Predicting hotel booking cancellations

Team 11

Motivation for study

Hospitality businesses have a **fixed inventory** in a way (Rajopadhye et al., 2001), since the supply of a hotel can't be adjusted with agility. Huge investments are needed if inventory needs to be increased and if a smaller inventory is wanted, property needs to be sold. World after COVID-19 restrictions is near and businesses including hotels are getting back to normal ways of operating. The nature of the hospitality business and the market situation amplifies the known fact that **forecasting tourism demand is extremely important** (Hassani et al., 2017). This calls for measures **to ensure profitability to cover for losses inflicted by restrictions** to travel and accommodation services. Hotels are also deploying more and more campaigns to acquire customers. It is paramount in this kind of market situation to **minimize the number of empty rooms** and **optimize staff and other resources** for the actual level of occupancy. To achieve this, predicting whether a booking realizes as a stay or **is the booking canceled** is a valuable tool.

Hypothesized benefits

Hotels are to benefit from predicting cancellations accurately by being able to **staff their establishment properly**. Need of housekeeping **staff per room is considered one of the largest contributions to hotel room prices** (Vives et al., 2018). Avoiding overstaffing due to bookings not realizing to stays by accurate predictions **saves the hotel costs in the form of salaries**. These aspects in mind it's feasible to think that **pricing can be optimized** if certainty about occupancy can be obtained. Other benefits to gain are **material savings** in the form of for example breakfast items. Clear benefit to be gained is first and foremost being able to **book the rooms of the accurately predicted no-shows to other customers**. With an accurate prediction model and a proactive marketing function the **predicted no-shows could also be persuaded not to cancel**. Possibilities of an accurate model are numerous, but one more worth mentioning is the possibility of **optimizing cancellation policies** so that cancellation windows and adjusting fees according to the customer class. Benefits of the customers include **better availability** of hotel rooms and **better customer experience**. The latter would be achieved by the proactive marketing functions as **risky customers could be contacted and offered additional services and discounts**. Models to predict cancellations are said to enable hotels to **loosen their cancellation policies without increasing uncertainty** (Antonio et al., 2017), but there is also a possibility to use the predictions to **tailor the policies to fit different customer classes**. If at the time of booking the customer falls into risky class, a different policy could be used for them. However, an obvious flaw in this kind of approach would be discrimination of certain customer groups.

Dataset

- From Kaggle[1]
- Contains booking information of a city hotel and a resort hotel in Portugal between 2015-2017
- 119 390 data points
- 32 different variables

Variable	Description	Values	Type	Action
hotel	Hotel type	Resort Hotel, City Hotel	Categorical	Change to Flag
is_canceled	Canceled (1) or not (0)	0, 1	Flag	Nothing
lead_time	Number of days between the booking action and the arrival date	0 to 737	Integer	Nothing
arrival_date_year	Year of arrival date	2015 to 2017	Continuous	Remove
arrival_date_month	Month of arrival date	January to December	Integer	Nothing
arrival_date_week_number	Week number of arrival date	1 to 53	Integer	Nothing
arrival_date_day_of_month	Day of arrival date	1 to 31	Integer	Nothing
stays_in_weekend_nights	Number of weekend nights stayed	0 to 19	Integer	Nothing
stays_in_week_nights	Number of week nights stayed	0 to 50	Integer	Nothing
adults	Number of adults	0 to 55	Integer	Clean up
children	Number of children	0.0 to 10.0	Float	Clean up
babies	Number of babies	0 to 10	Integer	Clean up
meal	Type of meal booked	SC, BB, HB, FB	Categorical	Encode to 0, 1, 2, 3
country	Country of origin	PRT, GBR, ...	Categorical	Removed
market_segment	Market segment designation	Online TA, Offline TA/TO, ...	Categorical	Clean up, Convert into Dummies
distribution_channel	Booking distributon channel	TA/TO, Direct, Corporate, ...	Categorical	Clean up, Convert into Dummies
is_repeated_guest	Repeated guest (1) or not (0)	0, 1	Flag	Nothing
previous_cancellations	Number of previous cancelled bookings	0 to 26	Integer	Nothing
previous_bookings_not_canceled	Number of previous not cancelled bookings	0 to 72	Integer	Nothing
reserved_room_type	Code of room type reserved	C, A, D, E, G, F, H, L, P, B	Categorical	Convert into Dummies
assigned_room_type	Code for the type of room assigned to the booking	C, A, D, E, G, F, I, B, H, P, L, K	Categorical	Remove
booking_changes	Number of changes made to the booking	0 to 21	Integer	Remove
deposit_type	Type of deposit made to the booking	No Deposit, Non Refund, Refundable	Categorical	Convert into Dummies
agent	ID of the travel agency	1.0 to 535.0	Float	Change to Flag
company	ID of the company that made the booking	8.0 to 543.0	Float	Change to Flag
days_in_waiting_list	Number of days the booking was in the waiting list	0 to 391	Integer	Nothing
customer_type	Type of booking	Transient, Contract, Group	Categorical	Convert into Dummies
adr	Average Daily Rate	-6.38 to 5400.0	Float	Nothing
required_car_parking_spaces	Number of car parking spaces required by the customer	0 to 8	Integer	Nothing
total_of_special_requests	Number of special requests	0 to 5	Integer	Nothing
reservation_status	Reservation's last status	Check-Out, Canceled, No-Show	Categorical	Remove
reservation_status_date	Date at which the last status was set	2015-10-17 to 2017-09-14	Categorical	Remove

1. https://www.kaggle.com/jessemostipak/hotel-booking-demand

Data Quality & Preparation

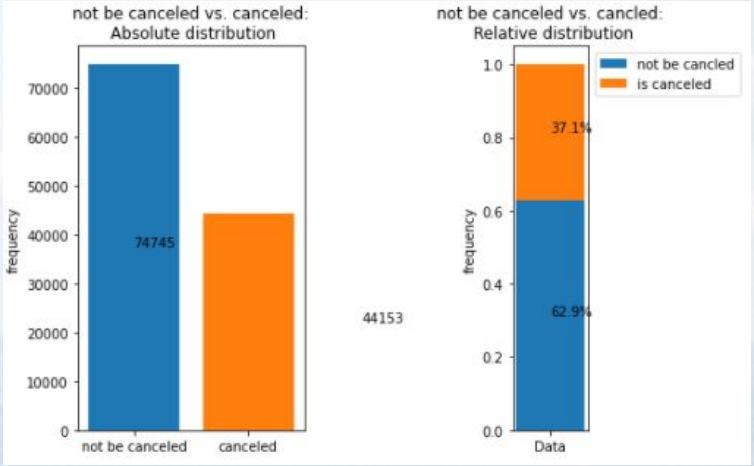
- Missing values:
  - country: 488
  - agent: 16 340
  - company: 112 593

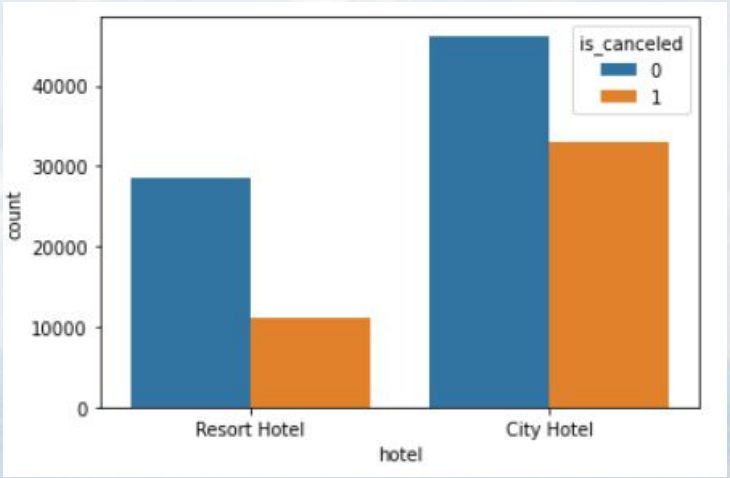
We did the following to prepare the data for the modelling:

- Error observations removed:
  - e.g. entries with 0 adults, 0 children and 0 babies
- Some columns were also excluded to make the model more applicable and to prevent data leakage
  - e.g. country, reservation status, arrival date year, booking changes
- Agent and company columns changed so that they only have values 1 or 0 based on whether a agent/company was associated with the booking
- Most categorical values were converted into dummies

After the data preparation, we had 119 205 data points and 50 features.

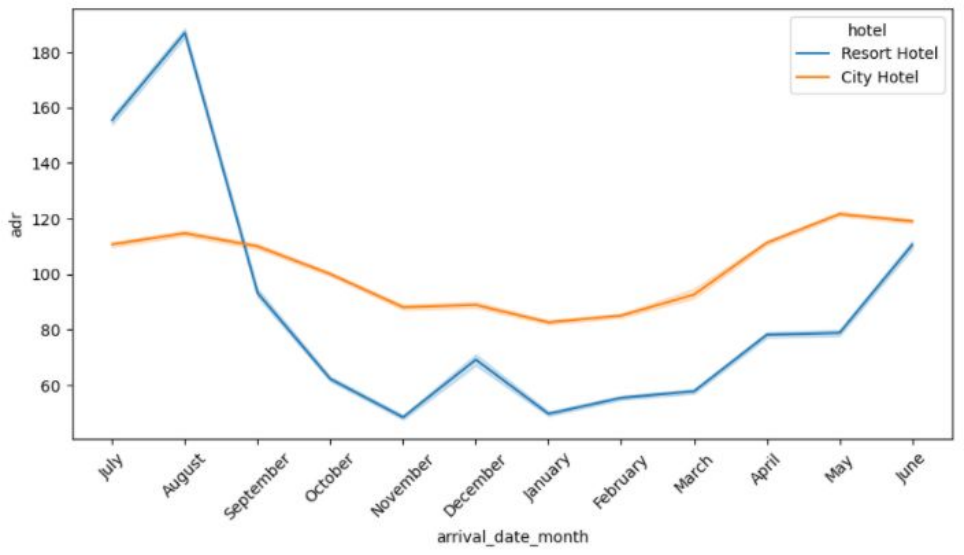
Descriptive data analysis

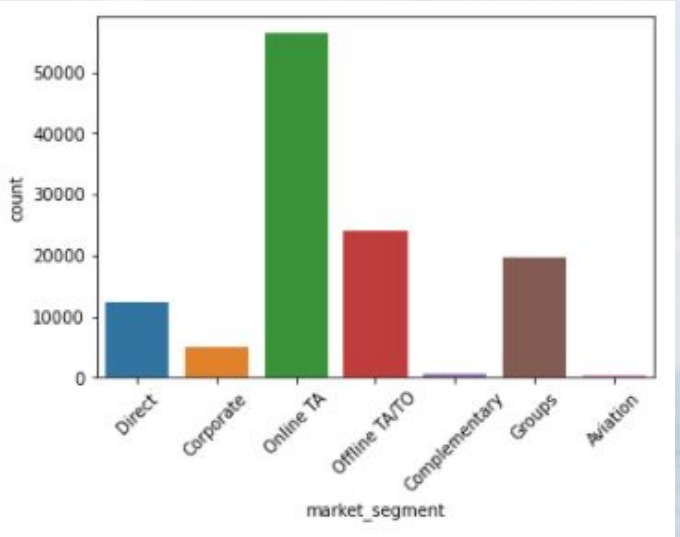




- The cancelation rate is about 37.1%, out of 119390 observations
- The City Hotel has a larger amount of booking order and cancelation, compared with Resort Hotel
- The City Hotel has cancelation rate with 41.73%, while the number of alternative is just 27.76%. Hence, the cancelation rate of City hotel is around 1.5 times higher than the Resort Hotel.

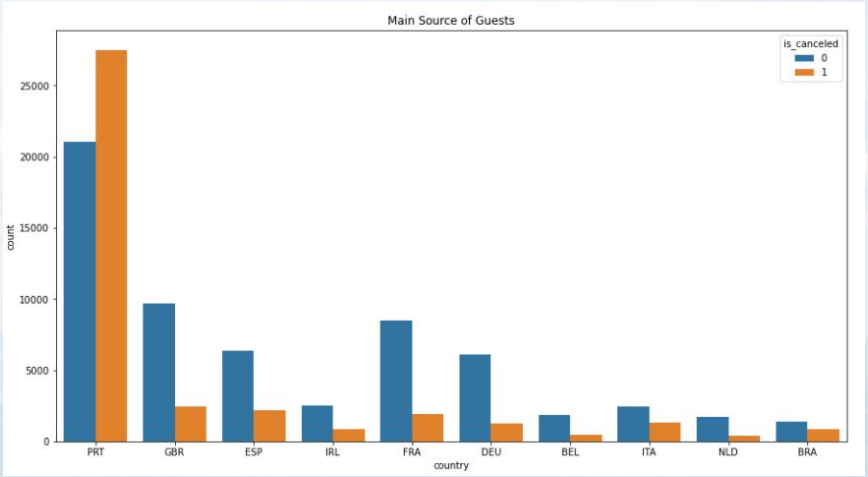
Descriptive data analysis

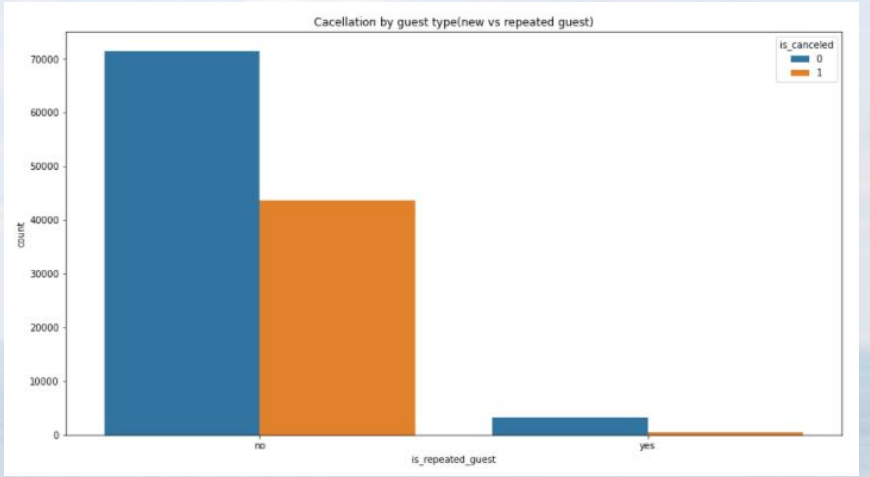




- Resort Hotel is popular in June, July and August. It implies that most of the customers prefer to go traveling in summer. Another booking peak time is in Christmas. The order for the Resort Hotel is depended on holiday and weather.
- Compared with Resort Hotel, City Hotel has steady booking volume during the year. Yet, as it has a stable volume and relatively higher canceltation rate, our cancelation prediction model is handy for saving the cost for City Hotel, especially in in May, when the company receive most orders.
- Among all different channels, most customers ordered booking online. Thus, the company can attract new customers and retain customers by giving a seamless ordering experience with intuitive user interface.

Descriptive data analysis





- Portugal is the main source of getting guests for the hotel, but it also has the highest cancelation rate, with 56.63%
- The UK, Spain and France also have relatively high cancellation volume
- The cancelation rate is associated with customers order history. The repeated customers has a smaller cancelation rate, comparing with new customers. The ratio of old customers is about 14.49%, while the one of new customers is 37.79%. Thus, it is important for the company to retain customers.

Modeling

We decided to use four different machine learning methods to for modeling:

- Support Vector Machine (SVM), SGD
- Logistic Regression
- Random Forest
- Decision Tree

These models were chosen because we had some experience and knowledge in using them. Also, we wanted to use some basic and well-known classification methods, which could be used as a benchmarks for further research. Because of the huge amount of training data, SVM model turned out to be really slow to compute, so we decided to instead use Stochastic Gradient Descent (SGD) with modified huber loss. Mostly default parameters were used in all of the models.

The dataset was splitted into three parts; training data (80.10%), validation data (9.90%) and testing data (10.00%). The models were trained using the training data. After that, they were evaluated using the validation data and some of the parameters were changed based on those results. Finally, one model is chosen to be enhanced and at the end tested with the testing data. This kind of process is done to avoid overfitting the model when optimizing it with the training and validation data. This three-way data split is possible because of the large amount of data we can still use for training the models.

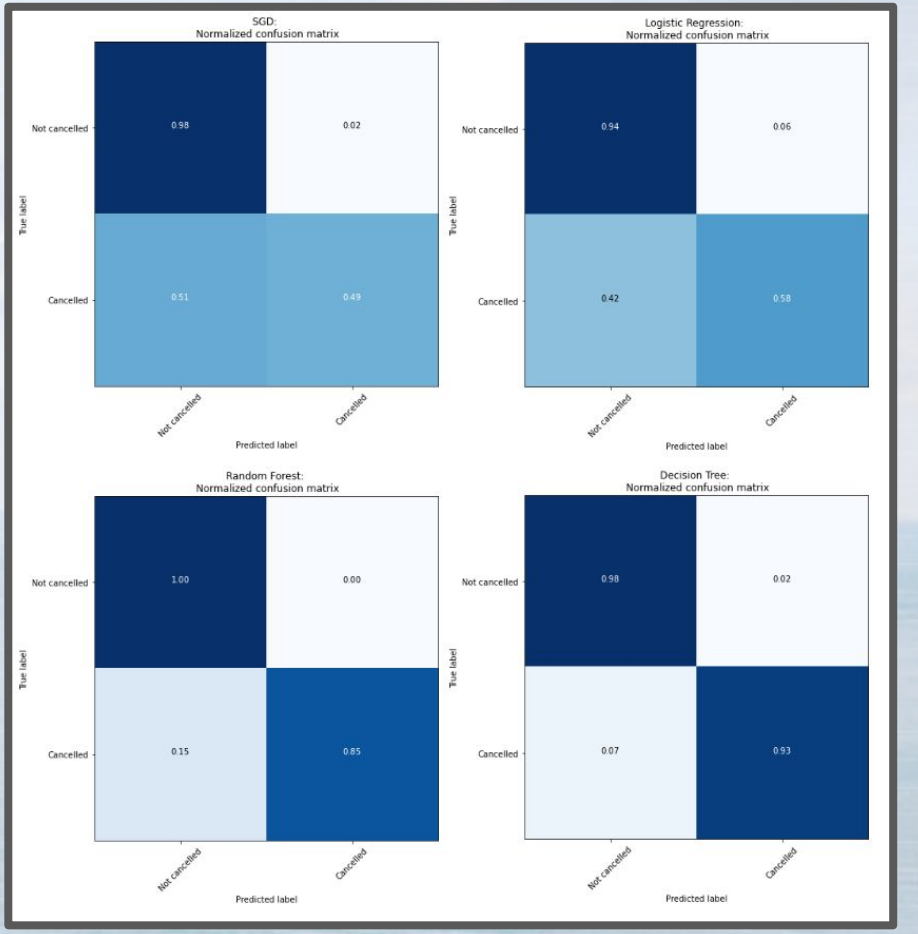
Evaluation of the models

The four different models were evaluated using *accuracy_score* function from sklearn. Because our problem is binary classification this function is equal to the *jaccard_score* function.

Accuracy scores on the validation data:

- SGD: 80.07%
- Logistic Regression: 80.81%
- Random Forest: 94.41%
- Decision Tree: 95.77%

Also, we computed the normalized confusion matrices for the models. As can be seen from the results, the tree based models (Random Forest and Decision Tree) clearly outperformed SGD and Logistic regression, which had hard time predicting positive cases. This behaviour can be seen as light blue colors in the bottom half of the matrices. However, all of the models had quite good true negative rates.



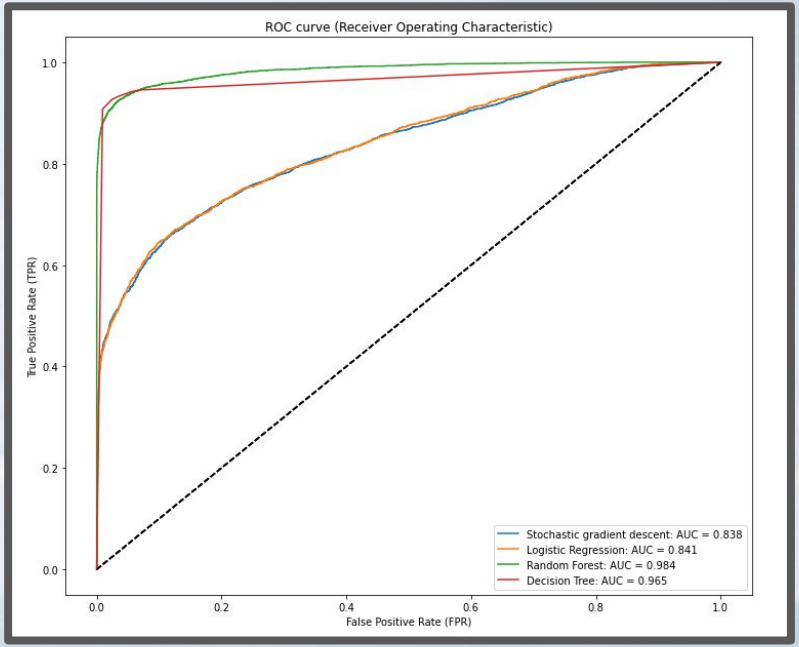
Evaluation of the models

The performance difference between the tree based models and the other two, is also clearly seen in ROC curve and AUC values. SGD and Logistic regression are behaving very similarly.

After training the Random Forest, we decided to also see what features had the highest impact on the model. This was done by calculating feature importance which is based on the mean of the impurity decrease within each tree for every feature. (Appendix 1)

Based on the results and already known correlations (Appendix 1) we can conclude that for example:

- High lead time and adr (average daily rate) and non refund deposit type had significant positive impact on the probability of cancellation
- No deposit and high number of special requests and required car parking spaces most of the time led to not cancelling




```
Feature: dep_type_Non Refund            Importance: 0.107
Feature: res_status_month               Importance: 0.103
Feature: dep_type_No Deposit            Importance: 0.097
Feature: lead_time                      Importance: 0.089
Feature: arrival_date_week_number       Importance: 0.075
Feature: res_status_day                 Importance: 0.065
Feature: arrival_date_day_of_month      Importance: 0.056
Feature: arrival_date_month             Importance: 0.055
Feature: total_of_special_requests      Importance: 0.051
Feature: previous_cancellations         Importance: 0.05
Feature: adr                            Importance: 0.045
Feature: required_car_parking_spaces    Importance: 0.023
Feature: mar_seg_Online TA              Importance: 0.023
Feature: stays_in_week_nights           Importance: 0.018
Feature: cust_type_Transient            Importance: 0.018
Feature: mar_seg_Groups                 Importance: 0.016
Feature: mar_seg_Offline TA/TO          Importance: 0.014
Feature: cust_type_Transient-Party      Importance: 0.013
Feature: stays_in_weekend_nights        Importance: 0.011
Feature: dist_ch_TA/TO                  Importance: 0.009
```

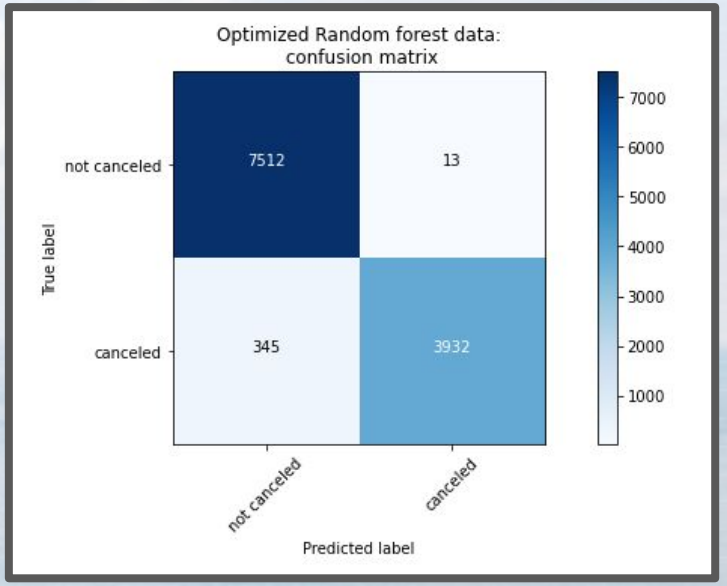
Choosing one model

Random Forest showed and Decision Tree models showed the most promising results. We decided to optimize the **Random Forest model** as that could be more susceptible to improvement by being more complex method.
Based on the feature importances, top 15 features were used to train the Random Forest again. This would make the model simpler by removing features which had little to now effect on the performance. After simplifying the model, we got 95.97% accuracy on the validation data, so 1.56% improvement on the base model.

After that, we decided to use grid search model to find better parameters for the model. Parameters used in the grid were:

- min_samples_split
- min_samples_leaf
- n_estimators

This gave us 96.97% accuracy on the validation data, so 1% improvement on the previous model. We also got 96.65% accuracy on the testing data which means the model is not overfitted and should be quite good on unseen data. It is good to note that small improvements in the model could have a high impact in a business application especially when dealing with large amounts of data.



Confusion matrix of the optimized Random forest model with validation data

Costs of false predictions

When discussing business implications of our models and predictions provided by them, it is assumed that we are consulting a highly popular hotel chain with nearly 100% occupancy rate. With this in mind we assume that if a booking is canceled, another customer will want to book the vacant room.

- Overbooking → Compensation to customer
  - Caused by falsely predicting that a customer will cancel
  - If our model is proved to be as accurate as it has been in our tests, this cost, being a large sum per instance, will not be significant, because the incorrect prediction of cancellation is so rare.
- Overstaffing and excess materials → Salaries to employees and material costs (e.g., breakfast ingredients)
  - Caused by falsely predicting that a booking will not be canceled
  - The share of the salary of cleaning, reception and kitchen staff per room is fairly low

Benefits of true predictions

- Optimization off staff and other resources → Savings
  - When the future occupancy of the hotel is accurately predicted, level of staff is optimal and the right amount of salaries are paid
  - No food goes to waste
  - Employees' well being is improved by increased certainty and customer experience is improved caused by satisfied workforce
- Persuading the customer not to cancel or selling the to-be-canceled room to another customer → Full price of booking
  - If it is correctly predicted that the booking is to be canceled, the customer can be contacted and persuaded to come.
  - If the persuasion is not successful, the room can be sold to someone else.
  - Since the hotel chain we are consulting has such a high demand, the room predicted to be vacant even if booked can be sold to another customer with full price.

Monetary benefits of chosen model

- Average value of booking in target hotel: 200€
- Compensation to customer in overbooking situations according to hotel's terms: 300€
- Average share of work and material costs per booking: 20€

- At the cancellation rate of the validation dataset (36%) expected monetary benefit of the chosen model (optimized random forest) is 65,72€ per customer (Appendix 6)
    - Cost/benefit analysis of each explored model are in appendices (Appendix 2 - 6)

Model	Expected benefit per customer
Random Forest	59,95 €
Optimized Random Forest	65,72 €
Decision Tree	61,92 €
SGD	27,42 €
Logistic regression	27,22 €

Conclusion

- By analyzing the data, we found out that two different types of hotels have different demands fluctuation during the year. Thus, during the peak time, the company can improve the sales and reduce the cost if they apply our predictive model.
- Based on the calculation, it can be predicted that the expected monetary benefit of the chosen model (optimized random forest) is 65,72€ per customer.
- The model we developed is based on Random Forest method, as it provides high accuracy score, better AUC and ROC results, and more susceptible to improvement by being more complex method.
- In practise, the company can save costs by optimizing staff and other resources. Besides, the company may take actions to retain customers who could be potentially cancelled the booking order by the prediction.
- Yet, despite the fact that the model gave up satisfying results, we can still improve the model by separating the hotel types to predict the cancelation rate. Thus, we can have 2 different prediction models for the resort hotel and city hotel, as they have slightly different customer demand.

Appendices

Appendix 1
Sorted correlation matrix values and feature importances

Correlation matrix values for numerical features

lead_time	0.292930
total_of_special_requests	0.234917
required_car_parking_spaces	0.195684
booking_changes	0.144824
previous_cancellations	0.110147
is_repeated_guest	0.083741
company	0.083589
adults	0.058157
previous_bookings_not_canceled	0.057364
days_in_waiting_list	0.054308
agent	0.046748
adr	0.046559
babies	0.032567
stays_in_week_nights	0.025551
arrival_date_year	0.016684
arrival_date_week_number	0.008299
arrival_date_day_of_month	0.005910
children	0.004877
stays_in_weekend_nights	0.001309

Correlation matrix values for categorical features

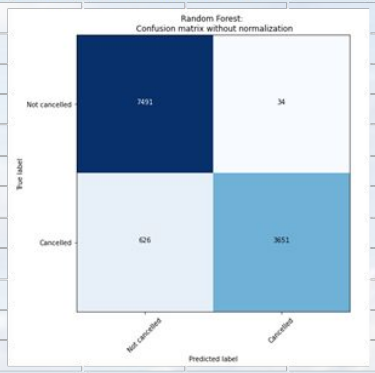
dep_type_Non Refund	0.481537
dep_type_No Deposit	0.477988
mar_seg_Groups	0.222014
dist_ch_TA/TO	0.176107
mar_seg_Direct	0.154384
dist_ch_Direct	0.151574
hotel	0.137049
cust_type_Transient	0.133325
cust_type_Transient-Party	0.124370
mar_seg_Corporate	0.081639
dist_ch_Corporate	0.075583
res_room_A	0.069197
res_status_month	0.068573
res_room_D	0.047722
mar_seg_Complementary	0.040329
res_room_E	0.038874
cust_type_Group	0.038841
res_status_day	0.033997
mar_seg_Offline TA/TO	0.028654
cust_type_Contract	0.023683
res_room_F	0.021773
dist_ch_GDS	0.014927
mar_seg_Aviation	0.013754
dep_type_Refundable	0.011344
arrival_date_month	0.011160
res_room_B	0.008839
res_room_C	0.007333
mar_seg_Online TA	0.006219
res_room_H	0.005439
meal	0.003502
res_room_G	0.001667
res_room_L	0.000550

Feature importances

Feature: dep_type_Non Refund	Importance: 0.107
Feature: res_status_month	Importance: 0.103
Feature: dep_type_No Deposit	Importance: 0.097
Feature: lead_time	Importance: 0.089
Feature: arrival_date_week_number	Importance: 0.075
Feature: res_status_day	Importance: 0.065
Feature: arrival_date_day_of_month	Importance: 0.056
Feature: arrival_date_month	Importance: 0.055
Feature: total_of_special_requests	Importance: 0.051
Feature: previous_cancellations	Importance: 0.05
Feature: adr	Importance: 0.045
Feature: required_car_parking_spaces	Importance: 0.023
Feature: mar_seg_Online TA	Importance: 0.023
Feature: stays_in_week_nights	Importance: 0.018
Feature: cust_type_Transient	Importance: 0.018
Feature: mar_seg_Groups	Importance: 0.016
Feature: mar_seg_Offline TA/TO	Importance: 0.014
Feature: cust_type_Transient-Party	Importance: 0.013
Feature: stays_in_weekend_nights	Importance: 0.011
Feature: dist_ch_TA/TO	Importance: 0.009
Feature: adults	Importance: 0.007
Feature: hotel	Importance: 0.007
Feature: meal	Importance: 0.007
Feature: dist_ch_Direct	Importance: 0.005
Feature: previous_bookings_not_canceled	Importance: 0.004
Feature: agent	Importance: 0.004
Feature: children	Importance: 0.003
Feature: company	Importance: 0.003
Feature: days_in_waiting_list	Importance: 0.003
Feature: mar_seg_Direct	Importance: 0.003
Feature: res_room_A	Importance: 0.003
Feature: is_repeated_guest	Importance: 0.002
Feature: mar_seg_Corporate	Importance: 0.002
Feature: res_room_D	Importance: 0.002
Feature: cust_type_Contract	Importance: 0.002
Feature: dist_ch_Corporate	Importance: 0.001
Feature: res_room_E	Importance: 0.001
Feature: res_room_F	Importance: 0.001
Feature: babies	Importance: 0.0
Feature: mar_seg_Aviation	Importance: 0.0
Feature: mar_seg_Complementary	Importance: 0.0
Feature: dist_ch_GDS	Importance: 0.0
Feature: res_room_B	Importance: 0.0
Feature: res_room_C	Importance: 0.0
Feature: res_room_G	Importance: 0.0
Feature: res_room_H	Importance: 0.0
Feature: res_room_L	Importance: 0.0
Feature: dep_type_Refundable	Importance: 0.0
Feature: cust_type_Group	Importance: 0.0

Appendix 2
Classifier 1 cost/benefit analysis

CLASSIFIER 1: Random forest



PYTHON OUTPUT

		Predicted	
Partition = Testing		0	1
Actual	0	7 491	34
	1	626	3651

Confusion matrix (reformulated)

		Actual class			Estimated probabilities		
		p	n			p	n
Predicted class	Y	3651	34	3685	Y	31 %	0 %
	N	626	7 491	8117	N	5 %	63 %
		4277	7525	11802			

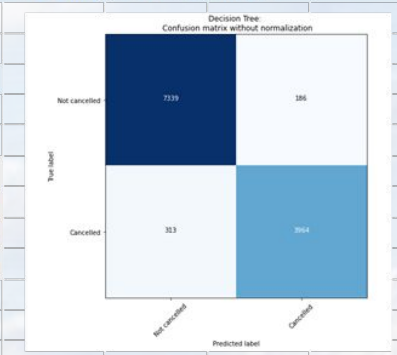
Cost-benefit information

		Actual class		Test-set priors		Conditional probabilities	
		p	n				
Predicted class	Y	200,00 €	-300,00 €	P(p)	36 %	P(Y|p)	85 %
	N	-20,00 €	0,00 €	P(n)	64 %	p(N|p)	15 %
						p(Y|n)	0 %
						p(N|n)	100 %

Expected benefit with test set priors = **59,95 €** (per customer)

Appendix 3
Classifier 2 cost/benefit analysis

CLASSIFIER 2: Decision tree



PYTHON OUTPUT

		Predicted	
Partition = Testing		0	1
Actual	0	7 339	186
	1	313	3964

Confusion matrix (reformulated)

		Actual class			Estimated probabilities		
		p	n			p	n
Predicted class	Y	3964	186	4150	Y	34 %	2 %
	N	313	7 339	7652	N	3 %	62 %
		4277	7525	11802			

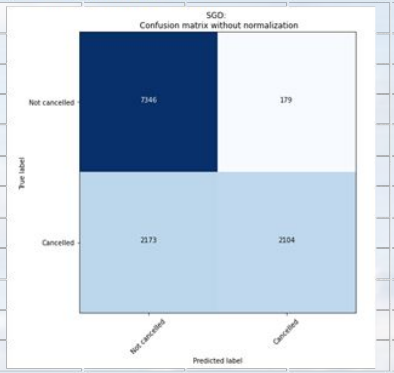
Cost-benefit information

		Actual class		Test-set priors		Conditional probabilities	
		p	n	P(p)	36 %	P(Y|p)	93 %
Predicted class	Y	200,00 €	-300,00 €	P(n)	64 %	p(N|p)	7 %
	N	-20,00 €	0,00 €			p(Y|n)	2 %
						p(N|n)	98 %

Expected benefit with test set priors = **61,92 €** (per customer)

Appendix 4
Classifier 3 cost/benefit analysis

CLASSIFIER 3: SGD



PYTHON OUTPUT

		Predicted	
Partition = Testing		0	1
Actual	0	7 346	179
	1	2173	2104

Confusion matrix (reformulated)

		Actual class		
		p	n	
Predicted class	Y	2104	179	2283
	N	2173	7 346	9519
		4277	7525	11802

		Estimated probabilities	
		p	n
Y		18 %	2 %
N		18 %	62 %

Cost-benefit information

		Actual class	
		p	n
Predicted class	Y	200,00 €	-300,00 €
	N	-20,00 €	0,00 €

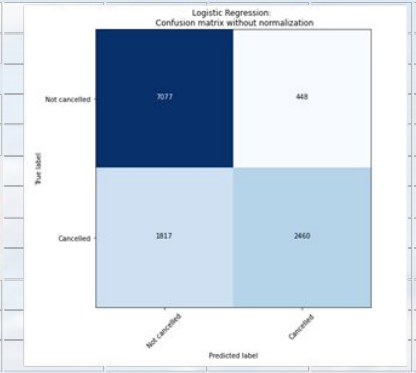
Test-set priors	
P(p)	36 %
P(n)	64 %

Conditional probabilities	
P(Y|p)	49 %
p(N|p)	51 %
p(Y|n)	2 %
p(N|n)	98 %

Expected benefit with test set priors = **27,42 €** (per customer)

Appendix 5
Classifier 4 cost/benefit analysis

CLASSIFIER 4: Logistic regression



PYTHON OUTPUT

		Predicted	
Partition = Testing		0	1
Actual	0	7 077	448
	1	1817	2460

Confusion matrix (reformulated)

		Actual class			Estimated probabilities		
		p	n			p	n
Predicted class	Y	2460	448	2908	Y	21 %	4 %
	N	1817	7 077	8894	N	15 %	60 %
		4277	7525	11802			

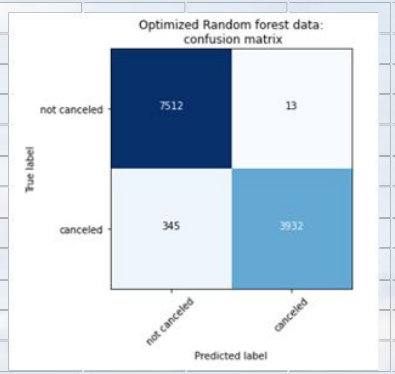
Cost-benefit information

		Actual class		Test-set priors		Conditional probabilities	
		p	n	P(p)	36 %	P(Y|p)	58 %
Predicted class	Y	200,00 €	-300,00 €	P(n)	64 %	p(N|p)	42 %
	N	-20,00 €	0,00 €			p(Y|n)	6 %
						p(N|n)	94 %

Expected benefit with test set priors = **27,22 €** (per customer)

Appendix 6
Classifier 5 cost/benefit analysis

CLASSIFIER 5: Random forest (Optimized)



PYTHON OUTPUT

		Predicted	
Partition = Testing		0	1
Actual	0	7 512	13
	1	345	3932

Confusion matrix (reformulated)

		Actual class		
		p	n	
Predicted class	Y	3932	13	3945
	N	345	7 512	7857
		4277	7525	11802

		Estimated probabilities	
		p	n
	Y	33 %	0 %
	N	3 %	64 %

Cost-benefit information

		Actual class	
		p	n
Predicted class	Y	200,00 €	-300,00 €
	N	-20,00 €	0,00 €

Test-set priors	
P(p)	36 %
P(n)	64 %

Conditional probabilities	
P(Y|p)	92 %
p(N|p)	8 %
p(Y|n)	0 %
p(N|n)	100 %

Expected benefit with test set priors = **65,72 €** (per customer)

References

Antonio, N., De Almeida, A., & Nunes, L. (2017). Predicting hotel booking cancellations to decrease uncertainty and increase revenue. *Tourism & Management Studies*, 13(2), 25-39.

Hassani, H., Silva, E. S., Antonakakis, N., Filis, G., & Gupta, R. (2017). Forecasting accuracy evaluation of tourist arrivals. *Annals of Tourism Research*, 63, 112-127.

Rajopadhye, M., Ghalia, M. B., Wang, P. P., Baker, T., & Eister, C. V. (2001). Forecasting uncertain hotel room demand. *Information sciences*, 132(1-4), 1-11.

Vives, A., Jacob, M., & Payeras, M. (2018). Revenue management and price optimization techniques in the hotel sector: A critical literature review. *Tourism economics*, 24(6), 720-752.