Big Data, Machine Lerning and Econometrics

Un tour por el mundo del Machine Learning

Gabriel Martos
gmartos@utdt.edu

Agenda

Objetivo: Introducir los modelos de aprendizaje supervisado y aprender a entrenar los mismos en R (casos estudio: Economía, Finanzas y Negocios).

A– **Introducción**:

- Contextualización del curso e introducción a R.
- Breve intro a métodos numéricos de optimización en ML.
- Minimización del riesgo empírico y descomposición del riesgo predictivo.

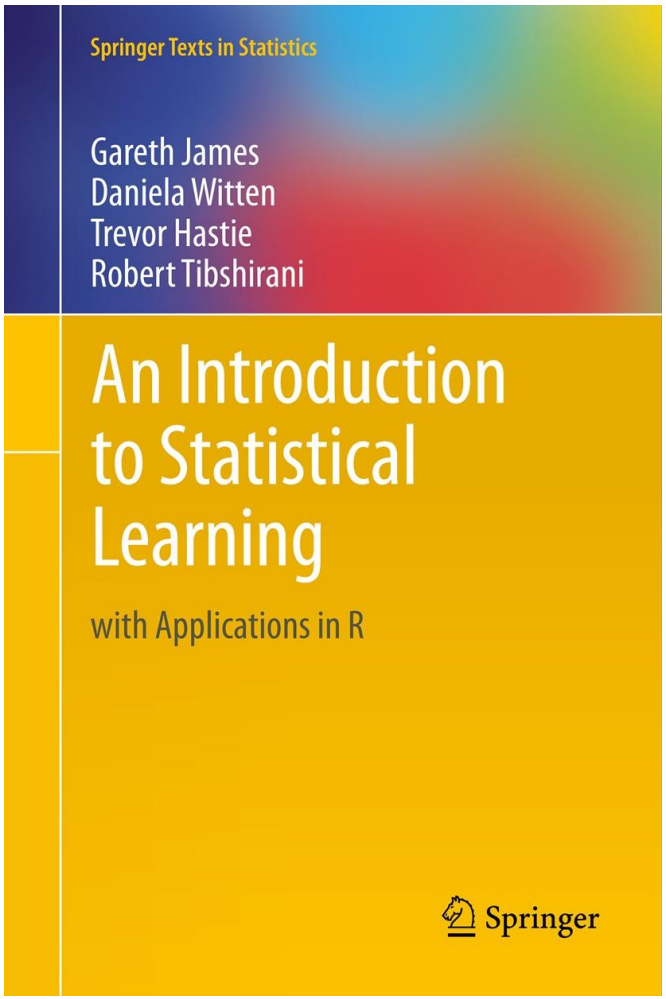
B– **Clasificación y Regresión**:

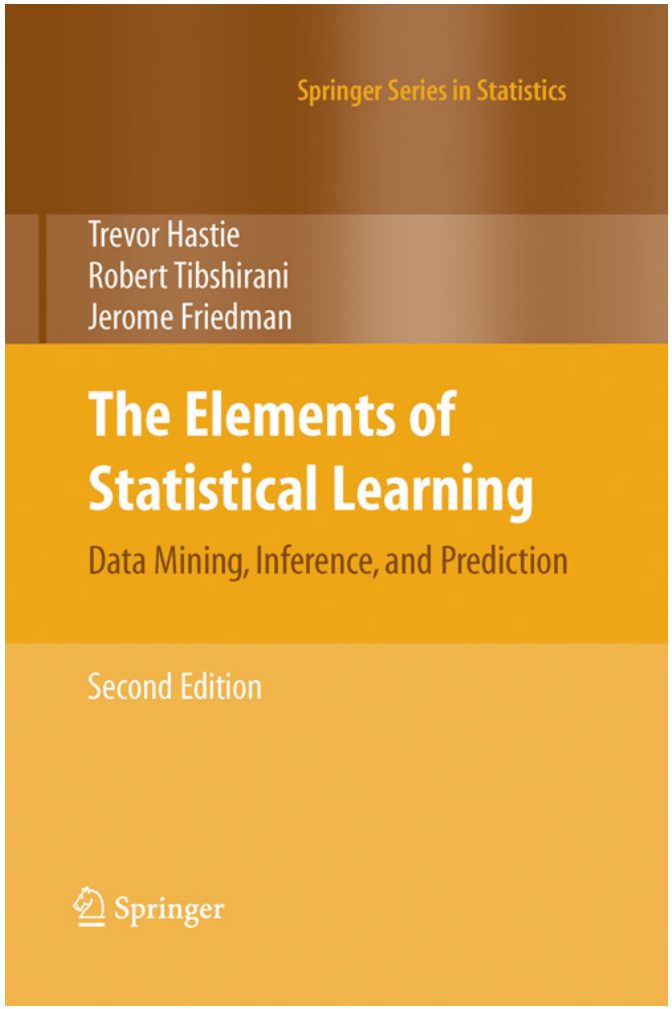
- Validación cruzada y modelo de K-vecinos.
- Modelos Aditivos, Selección y Regularización (Ridge, Lasso, ENets).
- Árboles y ensamble de modelos (Bagging, Random Forest, Boosting).
- Métodos de Kernel y Máquinas de Vector Soporte (SVM).
- Introducción a la modelización con Redes Neuronales.

C– **Modelos predictivos para series temporales**:

- Singular Spectrum Analysis y Procesos Gaussianos para modelar y predecir series temporales (VC para modelos de series).

Bibliografía





(Indicada al principio de cada tema)

Evaluación

- Trabajo práctico individual.
- Fecha límite de entrega: Febrero del 2025.
- Aspectos a evaluar:
  - Motivación y contextualización del problema predictivo.
  - Revisión bibliográfica relevante sobre el problema.
  - Originalidad de la presentación.
  - Contrastación de métodos y comparativa con modelos econométricos.
  - Claridad y profundidad en la presentación de la metodología que utilizaste para resolver el problema. Claridad en la exposición de los principales resultados obtenidos (ver el TP de referencia).

Algunos repositorios que puedes comenzar a explorar en búsqueda de datos / problemas de aprendizaje supervisado:

- AcademicTorrents.
- Open Data (PLOS).
- Datasetsearch.
- Kaggle.
- UCI Machine Learning repository.
- OpenML.
- DrivenData.
- CrowdAnalytix.
- ...

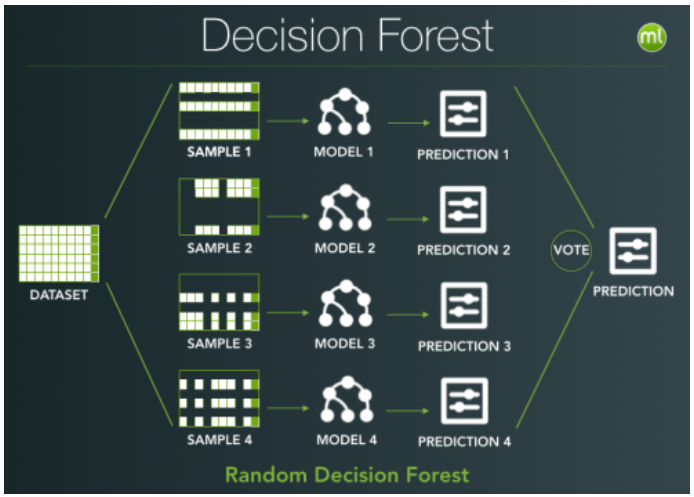
Agenda

Fact: Las TIC's producen una cantidad ingente de DATOS.

- *Bigdata*: Conjunto de datos grande y complejo.
- ¿Cuanto "big" es el Bigdata?:
  - Empresas Tecnológicas (Google) almacena las consultas en su buscador de manera diaria (BD del orden de peta-bytes–10^6GB– diarios!)
  - Bancos: Registran "eventos" diarios de clientes por distintos canales (cajero, homebanking, canales móvil, call-center, etc)... solo estas transacciones representan del orden de varios TB (10^3GB) por día.
  - Organismos públicos registran pagos de impuestos, trámites, consultas y reclamaciones, *vacunaciones y desplazamiento*, seguridad, etc.
- Necesitamos adaptar las **herramientas de análisis** de datos tradicionales para transformar los Bigdatos en información útil.
  - La maldición de la dimensión ($p \gg 0$).
  - *Fatdata* ó $p > n$: Más regresores que observaciones.
  - NO linealidad en las relaciones, el problema de la multicausalidad, etc.
- ML como complemento del toolkit "estadísticonométrico".

Aplicaciones en Finanzas

- Instituciones financieras recopilan información de sus clientes.
- Modelos de "Score Crediticio".
- La rentabilidad del negocio depende de manera crucial de la medición acertada del riesgo.

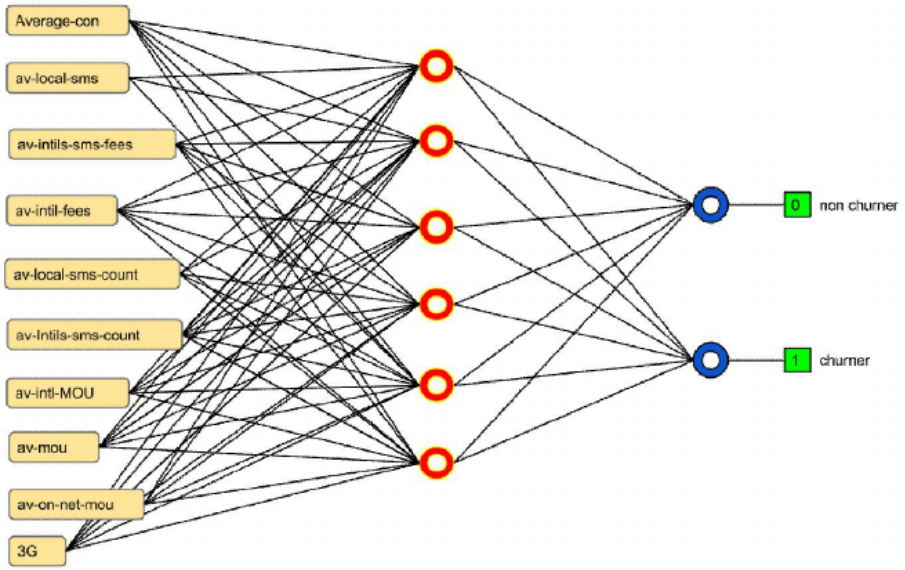


- Modelizaremos el riesgo crediticio con: Bagging, Random Forest, y Boosting.
- Capturamos efectos no lineales.
- Reducimos variabilidad ensamblando.

Aplicaciones en Marketing (churn)

- Modelizamos la probabilidad de fuga.
- Para las firmas el output de estos modelos es esencial para planificar su oferta de bienes y servicios; como así también para diseñar estrategias de retención de clientes (promociones).

Redes neuronales para modelizar la fuga de clientes de telefonía.

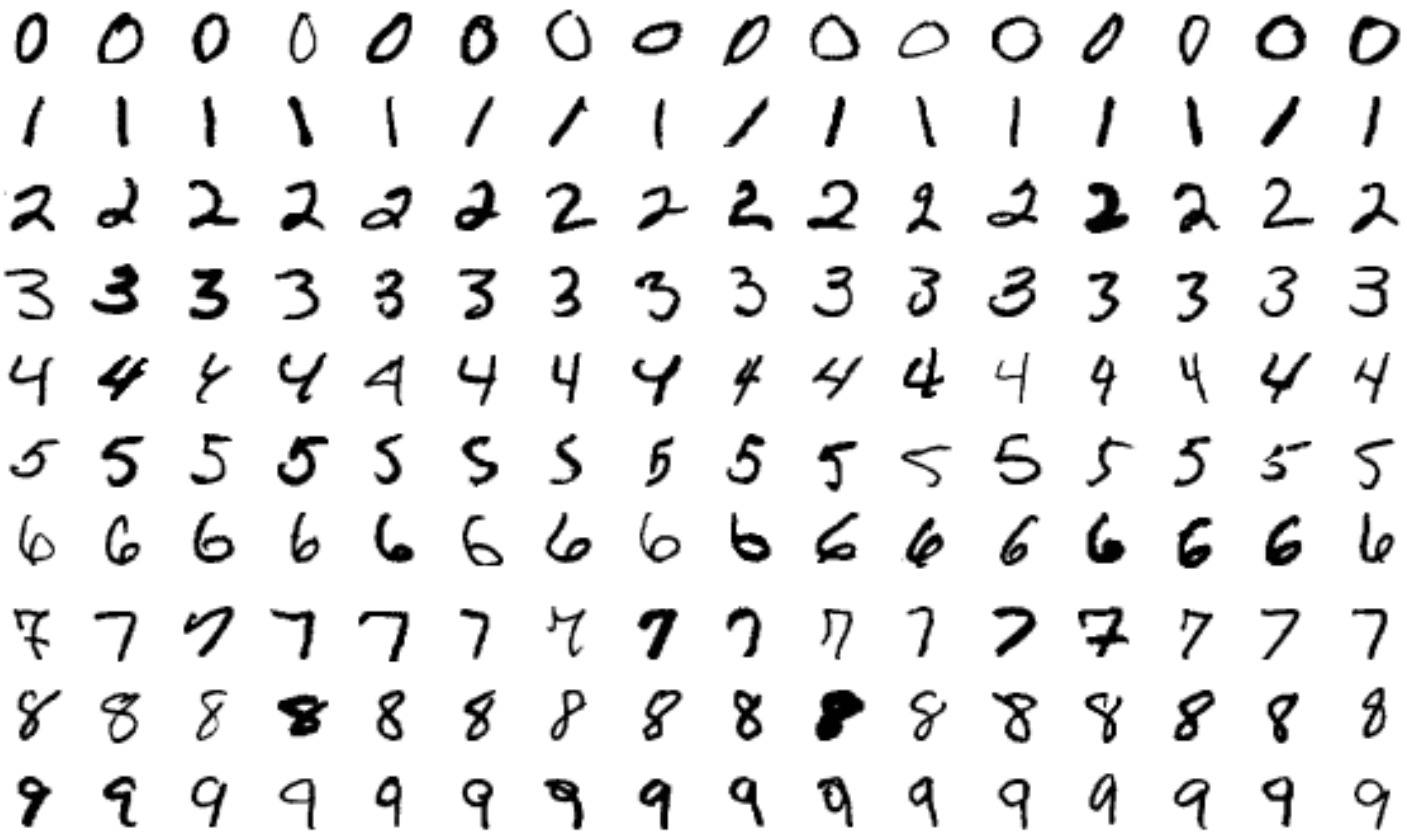


Aplicaciones en Transporte y Energía

Objetivo: Modelizar la demanda para gestionar una producción y tarifación adecuada a lo largo del tiempo (cada hora, día, semana, mes, año, etc).

- Urban mobility lab (MIT): https://mobility.mit.edu/machine-learning

Aplicaciones en Ciencia y Tecnología



Reconocimiento de imágenes para el diagnóstico medico, la seguridad informática, la automatización de procesos productivos, etc.

Otras aplicaciones en negocios, economía

- Customer Segmentation: Detectar nichos de clientes con características específicas para orientar acciones comerciales.
- Pricing: Modelizar la demanda de productos para gestionar una producción y tarifación adecuada a lo largo del tiempo.
- Recommendation Engines: Predecir de un conjunto muy grande de datos las preferencias/gusto de los clientes:

  `https://www.netflixprize.com/`

- ...

Agenda

... my advice to grads (in Economics) is "go to the computer science department and take a course in machine learning."

H. Varian (Chief economist at Google)

- Varian, H. R. (2014): "*Big data: New tricks for econometrics*". Journal of Economic Perspectives, 28(2), 3-28.

$\widehat{\beta}$ vs $\widehat{y}$

Estadística y Econometría están orientadas a:

- Hacer *Inferencia* apoyándose en **modelos probabilísticos**.
- Los modelos sirven para describir y cuantificar como se relación las variables ($\widehat{\beta}$). Nos interesan los parámetros del modelo per–se.
- En menor medida a hacer predicciones con el modelo.

$\widehat{\beta}$ vs $\widehat{y}$

Estadística y Econometría están orientadas a:

- Hacer *Inferencia* apoyándose en **modelos probabilísticos**.
- Los modelos sirven para describir y cuantificar como se relación las variables ($\widehat{\beta}$). Nos interesan los parámetros del modelo per–se.
- En menor medida a hacer predicciones con el modelo.

Machine Learning orientado a:

- **Predicción** de las variables de interés ($\widehat{y}$). Los parámetros de los modelos interesan poco (son sólo un medio para obtener $\widehat{y}$).
- Filosofía del "*dejar que los datos hablen por si mismos*" (modelos no–paramétricos y con escasa/nula estructura probabilística).

¿Qué pueden aprender los econometristas del ML?

- Overfitting (explicitar el trade–off sesgo y varianza de cada modelo).
- Técnicas de Validación Cruzada:
  - Aprender valores adecuados de ciertos (hiper/meta)parámetros.
  - Estimar la performance out–of–sample.
- Regularización: Ridge, Lasso "and friends" (HD data).
- Ensamble de modelos (otro mecanismo de regularización).
- Modelos no lineales: Árboles, SVM, Redes Neuronales, etc.
- Herramientas computacionales avanzadas y escalabilidad de los métodos de análisis al contexto del Bigdata.
- Herramientas para explorar nuevos dominios: Análisis semántico, de imágenes, de redes sociales, análisis de sentimientos, etc.

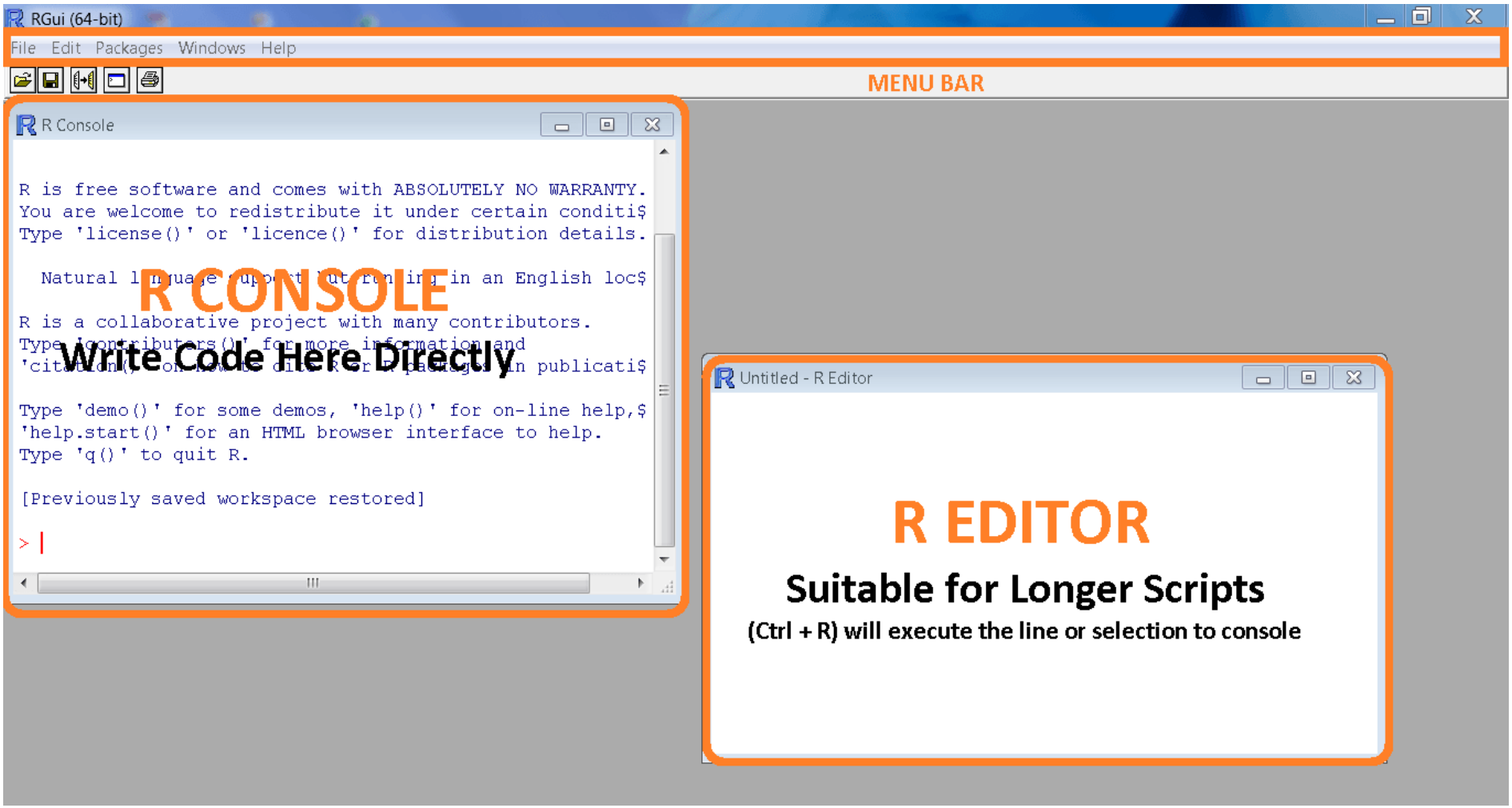
Algunos economistas que trabajan con modelos de ML:

- Susan Athey (Stanford Graduate School of Business).
- Guido Imbens (Stanford Graduate School of Business).
- Jann Spiess (Harvard).
- Andrew Gelman (Universidad de Columbia).
- Interesante para seguir leyendo:
  - *Why a leading economist is embracing machine learning*.
  - *Machine Learning Methods Economists Should Know About*

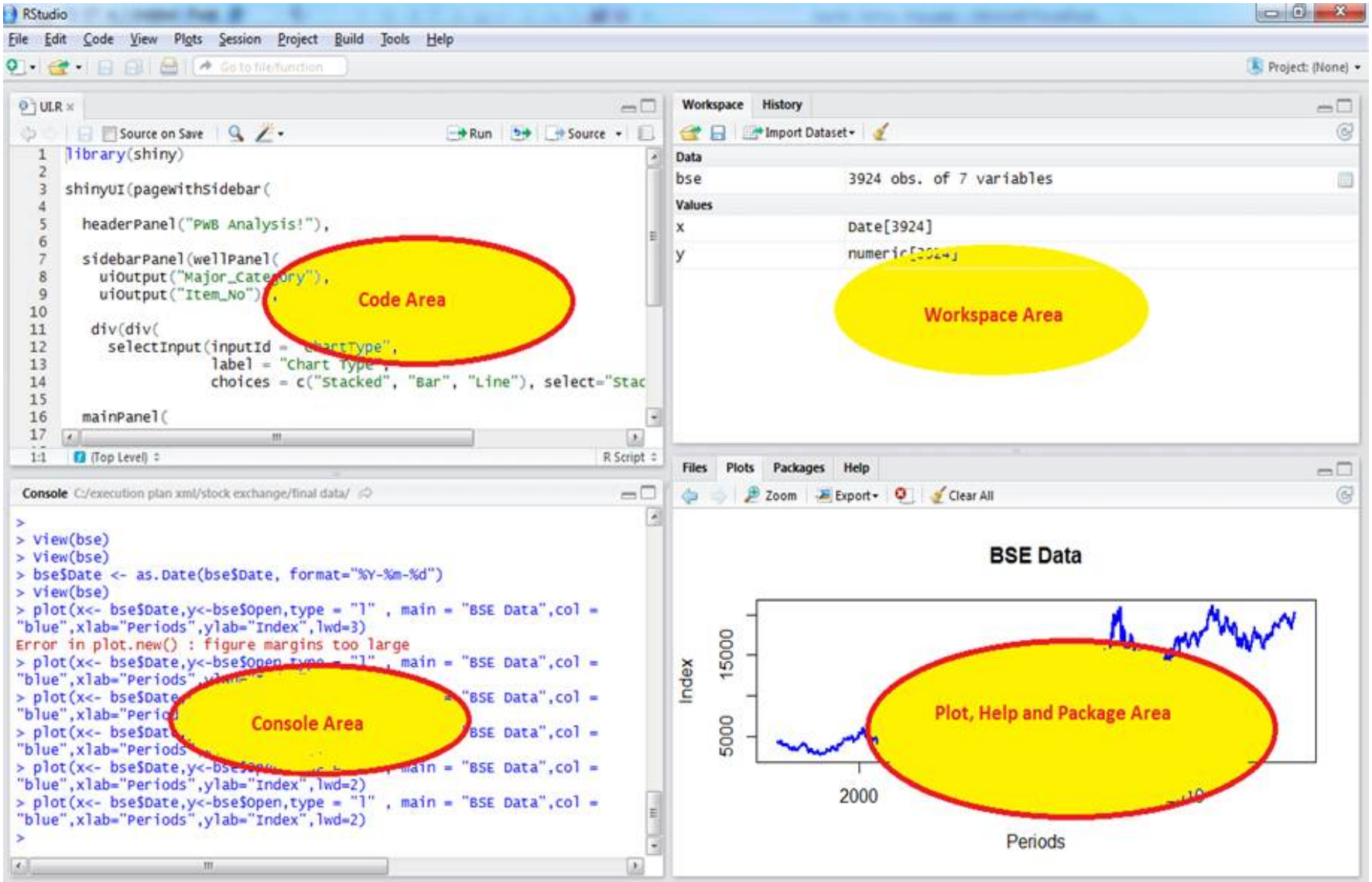
Agenda

- En el curso vamos a utilizar un enfoque "metodológico–práctico" para presentar cada uno de los modelos y temas.
- Para correr estos modelos, vamos a utilizar `R`.
- `R` es un lenguaje GRATUITO de programación ampliamente utilizado (al igual que Python) en Ciencia de Datos. Links para descargarlo:
  - Baja primero `R` de su web oficial.
  - Aconsejable descargar también `RStudio` (yo lo voy a utilizar en clase).
- Vamos a invertir el resto de la clase en aprender las instrucciones y comandos básicos para:
  - Crear, cargar y manipular datos (esto último es lo más trabajoso).
  - Utilizar estructuras de control que luego serán de muchísima utilidad (por ejemplo para implementar validación cruzada con un modelo).
- Una vez instales ambos programas verás lo siguiente...

R


RGui (64-bit)
File Edit Packages Windows Help
MENU BAR
R Console
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditi$
Type 'license()' or 'licence()' for distribution details.
Natural l
in an English loc$
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publicati$
Type 'demo()' for some demos, 'help()' for on-line help,$
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
[Previously saved workspace restored]
>
R CONSOLE
Write Code Here Directly
Untitled - R Editor
R EDITOR
Suitable for Longer Scripts
(Ctrl + R) will execute the line or selection to console

RStudio


RStudio
File Edit Code View Plots Session Project Build Tools Help
Project: (None)
ULR
Source on Save
Run
Source
1 library(shiny)
3 shinyUI(pageWithSidebar(
5 headerPanel("PWB Analysis!"),
7 sidebarPanel(wellPanel(
8 uiOutput("Major_Category"),
9 uiOutput("Item_No")),
11 div(div(
12 selectInput(inputId = "chartType",
13 label = "Chart Type",
14 choices = c("Stacked", "Bar", "Line"), select="Stac
16 mainPanel(
Code Area
Workspace
History
Import Dataset
Data
bse 3924 obs. of 7 variables
Values
x Date[3924]
y
Workspace Area
Console C:/execution plan xml/stock exchange/final data/
> View(bse)
> View(bse)
> bse$Date <- as.Date(bse$Date, format="%Y-%m-%d")
> View(bse)
> plot(x<- bse$Date,y<-bse$Open,type = "l" , main = "BSE Data",col =
"blue",xlab="Periods",ylab="Index",lwd=3)
Error in plot.new() : figure margins too large
> plot(x<- bse$Date,y<-bse$Open,type = "l" , main = "BSE Data",col =
"blue",xlab="Periods",ylab="Index",lwd=2)
> plot(x<- bse$Date,y<-bse$Open,type = "l" , main = "BSE Data",col =
"blue",xlab="Periods",ylab="Index",lwd=2)
>
Console Area
Files
Plots
Packages
Help
Zoom
Export
Clear All
BSE Data
Index
5000
15000
2000
Periods
Plot, Help and Package Area