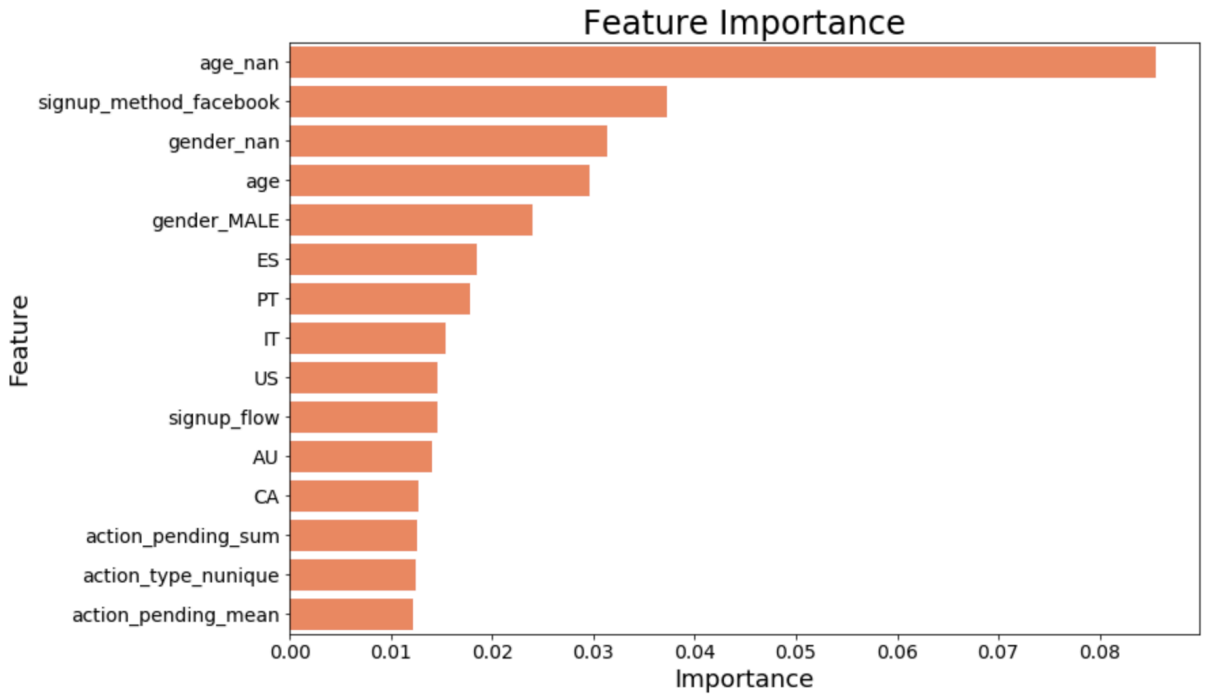
Airbnb New User Bookings Project – Machine Learning

A multi-class classification is implemented using users' destinations as the outcome variables and other variables related to users' demographic data, web session records, and some summary statistics of different countries as features. For each user, the top 5 countries with the highest predicted probability are chosen as the predicted output. Predictions based on the test data are submitted to Kaggle and are evaluated using NDCG (Normalized discounted cumulative gain).

Classification Models:

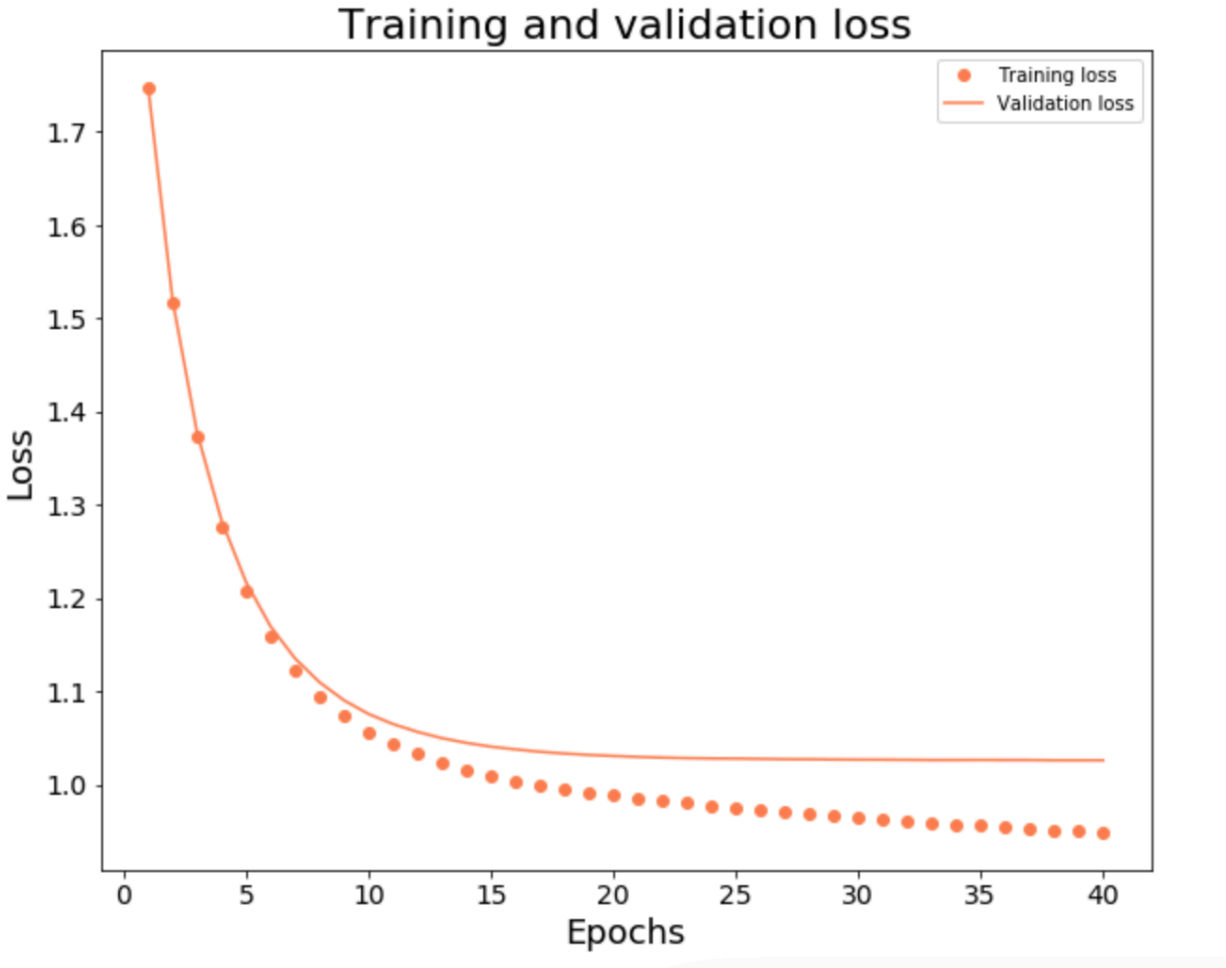
Random Forest with n_estimator=200

- Random forest is used because it can reduce overfitting by introducing randomness through bagging with different combinations of features.
- Grid search with K-Fold cross validation (k=3) is implemented to determine the best parameters for max_features (between 'auto' and 'log2') and min_samples_split (ranging from 0.0001 to 0.05).
- The best parameters are max_features= 'auto' and min_samples_split=0.0005 and the best model gives an NDCG of 0.87521.
- Looking at the feature importance of the best random forest model, the most important feature for predicting destination is whether the user leaves 'age' blank, which makes sense because those who did not intend to book a trip would probably leave 'age' blank. Other important features are whether the user signs up through Facebook, gender and age.



XGBoost

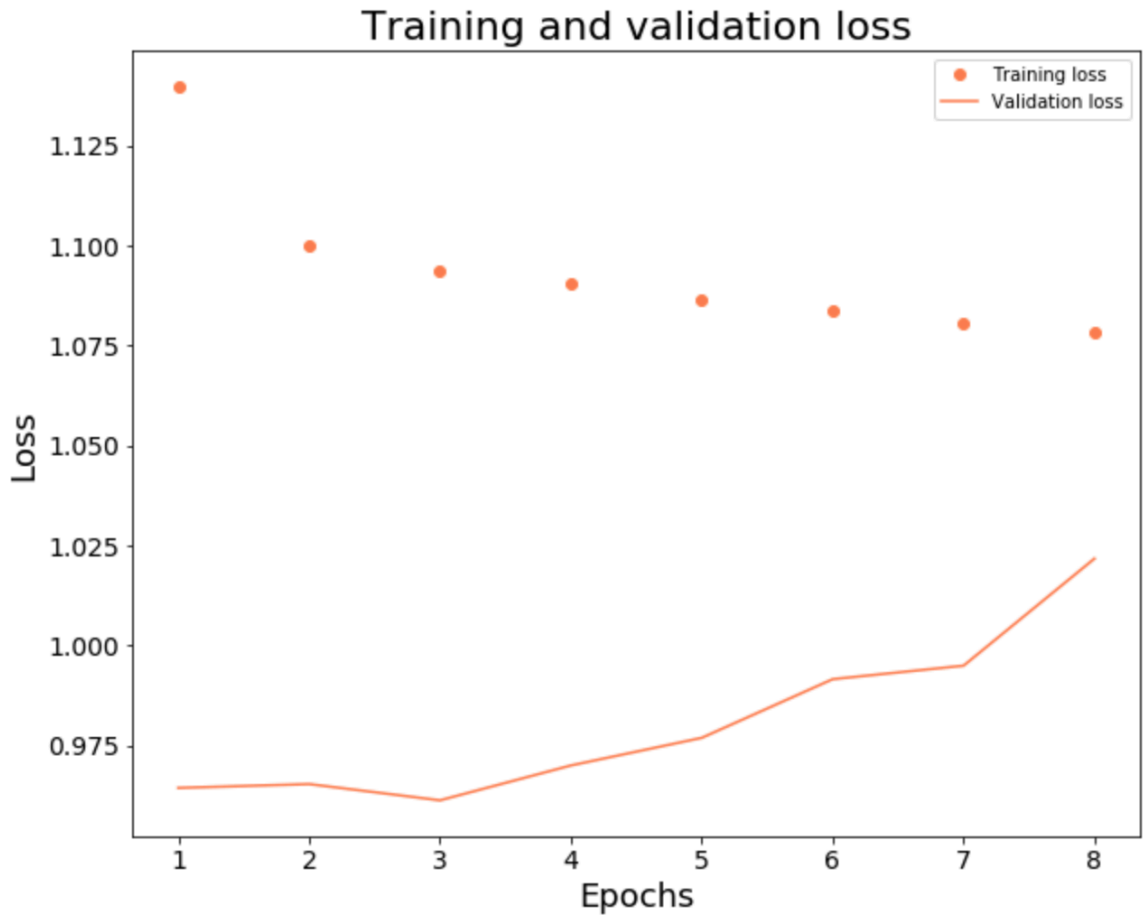
- I try XGBoost because it outperforms other algorithms and is usually one of the winning solutions in Kaggle competitions.
- XGBoost with number of rounds = [10, 20, 40] are run.
- Training and validation loss is monitored, and the training process will stop if the validation loss has not improved for 5 rounds to avoid overfitting.
- The model with number of round = 40 gives the best NDCG of 0.87837 although the validation performance doesn't seem to improve much after 15 rounds. Early stopping didn't kick in because the validation loss was still decreasing slightly (although it is hardly noticeable).



Feedforward Neural Networks

- Neural Networks are considered as they are flexible - by varying the number of layers and nodes, neural networks can fit data with different complexity.
- Feedforward Neural Networks with varying depths (from 1 to 5) and number of nodes (from 50 - 400) are tried. Features are normalized before feeding into the Neural Networks.

- Again, early stopping is implemented by monitoring the training and validation loss.
- Batch size = 50 and no. of epoch = 20 are used, and the model hits early stopping after 10 epochs.
- The best model is the one with 4 layers and number of nodes = [250, 125, 125, 12]. It gives an NDCG of 0.87414.



Ensemble Model (Soft Voting)

- Weighted mean of the predicted probabilities from the best Random Forest, XGBoost and Neural Network model are calculated. For each user, the top 5 countries with the highest weighted mean are chosen as the predicted output.
- Different combinations of weights are tried and the combination that gives the best model is [Random Forest, XGBoost, Neural Networks] = [1, 2, 1]. It gives an NDCG of 0.8780.

My final model is XGBoost Model with 40 rounds of training with a NDCG score of 0.87837. It ranks 325 out of 1462 on Kaggle Leaderboard.