Destination Seer for Airbnb Using DeepLearning

Aman Arora (1001667638), Krutiben Savaliya(1001660964)
Computer Science Department
University of Texas at Arlington
Texas, 76010

Abstract

The report describes the dataset that has been provided by the Airbnb on kaggle for the competition on predicting destination country of the user and the deep learning model we created for making our prediction. First of all we describe the dataset we used, exploring all the features for our model that we thought were important for training our model and increase its accuracy. After carefully going through various research papers and learning about the models that they used and learning about the accuracies they obtained, we used Neural Networks for our project to improve the accuracy of the prediction. Then we carry out the self analysis on the project that we carried out. Finally the conclusion is provided in the report along with the accuracy that we achieved through our model.

1. Introduction

The project makes the prediction of the top five countries to which the first time user on the Airbnb website, will like to visit. For exploring how the data can be used to enhance user experience and can be optimized for converting the new registration into business we went through various research papers. (Subramanium, 2015) models the problem as the classification problem and random forest classifier pruned with importance of features to be a very good model for predicting the destinations preferred by the users. Further readings in (Zhange, 2015) makes it more cogent to how to use the data for making the accurate prediction using the two level classification model. This lead us to explore the data in an efficient way and cast out the important features for our model and get rid of the data that was least important or didn't make any significant contribution for basis our prediction. We made use of Neural Networks for making our prediction, and increased its accuracy by keeping the relevant set of features and adjusting the hyperparameters to increase its efficiency.

2. Dataset description

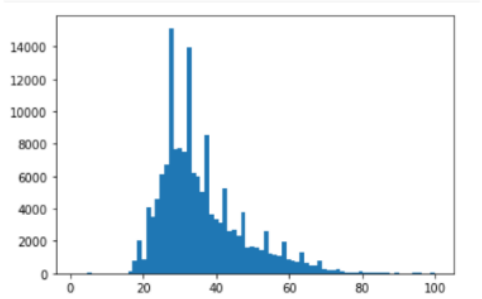
The dataset that we use for carrying out over research is provided by Airbnb for a competition on Kaggle. The files contains the information about the users, records of their sessions and various other statistics that can be vital for making our prediction. The data has been split up into five csv files named as train-users, test-users,sessions, countries, age-gender-bkts.

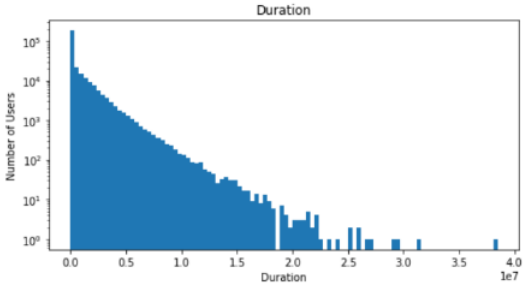
- **train-users and test-users:** The train-dataset contains the data for 213,451 users and the test-users data has the data of 62,096 users for training and testing our model. The fields that are included in these datasets are:- id, date-account-created, date-first-booking, gender, age, signup-method, signup-flow, language, affiliate-channel, affiliate-provider, first-affiliate-tracked, signup-app, first-device-type, first-browser, country-destination, timestamp-first-active.
- **sessions:**This file provides the logged session records of the users and has got the fields such as user-id, action, action-type, action-detail, device-type, secs-elapsed.
- **Countries:** This dataset has got the list of the destination countries of the users and has certain properties associated with them such as their longitude, latitude, languages spoken commonly in the respective countries and more.
- **age-gender-bkts:** This file lists the age-groups of the user, gender and the destination country. It contains the destination countries based on the age bracket, it has also got the gender information of the users.

Further exploring the data it was analysed that the most common language that was used by the users was english and it was observed that the people who travelled were in the age bracket of 24 and 36.

We summarize the session's data so that we can extract the important features out of it and then after summarizing the data we then join it with the train dataframe.

- In summarizing the sessions we create the value counts for the userid which returns the count of unique userids. Further we set the idcount to the count of the number of times the unique userids have been in the session and we set its index to the user id. And then we reset the index to userid. Further we get the duration for which the user remained in the session. For this we create a dataframe useractive and we group it by the id of the user and sum out the total time spent by each user in seconds. We use a function to get the most common values for all other





features that uniquely relate with the userid in our sessions dataset.

- Also we obtain the number of unique values for each feature associated with each userid so as to take all the unique values into account before making our prediction. Time spent by the user on the website in the session plays a crucial role while making our prediction thus obtain the minimum and maximum time spent by the user in each session and after obtaining the results we merge it as another feature in our dataframe which adds more value to our dataset for making the correct prediction. Also in our dataframe we include the averagetime spent.

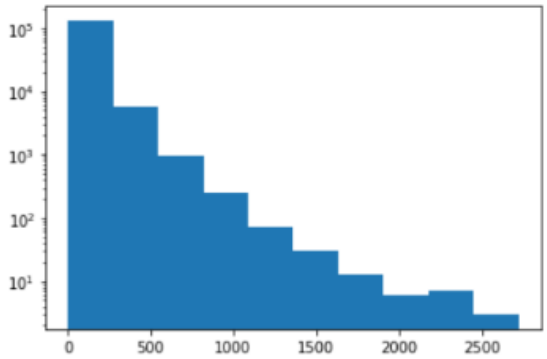
  The type of the device that the user uses also makes an impact in making our prediction so for this we add more features to our dataframe which are of the categories such as appledevice, desktopdevice, tabletdevice or mobiledevice. For this we do the one hot encoding.

- Finally we merge the session summary dataframe with the training set and also the userids that didn't match with the ones in the sessionsum we extract them out from the original training set and then concatenate it to obtain the train data and same we do with the test dataset.

  We remove all useless and missing data from the frame. One of the other important things in the data is to make sure that the accountcreated and firstbooking features are in the proper format of date. Hence, converting these features in datetime data type is important.

- When these features are converted in proper format, it will be easy to extract year, month and year including some special cases like holidays and business days the account was created or booked for the first time.

- Determining type of browsers can also be a factor to consider while learning the model. The Browsers are categorized in mobile browser or other. Same goes for language. The languages which are rarely selected by users are grouped in the same category as other. The cut-off set for languages to be considered as other is 275.



- Further we normalization of all features by making use of MinMaxScaler() method. After processing everything on data, we split the data back to training and testing set.

- After splitting the data, The deep learning model is the only final step left to find accuracy and predicting countries. Here we add inputnodes which is same as the number of columns train and test data. The model is implemented using tensorflow library. It has one input layer, two hidden layers and one output layer. .

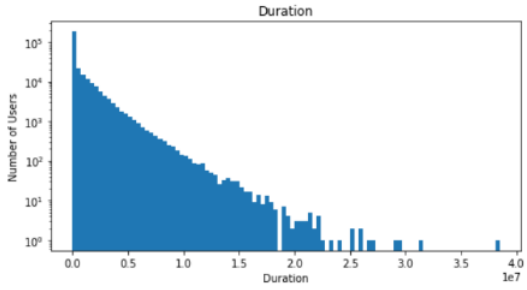
3.Project description

3.1 Description of the project:

The project is about predicting destination countries for first time users of AirBnB. We found the dataset for the project from Kaggle. The dataset includes bunch of csv files. We used four csv files out of all. Pre-processing the data was the hardest of everything to implement. Pre-processing includes filing the missing data, normalizing features, truncating redundant values, etc. For example, date feature was not to the proper format. Conversion of date in the format was done followed by fetching some uncovered features from date like holidays the first booking has been done or holidays the account was created. Pre-processing has covered 70 percent of the work. The next step was to convert the data in such a way that it goes to tensorflow model. The project is made with the help of Tensorflow library. The model has 1 input layer, 2 hidden layers and 1 output layer. The model has 2000 number of epochs and multiplier factor as 1.33. The number of inputs going to the model is 186, which is the same as the number of columns in dataframe. Over the loops for both testing and training data, we computed accuracy and loss for every epoch and observed the result linearly increasing.

3.2 Main references used for the project

The main reference for the project was a blog which is publicly available on the internet. The blog is by Tanmayee W. It uses XGBoost model and since the data preprocessing and



cleaning plays the major role thus walking through the steps that were taken by the author to excute this project was helpful to get the understanding of the dataset and to lay out foundation for us to how to go about cleaning our dataset. Kaggle is another good source that helped a lot in finding ideas, specially the thread which includes this as the competition has got the discussion section which, further helped to get much more in-depth understanding about the approach to follow. And knowing which models that were used by the other people and what results they obtained through their implementation, it became easier to focus in the right direction rathar than trying out our hand on different models and wasting the time to enhance the accuracy using the models that have already been implemented and are less accurate. Also some research papers are available on google scholar for the same topics. To search more about every reference, you will be able to see the links at the end of project and go through them.

3.3 Difference in APPROACH / METHOD between your project and the main projects of your references

The approach used in primary reference is way different than the approach we used for our implementation. Further explicating about the approach, the references that we gone through uses a machine learning model called XGBoost, Random forest,SVM, multi-class-one-against-rest logistic classification which provides the less accuracy in predicting the top most 5 destination countries that the user would like to visit. And for our project in order to increase the accuracy and surpass the accuracy that has been already achieved by the existing models we make use of the deep Learning model using tensorflow library. However, data preprocessing that has been carried out in the references and in our project is almost similar.

3.4 Difference in ACCURACY/PERFORMANCE between your project and the main projects of your references

The statistics show that the main reference gives accuracy 87 percent. We Were able to get almost 7 percent more than the reference, which makes it 94 percent. The exact statistics for both the training and testing data for our version of implementation are 97 percent for training data and 94 percent for testing data.

3.5 List of your contributions in the project (your work)

As mentioned earlier in one of the other sections, the model we implemented for prediction purpose is way different than what is previously implemented. After careful reading of various research papers, we got to understand that neural network fits better than machine learning for the project. Hence, we thought to work on deep learning. That resulted in a successful project. One of other different things we implemented is that we predicted top five countries for users, not just one country.

4. Analysis

4.1 What did I do well?

Getting the full understanding about the dataset that is involved in making the prediction was the foremost thing that was required to be done. Thus the approach to go through different research papers and online blogs, the discussions that were carried out over the kaggle platform where this competition was listed gave us the cogent understanding of how to go about the dataset. After analyzing the dataset it was easier to figure out which features that would have contributed the most and which were of no use to us. Further, making our knowledge strong about the data that had been provided to us, gave an advantage about correlating the missing data and fill out the gaps between them so as to make the data a reliable source for making our prediction. Further carrying out the research also helped us evading some of the machine learning models that we could have used in our project, but since they were not surpassing the accuracy of the model that won the competition, thus it was of no use to consider them and hence it helped us to save a lot of time to carry our research in the right direction and come up with an efficient model that was able to surpass the existing highest accuracy. Further coming to our model, the selection of the hyperparameters to train our model played a crucial role such as number hidden layers to keep, number of nodes to pass in the input layer, the batch size to keep, number of epochs to take so to make sure that all the nodes get to participate in our model making the prediction much more reliable. Also we tried out different optimizers in training our model and compared the different results that we obtained until we settled down for the best result that we could obtain.

4.2 What could I have done better?

The idea of going through various articles and analyzing the dataset was wise enough to save us from putting a lot of effort and still finding the results that were not up to the mark. Further enhancing the accuracy of the model that we came up with, by pruning the data further and experimenting with dropping the data that seemed less relevant might would have increased the accuracy but it would have overfitted our model. Further if time allowed we would have experimented our model with dataset that would have involved more number of destination countries, and see what results we could have obtained.

4.3 What is left for future work?

The extended thought after the successful implementation of the project is to go further and find the destination cities from the top selected countries. As per the thought, it should work something like the following. The project is already making prediction on top five countries. In addition, we have thought that we will run deep learning again on those countries to search for top five cities which user would want to visit. We have a thought to add a feature called feedback, which will be used by users who have already visited some places. We will try to collect as many feedbacks as possible. It is not sure what model to use to process the feedback and get the best output. The feedback will give the best suggestion to first time users of AirBnB besides prediction of destination countries and destination cities.

5. Conclusion

Based on Kaggle's evaluation method (https://www.kaggle.com/c/airbnb-recruiting-new-user-bookingsevaluation), the neural network scores just under 0.86 when all of the training data is used. The winning submission scored 0.88697 and the sample submission scored 0.68411. Looking at the leaderboard and some kernels, the most common algorithm for the better scores was XGBoost. Given that the purpose of this analysis was to further our knowledge of TensorFlow (in addition to the other aspects of machine learning - i.e. feature engineering, cleaning data, etc.), We do not feel the need to use XGBoost to try to make a better prediction. We are rather pleased with this model on a whole, given its ability to accurately predict which country a user will make his/her first trip in. The 'lazy' prediction method would be to use the top and top 5 most common countries for the predictions. This would equal an accuracy score of 58.35 percent for the top predictions and 93.9 percent for the top 5 predictions. For the testing data, my top predictions scored a higher accuracy of 62.37 percent, as well as for the top 5 predictions, at 96.87 percent. Our predictions are also more useful since they make use of all twelve countries, instead of just the five most common.

6. References

W, T. (2018). Predicting destination countries for new users of Airbnb. [online] Medium. Available at:https://towardsdatascience.com/predicting-destination-countries-for-new-users-of-airbnb-eb0d7db7579f [Accessed 5 Oct. 2019].

Kaggle.com. (2015). Airbnb New User Bookings — Kaggle. [online] Available at: https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/overview [Accessed 10 Oct. 2019].

Thanapalasingam, T. (2016). AirBnb: Customer Destination Prediction — Kaggle. [online] Kaggle.com. Available at: https://www.kaggle.com/thiviyan/airbnb-customer-destination-prediction [Accessed 7 Oct. 2019].

Winfield, M., Castellano, R., Kimmel, Y., Jiang, J., Li, H., Zhong, X., Srivastava, P., Guo, W., Liu, Z., Liu, A., Zhang, H., Galgali, P., Dogan, T., Guyton, A. and Galgali, P. (2016). Predicting a New User's First Travel Destination on AirBnB (Capstone Project) — NYC Data Science Academy Blog. [online] Nycdatascience.com. Available at: https://nycdatascience.com/blog/student-works/predicting-new-users-first-travel-destination-airbnb-capstone-project/ [Accessed 8 Oct. 2019].

Avireddy, Srinivas, Sathya Narayanan Ramamirtham and Sridhar Srinivasa Subramanian. "Predicting Airbnb user destination using user demographic and session information." (2015).

Zhang, Ke. "The Prediction of Booking Destination On Airbnb Dataset." (2015)

Dissertation or Thesis
Clancey, W. J. 1979b. Transfer of Rule-Based Expertise through a Tutorial Dialogue. Ph.D. diss., Dept. of Computer Science, Stanford Univ., Stanford, Calif.

Forthcoming Publication
Clancey, W. J. 1986a. The Engineering of Qualitative Models. Forthcoming.