NOVA IMS Information Management School

MDSAA

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

BUSINESS CASES WITH DATA SCIENCE

Group: R
Henrique Costa M20200652
Pedro Mendes M20200648
Rafael Soromenho M20200969
Rui Soromenho M20200578

March 2021

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

Table of Contents

List of figures

1. Introduction

The goal of this project is to develop a predictive model that supports the management of a hotel chain in Portugal in handling cancelations more effectively. The model should be able to identify with a high degree of accuracy which bookings are likely to result in cancelations, allowing for the hotel to set appropriate policies to limit its negative impact. To do this a dataset containing information on 79,333 bookings made between 2015 and 2017 will serve as basis for the analysis. Below is a high-level overview of the steps that will be taken to produce the desired output:

- **Business Understanding and Data Exploration:** Developing intuition and understanding the data considering the business context.
- **Data Cleaning, Selection and Transformation**: Identification of key dimensions that best capture relevant information to predict cancelations.
- **Modelling alternative machine learning techniques** to maximize the capability to predict future cancelations.
- **Evaluate results and select best model** to apply.
- Provide input on the **deployment** strategy to implement the model and consider its **business implications.**

2. Business Understanding

2.1 Background

Cancelations are a key issue in the hospitality industry causing significant revenue loss. In the specific case addressed in this project, the cancelation rate for the chain's city hotel in Lisbon has averaged around 42% in the period between 2015 and 2017. There are two common management approaches that attempt to limit the impact of cancelations: overbooking and restrictive cancelation policies. While overbooking often comes with negative side effects on customer trust and often results in additional costs, applying restrictive cancelation policies tends to lead to a decrease in demand and, thus, a potential decrease in revenue. The context previously described makes clear the potential for value creation of an increased ability to identify bookings with high probability of being canceled.

2.2 Business Objectives, Data Mining Goals and Success Criteria

The primary objective of the hotel is to decrease cancelations. For management, success in this task is defined as achieving a reduction of the cancelation rate from the current 42% to a number below 20%. There are several approaches to achieve the referred goal. On one side this would be made possible by taking preventative actions directed at the bookings that are likely to cancel. On the other, if the hotel can identify early on which bookings are likely to be canceled as well as understand cancelation drivers, better decisions can be made on how to strategically allocate capacity (e.g., reduce bookings accepted for certain segments with high cancelation rates while growing others). In addition, being able to predict net demand – gross bookings subtracted of cancelations - more accurately is crucial for the business at it would allow for improvements in current practices such as overbooking contributing for increasing the revenue stream while minimizing the negative impact of cancelations.

To support the achievement of the referred goals, the intended data mining output is a binary classification model that can predict if a booking is going to be canceled or not. The model should provide a balance of performance metrics as there are trade-offs between avoiding revenue loss – which would happen when the number of False Negatives is high (e.g. the model predicts that the booking will not cancel but it does) - and managing capacity responsibly – which would be difficult when the number of False Positives is high (e.g. the model predicts the booking will be canceled but it is not). This trade-off will be further discussed later in the report.

As per the discussed above, the outputs of the project should focus on three key areas:

- Identify what variables better explain cancelation patterns.
- Produce an accurate predictive model and bridge the model performance with business impact.
- Explore how the model could be employed to enhance cancelation management.

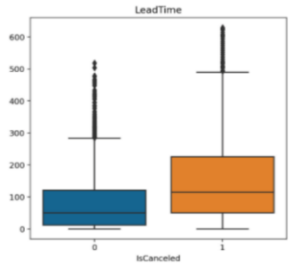
3. Predictive Analysis

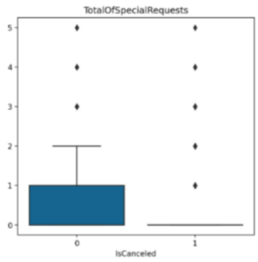
3.1 Data understanding and exploration

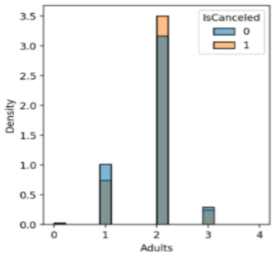
The data is made of 79,330 bookings of which 42% are labeled as cancelations – a balanced dataset. Each booking is described across 31 variables. An initial verification shows that there is a minimal number of missing values – which were set to the mode of the respective column. Other variables such as ‘Country’ or ‘Agent’ contained a significant number of ‘NULL’ values that seem to indicate that the observation does not correspond to any company or agent.

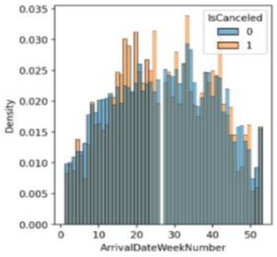
- Several variables are not suitable to be used in the prediction model as they contain information that is equivalent to the target variable: ‘ReservationStatus’ and ‘ReservationStatusDate’.
- Variable ‘DepositType’ raises data collection quality concerns and should not be included in the model.

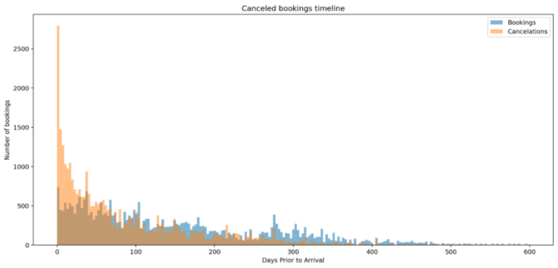
Visualizations were done to identify the need for data cleaning as well as to build intuition on which variables seem to have greater predictive power.

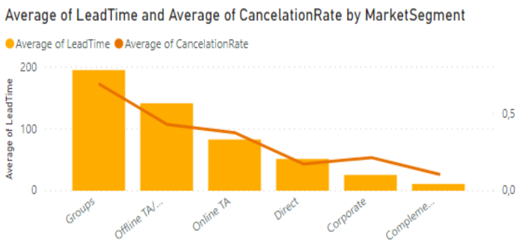












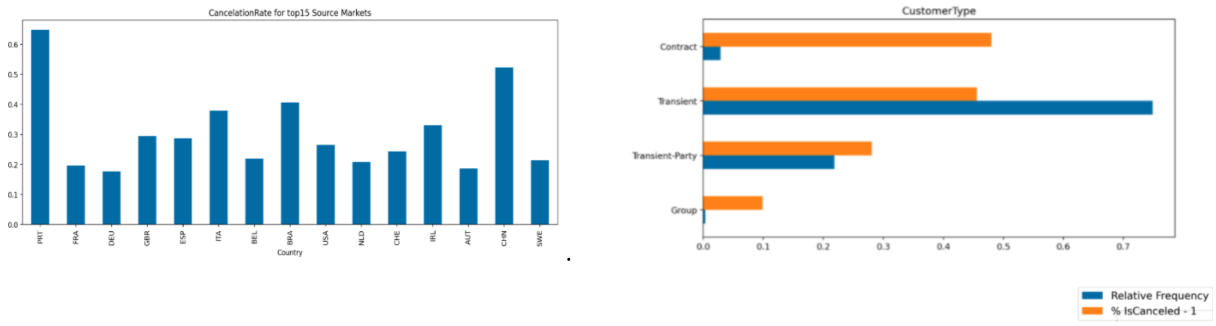


Figure 1 - Examples of visualizations for data explorations

- There are several groups of variables containing redundant information (see Pairplot in the notebook as well as Phik Matrix and Multiple Correspondance Analysis).
- The variables 'LeadTime,' 'MarketSegment,' 'TotalSpecialRequests,' 'BookingChanges,' 'ArrivalDateWeekNumber,' 'RequestedCarParkingSpace' , 'CustomerType' seem to have the highest predictive power.

A critical issue to address is the considerable number of duplicated values found in the dataset. Given that the data does not contain any variables that can be used as a unique identifier of the respective booking its analysis combined with business intuition seems to indicate duplicate rows correspond to different bookings. Given that it is expected that the model would be facing similar patterns when deployed it was decided that the duplicate observations would not be removed from the training dataset.

3.2 Data preparation

3.2.1 Removing Outliers

Given its ability to detect multivariate outliers, the density-based clustering algorithm BDScan was used for outlier removal. The two key parameters that must be defined to perform DBSCan are the MinPts and epsilon - defined as 32 and 4, respectively. MinPts should be around 2xdim and eps can be estimated based on the visualization of the distance between observations in ascending order.

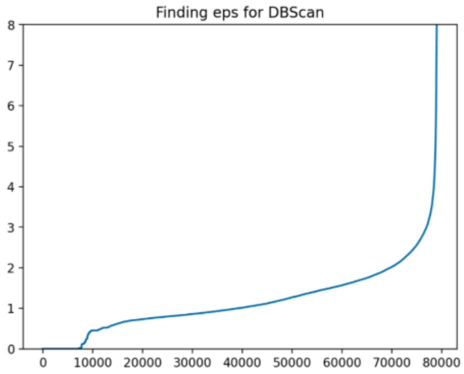


Figure 2 - Using DBSCan for Outliers removal

The algorithm was applied to the selected features, resulting in the identification of 0.5% of observations as outliers. There was a subsequent verification of the data in search of any remaining univariate outliers that were not classified as such by DBScan in which an additional 0.1% of the data was deleted.

3.2.2 Construct Data

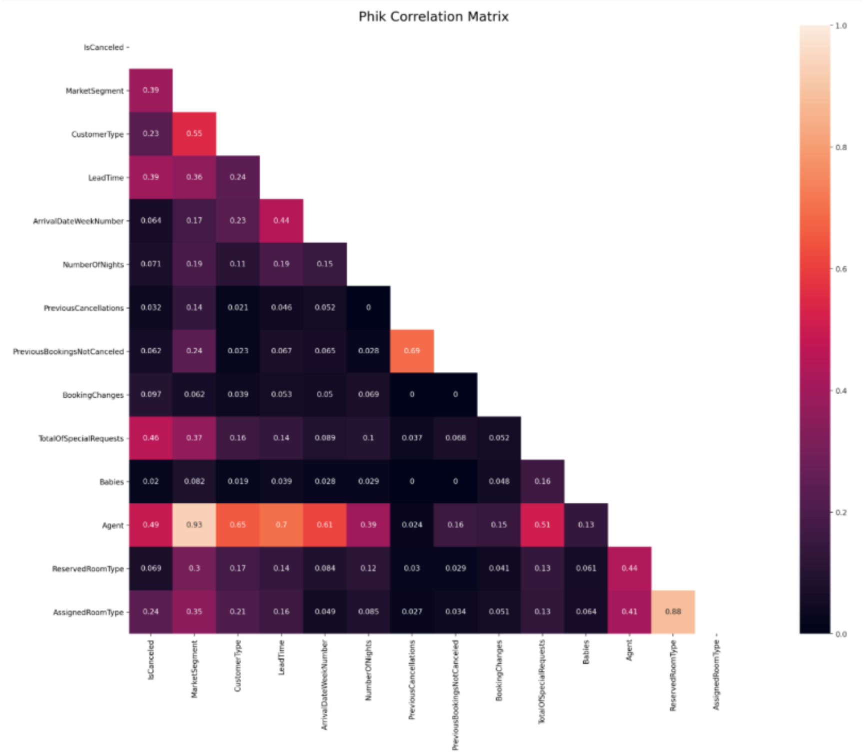
Several new variables were engineered to better capture relevant drivers of cancelations:

- NumberOfNights: Obtained by summing the weekend and weeknights.
- TotalRevenue: Obtained by multiplying number of nights by ADR.
- PrevCancelationRate: Dividing the number of previous cancelations by the number of previous bookings.
- ReservedVsAssigned: Comparing reserved and assigned room type.

An attempt to reduce cardinality by grouping both categorical (e.g., creating country groups) and numerical features (e.g., changing number of children to binary) was also made but not included in the final model as results were inferior to the ones obtained with the original data.

3.2.3 Data selection

Together with the observations from the data exploration phase, several techniques were used to facilitate the selection of the final group of features to be included in the model. PCA (Principal Component Analysis) was used to reduce dimensionality to 6 principal components - explaining 70% of variability in the data. Subsequently, the correlation between the referred principal components and the original variables was assessed to identify which characteristics are most relevant in explaining the patterns contained in the data. Other tools such as Multiple Correspondence Analysis, Pearson and Phik Correlation Matrixes or the Chi-squared coefficient were also utilized to understand the variables discriminatory power and potential redundancy between features.



	IsCanceled	MarketSegment	CustomerType	LeadTime	ArrivalDateWeekNumber	NumberOfNights	PreviousCancellations	PreviousBookingsNotCanceled	BookingChanges	TotalOfSpecialRequests	Babies	Agent	ReservedRoomType	AssignedRoomType
IsCanceled														
MarketSegment	0.39													
CustomerType	0.23	0.55												
LeadTime	0.39	0.36	0.24											
ArrivalDateWeekNumber	0.064	0.17	0.23	0.44										
NumberOfNights	0.071	0.19	0.11	0.19	0.15									
PreviousCancellations	0.032	0.14	0.021	0.046	0.052	0								
PreviousBookingsNotCanceled	0.062	0.24	0.023	0.067	0.065	0.028	0.69							
BookingChanges	0.097	0.062	0.039	0.053	0.05	0.069	0	0						
TotalOfSpecialRequests	0.46	0.37	0.16	0.14	0.089	0.1	0.037	0.068	0.052					
Babies	0.02	0.082	0.019	0.039	0.028	0.029	0	0	0.048	0.16				
Agent	0.49	0.93	0.65	0.7	0.61	0.39	0.024	0.16	0.15	0.51	0.13			
ReservedRoomType	0.069	0.3	0.17	0.14	0.084	0.12	0.03	0.029	0.041	0.13	0.061	0.44		
AssignedRoomType	0.24	0.35	0.21	0.16	0.049	0.085	0.027	0.034	0.051	0.13	0.064	0.41	0.88	

Figure 3 - Phik Correlation Matrix for selected features

The data selection approach was also validated using a recursive feature elimination approach through which several feature combinations were tested in the modeling phase to increase confidence in the prediction power of the final selection and ensure redundant or irrelevant features were excluded from the model while all relevant information was used.

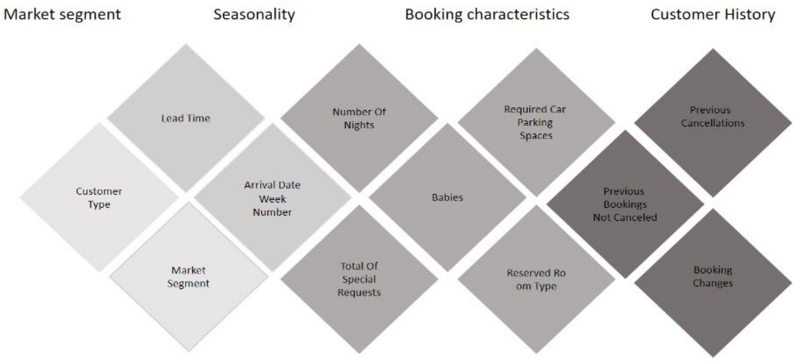


Figure 4 - The most relevant features

3.3 MODELING AND RESULTS EVALUATION

As expected, when applying the CRISP-DM methodology some steps were not sequential and required going back and forth. It was the case of the modeling and evaluation steps. During this step was important to understand the feature relevance for each model and match those findings with knowledge acquired from data understanding. Initially a benchmark with several binary classification algorithms such as K-nearest neighbors, Decision Trees, Neural Networks and Boosting classifiers and LightGBM. The initial benchmarking was followed by a more thorough optimization step for each model.

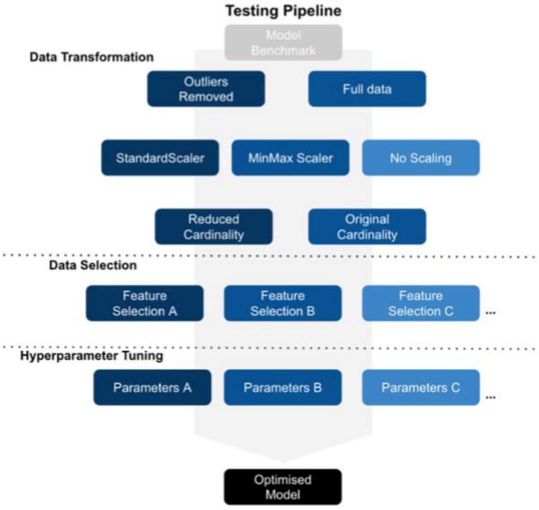


Figure 5 - Testing Pipeline

As expected, and identified due to a performance metric's comparison, the models with higher scores correspond to embedded methods using decision tree classifiers. The model scores were obtained using K-Fold Cross Validation with 10 folds:

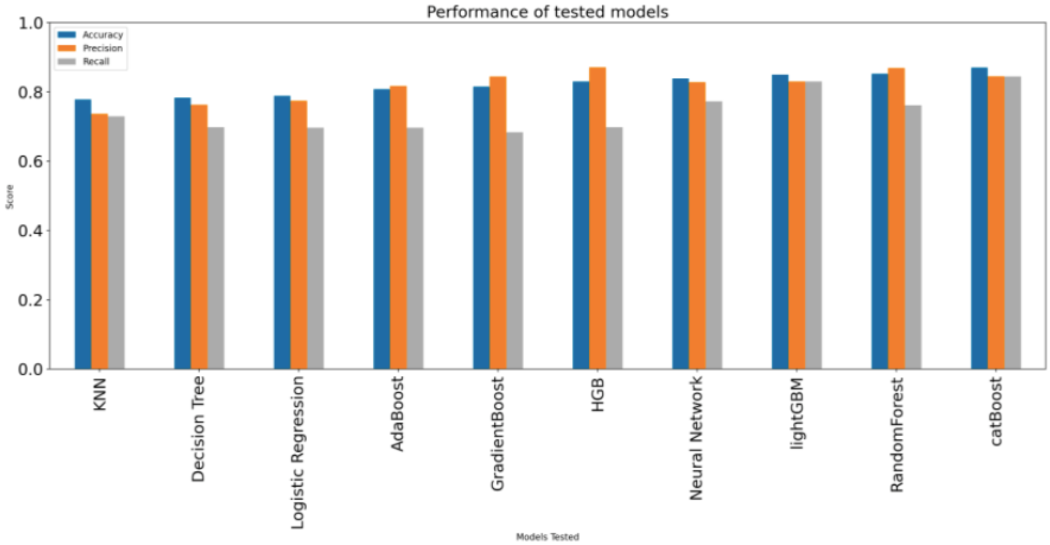
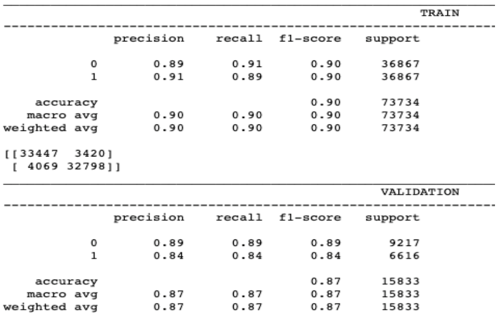


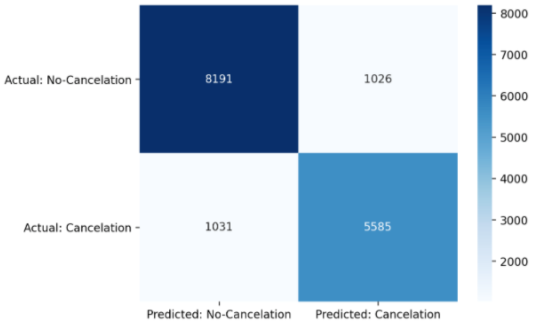
Figure 6 - Performance of tested models

CatBoost algorithm shows better and more consistent results among the three key performance metrics more relevant for the business objectives:



```
                                                           TRAIN
------------------------------------------------------------------
              precision    recall  f1-score   support

           0       0.89      0.91      0.90     36867
           1       0.91      0.89      0.90     36867

    accuracy                           0.90     73734
   macro avg       0.90      0.90      0.90     73734
weighted avg       0.90      0.90      0.90     73734

[[33447  3420]
 [ 4069 32798]]
                                                      VALIDATION
------------------------------------------------------------------
              precision    recall  f1-score   support

           0       0.89      0.89      0.89      9217
           1       0.84      0.84      0.84      6616

    accuracy                           0.87     15833
   macro avg       0.87      0.87      0.87     15833
weighted avg       0.87      0.87      0.87     15833
```

Figure 8
-



Performance of tested models

Figure 7 - Classification report and confusion matrix

In the chosen model, a fixed threshold of 50% was used, meaning bookings with probability below 50% were classified as 0 (non-canceled) and all others as 1 (canceled). Moreover, it is important to highlight that the threshold may be changed which would result in different performance metrics, since the cut-off for classification would start to bias results towards one of the classifications. For example, if the threshold were to be decrease to 40% the model would be more sensitive to cancelations (1). As a result, recall would be higher while both accuracy and precision would be penalized. This topic will be further discussed during business implementation chapter as the threshold choice is partly dependent on its intended use and specific policies the hotel will take.

4. Business Implementation

4.1 Implications of performance metrics for Model Employment

The observed results indicate that the data available is a useful source to predict with high accuracy if bookings are going to be canceled. Accuracy reached 87%, meanwhile precision and recall reached 84%. These results confirm that it is possible to identify bookings with a high likelihood of being canceled.

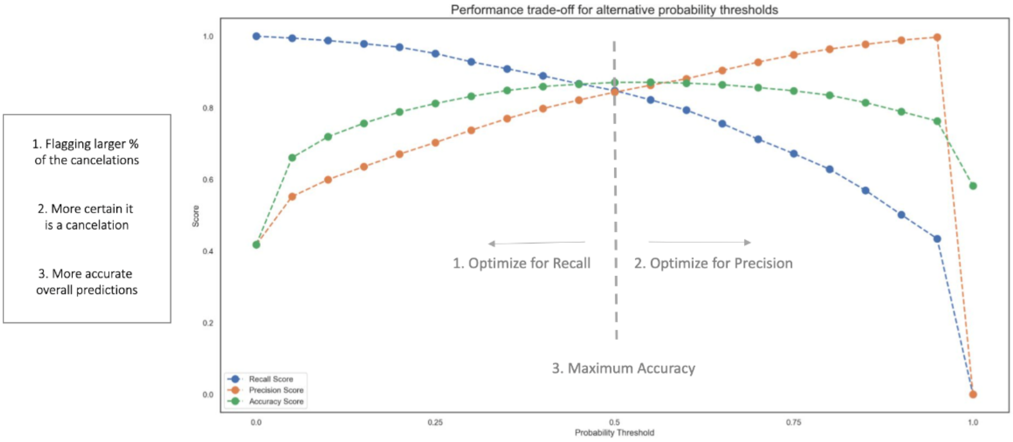


Figure 8 - Performance trade-off for alternative probability thresholds

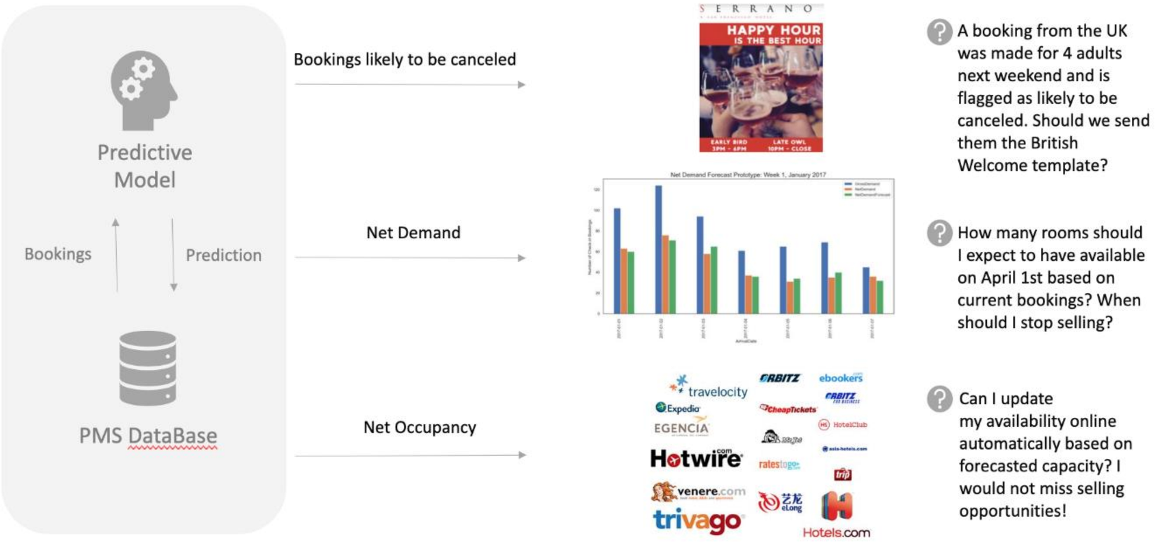
It is a crucial piece to understand what information is hidden in these metrics - Accuracy, Precision and Recall - to be able to understand which is the best fit to a particular business strategy. For instance, if the hotel would prefer to implement an ambitious preventive approach, then the number of false positives could be of extreme importance to prevent the hotel from spending in cash/services with bookings that would not have been canceled. Therefore, for this kind of approach precision would be the best fit. On the other hand, if the hotel would want to have a better net demand forecast, it is important to know how many of the actual cancelations could our model identify (recall) and how many correct predictions we have (accuracy). The mix of these two metrics would be interpreted as having a high understanding of predicted cancelations and having a high accuracy in the predictions, which would allow hotel's management to sell above capacity without fearing overbooking consequences. The final threshold choice would depend on the models intended use and risk preferences of the management team.

4.2 Deployment and Business Processes

To move forward with implementing the project there are 2 key factors to consider.

The first is integration. To effectively deploy the prediction model, it is proposed that it should be integrated into the hotel bookings management system. This would allow for real time updates of a database containing probable cancelations that could feed different interfaces for improving cancelations management. This is important because some features such as booking changes, special requests may vary over time, which would change the likelihood of the booking being a cancelation. It is also extremely important to periodically understand if the model performance has changed, since a change in the cancelation patterns of the customers could also decrease the model metrics.

Second is operationalization. There are several ways in which the model can be used to enable better approaches to cancelations management including support for preventive actions or improved net demand forecasting for overbooking policies.



A booking from the UK was made for 4 adults next weekend and is flagged as likely to be canceled. Should we send them the British Welcome template?

How many rooms should I expect to have available on April 1st based on current bookings? When should I stop selling?

Can I update my availability online automatically based on forecasted capacity? I would not miss selling opportunities!

Figure 9 – Action enabling interfaces with data integration.

In practice, there is a two-way constant information flow. When getting a booking the system would be connected to the predictive model and the model would feed the database with information on the cancelation likelihood. These outputs could then be connected to different interfaces to enable action.

- Preventive measures for bookings likely to be canceled: We could set other rules besides the cancelation likelihood for different triggers to be activated (e.g., contact a customer with a customized email, etc...)
- Enabling Forecasting Tools: allow for capacity planning based on forecasted net demand.
- Managing channel availability: Using the model predictions to update online availability both in our direct channels and connected to the online Travel Agencies would guarantee that you would not miss selling opportunities.

The above are common use examples but there are many other potential applications and room for customizing implementation to the needs of the hotel.

5. Conclusions

The results obtained in the project indicate a high potential for machine learning in contributing for managing cancelations more effectively in the hospitality industry. The three overarching goals of the project were achieved. Key factors in determining cancelation probability were identifies and seem to be connected to 4 key subgroups: market segment, customer history, booking features and seasonality. In addition, the GradientBoosting model proposed as well as most other models tested showed high accuracy in the predicting cancelations. This shows the data currently collected by most hotel booking systems allows for more sophisticated and effective approaches than the ones currently in use. Finally, several opportunities for embedding the predictive model in the hotel's practices were discussed highlighting the potential for real world impact of effectively transforming data into knowledge.

6. References

- Abbott, D. (2014). Applied predictive analytics: Principles and techniques for the professional data analyst. Indianapolis, IN, USA: Wiley.
- Rabianski, J. S. (2003). Primary and secondary data: Concepts, concerns, errors, and issues. Appraisal Journal, 71(1), 43 (13).
- Medium. 2020. Understanding Random Forest. [online] Available at: [Accessed 26 December 2020].
- Tingle, M., 2020. Preventing Data Leakage In Your Machine Learning Model. [online] Medium. Available at: [Accessed 10 November 2019]
- Scikit-learn.org. 2020. 1.13. Feature Selection — Scikit-Learn 0.24.0 Documentation. [online] Available at: [Accessed 26 December 2020].
- Scikit-learn.org. 2020. 1.13. Feature Selection — Scikit-Learn 0.24.0 Documentation. [online] Available at: <https://scikit-learn.org/stable/modules/feature_selection.html> [Accessed 26 December 2020].
- Ofori, Martinson,A First Look at Sklearn's HistGradientBoostingClassifier, medium.com , accessed 27thDecember2020,available from <https://medium.com/@mqofori/a-first-look-at-sklearns-histgradientboostingclassifier-9f5bea611c6a>
- Scikit-Learn2020, sklearn.ensemble.HistGradientBoostingClassifier, accessed 27thDecember 2020, available from https://scikit-learn.org/dev/modules/generated/sklearn.ensemble.HistGradientBoostingClassifier.html#sklearn.ensemble.HistGradientBoostingClassifier