# **Time Window Exercise**

Today you will be learning how to define appropriate time windows. To do so, it is important to consider the very basics of predictive modelling (such as churn modelling). Predictive modelling looks at patterns observed in the past and then uses these patterns to predict similar behaviour in the future. So, we will use **historic** data to train a model on, that we will then use on **present day** data to predict **future** behaviour. Take following example:

A firm (Newspaper Company) is afraid they will lose a lot of customers during the Summer months (July and August) of 2022 as this a recurrent problem in Summer. They ask you to tackle this problem. You start today, with a *data dump*[[1]](#footnote-1) of today which is the **26th of October 2022**. The company wants to have **1 month** for the deployment of the campaign (restructure data to usable format, contact people …).

Okay, now we know what we want to do, predict customer churn in Summer 2022. This is **ALWAYS** the start of your time windowing, always first start with the goal of your project. This is called the **purpose model**. Whether a customer will churn or not, is the dependent variable of your model.

Shape, rectangle

Description automatically generated

The Newspaper Company also wants a period to **deploy** the campaign. During this period, you have no longer new (predictor) data coming in, nor new information about your dependent variable (churn). Therefore, this period is called the **gap**.

Shape, rectangle

Description automatically generated

You use all available information before this gap period to predict churn behaviour. These features are known as **predictors** or **independent variables**. Your **dependent variable** *depends* on your independent variables (cf. Regression analysis in Statistics 101). Therefore, we call the period where you get this information the **independent period**.

Shape, rectangle

Description automatically generated

Okay, we now completely modelled the purpose model. This is how we are going to use our model. But remember, **first we need to train our model on past data**. We want to mimic our purpose model as closely as possible. So, since we are interested in churn behaviour in the Summer, we should also look for this in our past data. We have data up until October 26, 2022. Therefore, the most recent Summer (before our purpose model) was in 2021, so we will use churn behaviour from this Summer as the dependent period of our **model building phase.** The same would hold if we were modelling churn for the entire year of 2022. We would then look for the most recent entire year for which we have data. This would be 2021.

Shape, rectangle

Description automatically generated

A likewise pattern exists for the definition for the **gap** and the **independent period**. To make sure our independent variables for churn behaviour prediction are the same as the ones we will be using in our purpose model, we also need to maintain the one month gap.

Shape, rectangle

Description automatically generated

1. A data dump is all available data up until a certain date. [↑](#footnote-ref-1)