National Testing Agency

Question Paper Name :	Predictive Analytics 30 Sep 2020 Shift 2
Subject Name :	Predictive Analytics
Creation Date :	2020-09-30 18:37:40
Duration :	180
Number of Questions :	40
Total Marks :	100
Display Marks:	Yes

Predictive Analytics

Group Number :	1
Group Id :	899514200
Group Maximum Duration :	0
Group Minimum Duration :	120
Show Attended Group? :	No
Edit Attended Group? :	No
Break time :	0
Group Marks :	100
Is this Group for Examiner? :	No

Predictive Analytics

Section Id :	899514280
Section Number :	1
Section type :	Online
Mandatory or Optional :	Mandatory
Number of Questions :	40
Number of Questions to be attempted :	40

Section Marks :	100
Mark As Answered Required? :	Yes
Sub-Section Number :	1
Sub-Section Id :	899514325
Question Shuffling Allowed :	Yes

Question Number : 1 Question Id : 89951416959 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

The best simple linear regression model is the one for which ____
1.The R-square (coefficient) is the highest
2.The residuals follow normal distribution.
3.The p-value corresponding to t-test is less than the significance value α.
4.The p-value corresponding to t-test is less than the significance value α and the residuals follow normal distribution and the residual are homoscedastic.

Options :

89951466238. 1

89951466239. 2

89951466240. 3

89951466241. 4

Question Number : 2 Question Id : 89951416960 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

A high street jewellery shop uses a regression model $Y = -10.5 + 95 \times$ carat to predict the price of a diamond as a function of carat, where carat is the weight of the diamond. The value of $\beta 0$ is negative because:
1.Regression model is incorrect since the value of diamond cannot take negative value.
2.The regression models cannot be extrapolated beyond the range of the data used for building the model.
3.The regression model is valid only for carat values greater than 0.1106 since the value of Y will be positive when carat is greater than 0.1106.
4.The value of $\beta 0$ (= −10.5) should be ignored while calculating the price of the diamond.

Options :

89951466242. 1

89951466243. 2

89951466244. 3

89951466245. 4

Question Number : 3 Question Id : 89951416961 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

If the correlation between a predictor variable and the outcome variable is 0.8, the proportion of variation in the outcome variable explained by the predictor variable is:
1.0.9
2.0.72
3.0.89
4.0.64

Options :

89951466246. 1

89951466247. 2

89951466248. 3

89951466249. 4

Question Number : 4 Question Id : 89951416962 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

In a model $\ln(Y) = \beta 0 + \beta 1\ X$, the value of $\beta 1$ is
1.Change in value of Y for unit change in value of X
2.Change in value of X for unit change in value of Y.
3.Percentage change in value of X for unit change in value of Y
4.Percentage change in value of Y for unit change in value of X

Options :

89951466250. 1

89951466251. 2

89951466252. 3

89951466253. 4

Question Number : 5 Question Id : 89951416963 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

In multiple regression models, multi-collinearity may result in
1.Removing a statistically significant explanatory variable from the model.
2.The regression coefficient may have opposite sign
3.Adding a new variable to the model may cause huge change to the regression coefficient.
4.All of above

Options :

89951466254. 1

89951466255. 2

89951466256. 3

89951466257. 4

Question Number : 6 Question Id : 89951416964 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

When a new variable is added to the regression model, the R-square value increases by
1.Square of the semi-partial coefficient between added variable and the response variable
2.Correlation coefficient between the added variable and the response variable
3.Partial correlation coefficient between added variable and the response variable
4.Semi-partial coefficient between added variable and the response variable.

Options :

89951466258. 1

89951466259. 2

89951466260. 3

89951466261. 4

Question Number : 7 Question Id : 89951416965 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

If there is an auto-correlation between the successive errors in a time series regression ther
1.A statistically insignificant variable may be added to the model.
2.A statistically significant variable may be removed from the model.
3.The standard error of estimate of the regression parameter is underestimated.
4. The Durbin–Watson test statistic value will be close to 2.

Options :

89951466262. 1

89951466263. 2

89951466264. 3
89951466265. 4

Question Number : 8 Question Id : 89951416966 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

When a stepwise regression model is developed, the first variable that is added is
1. The variable with highest variance.
2. The variable that has the least variance.
3. The variable that has highest correlation with the dependent variable.
4. The variable with least covariance with the dependent variable.

Options :

89951466266. 1
89951466267. 2
89951466268. 3
89951466269. 4

Question Number : 9 Question Id : 89951416967 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

Variance inflation factor is
1.Factor by which the regression coefficient is increased.
2. Factor by which the t-statistic value is inflated.
3. The t-statistic is deflated by a factor of Sqrt $\sqrt{VIF}$
4. The t-statistic value is inflated by a factor of $\sqrt{VIF}$.

Options :

89951466270. 1
89951466271. 2
89951466272. 3
89951466273. 4

Question Number : 10 Question Id : 89951416968 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

The area under the ROC curve (AUC) represents
1.The maximum accuracy of the logistic regression model
2.Ratio of sensitivity to the specificity.
3.Difference between sensitivity and specificity
4.Proportion of concordant pairs in the dataset.

Options :

89951466274. 1

89951466275. 2

89951466276. 3

89951466277. 4

Question Number : 11 Question Id : 89951416969 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

Refer Linear Regression Model Output to answer this Question.

Which of the following statements about model 1 is incorrect?
1.The model explains 42.25% of variation in box office collection.
2.There are outliers in the model.
3.The residuals follow a normal distribution.
4.Box office collection increases as the budget increases.

Options :

89951466278. 1

89951466279. 2

89951466280. 3

89951466281. 4

Question Number : 12 Question Id : 89951416970 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

Refer Linear Regression Model Output to answer this Question.

The value of the constant in table 3 is negative (-8.354). So, we can conclude that _____
1.The model is incorrect, since the box office collection cannot be negative.
2.The value of the constant is negative due to heteroscedastcity.
3.The revenue is negative because the production house may have paid theaters money to release the movie.
4.Regression model cannot be extrapolated, so the value of constant should be incorporated only within the range of budget that was used for developing the model.

Options :

89951466282. 1

89951466283. 2

89951466284. 3

89951466285. 4

Question Number : 13 Question Id : 89951416971 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

Refer Linear Regression Model Output to answer this Question.

What is the average difference in the box office collection when a movie is released during a holiday season (Releasing_Time_holiday_season) versus movies released during normal season (Releasing_Time_Normal_Season)? Use model 2 to answer this question and a significance value of 5%.
1.1.15 Crores
2.16.97 Crores
3.No difference
4.2.32 Crores

Options :

89951466286. 1

89951466287. 2

89951466288. 3

89951466289. 4

Question Number : 14 Question Id : 89951416972 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

Refer Linear Regression Model Output to answer this Question.

What is the variation in response variable, ln(Box office collection), explained by the model after adding all 6 variables?
1.0.6669
2.0.8202
3.0.706
4.0.4242

Options :

89951466290. 1

89951466291. 2

89951466292. 3

89951466293. 4

Question Number : 15 Question Id : 89951416973 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

Refer Linear Regression Model Output to answer this Question.

Which factor has the least impact on the box office collection of a movie?
1.Music_Dir_Cat C
2.Director_Cat C
3.YouTube Views
4.Genre Comedy

Options :

89951466294. 1

89951466295. 2

89951466296. 3

89951466297. 4

Question Number : 16 Question Id : 89951416974 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

Refer Linear Regression Model Output to answer this Question.

Among the variables in Table 6, which variable is not useful for practical application of the model?
1.Budget_35_Crore
2.YouTube Views
3.Genre Comedy
4.Music Director Cat C

Options :

89951466298. 1

89951466299. 2

89951466300. 3

89951466301. 4

Question Number : 17 Question Id : 89951416975 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

Refer Linear Regression Model Output to answer this Question.

A regression model is developed for salary of employees of a company using gender (G), work experience (WE) and the interaction variable G x WE.G = 1 is coded as female and G = 0 is male. The corresponding regression equation is shown below (assume that all predictors are significant):

$Y = 45490.50 + 3000.900 \times G + 1497.89\ WE - 990.75\ G \times WE$

Which of the following statements are true?
1.Average salary of female employees is higher than male employees.
2.Female employees earn 3000.90 more than male employees.
3.Increase in salary with work experience for male employees is higher than female employees.
4.In the long run, male employees earn more than female employees.

Options :

89951466302. 1

89951466303. 2

89951466304. 3

89951466305. 4

Question Number : 18 Question Id : 89951416976 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No

Correct Marks : 2.5 Wrong Marks : 0

The Independent variable that has the highest impact on the dependent variable is given by ____
1. The variable with largest coefficient value.
2. The variable with largest absolute coefficient value.
3. The variable with largest standardized coefficient value.
4. The variable with largest absolute standardized coefficient value.

Options :

89951466306. 1

89951466307. 2

89951466308. 3

89951466309. 4

Question Number : 19 Question Id : 89951416977 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No
Correct Marks : 2.5 Wrong Marks : 0

Refer Logistic Regression Model Output to answer this Question.

Which of the following statements is incorrect about the model in tables 1 and 2?
1. Sensitivity is higher than specificity at a classification cut-off probability of 0.7
2. Sensitivity is higher than specificity
3. Probability of accepting the offer is more for those candidates who are willing to work in shifts
4. Probability of acceptance of job offer is more for female candidates than male candidates.

Options :

89951466310. 1

89951466311. 2

89951466312. 3

89951466313. 4

Question Number : 20 Question Id : 89951416978 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No
Correct Marks : 2.5 Wrong Marks : 0

Refer Logistic Regression Model Output to answer this Question.

The Probability of a female candidate willing to work in shifts is _____
1. 0.6357
2. 0.5319
3. 0.5914
4. 0.7413

Options :

89951466314. 1

89951466315. 2

89951466316. 3

89951466317. 4

Question Number : 21 Question Id : 89951416979 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

Refer Logistic Regression Model Output to answer this Question.

Which interview type increases the probability of accepting job offer?
1. Online test
2. Face to Face Interview
3. Telephone Interview
4. None of them

Options :

89951466318. 1

89951466319. 2

89951466320. 3

89951466321. 4

Question Number : 22 Question Id : 89951416980 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

Refer Logistic Regression Model Output to answer this Question.

Among the following 4 customers who is most likely to join?

Note: PerHike value is 10 for 10%.

1. 10 years of work experience, 10% hike in CTC, face to face interview and not willing to work in Shift.
2. 2 years of work experience, 20% hike in CTC, online test and willing to work in Shift.
3. 5 years of work experience, 15% hike in CTC, Telephone interview and not willing to work in Shift.
4. 6 years of work experience, 0% hike in CTC, face to face interview and willing to work in Shift.

Options :

89951466322. 1

89951466323. 2

89951466324. 3

89951466325. 4

Question Number : 23 Question Id : 89951416981 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

In a logistic regression, the regression coefficient corresponding to a predictor variable is interpreted as _____

1. Change in P(Y = 1) for unit change in the predictor variable value.
2. Change in odds for unit change in the predictor variable value.
3. Change in odds ratio for unit change in the predictor variable value.
4. Change in ln-odds ratio for unit change in the predictor variable value

Options :

89951466326. 1

89951466327. 2

89951466328. 3

89951466329. 4

Question Number : 24 Question Id : 89951416982 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

If in a data set with 250 positives, an LR model classifies 200 positives correctly then the specificity is _____

1. 0.8
2. 0.2
3. 1.25
4. Can't say

Options :

89951466330. 1

89951466331. 2

89951466332. 3

89951466333. 4

Question Number : 25 Question Id : 89951416983 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

Deviance in a logistic regression model should be _____

1. Maximum
2. Minimum
3. Less than the chi-square critical value with df equal to number of variables added.
4. Greater than the chi-square critical value with df equal to number of variables added.

Options :

89951466334. 1

89951466335. 2

89951466336. 3

89951466337. 4

Question Number : 26 Question Id : 89951416984 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

Refer Decision Tree Model Output to answer this Question.

Which of the following statement is correct?

1. Housing loan and subscription to term deposit are statistically independent
2. Housing loan and subscription to term deposit are statistically dependent
3. In a CHAID tree, the variable "housing loan" will not be selected for splitting the node.
4. The probability of subscription to term deposit increases when the customer has taken a housing loan.

Options :

89951466338. 1
89951466339. 2
89951466340. 3
89951466341. 4

Question Number : 27 Question Id : 89951416985 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

Refer Decision Tree Model Output to answer this Question.

Using CHAID tree we can conclude that _____

1. When there is no previous customer contact and the customer has taken a housing loan then there is a 93% chance that the customer will subscribe to the term deposit.
2. When there is no previous customer contact and the customer has taken a housing loan then there is a 93% chance that the customer will not subscribe to the term deposit.
3. Variables other than "previous" and "housing loan" are not statistically significant.
4. The customers who have been contacted previously are less likely to respond to the marketing campaign compared to those who were not contacted before.

Options :

89951466342. 1
89951466343. 2
89951466344. 3
89951466345. 4

Question Number : 28 Question Id : 89951416986 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

In a Classification and Regression Tree (CART), the splitting of node is based on _____

1. Gini impurity index or entropy.
2. Impurity measures such as Gini index or entropy for classification tree and Sum of Squared Errors (SSE) for regression tree.
3. Sum of Squared Errors (SSE).
4. F-test

Options :

89951466346. 1
89951466347. 2

89951466348. 3
89951466349. 4

Question Number : 29 Question Id : 89951416987 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

For a Classification problem with two classes _____
1. Gini index value is greater than or equal to entropy
2. Gini index value is less than or equal to entropy
3. Gini index value is greater than or equal to entropy when proportion of classes are equal
4. Can't say

Options :

89951466350. 1
89951466351. 2
89951466352. 3
89951466353. 4

Question Number : 30 Question Id : 89951416988 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

A Bonferroni correction is used in a CHAID tree development to _____
1. ensure that the tree size is optimized and Type-II error is minimized
2. Select only statistically significant variables for splitting a node.
3. adjust the significance value to maintain the overall statistical significance/Type-I error of the model at desired level.
4. test the statistical significance using the Chi-square test of independence

Options :

89951466354. 1
89951466355. 2
89951466356. 3
89951466357. 4

Question Number : 31 Question Id : 89951416989 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

For a classification problem with two classes, the proportion of positives at a node is 20%. The value of the Gini index at this node is _____

1. 0.16
2. 0.13
3. 0.26
4. 0.32

Options :

89951466358. 1

89951466359. 2

89951466360. 3

89951466361. 4

Question Number : 32 Question Id : 89951416990 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

Refer Unstructured Data Analysis model output to answer this Question.

The probability of finding word 1 given the document is positive is _____

1. 4/8
2. 6/16
3. 6/8
4. 4/16

Options :

89951466362. 1

89951466363. 2

89951466364. 3

89951466365. 4

Question Number : 33 Question Id : 89951416991 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

Refer Unstructured Data Analysis model output to answer this Question.

A Bernoulli document model is used to convert a comment about a product into a binary vector and is given by: [0, 1, 0, 1, 0, 0, 0, 1]. Which of the following statements is incorrect?
1. The vocabulary set has 8 words
2. Three words from the vocabulary set are present in the comment.
3. The comment is a positive comment.
4. The vocabulary set has more than 5 words.

Options :

89951466366. 1

89951466367. 2

89951466368. 3

89951466369. 4

Question Number : 34 Question Id : 89951416992 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

The maximum value of the entropy at node K of a classification tree with J classes is _____
1. 0
2. 0.5
3. 1
4. can't say

Options :

89951466370. 1

89951466371. 2

89951466372. 3

89951466373. 4

Question Number : 35 Question Id : 89951416993 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

Decision trees such as CHAID and CART can be used only when the dependent variable is _____
1. Discrete
2. Continuous
3. Interval
4. All three (Discrete, Continuous, Interval)

Options :

89951466374. 1
89951466375. 2
89951466376. 3
89951466377. 4

Question Number : 36 Question Id : 89951416994 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

Refer Forecasting Model Output to answer this Question.

Which season has the highest fluctuation from the trend line?
1. Q1
2. Q2
3. Q3
4. Q4

Options :

89951466378. 1
89951466379. 2
89951466380. 3
89951466381. 4

Question Number : 37 Question Id : 89951416995 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

Refer Forecasting Model Output to answer this Question.

The forecasted demand for Q1 of 2016 using AR(1) model is _____
1. 98.39
2. 115.32
3. 149.04
4. 174.68

Options :

89951466382. 1
89951466383. 2
89951466384. 3
89951466385. 4

Question Number : 38 Question Id : 89951416996 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

In a simple exponential smoothing method, the low value of smoothing constant α is chosen when _____

1. The data has high fluctuations around the trend line
2. There is seasonality in the data
3. The data is smooth with low fluctuations.
4. There are variations in the data due to cyclical component.

Options :

89951466386. 1

89951466387. 2

89951466388. 3

89951466389. 4

Question Number : 39 Question Id : 89951416997 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

Seasonality in time-series data is caused due to _____

1. Changes in macro-economic factors such as recession, unemployment, and so on
2. Festivals and customs in a society
3. Random events that occur over a period of time
4. Changes in customer behaviour driven by new products and promotions

Options :

89951466390. 1

89951466391. 2

89951466392. 3

89951466393. 4

Question Number : 40 Question Id : 89951416998 Question Type : MCQ Option Shuffling : No Is Question Mandatory : No Correct Marks : 2.5 Wrong Marks : 0

A necessary condition for accepting a time series forecasting models is _____

1. The residuals should follow a normal distribution
2. The residuals should be white noise
3. The residuals should be black noise
4. The residuals should follow a normal distribution and the R-square should be high

Options :

89951466394. 1

89951466395. 2

89951466396. 3

89951466397. 4