**Trends of movie Covid-19 Before and After**

Donghyeon KIM(814762)

1. Abstract

The data analysis process can be important to make a suitable model and new insight. The data analysis has four types and the data analytics process has six steps. Before analysing data and making model, the given data set is necessary to clean by filtering unnecessary data and removing incorrect, outliers, missing and duplicate values as well as normalisation. Then, exploratory data analysis is one of the critical parts of the process by identifying descriptive statistics such as the measure of tendency, frequency, variation and position, and suitable plots and graphs. According to each phase performed, the model can be developed to predict by using the numerous movie data by scraping from IMDB. The candidate models are created through diverse methods. The candidate models are continually developed and compared each other, then the most satisfied model is selected. After that, the selected model is evaluated. When this performance evaluation meets the initial purpose and follows significant results, the next phase is moved as communicating results and to operationalising such as briefing, final reports and technical documentation including code. If the performance evaluation is dissatisfied, the past phase is iterated. This report performs in accordance with the data analytics process then obtains the satisfied model in the R application.

1. Introduction

Many methods are existed to analyse trends for movies throughout the real world. Many companies would like to analyse movie trends in order to increase their profit. The movie market contracts because of Covid-19 but people tend to watch movies online than going to the theater. Then online movie platform such as Netflix and IMDB collects much information which movies people wish to watch. This information contributes what movie can be popular and followed the trends. It can be identified to the trends by using genre, user rating. The diverse a lot of variables in the given data set from website to make the best model can be possible. Hence, the solution is that a suitable model is selected by using effective methods, tools and techniques according to the data analysis process. Hence that the final satisfied model has more accurate prediction and identification of trends despite of numerous data.

The types of data analysis consist of data scraping, descriptive analytics, decision analytics, predictive analytics and finally prescriptive analytics. In addition, the components of the data analytics process are discovery, data preparation, model planning, model building, communicate results and final phase is operationalization. As the process of data analysis is departmentalized, the first step is data scraping and then pre-processing data for data analytics. When this step proceeds, the optimised can be created to analyse by data cleansing. The removed six features of dirty data such as incomplete, duplicate, incorrect, inaccurate, business rule violations and inconsistent data, are implemented to make model through techniques.

The compositions of procedure of technical implementation are data clearing methods, principal component analysis (PCA), correlation, linear regression modelling, classifying between testing, validation and testing, artificial neural networks (ANN). The making model is used by supervised and unsupervised learning. For instances of using learning in this report, supervised learnings such as a linear regression model and ANNs are performed. After the results are revealed by using that, candidate models are evaluated and compared by using statistical metrics, tables and plots. The final model is identified as close as the prediction and the accurate.

1. Technical implementation

The first phase is to understand the problem and the purpose to create a model so it is necessary to recognise the desired results. The framing this problem is the significant model to identify the trend of films before and after Covid-19 according to the procedure of data analytics. Before pre-processing the data set, the types of variable and data, as well as the definitions of the independent and dependent variable are identified such as data, numeric, character.

Second, this phase is to scrap the data in the IMDB. IMDB can be related to the films about the online database format. By using R studio, the variables can be scrapped such as title, genre, year the films launched and rating. However, the tile cannot be necessary and rating cannot be located in all of rows so the genre and year can be used in order to analyse the trend.

The third step is to prepare data in order to optimise the environment of analysis by cleansing data. The data is film information between 2019 and 2020 to identify the trends and predict. This scrapped data is secondary data so it is already collected and recorded. The data is necessary to clean and identify before analysis and creation of the model. It can help to improve the accurate model and results. Additionally, cleansing data removes and decreases anomaly and skew.

By using gsub function, the unnecessary parts are removed. Before over the next step, the consistency of data types is assessed so one of the columns as the label is necessary to be changed string to numeric in order to analyse to be useful and make model. The results of change are 2019 year’s rows represent by zero and 2020 year represent by one. Therefore, the 34771 observations in 3 variables change the 34585 observations in 2 variables by cleaning the data set.

After removing outliers and cleaning data set, the values in the data set can be able to standardise or normalize because they help to easily understand quantity data. The standardisation identifies how far apart the data value is from the mean by rescaling data to have a mean of 0 and a standard deviation of 1. This data can be categorical values so the data can be processed by standardising.

The next phase is model planning. This phase can be related to data analysis and dimensionality reduction. Firstly, for the below Figure 1, the table is indicated the factors of descriptive statistics. The factors are the number of valid observations (nbr.val), the number of null (nbr.null), the number of not available (nbr.na), the minimum value (min), the maximum value (max), the range, the sum, the median, the standard error of the mean (SE.mean), the 95% confidence interval of the mean (CI.mean.0.95), the variance(var), the standard deviation (std.dev) and the coefficient of variation (coef.var). These factors are important to understand each variable and to compare between variables. This table is able to identify the overall tendency. The values are rounded to the three-decimal place.

However, this model can be to differentiate by year so the frequency table can be created by year to identify how to be the difference in each genre. For the measure of tendency, the frequency of 2019 year can have higher than the 2020 year. Moreover, drama, comedy, action can be higher frequency regardless of year.

(Figure 1: descriptive statistics table 1)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | year | Genre | PC1 | PC2 |
| nbr.val | 34585 | 34585 | 34585 | 34585 |
| nbr.null | 0 | 0 | 0 | 0 |
| nbr.na | 0 | 0 | 0 | 0 |
| min | 0 | 1 | -2.771 | -1.506 |
| max | 1 | 26 | 1.790 | 3.055 |
| Range | 1 | 25 | 4.561 | 4.561 |
| Sum | 13888 | 291423 | 0.692 | -0.130 |
| median | 0 | 8 | -0.152 | -0.063 |
| Mean | 0.402 | 8.426 | 0.000 | 0.000 |
| SE.mean | 0.003 | 0.030 | 0.005 | 0.005 |
| CI.mean.0.95 | 0.005 | 0.060 | 0.011 | 0.010 |
| Var | 0.240 | 32.136 | 1.022 | 0.978 |
| std.dev | 0.490 | 5.669 | 1.011 | 0.989 |
| coef.var | 1.221 | 0.673 | 50560.704 | -264096.532 |

(Figure 2: frequency table 1)

|  |  |  |  |
| --- | --- | --- | --- |
|  | 2019 | 2020 | Percentage (2019/2020 \*100) |
| Action | 2021 | 1441 | 140% |
| Adventure | 456 | 234 | 195% |
| Animation | 682 | 427 | 160% |
| Biography | 322 | 245 | 131% |
| Comedy | 3749 | 2551 | 147% |
| Crime | 708 | 522 | 136% |
| Documentary | 14 | 7 | 200% |
| Drama | 7461 | 5224 | 143% |
| Family | 420 | 291 | 144% |
| Fantasy | 298 | 233 | 128% |
| Game-Show | 4 | 2 | 200% |
| History | 123 | 63 | 195% |
| Horror | 1619 | 905 | 179% |
| Music | 256 | 128 | 200% |
| Musical | 118 | 59 | 200% |
| Mystery | 208 | 104 | 200% |
| News | 12 | 6 | 200% |
| Reality-TV | 148 | 74 | 200% |
| Romance | 602 | 544 | 111% |
| Sci-Fi | 325 | 248 | 131% |
| Short | 24 | 12 | 200% |
| Sport | 252 | 126 | 200% |
| Talk-Show | 42 | 21 | 200% |
| Thriller | 838 | 425 | 197% |
| War | 68 | 34 | 200% |
| Western | 50 | 25 | 200% |
| Overall | 20820 | 13951 | 149% |

By using Principal Components Analysis (PCA), the dimension is reduced and the number of variables is decreased to analyse so it helps to improve accuracy and to apply in the candidate models to the data for classifying, clustering or finding relationships. The PCA is important when the data set has numerous variables so this data set needs to be applied.

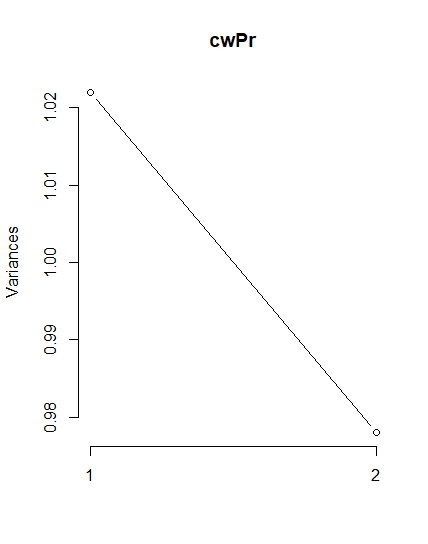
As seen from the below Figure 3, this table is the results of attributes of PCA using R. These numbers each principal component represent by standard deviation, the proportion of variance and cumulative proportion. That proportion is important to reflect how much make important variable up between 0 to 1. When the proportion can be as most closely as 1 and the highest value of variance, it is the first principal component. The next order of components is the descend dependent on the value of variance. This order can be easy to analyse to reduce dimension but this data have 2 variables so 2 variables are maintained. For the below Figure 4, the cumulative proportion is depicted at a glance easily so it is useful to choose components to analyse. For this data set, PC1 to PC2 is chosen.

(Figure 3: PCA table 1)

|  |  |  |
| --- | --- | --- |
|  | PC1 | PC2 |
| Standard deviation | 1.011 | 0.989 |
| Proportion of Variance | 0.511 | 0.489 |
| Cumulative Proportion | 0.511 | 1.000 |

(Provided by R Studio)

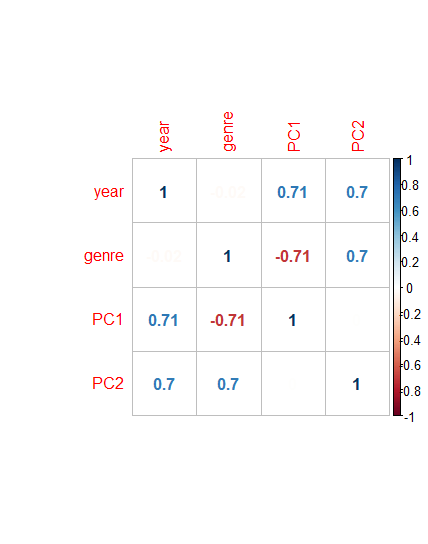
(Figure 4: PCA table 1)



(Provided by R Studio)

After PCA, as seen from the below Figure 5, this plot depicts by the number how much close as the relationship between variables. When this number is a strong relationship in two variables between vertical and horizontal value each other, the number is as close 1 or -1 as possible so if the vertical and horizontal value is same, this correlation’s value is 1. PC1 and PC2 can represent the principal component, so these can have strong relationship with 2 variables because of more than 0.7. In addition, the year and genre cannot have a significant relationship. Therefore, when the model can be put from the contributes, the principal components need to put into models.

(Figure 5: correlation plot 1)



(Provided by R Studio)

The first model, linear model, is made by using R. The resulting of the previous phase on analyzing PCA and correlation is referred and contributes to this model by choosing component and variables as the most related with gender variable. The linear model can be made by the variables and the principal components as the highest and next higher correlation obtained an analysis of PCA. The created linear model has calculated the quantity values so when the value is close to zero or closed to zero, it can be determined 2019 years and when the value is close to one or closed to one, it can be decided 2020 years. Then, the test of linear model can be evaluated by using the confusion matrix. This model is created manually so after the factors of the linear model such as slope and intercept and the number of testing data is identified, the results of slope and intercept in each model put in the linear equation by using the function methods.

As seen from the table, this explains the linear model. The coefficients can be shown.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std.Error | t value | p-value |
| (Intercept) | 34585 | 34585 | 34585 | 34585 |
| genre | 0 | 0 | 0 | 0 |
| PC1 | 0 | 0 | 0 | 0 |
| PC2 | 0 | 1 | -2.771 | -1.506 |

The ANNs model is tested, performance resilience through cross validation is necessary to examine. The second model is made by using an artificial neural network in R studio. An Artificial Neural Networks (ANN) is a computational model as perceptron that comprises a combination of simple processing units that mutually exchange information by transmitting the signal to each other over a large number of weighted.

1. Performance Evaluation

For the below tables, these are confusion matrix. The 2019 year is zero and the 2020 year is one. Furthermore, the vertical can be actual and the horizontal is prediction. It is useful to evaluate the model by identifying accuracy, precision, recall and false positive rate. The linear models’ evaluations have the four types for the below confusion matrix. The linear model is made by using variables and principal component because the linear model is composed one independent and one dependent variable. For below Figure 6, it is errors of prediction so its value will be able to classify type 2 error when 2019 year is true but when 2020 year is true, its errors can be classified type 1 error. when the 2019 year is positive and 2020 year is negative, the true positive is 5393, the true negative is 3739, the false positive is 509 and the false negative is 735. The accuracy is 0.8801 as rounded the fourth decimal place so it means accuracy is same as representing 88%. The precision is 0.9138, 91%. The recall is 0.8800 and the false positive rate is 0.1198. Otherwise, when 2020 year is positive, the precision is 0.8357. The recall is 0.8800 and the false positive rate is 0.1199.

(Figure 19: confusion matrix table 1)

|  |  |  |  |
| --- | --- | --- | --- |
| Confusion matrix | | Actual | |
| 0 | 1 |
| Prediction | 0 | 5393 | 509 |
| 1 | 735 | 3739 |

(Provided by R Studio)

There are four types of results in this table when classification predictions are performed by using ANNs. The linear models’ evaluations have the four types for the below confusion matrix. The linear model is made by using variables and principal component because the linear model is composed one independent and one dependent variable. For below Figure 6, it is errors of prediction so its value will be able to classify type 2 error when 2019 year is true but when 2020 year is true, its errors can be classified type 1 error. when the 2019 year is positive and 2020 year is negative, the true positive is 5393, the true negative is 3836, the false positive is 412 and the false negative is 582. The accuracy is 0.8990 as rounded the fourth decimal place so it means accuracy is same as representing 90%. The precision is 0.9308, 93%. The recall is 0.9050 and the false positive rate is 0.0969. Otherwise, when 2020 year is positive, the precision is 0.8683. The recall is 0.9030 and the false positive rate is 0.0950.

(Figure 19: confusion matrix table 1)

|  |  |  |  |
| --- | --- | --- | --- |
| Confusion matrix | | Actual | |
| 0 | 1 |
| Prediction | 0 | 5546 | 412 |
| 1 | 582 | 3836 |

(Provided by R Studio)

1. Results

The trend can be identified about the film. Moreover, the prediction model can be created as discrimination of the years such as 2019 and 2020. Because of Covid-19, the film market can be contracted so the data set reflected this circumstance. The

1. Reference

Cady, F. (2017). *The Data Science Handbook.* Wiley

R (Accessed: 18 Nov 2019)

R documentation[Online] Available at:  [https://www.rdocumentation.org//](%20https://www.rdocumentation.org//) (Accessed: 18 Nov 2019)

Schutt, R and O’Neil, C (2013). *Doing Data Science*. O’Reilly

Stack Overflow[Online] Available at: <https://stackoverflow.com/> (Accessed: 18 Nov 2019)