:
Study patterns in customer order cancellations, determine order importance and predict when a cancellation is likely to occur so that venues are prepared to find new bookings for empty space.

2) **Classify Bookings:**
Segment customers into clusters to effectively market to them so they are motivated to book venues which meet their personalized needs.

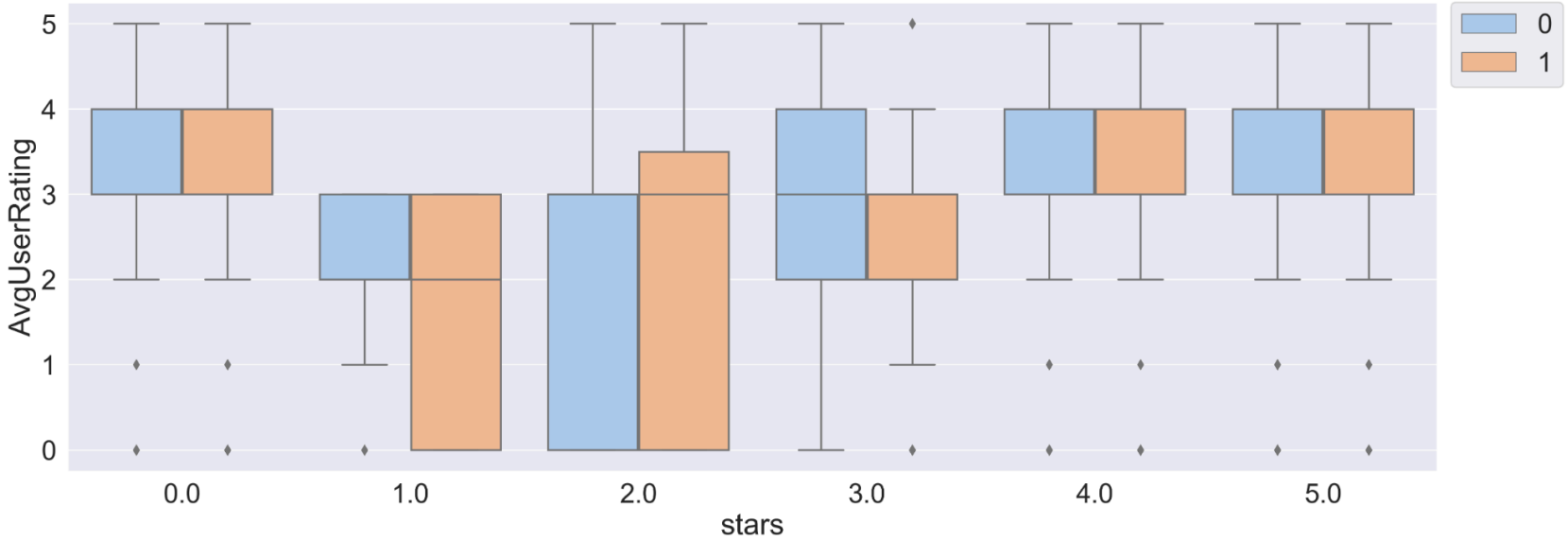
Part 1: Cancellation Predictions

CANCELLATION BY RATING AND TRAVEL TYPE



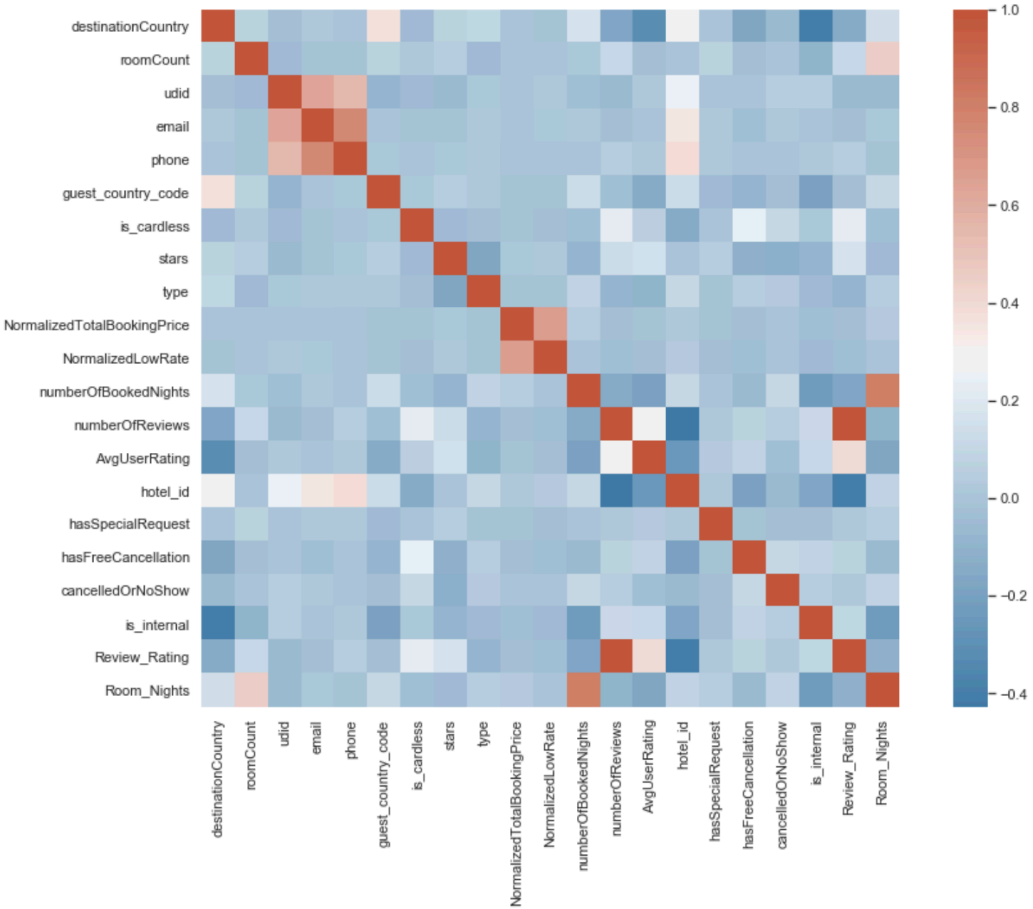
Cancellations are less likely for venues booked with higher ratings and booked for international travel

DEDUCED MISSING RATINGS



Hotels with no star ratings are comparable to 3 - 4 average user rating hotels

HEATMAP CORRELATIONS



Cancellations are correlated with number of booked nights, room counts, types of travel and venue types

F1 SCORE OF ALGORITHMS

Baseline	Logistic Regression	Support Vector Machine	Decision Trees	Random Forests	K-Nearest Neighbors	Neural Network
0.0	.22	.30	.37	.32	.23	.26

A cancellation happens 32% of the time and the Decision Trees model was best at predicting it, with an F1 score of 37% after cross-validation

PART 1 RECOMMENDATIONS

Develop strategies to effectively target buyers for specific empty rooms:

- Provide incentives for planners interested in rooms of a specific size or type
- Offer greater discounts for rooms with a larger availability gap from the cancellation date
- Analyze factors contributing to frequent cancellations and implement ways to avoid them
- Prioritize post-cancellation plan for rooms with a cancellation probability greater than 75%

Part 2: Segmentation

CLASSIFYING CLUSTERS

Clusters 0 and 4 were the most differentiated in terms of reviews and price

F1 SCORE OF ALGORITHMS

Dummy Classifier	Decision Trees	Random Forests	Logistic Regression	K Nearest Neighbors
.44	1	.67	.44	.40

The Decision Trees model was 56% more predictive than the dummy, making it best positioned to know which of the 4 clusters a new transaction belongs in

PART 2 RECOMMENDATIONS

Clusters 0 and 1 are rated equally, but cluster 1 charges higher prices.

- Venues in cluster 0 might be better purchases and could market themselves as more fairly-priced
- Venues in cluster 1 might need to justify higher prices or adjust prices

Cluster 3 has the the lowest ratings and lowest prices

- Venues in cluster 3 could use sentiment analysis on their reviews to determine biggest pain points to improve first, and by how much reviews need to increase for them to be able to charge equivalent to those in cluster 2

- All venues could use sentiment analysis to analyze text in their reviews for insights into how to distinguish themselves from competitors in their cluster

Thank you