I.M.L Hackathon 2023

Amit Benbenishti, Adi Shnaidman, Shoham Mazuz, Omer Ventura

The Data Set

The data set - Agoda_training.csv, consists of 58,659 records and includes 38 features related to booking details. Among these features is the variable of interest, "cancellation_datetime," which indicates the date of cancellation if applicable.

The data set is high dimensional, making analysis and extracting meaningful insights more complex (like determining feature significance). The large data size of 58,659 records requires efficient processing and it also contains missing values, requiring careful handling during preprocessing. The data consists of outliers and accuracy can be affected. The temporal and geographical aspect adds complexity in analyzing time and currency-dependent patterns.

The data was split into two separate datasets: 70% of it was retained as the training set, while the remaining 30% was allocated as the test set.

1.2.1 Cancelation Prediction

Data Cleaning and Pre-Processing

Checking the train set, we encountered several parameters we found interesting:

- Our initial goal was to validate the values of all features by addressing NULL values and handling invalid entries. For example, when we encountered cases with a rating of -1 stars, we replaced this value with the mean rating value.
- Additionally, we identified and eliminated irrelevant features, such as the customer's ID, that we believe were not useful for predicting cancellations.
- To handle categorical variables, we utilized dummy variables and transformed them accordingly. We also identified and computed features that were deemed to have a significant correlation with cancellations, such as the penalty amount and the distance between the check-in date and the date of booking.
- As part of our feature exploration, we conducted various data manipulations to identify relevant factors. We specifically focused on investigating the correlation between the rate of penalty relative to the payment and the actual penalty itself.

Decisions regarding the learning method

Initially, we experimented with several classification learners, including logistic regression, soft-SVM, and decision trees. However, their prediction performance was relatively poor. Recognizing the need for a more effective approach, we decided to employ an ensemble of weak learners. Consequently, we opted for an ensemble of 500 decision-stumps with a depth of 2 after experimenting with hyperparameters. We used Adaboost method to determine the trees thresholds. This decision was based on the understanding that the dataset consisted largely of binary questions represented by dummy values. As a result, we observed significant improvement and stability in the performance of our model.

Score Prediction

Finally, we used our fitted model to predict the cancellation for our test set, the F1-macro score was: 0.71
We attribute the achieved score of our model, which falls within the mentioned range and not beyond, to several factors. Firstly, it is possible that the given features are insufficient in capturing the intricacies and variations of the reservations and the individuals making the bookings. The limited scope of the provided features may have hindered our model's ability to achieve higher performance.
Additionally, we acknowledge that time constraints played a role in our results. Given more time, we could have devoted additional efforts to enhance data preprocessing and explore a wider range of model performances. The time limitation prevented us from delving deeper into the data and potentially uncovering more effective approaches.

1.2.2 Cost of Cancellation

Data Cleaning and Pre-Processing

- Just like in the previous question, we took care to exclude null and invalid entries.
- To enhance our analysis, we introduced various dummy variables for the explanatory factors such as hotel country, brand, city, and more.
- Recognizing their limited explanatory value, we decided to eliminate all features related to the booking date and user requests.
- Following the approach from the previous question, we conducted a thorough investigation of the correlation between the independent and dependent variables.

Decisions regarding the learning method

Initially, we considered employing various regression learners, including simple linear regression and polynomial regressor. Through experimentation, we observed that the linear regressor outperformed the others. Consequently, we decided to proceed with the ridge regressor, utilizing a lambda hyperparameter value of 0.1, which significantly improved performance. Comparing it to the lasso regressor, we determined that the ridge regressor yielded superior results.
Once we trained the initial model to predict booking cancellations, we proceeded to predict prices. Subsequently, we assigned a value of -1 to bookings that were not predicted to be canceled.

Score Prediction

Our predictions resulted in an RMSE metric of around 160. This metric provides insight into the cumulative errors of our model's predictions. Considering the relatively low value of the RMSE in relation to the sample size, we can infer that our model exhibited reasonably accurate predictions.
It's important to note that we faced time limitations during the project. Given more time or a larger sample size, we believe our model would have achieved even better prediction performance. The additional time or increased sample size would have allowed us to refine the model and potentially reduce the errors further, leading to more precise predictions.
The answers to the remaining tasks can be found in two separate files labeled "agoda_chrun_prediction_model.pdf" and "agoda_cancellation_policy.pdf".