Machine Learning in Python - Project 2

Yin Lau (s1735270)

Ivan Lim (s1657846)

Aditya Mardjikoen (ss2264710)

0. Setup

This is the report for Project 2 - most code has been edited out to meet the page limit.

1. Introduction

Preamble.

In tourism and travel-related industries, most of the research on Revenue Management demand forecasting and prediction problems employ data from the aviation industry, in the format known as the Passenger Name Record (PNR). This is a format developed by the aviation industry. However, other tourism and travel industries such as hospitality, cruising, theme parks, etc., have different requirements and particularities that cannot be fully explored without industry's specific data. Hence, two hotel datasets with demand data are use to aid in overcoming this limitation.

Nature of datasets.

The datasets now made available were collected for the development of prediction models to classify a hotel booking's probability to be canceled. Nevertheless, due to the characteristics of the variables included in these datasets, their use goes beyond this cancellation prediction problem.

The data that we are using is this project was taken from a published study conducted by Nuno Antonioa, Ana de Almeida, and LuisNunes. The data comes from the booking systems of two real hotels and reflect bookings made between July 1st, 2015 through August 31st 2017. This is a large dataset (119,390 observations) and real world (i.e. messy) data set; as such, we conduct feature engineering and data cleaning to remove the most negligible variables and manage the incomplete cases (rows) with missing values. The data can be downloaded from this following link: https://www.sciencedirect.com/science/article/pii/S2352340918315191#f0010.

One of the most important properties in data for prediction models is not to promote leakage of future information. In order to prevent this from happening, the timestamp of the target variable must occur after the input variables' timestamp. Thus, instead of directly extracting variables from the bookings database table, when available, the variables' values were extracted from the bookings' change log, with a timestamp relative to the day prior to arrival date (for all the bookings created before their arrival date).

Goals of this project.

Overall, there are two goals of this project. These goals can be formulated as two questions of interest as follows:

1. How accurate does the predictive model that we proposed for this project to predict the probability of a hotel booking being cancelled?
2. What variables in the hotel booking dataset that we used in this project which affect the probability of the hotel booming being canceled?

2. Exploratory Data Analysis and Feature Engineering

We first examine the nature of the data.

Variable identification.

`is_canceled` is the dependent variable in our model. The categorical variables are as follows:

`hotel` , `arrival_date_year` , `arrival_date_month` , `arrival_date_week_number` , `arrival_date_day_of_month` , `meal` , `country` , `market_segment` , `distribution_channel` , `is_repeated_guest` , `reserved_room_type` , `assigned_room_type` , `deposit_type` , `agent` , `company` , `customer_type`

The rest are numerical.

Given that hotels often experience seasonal patterns in their bookings (e.g. bookings would normally rise during the summer holidays), we therefore select `arrival_date_month` to best reflect the potential seasonality of the data, and hence drop `arrival_date_year` (since there are only 3 unique years in the data), `arrival_date_week_number` and `arrival_date_day_of_month` .

While the research article defines these variables as integer, we treat `arrival_date_month` as categorical as the trend is likely to be non-linear with each week; moreover, it would be more meaningful to deal with months as a categorical rather than ordinal variable. For example, August would be associated with more bookings for Edinburgh hotels due to the Festival Fringe; January would be associated with less bookings due to the cold temperatures and generally miserable atmosphere due to the grey skies, lack of sunlight and constant rain.

	month	% Rate of cancellation
0	July	37.453598
1	August	37.753117
2	September	39.170156
3	October	38.046595
4	November	31.233441
5	December	34.970501
6	January	30.477315
7	February	33.415964
8	March	32.152338
9	April	40.797186
10	May	39.665847
11	June	41.457172

We see that lower rates of cancellation tend to cluster around the months of November to March, with a minimum at January, and that higher rates of cancellation tend to cluster around the months of April to October, reaching a peak at June.

This dataframe and the graph above clearly demonstrates the seasonality of our data with respect to the rate of cancellation. The `month` variable is hence not irrelevant to our model.

Missing data.

We now check for missing data.

```
Index(['children', 'country', 'agent', 'company'], dtype='object')
```

```
array([     4,    488,  16340, 112593])
```

```
array([3.40000e-05, 4.08700e-03, 1.36862e-01, 9.43069e-01])
```

Using $0.5\% \approx 597$ of values as a threshold for a complete case analysis, we can therefore ignore the missing values for `'children','country'` and drop the corresponding rows.

Sanity checks.

We now consider sanity checks.

We note that in any one booking, there should be at least one adult or child. Therefore, it would not make sense to have zero of either associated with a booking.

Another possible sanity check is checking the sum of weekday and weekend nights - however, it is possible that an individual or group may book rooms for only a few hours and not stay overnight, so we cannot assume that a booking with 0 week and weekend nights is invalid.

We now consider the variables `'agent','company'` which contain missing data values.

From the source of the data:

The PMS assured no missing data exists in its database tables. However, in some categorical variables like Agent or Company, "NULL" is presented as one of the categories. This should not be considered a missing value, but rather as "not applicable". For example, if a booking "Agent" is defined as "NULL" it means that the booking did not came from a travel agent.

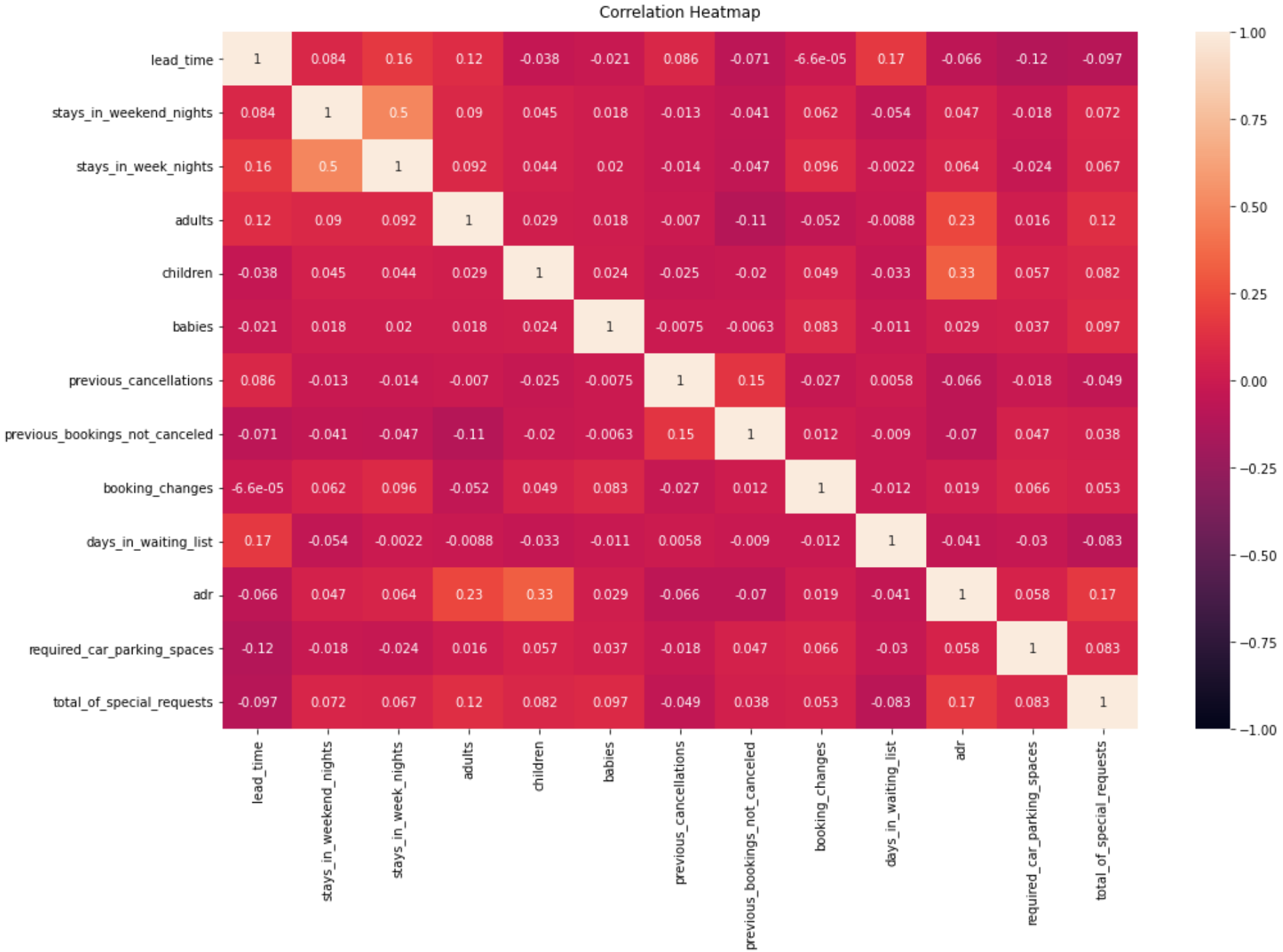
We can therefore replace all NULL values for the Agent variable with the integer 0, signifying that all such bookings were not made by a travel agent. Similarly, we do the same for the Company variable, where 0 signifies that a booking was not made by a company/entity or that a company/entity was responsible for making the payment. As no further information is provided, we are forced to guess that NULL values correspond to bookings made by private individuals.

However, we note that there is a significant proportion of NULL, NULL pairs in the `agent` and `company` variables: $\approx 58.7\%$ and $\approx 8.5\%$ of all missing values respectively. Considering the nature of bookings, it is likely that if a booking was not made by a company/entity or an agent, it must have been made by a private individual (or a private individual representing a group of people). We therefore change all such NULL, NULL pairs to 97, 97, as the ID number 97 is not present in either variable's data - this allows us to use it to represent a private individual.

We now have 13 numerical and 13 categorical variables (each with varying amounts of levels), with one binary dependent variable.

Numerical variables

We now conduct a correlation analysis on the numerical variables and examine highly-correlated variables, setting our correlation coefficient threshold at 0.6.


Correlation Heatmap
lead_time 1 0.084 0.16 0.12 -0.038 -0.021 0.086 -0.071 -6.6e-05 0.17 -0.066 -0.12 -0.097
stays_in_weekend_nights 0.084 1 0.5 0.09 0.045 0.018 -0.013 -0.041 0.062 -0.054 0.047 -0.018 0.072
stays_in_week_nights 0.16 0.5 1 0.092 0.044 0.02 -0.014 -0.047 0.096 -0.0022 0.064 -0.024 0.067
adults 0.12 0.09 0.092 1 0.029 0.018 -0.007 -0.11 -0.052 -0.0088 0.23 0.016 0.12
children -0.038 0.045 0.044 0.029 1 0.024 -0.025 -0.02 0.049 -0.033 0.33 0.057 0.082
babies -0.021 0.018 0.02 0.018 0.024 1 -0.0075 -0.0063 0.083 -0.011 0.029 0.037 0.097
previous_cancellations 0.086 -0.013 -0.014 -0.007 -0.025 -0.0075 1 0.15 -0.027 0.0058 -0.066 -0.018 -0.049
previous_bookings_not_canceled -0.071 -0.041 -0.047 -0.11 -0.02 -0.0063 0.15 1 0.012 -0.009 -0.07 0.047 0.038
booking_changes -6.6e-05 0.062 0.096 -0.052 0.049 0.083 -0.027 0.012 1 -0.012 0.019 0.066 0.053
days_in_waiting_list 0.17 -0.054 -0.0022 -0.0088 -0.033 -0.011 0.0058 -0.009 -0.012 1 -0.041 -0.03 -0.083
adr -0.066 0.047 0.064 0.23 0.33 0.029 -0.066 -0.07 0.019 -0.041 1 0.058 0.17
required_car_parking_spaces -0.12 -0.018 -0.024 0.016 0.057 0.037 -0.018 0.047 0.066 -0.03 0.058 1 0.083
total_of_special_requests -0.097 0.072 0.067 0.12 0.082 0.097 -0.049 0.038 0.053 -0.083 0.17 0.083 1
lead_time
stays_in_weekend_nights
stays_in_week_nights
adults
children
babies
previous_cancellations
previous_bookings_not_canceled
booking_changes
days_in_waiting_list
adr
required_car_parking_spaces
total_of_special_requests
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

There are no highly correlated features - we therefore need not do anything about this. However, there are a few pairs of variables which merit further consideration. We can therefore proceed to scale these variables in preparation for the Principal Component Analysis and thereafter, the model fitting.

Categorical variables

We now examine each of the categorical variables.

```
Index(['arrival_date_month', 'hotel', 'meal', 'country', 'market_segment',
       'distribution_channel', 'is_repeated_guest', 'reserved_room_type',
       'assigned_room_type', 'deposit_type', 'agent', 'company',
       'customer_type'],
      dtype='object')
```

```
City Hotel      79144
Resort Hotel    39584
Name: hotel, dtype: int64
```

```
Online TA        56320
Offline TA/TO    24126
Groups           19787
Direct           12423
Corporate         5101
Complementary      734
Aviation           237`
Name: market_segment, dtype: int64
```

```
0    114928
1      3800
Name: is_repeated_guest, dtype: int64
```

```
No Deposit    104002
Non Refund     14564
Refundable       162
Name: deposit_type, dtype: int64
```

```
Transient          89045
```

```
Transient-Party    25043
Contract            4072
Group                568
Name: customer_type, dtype: int64
```

```
BB           91737
HB           14417
SC           10612
Undefined     1164
FB             798
Name: meal, dtype: int64
```

Since `hotel`, `market_segment`, `is_repeated_guest`, `deposit_type`, `meal` and `customer_type` have relatively few categories and/or have non-negligible counts for each of these categories, we do not modify them in any way. `arrival_date_month` has non-negligible counts for each of its categories, but we do not show it for brevity.

```
TA/TO        97600
Direct       14456
Corporate     6480
GDS            191
Undefined        1
Name: distribution_channel, dtype: int64
```

As the `Undefined category` for `distribution_channel` has only one count, we can therefore ignore this as a categorical variable since we would not be able to determine its effect on whether a booking is canceled or not.

Given that there are numerous countries with extremely small quantities of bookings, we ignore any country with less than $0.1\% \approx 119$ bookings - this gives us 35 countries in total with non-negligible quantities of bookings. Moreover, the total number of bookings for countries below this threshold is only $\approx 2.3\%$ of all bookings.

Feature addition.

Next, consider `assigned_room_type` and `reserved_room_type` - if a customer should be assigned a room type by the hotel different from the room type that they reserved, they would be more likely to cancel the booking as the assigned room type may not satisfy their requirements. We therefore add the binary feature `room_type_change`, taking the value 0 if there is no difference in assigned or reserved room type, and 1 if there has been a change.

Finally, we examine `agent` and `company` in the same way as `country`. We remove agent and company levels which are below the threshold - 82

and 11 respectively.

Encoding of categorical variables

Given the large amount of variables, we originally considered different types of categorical encoding other than one-hot encoding - ordinal and binary encoding.

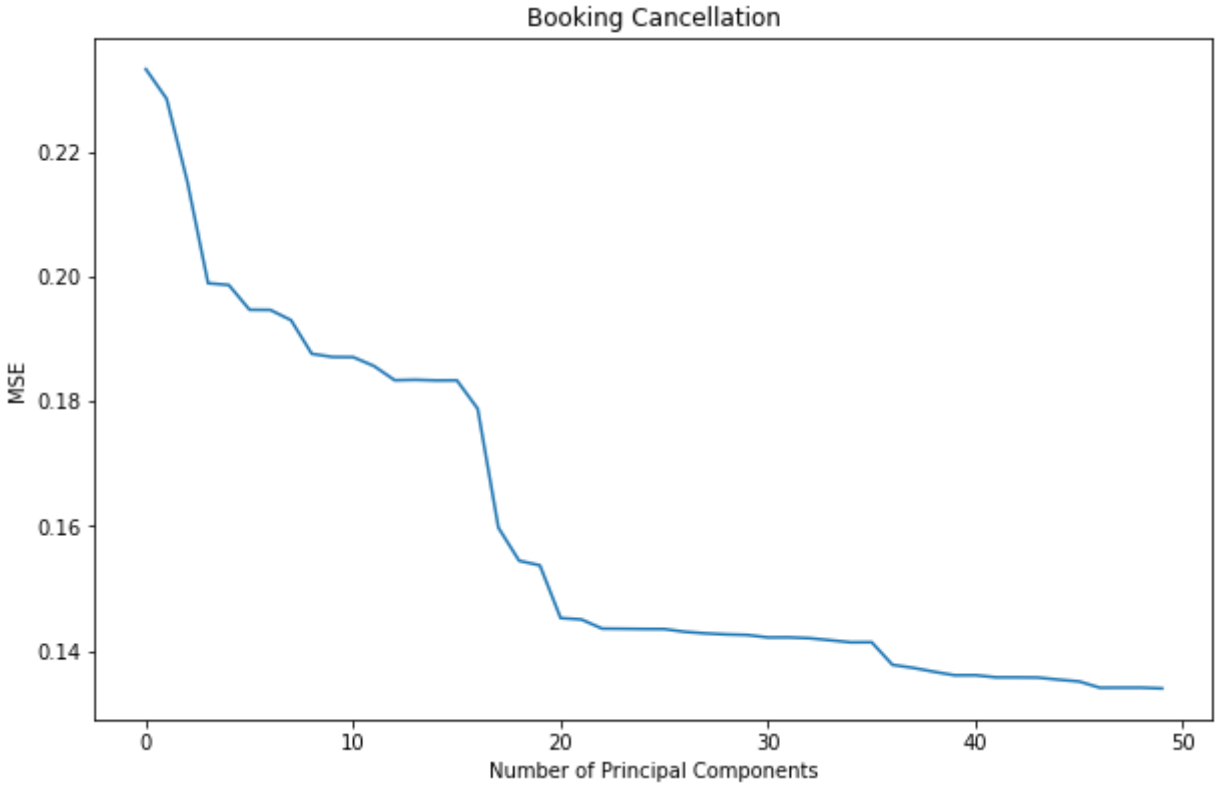
Ordinal encoding would induce a (nonsensical) order on the categorical variables; for example, it would not make sense for 'PRT' - Portugal - to be associated with a higher integer than 'GBR' - Great Britain. It was therefore deemed too risky to use as it would introduce relationships between levels of variables that most likely did not exist.

Binary encoding would massively reduce the number of new dummy variables needed to encode a categorical variable - for example, a categorical variable with 100 levels can be expressed with just seven binary variables, as 100 in decimal is 1100100 in binary. However, the same issue is present here as in ordinal encoding: we would introduce a non-negligible amount of dependence between levels of a categorical variable.

We therefore ultimately decided to one-hot encode all categorical variables, which gives us 188.

Principal Component Analysis

As we currently have a large number of variables, we conducted Principal Component Analysis (PCA) in order to reduce dimensionality and examined the results. We fitted and transformed the training data, and examined the Mean Squared Errors under a cross-validation process and the Maximum Likelihood Estimate in order to determine an optimal number of principal components.



Setting `n_components` to `'mle'` gives the same number of categorical variables, which is not helpful.

If we consider the graph of Mean Squared Error versus number of principal components, we observe that there is also no clear point where MSE is minimised (as it is achieved at the maximum number of components).

We therefore must consider the percentage of explained variance as our main metric for determining our threshold.

Desired % of explained variance	No. of principal components	Actual % of explained variance
0.9	30	0.901106
0.95	44	0.950943
0.975	62	0.975364
0.99	96	0.990019
0.999	166	0.999017

We thus select $n = 44$ principal components as a compromise between minimising the number of variables and maximising proportion of explained variance - this explains $\approx 95.1\%$ of the variance in the data.

Feature selection.

Instead of reducing dimensionality by PCA, we explored the feature selection methods provided by sklearn to select the most relevant features and discard the rest.

There are 3 methods we explored: the chi-squared test, the F-test in ANOVA, and the mutual information between variables. Given that the F-test estimates the degree of linear dependency between random variables, we opted not to make an assumption of linearity. As the mutual information method can capture most types of statistical dependencies as per the `sklearn.feature_selection` documentation and is also non-parametric, we ultimately decided to implement the mutual information method.

Given that the mutual information between two random variables is 0 if and only if two random variables are independent, we selected the threshold of 0 and removed all variables that were independent of `is_canceled` with respect to the mutual information method.

We thus remove 55 features.

Final steps.

We have finished all feature engineering steps. We now scale the numerical data in preperation for model fitting and tuning, and split the final dataset into training and test data.

In conclusion, we have $118,728$ rows of complete data cases and 147 binary (dummy and non-dummy) and numerical variables.

3. Model Fitting and Tuning

Approaches.

We initially fitted and trained four different models: Random Forest, Decision Tree, Naive Bayes and Logistic Regression. This was done for three different datasets - the full feature dataset without any modifications, the PCA transformed dataset and the dataset with 0% mutual information features removed. Ultimately, we decided to fit and tune the Random Forest model, using the 0% mutual-information (MI) dataset. The analysis and justifications are below.

Model fitting on full feature matrix without grid search, 5 fold CV

Metric	Random Forest	Decision Tree	Naive Bayes	Logistic Regression
Accuracy	0.885764	0.853373	0.718683	0.833929
Precision	0.875087	0.799776	0.595443	0.813038
Recall	0.806687	0.805745	0.751056	0.716221
F1 Score	0.839476	0.802745	0.664084	0.761550
FPR	0.067737	0.118620	0.300353	0.096857

We observed that in terms of Recall and Accuracy score, Random Forest and Decision Tree perform best. In terms of FPR, Random Forest is the best scorer. Naive Bayes is generally the worst performer in most of the metrics.

Model fitting on PCA transformed feature matrix without grid search, 5 fold CV

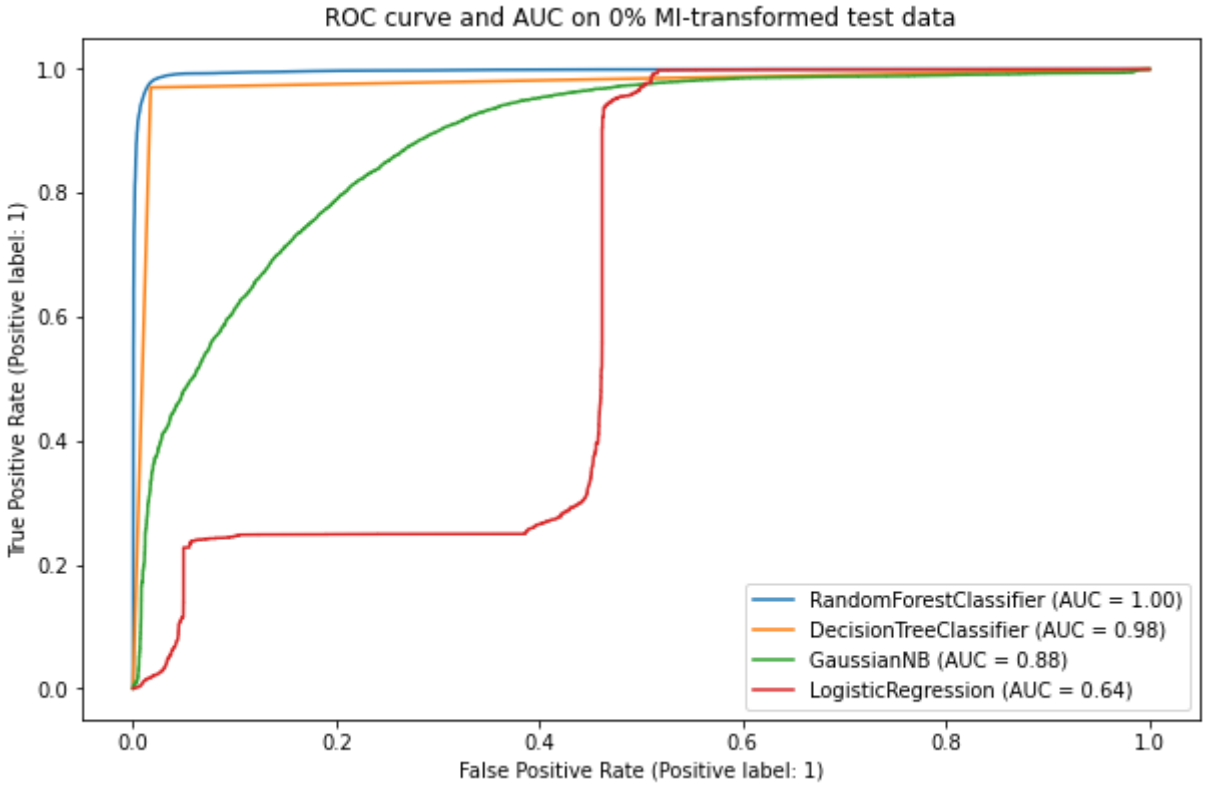
Metric	Random Forest	Decision Tree	Naive Bayes	Logistic Regression
Accuracy	0.874189	0.823918	0.757439	0.823629
Precision	0.879863	0.757987	0.669581	0.808475
Recall	0.764671	0.770520	0.681452	0.686294
F1 Score	0.818214	0.764183	0.675406	0.742381
FPR	0.061412	0.144683	0.197879	0.095615

Again, Random Forest and Decision Tree are the best performers.

Model fitting on 0% MI-removed feature matrix without grid search, 5 fold CV

Metric	Random Forest	Decision Tree	Naive Bayes	Logistic Regression
Accuracy	0.981578	0.978835	0.794631	0.629715
Precision	0.977849	0.970314	0.667510	0.498947
Recall	0.972282	0.972607	0.887860	0.474259
F1 Score	0.975055	0.971457	0.762019	0.486195
FPR	0.012955	0.017503	0.260189	0.357184

Once again, we observe that Random Forest and Decision Tree are the best performers.



Observations on models (without grid search)

If we consider the results of all three datasets, it is clear that using the 0% MI-removed dataset is best in terms of performance (Recall, Accuracy and FPR scores) balanced with the runtime, as the number of variables is between that of the full feature dataset and the PCA-transformed dataset.

Dataset	Random Forest	Decision Tree	Naive Bayes	Logistic Regression
Full (Training)	1.00	1.00	0.80	0.91
Full (Test)	0.96	0.85	0.80	0.91
---	---	---	---	---
PCA-Transformed (Training)	1.00	1.00	0.82	0.90
PCA-Transformed (Test)	0.95	0.82	0.83	0.90

Dataset	Random Forest	Decision Tree	Naive Bayes	Logistic Regression
---	---	---	---	---
0% MI-Transformed (Training)	1.00	1.00	0.88	0.65
0% MI-Transformed (Test)	1.00	0.98	0.88	0.64

By observing the above tables and ROC curves, we find that the random forest and the decision tree model on the 0% mutual information transformed feature matrix produces very close recall values - 0.972282 and 0.972607 respectively. However, we obtain a much lower false positive rate from the random forest model with a value of 0.012955 compared to the decision tree model's value of 0.017503. Moreoever, we note that the test AUC is valued at 1.00 compared to decision tree's 0.98 - we thus proceed with the random forest model from this point onwards.

Random Forest grid search with cross-validation.

We note that k-fold cross validation is usually performed for small datasets. Given that our dataset contains over $100,000$ rows, we therefore set $k = 2$, as we have enough data.

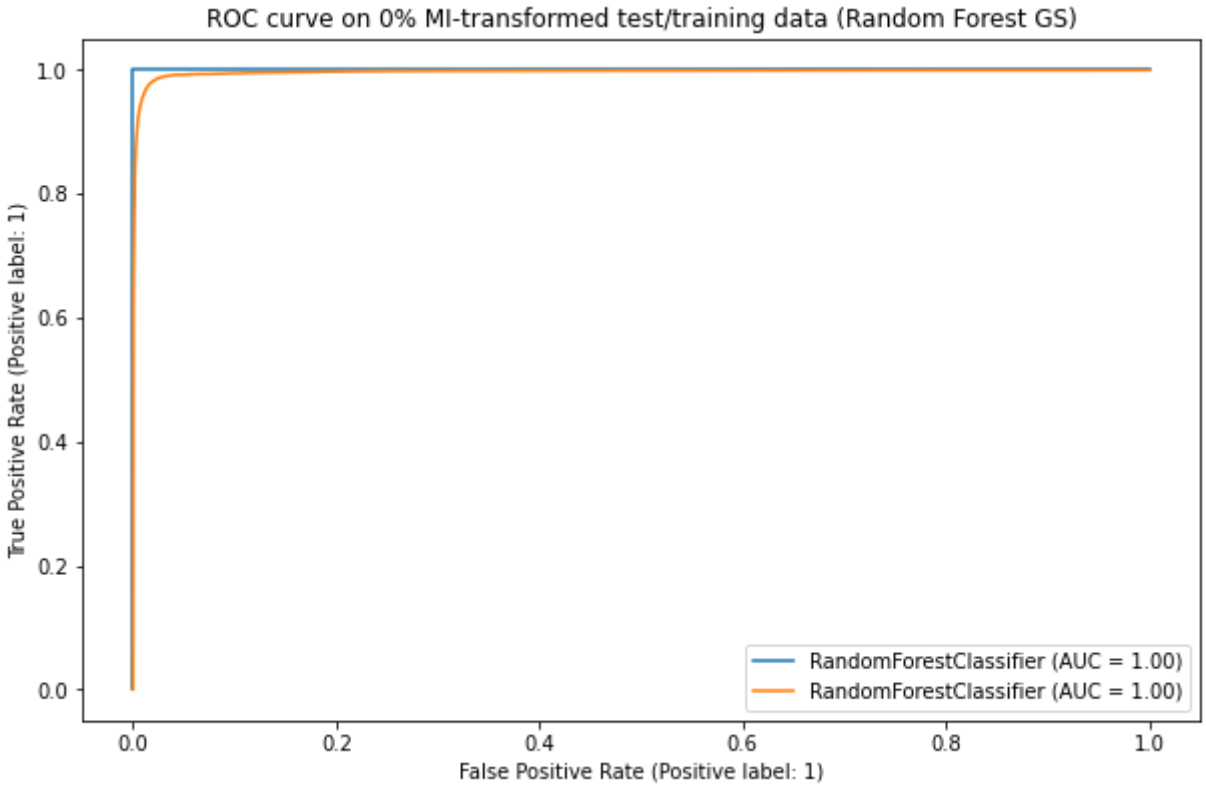
The best parameters from our grid-search are thus:

```
{'criterion': 'gini', 'max_depth': None, 'max_features': 'auto', 'n_estimators': 500}
```

We then evaluate our model with the specified parameters.

Metric	Random Forest without GS	Random Forest with GS
Accuracy	0.981578	0.981566
Precision	0.977849	0.976912
Recall	0.972282	0.973224
FPR	0.012955	0.013528

By looking at the table above, we find that random forest with grid produces a higher recall score. Hence, we proceed with the parameters.



We observe that the test/training AUCs are still ≈ 1.00.

4. Discussion & Conclusions

The performance of our model is as follows:

Metric	Random Forest
Accuracy	0.977961
Recall	0.967781
FPR	0.015939

In terms of predictive performance, our model was able to successfully predict the status (canceled or not canceled) of a booking $\approx 97.8\%$ of the time.

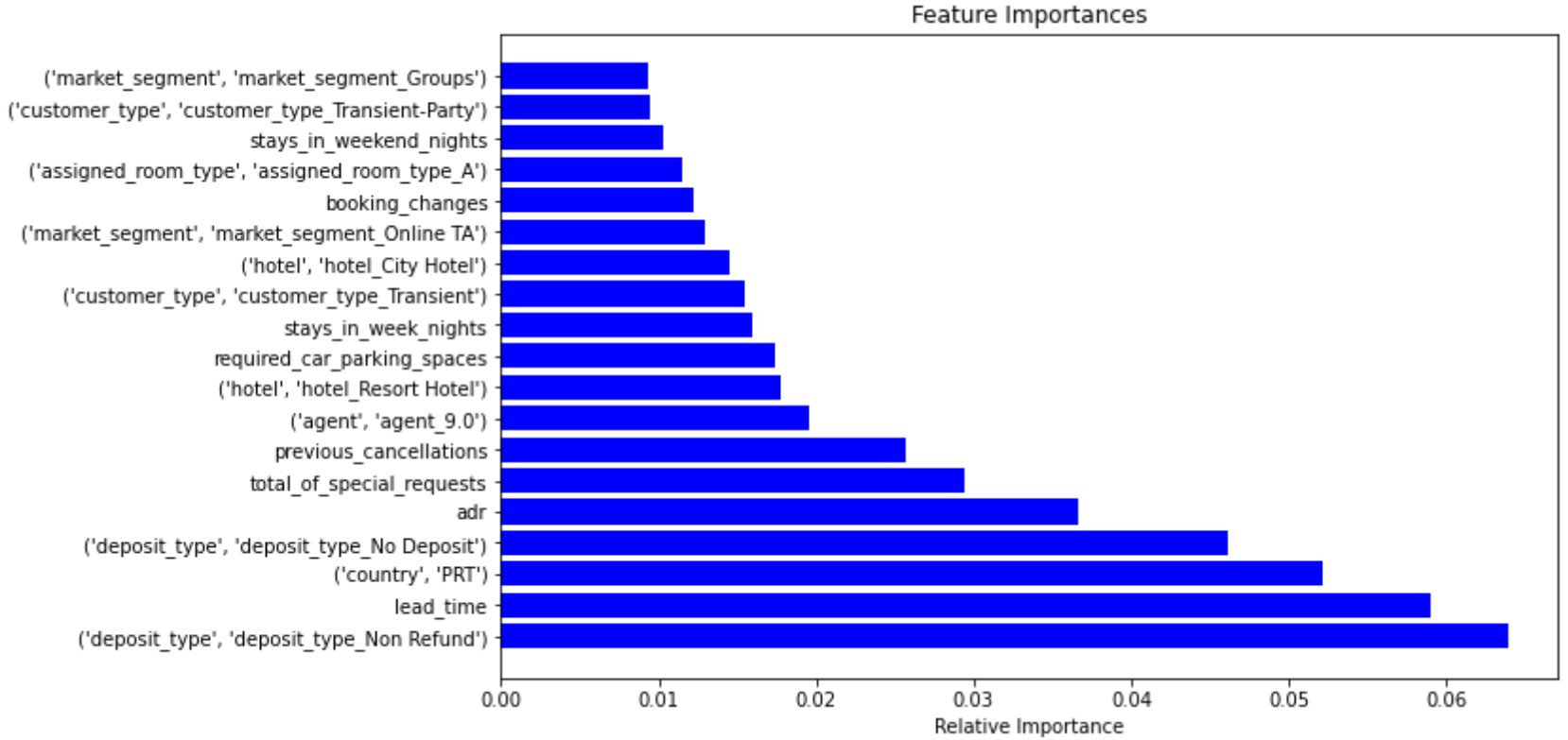
More importantly, our Recall score is ≈ 0.967781; in the context of the bookings, this implies that out of all cancelled bookings, our model is able to predict that a booking will be cancelled $\approx 96.8\%$ of the time, and not identify it as a booking that will not be cancelled (i.e. a false negative). This is key to our problem at hand as the cost associated with a cancelled booking is much higher when it goes unidentified.

Given that customer service is key in the hospitality industry, we use the False Positive Rate as a secondary metric to evaluate our model. In this context, we wish to avoid predicting that customers will cancel their bookings when they actually will not (a false positive result); doing so may alienate the customer and lower their opinion of the hotels. Our model is relatively successful in preventing such occurences, and such a false positive result would only occur $\approx 1.6\%$ of the time.

Feature Importance

We now move on to discuss the most and least important features that the company should focus on.

In the Feature Importances graph below, we display the most important features in ascending order of relative importance to the overall model.



In order to identify how they affect the probability of a booking being canceled, we examine the numerical and categorical variables separately.

Categorical Variables

For the categorical variables, we examine the conditional probabilities that a booking will be canceled given that the booking is or is not associated with a variable. If the conditional probability for when the variable has value 1 is higher than that when the value is 0, this implies that the 'presence' of the variable is likely to increase the probability of a booking being canceled, meaning that the company should pay more attention to this variable.

	Country: Portugal	Non-Refundable Deposit	No Deposit	Agent ID: 9	Transient Customer Type	Resort Hotel	City Hotel	Market Segment: Online TA	Assigned Room Type A	Market Segment: Groups
P (cancelled | var = 1)	0.231571	0.121884	0.249419	0.111583	0.306853	0.093264	0.278342	0.174525	0.276936	0.101872
P (cancelled | var = 0)	0.140034	0.249722	0.122187	0.260023	0.064753	0.278342	0.093264	0.197081	0.094670	0.269734
Difference in probabilities	0.091537	-0.127838	0.127232	-0.148440	0.242100	-0.185078	0.185078	-0.022556	0.182265	-0.167863

Country.

The only `country` variable present here is Portugal, which is the location of the two hotels as per the research article. We observe that an origin of Portugal does increase the probability of a booking being canceled.

Deposit Type.

Requiring a non-refundable deposit decreases the probability of a booking being canceled, while requiring no deposit increases the probability of a booking being canceled. This makes intuitive sense, as we can view a deposit as a sunk cost that may or may not be partially retrievable, and hence act as incentive for the booking to be not canceled.

Agent Type.

The travel agent with ID number 9 is associated with a decreased probability of a booking being canceled. This suggests that out of all agents, this agent is most reliable in terms of bookings.

Customer type.

Transient bookings are not part of a group or contract, and are not associated with other transient bookings. They are associated with an increased probability of cancelation. Transient bookings tend to be associated with walk-in guests, last minute guests, or simply people that require a very short term stay.

Hotel type.

The resort hotel is associated with a decreased probability, while the city hotel is associated with an increased probability.

Market segment.

Both 'Groups' and 'Online Travel Agent' are associated with decreased probabilities.

Assigned room type.

The room type 'A' is associated with an increased probability of cancelation. This suggests a number of causes - for example, that it may be the least desirable room type for guests.

Numerical Variables

	Weekday nights stayed	Weekend nights stayed	Booking changes	Car parking spaces	Previous cancellations	Special Requests	ADR	Lead time
Direction	Positive	Negative	Negative	Negative	Positive	Negative	Positive	Positive

In order to examine the effect of numerical variables on the probability of cancellation, we utilise the point-biserial correlation method and examine the direction of the coefficients associated with each variable. If a direction is positive, this implies that the corresponding variable contributes to an increase probability of cancellation.

Nights stayed.

We see that as the number of weekday nights stayed increases, so does the probability of cancellation; however, the probability decreases as the number of weekend nights increases.

Previous cancellations.

The probability increases as the number of previous cancellations increases. This makes some sense; non-zero previous cancellations imply that the behaviour of the customer tends towards cancelling bookings, instead of fulfilling the bookings.

ADR.

As the Average Daily Rate increases, so does the probability of cancellation. This makes sense, as the more expensive a booking is, the more likely a customer is to search for cheaper alternatives and hence cancel the booking.

Lead time.

As the lead time increases, the probability of cancellation increases. The lead time is the number of days that elapsed between the entering date of the booking into the PMS and the arrival date - we surmise that this may be due to the fact that plans are more likely to change over the longer lead time, and that more events contributing to the cancellation (such as an unforeseen accident, or personal issues) will happen over a longer period of time.

Number of booking changes.

We see that the probability decreases as the number of booking changes increases; this may imply that a customer is more invested in the booking if they have to request multiple changes to the booking to suit their changing requirements, instead of simply cancelling the booking.

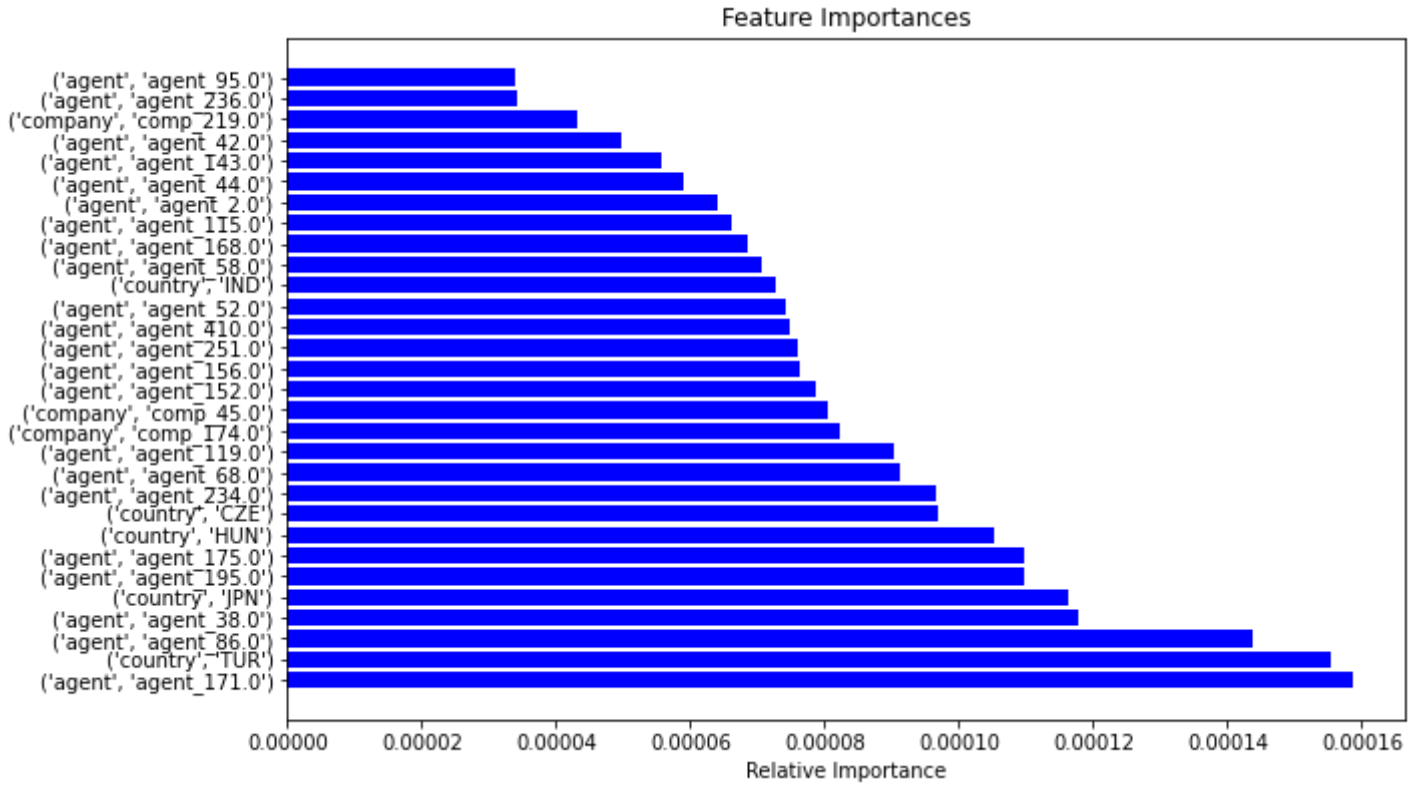
Car parking spaces requested.

The probability decreases as the number of requested car parking spaces requested increases. The rationale may be similar to the number of booking changes, as both demonstrate additional investment into the booking.

Special requests.

The probability decreases as the number of special requests increases. The rationale may be similar to the first two.

Least important



These are the least important features in our model. We note that for the most part, these features consists of travel agents or booking companies, and the rest consisting of countries such as Japan ('JPN') or Turkey ('TUR'). This implies that the large majority of agents and companies and countries have little to no effect on the rate of cancellation.

Recommendations and conclusions.

In conclusion, we recommend that the company should focus on these aspects.

Reducing cancellation rates.

In order to reduce the rate of booking cancellation, we recommend that the company should:

- Attempt to make more bookings require a non-refundable deposit
- Reduce the number of bookings that do not require a deposit
- Improve the quality of room type 'A'

- Explore ways of reducing the Average Daily Rate in order to be more competitive
- Improve customer service

The last point is more general, but in our model, variables associated with customer service, especially in the context of fulfilling requests, such as `number of special requests`, `number of bookings changes` or `car parking spaces requested`, are themselves associated with a decreased probability of cancellation. This suggests that a hotel which is able and willing to accomodate such requests will experience a decrease in cancellations.

Increasing profits.

The market segments 'Groups' and 'Online Travel Agent' are associated with decreased probability. We therefore suggest that the company should market more aggressively towards these specific segments as a way of ensuring more reliable profits (given a decreased cancellation rate).

The travel agent with ID 9 is also associated with a decreased probability. We thus recommend that the company should liaise more closely with this travel agent. However, in general, the agent or company associated with a booking does not really affect the probability; we thus advise caution when pursuing this route.

Areas of special interest.

The company should pay attention to bookings originating from Portugal, as they are associated with an increased probability of cancellation. We can only guess as to the reasons, but a possible reason is that since the hotels are themselves located in Portugal, Portuguese customers would in general have better knowledge and judgement about hotels and feasible alternatives in Portugal than customers from other countries.

The characteristics of a booking associated with an increased probability of cancellation are as follows: non-zero previous cancellations, transient, longer lead times. The company should consider any potential profits within these categories to be more unreliable than bookings without.

Lastly, the company should expect that the rate of cancellation of bookings is higher for city hotels than resort hotels.