Michael Dimitras
A12465780
CSE 190
Assignment 2

Predicting Airbnb Bookings by Country

1: Dataset Description

For this assignment, I selected the Airbnb "New User Bookings" set from Kaggle. The dataset is from a competition which seeks to predict which country a user will book their next trip in based on information Airbnb has gathered about the user. I selected this dataset because Airbnb is a company I've used several times in the past, and felt I had some insights that would help me with this predictive task. I was also interested in learning more about Kaggle competitions in general.

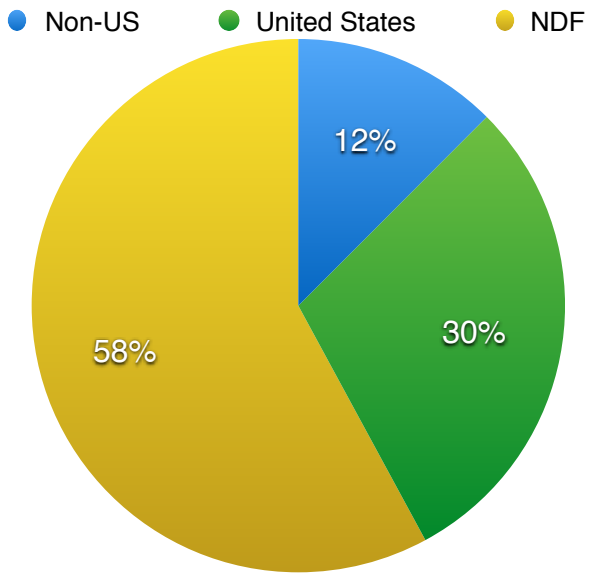
The training set provided on the website included 171,240 user profiles and asks to predict up to 5 possible travel destinations. The predictions are scored by order, such that the first position is weighted more heavily than the second, second more highly than the third, etc. The scores were evaluated using Normalized Discounted Curve Gain (Airbnb recruiting evaluation)

There are 16 features used to describe each user in the dataset:

user id
the date of account creation
timestamp of the first activity, note that it can be earlier than
date of first booking
gender
age
signup_method
the page a user came to signup up from
international language preference
what kind of paid marketing
where the marketing is e.g. google, craigslist, other
whats the first marketing the user interacted with before the signing up
signup_app
first_device_type
first_browser

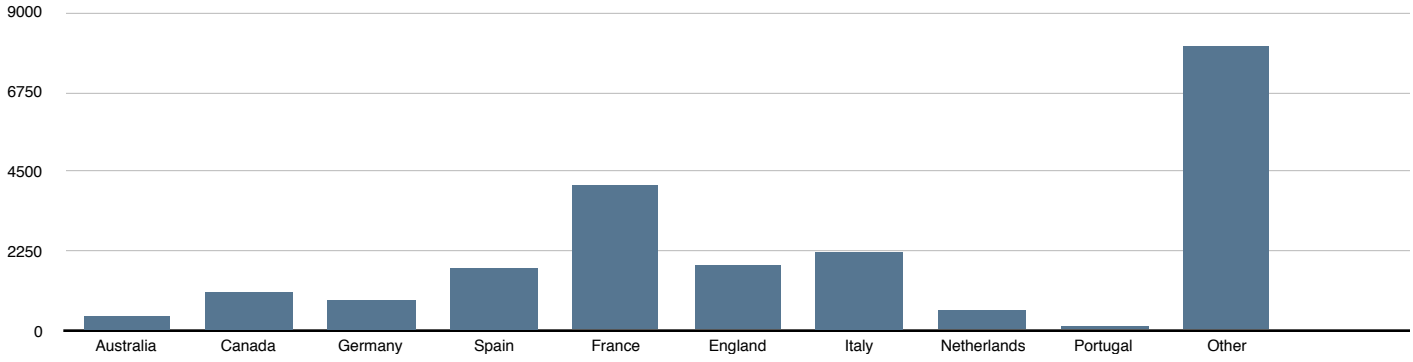
There were 11 possible destinations in the dataset. 99,151 of the instances in the training set were labeled as "No Destination Found". Of the remaining 72,809 training examples. The vast majority of the bookings (50,698) were in the US:

NO DESTINATION vs. UNITED STATES vs. INTERNATIONAL BOOKINGS IN TRAINING SET



The Non-US destinations in the dataset were somewhat more evenly distributed, but troublingly contained approximately 40% trips to countries labels as ‘other’

DISTRIBUTION OF TRIPS TO NON-US DESTINATIONS



Because of the disproportionate distribution of target values into two categories, I decided to train naive classifiers to simply predict ‘US’, ‘NDF’, and ‘Other’ to serve as a baseline to gauge other predictions on. I also attempted to train models to first predict one of the three main categories and then predict among the Non-US categories if it was determined to be not ‘US’ or ‘NDF’

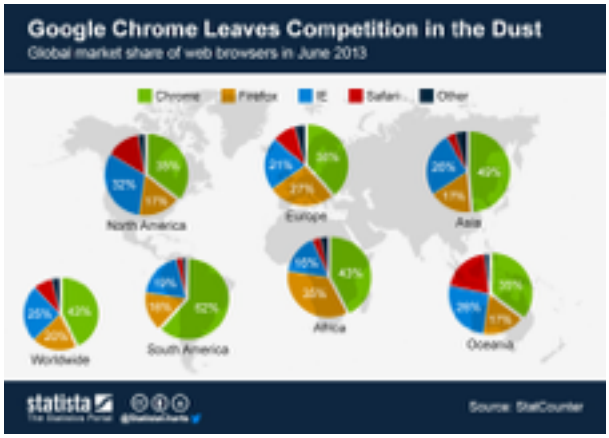
The feature in the dataset that initially seemed most useful was ‘international language setting’ . It seems natural that people will more than likely book trips within their home country rather than internationally, and language is obviously strongly correlated with country of residence. English was by far the most prevalent language preference in the training set, with 70,289 out of the 99,151 users speaking english. The number of english speakers was roughly the same as the number of people booking a trip to the US, and i was curious to see how much of a direct correspondence there was. I looked to see how many instances in the training set matched both language preference ‘english’ and

destination ‘US’. This test found that there were 49,444 such pairs, or approximately half of all training instances.

2: Feature Selection

Language was the only feature in the dataset that I felt was obviously helpful, and analyzed each of the other features more thoroughly to determine whether or not to use them.

The first feature I examined was ‘first browser used’, thinking that would be another strong way to get at the user’s country of origin. Examination of the dataset predictably found that the vast majority of users accessed Airbnb though the major browsers (Chrome, safari, firefox, safari mobile, etc). However there were some infrequently used browsers, such as the Russian “Yandex browser” and Chinese ‘Sogu Explorer”, that are very likely extremely predictive of their user’s origin. Even the major internet browsers are helpful predictors, as different global regions tend to have different preferences :



Source: (Infographic: Google Chrome…)

‘Affiliate provider’ was another feature that I felt would be strongly correlated with the user’s geolocation. Among the different providers in the dataset were predictable large, international services like Google and Facebook, but there were once again certain uncommon, but very descriptive member as well. Sites like China’s Baidu, South Korea’s Naver, and Russia’s Yandex were found in the dataset. As with the browsers, I suspected that the combination of slightly varying international popularity of big services, and highly informative small features would be very valuable in prediction.

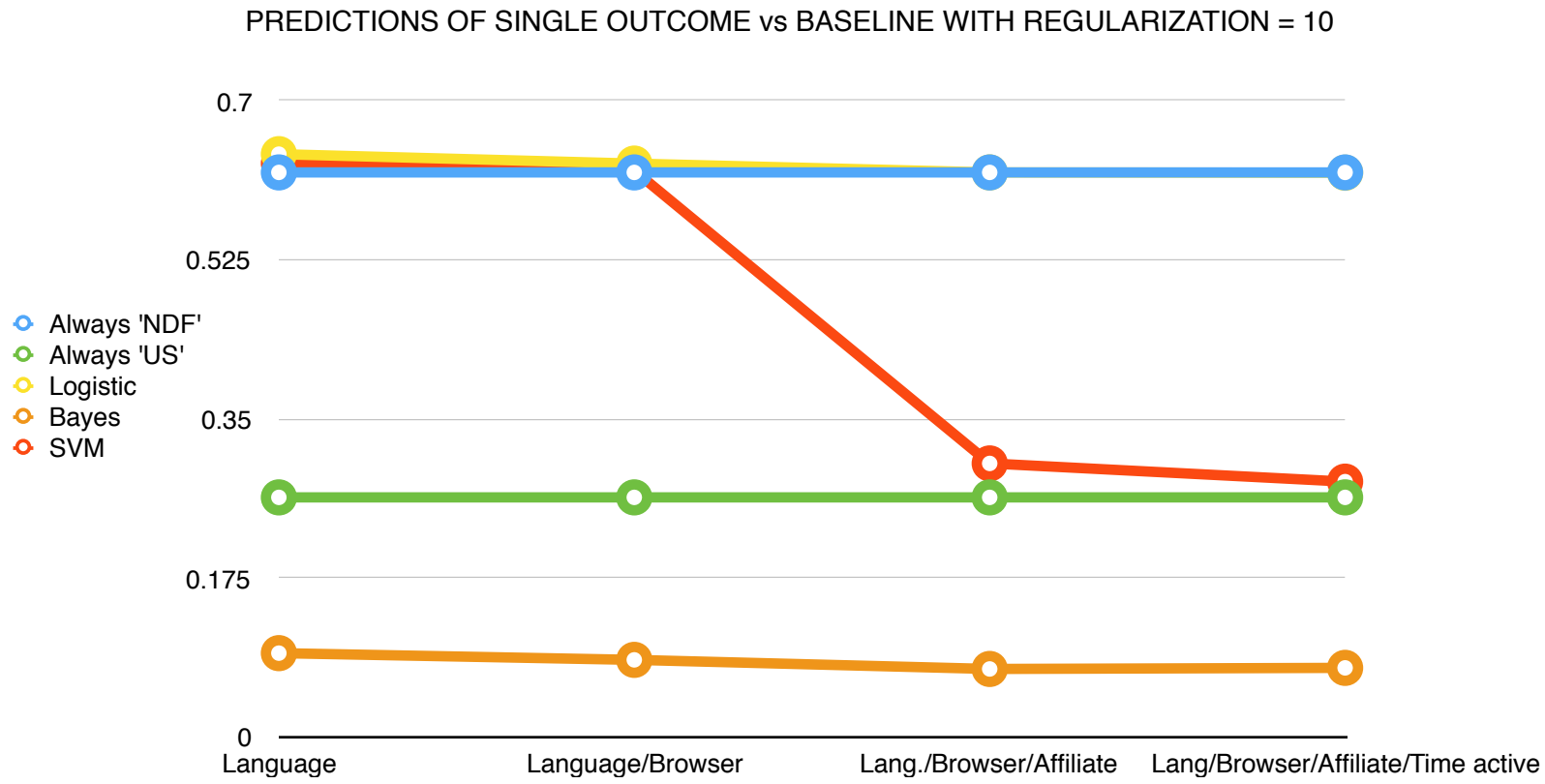
I felt the time related features, ‘date account created’ and ‘timestamp of first activity’. The dataset contains instances that date between 2010 and 2014. From reading histories of Airbnb online and particularly, the timeline at: https://en.wikipedia.org/wiki/Timeline_of_Airbnb - I found that Airbnb’s has become more popular in different regions in that time. It seems likely that there will be some change in the pattern of bookings over time. AdditionallyFrom these features, I chose to use ‘timestamp of first activity’ so not to double count what will likely be similar features.

I incorporated the features mentioned above into my initial models, and experimented with additional features in subsequent ones. I encoded all off the categorical features into binary matrices using sklearn’s DictVectorizer and converted and date features into seconds since January, 1 1970.

3: Model Selection

In this assignment, I employed 3 different machine learning algorithms: Support Vector Machines for Classification, Logistic Regression, and Naive Bayes. Each had various strengths and weakness that I believed would help provide insight into the problem. In all experiments I used the sklean SVM for Classification, and Gaussian Naive Bayes algorithims.

I began by training models to predict only a single country destination rather than all five possibilities, to get a handle on which features/model types would be most useful. Unfortunately, using this approach with several different feature combinations led both the SVM and Logistic Regression models to heavily biased predictions. Different combinations of features caused these models to predict 'US' or 'NDF' in over 95% of cases.



Unsurprisingly, these tests deviated only slightly from simple baseline predictors that always classified instances as always 'US' or always 'NDF'. This trend persisted after splitting the training instances 80%/20% and using the smaller portion for validation. Setting the regularization promoter to 10 and then to 50 had only a minor effect on the results.

The Naive bases model performed significantly worse in all instances. I expected it to perform somewhat worse on the tests compared to the other algorithms, but was surprised how poorly it performed and disregarded it in future experiments.

I was disappointed by how poorly my initial models performed compared to relatively simple classifiers, and decide to reanalyze the data to see how I could improve them. I found that all

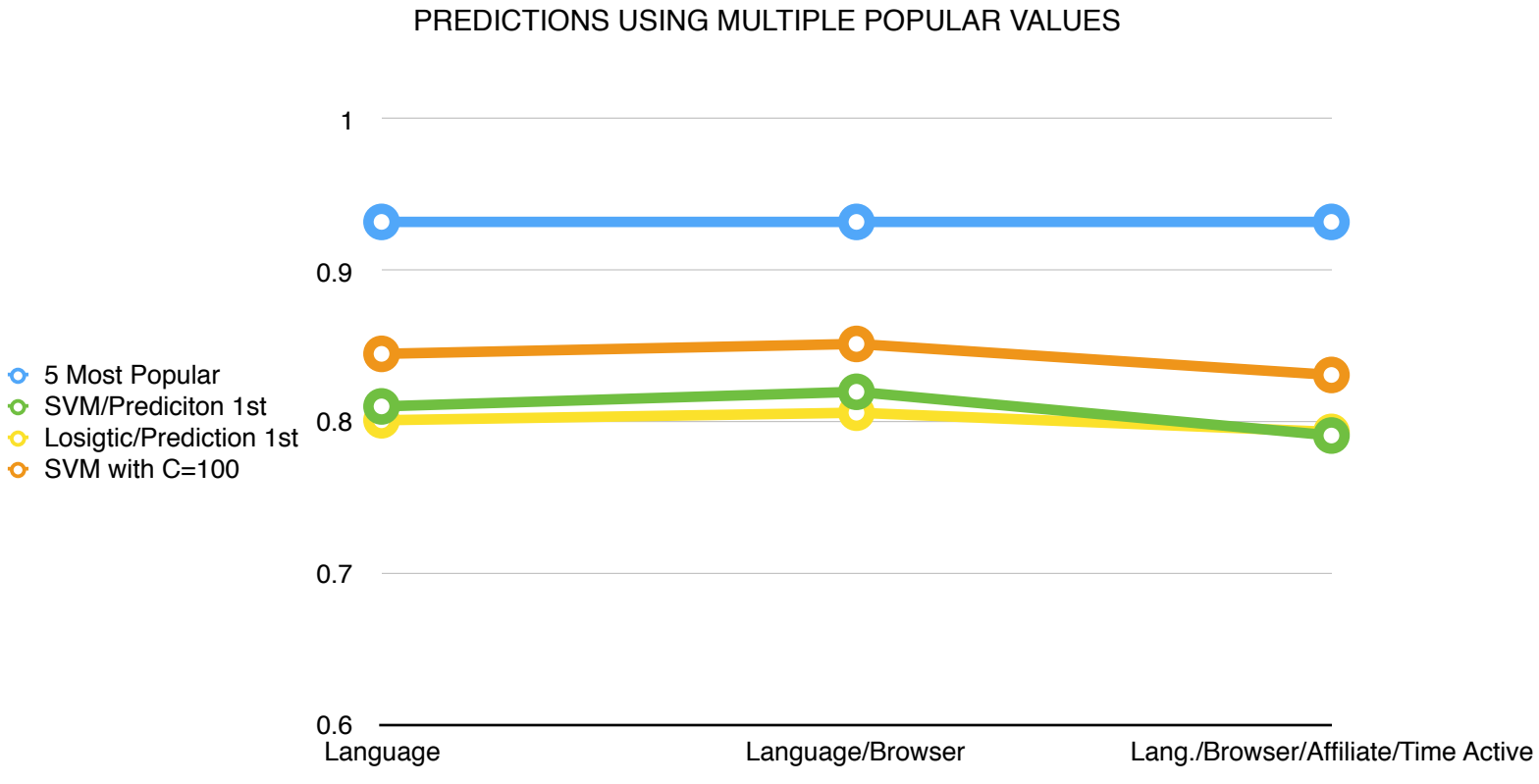
instances in which the target value was ‘NDF’ also had ‘date first booking’ field was null. To correct for this bias in the dataset, I removed all ‘NDF’ instances from the training and validation sets, and simply predicted ‘NDF’ on all instances in the test set with null booking dates.

Predictably, the results of the tests became much more accurate using this methodology. Re-running the experiments for the 4 feature combinations above yielded results above 70% accuracy in subsequent submissions. However, this was only slightly better than the baseline simply predicting ‘NDF’ and was worse than competition baseline of 76% derived from predicting both ‘NDF’ and ‘US’.

It was clear that it would be necessary to make use of all five available predictions in order to achieve the best possible results. I once again trained a basic predictor to serve as a baseline for the experiment. This predictor ranked all ‘NDF’ in the first position if ‘date first booking’ was Null, ‘US” first if it was not. It swapped ‘US’ and ‘NDF’ in the opposite case and ranked the remaining three positions in ‘other’, ‘France’ and ‘Italy’ in both cases. This predictor achieved an accuracy of 93.230%.

Unfortunately, I was unable to to train a predictor that even came close to beating this baseline. Since my pervious predictions were so heavily biased to ‘US’ and ‘NDF’ outcomes, I decided to train predictors on only instances in the training set in which those target values did not appear. I then experimented putting my prediction in the first rank for all entries where ‘date of first booking’ was non-null and ‘NDF’ otherwise. I ranked the remaining entries in order of popularity as in the baseline. This set of predictions performed noticeably worse than the baseline, likely due to the popular destination of ‘US’ being related to the 2nd spot.

I then tried training a predictor with a much higher regularization parameter on all non-‘NDF’ instances. Setting the parameter to C=100 gave a much more diverse set of results, and produced my most accurate result using machine learning algorithms.



4: Literature and Concepts for further study

I found several articles in which similar travel prediction tasks had been undertaken in the past. A prior Kaggle competition (How Machine Learning Will Affect Your Next Vacation) asked users to recommend hotels to users bases on similar website profiles. While this task called for hotel popularity to be ranked using regression, the winning team began by treating it as a binary classification problem (whether the user would say yes or no so a particular hotel). The winning team's experiments relied on neural networks and classification trees for the classification portion, and I plan to experiment with those tools on subsequent trials with this dataset.

I was also intrigued by the ability to attribute multiple target values to a particular user. While each instance in the dataset cannot truly "belong" to multiple classes, the improved accuracy from using all 5 possible target values instead of one might indicate a multi-label classification approach would be useful. *Multiclass classification* described how such a problem "decomposes into a set of unlinked binary problems" in which a collection of binary predictors can be trained on all possible target outcomes.

5: Results and Conclusions

Given that no set of features was better at predicting travel destinations than simply predicting destinations based on popularity, I am led to believe that popularity itself might be much more predictive than any other factor. At the time of this writing, the 93.230%. accuracy of the simple baseline model predicting popularity was 0.108% behind the first place solution after more than 1,000 entries.

Works Cited

"Airbnb Recruiting: New User Bookings." Evaluation -. N.p., n.d. Web. 30 Nov. 2015.

"How Machine Learning Will Affect Your Next Vacation." How Machine Learning Will Affect Your Next Vacation. N.p., n.d. Web. 30 Nov. 2015.

"Infographic: Google Chrome Leaves Competition in the Dust." Statista Infographics. N.p., n.d. Web. 30 Nov. 2015.

"Multiclass Classification." Hamel/Knowledge Discovery with Support Vector Knowledge Discovery with Support Vector Machines (2009): 183-92. Web.