NOVA
IMS
Information
Management
School

BUSINESS CASES WITH DATA SCIENCE

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS – MAJOR IN BUSINESS ANALYTICS

Business Case #2 – Hotel Booking Cancellations

D4B
Consulting

Data4Business Consulting

Débora Santos, number: 20200748

Diana Furtado, number: 20200590

Pedro Henrique Medeiros, number: 20200742

Rebeca Pinheiro, number: 20201096

March, 2021

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

INDEX

1. INTRODUCTION 1
2. BUSINESS UNDERSTANDING 1
    2.1. Background..1
    2.2. Business Objectives 1
    2.3. Business Success criteria 2
    2.4. Situation assessment...2
        2.4.1. Inventory of resources 2
        2.4.2. Requirements, assumptions, and constraints...2
        2.4.3. Risks and contingencies...2
    2.5. Determine Data Mining goals...3
    2.6. Project Plan..3
3. Predictive Analysis...4
    3.1. Data understanding..4
    3.2. Data preparation 5
    3.3. modelling and evaluation...6
4. RESULTS EVALUATION..8
5. DEPLOYMENT AND MAINTENANCE PLANS 9
6. CONCLUSIONS 9
7. REFERENCES..9

1. INTRODUCTION

Thank you for choosing **Data4Business Consulting (D4B)** to help you with the challenge of better understanding your customers characteristics. Our main objective is helping Hotel Chain C to identify high cancellation likelihood bookings and consequently increase your gains and customers' satisfaction.

The world is experiencing a great technological and digital revolution where understanding business data, customers and their needs is essential for the business success. The exponential technological advances, such as data mining techniques, artificial intelligence, internet of things, can help taking the business to the next level.

Through innovative technological programs, well-referenced data mining methods and insights of digital marketing, the present report intends to provide an overview of the process behind the analysis, presents the results and insights you need to be successful in this new era.

In addition to the present report, the following deliverables will be submitted:

- Outcomes presentation to C.
- Jupyter Notebook with the code of the entire process.

All files can be accessed in Github:

https://github.com/Debs86/Business_Cases_Projects/tree/main/BC2

We are excited to take part of this challenge.

2. BUSINESS UNDERSTANDING

2.1. BACKGROUND

Hotel chain C (C) has 2 hotels in Portugal: One is a resort located at the region of Algarve (H1) and the other one is a city hotel located at the city of Lisbon (H2).

Currently, customers can book through Travel Agencies, Tour Operators, Corporate or Directly with the hotel. The advancement of the internet not only brought more exposure, but also more competitiveness. With the appearance of the online travel agencies (OTAs), the number of the "deal-seeking" customers have grown immensely. "Deal-seeking" customers tend to make multiple bookings for the same trip to find the best deal. Consequently, it increases cancellations in the hotels.

C was severely impacted by cancellations, representing almost 28% of the bookings in H1 and nearly 42% in H2. The Revenue Manager Director of C, Michel, has already implemented several approaches to reduce cancellations, with no significant improvement.

Michael wants to implement prediction models to allow the chain's hotels to forecast net demand based on reservations on-the-books, more specifically in H2, therefore he has reached D4B.

2.2. BUSINESS OBJECTIVES

The customer's primary objective is to implement prediction models to forecast net demand based on reservations on-the-books. With these models, the customer expects:

- To implement better price and overbooking policies.
- To identify high cancellation likelihood bookings.
- To implement actions to prevent cancellation.
- To reduce cancellations to a rate of 20%.

2.3. BUSINESS SUCCESS CRITERIA

The expected outcome will be the development of a predictive model to forecast the net demand based on the bookings. The success of the proposed task will be evaluated by C revenue manager director and, if needed, we will go back to the model until we get an outcome that matches with his expectation.

2.4. SITUATION ASSESSMENT

2.4.1. Inventory of resources

This project was made following the CRISP-DM reference model (Cross Industry Standard Process for Data Mining). CRISP-DM is a standard process built in the end of 90's and it was built by more than 200 members lead by a consortium of big companies. *CRISP-DM succeeds because it is soundly based on the practical, real-world experience of how people conduct data mining projects.*[1]

This project has the support of C's Management and team.

On the D4B Consulting side, this project will be conducted by a team of 4 Data Scientists and Business Analysts.

We have been provided by the C team with a dataset of H2 resort bookings, of customers due to arrive between July 1, 2015 and August 31, 2017. Along with this dataset, we were also provided with its metadata file.

The main technology used to achieve the objectives of this report was Python. Python is one of most important and commonly used program languages in data science projects.

2.4.2. Requirements, assumptions, and constraints

The completion date of the present phase of the project is March 15, 2021, but we expect to continue giving support and helping C to achieve the next goals for the growth of the business.

Even if the booking has as *Reservation Status*, *No-Show* and the *Deposit Type* is *No deposit*, the customers will be charged on their credit card.

Some bookings provided seems to be duplicated, but we are going to disregard that and keep them to test the model.

2.4.3. Risks and contingencies

Table 2.1 identifies a list of risks and contingency proposed.

Risk	Contingency
Insufficient number of features	Work with remaining features or ask for different variables
Bookings with very similar characteristics	Split the data (train and test) avoiding bias.
Model overfitting	Ask for more observations (bookings)

Table 2.1 - Risks and contingency.

2.4.3.1. Terminology

Business glossary

- "Deal-seeking" customers: tend to make multiple bookings for the same trip to find the best deal.
- Net demand is defined as demand minus cancellations.

Data mining glossary

- Classification problem: In this type of problem, the objective in implementing machine learning techniques and algorithms is to predict a class of the target variable. On this specific problem, the variable *IsCanceled* is the target. Also, it is a supervised problem when the data provided already included desired outputs.
- Accuracy: A rate between the true outputs against the total. In other words, the proportion of correctly predicted cancellations or no cancelations against the total of bookings. $Accuracy = \frac{True\ Positives\ (TP)+True\ Negatives(TN)}{Positives\ (P)+\ Negatives\ (N)}$
- Precision: measure how precise the model is out of those predicted positives. The proportion of correctly predicted cancelations against the proportion of total correctly predictions. $Precision = \frac{True\ Positives}{True\ Positives+Tr\quad Negatives}$
- Recall: how many actual positives, the model captures as being positive. The proportion of correctly predicted cancelations against the total of actual cancelations. $Recall = \frac{TP}{TP+}$
- F1 Score [3]: Measure that represents a balance between recall and precision. This is measure of model quality is especially important for this analysis as it represents the balance between the false positive (precision) and false negative (recall). In this case, False positive is about to predict that a customer will cancel when, in reality, he will not cancel the booking. This kind of situation generates problems to the hotel such as: 1) overbooking and consequently extra expenses with rent a room in another hotel and reviews. 2) costs offering services to avoid a cancel that it would not happen. On the other hand, the false negative is about to predict that a booking will not be cancelled when, in reality, it will be cancelled. This kind of situation generates lower revenue to the hotel because the room will be empty. $F1\ score = 2 \times \frac{Precision \times Recall}{Precision+Reca}$
- AUC - ROC Curve (Area Under Receiver Operating Characteristics) [2]: This is a performance measurement in which ROC is a probability curve and AUC represents the measure of separability. It shows how capable the model is of distinguish cancellations and no cancellations. The higher the AUC is, the better the model is at predicting the true positives and negatives.

2.5. Determine Data Mining goals

The data mining goals states project objectives in technical terms:

1. Create a model that will be able to predict the probability of new bookings be cancelled or not.
   <u>*Success criteria*</u>: High percentages of accuracy, precision, F1-score and AUC.
2. Understand the main characteristics of the bookings.
   <u>*Success criteria*</u>: Give some insights about the characteristics of the bookings helping to identify if it has or not a high probability to be cancelled.

2.6. Project Plan

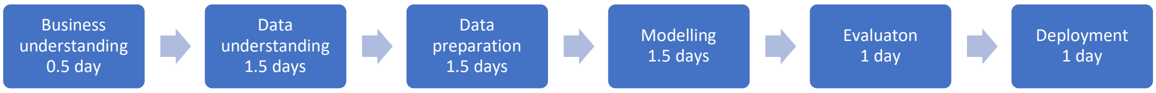
Business understanding 0.5 day
Data understanding 1.5 days
Data preparation 1.5 days
Modelling 1.5 days
Evaluaton 1 day
Deployment 1 day

Figure 2.1 - Project's timeline.

Resources wise, for the business understanding we plan to use all the information provided in the kickoff meeting's presentation. For the core stages of the project, we plan to use Python to work the data provided. To present the results, we expect to use Word for the report and Power point for the presentation.

The performance of the model will be directly connected with the quality of the input data. For this reason, we identify the Modelling stage as dependent of the Data preparation stage. During the project, we must go and back between Data preparation and Modelling many times, repeat this iteratively until we get the desired outcome.

For the Modelling stage we aim to build a supervised model (predictive) using Random Forecast classifier algorithm. We opted for this model because it presented the best results compared with other algorithms. Further details will be presented in section 3.3 – Modelling and evaluation. The model evaluation will be made using accuracy, precision and F1 score metrics. The accuracy metric is good to see the overall performance of the model. Precision to measure how good the model is predicting the cancellations. Finally, F1 to check the balance between recall and precision. It will also be presented a confusion matrix and an ROC curve.

3. PREDICTIVE ANALYSIS

In this section we go through the process of understanding and preparing the data for modelling, the modelling itself, the different algorithms used and, finally, the results evaluation.

3.1. DATA UNDERSTANDING

At this stage we analyzed the dataset to understand its potential and limitations. We have used the Pandas profiling to have an overview of the dataset: what variables are in the dataset, what they mean, number of variables (31 features, from which 13 categorical and 18 numerical as shown on Table 3.1) and observations (79.330 customers), how the variables are distributed (there are some skewed variables) , if there is noise, if there are missing values (28 missing entries) and/or duplicated values (none) , which of these features are relevant for the final goal and which features are redundant.

We have also looked at the metadata file provided to understand the meaning of each feature to understand their relevancy in the project.

Numeric	Categorical
ADR, Adults,Babies, BookingChanges,Children, DaysInWaitingList, LeadTime, PreviousBookingsNotCanceled, PreviousCancellations RequiredCardParkingSpaces, StaysInWeekendNights, StaysInWeekNights, TotalOfSpecialRequests	*Agent, ArrivalDateDayOfMonth, ArrivalDateMonth, ArrivalDateWeekNumber, ArrivalDateYear, AssignedRoomType, Company, Country, CustomerType, DepositType, DistributionChannel, IsCanceled, IsRepeatedGuest, MarketSegment, Meal, ReservationStatus, ReservationStatusDate, ReservedRoomType,*

Table 3.1 - Numerical and categorical features.

As we are dealing with a classification problem, it is important to check the distribution of the data according to the target variable (IsCanceled). Almost 42% of the total bookings are cancelled. From those cancelled bookings, 2.8% are no-show (meaning the customer got charged).

Going into more details, April to June are the months where there is a higher proportion of cancellations. Most of the bookings with no children and/or babies have a higher percentage of cancellations. Also, Bed & Breakfast bookings tend to cancel more than average. On the other hand, guests those requiring parking space, booking changes and/or special requests are more willing to show up. Repeated guests also have low percentage of cancellations, but they represent less than 3% of bookings.

Bookings through Travel Agents or Tour Operators (representing nearly 87% of the total bookings) tend to cancel more often, while groups reservations tend to cancel less than others.

Customer's bookings made far in advance are more willing to cancel the reservation.

3.2. Data preparation

A classification algorithm can be used to analyze discrete data, however, the dataset should be prepared to ensure it meets the requirements for this purpose. The dataset must be as complete as possible, missing values must be treated (either removed or filled in) and outliers should also be checked for. To avoid underfitting (when the model cannot detect any relationships in the data), relevant variables must be included; on the other hand, to avoid overfitting, variables with no predictive power should be eliminated. Because the quality of the data can make a model perform well or not so well, this data preparation stage takes the majority of the time in the project.

On the first step of data preparation, we drop some observations with missing values (28 observations, representing 0,035% of the dataset) and one observation in which the ADR is too high and clearly noisy (ADR=5.400). We have also eliminated the bookings with zero adults, as we believe it is an error, once in those same case there were children and/or babies included (0,48% of the dataset). The next step was feature engineering which we built 3 new features, for data analysis purpose:

New variables	Description
Days_before_cancel	*Number of days the booking is canceled before the entry date (=0 if not canceled)*
Days_until_cancel	*Number of days between the reservation is made until it's canceled (=0 if not canceled)*
RoomType_change	*Binary variable showing if the customer will get what he/she reserved (1 if ReservedRoomType=AssignedRoomType; 0 otherwise)*

Table 3.2 - New variables.

The next stage was to check the correlation matrices to identify redundant variables. We could not identify any strong correlation on the Spearman's matrix, but we identified strong correlations on Phi-k matrix, as we are dealing with a dataset with many categorical features. The following features were dropped, based on redundancy and their correlation between each other: *MarketSegment*, *RoomType_change*, *Agent*, *Company*, *ReservationStatusDate*, *AssignedRoomType*, *ReservationStatus*, *Days_before_cancel, Days_until_cancel*. We have also dropped 8 variables considered non-relevant for the model (*ArrivalDateYear*, *ArrivalDateWeekNumber*, *ArrivalDateDayOfMonth*, *ArrivalDate*, *Country*, *BookingChanges*, *RequiredCarParkingSpaces*, *TotalOfSpecialRequests*).

We have then reproduced the Phi-k correlation matrix to confirm it looked reasonable, once more. For further details, the correlations are available on the notebook.

Looking at the boxplots for each of the numeric variables, we have identified 5 variables (*ADR*, *DaysInWaitingList*, *StaysInWeekNights*, *Babies* and *StaysInWeekendNights*) with outliers (Figure 3.1).

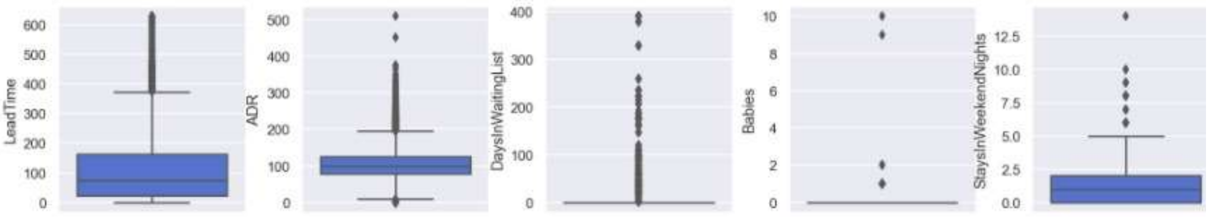

Figure 3.1 – Boxplot for numeric feature before outliers removal.

To remove these outliers, we tested applying five different methods on the entire dataset and using them individually or in combination. However, after testing these different approaches, and realizing that, outliers wise, it did not seem to make a big difference on the box and whiskers plots, we have opted for manually removing the outliers (considering it was feasible to do it for the five features where outliers were found). This way we have removed 2.723 observations, representing 3,45% of the dataset.

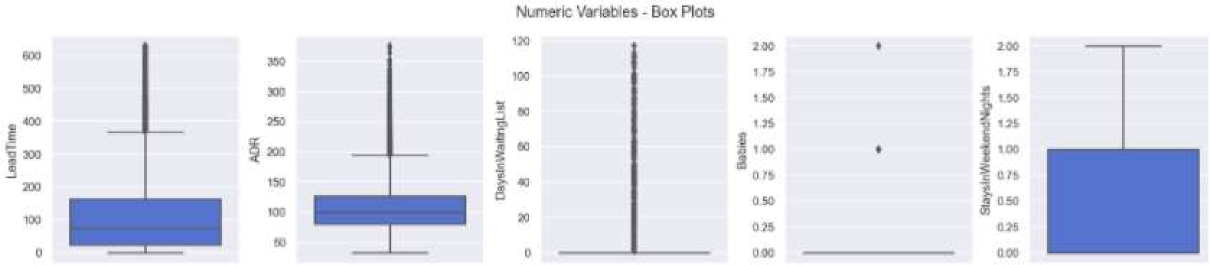

Figure 3.2 - Boxplot for numeric features after outliers removal.

In order to prepare the data for modelling, we have one-hot-encoded categorical variables (to transform in binary values).

3.3. MODELLING AND EVALUATION

The first stage of this process was to split the dataset into train and test, 80% and 20% of the dataset, respectively. This dataset has entries that are exactly the same and, even though they are real observations, to the model they can be considered as duplicates. Because of this, we had to be especially careful when splitting the dataset, to ensure the same pattern was followed both in the train and test datasets, so we split the repeated values equally between both datasets.

We start this process creating a pipeline to perform a baseline modelling. The pipeline includes two steps: excluding the features with variance low than 10% and running 4 different algorithms (Decision Tree, Random Forest, AdaBoost and Gradient Boosting) with the default parameters. We chose these algorithms based on their ability to predict the positive results (i.e. cancelations) and its ease of interpreting the results. We did not scale the data, because the algorithms chosen don't need it.

To measure the performance of each algorithm, we used a 10-fold cross validation only on the train set and calculated the average score for each fold. The metrics used were Accuracy, Precision and F1-Score. The results are presented on Table 3.3.

Model		**Mean**			**Std**	
	Accuracy	***Precision***	***F1-Score***	***Accuracy***	***Precision***	***F1-Score***
DT	*78.77%*	*74.74%*	*74.79%*	*0.52%*	*0.68%*	*0.57%*
GB	*77.57%*	*93.26%*	*65.30%*	*0.42%*	*1.07%*	*0.99%*
AB	*76.97%*	*95.79%*	*63.29%*	*0.43%*	*0.84%*	*1.04%*
RFC	*81.64%*	*81.22%*	*76.87%*	*0.43%*	*0.71%*	*0.64%*

Table 3.3 - Metrics results for the different algorithms.

The next step was to find the optimal number of features and get their importance; this analysis was made applying the Mutual information coefficient, as most of the variables have non-linear relationships. Features *DepositType_Non Refund, DepositType_No Deposit* and *ADR* stand out as the most relevant features to the target (altogether weighting 57%). The ranking with the features which weight>1% is presented in Figure 3.3.

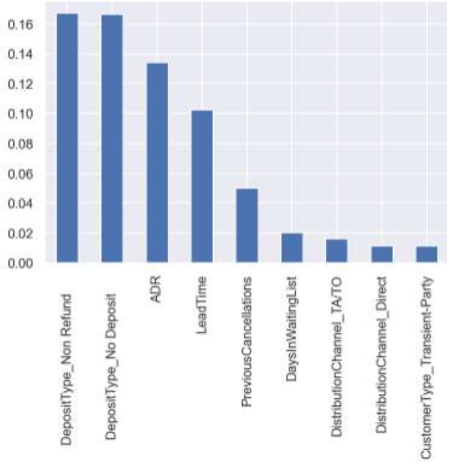

Figure 3.3 - Features importance from Mutual Information Coefficient.

We repeated the process of applying the same algorithms with the default parameters and cross validation on the train set, but now only with the top 9 features selected by mutual coefficient. The results are presented on Figure 3.4.

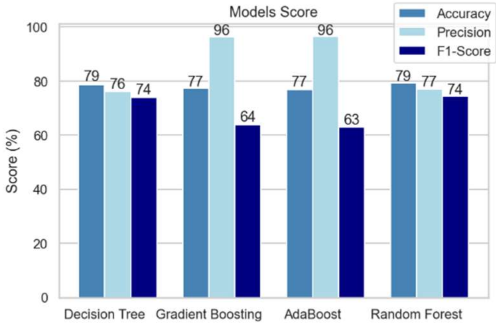

Figure 3.4 - Models score.

Analyzing the results, we chose as our final algorithm the Random Forest Classifier. Although in one hand Gradient Boosting and Adaboost presented very good scores for precision metric, on the other hand they presented low score for F1 metric, so we decided to not use them. Compared to Decision trees, Random Forest presented better accuracy and precision.

The next stage was trying to improve the scores of Random Forest tunning the parameters. At this time, we applied the model in the test set. We tested 6 different types of parameters. The table with the parameters of each model and the graphic with the results are presented below:

Model	Parameter changed	Value
modelRF	random_state	5
modelRF_ent	Criterion	Entropy
modelRF_10trees	n_estimators	10
modelRF_b	bootstrap	False
modelRF_s	max_samples	0.7

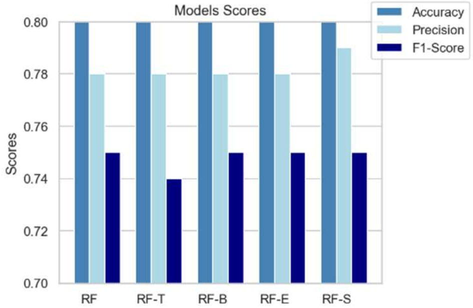

Table 3.5 - Tuning RF parameters and results.

The model with the best performance was the one with the parameter *max_samples = 0.7*. All models have very similar results, but the one chosen has a better precision when compared to the others.

You can see below the ROC curve and the confusion matrix for the final algorithm .

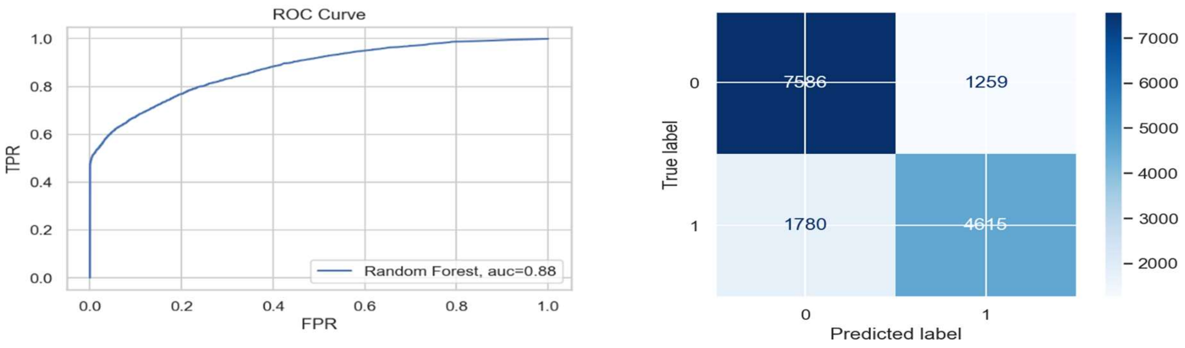

Figure 3.6 – ROC Curve and Confusion matrix to RFC.

The classification report and confusion matrix for all models with Random forest classifier can be found on the notebook.

The features that stood out as the most relevant for the model to predict the target are: *ADR*, LeadTime, *DepositType_No Deposit* and *DepositType_Non Refund*. It is important to note that these four features together weight almost 90%. The full ranking is presented in Figure 3..

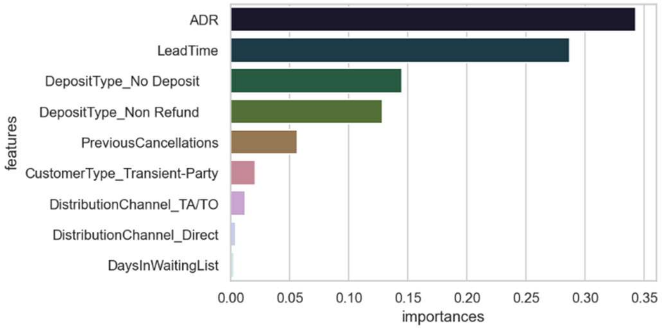

Figure 3.7 - Random Forest Classifier - Features importance.

4. RESULTS EVALUATION

The model developed by B4C reached good values on many important measures which will certainly help Hotel Chain C to implement actions to prevent cancellations.

		Predicted label	
		No cancel	Cancel
True label	No cancel	50%	8%
	Cancel	12%	30%

Random Forest Classifier predicts cancellation/no cancellation correctly 80% of the times. However, 12% of the bookings will be cancelled but the model will wrongly predict them as not cancelled, which may cause that some rooms are not occupied, leading to loss on revenue. 8% of the bookings will wrongly be labelled as cancelled, however the customer will show up, which may lead to overbooking or extra expense costs on trying to reverse the suppose cancellation.

We could say we found a good balance between the false cancellations and false no-cancellations reducing the costs generated by this.

To conclude, because the model achieves 80% of accuracy, it will probably lead to a reduction on cancellations to a rate of 20%, which was stated as one of the business goals.

5. DEPLOYMENT AND MAINTENANCE PLANS

We understand that the process of implementing the model is very critical. Hotel's market is very dynamic and bookings are changing daily. Every day, there are new bookings on the systems, cancellations or/and changes on the reservations on-the books. So, we understand that it would be very important to implement the model integrated with the hotel reservation system. Integrating both, the results will be generated in a daily basis and the hotel could react faster to the cancellations.

Another important point is the data is that we have some problems of quality in the data. We advise to improve the quality of variables Deposit Type and country. We also find some reservations with no adult guests and believe this might be an issue on the records. Also, on the data provided we found a good percentage of very similar bookings that can reduce the performance of the model.

We are going to monitor the performance of the model after it is implemented for some months and make some adjustments if necessary.

We also suggest updating the model from time to time as the behavior of the customers, and even the market trends, can change, reducing the model performance.

Lastly, considering this model was built with only few thousands of observations, we consider that re-training the model when new data is available would potentially improve its quality.

6. CONCLUSIONS

As state in the section 2.2 – Business objectives, the main objective of this project was to implement a predictive model to forecast net demand based on reservations on-the-books.

The model developed by B4C reached good values on many important measures such as 80% of accuracy and 87% of AUC. It would certainly help Hotel Chain C to implement actions to prevent cancellations.

We were also able to describe the key characteristics of bookings to help to identify high cancellation likelihood bookings.

In addition, we set some risks on this project. One of these risks is the model performance, as we have been working with almost 30% of duplicated observations. We implemented some actions to reduce this risk, but the model will have improvement margin if we get additional datasets to test its performance.

We hope Hotel Chain C is satisfied with D4B work and we can continue working together

7. REFERENCES

[1] Chapman, P, Clinton, J, Kerber, R., Khabaza, T., Reinartz, T, Shearer, C. & Wirth, R. (2000). *CRISP-DM 1.0*, CRISP-DM consortium

[2] Narkhede, S 2016, Understanding AUC - ROC Curve, viewed 11 March 2021, <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>

[3] Randolfo, M 2020, Como avaliar seu modelo de classificação, viewed 13 March 2021, <https://medium.com/data-hackers/como-avaliar-seu-modelo-de-classifica%C3%A7%C3%A3o-34e6f6011108 >