Estimation of Hotel Demand: Prediction of booking cancellation

Pricing and Revenue Management Project

Vaishnavi Soundararajan

TABLE OF CONTENTS

1. Background of the industry

The hotel industry is a part of the larger travel and tourism sector. The global hotel industry was thriving and worth over $570 billion and made up of 18 million rooms in 2019. The US is the largest market with around 51% of the global revenue. The industry has been growing for the past nine years owing to rising income levels and growing demand for travel. Technological advancements and hence accessibility have also played a key role in the growth of the industry.

The **average daily rate** (ADR) is a metric in the industry to measure the average rental income per paid occupied room. In the US, this has increased from $124 in 2015 to $129.83 in 2018. With increasing demands, this number seems to have only been rising in the past years. The revenue per available room (RevPAR) and occupancy rate in the industry has been $85.96 and 50-80% respectively. These numbers show us that the industry as a whole has been healthy and performing well. However, with the current health crisis impacting travel, the ADR has dropped by 15% and RevPAR has dropped by 50.6% in the US.

With the increase of competition and advent of comparison shopping in hotels, pricing and excellent customer service would create strong differentiators and build a loyal customer base. The rise of industry post the economic stress right now is going to be challenging, but it's important that governments look into this as this global industry is responsible for one tenth of the world GDP and 1 in 25 jobs in the US.

2. Business Problem

The hotel demand stems from business travelers and leisure tourists who travel frequently. With information on the hotel bookings, our goal is to predict booking cancellation. The business questions we are trying to answer are:

1) Would a customer cancel his booking?
2) If yes, what are the driving factors of cancellations?

Importance of the Problem

Predicting hotel demand is useful in allocating staff to cover the front desk, restaurant, and cleaning. In addition to that, knowing the expected hotel demand helps to estimate the demand of subsidiary businesses like gift shops, gyms and spas that are housed in hotels.

With the increase in flexible cancellation policy through online services, estimating the cancellation rates becomes important so that the hotel is not left with spare rooms to sell at a lower rate. Also, the optimal overbooking rate for the hotel with seasonality in mind could be derived by understanding the expected cancellation rate. So, predicting the likelihood of cancellation feeds into the demand analysis of the hotel in the considered time period.

3. Data Overview

The data we worked on is the hotel demand dataset on Kaggle that contains information on two types of hotel, the resort and city hotel. The dataset has 1,19,390 rows and 32 features. Each observation is a hotel booking done by a customer. The dataset includes bookings by customers who would arrive between 1st July 2015 and 31st August 2017, including information on cancellation status. It also has features indicating the stay information like length of stay, number of adults, the facilities provided by the hotel like parking and details on the country, market segment and arrival date of the booking.

Label	Is_Cancelled
Predictors	Hotel,Lead time,Arrival date year, Adults, Children, Babies, Meal, Country, Stays in weekend night, Stays in weeknights and 22 others.

Table 1: Label and predictor set

Detailed data dictionary has been included in the Appendix.

4. Approach and Methodology

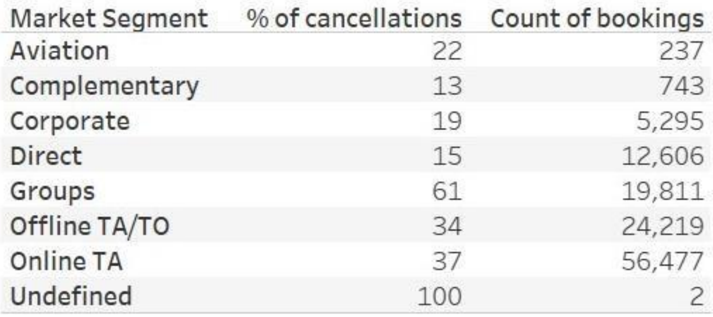
4.1 Exploratory Data Analysis

We explored the data further using Tableau to understand the relationship between the predictor variables and the label. This helped us develop hypotheses on whether the feature could be an influencing factor for cancellation.

4.1 Market segment

The table below gives the percentage of cancellations and the number of bookings across market segments. In terms of bookings, the leading segment is the online travel agents (OTA) segment and the least contribution is from the aviation segment. As discussed in the industry overview, with the increased use of technology in the booking process, it is no surprise that OTA is high on bookings.

The percentage of cancellation of this segment is almost 23% lesser than the group segment. With consumers making the choice of hotel after a lot of deliberation comparing various options, it is expected to see lesser cancellations. The groups segment is the highest on cancellations at 61%. This could be so high since the bookings in this segment is done in buik for events, conferences, etc. So,if the event gets cancelled, then a high number of bookings would be cancelled too.



Market Segment	% of cancellations	Count of bookings
Aviation	22	237
Complementary	13	743
Corporate	19	5,295
Direct	15	12,606
Groups	61	19,811
Offline TA/TO	34	24,219
Online TA	37	56,477
Undefined	100	2

Fig 1: Distribution of percentage cancellations and number of bookings across market segments

4.1.2 Type of hotels

For the market segments observed above, we further explored the percentage of cancellations by the type of hotel for which the booking was made. The dataset primarily consists of two types of hotel, city, and resort. The percentage of cancellations is higher for city hotels when compared to resort hotels. Intuitively, this makes sense as resort hotels are usually booked by leisure travelers who have planned a good vacation. The likelihood of them cancelling the trip is much lesser when compared to city hotels that are booked by business travelers. Across all market segments except complementary, we see that city hotels have a higher cancellation percentage when compared to resort hotels. The OTA segment is the only segment that has almost equal proportion of cancellation across both hotel types.

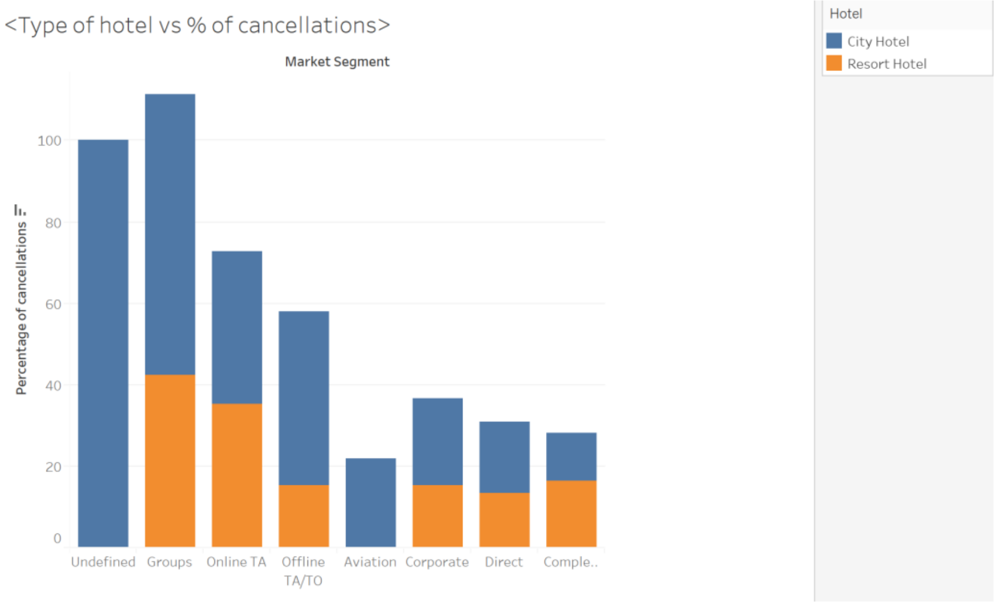


Fig 2: Distribution of percentage cancellations by type of hotel across market segments

4.1.3 Customer Type

In the below tree map, we have visualized the percentage of bookings across the various customer segments in the dataset . It was observed that transient customers account for the highest percentage of cancellations (40.75%) followed by contract customers (30.96%). The transient segment are the customers who are walk-ins/last minute bookings by both individuals and corporates for short stay. The nature of their booking shows that it’s not a pre-planned event and the risk of getting cancelled is high.

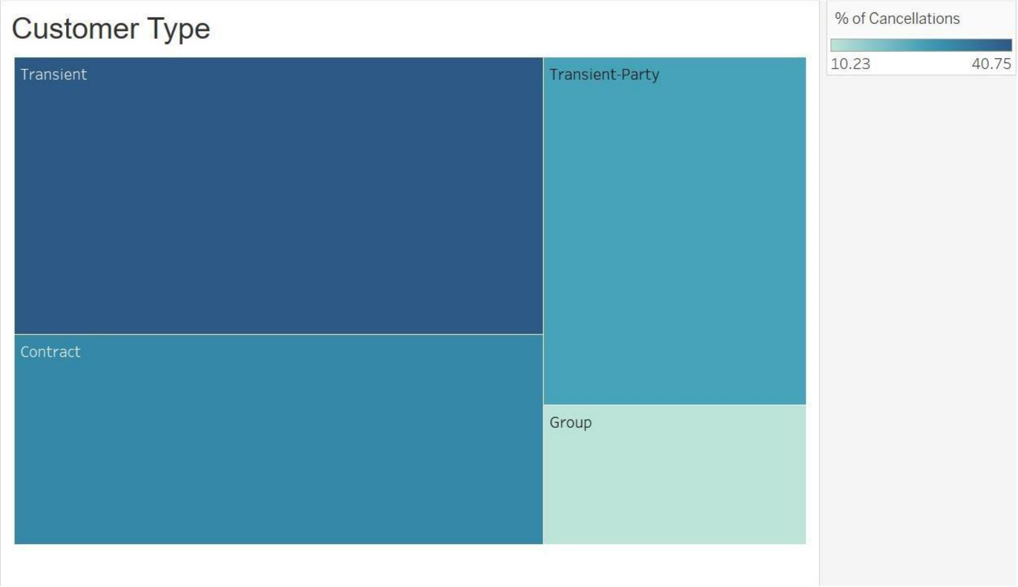


Fig 3: Distribution of percentage cancellations by customer type

4.1.4 Country wise Cancellations

To find the countries with high cancellations, we built the below map. It was observed that a few countries in Southeast Asia and Middle East had high percentages of cancellations. However, the number of bookings for them is very low in the order of ones. So, observing the countries with the number of bookings greater than the average of 670 bookings, we found that Portugal (57%) and China (46%) had higher percentages of cancellations. It is important to note that we are unaware of the number of properties considered in those areas.

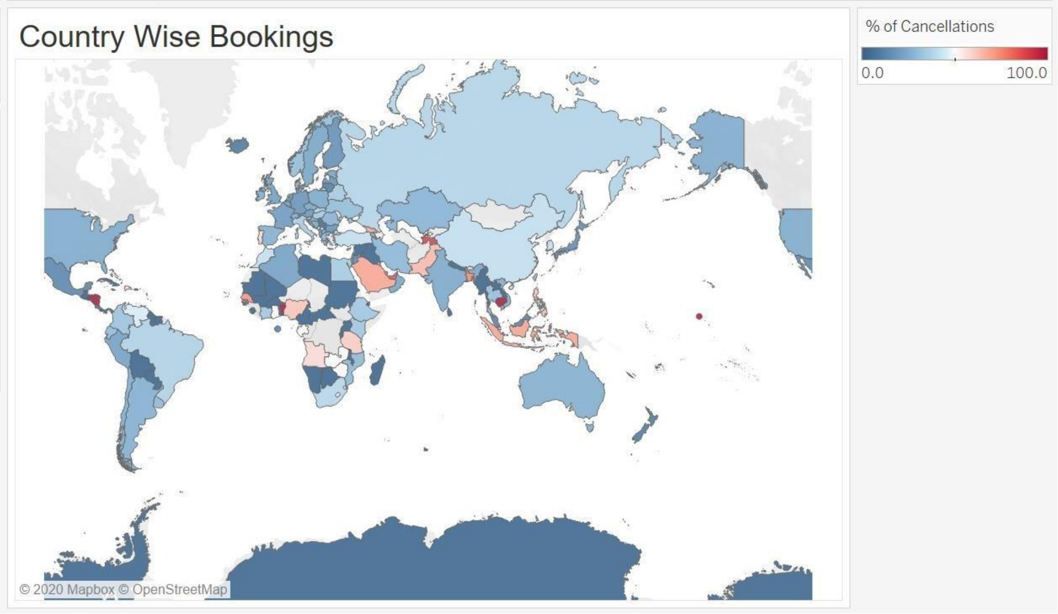


Fig 4: Distribution of percentage cancellations by countries

4.1.5 Days of the Week

When considering the day of week, given that people travel more on weekends compared to weekdays, it is expected to see a rise in bookings over the weekends. We expected to see this drive cancellations as well, but it is observed from the plot below that the highest % of cancellations happen on Wednesday and Thursday. In fact, the cancellations are lesser during weekends. However, there is not a huge variation in cancellation percentage seen across days and so this might not be an important factor driving cancellations.

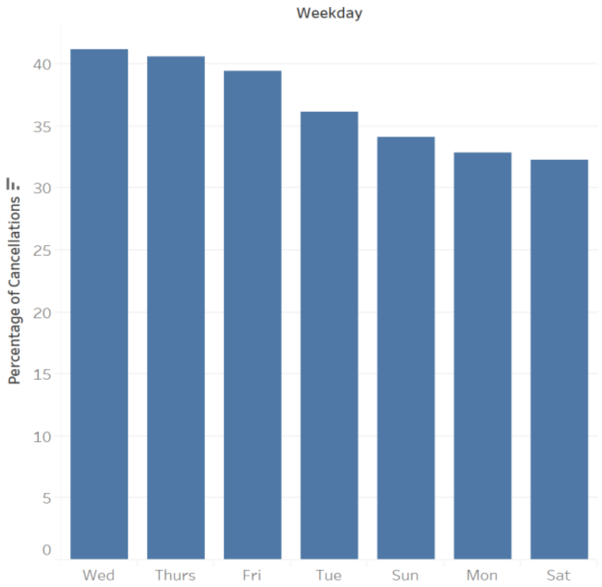


Fig 5: Distribution of percentage cancellations by days of week

4.2 Data Preprocessing

The challenges we faced with the data set were missing values, multicollinearity and class imbalance. The below steps were done to handle them.

- **Reduce Multicollinearity:** We dropped variables with correlation scores of 90% and higher.
- **Handle Missing values:** With missing values in four features, two variables (agent and company) that had a high percentage of missing values were dropped. The missing values in the country column were grouped into a new category named ‘Missing’ and the children feature was imputed with the mode of the values. The same steps were repeated on the test set separately to avoid data leakage.
- **Feature Engineering:** The model results are as good as the features we give into them. So, with our understanding of the data set, a few features were engineered. We expected these features to be good indicators of cancellations. The features have been listed below.

    1. Arrival_date: Combined the individual arrival_date year, month, and day columns to create a single date column
    2. Arrival_date_weekday: Created a column to store the weekday of the arrival day since it would help in forecasting demand on weekdays vs weekends

3. Family_or_not: Created a flag to check if the booking was made by a family using the adults, children,babies columns
4. Room_type_changes: Created a flag to check whether they were assigned the same room type as they requested. This would also help in predicting cancellations and would help us understand customer experience.
5. Total_No_Of_Nights_Stayed: Combined the number of nights stayed during weekdays and weekends to get Total Nights stayed
6. Non_Refund_Flag: Created a flag to check if the booking was non refundable. This usually is an important factor which customers consider before cancelling their reservations

- **One Hot Encoding:** The categorical features were one hot encoded with one dummy column created for each of the N-1 categories (where N = number of categories in the feature). The reason for considering N-1 categories is to avoid multicollinearity. While tree based models would handle categorical variables, this step is important for models such as logistic regression that require numerical independent variables.
- **Handle class imbalance:** The dataset was imbalanced implying that the label had a very small proportion of customers who cancelled their bookings (37%) when compared to the other class - customers who did not cancel (63%). The class of interest to us was the cancellation class and in order to create an effective and accurate model, the data needed to be balanced. We used downsampling keeping in mind computation time and resources.

The dataset now had 61,914 rows and 239 columns.

4.3 Feature Selection

We then used Recursive Feature Elimination (RFE) to select the best feature set to give into the models. This step is important for two reasons:

- Reduces computational time and complexity
- Reduces overfitting in models that are prone to overfit; removing misleading data leads to increase in model accuracy

RFE works by recursively removing attributes based on a metric and building a model on those attributes that remain. Random forest model with scoring as ‘f1’ and 3-fold cross validation was done to obtain the best features. The choice of metric is discussed in detail later. It was observed that 43 features were selected. The data set now had 61,914 rows and 43 features.

4.4 Model Building

The data was now ready to be used in the models. We now wanted to build model s and use it to find the best fit for our business problem. The data was split into training and validation set to control for underfitting and overfitting.

4.4.1 Choice of models

The goal being to predict likelihood of cancellations, we treated this as a classification problem and chose models accordingly. We built the following models for the below reasons:

- **Logistic Regression** - The log link bounds the value between 0 and 1 making it suitable for a classification problem. It’s a simple model with an interpretable equation and the coefficients give us the direction of influence on the dependent variable.
- **Decision Tree** - This tree model is high on explainability. The step of caution is that the predictions are sometimes coarse as one prediction is made for an entire subdivision of the feature space and is also susceptible to overfit.
- **Random Forest** - This tree model gives more accurate results as it is an ensemble of many individual models. The random selection of features to build each model makes the trees built less correlated and so this could improve the results. Also, it gives the important features by using information gain.
- **CatBoost** - This tree model works best with categorical information. As our dataset had a lot of categorical features, this model was built to improve the results. This model also gives the important features.

4.4.2 Choice of metric

The metric considered to evaluate the performance of the model was Recall. It’s defined as,

Recall = True Positive / (True Positive + False Negative)

where, True Positive is when the model correctly predicts the actual positive class (cancellation class here). False Negative is an outcome when the model misclassified the positive class. This score is high when true positive is high and false negative is low.

With more emphasis on the hotel owners being able to overbook, we wanted the recall to be high as the cost of mis predicting a cancellation is much higher than mis predicting a non-cancellation. At the same time, we kept a tab on the precision value to maintain the readiness of the hotel for customers.

4.4.3 Logistic Regression

This model requires there to be little or no multicollinearity among the independent variables to perform better. Also, it assumes linearity of independent variables and log odds. Using L2 regularization, the best performing version of this model gave us a recall of 0.86. This was used as the baseline model to which further model performances were compared with.

Feature Importance

The coefficients of the logistic regression model were looked at to understand the effect of the features on the label. The below plot shows the twenty important features with high weights, both in the positive and negative direction. This shows that if there is non-refund, then the likelihood of cancellation is high.

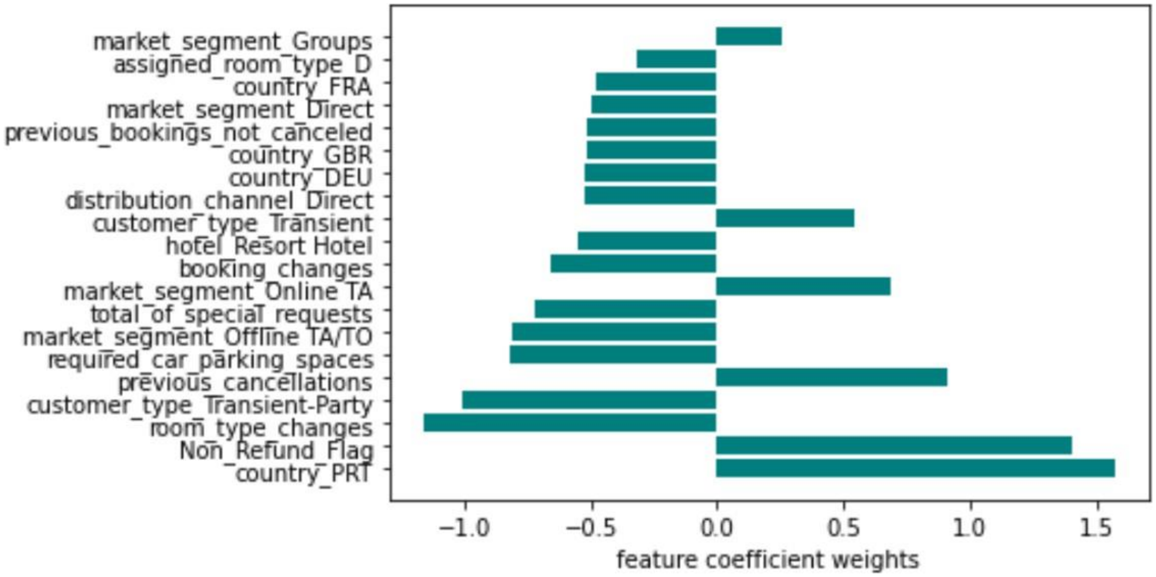


Fig 6: Important features of Logistic Regression

4.4.4 Decision Trees

We then used the Decision Tree algorithm to classify cancellation cases from non-cancellation. Decision Tree uses recursive partitioning to split the feature space based on decision thresholds and identify important features using information gain from the feature. Both categorical and numerical data can be handled by Decision Tree and for this data a tree with max depth of 10, gave us a recall of 0.93.

Feature Importance

The feature importance derived from this model has determined that bookings have a higher chance of getting cancelled if that booking is nonrefundable, when the booking is done via an online Travel Agent and if the lead time is high i.e. if the booking was made much earlier in advance.

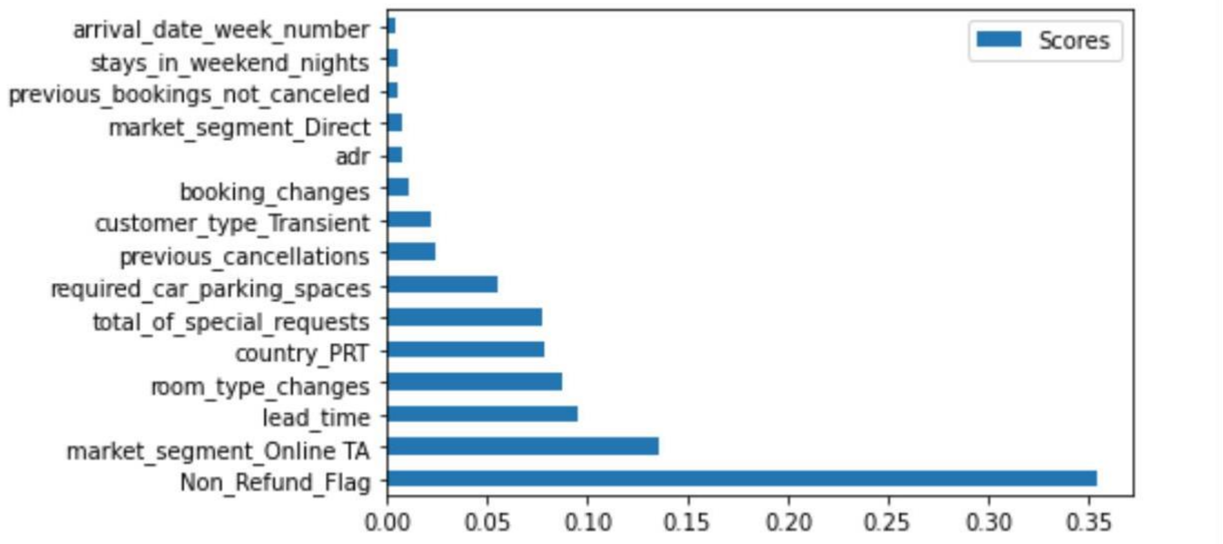


Fig 7: Feature importance of decision tree

4.4.5 Random Forest

Random Forest is an ensemble tree-based algorithm which creates multiple decision trees from randomly selected subsets of data from the training set. These decision trees are ultimately aggregated based on votes from the different trees to finalize the best tree. The untuned default model gave us a recall score of 0.99 on the train set and 0.91 on the validation set. The model was clearly over-fitting. So, tuning the model further using GridsearchCV best parameters of depth 13, number of trees as 500 and min sample split of 2 gave us a recall score of 0.93. This handled the overfitting and improved the results.

Feature Importance

The feature importance derived from this model has determined that bookings have a higher chance of getting cancelled in Portugal when the deposit type is non-refundable and if the lead time is high. These are like the features derived using decision trees.

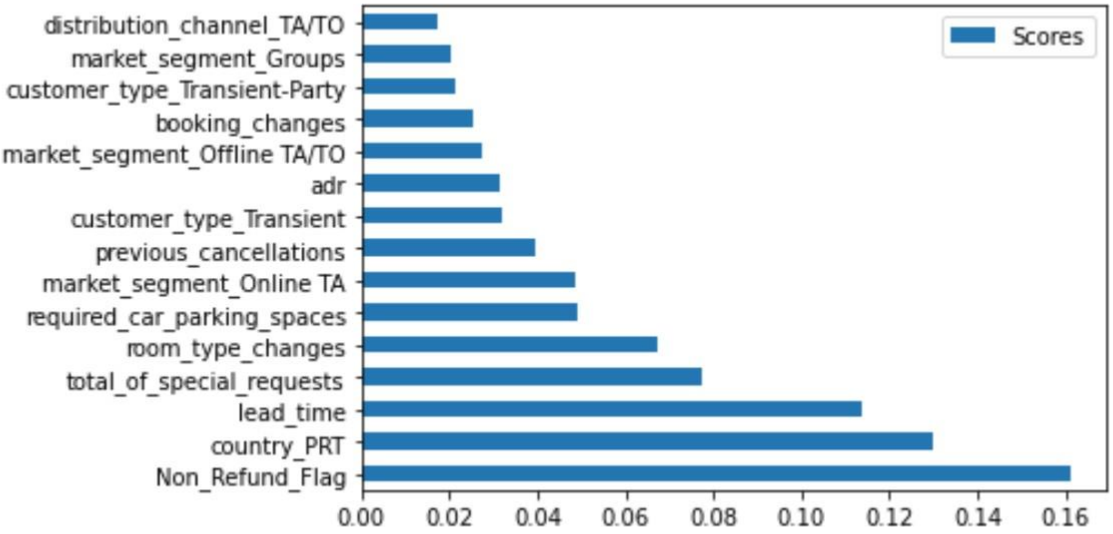


Fig 8: Feature importance of random forest

4.4.6 CatBoost

CatBoost stands for 'Category Boosting' and is an ensemble algorithm that uses the principle of Gradient Boosting. The model can work with a diverse set of data types while providing best in class performance. Two of the main advantages of the CatBoost package are its out of the box handling of data types such as categorical variables and its highly competitive performance without requiring extensive data training.

Feature Importance

The important features are again similar to those of Decision Tree and Random forest. The below plot shows the important features.

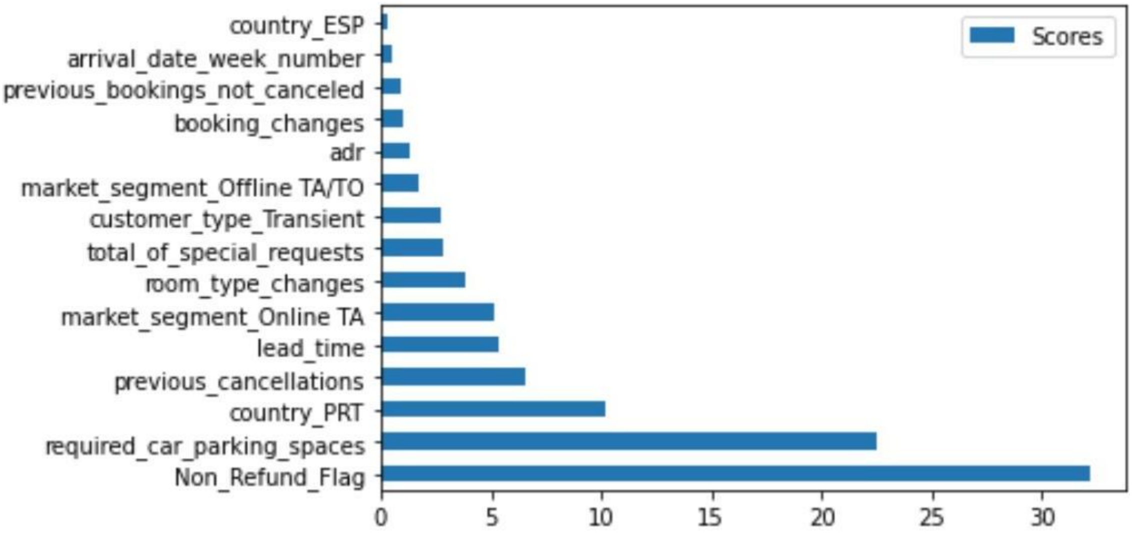


Fig 9: Feature importance of catboost

5. Model Comparison

Comparing the scores on the validation set for the various models, we see that Random Forest and Decision Tree have the highest recall score of 0.93, but Random forest has a higher precision score as well. So, we select the Random Forest model to base our recommendations on.

Model	Precision	Recall	F1-Score
Logistic Regression	0.61	0.86	0.72
Decision Tree	0.63	0.93	0.75
Random Forest	0.66	0.93	0.77
CatBoost	0.68	0.90	0.78

Table 2: Result scores from models

6. Recommendations

Studying the results of Random Forest and the inferences we got using the feature importance values, we recommend the following to hotel business owners.

- Get the expected number of cancellations on a day from the model and use it to calculate the overbooking limit that would give the highest returns with the capacity and booking limit in mind
- Realize the true demand for the day and plan staffing and other supplies accordingly. For example, the number of bookings for Jan 2017 is 1075 in the test set and the predicted number of cancellations for the same time period is 512. So, the expected demand is only 563. Now, the hotel could prepare for supplying 563 customers and open the unoccupied rooms for booking.
- Room type changes reduces the probability of cancellation. So, if the hotel is not fully occupied, offer upgrades to customers. This reduces the likelihood of cancellation and ensures a part of the revenue is realized.
- Total number of special requests reduces the likelihood of cancellation. So, being flexible on offerings seems to encourage customers to tay. Knowing the true demand, which now can be calculated from our models would help us accommodate these requests.

7. Appendix

Data Dictionary

Column Name	**Data Type**	**Entry Example**	**Description**
Hotel	string	‘Resort Hotel’	Hotel is a city hotel or a resort hotel
Is_cancelled	binary	0	Booking was cancelled (1) or not
Lead_time	Int	342	The number of days elapsed between the entering date of the booking in the property management system and the arrival date
Arrival_date_year	int	‘2015’	Year of arrival date (ranges from 2015 to 2017)
Arrival_date_month	string	‘July’	Month of arrival date
Arrival_date_week	int	27	Week number in a year for the arrival date (ranges from 1 to 53)
Arrival_date_day	int	1	Day of the month for the arrival date (ranges from 1 to 31)
Stays_in_weekend_nights	int	0	Number of weekend nights (Saturday or Sunday) that the guest stayed or booked to stay at the hotel
Stays_in_week_nights	int	0	Number of weeknights (Monday to Friday) that the guest stayed or booked to stay at the hotel
Adults	int	2	Number of adults
Children	int	0	Number of children
Babies	int	0	Number of babies
Meal	string	‘BB’	Type of meal booked
Country	string	‘PRT’	Country of origin
Market_segment	string	‘Direct’	Market segment designation
Distribution_channel	string	‘Direct’	Booking distribution channel

Is_repeated_guest	binary	0	Binary value indicating if the booking name was from a repeated guest (1) or not (0)
Previous_cancellations	int	0	Number of previous bookings that were cancelled by the customer prior to the current booking
Previous_bookings_not_cancelled	int	0	Number of previous bookings not cancelled by the customer prior to the current booking
Reserved_room_type	string	‘C’	Code of room type reserved
Assigned_room_type	string	‘C’	Code for the type of room assigned to the booking
Booking_changes	int	3	Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
Deposit_type	string	‘No Deposit’	Indication on if the customer made a deposit to guarantee the booking
Agent	int	304	ID of the travel agency that made the booking
Company	int	110	ID of the company/entity that made the booking or responsible for paying the booking
Days_in_waiting_list	int	0	Number of days the booking was in the waiting list before it was confirmed to the customer
Customer_type	string	‘Transient’	Type of booking, assuming one of four categories: Contract, Group, Transient, Transient Party
Adr	int	0	Average Daily Rate as defined by dividing the sum of all lodging

			transactions by the total number of staying nights
Required_car_parking_spaces	int	0	Number of car parking spaces required by the customer
Total_of_special_requests	int	0	Number of special requests made by the customer (e.g. twin bed or high floor)
Reservation_status	string	‘Check-Out’	Reservation last status
Reservation_status_date	date	‘7/1/2015’	Date at which the last status was set

Table 3: Data dictionary

8. References

https://www.condorferries.co.uk/hotel-industry-statistics
https://www.ihgplc.com/-/media/FF2DB7BB29C54FF2824393006F15A08F.ashx
https://www.xotels.com/en/glossary
https://www.sciencedirect.com/science/article/pii/S2352340918315191