Final Report

Modeling and Forecasting Hotel Booking Cancellations

Scheller College of Business, Georgia Tech OMSA

Management 6203: Data Analytics in Business

Team 60

Andrea Kirksey, Dakota Coomes, Jiangqin Ma, Gurman Singh, Duarte-Nuno de Sousa

April 16th, 2023

I. INTRODUCTION

A. Project Overview and Background Information/Business Justification

Increased hotel reservation options have changed booking behavior among potential customers. Due to the ease of bookings and cancellations associated with online reservation channels, hotels take on more booking risk, often with no financial recourse if bookings are cancelled. This risk can be mitigated substantially if it can be modelled and forecasted. To address this consideration, we are analyzing two related datasets that include booking profiles, hotel profiles, and time-series data along with cancellation status for multiple bookings.

B. Problem Statement and Primary Research Questions

Problem Statement: Determine what factors impact hotel booking cancellations and forecast future bookings.

Primary Research Question: Can hotel bookings and cancellations be modelled and forecasted using data on booking profiles and time of year?

Supporting Research Questions:

- Are hotel booking cancellations seasonal?
- Are hotel booking cancellations affected by hotel profile?
- Does family size/make-up impact booking cancellation propensity?
- Do hotel bookings show an increasing/decreasing trend?

C. Initial Hypotheses

We hypothesize that more cancellations happen during the winter (December, January, February) due to inclement weather affecting travel itineraries. Potential customers have more booking options in the city, and thus, we hypothesize that booking cancellations will be more prevalent at city hotels than resort/rural ones. With less booking options in rural areas, customers, we hypothesize, are less likely to cancel their respective bookings in these areas. We hypothesize that hotels will show an increasing bookings trend over time due to general understanding of increased travel and population increase over time. Large groups or families may have less booking cancellations than individuals due to logistical complexity associated with large-group planning.

II. OVERVIEW OF DATA

A. What is it and where did the dataset come from?

Our primary Kaggle dataset was originally obtained via queries of hotel industry SQL databases for two hotels in the country of Portugal. One hotel was located in the capital city of Lisbon; the other was a resort hotel located in the southern coastal region of Algarve, a popular vacation destination for residents of Portugal and other Europeans. The secondary dataset, although cleaner, had fewer variables and datapoints (119k vs 36k). The hotel type was not known, and the data source was not identified. However, the dataset variables are very similar to the primary dataset, suggesting the data may have been sourced via a hotel industry database.

B. Key Variables (Hypothesized)

Descriptive and predictive analytics - Logistic regression and SVM and KNN classification

- Potentially impactful predictors: number of adults, number of children, lead time for reservation, arrival month, arrival date, booking price, location, deposit type
  - Most likely impactful predictors: Arrival month, arrival date, lead time, booking price, location, deposit type
- Response: Booking status (cancelled or not)

Predictive analytics – time-series analysis/Holt-Winters forecasting

- Arrival date for cancelled hotel booking reservations
  - Total hotel booking reservations by week or month and hotel profile

C. Exploratory Data Analysis

Initial data clean-up concerned the removal of NA and Null values. Deeper inspection resulted in the removal of datapoints where the "Adult" value equaled zero, as well as datapoints where weeknights and weekend-nights both equaled zero. The final clean-up resulted in replacing meal plan values of "Unidentified" with the equal meaning value of "SC".

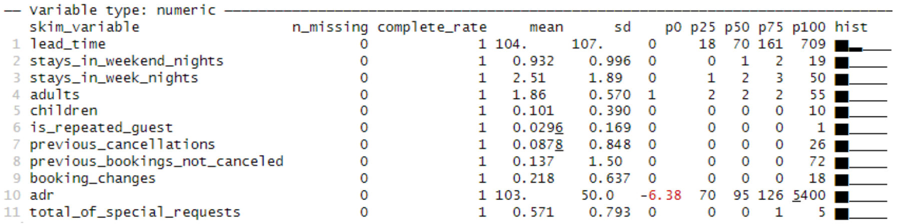
We added two mutated values:

- Season – to test our Seasonal hypothesis, we translated Month values of December, January, and February to Winter; March through May to Spring etc.
- Is_Domestic – there are 178 countries. The initial logistic regression deemed this variable to be not significant. However, translating this variable into a binary variable of Is_Domestic; where values of PRT = 1 and all others equal 0, revealed this to be a very significant variable in a subsequent regression.

After the clean-up and mutation processes, we then began inspecting our data more thoroughly. This included:

- Linear regression of numerical variables, which deemed the "babies" variable as insignificant.
- Skim (from the skimr package) provides a statistical summary which includes histograms. The histograms illustrate our numeric variables have long tails, thus requiring scaling.

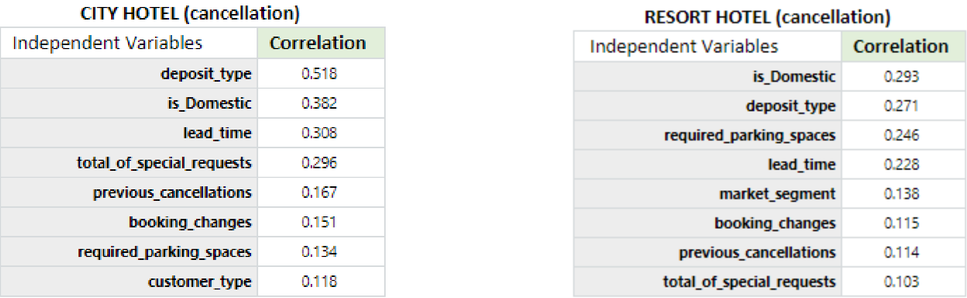
Output 1: Numeric Variable Summary



```
── Variable type: numeric ─────────────────────────────────────────────────────────────────
   skim_variable                  n_missing complete_rate    mean     sd    p0 p25 p50 p75 p100 hist
 1 lead_time                              0             1 104.     107.     0     18  70 161  709 ▇▂▁▁▁
 2 stays_in_weekend_nights                0             1   0.932    0.996 0     0   1   2   19 ▇▁▁▁▁
 3 stays_in_week_nights                   0             1   2.51     1.89  0     1   2   3   50 ▇▁▁▁▁
 4 adults                                 0             1   1.86     0.570 1     2   2   2   55 ▇▁▁▁▁
 5 children                               0             1   0.101    0.390 0     0   0   0   10 ▇▁▁▁▁
 6 is_repeated_guest                      0             1   0.0296   0.169 0     0   0   0    1 ▇▁▁▁▁
 7 previous_cancellations                 0             1   0.0878   0.848 0     0   0   0   26 ▇▁▁▁▁
 8 previous_bookings_not_canceled         0             1   0.137    1.50  0     0   0   0   72 ▇▁▁▁▁
 9 booking_changes                        0             1   0.218    0.637 0     0   0   0   18 ▇▁▁▁▁
10 adr                                    0             1 103.      50.0  -6.38 70  95 126 5400 ▇▁▁▁▁
11 total_of_special_requests              0             1   0.571    0.793 0     0   0   1    5 ▇▁▁▁▁
```

Because our primary data set consists of two specific hotel types, it was prudent to inspect whether there were patterns distinct to each. A correlation matrix revealed similarities among the top features for both hotels: deposit_type, is_Domestic, and lead_time. The top 8 correlated variables for each resort type are shown below (using absolute values). After the top similar features, the remaining variables diverge in importance.

Figure 1: Correlation Matrices for Primary Dataset



CITY HOTEL (cancellation)

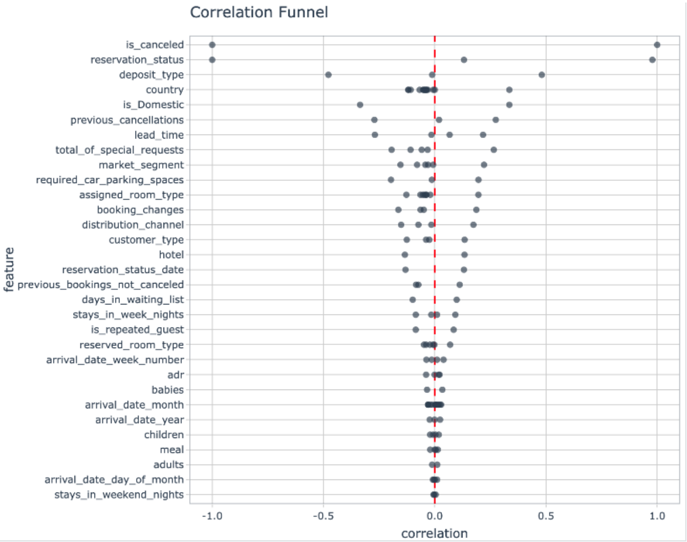
Independent Variables	Correlation
deposit_type	0.518
is_Domestic	0.382
lead_time	0.308
total_of_special_requests	0.296
previous_cancellations	0.167
booking_changes	0.151
required_parking_spaces	0.134
customer_type	0.118

RESORT HOTEL (cancellation)

Independent Variables	Correlation
is_Domestic	0.293
deposit_type	0.271
required_parking_spaces	0.246
lead_time	0.228
market_segment	0.138
booking_changes	0.115
previous_cancellations	0.114
total_of_special_requests	0.103

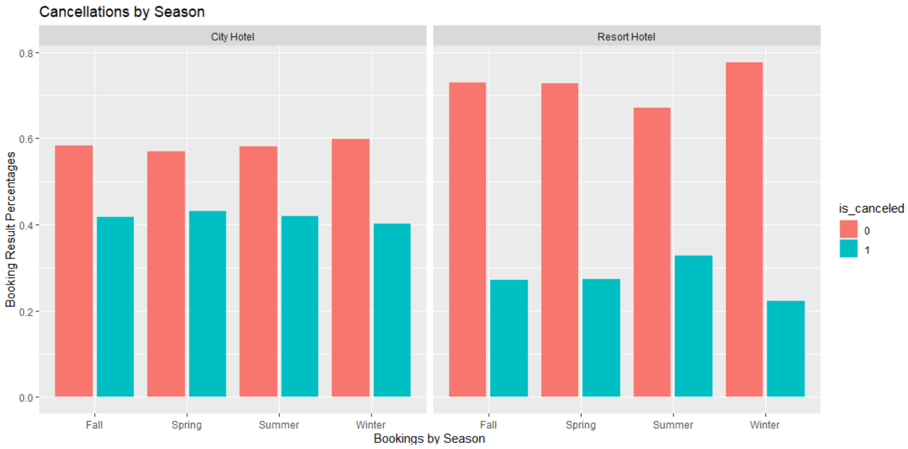
Our secondary dataset was analyzed to determine correlations between the booking cancellation response and other predictors. A correlation funnel of these results is shown below.

Figure 2: Correlation Funnel for Secondary Dataset



Additional comparisons between the hotel types illustrated a marked difference in the overall cancellation rates. The side-by-side graphs below clearly depict a sizable variance in cancellation rates. The City Hotel's overall cancellations hovers at 40% for all seasons, while the Resort Hotel's cancellations were below 30% percent for all seasons, except at their peak during Summer. These results support our initial hypothesis that urban hotels are likely to have greater rates of booking cancellations as compared to rural ones.

Figure 3: Booking Cancellations by Season

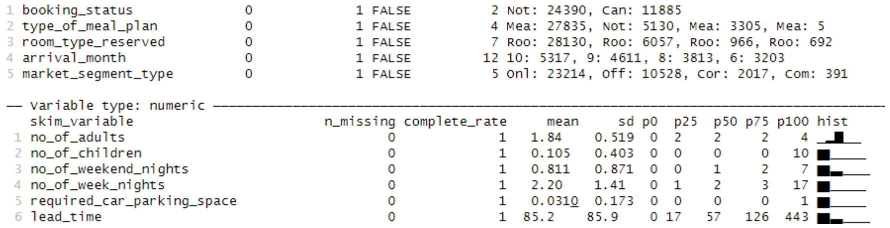


However, this chart does show that another of our initial hypotheses is not supported – that cancellations are more likely in the winter. Hotel cancellations were not more prevalent in Winter months. Plausible reasons include the additional flexibility customers have during the summer vacation months as well as the fact that Portugal is not notorious for harsh winter weather.

Besides comparisons to cancellations, an overall correlation matrix identified variables that are highly correlated to each other. An obvious find are reservations with more adults and children have a higher "adr" value (cost). The Market_Segment and Distribution_Channel variables are also highly correlated, and this is to be expected. Distribution refers to whether the booking was made via a travel agent or tour operator (82% of all bookings), while Market refers to whether that booking was made via online travel agent, offline travel agent/tour operator, or other. With such a high correlation one should be removed in our modeling, since both datasets use Market_Segment, the Distribution_Channel would be the drop candidate.

Because the variables in the secondary dataset are very similar to the primary dataset, this allowed our team to experiment modeling on both datasets with minor code adjustments.

Output 2: Variable Summary



```
1 booking_status            0             1 FALSE          2 Not: 24390, Can: 11885
2 type_of_meal_plan         0             1 FALSE          4 Mea: 27835, Not: 5130, Mea: 3305, Mea: 5
3 room_type_reserved        0             1 FALSE          7 Roo: 28130, Roo: 6057, Roo: 966, Roo: 692
4 arrival_month             0             1 FALSE         12 10: 5317, 9: 4611, 8: 3813, 6: 3203
5 market_segment_type       0             1 FALSE          5 Onl: 23214, Off: 10528, Cor: 2017, Com: 391

── Variable type: numeric ──────────────────────────────────────────────────────────────────────
  skim_variable                 n_missing complete_rate    mean     sd p0  p25  p50 p75 p100 hist
1 no_of_adults                          0             1   1.84   0.519  0 2    2     2    4
2 no_of_children                        0             1   0.105  0.403  0 0    0     0   10
3 no_of_weekend_nights                  0             1   0.811  0.871  0 0    1     2    7
4 no_of_week_nights                     0             1   2.20   1.41   0 1    2     3   17
5 required_car_parking_space            0             1   0.0310 0.173  0 0    0     0    1
6 lead_time                             0             1  85.2   85.9    0 17   57   126  443
```

III. OVERVIEW OF ANALYTICS AND MODELING AND DISCUSSION OF RESULTS

A. Analytics Methodology and Types of Models Used

Logistic regression to determine how parameters affect the likelihood of hotel booking cancellation.

- Plot ROC Curve and AUC to assess model performance

KNN and SVM model for classification

- Tune C and K hyper-parameters
- Split the dataset into training, validation, and test sets
- Compare accuracy across models

Holt-Winters forecasting and exponential smoothing to determine periodicity and trend in booking cancellations throughout the year

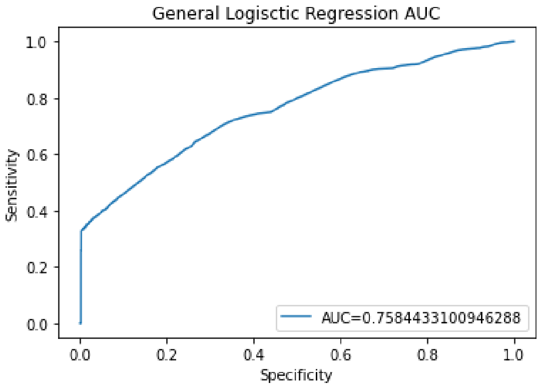
B. Model Performance, Comparisons, and Optimizations

Logistic Regression Model

Logistic regression is a supervised (outcome class for each observation is known) machine learning approach that estimates the probability of an outcome from a dataset of predictors (and their respective outcomes). As we were tasked with determining the likelihood of cancellation of hotel bookings based on booking profile, we determined that analysis via logistic regression would be appropriate and likely insightful for predictive purposes.

Because we were interested in determining whether there were differences in cancellation propensity between different types of hotels (rural versus urban), we set up a few different models. Our first model was an attempt to limit predictors for maximum interpretability while also retaining predictive power. Based on the correlation coefficients between predictors and response (cancellations) determined in our EDA along with SME knowledge on high-impact variables, we chose what we believe to be the most impactful predictors: hotel type, deposit type, lead time, and total number of special requests associated with the booking.

Figure 4: EDA/SME Generalized Hotel ROC and AUC



To determine model performance, we constructed an ROC curve and evaluated the associated AUC, which provides the model's performance across all possible classification thresholds. Our greatest AUC was approximately 0.76. It should be noted that within this analysis, a confusion matrix and associated accuracy was calculated, but we determined that because our response variable (booking status) is so imbalanced toward not cancelled versus cancelled, AUC would be a more fair and better evaluation of our model's performance as opposed to accuracy (which could be easily overfit with this imbalance).

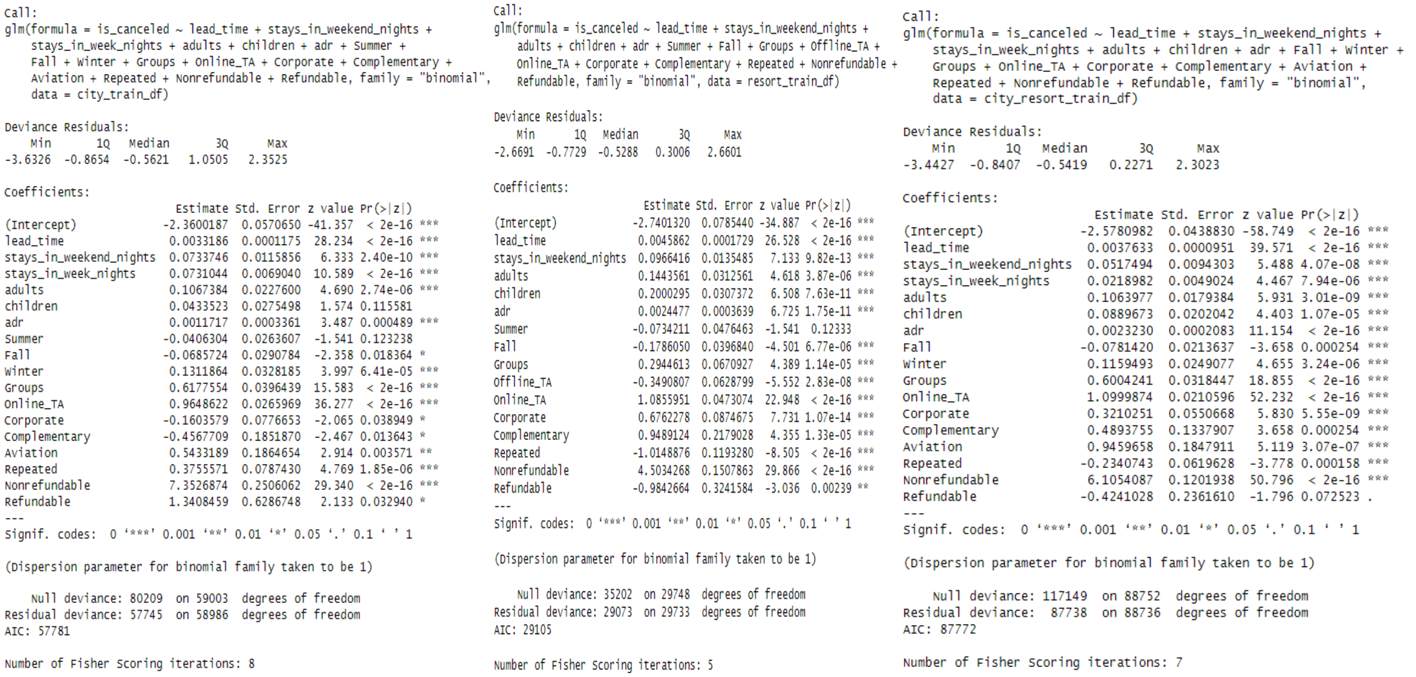
Our second approach within logistic regression modelling was to use factor selection to determine the most important predictors of booking cancellation while maximizing explained error. For this approach, we split our approach into three classes.

One class of model used only resort/rural hotel data and, as such, was devised to get more fine-grain insights into that market segment. Another class of model used only city hotel data for the same purpose for urban hotels. The final class used data from both a rural and city hotel to construct a more generalized model for hotel booking cancellations across the market.

Our first logistic regression model for each class was executed with all parameters simply to have a baseline understanding of relationships in the data. In general, across all three model classes, there were a fair amount of significant predictors at the p = 0.05 significance level (to be expected with as much power and number of observations our datasets have). Also of note is the fact that a nonrefundable deposit as part of a hotel booking unexpectedly coincides with an increased likelihood of cancellation as compared to no deposit or a refundable one in all three classes.

The second logistic regression model for each class was setup by first using stepwise regression to eliminate insignificant predictors at the p = 0.05 level using a greedy algorithm. However, even with this factor selection protocol in place, likely due to our large sample size, we still retained many significant predictors. Results from these models show the same increased likelihood of booking cancellation with a nonrefundable upfront payment, but we also see another significant and high-magnitude predictor of booking cancellation – bookings made through an online travel agent. This predictor, in all three models, increased the log odds of cancellation by around 1 (compared to lower magnitudes in all other significant predictors aside from a nonrefundable deposit).

Output 3: Logistic Regression Model Summaries



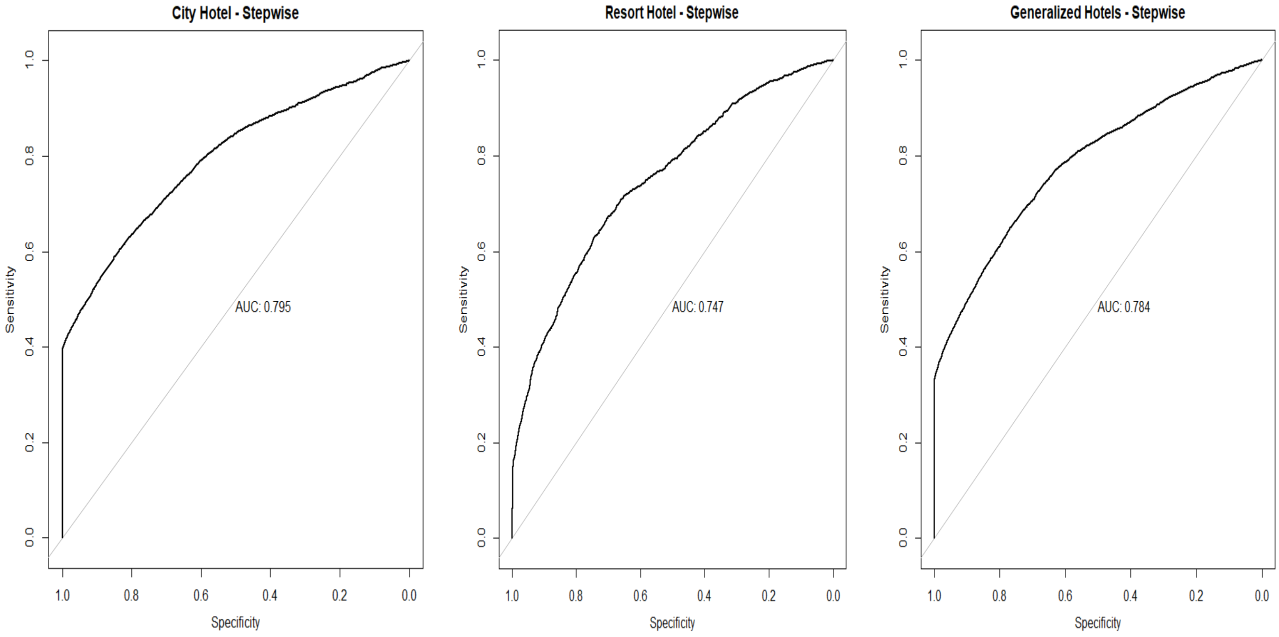
```
Call:
glm(formula = is_canceled ~ lead_time + stays_in_weekend_nights + 
    stays_in_week_nights + adults + children + adr + Summer + 
    Fall + Winter + Groups + Online_TA + Corporate + Complementary + 
    Aviation + Repeated + Nonrefundable + Refundable, family = "binomial", 
    data = city_train_df)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.6326  -0.8654  -0.5621   1.0505   2.3525  

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)    
(Intercept)             -2.3600187  0.0570650 -41.357  < 2e-16 ***
lead_time                0.0033186  0.0001175  28.234  < 2e-16 ***
stays_in_weekend_nights  0.0733746  0.0115856   6.333 2.40e-10 ***
stays_in_week_nights     0.0731044  0.0069040  10.589  < 2e-16 ***
adults                   0.1067384  0.0227600   4.690 2.74e-06 ***
children                 0.0433523  0.0275498   1.574 0.115581    
adr                      0.0011717  0.0003361   3.487 0.000489 ***
Summer                  -0.0406304  0.0263607  -1.541 0.123238    
Fall                    -0.0685724  0.0290784  -2.358 0.018364 *  
Winter                   0.1311864  0.0328185   3.997 6.41e-05 ***
Groups                   0.6177554  0.0396439  15.583  < 2e-16 ***
Online_TA                0.9648622  0.0265969  36.277  < 2e-16 ***
Corporate               -0.1603579  0.0776653  -2.065 0.038949 *  
Complementary           -0.4567709  0.1851870  -2.467 0.013643 *  
Aviation                 0.5433189  0.1864654   2.914 0.003571 ** 
Repeated                 0.3755571  0.0787430   4.769 1.85e-06 ***
Nonrefundable            7.3526874  0.2506062  29.340  < 2e-16 ***
Refundable               1.3408459  0.6286748   2.133 0.032940 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 80209  on 59003  degrees of freedom
Residual deviance: 57745  on 58986  degrees of freedom
AIC: 57781

Number of Fisher Scoring iterations: 8
```

```
Call:
glm(formula = is_canceled ~ lead_time + stays_in_weekend_nights + 
    adults + children + adr + Summer + Fall + Groups + Offline_TA + 
    Online_TA + Corporate + Complementary + Repeated + Nonrefundable + 
    Refundable, family = "binomial", data = resort_train_df)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.6691  -0.7729  -0.5288   0.3006   2.6601  

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)    
(Intercept)             -2.7401320  0.0785440 -34.887  < 2e-16 ***
lead_time                0.0045862  0.0001729  26.528  < 2e-16 ***
stays_in_weekend_nights  0.0966416  0.0135485   7.133 9.82e-13 ***
adults                   0.1443561  0.0312561   4.618 3.87e-06 ***
children                 0.2000295  0.0307372   6.508 7.63e-11 ***
adr                      0.0024477  0.0003639   6.725 1.75e-11 ***
Summer                  -0.0734211  0.0476463  -1.541  0.12333    
Fall                    -0.1786050  0.0396840  -4.501 6.77e-06 ***
Groups                   0.2944613  0.0670927   4.389 1.14e-05 ***
Offline_TA              -0.3490807  0.0628799  -5.552 2.83e-08 ***
Online_TA                1.0855951  0.0473074  22.948  < 2e-16 ***
Corporate                0.6762278  0.0874675   7.731 1.07e-14 ***
Complementary            0.9489124  0.2179028   4.355 1.33e-05 ***
Repeated                -1.0148876  0.1193280  -8.505  < 2e-16 ***
Nonrefundable            4.5034268  0.1507863  29.866  < 2e-16 ***
Refundable              -0.9842664  0.3241584  -3.036  0.00239 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35202  on 29748  degrees of freedom
Residual deviance: 29073  on 29733  degrees of freedom
AIC: 29105

Number of Fisher Scoring iterations: 5
```

```
Call:
glm(formula = is_canceled ~ lead_time + stays_in_weekend_nights + 
    stays_in_week_nights + adults + children + adr + Fall + Winter + 
    Groups + Online_TA + Corporate + Complementary + Aviation + 
    Repeated + Nonrefundable + Refundable, family = "binomial", 
    data = city_resort_train_df)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.4427  -0.8407  -0.5419   0.2271   2.3023  

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)    
(Intercept)             -2.5780982  0.0438830 -58.749  < 2e-16 ***
lead_time                0.0037633  0.0000951  39.571  < 2e-16 ***
stays_in_weekend_nights  0.0517494  0.0094303   5.488 4.07e-08 ***
stays_in_week_nights     0.0218982  0.0049024   4.467 7.94e-06 ***
adults                   0.1063977  0.0179384   5.931 3.01e-09 ***
children                 0.0889673  0.0202042   4.403 1.07e-05 ***
adr                      0.0023230  0.0002083  11.154  < 2e-16 ***
Fall                    -0.0781420  0.0213637  -3.658 0.000254 ***
Winter                   0.1159493  0.0249077   4.655 3.24e-06 ***
Groups                   0.6004241  0.0318447  18.855  < 2e-16 ***
Online_TA                1.0999874  0.0210596  52.232  < 2e-16 ***
Corporate                0.3210251  0.0550668   5.830 5.55e-09 ***
Complementary            0.4893755  0.1337907   3.658 0.000254 ***
Aviation                 0.9459658  0.1847911   5.119 3.07e-07 ***
Repeated                -0.2340743  0.0619628  -3.778 0.000158 ***
Nonrefundable            6.1054087  0.1201938  50.796  < 2e-16 ***
Refundable              -0.4241028  0.2361610  -1.796 0.072523 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 117149  on 88752  degrees of freedom
Residual deviance:  87738  on 88736  degrees of freedom
AIC: 87772

Number of Fisher Scoring iterations: 7
```

We can see from the below output that our ROC curves and associated AUCs for each of our six models show a non-random level of prediction accuracy, and in fact, have a decent amount of prediction accuracy in general. There was little to no advantage in predictive power of our stepwise-regressed models versus our models with all predictors, so we can conclude that the predictors excluded from our stepwise-regressed models were, indeed, superfluous.

Figure 5: Stepwise Logistic Regression ROC and AUC



Together, these models suggest that there are many factors that impact the likelihood of booking cancellations across market segments and hotel types, but from this approach, we posit that deposit status (specifically a non-refundable deposit) and booking through an online travel agent are the predictors most likely to significantly raise the likelihood of cancellation.

We contend that further research into this topic should focus on enhanced factor selection potentially through principal component analysis or an elastic net approach to more confidently determine superfluous predictors and remove them from analysis. This will allow for the derivation of models that are easier to understand and interpret for hotel finance managers looking to make actionable decisions to decrease monetary fallout associated with booking cancellations. With more advanced factor selection, it is likely that more impactful predictors will surface giving managers more insight into how to best handle booking cancellation risk.

KNN and SVM Classification Models

Our second set of classification models involved training and testing our cancellation data using the k-nearest neighbors algorithm (KNN) and a support vector machine (SVM) approach. Both of these machine learning models are supervised (KNN is parametric while SVM can be either parametric or nonparametric depending on whether the classifier used is linear – we use both approaches in our models). In order to make comparisons between these two models, we determined that analyzing confusion matrices and associated accuracies was acceptable as this dataset showed significantly less imbalance in response prevalence.

We first performed feature selection as with our linear regression approach. By doing so, we hoped to reduce the training time of the model and improve its interpretability. From our EDA analysis of correlations between response and predictors, we determined that the most important variables to include in our models were deposit_type, previous_cancellations, lead_time, total_of_special_requests, market_segment, required_car_parking_spaces, booking_changes, is_Domestic, and customer_type.

To evaluate the performance of our KNN and SVM models, we split a revised dataset (original dataset was prohibitively computationally expensive to analyze) into training and test sets. Using the training data, we built the models and then tested their performance on the test data set. This approach allowed us to compare the accuracy and efficiency of both models to determine which is better suited for the task at hand.

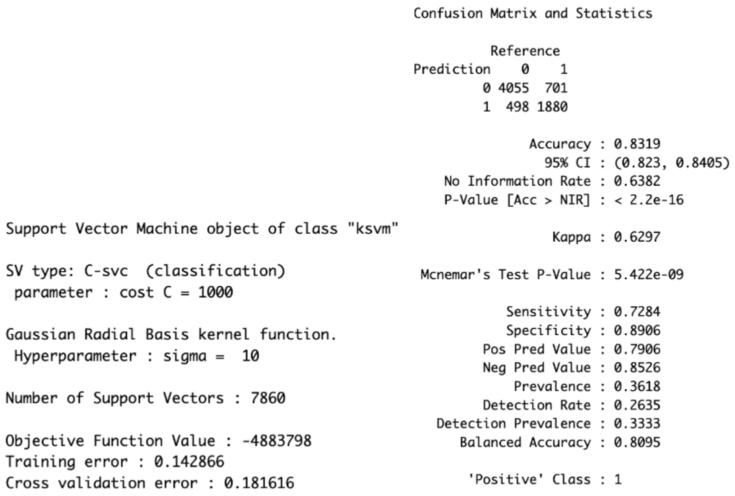
SVM

SVM is sensitive to the choice of kernel function and the regularization parameter which can affect the accuracy of the model. Our initial attempt used the linear kernel "vanilladot", but we encountered error messages. This led us to suspect the data may not be well-suited for a linear kernel; the vanilla dot product kernel assumes the input variables are linearly separable – that the classes can be separated by a straight line or plane in the input space. We then decided to approach the modeling with a nonlinear kernel. The radial basis function (RBF) kernel is a popular choice for SVM in classification and regression tasks, as it can capture complex nonlinear relationships between input variables and can handle high-dimensional feature spaces. Given the many features in the hotel reservation dataset, it is likely there are complex nonlinear relationships between the input variables, which can be modeled by the RBF kernel (kernel="rbfdot"). This tactic was successful in building a SVM model.

We used the ksvm() method to optimize the SVM model's hyperparameters – C and sigma. In ksvm(), the cross argument specifies the number of folds to use in cross-validation. Here, we set cross = 5, which means that the data is split into 5 folds, and the SVM model is trained and tested on each fold in turn, with the final performance being the average of the 5 iterations. Using cross-validation provides a more reliable estimate of the model's performance because it uses more of the data for both training and testing and ensures each data point is used for testing exactly once. This helps to avoid overfitting.

The sigma parameter controls the width of the RBF kernel, which determines the degree of influence that each training point has on the classification of new points, and the C parameter determines the trade-off between maximizing the margin between the hyperplane and the support vectors and minimizing the classification error. We experimented with different values until the best combination was identified. The optimal C was 1000, and the optimal sigma was 10. The model’s characteristics and performance display are as follows:

Output 4: SVM Model Summary



```
Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 1000

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  10

Number of Support Vectors : 7860

Objective Function Value : -4883798
Training error : 0.142866
Cross validation error : 0.181616
```

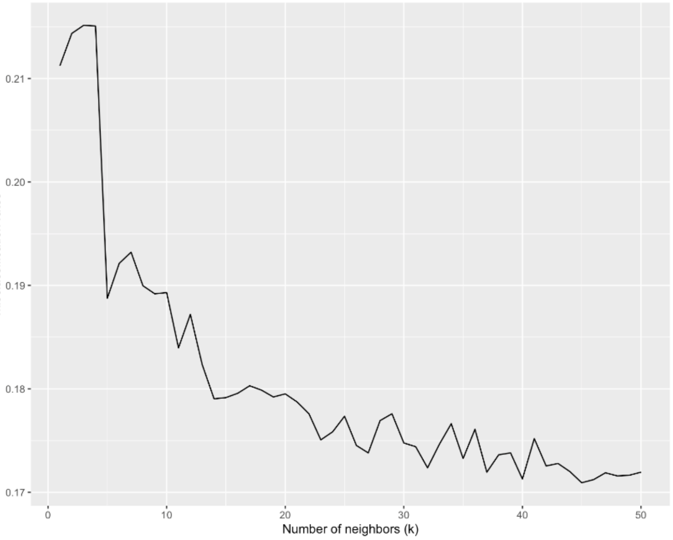
```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4055  701
         1  498 1880

               Accuracy : 0.8319
                 95% CI : (0.823, 0.8405)
    No Information Rate : 0.6382
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6297

 Mcnemar's Test P-Value : 5.422e-09

            Sensitivity : 0.7284
            Specificity : 0.8906
         Pos Pred Value : 0.7906
         Neg Pred Value : 0.8526
             Prevalence : 0.3618
         Detection Rate : 0.2635
   Detection Prevalence : 0.3333
      Balanced Accuracy : 0.8095

       'Positive' Class : 1
```

KNN

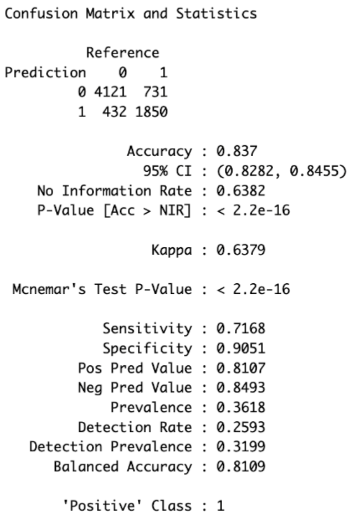
The performance of KNN can be affected by the choice of k and the distance metric used to calculate the distances between data points. To simplify the analysis, we are using the default distance metric. To determine the optimal value of the k parameter for the KNN model, we performed cross-validation. First, we set the maximum value of k to 50 and created an array to store the misclassification rates.

Next, we looped through each value of k to run 5-fold cross-validation using the cv.kknn() function. For each fold, we calculated the misclassification rate and then averaged it across all folds to estimate the model's performance for that value of k. The plot of k and corresponding misclassification rates is displayed as follows. The optimal k was 45. The performance display is as follows:

Figure 6: KNN Misclassification Plot



Output 5: Confusion Matrix Summary



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 4121  731
         1  432 1850

               Accuracy : 0.837
                 95% CI : (0.8282, 0.8455)
    No Information Rate : 0.6382
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6379

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7168
            Specificity : 0.9051
         Pos Pred Value : 0.8107
         Neg Pred Value : 0.8493
             Prevalence : 0.3618
         Detection Rate : 0.2593
   Detection Prevalence : 0.3199
      Balanced Accuracy : 0.8109

       'Positive' Class : 1
```

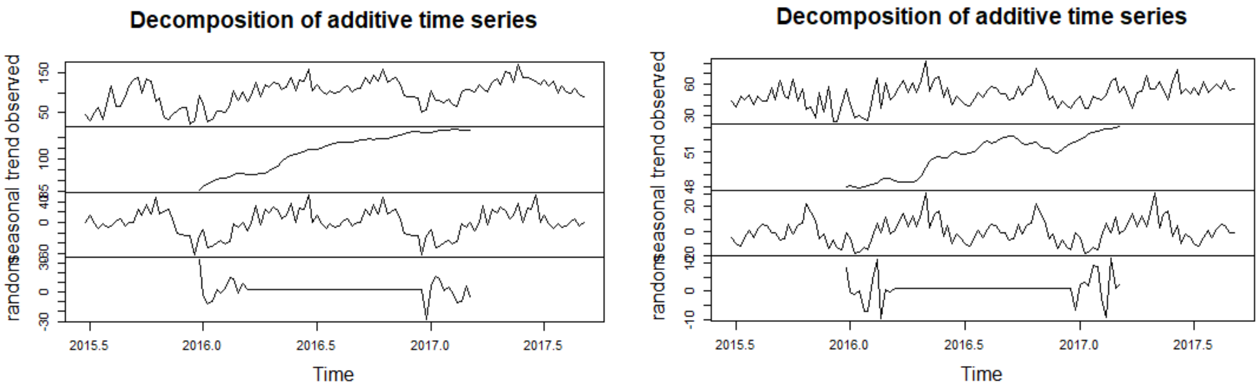
We can see from our above outputs that both the SVM and KNN model show relatively strong accuracy. Given that our accuracy after testing for KNN was slightly higher coupled with the remarkably quicker computational time of KNN as compared to SVM (4 minutes as opposed to 2 hours for KNN), we suggest that our KNN model is likely the better of the two models for this (and other) datasets. It is easy to understand, quick to compute, and a strong model in terms of accuracy.

Holt-Winters Model

Our third and final modeling approach was to forecast future bookings using the Holt-Winters exponential smoothing method. Holt-Winters is a time series analysis approach that looks at both trend and seasonality in data for descriptive analytics as well as forecasting.

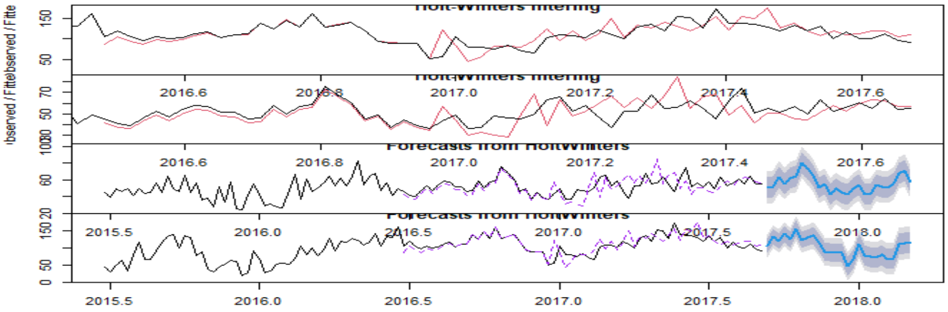
We started by plotting the top ten countries where hotel customers came from, the different types of meals booked, the distribution channels guests used to book, and the different types of guests who made bookings (based on earlier EDA analysis of important predictors). Since the dataset only provided data for about 2 years, we determined that a long-term forecast was neither reliable nor relevant. Therefore, we decided to transform the data to a weekly frequency to get each week's average arrivals for both the resort and city hotel. We used a frequency of 52. Before doing the forecast, we used a decomposition plot to try and gain some insight into the time series. This allowed us to separate the series into seasonality, trend, and random components.

Figure 7: Decomposition of City Hotel and Resort



We can see from the above output that the city hotel (on the left) had somewhat of a constant positive trend in bookings whereas the resort decomposition (on the right) didn't show as much of a constant trend. However, it appears that both have weekly seasonality. Once done, we fitted an exponential smoothing model with trend and additive seasonal components using Holt-Winters to forecast future values. Then, we plotted the fitted model and forecasts with a confidence interval of 80% to 95% for 6 months. The results are below.

Figure 8: Holt-Winters for Arrivals at City Hotel and Resort



We can see from the forecast output that the error associated with our forecast is moderately high. This is due to the limited timespan in our reference dataset coupled with the inherent nature of increasing error as forecast length increases. In future analysis, we suggest that any forecasting reference a larger amount of data with a greater timespan to reduce predictive error and improve confidence in projections. We also posit that any forecasting analyzed through a Holt-Winters approach should likely be performed on a monthly or even weekly basis as even a half-year forecast exhibited a suboptimal amount of error.

IV. CONCLUSION

Overall, we were able to model and predict hotel booking cancellations with various models such as logistic regression, SVM, KNN, and Holt-Winters. We employed different variable selection methods such as high correlation and feature selection to select the best variables for model training and analysis. The logistic regression and KNN models gave us good prediction AUC/accuracy of ~80%, and these models are easier to implement and optimize. After initial data exploration we determined that hotel cancellations were not prevalent in Winter months as initially predicted. The type of hotel (City Hotel and Resort Hotel) did influence cancellation rates – 40% for city hotels as opposed to 30% for resort/rural hotels. This result is aligned with our initial hypothesis that customers have more booking options and flexibility in urban areas. Family make-up did not have a significant impact on booking cancellations as determined by our EDA. Holt-Winters analysis showed an upward trend in city hotel bookings (and to a lesser degree in resort hotel bookings), and with seasonality, we were able to forecast bookings up to six months in the future.

It should be considered that the results and analysis we see here are representative only of the dataset of two hotels in Portugal. Hotels in other countries could have different significant variables, seasonality, and hotel types that could affect bookings and cancellations. Also, given the limited amount of data, we had some challenges identifying trends and forecasting long-term hotel bookings. However, the next step would be to forecast bookings and cancellations in the short-term and utilize other variables such as weather and competitive intelligence. We reviewed published and on-going research where the focus was predicting short-term (4-5 days) booking cancellations. Researchers have employed complex machine learning models (decision trees and random forests) to forecast cancellations with 85-90% accuracy. As hotels obtain more data, short-term forecasts would be instrumental in operational planning to maximize room utilization.

WORKS CITED

Antonio, N., de Almeida, A. M., & Nunes, L. (2017). Predicting Hotel Booking Cancellation to Decrease Uncertainty and Increase Revenue. *Tourism & Management Studies*, *13*(2), 25–39. https://doi.org/https://www.researchgate.net/publication/310504011_Predicting_Hotel_Booking_Cancellation_to_Decrease_Uncertainty_and_Increase_Revenue

Falk, M., & Vieru, M. (2018). Modelling the cancellation behaviour of hotel guests. *International Journal of Contemporary Hospitality Management*, *30*(3). https://doi.org/https://www.researchgate.net/publication/326824252_Modelling_the_cancellation_behaviour_of_hotel_guests

Sánchez, E. C., Sánchez-Medina, A. S.-M. J., & a Pellejero, M. (2020). Identifying critical hotel cancellations using artificial intelligence. *Tourism Management Perspectives*, *35*. https://doi.org/https://www.sciencedirect.com/science/article/abs/pii/S2211973620300854