Machine Learning I

HOTEL CALIFORNIA

Group 11

12/12/2022

António Oliveira - 20211595

David Martins - r20201628

Inês Graça - 20211598

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

Table of Contents

ABSTRACT

Before the new grand opening in 2025, the new investors of Hotel California are making arrangements that include the development of an appropriate overbooking strategy to minimize cancellation vacancies. As that is one of their priorities, a predictive model that is able to identify whether a given Booking will be cancelled or not needs to be created. In order to do that, we started by exploring the dataset and understanding how each feature may influence the customers choice. We then dropped irrelevant features and split the dataset into training and validation, and variables into numerical and categorical. The next step was to apply three scalers and only then the Feature Selection. Here we performed tests for the numerical and categorical data, which along side with trial-and-error procedures helped us decided which features were the most important. When it came to modelling we used not only models given in theoretical and practical classes, such as *Logistic Regression* and *Random Forest*, but also some that we found online, for example, the *Stochastic Gradient*. Lastly, we evaluated each model's performance using metrics like *classification_report*, *accuracy_score* and *f1_score*. As a result, we were able to judge variables based on their importance for prediction purposes, which allowed us to drop variables like *ArrivalHour*, *PreviousReservations*, *DaysUntilConfirmation* and *DailyRateUSD*. After trying 10 different models, each with several parameters, we obtained the highest score with the *Random Tree Classifier*. Unlike we expected, the Feature Selection and Parameter Tuning only worsened the model performance. On the other hand, when comparing our results with other papers [1] [2], we realised that their model performance is similar to ours, as well as the highest performing model. In conclusion, with a 79% accuracy when predicting whether a given booking will be canceled or not, we believe our selected model should be able to minimize cancellation vacancies and avoid some of the problems that caused its demise. However, due to a gap of about a decade between the year of the training data and the year when this model is set to be implemented, we suggest a reassessment of the model as soon as there is enough new data to do so.

KEYWORDS

Model; Random Forest; Canceled; Performance; Machine Learning;

INTRODUCTION

Hotel California had been running for over 40 years before it filed for insolvency during the Covid-19 lockdowns. Before touristic operators, overbooking and free cancellation policies became standard practice in the industry. The Hotel had rare vacancies and successful bookings were made months in advance. However, during the 2010s, that started to change with cancellations becoming more and more frequent. This became a problem that, at the time, neither management nor staff were able to deal with, which was one of the reasons behind the Hotel closing its doors. In order to prevent that to ever be a problem again, the new investors want to prioritise the development of an appropriate overbooking strategy to minimize cancellation vacancies (before a grand new opening in 2025). For that reason, our job was to create a predictive model that is able to identify whether a given Booking will be cancelled or not. This report will describe the analytical processes and the conclusions obtained from that model.

BACKGROUND

In addition to the methods and algorithms learnt in class, we also searched the internet with the intention of finding better-performing algorithms, as well as methods that helped us to easily and in more detail, analyze the dataset.

PROFILEREPORT

This package provides a summarized report for the features in the selected dataset. This report gives alerts in relation to constant features, missing values, skewed distributions, imbalanced features, highly-correlated pairs, as well as features that have a lot of zeros.

GRIDSEARCH

To improve model performance, GridSearch method was used. This method, given parameters and a set of values for each parameter, computes all possible combinations of the given set, returning the best option for each inputed parameter. This allowed for a more efficient parameter tuning, even though it is computationally expensive.

EXTRATREESCLASSIFIER

This is an ensemble method similar to *RandomForestClassifier*, with the difference that randomness goes one step further in the way splits are computed. As in random forests, a random subset of candidate features is used, but instead of looking for the most discriminative thresholds, thresholds are drawn at random for each candidate feature and the best of these randomly-generated thresholds is picked as the splitting rule. This usually allows to reduce the variance of the model a bit more, at the expense of a slightly greater increase in bias.

SUPPORTVECTORMACHINE

Support vector machines (SVMs) are a set of supervised learning methods used for classification, regression and outliers detection. It performs classification by finding the hyperplane that best separates the classes. To identify this mentioned hyperplane, SVMs will maximize the distances between the nearest data point (either class) and the hyperplane. This distance is called margin.

SGDCLASSIFIER

This estimator implements regularized linear models with Stochastic Gradient Descent (SGD) learning. The main advantages of SGD are the efficiency and ease of implementation. On the other hand, the main disadvantages include its' sensitivity to feature scaling and that it requires a number of hyperparameters, such as the regularization parameter, and the number of iterations.

METHODOLOGY

OUR DATASET AND DATA EXPLORATION

In order to discover if the customers of Hotel California are likely to cancel their reservations, our group started by exploring the dataset, understanding how each feature may influence the customers choice, only based on their interpretation by ourselves. We based our steps in a CRISP-DM methodology (Figure 3).

Then we decided to drop the features that were always contant (*ArrivalYear* and *CompanyReservation*), hence irrelevant for prediction purposes, having also split the training and validation datasets as well as numerical and categorical variables.

FEATURE SELECTION AND ENGENEERING

Before performing Feature Selection, we took advantage of what was learnt in class and applied three scalers: *MinMax*, *Robust* and *Standard* Scalers.

In the Feature Selection part we performed three tests for numerical data and two for categorical. For numerical data, having used the Spearman Correlation's Matrix, Recursive Feature Elimination (RFE) and Lasso Regression. Spearman's allowed us to remove the highly correlated features. For categorical data, the Independence test was performed, followed by the Chi-Squared & Mutual Criterion Test. We decided to use these methods as they were the ones used during class, and in both homeworks.

Having decided which features are the most important, and performing a trial-and-error procedure for the ones we were not sure to use, the next step would be Modeling.

MODELING

Models given in theoretical and practical classes were used (Decision Tree, Logistic Regression, Naïve Bayes, Random Forest, Gradient Boosting and Neural Networks). In addition to these ones, after some research we also decided to try out Stochastic Gradient, Support Vector Machine (SVM) and Extra Tree, all in the form of Classifiers. These methods have already been described in the Background section.

PERFORMANCE EVALUATION

To evaluate each model's performance, metrics like *classification_report*, *accuracy_score* and *f1_score* were used.

To improve model performance and make it faster to test different combinations, Grid Search was used. When we just wanted to test a parameter that only had integer values, a for loop was used. This allowed for a more efficient parameter tuning.

EXPORTING

In the end the best-performing model was applied to the *hotel_test* dataset, having then exported the results to a csv file to be introduced into Kaggle.

RESULTS

Having started with all features and after carefully analyzing them, the first step would be to explore the data and look for missing and unlikely values, as well as inconsistencies.

There are no missing values in the dataset. However, there are some features with doubtful values. For example, it is unlikely that 10 babies are registered in the same booking, or a customer that made 47 previous reservations in this hotel. There is even a customer that took 224 days to confirm their reservation, and one that changed its booking 21 times. These are not common events. Still, after a careful consideration and a discussion with the professors, we decided to keep these observations as we concluded they are pretty unlikely, but can happen.

Following this, we decided to scale our data and perform Feature Selection methods for numerical and categorical data. To do this 3 scalers were used, and the one that produced the best results was *MinMax* Scaler, having decided to perform the following steps according to it.

Starting with numerical data, our tests allowed us to remove *ArrivalHour, PreviousReservations, DaysUntilConfirmation* and *DailyRateUSD*, as it is shown in Figures 1 and 2. There were some features that generated some doubts, as for example *CountryofOriginAvgIncomeEuros (Year-2), CountryofOriginAvgIncomeEuros (Year-1)* and *CountryofOriginHDI (Year-1)* ,which were highly correlated between themselves, as showed by Spearman's (Figure 1), but were also quite important for prediction (Figure 2). Our solution would be to try different combinations with these three features, having decided to remove *CountryofOriginAvgIncomeEuros (Year-2).* The same issue was repeated with *FloorReserved* and *FloorAssigned*, where again, both had a high correlation (Figure 1) with each other, but both were also quite important for prediction purposes (Figure 2).

For categorical data, the performed tests allowed us to decide to keep *Children, AffiliatedCustomer, OnlineReservation,* and *ParkingSpacesBooked*. In the Chi-Squared and MIC test, our decision was to use a threshold equal to 6, as this was the one that allowed a higher performance when evaluating the models.

After trying 10 different models, each with several parameters, we obtained a maximum of 0.7909 in Kaggle, with the Random Tree Classifier, with *['Children', 'OnlineReservation', 'ParkingSpacesBooked', 'ArrivalMonth', 'ArrivalWeekNumber', 'WeekendStays', 'WeekdayStays', 'BookingChanges', 'BookingToArrivalDays', 'SpecialRequests', 'FloorReserved', 'FloorAssigned', 'DailyRateEuros', 'CountryofOriginAvgIncomeEuros (Year-1)', 'CountryofOriginHDI (Year-1)']* and three parameters (max_depth = 20, n_estimators = 500, random_state = 34). Our f1-score with these was 0.6991005723630418 in the validation (Figure 4).

DISCUSSION

The obtained results were not what we expected, as it appeared that *Feature Selection* and *Parameter Tuning* only worsened the model performance. To better understand which were the most important features, we used one of the *RandomForestClassifier()* attributes, *feature_importances_*. This knowledge allowed us to try and remove features with low importance, like *AffiliatedCustomer*, while trying to improve the model performance. After this process, we just managed to get a better result in Kaggle.

We consider our findings significant, but also believe that in the time between the dataset's collection and the predictions much has changed. Having this in mind, we believe that we successfully predicted the majority of cancelations.

Even when comparing with other papers [1] [2], written about the same subject, one may conclude that their model performance is similar to ours, as well as the highest performing model.

CONCLUSION

In conclusion then, our selected model had approximately 79% accuracy when predicting whether a given booking will be canceled or not on the testing data. By using this model the Hotel California should be able to minimize cancellation vacancies and avoid some of the problems that caused its demise.

However our team has a little skepticism regarding the validity of our training data. The model was trained using data from 2016, but it will have to predict whether or not bookings from 2025 onwards will be canceled or not. This is a gap of about a decade, therefore we fear that tourist patterns will have changed by then. We suggest for a reassessment of the model as soon as there is enough data after the grand opening in 2025, as failure to adapt to new costumer patterns was one of the reasons for the fall from grace of Hotel California in the 2010's.

We hope that this research will be helpful for the team of machine learning experts that does the reevaluation after the grand opening, as most findings in our research should still be relevant by then.

REFERENCES

[1] N. Antonio, A. de Almeida and L. Nunes, "Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 2017, pp. 1049-1054, doi: 10.1109/ICMLA.2017.00-11.

[2] M. S. Satu, K. Ahammed and M. Z. Abedin, "Performance Analysis of Machine Learning Techniques to Predict Hotel booking Cancellations in Hospitality Industry," 2020 23rd International Conference on Computer and Information Technology (ICCIT), 2020, pp. 1-6, doi: 10.1109/ICCIT51783.2020.9392648.

APPENDIX

Figure 1 - Spearman's Correlation Matrix with all Numeric features

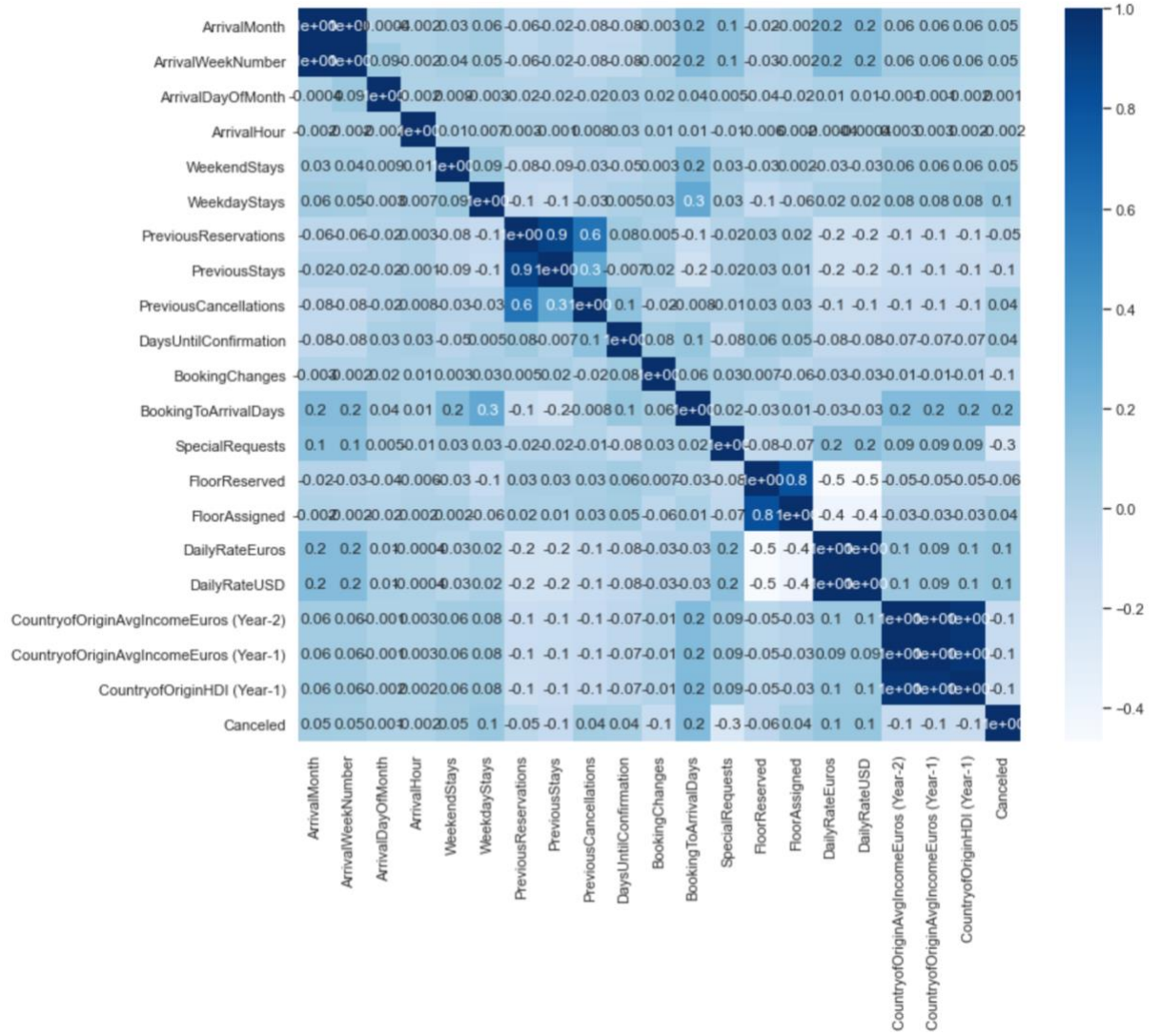

Figure 2 - Lasso with all Numeric features

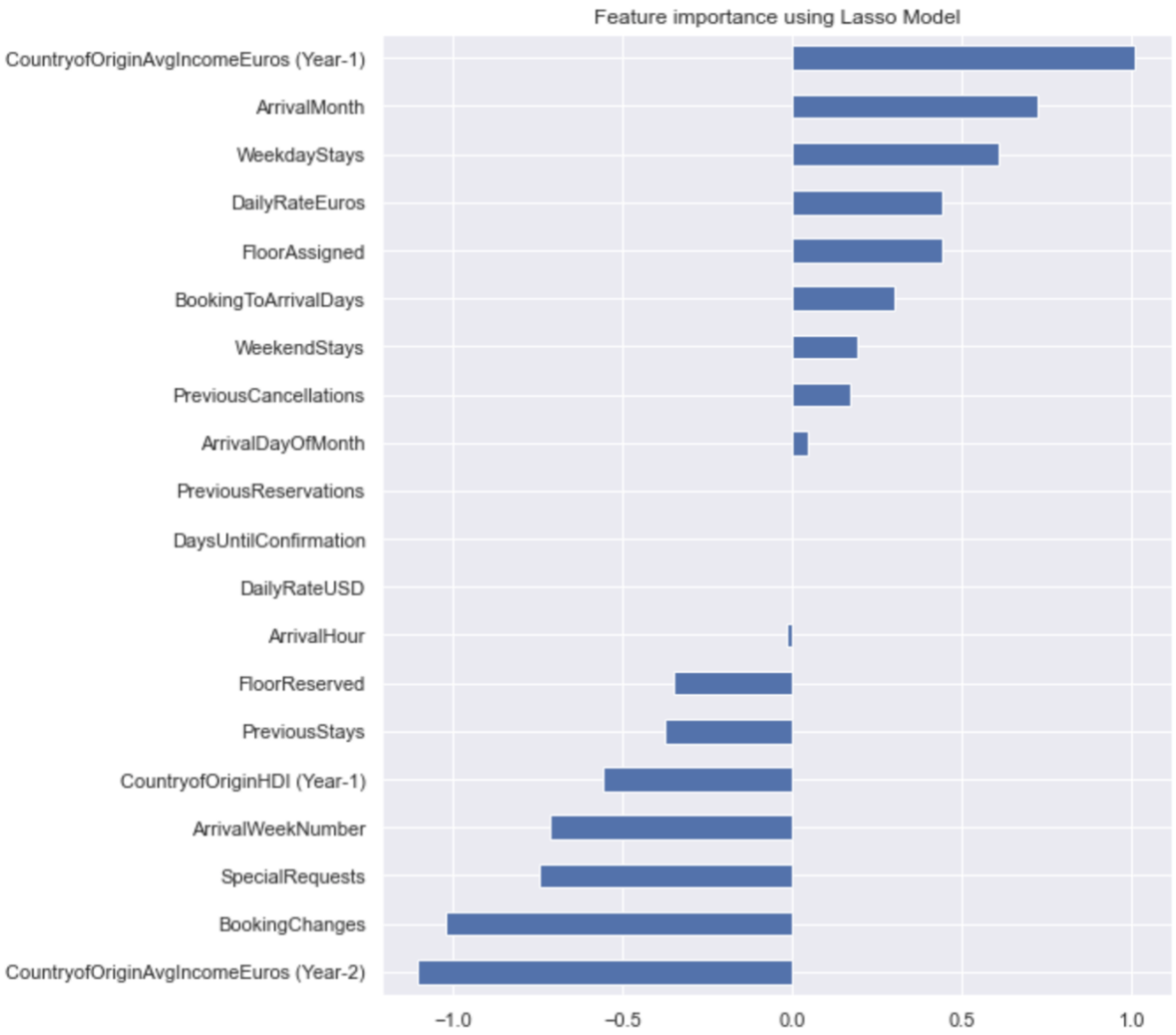

Figure 3 – CRISP-DM methodology

Figure 4 – Classification Report of the best prediction (validation)

	precision	recall	f1-score	support
0	0.87	0.81	0.84	2346
1	0.65	0.75	0.70	1139
accuracy			0.79	3485
macro avg	0.76	0.78	0.77	3485
weighted avg	0.80	0.79	0.79	3485

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa