Introduction to Logistic Regression

Logistic Regression I

All classifications in this world lack sharp boundaries, and all transitions are gradual.
-Aleksandr Solzhenitsyn

Learning Objectives

1. Understand classification problems in real life scenarios.

Learning Objectives

1. Understand classification problems in real life scenarios.
2. Understand why Linear Regression is not suitable for predicting categorical response variables.

Learning Objectives

1. Understand classification problems in real life scenarios.
2. Understand why Linear Regression is not suitable for predicting categorical response variables.
3. Understand how Logistic Regression predicts the probability of an outcome.

Classification Problems

Categorical Response Variables

- Response variable is categorical:
  - **binary** yes/no
  - **multinomial** more than 2 classes or categories
- Classification scenarios:
  - Purchase Propensity: Given their online transaction data, will this customer buy a particular product, or not.

Classification Problems

Categorical Response Variables

- Response variable is categorical:
  - **binary** yes/no
  - **multinomial** more than 2 classes or categories
- Classification scenarios:
  - Purchase Propensity: Given their online transaction data, will this customer buy a particular product, or not.
  - Job Attrition: Given an employee profile, will this employee leave the company, or not.

Classification Problems

Categorical Response Variables

- Response variable is categorical:
  - **binary** yes/no
  - **multinomial** more than 2 classes or categories
- Classification scenarios:
  - Purchase Propensity: Given their online transaction data, will this customer buy a particular product, or not.
  - Job Attrition: Given an employee profile, will this employee leave the company, or not.
  - Credit Default: Given the account details, will this customer default on their next credit card payment, or not.

Classification Problems

Categorical Response Variables

- Response variable is categorical:
  - **binary** yes/no
  - **multinomial** more than 2 classes or categories
- Classification scenarios:
  - Purchase Propensity: Given their online transaction data, will this customer buy a particular product, or not.
  - Job Attrition: Given an employee profile, will this employee leave the company, or not.
  - Credit Default: Given the account details, will this customer default on their next credit card payment, or not.
  - Disease Diagnosis: Given the blood sample results, does this patient have a particular disease, or not.

Classification Problems

Categorical Response Variables

- Response variable is categorical:
  - **binary** yes/no
  - **multinomial** more than 2 classes or categories
- Classification scenarios:
  - Purchase Propensity: Given their online transaction data, will this customer buy a particular product, or not.
  - Job Attrition: Given an employee profile, will this employee leave the company, or not.
  - Credit Default: Given the account details, will this customer default on their next credit card payment, or not.
  - Disease Diagnosis: Given the blood sample results, does this patient have a particular disease, or not.
  - Academic Program Selection: Given the high school subject grades, which program will this student choose in college?

Example: Credit Default

- In a bank that issues credit cards to its customers, it was found that a few customers were unable to pay their credit card bills in time, and finally they defaulted on their credit card payments.

Example: Credit Default

- In a bank that issues credit cards to its customers, it was found that a few customers were unable to pay their credit card bills in time, and finally they defaulted on their credit card payments.
- The bank wants to control the credit card debt among its customers.

Example: Credit Default

- In a bank that issues credit cards to its customers, it was found that a few customers were unable to pay their credit card bills in time, and finally they defaulted on their credit card payments.
- The bank wants to control the credit card debt among its customers.
- If the bank can predict whether a customer will default or not, they can intervene early on, and decide if they should continue to offer credit to the customer.

Example: Credit Default

- In a bank that issues credit cards to its customers, it was found that a few customers were unable to pay their credit card bills in time, and finally they defaulted on their credit card payments.
- The bank wants to control the credit card debt among its customers.
- If the bank can predict whether a customer will default or not, they can intervene early on, and decide if they should continue to offer credit to the customer.
- **ASK**: Can we predict whether a customer is likely to default on their credit card payment or not, given their profile and account details?

Example: Credit Default

- **ACQUIRE**: Use credit_default dataset [James et al., 2013].
  - Predictor variables: Credit card balance (`balance`), Monthly income(`income`), and whether the customer is a student or not (`student`)
  - Response variable: Whether the customer defaults or not (`default`)

```
library(tidyverse)
default_data <- read.csv("../data/credit_default_os.csv")
default_data <- default_data %>%
                 mutate(default = as.factor(default),
                        student = as.factor(student))
head(default_data)
```

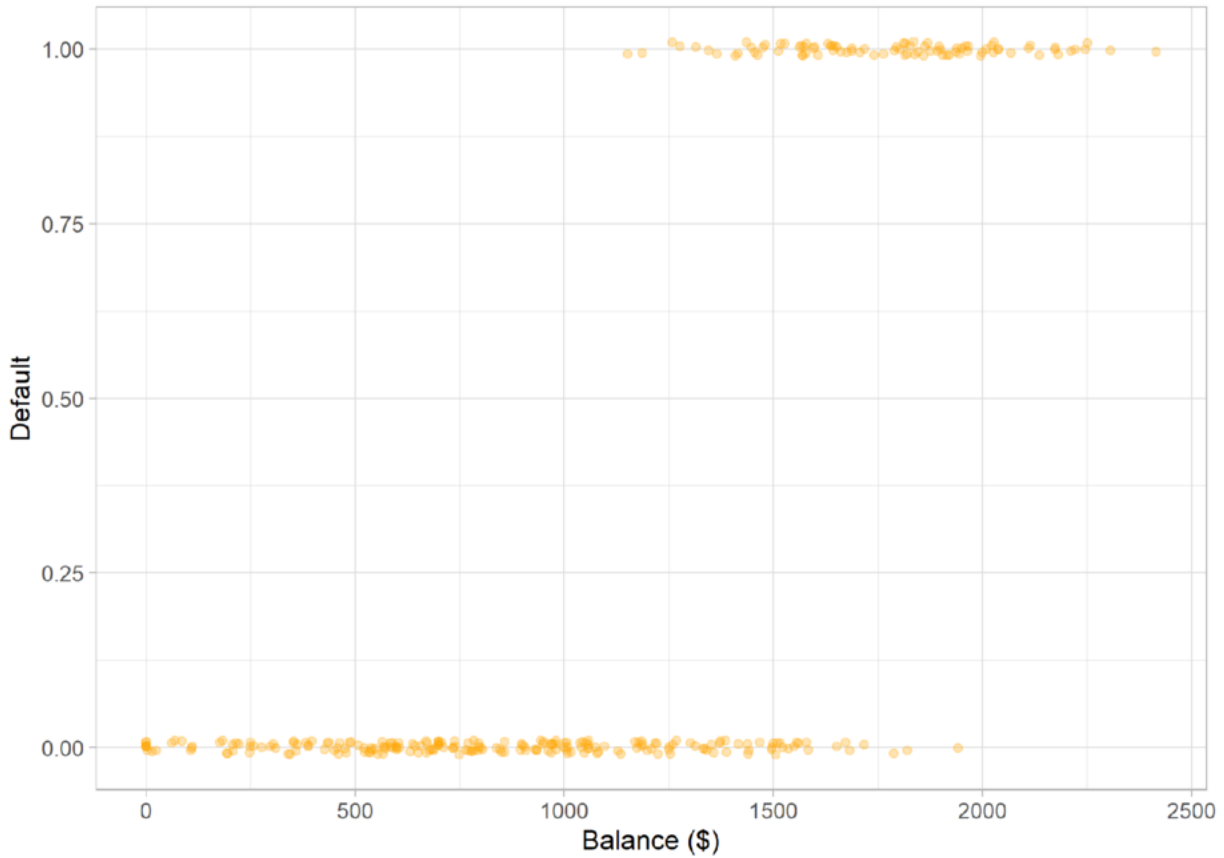
Dataset

	balance	income	student	default
1	1288.407	44253.31	No	Yes
2	1237.622	14862.71	Yes	Yes
3	1530.353	30003.82	No	Yes
4	1627.898	17547.00	Yes	Yes
5	1465.210	58699.98	No	Yes
6	2216.018	20911.70	Yes	Yes

Categorical Response Variable

Scatter Plot

- Encode categorical response variable, `default`:
  - `Y = 1`, if `default = YES`
  - `Y = 0`, if `default = NO`
- Y represents the probability of customer defaulting, i.e. `default = Yes`, given their credit card balance.



Dataset

Encoding categorical response variable

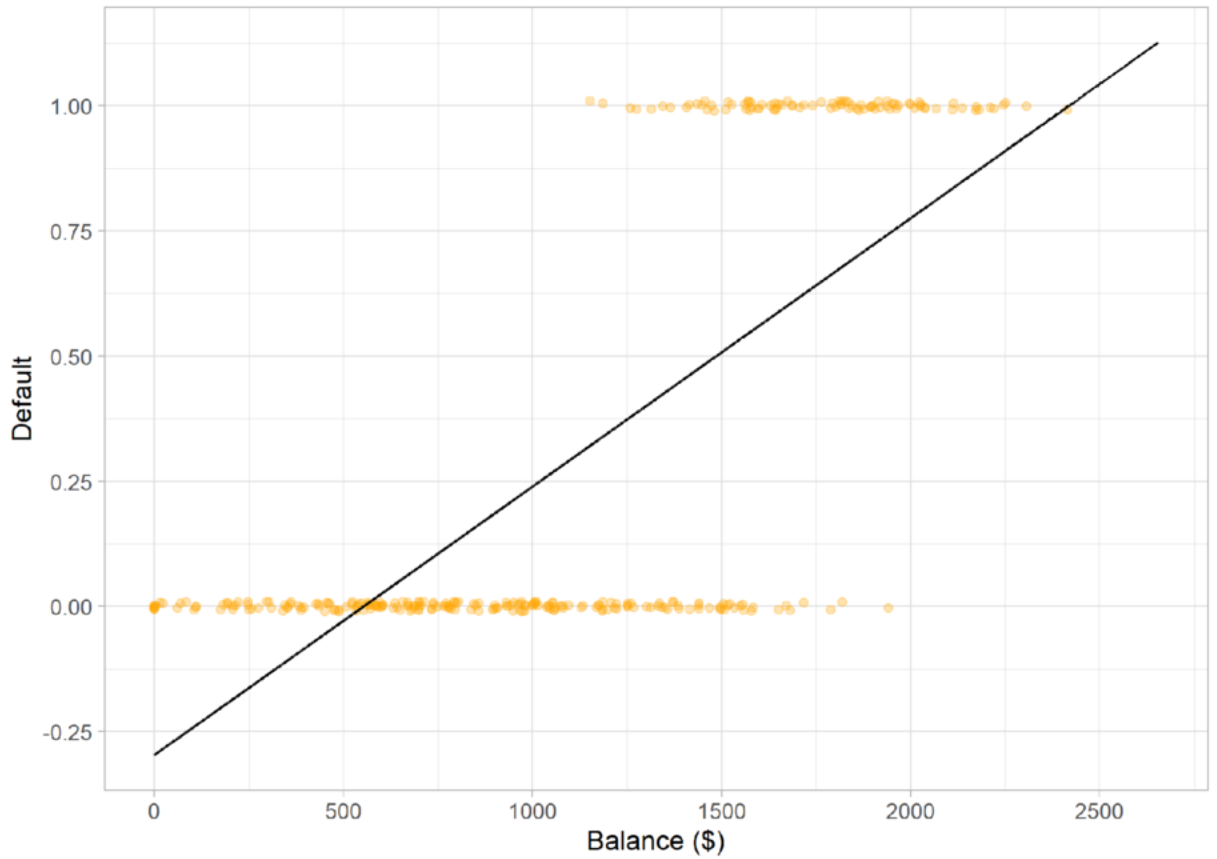
```
default_data <- default_data %>%
        mutate(p_default = ifelse(default == "Yes", 1, 0))
```

```
   balance   income student default p_default
1 1288.407 44253.31      No     Yes         1
2 1237.622 14862.71     Yes     Yes         1
3 1530.353 30003.82      No     Yes         1
4 1627.898 17547.00     Yes     Yes         1
5 1465.210 58699.98      No     Yes         1
6 2216.018 20911.70     Yes     Yes         1
```

Predicting Probabilities

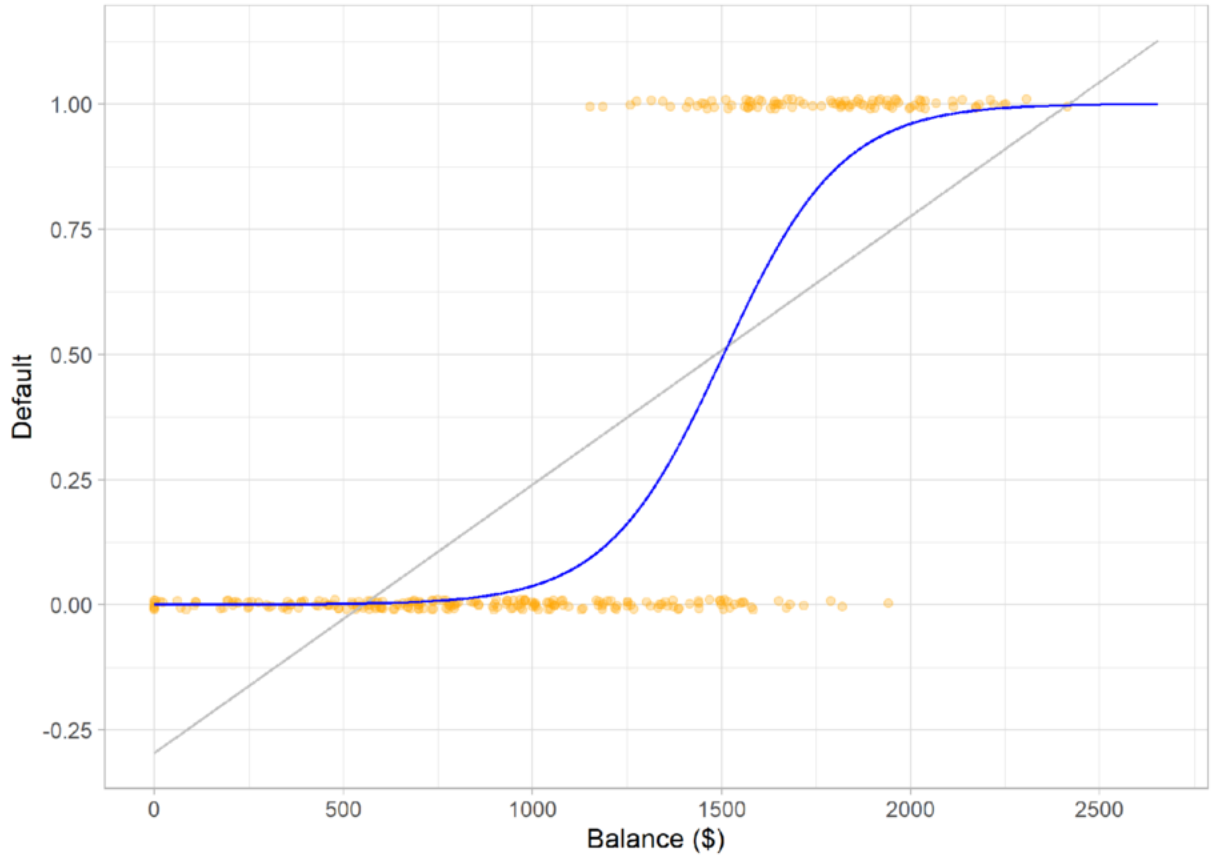
Linear Regression?

- Apply Linear Regression to predict Y as the probability of default, p_default $= \beta_0 + \beta_1 \times$ balance.
- But Linear Regression predicts values for Y, below 0 and above 1, which cannot be interpreted as probability.



Logistic Regression

- Logistic Regression measures relationship between categorical response variable and one or more explanatory variables.
- The logistic model is an S-curve and forces its output Y to be within the range of 0-1, so that we can interpret it as a probability.



Topic Outline

- Logistic model formulation.

Topic Outline

- Logistic model formulation.
- Apply logistic regression model using `glm()` function in `R`.

Topic Outline

- Logistic model formulation.
- Apply logistic regression model using `glm()` function in `R`.
- Interpret and assess training model results.

Topic Outline

- Logistic model formulation.
- Apply logistic regression model using `glm()` function in `R`.
- Interpret and assess training model results.
- Evaluate model performance on classification, with accuracy metrics.

References I


James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013).
An Introduction to Statistical Learning: with Applications in R.
Springer.