Airbnb New User Booking Prediction

Helnaz Soltani

May 15, 2020

Galvanize DSI

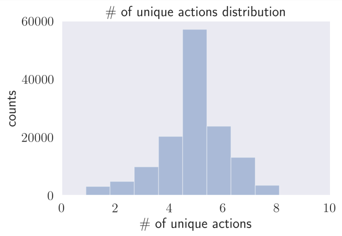
Project Overview

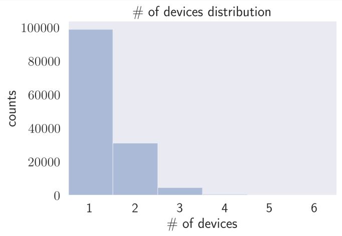
- Objectives:

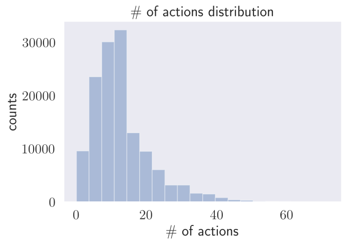
  Predicting new Airbnb users' first reservation location
- Main dataset information:
  - Number of instances: 200k
  - Number of features: 15
  - Data analysis tasks: Multivariate classification with 12 classes
  - The data is collected from US users.
- Data cleaning issues:
  - Duplicated entries
  - Null values
  - Incorrect datatype
  - Inconsistent values for some of the columns
- One-hot encoding for categorical features (total of 68 features)
- Feature engineering (total of 78 features)

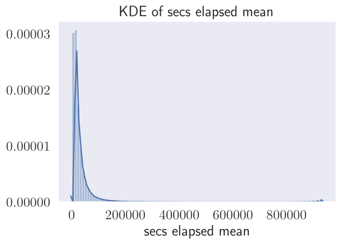
Feature Engineering

- 'date' columns were converted to day, month, and year features.
- The below features were calculated from users' web session records:



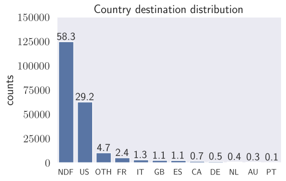


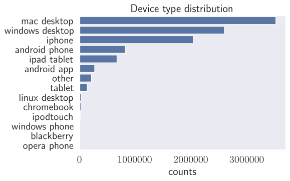


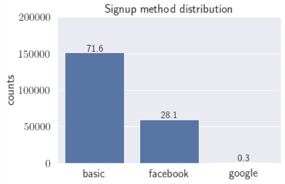


Explanatory Data Analysis

- 58% of users did not book ('NDF': no destination found).
- Airbnb users tend to use Apple products more.
- Signup method is dominated by 'basic' and 'facebook'.



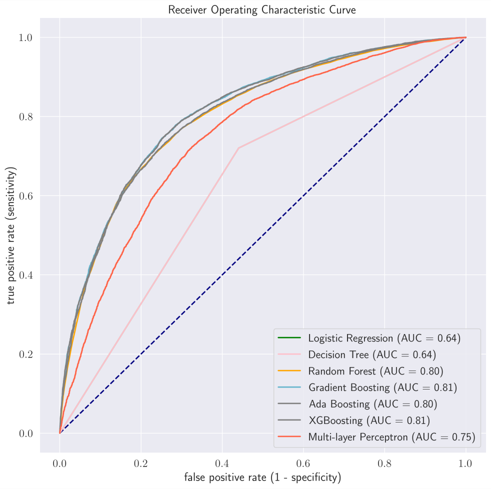


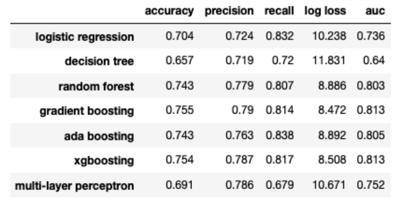


- Problem statement:
  - Question 1: Does a new Airbnb user book?
  - Question 2: If yes, where is the destination?

Machine Learning Modeling - Question 1

- Considering 'NDF' as one class and the rest as another class
- Binary classification



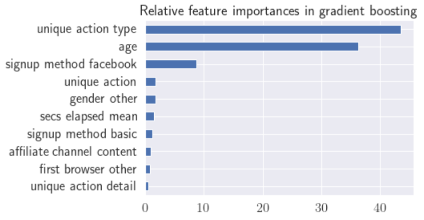


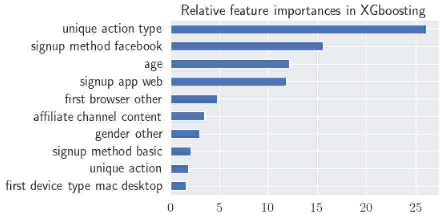
	accuracy	precision	recall	log loss	auc
logistic regression	0.704	0.724	0.832	10.238	0.736
decision tree	0.657	0.719	0.72	11.831	0.64
random forest	0.743	0.779	0.807	8.886	0.803
gradient boosting	0.755	0.79	0.814	8.472	0.813
ada boosting	0.743	0.763	0.838	8.892	0.805
xgboosting	0.754	0.787	0.817	8.508	0.813
multi-layer perceptron	0.691	0.786	0.679	10.671	0.752

- Best performance:
  - Gradient boosting
  - XGboosting

Feature Importance (10 tops) - Question 1

- Gradient boosting
- XGboosting

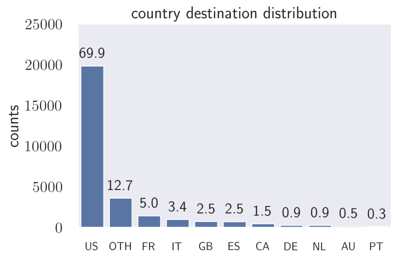




- Top 3 important features are shared in both models.
- The most important feature in both models was constructed in feature engineering step.

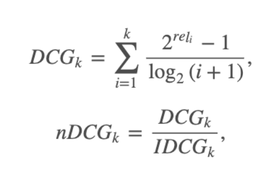
Machine Learning Modeling - Question 2

- Excluding 'NDF' and considering the rest of classes as individual classes.
- Multiclass classification



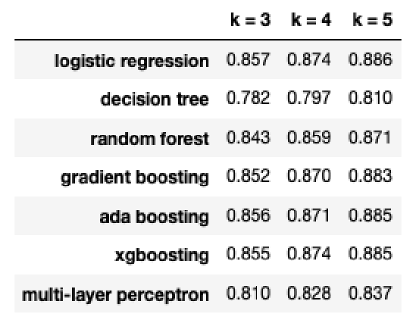
- Handling imbalanced dataset:
  - undersampling the observations with target variable = 'US'.
  - then, oversampling for the rest of the target variables.

Machine Learning Modeling - Question 2



$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2 (i+1)},$$

$$nDCG_k = \frac{DCG_k}{IDCG_k},$$



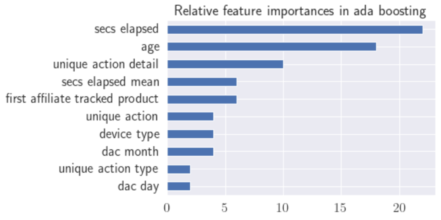
	k = 3	k = 4	k = 5
logistic regression	0.857	0.874	0.886
decision tree	0.782	0.797	0.810
random forest	0.843	0.859	0.871
gradient boosting	0.852	0.870	0.883
ada boosting	0.856	0.871	0.885
xgboosting	0.855	0.874	0.885
multi-layer perceptron	0.810	0.828	0.837

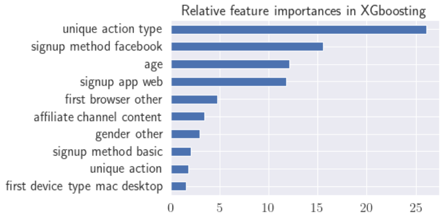
- Evaluation metric: NDCG (normalized discounted cumulative gain)
- k is the # of predictions for each new user

- k $\uparrow \;\Rightarrow\;$ NDCG $\uparrow$
- Best performance:
  - Logistic regression
  - XGboosting
  - Ada boosting

Feature Importance (10 tops) - Question 2

- Ada boosting
- XGboosting





- For Ada boosting, 8 important features out of 10 tops were constructed in feature engineering step.

Conclusions

- Predicting whether or not a user will book a reservation combined with the marketing strategy is beneficial for business purposes.
- Predicting which countries are more popular for accommodation booking is beneficial for demand forecasting.
- Also, Airbnb can build a recommendation system based on these predictions and suggest the destinations which are similar to the user's choice (in terms of climate, nature, and things-to-do)

Thank you!