Big Data

(BUDT758B)

Project Title: Expedia Hotel Recommendation

Team Members:

Sree Pradyumna Davuloori, LakshmiNarasimhan Narayanan, Sathya Anurag Siruguppa

ORIGINAL WORK STATEMENT

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

TABLE OF CONTENTS

I. EXECUTIVE SUMMARY

Predicted hotel clusters using expedia hotel recommendation kaggle dataset. Used Big Data Technologies such as SparkR, Pyspark, SparkSQL. Products were built on using multiple platforms such as Databricks, AWS and RShiny. Used Machine learning algorithms such as Random Forests, k-means clustering, logistic regression along with a custom algorithm for predicting the hotel_clusters.

II. DATA DESCRIPTION

Data source: https://www.kaggle.com/c/expedia-hotel-recommendations/data

We have taken the dataset from kaggle expedia hotel recommendation data set. The data set primarily consists of two datasets, training.csv which contains 24 columns of which most of the columns are int, tinyint and double. Date_time, search_ci and search_co are string, which are actually the timestamps. All the country, continent names, hotel cluster and user ids are anonymized in the form of numerical values

Data attributes like site_name, posa_continent are ID related to the expedia point of sale giving the country and continent of the sale. User_location_country, user_location_region and user_location_city are user location related attributes. Attributes like hotel_country and srch_destination_id are hotel related attributes. Here there are some binary attributes like is_booking which tells us if the user has booked in that particular hotel or not, similarly we have is_package, is_mobile which are yes or no (1,0) binary attributes.

Destinations.csv has 150 columns and srch_destination_id. The 150 columns don't have specific column names as the data is anonymized. These are some attributes related to the hotels which have been anonymized and normalized. The destination.csv is not very insightful for creating visualizations or getting any business insight.

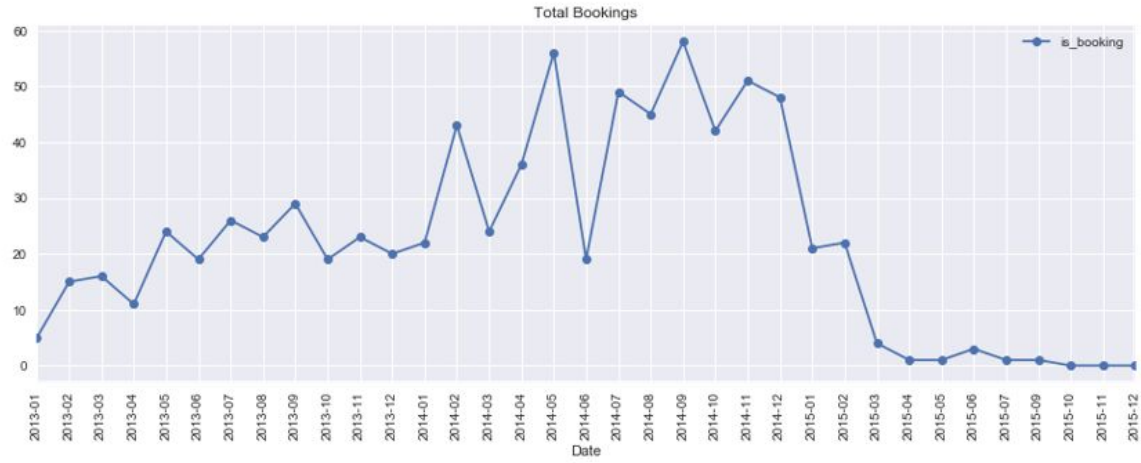
III. DATA PREPARATION AND EXPLORATION

Principal Component Analysis

In destinations.csv there are high number of attributes which of which many are highly correlated. Since we wanted to decrease the number of attributes used for the analysis, we implemented principal component analysis to explain most of the data using reduced number of components. So after performing PCA, we reduced the number of attributes to a total of 3 columns. This is an useful tool as it reduces the computation power and also improves the model performance by eliminating correlations.

Search_co and search_ci were in string format which were not usable for analysis. Hence we separated the date by day, month and year by creating new columns to store the check in and check out dates for analysis purposes

Exploratory data analysis



We did a basic timeline analysis of the bookings done by using the total count of the is_booking attribute against the month-year combination which was extracted from the date_time time stamp. We observed peaks in certain months of the data.

IV. MACHINE LEARNING MODELS

After the required pre-processing to the data was done, which included loading the data into SparkR, cleaning the data, performing a principal component analysis, we went on to fit a few important machine learning models that we felt would perform the best on our particular data.

The aim for this dataset was to accurately predict a hotel cluster based on a variety of information including user and hotel based information. This was a classification problem with the target variable being a categorical variable. Since we are trying to predict clusters and there can be a lot of important information gained by clustering and finding out similar characteristics among hotels and users. These similar characteristics maybe extremely useful for expedia when recommending certain hotels to certain users.

The language used for most of the machine learning was SparkR. It was hosted on the databricks platform to enable distributed computing and utilize the spark environment.

The final three models we decided on, for the classification task are:

1. Multinomial Logistic Regression

We built a multinomial logistic regression function in the SparkR environment. It is used to estimate the probability of a response based on the predictors.

It gave an accuracy of 0.94% over 150 class variables, each corresponding to one hotel cluster. This accuracy of 0.94% is pretty good and performs decently well compared to the random guessing model(baseline) model whose accuracy is 0.66%.

But there was room for improvement and we decided to go for better and more advanced model.

2. Random Forest

Random forest belongs to a class of methods called ensemble methods that have been known to perform very well on a wide range of problems. Random forest constructs a large number of trees on various samples of data using a subset of predictors. These trees constructed are

uncorrelated as it takes a subset of the predictors and therefore we get a low variance low bias model which performs well.

Random forest gave an accuracy of 18.14% which is a performance much better than the simple logistic regression model.

3. K-means Clustering

Although the performance given by the random forest model was excellent, we went ahead and decided to perform a k means clustering model anyway. This would allow us to cluster the data and find any general characteristics among users and what sort of hotels they book.

Although k-means clustering gave an accuracy of 0.73% which was just marginally better than the baseline model, it will allow expedia to gain valuable information on user profiles.

V. RESULTS AND FINDINGS

We found that using big data technologies and distributed computing to process large amounts of data, perform exploratory data analysis and run machine learning algorithms to gain valuable insights into the data.

VI. PRODUCT DEMO

We built two products for Demonstration:

1. **Product1**: Uses traditional machine learning algorithms such as randomforest to predict the hotel cluster.
2. **Product2**: Uses custom algorithm to predict hotel clusters and provides visualizations that could be used by Expedia to improve their business.

Product1: Traditional machine learning algorithms

The product runs randomforest algorithm in the backend to generate predictions. The product is built using SparkR in the backend and frontend is built using Rshiny. It takes “site_name”, ”posa_continent”, “user_location_country”, “is_booking” as input and it predicts the hotel cluster as output.

Product2: Custom Algorithm

The product was built in Databricks that uses SparkSQL in the backend. The applications takes data from the user and dynamically displays results in accordance with user queries.

The most important features for popularity of a hotel cluster is based on Bookings and clicks. “Is_booking” is a column in the dataset, when it corresponds to 1, booking is made and when it corresponds to 0, only clicks are made.
So we built an algorithm that finds the best hotel cluster for a destination. This we did by grouping the dataframe on srch_destination_id (Destination) and the hotel_cluster. For each group we found how many clicks are made and how many bookings are made. Weights of booking and clicks were derived based on an iterative process and trying out different values for them and finding when accuracy was the maximum. After performing this process we gave a weight of 1 for each booking and 0.2 for each click. Total rating for a particular group was calculated based on #booking + 0.2*#clicks.
For a particular srch_destination_id there will be multiple hotel clusters with various ratings. We sorted the hotel clusters in descending order based on rating and displayed only the top 5 hotel clusters for a given srch_destinaiton_id.
We implemented the same algorithm in PySpark to check the accuracy of the model. Model gives an accuracy of ~14% on the test set which is way more than the baseline model (0.667 %) (Refer to Prediction_Custom.ipynb/ Prediction_Custom.html file for accuracy calculations)

We also created other visualizations keeping expedia senior management team as our end customer. These visualization would help understand trends in booking. All the visualizations are dynamic and are connected to databricks in the backend and take input from the user and generates visualizations. The visualization we developed are:

1. Booking across months by countries: For each country we can find trends in booking across the months. We can also how countries are performing in comparison to each other.
2. Marketing Channel Effectiveness: We can see which kind of marketing needs to be done for a particular destination. Each marketing channel has different returns and choosing an effective marketing channel is of great importance.

3. Hotel Cluster performance: As discussed earlier, hotel cluster performance can be tracked based on number of bookings and number of clicks, so this visualization allows users to track bookings and clicks across months for a particular hotel cluster

VII. REFERENCES

1. https://www.kaggle.com/c/expedia-hotel-recommendations/data
2. http://spark.apache.org/docs/latest/sparkr.html
3. https://spark.apache.org/docs/0.9.0/python-programming-guide.html
4. http://spark.rstudio.com/mllib.html
5. https://beta.rstudioconnect.com/content/1518/notebook-classification.html