Project: Hotel Bookings Cancellation Prediction

The goal of this project is to find out the characteristic of customers who cancelled and finding a pattern in cancelled booking by doing an exploratory data analysis & building classification machine learning model to predict cancellation, that has accuracy score around 0.8 - 0.9

We have a data set of two hotels situated in Portugal over a period of 24 months from July 2015 to June 2017 obtained from hotel_bookings_dataset.

Background Knowledge:

- Hotel industry is one of the faster growing businesses of tourism sector, especially with the rise of giant OTA that make booking a hotel as easy as is ever been.

Problem Statement:

- The increased trend of cancellation from year to year has affected hotel not being able to accurately forecast occupancy within their revenue management, and the trend of cancellation also have causes hotel loss in opportunity cost (unsold room due to cancellation).

I performed the following sequentially one after the other:

Exploring the data

1. Checked dimensions: 119390 rows, 31 columns
2. Checked datatypes: Both numerical and categorical datatypes
3. Target Variable: is_cancelled
4. No null column in the dataset
5. Is_cancelled value_count:
    a. Around 63% are not cancelled
    b. Around 37% are cancelled
6. Statistical analysis using describe() function:
    a. Minimum “adr” value is negative
    b. No. of adults in one row is 0
    c. Almost half of the books were from Portugal
    d. City hotel(Lisbon) is more favoured than Resort hotel(Algarve region)
    e. Discripancy found in reserved room type and assigned room type
    f. “No Deposit” is the most popular deposit type
    g. Most popular room type is type “A”

Data Cleaning, Preparation and Outliers Handling

A. Cleaning:

1. Checked count of null values using df.isna().sum() function

- Children: 4 null values
- Country: 488 null values
- Agent: 16340 null values
- Company: 1,12,593 null values

We'd try to impute and fill the null values and might need to drop the columns like "agent" and "company"

2. Filled null values:
    - "Children" null values with 0
    - "Country" null values with most frequent country

   Dropped "agent" and "company" columns
3. Checked Outliers using Boxplot:
    - Many columns had outliers
    - I'd handle outliers by binning the columns
    - There was one "adult" value 0 and one "adr" value was beyond 5000 while one "adr" value was negative. So I dropped the rows with these values

B. Preparation:

1. Arrival & Total Stays Columns:
    - Created new columns "arrival_date" from "arrival_date_year", "arrival_date_month", "arrival_date_day" and from this new column, we created "arrival_month", "arrival_day", "arrival_year" .
    - Created "total_stays" by adding the columns "stays_in_weekend_nights" and "stays_in_week_nights"
    - Checked for rows with "total_stays"=0
      (Some hotels may not allow arrival_date and checkout same day. But some hotels may allow it. So we could not drop these rows)
2. Kids, Guests and Children:
    - Created "guests" by combining "adults", "babies" and "children"
    - Created "kids" by combining "babies" and "children"
3. Meal:
    - Replaced "undefined" with "SC"
4. Dropped redundant columns:
    - Dropped "arrival_date_year", "arrival_date_month", "arrival_date_day", "stays_in_weekend_nights", "stays_in_week_nights", "babies" and "children" columns
5. Boxplot for Outliers:
   Still have outliers in columns like parking space, kids, repeated_guests etc. But we could not drop these since they could give important information.

Exploratory Data Analysis (EDA)

A. Univariate Analysis:

1. Hotel Type:
    - City hotel: Around 66% bookings
    - Resort hotel: Around 34% bookings
2. Is_cancelled:
    - Not cancelled: Around 63% bookings
    - Cancelled: Around 37% bookings
3. Lead_time:
    - After converting lead time to months, I found that majority of the bookings (31%) were made during same month as planned arrival.
    - While some bookings were made even before a year
4. Distribution_channel:
    - TA/TO: 82% bookings (largest)
    - Direct: 12% bookings (second largest)
5. Market_segment:
    - This is similar to Distribution_channel.
    - TA/TO (online & offline): 67% bookings (largest)
    - Group: 17% bookings (second largest)
    - Direct: 11% bookings
6. Meal_type:
    - BB: 77% (most popular)
    - FB: 0.67% (least popular)
7. Country:
    - 41% bookings were from Portugal (i.e. almost half).
    - I grouped the data in two categories Local and International.
    - Local: 41% bookings
    - International: 59% bookings
8. Reserved_room_type:
    - Type A room: 72% bookings (the most popular)
9. Deposit_type:
    - No_deposit: 87%
    - non_refund: 12%

    This could be the reason of higher rate of cancellation.
10. Is_repeated_guest:
    - Only 3% guests are repeated
11. Previous_cancellation:
    - There are many categories like 0 times, 1 time, 2 times cancelled.
    - I combined in two categories: previously_cancelled or not.
    - Around 5% bookings were previously cancelled.
12. Bookings_changes:
    - I combined different categories into two categories: booking_changes or not.

- Around 15% bookings were previously cancelled.

13. Special_requests:
    - More than half of the customers don't have any special requests (58%).
14. Customer_type:
    - Transient: 75% bookings (most popular)
    - Transient Party: 21% bookings (second)
15. Arrival_day:
    - Friday and Thrusday has the highest bookings rate with around 16% each.
    - Sunday and Tuesday has the least bookings rate with around 12% each

B. Multivariate Analysis I:

1. Hotel type and cancellation:
    - City hotel cancellation: 41%
    - Resort hotel cancellation: 27%

    Reason: Increase in number of bookings implies increase in cancellation rate

2. Lead time and cancellation:
    - Lead time of more than 7 months has more than 50% cancellation rate.
    - Lead time is positively correlated with cancellation rate.
    - There is single booking in month 23 and 24 each with 100% confirmed rate
3. Arrival year and cancellation:
    - The cancellation rates for every year are between 35% and 38% which are similar to industry standard reported in2018
4. Arrival month and Cancellation:
    - June: 41.5%
    - April: 40.8%
    - May: 39.6%

    These are summer/school holidays months implies more bookings which further imply more cancellation rate.

5. Adults and cancellation:
    - More than 4 adults in a single was always cancelled.
6. Kids and cancellation:
    - Kids don't have any correlation with cancellation rate.
    - Cancellation rate varies from 25%-50%.
7. Guests and cancellation:
    - Guests have effect on cancellation rate similar to adults.
8. Meal and Cancellation:
    - FB : 59.8%
    - Other meal types: 34-37%
9. Prev_cancelled and cancellation:

- Prev_cancelled: 92%
- Never cancelled: 34%

10. Location and cancellation:
    - International: 24%
    - Local: 56%

    Further analysis for why local bookings have high cancellation rate

    - Booking location and previously cancelled:
        - 99.5% of International bookings were never cancelled before.
        - 87% of Local bookings were never cancelled before.

    Since prev cancelled bookings have 92% cancellation rate implying local bookings have higher cancellation rate.

11. Distribution channel and cancellation:
    - TA/TO: 41%
    - Direct: 17%
12. Market segment and cancellation:
    - Groups: 61%
    - TA (Online+Offline): almost same to groups
    - Direct : 30% (least)
13. Booking changes and cancellation:
    - Already changed: 15%
    - Never changed: 41%
14. Deposit_type and cancellation:
    - Non refund: 99.4%
    - No deposit: 28%
    - Refundable: 22%

    Further analysis for why Non refund have highest cancellation rate:

    - Deposit type and lead time:
      Median lead time for
        1. Non refund: 183 days
        2. Refundable: 169 days
        3. No deposit: 56 days

    No refund has highest median lead time compared to other types and analysis showed that higher lead time is more prone to cancellation.

15. Special requests and cancellation:
    - 0 requests: 48%
    - 1 request: 22%

Higher the requests, lower the cancellation rate

16. Parking space and cancellation:
    - 0 space: 39%
    - 1+ space: 0%

    Guests needed parking space never cancelled bookings

    A total of 6.2% (7407) guests required parking space

C. Multivariate Analysis II:

1. Hotel type, Lead time and Cancellation:

   For both the hotels, median lead time for
   - Cancelled bookings = 3 months
   - Confirmed bookings = 1 month
2. Hotel type, Median Adr and cancellation:
   - City hotel:

     Median adr for:
     - cancelled bookings: 99.9%
     - confirmed bookings: 100%
   - Resort hotel:

     Median adr for:
     - Cancelled bookings: 84%
     - Confirmed bookings: 72%
3. Hotel type, meal and cancellation:
   - City hotel:

     FB has high cancellations
   - Resort hotel

     Higher confirmed bookings than cancelled ones in any meal category
4. Market segment and deposit type:
   - More than 50% group bookings were made with No deposit
   - Almost all online TA bookings are done with No deposit

   Hotels still have full control in determining what kind of deposit policy need to be implemented on online TA
5. Prev cancelled and Deposit type:
   - 56% of bookings with nonrefundable deposit type were cancelled
   - 43% of bookings with no deposit type were cancelled

   Since 92% of prev cancelled bookings are cancelled. So there should be only nonrefundable bookings or refundable (with charges) who had cancelled their previous bookings.
6. Prev cancelled and market segment:
   - More than 50% of previous cancelled bookings were made by groups

Visualization

I visualized the various univariate and multivariate features with the help of pie charts, histograms, bar plots, line plots and count plots and found some more characteristics of customers who cancelled the bookings and the other factors affecting the cancellation.

1. Booking Locations:
   I visualized by pie chart and found that
   - International bookings: 58.8%
   - Local bookings: 41.1%
   - City hotel: 66.34%
   - Resort hotel: 33.66%
2. Booking market segment:
   I visualized by pie chart and found that:
   - Online TA: 47.25%
   - Offline TA: 20.32%
   - Direct: 10.55%
3. Distribution Channel:
   I visualized by pie chart and found that:
   - TA/TO: 82%
   - Direct: 12.25%
4. Deposit type:
   I visualized by pie chart and found that:
   - No deposit: 87.61%
   - Non refund: 12.26%
   - Refundable: 0.14%
5. Distribution of ADR:
   I visualized by histogram and found that:
   - Highest frequency of adr is where adr is around 61.5 - 62.49
6. Arrival month:
   I visualized by bar plot and found that:
   - August has highest total bookings followed by July, May and October

   Summer holidays in Portugal were between 09 June and 12 Sep 2021. I assumed that summer holidays happened around the same time period in 2015-2017.

   Also August and July appeared 3 times while the rest appeared for only two times.

7. Arrival Month year and ADR:
   I visualized by bar plot and line plot and found that:
   - % of bookings in winter (Nov, Dec and Jan) are low
   - % of bookings in summer (March, Apr, May) are high
   - Line plot for median price gown down in winter while goes up in summer
8. Arrival month year for each hotel:

I visualized by normalized bar plots and found that:

- In winter, Resort hotel outperforms.
  Reason:
  City hotel offered possibility of exploring around but winter is not a good time to explore
  Resort hotel is situated in Algrave region where temp. is 17-20 Celsius in winters while that in Lisbon is 08-15 Celsius.

9. ADR for each hotel:
   I visualized by line plot and found that:
   - Resort hotel price is more fluctuating than city hotel
   - Resort hotel price increases during summer and goes down during winter (Aug to Nov) and then increases in Dec (may be because of Christmas)
10. Number of cancellations towards number of bookings:
    I visualized by countplot (monthyear and count) and barplot (is_cancalled and not cancelled) and found that:
    - Number of cancellations increases with number with number of bookings, but no linear effect towards number of bookings
    - Some months have high % of confirmed bookings than cancelled
    - Nov 2015 and Jan 2016 have cancellation rate lower than 30%
11. ADR effect on cancellation:
    I visualized by bar plot and line plot and found that:
    - Bookngs increases implies adr and cancellation increases. Thus adr increases with increase in cancellations
12. Median lead time for cancelled and confirmed booking each month:
    I visualized by bar plot and found that:
    - Dec 2015 has the lowest median lead time for cancelled and confirmed bookings
    - March to August have longer median lead time than Sep to Dec
    - For each year, Sep has the highest median lead time for cancelled bookings
    - Lead time for cancelled and confirmed bookings are positively correlated
    - Cancelled bookings median lead time is always higher than that of confirmed bookings
13. Deposit type and month year:
    I visualized by bar plot and found that:
    - More nonrefundable deposit type for longer advance reservation (13 months and above) compared to no deposit
14. Booking location and cancellation:
    I visualized by bar plot and found that:
    - More than 50% local bookings are cancelled

    Why local bookings cancelled more?

- Higher previous cancelled bookings rate in local bookings (prev cancelled vs location bar plot)
- More nonrefundable deposit type bookings by locals (deposit type vs location bar plot)
- Almost all international bookings made without deposit (deposit type vs location bar plot)
- Almost every nonrefundable booking was cancelled (deposit type vs % bookings bar plot)
- Local bookings have Group as the biggest market segment (market segment vs % bookings bar plot)

15. Deposit type for each market segment:
    - No deposit type is trending
    - Most Group bookings with no deposit type and nonrefundable
    - Refundable part is negligible
16. Repeated guests and cancellation:

    I visualized by bar plot and found that:
    - Non repeated guests: 37% cancelled
    - Repeated guests: 14% cancelled

    Where do repeated guests come from?
    - Corporate has highest repeated bookings (23%) then direct with 6%

Machine Learning

A. Feature Selection:

1. Labeled Booking locations as
    - Local: 0
    - International: 1
2. Labeled hotels as
    - Resort hotel: 0
    - City hotel: 1
3. Checked association between all the features:
    - I found that "reservation_status" had 1.00 association with the target variable. Thus I had to drop this for result to be unbiased.
    - Many features were highly correlated (>.70) and some were moderately correlated (>.50) which also needed to be dropped to avoid biased results.
    - Also EDA showed that there were some features not much significantly affecting the bookings cancellation which also dropped for the further process

B. Model Building:

1. After dropping the features, we got the following features for machine learning model:
    1) Hotel_encoded
    2) Booking_location_encoded
    3) Lead_time

4) Market_segment
5) Deposit_type
6) Parking_space
7) Total_of_special_requests
8) Is_prev_cancelled
9) Is_repeated_guest
10) Is_booking_changes
11) Customer_type
12) Total_stays
13) Guests

2. I splitted the data into train and test categories using test_train_split
3. I found that the following are categorical features:
    - Market_segment
    - Deposit type
    - Customer type

    So I used one hot encoding to convert these features to numerical features.

4. I scaled all the numerical features using RobustScaler.

C. Machine learning Algorithms:

1. I used pipeline for model building.
2. I used the following ML algorithms for training the model:
    1) Logistic Regression
    2) KNN
    3) Decision Tree Classifier
    4) Random forest Classifier
3. The following metrices were used for the evaluation of performance:
    - Accuracy
    - Recall
    - Precision
    - F1 Score
4. I found that
    - Decision Tree and Random Forest had the highest score in training (94.19%)
    - Random Forest (84.04%) had slightly higher score than Decision tree (83.18%) in testing.
    - Logistic regression had almost same accuracy for test data (80.41%) and train data (80.24%)
    - KNN had training data accuracy (87.95%) slightly higher than test data accuracy (83.54%)
5. I made predictions for same sample data using Logistic regression and Random Forest and got the following results:
    - Logistic Regression: 75% chances of confirmed booking

- Random forest: 57% chances of confirmed booking

Since both the probabilities are greater than .5, we may consider that there are higher chances of confirmed booking by the person.

And since both the algorithms gave positive results, therefore the model is working with a good efficiency.

How this model will help hotels?

- This model will allow hotel managers / revenue manager to take actions on bookings that's identified as "potentially going to be canceled"
- These prediction models enable hotel managers to mitigate revenue loss derived from booking cancellations and to mitigate the risks associated with overbookings (reallocation costs, cash or service compensations, and particularly important today, social reputation costs)**.**
- Booking cancellations model also allows hotel managers to **implement less rigid cancellation policies, without increasing uncertainty**. This has the potential to translate into more sales, since less rigid cancellation policies generate more bookings.

Limitations of Projects:

- This hotel booking cancellation project only applied for hotel bookings in Lisbon Region and Algarve Region both location are located in portugal.
- Predicting cancellation with this web application outside both region might have not so accurate result due to the location constraint, different pattern of cancellation.