Ace the upcoming Data Science Interview

You can't anticipate every question an interviewer will ask. However, there are many **critical questions** that you can prepare before the interview.

Our hiring partners have helped us curate a set of interview questions on key skills, which will help you prepare better for the data science job roles.

Filters

1. How do you compare categorical values, how would you know that a categorical value is related to target variable?

Basic Advanced Stats

Comparing categorical Values: When there are three or more levels/categories for the predictor & Target variable is nominal, the degree of association between the predictor and target variable can be measured with statistics such as chi-squared tests

Categorical value is related to the target variable:

- When there is only one continuous target variable, there are one plus categorical independent variables, and there is no control variable at all, then you can go for ANOVA variables, and there is no control variable at all, then you can go for ANOVA.

- Similarly, when there is only one continuous target variable, there is only one categorical independent variable (i.e. dichotomous, e.g. pass/fail), and no control variable, then go for t-Test

2. What is Linear regression? Explain the assumptions.

Basic Advanced Stats

Linear regression is an analysis that assesses whether one or more predictor variables explain the dependent (criterion) variable. The regression has five key assumptions:

1) Linear relationship: Linear regression needs the relationship between the independent and dependent variables to be linear. The linearity assumption can best be tested with scatter plots.

2) Normality: The error terms must be normally distributed (To check normality, one can look at QQ plot, can also perform statistical tests of normality such as Kolmogorov-Smirnov test, Shapiro-Wilk test.

3) Multicollinearity: Linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other.

Multicollinearity may be tested with three central criteria: Correlation matrix, Tolerance, VIF

4) No auto-correlation: Linear regression analysis requires that there is little or no autocorrelation in the data. Autocorrelation occurs when the residuals are not independent of each other. For instance, this typically occurs in stock prices, where the price is not independent of the previous price.

5) Homoscedasticity: The error terms must have constant variance. This phenomenon is known as homoskedasticity. The presence of non-constant variance is referred to as heteroskedasticity.

3. Explain mathematically how Linear Regression works?

Basic Advanced Stats

The idea behind simple linear regression is to "fit" the observations of two variables into a linear relationship between them. Graphically, the task is to draw the line that is "best-fitting" or "closest" to the points (x_i,y_i), where x_i and y_i are observations of the two variables which are expected to depend linearly on each other.

Although many measures of best fit are possible, for most applications the best-fitting line is found using the method of least squares. The method finds the linear function L which minimizes the sum of the squares of the errors in the approximations of the y_i by L(x_i)

For eg: To find the line y=mx+b of best fit through N points, the goal is to minimize the sum of the squares of the differences between the y-coordinates and the predicted yy-coordinates based on the line and the x-coordinates.

4. In your project, why classification was chosen over regression ?

Basic Advanced Stats

Classification is used when the output variable is a category such as “red” or “blue”, “spam” or “not spam”. It is used to draw a conclusion from observed values. Differently from regression which is used when the output variable is a real or continuous value like “age”, “salary”, etc.

When we must identify the class the data belongs to, we use classification over regression. Like when you must identify whether a name is male or female instead of finding out how they are correlated with the person.

5. Explain the working of logistic regression?

Basic Advanced Stats

Hint?

Highlight the pros and cons. Also the scenarios for which the algorithm is applicaple. Also brush-up all algorithm equations and assumptions here https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

6. Evaluation metrics of regression/classification model?

Basic Advanced Stats

Hint?

Regression: R2, adjusted R2, p-value, RMSE, MAD, refresh OLS model Classification: Confusion matrix, accuracy score, precision, recall, F1 Score, ROC, AUC curves USL: Kmean, hirarchical, dendogram, Dimensionality reduction - PCS

7. Build a credit card fraud detection model

Advanced Advanced Stats

8. Evaluation Metrics (Difference between R-Square and Adjusted R-Square)

Basic Advanced Stats

R-squared (coefficient of determination) measures the proportion of the variation in your dependent variable (Y) explained by your independent variables (X) for a linear regression model.

R^2 = Explained variation / Total Variation

Adjusted R-squared adjusts the statistic based on the number of independent variables in the model.

It is possible that R Square has improved significantly yet Adjusted R Square is decreased with the addition of a new predictor when the newly added variable brings in more complexity than the power to predict the target variables.

Adj. R^2 = 1 - ((1 - R.squared) * (n - 1)/(n-p-1)) where p: no. of predictors, n: no. of observations

9. Difference between logistic regression and CART?

Basic Advanced Stats

1. Cart works best locally, Logistic regression works best Globally
2. Cart is Useful for identifying interactions between variables
3. Cart can predict both categorical and quantitative data while logistic can only predict categorical/ordinal
4. Cart is Easy to run & interpret
5. Cart can lead to overfitting as it has a disadvantage over stop splitting
6. CART works best with a larger dataset, while Logistic regression on a smaller dataset
7. Cart is non-parametric while logistic is parametric

10. What are the limitations of Logistic Regression

Basic Advanced Stats

1. The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables.

2. It can only be used to predict discrete functions. Hence, the dependent variable of Logistic Regression is bound to the discrete number set.

3. Non-linear problems can't be solved with logistic regression because it has a linear decision surface. Linearly separable data is rarely found in real-world scenarios

4. Logistic Regression requires average or no multicollinearity between independent variables

5. If the number of observations is lesser than the number of features, Logistic Regression should not be used, otherwise, it may lead to overfitting.

11. Name the library used to implement logistic Regression

Basic Advanced Stats

Python:

from sklearn.linear_model import LogisticRegression

R:

glm(Target ~.,family=binomial(link='logit'),data=train)

12. What is confusion matrix?

Basic Advanced Stats

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making.

True Positive (TP): The actual value was positive and the model predicted a positive value

True Negative (TN): The actual value was negative and the model predicted a negative value

False Positive (FP) – Type 1 error: The actual value was negative but the model predicted a positive value

False Negative (FN) – Type 2 error: The actual value was positive but the model predicted a negative value

13. What is vif? What is the precision of Vif ?

Basic Advanced Stats

VIF, the Variance Inflation Factor, is used during regression analysis to assess whether certain independent variables are correlated to each other and the severity of this correlation. If your VIF number is greater than 10, the included variables are highly correlated to each other. Since the ability to make precise estimates is important to many companies, generally people aim for a VIF within the range of 1-5. A cutoff number of 5 is commonly used.

14. How do you deal with multi-colinearity and conditional probability?

Intermediate Advanced Stats

Potential solutions to deal with multicollinearity:

- Remove some of the highly correlated independent variables.

- Linearly combine the independent variables, such as adding them together.

- Perform an analysis designed for highly correlated variables, such as principal components analysis or partial least squares regression.

15. Is logistic regression a part of Linear regression?

Basic Advanced Stats

Logistic regression is considered a generalized linear model because the outcome always depends on the sum of the inputs and parameters.

The actual value of the dependent variable is yi.

The predicted value of yi is defined to be y^i = a xi + b, where y = a x + b is the regression equation.

The residual is the error that is not explained by the regression equation:

ei = yi - y^i.

A residual plot plots the residuals on the y-axis vs. the predicted values of the dependent variable on the x-axis. We would like the residuals to be unbiased: have an average value of zero in any thin vertical strip, and homoscedastic, which means "same stretch": the spread of the residuals should be the same in any thin vertical strip.

The residuals are heteroscedastic if they are not homoscedastic.

16. Write the equation of the linear Regression? Explain residuals?

Basic Advanced Stats

The actual value of the dependent variable is yi.

The predicted value of yi is defined to be y^i = a xi + b, where y = a x + b is the regression equation.

The residual is the error that is not explained by the regression equation:

ei = yi - y^i.

A residual plot plots the residuals on the y-axis vs. the predicted values of the dependent variable on the x-axis. We would like the residuals to be unbiased: have an average value of zero in any thin vertical strip, and homoscedastic, which means "same stretch": the spread of the residuals should be the same in any thin vertical strip.

The residuals are heteroscedastic if they are not homoscedastic.

17. Explain homoscedasticity ?

Intermediate Advanced Stats

The assumption of homoscedasticity (meaning “same variance”) is central to linear regression models. Homoscedasticity describes a situation in which the error term (that is, the “noise” or random disturbance in the relationship between the independent variables and the dependent variable) is the same across all values of the independent variables. Heteroscedasticity (the violation of homoscedasticity) is present when the size of the error term differs across values of an independent variable. The impact of violating the assumption of homoscedasticity is a matter of degree increasing as heteroscedasticity increases.

18. Performance measures of linear Regression?

Basic Advanced Stats

Most commonly known evaluation metrics include:

R-squared (R2), which is the proportion of variation in the outcome that is explained by the predictor variables. In multiple regression models, R2 corresponds to the squared correlation between the observed outcome values and the predicted values by the model. The Higher the R-squared, the better the model.

Root Mean Squared Error (RMSE), which measures the average error performed by the model in predicting the outcome for an observation. Mathematically, the RMSE is the square root of the mean squared error (MSE), which is the average squared difference between the observed actual outcome values and the values predicted by the model. So, MSE = mean((observeds - predicteds)^2) and RMSE = sqrt(MSE). The lower the RMSE, the better the model.

Residual Standard Error (RSE), also known as the model sigma, is a variant of the RMSE adjusted for the number of predictors in the model. The lower the RSE, the better the model. In practice, the difference between RMSE and RSE is very small, particularly for large multivariate data.

Mean Absolute Error (MAE), like the RMSE, the MAE measures the prediction error. Mathematically, it is the average absolute difference between observed and predicted outcomes, MAE = mean(abs(observeds - predicteds)). MAE is less sensitive to outliers compared to RMSE.

Additionally, there are four other important metrics - AIC, AICc, BIC and Mallows Cp

The lower these metrics, the better the model.

AIC stands for (Akaike's Information Criteria): Basic idea of AIC is to penalize the inclusion of additional variables to a model. It adds a penalty that increases

the error when including additional terms. The lower the AIC, the better the model.

AICc is a version of AIC corrected for small sample sizes.

BIC (or Bayesian information criteria) is a variant of AIC with a stronger penalty for including additional variables to the model.

Mallows Cp: A variant of AIC developed by Colin Mallows.

19. Explain prior probability, likelihood and marginal likelihood in context of naiveBayes algorithm?

Basic Advanced Stats

20. Derive logistic regression equation.

Intermediate Advanced Stats

In Logistic Regression, the Probability should be between 0 to 1 and as per cut off rate, the output comes out in the form of 0 or 1 where the linear equation does not work because value comes out inform of + or - infinity and that the reason we have to convert a linear equation into Sigmoid Equation.

Transformation of Linear Regression Equation into Logistic Regression Equation.

Transformation of Linear Regression Equation into Logistic Regression Equation:

1. Linear Regression Equation is Y = b0 + b1*X

Converting into Sigmoid Equation:

2. Probability should not be less than 0 i.e. eliminating -infinity

converting into the exponential form: E^Y

3. Probability should not be greater than 1 i.e. eliminating +infinity

Dividing value with 1:

P = E^Y/E^Y+1

Odds Ratio:

4. Taking Odds Ratio which is used for calculating Probability

P = Probability of Success and 1-P= Probability of Failure

P/1-P

Sigmoid Equation put into Odd Ratio:

5. Substituting the value of P with equation E^Y/E^Y+1

P/1-P = (E^Y/E^Y+1) / (1-E^Y/E^Y+1)

=(E^Y/E^Y+1) / (1/E^Y+1)

=(E^Y/E^Y+1) x (E^Y+1/1)

=E^Y

Odds Ratio in the form of Sigmoid:

6. We can say P/1-P = E^Y

Log Transformation:

7. Converting into Log

P/1-P = E^Y

Log(P/1-P) = Y (When it converts into a log, Exponential naturally removed)

Log(P/1-P) = b0+b1*X