Hotel Booking Cancellation Prediction

- Overview :

Problem statement

Goal

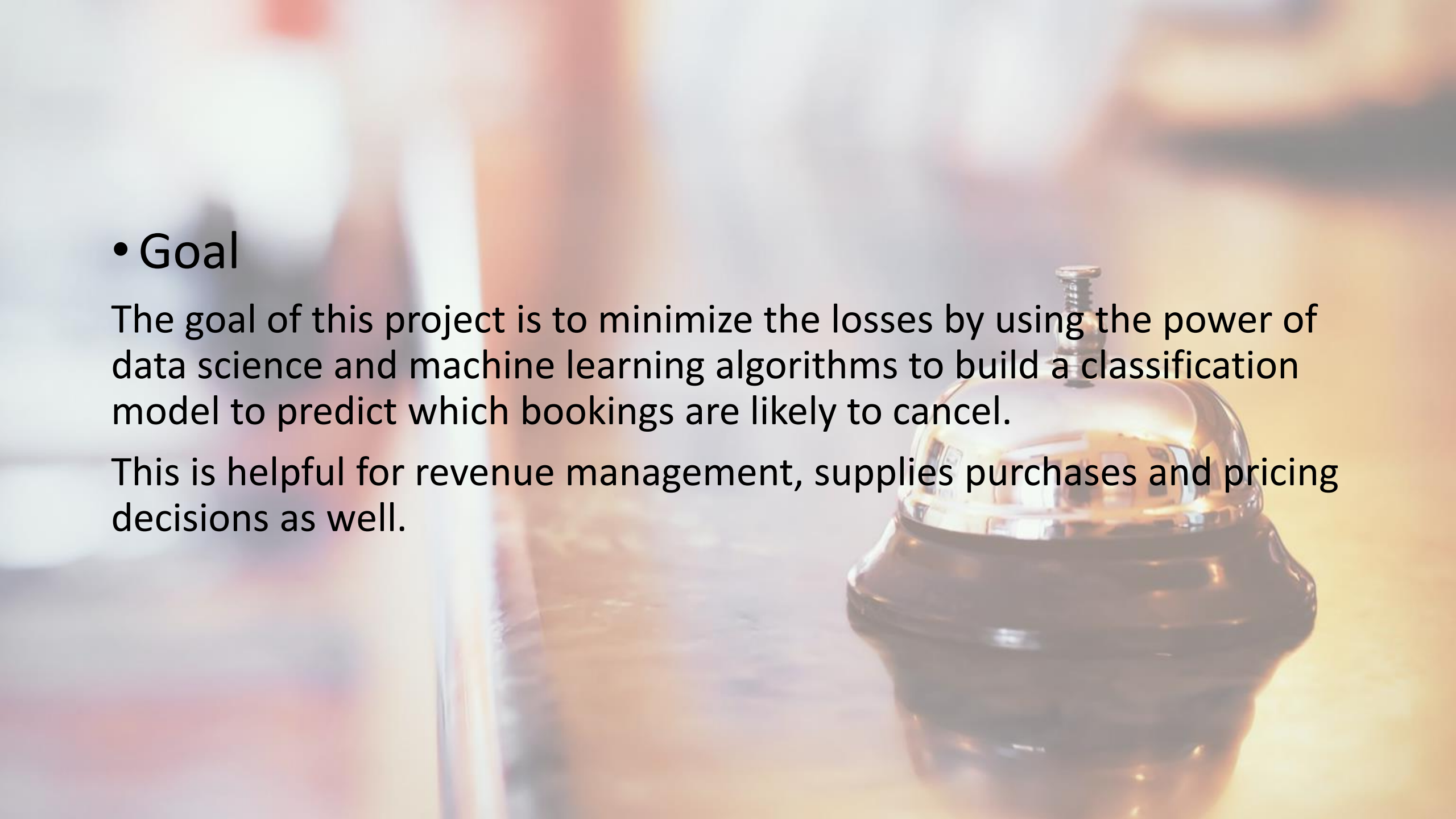
Data description

Models Implementation

- Problem Statement

Booking cancellations have substantial impact on hotel management systems. It affects the services, decrease revenue, and limits the production of accurate booking forecasts.

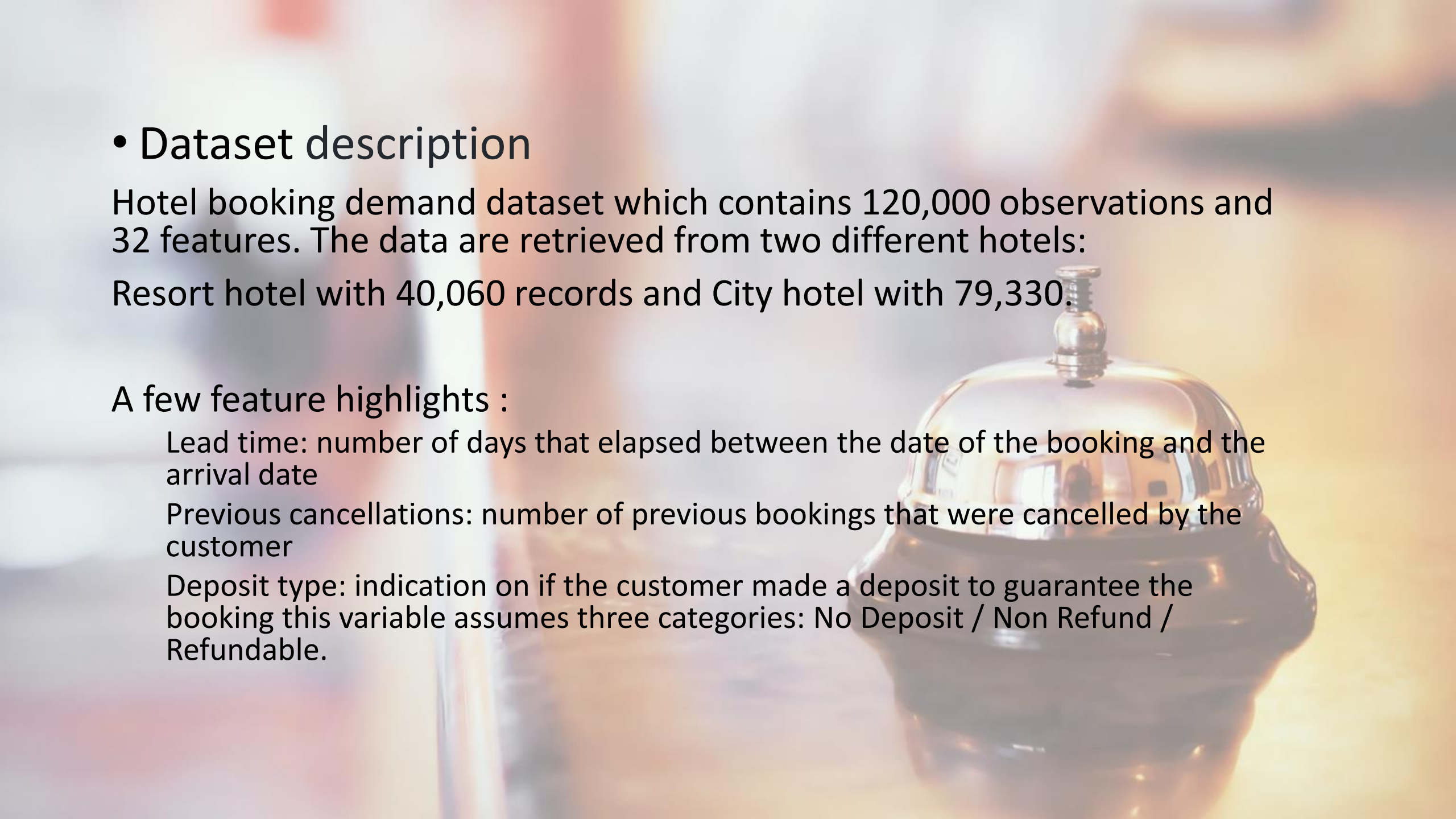
To overcome the effect of cancellations, some hotels use strict cancellations policies and define overbooking tactics. However, these regulations have a negative impact on number of bookings and on hotel reputation which in turn leads to revenue loss.



- Goal

The goal of this project is to minimize the losses by using the power of data science and machine learning algorithms to build a classification model to predict which bookings are likely to cancel.

This is helpful for revenue management, supplies purchases and pricing decisions as well.



- Dataset description

Hotel booking demand dataset which contains 120,000 observations and 32 features. The data are retrieved from two different hotels:

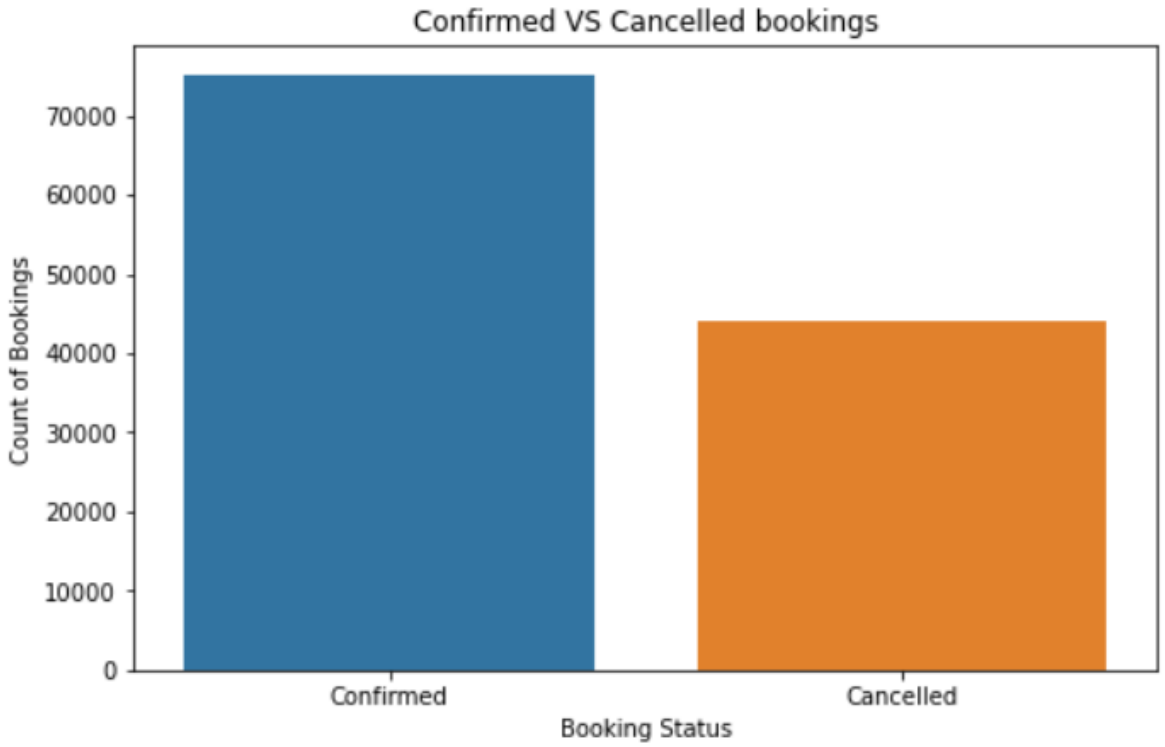
Resort hotel with 40,060 records and City hotel with 79,330.

A few feature highlights :

Lead time: number of days that elapsed between the date of the booking and the arrival date

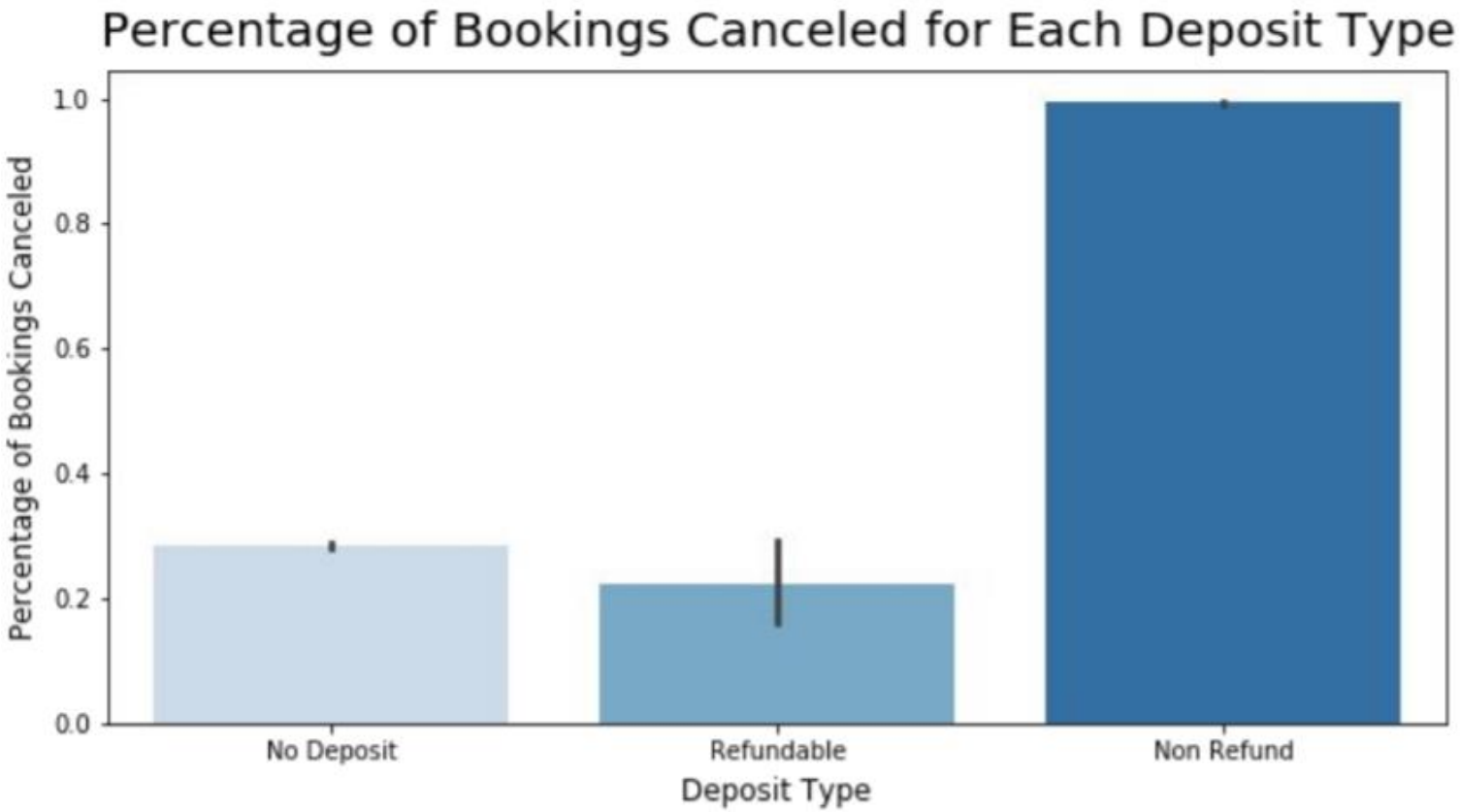
Previous cancellations: number of previous bookings that were cancelled by the customer

Deposit type: indication on if the customer made a deposit to guarantee the booking this variable assumes three categories: No Deposit / Non Refund / Refundable.

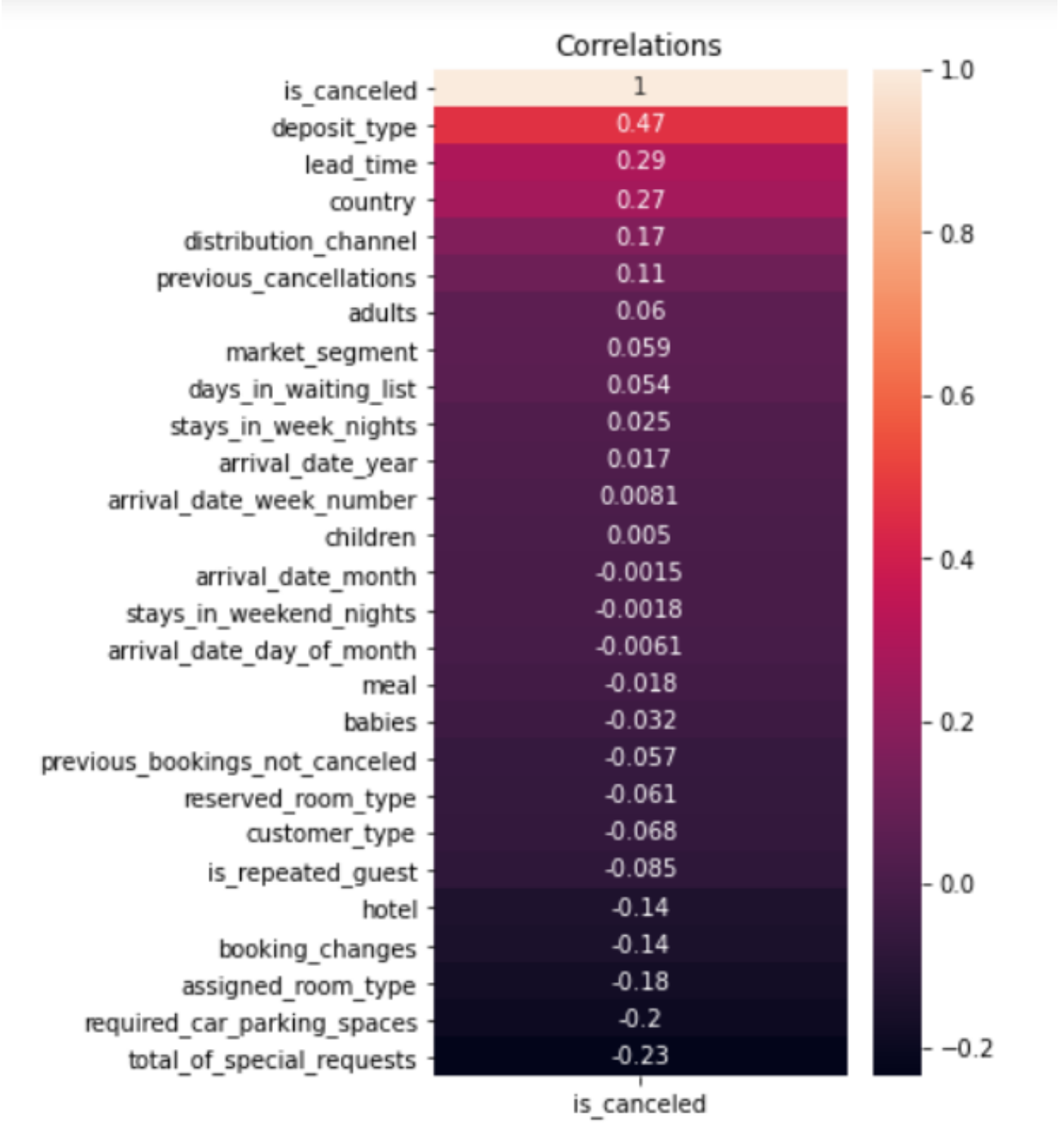


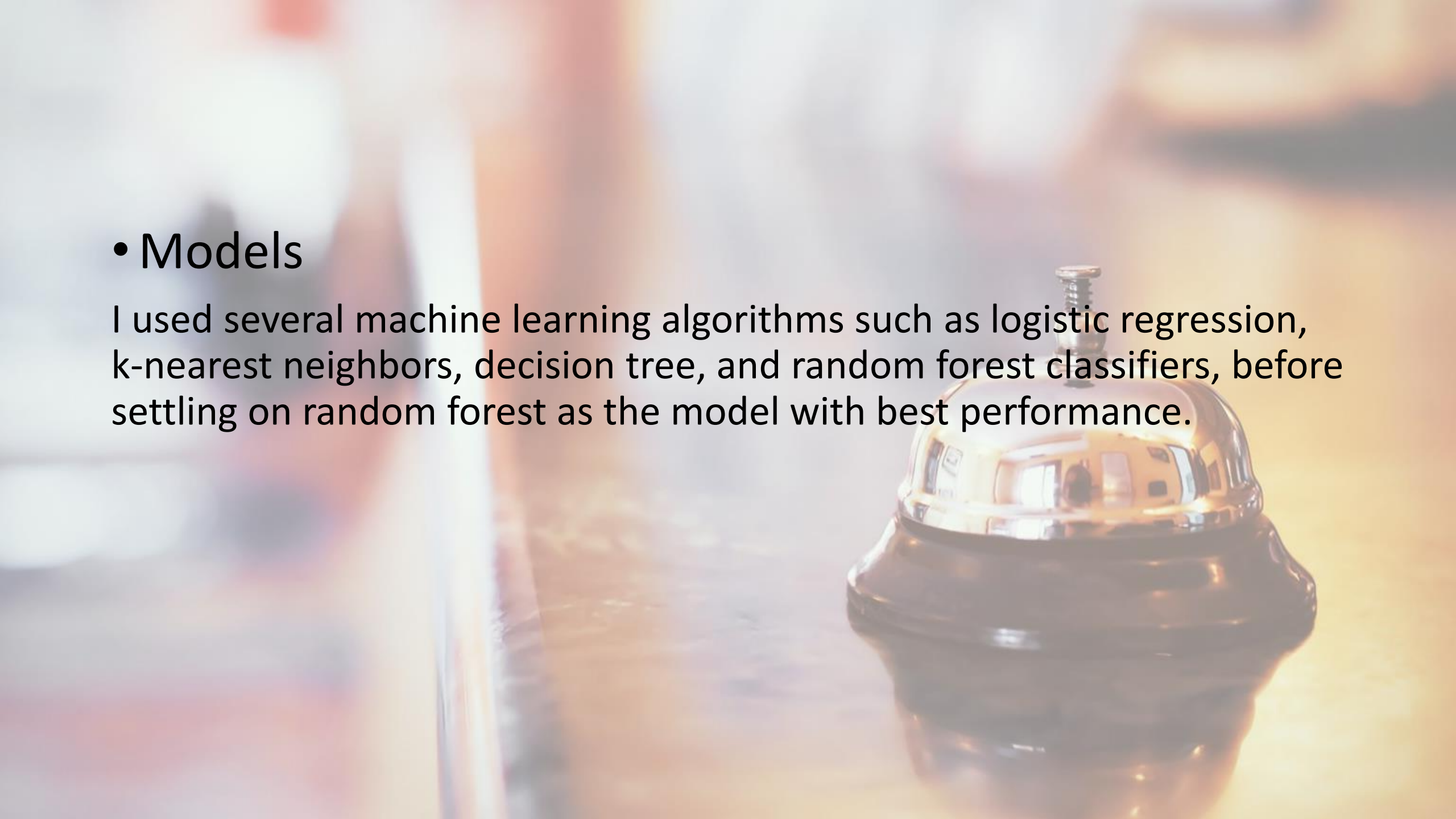
Almost 40% of bookings were canceled

Non-refundable deposits have much higher percentage of cancellation !



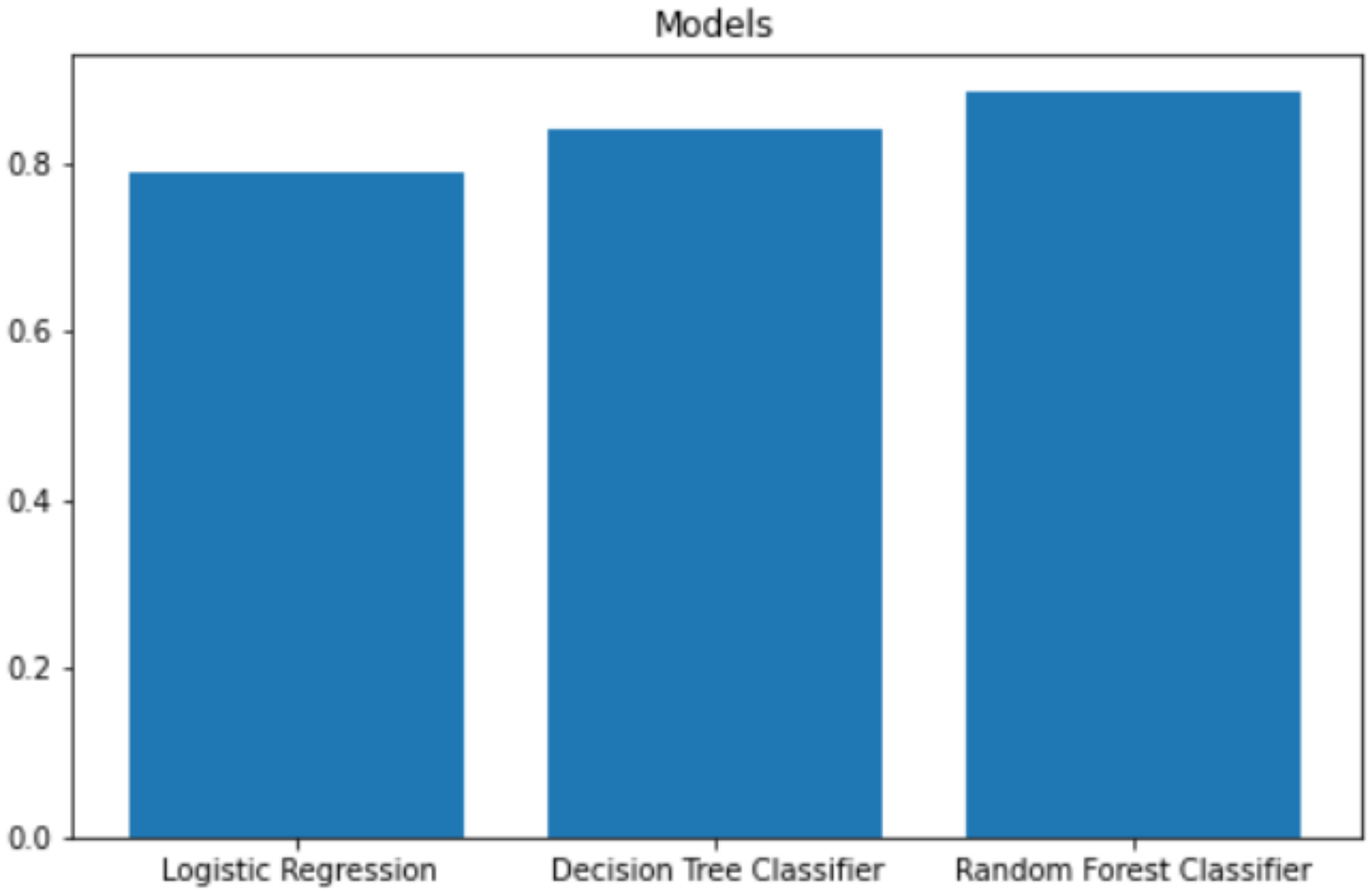
Correlations





- Models

I used several machine learning algorithms such as logistic regression, k-nearest neighbors, decision tree, and random forest classifiers, before settling on random forest as the model with best performance.


Models
0.8
0.6
0.4
0.2
0.0
Logistic Regression
Decision Tree Classifier
Random Forest Classifier

• Model Evaluation

The Classification Report of random forest model results are:

- Accuracy 0.885
- F1 0.87 macro
- precision 0.88 macro
- recall 0.87 macro