Project Proposal

CHEME6880 / SYSEN5880

Hotel Booking Optimization

Zhiming Xie (zx284) and Jaeha Kim (jk2894)

# Introduction

Hotel booking is ubiquitous in our daily life. With the dropping frequencies of Covid-19 lockdown, local governments encourage more people to travel so as to boost the local hospitality industry. According to TourismReview, hotel reservations were currently on an upward trend with a 46% global occupancy rate in April 2021. This was a significant increase compared to a low of 13% in the same month in 2020. Leave alone that the occupancy rate in January 2022 had increased up to 47.8%. With the increase in number of hotel bookings, our team’s common goals are to find out the optimal prediction models to address the following problems:

**Original Questions:**

1. **Which type of customer has the highest cancellation rate in terms of hotel booking?**
2. **What type of distribution channel has the highest cancellation rate?**
3. **What is a reasonable price range should a hotel set for traveling seasons to maintain a high occupancy rate?**
4. **Make a cancellation rate prediction on the validation dataset.**

**Corrected problems:**

1. **Which type of customer has the highest probability of canceling the booking?**
2. **What type of distribution channel results in the highest probability of canceling the booking?**
3. **With the dataset, which machine learning model gives the most accurate prediction of the hotel booking cancellation?**

The goals listed above are vital to the hospitality industry. They are the important factors in allocating human and financial resources to revive the local economy in the aftermath of the global pandemic. Our team hypothesized that certain customer profiles, booking timeline, and travel type will have a strong correlation to booking cancellation rate. Based on the datasets from Kaggle, we found “Hotel Booking Demand Datasets” which contains many variables related to the hotel booking process.

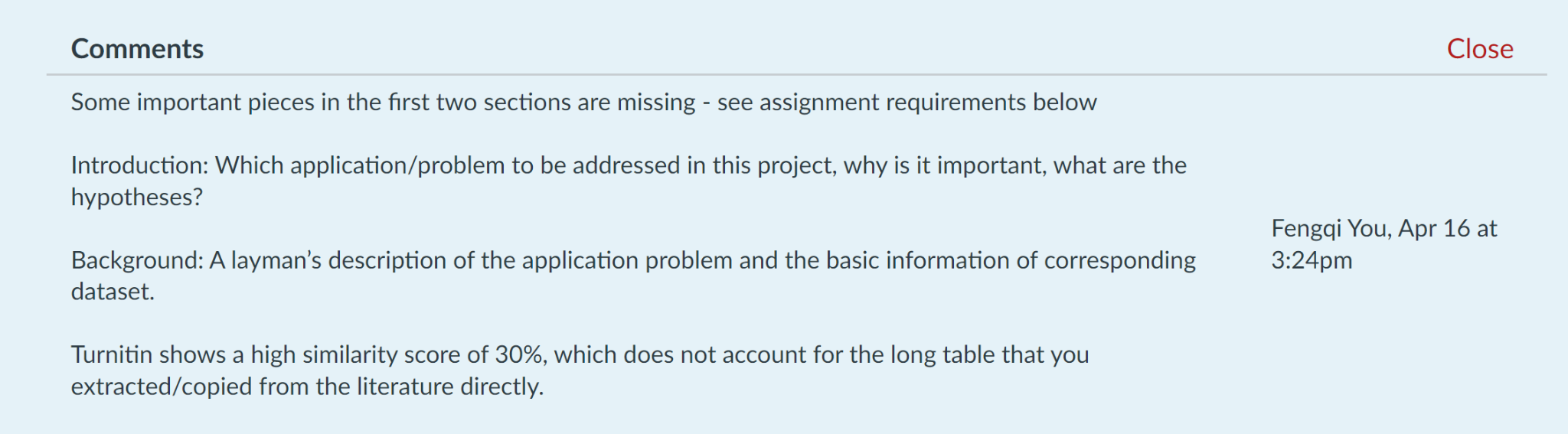
Hypothesis : since the data ~ binary, regression problem, polynomial ~ /// after plotting data set (observation, technical description) there are outliers or not, dimensions, noise, high-correlation between features, pca,

machine learning hypothesis: regression or classification; technical terms in machine learning; not the guessing

Why is it important? How many dimensions, noise, trend, outliners, correlation?

Background of the dataset is missing same as “why is it important”

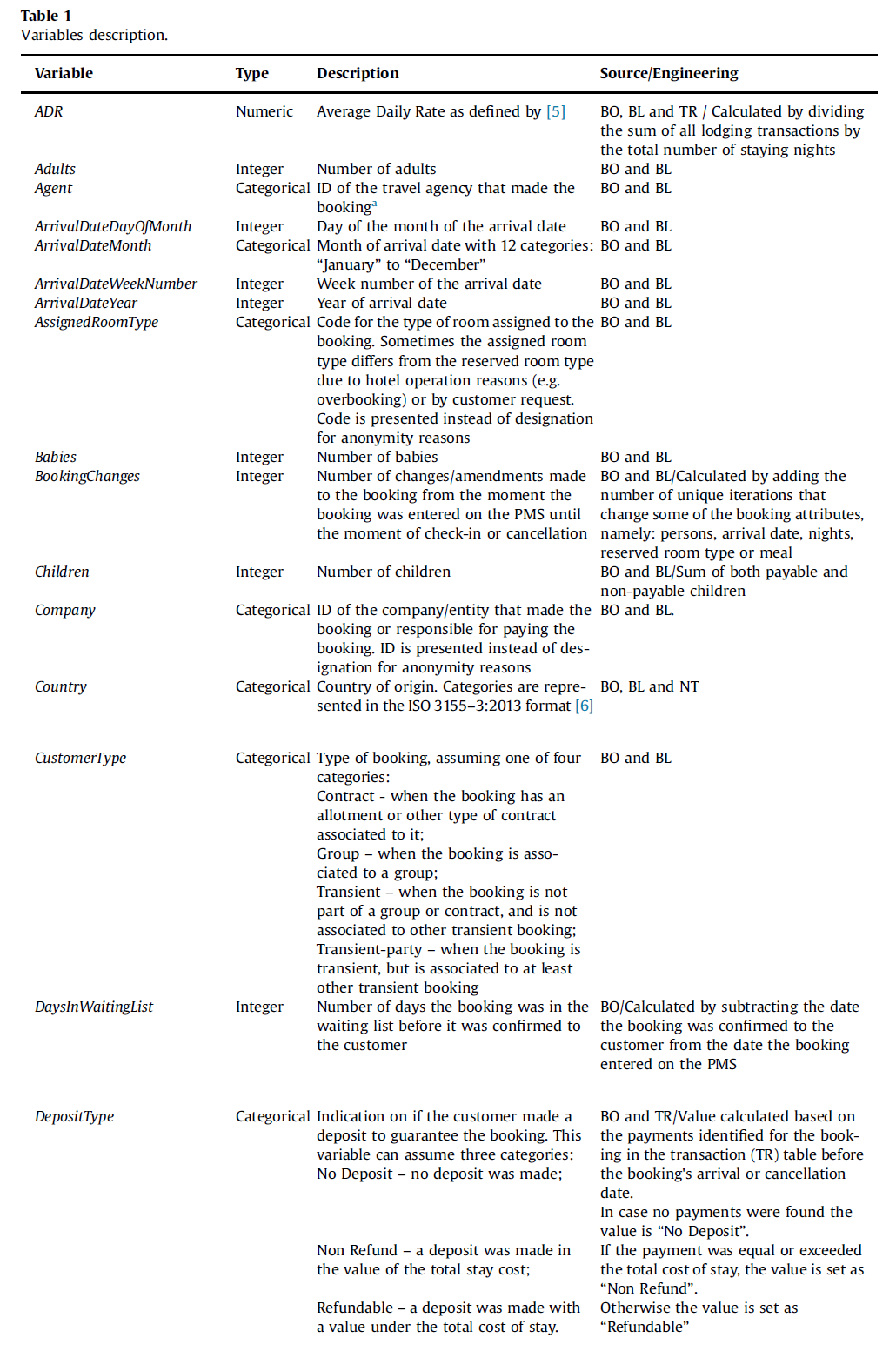
PCA also gives the PC

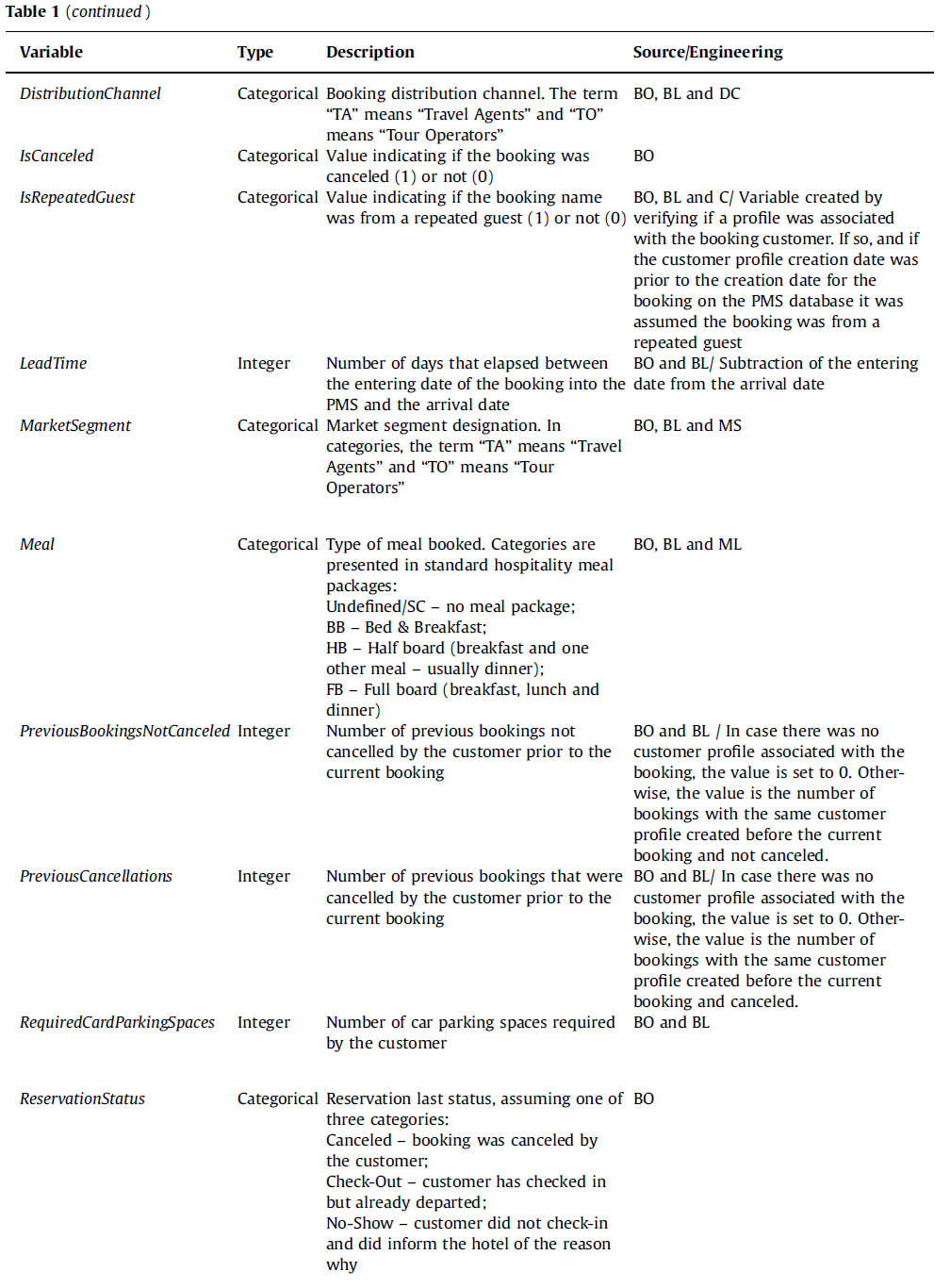


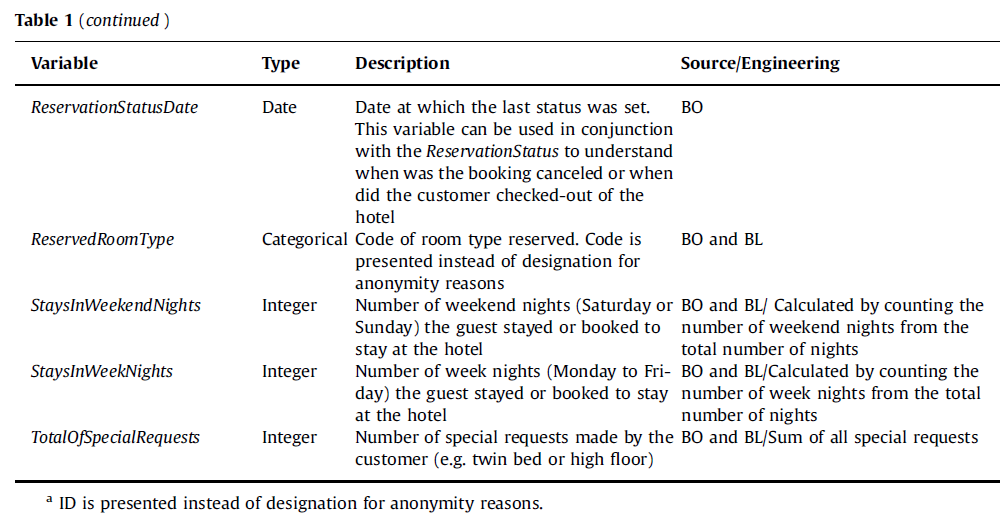
# Background of Hotel Booking Demand Dataset

In the hotel booking process, the most important goal for hotels is minimizing cancellation rates. No-show and cancellation is related to the hotels’ profit and to minimize this, hotels get some portion of extra bookings. However, it is difficult to decide how many extra bookings they need and it is directly related to predicting the cancellation rate. Moreover, it is also important to decrease the number of cancellations. In which conditions, customers are likely to cancel their bookings? This is an interesting question to both sides, hotels and customers. Cancellation might happen because a customer finds a better option, or dislikes something on the booking, or changes the travel plan. Understanding the reason behind the cancellation will help hotels to improve their services and make customers more satisfied. And this important problem can be solved by tracking booking records using classification algorithms.

We used the “Hotel Booking Demand Dataset” generated by Nuno Antonio, Ana de Almeida, Luis Nunes in 2019. This dataset describes 40,060 observations on resort hotel bookings and 79,330 observations on city hotel bookings with 31 variables. The data is collected on bookings arrived between the 1st of July of 2015 and the 31st of August 2017. Since the data was collected before Covid-19 outbreak, it didn’t include Covid-19 related variables. There are 14 categorical variables among 31 variables. More detailed descriptions on the variables are presented in the Table1 (2019, Antonio et al.).







# Method

Method1: Logistic Regression

Description: It is a classification algorithm and a process of modeling a binary outcome given an input variable. The dataset used in this project comes with labels and some problems we focus on are classification problems.

Method2: Decision Tree

Description: Decision trees are defined by recursively partitioning the input space by a set of nested decision rules, and defining a local model in each resulting region of input space. The overall model can be represented by a tree, with one leaf per region. It is easy to interpret and also can easily handle mixed discrete and continuous input values. However, the models do not predict very accurately compared to other models and it is unstable. The value of a decision tree to our study is to visualize the tree and understand the significance of each feature.

Method3: KNN

Description: To classify a new input, KNN finds the K closest examples to the input in the training set, and look at their labels, to derive a distribution over the outputs for the local region around the input. However, KNN requires the dataset to be linearly separable.

Method4: Random Forest

Description: Random Forests try to decorrelate the base learners by the same algorithm learning trees based on randomly chosen subset of input variables, with a randomly chosen subset of data cases. By modifying an equation, the feature split dimension is optimized over a random subset of the features.

Method5: Ada Boost

Description: AdaBoost is an iterative ensemble method which builds a strong classifier by combining multiple poor-performing classifiers. It is less prone to overfitting as the input parameters are not jointly optimized. To our benefit, we can use AdaBoost to reduce variance and bias to achieve higher accuracy in prediction.

Method 6: Gradient Boosting

Description: Gradient boosting can be used for predicting continuous target variables and also categorical target variables. It gives us the advantage of understanding regressor other than just classification. It minimizes the loss function of the model by adding weak learners using gradient descent. In our case, there may exist several weaker learners such as “Baby”, “Breakfast”, and etc. Gradient boosting comes in handy by reducing the loss function.

Method 7: XgBoost (Extreme Gradient Boosting)

Description: XgBoost is an efficient and very widely used implementation of gradient boosted trees. It adds a regularizer on the tree complexity, it uses a second order approximation of the loss instead of using a linear approximation.

Method 8: ANN

Description: ANN is capable of taking sample data rather than the entire dataset to provide the output result. This gives us the advantage of lowering the computational cost and coming out with an output in a shorter time.

Method 9: Support Vector Regression

Description: Similar to SVM, SVR is the adapted form of SVM when the input variables are numerical. It can use multiple classifiers trained on various types of data and improve the prediction accuracy by measuring the confidence in classification.Moreover, SVR can work with non-parametric data which fit in our case.

# References

Europe Hotel Satisfaction Score

: https://www.kaggle.com/datasets/ishansingh88/europe-hotel-satisfaction-score

Reference Project

: https://www.kaggle.com/code/steinerhslu/europe-hotel-review-gioele

Hotel Booking Prediction

: https://www.kaggle.com/code/niteshyadav3103/hotel-booking-prediction-99-5-acc

Dataset Source, Feature Table

: Antonio, Nuno, Ana de Almeida, and Luis Nunes. 2019. “Hotel Booking Demand Datasets.” *Data in Brief* 22 (February): 41–49. <https://doi.org/10.1016/j.dib.2018.11.126>.