BUSINESS ANALYST FOR A

1)Problem statement

First Binary Classification Model

Data_Final Project.xlsx

You work for a bank as a business data analyst in the credit card risk-modeling department. Your bank conducted a bold experiment three years ago: for a single day it quietly issued credit cards to everyone who applied, regardless of their credit risk, until the bank had issued 600 cards without screening applicants.

After three years, 150, or 25%, of those card recipients defaulted: they failed to pay back at least some of the money they owed. However, the bank collected very valuable proprietary data that it can now use to optimize its future card-issuing process.

The bank initially collected six pieces of data about each person:

- Age
- Years at current employer
- Years at current address
- Income over the past year
- Current credit card debt, and
- Current automobile debt

In addition, the bank now has a binary outcome: default = 1, and no default = 0.

You will combine data from the above six inputs to output a single “score.”

At first you are not told what your bank’s own best estimate for its cost per False Negative (accepted applicant who becomes a defaulting customer) and False Positive (rejected customer who would not have defaulted) classification.

Therefore, the best you can do is to design your model *to* maximize *the Area Under the ROC Curve, or AUC.*

You are told that if your model is effective (“high enough” AUC, not defined further) and “robust” (again not defined, but in general this means relatively little decrease in AUC across multiple sets of new data) then it may be adopted by the bank as its predictive model for default, to determine which future applicants will be issued credit cards.

Your first assignment is to analyze the data and create a binary classification model to forecast future defaults.

You are first given a “Training Set” of 200 out of the 600 people in the experiment. The Data_For_Final_Project (below) has both the training set and test set you will need.

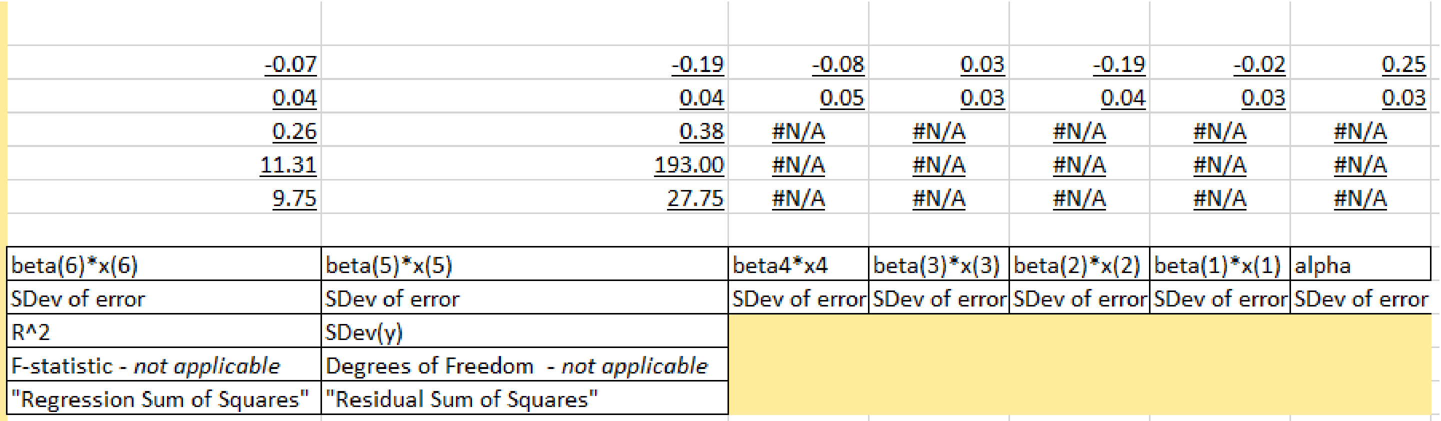
Design your model using the Training Set. Standardized versions of the input data also provided for your convenience. You may combine the six inputs by adding them to, or subtracting them from, each other, taking simple ratios, etc. Exclude inputs that are not helpful and then experiment with how to combine the most informative inputs.

Question: What is your model? Give it as a function of the two or more of the six inputs. For example: (Age + Years at Current Address)/Income [not a great model!].

Your model should have at least two inputs.

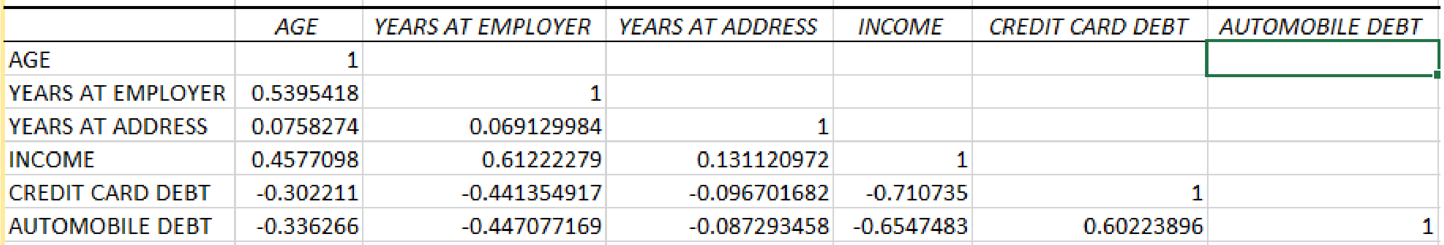
ANSWER)

The first thing that I did was USE The linest function in excel to check out the predictor variables which are important.



-0.07	-0.19	-0.08	0.03	-0.19	-0.02	0.25
0.04	0.04	0.05	0.03	0.04	0.03	0.03
0.26	0.38	#N/A	#N/A	#N/A	#N/A	#N/A
11.31	193.00	#N/A	#N/A	#N/A	#N/A	#N/A
9.75	27.75	#N/A	#N/A	#N/A	#N/A	#N/A
beta(6)*x(6)	beta(5)*x(5)	beta4*x4	beta(3)*x(3)	beta(2)*x(2)	beta(1)*x(1)	alpha
SDev of error	SDev of error	SDev of error	SDev of error	SDev of error	SDev of error	SDev of error
R^2	SDev(y)					
F-statistic - *not applicable*	Degrees of Freedom - *not applicable*					
"Regression Sum of Squares"	"Residual Sum of Squares"					

Then I made the correlation matrix for all the inputs to see if any of the variables are highly correlated with each other.



	AGE	*YEARS AT EMPLOYER*	*YEARS AT ADDRESS*	*INCOME*	*CREDIT CARD DEBT*	*AUTOMOBILE DEBT*
AGE	1					
YEARS AT EMPLOYER	0.5395418	1				
YEARS AT ADDRESS	0.0758274	0.069129984	1			
INCOME	0.4577098	0.61222279	0.131120972	1		
CREDIT CARD DEBT	-0.302211	-0.441354917	-0.096701682	-0.710735	1	
AUTOMOBILE DEBT	-0.336266	-0.447077169	-0.087293458	-0.6547483	0.60223896	1

Observation

1)there is a significant positive correlation between Credit Card Debt and automobile debt.

2)There is a significant negative correlation between automobile debt and income.

3) there is a significant positive correlation between income and automobile debt Years at employer.

Now to make a model with a good AUC, I tried various model by interpreting the outcomes from linest function.

As we can see from linest function beta 2 and beta 5 are weighted the most

The two models TO calculate score for binary classification with the best AUC are

0.25-0.19*Years at a current employer -0.08*income over the past year -0.19*Current credit card debt-0.07* Current automobile debt = SCORE

OR

0.19*Years at a current employer -0.08*income over the past year -0.19*Current credit card debt-0.07* Current automobile debt = SCORE

Both with area under curve = 0.84

2)

What is your model's AUC on the Training Set? Use two digits to the right of the decimal place.

0.84

3.

Initial Assessment for Over-fitting (testing your model on new data)

Next test your model, without changing any parameters, on the Test Set of 200 additional applicants. See the Test Set spreadsheet. It is part of the Data_For_Final_Project (below) and has both the training and test set.

Data_Final Project.xlsx

Hint: Make and use a second copy of the AUC Calculator Spreadsheet so that you can compare Test Set and Training Set results easily.

AUC_Calculator and Review of AUC Curve.xlsx

What is your model's new AUC on the *Test Set*? Give two digits to the right of the decimal place.

ANSWER

0.84

4.

Finding the Cost-Minimizing Threshold for your Model

Now that you have, hopefully, developed your model to the point where it is relatively "robust" across the training set and test set, your boss at the bank finally gives you its current rough estimate of the bank's average costs for each type of classification error.

[Note that all bank models here include only profits and losses within three years of when a card is issued, so the impact of out-years (years beyond 3) can be ignored.]

Cost Per False Negative: $5000

Cost Per False Positive: $2500

For the 600 individuals that were automatically given cards without being classified, the total cost of the experiment turned out to be 25%*($5000)*600 or $750,000. This is $1,250 per event.

Only models with lower cost per event than $1,250 should have *any* value.

Question: What is the *threshold score* on the Training Set data for your model that minimizes Cost per Event? *You will need this number to answer later questions.*

Hint: Using the**AUC Calculator Spreadsheet**, identify *which Column* displays the *same cost-per-event* (row 17) as the overall minimum cost-per-event shown in Cell J2. The threshold is shown in row 10 of that Column. What the threshold means is that at and above this number everything is classified as a "default."

AUC_Calculator and Review of AUC Curve.xlsx

ANSWER

0.1

5.

Finding the Minimum Cost Per Event

Question: Again referring only to the Training Set data, what is the overall minimum cost-per-event?

Hint: You will need this number to answer later questions. If you used the AUC Calculator, the overall minimum cost per event will be displayed in Cell J2.

Note: for Coursera to interpret your answer correctly you must give your answer as an integer - no decimals or dollar sign.

For Example - enter $800.00 as "800"

ANSWER

725

6. Comparing the New Minimum Cost Per Event on Test Set Data

When you compared AUC for the Training and Test Sets, all that is necessary is to look up the two different values in Cell G8. But to get an accurate measure of the *cost-savings*using *the original model on new data*, you can *not* automatically use the new threshold that results in the overall lowest cost-per-event on the Test Set.

Remember that your model is being tested for its ability to *forecast* - but the new optimal threshold will be known only *after* the outcomes for the entire Test Set are known.

All you can use is the model you developed on the Training Set data and the *threshold from the Training Set* that you should have recorded when answering **Question 4**.

Question: At that *same* threshold score (NOT the threshold score that would minimize costs for the new Test Set, but the "old" threshold score that minimized costs on the Training Set) what is the cost per event on *the test set*?

Hint: Using the AUC Calculator Spreadsheet previously provided, locate the column on the Training Set data that has the lowest-cost-per event. That same column and threshold in the *Test Set* copy of the AUC Calculator will have a new cost-per-event, displayed in row 17. This is almost always *higher* than the minimum cost-per-event on the Training Set, and a*lso higher* than what the minimal cost-per-event would be on the Test Set, if one could know the new optimal threshold in advance. This number is the *actual* cost per event when applying the model-and-threshold developed with the Training Set to the new, Test Set data.

Note: for Coursera to interpret your answer correctly you must give your answer as an integer - no decimals or dollar sign.

For Example - enter $800.00 as "800"

Answer
775

7. Putting a Dollar Value on Your Model Plus the Data

Assume your Test Set cost-per-event results from **Question 6** are sustainable long term.

Question: How much money does the bank *save*, per event, using your model and its data-inputs, instead of issuing credit cards to everyone who asks?

Hint: the cost of issuing credit cards to everyone (no model, no forecast) has been determined to be 25%*$5000 = $1,250 per event. Dollar value of the model-plus-data is the difference between $1,250 and your number.

Note: for Coursera to interpret your answer correctly you must give your answer as an integer - no decimals or dollar sign.

For Example - enter $800.00 as "800"

Answer
475

8.

Payback Period for Your Model

Question: Given that it apparently cost the bank $750,000 to conduct the three-year experiment, if the bank processes 1000 credit card applicants per day on average, how many days will it take to ensure future savings will pay back the bank's initial investment?

Give number rounded to the nearest day (integer value).

Hint: multiply your answer to **Question 7** - the cost savings per applicant - by 1000 to get the savings per day.

Answer

475 * 1000 is per day
Therefore to cover the cost
Days requires = **750,000/(**475 * 1000) = 2

9) Any model that is reducing uncertainty will have a True Positive Rate...

○

...Greater than the Test Incidence (% of outcomes classified as "default")

●

...Less than the Test Incidence (% of outcomes classified as "default")

○

...Equal to the Test Incidence (% of outcomes classified as "default")

10. Question 10

Given that the base rate of default in the population is 25%, any test that is reducing uncertainty will have a Positive Predictive Value (PPV)...

○

...Equal to .25

●

...Greater than .25

○

...Less than .25

Given that the base rate of default in the population is 25%, any test that is reducing uncertainty will have a Negative Predictive Value (NPV)...

○

...Less than .75

○

Equal to .75

○

...Greater than .75

12. Question 12

Confusion Matrix Metrics. To determine all performance metrics for a binary classification, it is sufficient to have three values

1. The Condition Incidence (here the default rate of 25%)
2. The probability of True Positives (the True Positive rate multiplied by the Condition Incidence)
3. The “Test Incidence” (also called “classification incidence” - the sum of the probability of True Positives and False Positives)

These three values can all be obtained from the AUC Calculator Spreadsheetand and then used as inputs to the Information Gain Calculator Spreadsheet to determine all other performance metrics.

AUC_Calculator and Review of AUC Curve.xlsx

Information Gain Calculator.xlsx

Question: What is your model’s True Positive Rate?

Save this answer as it will be needed again for Part 3 (Quiz 3)

0.7

13. Question 13

Question: What is your model’s “test incidence”?

Save this answer as it will be needed again for Part 3 (Quiz 3)

0.3349

Hi Dimitry,

I donno if you are still looking for this answer or not, I'll write what I think may be the problem here. It may help somebody else looking for similar question.

your "c" figure of 0.81 is in fact the test incidence but it is a very high figure when compared to the condition incidence, that tells us that the model used to get these numbers is not very good.

In the dataset only 25% of people are listed as defaulters(condition incidence). What your model is going to do is classify 81% of total applicants as defaulters(test incidence). Then since your model's TP is 12% that will result in 69% of False positives, which is very very high. A high FP rate adds significant cost.

We should try to tweak our model to make make the test incidence not very high as compared to condition incidence, so that our FP rate is not very high.

0 Upvotes

QUIZ 2

1. Question 1

Comparing the Information Gain of Eggertopia Scores and Your Model

Both the Eggertopia Scores and your binary classification model can be thought of as *tools to reduce uncertainty* about future default outcomes of credit card applicants.

Your own model, developed in Part 1, identifies *dependencies* between, on the one hand, the six types on input data collected by the bank, and on the other hand, the binary outcome default/no default.

If we assume that the dependencies identified by Eggertopia Scores and by your model on the Test Set are stable and representative of all future data (a big assumption) we can draw some further conclusions about how much information gain, or reduction in uncertainty, is provided by each.

Definitions are given in the Information Gain Calculator Spreadsheet, provided below.

Information Gain Calculator.xlsx

Question: On your model's Test Set results, what is the conditional entropy of default, given your test classifications?

Hint: you need your model's true positive rate from Part 1, Question 12, and "test incidence" [proportion of events your model classifies as default] from Part 1, question 13. Use the condition incidence of 25% and your model's True Positive rate to calculate the portion of TPs. Then you have the inputs needed to use the Information Gain Calculator Spreadsheet.

The conditional entropy of default, given my test classifications, is
0.6725.

Conditional entropy of default is H(X | Y)

2.

Recall that the entropy of the original base rate, minus the conditional entropy of default given your test classification, equals the Mutual Information between default and the test.

$I(X;Y) = H(X) – H(X|Y)$.

The population of potential credit card customers consists of 25% future defaulters. The base rate incidence of default (.25, .75) has an uncertainty, or entropy, of H(.25, .75) = .25*log4 + .75*log1.333 = .8113 bits.

Question: On your test set results, what is the Mutual Information, or information Gain, in average bits per event?

The Information gain in average is 0.1388 bits per event.

Mutual information (Info. gain) calculated in units of bits per event and information Gain, in average bits per event are the same. The question was repeating itself to make it more clear.

3. Question 3

Recall that Percentage Information Gain (P.I.G.) is the ratio of I(X;Y)/H(X).

Question: on your Test Set results, what is the Percentage Information Gain (P.I.G.) of your model?

P.I.G is 17.11%

4.

Since you have, for you model on the Test Set, a savings-per-event, and a bits-per-event (Mutual Information) you can calculate a savings-per-bit. This is a powerful concept, because it places a financial value directly on the information content of a model (or additional data source, like the Eggertopia scores).

Question: How many dollars does the bank save, for every bit of information gain achieved by your model?

ANS Dollar saving = your dollar saved per event / your information gain. The numbers given in the question are used as samples, just replace those numbers with the numbers you have.

475/I.G

475/0.1388

5. Question 5

Information Gain of Eggertopia Scores over the Base Rate

For questions in this section, assume your model and the data it uses are not available – the bank's choice is between Eggertopia scores and the base rate.

Question: What is the Mutual Information of the Eggertopia Scores?

In other words, on the Test Set, What is the information gain, in average bits per event, over the base rate of (.25, .75) offered by the Eggertopia Scores?

○

.1305 bits per event

○

.1255 bits per event

○

.1243 bits per event

●

.1205 bits per event

Question 5 Information Gain of Eggertopia Scores over the Base Rate On the test set, What is the information gain, in average bits per event, over the base rate of (.25, .75) offered by the Eggertopia Scores?

Given

Base Rate = (0.25, 0.75)

PPV = 0.48 [from part 2]

NPV = 0.888 [from part 2]

Find

Information Gain

Calculate

*Base rate = H(X) = .25*log[2](1/.25)+.75*log[2](1/.75)*

Model rate = H(X|Y) = c*H(PPV, 1-PPV) + d*H(1-NPV, NPV)

Information Gain = Base rate - Model rate = H(X) - H(X|Y)

6. Question 6

On the test set, what is the Eggertopia scores’ Percentage Information Gain (PIG)?

○

13.95%

○

15.25%

●

14.85%

○

15.35%

Question 6 On the test set, what is the Eggertopia scores’ Percentage Information Gain (PIG)?

Percentage Information Gain (PIG) = Information Gain/initial entropy

7. Question 7

If Eggertopia data were free, and your model was unavailable, what would the dollar savings per bit of information extracted be?

Dollar savings are $412 rounded to the nearest dollar- from quiz 2, question 6

○

Value would be $3,627 per bit.

●

Value would be $3,427 per bit.

○

Value would be $427 per bit.

Question 7 If Eggertopia data were free, and your model were unavailable, what would the dollar savings per bit of information extracted be?

Saving per bit = $ saved / bits extracted

8.

Incremental Information Gain of Eggertopia Scores Compared to Your Model and Available Data (any answer scores)

(For this section, assume your Model and the Data it uses *are* available).

Question: What is the *incremental* information gain of the Eggertopia scores, over your model from Part 1, in average bits per event, if any?

My model had information gain of 0.1388 bits per event where as eggertopia has information gain of 0.1205 bits per event.Hence my model is better with 0.0183 bits more information gain.

If your model has higher information gain that eggertopia, Just give a negative number.

9. Question 9

What is the maximum (break-even) price the bank should pay for Eggertopia scores, per score, if your model from Part 1 and data are already available?

0 $ since my model is better

Part 4: Modeling Profitability Instead of Default

Question 1

Correct

1 / 1

point

1. Question 1

Modeling Profitability Instead of Default

Modeling Profitability Level as a Continuous Output (Instead of Binary Classification Default/No Default)

Introduction

Both your own model and the forecast based on Eggertopia scores are binary classifications: they forecast one of just two outcomes: “Default” or “No Default.” Your boss is interested in the idea that it might be preferable instead to model and forecast profits and losses as continuous values, using a a multivariate linear regression model on the same six input variables. This idea has arisen because the bank has been reviewing individual profit and loss numbers for each customer over the three-year period and has made an interesting discovery: some defaulting customers carried so much debt for so long, and paid so much interest on it, that they were profitable for the bank even though they defaulted! Many customers who seem to have risky spending behaviors are also among the most profitable for a lending business. And, at the opposite extreme,customers who always paid off their cards in full each month never defaulted but were not very profitable: the bank barely broke even, or even lost money, on its“safest” borrowers.

Your boss asks you to forecast each applicant’s expected profitability, in dollars,before deciding whether or not to issue them a credit card. He wants to know how reliable this type of forecast would be: what is the range above and below the point estimate that will be correct 90% of the time?

Although it might be possible to combine the six inputs in other ways, in the interests of time and focusing on the key learning objectives, we will use only a simple linear combination of the six input variables for Part 4 of this Project. (You should not include the Eggertopia Scores as an input variable).

Question 1 is about the coefficients or “betas” used to combine the standardized inputs to get the best-fit-line on standardized outputs on the Training Set. We then use those fixed betas to measure the observed residual error of the model on the Test Set.

Questions 2 through 6 concern the forecasts on the Test Set.

Questions 7 through 11 look at the Training Set results so that they can be compared (for possible over-fitting) against the Test Set Results.

Questions 12 through 14 are about the uncertainty that remains in a new individual forecast of profitability.

Use the Excel “Linest” function on the six inputs and profitability output on the 200 Training Set applicants to calculate the coefficients (the “betas”) that result in the best-fit line.

Question: Do you feel prepared to take this quiz?

●

Yes

○

No

Question 2

Correct

1 / 1

point

2. Question 2

Question: What are your values for each “beta” on the Training Set?

- Age
- Years at current employer

- Years at current address
- Income over the past year
- Current credit card debt
- Current automobile debt

●

.01, .19, -.07, .64, -.06, 0

Correct

○

.01, .19, .07, .64, .06, 0

○

01, -.19, -.07, -.64, -.06, 0

Question 3

Correct

1 / 1

point

3. Question 3

For this question, use the Liner Regression Forecasting explanation and Excel spreadsheet.

Question: What is the root-mean-square residual (the standard deviation of model error) on Standardized output for the Test Set?

○

.8109

○

.5835

●

0.6750

Correct

If you have set up the betas correctly in cells C7:H7, the standard deviation of model error on standardized outputs should appear in [Cell W8]

○

.6875

○

.3250

Question 4

Correct

1 / 1

point

4. Question 4

For this question, use the Linear Regression Forecasting Explanation and Spreadsheet.

Question: What is the observed correlation R on the Test Set?

○

.8095

○

.7590

○

.7332

●

0.7378

Correct

Correct!

Question 5

Incorrect

0 / 1

point

5. Question 5

For this question, use the Linear Regression Forecasting explanation and Excel spreadsheet.

Question: What is the Standard deviation of model error, in Dollars, for the Test Set?

○

$3,379.36

○

$3,996.81

○

$3,885.14

○

$3,411.80

This should not be selected

The standard deviation of the original Training Set outputs in dollars is $5,755.91. [Cell AE6] Multiplying the standard deviation of error on standardized data from the Test Set – Answer to Question 3 above - by $5,755.91 should give the correct answer, which is displayed in [Cell Y7].

Question 6

Incorrect

0 / 1

point

6. Question 6

For this question, use the Linear Regression Forecasting explanation and Excel spreadsheet:

Question: What is the 90% confidence interval, in dollars, for the Test Set?

○

$5,611.91 above the point estimate, and $5,611.91 below the point estimate

This should not be selected

The 90% two-sided confidence interval means estimates 5% of values are excluded at each of the top and bottom of the distribution of possible estimation errors. Z-score = Normsinv(.95) = 1.6448. Multiplying the standard deviation of model error in dollars – from Question 4 above – by 1.6448 gives the correct answer, which should appear in [Cell Y10].

○

$5,558.55 above the point estimate, and $5,558.55 below the point estimate

○

$6,574.17 above the point estimate, and $6,574.17 below the point estimate

○

$6,390.49 above the point estimate, and $6,390.49 below the point estimate

Question 7

Correct

1 / 1

point

7. Question 7

What is the Percentage Information Gain (P.I.G.) on the Test Set?

○

37.2%

○

26.4%

○

27.7%

Correct

Correct!

○

18.9%

Question 8

Correct

1 / 1

point

8. Question 8

For this question, use the Linear Regression Forecasting explanation and Excel spreadsheet:

Question: What is the Correlation, R, of your model on the Training Set?

○

.8095

Correct

Correct! R^2 is given by the Linest function on Training Set data. Enter R^2 in [Cell AC4] to get R as output of [Cell AA4].

○

.7805

○

.7505

Question 9

Correct

1 / 1

point

9. Question 9

For this question, use the Linear Regression Forecasting explanation and Excel spreadsheet:

You need to quantify the uncertainty in a regression model forecast of applicants' future profitability. Assume that both the forecast profits and the errors have a Gaussian distribution. You will calculate the standard deviation of model error on standardized data, the standard deviation in dollars of the model error, and the 90% confidence interval for profitability estimates.

Question: What is the standard deviation of your model error on the standardized Training Set output?

○

.587

Correct

R^2 is given by Linest function on the Training Set. Standard deviation of model error = Sqrt(1-R^2). Enter R^2 in [Cell AC4] and answer appears in [Cell AA6].

○

.487

○

-.487

○

-.587

Question 10

Correct

1 / 1

point

10. Question 10

For this question, use the Linear Regression Forecasting explanation and Excel spreadsheet.

Question: What is the standard deviation of model error in dollars on the Training Set?

**This may seem similar to question 5, but Q5 refers to the Test Set.

○

$4,379.36

○

$4,312.91

○

$5,500.87

○

$3,379.36

Correct

Correct!

Question 11

Correct

1 / 1

point

11. Question 11

For this question, use the Linear Regression Forecasting explanation and Excel spreadsheet.

Question: What is the 90% confidence interval, in dollars, on the Training Set?

**This may seem similar to question 6, but Q6 refers to the Test Set.

○

$5,558.55

Correct

Correct!

○

$5,328.93

○

$6,211.18

○

$7,128.55

Question 12

Correct

1 / 1

point

12. Question 12

For this question, use the Linear Regression Forecasting explanation and Excel spreadsheet.

Question: What is the Percentage Information Gain (P.I.G.) on the Training Set?

**This may seem similar to question 7, but Q7 refers to the Test Set.

○

41.4%

○

37.5%

Correct

Correct!

○

32.4%

○

36.5%

Question 13

Correct

1 / 1

point

13. Question 13

Questions 13 through 15 use the same example applicant.

The following data are known about the sample applicant:

Age: 42.00

Years at Employer: 12.44

Years at Address: 0.9

Income: $121,400

CC debt: -34,228

Auto debt: -23,411

To convert above inputs to standardized form, locate the **Training Set** Spreadsheet (first bottom tab of workbook) in the **Data for Final Project** Workbook.

Data_for_Final_Project.xlsx

Use the input means [Cells C207:H207] and standard deviations [Cells C209:H209].

Use the Training Set profitability mean [$1,905.51] and standard deviation [$5755.91] from the **Profit and Loss** (last bottom tab) Spreadsheet.

Use the Test Set standard deviation of error on standardized outputs of .6750

Question: What is the point estimate of profitability, in dollars?

○

-$10,683.61

○

$11,109.61

○

$10,683.61

Correct

Standardize each input variable separately, using the appropriate mean and standard deviation from the *Data for Final Project* Workbook [mean, cells C207:H207 and standard deviation, cells C209:H209].

Then multiply each individual z-score by its “beta” coefficient from the original Excel “Linest” Calculation on the (standardized) Training Set.

Then sum the results. That sum is the point forecast of profitability, expressed as a standardized output (z-score). The correct z-score estimate for y = 1.525059.

Multiply the z-score by the standard deviation of profits ($5755.91), then add the mean profit ($1,905.51). You should get $10,683.61.

○

$8,451.61

Question 14

Correct

1 / 1

point

14. Question 14

The following data are known about the sample applicant:

Age: 42.00

Years at Employer: 12.44

Years at Address: 0.9

Income: $121,400

CC debt: -34,228

Auto debt: -23,411

To convert above inputs to standardized form, locate the **Training Set** Spreadsheet (first bottom tab) in the **Data for Final Project** Workbook.

Use those means [Cells C207:H207] and standard deviations [Cells C209:H209].

Use the Training Set profitability mean [$1,905.51] and standard deviation [$5755.91] from the **Profit and Loss** (last tab on bottom) Spreadsheet

Use the Test Set standard deviation of error on standardized outputs of .6750

Question: With 50% confidence, what is the range of profitability?

○

Range from $13,304.16 to $8,063.06.

Correct

The mean of the interval is $10,683.61 (from Question 13).

This is a left-sided 25% confidence interval.

The normsinv(p = .25) = -0.67448975.

The interval from p= .25 to p = .75 is +- plus or minus

(the standard deviation of error as a fraction of the standard deviation of profitability)*(normsinv(.25))*(standard deviation of profits)

= (.675)*(0.67448975)* ($5755.91)

= +- $2,620.55.

The 50% confidence interval range is from ($10,683.61 + 2,620.55) to ($10,683.61 - $2,620.55).

○

Range from $12,962.61 to $10,683.61

○

Range from $11,823.28 to $9,543.94

○

Range from $10,683.61 to – $2,278.99

Question 15

Correct

1 / 1

point

15. Question 15

The following data are known about the sample applicant:

Age: 42.00

Years at Employer: 12.44

Years at Address: 0.9

Income: $121,400

CC debt: -34,228

Auto debt: -23,411

To convert above inputs to standardized form, locate the **Training Set Spreadsheet**(bottom tab) in the **Data for Final Project Workbook.**

Use those means [Cells C207:H207] and standard deviations [Cells C209:H209].

Use the Training Set profitability mean [$1,905.51] and standard deviation [$5755.91] from the **Profit and Loss** (bottom tab) Spreadsheet

Use the Test Set standard deviation of error on standardized outputs of .6750 .

Question: With 99% confidence, what is the range of profitability?

○

Range from $19,388.27 to 10,683.61.

○

Range from $16,388.27 to -$7,704.31

○

Range from $20,691.32 to $675.90.

Correct

The mean of the interval is $10,683.61 (from Question 13).

The left-sided confidence interval is .5% or .005. The normsinv(p = .005)

= -2.575829304.

The interval from p= .005 to p = .995 is:

+- plus or minus

(the standard deviation of error as a fraction of the standard deviation of profits)*(normsinv(p = .005))*(standard deviation of profits)

= (.675)*(2.575829304)* ($5755.91)

= +- $10,007.71.

The range is from ($10,683.61 + $10,007.71)

to ($10,683.61 - $10,007.71).

○

Range from $10,683.61 to -$8,704.31

Question 16

Correct

1 / 1

point

16. Question 16

Comparing Test Set and Training Set Performance

Question 15: Between the Training Set and the Test Set, the dollar value of the standard deviation of model error…

○

Decreased by about 15%, which suggests a very strong model on Test Set data.

○

Increased by less than 20%, which suggests minimal model over-fitting.

Correct

Correct. The actual change was from $3,379.36 to $3,885.14, an increase of less than 15%.

○

Increased by more than 50%, which leads to the conclusion of model over-fitting.

○

Increased by more than 25%, which suggests possible model over-fitting.

Thank you Elizabeth Wagner-Badrov. To break it down. You are given the 6 measured values - Age, Year at employer...etc. Given this you want to determine the point estimate of profitability of this new data (or person). Profitability is dependant on the 6 variables. The score model or the formula for profitbility is given by the Linest function (which is already worked out). (a)Linest functions gives the Beta values x6 to x1 (Left to right), where x1 corresponds to AGE. (b) x1, x2..x6 are the standardized values of the new data. The formula is betax1*x1+ betax2* x2+...+betaxb* x6. This results in the z-score of Profitability. Next we want to determine the Point estimate of profit. We already know the Mean and Standard Deviation of profit from Training set. Therefore the point estimate of profit is the the z-score point from the mean, which is, mean+ z-score * standard deviation (of profit from training set). This point estimate is a mix of true value + noise. Finally, we can tell with 50% or 99% of C.I what the range of profit will be. The range of profit is *Original point estimate* (in $) + and - *Estimate of residual* (in $). Estimate of residual is already determined from the *Standard deviation of profit*and *root mean square residual. (No manipulation needed on your part). For the C.I. the only thing you need to manipulate is 50% and 99% in the formula.*

Quiz PART 5

What is your predictive model?

a. Describe the arithmetic clearly so that another learner could implement your model on new standardized input data if they wished.

b. Give an example of the score you would assign the following applicant, whether they would be approved or rejected for a credit card and why.

a) Years at a current employer: -0.19
income over the past year: -0.08
Current credit card debt: -0.19
Current automobile debt: -0.07

SCORE = 0.19*Years at a current employer -0.08*income over the past year -0.19*Current credit card debt-0.07* Current automobile debt

b)Using the above formulae the score that would be assigned to the following applicant will be -0.03.

Optimizing AUC, we have got the threshold for the min cost /event as 0.25

Since this score lies below 0.25- it will be classified as a negative test which basically means that this person will be profitable for the bank and the credit card will be approved.

Q

What would the bank's *average profit per applicant* be (net profits divided by 200) when using your predictive model on the Training Set?

The net profit per applicant will be 794$

What is the *incremental financial value* per applicant of your model over no model on the Training Set?

654.41$ is the incremental financial value per application of model over no model.

Evaluate your model on the Test Set data. How confident are you that your model does not over-fit the Training Set data? The only basis to evaluate over-fitting is to give the same metrics on the Test Set and Training Set, and compare them.

My model does not overfit the training data as the AUC for both the datas is 0.84. Other than that the Cost per event at the same threshold for testing data is 778$ which is just 4$ more than the minimum cost at training data.

valuate your model on the Test Set data. How confident are you that your model does not over-fit the Training Set data?

A. Choose between three broad degrees of confidence: “very” “somewhat” or “not at all.” (Note that “not at all” is still an acceptable answer if you give persuasive reasons for why you chose this answer).

B. Explain the evidence your degree of confidence is based upon. Your explanation should include the test set profits and training set profits per applicant.

How much confidence to have in the model must relate to the relationship between the profits-per-applicant on the Training Set and the Test Set

A)"Very"

B)I am very confident that my model does not overfit the training data as the AUC for both the datas is 0.84. Other than that the Cost per event at the same threshold for testing data is 778$ which is just 4$ more than the minimum cost at training data.

The profit per applicant of the training set is 2058 $ per applicant

and profit per applicant on the testing set is 2087 $ per applicant which is infact more.

Hence my model is working even better on testing data.

Profit per applicant is calculated by using

calculate the profit per event = ((TP*0)+(TN*4000)+(FN*-4900) + (FP*0)) / Total Events

1.The template spreadsheet counts the minimum cost but what we need to do is calculate the maximize profit,so you will have to adjust the spreadsheet.

2.The profit is (TP*0)+(TN*4000)+(FN*-4900) + (FP*0)) / Total Events. TN is those approved applicants do bring profit to the bank and the FN is those approved applicants actual do not make the bank lose money.

In the spreadsheet given to us, there is no TN, my suggestion is: add one line above the "False positive rates at the threshold" and calculate the TN by "TN = Negative - False Negetive = 150 - FP (since the instruction told us that 75% is negative/profitable) and then you adjust the line 18 accordingly.