Canceled or Not? Machine Learning Tells You!

Han Chen[1], Supervised by Hien Tran[2]
[1]Tanwei College, Tsinghua University
[2]Department of Mathematics, NC State University

Introduction

In one of his paper in 1959, Arthur Samuel coined the term “machine learning" (ML) as “the field of study that gives computers the ability to learn without being explicitly programmed" [1]. Today, ML is a powerful tool to extract information from data and have a wide range of real-world applications.

The online hotel reservation channels caused increased cancellations, which is a revenue-diminishing factor for the hotels to deal with. In this work, we built ML models to make prediction on whether a customer is going to cancel the reservation, in an attempt to offer suggestions on room management for hotels.

Objectives

The goal is to apply established ML algorithms to predict the cancellation of hotel reservations.

Besides, we also want to figure out the most important features that affected a customer’s decision, so that hotels can have a quick judgement.

Data Description

The hotel reservations data are from *Kaggle*. There are 14 categorical variables and 4 numerical variables. The variable `booking_status` is used as label. The distribution of some variables is shown in Figure 1.

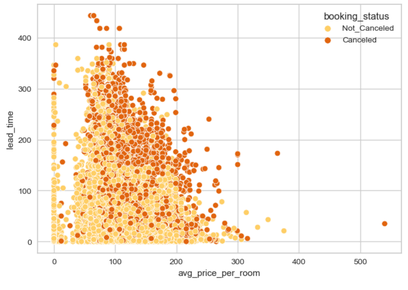

Figure 1: Distribution of room price, lead time, and their relations with booking status

It is difficult to draw conclusions from such a single figure. Therefore, ML models are applied to make prediction.

Methods

I. Multilayer Perceptron (MLP) A general neuron-like processing unit is:

$$a_i = \varphi(\sum_j w_{i,j}x_j + b_i), \tag{1}$$

where x_j’s are the inputs, $w_{i,j}$’s the weights, b_i the bias, φ the activation function, and a_i the unit’s activation. A neural network is a combination of those units (Figure 2).

We obtain w_{ij} by solving an optimization problem:

$$w_{ij} = \text{argmin}_{w_{ij}} \left\{ \frac{1}{d} \sum_{i=1}^{d} (Out_i - Label_i)^2 \right\} \tag{2}$$

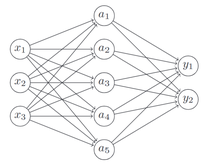

Figure 2: A multilayer perceptron with one hidden layer

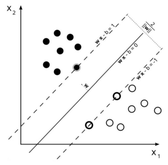

Figure 3: SVM. *Source: Wikipedia.*

II. SMV Support Vector Machine (SVM, Figure 3) tries to find the best hyperplane that separate the data as much as possible. We get $\mathbf{w}$ by solving an optimization problem:

$$\min_{b,\mathbf{w},\mathbf{e}} \|\mathbf{w}\| + C \sum e_i, \text{ Subject to}$$
$$y(i)(\mathbf{w}'\mathbf{x}(\mathbf{i}) + b) \leq -1 + e_i,\ e_i \geq 0,\ i = 1, \ldots, d. \tag{3}$$

C is a parameter we need to tune via cross-validation.

III. Decision Tree & Random Forest A decision tree maps a series of features to the labels. Two algorithms, CART and ID3, are mainly used in practice. The former is based on entropy and the latter is based on Gini impurity.

A random forest fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. [2]

Results

The correlation of variables in hotel reservations after data preprocessing is shown in Figure 4.

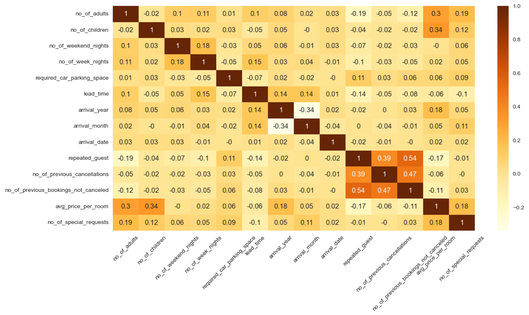

Figure 4: Correlation heat map

Notice that all the correlation are not very significant. Therefore, we include all the features in the model at the beginning.

Ten different classification models were fitted and tuned to get the highest precision.

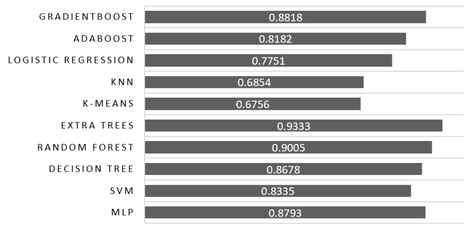

Figure 5: Accuracy of different models.

Among all the 10 models, extra trees was the best, with accuracy reached 0.9333. MLP, decision tree, random forest, gradientboost are also good models. KNN and K-means had relatively low accuracy, which means they are not propriate choices in this case.

Feature Selection & Model Optimization

In the original data, points are in a high-dimensional space. To make training more effective and the model more explainable, we selected several important features based on several methods. The selected feature and their importance score is shown in Figure 6.

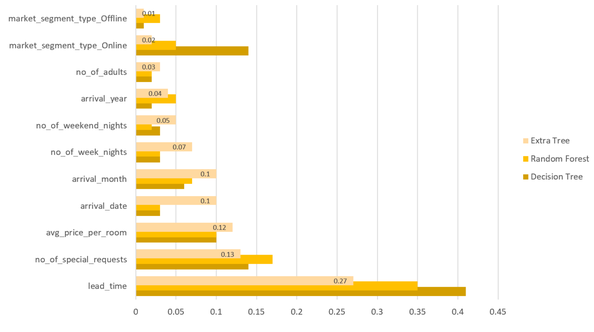

Figure 6: Important scores of selected features

We noticed that lead time is the most important among all the features. Actually, customers who take reservations with a long lead time are more likely to cancel it, which is consistent with the commonsense.

Then we used selected features to optimize established model. Results are shown in Figure 7.

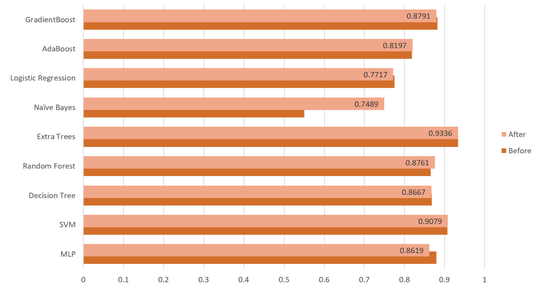

Figure 7: Comparison of model accuracy before and after feature selection

In general, the accuracy scores before and after feature selection are similar, with the Naive Bayes model even better (although still not satisfactory), confirming that little had lost due to the dropping of unimportant features. Besides, the training time decreased significantly. In the case of SVM, the training time shrank from 388 min to 347 min, a decrease about 10.6%.

References

[1] A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3:210–229, July 1959.

[2] S.S. Haykin. *Neural networks and learning machines*. Prentice Hall, 2009.

Discussion

In this work, we made prediction in hotel reservation via ML classification model with accuracy reached 0.93. Besides, ML models also reveals important factors that influence a cancellation decision. Our work also shows that dropping unimportant features results in no loss and is a more effective strategy in modeling.