

DSC 433/533 – Class 4 – Charity Case Example

Task: A major charity wishes to solicit a donation. Which donor characteristics are associated with large donations? Can we use this information to target the solicitations.

25 measured characteristics (features) pre-selected from an original database with 481 features.

What to do with nominal features?

STATE: 4 groups coded with 3 indicator variables, STGP_2, STGP_3, and STGP_4.

RFA: 4 groups coded with 3 indicator variables, RFA_2A_E, RFA_2A_F, and RFA_2A_G.

1176 donors responded to the last mailing (original database contains 95412 former donors):
588 (50%) *training*, 353 (30%) *validation*, 235 (20%) *test* (seed: 12345).

Consider predicting the donation amount for each donor.

Multiple linear regression model

- Select XLMiner > Prediction > Multiple Linear Regression.
- In step 1 move all the variables except ID, STATE, STGP, AGE, RFA_2, and RFA_2A to the “Input variables” box and “TARGET_D” to the “Output variable” box.
- In step 2 check “Fitted values.”
- On the “Advanced” dialog box check “Studentized residuals,” “Cook’s distance,” and “Hat matrix diagonals.”
- Select “Summary report” for “Score training data.”
- Select “Detailed report,” “Summary report,” and “Lift charts” for “Score validation data.”
- De-select “Summary report” for “Score test data” (we won’t be using test data for this example).

Select the “MLR_Resi-FitVal1” worksheet, and use the chart wizard to make the following scatterplots.


- Highlight the “Fitted Values” and “Stu. Residuals” columns (D5:E593) to make a plot with studentized residuals on the vertical axis and fitted values on the horizontal axis.
- Highlight the “Cooks” and “Hat Matrix” columns (F5:G593) to make a plot with leverages on the vertical axis and Cook’s distances on the horizontal axis.
- Copy the “Fitted Values” column (D5:D593) to column H and copy the “TARGET_D” column (AK19:AK607 in the Data_Partition1 worksheet) to column I, and then highlight the “Fitted Values” and “spend” columns (H5:I593) to make a plot with actual donation values on the vertical axis and fitted values on the horizontal axis.

Construct a Lift Chart for donations for the Validation sample. This has cumulative solicitations sent (on the horizontal axis) versus cumulative expected donations (on the vertical axis). You can find the predicted donations values on the “MLR_ValidScore1” worksheet in the column headed “Predicted Value” next to the column headed “Actual Value.”

- Select and copy these two columns (from D5 to E358), and paste them into the AI and AJ columns.

- Select Data > Sort to sort the AI and AJ columns in descending order using “Predicted Value” in column AI to sort on.
- Put consecutive integers from 1 to 353 into column AK (e.g., type “1” into cell AK6, then select and drag the bottom right corner of the cell while holding down the “Control” key to put consecutive integers into cells AK6 to AK358).
- Put cumulative expected donations from the linear regression analysis into column AL (e.g., type “=SUM(AJ\$6:AJ6)” into cell AL6, then copy this cell into AL6 to AL358).
- Put cumulative expected donations from a simple random sample of customers into column AM (e.g., type “=AL\$358*AK6/AK\$358” into cell AM6, then copy this cell into cells AM6 to AM358).
- Type suitable titles for columns AK, AL, and AM, e.g., rows 5 to 8 should be:

Predicted Value	Actual Value	Cumulative solicitations	Linear regression	Random sample
80.66523406	100	1	100	15.09773371
64.51030568	75	2	175	30.19546742
59.53783185	30	3	205	45.29320113

Then highlight cells AK5 to AM358, start the “Chart Wizard” by clicking on the chart icon () , select “XY (Scatter)” for the Chart type, and select “Scatter with data points connected by lines” for the Chart sub-type. Label the resulting chart by selecting “Chart > Chart Options,” and labeling the X-axis “solicitations sent,” labeling the Y-axis “donations,” and titling the chart “Lift Chart for Donations.” You can check your lift chart looks OK because XLMiner should have drawn one automatically on the “MLR_ValidLiftChart1” worksheet.

The lift in the first decile (the first 35 solicitations) is **2.235**. (This is the cumulative expected donations under the linear regression model for the first 35 solicitations divided by the cumulative expected donations for a random sample for the first 35 solicitations. You can check your calculation because XLMiner should also have drawn a “decile-wise” lift chart automatically on the “MLR_ValidLiftChart1” worksheet.)

The “root mean square error” (RMSE) in the validation sample if we use a random sample (i.e., use the sample mean to predict donations) is the sum of the squared differences between actual donations and the sample mean. It is **\$10.97**. To calculate this:

- Calculate the sample mean in cell E359 of the “MLR_ValidScore1” worksheet as “=AVERAGE(E6:E358).”
- Put squared differences into column A (e.g., type “=(E6-E\$359)^2”) into cell A6 and copy this cell into cells A6 to A358).
- Calculate the root mean square error in cell A359 as “=SQRT(AVERAGE(A6:A358))”

The “root mean square” error in the validation sample if we use the linear regression model to predict donations is the sum of the squared differences between actual donations and predicted donations. It is **\$6.24** (also in the “MLR_Output1” worksheet). To calculate this:

- Put squared residuals into column AO (e.g., type “=(AI6-AJ6)^2”) into cell AO6 and copy this cell into cells AO6 to AO358).
- Calculate the root mean square error in cell AO359 as “=SQRT(AVERAGE(AO6:AO358)).”

Missing data for AGE

Re-open original Charity.xls file.

XLMiner > Data Utilities > Missing Data Handling.

Age – median – apply (overwrite).

Then partition: 588 (50%) *training*, 353 (30%) *validation*, 235 (20%) *test* (seed: 12345).

- Select XLMiner > Prediction > Multiple Linear Regression.
- In step 1 move all the variables excepted, STATE, STGP, RFA_2, and RFA_2A to the “Input variables” box and “TARGET_D” to the “Output variable” box.
- Select “Summary report” for “Score training data.”
- Select “Detailed report,” “Summary report,” and “Lift charts” for “Score validation data.”
- De-select “Summary report” for “Score test data” (we won’t be using test data for this example).

Lift in first decile: **2.235**.

RMSE: **\$6.24**.

p-value for AGE: **0.68**.

Conclusion: no need for AGE in the model.

Transform TARGET_D to ln(TARGET_D)

Re-open original Charity.xls file.

Add LTARGET_D (=ln(TARGET_D)) column.

Then partition: 588 (50%) *training*, 353 (30%) *validation*, 235 (20%) *test* (seed: 12345).

- Select XLMiner > Prediction > Multiple Linear Regression.
- In step 1 move all the variables excepted, STATE, STGP, AGE, RFA_2, and RFA_2A to the “Input variables” box and “LTARGET_D” to the “Output variable” box.
- In step 2 check “Fitted values.”
- On the “Advanced” dialog box check “Studentized residuals.”
- Select “Summary report” for “Score training data.”
- Select “Detailed report,” “Summary report,” and “Lift charts” for “Score validation data.”
- De-select “Summary report” for “Score test data” (we won’t be using test data for this example).

Select the “MLR_Resi-FitVal1” worksheet, and use the chart wizard to make the following scatterplots.


- Highlight the “Fitted Values” and “Stu. Residuals” columns (D5:E593) to make a plot with studentized residuals on the vertical axis and fitted values on the horizontal axis.

Construct a Lift Chart for donations for the Validation sample. This has cumulative solicitations sent (on the horizontal axis) versus cumulative expected donations (on the vertical axis). You can find the predicted log-donations values on the “MLR_ValidScore1” worksheet in the column headed “Predicted Value” next to the column headed “Actual Value.”

- Exponentiate these two columns (D5 to E358) in the AI and AJ columns and “paste special” as values..
- Select Data > Sort to sort the AI and AJ columns in descending order using “Predicted Value” in column AI to sort on.

- Put consecutive integers from 1 to 353 into column AK (e.g., type “1” into cell AK6, then select and drag the bottom right corner of the cell while holding down the “Control” key to put consecutive integers into cells AK6 to AK358).
- Put cumulative expected donations from the linear regression analysis into column AL (e.g., type “=SUM(AJ\$6:AJ6)” into cell AL6, then copy this cell into AL6 to AL358).
- Put cumulative expected donations from a simple random sample of customers into column AM (e.g., type “=AL\$358*AK6/AK\$358” into cell AM6, then copy this cell into cells AM6 to AM358).
- Type suitable titles for columns AK, AL, and AM, e.g., rows 5 to 8 should be:

Predicted Value	Actual Value	Cumulative solicitations	Linear regression	Random sample
75.11142202	100	1	100	15.09773371
66.46086366	75	2	175	30.19546742
53.97751088	25	3	200	45.29320113

Then highlight cells AK5 to AM358, start the “Chart Wizard” by clicking on the chart icon () , select “XY (Scatter)” for the Chart type, and select “Scatter with data points connected by lines” for the Chart sub-type. Label the resulting chart by selecting “Chart > Chart Options,” and labeling the X-axis “solicitations sent,” labeling the Y-axis “donations,” and titling the chart “Lift Chart for Donations.” You can check your lift chart looks OK because XLMiner should have drawn one automatically on the “MLR_ValidLiftChart1” worksheet.

The lift in the first decile (the first 35 solicitations) is **2.207**. (This is the cumulative expected donations under the linear regression model for the first 35 solicitations divided by the cumulative expected donations for a random sample for the first 35 solicitations. You can check your calculation because XLMiner should also have drawn a “decile-wise” lift chart automatically on the “MLR_ValidLiftChart1” worksheet.)

The “root mean square” error in the validation sample if we use the linear regression model to predict donations is the sum of the squared differences between actual donations and predicted donations. It is **\$6.64** (also in the “MLR_Output1” worksheet). To calculate this:

- Put squared residuals into column AO (e.g., type “=(AI6-AJ6)^2”) into cell AO6 and copy this cell into cells AO6 to AO358).
- Calculate the root mean square error in cell AO359 as “=SQRT(AVERAGE(AO6:AO358)).”

Conclusion: no need to use log-transform here.