Cross-Validation
&
Logistic Regression

Assessment

Type	Weight
Homework's (four)	20%
Midterm Quiz 1	20%
Midterm Quiz 2	20%
Project (Report + Presentation)	30 + 10%
	100%

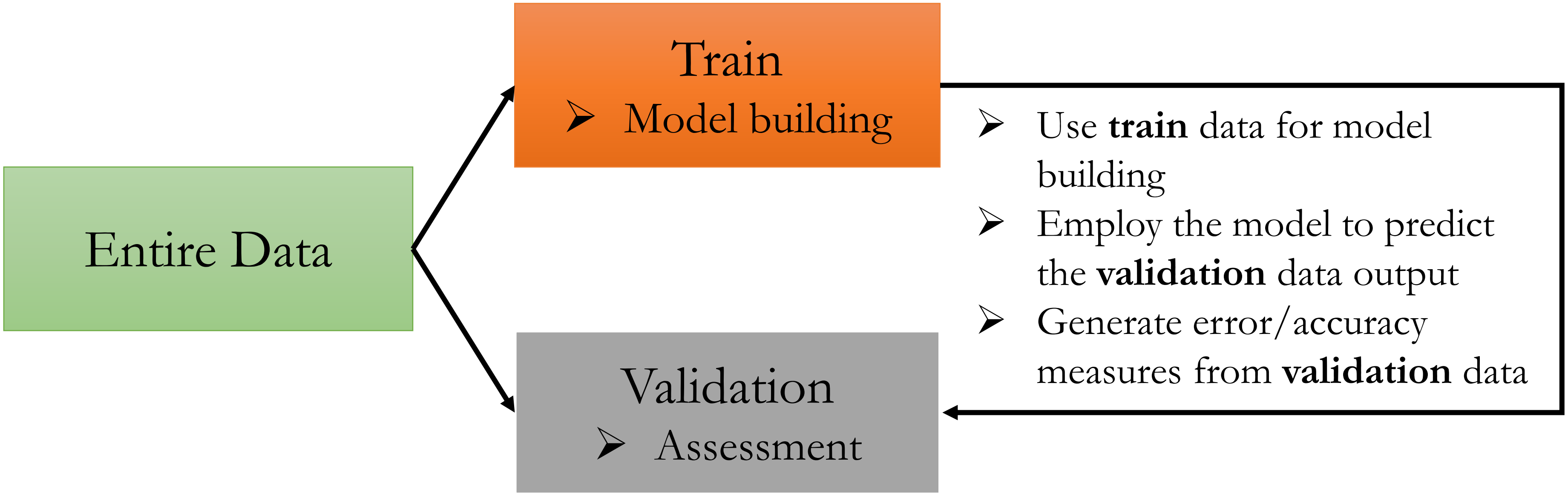
Midterm2 (20%)

- Canvas quiz
  - **Thursday 12th May 2022, 8 am - 9:45 am (105 minutes)**
  - 49 questions, 60 points
  - Path: Canvas → Assignments → Midterm2
- Content
  - Linear regression, Logistics regression
  - Model evaluation (classification & regression) and Cross-validation
- Open book
- Exam in class

Previous class(es)

- Data Partition
  - Train and Validation datasets
- Model Evaluation measures for Regression
  - ME, MAE, MPE, MAPE, RMSE
- Model Evaluation details and measures for Classification
  - Confusion Matrix
  - Misclassification rate, Accuracy
  - Sensitivity and Specificity

Data Partition : Training & Validation



- Assuming 80-20 partitions, how many exhaustive partitions are possible for a dataset with 100 rows?
- $\binom{100}{80} = \frac{100!}{80! * 20!} = 5.36 * 10^{20}$
- We are analyzing only one partition of $5.36 * 10^{20}$
- What about other partitions?

Drawbacks

- What are the drawbacks of analyzing one random partition?
    - Model fit is analyzed on **one** training data partition
    - Error/Accuracy measures are evaluated on **one** validation data partition
    - Likelihood of an excellent model fit and performance on this **one** partition is possible
- Analyzing on a different partition can lead to an unfavorable result
- How to overcome this drawback?

Resampling

- Indispensable tool in Statistics/Machine Learning
- Idea
  - Repeatedly draw a sample from the data
  - Fit model of interest on each sample
- Example
  - Fit Linear Regression on each repeated sample
  - Examine the extent to which results/accuracy measures differ across multiple validation datasets
- Computationally expensive
- Methods: **Cross-Validation** and **Bootstrap**

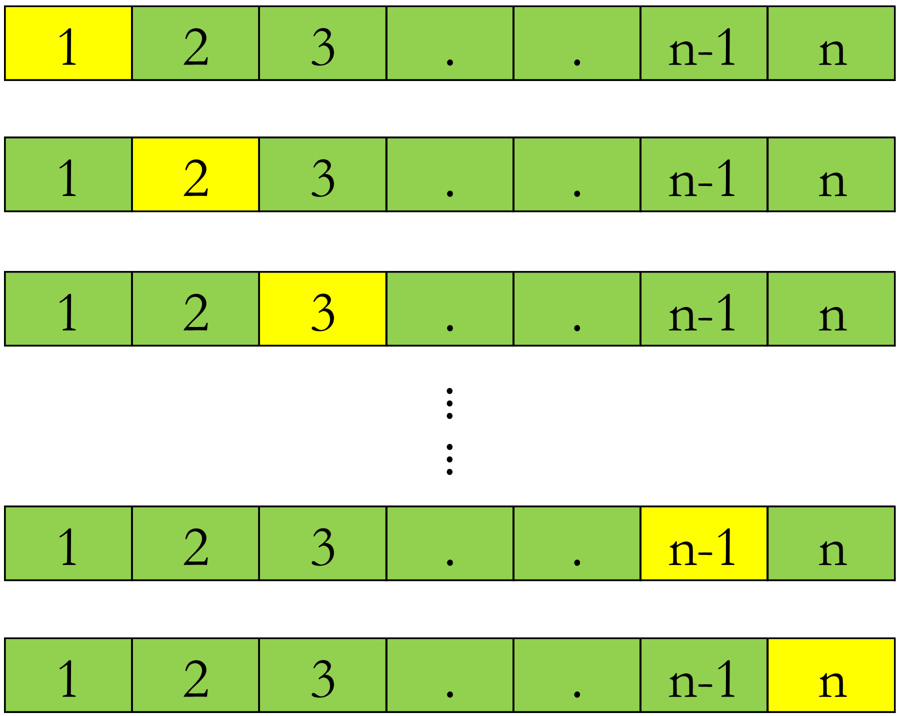
Methods

- Cross-Validation
  - Leave-One-Out Cross-Validation (LOOCV)
  - K-Fold Cross-Validation

Leave-One-Out Cross-Validation (LOOCV)

- Splits the data into two parts but **not** the comparable size
- If you have "n" observations, split into 1 observation and (n-1) observations (iteration)
  - Training data: n-1 observations
  - Validation data: 1 observation
- Fit model on n-1 observations & evaluate on the remaining observation
- How many such iterations are possible?
  - "**n**"
- let's visualize it pictorially

Leave-One-Out Cross-Validation (LOOCV)



Accuracy measure for observation 1

Accuracy measure for observation 2

Accuracy measure for observation 3

⋮

Accuracy measure for observation n-1

Accuracy measure for observation n

Report the Mean/Standard deviation of the accuracy measures

Today's class mandatory steps

- Create a folder name "**j.cross_validation**" within the folder "**oba_455_555_ddpm_r/rproject**"
- Download "**cv_logistics_reg_code.R**", and all **csv** files from canvas
- Place all downloaded files in "**oba_455_555_ddpm_r/rproject / j.cross_validation**"
- Open RStudio project
- Open "**cv_logistics_reg_code.R**" file within RStudio

K - fold Cross-Validation

- Randomly divide the entire data into "**K**" groups (folds), each of approximately the **same** size
- Model is fit on **K**-1 folds and evaluated on the remaining one-fold
- How many such combinations are possible?
    - "**K**"
- let's visualize it pictorially

K-fold Cross-Validation

Fold 1	Fold 2	Fold 3	.	.	Fold K-1	Fold K
Validation	Training	Training	.	.	Training	Training
Training	Validation	Training	.	.	Training	Training
Training	Training	Validation	.	.	Training	Training
			⋮			
Training	Training	Training	.	.	Validation	Training
Training	Training	Training	.	.	Training	Validation

Report the Mean/Median of the accuracy measures obtained for **K** iterations

Generally, K is chosen 5 or 10

LOOCV and K-Fold CV Summary

LOOCV	Measure	Mean	Standard Deviation
	MAPE	9.95	12.7
	RMSE	1,026	1,832

K-Fold	Measure	Mean	Standard Deviation
	MAPE	?	?
	RMSE	?	?

Comparison

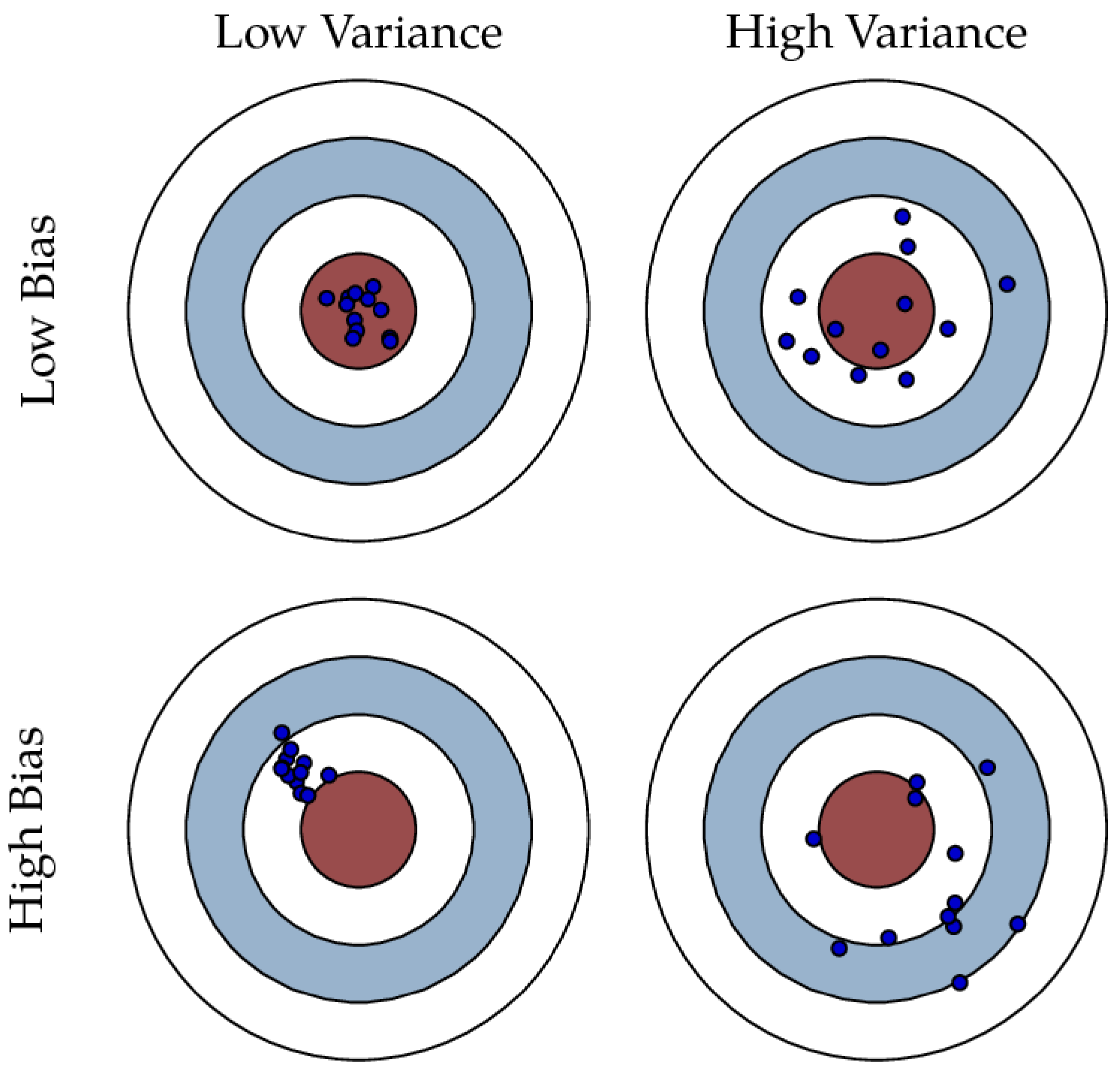
LOOCV

- No randomness in the process
- Time-consuming when "n" is large
- Special case of K-fold CV when K = n
- Less bias compared to **true** validation error measures
- Higher variance

K-fold CV

- Incorporates randomness
- Less time consuming as the process requires to run only K times
- More bias compared to **true** validation error measures
- Less variance

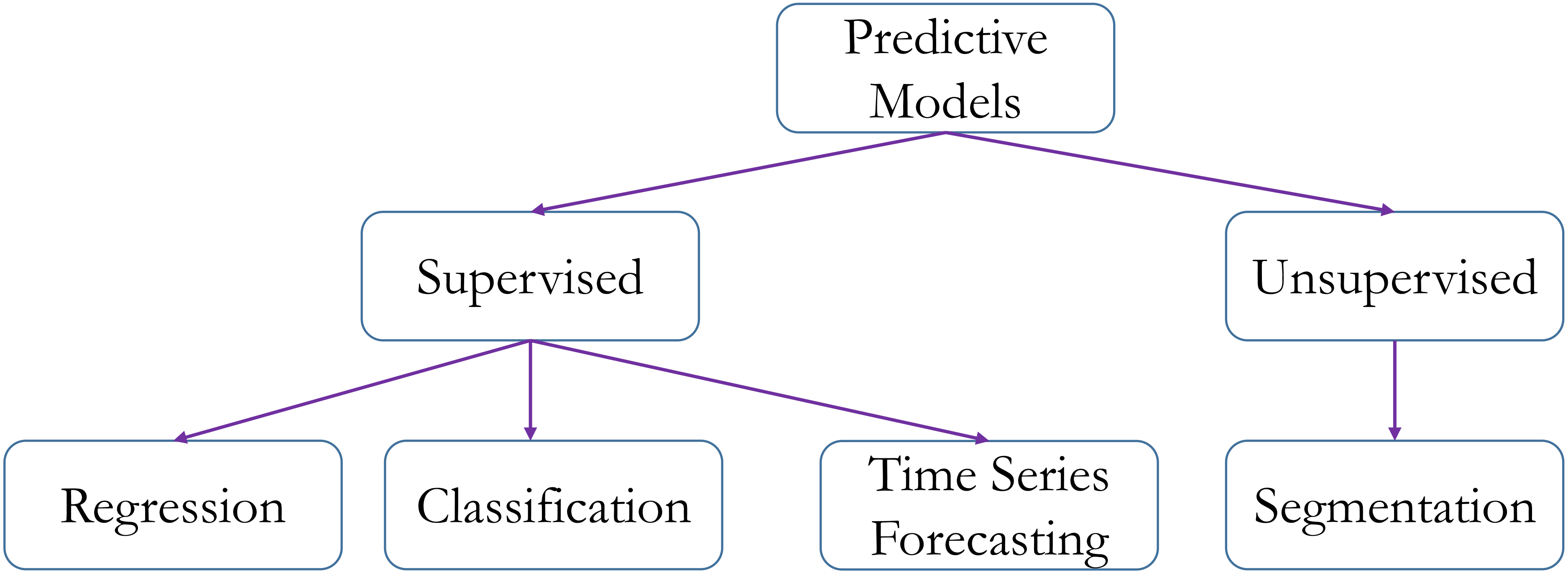
Bias-Variance Trade-off



LOOCV

K-fold CV

Logistic Regression



- ***k*-Nearest Neighbor**
- **Linear Regression**
- **Regression Trees**
- Neural Networks
- Ensembles
- ……

- ***k*-Nearest Neighbor**
- Naïve Bayes
- **Logistic Regression**
- **Classification Trees**
- Neural Networks
- Discriminant Analysis
- Ensembles
- ……

- **Regression-based**
- Smoothing methods
- ……

- **Clustering**
- ……

Logistic Regression

- Prevalent and powerful classification method
- Computationally fast
- Example
  - Let Y denotes recommendation on holding/selling/buying a stock
  - Three categories – **hold**, **sell** and **buy** class
- Goal is to classify a new record whose class is unknown

Logistic Regression

- Non-Linear model
- Like Linear Regression, the method fits a relationship between a categorical variable Y and set of "q" predictors $X_1, X_2, X_3, \cdots\cdots X_q$
- The outcome variable Y is categorical
- Predictors $X_1, X_2, X_3, \cdots\cdots X_q$ can be categorical or numerical
- Prediction is a probability that the new record belongs to a category
- What is the difference compared with *k*-NN?
    - *k*-NN prediction is 100% belonging to a class
    - Logistic Regression prediction is probability belonging to a class

Example : Acceptance of Personal Loan

- Response: Bank customer accepting a loan (1) or not (0)

- Predictors (X)

    - Age, Experience, Income, Family Size, Education
    - Spending on Credit cards
    - Mortgage, Securities account
    - Online banking
    - ……

Example : Predicting delayed flights

- Response: On-time (0) or Delayed (1)

- Predictors (X)
  - Carrier
  - Day of the week
  - Origin
  - Destination
  - Weather
  - ……

Example : Financial condition of Banks

- Response: Weak (0) or Strong (1)

- Predictors (X)

  - Total capital/Assets
  - Total expenses/Assets
  - Total Loans & Leases/Assets
  - ……

Example : Competitive Auctions on e-commerce

- Response : Competitive auction (1) or Non-competitive auction (0)

- Predictors (X)

  - Category (Music, Automotive, etc.)
  - Seller and their rating
  - Auction duration
  - Open price
  - Currency
  - Day of the week of auction close
  - ……

More applications

- Classifying customers as returning or non-returning
- Finding factors that differentiate between male and female top executives (profiling)
- Predicting the approval or disapproval of a loan based on information such as credit scores
- Consumer purchasing behavior
- Choice modeling in Econometrics

Model

- Model in Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q + \epsilon$$

Numeric

Noise or
Unexplained part

- Can we use a similar approach?

- But we have a problem

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q + \epsilon$$

Categorical

- How to address this? (let's say Y has two categories 1, 0)

Transformation

- Logistic response function

Value of e is 2.718

$$p = \Pr(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q)}}$$

- For any values of $X_1, X_2, X_3, \cdots\cdots X_q,$ the right-hand side is always between 0 and 1

- Odds: Ratio of the probability of belonging to class 1 to the probability of belonging to class 0

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

- Odds word is much popular in horse races, sports, gambling…

- Instead of using probability of winning, people quote odds of winning

- If $p = 0.5,$ then $\text{Odds} = 1$

Estimation

- Log Odds

$$\log(\text{Odds}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_q X_q$$

- Information on both $X's$ & Y is available
- $\beta_0, \beta_1, \beta_2 \cdots\cdots \beta_q$ are coefficients
- Required to estimate the coefficients
- Underlying estimation process: **Maximum Likelihood Estimation (MLE)**
    - Find estimates that maximize the chance of obtaining the data we have

Example : Acceptance of Personal Loan

- Response: Bank customer accepting a loan (1) or not (0)

- Predictors (X)
  - Age (years), Experience (years), Income($000s)
  - Family Size
  - Education (undergrad, graduate, advanced)
  - Ccavg (Spending on Credit cards)
  - Mortgage (value of house mortgage in $000s)
  - Securities account (1 if the customer has securities account with the bank)
  - CD account ((1 if the customer has a certificate of deposit account with the bank)
  - Online banking (1 if the customer uses Internet banking facilities)
  - Credit card (1 if the customer uses credit card issued by the bank)

- 5000 customers, 480 accepted (9.8%)

Personal loan data partition

- Let's us consider 70-30 partition
- **Train**: Randomly filter 70% of the entire data
- **Validation**: Extract the remaining 30% of the entire data

Logistic Regression on training data

```
glm(formula = loan_status_actual ~ age + experience + income +
    family + ccavg + education_graduate + education_advanced +
    mortgage + securities_account + cd_account + online + credit_card,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1580  -0.1806  -0.0698  -0.0223   4.1862

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.309e+01  2.198e+00  -5.957 2.57e-09 ***
age                -9.487e-03  8.084e-02  -0.117 0.906585
experience          2.162e-02  8.014e-02   0.270 0.787312
income              5.939e-02  3.500e-03  16.970  < 2e-16 ***
family              6.998e-01  9.638e-02   7.261 3.86e-13 ***
ccavg               1.529e-01  5.218e-02   2.930 0.003394 **
education_graduate  3.724e+00  3.197e-01  11.647  < 2e-16 ***
education_advanced  3.944e+00  3.228e-01  12.218  < 2e-16 ***
mortgage            6.233e-04  7.057e-04   0.883 0.377107
securities_account -1.155e+00  3.876e-01  -2.980 0.002882 **
cd_account          3.833e+00  4.281e-01   8.954  < 2e-16 ***
online             -6.788e-01  2.010e-01  -3.376 0.000734 ***
credit_card        -1.093e+00  2.667e-01  -4.099 4.15e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

logistic regression is run on train data

Results

```
glm(formula = loan_status_actual ~ age + experience + income +
    family + ccavg + education_graduate + education_advanced +
    mortgage + securities_account + cd_account + online + credit_card,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1580  -0.1806  -0.0698  -0.0223   4.1862

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.309e+01  2.198e+00  -5.957 2.57e-09 ***
age                -9.487e-03  8.084e-02  -0.117 0.906585
experience          2.162e-02  8.014e-02   0.270 0.787312
income              5.939e-02  3.500e-03  16.970  < 2e-16 ***
family              6.998e-01  9.638e-02   7.261 3.86e-13 ***
ccavg               1.529e-01  5.218e-02   2.930 0.003394 **
education_graduate  3.724e+00  3.197e-01  11.647  < 2e-16 ***
education_advanced  3.944e+00  3.228e-01  12.218  < 2e-16 ***
mortgage            6.233e-04  7.057e-04   0.883 0.377107
securities_account -1.155e+00  3.876e-01  -2.980 0.002882 **
cd_account          3.833e+00  4.281e-01   8.954  < 2e-16 ***
online             -6.788e-01  2.010e-01  -3.376 0.000734 ***
credit_card        -1.093e+00  2.667e-01  -4.099 4.15e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Higher income, family
- Higher ccavg
- Graduate
- Advanced degree
- Holding a cd account

Associated with higher probability of accepting a loan offer

Results

```
glm(formula = loan_status_actual ~ age + experience + income +
    family + ccavg + education_graduate + education_advanced +
    mortgage + securities_account + cd_account + online + credit_card,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1580  -0.1806  -0.0698  -0.0223   4.1862

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.309e+01  2.198e+00  -5.957 2.57e-09 ***
age                -9.487e-03  8.084e-02  -0.117 0.906585
experience          2.162e-02  8.014e-02   0.270 0.787312
income              5.939e-02  3.500e-03  16.970  < 2e-16 ***
family              6.998e-01  9.638e-02   7.261 3.86e-13 ***
ccavg               1.529e-01  5.218e-02   2.930 0.003394 **
education_graduate  3.724e+00  3.197e-01  11.647  < 2e-16 ***
education_advanced  3.944e+00  3.228e-01  12.218  < 2e-16 ***
mortgage            6.233e-04  7.057e-04   0.883 0.377107
securities_account -1.155e+00  3.876e-01  -2.980 0.002882 **
cd_account          3.833e+00  4.281e-01   8.954  < 2e-16 ***
online             -6.788e-01  2.010e-01  -3.376 0.000734 ***
credit_card        -1.093e+00  2.667e-01  -4.099 4.15e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Holding securities account
- Holding a credit card

Associated with lower probability of accepting a loan offer

Results

```
glm(formula = loan_status_actual ~ age + experience + income +
    family + ccavg + education_graduate + education_advanced +
    mortgage + securities_account + cd_account + online + credit_card,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1580  -0.1806  -0.0698  -0.0223   4.1862

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.309e+01  2.198e+00  -5.957 2.57e-09 ***
age                -9.487e-03  8.084e-02  -0.117 0.906585
experience          2.162e-02  8.014e-02   0.270 0.787312
income              5.939e-02  3.500e-03  16.970  < 2e-16 ***
family              6.998e-01  9.638e-02   7.261 3.86e-13 ***
ccavg               1.529e-01  5.218e-02   2.930 0.003394 **
education_graduate  3.724e+00  3.197e-01  11.647  < 2e-16 ***
education_advanced  3.944e+00  3.228e-01  12.218  < 2e-16 ***
mortgage            6.233e-04  7.057e-04   0.883 0.377107
securities_account -1.155e+00  3.876e-01  -2.980 0.002882 **
cd_account          3.833e+00  4.281e-01   8.954  < 2e-16 ***
online             -6.788e-01  2.010e-01  -3.376 0.000734 ***
credit_card        -1.093e+00  2.667e-01  -4.099 4.15e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- A \$1000 increase in income, holding others constant increases the odds that the customer accepts the loan offer by a factor of 1.061 ($2.718^{0.05939}$)

Results

```
glm(formula = loan_status_actual ~ age + experience + income +
    family + ccavg + education_graduate + education_advanced +
    mortgage + securities_account + cd_account + online + credit_card,
    family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1580  -0.1806  -0.0698  -0.0223   4.1862

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.309e+01  2.198e+00  -5.957 2.57e-09 ***
age                -9.487e-03  8.084e-02  -0.117 0.906585
experience          2.162e-02  8.014e-02   0.270 0.787312
income              5.939e-02  3.500e-03  16.970  < 2e-16 ***
family              6.998e-01  9.638e-02   7.261 3.86e-13 ***
ccavg               1.529e-01  5.218e-02   2.930 0.003394 **
education_graduate  3.724e+00  3.197e-01  11.647  < 2e-16 ***
education_advanced  3.944e+00  3.228e-01  12.218  < 2e-16 ***
mortgage            6.233e-04  7.057e-04   0.883 0.377107
securities_account -1.155e+00  3.876e-01  -2.980 0.002882 **
cd_account          3.833e+00  4.281e-01   8.954  < 2e-16 ***
online             -6.788e-01  2.010e-01  -3.376 0.000734 ***
credit_card        -1.093e+00  2.667e-01  -4.099 4.15e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Customer who has cd account will accept the offer with an odds of 46.2 ($2.718^{3.833}$) relative to a customer who does not have cd account holding all other variables

Requirement for Error/Accuracy measures

- The actual outcome in the validation data takes - 0 (reject), 1 (accept)
- In logistics regression, however the prediction is $p = \Pr(Y = 1)$
- How to compare the actuals and prediction?
- Set the cutoff value and classify the record into the choice of class
    - Set a cutoff value to 0.5
    - If $p \geq 0.5$, the new record is classified to the category "1"
    - If $p < 0.5$, the new record is classified to category "0"
- Build error/accuracy measure from Confusion matrix

Final Project (40%)

- Specify a business problem
- Identify a relevant dataset
- Business context could be in any area or function
- Assessment
  - Report (30%) + Presentation (10%)
- Presentation
  - 10–15-minute presentation on one of the classes in last week
  - **Presentation date(s) i**n the syllabus file

Final Report

- Formal report
  - Introduction, Problem description, Approach (Regression / Classification)
  - Data Analysis, Results, Inference
  - Conclusions, recommendations
- Regression: *k*-NN as Regression, Linear Regression & Regression Tree
- Classification: *k*-NN as classification, Logistic Regression & Classification Tree
- Assess the performance & recommend the best predictive model
- 8-10 pages including any tables and graphs (excluding code)
- Two or Three key insights from the entire analysis
- Submit the code with comments at end of the report

Public datasets for final project

kaggle

- https://www.kaggle.com/
- Online community of data scientists and machine learners
- Owned by Google Inc.
- Register yourself, and you can download datasets for free
- As of June 2017, Kaggle passed over 1,000,000 registered users
- Variety of datasets
- Your imagination only limits possibilities

Next Class

- Logistics Regression on a different dataset
- Grouping categories of input variables in different models

Thank You