Predicting Credit Card Fraud with R

Background

- **Instructor:** John Garcia
  - I teach advanced data analytics at the University of North Texas

- **Project:** We will use R to fit three classification model to a highly imbalanced dataset:

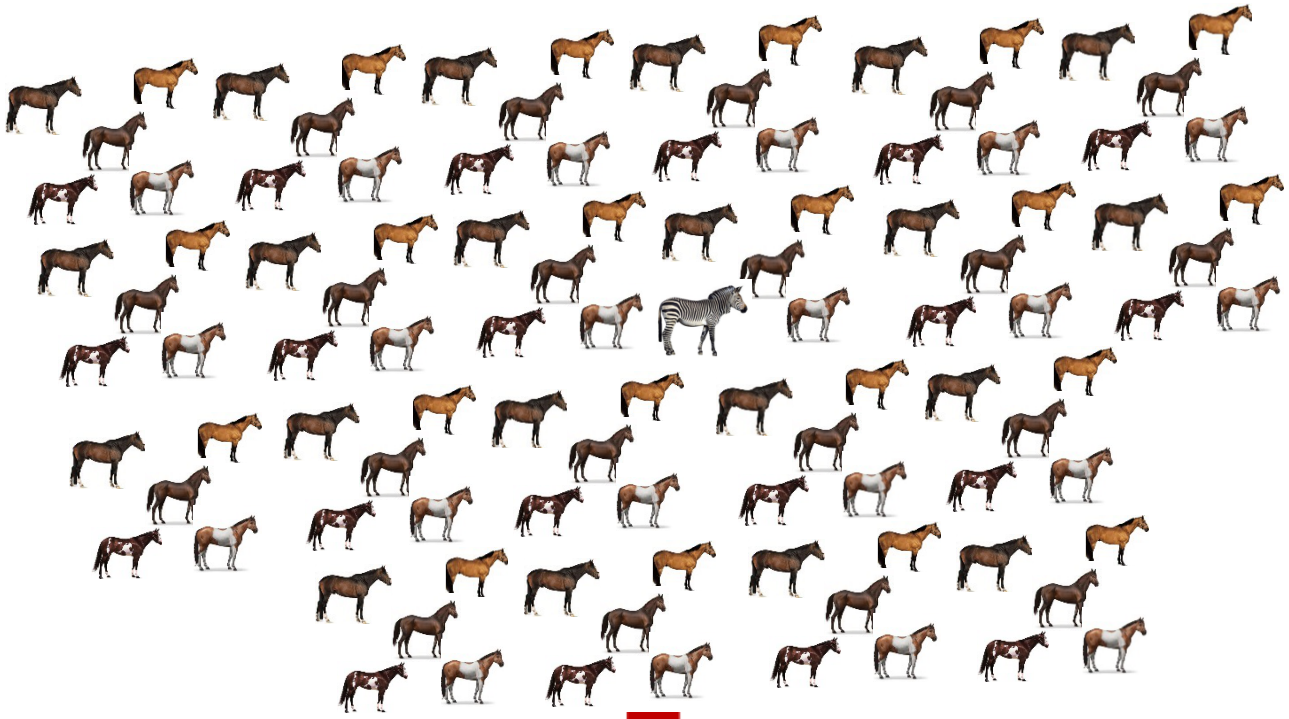
1	**Decision Tree:** Uses a tree-like model of decisions to arrive at a classification prediction.
2	**Naïve Bayes classifier:** Uses Bayes' theorem to use probability to arrive at a classification prediction.
3	**Linear Discriminant Analysis**: Finds a linear combination of features that is used to separate the classes.

We will use R to…..

Explore	Perform Exploratory Data Analysis
Split Data	Randomly split our data into a training / test set
Balance	Apply synthetic balancing techniques to balance the highly imbalanced training dataset
Train	Train our three classification models.
Evaluate	Evaluate their performance on the test data set.

Why Balance our Dataset?

Dataset: 99 Horses…..1 Zebra



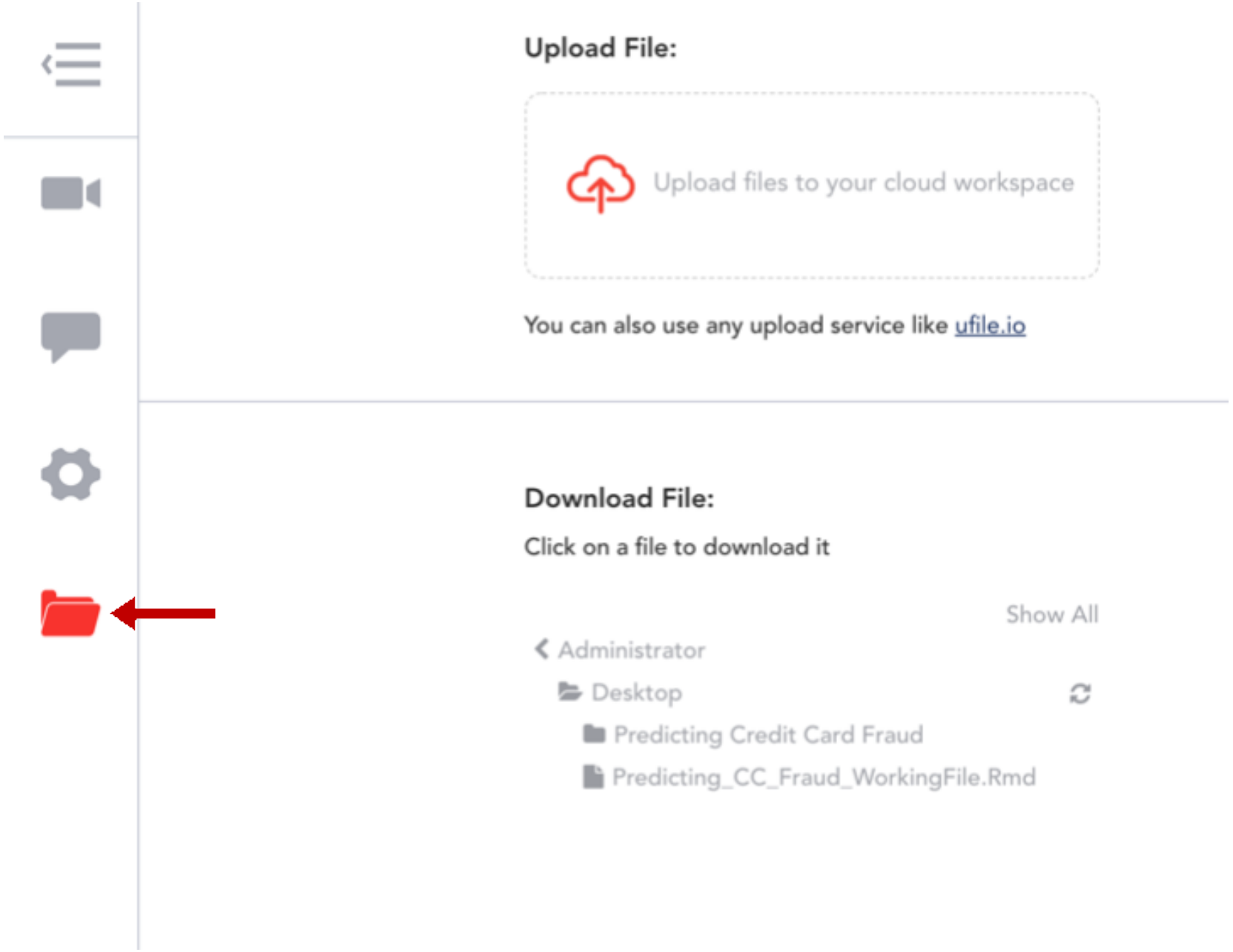
Prediction Algorithm
All Horses

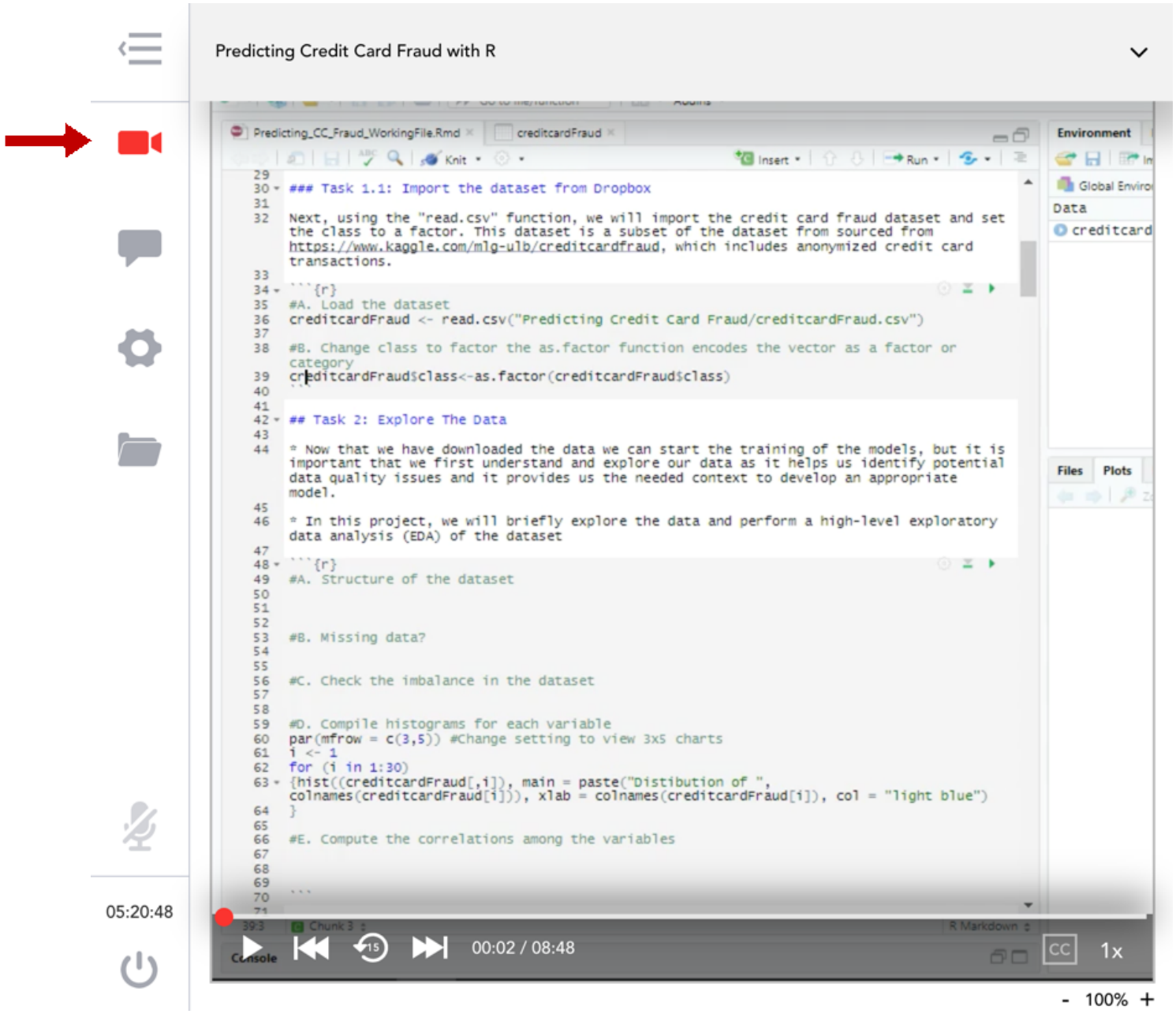
100 Horses

99% Accuracy

Predicting Credit Card Fraud with R

1	**Introduction**	07:49
2	Exploratory Data Analysis	08:48
3	Create Training / Test Datasets	06:29
4	Generate Synthetic Samples	08:36
5	Train Classifiers on Original Imbalanced Dataset	09:17
6	Train Classifiers on SMOTE Balanced Dataset	06:07
7	Train Classifiers on ADASYN Balanced Dataset	04:45
8	Train Classifiers on DB SMOTE Balanced Dataset	05:56
9	Compare All The Trained Models	05:59


Upload File:
Upload files to your cloud workspace
You can also use any upload service like ufile.io
Download File:
Click on a file to download it
Show All
Administrator
Desktop
Predicting Credit Card Fraud
Predicting_CC_Fraud_WorkingFile.Rmd


Predicting Credit Card Fraud with R
Predicting_CC_Fraud_WorkingFile.Rmd
creditcardFraud
Knit
Insert
Run
Environment
Global Enviro
Data
creditcard
Files
Plots
Task 1.1: Import the dataset from Dropbox
Next, using the "read.csv" function, we will import the credit card fraud dataset and set the class to a factor. This dataset is a subset of the dataset from sourced from https://www.kaggle.com/mlg-ulb/creditcardfraud, which includes anonymized credit card transactions.
```{r}
#A. Load the dataset
creditcardFraud <- read.csv("Predicting Credit Card Fraud/creditcardFraud.csv")
#B. Change class to factor the as.factor function encodes the vector as a factor or category
creditcardFraud$class<-as.factor(creditcardFraud$class)
```
Task 2: Explore The Data
* Now that we have downloaded the data we can start the training of the models, but it is important that we first understand and explore our data as it helps us identify potential data quality issues and it provides us the needed context to develop an appropriate model.
* In this project, we will briefly explore the data and perform a high-level exploratory data analysis (EDA) of the dataset
```{r}
#A. Structure of the dataset
#B. Missing data?
#C. Check the imbalance in the dataset
#D. Compile histograms for each variable
par(mfrow = c(3,5)) #Change setting to view 3x5 charts
i <- 1
for (i in 1:30)
{hist((creditcardFraud[,i]), main = paste("Distibution of ", colnames(creditcardFraud[i])), xlab = colnames(creditcardFraud[i]), col = "light blue")
}
#E. Compute the correlations among the variables
```
Chunk 3
R Markdown
Console
00:02 / 08:48
CC
1x
- 100% +
05:20:48