LUISS

Department of Business and Management

Course of Data Analysis for Business

Performance Comparison of Time-Dependent Variables Predictive Models for Hotel Booking Cancellations

Prof. Francesco Iafrate
SUPERVISOR

Martina Serandrei
270391
CANDIDATE

Academic Year 2023/2024

TABLE OF CONTENTS

1. INTRODUCTION

1.1. Importance of Demand and Cancellation Forecasting

In the hospitality industry, successful revenue management depends on the ability to make the right room available to the right guest, at the right price, at the right time, and through the right channel. (R. Mehrotra and J. Ruttley, 2006.) Having an optimal strategy of management maximizes profitability for hotels (Rajopadhye et al., 2001) and reservation systems play a pivotal role in this process by facilitating bookings. Though, when customers have the option to cancel reservations it introduces a significant risk for revenue management (Tekin et al., 2021). Forecasting demand in an accurate and realistic way, becomes challenging when cancellations are frequent, leading to potential revenue losses for hotels (Rothstein, 1985). Overall, researches indicate that the average cancellation rate in hotel bookings is around 20%, with this rate escalating to 60% for hotels situated near airports or along major travel routes (Iliescu et al., 2008).

In order to mitigate the impact of cancellations, the concept of "net demand" is introduced; it is basically the number of demand requests minus the number of cancellations, and it is deemed as essential for maintaining an efficient revenue management system (Rajopadhye et al., 2001). Moreover, net demand is also fundamental to implement strategies like overbooking, which are aimed at counter-acting the potential risk of customers cancelling their bookings or even not showing up. The former is a common practice in both the airline and hotel industries, which involves accepting more bookings than the actual capacity, based on estimated cancellation rates. This strategy ensures that hotels can manage capacity without significant early sell-outs by employing dynamic pricing and capacity allocation controls (Chatterjee, 2001). Typically, overbooking decisions are made towards the end of the booking horizon, when the likelihood of cancellations is clearer (Romero Morales & Wang, 2010).

Determining overbooking levels is not the only measure where forecasting cancellation rates is crucial, but also for estimating net demand throughout the booking period (Rajopadhye et al., 2001; Chatterjee, 2001). The revenue management system

continuously needs to evaluate both the anticipated net demand and the net demand from ongoing bookings. Hoteliers that rely on improved demand forecasting, can reach a better understand of their net demand, reinforcing globally the revenue management strategies.

This thesis' aim is to explore the performance of different regression algorithms in order to predict hotel booking cancellations rates. By comparing these algorithms, this research seeks to identify the most effective methods for enhancing the accuracy and interpretability of cancellation forecasts, thereby contributing to more robust revenue management practices in the hospitality industry.

1.2. Thesis Organization

The paper is structured in such a way that it guides the reader through the research process. The introduction sets the stage for the research by defining the study's scope, context and objectives.

The literature review part provides an overview of current research and studies on booking cancellations in the revenue management context, also retrieving some studies performed outside the hospitality industry. Some comparison will be made on the different methods and interpretations when cancellation forecasting is performed (whether it is seen as a classification or a regression problem). This first part of the review will be guiding the reader to the possible research gap that this thesis will try to cover. The second section instead, will give a theoretical explanation of all the forecasting methods used subsequently to carry out the analysis, ending also with a brief mention to the performance comparison measure that will be used.

The methodology and analysis section starts by outlining the data that will be used in the study and the source where it has been retrieved. After that, the analysis will open with an overview on the variables that are present in the dataset and will immediately dive into a data preprocessing. Here dataset will be checked for inconsistencies, addressing duplicates and missing values; then it will proceed with the feature engineering explanations. A very important part will be formed by the steps of the dataset transformation from a PNR dataset to a time series. To complete this first part there is the EDA combined with data visualization to extrapolate the main features of the data.

Moving on to the modelling, each model application will be explained as well as its result, ending with a final section on performance comparison, through the analysis of the applied measure.

The results and discussion part will provide the interpretation of the findings and trying to capture what their practical implications are. Lastly, the conclusions will draw the key takeaways from the project and the main results. Obviously, there will be a bibliography section where all the references that were useful throughout this thesis will be included.

2. LITERATURE REVIEW

2.1. Previous Studies on Predictive Modeling for Hotel Booking Cancellation

The analysis of prior relevant literature is essential to identify the areas covered within a body of research and uncover areas where further research is necessary. The importance of forecasting in revenue management (RM) research is recognized in the latest literature reviews (Chiang, Chen, & Xu, 2007; Denizci Guillet & Mohammed, 2015; Ivanov & Zhechev, 2012).

Starting from a broader analysis of the literature, earlier studies focused on forecasting bookings in terms of demand, primarily using time series models. Non-causal time series models aim to uncover future patterns from historical data. Historically, the integrated autoregressive moving average (ARIMA) model has been the most widely used to predict demand (C.-Sánchez et al., 2022). However, in recent years, seasonal ARIMA models (e.g., SARIMA) have gained popularity due to the close relationship between tourism and seasonality (Claveria & Datzira, 2010). For example, Claveria and Datzira included consumer expectations in time series models to predict tourist demand from four different European markets in Catalonia, Spain, and found that ARIMA and Markov switching regime (MKTAR) models performed best, while models including consumer expectations did not offer better outcomes (Claveria & Datzira, 2010). Pfeifer and Bodily (1990)

utilized a space–time ARIMA (STARIMA) approach to predict arrivals in eight hotels of the same chain in an American city and concluded that STARIMA, which assumes dependence among points and gives a higher weight to closer ones, performed better than a single ARIMA time series model.

Shifting the attention to booking cancellations: they already have a well-known body of knowledge in the scope of revenue management applied to service industries, particularly the hospitality industry. With the increasing influence of the internet on how customers search and buy travel services, research on this topic has increased, especially on controls, used to mitigate the effects of cancellations on revenue and inventory allocation, cancellation policies, and overbooking (Anderson, 2012; Chen et al., 2011; Noone & Lee, 2010; Hayes & Miller, 2011; Ivanov, 2014; Talluri & Van Ryzin, 2004).

The development of a booking cancellation prediction model aligns with Chiang et al. (2007), who emphasized that revenue management should utilize mathematical and forecast models to better leverage available data and technology. The literature on bookings cancellation prediction for travel-related service industries is quite sparse and relatively recent. Antonio et al. (2019) found 16 publications on the subject, all published within the last 15 years. Notably, only five concerned the hotel industry, and all implemented classification algorithms, reflecting a trend towards using detailed booking data in the PNR format rather than time-series aggregated data. Detailed booking data enhances forecast accuracy (Hueglin & Vannotti, 2001; Petraru, 2016) and facilitates the development of classification prediction models. Cancellation prediction models classify the cancellation outcome of each booking, allowing an understanding of cancellation drivers (Morales & Wang, 2010; Petraru, 2016). Antonio et al. (2019) applied several two-class classification algorithms (e.g., boosted decision trees, decision forests, decision jungles, locally deep support vector machine, and neural networks) to predict cancellation rates for four hotels in the Algarve region, Portugal, and found that decision forests were particularly effective.

As just said, recent studies have primarily approached the problem as a classification one. Depending on the goals, booking cancellations can also be predicted using regression. When the objective is solely to estimate the cancellation rate, the problem should be considered a forecasting problem. When aiming to estimate the likelihood of a booking

being canceled and understanding the cancellation drivers, it should be considered a classification problem. In this context, booking cancellation predictions enable the estimation of the overall cancellation rate. (Antonio et al., 2019)

A novel approach, instead, could involve aggregating detailed PNR dataset data to create a time series for more accurate regression forecasts of booking cancellation rates. The only study to have done something similar is "Forecasting cancellation rates for services booking revenue management using data mining". Morales and Wang (2010) used data mining to predict cancellation rates for service-booking revenue management, considering variables like price, room category, and booking channel. They found tree-based and kernel methods (particularly, support vector machine, SVM) to be the most robust for forecasting hotel cancellations.

2.2. Time-series Forecasting Models Applied in the Study

Since the data is transformed into a time series, the models that are going to be applied are suitable for those type of data.

To start with, it is important to give also a proper definition for time series. A time series is a collection of well-defined data points gathered through repeated measurement in intervals of time and can be decomposed into three components: the trend (long-term direction), the seasonal (systematic, calendar related movements) and the irregular or residuals (unsystematic, short-term fluctuations).

This study compares four different models for time series regression, aimed at forecasting booking cancellation rates.

ARIMA

Traditionally, time series predictions are performed using the autoregressive integrated moving average (ARIMA) models, which attempt to filter out high-frequency noise in the data to detect local trends based on linear dependence in observations in the series.

Moreover, ARIMA is a popular statistical method for time series forecasting and is designed to predict future points in a series by considering the dependencies between observations and their lags. Essentially, ARIMA captures the dynamics in time series data through three key components:

1. Autoregression (AR), which explains the variable of interest using its own previous values through the lags.
2. Integration (I), which is used to make the time series stationary, meaning that statistical properties like mean and variance are constant over time. This is achieved by differencing the series or subtracting an observation from a previous one.
3. Moving Average (MA), which uses past forecast errors in a regression-like model. The idea is that the error in the prediction can be adjusted based on previous errors to improve future forecasts.

The ARIMA model is specified by three parameters: p (number of lag observations), d (number of times the data is differenced), and q (size of the moving average window) (https://otexts.com/fpp3/non-seasonal-arima.html).

STL

STL stands for "Seasonal and Trend decomposition using LOESS". It is a versatile and robust method for decomposing time series into its three main components:

- Trend: The underlying trend of the data.
- Seasonal: Seasonal effects.
- Residual: The remainder after accounting for the trend and seasonal components.

LOESS (Locally Estimated Scatterplot Smoothing) flexibly decomposes a time series, adapting to changing trends and seasonality and making it suitable for more complex datasets. When dealing with non-linear and evolving seasonal patterns trends, it is particularly valuable. (https://mlpills.dev/time-series/time-series-forecasting-with-stl/)

Forecasts of STL objects are obtained by applying a non-seasonal forecasting method to the seasonally adjusted data and re-seasonalizing using the last year of the seasonal component. (Hyndman & Athanasopoulos, 2021; Najera)

MULTIPLE LINEAR REGRESSION for TIME SERIES

Multiple linear regression extends the concept of simple linear regression to include more than one explanatory variable. In both cases, we still use the term 'linear' because we assume that the response variable is directly related to a linear combination of the explanatory variables. The equation for multiple linear regression has the form:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + e_i$$

As for the simple case, β_0 is the constant – which will be the predicted value of y when all explanatory variables are 0. In a model with p explanatory variables, each explanatory variable has its own β_coefficient. In general, the analysis does not allow us to make causal inferences, but it does allow us to investigate how a set of explanatory variables is associated with a response variable of interest. (Tranmer et al., 2020)

For the time series application, lagging of independent variables is often necessary in order for the regression model to be able to predict the future – i.e., to predict what will happen in period t based on knowledge of what happened up to period t-1. (Nau)

RANDOM FOREST for TIME SERIES

For classification and regression problems, random forest is a collection of decision tree techniques and an expansion of bootstrap aggregation (bagging) of decision trees. Several decision trees are constructed during bagging, each one starting from a distinct bootstrap sample taken from the training dataset. Because each decision tree is fitted on a slightly different training dataset and performs slightly differently as a result, bagging is an efficient ensemble approach. This is preferable since it makes each tree more unique and reduces prediction mistakes and correlations in the predictions. When the average of all the decision trees' predictions is used, the model performs better than when any single tree is used alone. The average of the predictions made by each tree in the ensemble is a prediction on a regression problem.

Random forest, similarly to bagging, builds a huge number of decision trees using bootstrap samples from the training dataset. On the other hand, in random forest, a subset of input characteristics (variables or columns) is chosen at each split point for building the trees. Each decision tree in the ensemble is forced to be more distinct from the others by condensing the features to a random subset that may be taken into consideration at each split point.

As a result, there is a greater or lesser degree of correlation between the predictions and prediction errors made by each tree in the ensemble. When the predictions from these less correlated trees are averaged to make a prediction, it often results in better performance than bagged decision trees. (Brownlee, 2023)

EVALUATION METHODS

As model evaluation measures, RMSE will be considered. The root mean square error (RMSE) measures the average difference between a statistical model's predicted values and the actual values. Mathematically, it is the standard deviation of the residuals. Residuals represent the distance between the regression line and the data points. RMSE quantifies how dispersed these residuals are, revealing how tightly the observed data clusters around the predicted values. As the data points move closer to the regression line, the model has less error, lowering the RMSE. A model with less error produces more precise predictions. (Frost, 2023)

3. METHODOLOGY AND ANALYSIS

3.1. Data Description

For our research, we utilized the "Hotel Booking Demand" dataset published on Kaggle, an online community for data scientists and machine learning practitioners that houses a vast repository of datasets. This dataset, which was downloaded and cleaned by Thomas Mock and Antoine Bichat for #TidyTuesday during the week of February 11th, 2020, contains information about reservations for both city and resort hotels, with all personally identifying information removed (Mostipak, 2020). It includes 32 attributes:

1. **hotel:** Resort Hotel or City Hotel.
2. **is_canceled:** Value indicating if the booking was canceled (1) or not (0).
3. **lead_time:** Number of days that elapsed between the entering date of the booking into the PMS and the arrival date.
4. **arrival_date_year:** Year of arrival date.
5. **arrival_date_month:** Month of arrival date with 12 categories: "January" to "December".
6. **arrival_date_week_number**: Week number of year for arrival date.
7. **arrival_date_day_of_month:** Day of the month of the arrival date.
8. **stays_in_weekend_nights:** Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel.

9. **stays_in_week_nights:** Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel.
10. **Adults:** Number of adults.
11. **Children:** Number of children.
12. **Babies:** Number of babies.
13. **Meal:** Type of meal booked. Categories are presented in standard hospitality meal packages: Undefined/SC– no meal package; BB– Bed & Breakfast; HB– Half board (breakfast and one other meal– usually dinner); FB– Full board (breakfast, lunch and dinner).
14. **Country:** Country of origin. Categories are represented in the ISO 3155–3:2013 format.
15. **market_segment:** Market segment designation. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators".
16. **distribution_channel:** Booking distribution channel. The term "TA" means "Travel Agents" and "TO" means "Tour Operators".
17. **is_repeated_guest:** Value indicating if the booking name was from a repeated guest (1) or not (0).
18. **previous_cancellations:** Number of previous bookings that were cancelled by the customer prior to the current booking.
19. **previous_bookings_not_canceled:** Number of previous bookings not cancelled by the customer prior to the current booking.
20. **reserved_room_type:** Code of room type reserved. Code is presented instead of designation for anonymity reasons.
21. **assigned_room_type:** Code for the type of room assigned to the booking. Code is presented instead of designation for anonymity reasons.
22. **booking_changes:** Number of changes made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation.
23. **deposit_type:** Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit– no deposit was made; Non Refund– a deposit was made in the value of the total stay cost; Refundable– a deposit was made with a value under the total cost of stay.
24. **agent:** ID of the travel agency that made the booking.

25. **company:** ID of the company that made the booking or responsible for paying the booking.
26. **days_in_waiting_list:** Number of days the booking was in the waiting list before it was confirmed to the customer.
27. **customer_type:** Type of booking, assuming one of four categories: Contract - when the booking has an allotment or other type of contract associated to it; Group– when the booking is associated to a group; Transient– when the booking is not part of a group or contract, and is not associated to other transient booking; Transient-party– when the booking is transient, but is associated to at least other transient booking.
28. **adr:** Average Daily Rate as defined by dividing the sum of all lodging transactions by the total number of staying nights. It measures the average rental revenue earned for an occupied room per day.
29. **required_car_parking_spaces:** Number of car parking spaces required by the customer.
30. **total_of_special_requests**: Number of special requests made by the customer (e.g. twin bed or high floor).
31. **reservation_status:** Reservation last status, assuming one of three categories: Canceled– booking was canceled by the customer; Check-Out– customer has checked in but already departed; No-Show– customer did not check-in and did inform the hotel of the reason why.
32. **reservation_status_date:** Date at which the last status was set. This variable can be used in conjunction with the reservation_status.

(Antonio et al., 2017; Novakovic & Turina, 2021)

The dataset, based on real Portuguese market research data, provides a valuable resource for studying hotel market trends. Collected from hotels in Portugal, it comprises 119,390 reservation records spanning from July 1, 2015, to August 31, 2017, with records from resort hotels in the Algarve region and city hotels in Lisbon. All data are anonymized to protect the privacy of hotels and guests. (Tekin & Gök, 2021)

3.2. Data Preprocessing

All the models and data manipulation executed for this work were programmed in R (R Core Team, 2016).

The data preprocessing part can be divided into two sections: the first involves the analysis of the original dataset, and the second concerns the aggregation of the dataset to transform it into a time series.

By looking at the summary, the first thing that it is evident is the need to create a dummy variable for the “hotel” column, which attributes 0 to "City hotel", 1 to "Resort hotel". The "company" variable contained mostly NULL values, and the "agent" variable also had a significant number of NULL values. Since "agent" represents the ID of the travel agency that made the booking and "company" corresponds to the ID of the entity responsible for the booking, both were deemed non-significant and excluded from the analysis. The "children" variable had four NULL values, which were replaced with 0 under the assumption that no data indicated no children in those bookings. Subsequently, duplicates have been removed with the function hotel_data = hotel_data[!duplicated(hotel_data),]. Looking again at the summary, clearly some variables required format conversion: "arrival_date_year," "arrival_date_month," "arrival_date_day_of_month," "is_canceled," and "is_repeated_guest" were transformed into factors. Lastly, the "adults" column presented a maximum of 55 and a minimum of 0, which was interpreted as problematic since hotel reservations should have at least one adult (and, obviously children cannot book hotel rooms). Rows with 0 adults were thus eliminated.

Successively, some feature engineering has been performed on “children” and “babies” by unifying them into a single column, called “kids”; the former two were then dropped from the dataset. A column was added to represent the total number of guests by summing the "kids" and "adults" columns, and the "arrival_date" variable was created by unifying "arrival_date_year," "arrival_month_num," and "arrival_date_day_of_month." Lastly, a new variable, "same_room_type," was created by comparing "reserved_room_type" and "assigned_room_type" to check for consistency between the reserved and assigned rooms.

The second part of the data preprocessing involved transforming the dataset into a time series format. This study, indeed, will be performed on the time series dataset obtained by aggregating the PNR original data.

The first step regarded dealing with the categorical variables present in the dataset: they were encoded as dummy variables. However, due to the high number of levels in some variables, a Cramer's V score analysis through a chi-square test was carried out to identify significant associations with the "is_canceled" variable.

Based on the analysis, several categorical variables showed significant associations with the is_canceled variable, as indicated by their very low p-values (all below 0.05, many approaching 0). The strongest association is observed with reservation_status (Cramér's V = 1.0000), indicating a direct relationship with cancellations. Other notable features include market_segment (Cramér's V = 0.2213, p-value = 0.00), same_room_type (Cramér's V = 0.2132, p-value < 0.01), country (Cramér's V = 0.1980, p-value < 0.01), deposit_type (Cramér's V = 0.1653, p-value = 0.00), distribution_channel (Cramér's V = 0.1522, p-value = 0.00), and customer_type (Cramér's V = 0.1277, p-value = 3.85e-307). These findings suggest that factors such as room type consistency, guest origin, deposit conditions, booking methods, and customer type significantly influence cancellation rates. As a result, the variables "market_segment", "deposit_type", "distribution_channel" and "customer_type" are chosen to be encoded in the new dataset, whereas "country" has been excluded due to the enormous number of different countries present (almost all the world's countries). The "reservation_status" variable was excluded as well, to avoid overfitting.

To create a time series dataset, the numerical variables were aggregated by calculating the mean of bookings for each day. For the dummy variables, the sum of occurrences for each day was calculated and then converted into a percentage of occurrences for that day. For example, the percentage of guests paying with a credit card on a certain day was computed.

The final step in data preprocessing was transforming the dataset into a tsibble (Tidy Temporal Data Frames and Tools) object. A tsibble object in R preserves time indices as essential data columns and supports heterogeneous data structures, providing advantages over conventional time series objects like ts, zoo, and xts (Hyndman & Athanasopoulos, 2021; Wang).

3.3. EDA and Data Visualization

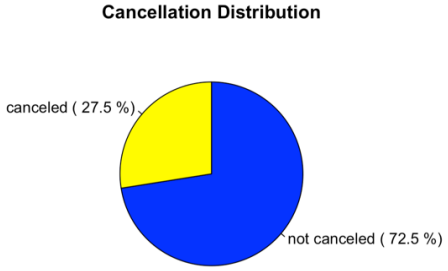
The exploratory data analysis (EDA) and data visualization for this study can be divided into two sections: before and after the aggregation of the dataset.

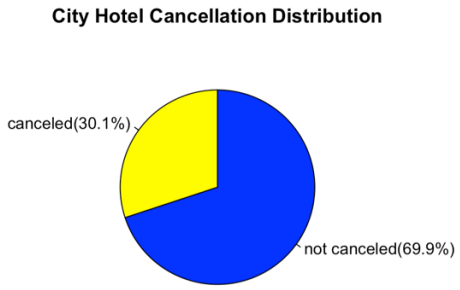
Before the aggregation

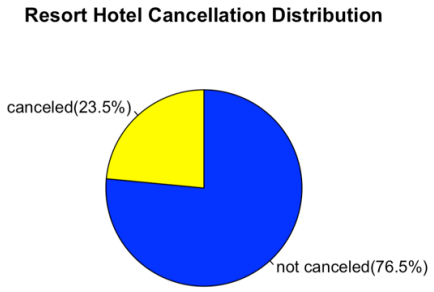
The first step involved plotting boxplots for numerical variables to identify outliers that could impact the analysis. In particular, they were detected in the "kids" and "adr" columns, leading to their removal. Subsequently, boxplots and stripplots were used to analyse how numerical variables affect cancellation behaviour. Several key observations emerged:

- Being lead time the number of days elapsing between the booking and the arrival, from the boxplot emerged that bookings made a few days before the arrival date are rarely cancelled, whereas bookings made over one year in advance are cancelled very often. The same is for days in the waiting list, meaning that the more the request remain pending, the more probable is the cancellation.
- Bookings that eventually get cancelled often have a history of previous cancellations, which could be indicative of habitual behaviour or booking strategies that include speculative bookings.
- There's a clear concentration of bookings involving fewer guests (up to about 4 guests) across both cancelled and not cancelled bookings, with this range appearing quite densely populated. Notably, there are outliers with a larger number of guests, particularly in the cancelled bookings, which could imply that larger group bookings are more prone to cancellation – it could be due to the logistical complexities or changing circumstances impacting larger groups more significantly.

The distribution of booking cancellations was then visualized both generally and specific to hotel types.

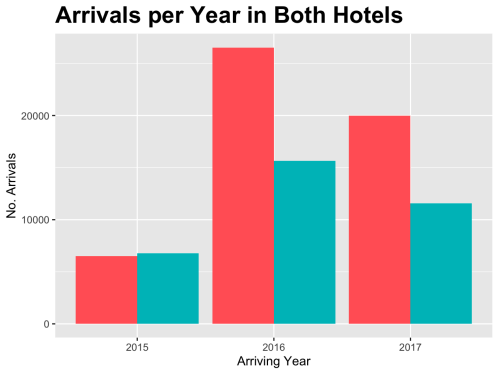


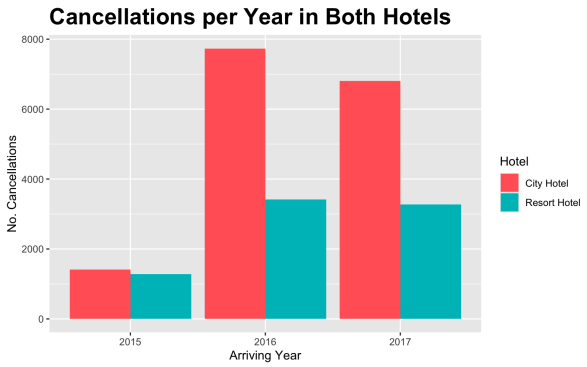




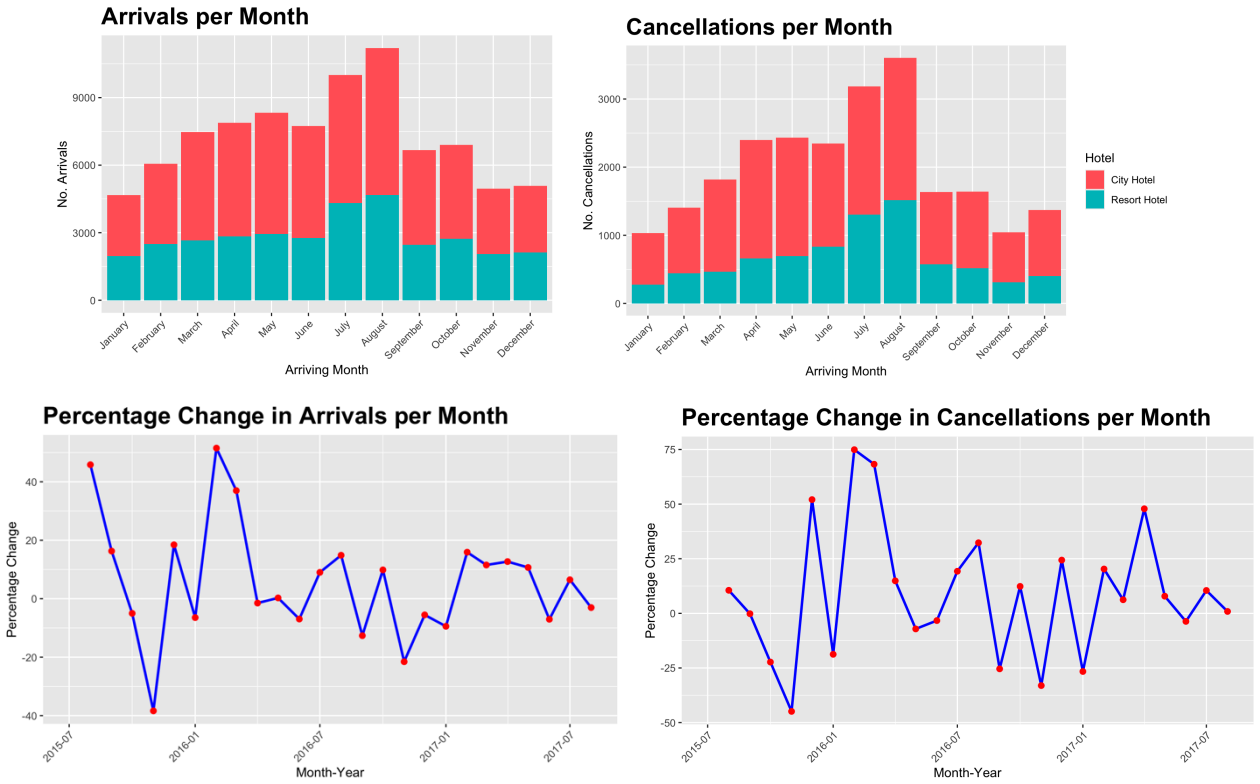
Approximately 27.5% of all bookings are cancelled, while 72.5% are not cancelled. This provides a baseline for comparing cancellation rates across different hotel types. Cancellation rates are slightly higher in city hotels, at 30.1%, this may be due to the dynamic nature of city travel, where plans can change rapidly. Resort hotels show a lower cancellation rate of 23.5%, with a more stable booking pattern, because resort stays are often planned well in advance and less subject to last-minute changes.

The number of arrivals and cancellations was visualized by year and month, based on hotel type. City hotels consistently experience higher traffic compared to resort hotels, likely due to their accessibility and appeal to both business travelers and tourists.





Looking at yearly data, a peak in 2016 can be noticed, especially for city hotels. A detailed look at the month-by-month arrivals shows a pronounced seasonality effect.



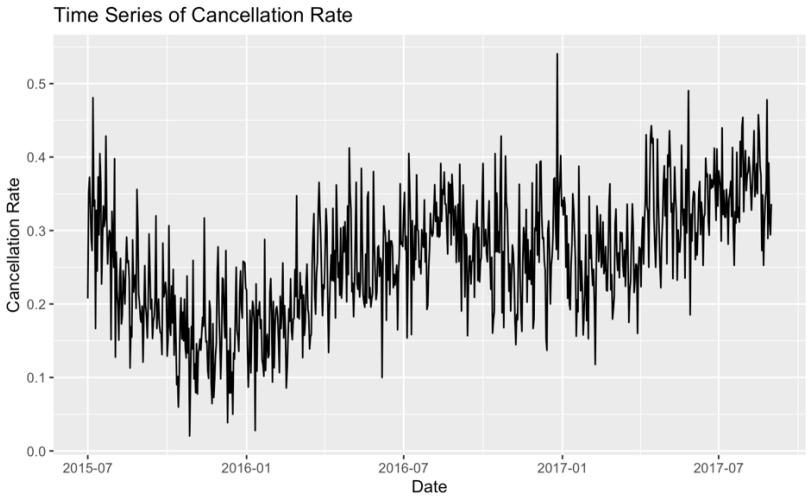
For both types of hotels, the summer months and December produce a strongly positive increase of arrivals and cancellations, coinciding with holiday seasons and warmer weather, which is particularly favourable for resort hotels. Interestingly, cancellations seem to be highly correlated to arrivals, indeed they are subject to the same variations.

Cancellation trends also underscore seasonality, with both types of hotels showing peaks in cancellations around the summer months, which could be attributed to the high volume of bookings and possibly more uncertain travel plans during this peak period.

After aggregation

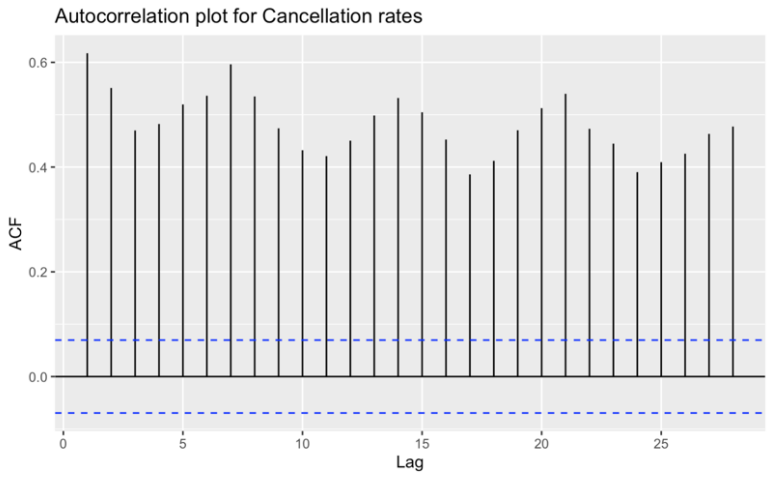
Once the aggregation has been made, the dataset is now a time series, so it is visualized with graphics specifically for time series.

The first plot has been executed through the autoplot function, which should give back the most suitable plot for the input type of data. For time series data, the obvious graph to start with is a time plot. That is, the observations are plotted against the time of observation, with consecutive observations joined by straight lines. (Hyndman & Athanasopoulos, 2021)



In this case, it resulted in a daily plot of the cancellation rates throughout the entire period of the study. The time series graph of the cancellation rate from mid-2015 through mid-2017 shows a marked variability, with fluctuations ranging from about 10% of bookings cancelled to over 40%. The data points suggest a lack of stable seasonality, although there are noticeable spikes, which may indicate periods of higher cancellation rates possibly related to specific events or seasonal trends.

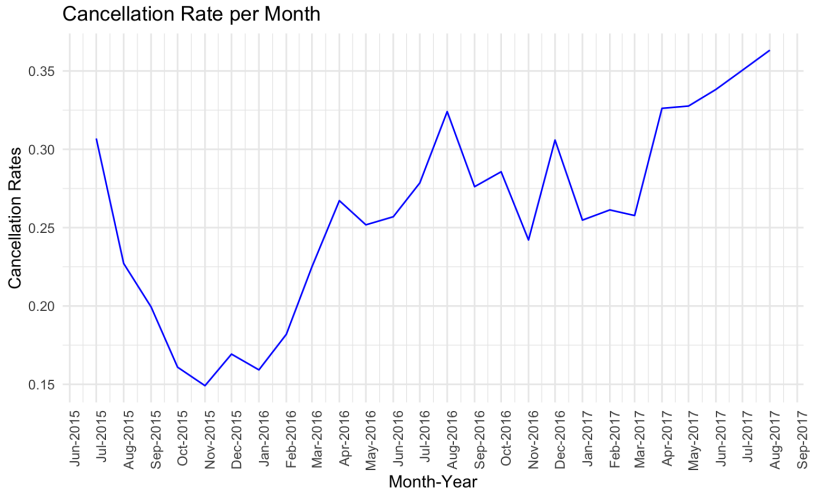
The plot also shows some extreme outliers, particularly high peaks (around January 2017), which could be instances of large-scale cancellations possibly due to specific external events (e.g., natural disasters, strikes, or political instability) impacting travel plans during those periods. The overall trend seems to exhibit a gradual increase in cancellation rates over the two years, especially noticeable from early 2017 onwards.



The autocorrelation function (ACF) graph shows the autocorrelations of the dependent variable "is_canceled_rate" time series at different lags.

In particular, from this graph it is evident that there is seasonality because there is a repeating pattern in the peaks, which repeat every 7 lags. Yet, there is not a clearly recognizable trend because ACF of a trended time series tends to have positive values that slowly decrease as the lags increase (Hyndman & Athanasopoulos, 2021), here instead it is not decreasing nor increasing.

The line chart showing the cancellation rate per month from June 2015 to September 2017 illustrates significant variability in cancellation rates over time. Notably, there are peaks and troughs that could indicate seasonal patterns or specific external events influencing cancellation behaviour.



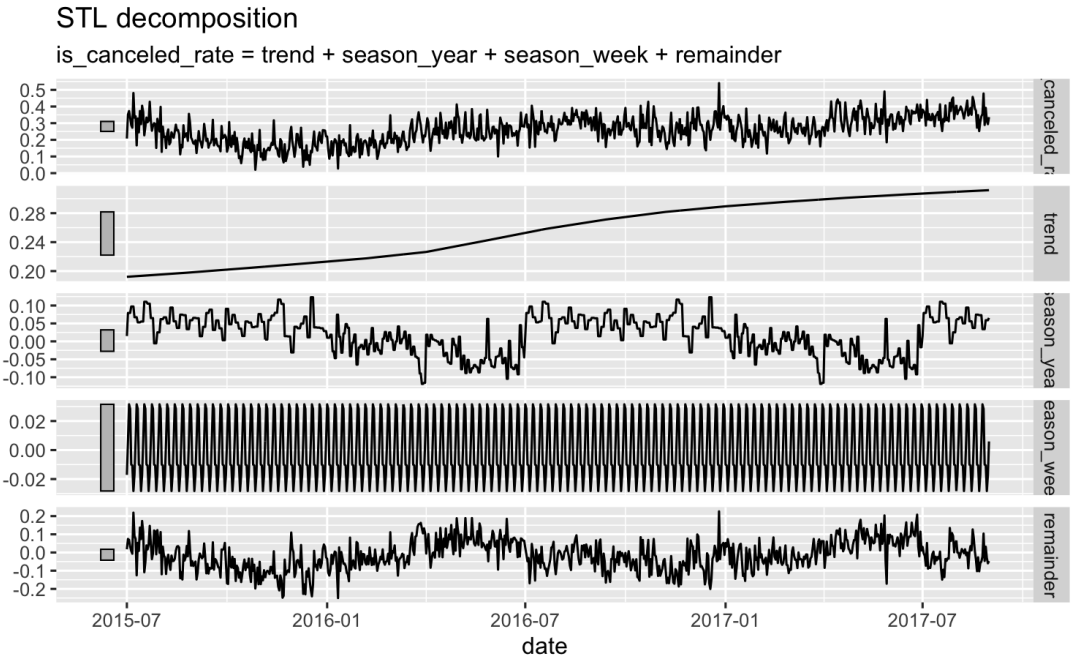
The graph starts with a sharp decline from June to December 2015, suggesting a period of higher stability or effective strategies in reducing cancellations. From early 2016, the cancellation rate gradually increases, peaking around mid-2016 and again in mid-2017. These peaks could be associated with seasonal factors where certain times of the year, possibly summer or major holiday seasons, experience higher cancellation rates due to changes in traveler plans. The dip in early 2017 followed by a steep rise through to September 2017 might indicate external factors or changes in booking or cancellation policies that affected guest behaviour. Alternatively, this could also reflect broader economic or travel industry trends affecting consumer confidence and decision-making. Understanding these trends is crucial for anticipating periods of high cancellation rates and could help in strategizing better booking policies, pricing adjustments, and promotional activities to mitigate potential revenue losses.

3.4. Modelling and Validation

As is conventionally done in the construction of machine learning predictive models, the data set was divided into two stratified subsets, one with 70% of data for training (model learning) and another with the remaining 30% to test the developed model. (Antonio et al., 2017). To perform a 70/30 split of the time series data spanning from 1 July 2015 to 31 August 2017, I first calculated the total time span in days, which amounted to 792 days. By multiplying this total by 0.7, I determined the split point to be approximately 554 days from the start date. Adding these 554 days to the start date, 1 July 2015, gave a split date of 5 January 2017. Using this split date, I divided the data into a training set containing observations up to and including 5 January 2017, and a testing set containing observations after this date. This method ensured that approximately 70% of the data was used for training and 30% for testing, maintaining the temporal order of the time series.

STL

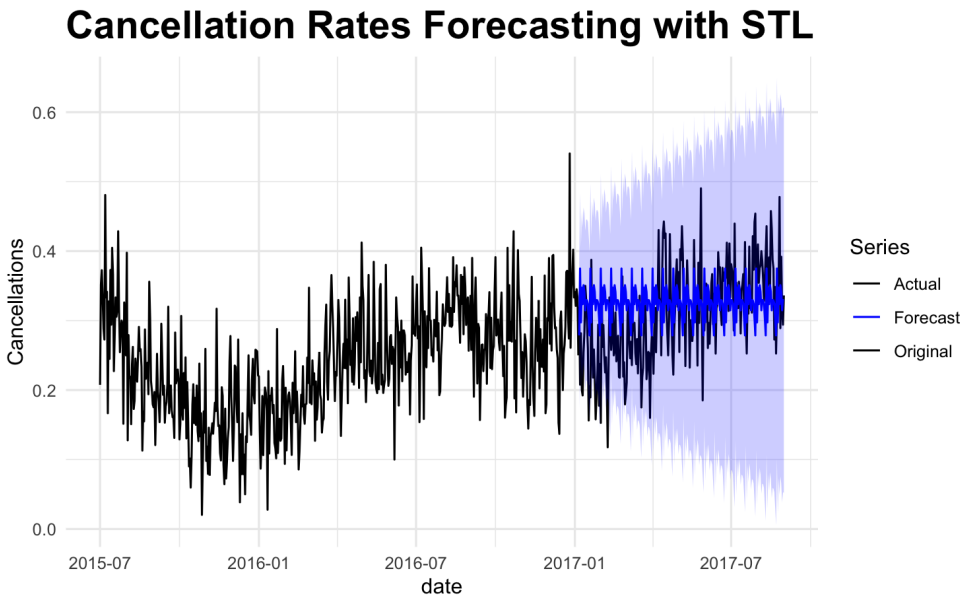
The first thing that has been done is plotting the three components (trend, seasonal, and remainder) to analyze them.



In the first plot there is simply the distribution of the dependent variable, which was already analyzed before. What's interesting are the three plots underneath. In particular, the trend plot shows a gradual increase over time, suggesting that the cancellation rate

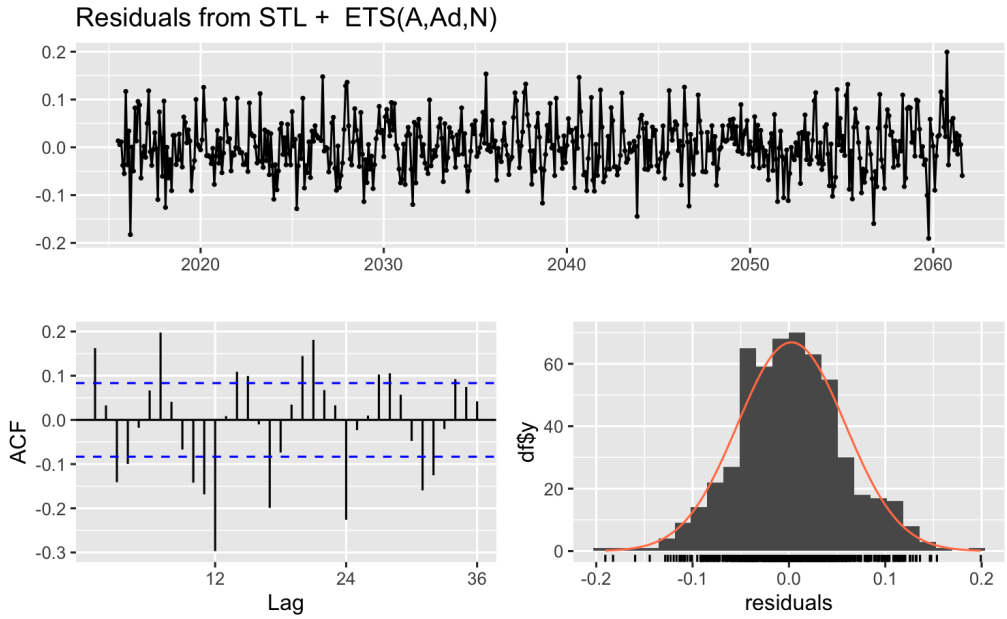
has gradually risen throughout the observed period. Then for the seasonal component we can notice something very valuable: there is not a repeating annual pattern, which would indicate a regular influence of specific months or seasons on cancellation rates; however, it shows very strong and consistent weekly patterns. The clear, regular oscillations suggest that the cancellation rate is significantly influenced by the day of the week. This could be attributed to behaviors such as weekend travels or weekly booking reviews by customers. Finally, the remainder or residuals don't show any additional obvious pattern or structure, suggesting that the trend and seasonal components have captured most of the systematic information in the data.

The forecasting was conducted using the "stlf" function, which calculates the seasonal component by selecting a model suitable for this task (Najera). The most common methods are the ARIMA and ETS (Exponential Smoothing) models. These models effectively facilitate the calculation of seasonality once the trend has been established. In this study, the ETS method was specified within the function. The use of the ETS model allowed to capture, in an effective way, the behavior of the component of the trend, in such a way that the only thing that "remains" of the series is white noise, that is, random variations that cannot be predicted.



From the plot, the STL (Seasonal and Trend decomposition using Loess) method seem effectively captured the seasonality in the cancellation rates. The red forecast line exhibits consistent seasonal patterns that align quite well with the observed seasonal fluctuations

in the historical data. This performance suggests that the STL method is suitable for time series data with strong seasonal components. Further validation through accuracy metrics provides a comprehensive assessment of the STL model's efficacy: RMSE = 0.07533898



Eventually, a residual diagnostic of the model is computed, to check that the model is utilizing all available information and is providing unbiased, efficient forecasts. From the first image, the residuals fluctuate around zero without any systematic deviation, which indicates a good model fit without apparent bias. The second plot measuring the autocorrelation, suggests that the residuals are white noise – i.e., there is no autocorrelation in the residuals. Finally, the distribution of the residuals seems to be approximately with a normal distribution (red curve), being a good sign for the reliability of the model's prediction intervals.

ARIMA

First of all, we check if the dataset is stationary or not with a couple of tests, to decide whether differencing is needed on the dataset.

Augmented Dickey-Fuller Test

If a time series has no trend, constant variance over time, and a consistent autocorrelation structure across time, it is considered to be “stationary.” An augmented Dickey-Fuller test, which uses the following null and alternative hypotheses to determine whether a time series is stationary, is one technique to do so.

H0: The time series is non-stationary. To put it another way, it has some time-dependent structure and does not exhibit constant variance over time. HA: The time series is stationary.

The p-value < 0.01 so the null hypothesis can be rejected, and the result indicates that the time series is stationary. In other words, it doesn't show some time-dependent structure and does not exhibit constant variance over time.

Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test

This test uses the following null and alternative hypothesis:

$\mathbf{H_0}$: The time series is trend stationary. $\mathbf{H_A}$: The time series is ***not*** trend stationary.

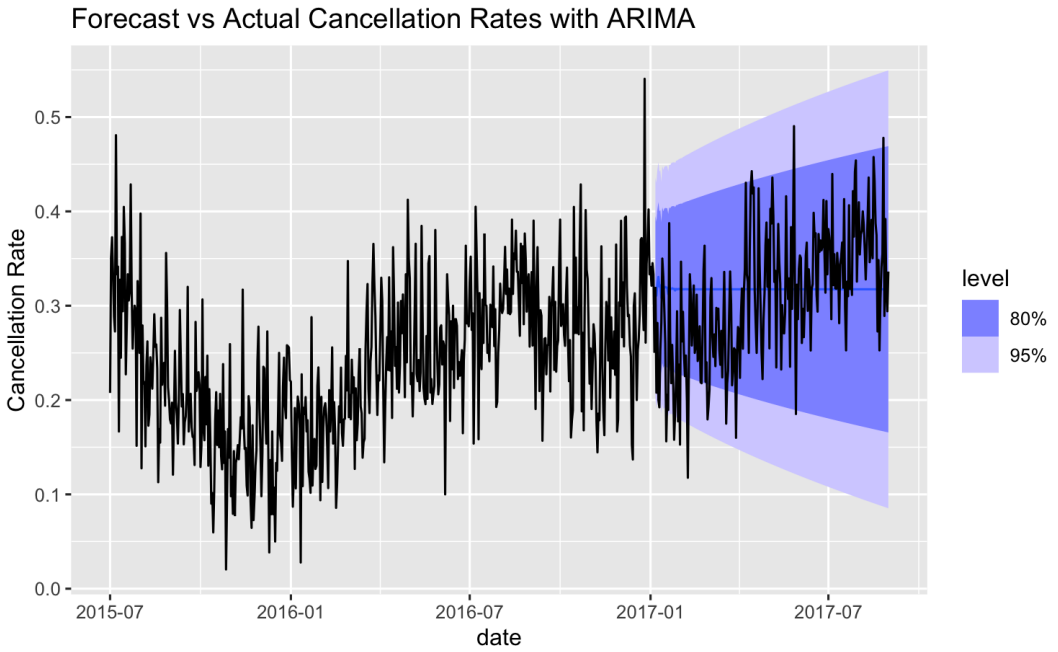
If the p-value of the test is less than some significance level (e.g. $\alpha = .05$) then we reject the null hypothesis and conclude that the time series is not trend stationary

The p-value is 0.01, therefore the null hypothesis is rejected again of the KPSS test. This means we can assume that the time series is not trend stationary.

Time series is stationary according to ADF test, but not according to KPSS test. Therefore, we need differencing to make the time series stationary. Many statistical models require the time series to be stationary to make effective and precise predictions. This is the case of ARIMA models. A very common way to make a time series stationary is differencing: from each value in our time series, we subtract the previous value.

Using the function auto.arima, it returns best ARIMA (p, d, q) model according to either AIC, AICc or BIC value. (Hyndman)

In this case we have ARIMA (3,1,1), meaning that it used 3 lags, it differenced only once, and the size of the moving average window also is 1. The plot resulting from forecasting on the test set:

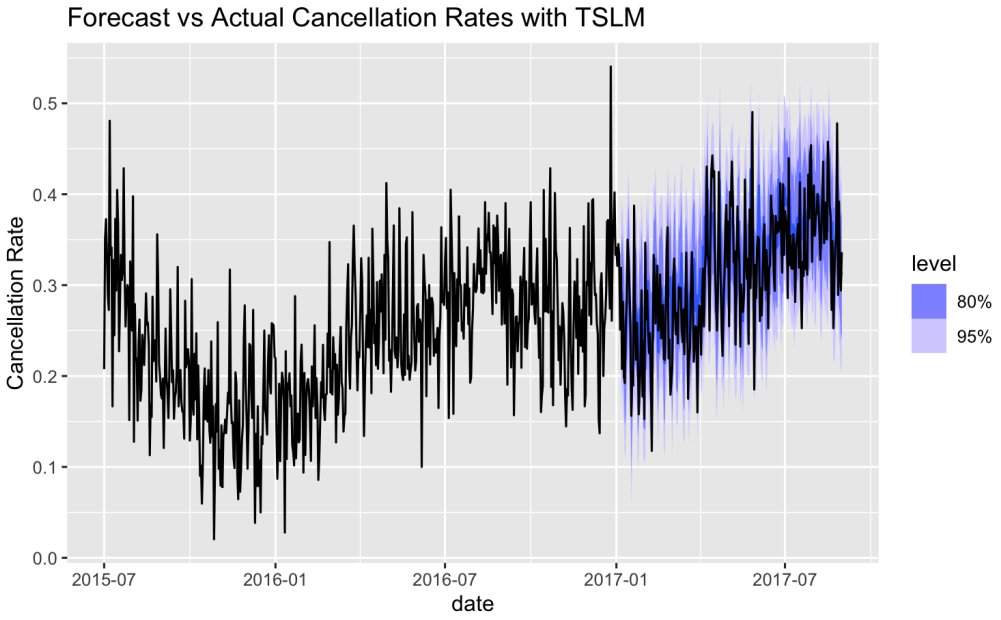


We can see from the chart that most of the test observation values lie within the 95% confidence band, even if the predicted forecasts are consistently different from the actual values because the first ones seem to mimic the mean of the actual values.

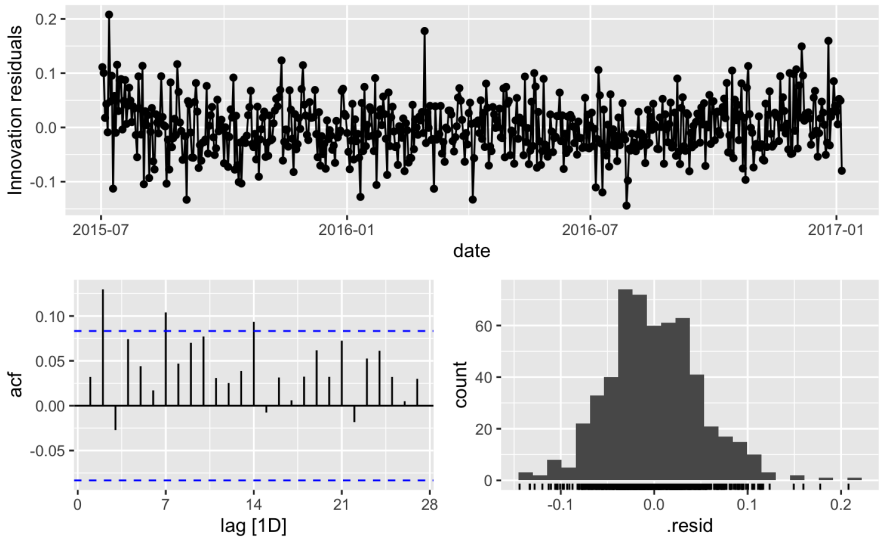
The RMSE for the Arima model is 0.06981135.

REGRESSION for TIMESERIES

When computing a multiple linear regression, we need some strategy for selecting the best predictors to use in the model. If this case there are too many predictors, to fit all possible models and evaluate them one by one with a measure like AIC or BIC. Therefore, the strategy applied was the stepwise regression, in particular the 'both-direction' one, where the algorithm combines both forward and backward steps, optimizing the model by adding significant variables and removing insignificant ones. Each model was evaluated automatically with the Akaike's Information Criterion and the best model found included variables like the lagged variable, the number of total guests, the lead time, the required car parking spaces, etc. for a total of 17 predictors.



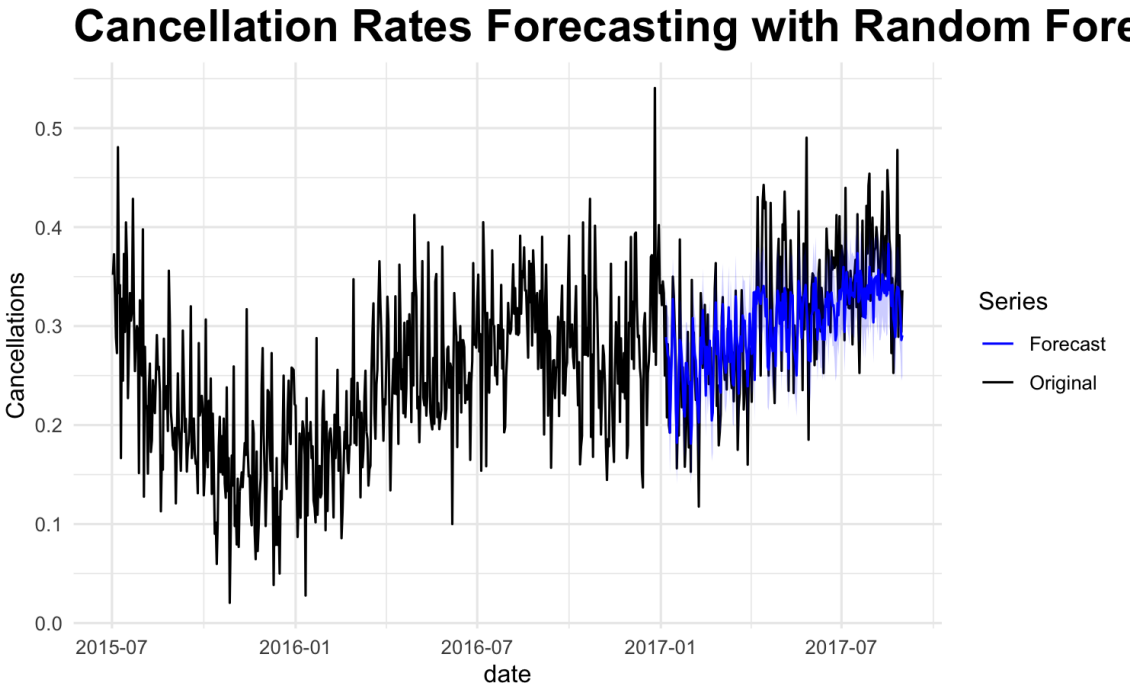
By looking at the visualization of the forecasting, it is evident how the model performs well. Indeed, the forecasted values seem to capture almost perfectly the actual ones. Moreover, doing the residual diagnostics, they seem to respect all the condition needed to be considered as a good fitting model.



The RMSE for the regression model is 0.04788151.

RANDOM FOREST

The random forest model has been built with the randomForest function with the training set and with 500 trees to be grown, chosen arbitrarily. Then forecasting on the test set is performed and the visualization of the result is the following:



The forecasted values achieve very good result in predicting the actual ones. However, the model seems to perform a little bit worse than the previous one. Unfortunately for the random forest as well as for the Arima model is not possible to check residuals, and there is only the RMSE that gives a comparison measure, which in this case is equal to 0.04960623

3.5. Results and Performance Comparison

To resume the results of the previous paragraph a table is provided.

MODEL	RMSE
STL Decomposition	0.07533898
ARIMA	0.06981135
Regression for Timeseries	0.04788151
Random Forest	0.04960623

Table 1 Source: Self-Elaboration

As stated before, the RMSE values provided for different forecasting models reveal the comparative accuracy of each model applied to the same dataset. STL Decomposition and ARIMA models show moderate accuracy with RMSEs of 0.07533898 and 0.06981135, respectively, so they are performing poorly with respect to the other two.
Conversely, the Regression for Timeseries and Random Forest models exhibit enhanced predictive performance, with RMSEs of 0.04788151 and 0.04960623, indicating more precise forecasts. This suggests that both the Regression for Timeseries and Random Forest are better at capturing the complexities of the dataset, with the Regression for Timeseries model being slightly more accurate. In this case, if selecting a model for deployment, one might be inclined towards these two, particularly towards the Regression for Timeseries model due to its slightly lower RMSE, indicating higher accuracy on this specific dataset. Surprisingly, the Multiple Linear Regression is the model that performed better and probably also the fact of predicting with a lagged independent variable was useful, meaning that for this type of problem is useful forecasting based on past data.

4. CONCLUSIONS

Through this empirical thesis, the goal to explore and compare the performance of various regression algorithms in predicting hotel booking cancellation rates was pursued; the final aim was to provide valuable insights to offer to the hospitality industry in particular. As acquired above, accurate forecasting of cancellations is crucial for effective revenue management in the hospitality industry, allowing hotels to optimize their booking strategies, mitigate potential revenue losses, and improve overall profitability. Through a detailed analysis involving data preprocessing, exploratory data analysis, and model training and validation, the study provides significant results about the most powerful method among the proposed ones, for forecasting booking cancellation rates.

The dataset chosen for this research, the "Hotel Booking Demand" dataset, was meticulously processed and transformed into a time series format – through aggregation – to facilitate the application of time series forecasting models. Initial data exploration was fundamental to handle the data inconsistencies.

At the core part of this document, four different models were applied and compared: ARIMA, STL decomposition, multiple linear regression for time series, and random forest. Each model was built in order to exploit its unique advantages, offering insights into the data's underlying patterns. The ARIMA model, a traditional method for time series forecasting, proved moderately effective with an RMSE of 0.06981135. STL decomposition, which leverages LOESS for flexible decomposition, captured the seasonal patterns well, yet not the trends, indeed had a higher RMSE of 0.07533898.

The multiple linear regression model for time series, availing of lagged variables, exhibited the highest accuracy with an RMSE of 0.04788151. This model's ability to account for various predictors and their interactions likely contributed to its superior performance. The random forest model also performed well, with an RMSE of 0.04960623, highlighting its robustness and adaptability to complex datasets.

The findings indicate that while specific time series models like ARIMA and STL are valuable, the multiple linear regression model for time series and random forest models offer enhanced accuracy for predicting hotel booking cancellations. The regression model's slight edge suggests that incorporating lagged predictors and understanding the relationships between variables can significantly improve forecasting precision.

This study's results underlines the importance of selecting appropriate models based on the specific characteristics of the dataset and the forecasting objectives. For hotel managers and revenue management professionals, these insights can inform better decision-making processes, enabling more effective booking strategies and ultimately enhancing profitability.

In conclusion, the multiple linear regression model for time series emerged as the most accurate predictor of hotel booking cancellations in this study. However, the predictors should always be properly evaluated and, in general, the choice of model should consider the context and specific requirements of the forecasting task. Future research could explore the improvement of more sophisticated machine learning techniques by parameter tuning and the incorporation of additional external factors to further improve forecasting accuracy.

5. BIBLIOGRAPHY

Anderson, C. K. (2012). *The Impact of Social Media on Lodging Performance. Cornell Hospitality Report, 12(15), 4–11.*

Antonio, N., Almeida, A. de, & Nunes, L. (2017). Predicting hotel booking cancellations to decrease uncertainty and increase revenue. *Tourism & Management Studies*, *13*(2), 25–39. https://doi.org/10.18089/tms.2017.13203

Antonio, N., de Almeida, A., & Nunes, L. (2017). Predicting Hotel Bookings cancellation with a machine learning classification model. *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. https://doi.org/10.1109/icmla.2017.00-11

Antonio, N., de Almeida, A., & Nunes, L. (2019a). Big Data in Hotel Revenue Management: Exploring cancellation drivers to gain insights into booking cancellation behavior. *Cornell Hospitality Quarterly*, *60*(4), 298–319. https://doi.org/10.1177/1938965519851466

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

António, N., Almeida, A. de, & Nunes, L. (2019). Predictive models for hotel booking cancellation: A semi-automated analysis of the literature. *Tourism & Management Studies*, *15*(1), 7–21. https://doi.org/10.18089/tms.2019.15011

Brownlee, J. (2023, November 18). *How to create an Arima model for time series forecasting in Python*. MachineLearningMastery.com. https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/

C.-Sánchez, E., Sánchez-Medina, A. J., & Romero-Domínguez, L. (2022). Forecasting hotel-booking cancelations using personal name records: An Artificial

Intelligence Approach. *Marketing and Smart Technologies*, 3–14. https://doi.org/10.1007/978-981-16-9268-0_1

Chatterjee, S. (2020). Drivers of helpfulness of online hotel reviews: A sentiment and emotion mining approach. *International Journal of Hospitality Management*, *85*, 102356. https://doi.org/10.1016/j.ijhm.2019.102356

Chen, C.-C., Schwartz, Z., & Vargas, P. (2011). The search for the best deal: How hotel cancellation policies affect the search and booking decisions of deal-seeking customers. *International Journal of Hospitality Management*, *30*(1), 129–135. https://doi.org/10.1016/j.ijhm.2010.03.010

Chiang, W. C., Chen, J. C. H., & Xu, X. (2007). An overview of research on revenue management: Current issues and future research. *International Journal of Revenue Management*, *1*(1), 97. https://doi.org/10.1504/ijrm.2007.011196

Claveria, O., & Datzira, J. (2010). Forecasting tourism demand using consumer expectations. *Tourism Review*, *65*(1), 18–36. https://doi.org/10.1108/16605371011040889

Denizci Guillet, B., & Mohammed, I. (2015). Revenue management research in hospitality and tourism. *International Journal of Contemporary Hospitality Management*, *27*(4), 526–560. https://doi.org/10.1108/ijchm-06-2014-0295

Frost, J. (2023, May 28). *Root mean square error (RMSE)*. Statistics By Jim. https://statisticsbyjim.com/regression/root-mean-square-error-rmse/

Hayes, D. K., & Miller, A. A. (2011). *Revenue Management for the Hospitality Industry. Hoboken, NJ: John Wiley.*

Hueglin, C., & Vannotti, F. (2001). Data mining techniques to improve forecast accuracy in airline business. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/502512.502578

Hyndman, R. (n.d.). *Fit best arima model to Univariate time series - auto.arima*. - auto.arima • forecast. https://pkg.robjhyndman.com/forecast/reference/auto.arima.html

Hyndman, R. J., & Athanasopoulos, G. (2021, May). *Forecasting: Principles and practice (3rd ed)*. OTexts. https://otexts.com/fpp3/

Iliescu, D. C., Garrow, L. A., & Parker, R. A. (2008). A hazard model of US airline passengers' refund and Exchange behavior. *Transportation Research Part B: Methodological*, *42*(3), 229–242. https://doi.org/10.1016/j.trb.2007.10.005

Ivanov, S. (2014). *Hotel Revenue Management: From Theory to Practice. Varna, Bulgary: Zangador.*

Ivanov, Stanislav, & Zhechev, V. S. (2011). Hotel Revenue Management – A Critical Literature Review. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1977467

Mehrotra, R., & Ruttley, J. (2006). Revenue management(second ed.). Washington, DC, USA: American Hotel & Lodging Association (AHLA). (n.d.).

Mostipak, J. (2020, February 13). *Hotel Booking Demand*. Kaggle. https://www.kaggle.com/jessemostipak/hotel-booking-demand/data#

Najera, J. (n.d.). Time Series forecast using the STL model. https://openreview.net/pdf?id=rylYCmQ_1H

Nau, R. (n.d.). *Fitting time series regression models*. Why do simple time series models sometimes outperform regression models fitted to nonstationary data? https://people.duke.edu/~rnau/timereg.html

Noone, B. M., & Lee, C. H. (2011). Hotel overbooking. *Journal of Hospitality & Tourism Research*, *35*(3), 334–357. https://doi.org/10.1177/1096348010382238

Novakovic, J., & Turina, S. (2021). Hotel reservation cancellations: analysis and prediction using machine learning algorithms. *International Academic Journal, Vol. 2, Issue 1. 2021*, 4–13.

O., P. (2016). *Airline Passenger Cancellations: Modeling, Forecasting and Impacts on Revenue Management. (M.Sc. Thesis). Massachusetts Institute of Technology, Boston.*

http://hdl.handle.net/1721.1/104325

Pfeifer, P. E., & Bodily, S. E. (1990). A test of space-time Arma Modelling and forecasting of Hotel Data. *Journal of Forecasting*, *9*(3), 255–272. https://doi.org/10.1002/for.3980090305

Rajopadhye, M., Ben Ghalia, M., Wang, P. P., Baker, T., & Eister, C. V. (2001). Forecasting uncertain hotel room demand. *Information Sciences*, *132*(1–4), 1–11. https://doi.org/10.1016/s0020-0255(00)00082-7

Romero Morales, D., & Wang, J. (2010). Forecasting cancellation rates for services booking revenue management using data mining. *European Journal of Operational Research*, *202*(2), 554–562. https://doi.org/10.1016/j.ejor.2009.06.006

Talluri, K. T., & Van Ryzin, G. J. (2004). The theory and practice of Revenue Management. *International Series in Operations Research & Management Science*. https://doi.org/10.1007/b139000

Tekin, M., & Gök, M. (2021). *Performance Comparison of Classification Algorithms in Hotel Booking Cancellation Prediction. Artificial Intelligence Theory and Applications, 1(1), 8-19.*

Tranmer, M., Murphy, J., Elliot, M., & Pampaka, M. (2020). Multiple Linear Regression (2nd Edition). *Cathie Marsh Institute Working Paper 2020-01*.

https://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/working-papers/2020/multiple-linear-regression.pdf

Wang, E. (n.d.). *Introduction to tsibble*. cran.rstudio. https://cran.rstudio.com/web/packages/tsibble/vignettes/intro-tsibble.html