Machine Learning Final Project

No *Room* for Doubt: Demystifying Hotel Cancellations

Aayoshi Dutta

Executive Summary

This project is about exploring the factors that influence hotel cancellation rates, a problem affecting revenue and forecasts in a market valued at over $230 Billion. We applied a range of machine learning models to our 36,000+ observations to predict cancellations based on a variety of guest and booking attributes. Our Random Forest model provides the highest accuracy at 8.87%. Further analysis uncovered key factors that influence cancellation likelihood, such as lead time, average price, length of stay, and medium of booking.

Our recommendations to decrease cancellation rates include focusing on generating guest buy-in and commitment to seeing through their stay, and having favorable rebooking policies for those guests who cannot stick to their original booking. Hotels can use our random forest model to predict cancellation risks, and use personalized communications to target high-cancellation-risk guests. By leveraging our statistical model, hotels can better manage cancellations and use data-driven insights to address one of the top industry challenges.

Word Count: 2215

Project Background

Hotel Cancellations in the U.S.

While some hotel room cancellations are inevitable, with guests not being able to travel due to sickness or other commitments, cancellations can affect revenue, demand forecasts, and general operations as they relate to staffing and hotel inventory management. Recent travel data has shown that hotel cancellation rates have skyrocketed to 20%[1]. Needless to say, for an industry with a market size of $230+ Billion in the U.S. alone[2], hoteliers and hotel consultants are tirelessly working on solutions to mitigate hotel cancellations. Nonetheless, hotels have a wealth of stored reservation records, data that has amounted to thousands of records of booking-specific information. Our project revolves around uncovering the story within the data with the goal of making recommendations to tackle the liability of hotel cancellations. If hotels can predict cancellations, based on certain guest or booking attributes, and hedge against those issues by implementing policies that encourage less cancellations, or encourage canceling bookings early enough the rebook the rooms, then hotels can safeguard a bigger proportion of revenue and derisk booking demand models which are affected by cancellations.

Current Mitigation Strategies

Hotels are already pursuing mitigation strategies, which help with cancellations, but may hurt new business. For example, having a cancellation deposit or penalty fee for late cancellations helps with cancellation rates but it may hurt booking rates; certain hotels are

[1] https://experience-crm.fr/en/where-do-cancellations-come-from/

[2] https://www.grandviewresearch.com/industry-analysis/us-hotels-resorts-cruise-lines-market

offering penalty-free cancellation policies, and for guests who value flexibility, those competitors are more appealing[3]. Hotels are also recommended to focus on direct booking rather than third party bookings because the latter have a historically higher cancellation rate. No resource, however, seems to be showing specific recommendations on other influencing factors for hotel reservation cancellations, specifically as they relate to booking specifics or guest attributes.

Data Description

Our dataset is a newly published dataset from Kaggle (originally posted December 2023)[4]. Some variables are already numerical by class, such as number of adults, number of children, number of weekend nights, number of week nights, lead time, repeated (whether it's a repeat customer), P-C (count of previous bookings canceled), P-not-C (count of previous bookings not canceled), average price, and number of special requests. Categorical data is converted to indicator variables, including type of meal, car parking space (yes/no), room type, market segment (online, offline, corporate, aviation, complementary etc), and our dependent variable, booking status (canceled vs not canceled). Our dataset contains 36,285 records, which allows for a sizable set of training as well as testing data.

Model Development

The dataset does not have records with missing values. Aside from categorical variable encoding, we prepared the dataset by updating some of the invalid reservation date records

[3] https://www.linkedin.com/pulse/protect-your-revenue-guide-how-hotels-can-prevent-manage-chargebacks
[4] https://www.kaggle.com/datasets/youssefaboelwafa/hotel-booking-cancellation-prediction

(substituting 02/28/2019 to 03/01/2019). Our exploratory data analysis includes some variable plotting, where we immediately notice clearly distinguishable differences between the boxplots for the attributes of canceled bookings, compared to those that were not canceled. On average, bookings affected by cancellations seem to involve weekend bookings. Furthermore, cancellations seem to be more frequent for reservations booked longer in advance, with greater lead time. Some variables are moderately to highly correlated with one another. What stands out the most, is that our outcome variable seems to be highly correlated with lead time, special requests, and average price.

Additional data preprocessing included scaling where applicable (for example, for our shrinkage methods, since the penalties are imposed on the coefficient values, scaling is essential, SVC, K nearest neighbor and logistic regression). The scaled data have a distribution of mean of 0 and standard deviation of 1. The standardization of scale helps each feature in SVC and KNN have equal measurement of distance among different data points, speed up convergence in gradient descent and standardized coefficient and therefore penalty in LASSO regressions. Overall, it gives all the data points an standardized proportion in the optimization process regardless of their inherent size. Regressions such as random forest and AdaBoost don't require scale for x variables.

Models Considered

Since the goal of our project is two-fold, on one hand predicting cancellation liabilities and on the other hand uncovering potential key influences of cancellation liability, our model of choice balances these aspects. We aimed for a model with high accuracy rates, as well as one that allowed us to interpret coefficients - with that, key drivers of cancellations.

As we predict our dependent variable, cancellation or not, our work is in the area of supervised learning. We developed 12 different models using 27 selected features/columns for cancellation prediction, including a regular logistic regression model, PCA-transformed logistic regression, shrinkage regression models (partial, Ridge, Lasso, Elastic Net), K nearest neighbors classification model, a classification regression tree, a random forest model, as well as XGBoosted, AdaBoost, and SVC models.

The key metrics we considered include mean squared error, a commonly used metric to quantify the errors our model makes looking at the out of sample predictions and comparing those to the out of sample actuals, as well as accuracy score, which is the proportion of correct predictions over all total out of sample predictions. Finally, to help visualize the performance we also included a visual on the confusion matrix, which would further clarify where our model was making the most mistakes / if it was correctly classifying cancellations vs non-cancellations better.

Our results, splitting our total records into 80% training data, and 20% testing data, are as follows:

Rank	Model	Mean Squared Error	Accuracy
1	**Random Forest**	**0.1113407744246934**	**88.87%**
2	XGBoost	0.1185062698084608	88.15%
3	Classification Regression Tree	0.14124293785310735	85.88%
4	AdaBoost	0.14330990767534793	85.67%
5	K Nearest Neighbors	0.15047540305911533	84.95%
6	Support Vector Machine	0.17913738459418493	82.09%
7	Ridge Logistic Regression	0.20118506269808462	79.88%
8	Lasso Logistic Regression	0.20159845666253273	79.84%
9	Non-penalized Logistic Regression	0.20187405263883149	79.81%
10	Elastic Net Logistic Regression	0.20118506269808462	79.77%

11	Partial Lasso Logistic Regression	0.2077993661292545	79.22%
12	PCA Logistic Regression (5 components)	0.2691194708557255	73.09%

This leads us to opt for the random forest model, though XGBoosted Tree, and AdaBoost are close running-ups with lower prediction errors as well. These models do utilize similar regression techniques so that they align with an almost equivalent performance. That said, random forest had the highest accuracy and lowest misclassification rate (Please see Figure 1 in the appendix).

Random Forest Interpretations

Among all the models, we have Random Forest as the winner for the highest accuracy score of 88.87% and the lowest mean squared error of 0.11134. We took a deeper dive with the Random Forest model to see the Gini importance and permutation feature importance. The top 5 Gini important features are lead time, average price, special requests, number of weeknights, and number of weekend nights for Mean Decrease in Impurity which means that those are the top parameters that reduce the impurity of the children nodes the most. The top 5 permutation feature importance features are lead time, special requests, average price, online segment, and number of weekend nights. These features are confirmed to improve model accuracy for out-of-sample tests.

In addition, we completed a perturbation analysis for the top 10 variables from the feature importance and marginal effect analysis to see how each of the variables will affect the cancellation outcome holding other variables constant. Even though cancellation rate is not extremely sensitive to perturbed data among count of special requests, online and offline segments, number of weekend nights, parking space, and number of children according to our perturbation analysis, the marginal effect of weeknights, average price, and lead time are more likely to have a positive effect on the outcome, or, in other words, can lead to reservation cancellations. (Please see Figure 2.1 in the appendix).

We also used partial dependence plots to further understand the strength and direction of our most impactful predictor variables in a more holistic view. The benefit of these plots is that we can visualize the relationship between an independent variable and the dependent variable while averaging out all other variable effects. We conclude that a longer lead time leads to higher cancellation rates and more special requests lead to a decrease in cancellation rates. The plots also show a positive and moderately strong relationship between average price and cancellation - we conclude that higher-priced rooms are more likely to be canceled. Online bookings are more likely to be canceled, and bookings with more weekend nights are also more likely to be canceled. Lastly, total days of stay also show a positive albeit non-linear relationship with cancellation likelihood. (Please see figures 3.1, 3.2, 3.3, 3.4, and 3.5 in the appendix).

While some of these relationships do not exhibit a linear pattern, the plots still provide insights into how the predictor affects our cancellation likelihood, as well as a general strength and direction for each predictor. Knowing these predictors, we can make targeted recommendations for hotels to decrease their cancellation rates.

Recommendations

In our findings, we see that lead time, average price, and length of stay are positively correlated with higher cancellation likelihood. We cannot directly translate our findings around average price and length of stay into an actionable recommendation, as these are not attributes we can directly influence. However, given these higher priced bookings can be considered cancellation-risky, we do recommend hotels to focus their efforts on these bookings by utilizing personalized marketing to keep the guests interested and engaged all the way up to their actual stay. If all else fails, redirecting guests to rebook instead of cancel, by for example offering a flexible rebooking policy, can be a last resort to at least retain the expected revenue from a high-value booking, albeit it at a later date. When it comes to lead time, hotels can directly decrease cancellation risk by opening up reservations less time in advance of the actual stay date. While they may lose out on some new business with that, this revenue already comes with more uncertainty and can throw a wrench in booking forecasts, so this cost may actually be an investment in the long term.

Notably, online bookings are also associated with a higher risk of cancellation. Since online bookings are the preferred way to go for many customers, no recommendation on changing the accessibility to book online can be made in good faith. Nonetheless, we can recommend tweaking some online booking practices, for example offering nonrefundable rates at a discount, which forces guests to have a higher commitment on following through with their stay. We also suggest offering loyalty programs that are visible in the early stages of booking - this is another attempt at generating that guest buy-in that will decrease cancellation likelihood.

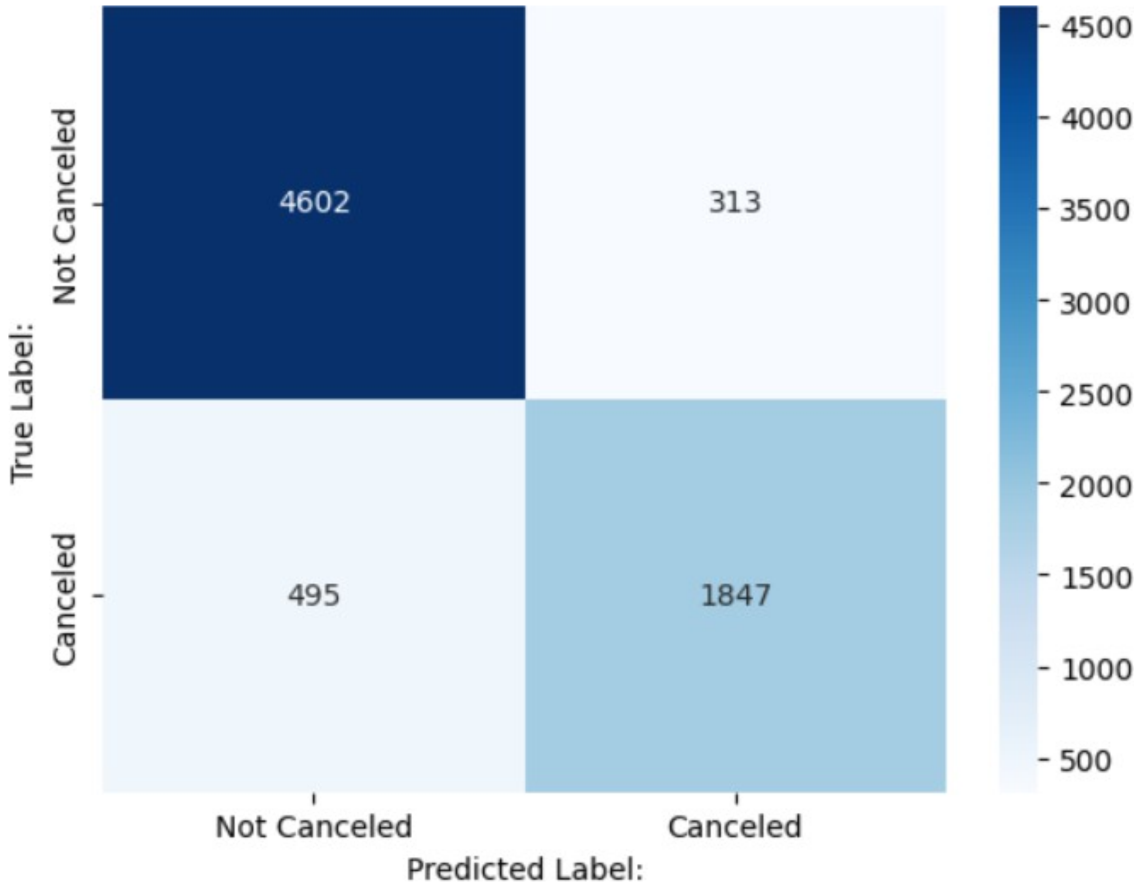
Conclusion

While hotels continue to deal with a cancellation rate spike issue, simply imposing harsh cancellation penalties will not do the trick. In our project, we examined 35,000 records of bookings to uncover any useful predictors of cancellation rates. We preprocessed this data appropriately to be able to run it on different models, including but not limited to shrinkage regressions, random forest and boosted trees, and K nearest neighbor. Our models were trained and tested on the same splits, and we compared mean squared errors as well as accuracy scores to select the best model. The random forest prediction model makes 88% accurate predictions on unseen data, and could be leveraged by hotels to predict individual customer cancellation rates and apply targeted mitigation strategies.

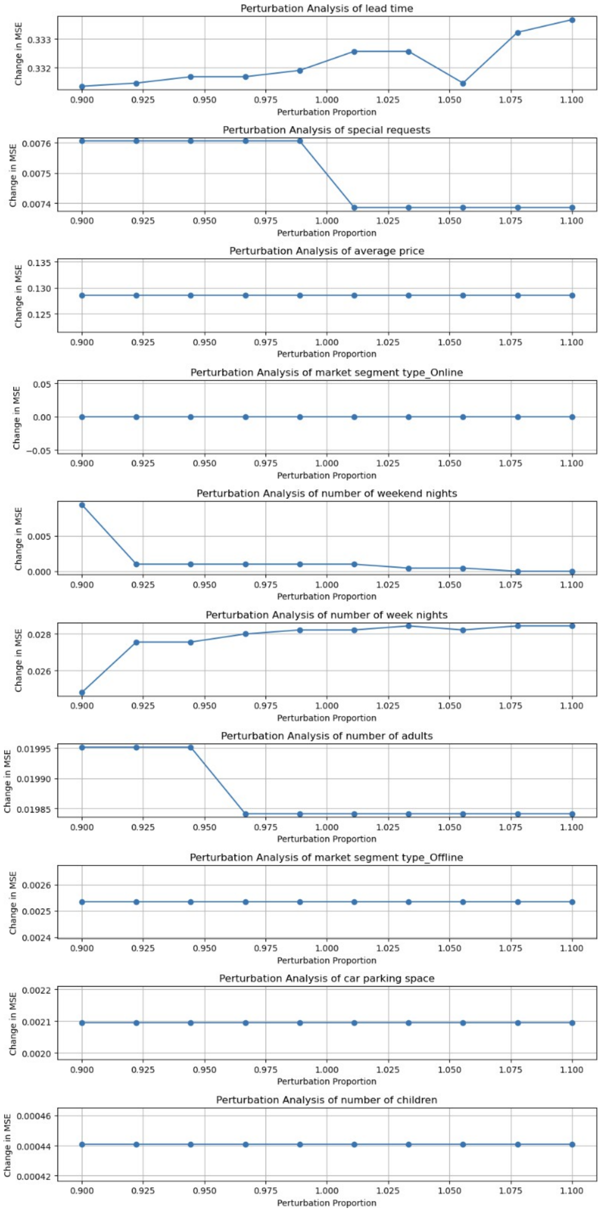
From our model analysis, we also learn that hotels need to diversify their cancellation prevention efforts, with the objective of having a favorable public image, making a flexible impression, while still decreasing overall cancellation rates. Cancellation policies need to focus on offering incentives for customers to rebook instead of cancel, and other guest buy-in initiatives such as nonrefundable but discounted room offerings, or better marketing for loyalty programs, are intended to help decrease cancellation rates.

Appendix

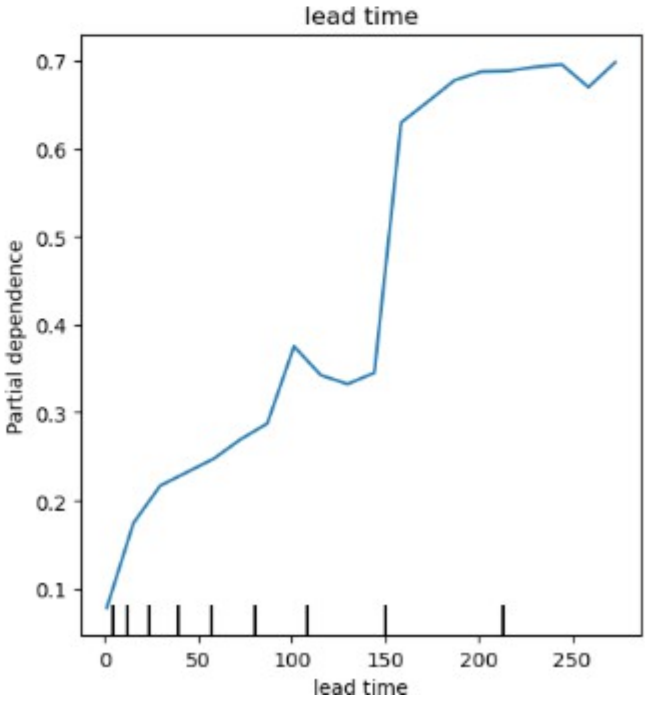
Figure 1 - Random Forest Model Confusion Matrix



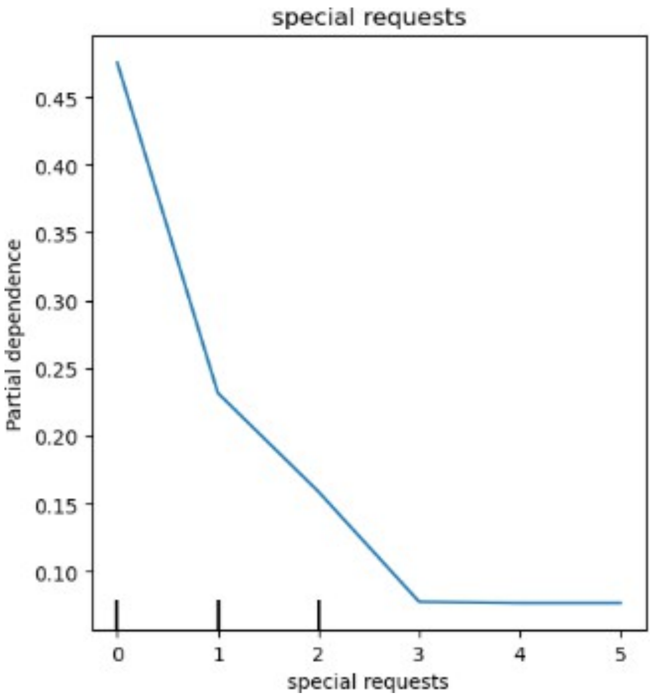
2.1


Perturbation Analysis of lead time
Change in MSE
0.332
0.333
Perturbation Proportion
0.900
0.925
0.950
0.975
1.000
1.025
1.050
1.075
1.100
Perturbation Analysis of special requests
Change in MSE
0.0074
0.0075
0.0076
Perturbation Proportion
0.900
0.925
0.950
0.975
1.000
1.025
1.050
1.075
1.100
Perturbation Analysis of average price
Change in MSE
0.125
0.130
0.135
Perturbation Proportion
0.900
0.925
0.950
0.975
1.000
1.025
1.050
1.075
1.100
Perturbation Analysis of market segment type_Online
Change in MSE
−0.05
0.00
0.05
Perturbation Proportion
0.900
0.925
0.950
0.975
1.000
1.025
1.050
1.075
1.100
Perturbation Analysis of number of weekend nights
Change in MSE
0.000
0.005
Perturbation Proportion
0.900
0.925
0.950
0.975
1.000
1.025
1.050
1.075
1.100
Perturbation Analysis of number of week nights
Change in MSE
0.026
0.028
Perturbation Proportion
0.900
0.925
0.950
0.975
1.000
1.025
1.050
1.075
1.100
Perturbation Analysis of number of adults
Change in MSE
0.01985
0.01990
0.01995
Perturbation Proportion
0.900
0.925
0.950
0.975
1.000
1.025
1.050
1.075
1.100
Perturbation Analysis of market segment type_Offline
Change in MSE
0.0024
0.0025
0.0026
Perturbation Proportion
0.900
0.925
0.950
0.975
1.000
1.025
1.050
1.075
1.100
Perturbation Analysis of car parking space
Change in MSE
0.0020
0.0021
0.0022
Perturbation Proportion
0.900
0.925
0.950
0.975
1.000
1.025
1.050
1.075
1.100
Perturbation Analysis of number of children
Change in MSE
0.00042
0.00044
0.00046
Perturbation Proportion
0.900
0.925
0.950
0.975
1.000
1.025
1.050
1.075
1.100

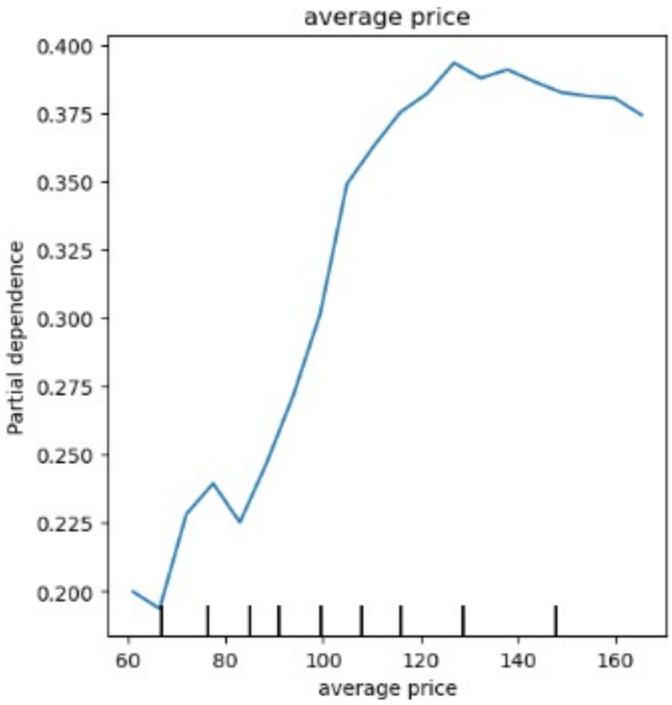
3.1 Partial Dependence Plot for Lead Time



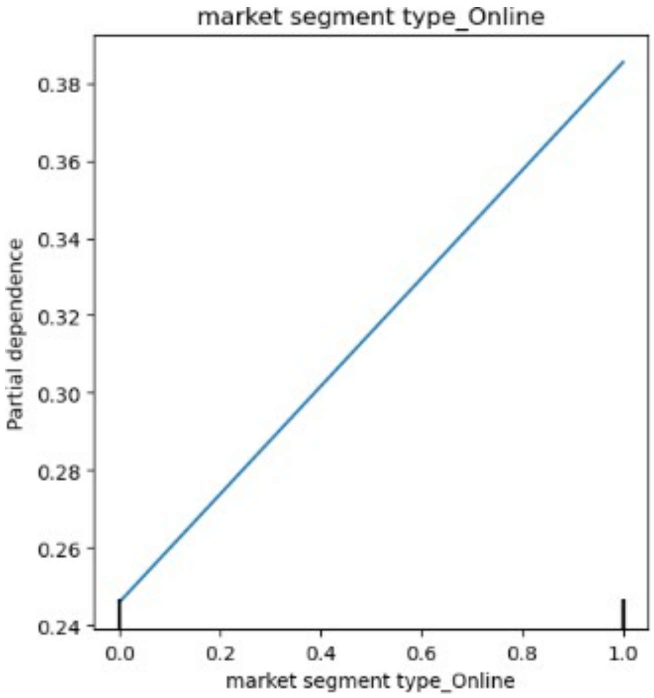
3.2 Partial Dependence Plot for Number of Special Requests



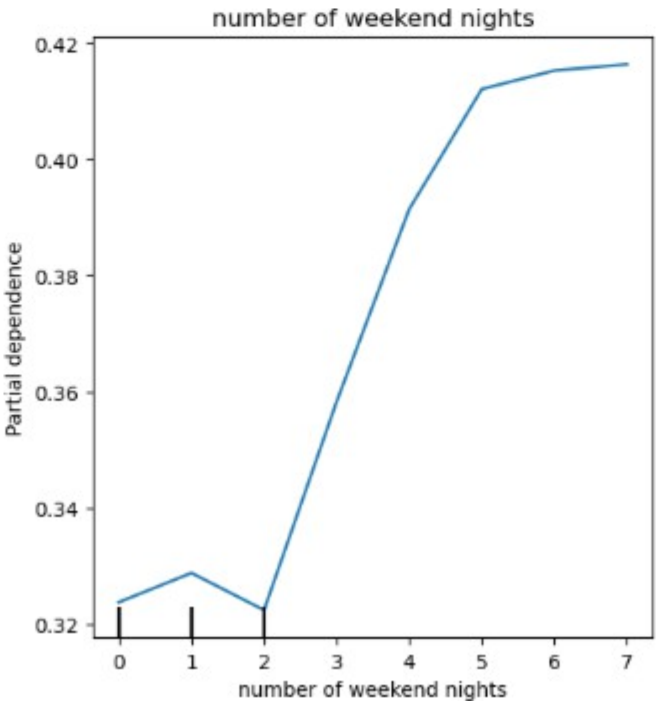
3.3 Partial Dependence Plot for Average Price



3.4 Partial Dependence Plot for Online Booking Segment



3.5 Partial Dependence Plot for Number of Weekend Nights


number of weekend nights
Partial dependence
number of weekend nights