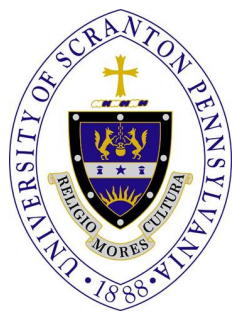
Classification: Logistic Regression
Part 1: Introduction to Classification and Logistic Regression

Aram Balagyozyan

Department of Operations and Information Management
Kania School of Management
The
University of Scranton

January 13, 2020



1. Introduction
2. The Logistic Regression Model

Starter Case

- Suppose a bank is attempting to identify customers who are likely to accept a loan offer in the future.
- The bank has a dataset that includes 5000 customers. The data include the customers' response to the last personal loan campaign (`Personal Loan`), as well as customer demographic information (`Age`, `Education`, `Income`, etc.), and the customer's relationship with the bank (`Mortgage`, `SecuritiesAccount`, etc.).
- The data are stored in file *UniversalBank.csv*.
- Among these 5000 customers, only 480 ($= 9.6\%$) accepted the personal loan offered to them in a previous campaign.

Starter Case

- Let's download the data and run a quick preprocessing and visualization

```
bank.df<-read.csv("UniversalBank.csv")
#Drop ID and Zip Code columns
bank.df<-bank.df[,-c(1,5)]
```

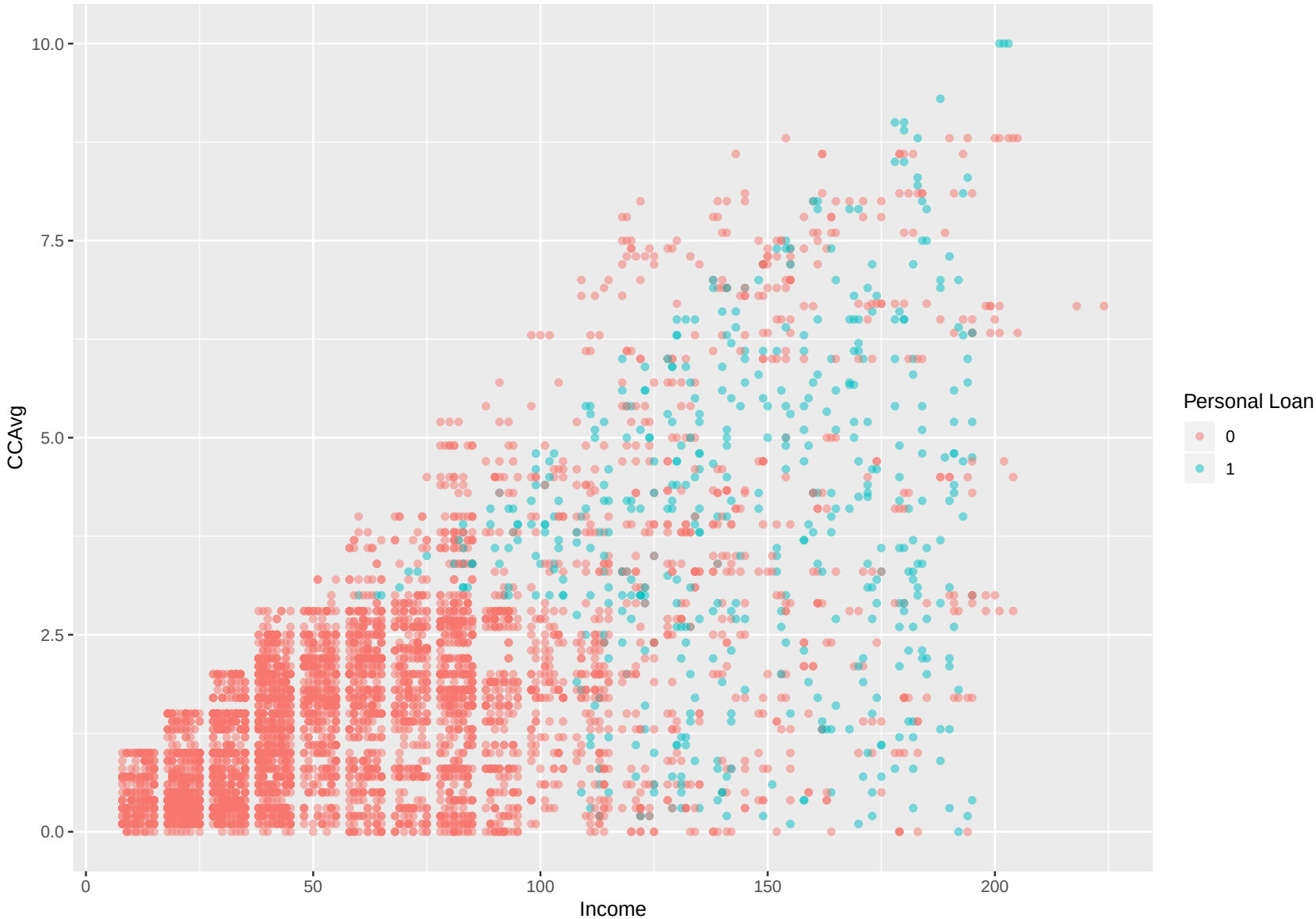
Starter Case

- Below is the data dictionary

Variable	Description
Age	Customer's age in completed years
Experience	Number of years of professional experience
Income	Annual income of the customer ($000s)
Family	Family size of the customer
CCAvg	Average spending on credit cards per month ($000s)
Education	Undergrad; Graduate; Advanced/Professional
Mortgage	Value of house mortgage if any ($000s)
PersonalLoan	1 if customer has accepted a personal loan offer in the past
SecuritiesAccount	1 if customer has securities account with bank
CDAccount	1 if customer has certificate of deposit (CD) account with the Bank
Online	1 if customer uses Internet banking facilities
CreditCard	1 if customer uses credit card issued by the Bank

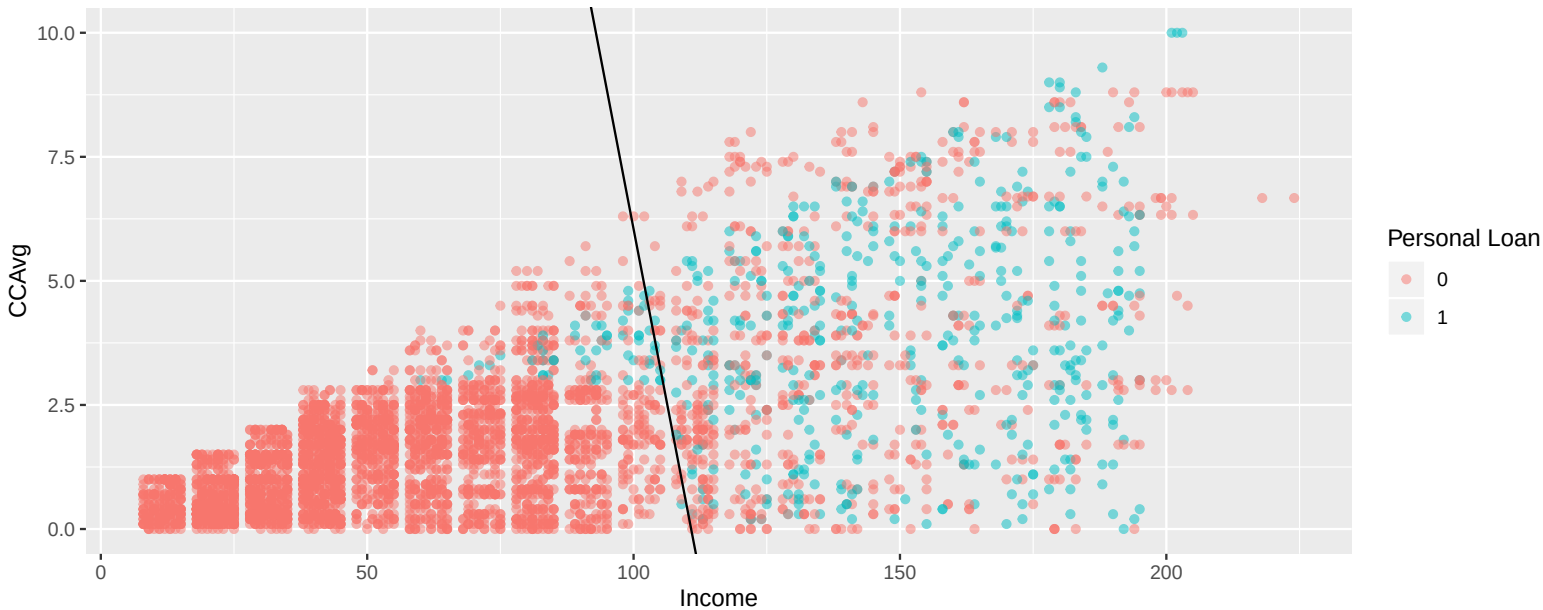
- Suppose the analytics department at the bank is interested to find out if knowledge about the customer's income and family size can help to classify the customer as someone who will likely accept a personal loan offer. The plot on the following slide may shed light.

Starter Case



Starter Case

- It would be great to have a separation rule such that given a customer's income and the average credit-card spending, we could classify the customer s someone who will likely accept a loan offer.
- In other words, it would be great to have the formula of of the black line below.



- If for a given new customer, the intersection of the Income and CCAvg coordinates falls above the black line then the person will likely accept the loan offer.

Introduction to Logistic Regression

- In the case above, logistic regression can be used to classify new customers as likely borrowers or non-borrowers.
- Logistic regression extends the ideas of linear regression to the situation where the outcome variable, Y, is categorical. We can think of a categorical variable as dividing the records into classes, such as borrower or non-borrower.
- Classes can be binary or contain more than two categories. For example, a stock broker may want to categorize each of the stock in the dataset as belonging to one of the three classes: the *hold* class, the *sell* class, and the *buy* class.
- Logistic regression can be used for classifying a new record, where its class is unknown, into one of the classes, based on the values of its predictor variables (called classification).
- It can also be used in data where the class is known, to find factors distinguishing between records in different classes in terms of their predictor variables, or "predictor profile" (called profiling).

Introduction to Logistic Regression

- While in multiple linear regression the aim is to predict the value of the continuous Y for a new record, in logistic regression the goal is to predict which class a new record will belong to, or simply to classify the record into one of the classes.
- Logistic regression is used in applications such as
  1. Classifying customers as returning or non-returning (classification)
  2. Finding factors that differentiate between male and female top executives (profiling)
  3. Predicting the approval or disapproval of a loan based on information such as credit scores (classification)
- Thus, in the case of the binary logistic regression model, we only deal with a binary outcome variable having two possible classes, such as success/failure, buy/don't buy, default/don't default. We often code the values of the binary outcome variable as 0 or 1.
- The predictor variables X_1, X_2,. . . , X_k may be categorical variables, continuous variables, or a mixture of these two types.

Introduction to Logistic Regression

- In logistic regression, we take two steps:
  1. The first step yields estimates of the propensities or probabilities of belonging to each class. In the binary case, we get an estimate of $p = P(Y = 1)$, the probability of belonging to class 1 (which also tells us the probability of belonging to class 0).
  2. In the next step, we use a cutoff value on these probabilities in order to classify each case into one of the classes. For example, in a binary case, a cutoff of 0.5 means that cases with an estimated probability of $P(Y = 1) \geq 0.5$ are classified as belonging to class 1, whereas cases with $P(Y = 1) < 0.5$ are classified as belonging to class 0.
- This cutoff does not need to be set at 0.5. When the event in question is a low probability but notable or important event a lower cutoff may be used to classify more cases as belonging to class 1.

The Logistic Regression Model

- In the logistic regression model, the outcome variable is $p = P(Y = 1)$ is the probability that a given record belongs to class 1 (as opposed to class 0).
- It must be obvious that p can take any value between 0 and 1. However, if we express p as a linear function of the predictors of the form

$$p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_q x_q$$

  it is not guaranteed that the right-hand side will lead to values within the interval $[0, 1]$.
- The solution is to use a non-linear function of the predictors in the form of the *logistic response function*.

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_q x_2)}}$$

- Logistic regression coefficients can be easily interpreted when the odds rather than the probability is considered.

Odds vs. Probability

- The odds of belonging to class 1 are defined as the ratio of the probability of belonging to class 1 to the probability of belonging to class 0:

$$Odds(Y = 1) = \frac{p}{1 - p}$$

- If, for example, the probability of winning is 0.25, the odds of winning are $0.25/0.75 = 1/3$.
- The formula above computes the odds of an event given its probability. Conversely, we can compute the probability of an event given its odds:

$$p = \frac{Odds(Y = 1)}{1 + Odds(Y = 1)}$$

Odds vs. Probability

- Substituting the above formula into the formula of the logistic response function and solving for $Odds(Y = 1)$ yields:

$$Odds(Y = 1) = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_q x_2}$$

- If we take a natural logarithm on both sides, we get:

$$\log(Odds(Y = 1)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_q x_2$$

- The last two equations describe a multiplicative (proportional) relationship between the predictors and the odds.
- Such a relationship is interpretable in terms of percentages. For example, a unit increase in predictor X_j is associated with an average increase of $\beta_j \times 100\%$ in the odds (holding all other predictors constant).