BUSINESS CASES WITH DATA SCIENCE

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

Business Case 2 – Predict Hotel Booking Cancellations

Authors (Group AA):

Addex Consulting
Emil Ahmadov (m20201004)
Doris Macean (m20200609)
Doyun Shin (m20200565)
Anastasiia Tagiltseva (m20200041)

March 15, 2021

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

INDEX

1. INTRODUCTION

Our team was hired by Hotel chain C to develop a model that can predict the likelihood that a booking will be canceled. The goal is to improve cancellation policies, define better overbooking tactics and thus implement better pricing. The group was given a data set of 79 330 city hotel bookings, which were due to arrive between July 1, 2015, and August 31, 2017.

2. METHODOLOGY

Cross-Industry Standard Process for Data Mining (CRISP-DM) [1] was the applied methodology in the execution of this research. CRISP-DM [2] provides a six-step process: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The following sections offer a succinct description of the execution of each of these steps.

Our team followed a predictive analytic process to better understand our problem and the relevance of the feature set. We began with some descriptive analytics to understand what happened in this dataset. We then wanted to explore possible reasons why the cancellations happen. Once our data exploration was complete, we proceeded to our predictive analysis where models were tested, and the optimal predictive model was chosen. Finally, we obtained an understanding of which features have the highest explanatory power.

Our team took a systematic approach to find the most appropriate predictive model. We began by extensively reviewing and cleaning the data. At this preliminary stage, we ensured that there was no missing data, transformed features so that they would be useful for our analysis and addressed outliers. Secondly, we performed an extensive exploration of the data to understand how each feature relates to the target variable. Here we reviewed the distribution of the target variable, as well as correlations between all the features. Finally, we preprocessed the data by mapping categorical features to numeric, scaling the data and splitting the data into a train and validation set. At this point, our data was clean and ready to have the machine learning models applied and tested. Below we have detailed the steps we took to prepare the data for modelling.

3. BUSINESS UNDERSTANDING

3.1. BUSINESS OBJECTIVES

Our client's business objectives essentially come down to the following:
1. Tuning the balance of overbooking and cancellation policies, and reduction of cancellation rate through offerings
2. Reduction of cancellation rate through offerings, to those that have high likelihood of cancellation.

3.2. BUSINESS SUCCESS CRITERIA

The client's goal is to reduce cancellations to a rate of 20%. To accomplish this, model(s) should achieve a prediction accuracy and an area under the curve (AUC) above 0.8, which is commonly considered a good prediction result [3].

3.3. SITUATION ASSESSMENT

For this project, we utilized mid-range PC of our own and a nearly 80,000 sample of booking history data from Hotel Chain C. Addex Consulting consists of four data scientists and python as the primary

coding language. Various python modules were used to analyze the data, and develop a prediction model for cancellation. The specific tools are addressed in our enclosed notebook.

We were able to develop a model that could accurately predict customer cancellations with the data provided. However, the utility of our algorithm in other area was limited by limited data. Namely, the hotel capacity data; current overbooking and cancellation policies; and reliable deposit type data could help us to develop applications that could be useful to the hotel. Furthermore, these data could help us with many decisions we made during our process. Additionally, our computing power limited usage of some tools, such as grid search method for hyperparameter tuning.

Please note in this report, we refer to Hotel Chain C as Hotel C.

3.4. DATA MINING GOALS

The aforementioned first business objective can be translated to 'developing robust and accurate algorithms for predicting if a specific customer is likely to cancel. The second objective can be supported with analyzing and identifying the factors most correlated with cancellation tendencies.

4. DATA UNDERSTANDING

To obtain a general understanding of the hotel's customers and their behavior, we conducted a descriptive analysis. The figure below shows a total of 42% of bookings that were canceled.

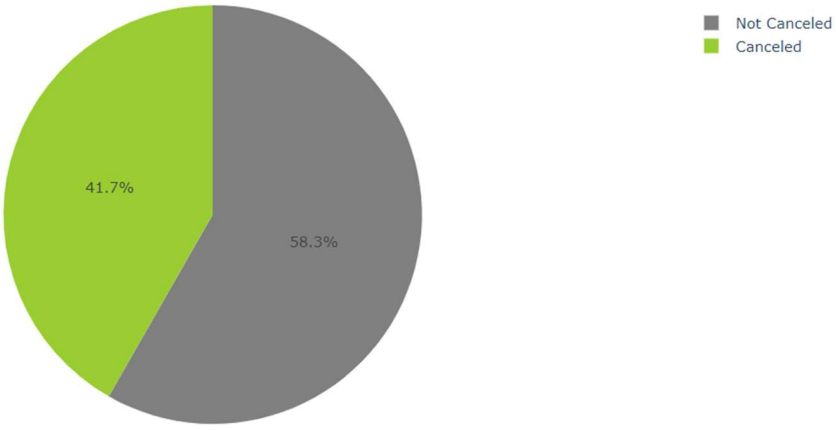


Figure 1: Proportion of booking that were cancelled.

We have identified that the greatest number of bookings occurred in the month of August and the least number of bookings in January. August sees the highest number of cancellations. As we can see on the graphs below, the number of cancellations appears to be a reasonably constant proportion of the number of bookings per month.

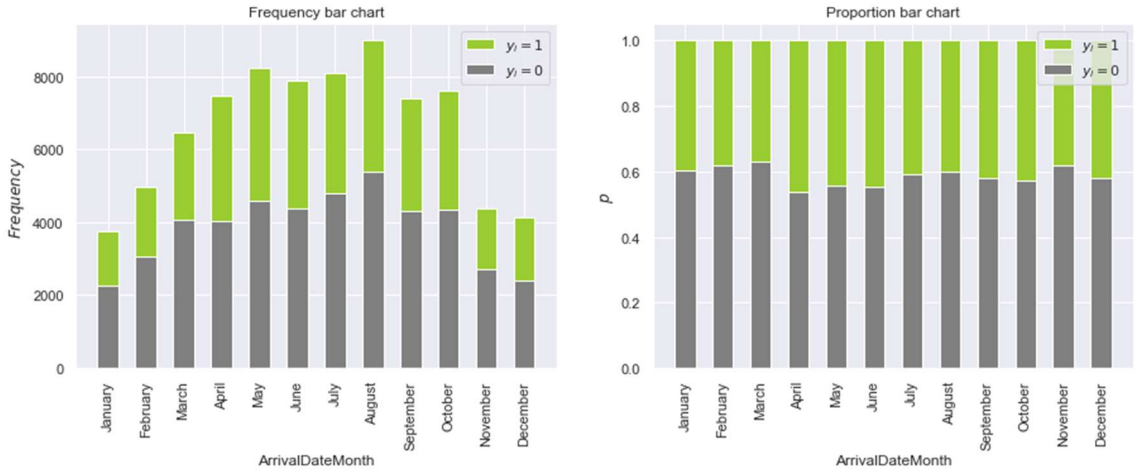


Figure 2: Frequency and proportion of cancellation by month.

We assessed the geographic distribution of the customers in the dataset. The majority of the hotel's clients reside in Portugal. Many guests also come from countries around Europe, followed by USA and Brazil. Please see the figure below showing a heatmap accordingly. We observed that the proportion of cancellations done by Portuguese customers is significantly higher than international clients.

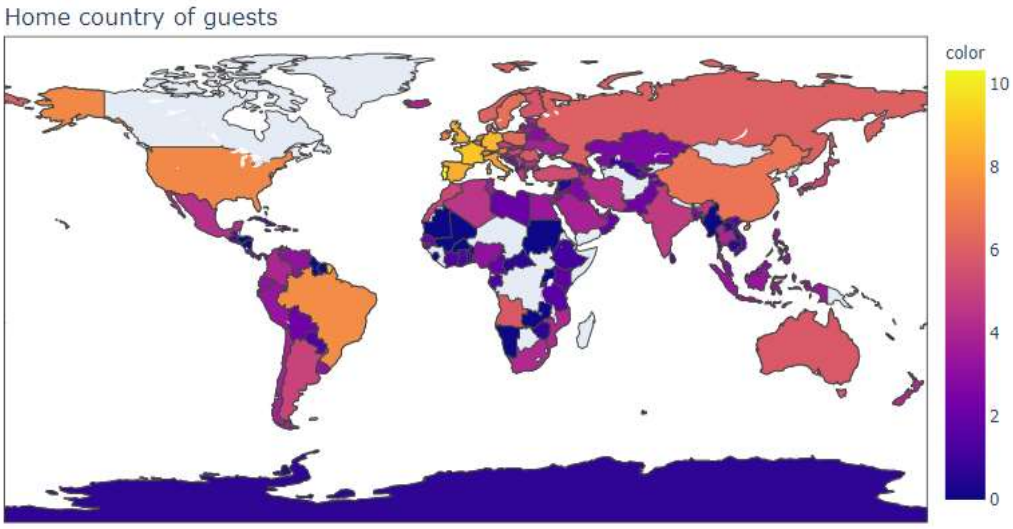


Figure 3: Heatmap of client's customers by country.

An extensive exploration of the data was then done to understand how each feature relates to the target variable. As you can see in the below example, the proportion of bookings that were canceled when no special requests were made is significantly higher than those bookings with special requests. We also looked at the average values of features within each class. Here we found variables that did not vary significantly and others that really differed by class such as Booking changes and LeadTime. Below is a sampling of some of the variables we reviewed.

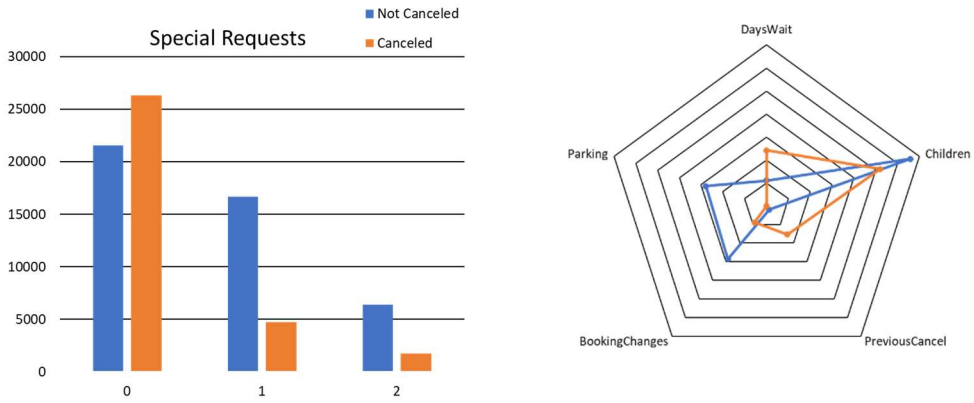


Figure 4: Comparison of proportion and canceled and non canceled bookings across several features.

5. DATA PREPARATION

5.1. CHECKING AND HANDLING OF MISSING AND CORRELATED VALUES

To eliminate multicollinearity and limit redundancy, we plotted the correlation matrix of all the independent variables. As our data set contains both categorical and interval variables, we obtained the ϕK correlation coefficient. Phik has several refinements on Pearson's hypothesis test of independence of two variables [4].

When independent variables are highly correlated, change in one variable would cause change to another and so the model results become biased toward the correlated variables. This will create the following problems. It would be hard to choose an appropriate list of significant variables for the model. Coefficient Estimates would not be stable, and it would be hard to interpret the model. From the heatmap below we can see a strong correlation (>0.8) between ReservationStatusDate with others. The most straight-forward method is to remove some variables that are highly correlated to others and leave the more significant ones in the set.

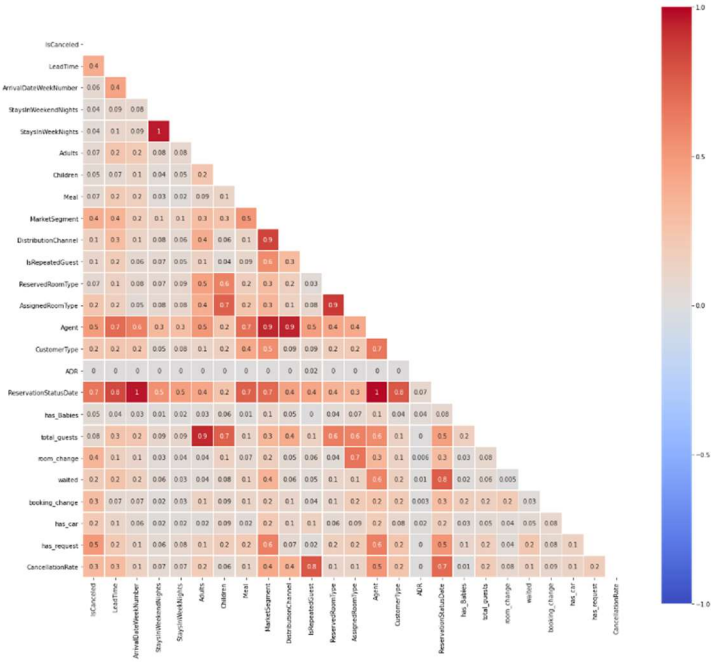


Fig.5 φK correlation matrix heatmap

We have observed that 'Company' has a high proportion (95%) of missing values and as such will likely not be useful in our modeling. As such, we dropped the 'Company' column. Furthermore, 'DistributionChannel' is highly correlated with 'MarketSegment' and was dropped to increase generalisability. In addition, our team was advised that the country is only confirmed on arrival. We have deemed the feature unreliable and have not proceeded to use it in our analysis.

We replaced null values in the 'Children' column with 0 as this appears to be the most frequent value found in our dataset. In addition, we replaced rows in the 'Meal' column that were labeled "Undefined' with label 'SC', as they represent the same circumstance of no meal package.

5.2. HANDLING OUTLIERS

To identify outliers, our team conducted an analysis of the boxplots of the numeric variables. Some anomalies were detected in the variables of "Babies", "PreviousCancellations" and "Adults". As these circumstances are infrequent, they are unlikely to be useful in a predictive model. Our team proceeded to drop rows with 9 and 10 Babies, as well as those with more than 10 Cancellations.

Several bookings showed 0 customers – no adults or children. The client advised that Bookings with no assigned people can happen due to expenses in other hotel services like spa and restaurant. As these represented as small percentage of the dataset and will not impact the vacancy rate, we dropped these rows from our analysis.

5.3. FEATURE ENGINEERING AND ENCODING

We engineered some features which we considered will be important to our analysis:

- "has_Babies": indicates if guest is travelling with his/her baby
- "room_change": indicates if assigned room is different than reserved room
- "waited": if customer waited in waiting list more than a day
- "booking_change": if customer made any change to their booking
- "has_car": if customer asked for parking spot
- "has_request": if customer has any special request
- "CancellationRate": percentage of the previous bookings that were cancelled. For the first time customers we used 0.42, which is the proportion of first time customers cancelling their bookings among the dataset

For encoding categorical variables, we used mean encoding. This method transforms categories to numerical values by assigning the cancellation probability of that category. To prevent bias based on the number of occurrences of some categories being low, additive smoothing was used[1]. This method soothes the probabilities according to general mean.

[1] Halford Max (2018, October), Target encoding done right way, https://maxhalford.github.io/blog/target-encoding/

$$\mu = \frac{n \times \bar{x} + m \times w}{n + w}$$

Where:

$\bar{x}$: average cancellation rate per category

m: general average of cancellation in dataset

n: number of occurrences for a category

w: smoothing weight

5.4. FEATURE SELECTION

After inspecting the dataset and taking into account the correlation of each feature with 'IsCanceled', we feel that the 'ArrivalDateYear', 'ArrivalDateMonth', ''ReservationStatus', 'ArrivalDateDayOfMonth' and 'DepositType' columns are irrelevant to our analysis, and hence we dropped those columns.

6. MODELING

6.1. SCALING

We applied Robust Scaler, Min-Max scaler, and Standard Scaler to our prediction model. After multiple alteration of feature selection throughout the modelling process, min-max scaler consistently showed reliable results in our final prediction, hence min-max scaler was used.

6.2. MODEL SELECTION

Using stratified 10-fold cross-validation (CV), eight classifiers were evaluated for their suitability with the dataset. The classifiers we tested are: Logistic Regression, Stochastic Gradient Descent, K-Nearest Neighbors, Decision Tree Classifier, Random Forest, Linear Support Vector Machine, Gradient Boosting and XGBClassifier. Mean score, standard deviation, precision, and average training time were recorded for comparison (as shown in table below).

	Average score	Std	Precision	Average training time
LogRegression	0.7998	+/-0.007	0.8559	1.230
SGDC	0.8009	+/-0.009	0.8538	0.528
KNN	0.8284	+/-0.005	0.8541	11.877
DT	0.8266	+/-0.006	0.7210	0.618
RandomForest	0.8633	+/-0.004	0.9255	11.060
LinearSVC	0.8025	+/-0.008	0.8556	4.854
GradientBoost	0.8222	+/-0.005	0.8854	12.134
XGBoost	0.8506	+/-0.006	0.9149	2.855

Note on abbreviation: SGDC: Stochastic Gradient Descent Classifier; KNN: K-Nearest neighbors; DT: Decision Tree Classifier

Table 1. 10-fold cross-validation results for each model used.

We identified 3 classifiers - Random Forest, Gradient Boosting and XGBoost – that performed well and were chosen for further exploration.

6.3. TRAINING THE MODEL

GridSearchCV was used on the training dataset to develop the best hyperparameter for each selected model. The models were then trained with the best hyperparameters and performance on the test dataset was assessed.

Model	Random Forest	Gradient Boosting	XGBoost
Hyperparameter	bootstrap=False, max_depth=9, max_features=10, min_samples_leaf=3, min_samples_split=10, n_estimators=300	max_depth=4, max_features=7, min_samples_leaf=15, min_samples_split=120, n_estimators=700, random_state = 10, subsample=0.8	max_depth=10, learning_rate=0.1
Train data score	**0.8305**	**0.8590**	**0.8752**
Test data score	**0.8289**	**0.8478**	**0.8552**

Table 2. GridSearchCV hyperparameter tuning result

It can be concluded that the models are not overfitting given that the training accuracy doess not exceed the validation accuracy by a significant amount. The confusion matrices of all models given are shown in Table 3. The number of false positives is important as overbooking according to a false prediction will result in the hotel's relationship with its customers being damaged. As such, we would like our model to have a low false positive rate.

Computed confusion matrix to evaluate the accuracy of a classification			Predict					
			Random Forest		Gradient Boosting		XGBoost	
			0	1	0	1	0	1
Actual	Train data	0	30.118	2.139	29.832	2.425	30.213	2.044
		1	7.253	15.902	5.386	17.769	4.873	18.282
	Validation data	0	12.850	975	12.635	1.190	12.757	1.068
		1	3.088	6.836	2.424	7.500	2.371	7.553

Table 3. Confusion matrix of each chosen model

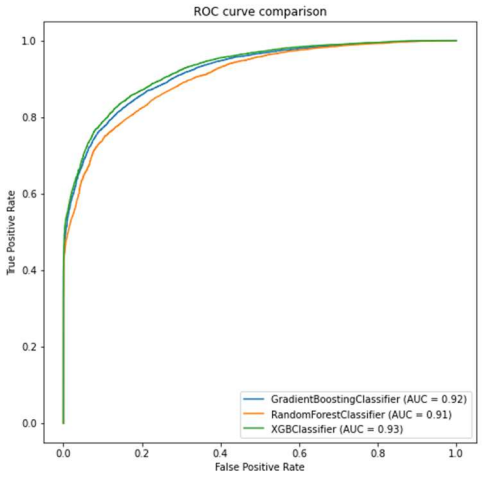
We attempted to combine these models in a number od ensemble models, in an attempt to obtain a better result. We tested all the possible combinations of the three models from two-model stacked to three-model stacked models. The ensemble model called Super Learner from ml-ensemble [6] where a 10-fold cross-validation was executed for each of the stacking models was also applied. Subsequently, the Stacking Classifier was applied. The stacking method showed a similar result to XGBoost.

```
Accuracy score: 0.8470 - ['Random Forest', 'Gradient Boosting']
Accuracy score: 0.8552 - ['Random Forest', 'XGBoost']
Accuracy score: 0.8552 - ['Gradient Boosting', 'XGBoost']
Accuracy score: 0.8543 - ['Random Forest', 'Gradient Boosting', 'XGBoost']

Best stacking model is ['Random Forest', 'XGBoost'] with accuracy of: 0.8552
```

7. EVALUATION

The ROC curves revealed that XGBoost is slightly better than the remaining models. After looking at the results, we decided to implement the XGBoost. The results are shown in the table below.



TRAIN

	precision	recall	f1-score	support
0.0	0.8611	0.9366	0.8973	32257
1.0	0.8994	0.7895	0.8409	23155
accuracy			0.8752	55412
macro avg	0.8803	0.8631	0.8691	55412
weighted avg	0.8771	0.8752	0.8737	55412

```
[[30213  2044]
 [ 4873 18282]]
```

VALIDATION

	precision	recall	f1-score	support
0.0	0.8433	0.9227	0.8812	13825
1.0	0.8761	0.7611	0.8146	9924
accuracy			0.8552	23749
macro avg	0.8597	0.8419	0.8479	23749
weighted avg	0.8570	0.8552	0.8534	23749

```
[[12757  1068]
 [ 2371  7553]]
```

Fig.6 ROC curves for the models: Random forest, Gradient boosting and XGBoost classifier after hyperparameter tuning and Classification report for the XGBoost

7.1. FEATURE IMPORTANCE

Using the feature importances calculated from the training dataset, we then wrapped the model in a SelectFromModel instance. We used this to select features on the training dataset, train a model from the selected subset of features, then evaluate the model on the testset, subject to the same feature selection scheme. We can see that the performance of the model decreases with the decreased number of selected features. We have plotted features and sorted based on importance.

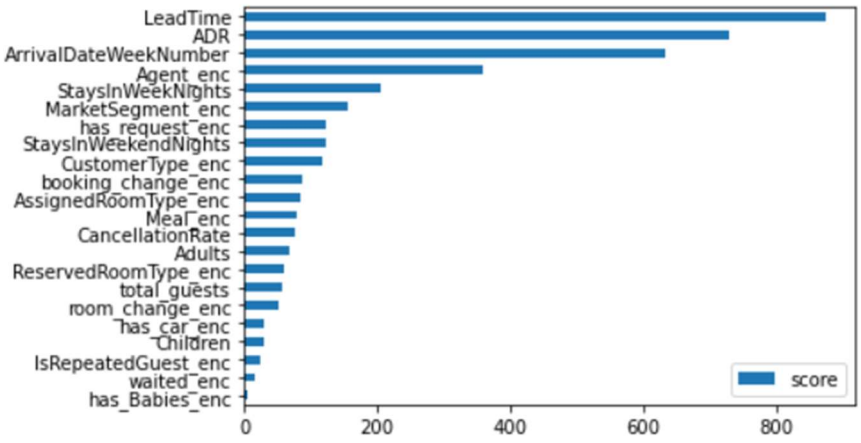


Figure 7: XGBoost Feature Importance Plot

If the Hotel is able to make use of factors that are within their control to implement measures such as improving their services or setting a favorable price, they will be able to reduce the frequency of booking cancellations.

8. DEPLOYMENT AND MAINTENANCE PLANS

8.1. PRE-DEPLOYMENT

We propose to deploy our cancellation forecast model as a module onto Hotel C's existing booking system. Our team will work on the module with user interface that closely resembles the UI of the existing booking system of Hotel C. This is to minimize the potential confusion among the first group of individuals that will be involved in testing out our pilot program.

We will also need access to other data to maximize the functionality of the module. With this data, we will be able to develop a module to display the expected occupancy rate for the given day in real-time. In turn Hotel C can make much more informed decision to overbook guests. This also entails Hotel C will need to conduct an additional cost-and-benefit analysis to find a new threshold and limit for overbooking rate and amend the current overbooking policies.

For the pilot program, we suggest working only with a limited number of employees on day-to-day. Focus group studies will be a highly useful tool to accommodate the needs of the end users. With their input, we can modify the module functions, especially the UI. The pilot program will be in a form of a non-biding simulation. While Hotel C conducts its business as it is, the selected booking staff will work with Addex Consulting's on-site analysts to make decisions based on our model and the new overbooking policies. Daily status update will be formed by the on-site analysts for both Hotel C's management and Addex Consulting. These analysts will also be responsible for real-time monitoring of the pilot program and training of the selected users.

We believe in-depth comparison assessment of the current and new practices needs to be done to allow further optimization of our module. This assessment will require participation from both Hotel C's management, the representative of the selected employees, and Addex Consulting analysts, as the final utility of the tool ultimately depends on Hotel C's needs. Furthermore, these assessment meetings will help us avoid proposing solutions that conflicts with Hotel C's business practices unbeknownst to us. In turn, Addex Consulting can aid Hotel C for deeper understanding of the results.

8.2. FULL DEPLOYMENT

The business environment, the pilot result, and Hotel C's confidence in our software will dictate the timeline of the full deployment. Depending on these aspects, we believe it is possible to cut the pilot program to 6 months from 12 months. Regardless of the result, we recommend hotel-by-hotel or group-by-group of users to conduct incremental stress testing of our module in real-life situation.

8.3. POST-DEPLOYMENT

With user feedback and business needs, Addex can help integrating Hotel C's CRM system. Namely, the hotel identifies customers who may cancel (based on the inefficient algorithm in place) and make offers to deter them from cancellation. We believe this decision on "which customers to contact, when to contact, and is it even necessary to contact?" can be automated. Additionally, as we are able to identify which factors are the most correlated with cancellation tendencies, we can make recommendations on what to offer these customers when the hotel wishes to avoid cancellation from these individuals.

9. CONCLUSIONS

9.1. CONSIDERATIONS FOR MODEL IMPROVEMENT

During our first meeting with Hotel C's management, we concluded their objectives can be summarized into two core aims. First, finding the balance of overbooking and cancellation. Second, reduce the cancellation rate. From data mining perspective, we identified reliable cancellation forecasting is vital. The second objective was less definitive as we are unaware of their cancellation deterrence efforts other than they contact customers and make offerings, such as free parking spot. Hence, we believed an insight into the characteristics most highly correlated with cancellation tendencies of the customers was a way to help Hotel C in making decisions related to cancellation rate reduction.

For our model, we experimented with multiple scaling methods, machine learning models, and different sets of features both given and engineered. Ultimately, our model was developed based on min-max scaling, XGBoost model, and 22 features as addressed in the Feature Importance section. LeadTime, ADR, and ArrivalDateWeekNumber are among the most important predictive features. The model scored accuracy of 0.8552 for our test data set, 0.8570 in precision, and 0.8534 in f1 score. We believe an improvement of the prediction can be achieved during the pilot program. We believe the main limitations for our study were lack of certain data/information and available computer power.

We also suggested a pre- and post- deployment plan in addition to the full deployment plan. These proposals primarily consist of deployment as a module; real-time monitoring; periodical assessment; on-site support; and potential integration into the existing CRM with further functionalities in the future.

10.REFERENCES

[1] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. Retrieved September 10, 2015, from https://themodeling-agency.com/crisp-dm.pdf

[2] Abbott, D. (2014). Applied predictive analytics: Principles and techniques for the professional data analyst. Indianapolis, IN, USA: Wiley.

[3] Zhu, W., Zeng, N., Wang, N., & others. (2010). Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations. NESUG Proceedings: Health Care and Life Sciences, Baltimore, Maryland, 1–9.

[4] Baak , M., Koopman, R., Snoek, H., Klousa, S., A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics. March 12, 2019 https://arxiv.org/pdf/1811.11440.pdf

[5] About Feature Scaling and Normalization. Sebastian Raschka'sWebsite. July 11, 2014 https://sebastianraschka.com/Articles/2014_about_feature_scaling.html