SAVITRIBAI PHULE PUNE UNIVERSITY

A
PROJECT REPORT
ON

"AIRBNB CLASSIFICATION"

A project work submitted to the Department of Computer Engineering, SITS , Narhe , Pune. In the fulfillment of the requirements for

SOFTWARE DEVELOPMENT LAB
Third Year (Computer Engineering)

By

1. **(Piyush Pachare)--------(44)**
2. **(Aditya Patil)--------(48)**
3. **(Gaurav Uke)--------(66)**

Under the guidance of : **Prof. Neelam Thorat**

Sinhgad Institutes

DEPARTMENT OF

COMPUTER ENGINEERING

SINHGAD INSTITUTE OF TECHNOLOGY & SCIENCE, NARHE,PUNE

(2020 – 2021)

SINHGAD TECHNICAL EDUCATION SOCIETY'S

SINHGAD INSTITUTE OF TECHNOLOGY AND SCIENCE, NARHE, Pune – 411 041

Sinhgad Institutes

DEPARTMENT OF

COMPUTER ENGINEERING

CERTIFICATE

This is to certify that final project work entitled “AIRBNB Classification”was successfully carried by

(Piyush Pachare)--------(44)
(Aditya Patil)--------(48)
(Gaurav Uke)--------(66)

in the fulfillment of the Software Development Lab course in third year Computer Engineering, in the academic year 2020-2021 prescribed by the Savitribai Phule Pune University.

Guide **Prof. Neelam Thorat**

H.O.D **Prof. Geeta Navale**

Guide’s Name **Dr. G.S. Navale**

Dept. of Computer Engg. Dept. of Computer Engg.

Principal

Dr. R. S. Prasad

INDEX

INTRODUCTION

Airbnb is an American online marketplace company, enabling people to lease or rent short-term lodging including vacation rentals, apartment rentals, homestays, hostels beds, or hotel rooms. New users on Airbnb can book a place to stay in 34,000+ cities across 190+ countries. The company does not own any of the real estate listings, nor does it host events, it acts as a broker, receiving commissions from each booking.

***Project overview*:**

This project focuses to find an answer to the question of how Airbnb comes to know where you are going to Book your First Travel Destination by Solving Airbnb New User Booking Prediction problem using Machine Learning.

Hardware/Software requirements:

Technical requirements:

Name	Details
Processor	Intel Core2Duo
RAM	4GB

Software Requirements:

Name	Windows/Linux
IDE	JupyterNotebook
Language	Python

Documentation created using Google Docs.

SOURCE OF DATA

The Dataset is collected from the Kaggle which is from Airbnb New User Bookings Competition.

- The Dataset contains a list of users along with their demographics, web session records, and some summary statistics 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'.
- The Dataset is in CSV formats Train_users.csv, Sessions.csv, Countries.csv, Age_gender_bkts.csv.

SOLVING APPROACHES

Existing Approach:

- Existing approaches include using only the data that is in Train.csv file and building two-level classifiers.
- Using Mean, Median, and most common value imputation for missing values and Building neural networks with limited hidden layers.
- Removing highly correlated features and building an ensemble by using accuracy as metrics for parameter tuning.

First-Cut Approach:

- Feature extraction is only done on Train.csv data with some changes from the existing approach.
- All Missing or Nan values in the data in Some Feature is imputed by Model-Based Imputation.
- Used all features with no feature selection with Conversion of all categorical variables to one hot-encoding type features.

Final Improved Approach:

- Use both data Train.csv and Session.csv for feature extraction and feature engineering.
- Use Xgboost for feature importance to perform feature selection.
- Use Feature binning and some advanced feature engineering techniques to improve performance.

DATA ANALYSIS AND VISUALIZATION

Data analysis helps in gaining insights from data to solve a problem and exploring data to understand it.

We have used visualization extensively to represent the data in an attractive and easy to grasp manner.

- The Packages used for visualization were seaborn and matplotlib-pyplot
- First Visualization Gives us age distribution we can see maximum booking came from age group of 20 to 40 years
- Second Visualization shows that most of user are booking from web application
- Third Visualization gives year wise distribution of bookings.It can be seen the consistent increase of booking from 2010 to 2014

DATA PRE-PROCESSING

Discovering the format of data that the machine learning model can understand and construction of features.

- Preprocessing Include Checking Nan/Null Values, Duplicates in Data and removing Unwanted Data from Features
- Grouping Data by user_id and obtaining unique value count of categorical features, Mean and Std of its occurrence.
- This Step includes Feature extractions like an hour, days, week, month, a year from date_account_created and first_active feature for each Datapoint.
- Extracting Season feature and difference in seconds from account_created and first_active Features.
- Converting all categorical variables to Binary encoding features.
- Feature binning for seconds_elasped and Age Feature to get a vector of some length.
- Performing Standardization on data for modeling.

ONE HOT ENCODING

- One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction.
- We use one hot encoder to perform “binarization” of the category and include it as a feature to train the model.
- In Our project we have created binary labels for our target variable 'gender', 'signup_method', 'signup_flow', 'language', 'affiliate_channel', 'affiliate_provider', 'first_affiliate_tracked', 'signup_app', 'first_device_type', 'first_browser'and split them into several target variables i.e - each category has a label ‘1’ or ‘0’.

MACHINE LEARNING MODELING

Modeling refers to the process of training machine learning algorithms, tuning its parameters, evaluating it using a metric, and finally making it fit to predict a new data point.

XG-Boost:

- XG-Boost performs effective tree pruning when compared to other ensemble algorithms.
- XG-Boost utilizes parallel processing by using all cores on the system at the same time.
- XG-Boost has an inbuilt capacity to handle missing values.

Testing and Maintenance

This project has been developed accordingly for having the least maintenance cost and testing purposes. Since only two modules are major which would require constant changes with increasing leads.

CONCLUSION

Depending on the data of previous booking done in a specific period of time by a particular group of people classified on the basis of gender and age we can use this project as a tool to predict the future bookings at a particular place with good mount of accuracy achieved by using various machine learning modules eg. XGboost.

This data will help to target the specific customers and increase the frequency even more.