MGT 6203 Project Progress Report

Modeling and Forecasting Hotel Booking Cancellations

Background and Problem Statement

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 customer profiles, hotel profiles, and time-series data along with cancellation status for multiple bookings.

Problem Statement

Determine what factors impact hotel booking cancellations and forecast future booking cancellations.

Primary Research Question

Can hotel booking cancellations be modelled and forecasted using data on customer profiles, 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 booking cancellations show an increasing/decreasing trend?

Initial Hypotheses

More cancellations happen during the winter (December, January, February) due to inclement weather affecting travel itineraries.

City hotels have more options. Therefore, the booking cancellations would be higher due to customers having more options in the city. Whereas, when traveling to a specific destination, they have fewer hotel options.

Hotels will show an increasing booking cancellations trend due to the gradual increase in booking reservations over time in general.

Large groups or families may have less cancellation than individuals due to logistical complexity associated with large-group planning.

Analysis Approach

We intend to compare logistic regression as well as SVM (Support Vector Machine) and KNN (K-Nearest Neighbor) classification to model booking cancellations. We also plan to perform time series analysis

involving a Holt-Winters method and exponential smoothing to forecast booking cancellations in the future.

Exploratory Data Analysis

Our primary Kaggle dataset was originally obtained via queries of hotel industry SQL databases for two hotels in the country of Portugal. One hotel located in the capital city of Lisbon; the other, a resort hotel located in the southern coastal region of Algarve, a popular vacation destination for residents of Portugal and other Europeans.

Initial data clean-up steps involved the removal of NA and Null values. Deeper inspection resulted in the removal of datapoints where the "Adult" value equaled zero. Further cleansing included replacing *meal plan* values of "Unidentified" with the equal meaning value of "SC".

To test our seasonal hypothesis, we created the *Season* variable and translated *Month* values of December, January, and February to Winter; March through May to Spring etc.

The primary dataset contains a *Country* variable with over 100 different values. The initial logistic regression deemed this variable to be not significant. However, we considered translating this variable into a binary variable of *Is-Domestic;* where values of PRT = 1, and all others equal 0. A repeat of the logistic regression proved this to be a very significant variable for both the City and Resort hotel.

Additional data wrangling revealed one of our initial hypotheses to be incorrect. Hotel cancellations were not more prevalent in Winter months (33.1%). In fact, quite the opposite was true, and summer months reported the highest cancellation rate (38.8%). A plausible reason for this could be the additional flexibility customers have during the summer vacation months. Notably, the city hotel had much higher overall cancellation rates for all seasons (over 40%) compared to the resort hotel (under 30%).

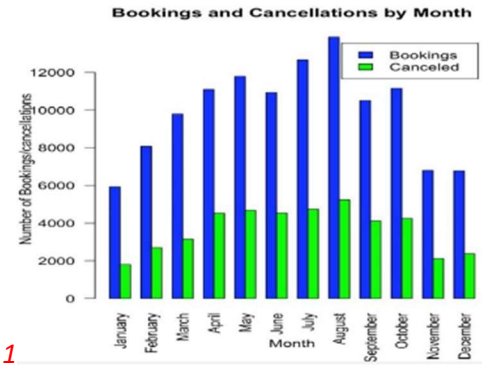


Fig. 1

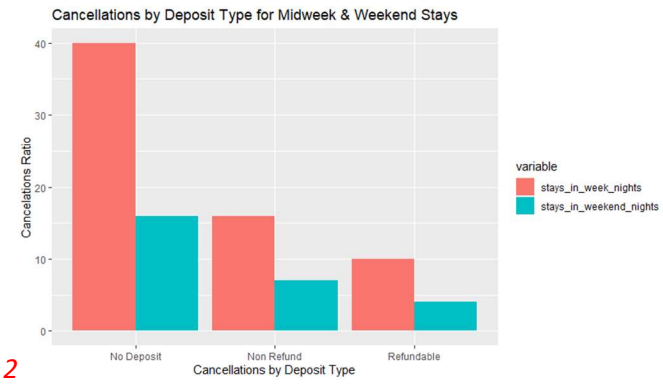


Fig. 2

Other analysis included boxplots which clearly defined the need to scale the data for our analytical modeling. A correlation matrix revealed the variable *Deposit Type* to be highly correlated with cancellations, and when a *No Deposit* datapoint was grouped with a *Weeknight stay*, the combination yielded a 40% overall cancellation ratio (*Fig. 2*).

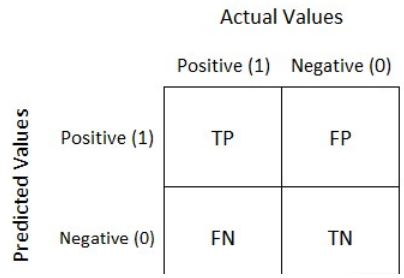
Analytical Method and Models

Logistic regression

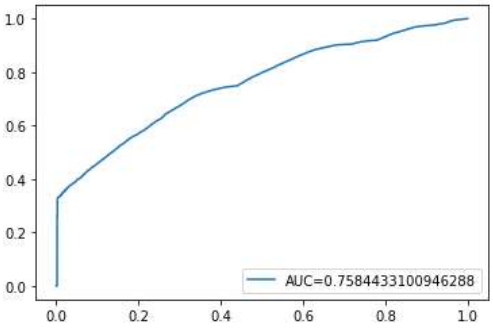
To determine how hotel booking parameters affect the likelihood of booking cancellation, we developed a logistic regression model. Based on the correlation coefficients between predictors and response (cancellations) determined in our EDA, we chose the most impactful predictors: hotel type, deposit type, lead time, and total number of special requests associated with the booking.

From this initial regression model, we constructed a confusion matrix to determine its accuracy; with a total of 26,609 correct predictions out of a possible 35,697, our accuracy came out to approximately 75%.

Confusion matrix



21697	608
8480	4912



Additionally, we constructed a ROC curve and determined the AUC, which provides the measure of our model's performance across all possible classification thresholds. Our greatest AUC was approximately 0.76.

SVM and KNN models for classification

To evaluate the performance of SVM and KNN models for hotel reservation prediction, we split the dataset into training and test sets. Using the training data, we built both SVM and KNN models and then tested their performance on the test data set. This approach allows us to compare the accuracy and efficiency of both models, to determine which is better suited for the task at hand.

SVM

It is true that SVM can be 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, i.e., 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.

Cross-validation is a widely used method for evaluating the performance of machine learning models. It involves dividing the dataset into two parts: a training set used to fit the model, and a testing set used to evaluate its performance. 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, where the model is too closely fitted to the training data and performs poorly on new, unseen data.

To optimize the SVM model's hyperparameters, C and sigma, we experimented with different values until the best combination was identified.

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 kknn model, we performed cross-validation. First, we set the maximum value of k to 30 and created an array to store the misclassification rates.

Next, we looped through each value of k to run 10-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.

Next Steps

Future analysis of **logistic regression** will center on maximizing accuracy and associated ROC AUC, potentially by including or excluding factors in our model. We will also interpret our models' coefficients and parameters.

For our other classification models, **SVM and KNN** should be further tuned with optimal hyperparameters. From here, accuracies and runtime efficiencies can be compared to arrive at an optimal model.

Holt-Winters forecasting and exponential smoothing will be performed to determine periodicity and trend in booking cancellations throughout the year. We have noticed in our initial models that the

higher the frequency, the higher the error. We are going to continue testing the model's accuracy for various forecasts in the future.

Additionally, we will be performing a cross-sectional analysis of all our classification models (logistic regression, KNN, and SVM) to balance maximized model performance with interpretability/meaningfulness.

Literature Survey

There are quite a few published research articles about hotel booking cancellations. Most of the time, hotels can predict cancellations due to unexpected events such as natural disasters, pandemics, and civil unrest. However, challenges arise when predicting booking cancellations near term (4-5 days). Researchers have employed various models to predict booking cancellations. For example, PNR (Personal Name Record) data has been utilized to create A.I. (Artificial Intelligence) models (SVM) to forecast near-term cancellations (4-7 days) with ~70% accuracy (Falk & Vieru, 2018). Another approach utilizes a Hotel PMS (Property Management System) (Antonio et al., 2017) and a customer PNR to create a machine learning model (decision trees & forest) to predict cancellations with AUC > 0.90 (Sánchez et al., 2020).

Future research should attempt to implement and utilize variables from data sources such as weather, competitive intelligence, and other hotel types to potentially improve the models' prediction accuracy.

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