Group Project (Airbnb Prediction Competition)

K353 - Spring 2025

Mohammad Zhalechian

Agenda

Background

Datasets

Prediction goal and submission

Evaluation, ranking, and grading

Agenda

Background

Datasets

Prediction goal and submission

Evaluation, ranking, and grading

Background: Airbnb

One of the biggest short-term home-rental online marketplaces

7M listings, across 220 countries (2023)

8.4B USD revenue (2022)

Primary source of Airbnb's revenue: commission fee per booking

- Guest: 6-12%
- Host: 3%

https://www.cityrealty.com/nyc/market-insight/resources-and-guides/airbnb

Background: Airbnb

What Airbnb needs: a model to predict properties bookings

- Revenue predictions for a market (e.g., New York City)
- Identify what types of properties attract more bookings than others
- Identify key factors that impact the bookings

Agenda

Background

Datasets

Prediction goal and submission

Evaluation, ranking, and grading

Datasets: Overview

Airbnb properties in New York City, and their daily listing information

Property information: 14,984 properties

- Location, type, number of reviews, ratings, average price, booking policy, etc.
- Data frame: property_info

Daily Listing information: Covers the first 2 quarters of 2016 with 2.7M observations

- Date, listing status, daily price, etc.
- Data frames: listing_2016Q1, listing_2016Q2

Datasets: Loading into R

Use the following code to open the Airbnb data:

- load(url("https://drive.google.com/uc?export=download&id=1mlJAYmo9TszSJsbYSWhhOY1a3fTJB_Ko"))

After loading, you should be able to see:

- property_info
- listing_2016Q1
- listing_2016Q2
- reserve_2016Q3_train
- propertyID_test

Agenda

Background

Datasets

Prediction goal and submission

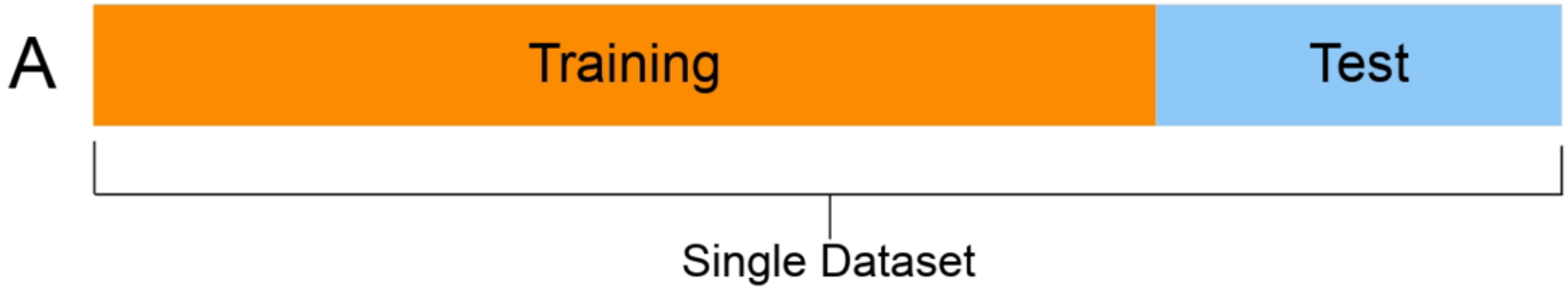
Evaluation, ranking, and grading

Prediction goal

Forecast properties' number of reservation days in Q3 of 2016

Training and Test Data

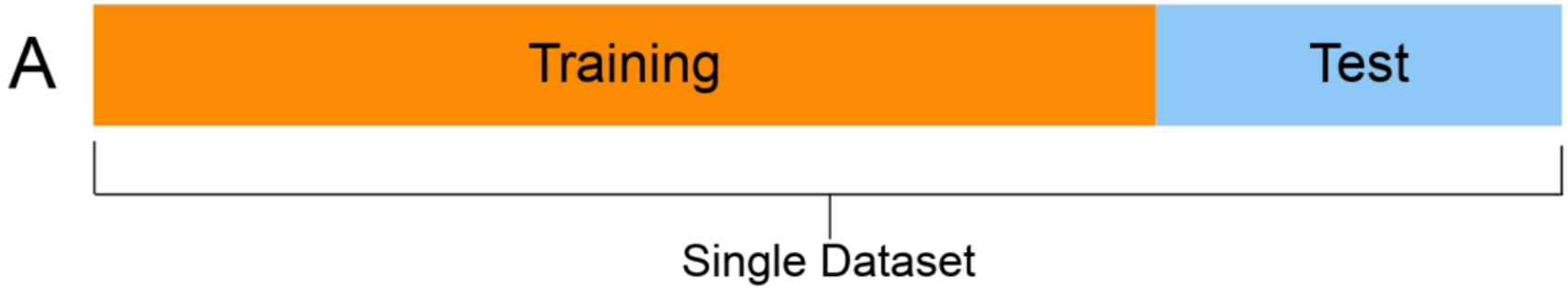
Single Model



https://commons.wikimedia.org/wiki/File:ML_dataset_training_validation_test_sets.png

Training and Test Data

Single Model



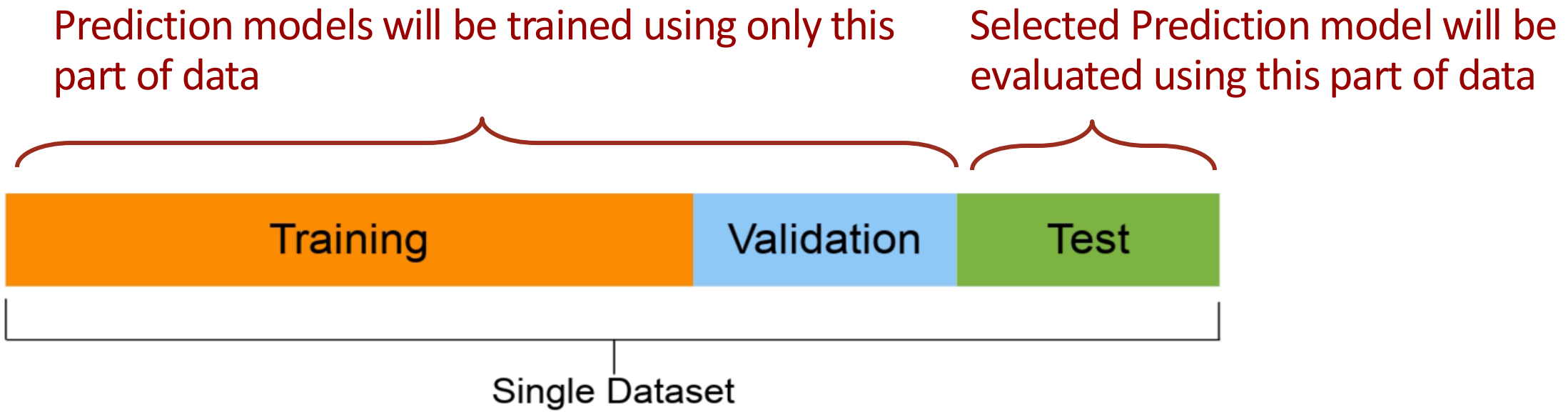
Multiple Models

OR

A single model with fine-tuning

https://commons.wikimedia.org/wiki/File:ML_dataset_training_validation_test_sets.png

Training and Test Data



Prediction goal

Forecast properties' number of reservation days in Q3 of 2016

We know NumReserveDays2016Q3 for half of the properties (7590 out of 14984)

- Train set: reserve_2016Q3_train

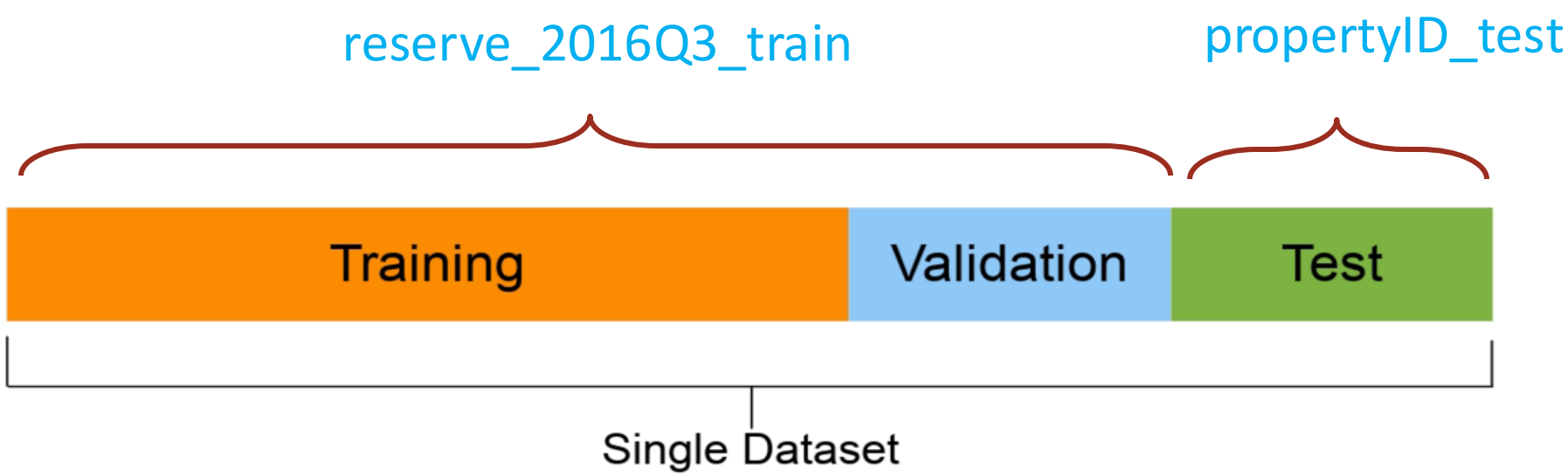
We need to forecast NumReserveDays2016Q3 for the other half of the properties (7394 out of 14984)

- Test set: propertyID_test

```
> head(reserve_2016Q3_train)
  PropertyID NumReserveDays2016Q3
1        105                    0
4       2534                    0
5       2539                   35
7       3330                   43
8       3831                   76
9       4611                   81
```

```
> head(PropertyID_test)
[1]  795 2515 2595 5099 5107 5110
```

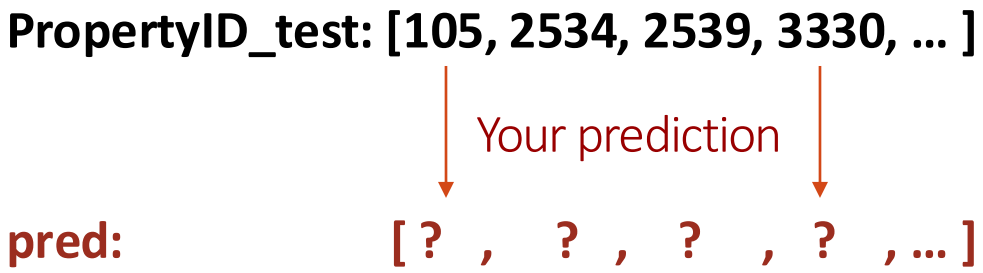
Training and Test Data



Submission requirements

Construct prediction vector:

- pred: a numeric vector of predicted number of bookings for ALL properties in PropertyID_test
- The i^{th} value in pred (i.e., pred[i]) is the booking prediction for property PropertyID_test[i]
- The length of pred should be the same as the length of PropertyID_test, which is 7394
- Missing values (i.e, NA) in pred are not allowed



Submission requirements

Construct prediction vector:

- pred: a numeric vector of predicted number of bookings for ALL properties in PropertyID_test
- The i^{th} value in pred (i.e., pred[i]) is the booking prediction for property PropertyID_test[i]
- The length of pred should be the same as the length of PropertyID_test, which is 7394
- Missing values (i.e, NA) in pred are not allowed

Email a file named your_group_name.rdata that includes vector pred to k353z@iu.edu

- To save pred in a .rdata file to your local drive, you can use something like:

```
save(pred, file = "your_group_name.rdata")
```

- Call ?save for details about how save function is used

Submission limits

No submission limits!

Feel free to submit each time you modify your code

Agenda

Background

Datasets

Prediction goal and submission

Evaluation, ranking, and grading

Presentation (50% of project points)

Data cleaning/management process

Feature selection

- How did you select features for the prediction?
- How did you construct them if they are not directly available?

Prediction models and methods

- What models and methods did you try?
- How did they perform?
- What are the most important features in your selected prediction model?

Presentation (50% of project points)

R implementation

- How efficient is your code (using functions, for loops, etc.)?

Reflection on the prediction challenge

- What have you learned about improving prediction accuracy?
- What additional analyses have you conducted?

Biggest challenge

- What was the most significant challenge you encountered while working on the project?
- How did you overcome it?

Evaluation: Prediction accuracy

Root of mean squared error (RMSE): (the lower the better)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(R_i - \hat{R}_i\right)^2}$$

- i denotes the property index in the test set ($N = 7394$)
- R_i is the actual booking quantity of property i in the test dataset (*undisclosed*)
- $\hat{R}_i$ is your prediction of the booking quantity of property i

Evaluation: Prediction accuracy

Team ranking and leaderboard on Canvas

- Updated based on the best RMSE of each team every day
- The best RMSE of each team will be posted
- Keep track of each of your submission and its performance

Ranking (40% of project points)

Based on the ranking on the last day of the competition

(**4/21/2025 at noon EST**)

First 3 teams get the full grade of the prediction (30 points)

No. i team: $\frac{RMSE\ of\ No.3\ team}{RMSE\ of\ No.i\ team} \times 30$

Peer Evaluation (10% of project points)

You will be asked to submit an evaluation of your contributions to the group project as well as the contributions of your group members.

You will rate each person's contributions as Excellent, Satisfactory, Poor, or Unsatisfactory.

If you receive Poor or Unsatisfactory ratings from all group members, your grade will receive a 30% penalty

Team information

Team size: no more than 3 students

Enter the team information to the spreadsheet on Canvas home page (**List of teams**) by **2/10/2025 at noon EST**

To do for next class

Form teams for Airbnb project

Check out the Airbnb dataset and contact me if you have questions