d











**INDEX**

[Introduction and Business Understanding](#_gjdgxs) **3**

[Data Preprocessing](#_1fob9te) **3**

[Data Understanding](#_mqy05uz4q99x) 3

[Feature Engineering](#_3znysh7) 3

[Duplicates in Data](#_2et92p0) 4

[Missing values](#_tyjcwt) 4

[Data Exploration](#_3dy6vkm) **4**

[Data Transformation](#_1t3h5sf) **5**

[Feature Selection](#_4d34og8) **6**

# Introduction and Business Understanding

In a world of high global competition and further ongoing digitalization the hospitality industry is facing not only opportunities, but also challenges. Due to the rise of online travel agencies that try to satisfy the high customer demands and putting pressure on hotels for loose cancellation rules, hotels are challenged with a high number of booking cancellations. The objective of this case is to develop a model that reduces the uncertainty about demand by predicting booking with a high likelihood of canceling. Therefore, we received a dataset from a Portuguese chain hotel company that would like to reduce uncertainty about demand, implement better pricing and overbooking policies and identify bookings that will likely be canceled to prevent cancellations by contacting the customers; the goal behind this is to reduce the cancellation rate from 41.7% to 20%.

It is important to understand how cancellation policies impact the business. Free cancellation policies are usually linked with overbooking strategies; however, overbooking has some issues such as re-allocation costs for customers, social reputation damage and loss of future or immediate revenue. The more conservative approach, restrictive cancellation policy, has some issues as well such as decrease in revenue and number of bookings. In our case, it is possible to observe how 43.1% of revenue comes from bookings being canceled.

| **Metric** | **Not Canceled** | **Canceled** | **Total** |
| --- | --- | --- | --- |
| Bookings | 46,228 (58.3%) | 33,102 (41.7%) | 79,330 (100%) |
| Room Revenue | 14,394,410€ (56.9%) | 10,885,060€ (43.1%) | 25,279,470€ (100%) |

# Data Preprocessing

## Data Understanding

The dataset consists of 79,330 booking entries recorded in the years between 2015 and 2017. Besides the information about the booking and the reservation date each record contains information about the guests (e.g. adults, origin, previous cancel behavior), the particular booking (e.g. Room Type, Stays in Weekend Nights), the booking agent and the target value if the booking was canceled or not. The distribution of the target value showed that we are dealing with a slightly unbalanced data set in favor of the non-canceled bookings (58.27%). In total the original dataset contained 31 columns with different data types (categorical, continuous, discrete). A listing of all features can be found in the Appendix.

## Feature Engineering

Engineering some feature was needed due to the fact that the original dataset showed some inconsistencies with the date formats. We created an *ArrivalDate* variable, that combined the information about the arrival date from different features (eg. ArrivalDateYear, *ArrivalDateDayOfMonth*) in one variable in datetime format. The exact opposite was done to the given feature *ReservationStatusDate* to get additional information from splitting it up into variables like *ReservationMonth* or *ReservationDayOfWeek*. Furthermore, new features that seemed logically connected were created: (1) *DurationOfStay* is the sum of *StaysInWeekNights* and *StaysInWeekendNights*. (2) *TotalValue* is the product of *DurationOfStay* and *ADR* (Average Daily Rate) and contains the amount of money a guest was paying for the whole stay. (3) *IsRoomChanged* is a binary variable that indicates if the guest got another room than specified on the reservation.

## Duplicates in Data

The investigation for duplicates in the dataset returned a value of 32%. This surprising high value was explained by the fact that most of these duplicates were within *MarketSegment* type “Group”. Each guest of a booking made by a group was recorded individually but with the same information as the whole group. Keeping all of the values would have resulted in overfitting, which led to the the decision of keeping one entry per group involved in duplicates and discarding the others. This treatment resulted in a filtered dataset 85.7% the size of the original data. Unfortunately, the removal of these duplicates further increased the imbalance in the target variable *IsCanceled* to 64.16%.

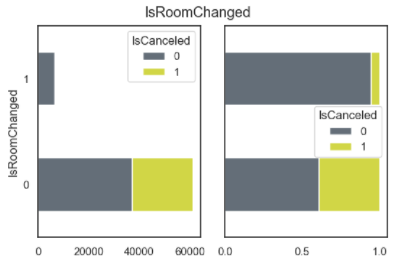
## Missing values

Missing values were detected in *Children* (4) and *Country* (24). On account of these small numbers and the urge to not lose any further information, the missing values were replaced by the most common value (mode) for the regarding feature.

# Data Exploration

The visual inspection of the prepared dataset gave some interesting insights but also further challenges to deal with. The plotted histograms and boxplots of the metric variables showed that some features contained outliers that ought to be taken care of. E.g. an unlikely booking with 10 *babies* or a booking with an *ADR* of 5,400 (*ADR* median was 99).

The correlation matrix , based on Spearman’s Rho, chosen in regard of existing numeric but not metric values (e.g. *Children*), showed the obvious strong correlations between the new created variables and their base variables. This correlation method was chosen by analyzing the scatterplot of metric features. Furthermore, slight positive correlations (**≈**0,3) were detected between *DurationOfStay* and *LeadTime* (the longer the stay, the more in advance guests tend to book) and between *TotalValue*/*ADR* and *Adults/Children*  (more guests lead to higher value). A moderate correlation (**≈**0,4) was discovered between *PreviousBookingsNotCanceled* and *PreviousCancellations*. This sounds surprising at first but can be explained by frequent travelers that have high absolute values in both, while occasional guests have very low values in both. The full correlation matrix can be found in the appendix.

The plotting of the numeric variables in histograms and boxplots with the target variable *IsCancelled* marked, gave a first insight into which variable might be useful for the target prediction. *LeadTime* was the variable with the highest indication since it showed that guests with a higher lead time (days in between booking and actual stay) are more likely to cancel (see figure on the right). Plotting the categoricals by their absolute and relative number of values and their influence on the target variable detected a few variables as promising. *IsRoomChanged*: If guests get another room type than they made the reservation for (probably through an upgrade) they are less likely to cancel. *DepositType*: all guests with non refundable deposit canceled. *IsRepeatedGuest*: first time guests are more likely to cancel. *CustomerType*: transients are more likely to cancel than e.g. groups. *DistributionChannel*: travel agencies have a higher rate of cancellations. *MarketSegment*: online travel agencies are more likely to cancel. *ReservationDayOfWeek*: bookings made on weekdays are more likely to get canceled than the ones made on weekends.

Some more interesting conclusions: ReservationDayOfWeek shows higher cancellation rates during the week. 1208 records (~5%) with ADR value = 0. Room P is always canceled. Visual analysis with Histograms of metric variables show us unbalanced data. By market segment we could analyze by far bigger cardinality in online TA.

# Data Transformation

* 1. **Scaling**

The first step in transforming the data was scaling the metric features with the Standard Scaler since the selected outlier detection method needed scaled data. This scaling procedure ensured that the mean is 0 and the standard deviation is 1 in respect of the feature’s variance. The Standard Scaler was chosen because it is a bit more resilient with existing outliers than other scalers (e.g. MinMaxScaler) and made a second scaling after the outlier removal obsolete.

* 1. **Outlier treatment**

For outlier detection a One-Class Support Vector Machine (OCSVM) with a contamination level of 2% (nu=0.02) was chosen. OCSVM is a Support Vector machine that can be used for outlier detection. By analyzing the records detected as outliers, this technique seemed to be an effective approach.

* 1. **Reducing Cardinalities**

Some categorical features we analyzed had many classes, and in some cases the cardinalities of these classes were not big enough to be considered relevant. We decided then to group those low cardinalities classes into a single one. *DistributionChannel* : “Undefined” and “GDS” into “Others”. *MarketSegment*: “Aviation”, “Undefined”, “Complementary” into “Others”. *ReservedRoomType* and *AssignedRoomType*: “G”, “P”, “C”, “K” into “Others”. *Company*: companies who made less than 100 bookings into “Others”. *Agent*: agents with less than 1000 records into “Others”. *Country*: countries showing less than 1000 times were grouped into “Others”

* 1. **Encoding**

To encode our categorical features we decided to implement a Label Encoder followed by the OneHotEncoder. The features being encoded are the ones selected by categorical feature selection.

# Feature Selection

* 1. **Categoricals**

The feature selection of the categorical variables was conducted with 3 different techniques (Chi2, F-Classifier, Mutual-Information-Classifier). It included all categorical variables except the target variable *IsCanceled* itself and the *ReservationStatus*, which is an information that is only known after the guest’s stay and not good for a predicting model. The results of the analysis are presented in the appendix. Based on these the chosen categorical features for building the model were: *IsRoomChanged, IsRepeatedGuest, ReservationDayOfWeek, ReservationMonth, CustomerType, Company, Agent, DepositType, AssignedRoomType, DistributionChannel, MarketSegment* and *Country.*

* 1. **Metrics**

For the metric variable feature selection three different approaches were chosen (Lasso regression / Least Absolute Shrinkage and Selection Operator, Mutual Info Classifier, Random Forest Classifier). The amount of features important for the target prediction according to Lasso with a value threshold of 0.02 was reduced from 15 variables to 8. The Mutual Info Classifier returned similar results. For the final selection of metric features, the Lasso features were chosen but *ADR* was removed, since it had a strong correlation with *TotalValue*.

1. **Cross-Validation**

To assess our models, we decided to use *TimeSeriesSplit* for cross validation in training data and *train\_test\_split* to divide 70/30 our data into training and testing.

We mainly had two issues to focus on: because of the time-series nature of our dataset, we had to be sure that train and validation data would be previous than the test data. This order has to be kept for train and validation as well, and it was obtained by sorting the values by *ReservationStatusDate*. The other issue is the imbalanced target feature for train and test data: to treat this, we decided to over-sample the minority classes with SMOTE. By doing that, some synthetic records were created using k-Nearest-Neighbors and we had a 50/50 balance in the target variable for both train and test data.

1. **Models**

We applied a variety of different models to obtain the most appropriate result of the case.

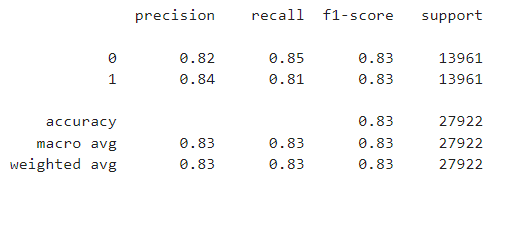
After the initial outcome of the models without much fine tuning, we were able to observe that it is very likely to have a greater accuracy; but yet alone this was not sufficient enough in helping to achieve the complete objective of the model which was to be able to predict right cancellations.

The performance of the classification model was examined with the confusion matrix as it is used to comprehend the essential predictive analytics. True positives are the cancellations that were predicted correctly and true negatives are the non-cancellations predicted correctly. Similarly, false positives are the non-cancellations that are predicted to be cancellations; it may result in overbooking which may be more costly to the hotel than to have an empty room. Contrarily, false negatives are the cancellation predicted to be non-cancellations; if there are alot of false negatives then the rooms will be empty and there is always an opportunity cost compromised which is to have the room rented. This overall will impact the revenue stream. Our goal was to not only predict the true positives and true negatives but was also to reduce the false positives and false negatives as these too had a major impact on the revenue stream.

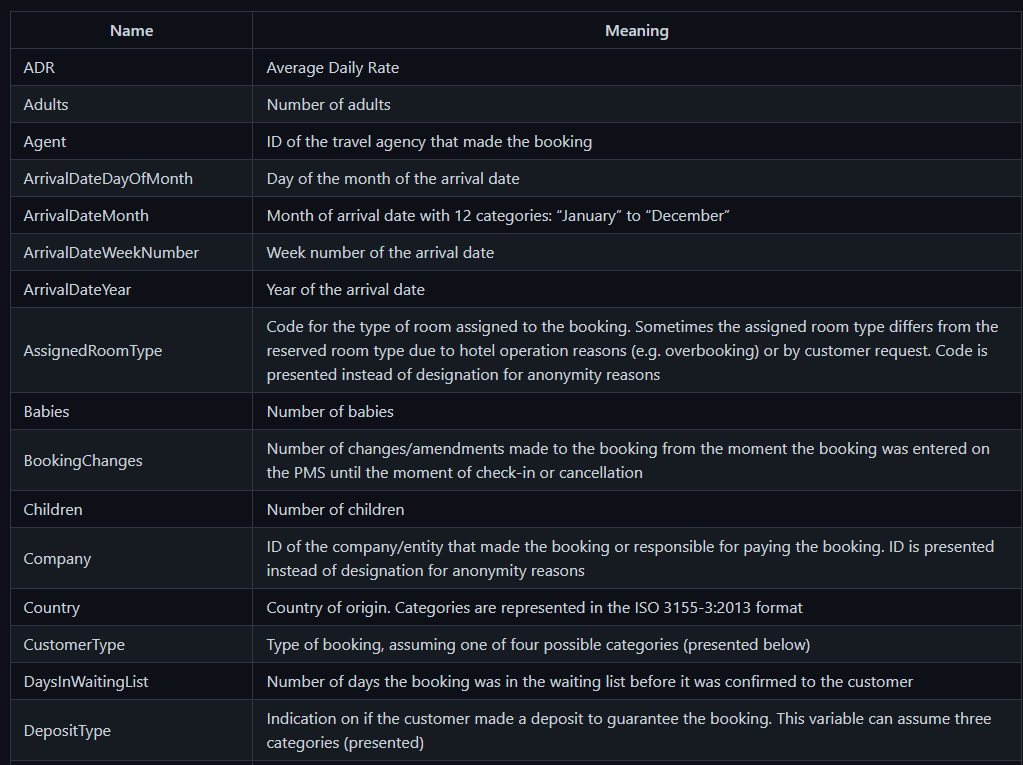
To begin with, we applied Logistics Regression and Decision Trees to our analysis but decided not to move forward with the results as the amount of false positives and false negatives were higher. The most adequate result obtained was from Random Forest with most correct true positive and true negative numbers and a balanced amount of false positives and false negatives.

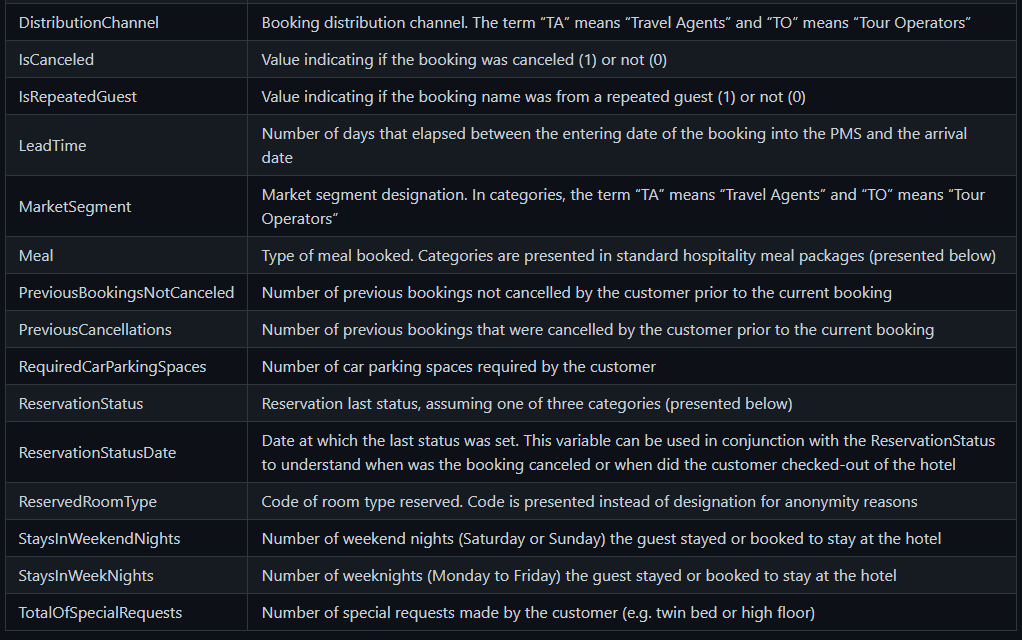
|  | Positive | Negative |
| --- | --- | --- |
| Positive | 11840 | 2121 |
| Negative | 2625 | 11336 |

All the classification reports were helpful in concluding that Random Forest performed better from all three models. The precision, recall and F1 score was better in Random Forest with a rounded value of 0.83 whereas for Decision Trees and Logistic regression it was 0.77 and 0.75 respectively.



1. **Appendix**

****

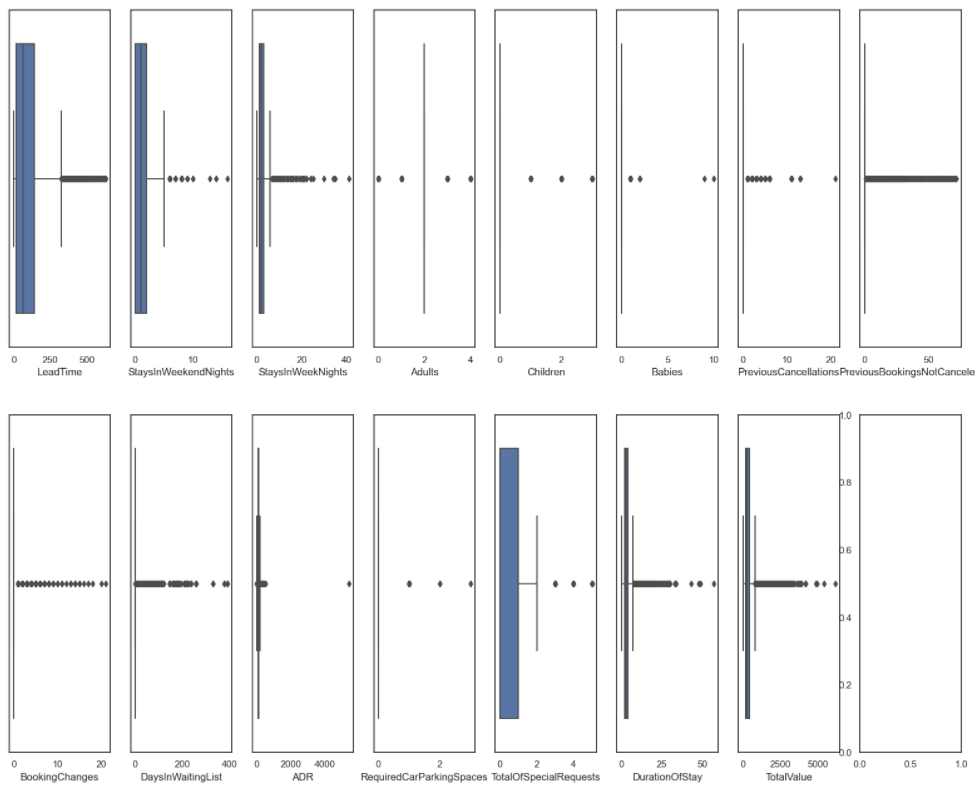
****

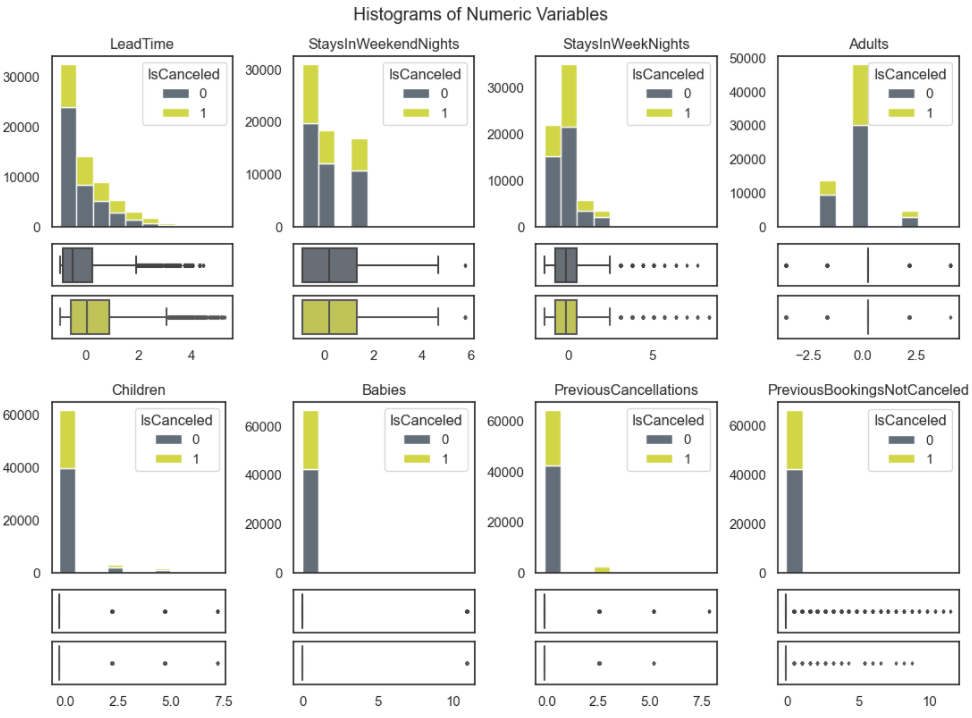
Gráfico, Gráfico de barras

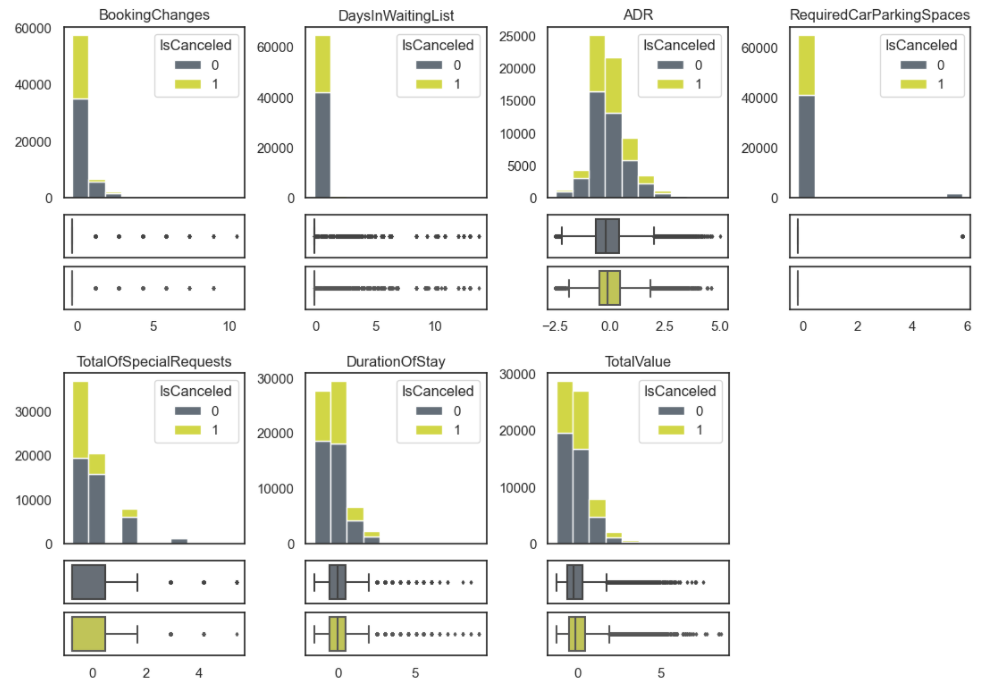
Descrição gerada automaticamente

Gráfico, Histograma

Descrição gerada automaticamente







Gráfico, Histograma

Descrição gerada automaticamenteGráfico, Gráfico de barras

Descrição gerada automaticamente

Interface gráfica do usuário, Aplicativo

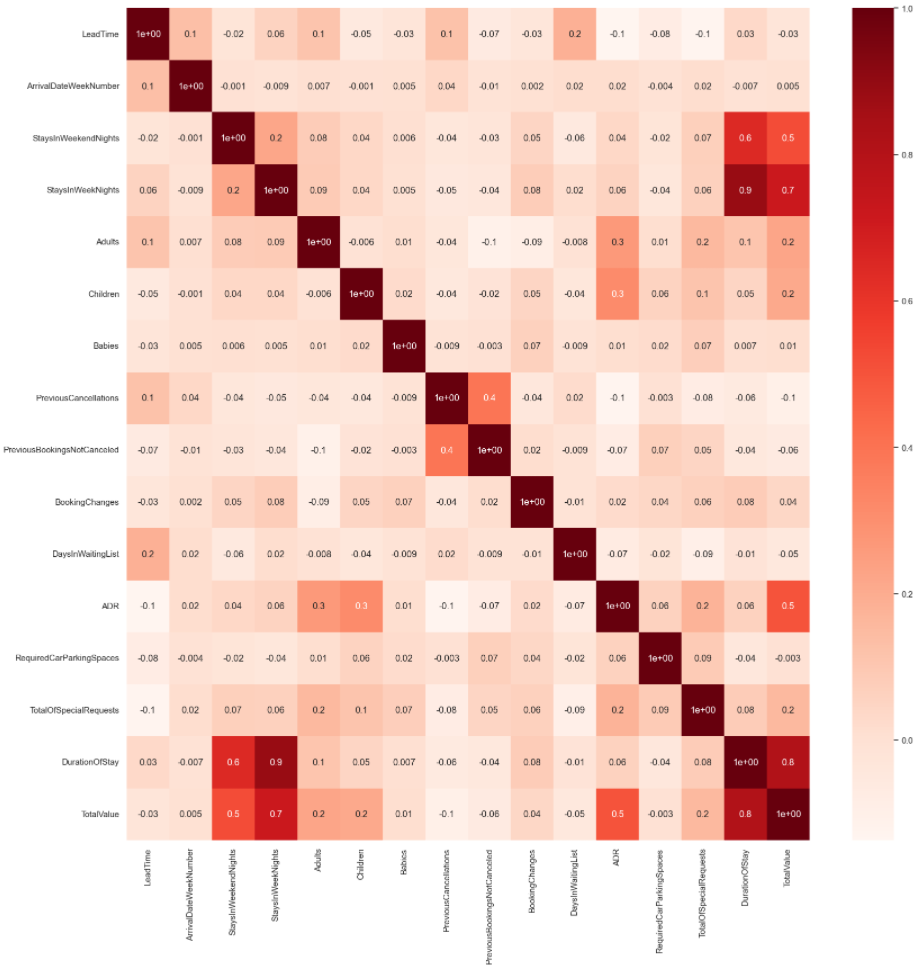
Descrição gerada automaticamente

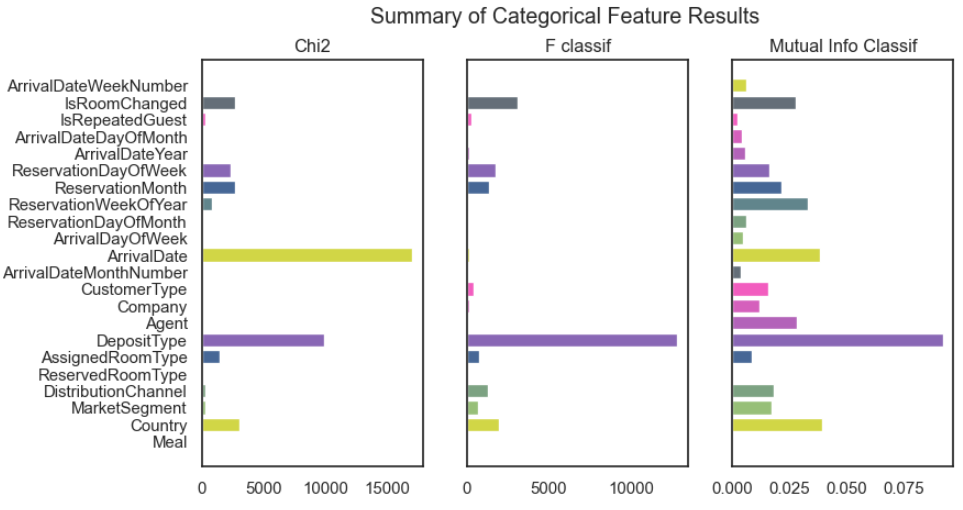
Interface gráfica do usuário, Gráfico

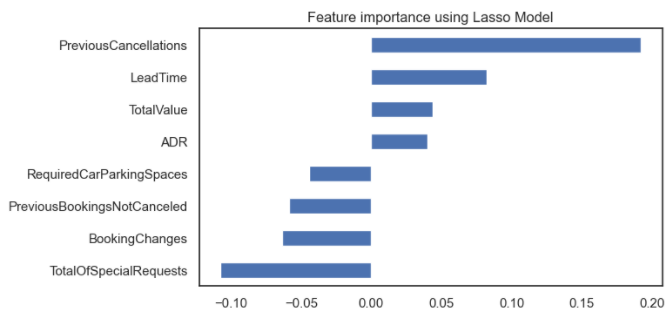
Descrição gerada automaticamente

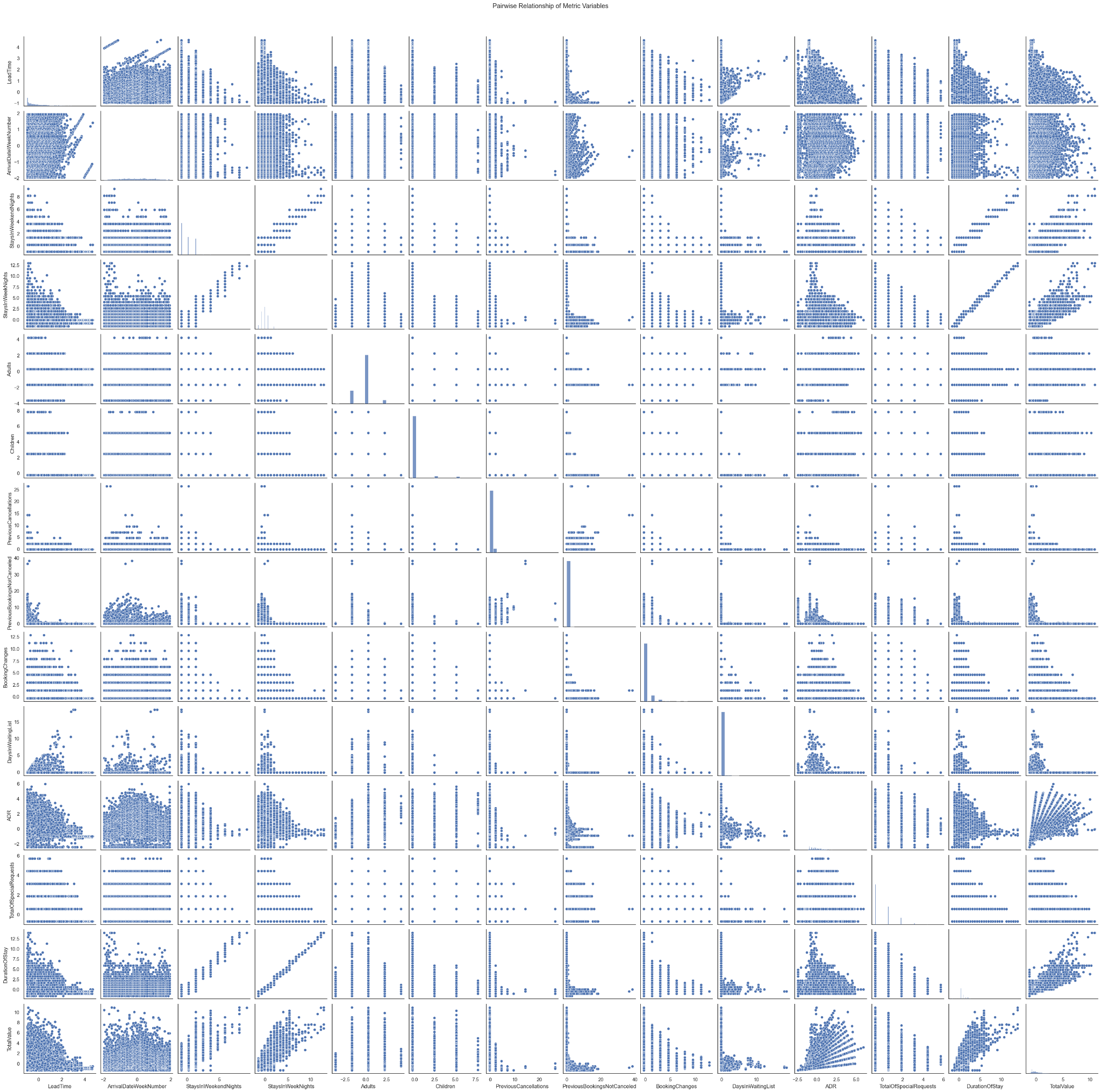
Gráfico, Gráfico de cascata

Descrição gerada automaticamente



****

****

****

**Random Forest Classification Report and Confusion Matrix**