**Individual Assignment 2: Quantitative Analysis of Credit**

This assignment is based on the data we used during class 4. There are two files:

* The credit data file contains data pertaining to 1,000 personal loan accounts.
* The data dictionary contains a description of what the various variables mean.

As a part of a new credit application, the company collects information about the applicant. The company then decides an amount of the credit extended (the variable CREDIT\_EXTENDED). For these 1,000 accounts, we also have information on how profitable each account turned out to be (the variable NPV). A negative value indicates a net loss, and this typically happens when the debtor defaults on his/her payments.

The goal in this assignment is to investigate how one can use this data to better manage the bank's credit extension program. **Specifically, our goal is to develop a classification model to classify a new credit account as “profitable” or “not profitable."**Secondly we want to compare its performance in the context of decision support to a linear regression model that predicts NPV directly.

Please answer all the questions. Supply supporting documentation and show calculations as needed. **Please submit a single, well-formatted PDF or Word file**. The instructor should not need to go searching for your answers! Feel free to add your scripts or other supporting information as an appendix (in the same word or pdf file) – the appendix will not be graded, but will help the instructor give you feedback, if your model differs substantially from the solutions.

**Data Preparation**

The data preparation repeats the steps from class 4:

1. The goal is to predict whether a new credit will result in a profitable account.  Create a new variable to use as the dependent variable.
2. Split the data into 2 parts, setting the seed to 1, 70% training and 30% validation.

**The Assignment**

1. ***Applying Logistic Regression***

If one fits a Logistic Regression Model using all the independent variables, one observes a) a gap in the classification performance between the training data and the validation data, and b) very high p-values for some of the variables. The performance gap between the training and validation may be a sign of overfitting, and the high p-values may be a sign of “useless” variables in the model, or of multicollinearity.

1. Our goal is to classify credit requests into “profitable” and “not profitable." To that end, fit a backwards elimination model. You may find the helpful guidance in the book (the discussion on Variable Selection following Figure 10.3, and our in-class activity).  
     
   **Note:** Exclude Credit Extended and any other variables not appropriate for the analysis.

Include the model (the variables and the corresponding regression coefficients) as an Exhibit.  
See exhibit 1 below

Table

Description automatically generated

1. Based on your model, and setting the cut-off value to 0.5, please provide the following information (based on the validation data):
   * The sensitivity of the model = 0.8009
   * The specificity of the model = 0.5316
2. ***ROC Curves***
3. We now want to compare the predictive performance of the model on the training sample and on the validation sample. Create a **single** figure that compares the ROC curves for both the training sample and the validation sample.

HINT: Either create the combined ROC curve figure in EXCEL, or you may find this R code helpful which does the same for our Beer data.

#libraries and data

library(tidyverse)

library(ROCR)

Beer <- read\_csv("beer.csv", show\_col\_types = FALSE)

#In order to be able to run the Logistic Regression model, we need to translate the preference into a binary variable, we will code the Preference as Light =1 and Regular = 0.

Beer<-Beer %>% mutate(Preference=as.factor(ifelse(Preference=="Light",1,0)))

#Add a counter so we can recycle our splitting code

Beer<-Beer %>% mutate(ObsNum=seq(1:100))

#split up the data

Train <- Beer %>% sample\_frac(0.7)

Validation <- Beer %>% anti\_join(Train, by="ObsNum")

#Now we will run our logistic regression model, and create the predictions:

lr1 <- glm(Preference ~ . -ObsNum, family = "binomial", data = Train)

summary(lr1)

Trainpredprob <- predict(lr1, type="response")

Validationpredprob <- predict(lr1, type="response",newdata=Validation)

#We are now ready to explore the ROC curve

#to get both lines on the same plot with the ROCR package we need to play a little trick

#Step0: gather predictions and lables into lists

AllPredictions=list(Trainpredprob,Validationpredprob)

AllLabels=list(Train$Preference,Validation$Preference)

#step 1 is to create a "prediction object" that keep the predictions and outcomes

Allpred<-prediction(AllPredictions,AllLabels)

#step 2 we plot:

plot(performance(Allpred,"tpr","fpr"))

abline(0,1)

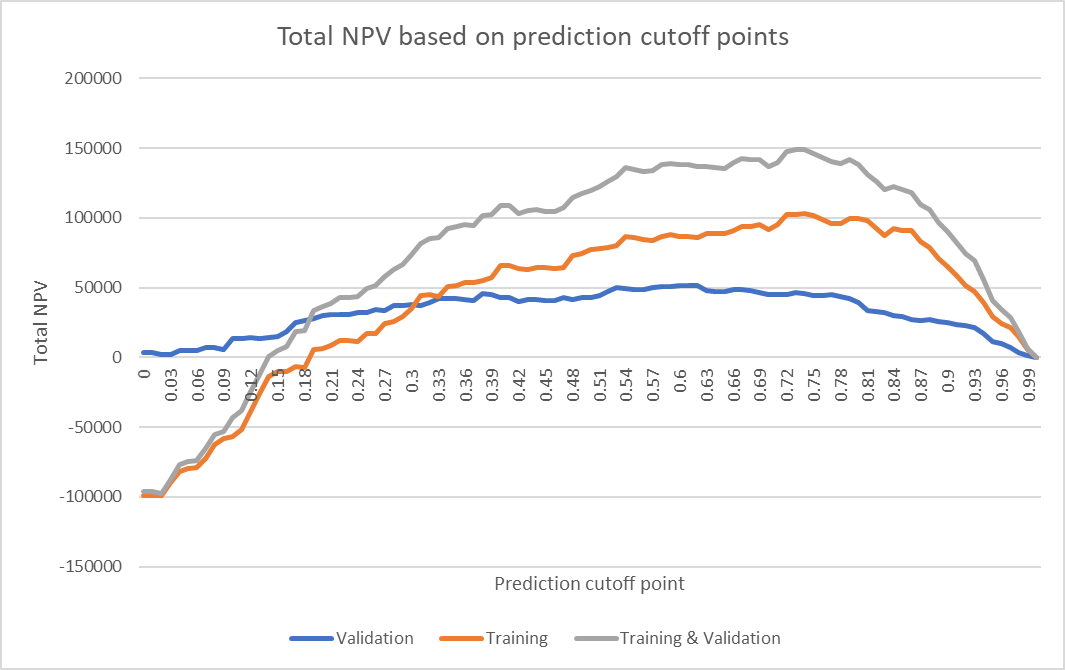
Include a ***clean*** figure as an Exhibit.  
 See Exhibit 2 below

Chart

Description automatically generated

1. ***Finding the "best" cut-off***
2. Either in R (using a loop) or in Excel (using a data-table) calculate the total NPV (assuming we extend credit to all applicants classified as “profitable” as a function of the cut-off based on the training data). Select the best cut-off.  Include a table or plot as an Exhibit.

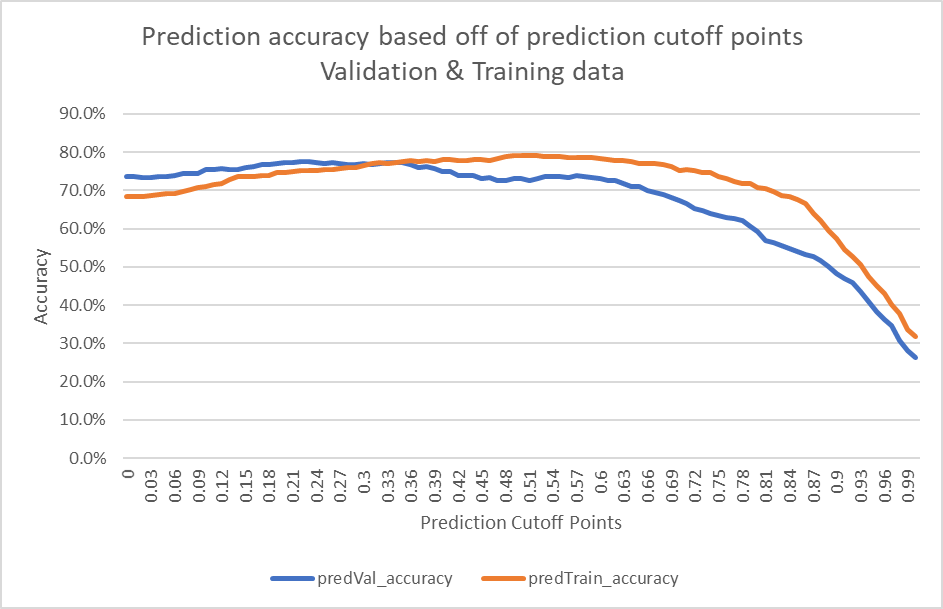
See Exhibit 3 below



1. What is your selected cut-off?

I would select cutoff = 0.5 for training data because accuracy is the highest at 0.5 cutoff, despite the fact that the predicted NPV might not be the highest.

\*\*\*But if accuracy is not a part of the consideration criteria and that total NPV output is the sole consideration criteria, I would select cutoff = 0.74 for training data.



1. Repeat the cut-off analysis for the validation data. Include a table or a figure as an Exhibit.

See Exhibit 3 above for combined plotting and figure presented in part b above.

I would also select cutoff = 0.5 for validation data because the tradeoff of selecting this particular cutoff yielding highest accuracy on training data is tolerable (validation’s prediction is still relatively accurate) for validation data.

\*\*\*But if accuracy is not a part of the consideration criteria and that total NPV output is the sole consideration criteria, I would select cutoff = 0.6 for validation data.

1. Apply the cut-off you selected based on the training data to the validation data. What is the total profit on the validation data?  
   cutoff = 0.5

$ $42,535.00

1. Provide a single figure that shows the total NPV as a function of the cut-off for each of the training and the validation data.

cutoff = 0.5

1. $77,187.00 + $42,535.00 = $119,722.00
2. ***Comparison with linear regression***
3. Repeat our model development from our first live session (feel free to use the script provided). Rerun a variable selection model to find a "good model”.

lr1NPV <- lm(NPV ~ CHK\_ACCT+SAV\_ACCT+NUM\_CREDITS+DURATION+HISTORY+EMPLOYMENT+OWN\_RES+REAL\_ESTATE+TYPE, data=Train)

Include the model (the variables and the corresponding regression coefficients) as an Exhibit.

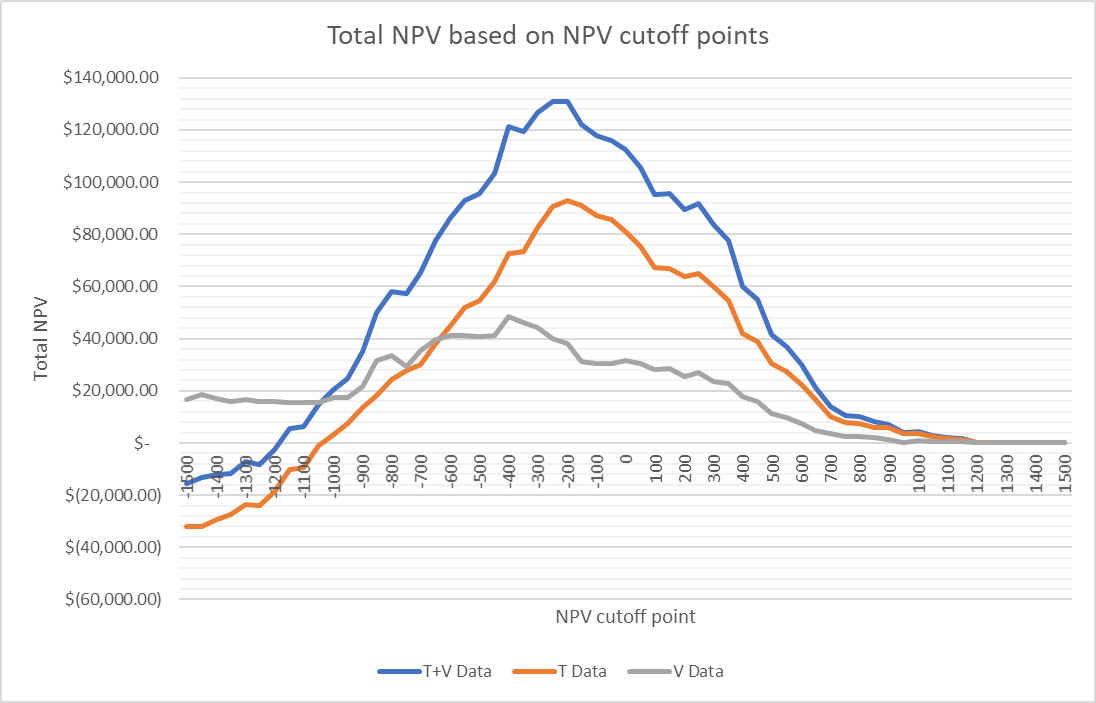
See Exhibit 4 below

Text, table

Description automatically generated with medium confidence

1. Either in R (using a loop) or in Excel (using a data-table) calculate the total NPV as a function of the NPV cut-off for extending credit on the training data (note that now your cut-off is in $ you will need to investigate what is a good cut-off, for example -$50 or $50, or something else). Select the best cut-off.   
     
   Include the table as an Exhibit.

See Exhibit 5 below



1. What is your selected cut-off?

NPV cutoff = -200

1. Repeat the analysis for the validation data and include it as an Exhibit.

See Exhibit 5 for combined plotting

1. Apply the cut-off you found using the training data to the validation data. What is the total profit on the validation data?

$38,076.00

1. Provide a single figure that shows the total NPV as a function of the cut-off for both the training data and the validation data.  
   $92,978.00 + $38,076.00 = $131,054.00
2. ***Model comparison***
3. Compare the performance of the logistic regression model and the linear regression model. How does the total NPV compare for the two models? Which model would you select as the foundation of a decision support system and why?

Log model will reflect the magnitude of changes in NPV as a result of % changes in variables, whereas linear model will only reflect numerical changes in NPV as a result of numerical changes in variable. I think both models are important because they serve different purposes. But for building a foundation of a decision support system, I will use linear regression because it will help identify an initial scope of factors that need to be reviewed deeper.