Hotel Cancellation Prediction

Zi Gu, Ivy Wang, Yiqi Ye

Introduction

Customers canceling reservations severely impact the results of hotels and dining facilities. It not only damages revenue but also wastes the food in stock and personnel expenses used to provide services.

The goal of this project is to use data consisting of over 100,000 reservations to categorize reservations into those that will and will not be canceled. The data uses actual reservations that were made in several cities across Portugal. This is a project that predicts cancellations. It's also very meaningful for society because it can later be used for various purposes such as analyzing the causes behind the cancellations.

According to our data, 40,325 reservations are canceled and 68,460 are not. The cancellation rate is 37%. We hope with our model and analysis, we could provide thorough and robust recommendations that could reduce the cancellation rate while finding out influential features in determining whether there would be a cancellation or not.

Methods

Dataset description

The original data has over 100,000 reservations which came with 27 features which are: hotel, lead_time, stays_in_weekend_nights, stays in weeks nights, adults, children, babies, distribution channel, is_repeated_guests, previous_bookings_not_canceled, reserved_room_type, assigned_room_type, booking_changes, agent, company, etc. And label as is_cancled. The data has a total of 40,000 reservations from resort hotels and 80,000 reservations from city hotels. The test file has over 10,000 reservations which are all provided by Quevico AI.

This machine learning project is performed based on the hotel dataset which consists of 40,000 reservations from resort hotels and 80,000 reservations from city hotels. The resort hotel data comes from the Algarve region and the city hotel data comes from Lisbon, the capital city of Portugal. Since the data was obtained from the hotels' property management system (PMS), we believe that the data is valid and reliable for us to study.

Task and baseline model

Our goal is to provide predictions on future customer cancellations based on historical data. We defined this as a binary classification problem since there are discrete binary outcomes between 0 and 1, we denoted 1 as succeed to show and 0 as failed to show.

We used logistic regression as our baseline model, to perform a binary classification task, the reason we choose logistic regression is that it is a widely used technique so as efficient. Logistic regression does not consume too many computational resources. As a baseline model, logistic regression does not need any hyperparameter tuning and is easy to be regularized and could also provide us some basic insights on variable importances. Based on the given reason, we choose logistic regression as our baseline model before we move to other more complex algorithms.

Other models

After building our baseline model, we decided to move on to tree-based models. Tree-based models are popular in classification problems due to their adaptability and several advantages. Tree-Based models can be used for any type of data, whether they are numerical or categorical and they are robust with handling missing values and outliers. Besides, using a

tree-based model requires no data preparation like standardization and other forms of transformation.

We perform two kinds of tree-based models using the ensemble method in the scikit-learn package. We implemented bagging by training random forest models and implemented boosting by training models like Gradient Boosting Classifier and Adaboost. All the tree-based models outperformed logistic regression but random forest performed the best.

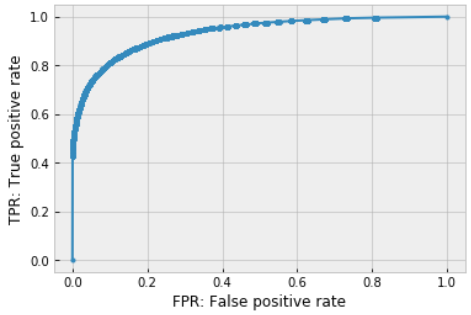
High-level code explanation

The actual coding process included data cleaning, preprocessing, feature selection, training, and validation. In order to select the most relative features to our model, we did a basic exploratory analysis and drew a couple of univariate and bivariate graphs between features. We found that lead_time has a strong correlation with the dependent variable and a couple of features like stays_in_weekends_nights, stays_in_week_nights do not interfere with cancellation much. Besides, drawing boxplots helped us detect outliers among the features, we ultimately removed the data which have a negative average daily rate and rate greater than 1000 dollars. We also removed some questionable data such as the reservation with only children and babies but no adults, or request_parking_spaces over five with only two adults. We processed our data cleaning part after consulting experienced people in the hotel industry and ended up with 21 total features of which 14 are numerical and 7 are categorical. For all of the 7 categorical features, we performed label encoding to transform the categorical features into readable data for our model. We put our data into our model to train by using 4-fold with 75% training data and 25% testing data. We achieved our best result using random forest with tuned hyperparameters.

Evaluation and Analysis

Evaluation

Both validation and test results are evaluated on areas under the AUC curve between the predicted probability and the observed target.



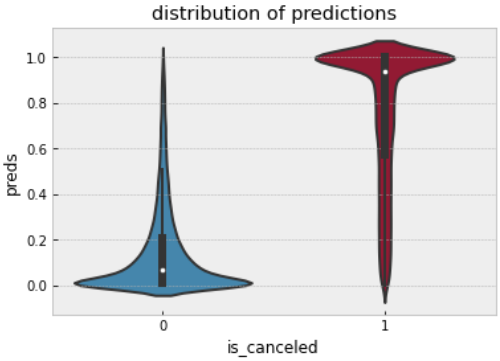


Figure 1 Figure 2

For business in hotel services, we do not want to make the following mistakes: 1) We falsely assume people will cancel the reservation which could inconvenience the customer upon a false cancellation or lead to overbooking the hotel. 2) sell the room to customers who would like to cancel reservations and leave many empty rooms. So both false negatives and false positives are expensive to us. Based on this understanding, the metric chosen for the task is the AUC (Area Under the Curve) score and precision score. Both of these evaluation metrics do a good job of evaluating classification problems. As a result, we achieved an AUC of 0.934 from

the random forest model (Figure 1), which is very close to the perfect AUC. Since the higher the AUC, the better the model is. It tells us our random forest model is highly capable of distinguishing whether a reservation will be canceled or not.

From the distribution of the predictions plot (Figure 2) we could see that our random forest model is very competent in predicting cancellations. The medians and interquartile range of predicted values are all aligned with the actual data points. Two violin plots are well separated.

Finally, the AUC on the testing data is approximately 0.85. This could seem slightly overfitting as the testing AUC is about 8% below the training data. However, there could be multiple reasons why this is the case -- a difference in data distribution between the training and testing set, for example, could explain such discrepancy. Is the model slightly overfitting due to a skewed distribution in seasonality? Questions such as these could be looked into further for improvements on our task.

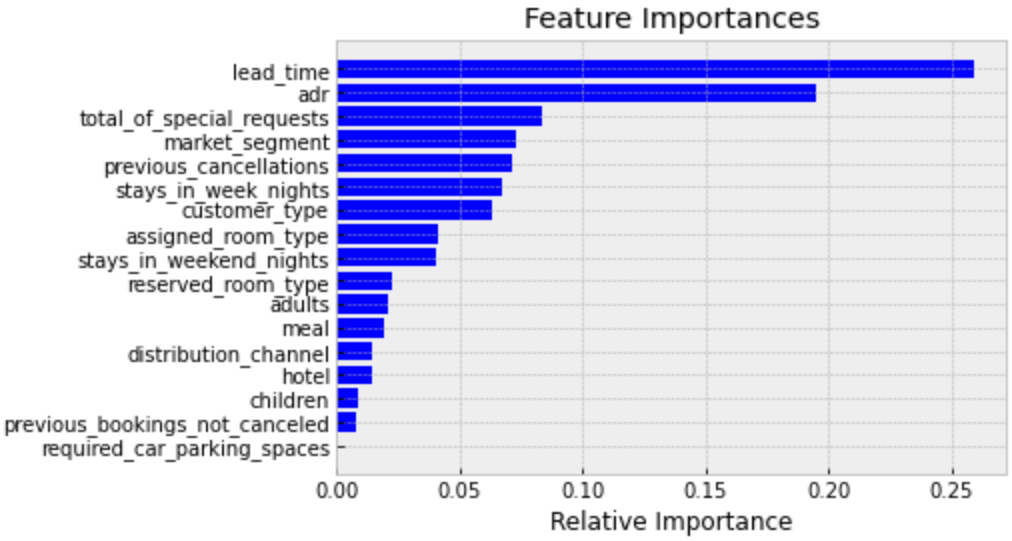
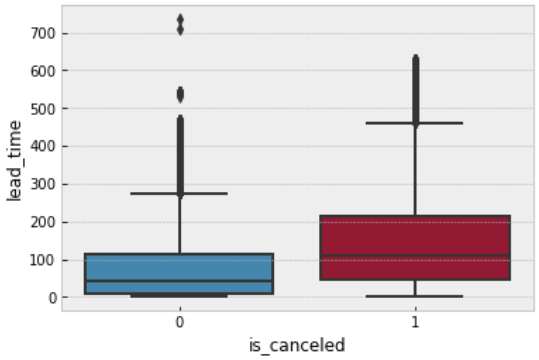


Figure 3

Analysis

Generating the model to make accurate cancellation predictions is not enough for answering the project questions. We also wonder why the cancellation is being made and which features are most important in determining the prediction. Knowing feature importance can provide multiple benefits. 1) It helps us get a deeper understanding of our model. 2) we can use this to do feature selection by removing less significant variables. 3) By removing the low significance feature can expedite the model running time. 4) It supports future analysis of the variables and setting ground_truth evidence for business process improvement. The way we plot a feature importance chart is by using the scikit-learn package.



As figure 3 shown above. The most important feature is the lead-time. The lead-time here is defined as the time between the date the customer makes the reservation to the reserved date. The figure on the right compares the distribution between cancellation. The boxplots reveal that the longer the lead-time is, the more likely the customer is going to cancel that reservation. Hotels can use this insight to maybe regulate how early customers can make a reservation. For example, customers cannot book a room with a check-in date 200 days later. The threshold here can be determined by considering both business tradeoffs and the third quartile of the boxplot above.

The second most important feature is the average daily rate (ADR). We plotted the distribution of ADR for both canceled and not-canceled reservations. Two distributions resemble each other. This means that there may be more hidden information that is unseeable from distributions. Thus we cannot recommend hotels to increase or decrease ADR for reducing cancellations.

The third most important is the total of special requests. In the exploratory data analysis, we found that there are more people that have at least one special request within the group that did not cancel the reservation. With this insight, hotels could try to increase their capability of providing more specialized services. This way does not only reduce the cancellation rate but also increase hotels' competitiveness in the market.

Related work

In order to perform a good predictive analysis in hotel cancellation, we referred to some similar research paper in health no show predictions. Laplan-Lewis [1] in *No-show to primary care appointments: why patients do not come* mentioned missing reservation in the medical environment, wasted health care dollars, and inefficiency use of provider's time. Not only in the Medical environment, but no-show also causes the same consequence in the hotel industry, she suggests using the outcome to make future adjustments in the reservation system.

Besides conducting research, we also consulted with professionals in the feature selection process, based on professional suggestions and experience, we were able to remove the outliers in multiple features such as negative ADR and extreme parking space. Efficiently removing outliers highly increased our model's performance.

Discussion and Conclusion

We ultimately implemented a random forest for the classification task. According to our model and the threshold we set, the result shows that 22% of people will cancel the reservation and 78% will stay. Every machine learning task should have a business problem behind it, our question is to study what leads to the results. There is no doubt that advanced technologies and fast speed internet simplified most of the booking services but the cost increased pressure for the hotel industry to face more cancellation rates than before. According to the evaluation section, we talked about the feature importance in our model. We learned that lead_time, ADR, a total of special requests interfere with the cancellation the most. In terms of the lead_time, hotels and agencies tended to be flexible with cancellation dates due to the previous economic crisis, they were willing to take everything in but once the cancellation happened close to the reservation date, there are higher risks to face financial loss. We recommend hotels to cut off the lead time and offer more specialized services.

By doing this project, we learned how to efficiently conduct feature engineering to optimize our model. We also practice our skills in supervised classification tasks, strengthening our understanding of evaluation metrics. Future work needed to be conducted in finding more insights on the business side. If we are able to utilize our model and adjust the reservation system accordingly, it will have a huge impact on the hotel industry. Some future work includes user research and heuristic evaluation as well as deep analysis in hotel profit models which can be planned for whoever is interested in this area. From our perspective, we believe this project has the potential to make some changes to the hotel industry.

References

[1]
Kaplan-Lewis, Emma and Percac-Lima, Sanja, “No-Show to Primary Care Appointments : Why Patients Do Not Come,” Journal of primary care & community health, vol. 4, no. 4, pp. 251–255, 2013.