Arshia Sharma, Betsy Heredia, Kayla Cho, Nawal Alasmari

Hotels Dataset Final Paper

Problem

Hotels are a vital part of travel and tourism for every country around the world; they provide travelers with comfortable accommodations, meals, laundry, occasionally entertainment, and a 24-hours staff available for every guest. As the industry continues to grow and with various different online platforms to book hotels such as Expedia and Kayak, hotels worldwide compete with competitors in their area to attract guests with their services, pricing, advertising, and ratings. However, since hotel rooms are typically booked well before the guest arrives, there are always risks of unforeseen changes in a guest's plan or spontaneous last-minute cancellations that puts the company in a tight situation where they may lose money rather than gain.

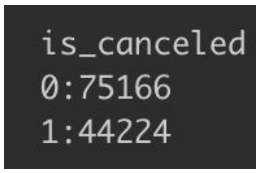
Motivation

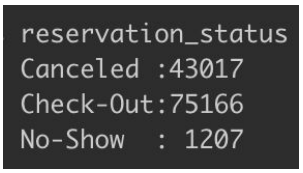
The more cancellation a hotel has may lead to a profit loss which may negatively impact a hotel's business. Therefore, predicting why customers would cancel their reservation is one of the best ways that will help to understand what causes people to cancel and how businesses could prevent that in the future. By studying this variable, we are aiming to determine which factors are the most important when it comes to their impact on a hotel business and their profit. This may also help a hotel business to have a better idea of which factors of their operation to improve on in order to keep customers from cancelling since the more committed customers a company has, the more profit they may earn.

Summary Statistics

The dataset that we used for this project was found on Kaggle and consisted of information from two main regions of the country Portugal: Lisbon and Algarve, from the

years 2014 to 2017. The raw dataset consisted of 19,390 observations and 32 features: ADR, adults, agent, arrival date day of the month, arrival date month, arrival date week number, arrival date year, assigned room type, babies, booking changes, children, company, country, customer type, days in waiting list, deposit type, distribution channel, hotel, is canceled, is repeated guest, lead time, market segment, meal, previous bookings not canceled, previous cancellations, required car parking spaces, reservation status, reservation status date, reserved room type, stays in weekend nights, stays in week nights, and total of special requests. Overall, our model had a fair mix of factor and numeric variables, with 15 factors, 16 numerical, and 1 Date variable types. The reservation status date was originally a character value and was changed to a date type using the lubridate package that R Studio provides. The 15-factor values were also converted from either character or numerical values to factor variables using as.factor() when first loading in our dataset. After analyzing our dataset descriptive statistics and producing a few graphic models, we quickly saw that there two very similar columns in your dataset Is Canceled and Reservation Type.





As shown above, we can see that the amount of people that checked out under reservation status and those who did not cancel their reservation in is canceled are the same value, 75,166. Additionally, those who did cancel from is canceled is the same number of guests that canceled or were no-shows under reservation status. To avoid these double variables in our models, we chose to drop the reservation status column from our dataset. However, along with reservation status also came reservation status date, which was to be "used in conjunction with the reservation status to understand when the booking was canceled or when the customer checked-out of the hotel." Since we chose to remove reservation status, we also removed the reservation status date as well. Furthermore, the variables agent, company, and country had multiple levels of factors in their respective columns which totaled out to be 865 coefficients generated from these 3 variables alone.

Although we did try to run our models with these 3 variables included, it was too computationally costly and some of our models did not finish running before facing multiple errors. Therefore, we also decided to remove these as well in order to save our computational speed of the model generations. After removing these 5 columns and running all of our models, we also saw that our accuracies, sensitivity, and specificity scores between the testing and training sets were extremely similar, which we believed may have been due to a class imbalance between our variables where one class heavily outweighed the other. To solve this issue, we also upsampled our model to include more repeated copies of our minority classes. When upsampling, we originally upsampled the whole dataset, but later realized that we should only upsample the training set since we still aim to test the model that we trained using our training set on the original testing set. Argo, the training and testing datasets were first split 80% to 20%, respectfully. The testing dataset remained at having 23, 876 observations, while the training set increased to 120,194 observations after the upsampling code was executed.

Analysis

To further analyze our question as to which attributes companies should focus on, or pay closer attention to, we chose to train three models: logistic regression, elastic net, and a decision tree. Each of these models all predicts the same outcome variables, is canceled, and were trained using the whole upsampled training dataset, when possible.

Logistic regression

The first model our group chose to test our outcome of predicting whether a guest cancelled their reservation started with a logistic regression model. Since our Is Canceled variable is a categorical one, a logistic regression model was the appropriate fit over a linear model. All variables of the dataset were used against our Is Canceled variable and after exponentiating all of our variables, we found that our top 3 most impactful coefficients was Required Car Parking Spaces, Assigned Room Type, and Reserved Room Type. We then found our residuals and predictions where we could create our confusion matrices to show how many true positives and true negatives our model can predict. With the changes from our presentation to only upsample the training set, our accuracy,

sensitivity, and specificity scores for both training and testing has been updated as well. For the training dataset, the accuracy was about 75.23% with a sensitivity score of about 67.87% and a specificity score of about 82.59% and for the testing set, there was an accuracy of about 73.48%, a sensitivity of 68.24%, and a specificity of about 82.44%. From these results, we can see that the model can better predict specificity, or when guests do not cancel their reservation opposed to those who do cancel. Since the accuracy and sensitivity scores from the training set were slightly higher than those from the testing set, our model is slightly overfit. Both accuracies were around 73%-75% range, which indicates that this is an average model at best. Additionally, when calculating our AUC-ROC score and r-squared we found that our training set had an AUC-ROC of about 86.40% and a r-squared of about 0.2603 and our testing set has an AUC-ROC score of 86.42% and a r-squared of 0.2391. From our AUC-ROC curve scores for both the training and testing sets, our model does fairly well when determining which guests cancel from those who keep their reservation but from our calculation of our r-squared, this is not the best model to use going forward with such a low r-squared and a fairly average accuracy score.

Elastic Net

The second model our group chose to create was an Elastic Net model. For our presentation to test the possible alpha values, we took a sample of 20,000 data points from the training set because it was too computationally expensive with this large data set. We visually observed the cross-validated error with different alphas ranging from 0 to 1 using the min loss plot. The optimal alpha was 0.95 because this alpha value produced the lowest cross-validated error. Since alpha was 0.95, it meant that the model was some ridge but mostly lasso regularization, where it penalizes more coefficients to precisely zero. We then created another model using alpha equal to 0.95 and the entire training set. We found that only 24 variables played significance in booking cancellations and reduced the number of coefficients from 80 to 61.

The top five most impactful variables on booking cancellations were total parking spaces, Required Car Parking Spaces, Deposit Type- Non-Refund, Reserved Room Type- P, Previous Cancellations, and Assigned Room Type - I. Based on the coefficients, we determined that the higher the number of required car parking spaces, the more likely the

customer will cancel. Customers are also more likely to cancel if the deposit type is considered non-refund, which the dataset's author defined as "a deposit was made in the value of the total stay cost." We found this odd because it seems counter-intuitive that someone would cancel their booking if considered Non-Refundable. To understand this better, we looked through the dataset's Kaggle discussions. We learned that the author discussed this issue in another paper where he found that bookings are usually made through Online Travel Agents using false or invalid credit card details. The term Non-Refund is for the hotel where they lose money, not the customer. These bookings were issued as support for visa requests to enter the country because a hotel booking is mandatory for applying for a Portuguese entry visa. We also found that if the guest had previously canceled before, they would be more likely to cancel again. If the Reserved Room is type P, there is a higher chance the booking is canceled. For privacy reasons, the hotels wanted to keep the specific type of rooms private, so we only have letter variables. If the Assigned Room Type is I, it is less likely that the booking is canceled. Since the variables "days in waiting list," "arrival date week number," "arrival date day of month" coefficients shrunk to zero, we concluded that these variables had no impact on if a booking is canceled or not.

We also used the factor_other() function to combine factors that the Elastic Net model shrunk to zero. We did this so that we could produce a more interpretable decision tree when we created our decision tree model because it has fewer factors. We tested the decision tree with and without this version of the dataset and had almost identical results but found that it was easier to interpret with the modified version.

In terms of how well our Elastic Net model performed, it performed slightly better than the logistic regression with a training r-squared of 0.2582 and accuracy of 75.2033% and testing r-squared of 0.2361 and accuracy of 75.2836%. We also found that since our sensitivity (Train: 67.8997%, Test:67.6208%) is lower than our specificity (Train: 82.5176%, Test: 82.9011%), this model is a lot better at predicting if someone does not cancel their booking.

After fixing how we upscaled the data, we reran the same model and followed the same process. We also ran another model using the entire train set because we wanted to represent the entire dataset better. We concluded that both models' optimal alpha value

was one because this alpha value produced the lowest cross-validated error. Since alpha equaled one, we created a model using lasso regularization. The binomial deviance plot slightly changed where lambda.min has approximately 70 nonzero coefficients, and lambda 1se had approximately 62 nonzero coefficients. We continued to use lambda 1se to analyze our variables and created another data table and graph of all of the coefficients. The top five coefficients that have the largest impact on booking cancellations remained the same. However, the coefficients “is_repeated_guest1”, “assigned_room_typeC”, “assigned_room_typeH”, and “deposit_typeRefundable'' were all shrunk to zero and “market_segmentDirect”, “distribution_channelCorporate”, and “reserved_room_typeC” no longer had a coefficient of zero. We regrouped the dataset using factor_other() after discovering these changes in the coefficients.

The lasso model performed slightly worse than the logistic regression with a training r-squared value of 0.2582 and accuracy of 75.1161% while the testing set had a r-squared of 0.2361 and accuracy of 73.2325%, which shows our model is slightly overfit. The lasso model's sensitivity (Train: 67.5608%, Test: 67.7661%) and specificity (Train: 82.6714%, Test: 82.5840%) changed slightly. This model’s test sensitivity and train specificity increased by a little more than 0.10% while the others decreased compared to the performance of the Elastic Net model (alpha = 0.95). However, this model is still better at predicting if someone does not cancel their booking.

Decision Tree

The third model we decided to create was a decision tree. We first wanted to prune the tree in order to know the optimal depth that will both decrease overfitting and error. We utilized the rpart function from the rpart package and found max depth of 12 to be ideal. Using the ctree function we created the model with all the variables using the training data set and capping the maximum depth at 12. When we plotted the visual representation of the decision tree, we found it to be too clustered to be easily interpretable. Therefore, we kept lowering the depth until we discovered that only at a maximum depth of 4, was the decision tree visually interpretable. So, we decided to continue with both versions of the model for different functions, the tree at depth 4 for visual interpretation and the tree at depth 12 for comparing predictive qualities with the other models.

When plotting the decision tree, we used a maximum depth of 4 on the model. Originally, from it we could see that the top node was "deposit_type" which split off by "Non Refund" versus "No Deposit" and "Refundable". The first split of non refund deposits lead to almost 100% cancellation. Intuitively this does not make sense and this is because, as stated before, the hotel found that people were booking rooms online using false credit cards which registered to the hotel as nonrefundable. Beside deposit type, the decision tree also showed that those who booked through an online travel agent, do not make any special request, and do not require a parking space were the second most likely to cancel. Meanwhile, those that booked through an online travel agent, don't make any special request, but do require a parking space were the least likely to cancel. However, since we later adjusted how we upsample the data the decision tree slightly changed. Non refund deposit types still mostly result in booking cancellations, but the booking was less likely to be canceled if the customer's "arrival_date_month" was in July or May or if the customer had not previously cancelled their booking. On the other split of the top node, which is no deposit or refundable deposit type, the least likely to cancel were customers who did require car parking space.

We decided to focus on the decision tree with a maximum depth of 12 for analyzing the performance metrics. The overall accuracy of the decision tree model when used against the training data set was 80.21% and 80.66% against the testing data set . In addition, when we looked at the fit through the r-squared values (Train: 0.3661, Test: 0.3485) we can see that although the model performed slightly better on the training set, it was not too overfit. This means that theoretically we could expect this model to perform similarly with an accuracy around 80.66% on any other data set from the hotel. When we plotted a confusion matrix for our model performed on the testing data set, we noticed that the weakness of our model is that it gave predicted 2,566 false negatives (false noncancellation predictions) and 2,051 false positives (false cancellation predictions).

Results and Conclusions

The test r-squared for the decision tree, the elastic net and the logistic regression models was at 0.3620, 0.2615 and 0.2609, respectively. The test accuracy for the decision

tree was 80.66%, 73.23% for the elastic net and 75.27% for the logistic regression. The results indicate that the decision tree had a higher r-squared and a higher accuracy when compared to the elastic net and the logistic regression models. Therefore, we can conclude

In conclusion, despite the low r-squared that we found, the decision tree performed better than the other two models. Across all three models , the top variables were assigned room type, reserved room type, and required car parking spaces. These variables and their relationship to booking cancellations are important for the company to keep that in consideration where this variable has one of the greatest impacts on cancellation. In addition to that, people who cancelled their reservation in the past, are more likely to cancel again. Therefore, the company might want to keep that in consideration as well and try to understand why people cancel to prevent that in the future.

Moving forward, we would like to test if using a different model type may produce a higher r-squared and accuracy score. In particular, we would want to create a random forest model for predicting if a booking is canceled. With the new model, we would want to observe which market segment and distribution channel has the least cancellations so that we can focus on reaching out to those markets and channels. Therefore, we would consider these adjustments for any future work with the dataset in order to get better results.

Sources

https://www.kaggle.com/jessemostipak/hotel-booking-demand

https://www.kaggle.com/jessemostipak/hotel-booking-demand/discussion/131787

Antonio, N., de Almeida, A., & Nunes, L. (2019). Big Data in Hotel Revenue Management: Exploring Cancellation Drivers to Gain Insights Into Booking Cancellation Behavior. Cornell Hospitality Quarterly, 60(4), 298–319. https://doi.org/10.1177/1938965519851466