Final project: Bookings

Gloria, Prada Moore

Bellevue University

FINAL PROJECT

Introduction

For hotels and resorts bookings is the daily bread, for those who are looking for a nice stay in the next vacations it is necessary to make plans, reservations, etc. How nice would be for hotels and resorts to know when a stay is going to be canceled, that way they can offer to some guests, the possibility to book after they already say that they are booked?. This analysis look to predict what bookings are going to be canceled.

Problem Statement:

I chose a data set available in Kaggle, the data set contains booking information for a city hotel and for a resort hotel and includes guest data and details like, day, month, year of arrival, number of children and babies, number of special requests made by the guests, length of the stay and some others. The data correspond to the years 2015 to 2017. Hotel booking cancellations always has been an issue specially in busy season, sometimes people want to have vacation in some hotels, and when we call, is not possible to make reservations because the hotel is already booked, but what happened when those booking that are made some days ahead gets canceled? The hotel loses potential clients and money because it might not find another guest to fill that canceled room

Question:

What reservations are going to be canceled?

FINAL PROJECT

Proposal:

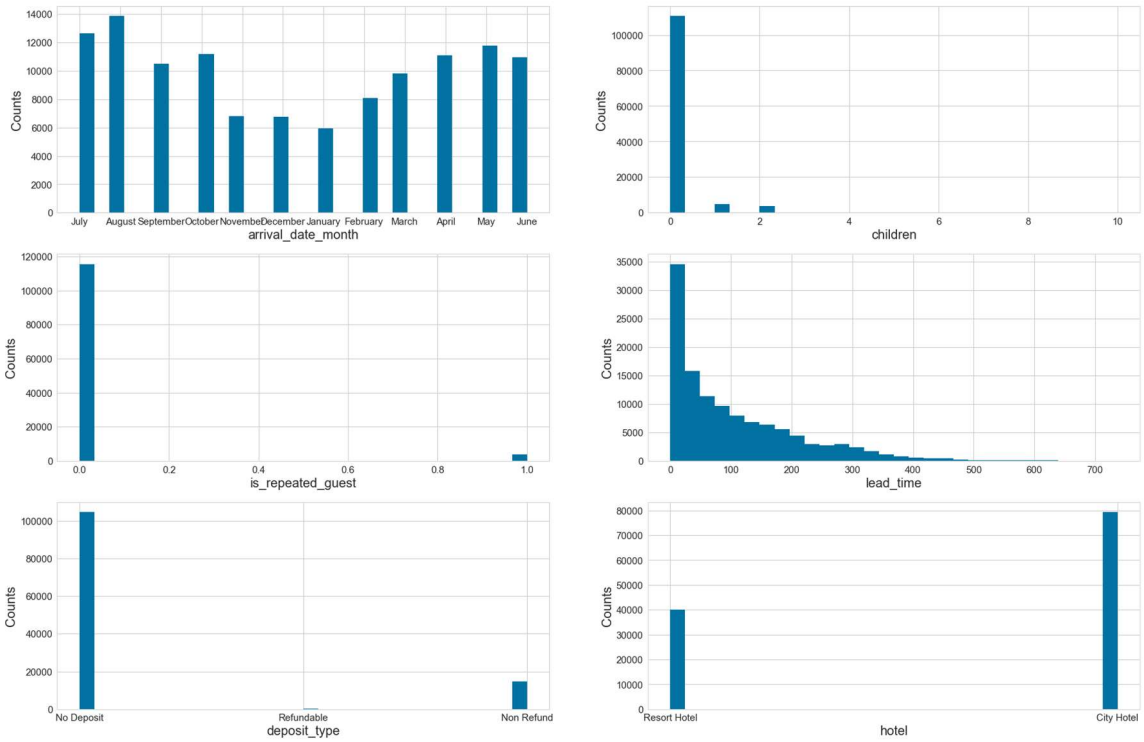
This case study will go through the features that are more suitable to predict if a reservation is going to be canceled or not, the data set have 32 variables, that I am going to describe, I will start to choose variables that I find interesting to analyze, I will build the histogram of each, bar chart and correlation values. I would like also to build pie charts for those variables with higher correlation values.

The features of interest that I chose to start are:

- 'arrival_date_month',
- 'children',
- 'is_repeated_guest'
- 'lead_time',
- 'deposit_type'
- 'hotel'

Graph Analysis

Histograms



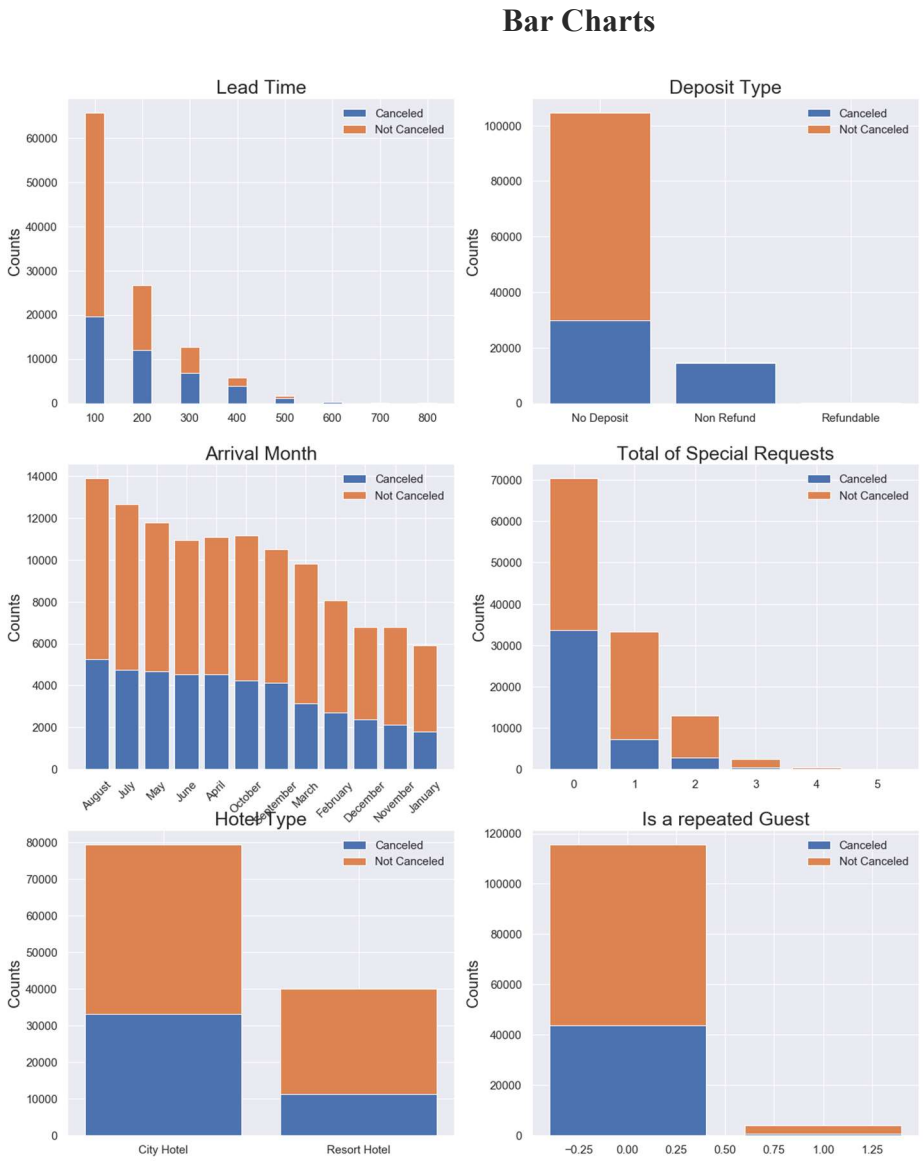
From the previous Histograms we can see the lead_time variable that represent the number of days that are between the date of entering the booking in the system and the arrival date, so basically, how many days in advance a guest makes a reservation, this histogram is skewed to the right, and present that most of the guest make reservations with less the 100 days previous to their arrival date. Also we can see that we have more reservations for city hotels than resort hotels

Most of the reservation does not require a deposit

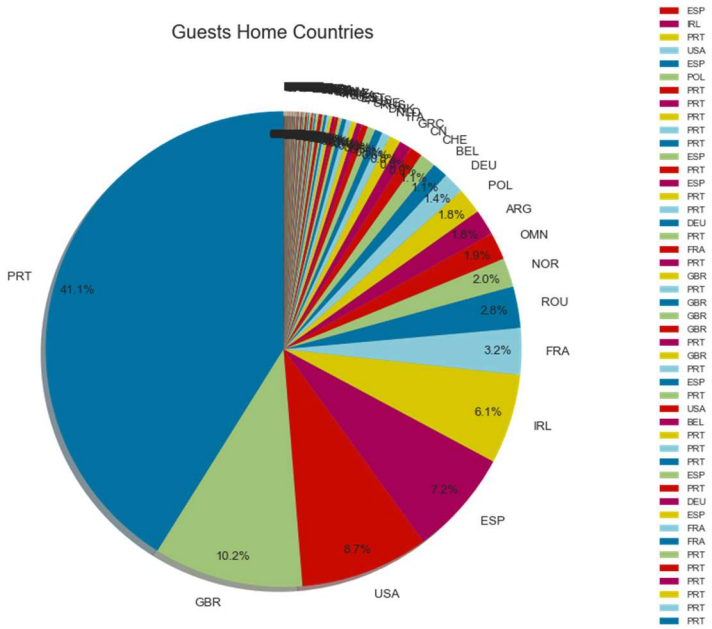
Most of the guest are new guests for those hotels

FINAL PROJECT

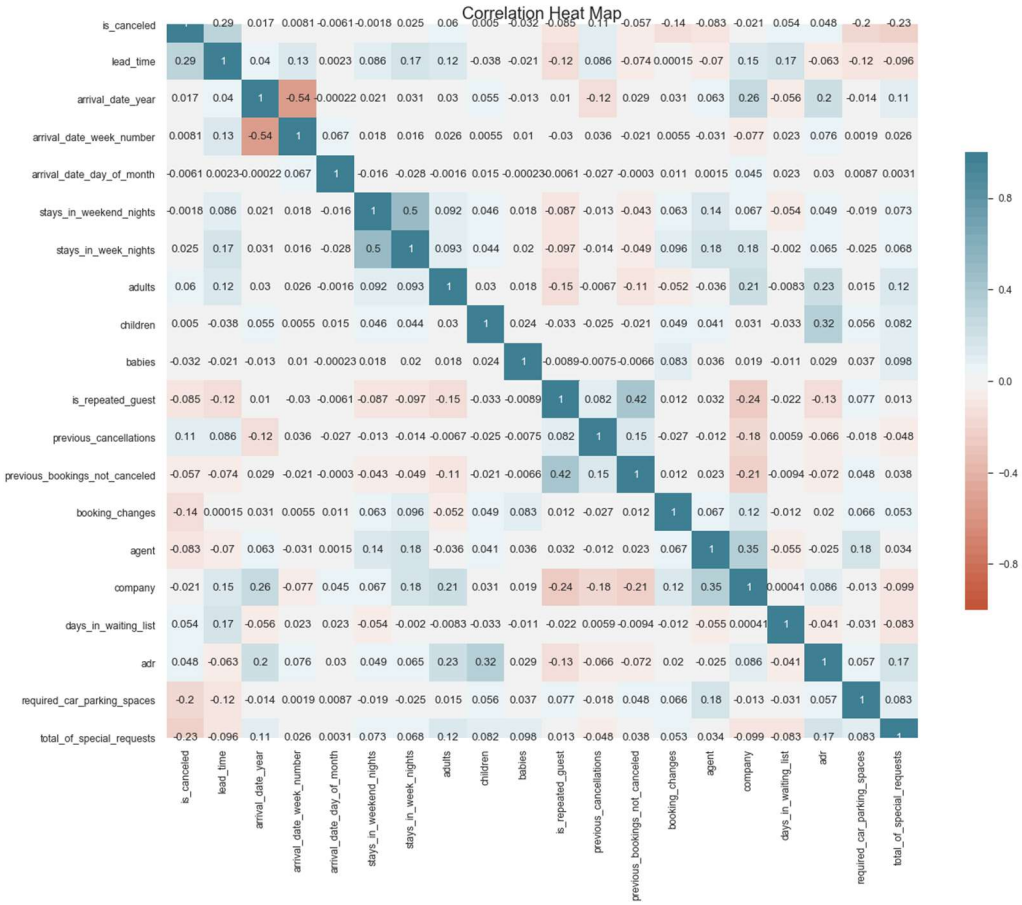
Most of the guest travel without children, and those who travel with children, most of them 1 or max 2. Finally, for we can see a higher number of bookings for the month of August, and have a "low season" in between November and February



FINAL PROJECT



Correlation Heat-Map



FINAL PROJECT

I used Pearson correlation to get the values for every single feature in the data frame. Checking that the higher positive value correlated with the cancelation rate, would be the lead time **0.29**, what means is that as higher the number of days prior to the arrival date that the booking is made, the higher the number of cancelations

On the other hand, we find the total of special request negative correlated with the cancelation rate **-0.23**, the higher the special request there are less cancelations

But in general the correlation values are low

A look of my Data

```
data.head()
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_i
0	Resort Hotel	0	342	2015	July	27	1	0	
1	Resort Hotel	0	737	2015	July	27	1	0	
2	Resort Hotel	0	7	2015	July	27	1	0	
3	Resort Hotel	0	13	2015	July	27	1	0	
4	Resort Hotel	0	14	2015	July	27	1	0	

5 rows × 32 columns

```
The dimension of the table is:  (119390, 32)
```

Next steps I will check the data and variable types and descriptions

After graph analysis, I filled missing values, create a logarithmic scale for the lead-time variable, create dummy variables for categorical features, and calculate correlations for each final feature

```
]: lead_time_log                    0.320063
   lead_time                        0.293123
   total_of_special_requests        0.234658
   required_car_parking_spaces      0.195498
   booking_changes                  0.144381
   previous_cancellations           0.110133
   is_repeated_guest                0.084793
   adults                           0.060017
   previous_bookings_not_canceled   0.057358
   days_in_waiting_list             0.054186
   adr                              0.047557
   babies                           0.032491
   stays_in_week_nights             0.024765
   company                          0.020642
   arrival_date_year                0.016660
   arrival_date_week_number         0.008148
   arrival_date_day_of_month        0.006130
   stays_in_weekend_nights          0.001791
   Name: is_canceled, dtype: float64
```

I chose to work with variables that have higher correlation value, I included some of the ones with high variance to do my feature reduction using first PCA and using LDA

Finally, after clean, select and organize my features I applied 3 models to my data to be able to check what model would perform the best.

- Logistic Regression
- K-Nearest Neighbor
- Random Forest Classifier

My numeric features were:

- 'lead_time_log'
- 'total_of_special_requests'
- 'previous_cancellations'
- 'adults'
- 'is_repeated_guest'
- 'days_in_waiting_list'

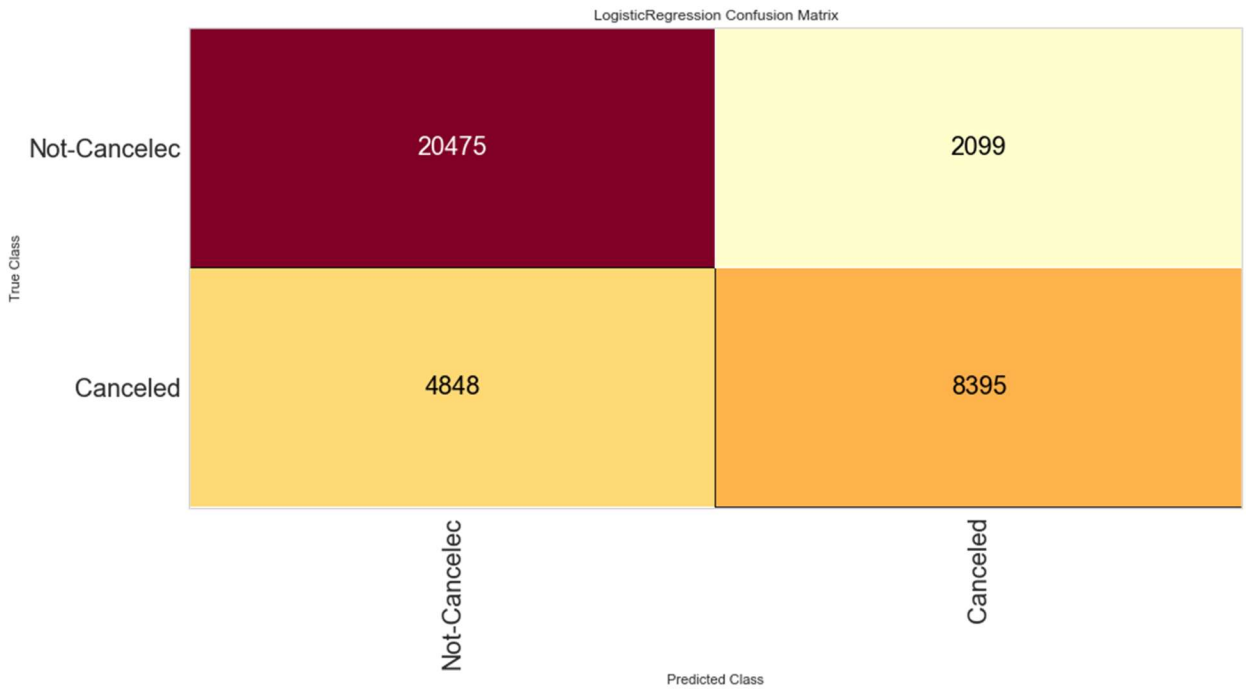
FINAL PROJECT

My categorical features:

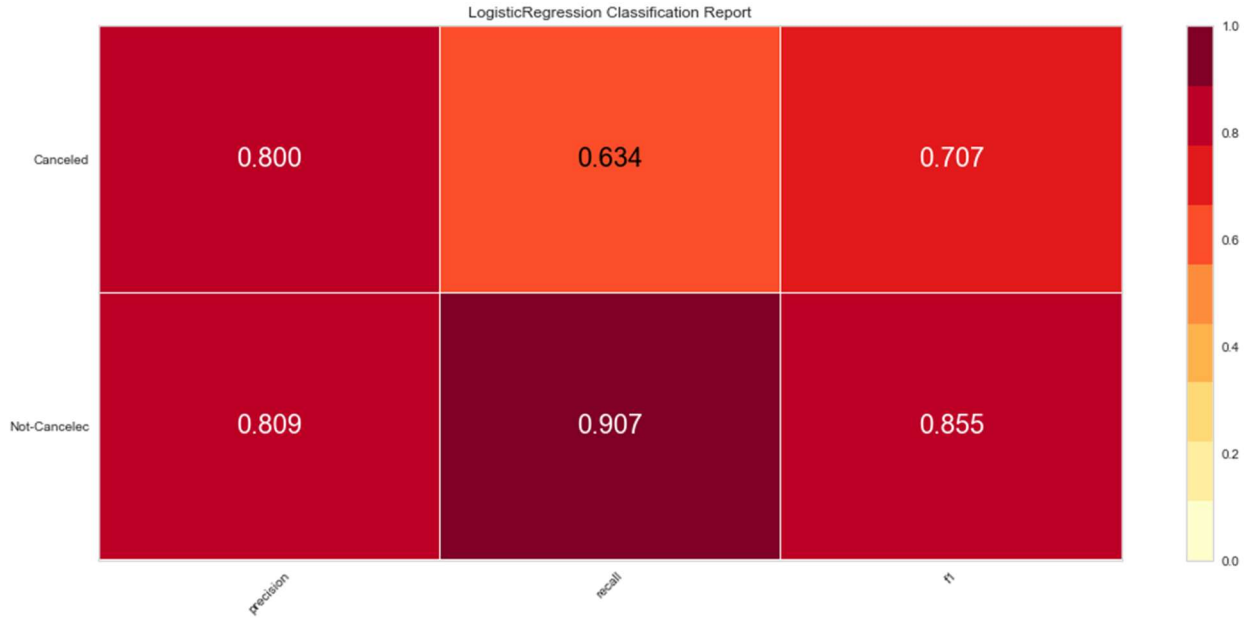
- 'hotel'
- 'arrival_date_month'
- 'country',
- 'deposit_type'

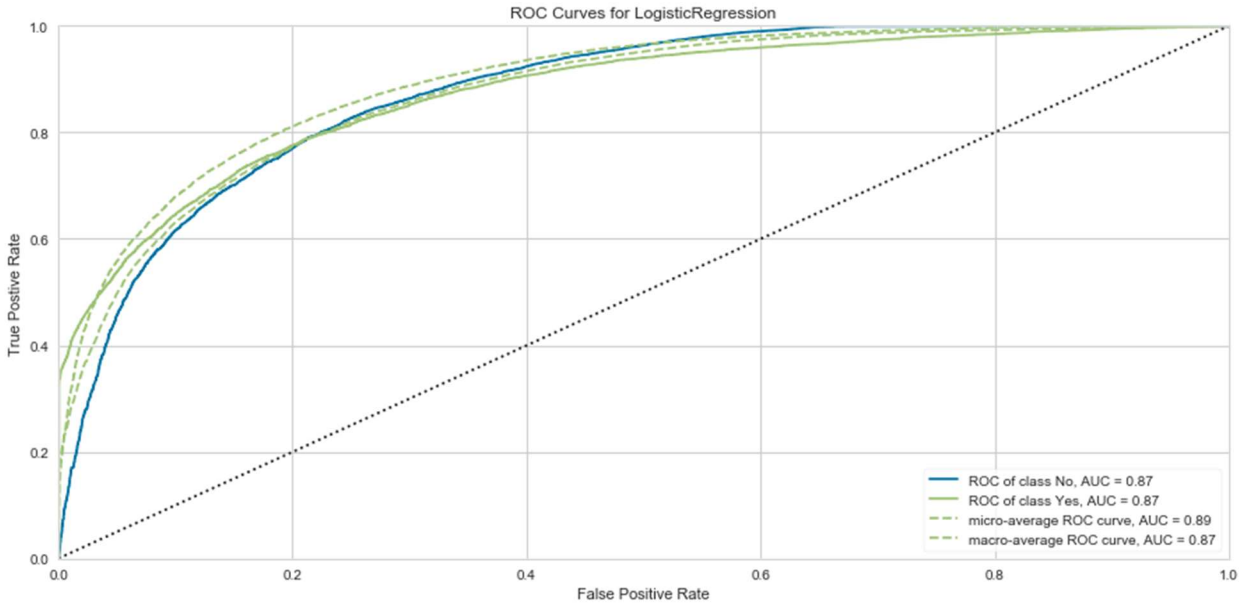
After convert my categorical variables in numbers and create my dummy feature data frame, I concatenate both

After this I Applied Logistic regression and these were my results:



Looking at these results, looks like models works better predicting not canceled bookings than they canceled ones, these last ones are the ones that I am interested to know about. I will check another model



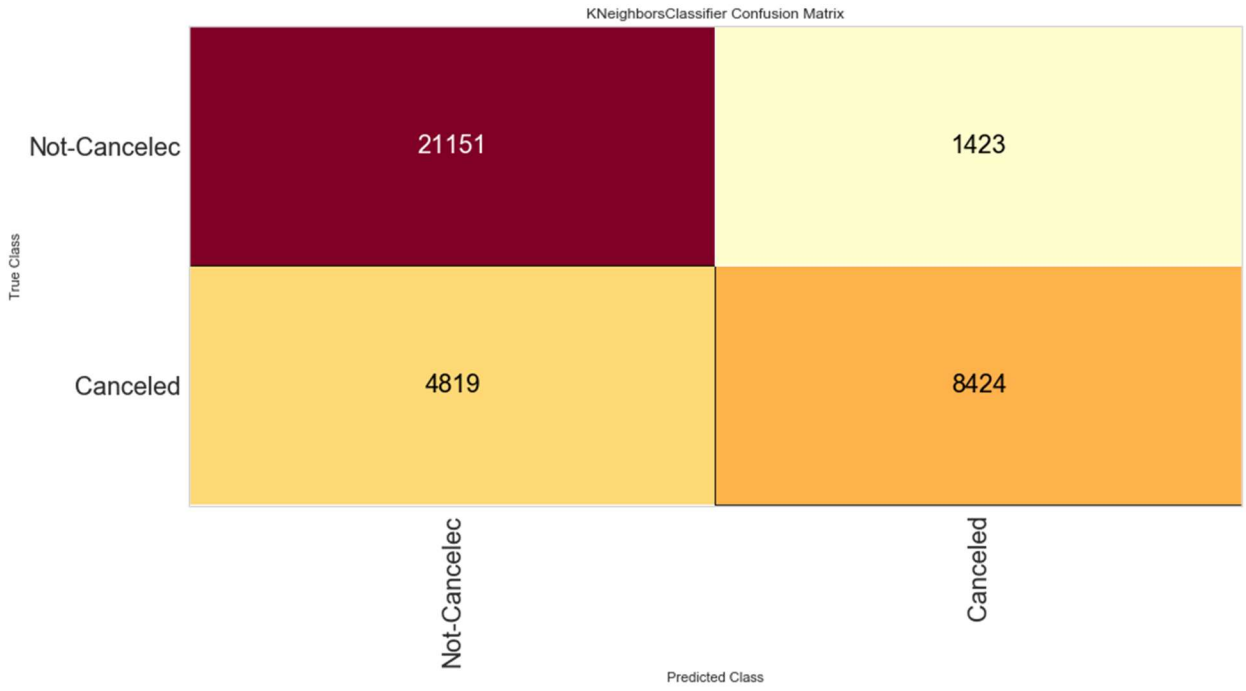


The recall for the canceled reservations is low in comparison with the precision and F1 metrics, what is related with the confusion matrix. The AUC is 0.87 that is not a bad number, is close to 1, so it is not a bad model however the high precision and low recall indicates that the first model it does a good job classifying the observations however about 40% of the canceled observation remains unidentified. https://filosophy.org/writing/visual-algorithms-precision-and-recall/

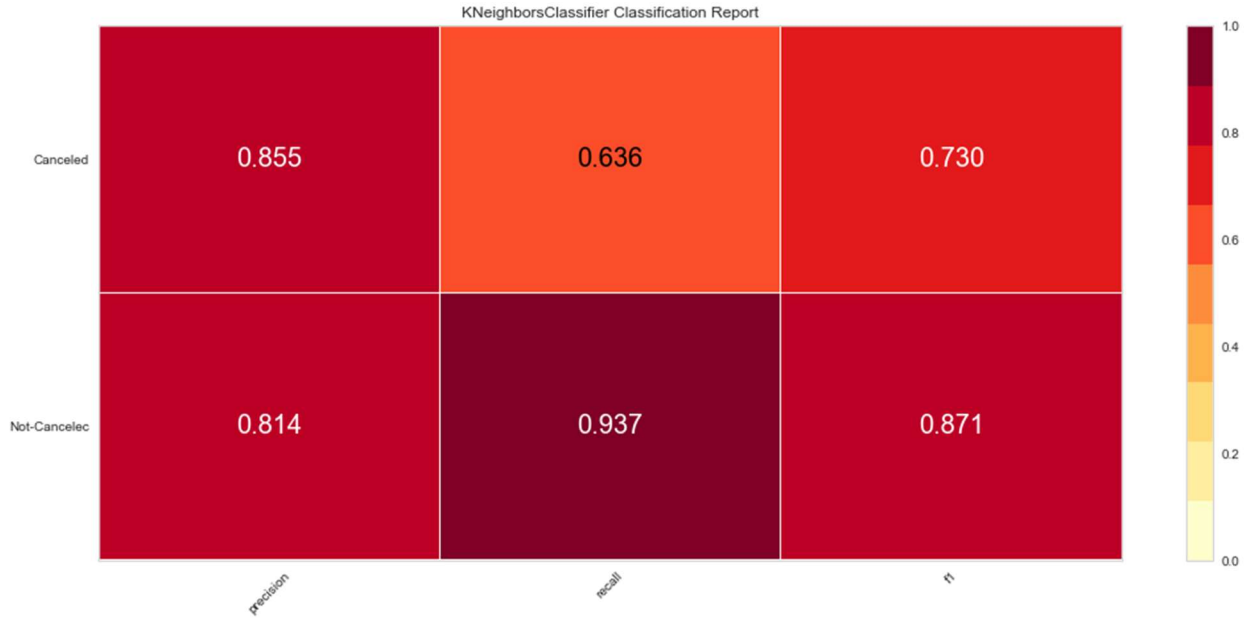
K-Nearest Neighbors

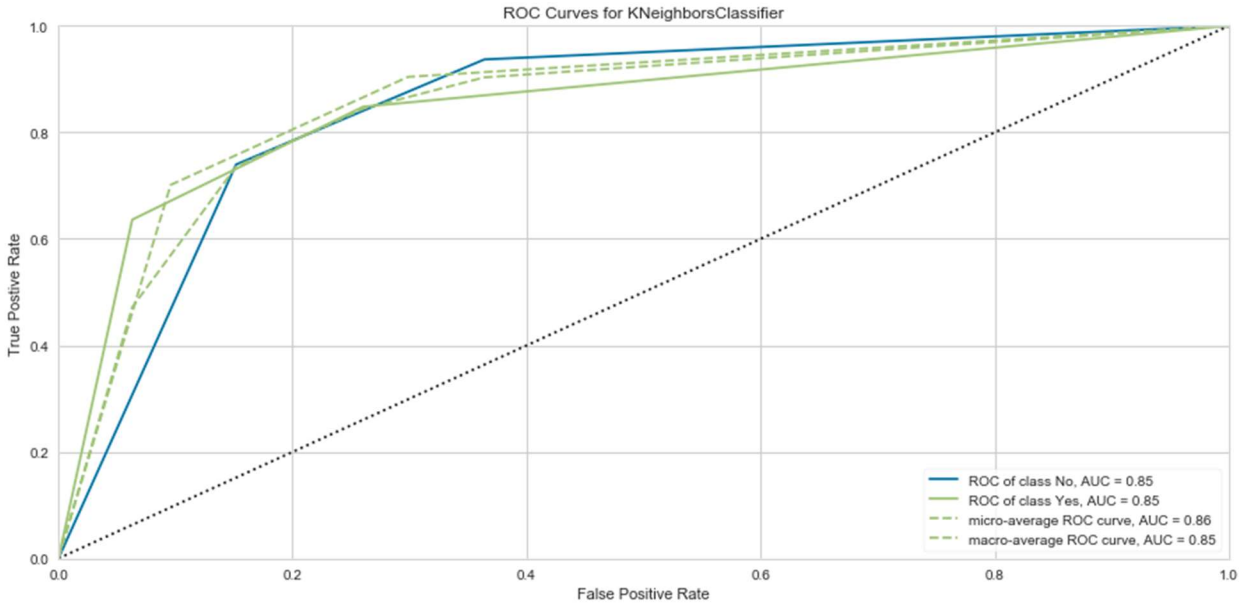
I used the K-nearest neighbor classifier in my problem, this algorithm will predict the observation to be in one class or the other one depening on the closer class of other obvservation. For this model I used our textbook (Albon. 2018. p254)

I have a number of new observations (My validation set) and want to know what observations will going to be canceled and which ones are not, I am going to use K-Nearest Neighbor Classifier, I define n_neighbors just 2, I just have 2 categories



This matrix looks better than the one for the Logistic regression, K-nearest neighbor perform better on the canceled category, it predicted more observations canceled that were actually canceled

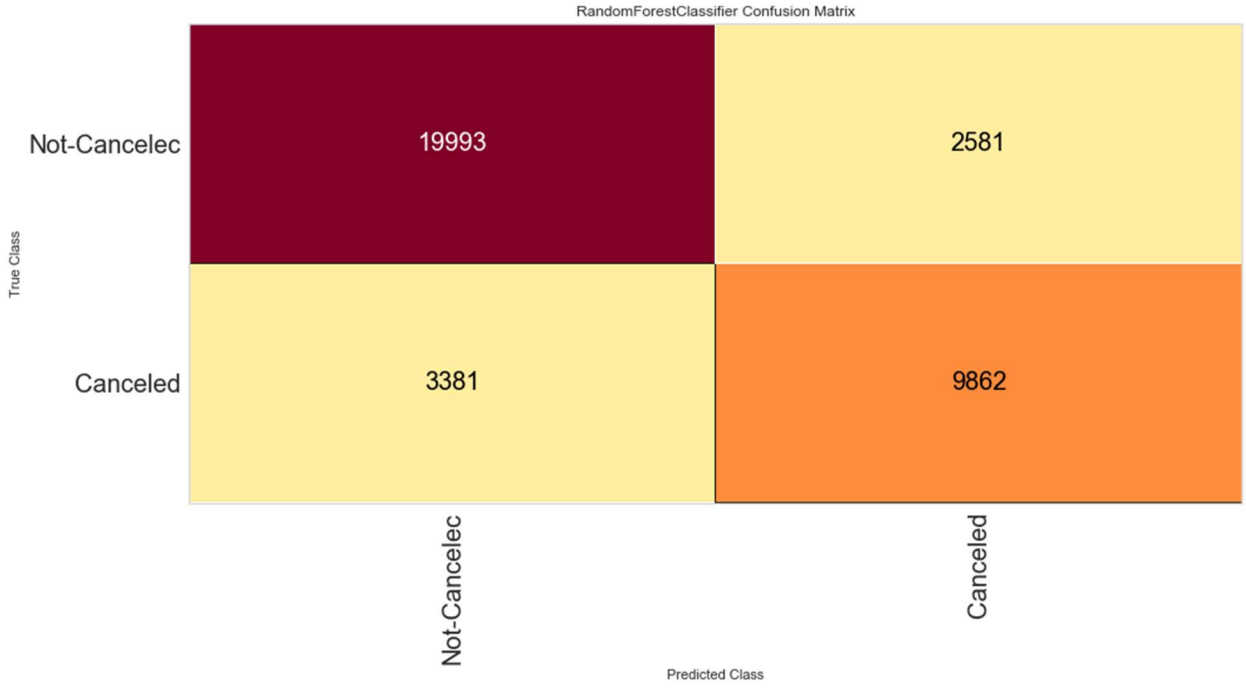


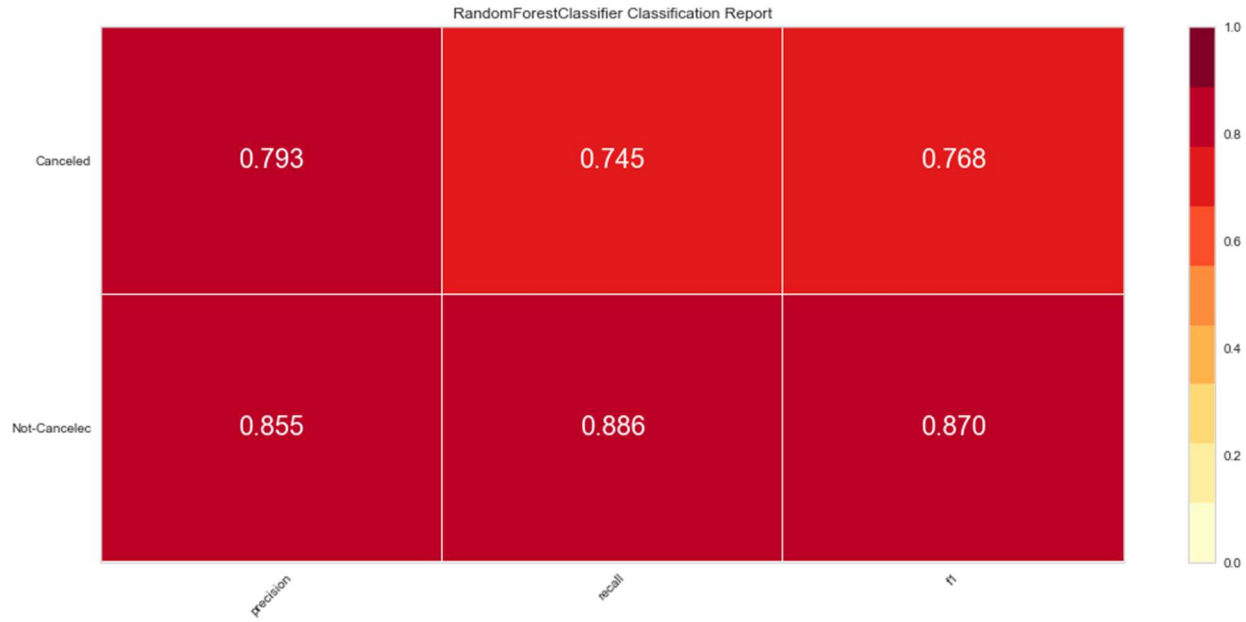


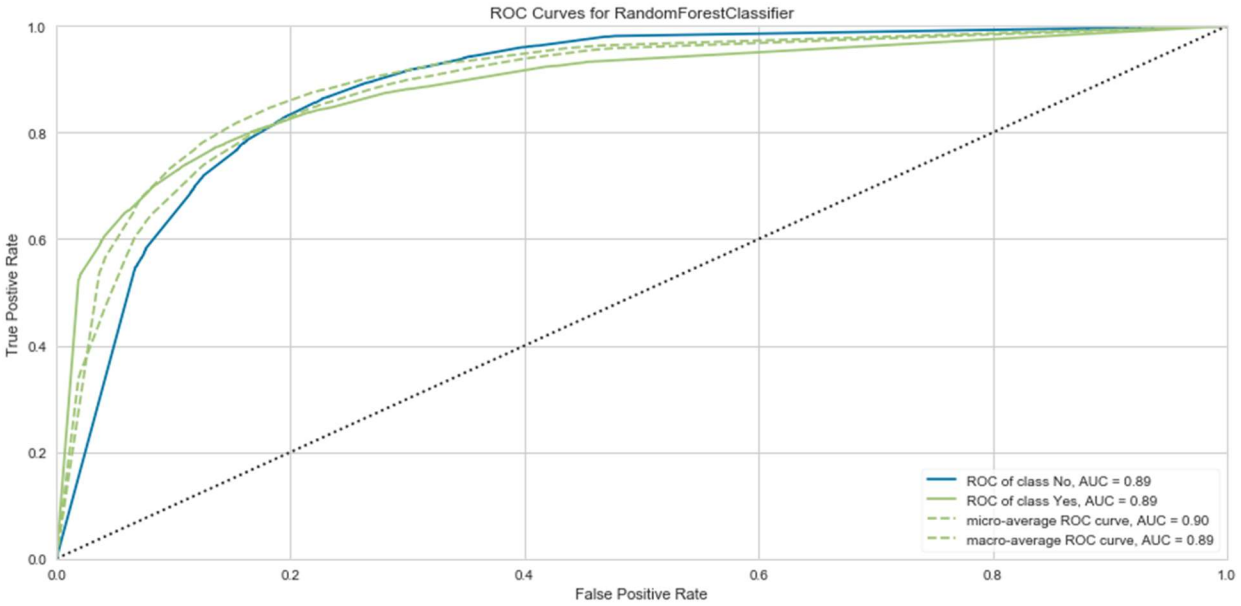
The results of this model look similar to the Logistic regression, this one present a lower AUC value 0.85. but the recall for canceled observations went up in about .4%

Random Forest Classifier

I'll use as third option a Random Forest Classifier, I am going to use this model because is another option for classification problem. I used our textbook (Albon. 2018 p 238)







In conclusion my model selection is the Random Forest model, this model presented better performance with better values of precision, recall and F1 for both classes and the AUC value of 0.89 what is higher than the models used before. Even that the Random forest present a lower precision, in general all metrics were better with Random Forest Classifier. Also the True Positive (9872) for Random Forest is better than the 2 model before

References

Trevino, A. (2016). Introudction to K-means clustering. Learn Data Science. Retrieve, September 24, 2020, from: https://blogs.oracle.com/datascience/introduction-to-k-means-clustering

DataRobot. (2018). H Unsupervised Machine Learning. Retrieved, September 24, 2020, from https://www.datarobot.com/wiki/unsupervised-machine-learning//

Bengfort, B. Bilbro, R and Ojeda, T. (2018). Applied Text Analysis with Python. . O'Reilly, Sabastopol, CA

https://matplotlib.org/3.1.3/gallery/statistics/hist.html#generate-data-and-plot-a-simple-histogram