Lecture 21: Survival analysis

Firas Moosvi

Announcements

- Test 3 update
  - One of the MCQ was mistakenly set to 4 marks instead of 1, need to fix that
  - Resolution will mean everyone's score will increase
  - Scores will be released after
- HW9 has been released!
- Practice Workspace question for the final exam will be posted on PrairieLearn this weekend to help you prepare

Announcements (cont’d)

- We are in the home stretch now! Monday will be the last class
- Tutorials are ON Tues/Wed
- NO tutorials on Thurs/Fri
- Office hours as usual next week

Wrap-up Time Series from last class

…

Clicker 21.1

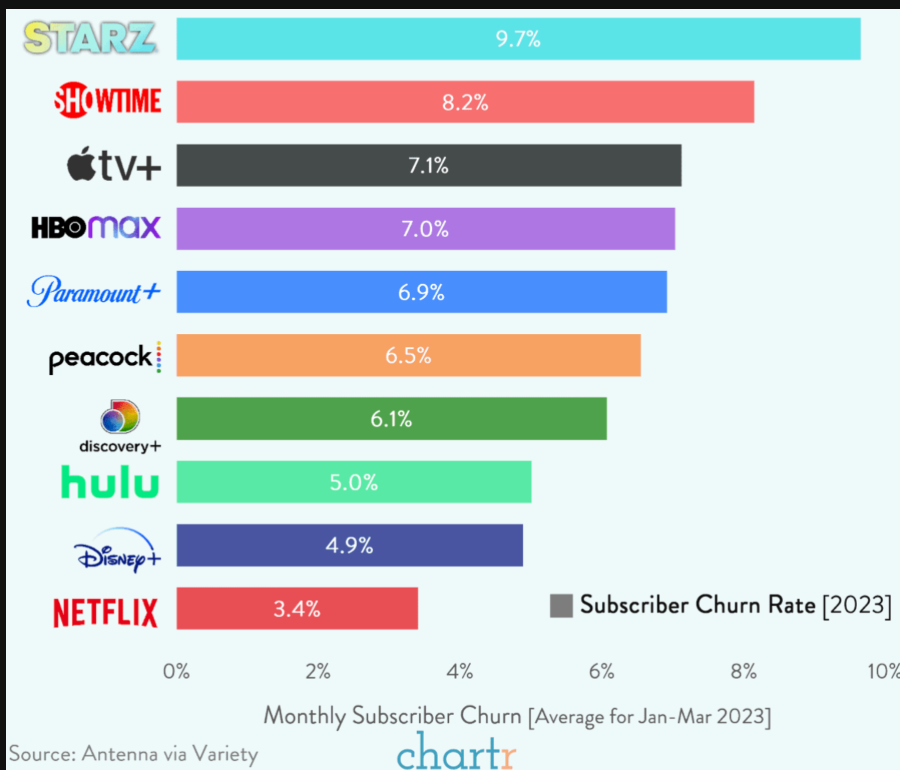
Select all of the following statements which are TRUE.

- a. We need to be careful when splitting the data when working with time series data.
- b. Cross-validation in time series can be randomly applied like in other machine learning tasks.
- c. In time series forecasting, the future value of a series can only be predicted based on its past values and cannot incorporate other variables.
- d. When we used RandomForestRegressor model on the POSIX time feature, it predicted a straight line on the test data because tree-based models are inherently unable to extrapolate (i.e., make predictions outside the - range of the training data).

Customer churn

Customer churn, also known as customer attrition, refers to the phenomenon where customers or subscribers stop doing business with a company or service.

Monthly subscriber churn rates for various streaming services



Source

Question: Is a smaller or a larger churn rate more desirable for a subscription-based company?

- A smaller churn rate is better (means fewer customers leaving)
- Lower churn = higher customer retention = more stable revenue

The challenge: Predicting *when*, not just *whether*

Imagine you work for a subscription-based telecom company.

- Your team wants to predict **when a customer will churn**, not just whether they churn.
- This helps the company:
  - Target retention strategies at the right time
  - Allocate resources efficiently to high-risk customers
  - Understand which factors accelerate or delay churn
- Our goal: **model time to churn** while accounting for customers who haven't churned yet.

Customer Churn Dataset

If you wanted to predict whether a customer churns, what kind of model from your ML toolbox would you use?

	customerID	gender	SeniorCitizen	Partner	Dependents	tenur
6464	4726-DLWQN	Male	1	No	No	50
5707	4537-DKTAL	Female	0	No	No	2
3442	0468-YRPXN	Male	0	No	No	29
3932	1304-NECVQ	Female	1	No	No	2

	customerID	gender	SeniorCitizen	Partner	Dependents	tenur
6124	7153-CHRBV	Female	0	Yes	Yes	57

5 rows × 21 columns

Churn prediction as binary classification

- When we treat churn as a binary classification problem, we only ask: **Has the customer churned by the time of data collection?**
- Limitations of this approach:
  - Answers only “Yes/No” and discards **when** churn occurred
  - Treats a customer who churned after 1 month the same as one who churned after 5 years
  - Ignores the time dimension entirely
- **Is that what we want?** Not if timing matters for business decisions!

Predicting tenure

	tenure	Churn
6464	50	No
5707	2	No
3442	29	No
3932	2	Yes
6124	57	No
301	4	Yes
3552	68	No
2874	64	No

In our dataset, the tenure column is the number of months the customer has stayed with the company. Can we use the techniques you learned so far (e.g., regression models) to predict the time (tenure in our case)?

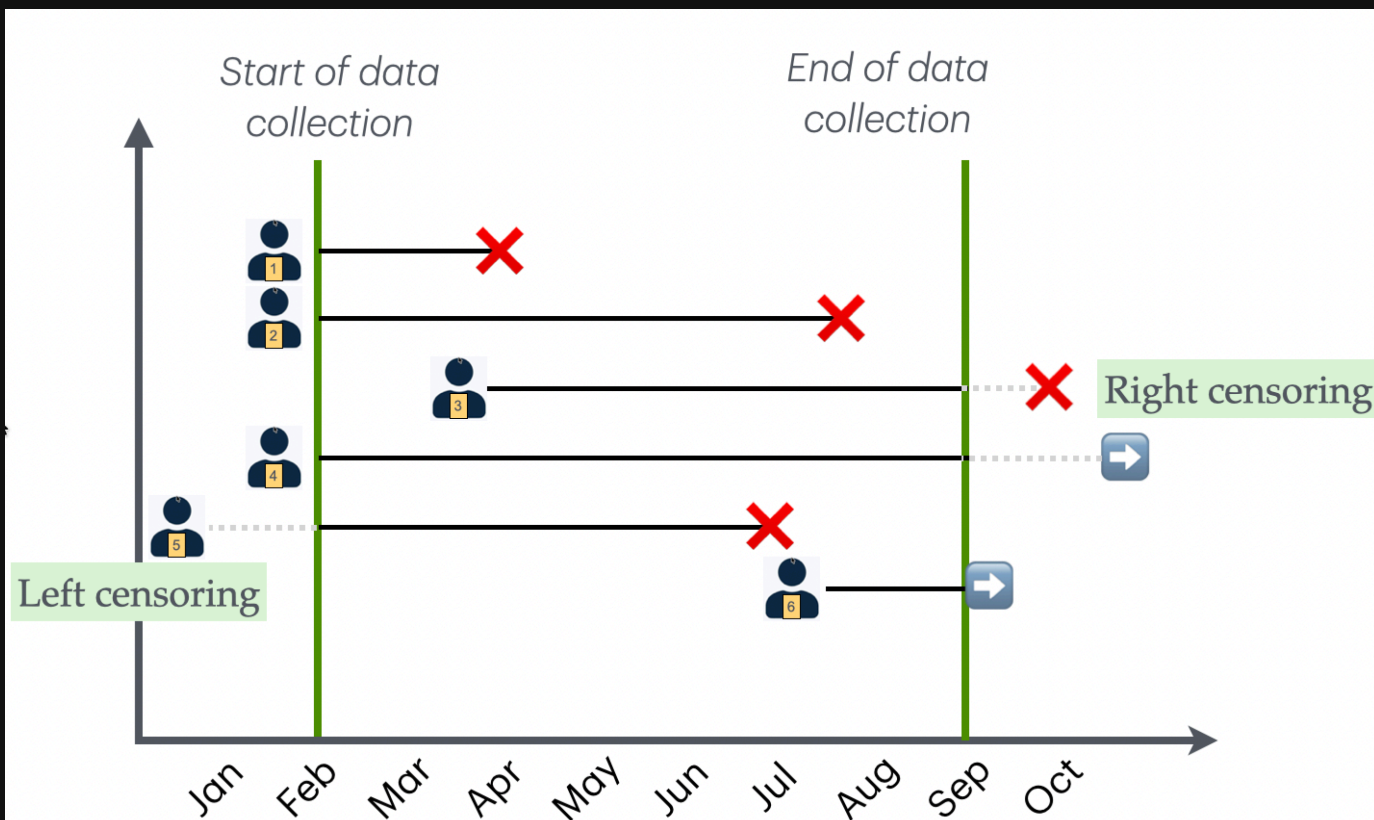
The problem: Incomplete information

	tenure	Churn
6464	50	No
5707	2	No
3442	29	No
3932	2	Yes
6124	57	No
301	4	Yes
3552	68	No
2874	64	No

- We only have information about tenure **up to the point we collected the data**.
- For customers who haven't churned:
  - We don't know their true "time to churn"
  - We only know they lasted **at least** this long
  - Their actual churn time is unknown (incomplete information)
- This is called **right-censoring** - the event hasn't occurred yet

Types of censoring

- **Right-censoring**: Event hasn't occurred yet (most common in practice)
- **Left-censoring**: Event occurred before observation started
- **Interval-censoring**: Event occurred within a time interval



Time to event and censoring

	tenure	Churn
6464	50	No
5707	2	No
3442	29	No
3932	2	Yes
6124	57	No
301	4	Yes
3552	68	No
2874	64	No

- Many customers in the dataset are still active (Churn = "No").
- They are **right-censored**: we only know their tenure **so far**, not their final tenure.

Time-to-event problems

Time-to-event problems appear everywhere. Examples include:

- Time until a customer leaves a subscription service
- Time until a disease causes death
- Time until equipment fails
- Duration of unemployment until finding a job
- Waiting time until a scheduled surgery

These all follow the same pattern: the event happens once, and we care about how long it takes.

Approaches

Approach 1: Only use churned customers ⛔

Suppose we only consider rows where Churn == “Yes” and throw away all right-censored customers.

	customerID	gender	SeniorCitizen	Partner	Dependents	tenur
3932	1304-NECVQ	Female	1	No	No	2
301	8098-LLAZX	Female	1	No	No	4
5540	3803-KMQFW	Female	0	Yes	Yes	1
4084	2777-PHDEI	Female	0	No	No	1

	customerID	gender	SeniorCitizen	Partner	Dependents	tenur

4 rows × 21 columns

This throws away valuable information from active customers!

Clicker

Would only considering rows where Churn == “Yes” and throwing away all right-censored customers overestimate or underestimate the average survival time?

- a. Overestimate
- b. Underestimate
- c. Cannot tell based on the provided information

Approach 2: Assume everyone churns now ⛔

Treat all current tenure values as final, even for active customers.

	tenure	Churn
6464	50	No
5707	2	No
3442	29	No
3932	2	Yes
6124	57	No

This assumes active customers (Churn = "No") will churn immediately.

Question for you:

Would assuming everyone churns now underestimate or overestimate tenure?

- a. Overestimate
- b. Underestimate
- c. Cannot tell based on the provided information

- **Key insight**: Ignoring or removing censored cases will bias our estimates:

Approach 3: Survival analysis ✅

Survival analysis explicitly models time until an event **and properly handles censoring**.

Common methods:

- **Kaplan–Meier estimator**: Non-parametric method for estimating survival curves
- **Cox proportional hazards model**: Semi-parametric regression model for survival data
- **Survival forests**: Random forest variant adapted for censored data

Survival analysis

These methods allow estimation of:

Survival function:

- $S(t) = P(T > t)$ (probability of surviving past time t)
- Imagine starting with 100 customers on Day 0. The survival function tells you: What fraction of them are still active (have not churned) at time t?
- $S(3) = 0.80$ (80% haven't churned by month 3)
- $S(12) = 0.40$ (40% haven't churned by month 12)

Hazard function: $h(t)$ (instantaneous risk of the event at time t)

- Imagine a customer who has not churned yet and is still active at month 10.
- The hazard tells us: How likely are they to churn right now, at month 10?
- Human mortality: hazard is low at age 20, higher at 80

Group Work: Class Demo & Live Coding

For this demo, each student should click this link to create a new repo in their accounts, then clone that repo locally to follow along with the demo from today.

Clicker 21.2

Select all of the following statements which are TRUE.

- a. Right censoring occurs when the endpoint of event has not been observed for all study subjects by the end of the study period.
- b. Right censoring implies that the data is missing completely at random.
- c. In the presence of right-censored data, binary classification models can be applied directly without any modifications or special considerations.
- d. If we apply the Ridge regression model to predict tenure in right censored data, we are likely to underestimate it because the tenure observed in our data is shorter than what it would be in reality.