IML: Booking

Hadar Meron
hadar.meron@mai.huji.ac.il

Eyal Weisman
eyal.weissman@mail.huji.ac.il

Liel Amar
liel.amar@mail.huji.ac.il

Gal Bahary
gal.bahari@mail.huji.ac.il

Introduction

This project is a self-supervised framework for detecting the likelihood of customers canceling their booking orders as well as the expected loss the establishment will incur. Our model, when trained with a sufficient number of training cases, can accurately predict 73% of cases.

Task 1 - Cancellation Prediction:

In the first task we tried to solve the problem of; given an input of booking information, predicting whether or not a booking order would be canceled. At first, we decided to get to know the data by playing with it and gathering statistics. This helped us get familiar with the task and the data, and prepared us to preprocess the data. We began by splitting our data into train, validation and test sets, and then decided to get rid of many features that seemed unnecessary, both logically and through different correlation calculations. For the classification features we chose to keep, we replaced them with dummy features. This allowed us to have a numeric dataset we could work with very easily. Following that, we attempted to find the model and hyper-parameters that would result in the best predictions over our validation set. During our research, we noticed that many tree-based models performed significantly better than any other models. After consolidating our different streams of thought and running the preprocess code multiple times, the model we found to be best was *XGBClassifier*. This is, once again, a tree-based classification model that is open-source.
It was chosen since it had a very good f1 score, in addition to being very simple and efficient. It seemed like a logical tradeoff for us, and chose it -over other models, that might have been more accurate but at a higher cost.

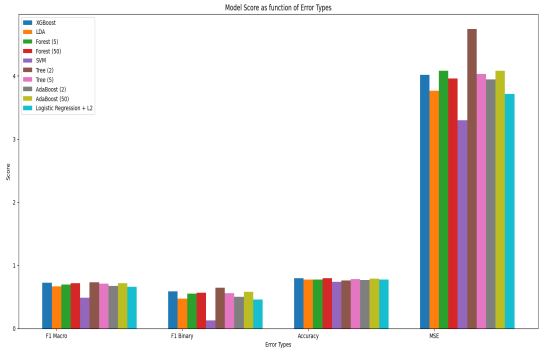

Figure 1. Model Scores for different functions

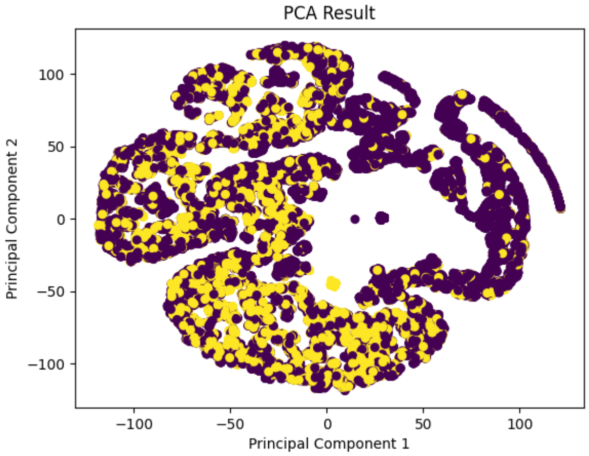

Figure 2. PCA of the train set

Task 2 - Cost of Cancellation:

In the second task, we attempted to find a model that could accurately predict the loss of money a hotel would suffer, if a customer were to cancel. We began working on this task after the first one was nearly completed, therefore, we were familiar with the data. This helped us efficiently preprocess the data and more easily realize which model would fit. The preprocess we decided to use was similar to the one used in the first task. Even so, we decided to remove now-unnecessary features. Next, we attempted to find the model and hyper-parameters that would result in the best predictions, over our validation set. It seemed logical that a regression model would perform the best, therefore we started to look for the best one. As seen in *Figure 3*, the Ridge Regression model outperformed Lasso Regression model, with the best lambda value being around 5.5. Thus, we decided to go with Ridge Regression for this task. We realized that we did not have the data of the cancellation dates in the test data set. This led us to use our model from the first task to predict whether a reservation would be canceled or not, to later be used in our new task.

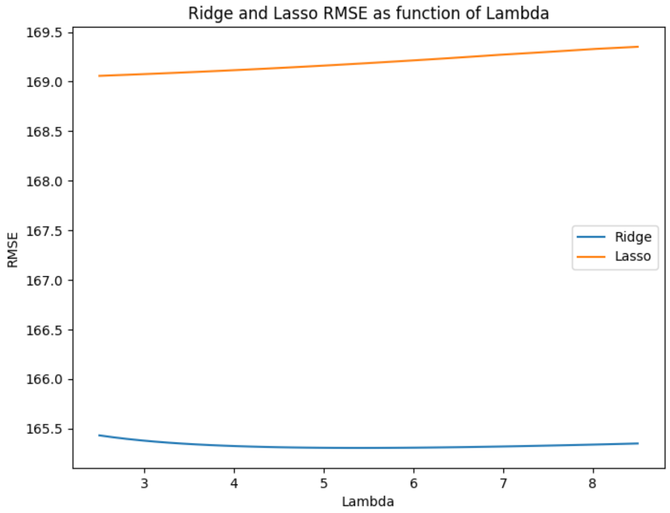

Figure 3. Ridge vs Lasso over different lambdas