CWRU DSCI353-353M-453: Class 03a Predictive Analytics

Profs: R. H. French, L. S. Bruckman, P. Leu, K. Davis, S. Cirlos

TAs: W. Oltjen, K. Hernandez, M. Li, M. Li, D. Colvin

31 January, 2023

Contents

3.1.2.1 Class Readings, Assignments, Syllabus Topics

3.1.2.1.1 Reading, Lab Exercises, SemProjects

- Readings:
  - For today: DL03, ISLR3
  - For next class: DL04, DL05
- Laboratory Exercises:
  - LE1 is **Due Saturday at Midnight**
  - LE2 will be given out Thursday Feb. 2nd
    - **LE2 is due Tuesday Feb. 14th**
- Office Hours: (Class Canvas Calendar for Zoom Link)
  - Wednesdays @ 4:00 PM to 5:00 PM
  - Saturdays @ 3:00 PM to 4:00 PM
  - **Office Hours are on Zoom, and recorded**
- Semester Projects
  - **Office Hours for SemProjs: Mondays at 4pm on Zoom**
  - DSCI 453 Students Biweekly Updates Due

    - Update #1 is Due ** This Friday **
  - DSCI 453 Students
    - Next Report Out #1 is Due ** Feb. '17th **
  - All DSCI 353/353M/453, E1453/2453 Students:
    - Peer Grading of Report Out #1 is Due ** **
  - Exams
    - MidTerm: **Thursday March 9th**, in class or remote, 11:30 - 12:45 PM
    - Final: **Thursday May 4th**, 2023, 12:00PM - 3:00PM, Nord 356 or remote

3.1.2.1.2 Textbooks

- Introduction to R and Data Science
  - For R, Coding, Inferential Statistics
    - Peng: R Programming for Data Science
    - Peng: Exploratory Data Analysis with R

Textbooks for this class

- OIS = Diez, Barr, Çetinkaya-Runde: Open Intro Stat v4
- R4DS = Wickham, Grolemund: R for Data Science

Textbooks for DSCI353/353M/453, And in your Repo now

- ISLR2 = James, Witten, Hastie, Tibshirani: Intro to Statistical Learning with R, 2nd Ed.
- ESL = Trevor Hastie, Tibshirani, Friedman: Elements of Statistical Learning
- DLwR = Chollet, Allaire: Deep Learning with R, 2nd Ed.

Magazine Articles about Deep Learning

- DL1 to DL6 are "Deep Learning" articles in 3-readings/2-articles/

3.1.2.2 Syllabus

3.1.2.2.1 Tidyverse Cheatsheets, Functions and Reading Your Code

- Look at the Tidyverse Cheatsheet
  - **Tidyverse For Beginners Cheatsheet**
    - In the Git/20s-dsci353-353m-453-prof/3-readings/3-CheatSheets/ folder
  - **Data Wrangling with dplyr and tidyr Cheatsheet**

  Tidyverse Functions & Conventions

  - The pipe operator `%>%`
  - Use `dplyr::filter()` to subset data row-wise.
  - Use `dplyr::arrange()` to sort the observations in a data frame
  - Use `dplyr::mutate()` to update or create new columns of a data frame
  - Use `dplyr::summarize()` to turn many observations into a single data point
  - Use `dplyr::arrange()` to change the ordering of the rows of a data frame
  - Use `dplyr::select()` to choose variables from a tibble,
    - keeps only variables you mention
  - Use `dplyr::rename()` keeps all the variables and renames variables
    - rename(iris, petal_length = Petal.Length)
  - These can be combined using `dplyr::group_by()`
    - which lets you perform operations "by group".
  - The `%in%` matches conditions provided by a vector using the c() function
  - The **forcats** package has tidyverse functions
    - for factors (categorical variables)

Day:Date	Foundation	Practicum	Readings(optional)	Due(optional)
w01a:Tu:1/17/23	Markov Cluster	R, Rstudio IDE, Git		(LE0)
w01b:Th:1/19/23	Stat. Learning, Approach	Bash, Git, Class Repo	ISLR1,2 (R4DS-1-3)	
w02a:Tu:1/24/23	Lin. Regr. Bias-Var.	SemProjs; Regr. Ovrvw	ISLR3,(R4DS-4-6)	**(LE0:Due)** LE1
w02b:Th:1/26/23	Train/Test, Bias vs. Vari.	Tidyverse Review	DL01 DL02 (R4DS-7,8)	
w02Pr:Fr:1/27/23	**ADD DROP**	**DEADLINE**		**453 Update 1**
w03a:Tu:1/31/23	Logistic Regr. Classif	Pred. Analytics, Regr.	DL03,ISLR4	
w03b:Th:2/2/23	LDA/QDA	ggPlot2, Code Expect.	DL04, DL05	**LE1:Due**, LE2
w03:Sa:2/4/23				**LE1:Due**
w04a:Tu:2/7/23	Resample Cross-Valid.	Multilevel Mod.	ISLR5	
w04b:Th:2/9/23	Bootstrap	Mixed Effects		
w04Pr:Fr:2/10/23				**453 Update 2**
w05a:Tu:2/14/23	Subset Selec., Shrink.	Bootstrap	ISLR6 (R4DS9-16)	**LE2:Due**, LE3
w05b:Th:2/16/23	Mod. Selec. Dim. Red.	Clustering, ggplot2	DL06	
w05Pr:Fr:2/17/23				**453 Rep. Out 1**
w06a:Tu:2/21/23	Beyond Linear Modls	Feature Select., Caret	ISLR7, DL07	
w06b:Th:2/23/23	PCA, PCR, FA	Tidy Modeling	ISLR10(R4DS22-25)	**LE3:Due**, LE4
w06Pr:Fr:2/24/23				**453 Update 3**
w07a:Tu:2/28/23	Dec. Trees, Rand. Forest.	Machine Learning	ISLR8, DL08,09	
w07b:Th:3/2/23	MidTerm Review, SVM	SVM, SVR, ROC	ISLR9 (R4DS26-30)	**Peer Review 1**
w08a:Tu:3/7/23	R-Keras/TensorFlow2	Perceptron, Neural Nets	ISLR10	
w08b:Th:3/9/23	**MIDTERM EXAM**		DL10,11	**LE4:Due** LE5
w08Pr:Fr:3/10/23				**453 Update 4**
Tu:3/14/23	**SPRING**	**BREAK**	ISLR10	
Th:3/16/23	**SPRING**	**BREAK**	DL12,13	
w09a:Tu:3/21/23	Deep Learning	TF2 Keras Intro	Pocket Perceptron	ISLR10, DLR3
w09b:Th:3/23/23	Computer Vision, CNN	CNN w/TF2, Overfit	DLR4	
w09Pr:Fr:3/24/23				**453 Rep. Out 2**
w10a:Tu:3/28/23	Deep Learn Intro	NN Types	DLR5	
w10b:Th:3/30/23	DL CNN,RNN ImageNet	NN Types, CNN wTF2	Hinton ImageNet	
w10Pr:Fr:3/31/23				**453 Upd.5 & PrRev 2**
Sa:4/1/23				**LE5:Due** LE6
w11a:Tu:4/4/23	Fitting NNs	AUC,Prec,Recall Fruit		
w11b:Th:4/6/23	NLP, Graphs & ML		LeCun DL Rev. 2015	
w12a:Tu:4/11/23	Graphs & ML	NLP with sequences	DLR6	
w12b:Th:4/13/23	NLP w attention	Graph Repr Proc Wrkflw		**LE6:Due** LE7
w13a:Tu:4/18/23	DL Frameworks	Explaining DL w Lime		
w13b:Th:4/20/23	Linux Distros XGBoost	Explain Preds	Deep Dream	
w13Pr:Fr:4/21/23				**453 Rep. Out 3 Due**
w14a:Tu:4/25/23	Tranformers			
w14b:Th:4/27/23	Final Exam Review	Torch NN & DeepLearn		**LE7:Due**
w14Pr:Fr:4/28/23				**Peer Rev 3 Due**
	FINAL EXAM	**Th. 5/4/23, 12-3pm**	Nord 356 & Zoom	
	453 Final PDF Report	**Fr. 4/29, 11:59pm**		

Figure 1: Modeling, Prediction and Machine Learning Syllabus

  - The **readr** package has tidyverse functions
    * to read_…, melt_… col_…, parse_… data and objects

Reading Your Code: Whenever you see

- The assignment operator `<-`, think **"gets"**
- The pipe operator, `%>%`, think **"then"**

3.1.2.3 What is Statistical (and Machine) Learning

- We will go far beyond classical inferential statistical methods,
  - such as linear regression.

As computing power has increased over the last 20 years

- many new, highly computational, regression, or "Statistical Learning",
- methods have been developed.

In particular the last decade has seen a significant expansion

- of the number of possible approaches.

Here we will provide a very applied overview to such modern non-linear methods as

- Generalized Additive Models,
- Decision (or Regression) Trees,
- Boosting,
- Bagging and
- Support Vector Machines

As well as more classical linear approaches such as

- Logistic Regression,
- Linear Discriminant Analysis,
- K-Means Clustering and Nearest Neighbors.

At the end of this course you should have

- a basic understanding of how all of these methods work
- and be able to apply them in real data analyses.

With the explosion of "Big Data" problems,

- statistical learning has become a very hot field in many areas.

People with statistical learning skills are in high demand!

To this end, approximately one third of the class time

- is dedicated to in lab exercises
- where the students will work through
- the latest methods we have covered,
- on their Open Data Science VDI.

These labs will ensure that every student

- has a full understanding of the
- practical and theoretical, aspects of each method.

3.1.2.4 Supervised and Unsupervised Learning

- Two broad families of algorithms will be covered:
  - Unsupervised learning algorithms

- Supervised learning algorithms

3.1.2.4.1 Unsupervised learning

- In unsupervised learning,
  - the algorithm will seek to find the structure that organizes unlabeled data.

3.1.2.4.2 Supervised learning

- In supervised learning,
  - we know the class or the level of some observations of a given target attribute.

3.1.2.5 Classification and Regresssion Problems

- There are basically two types of problems that predictive modeling deals with:
  - Classification problems
  - Regression problems

3.1.2.5.1 Classification

- In some cases,
  - we want to predict which group an observation is part of.

Here, we are dealing with a quality of the observation.

3.1.2.5.2 Regression

- In other cases,
  - we want to predict an observation's level on an attribute.

Here, we are dealing with a quantity, and this is a regression problem.

3.1.2.6 The critical role of domain knowledge

- in modeling and prediction

Domain knowledge informs and is informed by data understanding.

- The understanding of the data
  - then informs how the data has to be prepared.

The next step is data modeling,

- which can also lead to further data preparation.

Data models have to be evaluated,

- and this evaluation can be informed by field knowledge,
  - which is also updated through the data mining process.

Finally,

- if the evaluation is satisfactory,
  - the models are deployed for prediction.

3.1.2.7 Caveat: For Predictive Analytics

- Of course, predictions are not always accurate,
  - and some have written about the caveats of data science.

What do you think about the relationship between

- the attributes titled Predictor and Outcome on the following plot?

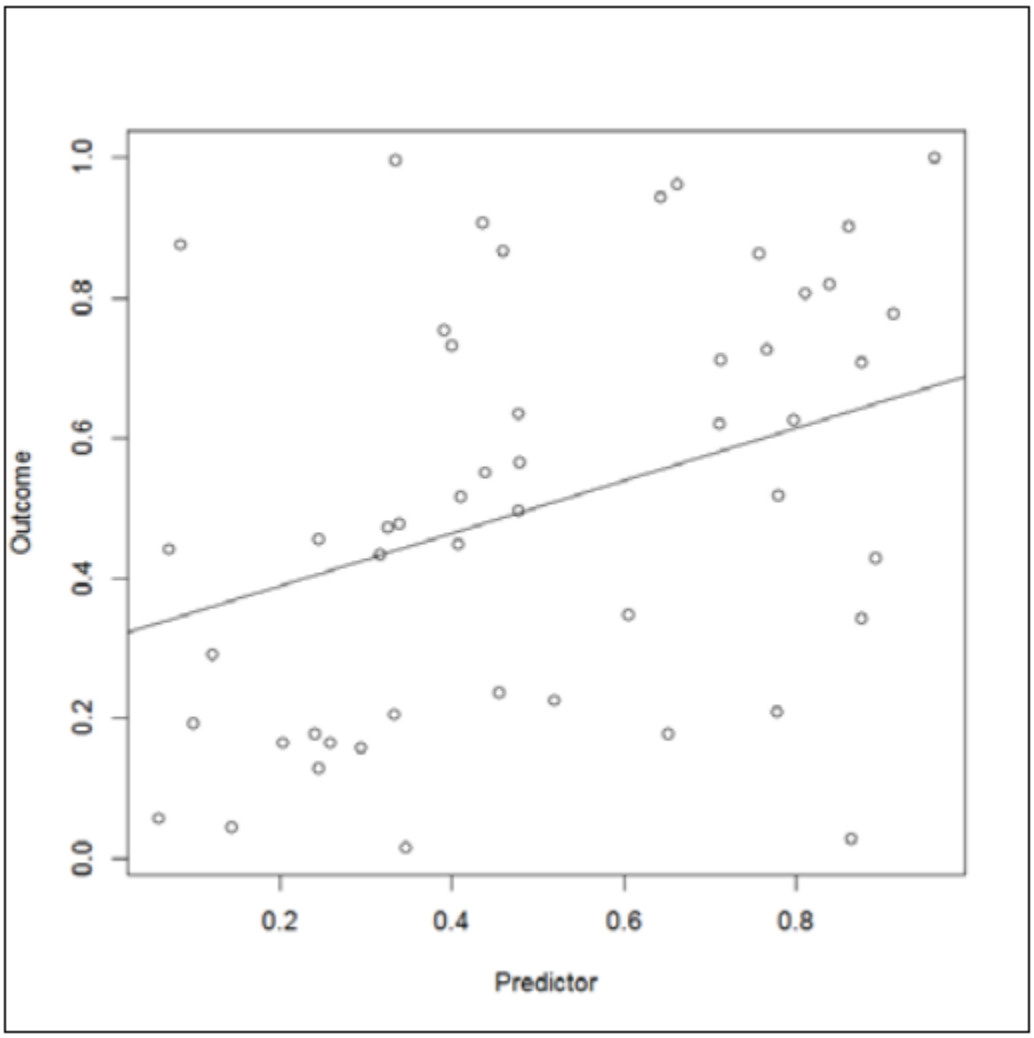


Figure 2: Relationship between Predictor & Outcome

It seems like there is a relationship between the two.

- For the statistically inclined,
  - I tested its significance:
    - r = 0.4195, p = .0024.
- The value p is the probability of obtaining a relationship of this strength or stronger
  - if there is actually no relationship between the attributes.
- (This is the p-value of hypothesis testing, if $p<0.05$
  - typically we assert we can reject the null hypothesis)
- We could conclude that the relationship between these variables
  - in the population they come from is quite reliable,

- **right?**

3.1.2.7.1 No: Lets think about this

- Believe it or not,
  - the population these observations come from
    - is that of randomly generated numbers.
  - We generated a data frame of 50 columns
    - of 50 randomly generated numbers.
  - We then examined all the correlations (manually)
    - and generated a scatterplot of the two attributes
    - with the largest correlation we found.

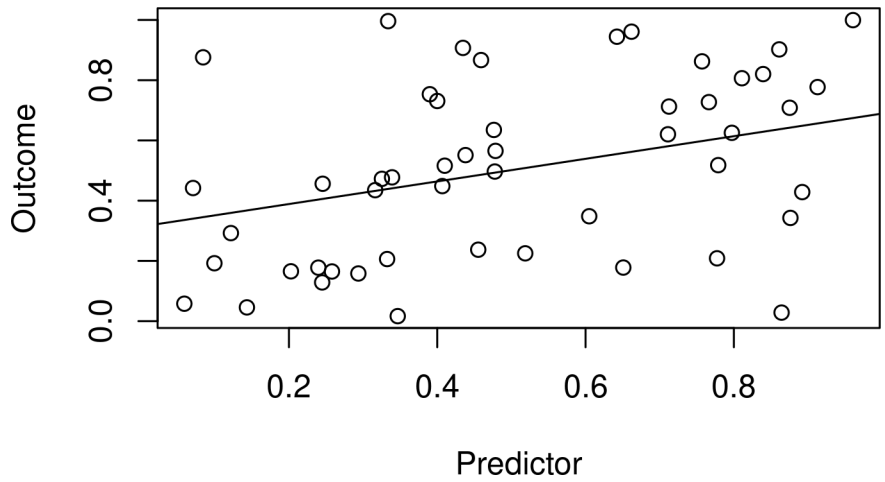
3.1.2.7.2 The code is provided here,

- We'll use runif()
  - help(runif)
    - The Uniform Distribution
    - Description

These functions provide information about the uniform distribution

- on the interval from min to max.
  - dunif gives the density,
  - punif gives the distribution function
  - qunif gives the quantile function and
  - runif generates random deviates.

```
set.seed(1)
DF <- data.frame(matrix(nrow = 50, ncol = 50))
for (i in 1:50)
  DF[, i] <- runif(50)
plot(DF[[2]], DF[[16]], xlab = "Predictor", ylab = "Outcome")
abline(lm(DF[[2]] ~ DF[[16]]))
```




```
cor.test(DF[[2]], DF[[16]])
```

```
##
##  Pearson's product-moment correlation
##
## data:  DF[[2]] and DF[[16]]
```

```
## t = 3.2023, df = 48, p-value = 0.002421
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1598919 0.6249314
## sample estimates:
##       cor
## 0.4195666
```

In case you want to check it yourself

- line 1 sets the seed so that you find the same results as we did,
- line 2 generates to the data frame,
- line 3 fills it with random numbers, column by column,
- line 4 generates the scatterplot,
- line 5 fits the regression line, and
- line 6 tests the significance of the correlation:

Normally we reject the null with a p-value of <0.05

- i.e. we'll be wrong 5% of the time
  - in a set of 20 trials

Here we did 50 trials

- And cherry picked the best correlation
  - But its all randomly generated numbers
  - There is no predictive or causal relationship
- And we'd only recognize this if we consider
  - That our p-value

3.1.2.7.3 Bonferroni Correction for multiple comparisons

- How could this relationship happen given that the odds were 2.4 in 1000 ?
  - Well, think of it;
    - we correlated all 50 attributes 2 x 2,
    - which resulted in 2,450 tests
    - (not considering the correlation of each attribute with itself).
  - Such spurious correlation was quite expectable.

The usual p-value threshold below which

- we consider a relationship significant is p = 0.05.

This means that we expect to be wrong once in 20 times.

- You would be right to suspect that there are other significant correlations
  - in the generated data frame (there should be approximately 125 of them in total).
- This is the reason why we should always correct the number of tests.
- In our example,
  - as we performed 2,450 tests,
  - our threshold for significance
  - should be 0.0000204 (0.05 / 2450).
- This is called the Bonferroni correction.

3.1.2.8 Overfitting: The need for Trainging and Testing Datasets

- Spurious correlations are always a possibility in data analysis
  - and this should be kept in mind at all times.

A related concept is that of overfitting.

- Overfitting happens, for instance,
  - when a weak classifier bases its prediction on the noise in data.
- We will discuss overfitting when discussing
  - Training datasets for fit a model to
  - Testing datasets for evaluating the goodness of fit
    - when using various types of cross-validation
  - And when evaluating $Predictive, Adjusted, R^2$

3.1.2.9 Citations

1. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2014. http://www.R-project.org/.
2. G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: 2nd Ed., with Applications in R, 2nd ed. 2021 edition. New York: Springer, 2021.
3. Diez, David M., Christopher D. Barr, and Mine Çetinkaya-Rundel. OpenIntro Statistics: Third Edition. 3 edition. S.l.: OpenIntro, Inc., 2015.
4. Al Sharif, IOM 530 – Applied Modern Statistical Learning Methods, USC
5. Mayor, Eric. Learning Predictive Analytics with R. Packt Publishing - ebooks Account, 2015.