PREDICTING HOTEL BOOKING CANCELLATIONS

AUTHOR: RAZVAN NELEPCU

MENTOR: DHIRAJ KHANNA

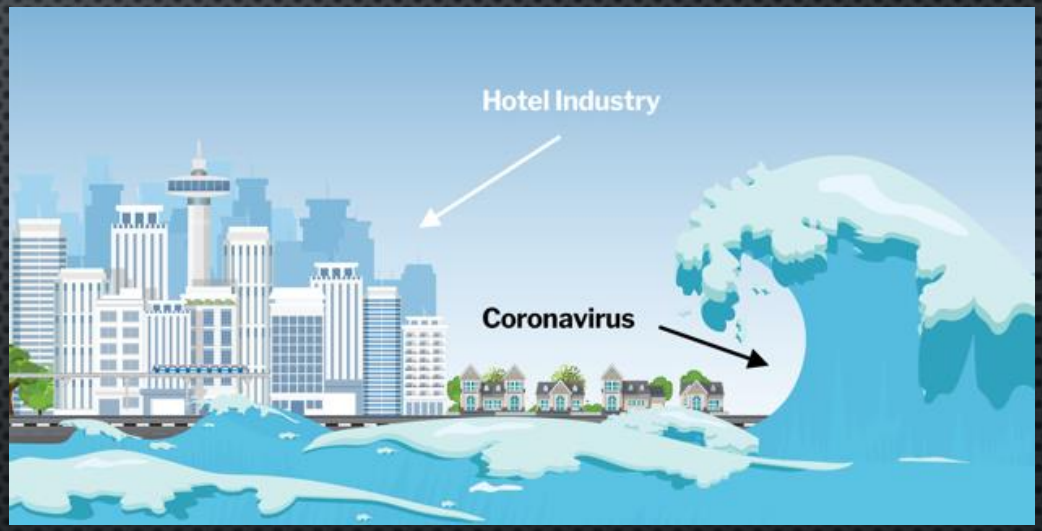
Project Organization

The project follows all steps of Data Science workflow divided in the following major steps:

- Problem Identification
- Data wrangling
- Exploratory Data Analysis
- Pre-processing and training data development
- Modeling
- Documentation

CONTEXT AND PROBLEM STATEMENT

We are at a point where nobody can say that their lives were not affected by the Coronavirus pandemic. Even if it was having someone close getting sick or even die, losing the job or having to change the working style completely to working from home, or just the fact that you had to see the friends and relatives over video.



The businesses were also heavily impacted by the pandemic. Bars and restaurants, airlines and hotels were the businesses that took the hardest hit.

With pre-pandemic cancelation rate around 40% and considering current situation, it became crucial for hotels to identify the customers who might cancel their reservation and establish a Business strategy to have them keep their reservation. For their survival in a pandemic and post-pandemic world, hotels must adapt and use all tools available to maximize profits.

For these reasons we considered working on a Project that uses hotels booking data with the goal to apply the Data Science Methods in order to predict future bookings cancelations, respectively the cancelations likelihood for new bookings.

DATA WRANGLING

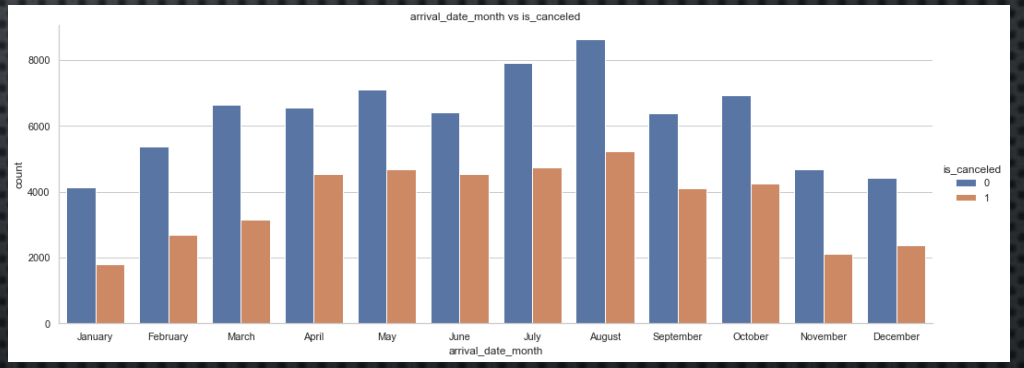
The data was downloaded from ScienceDirect
It contains more than 100,000 booking records from 2 hotels in Portugal, a hotel in Lisbon and a resort hotel in Algarve. The period covered was between 1st of July of 2015 and the 31st of August 2017.

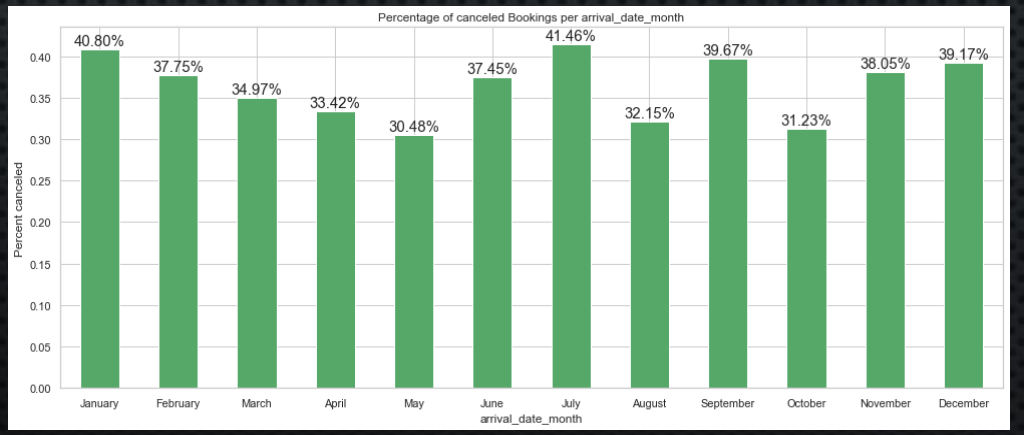
The dataset had a total of 32 features containing information such as: arrival date, number of adults, children or babies, if they are previous guests, how many prior cancellations they had and other.

Before proceeding with data cleaning and dealing with missing values and outliers, we considered necessary to have a deeper look into the data. This was done separately for numerical and for categorical features.

Some key findings after doing the initial feature exploration were:

- The average cancelation rate was 37.04%
- 99% of people with non-refundable deposits canceled their booking
- From all outliers only 1 was considered a wrong entry and removed





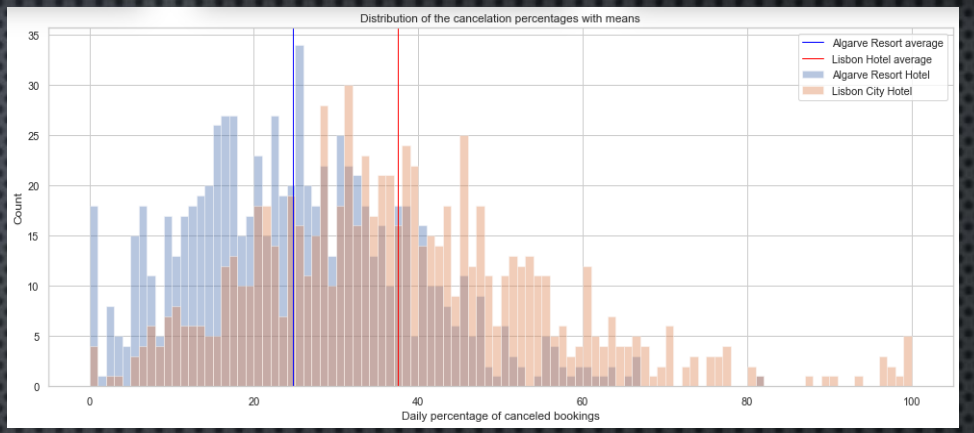
HYPOTHESIS TESTING

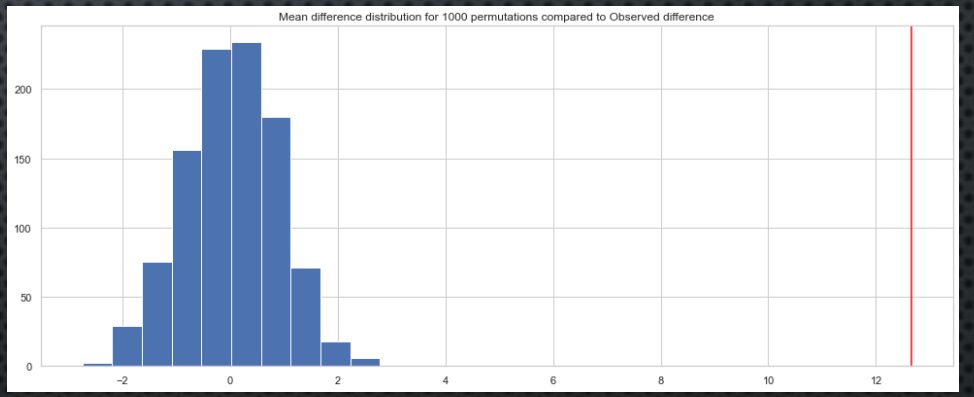
We conducted a statistical hypothesis testing to check if the 2 hotels' cancellations were significantly different.

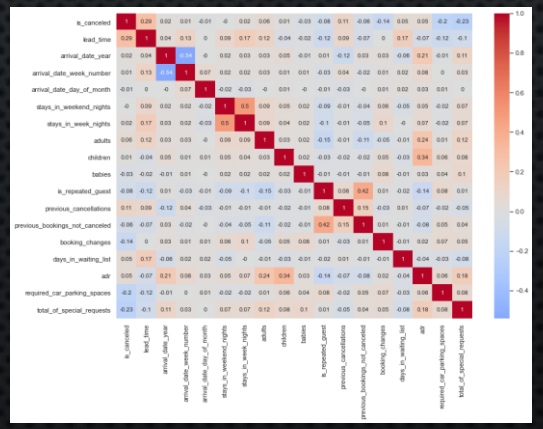
We visualized the Heatmap to visualize all correlations from our numerical features

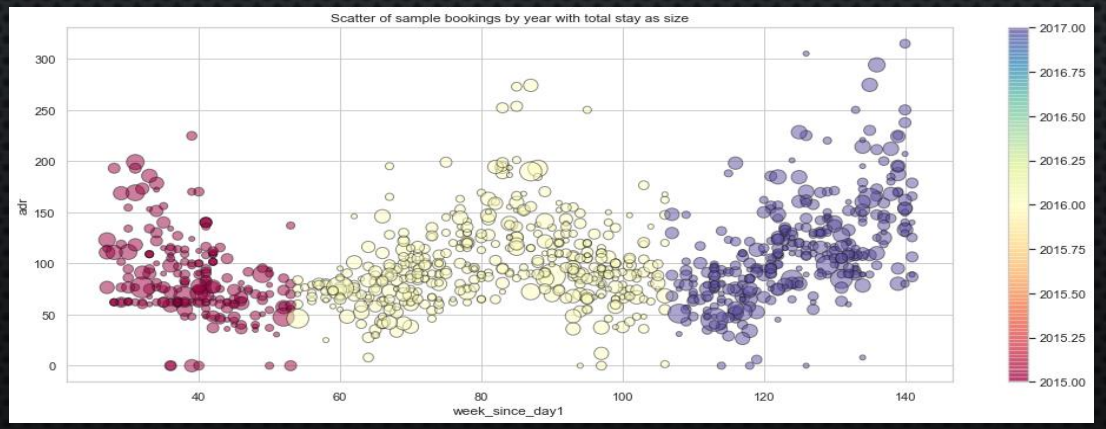
Using the correlations discovered we continued with bivariate EDA and multivariate EDA.

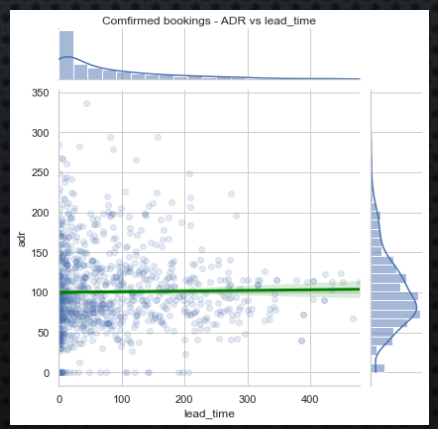
Lastly, we explored our data to find an answer to the question: *Why 99% of customers with non-refundable deposits are cancelling?*

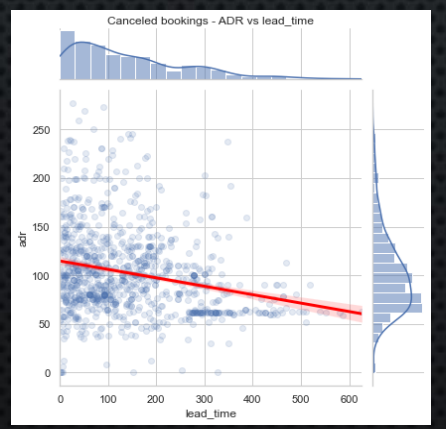












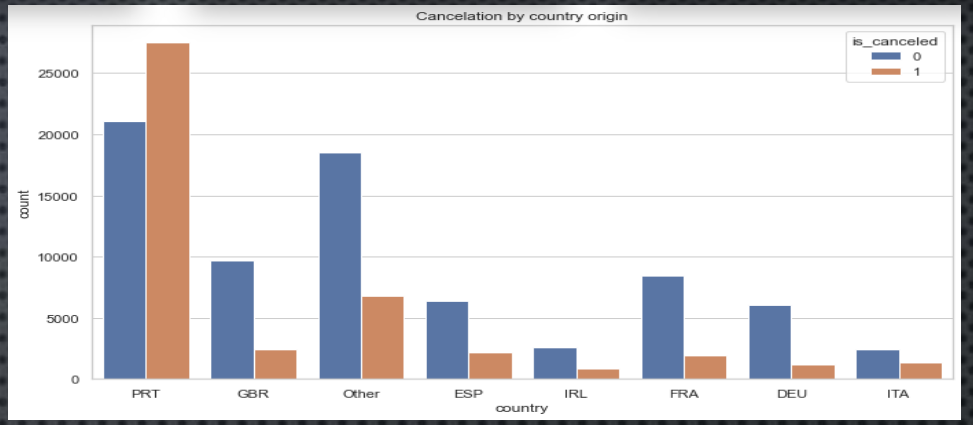
PRE-PROCESSING

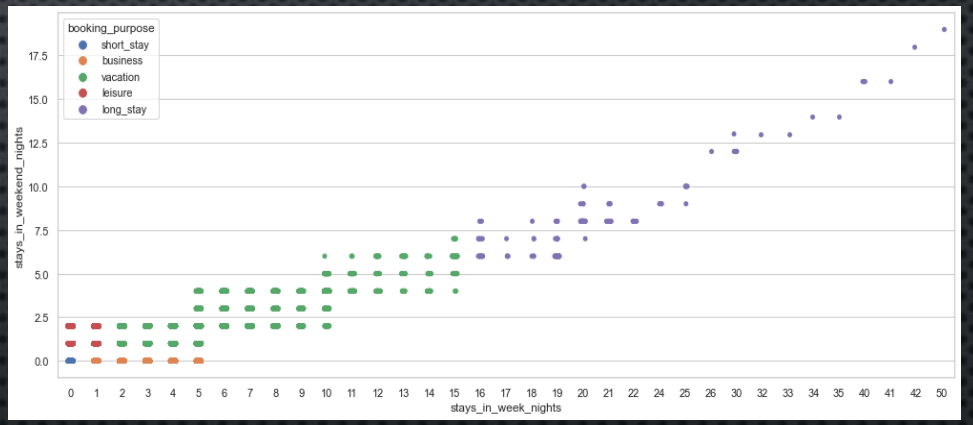
The pre-processing had 3 major parts:

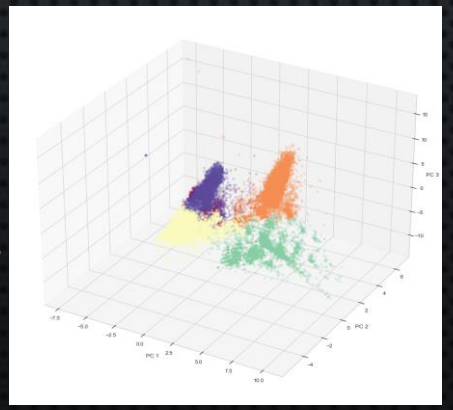
- **Data Leakage analysis.** We checked with our dataset authors to make sure there was no data leakage. Some features what might have constituted data leakage were analyzed, and we considered just the country origin to be source of data leakage and removed it.

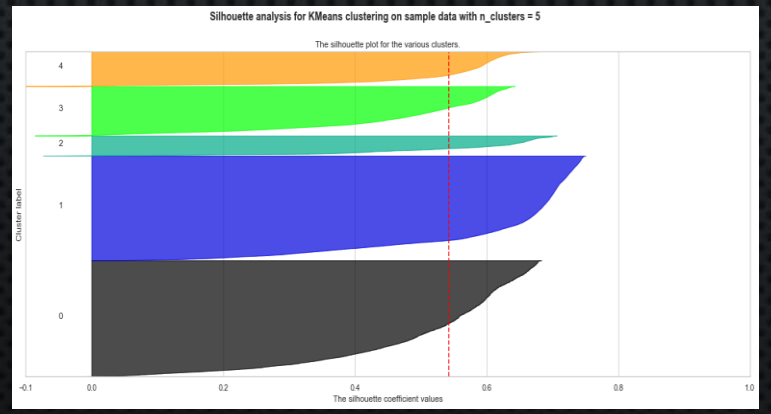
- **Feature engineering**. Using available data and domain knowledge we created 3 more features:
  1. arrival_day_of_week
  2. booking_purpose
  3. received_different_room

- **Customer segmentation using Unsupervised Learning** (Kmeans). We used elbow method and silhouette scores and plots to find the optimal number of cluster: 5.









MODELING

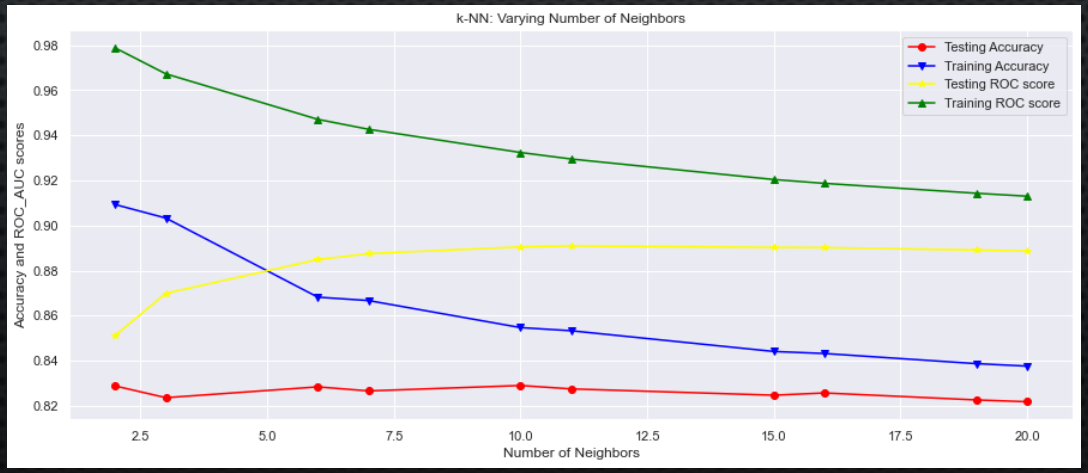
Metrics used: Accuracy and ROC's area under the curve for ranking and additional confusion matrix and classification report from visualization. We used a DataFrame to save the metrics.

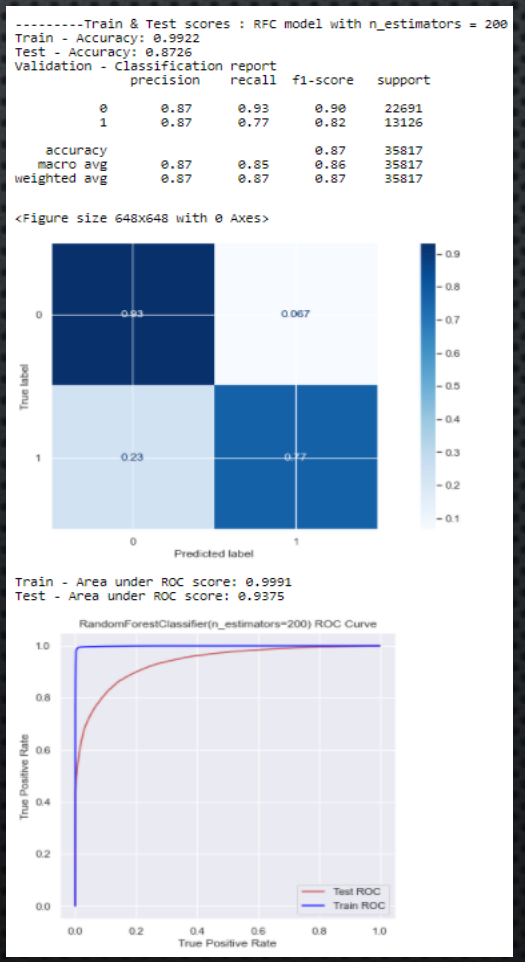
The models used were: LogisticRegression, KNN, Random Forest Tree and Catboost.

Other Tools used:

- Hyperparameter Tuning Using GridsearchCV and RandomSearchCV
- Regularization using Lasso or Ridge
- Feature Decomposition using PCA
- Feature Selection using permutation_importance and feature_importances_

Best model – RFT with n_estimators = 200
Accuracy: 0.8726
roc_auc score: 0.9375





RECOMMENDATIONS
MODEL IMPLEMENTATION AND USABILITY

One aspect had under consideration was to have the prediction as fast as possible.

Identified 2 possible models that can be deployed:

- *Front Desk Cancellations Percentage model.*
- *Monthly Cancellations Assessment.*

Manually created a new random booking and calculated the prediction.

The predicted probability is 0.52, so a 52% cancellation percentage.

new_booking

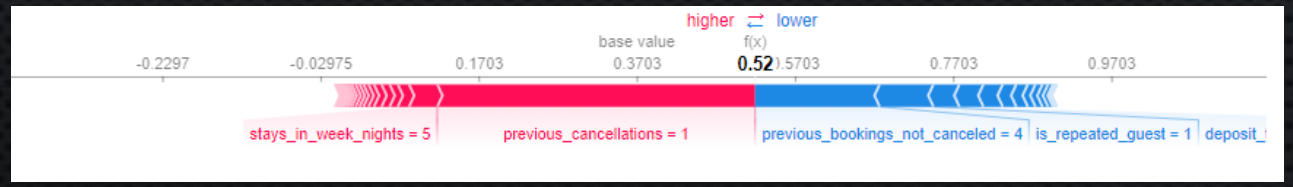
```
{'hotel': 'Algarve Resort Hotel',
 'arrival_date_year': 2019,
 'arrival_date_month': 7,
 'arrival_date_day_of_month': 12,
 'stays_in_weekend_nights': 4,
 'stays_in_week_nights': 5,
 'adults': 2,
 'children': 1,
 'babies': 0,
 'meal': 'BB',
 'market_segment': 'Direct',
 'distribution_channel': 'TA/TO',
 'is_repeated_guest': 1,
 'previous_cancellations': 1,
 'previous_bookings_not_canceled': 4,
 'reserved_room_type': 'A',
 'deposit_type': 'No Deposit',
 'agent': '250',
 'company': 'no_company',
 'arrival_day_of_week': 4,
 'arrival_date_week_number': 28,
 'booking_purpose': 'vacation'}
```

```
booking_prediction = best_model.predict(booking_w_dummy)
print('Customer is likely to cancel the booking' if booking_prediction[0]==1 else 'Customer is likely to keep the booking')
```

```
Customer is likely to cancel the booking
```

```
y_pred_prob = best_model.predict_proba(booking_w_dummy)[:,1]
print('Customer is {} % likely to cancel the booking'.format(round(y_pred_prob[0]*100,4)))
```

```
Customer is 52.0 % likely to cancel the booking
```



RECOMMENDATIONS
BUSINESS COURSE OF ACTION USING MODEL