CSE258 Assignment 2
brb Predicting on Airbnb

Arvind Rao
A10735113
a3rao@ucsd.edu

Behnam Hedayatnia
A09920117
bhedayat@ucsd.edu

Daniel Riley
A10730856
dgriley@ucsd.edu

Ninad Kulkarni
A09807450
nkulkarn@ucsd.edu

Abstract

This paper details the exploration of a dataset released by Airbnb in the form of a Kaggle competition the purpose of which was to predict the first country to which an Airbnb user books a trip. The features of this dataset were studied and multiple custom classifiers were created to exploit the structure of the dataset. After submitting to Kaggle, it was found that the best classifier attempted (according to their NDCG metric) was a 3-layer neural network. This classifier received an NDCG score of 0.875 which placed it in the top 25% of user submissions.

1 Introduction

With the advent of Big Data and powerful machine learning techniques we are able to design systems that can be tailored to specific users allowing for a more personalized product.

By accurately predicting where a new user will book their first travel experience, Airbnb can share more personalized content with their community, decrease the average time to first booking, and better forecast demand.

This leads us to predict in which country a new user will make his or her first booking based on session activity and user demographic information.

2 The Dataset

Our first csv file called train_users contains user information such as user ID, the date of account creation, timestamp of first activity, date first booking, gender, age, sign-up method, and language. We have a total of 213,451 users in our training set.

We also have a csv file containing session information for each user which states the actions that were taken and how long each action was taken on the site. We have 10,567,737 recorded sessions of which there are 135,484 unique users. There are 73,815 users from the training set that have session information.

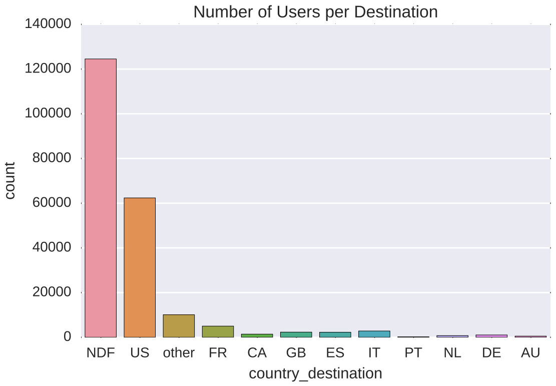

Figure 1: Histogram of Destinations Users Book. The unbalanced dataset is evident in this histogram

2.1 Dataset Exploration

2.1.1 Bookings over Time per Destination

The first exploration that was done was to view how the amount of bookings for each country changed over time. This is shown in Figure 2. The dataset contains samples from October 2010 through July 2015. As we can see, at around the summer time every year, there is a drop in the number of bookings for all countries. This makes sense since most of the trips would be during the summer so the amount of bookings would go down. The overall increase in bookings over time is due to the increasing number of users on Airbnb as the service grew over the years.

We also looked at the amount of bookings for

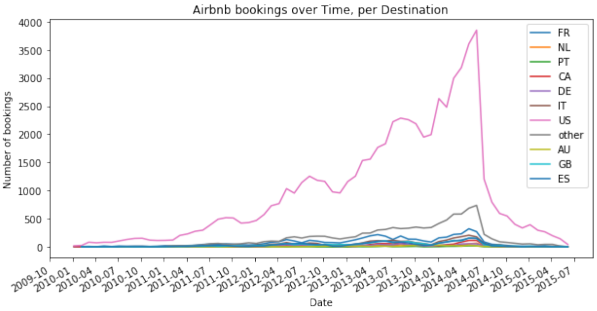

Figure 2: The amount of Bookings for each Country over time

each country with the United States removed. We removed the U.S., because it had significantly larger number of bookings (by about a factor of 10). This gives us a slightly better picture for other countries as shown in Figure 3. From the figure we can see there are more pronounced dips of the number of bookings during certain seasons.

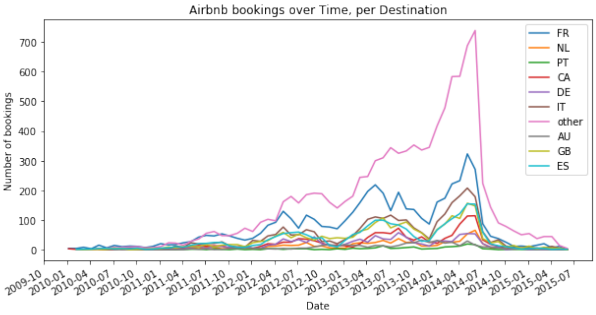

Figure 3: The amount of Bookings for each Country over time not including U.S.

2.1.2 Bookings based on Signup Method

We examined the number of bookings based on signup method: basic, facebook, or google. This is shown in Figure 4. The unbalanced nature of the dataset is again evident, as U.S. out numbers all the other countries. To get a better view we took out the U.S. and again viewed the number of bookings based on signup method which is shown in Figure 5.

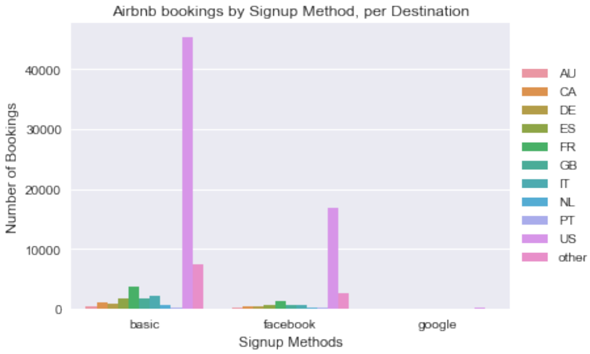

Figure 4: The number of bookings for each signup method for each country

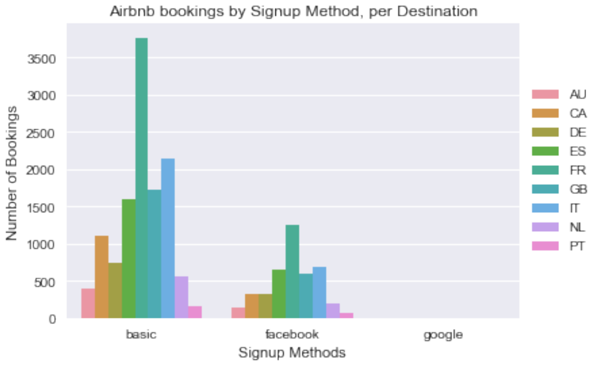

Figure 5: The number of bookings for each signup method for each country without the U.S.

2.1.3 Age vs Bookings

Age and how it affects the number of bookings for each country was also explored. The data was cleaned such that only users younger than 80 and older than 10 were kept, as it's possible that users outside of this range are not proper ages of Airbnb users. These are shown for destinations U.S.,NDF and other in figure 6. The rest is in figure 7. In all of the histograms, the distributions are skewed to the left, with long right tails. The ”center of mass” of the distributions is different for each destination, so it is possible that age is a good discriminator between countries.

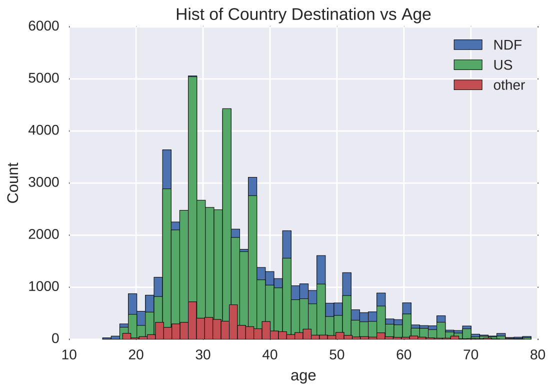

Figure 6: Histogram of Bookings Based on Age

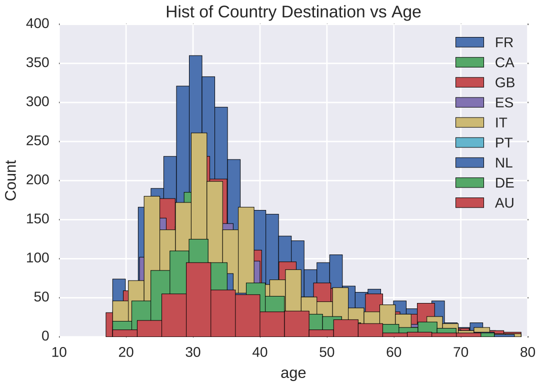

Figure 7: Histogram of Bookings Based on Age

2.1.4 User Language Preference vs Destination

We identified that user language may be helpful for categorizing which Non-US country a user has booked for. There tends to be a trend in which native speakers book a trip to the native countries. For example, ’fr’ language users like to go to France the most, as seen in figure 8.

In addition, users of languages that are in ’other’ countries tend to go to ’other’ countries. As an example, the majority of ’ko’ language users go to ’other’. This may likely be Korean language users going to Korea, which explains the higher count in ’other’ for ’ko’ users. Similar observations hold for other languages like ’ja’, ’zh’, ’sv’, ’ru’, etc.

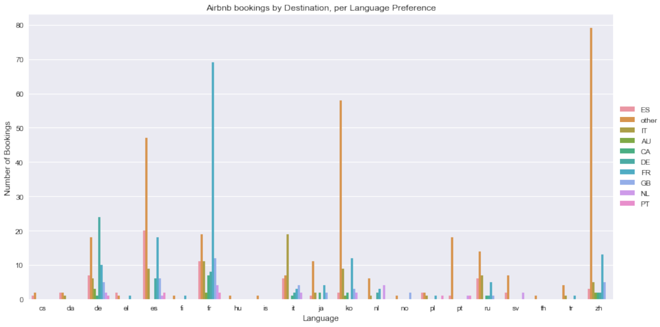

Figure 8: Number of bookings to different destinations, across users of different language preferences.

2.1.5 User Session Activities

For our data exploration on sessions we took a look at each action a user took, for the users that took said action we looked at the difference between people who booked and who did not book and took features with a percentage difference of over 10% for booking versus non booking. A sub selection of actions that were analyzed are listed below. The features that were gleaned from these sessions were the most discriminatory features given in the dataset.

Action	NDF	NonNDF
about us	72	49
active	14444	6716
ajax payout edit	238	89
authenticate	10470	7787
contact new	324	224
create	28638	20703
click	1969	1072

Table 1: A sample of actions analyzed from sessions.csv dataset containing user activity.

3 Predictive Task and Assessment Criteria

Given historical user data, the predictive task is to determine which destination a new user will book. Since the destinations in the training data are unbalanced, one strategy that was attempted included splitting the classification into multiple levels. The validity of the predictions for each stage of this strategy were assessed using area under the curve of the Receiver Operating Characteristic(ROC) curve and the Normalized Discounted Cumulative Gain(NDCG). In ROC curves, the True Positive Rate and False Positive is plotted at different thresholds for the decision function. $TPR = \frac{TP}{TP+FN}$ and $FPR = \frac{FP}{FP+TN}$. Both of these equations can be considered a probability conditioning on the true label, which is independent of the class balance. Therefore, the reason for using this metric is that at each stage, the distribution of classes is unbalanced, so using accuracy is a misleading measure, while the ROC curve is not affected by class balance (stated in Fawcett 2004). k-fold cross validation will be performed, and each fold will return the AUC under its ROC.

The metric used by Kaggle for ranking is NDCG. This is calculated as $DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i}-1}{log_2(i+1)}$, and $NDCG_k = \frac{DCG_k}{IDCG_k}$, in which k represents the number of guesses made. IDCG is the ideal DCG for a possible set of queries, which is 1.0 for the ground truth in the $k = 1$ positive. $k = 5$ predictions are made for each test user, and the NDCG penalizes predictions that have the true destination not in the first position. This metric will be used for evaluation of the model's performance on the test set, which is documented in Section 6.

Most of the features of the dataset are categorical, as seen in Section 2. Thus, each category was converted into a one-hot representation for each value that it could hold.
To deal with temporal information we binned when a user created an account into seasons, months and years.

4 Model Selection

From the Dataset exploration and the large amount of unbalanced datasets, one approach that was attempted was to predict which country a user is going to is using a multi-level classification approach:

1) Classify between No Destination Found (NDF) and Non-NDF to figure out if a user will book on Airbnb.
2) Classify between US and Non-US to predict if a user will book outside of the US, given that the user has booked on Airbnb .
3) Classify between the remaining countries to predict which country a user will book, given that the user has booked on the Airbnb site, and that the user has booked outside of the US.

For this strategy each of our classifiers uses either Random Forests or Neural Networks. Random Forests are a bagging algorithm, as it combines complex decision trees with low bias and high variance to form a predictor with low bias and low variance. As stated in (Galar 2012), ensemble methods are good solutions to imbalanced data sets, which is why Random Forests are used in this predictive task. Neural Networks are sensitive to class imbalance, because during the learning process, if the class is seen less often, the weights of the neural network will not update to account for that class. To account for this, in each stage the minority class is oversampled such that the classes appear equally. This technique is drawn from (He 2009).

The second approach attempted was to use a fully connected Neural Network to directly predict all 12 destinations. This classifier would likely perform less well for the ROC AUC but it better models the distribution of the data according to the metric given by Kaggle (since their metric has equal weight for each class).

4.1 Features for NDF and Non-NDF Prediction

The features we looked at for classifying between NDF and Non-NDF were mainly actions that were taken by a user. As not every user in the training set has session information, the training set is pruned such that all users will have session information. This reduces the number of samples from $N = 213451$ to $N = 73815$.

The features were user actions that had a large discriminatory difference between NDF and Non-NDF. Later it was determined that using all the actions had a better performance (instead of a subset of user actions). For each training sample, an action feature was given binary value True if the user had performed that action, and False otherwise. Using all user actions possible, there were a total of 359 features. Taking all 359 actions is explained by the data exploration done in section ”User Session Activities”. In table 1, showing the most performed actions with widest NDF/Not NDF spread, the number of users performing that action is much less than the total number of users. Therefore, only using a subset of actions would result in a very sparse feature set, as it would be possible for some users to have a feature set that is all False, as they might not have performed any of the subset actions. From this analysis, all actions were used.

After shuffling the data, two classification methods were attempted: 5-fold cross validation with a Random Forest Classifier with 100 trees and a 3-layer neural network with 400 hidden units in each layer. These classifiers were optimized for validation performance using area under the curve from our ROC curves.

The ROC curves for this first stage (NDF, Not-NDF) classifier are shown in figures 9 and 10. As is shown our first level RF classifier gets an AUC at around 0.7 and our NN gets 0.8. A random predictor will have an $AUC = 0.5$, so both the Random Forest and Neural Network classifiers achieve significantly above random chance.

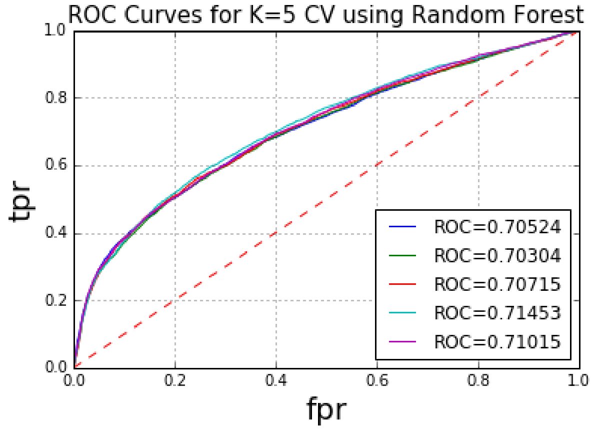

Figure 9: ROC curves for our level 1 RF classifier distinguishing between NDF and non-NDF

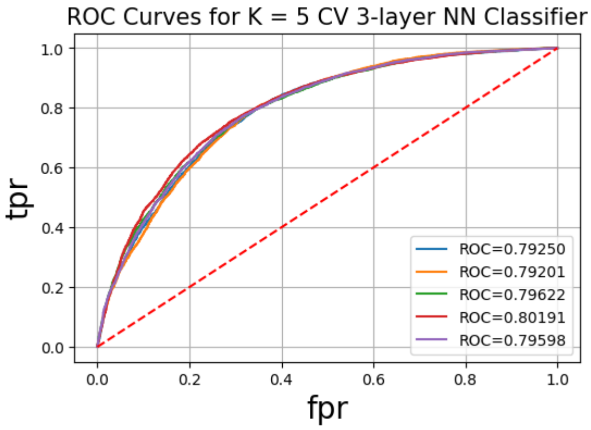

Figure 10: ROC curves for our level 1 NN classifier distinguishing between NDF and non-NDF

4.2 Features for US and Non-US Prediction

To classify between US and non-US we took users whose sessions activity was recorded. We then took into account all actions that users took along with the following features: signup method, signup flow, affiliate channel, affiliate provider, first affiliate tracked, signup app, first device type, first browser.

After shuffling the data, 5-fold cross validation with a Random Forest Classifier and optimized performance using Area under the curve from our ROC curves was performed.

The ROC curves for this second stage (US, Not-US) classifier is shown in figures 11 and 12 for the Random Forest and Neural Network models respectively. Both models achieve AUC around 0.59. As before, a random predictor will have $AUC = 0.5$, so both models perform better than random chance.

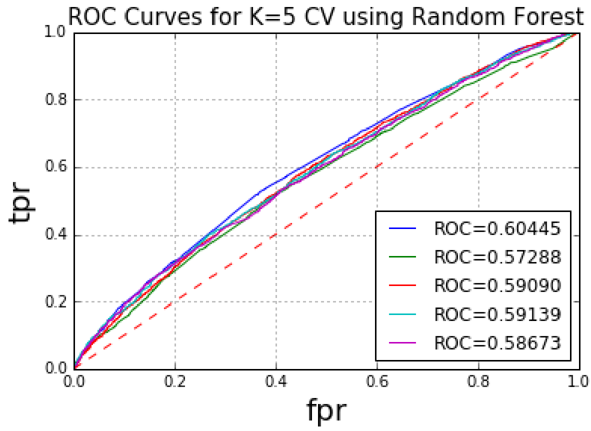

Figure 11: ROC curves for our level 2 RF classifier distinguishing between US and non-US

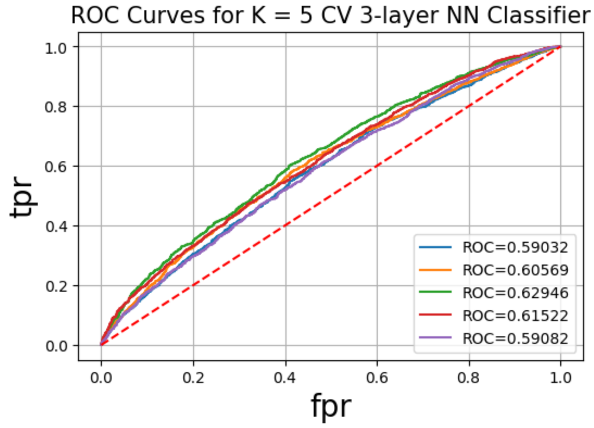

Figure 12: ROC curves for our level 2 NN classifier distinguishing between US and non-US

4.3 Features for Non-US Countries Prediction

The first two levels of classification were binary classification, in which a prediction is made between NDF or Non-NDF, and US or Non-US. Our third level model requires multi-class classification, to predict among the rest of the countries. There are a total of 10 categories for each of the countries, including 'other'. Due to this requirement, the natural approach of choice was to use the One-VS-Rest classifier. This method trains a classifier for each class to discriminate the membership of a given sample in the class that it is trained for.

Note that at this level of the model, the assumption that the model has already excluded NDF and US samples as best as it could. Thus, training was done using only the samples known to have Non-NDF and Non-US destinations. Our implementation uses a One-VS-Rest approach that trains a Gradient-Boosting Classifier for each of the 10 countries that belong in the Non-US class: AU', 'CA', 'DE', 'ES', 'FR', 'GB', 'IT', 'NL', 'PT', 'other'.

The features used at this level include the language, signup_flow, first_browser, as well as the session activity. The hyperparameters used for the Gradient Boosting Trees are as follows: 100 estimators, tree depth of 10, learning rate of 0.1, and the deviance loss function.

The ROC curves for this third stage (remaining countries) classifier using Gradient Boosting Trees is shown in figure11. AUC around 0.59. As before, a random predictor will have $AUC = 0.5$. This model performs better than random choice.

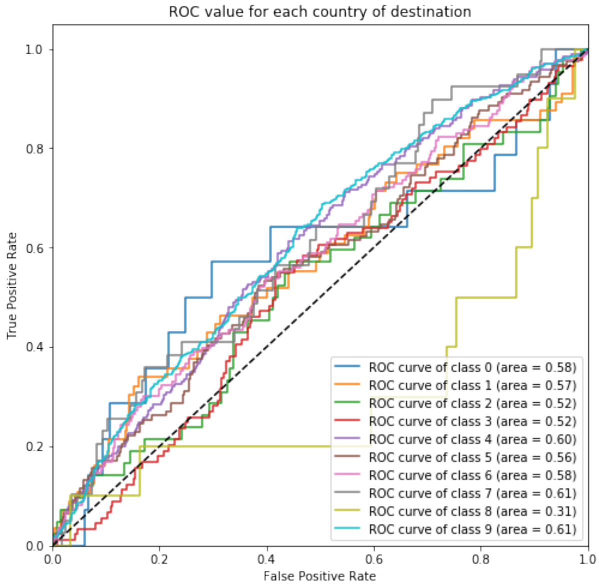

Figure 13: ROC curves for the level 3 model, with the sessions features included while training.

Figure 13 is the ROC curves obtained by the third level model, for each of the 10 classes. The ROC is calculated for in a One-Vs-Rest fashion for each class. This model achieves varying AUC for each of the classes. The rest of the classes have an average of 0.57 AUC, which is better than random choice. Only one class ('PT' = Portugal) performs worse than the baseline of random choice, with an AUC of 0.39. The ROC curve for this class contains a several of concavities, indicating some locally worse than random behavior. A better pruning of input features sometimes reduced the concavity for the PT class, but it came at a cost of lower AUC for the other classes.

Unfortunately one-vs.-all cross validation of the lowest level NN classifier proved too time intensive to produce.

5 Fully Connected NN Direct Classifier

A single level neural network classifier was trained on the same features as the multi-level classifiers. This network had 3 hidden layers with 400 hidden units each ending with a 12 way softmax classification layer. Each class was represented with a one-hot encoding vector of length twelve. L1 regularization, adam optimization, and cross-entropy loss were were utilized to train the neural network. Unfortunately generating cross-validation ROC AUC values requires training the neural network for a prohibitive amount of time so hyperparameters for this NN were taken from the first level hyperparameters of the 3-tier NN classifier. These hyperparameters included epochs trained, weight regularization parameter, layers, units per layer, and optimizer.

6 Result/Test Set Performance

It was found that the ROC values for each classifier were not indicative of performance on the Kaggle test set. The test set is composed of the same information as the training set, except the Time First Booking and Country Destination columns are removed. The metric used for ranking is the Normalized Discounted Cumulative Gain, which is described in Section 4. $k = 5$ predictions are made for each test user, and the NDCG penalizes predictions that have the true destination not in the first position. The trivial predictor for the test set is to predict NDF for all 5 predictions for all test users.

Trivial	Tiered N.N.	Single N.N.
0.6811	0.8391	0.875

Table 2: Test Set NDCG Performance. N.N. stands for Neural Network

The top five countries output from each NN classifier was based on the output probability by class of each test user.

We also constructed a pipeline for our multi-level Random Forest classifier. After evaluating on the test set the result was far worse than the Neural Net classifier. This might be due to us guessing only a top prediction and not a top 5 prediction. As a result our final classifier was to use a Neural Network.

7 Related Literature

As stated in (Fawcett 2004) ROC curves are not sensitive to unbalanced data sets, so using ROC will give a honest metric to compare models. In addition, (Galar, et al. 2012) show that ensemble methods are valid approaches to building models on unbalanced data sets. For the Neural Network approach (He, et al. 2009) states that oversampling the minority class is also a valid approach to building models on unbalanced data sets.

Specific to this predictive task, the second place winner (Kuroyanagi) used out-of-fold cross validation on 18 XGBoost models. The main feature that allowed the winner to achieve a high result for both the private and public sets was the introduction of time deltas between (time booked, time account created) and (time first active, time account created). The main difference in the winner's implementation and ours is that the winner binned the time deltas into (positive, negative, N/A), while our models used the raw values.

8 Summary

This paper detailed our exploration of the Airbnb dataset and the models that was made to predict the booking destination of first-time users. Key features were identified from the dataset, and due to the unbalanced nature of the dataset, we experimented with the single-level classification and a hierarchical approach, in which each layer was responsible for predicting a subset of the destination countries. The final classifier received an NDCG score of 0.875 which placed the submission in the top 25% of user submissions was the single-level classification using a Neural Network.

Acknowledgments

We would like to acknowledge Airbnb and Kaggle for providing the dataset. We would also like to thank Professor McAuley for provided knowledge.

References

[Galar, Mikel. et al. *A Review on Ensembles for the :*] Class Imbalance Problem Bagging-, Boosting-, and Hybrid-Based Approaches.
2012 IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS PART C: APPLICATIONS AND REVIEWS

[Fawcett, Tom. *ROC Graphs: Notes and Practical* .] Considerations for Researchers
2004 Kluwer Academic Publishers

[He, Haibo. Garcia, Edwardo *Learning from Imbalanced Data.*] 2009 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING

[Airbnb Recruiting: New User Bookings Evaluation.] https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data.

[Airbnb Recruiting: New User] Bookings Second Place Winner Interview. http://blog.kaggle.com/2016/03/17/airbnb-new-user-bookings-winners-interview-2nd-place-keiichi-kuroyanagi-keiku/

[Random Forest Classifier.] http://scikit-learn.org/stable /modules/generated/sklearn.ensemble.RandomForestClassifier.html

[Learning from Imbalanced Classes.] https://svds.com/learning-imbalanced-classes/

[Keras Neural Network Modeling] https://keras.io/