JAIPURIA INSTITUTE OF MANAGEMENT, NOIDA

PGDM / PGDM (M) / PGDM (SM)

IV TRIMESTER (Batch 2021-23)

END TERM EXAMINATION, NOVEMBER 2022

SET - 1

Course Name	Programming for Business Analytics (PBA)	Course Code	**20823**
Max. Time	**2 hours**	Max. Marks	**40**

INSTRUCTIONS:

a. Attempt all the questions on a single Jupyter Notebook
b. The data for the case is available on Moodle.
c. Write down your Roll no., course name and course code on top of Jupyter Notebook
d. Save your Jupyter notebook with .ipynb extension and as pdf file
e. Upload both the files on Moodle.
f. Label the files as PBA_roll no (for example: PBA_23)
g. This is an open book exam. Students may refer to the codes.

Read the case below and answer the questions given by analyzing the data using Python.

Case: Predicting Success of Bollywood Movies

The Indian film industry produces the maximum number of movies per year, higher than any other country's movie industry. However, very few movies taste commercial success. With 3.3 billion tickets sold annually, India also has the highest number of theater admissions. With so much at stake and highly uncertain nature of returns, it is of commercial interest to develop a model which can predict the success of a movie. Indian Hindi Movie industry popularly known as Bollywood has reached staggering proportions in terms of volume of business, employment, movies produced (more than 100 in a year) and its reach (more than 100 countries worldwide). Movies have been described as experience goods with very less shelf life, therefore, it is difficult to forecast the demand for a movie

There are number of parameters that may influence success of a movie like – time of its release, marketing, lead actors, director, producer, writer, music director – being some of the factors. Data scientist may develop the models and mechanisms to predict reliably the ranking and / or box office collections of a movie & can help de-risk the business significantly and increase average returns. Various stakeholders such as actors, financiers, directors etc. can use these predictions to make more informed decisions.

In a BBC article, film director Karan Johar is quoted as saying "only 45 of the 300 million (India's middle class) is reached by the movie industry. A vested effort in using data analytics to improve

India's film industry could be the way to reach the remaining 255 million. By making the filmmaking process more efficient, more profitable, more relatable and less risky, a great return on huge investments is guaranteed with data analytics."
The data of 140 Bollywood movies is given in Excel file labelled "PBA_Set1_2022". Karan Johar has hired you as the analytics consultant. He asked you to "identify" and "quantify" the factors responsible for estimating the "Box office collection" in a multivariate fashion. Let's help Mr. Johar in carrying out the analysis.

The data description is given below:

Variable(s)	Description
MovieName	Name of the Movie
ReleaseDate	Date of Release of Movie
Release_festive	Whether the movie is released during festive or holiday season 1 if released during festive season 0, otherwise
LeadActor	Lead actor in the movie
ActorNoofMovies	No. of movies done by the lead actor
StatusStar	Status of lead actor/ actress; debut, star or superstar
Sequel	Whether the movie is sequel 1 if yes 0, otherwise
ProductionHouse	Movie Production house
Genre	1. Action/adventure 2. Family/children 3. Comedy 4. Drama 5. Horror 6. Mystry/Suspense 7. Sci-fi/ fantasy
BudgetCrores	Budget in crore
CriticsRating	Critics' rating
BollyRating	Bollywood Hungama movie rating
IMDbRating	IMDB movie rating
TwitterRating	Twitter rating
@1stWeekBoxOfficeCollection	First week box office collection
TotalBoxOffice	Total Box-office collection

Analyze the data and write the answers of the following questions:

1. Mr. Johar wants to understand, what are the key factors that influence the first week and total box office collection? Can you develop a model using historical data to estimate the first week and total box office collection? Which data science technique will you use? Why? **(4 marks)**
2. Create a descriptive statistics report for all numeric as well as categorical data and interpret the results. **(5 marks)**

3. Is there a difference in the average total box-office collection of movies based on different genre? **(3 marks)**
4. Check whether there exist multi-collinearity amongst the features. How multi-collinearity can be handled? **(3 marks)**
5. Check for outliers in the total box office-collection. What treatment will you give in case there exist an outlier? **(3 marks)**
6. What else you suggest can be done as part of data pre-processing and exploration? **(2 marks)**
7. Build two multiple regression models. First with “first week box-office collection” as dependent variable and the second model with “total box-office collection”. Evaluate the results and answer the following questions: **(6 marks)**

    a. List important features for both the models separately. **(2 marks)**
    b. Comment upon the model fit of both the models. **(3 marks)**
    c. Validate the model fit with the test data set in both the cases. **(4 marks)**

8. Prepare a report by summarizing the usefulness of the study. **(5 marks)**

Note: Interpretation of all the outputs should be written by putting comments on the Jupyter notebook.

*****