

Modeling Hotel Demand

DNSC 4281: Pricing and Revenue Management Analytics

Matthew Wolf mattwolf@gwmail.gwu.edu, Charlie Buckman
charlie.buckman@gwmail.gwu.edu, Kaylyn Phung kaylyn.phung@gwmail.gwu.edu, Lauren
Kim lauren.kim1@gwmail.gwu.edu

Professor Zhengling Qi
George Washington University

December 8, 2025

Executive summary

The purpose of this project is to provide an avenue into how demand can be modeled for the hotel industry, with a focus on the significant features that contribute towards demand. The methodology utilized various machine learning techniques, and large language models to model demand and derive pricing information, given a dataset of booking entries for hotels within the industry. The findings of this work highlighted key factors that weigh into booking demand, and optimal pricing for a hotel in the industry. This project aims to pinpoint the variables that a hotel can manage in order to maximize their revenue.

Introduction

Pricing and revenue management has a critical role within the hospitality industry. Firms often have a multitude of factors to consider, both inside and outside of their control. The profit structure of a hotel in itself presents a unique challenge, as unlike physical goods, cannot be saved for future sale, a room unbooked for a night represents permanent loss of potential revenue. As a result of this, hotels must rely on accurate demand forecasting to make informed pricing, inventory, and capacity decisions. With complex confounding factors in action, such as seasonal fluctuations and various booking channels being utilized simultaneously, machine learning based models provide a solution through using past, real-world data to produce reliable predictions.

The research questions this project aims to answer are as follows: What variables carry the most strength in hotel booking demand? What is the optimal pricing strategy for a hotel in this industry? How can in-context learning optimize pricing for real world applications? The researchers aimed to identify exact variables that lead to a significant increase or decrease in hotel booking demand, provide a clear figure on what the optimal pricing strategy is for a firm within this industry, and leverage large language models to search for patterns and draw conclusions from the data in order to estimate the elasticity of demand for this industry. The final objective of this project is to generate actionable insights that can be applied within the real-world for hotels to enhance their forecasting accuracy, therefore improving revenue through optimizing pricing or potentially reducing variable costs.

Methodology and Data

This project uses a publicly available hotel booking dataset covering more than 40 features and 731 observations of vendors from around the world (after aggregation, 119,00 individual bookings). Since the data is recorded at a booking level, the first step was dedicated to aggregating all entries to a daily level, creating a time-series dataset. This allows the modeling of bookings per day as a function of price, or average daily rate (ADR), seasonality, hotel characteristics, and guest behavior. This daily panel includes mean ADR, total non-canceled bookings, arrival characteristics, room assignments, and customer types, which together provide a comprehensive view of hotel demand patterns and travel seasonality.

To clean and prepare the data, we used Python for transformation and feature engineering, and Matplotlib for visualizing key seasonal trends in the dataset, such as month-to-month patterns and weekday effects. The created seasonality indicators, month, day of the week, and holiday, along with numerical summaries such as average lead time and room-type proportions, and categorical encodings, were used to ensure that the models captured the operational realities of hotel demand. Several machine learning models were built, including a linear regression model, LASSO, Random Forest, Gradient Boosting, and ARIMA, using the same train–test splits and test metrics: RMSE, MAPE, and WMAPE. LASSO ended up being the strongest performing model due to its ability to shrink irrelevant coefficients and highlight the most predictive features of hotel demand, such as hotel type, room assignment, and seasonal factors.

In addition to building traditional machine learning models, In-Context Learning was also incorporated using a large language model to explore price elasticity. The model inferred the shape of the demand curve based on a small extracted sample from the hotel booking dataset, identifying optimal pricing for demand. This approach enhanced our understanding of what drives hotel bookings and how that interacts with pricing.

EDA

After data from the raw hotel demand dataset was aggregated to the daily level, exploratory data analysis was performed before modeling to determine which variables would be most influential on demand. First, an overall time series plot of bookings per day was created (A.1 in the appendix). There does not appear to be an upwards or downwards trend over time, and demand per day varies greatly day to day. The red line on the plot shows the largest outlier for bookings in the dataset, 448 (mean bookings per day was 150).

One of the most important aspects of the hotel industry is seasonality. Whether due to day of the week, month, or if a given day is a holiday or not, demand can vary significantly. Both ADR and bookings had boxplots created for comparison across day of week, month, and if the day was a holiday (A.2 in the appendix). Days were defined to be a holiday in the dataset if they were one of the holidays from a list of; New Year's Day, Good Friday, Easter Sunday, Labor Day, and Christmas Day. These holidays were chosen because of their widespread observation, and would allow us to capture the most information.

When looking at ADR by month, it appears to peak in the summer months of July and August, with lows in the Fall and Winter months of November, December, and January. Bookings by month did not have as much variability month to month when compared to ADR, however slight increases from the spring to early summer were exhibited, with low demand levels in the late fall and December and January. ADR by day of the week was relatively constant across all days (although there is a slight hike on Fridays). Bookings by day of week have stronger variation compared to ADR with the highest volumes coming on Thursday and Fridays. For both ADR and bookings, a given day being a holiday did not seem to suggest a significant difference in the level of price or demand.

In order to determine which variables would be used for the modeling stage, two matrices correlated with demand (bookings) were created. One contained variables carrying information about numerical columns (such as ADR, mean adults for bookings on a given day, average lead time for bookings on a given day, etc, A.3 in appendix) and the other contained information about categorical data (i.e. proportion of bookings which were for a resort hotel vs city hotel, A.4 in appendix). In order to extract the most useful information, variables whose absolute value of correlation with bookings were 0.15 or greater were used. 28 total numerical columns were used along with boolean indicators for seasonal factors.

Distributions for both bookings and ADR were observed by creating histograms of each. (displayed in appendix A.5). Both variables exemplified a relatively normal gaussian distribution pattern.

Analysis and Results (Models can be found in the modeling updated ipynb)

Models attempting to predict bookings (demand) were created for the following models, Standard Linear Regression, LASSO, Random Forest, Gradient Boosting, and ARIMA. Listed below are each models' respective parameters and performance. All models were fit on the same 80% split of training data and tested on 20% for evaluation. RMSE, MAPE, and WMAPE were the metrics used on the test set for comparison. Additionally, for each model, three different types of information were used. One with no confounder control (no monthly or daily or holiday control), one with daily and monthly information (but no holiday information), and one containing all seasonal information. Models with all the information performed best, so they will be shown.

Random Forest

The Random Forest model was trained with 100 trees, bootstrap sampling and sqrt feature-subsampling. Without confounder control, the model scored an RMSE of 43.09, but performance improved slightly when month and weekday controls were integrated, reducing RMSE to 41.02 and to 41.99 when holiday control was added. These parameters allowed the model to capture non linear relationships while still improving prediction accuracy.

ARIMA(1,1,1)

The ARIMA model learns how a previous days' change in bookings and forecast error helps predict the next day's change. This is done through autoregression, integration through first differencing and finally moving the average. This model received an RMSE of 45.68

Gradient Boosting

A gradient boosting regressor was utilized, with both confounder control and no confounder control. The uncontrolled model received an RMSE score of 40.66, while controlled achieved a slightly better RMSE score of 39.06. The hyperparameters utilized included a learning rate of 0.15 and 2,000 estimators.

Linear Regression

When the Linear Regression learns the input features without controls for months, weekdays, or holidays, the model overfit as it treats seasonality as true drivers of demand. After accounting for seasonal patterns, the model was more stable and produced reliable prediction of daily bookings, achieving RMSE scores of around 39 (confounder control) and 38.73 (holiday control).

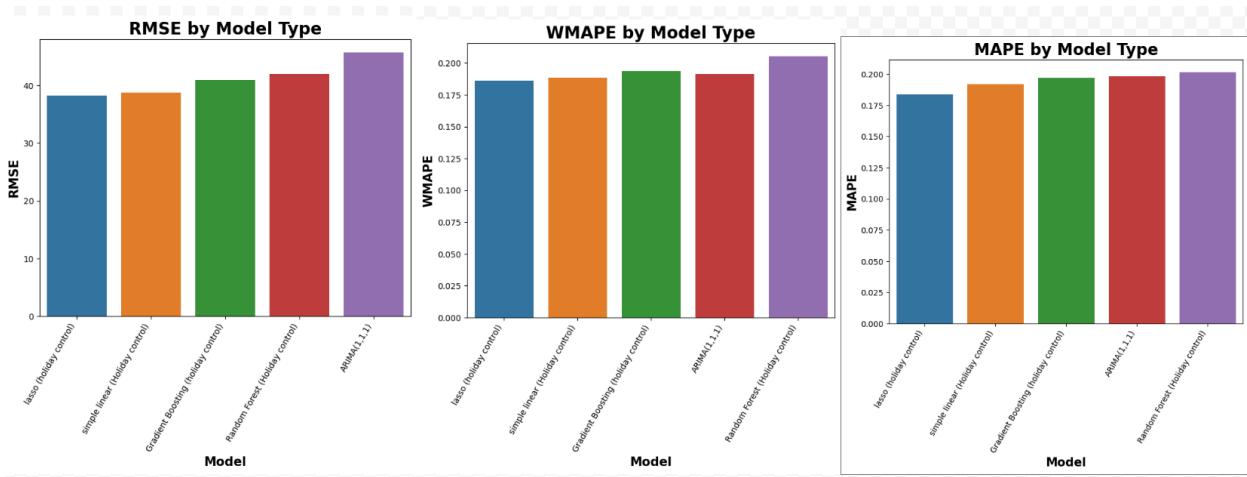
LASSO

The parameter which was used on the LASSO model was an alpha value of 0.1. The alpha controls the L1 regularization strength of the LASSO model, with 0 being closest to standard least squares, while 1 penalizes the most. After fitting the model and evaluating it on the unseen test set, the following metrics were output. A train R2 of 0.707, test R2 of 0.6456, test RMSE of 38.22, MAPE of 0.183, and WMAPE of 0.186. This was the best performing model created, largely due to the fact that LASSO's coefficient shrinking property significantly drives down variance without increasing bias relatively as much.

Additionally, due to the coefficient shrinking property of the LASSO, de-facto variable selection occurred. By examining the coefficients for variables used to fit the model, it was determined which features were most important for the task of predicting demand (more on that in the recommendations section and appendix). In total, of the 48 columns used to fit the model, 18 were shrunk to 0, and 30 were not shrunk to 0.

Model Comparison

The following bar charts summarizes performances across RMSE, MAPE, and WMAPE.



Surprisingly, the “simple” models (LASSO, linear regression) outperformed the more complex models such as gradient boosting, random forests, and ARIMA.

LLM Application

To better understand how hotel demand responds to changes in price, we implemented two complementary modeling approaches: an In-Context Learning (ICL) method based on

curated price–demand examples, and a traditional statistical model using LASSO regression. Although both approaches aim to capture the underlying relationship between nightly room rate (ADR) and demand, they differ significantly in methodology, interpretability, and the type of insight they provide.

The ICL approach operates without formal parameter estimation. Instead, it learns the structure of the demand curve directly from a small set of representative examples extracted from the hotel_bookings.csv dataset. We binned ADR into meaningful price ranges and computed a demand index representing the volume of non-canceled bookings within each bin. These pairs were provided to the model as part of the prompt. The model then inferred the price–demand relationship purely from these embedded examples, producing intuitive predictions that clearly reflected a downward-sloping demand curve: high booking volume around \$70–\$110, a moderate decline through \$120–\$150, and a steep drop beyond \$150. Because ICL relies entirely on pattern recognition within the prompt, the resulting “model” is transparent and easy to explain, though not statistical in the traditional sense and lacking formal measures of uncertainty or generalization performance.

In contrast, the LASSO regression model used in our main project follows a structured, data-driven approach grounded in regularized linear modeling. Unlike ICL, which focuses solely on price, LASSO considers multiple predictors simultaneously—such as ADR, lead time, distribution channel, arrival date, and other hotel characteristics. LASSO’s strength lies in variable selection: by shrinking less influential coefficients toward zero, it automatically identifies the most important drivers of demand while reducing risk of overfitting. This allows the model to quantify the marginal impact of price while controlling for confounding effects and capturing interactions among operational, seasonal, and customer variables. The LASSO model also provides formal performance metrics (e.g., RMSE, R²) that allow us to evaluate predictive accuracy objectively.

When comparing the two, both approaches detected the same core insight: price has a strong negative relationship with demand, with the highest sensitivity occurring once ADR rises above the mid-range. However, the LASSO model provides a more comprehensive understanding of demand behavior by situating price among many competing factors, whereas ICL provides a lightweight, example-driven summary of the demand curve. The ICL method is ideal for quick reasoning, scenario intuition, or communicating patterns to non-technical stakeholders, while the LASSO model is more appropriate for forecasting, optimization, and data-driven decision making.

Ultimately, the two methods complement each other. ICL offers a simple and interpretable demonstration of the core economic relationship between price and demand, making the pattern easy to visualize. The LASSO model deepens this understanding by quantifying effect sizes, selecting meaningful variables, and providing predictive rigor. Together, they reinforce the conclusion that pricing decisions must balance the trade-off between higher room rates and the significant decline in booking volume as prices increase.

Recommendations and Business Impact

As this project aims to enable a hotel to make informed operational decisions to maximize their revenue, recommendations have been set forth on how a firm within the hotel industry can utilize the analysis and results of the data. Managerial decisions can be optimized through staffing management. If a given day is forecasted to be of high, or low, demand, the firm can adjust the number of on-site staff, generating cost savings through optimizing hourly wage-based variable costs. The hotel can also manage their inventory assets, such as cleaning supplies, food or beverage, through an estimate of how many occupants will be utilizing its facilities in a given time period. When taking a look into daily rates, dynamic pricing can be optimized through adjusting the daily rate of a hotel based on projected demand. The LASSO model highlights how arrivals during the month of May demonstrated stronger demand, with 13.29 more bookings per day than the baseline month. The hotel can employ surge pricing for time periods such as the month of May. In contrast, the month of December has a small, negative impact on booking volume, with 3.32 fewer bookings when compared to the baseline month. This is an example of a time period where the hotel can employ discounts during periods of lower demand. Marketing and customer acquisition is also accounted for if the hotel utilizes the analysis to focus their promotions among customer types more likely to drive demand, for example with transient customers. Transient customers in particular were shown to boost demand by 74.50. The model also came to the conclusion that city hotel types see around 25 more hotel bookings, leading to the recommendation for the firm to focus promotional activity in these hotels to maximize this market segment (see appendix A.6 and A.7 for all LASSO coefficient outputs).

Limitations and Future Work

The dataset used provided a large amount of data, with realistic attributes that hotels most likely collect themselves, however, several limitations still affect the scope and accuracy of the analysis and results. First, the data has been fully anonymized, which removes valuable contextual information, such as hotel identity, geographic location and customer demographics. This limits modeling segment-specific behavior, and evaluating how specific markets influence demand. Another limitation in the data comes from the data spanning from 2015 to 2017, which reflects a period that may not represent the current industry, especially given the shift in travel behavior that many hotels experienced from 2019 onward due to the pandemic. This highlights that model insights may not generalize well to present-day demand patterns for this industry. Information of the competitive landscape of the hotels in this dataset is not known, also presenting a limitation within this project. With no data on nearby hotels or competitor accuracy, this project is limited in its ability to forecast demand within a market-based context, and provides forecasted demand down to an individual firm level instead. Lastly, this dataset lacks granular data, and therefore cannot be utilized for forecasting dynamic pricing. Since the data does not have information on day-to-day price adjustments, dynamic pricing strategies cannot be derived from the analysis in this project.

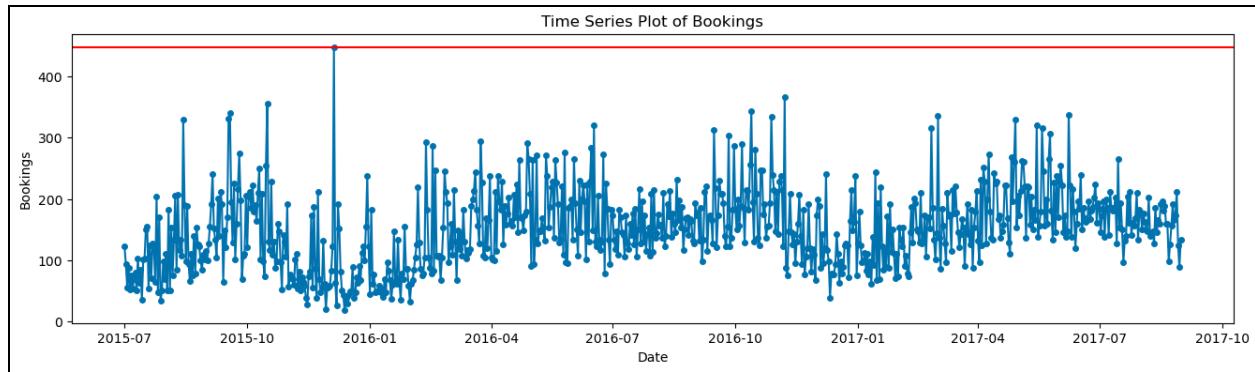
Future research can address these limitations by utilizing a more recent and granular database. A firm using this analysis using internal data may see more success and accuracy through having access to clearer, more detailed attributes. The addition of customer segmentation data can allow for more accurate demand forecasting, especially since the target market of a firm within this industry often comes from an international background, relative to the firm's geographic position. Additionally, neural network models, such as LSTM's can be utilized to help capture more long-term demand patterns, with the model learning seasonal demand patterns rather than controlling for them. This is especially pertinent, as a limitation within the data used came from the limited amount of observations within it.

References

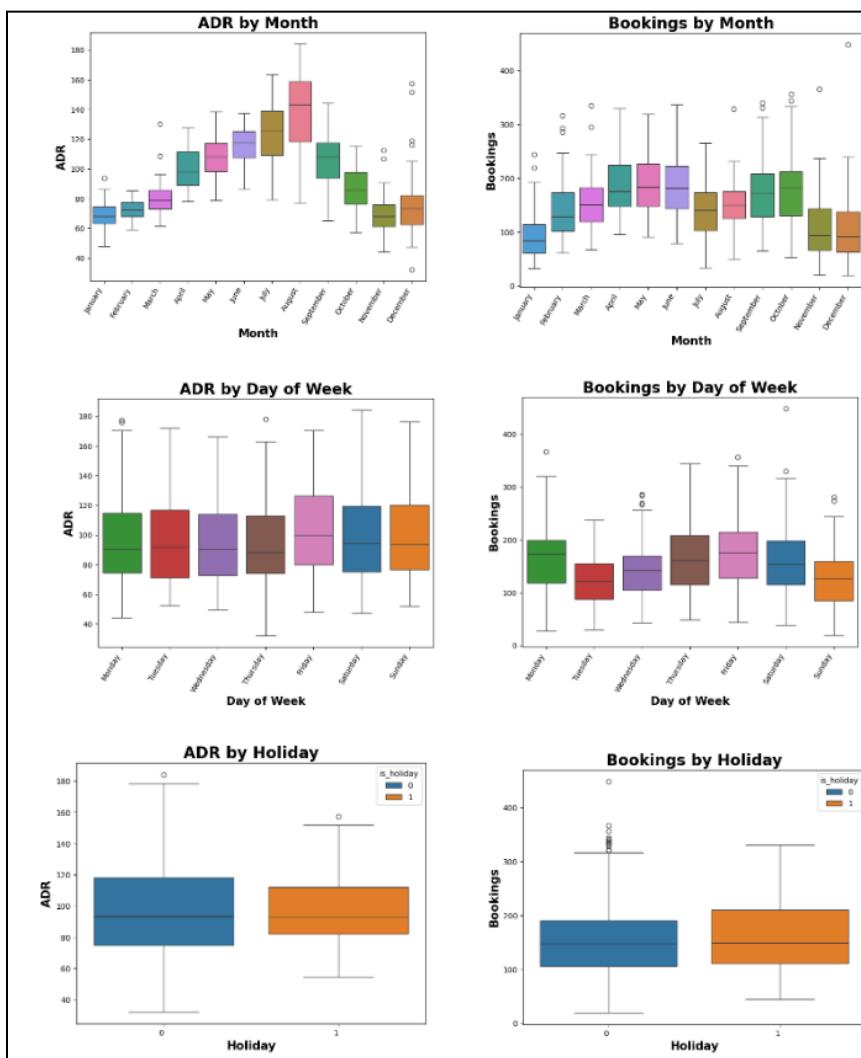
Antonio, N., de Almeida, A., & Nunes, L. (2019). *Hotel booking demand datasets*. Data in Brief, 22, 41–49. <https://doi.org/10.1016/j.dib.2018.11.126>

Appendices

A.1. Time Series Plot of Bookings



A.2. Seasonal Boxplots of ADR and Bookings



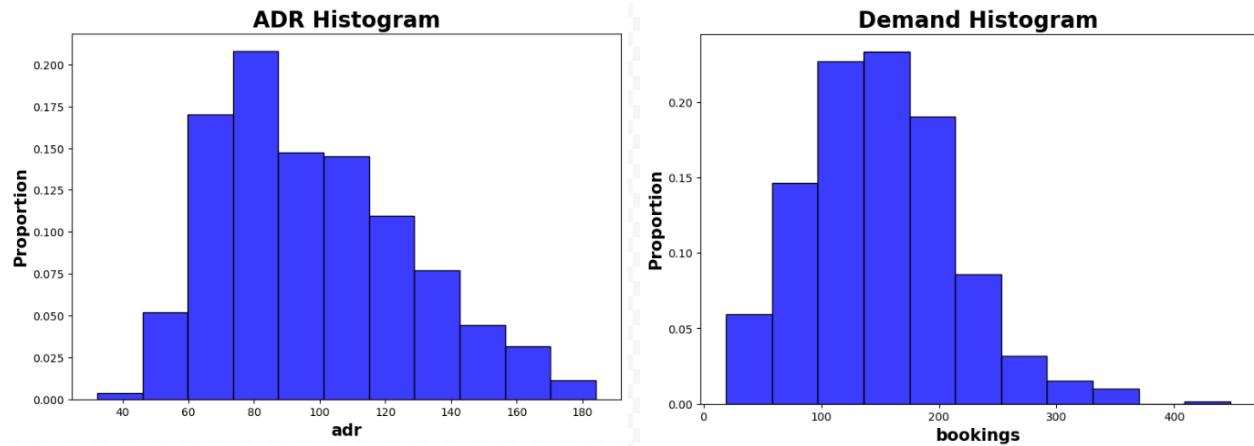
A.3. Significant Numerical Variables Correlated with bookings

	bookings
bookings	1.000000
mean_lead_time	0.529648
mean_required_car_parking_spaces	-0.457250
adr	0.296921
mean_total_special_requests	-0.259862
mean_babies	-0.177304
mean_adults	0.167166
mean_days_in_waiting_list	0.161465

A.4. Significant categorical Variables Correlated with bookings

	bookings
bookings	1.000000
market_Direct_proportion	-0.507497
deposit_No_Deposit_proportion	-0.487992
deposit_Non_Refund_proportion	0.483653
assigned_A_proportion	0.465651
hotel_City_proportion	0.458626
hotel_Resort_proportion	-0.458626
assigned_D_proportion	-0.439876
room_not_moved_proportion	0.394406
room_moved_proportion	-0.394406
market_Groups_proportion	0.382261
customer_Transient_Party_proportion	0.303278
market_Corporate_proportion	-0.270509
assigned_E_proportion	-0.252706
meal_BB_proportion	-0.236397
market_Complementary_proportion	-0.217285
customer_Group_proportion	-0.206361
customer_Contract_proportion	-0.203354
market_Online_TA_proportion	-0.200180
customer_Transient_proportion	-0.187779
market_Offline_TA_TO_proportion	0.172909
meal_HB_proportion	0.161493

A.5. Distributions of ADR and Bookings (demand)



A.6. Lasso Coefficients Not Shrunk

	coef	col
25	74.50	customer_Transient_Party_proportion
20	71.38	assigned_A_proportion
29	24.65	hotel_City_proportion
35	18.31	market_Groups_proportion
11	15.38	arrival_date_month_October
8	14.84	arrival_date_month_March
9	13.29	arrival_date_month_May
1	10.85	arrival_date_month_April
31	10.42	is_holiday
4	6.71	arrival_date_month_February
14	5.16	arrival_day_of_week_Monday
13	2.67	arrival_day_of_week_Friday
0	0.71	adr
43	0.21	mean_lead_time
30	-0.00	hotel_Resort_proportion
7	-0.12	arrival_date_month_June
42	-0.12	mean_days_in_waiting_list
3	-3.32	arrival_date_month_December
45	-13.83	mean_total_special_requests
5	-14.82	arrival_date_month_January
19	-18.91	arrival_day_of_week_Wednesday
18	-19.73	arrival_day_of_week_Tuesday
16	-19.76	arrival_day_of_week_Sunday
2	-24.81	arrival_date_month_August
6	-27.34	arrival_date_month_July
34	-47.41	market_Direct_proportion
39	-79.90	meal_HB_proportion
38	-122.40	meal_BB_proportion
27	-126.95	deposit_No_Deposit_proportion
23	-128.56	customer_Contract_proportion

A.7. Lasso Coefficients Shrunk

	coef	col
10	0.00	arrival_date_month_November
12	-0.00	arrival_date_month_September
15	0.00	arrival_day_of_week_Saturday
17	0.00	arrival_day_of_week_Thursday
21	-0.00	assigned_D_proportion
22	-0.00	assigned_E_proportion
24	-0.00	customer_Group_proportion
26	0.00	customer_Transient_proportion
28	0.00	deposit_Non_Refund_proportion
32	0.00	market_Complementary_proportion
33	-0.00	market_Corporate_proportion
36	0.00	market_Offline_TA_TO_proportion
37	-0.00	market_Online_TA_proportion
40	0.00	mean_adults
41	0.00	mean_babies
44	-0.00	mean_required_car_parking_spaces
46	-0.00	room_moved_proportion
47	0.00	room_not_moved_proportion