Google Discussions

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 91 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 91
Topic #: 1
[All Professional Data Engineer Questions]

You work for a bank. You have a labelled dataset that contains information on already granted loan application and whether these applications have been defaulted. You have been asked to train a model to predict default rates for credit applicants.
What should you do?

A. Increase the size of the dataset by collecting additional data.

B. Train a linear regression to predict a credit default risk score.

C. Remove the bias from the data and collect applications that have been declined loans.

D. Match loan applicants with their social profiles to enable feature engineering.

Show Suggested Answer

by [deleted] at *March 22, 2020, 4:37 p.m.*

Comments

Type your comment...

Submit

GHN74 Highly Voted 4 years, 8 months ago
A is incorrect as you need to work with the data you have available
C is an optimisation not a solution

C is an optimisation not a solution
D is ethically incorrect and invasion to privacy, there could be several legal implications with this
B although oversimplified but is a workable solution

upvoted 40 times

sergio6 3 years, 7 months ago

Information in social profiles are public

upvoted 1 times

sergio6 3 years, 7 months ago

according to the privacy settings and shareable informations

upvoted 1 times

sumanshu Highly Voted 4 years ago

We have labelled data that contains whether a loan application is accepted or defaulted - So Classification Problem Data.

We need to predict (Default Rates for applicants) - I think whether application will be granted or defaulted. - So Binary Classification.

No option matches the answer. - if we mark 'B' - It should be Logistic Regression, Instead of Linear Regression

upvoted 21 times

szefco 3 years, 4 months ago

Question says: "to predict default RATES for credit applicants".
It is not binary classification, so Linear Regression would work here.
I think B is correct answer.

upvoted 23 times

cchen8181 1 year, 11 months ago

Correct approach is to use logistic regression to predict default/not default, and then take the confidence/probability of the outcome as the "default rate". Linear regression doesn't make sense since we are not given a default rate label in our data, we are just given the labels default vs no default.

upvoted 1 times

Aaronn14 2 years, 1 month ago

You cannot predict rate. You predict a realization, which is either default or not. This question is terribly written.

upvoted 2 times

Skyw4lk3r Most Recent 3 months, 3 weeks ago

Selected Answer: A

A. Because the dataset is just of granted loans of the business, but is needed a grater database where to train the model in order to get aceptable results

upvoted 1 times

nairoh 6 months, 3 weeks ago

"social" does not mean Social Media. This could be linked to demograhic data, si it could improve the score.

upvoted 1 times

TVH_Data_Engineer 1 year, 4 months ago

Selected Answer: B

To predict default rates for credit applicants using the labeled dataset of granted loan applications, the most appropriate course of action would be:

B. Train a linear regression to predict a credit default risk score.

Here's the rationale for this approach:

Appropriate Model for Prediction: Linear regression is a common statistical method used for predictive modeling, particularly when the outcome variable (in this case, the likelihood of default) is continuous. In the context of credit scoring, linear regression can be used to predict a risk score that represents the probability of default.

Utilization of Labeled Data: Since you already have a labeled dataset containing information on loans that have been granted and whether they have defaulted, you can use this data to train the regression model. This historical data provides the model with examples of borrower characteristics and their corresponding default outcomes.

upvoted 5 times

rocky48 1 year, 5 months ago

Selected Answer: B

B. Train a linear regression to predict a credit default risk score.

upvoted 3 times

gaurav0480 1 year, 8 months ago

What would be the target variable if B is correct i.e. training a linear regression model? Default/No-Default is a categorical variable one cannot train a linear regression model with this target variable

upvoted 1 times

FDS1993 1 year, 9 months ago

Selected Answer: C

C - it is a typical approach in credit loans.
Keeping only the accepted loans leads to a bias in the application

upvoted 1 times

Mathew106 1 year, 9 months ago

Selected Answer: B

Linear regression is not the good way to solve such a problem, but you can totally apply linear regression to solve a classification problem. Just set the labels to numeric values 0 and 1 and linear regression will try to predict a value inbetween and round to the closest label (0 or 1).

Totally not the way to go about it, but actually it's possible.

upvoted 2 times

juliobs 2 years, 1 month ago

Selected Answer: D

Cannot be B. This is logistic regression, not linear regression.

D is the only acceptable option.
Social profile can include things like high or low income, for example.
When you apply for a credit you usually have to give this information, so totally legal.

upvoted 1 times

baimus 7 months, 1 week ago

Credit scores are numbers, so this is a regression. Whether or not a client defaults could be a classification, but the option specifies the use of scores, which is fine.

upvoted 1 times

Oleksandr0501 2 years ago

D. Matching loan applicants with their social profiles to enable feature engineering is not recommended as it raises privacy concerns and may not be legal in some jurisdictions. Additionally, social profiles may not be a good indicator of creditworthiness, and relying on them may introduce bias or discrimination.

upvoted 1 times

jin0 2 years, 2 months ago

Because there is no option to know what dataset schema is even though B is needed for this question's purpose Nobody can't select B. So there is none to answer

upvoted 1 times

musumusu 2 years, 2 months ago

all options are wrong: Still in favour of B
A: ofc its good to have more data but its not clear how much data we have
B: Linear can be a workable approach but current situation is not for linear approach, decision tree, random forest etc can be good for it.
C: DAta should be unbiased, removing bias is negative for tranining

upvoted 1 times

Besss 2 years, 3 months ago

Selected Answer: B

default rates can be predicted with linear regression.

upvoted 2 times

21c17b3 1 year, 5 months ago

default rates is classification probability

upvoted 1 times

Whoswho 2 years, 4 months ago

Answer should actually be a logistic Regression model

upvoted 2 times

zellck 2 years, 5 months ago

Selected Answer: B

B is the answer.

upvoted 3 times

ladistar 2 years, 5 months ago

The question asks about default RATES, as in you are predicting a continuous variable, not a discrete one (classification). This is a regression problem, so choice B.

upvoted 2 times

woyaolai 2 years, 5 months ago

I used to be a Credit Risk modeler and I think this question is stupid.

upvoted 8 times

Load full discussion...