2021

Big Data & Predictive Analytics: Classification & Clustering

UNIVERSITY OF LEICESTER

Group 31 Agbo A. Abra(aaa144), Harini Arumugam(ha310), Rakunanthan(rks36)

Table of Contents

Abstract 2

Cleansing, visualizing and understanding the data 2

Cleaning of the dataset 2

Data exploration 4

Model building……..…..7

Improved model……..…7

Conclusion……..7

References 9

SCHOOL OF INFORMATICS
University Road phone: +44 (0)116 252 3887
Leicester LE1 7RH fax: +44 (0)116 252 3915
United Kingdom http://www.cs.le.ac.uk

UNIVERSITY OF LEICESTER

Abstract

This document reports on our findings in exploring a given dataset, Oscar demographics. It describes the model of choice used and its evaluation. It also describes the clusters used and its justification, as well as any decisions or actions that may be taken following your analyses.

The following assignment will discuss how big data can predicatively be analyzed in python. The following project will introduce cleansing of data, visualizing of data and explore the provided data frame. Follow up by building a predictive model and evaluate it with providing proper justification of the project. Data cleansing is an important part of machine learning. It will play the most important role here in building up the model. Better data beats fancy algorithms is believed here. A well-maintained dataset can provide better results with simple algorithms.

Cleansing, visualizing and understanding the data

Getting basic data cleaning done that is removing NAs and blank spaces, imputing values to missing data points, changing the variable type and so on[1]. In the given dataset Oscar_demographics.csv dataset contains the data regarding race, age, birth place, award type and date_of_birth several other demographics details of all the Oscar winners from 1927 to 2014: such as best director, best actress, best actor, best support actor and best support actress. This model is going to predict the type of awards won based on some of the features such as their race, origin of country, age and so on.

Cleaning of the dataset

The given dataset Oscar_demographics contains some null values, empty spaces, inappropriate format of date of birth. Therefore, we prioritized to clean the data first. According to the assessment, we are importing the csv file and consider the subset of the dataframe formed by the following columns: birthplace, race_ethnicity, year_of_award, award using pandas.

- **df = pd.read_csv("oscars-demographics.csv", encoding='mac_roman')**// used to import the csv file. Since the csv file is not in the UTF-8 format, encoding= mac_roman is applied.

- Next step is to display the first 3 rows of the subset and displaying all the distinct values for the column award in the entire subset.
- **df1.head(3)** // used to display the first 3 rows.

- After displaying, the first 3 rows there are some inconsistencies on the date_of_birth have been recorded and that for some rows, the country of origin is missing. Adding a new column ldob to your current data frame to record the length of the date of birth for each row to avoid inconsistency and displaying the distinct values in the column dob.by using **Date** function we have cleaned with the format date-month-year.

- The birthplace ending with two characters is in USA by using a function **birth_place**(replace) method is used to add the country of birth to those rows that are missing the country of birth.

- By using the function **date_bir** we have cleaned the columns date_of_birth, birthplace and added a new column award age to in our data frame to display age of the individual when she or he received the award.
- This new column will be essentially show the difference between the year of award and the year of birth.

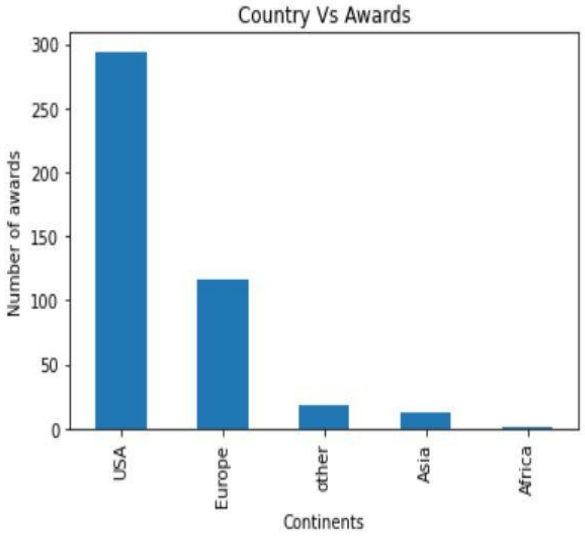
- By using the column birthplace we have added the column country that shows the origin of the country.
- The function **continent** is used to perform the above task.

Data exploration

Python language has made its analysis work more and more valuable. Many static analysis algorithms need to rely on the construction of call graphs[2]. In this part, the data needs to precisely set up the environment. First familiarizing with the built-in data structure such as list and dictionaries. Then using up the pandas python library here with the built-in function df.info() to direct the no. of rows and columns.by using the above cleaned data we need to access the impact of predictors(age, race and country of origin) on the outcome(award).we use graphs to prove or disprove the given hypotheses

1. **Most Oscar winners are from USA.**

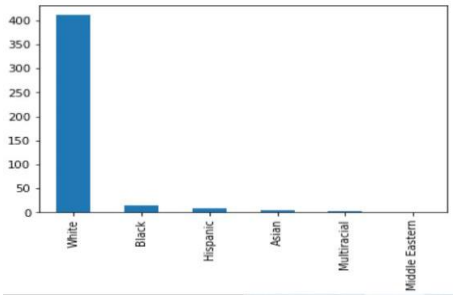
Continent												
Africa	1	2004	2004.000000	NaN	1	9	9.000000	NaN	1	29	29.000000	NaN
Asia	12	23638	1969.833333	27.408802	12	108	9.000000	0.000000	12	444	37.000000	11.847516
Europe	116	228448	1969.379310	24.944688	116	1075	9.267241	0.858287	116	5369	46.284483	12.110223
USA	294	580423	1974.227891	22.780578	294	2703	9.193878	0.629071	294	12746	43.353741	12.282679
other	18	35653	1980.722222	32.640295	18	165	9.166667	0.383482	18	753	41.833333	15.763883



2. Most Oscar winners are White.

race_ethnicity

Asian	4	7962	1990.500000	24.718414	4	36	9.000000	0.000000	4	185	46.250000	12.841988
Black	15	29937	1995.800000	20.358221	15	137	9.133333	0.516398	15	615	41.000000	10.099505
Hispanic	8	15833	1979.125000	25.709573	8	72	9.000000	0.000000	8	305	38.125000	4.517822
Middle Eastern	1	1985	1985.000000	NaN	1	9	9.000000	NaN	1	46	46.000000	NaN
Multiracial	2	3985	1992.500000	13.435029	2	18	9.000000	0.000000	2	76	38.000000	2.828427
White	411	810464	1971.931873	23.734615	411	3788	9.216545	0.697954	411	18114	44.072993	12.676059

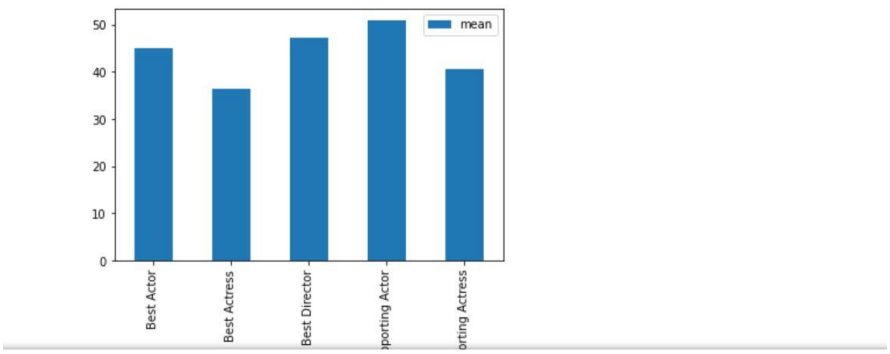


3. Best directors tend to be older than the best actor or actress.

```
In [14]: #Hypothesis 3 : Best Directors tend to be older than best Actors or Actresses.

         df4 = df1.groupby('award').agg(['mean'])
         df4 = df4['award_age']
         #for i in df4:
         df4.plot(kind='bar')
```

```
Out[14]: <AxesSubplot:xlabel='award'>
```



SCHOOL OF INFORMATICS
University Road phone: +44 (0)116 252 3887
Leicester LE1 7RH fax: +44 (0)116 252 3915
United Kingdom http://www.cs.le.ac.uk

UNIVERSITY OF LEICESTER

MODEL BUILDING

In the provided dataset age is a numeric variable and age here needs to be discretise by the help of buckets. The following buckets can be formed such as ;

Bucket 1: ***age < 35***

Bucket 2: ***35 ≤ age < 45***

Bucket 3: ***45 ≤ age < 55***

Bucket 4: ***age ≥ 55***

The data frame needs to upgrade accordingly and build more types of awards that can be added. Based on race, age, country, and origin. the data will be divided into a test and training set, showing the confusion matrix and that will help in building the model. By using cut method, we are separating into buckets 1,2,3,4 according to the above-mentioned age. The below code is used to do this above operation.

bins = [0,35,45,55,100]

df1['Bucket']=pd.cut(df1['award_age'],bins, labels=[1,2,3,4]).

Random Forests (RFs) are frequently used in many computer vision and machine learning applications. Their popularity is mainly driven by their high computational efficiency during both training and evaluation while achieving state-of-the-art results [3]. As per the task, "Random Forest" model is used to update the data frame accordingly and build up a model that predicts the award type based on age, race and country of origin. While predicting the race and country of origin, create a dummy columns for splitting into test and train data by using train_test_split() method because the race and country of origin is in the string datatype.

trainX, testX, trainY, testY = train_test_split(X, Y, test_size = 0.25, random_state = 2)// used to split the data into test and train sets.

Then using fit() method, adjusting the weights according to data values so that to get a better accuracy. Next step, for predicting the accuracy of the test and train datasets, predict() method is used.

format(metrics.accuracy_score(testY, predicted))// used to get accuracy score.

According to the task, next step is find the confusion matrix of the training and test data sets by using classify_for_threshold() classifier. Then plot a confusion matrix accordingly.

Improved Model:

Random forest doesn't give the improved accuracy for the training and test datasets, so just jump to cross validation and normalize it.

score1 = cross_val_score(clf, X, Y,scoring='accuracy',cv=crossvalidation,n_jobs=-1)// used to cross validate the train and test dataset.

Then again, split the dataset into train and training set. Then use fit() method for better accuracy in the X and Y training set. Then predit the accuracy score for the training and the set using predict() method.

K-means clustering algorithm:

This method is an unsupervised machine learning technique used to identify clusters of data objects in a data set[1]. First, input the number of clusters randomly initialize centres, then assign all the points to the closest cluster centre using KMeans() method.

model = KMeans(n_clusters=6)

Then perform normalization to find the cluster centre using

centroids = model.cluster_centers_.

Then, plot histogram, 2D plot and graph with the cluster centre.

Conclusion:

We improved the model from 0.3 to 0.5 accuracy by including the gender column and normalizing the dataset. Clusturing made it more normal and helped us visualize the grouped data.

.

SCHOOL OF INFORMATICS
University Road phone: +44 (0)116 252 3887
Leicester LE1 7RH fax: +44 (0)116 252 3915
United Kingdom http://www.cs.le.ac.uk

UNIVERSITY OF LEICESTER

References

(*1) AUTHOR Ashish Kumar PUBLISHER Packt Publishing, Limited PRINT PUB DATE 2016-02-15 Learning predictive analytics with python.*

(2) 2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE).

(3) 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops

Footnotes

[1][Add footnotes, if any, on their own page following references. For APA formatting requirements, it's easy to just type your own footnote references and notes. To format a footnote reference, select the number and then, on the Home tab, in the Styles gallery, click Footnote Reference. The body of a footnote, such as this example, uses the Normal text style. *(Note: If you delete this sample footnote, don't forget to delete its in-text reference as well. That's at the end of the sample Heading 2 paragraph on the first page of body content in this template.)*]

SCHOOL OF INFORMATICS
University Road phone: +44 (0)116 252 3887
Leicester LE1 7RH fax: +44 (0)116 252 3915
United Kingdom http://www.cs.le.ac.uk

UNIVERSITY OF LEICESTER

Tables

Table 1

[Table Title]

Column Head	Column Head	Column Head	Column Head	Column Head
Row Head	123	123	123	123
Row Head	456	456	456	456
Row Head	789	789	789	789
Row Head	123	123	123	123
Row Head	456	456	456	456
Row Head	789	789	789	789

Note: [Place all tables for your paper in a tables section, following references (and, if applicable, footnotes). Start a new page for each table, include a table number and table title for each, as shown on this page. All explanatory text appears in a table note that follows the table, such as this one. Use the Table/Figure style, available on the Home tab, in the Styles gallery, to get the spacing between table and note. Tables in APA format can use single or 1.5 line spacing. Include a heading for every row and column, even if the content seems obvious. A default table style has been setup for this template that fits APA guidelines. To insert a table, on the Insert tab, click Table.]

SCHOOL OF INFORMATICS
University Road phone: +44 (0)116 252 3887
Leicester LE1 7RH fax: +44 (0)116 252 3915
United Kingdom http://www.cs.le.ac.uk

UNIVERSITY OF LEICESTER

Figures title:

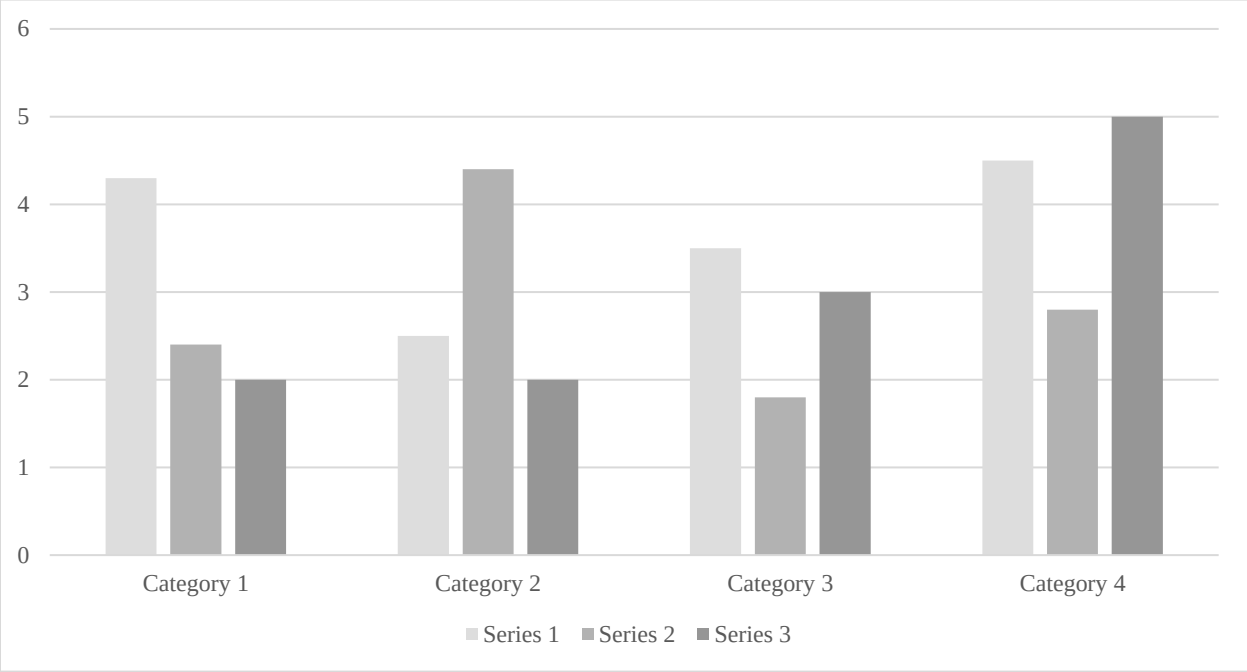


Figure 1. [Include all figures in their own section, following references (and footnotes and tables, if applicable). Include a numbered caption for each figure. Use the Table/Figure style for easy spacing between figure and caption.]

For more information about all elements of APA formatting, please consult the *APA Style Manual, 6th Edition*.