# **AB testing**

* when choosing how to aggregate by time we need to see which one it has the same
* Sometimes the trends in a time series might be hidden by seasonality and grouped data for example by gender, by age, by country, by device, by platform
* A/B test: test 2 different scenarios to which one is best for our intent
* How to decide the test sensitivity? You need to use your experience and intuition. Look at what your current data of the kpi is now, aggregated by the users that you wnat to influence
* How to conduct the test. Calculate your kpi by the 2 groups and calculate the pvalue

## **PREDICTING CUSTOMER CHURN**

* DROP VARIABLES THAT ARE CORRELATED
* Always start with a baseline model that normally could be a logistic, then try to beat it with a new model
* For unbalanced classes then you can apply a few techniques but if you don’t fix it then you cannot use accuracy, You should understand whether recall is more important than precision and also use the f1score
* Roc curve could be also a better measure->AUC>0.5 is better than randomly guessin
* Plot it with matplotlib
* Plot precision recall curve: this curve summarises the tradeoff between tpr and fpr
  + They tend to be more informative when you have imbalanced classes
  + High area means high precision and high recall
  + F1 score quantifies the trade off

## **CUSTOMER SEGMENTATION**

* Cohort data: rows are dates and columns are transitions, The values are the number of customers who made their first translation, second trans and so on on that date
* An example is to segment customers based on the month they made their first purchase. Then we need to represent the number of months since the first transaction
* First thing we need to create cohorts. In this example we use Month Cohorts, meaning for each transaction we calculate the month since the first transaction of that same customer. Then we want to know how many customers for each cohort.
* We also want to know the offset in days (how many days have passed since each transaction)
* We can calculate COHORT BUSINESS METRICS: retention rate- how many customers from each of the cohorts have returned in the subsequent months. We can just calculate it dividing every column by the first column (100%)
* Other metrics: we can create the same thing but with quantity
* Best way to visualise it it’s a heatmap
* Recency (days since last customer transaction), frequency (nr of trans in the past 12 months) and monetary value (tot spent in last 12 months) are good metrics to calculate for a customer and segment it or use them as features.
* Easy way to segment them can be assigning quartiles separately to each metric, then concatenating them, and summing these values
* Analyse the segmentation: see nr of customers in each. If needed also maybe reduce the number of segments (in this example 3 categories : low values are called bronze, high gold, middle silver). Trial and error to find the right cutoff
* Instead of writing your cutoffs then you can cluster these measures with kmeans or others methods
* INTERPRETATION:1 Average values, 2 snake plot on standardised measures, 3 relative importance of each segments attribute (mean of the cluster/mean\_population -1) the further from 0 the more important the attribute is for defining the specific cluster