Machine Learning Project Report

Prediction of New User Bookings in Airbnb

Team Members:

Thiagarajan Ramakrishnan – txr150430

Anandan Sundar – axs156730

Mohana Selvi Sriram – mxs144430

1. **Introduction**:

This project describes us about a set of users who book their apartments online using an Online Lodging Reservation System in Airbnb. In such systems, ticket reservations of various users of different age, gender would be done in various countries. Each and every booking would be considered as a session. Based on these given set of attributes, we would be predicting the location from which reservations would be done on larger numbers. New users of reservation system in Airbnb can book a place to stay in 34000+ cities across 190+ countries. By accurately predicting where a new user will book their first travel experience, we can share a more personalized content with their community, decrease the average time for booking, and better forecast demand.

2. **Problem Definition:**

2.1 Task Definition:

Large amount of reservation data can be interpreted to acquire knowledge about tasks that occur in the environment. Patterns in the data can be used to predict the future events. Knowledge of these tasks facilitates the automation of task components to improve the inhabitant's experience.

Input Dataset Details:

We have 12 classification labels and 15 features in our dataset. There are 213451 training instances in our Airbnb User Booking dataset.

Some of the attributes that are present in our set is as follows:

Airbnb_dataset.csv – Training set of users

Set of Independent Attributes

- gender
- signup_method – Which account is used for login
- signup_flow: the page a user came to signup up from
- language: international language preference
- affiliate_channel: what kind of paid marketing
- affiliate_provider: where the marketing is e.g. google, craigslist, other
- first_affiliate_tracked: the first marketing the user interacted with before the signing up
- signup_app – What app is used for signup?

- first_device_type – Android / iPhone?
- first_browser – First browser in which reservation has been made
- country_destination: this is the **target variable** we will predict

sessions.csv - web sessions log for users

- user_id: to be joined with the column 'id' in users table
- action
- action_type
- action_detail
- device_type
- secs_elapsed

countries.csv - summary statistics of destination countries in this dataset and their locations

age_gender_bkts.csv - summary statistics of users' age group, gender, country of destination

Based on the combination of the files present a consolidated set of file was generated from the list of files present. **In that file around 13000 instances were chosen and 9 attributes were chosen which is listed in the next section of the document.**

2.2 Algorithm Definition

Some of the Machine Learning algorithms and the definitions for them which we are going to implement in this account for computing the target variable which we are going to predict now are as follows:

<u>Decision Trees</u>
DTs use training instances to build a sequence of evaluations which can be used to permit the correct category (prediction). This algorithm hence can be used to identify the countries from which prominent number of users would be booking tickets upfront. Best attribute that can be used to split the attribute set is done based on information gain, which can be calculated based on Shannon's entropy.

<u>Neural Networks</u>
Artificial neural networks are relatively crude electronic networks of neurons based on the neural structure of the brain. They process records one at a time, and learn by comparing their classification of the record with the known actual classification of

the record. The errors from the initial classification of the first record is fed back into the network, and used to modify the networks algorithm for further iterations.

Naïve Bayes Classifier

We use Bayes probabilities to determine the most likely next event for the given instance for all the training data. Conditional probabilities are determined from the training data. Based on those values, classification would be done.

Support Vector Machines

Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

Random Forest

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement if bootstrap=True (default).

Bagging

It is a method which generates multiple versions of predictor by bootstrap samples and using them to get an aggregated predictor. The aggregation averages over the versions when predicting a numerical outcome and does a plurality vote when predicting a class.

Boosting

The weight of all training samples would assigned equally. Then training on the model is done. Based on the error calculated in the iteration, we would increase the weights of incorrectly classified data. This process would be repeated until accurate prediction of weights is done for the model. Boosting are of two types. Ada Boosting and Gradient Boosting. Both algorithms are implemented in our system.

k-NN algorithm was not used for implementation in our system as the dataset had lot of non-numerical values. Hence finding a nearest neighbor is complicated.

3. Experimental Methodology

3.1 Methodology

The above mentioned algorithms would be used to predict the country from which users would predominantly be making their room reservations. We are provided with list of users along with the demographics, web records and some summary statistics. **We will be predicting the country a new user's first booking destination.** There are 12 possible outcomes of the destination country: 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL','DE', 'AU', 'NDF' (no destination found), and 'other' (which are other countries that doesn't belong to this set).

We have predicted the values using following set of classifiers for addressing this problem:

- Bagging
- Ada - Boosting
- Gradient – Boosting
- Random Forest
- Naïve Bayes
- Support Vector Machines
- Decision Tree
- Perceptron
- Neural Networks

We had split the data into 80-20 ratio and ran it for predicting the values in the dataset. This helps us in predicting the country of the user from where he will be booking his reservation with high accuracy. The results which were obtained during the process has been presented below. We also use 3 fold cross validation technique to further improve the learning.

3.2 Results and Discussion

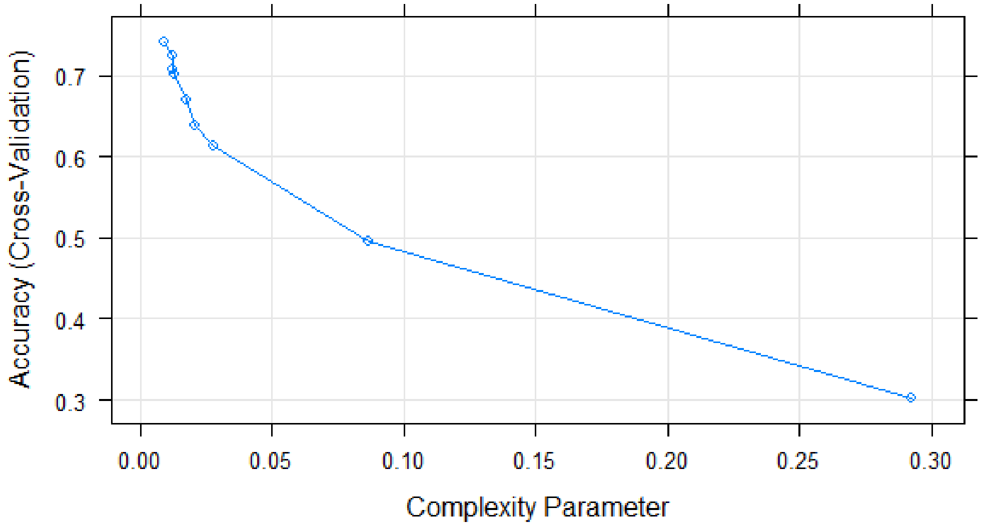
The accuracy values for ensemble methods have been calculated using cross validation. The non-ensemble methods accuracies have been calculated with and without cross validation have been tabulated as shown below:

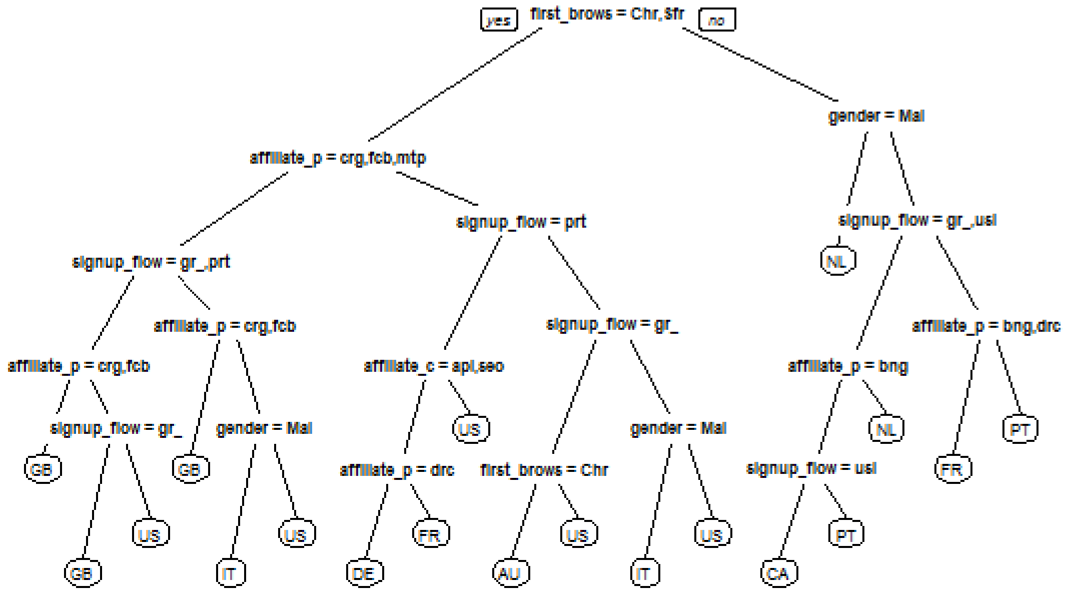
Decision Tree (80 – 20 split)	:	78.00926
Decision Tree (3 fold cross validation)	:	74.45729
Neural Network (80 – 20 split)	:	98.8209

Neural Network with (3 fold cross validation)	:	87.2268
Perceptron (80 – 20 split)	:	77.7392
Support Vector Machines (Linear Kernel) (80 – 20 split)	:	68.82716
Support Vector Machines (Radial Kernel) (With 3 fold Cross validation)	:	97.09881
Naïve Bayes (80 – 20 split)	:	73.95833
Random Forest (3 fold cross validation)	:	97.77008
Bagging (3 fold cross validation)	:	97.63114
Ada Boosting (3 fold cross validation)	:	98.00933
Gradient Boosting (3 fold cross validation)	:	93.91193

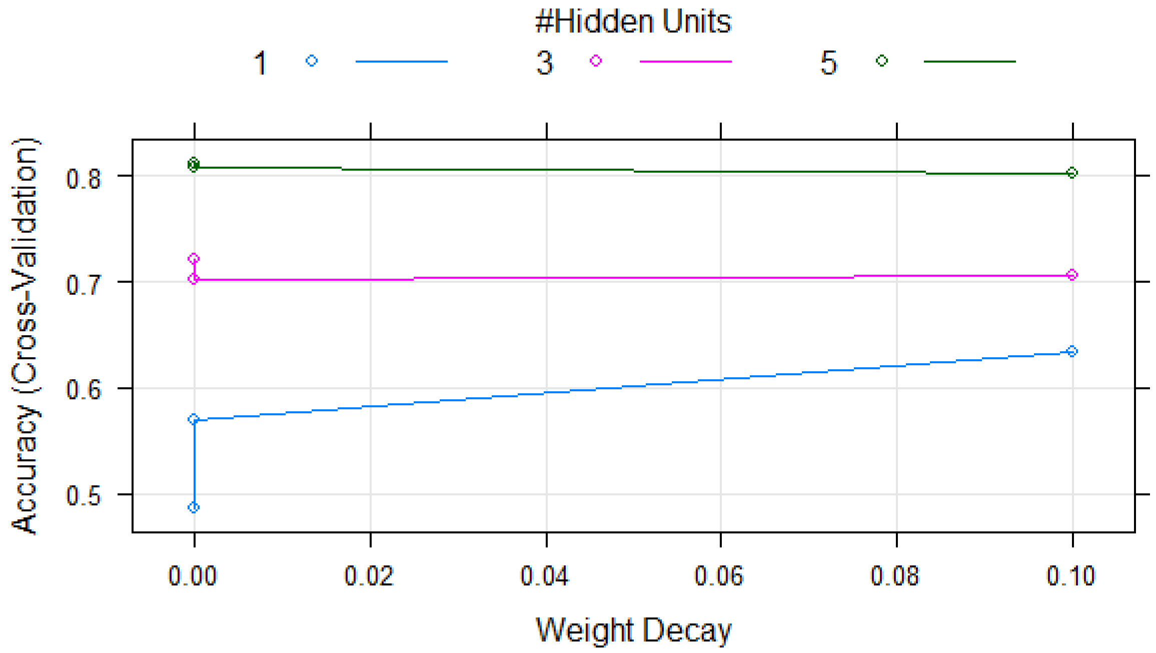
Few other interesting plots which we made during the process:

<u>Decision Tree</u>

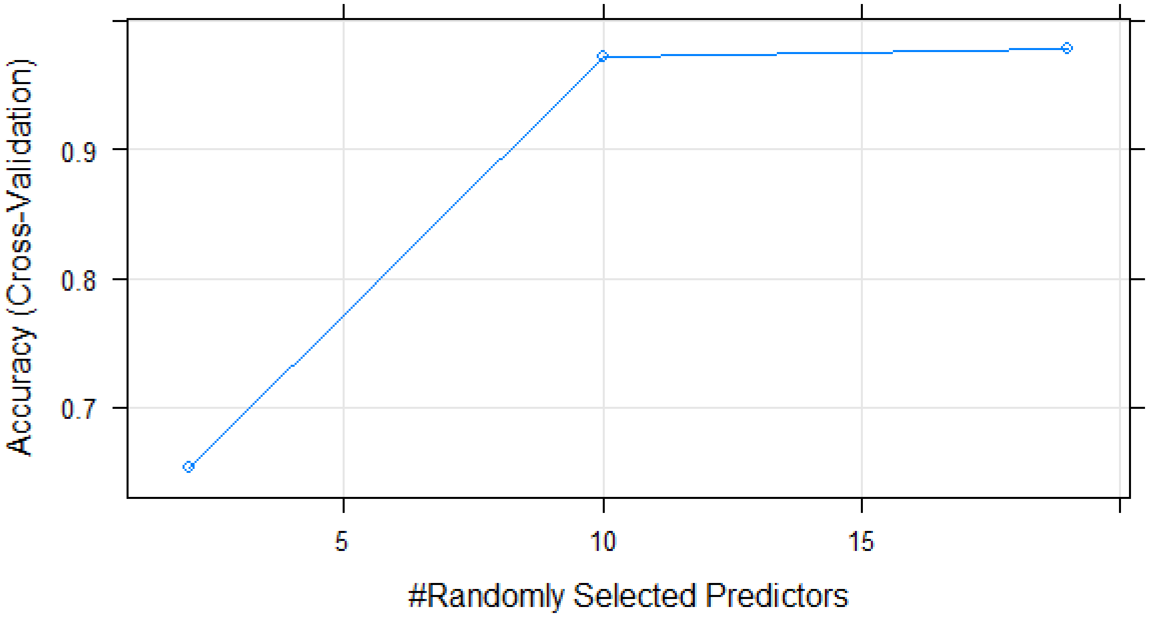




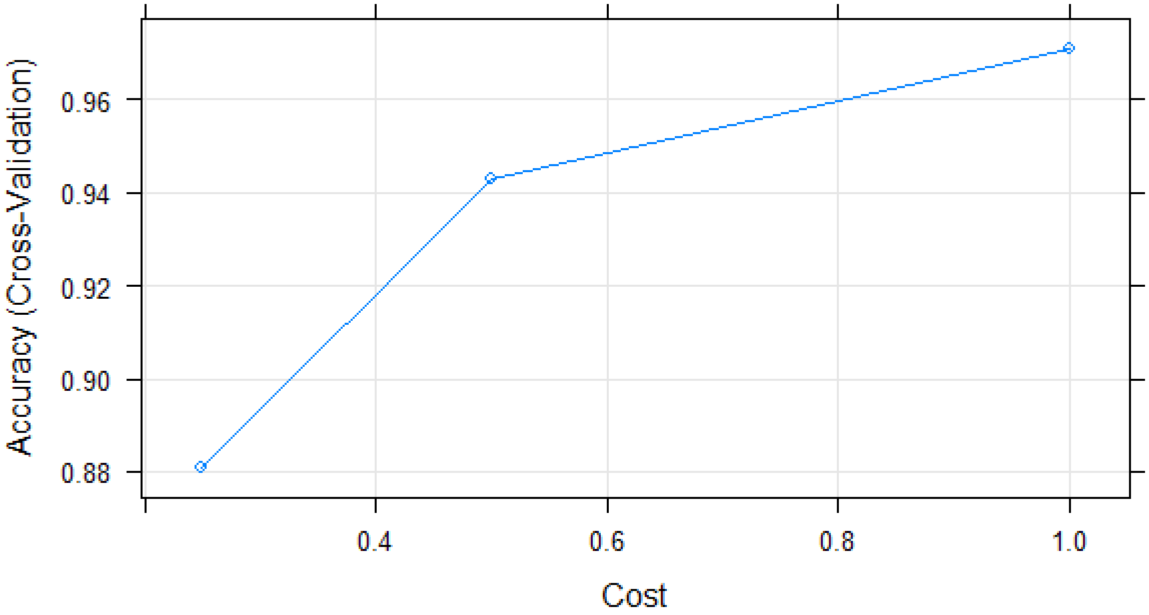
Neural Network



Random Forests



Support Vector Machines



4. Related Work

In this project, we were given a list of users along with their demographics, web session records, and some summary statistics. You are asked to predict which country a new user's first booking destination will be. All the users in this dataset are from the USA. There are 12 possible outcomes of the destination country: 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL','DE', 'AU', 'NDF' (no destination found), and 'other'. Please note that 'NDF' is different from 'other' because 'other' means there was a booking, but is to a country not included in the list, while 'NDF' means there wasn't a booking. When we process the data, **we have trimmed the NDF and other destination possibilities since they contain many empty values.**

5. Future Work

The data set contains lot of non-numerical values. Hence it is difficult to implement k-NN classifier for this dataset. In addition to that, most of the data that exists are having faulty records. Lot of destinations are not found for the present dataset. Using that for training the classifier degrades the performance. Hence we can try collecting efficient information for providing further accuracy. Other key issues in the account were already addressed during implementation.

6. Conclusion

From these experiments we can see that the classifier selection depends on the data. In addition to that, when we are dealing with data which has multiple classification attributes, using ensemble methods will be the ideal choice. The time taken for execution of the ensemble methods however, is large. It was also found that the order in which time it took for training the datasets were as follows: Neural Networks, Support Vector machines, Decision Tree, Perceptron and Ensemble methods. But the reverse was found for testing the datasets.