Hackathon Workflow Description

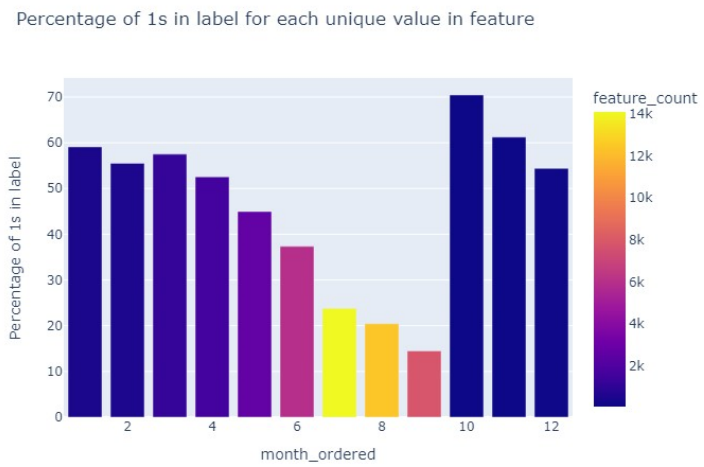
Submitters: Assaf Shiloah, Omri Roth, Elia Merran, Yoav Balkanay

We chose the first assignment- Agoda Cancellation.

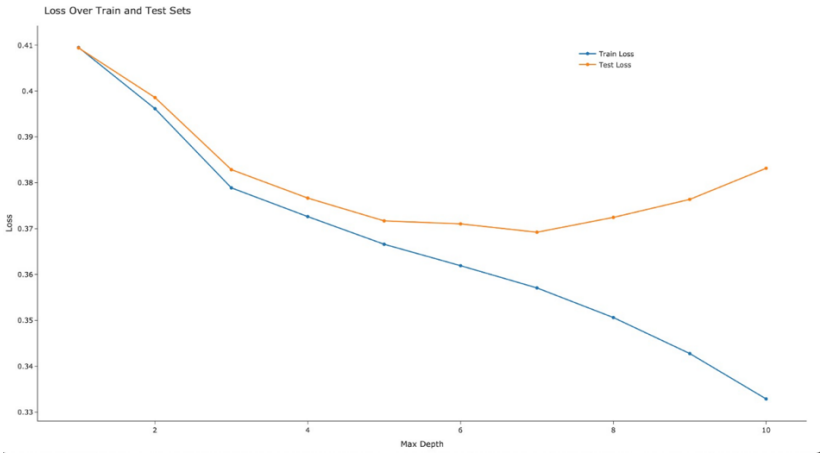
First we split our data into train and test set by a factor of 0.2 since the data seemed large enough.

For the first question we started the process by following the stages of model development suggested in class: Preprocess −→ EDA −→ Baseline −→ Model Selection.

- Preprocess – we assumed based on reality that cancellation policy has a major effect on our consideration of whether we will cancel or not our reservation. Thus we created categories of percentage of payment due to cancellation which indicates the days before check that will require payment. Also we created dummies for none numerical features, and handles time-based features. Lastly we changed the label (cancellation datetime) to include 0 or 1 – cancelled/not cancelled.
- EDA – we used feature evaluation and revisited our preprocess in order to delete and create additional features for exploring the data such as 'days before checkin'.
  We generate bar plots showing the percentage of cancellations (1's) in the canccilation for each unique value in a specified feature. These graphs provided us with visual representations of the relationship between features and cancellation percentages, aiding in understanding the impact of different factors on cancellations in the dataset. For example:



  We plotted many graphs as the one above plotting amounts of cancelation as function of normalized unique feature values, while plotting colors to feature count. We didn't get much insight but could differentiate between features that seemed flat all across, unlike month_ordered above.
- Baseline – we performed CV over various classifiers with differently hyper parameters including: Decision Tree, Random Forest, K-NN, SVM, Naïve Bayes, Gradient Boosting and Logistic Regression. For each model we examine the F1-macro score, from which we concluded that Random Forest gives us the best score considering a trial to avoid too complex model. We decided on the hyper-parameter of Random Forest using this graph (and other):

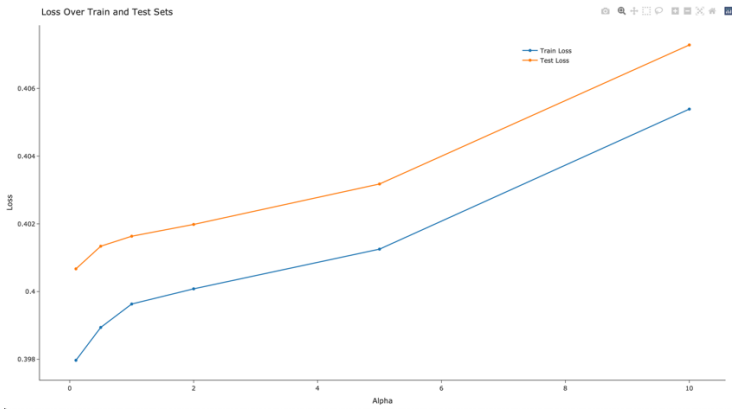


We noticed the bias-variance tradeoff – that's why we chosemax depth = 5.

- Model Selection – after we decided to use RF, we ran CV in order to select the number of estimators in the forest, and also we tested making regularization using Ridge and Lasso. We concluded that no regularization needed. Also we tried re-evaluating our preprocess and adjusting the train set.

After we done this process we got a score of 0.729 in prediction of cancellation.

In the second question, we made adjustments to our preprocessing steps to align with the required data input. Since the task in question 2 is for regression rather than classification (as in question 1), we decided to use Lasso regression and explore the hyper parameter of Lasso regression:



We noticed that increasing the regularization is increasing the loss, that's why we decided to stick to the theory and chose minimum regularization as we learned in HW4 it helps.

When addressing question 2, we recognized the need to split our training data into two halves. We trained the random forest model on one half of the data and used it to make predictions on the other half. This approach allowed the Lasso model to incorporate an additional feature that is relevant to predicting the price.

By employing this two-step process, we were able to leverage the strengths of both the random forest and Lasso regression models, enhancing our ability to capture meaningful features and predict the price accurately.

After we done this process our RMSE was 357.08 on our test set.