Credit Scoring Models and Logistic Regression

QFRM Course

Dr Svetlana Borovkova

Credit scoring models

- Are needed to assess “quality” of a borrower/loan
- Individual borrowers (mortgages, credit cards)
- SMEs
- Corporates

- These models compare individual characteristics of a specific borrower to a pool of existing borrowers for whom it is known whether they defaulted or not

- And in this way (statistically) these models try to forecast default behavior, i.e. assess the likelihood of default of a specific borrower

Use of credit scoring models

- Deciding on whether give or reject a loan/mortgage
- Assessing “health” of existing loans
- Understanding which person- or company-specific characteristics drive defaults

What is default?

- This is the most difficult and fundamental question
- Different definitions of defaults
- Illiquidity, insolvency, missed payments, …
- Recently, regulatory definition of default (of corporate clients) has changed, leading banks to great problems with their credit scoring models
- **This is because your default definition and that in the data you use to build credit model must match**

What are these individual characteristics?

- For individual borrowers:
  - Type of loan : amount, interest rate, maturity etc
  - Financial (FICO score, other loans, missed payments, previous loans etc)
  - Personal/social (age, gender, marital status, education, postcode, children etc)
  - Employment (type, salary, years in service etc)
  - For mortgages: also characteristics of mortgage and of property

- For corporates:
  - Financial ratios
  - Sector, region
  - Size and age of company

- For all models: global variables (interest rates, GDP, unemployment)

For corporates: famous Altman's Z-score model

Equation for Altman's Z-Score Model (1968):

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1X_5$$

X_1 = Working Capital / Total Assets

X_2 = Retained Earnings / Total Assets

X_3 = Earnings Before Interest & Tax (EBIT) / Total Assets

X_4 = Market Capitalisation / Total Liabilities

X_5 = Sales / Total Assets

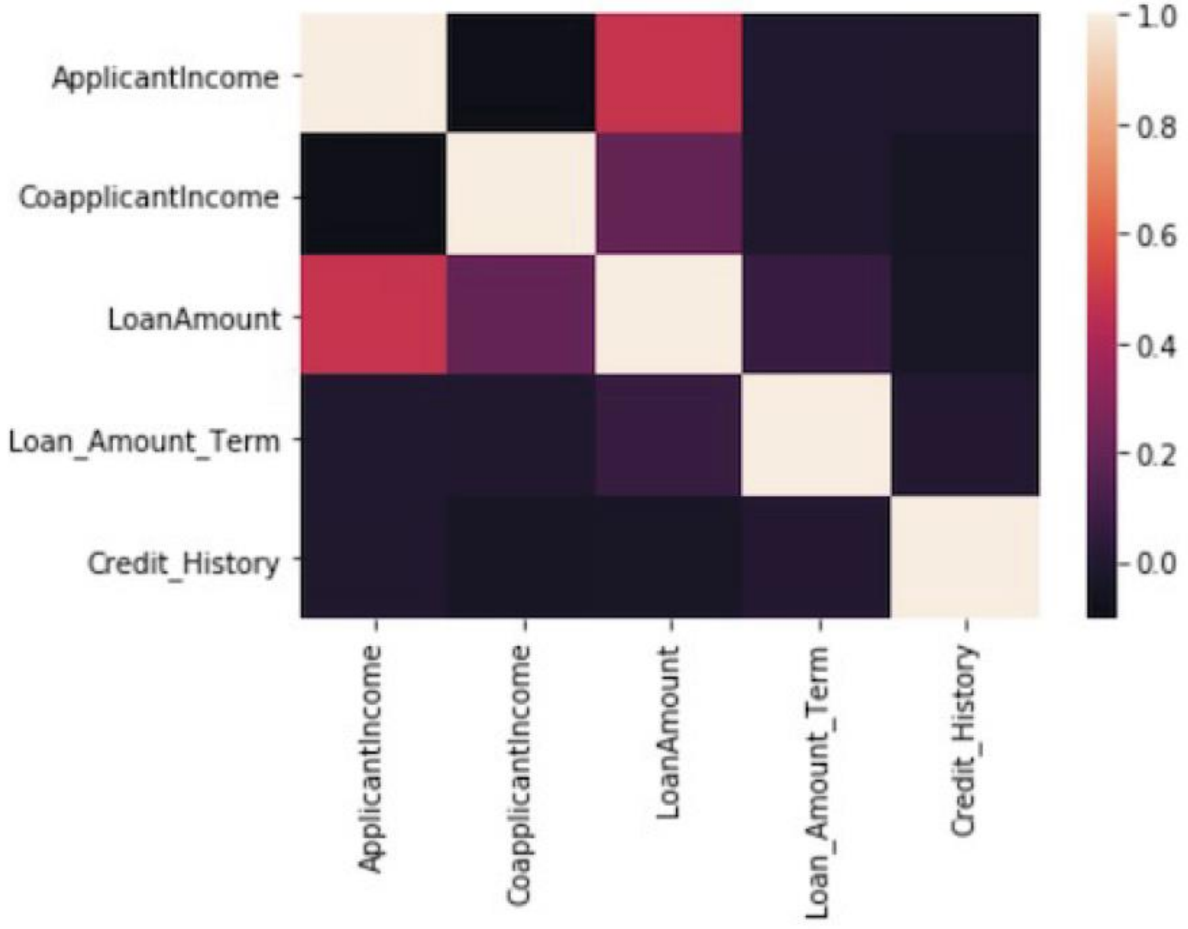
Data is the key

- Data issues are most time consuming and challenging question when developing credit scoring models
- Internal vs external data?
- Data quality and relevance
- Next, we summarize the main data issues

Data issues

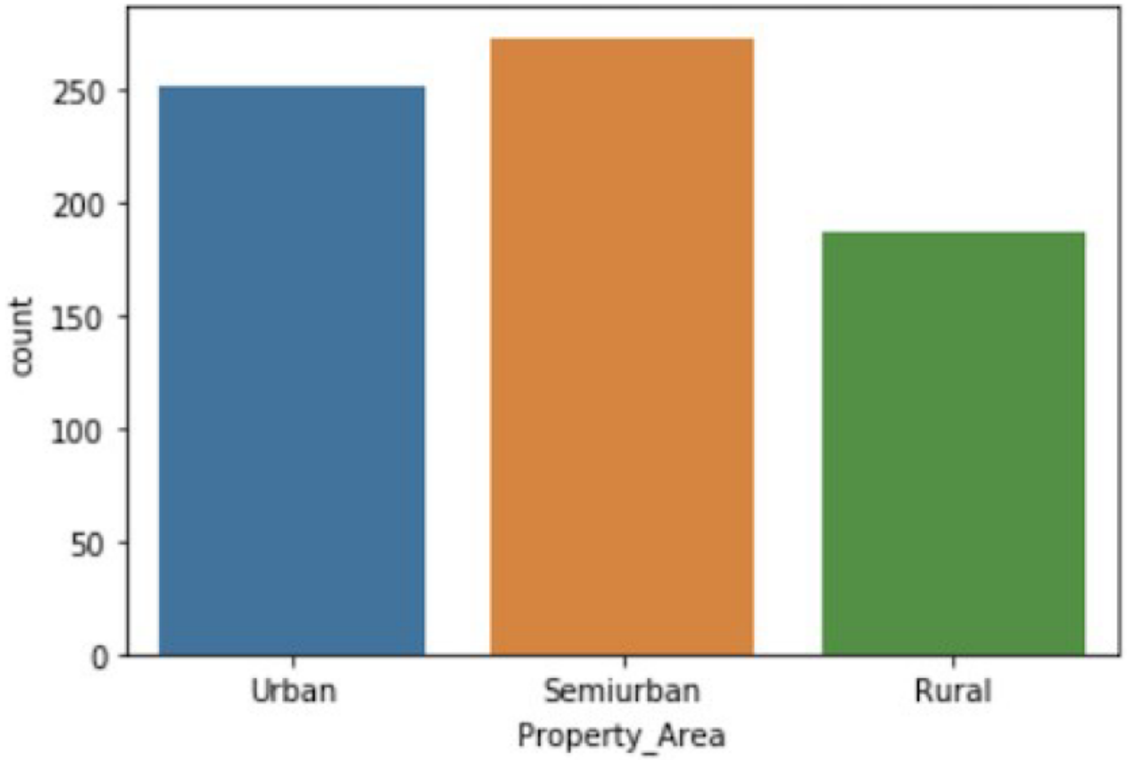
- Size and “representativeness” of the data in relation to actual credit portfolio
- Missing values: how many and why?
- Proportion of defaults in dataset (SMOTE)
- Frequency of values in each potential explanatory variable
- Proportion and reason of outliers
- Selection bias (“reject inference”)
- Transformation of variables: log, square, deviation from the mean , standardized,

Data pre-processing: relations between predictor variables: correlations or scatter plots

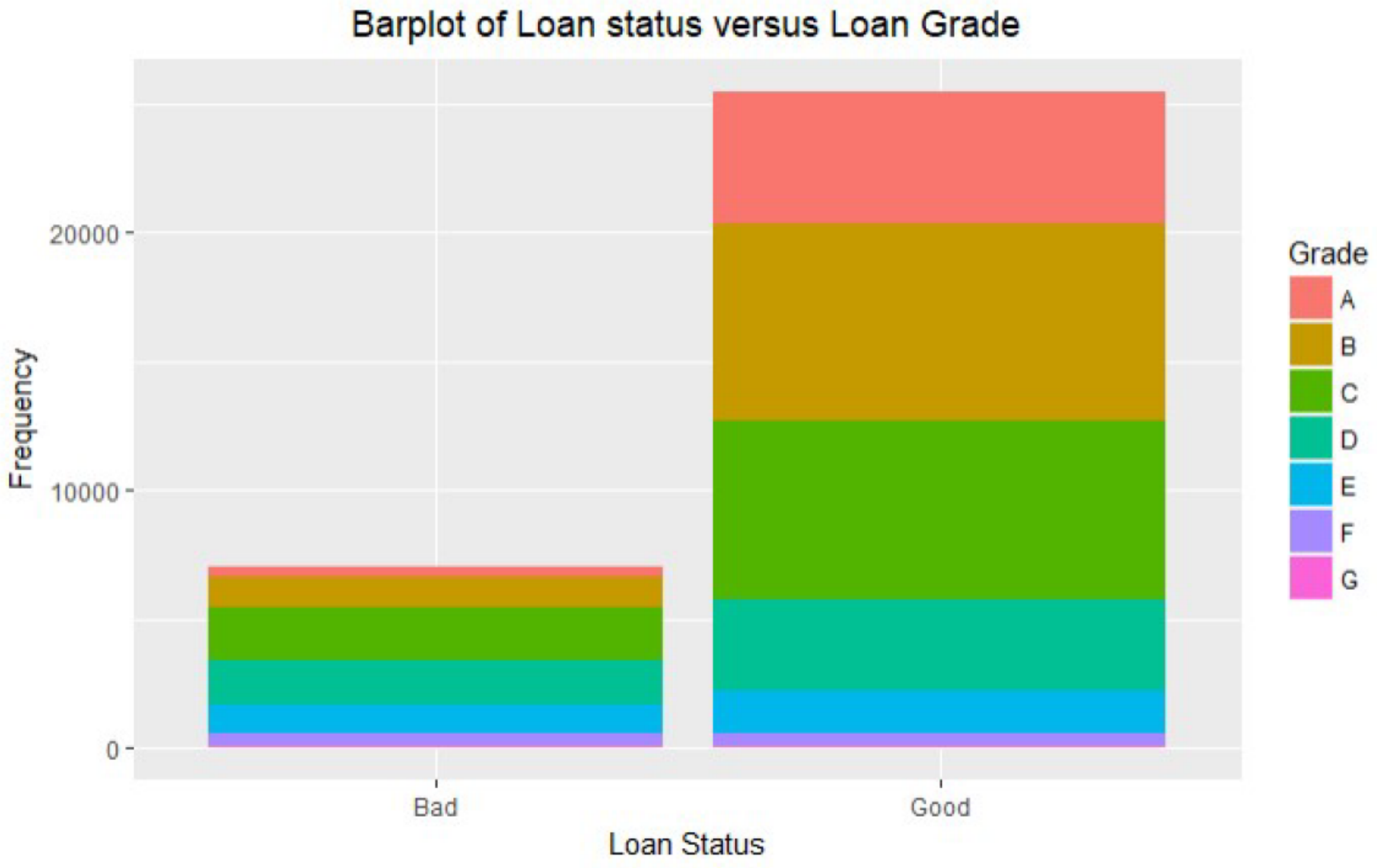


The Correlation Matrix for the Dataset

Frequency of values: bar or pie charts



Also per category



Credit scoring models

- This is fundamentally a classification problem (1 – defaulted, 0 – non-defaulted)
- Other related problems are similar (e.g., mortgage prepayments, non-maturing deposits withdrawals)

- Possible model choices:
- Logistics (logit) or probit regression
- ML classification methods: Neural Nets, Decision Trees, Gradient Boosting, Support Vector Machines, Random Forest

- I have never seen ML methods (or anything else) significantly outperform **logistic regression**

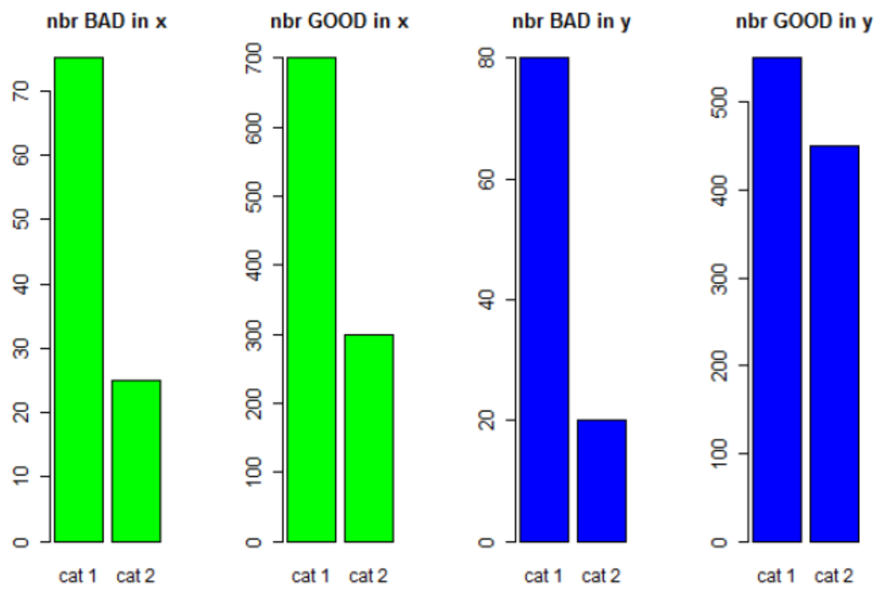
Issues in model development, performance testing and interpretation

- These issues are fundamentally the same or similar for any chosen model (except for interpretation, which is much more transparent in logistic regression)
- Training vs generalization error: 70% of data – training set, 30% - test set
- Variable selection: PCA, top-down or bottom-up, but need exploratory data analysis first to get some idea.
- Often Information Value criteria are used

Information value of a variable x

$$IV(x) = \sum_{i=1}^{N(x)} \left(\frac{g_i}{g} - \frac{b_i}{b}\right) \cdot log\left(\frac{\frac{g_i}{g}}{\frac{b_i}{b}}\right)$$

- *N(x)* is the number of levels in the variable *x*
- g_i represents the number of goods (no default) in category *i* of variable x_i
- b_i represents the number of bads (default) in category *i* of variable x_i
- *g* represents the number of goods (no default) in the entire dataset
- *b* represents the number of bads (default) in the entire dataset



VARIABLE x	GOOD	BAD	VARIABLE y	GOOD	BAD
Category 1 of x	700	75	**Category 1 of y**	550	80
Category 2 of x	300	25	**Category 2 of y**	450	20

$IV(x) = 0.0064$ and $IV(y) = 0.158$.

How do we use IV?

Classification power	Information Value
Poor	<0.15
Moderate	Between 0.15 and 0.4
Strong	>0.4

Assessing quality of the model

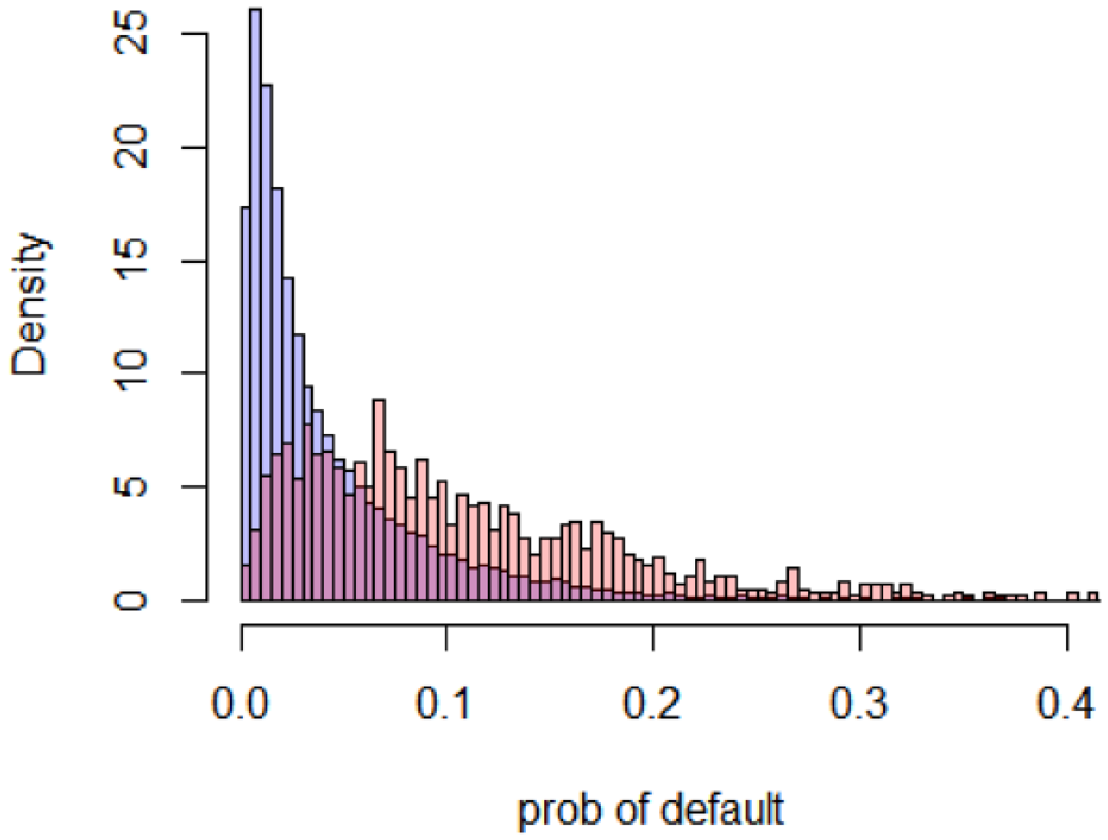
- For logistic regression, quantities analogous to linear regression residuals, goodness-of-fit are available → next time

- But there are also general tools for assessing quality of classification models

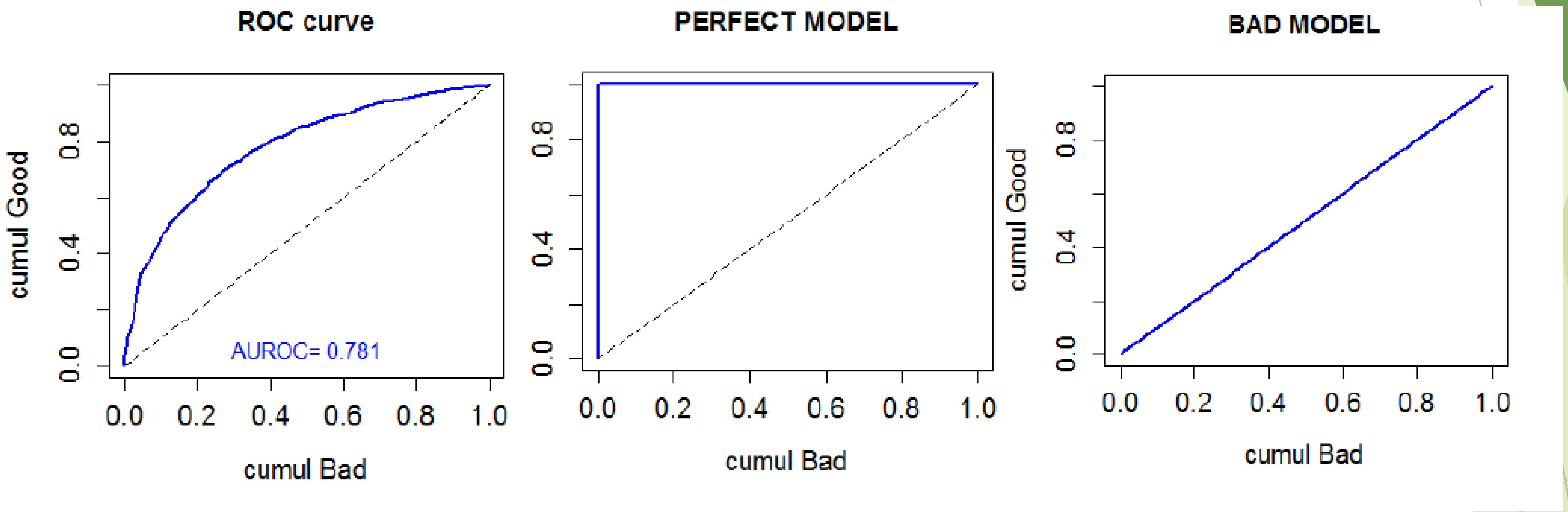
- These are:
- ROC (Receiving Operating Characteristics, or Area Under Curve (AUC))
- Confusion matrix
- Accuracy, Precision and Recall

Red: defaulted, blue: not defaulted

- So the problem is choosing correct cut-off value of PoD



ROC: includes in one graph the performance of the model for all cut-off PoD values



Quality: Area Under Curve (perfect model: AUC=1, bad model: AUC=0.5)

Predictive Power	Area Under ROC
Acceptable	>70%
Good	>80%
Very Good	>85%

Confusion matrix

	Predicted Bad	Predicted Good
Observed Bad	357	178
Observed Good	3171	7014

- Particular focus on
- True Positive (= 7014/(7014+3171)) (also called **Recall**)
- True Negative (=357/(357+178))
- Typical criteria:

Predictive Power	TP & TN rate
Acceptable	>60%
Good	>70%
Very Good	>85%

A closer look at confusion matrix

	Predicted Bad	Predicted Good
Observed Bad	357	178
Observed Good	3171	7014

- False positive: Type I error
- False negative: Type II error
- **Recall**: True Positive = true predicted positive / total positive
- **Precision**: true predicted positive / total predicted positive (=7014/(7014+178))
- **Accuracy**: (True Positive + True Negative) / Total Population

F1 score

- F1 score is the singular metric summarizing the confusion matrix and so model performance. It is harmonic sum of precision and recall:

$$F1 = 2 \cdot Precision \cdot Recall / (Precision + Recall)$$

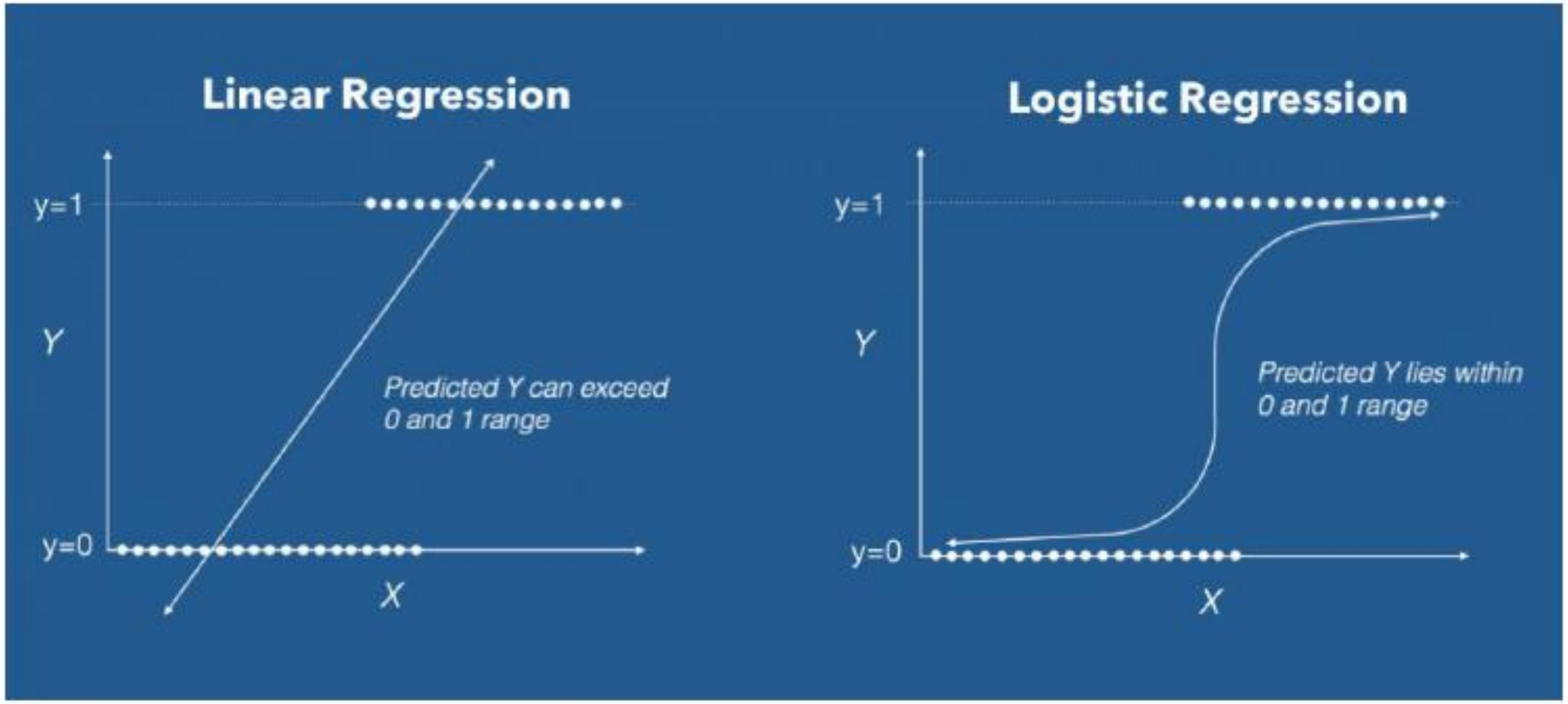
1. **Accuracy: 0.8116883116883117**
2. **Precision: 0.875**
3. **Recall: 0.8504672897196262**

F1 Score: 0.8625592417061612

Logistic regression

- Generalized Linear Model (GLM) where the response variable is either zero or one
- Aims to forecast the probability that the response is 1, given the set of regressors (explanatory variables)
- Main concept: not probability but **odds** (often used in gambling)

Why linear regression cannot be used?



Two alternative formulae for logistic regression

$$logit(p_X) = \log\left(\frac{p_X}{1-p_X}\right) = \beta_0 + \beta_1 X$$

$$p_X = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

Multiple logistic regression for defaults

$$p = \frac{\exp(\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n)}{1 + \exp(\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n)}$$

- p is the probability of default
- x_i is the explanatory factor i
- β_i is the regression coefficient of the explanatory factor i
- n is the number of explanatory variables

Estimation, coefficients interpretation, model diagnostic: next time