MSBA CAPSTONE PROJECT
HOTEL CANCELLATION PREDICTION

Professor Misuk Lee, PhD.
Co-authors: Ha (Hannah) Khuong, Yifan Xiang
December 10th, 2019

TABLE OF CONTENT

I. INTRODUCTION

Booking cancellations are one of the inevitable issues that arise, and they – especially last-minute cancellations – can wreak havoc with company's revenue management strategy if the company is not prepared for them. Thus, booking cancellation forecasting is of foremost importance to optimize revenue in the hospitality industry.

Customers can cancel room bookings due to a multitude of reasons, such as weather, rescheduled event, or illness. Understanding the cancellation patterns and predict this behavior can benefit firms in hospitality industry to minimize the possible damage on revenue. Once the cancellation prediction is known, hotels can counteract by using more aggressive customer acquiring strategies, such as room price reduction for products that are likely to have great number of cancellations. Better, once a cancellation pattern is understood, hotels can implement different policies to discourage this behavior as to how they see fit. Leeuwan (2018) and Petraru (2016) discussed about how firms can have overbooking strategies to cope with the cancellation behavior. However, this discussion about overbooking is outside of this paper's scope. Instead, we will suggest implementing some restricting cancellation policy based on our analysis results.

In this project, we tried to predict the cancellation rate of 2 international hotels located in New York city and Atlanta. Since these 2 hotels are of international scale, they have their bookings situation and cancellation pattern very different than middle and small scaled hotels. Their rooms are sold via multiple channels with different kind of contracts. The cancellation patterns also depend greatly on how the bookings were sold. In addition, due to the location difference, these 2 hotels demonstrated very unique cancellation patterns and will be discussed later in our paper.

After this introduction, this paper proceeds to the Literature Review to summarize prior research's discussion about statistical and machine learning application in predicting cancellation behavior. In Section III, you will get to know about our Data. First (part 1) is the General Information about our data set (New York city and Atlanta). Then, we dived into the Exploratory Analysis (part 2) to understand some important features in our dataset. Based on Exploratory Analysis, we would introduce some new features through Feature Engineering and explain how we processed our data (part 3). At the end of Section III, we provide tables to summarize features used in all of our models.

In Section IV, we explained about all models we used for prediction and provided result summary for each model. We discussed about the results of all our model in Section V, along with the project limitations and recommendation for hotels to have better revenue management strategies via controlling cancellation behavior.

II. LITERATURE REVIEW

Revenue management (RM) is key to any business that has relatively fixed capacity and time-variable demand. For the hospitality industry, successful revenue management strategies hinge on the ability to forecast demand and to control room availability. Specifically, accurate booking cancellation forecast is of foremost importance to estimate demand. Airline industry shares the same characteristics as hospitality industry in revenue management.

The literature regarding passenger cancellation in airline RM systems can be found Oren Petraru (2016) where traditional time series forecasting and Passenger Name Record (PNR) based forecasting for cancellation models have been stated. Past research found that cancellation fluctuates by time, type of passengers and passenger group size. All the models and tests run in Petraru's research is conducted in a passenger original destination simulator (PODS). Therefore, passenger behaviors are simulated, with the basis derived from prior studies. He then compared the behaviors simulated in the PODS model to real passenger behaviors collected from a real airline company and found that the cancellation rate generated is close to the real data. It is noticeable that in Petraru (2016) didn't include ticket price or any monetary value as variable in this study.

Eight tests were performed to compare the performance of the four cancellation methods in PODS. The results in all tests were consistent, and they raised several points. First, the common practice used by airlines today for cancellation forecasting is time series forecasting based on historical observations. Second, cancellation probabilities vary by time, by type of passengers and by number of passengers in group. Third, forecasting cancellation can help airlines increase their revenues. The benefits of cancellation forecasting are greater for airlines with higher cancellation rates. Among the four cancellation forecasting methods discussed by Petraru (2016), CM2 is the most favorable method for this project. It uses gross bookings and cancelation in each time frame and uses same cancellation estimates for both BIH and BTC. This method is more fitted in the dataset we have.

Rajopadhye (2001) also introduced a new method of giving the Revenue Management system "net demand", which is defined as the number of bookings minus the number of cancellations. The disadvantage of this method is that a part of the reality is neglected, which may result in adding uncertainty to the model. Leeuwen (2018) describes an approach how an industry, such as hospitality, can overcome this problem by creating a cancellation predictor. The key of this system is to controlled overbookings and relies on subsequent cancellations to keep the remaining number of bookings at the check-in date at, or just below, the capacity of the hotel by analyzing the behavior of cancellations.

Leeuwen (2018) does not discuss about overbooking strategy and its cost. However, as indicated by the literature, even with no overbooking applied, forecasting can help hotels to save 0.12% more revenue since they can adjust price to attract more demand.

Leeuwen (2018) used Passenger Name Record (PNR) approach to create classification models for 7 properties to predict cancellations of reservations. Four models were applied: Naïve Bayes, logistics regression, decision tree and random forest. According to the author, Naïve Bayes is efficient and easy to implement, but the algorithm may be too simplistic. Logistic regression provides probability outcome and is more robust to noise in the dataset, while the grade off is a higher bias. Decision Tree is intuitive to understand because of its graphical representation; however, the disadvantage could be over fitting. On the contrary, Random Forest can counter overfitting, but the run-time performance of Random Forest needs to be taken into consideration.

K-fold cross-validation was used to evaluate the performance of the models since it can overcome the threat of overfitting of a model. The author calculated confusion matrix, accuracy, precision, recall and F-score per model to compare the models. The results showed that the random forest is the algorithm with best performance. The dynamics within this algorithm are also consistent. The most important features to predict cancellations are Boolean Refundable and LeadTime. These findings confirm that feature engineering and selection are one of the most important steps in creating a useful model.

Implementation of the algorithms can benefit revenue managers in creating different sales strategy in different situations. The author suggested revenue managers to create a dashboard to indicate which reservations are likely to cancel for each single day in the future. The author also discussed the influence of competitors in the end of the paper. Although there is no information taken into account about this external factor and there is no literature about competitors influence, it could be innovative to see if there is a correlation between cancellations and competitors.

III. DATA

1. General Information

The dataset used in this research consisted of data from 2 hotels located in New York city and Atlanta. The booking information in a semi-aggregated form, with unique key as a combination of stay date, days prior and product type. Our booking data dated from 02/01/2012 to 04/29/2012. The last 3 weeks was isolated and used as testing data (04/09/2012 - 04/29/2012).

Variable Product Type is the product category that indicates the category that the room was booked under. They are Group, Membership Marketing, Tactical Marketing, Corporate, Government, Wholesale, Business Travel Agencies, Opaque, Other, Fenced, and Unfenced.

Product Type	Description
Group	Group booking. E.g. Rooms for conference, event venue, group travel.
Membership Marketing	Bookings made through membership marketing
Tactical Marketing	Tactical marketing offers for small group of customers
Corporate	Bookings based on contracted rates with big corporates
Government	Bookings based on contracted rates with government
Wholesale	Bookings based on contracted rates with wholesale agency
Business Travel Agencies	Business travel agency. E.g. Concur
Opaque	Selling unsold travel inventory at a discounted price. E.g. Priceline.com
Other	Other
Fenced	Advanced purchase reservations with non-refundable/non-cancellable condition
Unfenced	Regular refundable/Cancellable reservations

Table [1]. Description of product types.

Our variable for prediction is Cancelation Rate, which was calculated by on the book (OTB) to be cancelled divided by OTB. To validate prediction results, we used OTB to survive variable.

Room price was calculated based on daily revenue data but was end up not being included in our models for it can contain the effect of days prior and the hotels' price change policy to attract demands.

2. Exploratory Analysis

2.2. New York

2.2.1. Product Types

There are 43,920 observations in the period 02/01/2012 to 04/08/2012. The number of observations in each product type is comparable, yet they differ in their cancellation rate. This hotel has 11 product types (listed below). The product type with highest cancellation rate is Business Travel Agency (18.3%), and the product type with the lowest cancellation rate is Other (2.5%). The overall cancellation rate is 10.62% and was used as our Naïve prediction in our Metrics calculation.

Product Type	Records	Mean Cancellation Rate
Group	4,080	12.0%
Membership marketing	4,080	14.6%
Tactical marketing	4,080	6.0%
Corporate	4,080	14.8%
Government	3,660	11.7%
Wholesale	4,080	14.9%
Business travel agencies	4,080	18.3%
Opaque	4,080	2.7%
Other	3,540	2.5%
Fenced	4,080	3.0%
Overall	**43,920**	**10.62%**

Table [2]. Summary information of each Product Type (New York hotel).

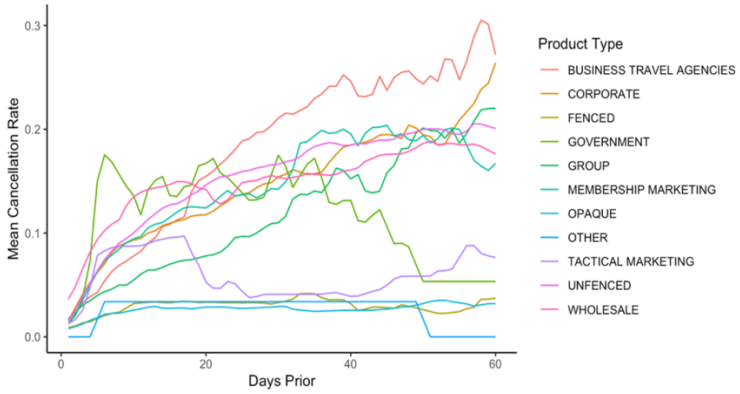

Figure [1]. Cancellation Rate trend of all Product Types in 60 days prior to Stay Date (New York hotel).

As expected, Product Types that have stricter cancellation policy such as Fenced and Opaque had lower cancellation rate all along 60 days prior. Product Types that have bookings through contracted rate like Corporate and Business Travel Agencies have really high cancellation rate. Yet, all Product Types have their cancellation rate trend decline at around 7 days prior to Stay Date.

In order to have Product groups that bear similar cancellation patterns, we decided to regroup Product Types. Based on the cancellation trend in 60 or 20 days prior, we regrouped the Product Types into 4 grouping with strategies as specified below.

Grouping Method	**Reasons**
Grouping 1	Grouped by trend level in 60 days
Grouping 2	Grouped by trend level in last 20 days
Grouping 3	Similar to grouping 2, but combined High Level and Low Level into 1 group
Grouping 4	Based on mean cancellation rate

Table [3]. Summary of Grouping methods (New York hotel).

Product Type	**Grouping 1**	**Grouping 2**	**Grouping 3**	**Grouping 4**
Group	Middle	Middle	High	High
Membership Marketing	High	Middle	High	Middle
Tactical Marketing	Middle	Middle	High	Low
Corporate	High	Middle	High	High
Government	Middle	High	High	Low
Wholesale	High	High	High	Middle
Business Travel Agencies	High	High	High	Middle
Opaque	Low	Low	Low	Middle
Other	Low	Low	Low	Low
Fenced	Low	Low	Low	Low
Unfenced	High	High	High	High

Table [4]. Product Type Grouping for New York hotel

2.2.2. Days prior

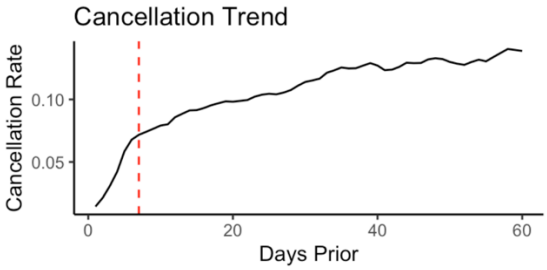

Figure [2]. Average cancellation rate across all product types (New York hotel).

There is a sharp decline in cancellation rate from day 7 priors to stay date. We created a dummy variable to capture this last week effect on cancellation. However, this variable did not reduce prediction error, so it was excluded from all models.

2.2.3. Days of Week

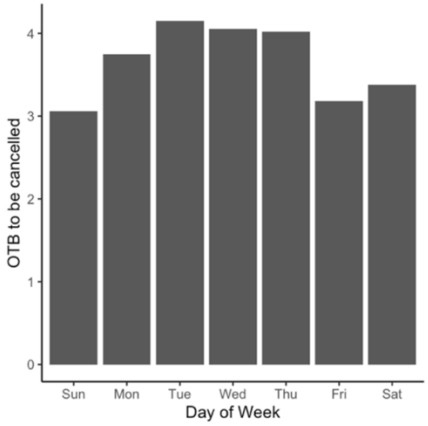

Figure [3]. Number of OTB bookings to be cancelled in different days of week (New York hotel).

For bookings with stay data on Friday, Saturday and Sunday, the number of bookings OTB that will be cancelled is lower than in other days of the week. We grouped into Weekend (Friday, Saturday, Sunday) and Weekday group to increase the number of observations in each day of week group for better prediction.

2.3. Atlanta

2.3.1. Product Types

There are 46,299 observations in the period 02/01/2012 to 04/08/2012. Observations with days prior equal then 0 have been removed, since they are meaningless. After removing those observations, we have 45540 observations left. Atlanta hotel has 9 product types (listed below). It doesn't have Tactical marketing product type and Wholesale product type. The product type with highest cancellation rate is

Membership Marketing (18.4%), and the product type with the lowest cancellation rate is Opaque (1.9%). The overall cancellation rate is 9.7%.

Product Type	Records	Mean Cancellation Rate
Membership Marketing	5,280	18.4%
Other	3,360	13.5%
Unfenced	5,340	13.4%
Corporate	5,340	13.3%
Group	5,340	11.1%
Government	5,340	8.2%
Business Travel Agencies	5,040	5.0%
Fenced	5,340	3.4%
Opaque	5,160	1.8%
Overall	**45,540**	**9.7%**

Table [5]. Summary information of each Product Type (Atlanta hotel).

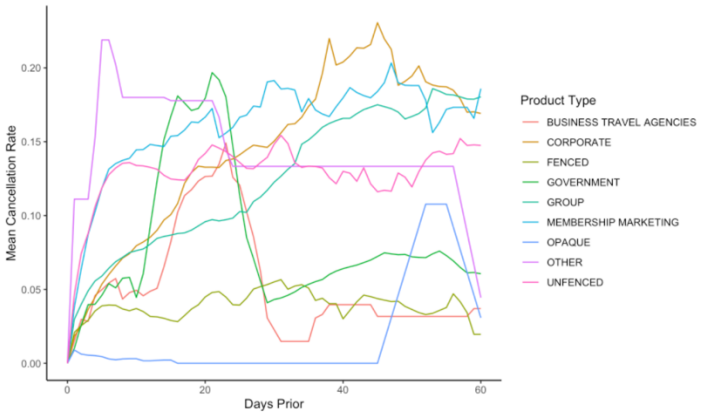

Figure [4]. Cancellation Rate trend of all Product Types in 60 days prior to Stay Date (Atlanta hotel).

As expected, all Product Types have their cancellation rate trend decline at around 7 days prior to Stay Date. However, unlike the hotel in New York, this hotel's cancellation rate patterns of different product type vary greatly as days getting closer to the stay date. We can see an irregular peak of cancellation rate for product type Government and Business Travel Agencies at days prior 20. Since it is known that this hotel usually holds big conferences and events, those irregular fluctuation can be explained as a sudden cancellation or date change of a big event.

In order to have Product groups that bear similar cancellation rate levels, we decided to regroup Product Types base on the cancellation rate level. Because the cancellation rate levels of different product type vary greatly as days getting closer to the stay date, and we want more accurate forecast as days prior gets closer to the stay date, we only consider the cancellation rate levels in last 30 or 10 days prior to the stay date.

It is also noticeable that Group product type has significantly high cumulative gross bookings. Since cumulative gross booking is an indicator of demand level, we didn't group Group product type with other product types.

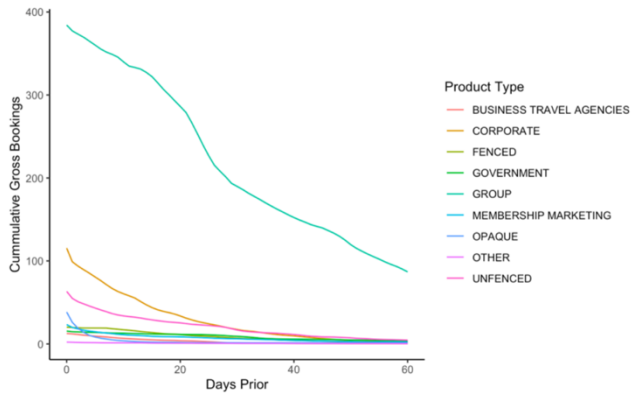

Figure [5]. Cumulative Gross Bookings trend of all Product Types in 60 days prior to Stay Date (Atlanta hotel).

Based on the cancellation rate trend in 30 or 10 days prior, we regrouped the Product Types into 4 grouping with strategies as specified below.

Grouping Method	**Reasons**
Grouping 1	Grouped by trend level in last 30 days
Grouping 2	Grouped by trend level in last 10 days (4 groups)
Grouping 3	Grouped by trend level in last 10 days (3 groups)
Grouping 4	Based on mean cancellation rate

Table [6]. Summary of Grouping methods (Atlanta hotel).

Product Type	Grouping 1	Grouping 2	Grouping 3	Grouping 4
Group	Group	Group	Group	Group
Membership Marketing	High	High	High	High
Other	High	High	High	High
Government	High	Middle	Low	Low
Unfenced	Middle	High	High	High
Corporate	Middle	Middle	Low	High
Business Travel Agencies	Middle	Middle	Low	Low
Fenced	Low	Middle	Low	Low
Opaque	Low	Low	Low	Low

Table [7]. Product Type Grouping for Atlanta hotel

2.3.2. Cancellation rate

We expected to see a strong correlation between room price and the cancellation rate, however, it turns out that there is nearly no correlation between these two variables. In fact, the graph below shows that cancellation rate of this hotel has weak correlation with all other factors. After we regrouped product types, we found out that the correlation between cancellation rate and other factors change by different days prior and by different groups. This indicates that the impactors of cancellation rate are different for different groups and in different time.

2.3.4. Days of Week

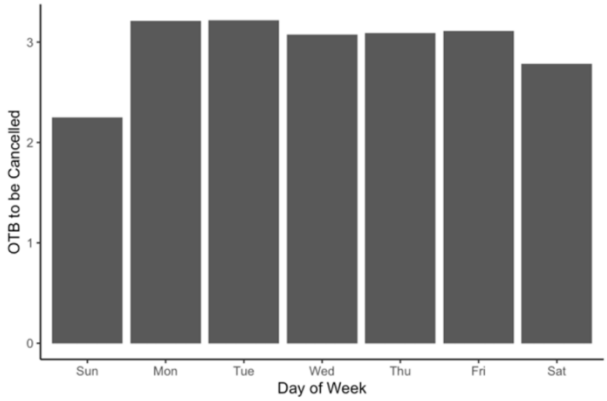

Figure [6]. Number of OTB bookings to be cancelled in different days of week (Atlanta hotel).

For bookings with stay data on Saturday and Sunday, the number of bookings OTB that will be cancelled is lower than in other days of the week. We grouped data into Weekend (Saturday, Sunday) and Weekday groups for better prediction.

3. Data Processing

3.1 Outliers Treatment

Abnormal trend of cancellation rate in Atlanta hotel are considered as outliers. An example of outliers is the peak of cancellation rate for product type Government and Business Travel Agencies at days prior 20. Those outliers are probably caused by the hold of emergency meetings or sudden cancel of special events or conferences. Those outliers can have great impacts on the forecasting. Most of those outliers happen in Government, Business Travel Agencies, and Opaque product groups. We used Tableau to find out stay dates stand out for Government and Business Travel Agencies in days prior 10 to 30, and also outliers for Opaque in days prior 40 to 60. We removed those data to avoid overfitting problem of our prediction models. After all, 2140 observations have been removed, we ended up having 33240 observations in our training dataset with a naïve prediction of cancellation rate equal to 9.91%.

3.2 Data Preprocessing

Scaling data is required for KNN, Neural Network and SVR models. Because our datasets contain features highly varying in magnitudes, units and ranges, and most of the algorithms use distance between two datapoints in computations, we need to normalize data. If left alone, these algorithms only take in the magnitude of features neglecting the units. The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes. Scaling data helps to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. StandardScaler from sklearn has been used to scale the data.

For KNN, Neural Network, SVR, Regression Tree, and Random Forests models, we need to change factors in the data frame into dummy variables. We used get_dummies function in Python to create dummy

variables. It can create one new variable for every level of the factor for which we are creating dummies. It appends the variable name with the factor level name to generate names for the dummy variables.

4. Features for Models

Field Name	Description	Included in Model?					
		Traditional	**Regression**	**KNN**	**Tree**	**Random Forest**	**Neural Network**
Product Type	Product Category	ATL & NYC	ATL & NYC	ATL & NYC	ATL & NYC	ATL & NYC	ATL & NYC
Day of Week	Day of week for stay date				ATL & NYC	ATL & NYC	ATL & NYC
Days Prior	Number of days between Stay Date and Booking Date		ATL & NYC	ATL & NYC	ATL & NYC	ATL & NYC	ATL & NYC
Days Prior Group	Days Prior divided into ranges 1-7, 8-14, 15-20, 21-27, 28-60	ATL & NYC			ATL & NYC	ATL & NYC	
Regrouped Day of Week	Weekend (Sat, Sun), and Weekday		ATL & NY	ATL			
Daily Gross Bookings	Daily Gross Bookings				ATL & NYC	ATL & NYC	
Daily Cancellations	Daily Total Cancellations				ATL & NYC	ATL & NYC	
Daily Net Bookings	Daily Gross Bookings minus Daily Cancellations				ATL & NYC	ATL & NYC	
Cumulative Gross Booking	Cumulative Gross Booking		NYC	ATL & NYC	ATL & NYC	ATL & NYC	ATL
Cumulative Cancellations	Cumulative Cancellations				ATL & NYC	ATL & NYC	
On-the-book (OTB) Bookings	Cumulative net bookings: cumulative Gross Bookings minus Cumulative Cancellations		ATL	ATL	ATL & NYC	ATL & NYC	ATL & NYC
Cancellation Rate	on the book (OTB) to be cancelled divided by OTB	This is dependent variable for all models					
OTB to be Cancelled	Bookings that are on the book that will be cancelled, which is unknown at the given days prior.	This is what we are going to predict					
OTB to Survive	Bookings that are on the book that will not be cancelled. Equals to OTB minus OTB to be cancelled.	Used for Metrics calculation					

Table [8]. Summary table of variables used in all predictive models (both hotel).

IV. MODEL

1. Error Metrics

Variable used in our error metrics is OTB to Survive, which is the number of bookings that are OTB and will not be cancelled. The reason for choosing this variable as our metrics variable is OTB to survive is a meaningful measure for hotels. Hotels need to know how many rooms that will be occupied and how many will be available to decide on appropriate actions to minimize financial damages from cancellation. The second reason is to weight variance of predicted cancellation rate by OTB of days prior. The cancellation rate in days further from Stay Date is smaller than in days closer to Stay Date. This fact makes the variance in cancellation rate become more significant than it should be in these days' prediction.

As discussed in the Exploratory Analysis, how many days prior to the stay date that customers book the hotel is highly correlated with the cancelation rate. To evaluate our models' predictive capabilities in a times series basis, we grouped days prior variables into 5 different groups. For the last 28 days, we grouped every 7 days into a group. For the first 32 days, because the accumulated booking amount is usually low in the first 32 days, we just simply grouped them into one group.

We evaluated both the in-sample and out-of-sample performance for different models. For the in-sample performance evaluation, we used the model we built to forecast the cancellation rate of the training dataset to see how effective the algorithms are in reproducing the data. For the out-sample performance evaluation, we used the model to forecast the cancellation rate of the test dataset. By comparing the in-sample and out-of-sample performance, the problem of the model caused by overfitting can be exposed.

We used 3 different error measures to assess the performance of the different models by different days prior categories. They are specified as follows:

1. A_t: The actual value in the time series at time t.
2. F_t: The forecasted value in the time series at time t.
3. N: The length of the time series.
4. F'_t: The forecasted value of a benchmark (Naïve) model at time t.

The following error measures are calculated:

- **Mean Absolute Error (MAE):** Measures average absolute deviation of forecast from the actual

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|A_t - F_t|\,.$$

- **Mean Absolute Percentage Error (MAPE):** The average absolute percentage of errors to the actual values.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|.$$

  It's worth noting that since the denominator can't be 0, we ignore those records with forecasting on the book to survive equal to 0.

- **Mean Absolute Scaled Error (MASE):** Measures the accuracy of forecasts. It is the mean absolute error of the forecast values, divided by the mean absolute error of the in-sample one-step naïve forecast.

$$MASE = \frac{\sum_{t=1}^{k}|A_t - F_t|}{\sum_{t=1}^{k}|A_t - F'_t|}$$

2. Traditional Model

This is a common way that hotels predict future cancellations. The predicted cancellation rate is the mean of cancellation rate by Product Type and Days Prior of the training data. The set of cancellation rate means are then applied on the testing data to make prediction.

2.1. New York

The traditional model performed best on the original Product Type grouping.

Days Prior	In-sample Metrics			Out-sample Metrics		
	MAE	MAPE	MASE	MAE	MAPE	MASE
0-7	1.51	3.76%	22.20%	1.85	3.10%	21.04%
08-14	2.00	6.26%	49.92%	2.63	4.96%	42.17%
15-20	2.02	6.97%	58.19%	2.65	5.71%	52.32%
21-27	1.80	8.21%	60.03%	2.69	7.10%	60.02%
28-60	1.55	12.46%	73.39%	2.56	12.70%	63.63%

2.2. Atlanta

The traditional model performed best on the first regrouping method.

Starting from days prior 8 to days prior 60, we found that MAE is not in consistent with MAPE for both in-sample forecasting and out-of-sample forecasting. This is because MAPE's denominator cannot be 0. Thus, we filtered out rows with no actual OTB to survive. There is 130 observations in test datasets with no actual OTB to survive in days prior 8 – 14. In this case, we mainly focus on the result of MAE and MASE, we will only use MPAE as a reference.

Usually, as days prior increase, the error rates will increase. However, this is not the pattern we observed for MAE. This is because in the first 30 days, the booking amount is small, so the forecast OTB to survive is small accordingly, and the absolute difference between actual OTB to survive and predication is small.

Days Prior	In-sample Metrics			Out-sample Metrics		
	MAE	MAPE	MASE	MAE	MAPE	MASE
0-7	1.70	6.41%	46.63%	1.63	5.26%	45.73%
08-14	1.83	9.04%	77.94%	1.77	7.73%	80.21%
15-20	1.96	13.56%	93.31%	1.98	12.00%	94.89%
21-27	1.75	14.45%	101.71%	1.80	13.55%	99.49%
28-60	1.19	16.01%	103.11%	1.61	14.47%	156.89%

3. Fractional Polynomial Regression Model

Regression analysis helps us understand how a set of variables influence on the value of a variable of interest (dependent variable). The regression line is found by finding the line that most closely fits the data by minimizing sums of squared distances between the true data and the regression line. The predicted

values are on this regression line. In this problem, we used Fractional Polynomial transformed regression of New York City hotel and Logarithmic regression of Atlanta hotel.

3.1. New York

We power transformed the regression with a fractional polynomial term as we observe curvature in cancellation rate trend (Royston & Altman, 1994). Multiple fractional power in was tried and the power 0.2 fitted our curve the best.

Because different product types have different patterns of cancellation rate along days prior axis, an interaction term of Product Type and Days Prior was included. Cumulative Gross Bookings had a negative correlation of 0.33 with Cancellation Rate. The formula included interaction term of the Regrouped Day of Week and Cumulative Gross Bookings to capture the effect of Weekdays versus Weekend.

The best regression model for New York hotel belongs to the Grouping 3 of product type. Even though this Grouping did not have the best R-squared, its performance in MASE Metrics was the best. The most important features in this regression models are Cumulative Gross Bookings, fractional polynomial transformed Days Prior, Product Type and the interaction of Product Type and Days Prior.

Variables	Estimate	t value	Pr(>|t|)
Cumulative Gross Bookings	-6.45E-05	-5.795	6.90e-09 ***
Regrouped Day of Week (weekend)	3.18E-03	1.802	0.0716 .
Days Prior$^{0.2}$	1.14E-01	14.937	< 2e-16 ***
Product Type Grouping 3 (Low Cancellation)	-4.42E-02	-12.885	< 2e-16 ***
Days Prior	8.48E-05	0.628	0.5298
Cumulative Gross Bookings * Regrouped Day of Week (weekend)	3.02E-05	-1.741	0.0817 .
Product Type Grouping 3 (Low Cancellation) * Days Prior	-2.03E-03	-20.861	< 2e-16 ***

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Adjusted R-squared: 0.122

As expected, the fractional polynomial regression somewhat captured the declining curve at days closer to Stay Date. The different level of 2 cancellation trends was also shown in the result of this regression.

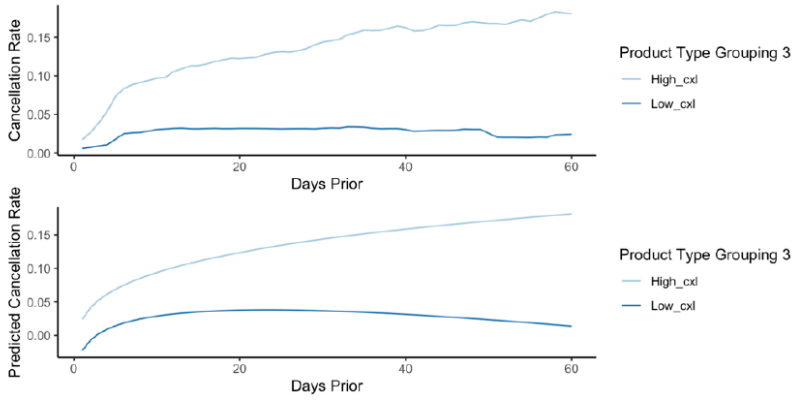

Figure [7]. Mean cancellation rate by Days Prior in actual data and Regression prediction (New York hotel).

3.2. Atlanta

Similar to the regression model of New York hotel, we powered transformed the regression with a logarithmic Regression term. The best power we found that fit best for Atlanta hotel is -0.21. We included the interaction term of Product Type and Days Prior, the interaction term of the Regrouped Day of Week and the Product type as well in the regression model for Atlanta hotel.

The regression model without regrouping product type, which is our benchmark, turned out to have the highest R-square (11.57%). This indicates this model can explain 11.57% of the variability of the response data around its mean. The regression model with best prediction result is the one with the first regrouping method. The R-square of this method is 6.2%. Even though this Grouping did not have the best R-squared, its performance in MASE Metrics was the best. The most important features in this regression models are Regrouping Product Type, Regrouping Day of Week, OTB, transformed Days Prior, and the interaction term of Days Prior and Regrouping Product Type, the interaction term of regroup Day of Week and Regroup Product Type.

```
Coefficients:
                                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 1.721e-01  1.386e-02  12.416  < 2e-16 ***
I(days_prior^(-0.21))                      -1.426e-01  1.627e-02  -8.767  < 2e-16 ***
product_type_regroupHigh_cxl                4.815e-02  8.981e-03   5.361 8.33e-08 ***
product_type_regroupMed_cxl                 8.237e-03  8.584e-03   0.960  0.33726
product_type_regroupLow_cxl                -5.112e-02  9.215e-03  -5.548 2.91e-08 ***
days_prior                                  1.418e-03  1.993e-04   7.112 1.17e-12 ***
dowgroupweekend                            -1.036e-03  6.242e-03  -0.166  0.86816
OTB                                        -7.187e-05  1.536e-05  -4.680 2.88e-06 ***
product_type_regroupHigh_cxl:days_prior    -1.841e-03  2.059e-04  -8.939  < 2e-16 ***
product_type_regroupMed_cxl:days_prior     -1.252e-03  1.994e-04  -6.280 3.43e-10 ***
product_type_regroupLow_cxl:days_prior     -2.236e-03  2.138e-04 -10.458  < 2e-16 ***
product_type_regroupHigh_cxl:dowgroupweekend -2.351e-02  7.361e-03  -3.194  0.00141 **
product_type_regroupMed_cxl:dowgroupweekend  -7.264e-03  7.223e-03  -1.006  0.31460
product_type_regroupLow_cxl:dowgroupweekend   5.352e-03  7.719e-03   0.693  0.48813
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.18 on 33226 degrees of freedom
Multiple R-squared:  0.06209,   Adjusted R-squared:  0.06172
F-statistic: 169.2 on 13 and 33226 DF,  p-value: < 2.2e-16
```

We found out that we have some negative prediction of cancellation rate in last 7 days. At beginning, we thought it was caused by Opaque product type because it is the only product type that the cancellation rate doesn’t decrease in the last 7 days . We tried to separate Opaque product type from rest of product types. We created regression model only based on the product types other than Opaque, and for Opaque we used naive cancellation rate of Opaque as its forecast cancellation rate. However, this didn’t improve our forecasting, we even got worse R-square and higher MASE. In the end, we chose not to isolate Opaque in our prediction, instead, we manually changed all negative prediction of cancellation rate into 0.

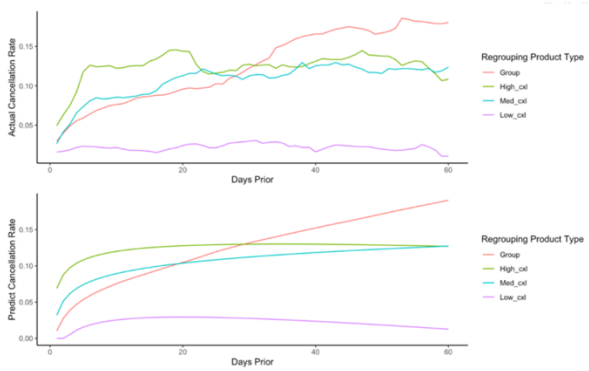

Figure [8]. Mean cancellation rate by Days Prior in actual data and Regression prediction (Atlanta hotel).

Days Prior	In-sample Metrics			Out-sample Metrics		
	MAE	MAPE	MASE	MAE	MAPE	MASE
0-7	1.94	6.31%	51.25%	1.74	5.30%	48.16%
08-14	1.79	8.75%	73.41%	1.81	7.89%	84.09%
15-20	1.84	11.51%	85.18%	1.99	11.88%	96.48%
21-27	1.66	13.33%	94.26%	1.98	13.35%	110.89%
28-60	1.06	15.19%	84.68%	1.52	13.97%	148.68%

4. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is considered a lazy learner as it does not generate a generalized pattern of the training data. It analyzed training data only when testing data is presented. KNN algorithm calculates Euclidean distance to determine closeness of datapoints. Based on this distance, k number of observations is considered one another's 'neighbor'. In Regression KNN, a predicted value is the mean value of the top k neighbors. KNN is a simple and fast algorithm, but susceptible to local patterns. The smaller the k, the more complex the model and poses a risk of overfitting.

As the algorithm uses distance to determine neighbors, categorical variables were transformed into dummy variables. Scaling is important for KNN algorithm because we want all features to be equally important in distance calculation, regardless of the variable scale. Therefore, all features were scaled for this algorithm.

4.1. New York

The best KNN model used Product Type with grouping method 1. The best k (chose based on lowest RMSE) was 20. The relatively small k indicates this is a quite complex model and have a tendency for overfitting. KNN seemed to overfit observation in Days Prior further from Stay Date (day 20-60) as out-sample MASE was higher than in-sample MASE at around 10 percentage-point.

Days Prior	In-sample Metrics			Out-sample Metrics		
	MAE	MAPE	MASE	MAE	MAPE	MASE
0-7	1.66	3.72%	24.30%	2.39	3.80%	27.17%
08-14	1.74	5.79%	43.08%	2.60	5.48%	41.78%
15-20	1.66	6.37%	47.67%	2.54	5.76%	50.04%
21-27	1.45	7.09%	47.95%	2.65	6.72%	58.77%
28-60	0.96	9.80%	45.08%	2.24	12.02%	55.26%

Pattern-wise, KNN generated a close approximation trend of 3 regrouped Product Types as seen in the figures below. KNN also successfully picked up the declining curve of lead day 7 to Stay Date which was prominent in the Middle Level Cancellation group. The higher mean variation seen in predicted trend in Low Cancellation Group might be due to the complexity of a low k KNN model.

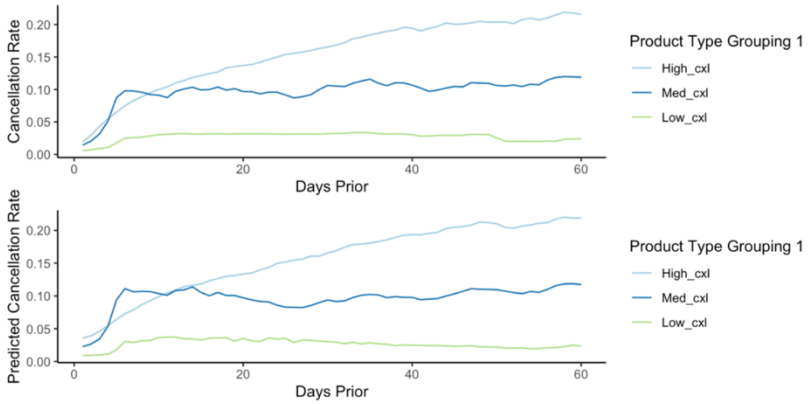

Figure [9]. Mean cancellation rate by Days Prior in actual data and KNN prediction (NewYork hotel).

4.2. Atlanta

The best KNN model used Product Type without regrouping method. The best k (chose based on lowest RMSE) was 61. KNN seemed to overfit observation in Days Prior further from Stay Date (day 20-60) as out-sample MASE was higher than in-sample MASE at around 30 to 90 percentage-point.

Days Prior	In-sample Metrics			Out-sample Metrics		
	MAE	MAPE	MASE	MAE	MAPE	MASE
0-7	1.72	5.58%	48.34%	1.89	5.2%	56.77%
08-14	1.46	7.34%	63.8%	1.70	6.9%	82.03%
15-20	1.47	9.95%	70.75%	1.84	11.24%	93.54%
21-27	1.31	11.67%	76.29%	1.64	12.68%	95.11%
28-60	0.77	13.73%	62.39%	1.57	13.79%	159.33%

Pattern-wise, KNN generated a close approximation trend of 9 Product Types as seen in the figures below. KNN also successfully picked up the peak of lead day 20 to Stay Date which was prominent in the Government and Other product types.

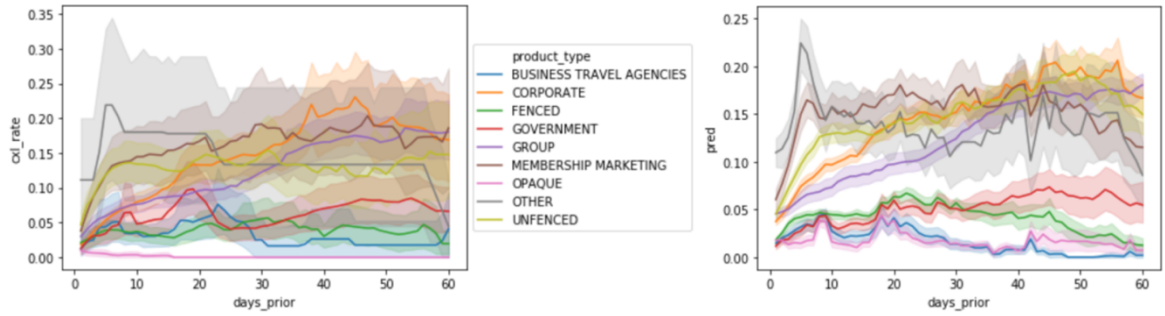

Figure [10]. Mean cancellation rate by Days Prior in actual data and KNN prediction (Atlanta hotel).

5. Regression Tree

Decision tree resembles a flow chart. The key of this algorithm is impurity reduction, which means the tree tries to increase the node homogeneity in each node. The tree starts with the feature that can reduce the most impurity, then continue until the tree stops growing. If the tree stops splitting, it forms a 'leaf node'. In Regression Tree, the predicted value is the mean value of all observations in the leaf that it belongs to. The impurity measure is node's variance.

Similar to KNN, Tree algorithm also has the risk of overfitting to training data. We can alleviate this problem by pruning the tree by implementing stopping rules. Tree's strength is that it and can help us visualize how decisions are made and the result of each decision. Because of this characteristic, we have better understanding of what features are more important in determining the outcome.

5.1. New York

This regression tree was pruned with maximum depth of 12 and minimum observations to be leaf node as 150 observations.

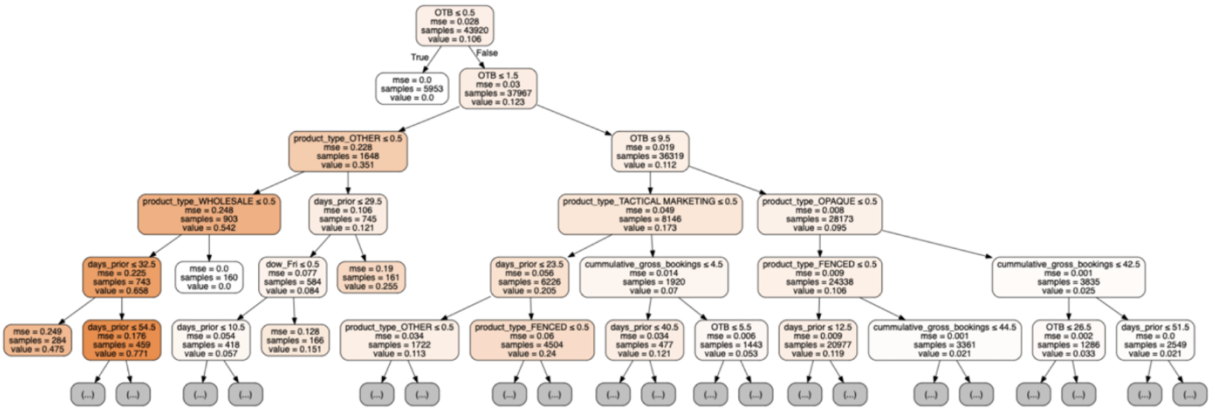

Figure [11]. Visualization of Regression Decision tree (New York hotel).

The most important variable is number of OTB bookings as it showed up in 3 top levels of the tree. The fact that OTB showed up on multiple leaf in the tree indicated that OTB had a non-linear relationship with Cancellation Rate. The importance scores (table below) also showed that OTB is the most important feature, which mean this feature decreased impurity (measured by variance) the most. In our tree, OTB decreased 37% of total impurity.

Product Types at the extreme ends of cancellation behavior also showed themselves as important factors in prediction. They are Product Types with extremely high cancellation rate (Tactical Marketing, Wholesale) and with extremely low cancellation rate (Fenced, Opaque, Other).

Cumulative Gross Bookings also showed up as an important feature. Even though this variable might be close to OTB, it differs from OTB in an important way that Cumulative Gross Bookings indicated

demand level while OTB took cancellations into account. This understanding of Cumulative Gross Bookings is important for model interpretation.

Variable	Importance Score
OTB	37.01%
Product Type (OTHER)	12.75%
Days Prior	12.14%
Product Type (WHOLESALE)	10.83%
Product Type (FENCED)	7.14%
Product Type (TACTICAL MARKETING)	5.74%
Product Type (OPAQUE)	5.03%
Cumulative Gross Bookings	2.75%
Cumulative Cancellation Bookings	2.12%
Days of Week (Thursday)	1.22%

Table [9]. Top 10 important variables in Regression Tree and their Score (New York hotel).

Days Prior	In-sample Metrics			Out-sample Metrics		
	MAE	MAPE	MASE	MAE	MAPE	MASE
0-7	1.53	4.51%	23.81%	1.89	3.08%	22.63%
08-14	1.85	6.23%	55.67%	2.45	4.76%	47.18%
15-20	1.78	6.33%	63.65%	2.50	5.85%	53.32%
21-27	1.58	7.64%	63.36%	2.31	7.88%	52.51%
28-60	1.09	11.14%	62.48%	2.13	14.70%	60.73%

5.2. Atlanta

This regression tree used default pruning method of rpart function in R with caret as backend, it has 27 nodes after pruning. Regrouping Method 2 turned out to perform best under regression tree.

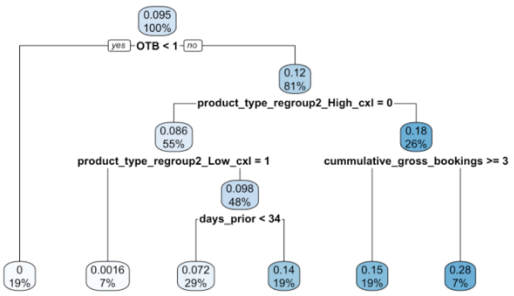

Figure [12]. Visualization of Regression Decision tree (Atlanta hotel).

The most important variable is number of OTB bookings as it showed up in the top levels of the tree. The importance scores (table below) also showed that OTB is the most important feature. The tree graph shows that whether there is any booking affect the cancellation rate a lot, which makes sense because if there is no bookings so far then the current cancellation rate will be 0. There are 19% of our training observations have no on the book bookings and have 0 cancellation rate. If there is any bookings then the cancellation rate will change from 0 to a non-zero rate.

The second and the third level of the tree are two regrouping product type groups. This shows that our second grouping method, regroup product types based on last 10 days cancellation level, work well in tree model.

Cumulative Gross Bookings also showed up as an important feature. It is important to understand that Cumulative Gross Bookings is an indicator of demand level. Moreover, 34 days prior to the stay date is marked as a watershed. The predict cancellation rate of bookings happened in last 34 days is 7.2%, however, if the bookings happened before the last 34 days, the predict cancelation rate is 19%.

Variable	Variable Importanc
OTB	30
Cumulative Gross Bookings	29
Product Type Regroup High Cancellation	16
Product Type Regroup Low Cancellation	5
Days Prior	5
Product Type Regroup Middle Cancellation	5

Table [10]. Top 6 important variables in Regression Tree and their Score (Atlanta hotel).

Days Prior	In-sample Metrics			Out-sample Metrics		
	MAE	MAPE	MASE	MAE	MAPE	MASE
0-7	2.65	7.9%	72.27%	2.76	6.99%	77.19%
08-14	1.83	8.7%	75.06%	1.80	7.43%	81.62%
15-20	1.84	11.82%	85.46%	1.66	11.05%	79.63%
21-27	1.65	13.82%	93.54%	1.51	13.01%	83.46%
28-60	1.16	16.88%	97.04%	1.37	14.63%	133.35%

Although the regression tree model shows that days prior 34 is an important factor, the error matrix shows that our prediction for days prior 28 – 60 is not better than the naïve prediction.

6. Random Forests

Random forests is an ensemble learning method for regression and classification. As we mentioned in above, decision trees that are grown very deep tend to learn highly irregular patterns: they overfit their training datasets. Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. This comes at the expense of a small increase in the bias and some loss of interpretability, but generally greatly boosts the performance in the final model.

Random forests surely do a good job at classification but not as for regression problem as random forests don't gives precise continuous nature prediction. In case of regression, it doesn't predict beyond the range in the training data, and that they may over fit data sets that are particularly noisy. Another disadvantage of random forests is that it can feel like a black box approach for a statistical modelers. We have very little control on what the model does.

6.1. New York

OTB, Days Prior are the 2 most important factors in our random forest. This score of the feature importance in random forest is measured by the mean decreased of impurity (variance), weighted by the probability of reaching that node, and averaged over all trees of the forest. Important features in this random forest were quite similar to decision tree's, with Days Prior became the second important feature. The

pattern of average cancellation trend in random forest resembles the actual cancellation rate trend more than the Tree's pattern.

Variable	Importance Score
OTB	33.77%
Days Prior	12.37%
Product Type (OTHER)	9.79%
Product Type (WHOLESALE)	9.17%
Product Type (FENCED)	7.42%
Cumulative Gross Bookings	6.15%
Product Type (TACTICAL MARKETING)	4.83%
Product Type (OPAQUE)	4.46%
Product Type (BUSINESS TRAVEL AGENCIES)	2.26%
Days of Week (Thursday)	1.57%

Table [11]. Top 10 important variables in Random Forest and their Score (New York hotel).

Days Prior	In-sample Metrics			Out-sample Metrics		
	MAE	MAPE	MASE	MAE	MAPE	MASE
0-7	1.49	3.90%	23.30%	1.92	3.04%	23.15%
08-14	1.78	5.48%	56.42%	2.43	4.69%	50.54%
15-20	1.73	5.82%	64.23%	2.45	5.84%	55.57%
20-27	1.52	6.98%	63.38%	2.43	7.35%	58.38%
28-60	1.03	10.13%	58.79%	2.09	13.80%	60.38%

Table [12]. Top 10 important variables in Random Forest and their Score (New York hotel).

6.2. Atlanta

The important variables that Random Forests indicated are in consistent with the result of Regression Tree. However, in Random Forests model, Regrouping Method 4 stand out and has the best performance in predicting cancellation rate.

Variable	Importance Score
Days Prior	32.44%
OTB	18.80%
Cumulative Gross Bookings	17.69%
Regroup Product Type Low Cancellation	12.15%
Sunday	2.93%

Table [13]. Top 5 important variables in Random Forest and their Score (New York hotel).

By comparing the Error Metrics of in-sample and out-sample models, it is clear that Random Forests has overfitting problem due to the reason discussed above. We ended up not consider Random Forest in our combine model.

Days Prior	In-sample Metrics			Out-sample Metrics		
	MAE	MAPE	MASE	MAE	MAPE	MASE
0-7	0.44	2.01%	12.32%	1.65	5.52%	49.52%
08-14	0.25	2.32%	10.68%	1.85	7.35%	89.16%
15-20	0.25	3.41%	12.16%	2.06	11.41%	104.62%
20-27	0.27	4.23%	15.63%	1.89	13.69%	109.47%
28-60	0.19	7.15%	15.79%	1.77	15.98%	179.60%

7. Neural Network

Neural Network are comprised of simple elements, called neurons, each of which can make simple mathematical decisions. Together, the neurons can analyze complex problems, emulate almost any function including very complex ones, and provide accurate answers. A shallow neural network has three layers of neurons: an input layer, a hidden layer, and an output layer. A Deep Neural Network has more than one hidden layer, which increases the complexity of the model and can significantly improve prediction power. In this project, we used Deep Neural Network. Two hidden layers were included in Neural Network Models for both hotels.

Neural Networks are reducible to regression models. It takes input parameters, which equal to dependent variables in regression, multiplies them by their weights, which equal to coefficients in regression, and run them through an activation function to get the output. When a Neural Network is trained, it will perform gradient descent to find coefficients that are better and fit the data, until it arrives at the optimal linear regression coefficients.

Neural Networks offer a number of advantages, including requiring less formal statistical training, ability to implicitly detect complex nonlinear relationships between dependent and independent variables, ability to detect all possible interactions between predictor variables, and the availability of multiple training algorithms. Disadvantages include its "black box" nature, greater computational burden, proneness to overfitting, and the convergence problem of weights.

Our neural network models for both hotels did not converge. We tried different combinations of activation function (Relu, Sigmoid) and optimizer (Adam, SGD, Adadelta). The standard deviation for 5-fold cross valuation is quite large (MASE's standard deviation can go up to 20%). The reason can be due to high variation in our dataset, and our small number of observations. We decided to leave this model out of our final combined model because it is not robust.

8. Support Vector Regression

In the same way as with classification approach there is motivation to seek and optimize the generalization bounds given for regression. The model produced by Support Vector Regression (SVR) depends only on a subset of the training data, because the epsilon intensive - cost function for building the model ignores any training data close to the model prediction. One of the main advantages of SVR is that its computational complexity does not depend on the dimensionality of the input space. Additionally, it has excellent generalization capability, with high prediction accuracy.

Usually SVR will perform better than KNN, however, in our project, SVR performed very bad for both hotels. This indicates that our dataset is not easily separable using the decision planes that we set in SVR. KNN can generate a highly convoluted decision boundary as it is driven by the raw training data itself. Given the bad performance of SVR, we didn't include SVR in our combine model.

9. Combined model

We used a meta-linear-regressor to stack the results from the selected models (sub-learners). For each Days Prior period, the meta-regression was run to combine different prediction patterns of these individual models. For example, for New York hotel, meta-regressors was run on 5 predictions provided by 5 models for data in Day 01-07 period, Day 08-14 period and so on. The predictions of all Days Prior periods were then combined to be the prediction for all 60 days prior to the Stay Date.

A strength of this combination method is the regression can combine pattern that one model can capture with a pattern another model captures that the first model failed to capture.

9.1. New York

The 5 models (sub-learners) used in this meta-regression are Traditional, Regression, KNN, Tree, and Random Forest. Our stacked model is a successful one because this model outperformed all the sub-learners' predictions in all Days Prior periods.

Days Prior	In-sample Metrics			Out-sample Metrics		
	MAE	MAPE	MASE	MAE	MAPE	MASE
Day 01-07	1.44	3.72%	21.15%	1.84	2.91%	20.97%
Day 08-14	1.74	5.59%	43.40%	2.43	4.67%	38.02%
Day 15-20	1.73	5.87%	49.64%	2.48	5.83%	48.99%
Day 21-27	1.54	7.09%	51.14%	2.36	7.42%	52.31%
Day 28-60	1.05	9.90%	49.20%	2.10	12.46%	51.97%

We obtained the best validation results from this model. In every Days Prior period, the out-sample MASEs were the smallest. This result indicates a much better predictive ability of this combined model compared to Naïve prediction. The best Days Prior period is Day 01-07, which sucessfully recued 80% error in Naïve predictions.

9.2. Atlanta

Similar to New York hotel, we used a meta-linear-regressor to stack the results from 4 models (sub-learners): Traditional, Regression, KNN, and Tree. We manually set cancellation rate equal to naïve prediction for days prior 28 – 60, since there is no model performs better than the benchmark.

Days Prior	In-sample Metrics			Out-sample Metrics		
	MAE	MAPE	MASE	MAE	MAPE	MASE
Day 01-07	1.96	5.79%	51.8%	2.05	5.33%	57.66%
Day 08-14	1.66	7.48%	68.05%	1.82	7.10%	82.62%
Day 15-20	1.53	10.01%	70.98%	1.79	11.34%	85.68%
Day 21-27	1.36	11.98%	77.33%	1.50	12.88%	83.09%
Day 28-60	1.19	16.83%	100%	1.01	15.11%	100%

The result we obtained from this model is not the best for each days prior categories. If we chose only one model for each days prior category, we will get the best MASE for each days prior. However, this is an extreme case, it is not practical for hotel to forecast cancellation rate by running different models for each days prior category and manually pick the best model for each days prior category. in the real world. The regression stacking method gives the best overall result, considering both the model performance and practicability.

V. DISCUSSION

1. Model Comparison

1.1. New York

For New York hotel, the out-of-sample metrics are usually better than the in-sample metrics. Our models often outperformed Naïve predictions more in testing set than in training set. This can be due to the higher variation nature of the training set than the testing set.

In general, the predictions for Days Prior closer to Stay Date were better than predictions for longer lead days to Stay Date. Again, the reason can be the high volatility in data and fewer observations the further away from the booked Stay Date. Closer to Stay Date, there are more observations and the pattern might be more prominent and easier for our model to pick up.

In Days Prior period 01-07, the in-sample metrics and out-sample metrics are very close to each other. The cancellation trend in training and testing set might be similar, which can constitute a last-week effect. Further investigation on the last-week effect might be helpful to understand the last-week cancellation behavior better, and potentially find a way to reduce cancellation in other Day Prior periods. It is possible that there is a policy restricting last-week cancellation, regardless of product type, that can explain for the steady decline in cancellation rate in the last week before Stay Date

The Traditional Model performed relatively well compared to other sub-learners in days closer to Stay Date while Regression Tree performed better than other sub-learners in Days Priors further from Stay Date. In Days Prior period 01-07, cancellation pattern might be more uniform, hence Traditional model performed particularly well in this time period. In days further away from the Stay Date, the values are more extreme (cancellation rate of 100% or 0%), and Regression Decision Tree was particularly good in Days Prior period 21-60. Investigating the Tree model, we can see that the top 2 levels both used OTB as splitting variable. In the top level, it already assigned the extreme 0% cancellation rate for 13.5% of total observations. Similarly, Tree model can predict extreme values such as 99% cancellation rate, while Traditional model and Regression never came up with such extremity. Different model performed well in different Days Prior period due to observation variation depending on days prior.

There are some features that proved their importance across models. Cumulative Gross Bookings indicates the demand level for a room and OTB bookings indicates the raw demand level adjusted by cancellations made. Days Prior is also an important factor. It has nonlinear relationship to cancellation rate, as indicated by Regression and Decision Tree model. Product Types is also an important factor. Product type interacted with Days Prior and have distinct cancellation pattern. For example, Government product had increasing cancellation rate as Days Prior approaching days 10 prior to Stay Date. Meanwhile, cancellation behavior for Corporate product manifested as a gradually declining trend line. As Product Type showed distinctive cancellation patterns, we suggest hotels to look deeper into each individual Product Type to find suitable policy for better cancellation control.

Our combined model by linear-regression stacking method was the best-performing model in all Days Prior periods. This result indicated that the meta-regressor generated predictions that resembled cancellation patterns that individual model did not fully capture.

Days Prior	Model	In-sample Metrics			Out-sample Metrics		
		MAE	MAPE	MASE	MAE	MAPE	MASE
Day 01-07	Naïve (benchmark)	6.83	8.68%	100%	8.78	8.19%	100%
	Traditional	1.51	3.76%	22.20%	1.85	3.10%	21.04%
	Regression	2.12	4.26%	31.01%	2.75	3.60%	31.31%
	KNN	1.66	3.72%	24.30%	2.39	3.80%	27.17%
	Tree	1.53	4.51%	23.81%	1.89	3.08%	22.63%
	Random Forest	1.49	3.90%	23.30%	1.92	3.04%	23.15%
	Combined	**1.44**	**3.72%**	**21.15%**	**1.84**	**2.91%**	**20.97%**
Day 08-14	Naïve (benchmark)	4.03	8.20%	100%	6.23	7.35%	100%
	Traditional	2.00	6.26%	49.92%	2.63	4.96%	42.17%
	Regression	2.17	6.56%	53.77%	2.79	5.29%	44.85%
	KNN	1.74	5.79%	43.08%	2.60	5.48%	41.78%
	Tree	1.85	6.23%	55.67%	2.45	4.76%	47.18%
	Random Forest	1.78	5.48%	56.42%	2.43	4.69%	50.54%
	Combined	**1.74**	**5.59%**	**43.40%**	**2.43**	**4.67%**	**38.02%**
Day 15-20	Naïve (benchmark)	3.49	8.82%	100%	5.07	7.40%	100%
	Traditional	2.02	6.97%	58.19%	2.65	5.71%	52.32%
	Regression	2.13	7.45%	61.06%	2.87	6.12%	56.63%
	KNN	1.66	6.37%	47.67%	2.54	5.76%	50.04%
	Tree	1.78	6.33%	63.65%	2.50	5.85%	53.32%
	Random Forest	0.96	9.80%	45.08%	2.24	12.02%	55.26%
	Combined	**1.73**	**5.87%**	**49.64%**	**2.48**	**5.83%**	**48.99%**
Day 21-27	Naïve (benchmark)	3.02	10.28%	100%	4.50	8.72%	100%
	Traditional	1.80	8.21%	60.03%	2.69	7.10%	60.02%
	Regression	1.91	8.80%	63.22%	2.85	7.73%	63.24%
	KNN	1.45	7.09%	47.95%	2.65	6.72%	58.77%
	Tree	1.58	7.64%	63.36%	2.31	7.88%	52.51%
	Random Forest	1.09	11.14%	62.48%	2.13	14.70%	60.73%
	Combined	**1.54**	**7.09%**	**51.14%**	**2.36**	**7.42%**	**52.31%**
Day 28-60	Naïve (benchmark)	2.13	14.92%	100%	4.05	15.13%	100%
	Traditional	1.55	12.46%	73.39%	2.56	12.70%	63.63%
	Regression	1.42	12.53%	66.81%	2.55	13.03%	63.01%
	KNN	1.45	7.09%	47.95%	2.65	6.72%	58.77%
	Tree	1.58	7.64%	63.36%	2.31	7.88%	52.51%
	Random Forest	1.03	10.13%	58.79%	2.09	13.80%	60.38%
	Combined	**1.05**	**9.90%**	**49.20%**	**2.10**	**12.46%**	**51.97%**

Table [14]. Summary Error Metrics Table (New York hotel).

.2. Atlanta

Atlanta hotel has very irregular patterns both in training dataset and test dataset, which makes it really hard to avoid overfitting problem and increase the predict accuracy. The nature of Atlanta hotel is that it usually holds large conference and events. If there is any change of the events, such as reschedule or a sudden cancellation, will cause fluctuation in hotel cancellation rate.

In general, the predictions for Days Prior closer to Stay Date were better than predictions for longer lead days to Stay Date. The MASE difference of In-sample and Out-sample decreases as the Days Prior closer to Stay Date. Again, the reason can be the high volatility in data and fewer observations the further away from the booked Stay Date. Closer to Stay Date, there are more observations and the pattern might be more prominent and easier for our model to pick up.

For Days Prior 0 – 7, since this days prior category is most close to the stay date, we want to predict this category as accurate as possible. The error matrix shows that all the models performed better than the benchmark, considering that they all have better MAE, MAPE, and MASE value. The model with the best performance is traditional model. It has smallest out-sample MASE, MAPE and MAE. The regression model also performs very well. This could be caused by a uniform cancellation rate pattern in days prior 0 – 7. Although Random Forest has good out-sample performance in this days prior category, by comparing it's in-sample performance with out-sample performance, we can see it has overfitting problem.

For Days Prior 8 – 27, as Days Prior increase, we can see that the performance of Regression Tree and KNN model become better. It's MAE, MAPE and MASE stand out as days prior increases. Days Prior 8 – 27 is where most chaos happened. KNN and Regression usually have the problem of overfitting. By pruning the Tree model and adjust the K value of KNN model, they have the ability to capture the pattern of the data and also avoid noises.

For Days Prior after 28, we see that no model performed better than the benchmark. For Days Prior after 28, we see that no model performed better than the benchmark. Because our dependent variable is cancellation rate, in days further away from the Stay Date, this values are more extreme (cancellation rate of 100% or 0%). This characteristic makes it really hard to predict cancellation rate for days prior after 28. It is easier to simply use naïve prediction for those days.

There are some features that proved their importance across models. Cumulative Gross Bookings indicates the demand level for a room and OTB bookings indicates the raw demand level adjusted by cancellations made. Days Prior is also an important factor. It has nonlinear relationship to cancellation rate, as indicated by Regression and Decision Tree model. Product Types is also an important factor. Most of our models performed best without regrouping product type, which means that product type contains information of different cancellation behavior. Product type interacted with Days Prior can also distinct cancellation pattern.

Our combined model by linear-regression stacking method was not the best-performing model in all Days Prior periods. Because for different Days Prior period, different model standing out. However, it is more practical to just use our combine model in real business to forecast cancellation rate.

Days Prior	Model	In-sample Metrics			Out-sample Metrics		
		MAE	MAPE	MASE	MAE	MAPE	MASE
Day 01-07	Naïve (benchmark)	3.78	8.87%	100%	3.56	8.22%	100%
	Traditional	1.70	6.41%	46.63%	1.63	5.26%	45.73%
	Regression	1.94	6.31%	51.25%	1.74	5.30%	48.16%
	KNN	1.72	5.58%	48.34%	1.89	5.2%	56.77%
	Tree	2.65	7.9%	72.27%	2.76	6.99%	77.19%
	Random Forest	0.44	2.01%	12.32%	1.65	5.52%	49.52%
	Combined	**1.96**	**5.79%**	**51.8%**	**2.05**	**5.33%**	**57.66%**
Day 08-14	Naïve (benchmark)	2.44	9.99%	100%	2.21	9.23%	100%
	Traditional	1.83	9.04%	77.94%	1.77	7.73%	80.21%
	Regression	1.79	8.75%	73.41%	1.81	7.89%	84.09%
	KNN	1.46	7.34%	63.8%	1.70	6.9%	82.03%
	Tree	1.83	8.7%	75.06%	1.80	7.43%	81.62%
	Random Forest	0.25	2.32%	10.68%	1.85	7.35%	89.16%
	Combined	**1.66**	**7.48%**	**68.05%**	**1.82**	**7.10%**	**82.62%**
Day 15-20	Naïve (benchmark)	2.15	12.6%	100%	2.09	12.48%	100%
	Traditional	1.96	13.56%	93.31%	1.98	12.00%	94.89%
	Regression	1.84	11.51%	85.18%	1.99	11.88%	96.48%
	KNN	1.47	9.95%	70.75%	1.84	11.24%	93.54%
	Tree	1.84	11.82%	85.46%	1.66	11.05%	79.63%
	Random Forest	0.25	3.41%	12.16%	2.06	11.41%	104.62%
	Combined	**1.53**	**10.01%**	**70.98%**	**1.79**	**11.34%**	**85.68%**
Day 21-27	Naïve (benchmark)	1.76	14.13%	100%	1.81	14.68%	100%
	Traditional	1.75	14.45%	101.71%	1.80	13.55%	99.49%
	Regression	1.66	13.33%	94.26%	1.98	13.35%	110.89%
	KNN	1.31	11.67%	76.29%	1.64	12.68%	95.11%
	Tree	1.65	13.82%	93.54%	1.51	13.01%	83.46%
	Random Forest	0.27	4.23%	15.63%	1.89	13.69%	109.47%
	Combined	**1.36**	**11.98%**	**77.33%**	**1.50**	**12.88%**	**83.09%**
Day 28-60	Naïve (benchmark)	1.19	16.83%	100%	1.00	15.11%	100%
	Traditional	1.19	16.01%	103.11%	1.61	14.47%	156.89%
	Regression	1.06	15.19%	84.68%	1.52	13.97%	148.68%
	KNN	0.77	13.73%	62.39%	1.57	13.79%	159.33%
	Tree	1.16	16.88%	97.04%	1.37	14.63%	133.35%
	Random Forest	0.19	7.15%	15.79%	1.77	15.98%	179.60%
	Combined	**1.19**	**16.83%**	**100%**	**1.01**	**15.11%**	**100%**

Table [15]. Summary Error Metrics Table (Atlanta hotel).

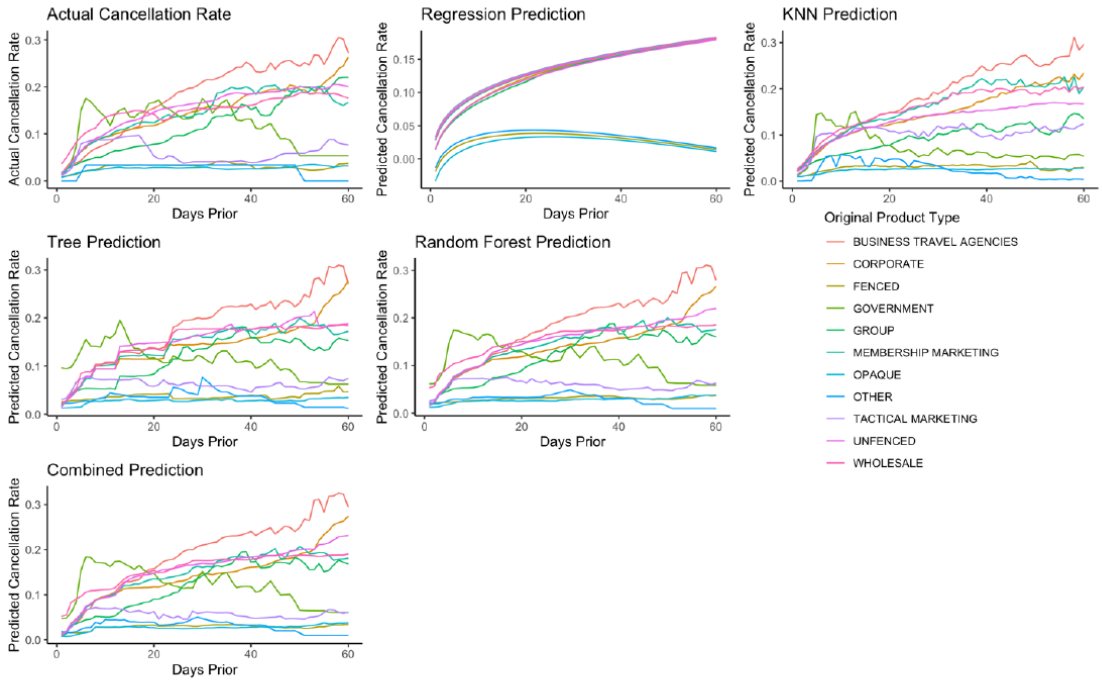

Figure [13]. Visualization of Models Prediction (New York hotel).

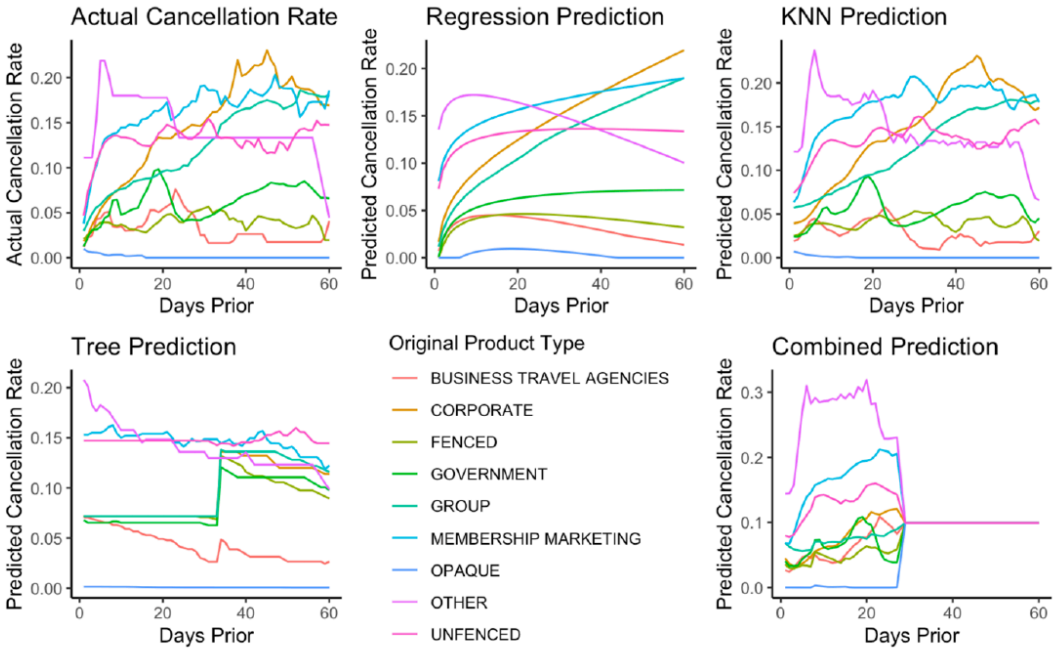

Figure [14]. Visualization of Models Prediction (Atlanta hotel).

2. Conclusion and Discussion

In hotel industry, a booking symbolizes a contract between the customer and the hotel, which gives the customer the right to use the service in the future at a settled price. Usually, an option to cancel the contract prior to the service provision is included. However, the option to cancel a service prior to its provision places all risk on hotels, which have to guarantee rooms to customers who honor their bookings but also bear the cost of vacant rooms when a booking is cancelled or a customer does not show (Talluri and Van Ryzin, 2005). Consequently, cancellations have a significant impact on demand-management decisions in the context of revenue management.

The exploratory data analysis in Section III provided insights in the dataset. It shows different cancellation pattern of two hotels. This indicates that the cancellation patterns of different hotels are not always similar. The cancellation pattern of a hotel really depends on its geography location, its business contracts with travel agencies and companies. Customer type of a hotel in New York City is very different from the customer type of a hotel in Atlanta. Hotel in Atlanta has more fluctuate cancellation rate pattern because it usually holds special events and conferences, while the hotel in New York city has a more stable cancellation rate pattern. Which can be the higher booking volume in other more stable product types in New York City hotel that dilutes much of irregular patterns. The hotel in New York city has customers from Wholesale and Tactical Marketing product type, which is because New York is a big tourist city, while the hotel in Atlanta doesn't have customers from these two product types. When predicting cancellation rate a hotel, the hotel should consider its own actual situation.

From the modeling results, presented in Section IV, the conclusion can be made that Cumulative Gross Bookings, which indicates the demand level for a room, OTB, Days Prior, and Product Type are the most important factors when predicting cancellation rates for hotels. Revenue managers can manage cancellation rate by analyzing the market demand level, providing different cancellation policy to different product type, making noncancelable rules for last minute bookings, etc.

The approach that is suggested in this paper is to use meta-linear-regressor to stack the results from the selected models (sub-learners). This approach provides the best result regarding the model performance, practicability, and cost efficiency.

The limitation of this paper is that we didn't consider the individual cancellation behavior since we were not provided with individual customer information. The revenue managers can implement the approach of cancellation rate forecast by considering Passenger Name Record (PNR) based forecasting. If the revenue managers want to forecast cancellation rates for not only one hotel or want to compare the cancellation rate patterns of different hotels, it is also worth taking the size of the hotel, the luxury level, the cancellation policy, and hotel's competitor's performance into account. Another point of discussion is that we didn't take overbooking strategies and booking seasonality into consideration in this paper.

REFERENCE

Royston, P., & Altman, D. (1994). Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. Journal of the Royal Statistical Society. Series C (Applied Statistics), 43(3), 429-467. doi:10.2307/2986270 https://www.jstor.org/stable/pdf/2986270.pdf?refreqid=excelsior%3A4e99b86cfc2b1e9afa6c7081d7ef32d3

Airline Passenger Cancellations: Modeling, Forecasting and Impacts on Revenue Management, By Oren Petraru (Petraru), Master Thesis, MIT. https://dspace.mit.edu/handle/1721.1/104325

Cancellation Predictor for Revenue Management applied in the hospitality industry, By R. van Leeuwen (Leeuwen), Master Thesis, Vrije Universiteit Amsterdam. https://beta.vu.nl/nl/Images/werkstuk-leeuwen_rik_van_tcm235-876479.pdf