Model improvement

viettel

01- Data preparation

1. Why?

Preparing the data is required for :

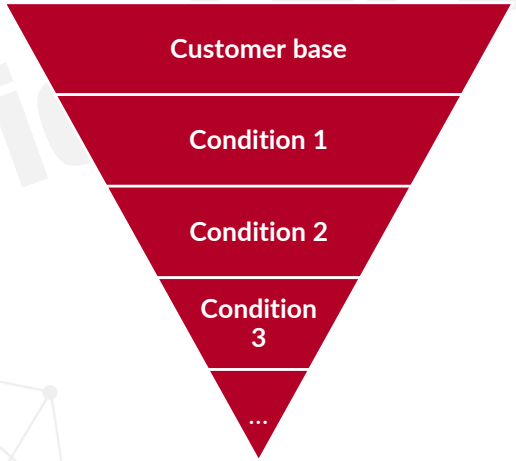
- Focusing on a **sample** which is **relevant** to the business question and getting a **meaningful** score so to improve the quality of the decision behind that sample
- Having a set of **cleaned and created features** so to catch the most from the phenomenon behind the data
- Limiting the **bias and the variance** of the model behind the data
- **Evaluating** the **data** in an independent fashion

Viettel Digital Talent 2023

2. Defining the scope

Determining the records relevant to the business question and for which the knowledge should be extracted:

- Excluding records which are irrelevant to the question (ex: customers who passed away, customers who are foreigners, old customers, customers part of specific clusters,...)
- Excluding records which might present some risks to the company (bad debtors)
- Excluding records which might cost money to the company (unprofitable customers)
- Excluding records already targeted by some marketing actions (avoiding customer harassment)



Condition	Value	no_customers	prop	no_target	prop	target_prop
customer base		10.000.000	100%	500.000	100%	5%
condition 1	point_id=1000001	8.000.000	80%	400.000	80%	5%
condition 2	status=1	7.000.000	70%	300.000	60%	4%
condition 3	partition=[20211201, 20220631]	4.000.000	40%	100.000	20%	3%

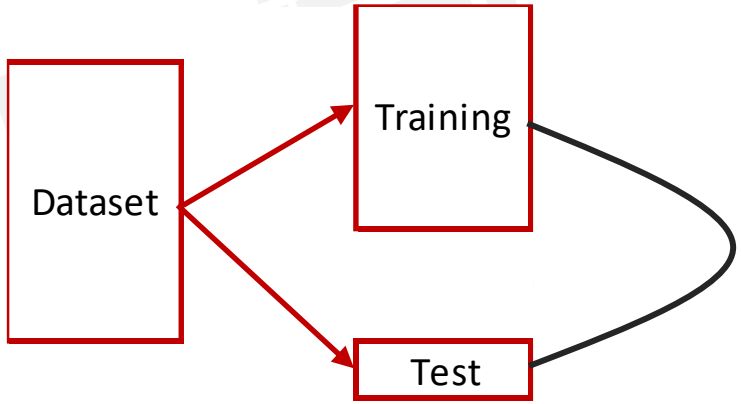
The idea is to have a target proportion similar to the record proportion, across the conditions, so to no suffer from a decreasing target proportion or to eventually consider the reason this situation would occur

Viettel Digital Talent 2023

3. Partitioning the data

Partitioning[1-2] the data into a training set and a test set:

- **Training set**: used to learn a function from the data, which usually contains around 70-90% of the records of the original dataset
- **Test set:** used to evaluate the performance of the function learned on unseen data, which contains the remaining records of the original dataset



The idea behind testing on unseen data is to get an unbiased estimation of the generalization error[3]

[1] Always prefer stratified sampling over random sampling so to preserve prior target probability between datasets

[2] Data partition is performed before any cleaning/transformation activities (value imputation, winsorization, scaling, binning, recoding,…) in order to avoid to make decisions about the training set from insights learned from the test set

[3] Using the training set for evaluation leads to an over-optimistic estimation of the generalization error (i.e. evaluation bias). But using the test set, from a data partition, is usually not sufficient to avoid the bias since the data distribution between the training and test sets is quite similar. As a result, it is advised to test the performance on independent data under which the underlying distribution is more likely to change

Viettel AI Talent 2023

4. Cleaning the data

Cleaning the data which are missing, extreme or erroneous.

Missing data refers to:

- Native language-based (R, Python,...) missing values: NA, NaN, numpy.nan, None,...
- Any default value input a priori in the database: ., -1, -10000,...

There are 3 types of missing data:

Missing type	Explanation	Cleaning method
Missing Completely At Random (MCAR)	**A feature's missing data doesn't depend neither on any other feature nor itself** => Probability of missingness is the same for all records => Assumption generally unrealistic	Removing missing (no bias) Using imputation methods
Missing At Random (MAR)	**A feature's missing data depends on other features which are observed in the dataset** => Probability of missingness varies between groups which are defined by other feature's observed data => Assumptiom somewhat plausible	Using imputation methods
Missing Not At Random (MNAR)	**A feature's missing data depends on other features which are unobserved** => Probability of missingness depends on the feature itself => Assumption often plausible	Getting more data about the cause of missingness

cost	brand	color
1000	audi	black
600	vw	.
800	audi	.
550	vw	blue
.	kia	.
.	kia	red
1000	audi	.
.	kia	.
400	vw	.

MAR MCAR

Viettel Digital Talent 2023

4. Cleaning the data

There are **2 main methods** for dealing with missing data:

- **Listwise deletion/Complete case analysis**[1-2]: deleting missing data
- **Value imputation**: replacing missing data with substituted values

cost	brand
1000	audi
600	vw
800	audi
550	vw
~~.~~	~~kia~~
~~.~~	~~kia~~
1000	audi
~~.~~	~~kia~~
400	vw

Listwise deletion

cost	brand	
1000	audi	
600	vw	
800	audi	
550	vw	
.	kia	→ 725
.	kia	→ 725
1000	audi	
.	kia	→ 725
400	vw	

Value imputation

[1] Listwise deletion is often discarded because it reduces drastically the size of the dataset

[2] If missing is MCAR, listwise deletion doesn't add any bias but decreases the power of the analysis by reducing the sample size

Viettel Digital Talent 2023

4. Cleaning the data

In case of value imputation, techniques will depend on:

- The **level of imputation**[1]: single imputation or multiple imputation
- The **dimension of data**[2]: univariate dimension or multivariate dimension
- The **type of data** for both missing and input data: qualitative data or quantitative data

There are **many value imputation techniques:**

Imputation technique	Imputation	Dimension	Missing	Input data	Explanation	Use if
Zero	Single	Univariate	Quantitative	Quantitative	**Replacing missing data with 0** Method can make sense if the record is not concerned with the feature (but data becomes skewed)	
Last Observation Carried Forward (LOCF)	Single	Univariate	Quantitative	Quantitative	**Replacing missing data with the last record-wise observed value** Usually used in longitudinal studies	
Mean/Median	Single	Univariate	Quantitative	Quantitative	**Replacing missing data with the mean or median** Mean imputation is sensitive to outliers Median imputation is robust to outliers Method might be straightforward and is very prone to bias	

[1] Single: a unique value is imputed to the missing data vs Multiple: multiple values are imputed to the missing data before taking a decision (aggregation, optimization,…)

[2] Univariate: imputation from the feature which has missing data vs Multivariate: imputation from other available features

Viettel Digital Talent 2023

4. Cleaning the data

Value imputation techniques

Imputation technique	Imputation	Dimension	Missing	Input data	Explanation	Use if
Regression[1]	Single	Multivariate	Quantitative Qualitative	Quantitative	**Replacing missing data with predicted values from a regression model trained on other features observed on complete records** Method is parametric and relies on assumptions (variance homoscedasticity, normality of residuals, linear relationship between X and Y,...) which might require data transformation Method is prone to bias as the relationship between X and Y might be not linear Method is sensitive to outliers in its non-penalized form Linear regression in case of missing quantitative data Logistic regression in case of missing qualitative data	**MCAR** **MAR**
K-nearest neighbors (KNN)	Single	Multivariate	Quantitative Qualitative	Quantitative	**Replacing missing data with averaged/modal value of the K nearest neighbors determined from the distances calculated on other features observed on complete records** Distance metric is usually either Euclidean or Mathattan Method might be sensitive to outliers if K is small Method needs data to be on the same scale (fair weight allocation) Method might be computationally expensive (distance calculation)	**MCAR** **MAR**

[1] Using other features linearly for value imputation might lead to collinearity so it is subsequently advised to use non-parametric models for estimating the function

Viettel Digital Talent 2023

4. Cleaning the data

Value imputation techniques

Imputation technique	Imputation	Dimension	Missing	Input data	Explanation	Use if
Random Forests (missForest)	Multiple	Multivariate	Quantitative Qualitative	Mixed	**Replacing missing data with predicted values from a random forests trained on non-missing data's other features, after imputing to missing data the mean (in case of quantitative feature) or the mode (in case of qualitative feature)** Method relying on an iterative process aiming at updating the imputation values until convergence is met Method is non-parametric and makes no assumption about data (no data transformation) Method might be computationally expensive (multiple trees)	**MCAR** **MAR**
Multiple Imputation by Chained Equation (MICE)	Multiple	Multivariate	Quantitative Qualitative	Quantitative	**Replacing missing data with predicted values from a GLM trained on non-missing data's other features, after imputing to missing data the mean (in case of quantitative feature) or the mode (in case of qualitative feature)** Method relying on an iterative process aiming at updating the imputation values until convergence is met Method is parametric and relies on assumptions (variance homoscedasticity, normality of residuals, linear relationship between X and Y,...) which might require data transformation Method is prone to bias as the relationship between X and Y might be not linear Linear regression in case of missing quantitative data Logistic regression in case of missing qualitative data	**MCAR** **MAR**

Viettel Digital Talent 2023

4. Cleaning the data

Value imputation techniques

Example of value imputation behind Random Forests/MICE:

Original dataset

salary	rank	age
1000	high	40
800	medium	32
.	low	.
600	low	.
.	.	26
800	**low**	**32.6**

First iteration[1]

salary	rank	age	
1000	high	40	train
800	medium	32	train
? => 500	low	32.6	predict
600	low	32.6	train
? => 400	low	26	predict

salary	rank	age	
1000	high	40	train
800	medium	32	train
800	low	32.6	train
600	low	32.6	train
800	? => low	26	predict

salary	rank	age	
1000	high	40	train
800	medium	32	train
800	low	? => 26	predict
600	low	? => 24	predict
800	low	26	train

Second iteration[2]

salary	rank	age	
1000	high	40	train
800	medium	32	train
? => 550	low	26	predict
600	low	24	train
? => 380	low	26	predict

salary	rank	age	
1000	high	40	train
800	medium	32	train
500	low	26	train
600	low	24	train
400	? => low	26	predict

salary	rank	age	
1000	high	40	train
800	medium	32	train
500	low	? => 25	predict
600	low	? => 23	predict
400	low	26	train

Final dataset

salary	rank	age
1000	high	40
800	medium	32
550	low	25
600	low	23
380	low	26

At each iteration, we get predictions for all columns with missing data, 1-by-1, while using other columns for which missing data is replaced by:

[1] The mean (quantitative data) or the mode (qualitative data)

[2] The prediction of the previous iteration

viettel

4. Cleaning the data

Missing data may cause severe consequences.

1. Missing data negatively impacts the predictive modeling activity:
   - By **creating a bias** in the data
   - By **preventing us from using some** predictive modeling **methods** (GLM,…)[1] which don't accept them

2. Missing data can be a severe problem when highly frequent :
   - Excluding missing data might **reduce massively the size of the dataset** and create a bias (sampling bias) although these data points might contains insights about the target feature
   - While imputing values to massive data might end with **poor imputation** (imputation bias)

[1] Some predictive models instead accept them and even derive insights from them (ex: decision tree, XGBoost,…)

Viettel Digital Talent 2023

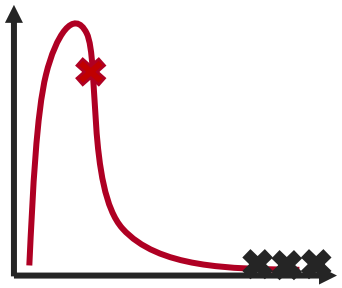
4. Cleaning the data

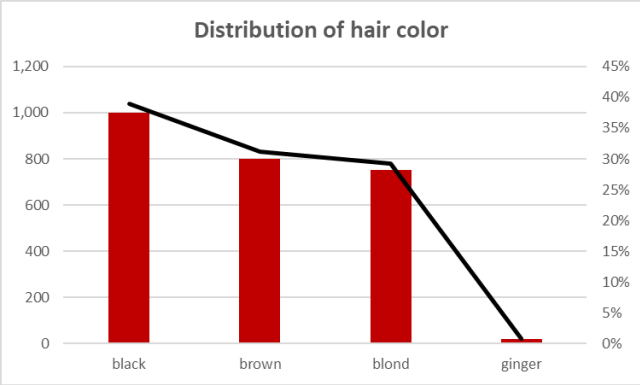
Cleaning the data which are missing, extreme or erroneous.

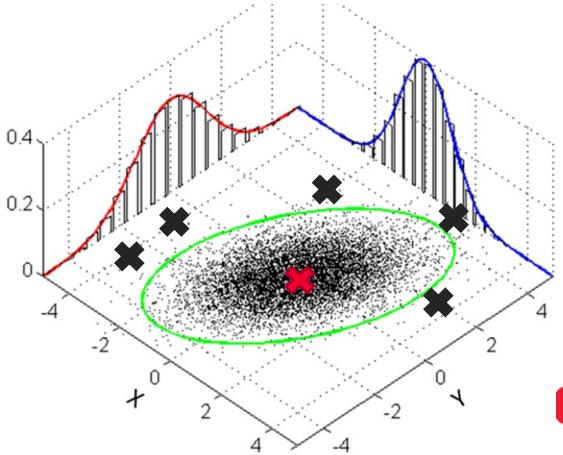
Outliers refer to extreme values that differ from the majority of observed values and result from error or fraud.

Outliers concern both :

- Quantitative data: small amount of data points with values far away from the center of the univariate distribution/the centroid of the multivariate distribution (located in the tails)
- Qualitative data: small amount of data points belonging to some specific levels







Viettel Digital Talent 2023

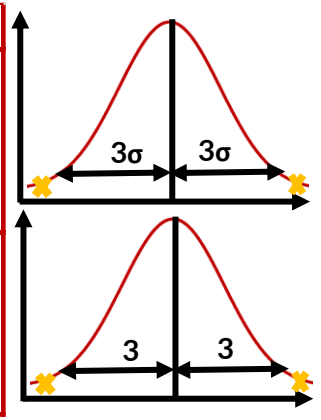
4. Cleaning the data

Outlier detection techniques will depend on:

- The **distribution of data**: normal or standardized or undefined
- The **dimension of data**[1]: univariate dimension or multivariate dimension
- The **type of data:** qualitative data or quantitative data

There are **many outlier detection techniques:**

Detection technique	Distribution	Dimension	Data type	Concept	Explanation
3σ	Normal	Univariate	Quantitative	Distance	**Points outside [μ - 3σ, μ + 3σ]** Around 0.27% of data is expected to fall outside this interval, more reflect outliers Method not robust to non-normality
Z score	Standardized	Univariate	Quantitative	Distance	**Points outside [-3, 3] after standardization** Around 0.27% of data is expected to fall outside this interval, more reflect outliers Method not robust to non-normality



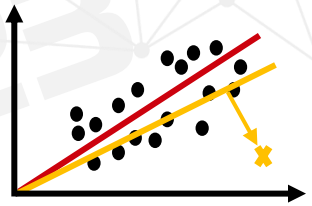
[1] Univariate: outlier detection in the univariate space vs Multivariate: outlier detection in the multivariate space (ex: so a point might not be a univariate outlier for any of 3 features but become one while combining them)

Viettel Digital Talent 2023

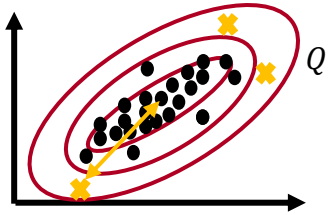
4. Cleaning the data

Outlier detection techniques

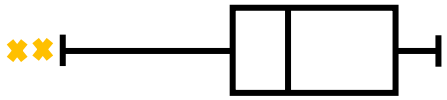
Detection technique	Distribution	Dimension	Data type	Concept	Explanation
Cook's distance[1]	Normal	Multivariate	Quantitative	Influence	**Points with distance D_i > 1 or 4/N** Since residuals are involved in the metric, linear regression needs to be run Method relies then on linear model assumptions (variance homoscedasticity, normality of residuals, linear relationship between X and Y,...)
Mahalanobis distance[2]	Undefined	Multivariate	Quantitative	Distance	**Points with distance beyond a specific percentile of the distribution of the distance $Chi2_{p-1,\ 97.5\%}$: $D_i > Chi2_{p-1,\ 97.5\%}$** Method is scale invariant
Whiskers	Undefined	Univariate	Quantitative	Distance	**Points beyond the whiskers** Whiskers may be defined according to different rules: **- closest values within [μ - σ, μ + σ]** **- closest values within some percentile [P1%, P99%]** **- closest values within [Q_1 - 1.5 IQR, Q_3 + 1.5 IQR]** IQR-based and percentile-based methods robust to non-normality



$$D_i = \frac{\sum_{j=1}^{N}\left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{MSE}$$



$$d(X_i, Q) = \sqrt{(X_i - \mu)^t S^{-1} (X_i - \mu)}$$



[1] Cook's distance measures the influence of a data point i by looking at how much the regression slopes (parameters) changes while removing it

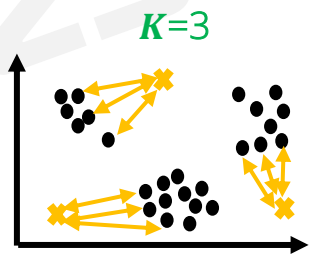
[2] Mahalanobis distance measures the number of standard deviations that a data point is from the center/centroid μ of a distribution Q (S being the covariance matrix)

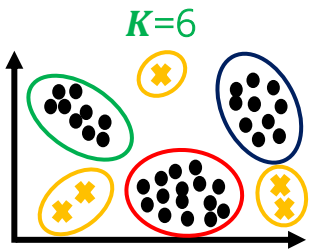
Viettel Digital Talent 2023

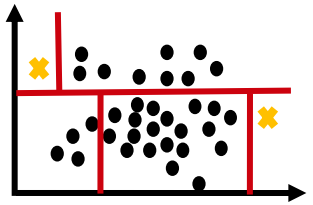
4. Cleaning the data

Outlier detection techniques

Detection technique	Distribution	Dimension	Data type	Concept	Explanation
Local Outlier Factor[1]	Undefined	Multivariate	Quantitative	Density (Distance)	**Points whose local density is substantially lower than their K closest neighbors's average local density: LOF_i higher than 1**[2] Distance metric is usually either Euclidean or Manhattan Since LOF is a ratio of neighbor's average local density to local density, threshold rules might be difficult to set for ratios (they depend on the data)
Clustering	Undefined	Multivariate	Quantitative Qualitative	Distance	**Points part of a specific cluster which is small and far from the center of gravity of cluters** Distance is usually Euclidean (quantitative data) but Gower might be reauired (qualitative or mixed data) Method might be computationally expensive (distance calculation) Method needs data to be decorrelated (unique information) Method needs data to be on the same scale (fair weight allocation)
Random Forests (IForest)[3]	Undefined	Multivariate	Quantitative Qualitative	Distance Isolation	**Points part of 1 record-leaves which are isolated and have the shortest path length: s_i close to 1**[2] Method is non-parametric and makes no assumption about data (no data transformation) Method might be computationally expensive (multiple trees)
Frequency	Undefined	Univariate	Qualitative	Density	**Points with level frequency below 5%** **Points with level concerned with less than 10 targets**







[1] The higher the distance to its closest neighbors, the lower the local density to these neighbors (link between distance and density)

[2] See research papers for details of the quantities

[3] Isolation Forest recursively partitions the sample by randomly selecting an attribute and then randomly selecting a split value in order to reach 1-record leaves and identify records far from the mass (isolation) for which the underlying leaf is built with the minimum number of split decisions (shortest path length)

Viettel Digital Talent 2023

4. Cleaning the data

There are **3 main methods** for dealing with outliers once identified:

- **Truncation/Trimming**: discarding outliers
- **Winsorization**: replacing the outliers with the nearest "non suspicious" data point's value
- **Transformation**[1]: transforming the feature for smoothing the data points

In case of **winsorization**, proposed values for replacement depend on the outlier detection technique:

Detection technique[2]	Winsorizing values
3σ	Outliers in the left tail: $\mu - 3\sigma$ Outliers in the right tail: $\mu + 3\sigma$
Z score	Outliers in the left tail: -3 Outliers in the right tail: 3
Whiskers	- Outliers in the left tail: $\mu - \sigma$, outliers in the right tail: $\mu + \sigma$ - Outliers in the left tail: P1%, outliers in the right tail: P99% - Outliers in the left tail: Q_1 - 1.5 IQR, outliers in the right tail: Q_3 + 1.5 IQR
Frequency	Level recoding

[1] See infra for data transformations

[2] Winsorization is only proposed for univariate detection techniques

Viettel Digital Talent 2023

4. Cleaning the data

Outliers may cause severe consequences.

1. Outliers negatively impact the predicting modeling/multivariate analysis activity:
   - By **increasing the variance** of the estimated function and thus leading to overfitting
   - By **leaving more weights** to the concerned **high magnitude features** in some analyses (PCA, clustering,...)

2. Outliers can be a severe problem when too frequent :
   - Removing outliers (=truncation/trimming) might **reduce the size of the dataset** and create a bias (sampling bias) although these data points might contains insights about the target feature
   - While replacing values to massive data might end with **poor value replacement**

Viettel Digital Talent 2023

5. Transforming the data

Transforming the data so they better fit the modeling algorithm.

There are **5 different methods** for transforming the data:

- **Scaling**: Giving quantitative data a new scale
- **Linear/non-linear transformation**: Creating a linear or non-linear transformation of quantitative data
- **Encoding**: Encoding qualitative data with numeric values
- **Binning**: Creating bins from quantitative data
- **Recoding**: Aggregating levels of qualitative data

Viettel Digital Talent 2023

5. Transforming the data

Data transformation techniques will depend on:

- The **purpose of transformation**: distribution[1] or scale[2] or type[3] of the data
- The **underlying method**: scaling or transforming or encoding or binning or recoding

There are **many data transformation techniques** :

Technique	From	To	Change	Explanation
Min-max normalization	Quantitative	Quantitative	Scale	**Scaling the range into [0, 1]** Method guarantees same scale across the features but doesn't smooth the outliers
Z-score normalization/ Standardization	Quantitative	Quantitative	Scale	**Scaling the range so the sample mean is 0 and standard deviation is 1** Method smooths the outliers but doesn't guarantee same scale across the features Method not robust to non-normality

$$X' = \frac{X - \min(X)}{\max(X) - \min(X)}$$

$$Z = \frac{X - \bar{X}}{\sigma}$$

[1] Some predictive models require data to be symmetric of to follow a normal distribution, otherwise assumptions behind are violated and those models are no more valid

[2] Some predictive models need data to be scaled, especially for distance-based algorithms, otherwise high magnitude features get more weight in the training process

[3] Some predictive models require data to be quantitative, otherwise they don't work

Viettel Digital Talent 2023

5. Transforming the data

Data transformation techniques

Technique	From	To	Change	Explanation
Linear transformation[1]	Quantitative	Quantitative	Distribution Scale	**Keeping the linear relationship between 2 features** Useful for correlation analysis
Non-linear transformation[2]	Quantitative	Quantitative	Distribution Scale	**Making distributions more symetric, even normal** A non-linear transformation changes the linear relationship between features and thus the correlation
Tukey's ladder of powers[3]	Quantitative	Quantitative	Distribution Scale	**Re-expressing the relationship between 2 features so it becomes linear** The tuning parameter λ is choosed so to maximize the Pearson correlation coefficient between the features Method cannot handle 0 and negative values depending on the tuning parameter λ

$$Y = \begin{cases} X^{\lambda} & if\ \lambda > 0 \\ \ln X & if\ \lambda = 0 \\ -(X^{\lambda}) & if\ \lambda < 0 \end{cases}$$

[1] A transformation f is linear iif: $f(x+y) = f(x) + f(y) \quad \forall x, y$

[2] See infra for examples of non-linear transformations

[3] Tukey's ladder of powers is a powerful technique that relies on power transform (family of power transformations) which helps at linearizing the relation between 2 features

Viettel Digital Talent 2023

5. Transforming the data

Data transformation techniques

Technique	From	To	Change	Explanation
One-hot encoding	Qualitative	Quantitative	Type	**Encoding L levels on L-1 boolean features** Method might lead to a high feature space and disregard the ordinality Method useful for nominal data (not ordinal)
Label encoding	Qualitative	Quantitative	Type	**Encoding L levels numerically from 0 to L-1** Method ensures a small feature space but imposes ordinality Method useful for ordinal data (not nominal)
Target encoding	Qualitative	Quantitative	Type	**Encoding levels with conditional target probability** Method ensures a small feature space but is prone to overfitting Method contains target-related information
Leave-1-out encoding	Qualitative	Quantitative	Type	**Encoding levels with conditional target probability while excluding the current record** Method ensures a small feature space and is less prone to overfitting Method contains target-related information

Viettel Digital Talent 2023

5. Transforming the data

Data transformation techniques

Technique	From	To	Change	Explanation
Width-based binning	Quantitative	Qualitative	Type	**Binning into bins of same width**
Size-based binning/ recoding	Both	Qualitative	Type	**Binning/recoding into bins of same size**[1] Percentiles might be used for binning
User-based binning/ recoding	Both	Qualitative	Type	**Binning/recoding based on user experience or business knowledge**
Target-based recoding	Qualitative	Qualitative	Distribution	**Aggregating levels whose target probability is similar**
k-score and Z-score recoding	Qualitative	Qualitative	Distribution	**Aggregating levels which don't meet k-score threshold** (level frequency < 5% or target count < 10) **and Z-score proportion test** (target probability significantly lower than prior target probability) **in a trashbin level**

[1] Same size implies same number of records

Viettel Digital Talent 2023

5. Transforming the data

Non-linear transformations

Transformation[1]	Use if	Limitations
Square/cube root	Feature with frequency counts (poisson distribution) Positively skewed distribution	Square root cannot handle negative values
Logarithmic	Positively skewed distribution	Cannot handle 0 and negative values
Power	Negatively skewed distribution	N/A
Inverse	Platykurtic distribution	Cannot handle 0
Arcsine	Feature with proportions or counts (not percentages)	Cannot handle absolute values > 1
Box-Cox[2]	Unknown distribution	Cannot handle 0 and negative values depending on the tuning parameter λ

$$Y^{\lambda} = \begin{cases} \dfrac{Y^{\lambda} - 1}{\lambda} & if\ \lambda \neq 0 \\ \ln Y & if\ \lambda = 0 \end{cases}$$

[1] We add sometimes 1 to the feature before transforming so to avoid impossible calculation with null values => especially, for logarithmic transformation, it makes null values remain null and values inside $]0, 1[$ positive rather than negative

[2] Box-Cox is a powerful technique that relies on power transform (family of power transformations) which might help at making a distribution more normal and/or stabilizing the variance

Viettel Digital Talent 2023

5. Transforming the data

Keeping the data untransformed may cause severe consequences.

1. Skewed or non-normal distributions negatively impact the predictive modeling activity:
   - By making some **predictive modeling methods** (GLM,…)[1] **unreliable** and preventing us from using them
2. Unscaled data negatively impacts the predictive modeling/multivariate analysis activity :
   - By leaving **more weights to the concerned high magnitude features** (PCA, clustering, KNN,…)

[1] Non-parametric models instead don't make strong assumptions about the distribution and can be used (ex: tree-based models, non-linear SVM,…)

Viettel Digital Talent 2023

6. Creating new features

Creating features[1] which we assume to be somehow correlated with the target feature.

Usually, there are **3 types of features**[2]:

- **Recency**: How recent is an event X[3]?
- **Frequency**: How many events X[3] occurred within a period of time?
- **Monetary**: What is the amount related to the event X[3]?

Viettel Digital Talent 2023

[1] Features of interest depend on the business problem: customer churn, cross-sell,...

[2] These features refer to "RFM" framework and are quantitative, but qualitative features should be considered too (demographics,...)

[3] Event might concern: a phone transaction (call, SMS, data usage), the purchase of a product, a complain, internet traffic, interactions with POS,...

6. Creating new features

Relying on existing features may have severe consequences.

1. Relying only on existing features negatively impacts the business understanding:
   - By **loosening our skills** to proceed extra activities **to derive insights from the data and serve the business**

2. Relying only on existing features negatively impacts the predictive modelling acitivity:
   - By **feeding input features** to the model that might **not catch the pattern behind the data**[1]

Viettel Digital Talent 2023

[1] The more features we create, the more chance to find one(s) that is/are of high importance to the predictive model

7. Excluding features

Excluding features which are not part of the problem:

- Features with important missing data
- Qualitative features with highly imbalanced levels[1]
- Features highly correlated together[2]
- Features uncorrelated with the target feature[3]
- Features with leakage from the future[4]
- Specific features (related to campaign, demographics,…)[5]

Viettel Digital Talent 2023

[1] These features may trigger overfitting

[2] Excluding features with correlation coefficients higher than a specific threshold (ex: 99%) or whose p-value is lower than some risk (ex: 5%) to get rid of redundant information

[3] Excluding features with correlation coefficients lower than a specific threshold (ex: 2%) or whose p-value is higher than some risk (ex: 5%) to get rid of useless information

[4] Excluding features which are subsequent to the target event and which leaked into the training set

[5] Excluding features we don't want to have an effect on the predictive modeling (campaign-free, gender-free,…)

7. Excluding features

Keeping useless features may have severe consequences.

1. Extremely unbalanced features negatively impact the predictive modelling activity:
   - By **increasing the variance** of the estimated function and thus leading to overfitting

2. Features with leakage from the future negatively impact the predictive modelling activity:
   - By **creating an evaluation bias**

3. Useless features negatively impact the computational cost behind data processing activities :
   - By **increasing the required resources and time** to reach the final desired output

Viettel Digital Talent 2023

8. Resampling the data

Resampling the dataset to make the target data smaller or more balanced[1].

Usual resampling techniques[2] are:

Technique	Comments
Random oversampling	**Sampling records randomly with replacement[3] from the minority class and adding them to the dataset** Method likely to increase the computational cost of predictive models (more records) Method likely to lead to overfitting and bias
Random undersampling	**Sampling records randomly from the majority class and removing them to the dataset** Method likely to decrease the computational cost of predictive models (less records) Method likely to lead to overfitting and bias

[1] In a binary classification problem, because the minority class is usually the class for which the predictions are the most important, it is thus difficult to get insights from it

[2] Usually, resampling techniques involve both bias and overfitting as the data distribution of X in the training set is changed, by creating/removing artificially records, and thus differs from the one in the test set

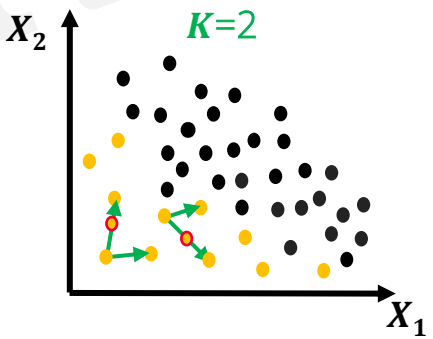
[3] See Annex 10 about Bootstrapping for further explanation

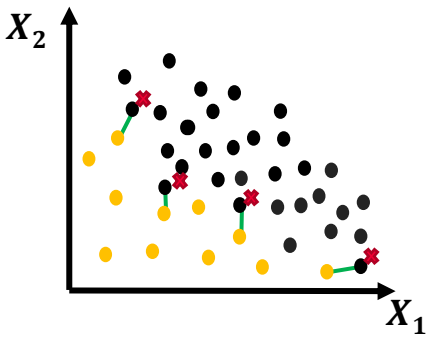
Viettel Digital Talent 2023

8. Resampling the data

Usual resampling techniques

Technique	Comments
Synthetic Minority Oversampling TEchnique (SMOTE)[1-2]	**Selecting randomly 1 of the each K nearest minority class neighbors of bootstrap samples from the minority class and creating a synthetic record from a point selected randomly between the 2 records in the feature space, adding the synthetic record to the dataset** Distance metric is usually either Euclidean or Mathattan Method might be sensitive to outliers if K is low Method needs data to be on the same scale (fair weight allocation) Method computationally expensive (distance calculation) and likely to increase the computational cost of predictive model (more records) Method likely to lead to overfitting and bias
Tomek Links[2]	**Selecting pairs of records which are each other's nearest neighbor and are from opposite class, removing the record of the majority class from the dataset** Distance metric is usually either Euclidean or Mathattan Method needs data to be on the same scale (fair weight allocation) Method might be computationally expensive (distance calculation) but likely to decrease the computational cost of predictive model (less records) Method likely to lead to overfitting and bias





[1] SMOTE leverages the idea of random oversampling without ending with exact copies of records from the minority class; the method refers then to data augmentation as new records are synthetized from existing records

[2] Empirical results show that combining SMOTE and Tomek Links achieve better performance

Viettel Digital Talent 2023

8. Resampling the data

Not resampling the data may cause severe consequences[1].

1. Too big data negatively impacts the computational cost behind data processing activities:
   - By **increasing the required resources and time** to reach the final desired output
2. Unbalanced data negatively impacts the predictive modeling activity:
   - By **training models** that either **perform poorly** on data **or that overfit**

Viettel Digital Talent 2023

[1] But resampling techniques also involve negative consequences: bias and overfitting

02- Hyper-parameter tuning

1. Hyper-parameter tuning

Refer to presentation about tree-based models

Viettel Digital Talent 2023

03- Evaluation

1. Why?

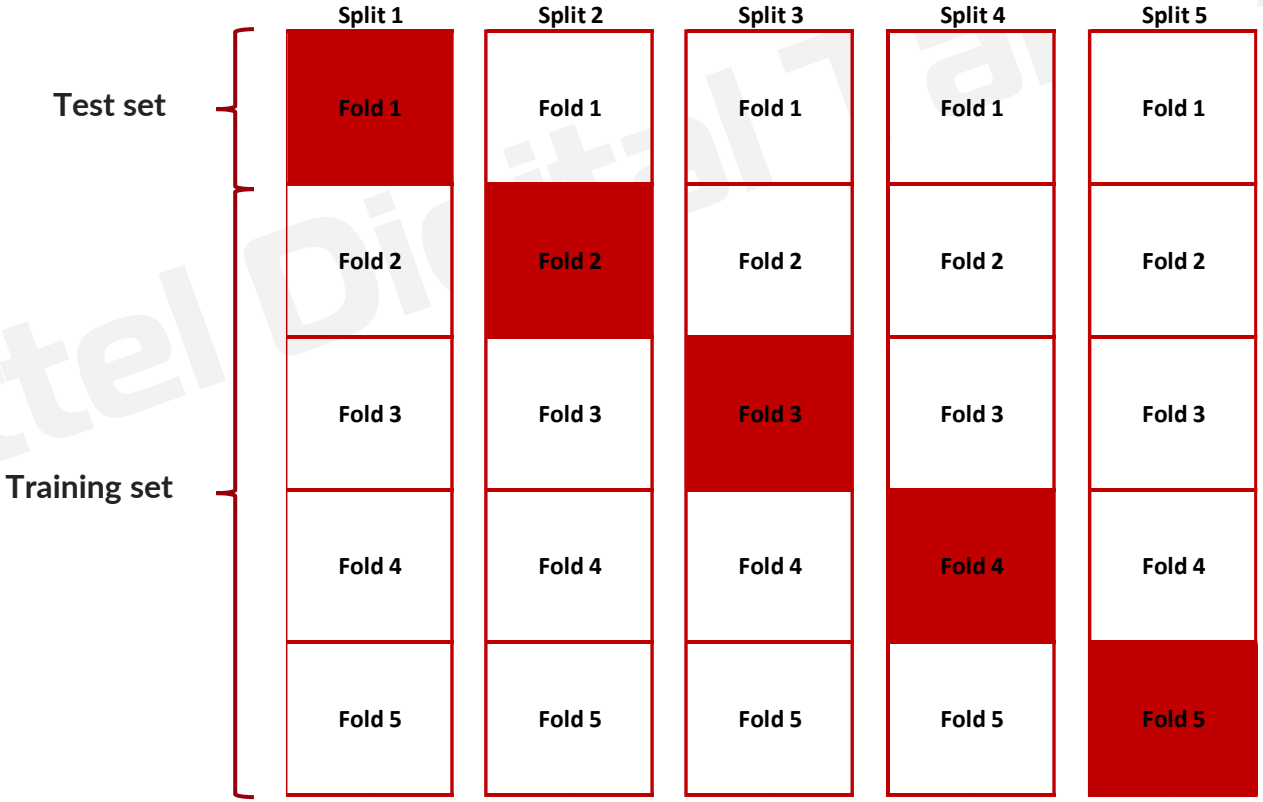
Evaluation is required for :

- **Testing the ability of the model** to predict new data which was not used for estimating the function
- Getting insights about **how well the model will generalize** to an independent dataset (overfitting)
- Getting insights about the set of **hyper-parameter's optimized values**, i.e. values that minimizes the test error

Viettel Digital Talent 2023

2. K-fold cross-validation

K-fold cross-validation is a **resampling method** that splits the dataset into K different portions of equal size and **rotates K times** over them to train a predictive model on $K-1$ portions and test it on the remaining portion



Viettel Digital Talent 2023

2. K-fold cross-validation

The **procedure of K-fold cross-validation** is as following, given a set of hyper-parameters θ:

1) Fit a predictive model $\hat{F}_k^\theta$ on the training set while excluding the kth portion
2) Test the model on the kth portion
3) Calculate the cross-validation error of the kth portion: $CVE_k^\theta = \sum_{i \in D_k}^{N} \frac{K}{N} Loss(\hat{F}_k^\theta)$
4) Calculate the cross-validation error of the model[1]: $CVE^\theta = \frac{1}{K} \sum_{k=1}^{K} CVE_k^\theta$
5) Calculate the standard error[2] of the cross-validation errors along the K portions: $SE^\theta = \sigma_{CVE_k^\theta}$
6) Choose θ^* that minimizes CVE^θ
7) Fit the model $\hat{F}^{\theta^*}$ on the training set and test it on the test set

[1] The cross-validation error is an unbiased estimation of the test error but the test error, resulting from a data partition, is usually over-optimistic because the data distribution between the training and test sets is quite similar. As a result, it is advised to test the performance on independent data under which the underlying distribution is more likely to change

[2] Standard error of a statistic is the standard deviation of its sampling distribution (=sample statistic's distribution): it gives an idea of the dispersion of a sample statistic around its mean

Viettel Digital Talent 2023

2. K-fold cross-validation

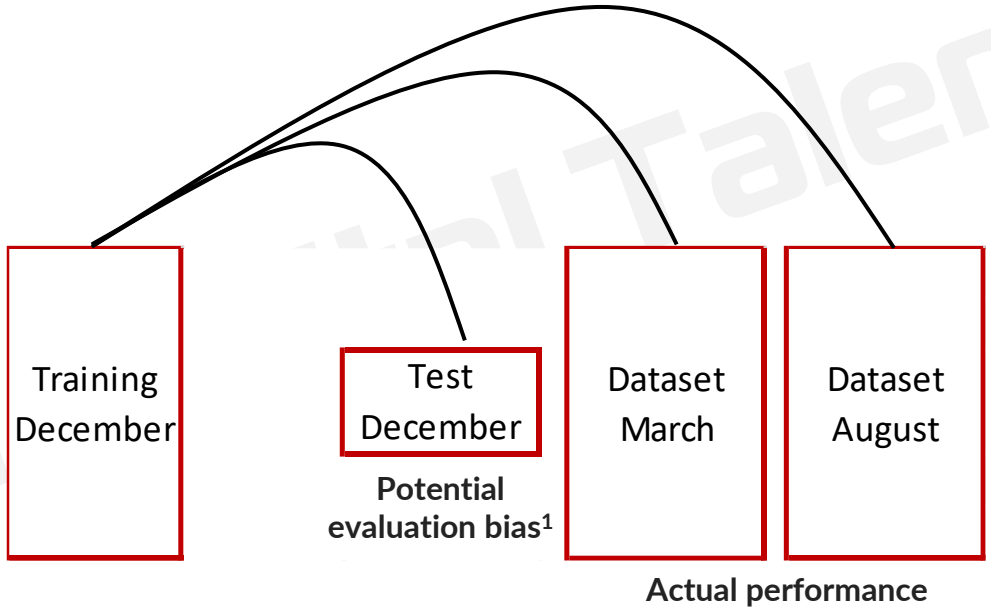
The $\boldsymbol{K}$-fold cross-validation is used for:

- Identifying values of parameters θ^* that minimizes the cross-validation error CVE^{θ}
- Getting insights about the variance of the error SE^{θ} and then overfitting risks
- Ensuring that the predictive model is tested on the whole dataset

Viettel Digital Talent 2023

3. Independent test set

The **independent test set** is used to evaluate the performance of the learned function on an **unseen and independent data** in order to avoid the evaluation bias.



[1] Using the test set, from a data partition, is usually not reliable since the data distribution between the training and test sets is quite similar. As a result, it is advised to test the performance on independent data under which the underlying distribution is more likely to change (usually another month)

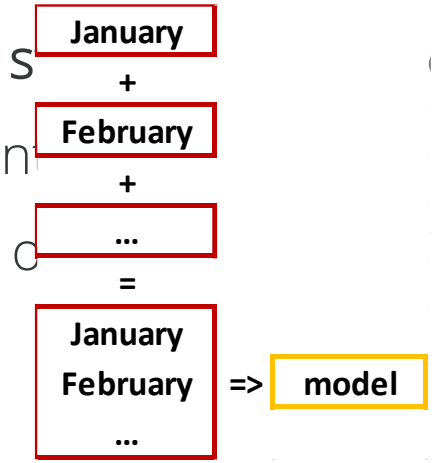
Viettel Talent 2025

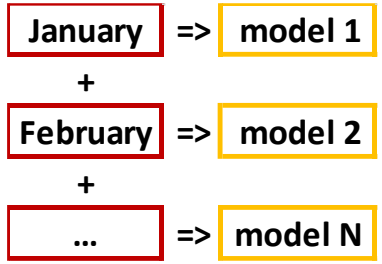
3. Independent test set

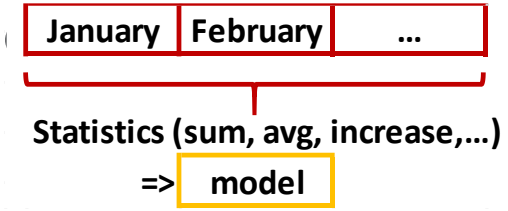
In case of seasonality either in the input features or in the output feature, the performance of the predictive model might fluctuate over months and so does the model overfit.

Some solutions include:

- Either estimating a function every month or at least for a group of successive months sharing some stability
- Or smoothing input features through s
  some period of time (3 months, 6 mon
- Or data appending over some period o







[1] Warning: records won't be assumed independent anymore so non parametric models will further be required

Viettel Digital Talent 2023

4. Evaluation metrics

Evaluation metrics are used to **quantify the performance of a predictive model,** while this model is trained using a loss function.

The **choice of metrics**[1] will depend on:

- **The type of the predictive task:** regression or classification task
- **The KPI's behind the project**[2]
- **The distribution of the output feature:** skewed[3] or symmetric distribution

Viettel Digital Talent 2023

[1] If we choose the wrong metric to evaluate predictive models, we might evaluate them improperly and, as a result, guide the predictive modeling improperly

[2] Which are determined, based on an agreement with the stakeholders

[3] In case of a classification task, skewed data refers to imbalanced data

4. Evaluation metrics

Common evaluation metrics

Metric	Task	Informativeness
Accuracy[1]	Classification	Focussing on majority class
TPR/Recall/Sensitivity[2]	Classification	Focussing on positive class
FNR/Missout	Classification	Focussing on positive class
TNR/Specificity	Classification	Focussing on negative class
FPR/Fallout	Classification	Focussing on negative class
Precision[2]	Classification	Focussing on positive class
Balanced accuracy[3]	Classification	Focussing on all classes
F_β score	Classification	Focussing on positive class
G-mean/Fowlkes-Mallows index	Classification	Focussing on positive class
Matthews correlation coefficient[3]	Classification	Focussing on all classes
ROC curve & AUC	Classification	Focussing on all classes
Cumulative Gains curve	Classification	Focussing on positive class
Precision-Recall curve	Classification	Focussing on positive class
MSE	Regression	N/A
RMSE	Regression	N/A
R^2	Regression	N/A

[1] Accuracy is not considered informative as it only focusses on the majority class and is misleading in case of imbalanced data

[2] Recall and precision only consider the positive class, as do the metrics which rely on them (G-mean, F score,…)

[3] Balanced accuracy and Matthews correlation coefficient are more informative because they consider all classes

Viettel Digital Talent 2023

4. Evaluation metrics

Common evaluation metrics

Most classification metrics rely on the **confusion matrix**[1], which is a 2x2 contingency table confronting actual class vs predicted class:

	$\hat{y} = 0$	$\hat{y} = 1$	
$y = 0$	N_{00}	N_{01}	$N_{0.}$
$y = 1$	N_{10}	N_{11}	$N_{1.}$
	$N_{.0}$	$N_{.1}$	N

- $Accuracy = \frac{N_{00}+N_{11}}{N}$
- $TPR^2 = \frac{N_{11}}{N_{1.}}$
- $FNR^2 = \frac{N_{10}}{N_{1.}}$
- $TNR^3 = \frac{N_{00}}{N_{0.}}$
- $FPR^3 = \frac{N_{01}}{N_{0.}}$

[1] Only the binary case is considered, with the minority/majority class representing the positive/negative records respectively

[2] TPR refers to the power $1-\beta$, the capacity to detect positive records, while FNR refers to type 2 error β so $TPR + FNR = 1$

[3] TNR refers to the confidence $1-\alpha$, the capacity to detect negative records, while FPR refers to type 1 error α so $TNR + FPR = 1$

[2-3] We wish we could maximize both TPR and TNR but, in reality, it is often impossible (see Annex 4 about the trade-off for explanation)

Viettel Digital Talent 2023

4. Evaluation metrics

Common evaluation metrics

Most classification metrics rely on the **confusion matrix**, which is a 2x2 contingency table confronting actual class vs predicted class:

	$\hat{y} = 0$	$\hat{y} = 1$	
$y = 0$	N_{00}	N_{01}	$N_{0.}$
$y = 1$	N_{10}	N_{11}	$N_{1.}$
	$N_{.0}$	$N_{.1}$	N

- $Recall = \frac{N_{11}}{N_{1.}}$
- $Precision = \frac{N_{11}}{N_{.1}}$
- $Balanced\ accuracy = \frac{recall + TNR}{2}$
- $F_\beta\ score^1 = (1+\beta^2)\frac{precision * recall}{(\beta^2 * precision) + recall}$
- $G - mean^2 = \sqrt{precision * recall}$
- $Matthews\ coefficient^3 = \frac{(N_{11} * N_{00}) - (N_{10} * N_{01})}{\sqrt{(N_{11}+N_{01}) * (N_{11}+N_{10}) * (N_{00}+N_{01}) * (N_{00}+N_{10})}}$

[1] $F_\beta\ score$ is the harmonic mean of $recall$ and $precision$ and uses a parameter β to determine how many times $recall$ is more important than $precision$; in case of $\beta = 1$, $recall$ and $precision$ receive equal importance

[2] G-mean is the geometric mean of $recall$ and $precision$

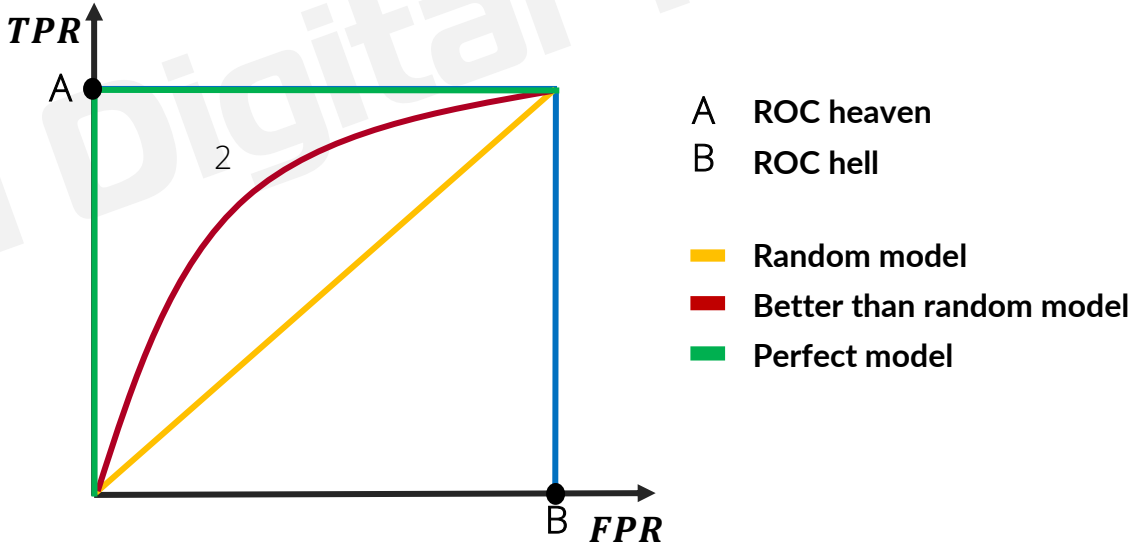
[3] Matthews correlation coefficient, ranging from -1 to 1, measures how much the actual class's records are correlated with predicted class's records

Viettel Digital Talent 2023

4. Evaluation metrics

Common evaluation metrics

The ROC curve (short for Receiver Operating Characteristics) is a graphical representation plotting $TPR(s)$ against $FPR(s)$, after sorting decreasingly the records by their scoring probabilities and varying the score threshold s from 1 to 0[1].



[1] While decreasing s, more records are positively classified so both FPR and TPR increase (see Annex 4 about the trade-off for explanation), but we prefer TPR to increase over FPR; A random model is one that discovers as many true positive cases as false positive cases

[2] While comparing models, we used to consider the ones which are forming the convex hull, that is a geometric construction of ROC curves which are above all others for a quantity $FPR(s)$

Viettel Digital Talent 2023

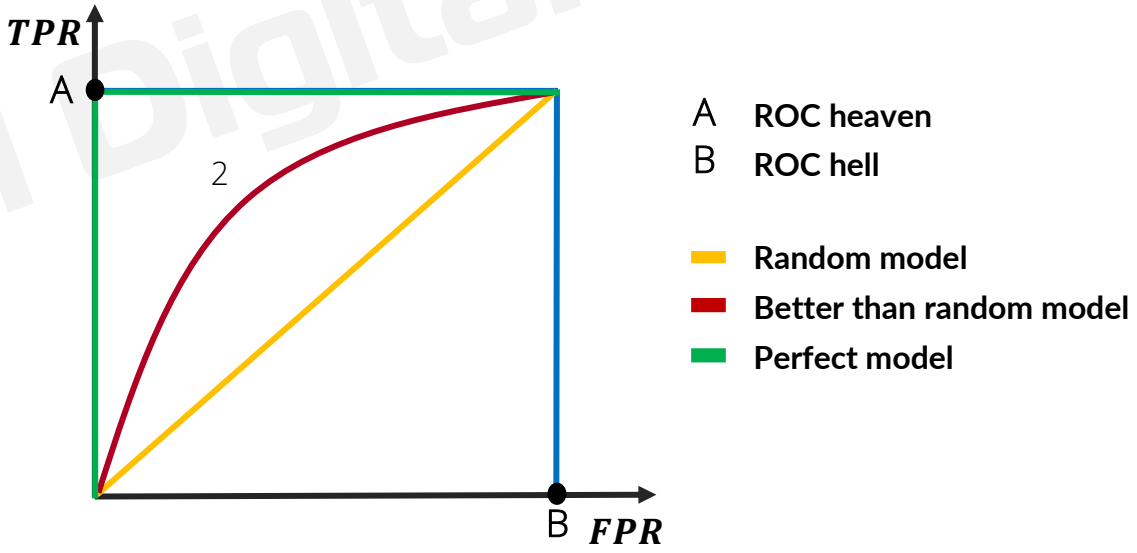
4. Evaluation metrics

Common evaluation metrics

Classification metrics

The AUC[1] (short for Area Under roc Curve), measured by the surface below the curve, estimates the probability that a positive record randomly sampled gets a higher scoring probability than a negative record randomly sampled.

$$AUC = P\left(S(X_{i,1}) > S(X_{j,0})\right)$$



[1] AUC is usually preferred over the ROC curve because it summarizes the information into a single number and makes the comparison between models easier

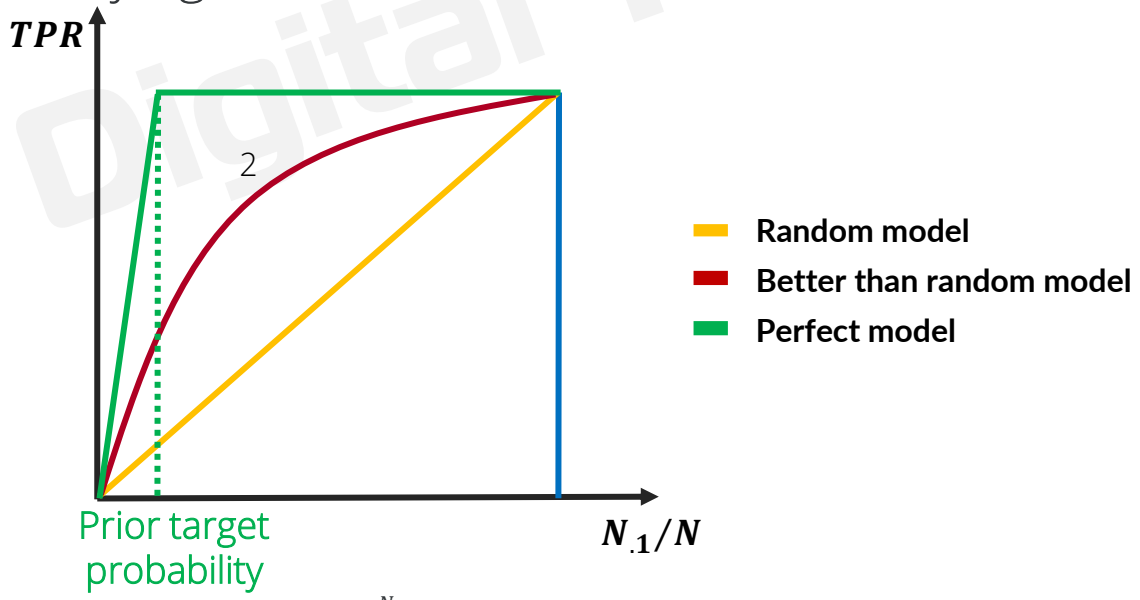
[2] While random model has an AUC of 50%, perfect model has an AUC of 100%; the higher the AUC, the better

Viettel Digital Talent 2023

4. Evaluation metrics

Common evaluation metrics

The Cumulative Gains curve is a graphical representation plotting $TPR(s)$ against the probability of positive prediction $N_{.1}/N\ (s)$, after sorting decreasingly the records by their scoring probabilities and varying the score threshold s from 1 to 0[1].



[1] While decreasing s, more records are positively classified so both $\frac{N_{.1}}{N}$ and TPR increase, but we prefer TPR to increase over FPR in $N_{.1}$; A random model is one that discovers as many true positive cases as false positive cases

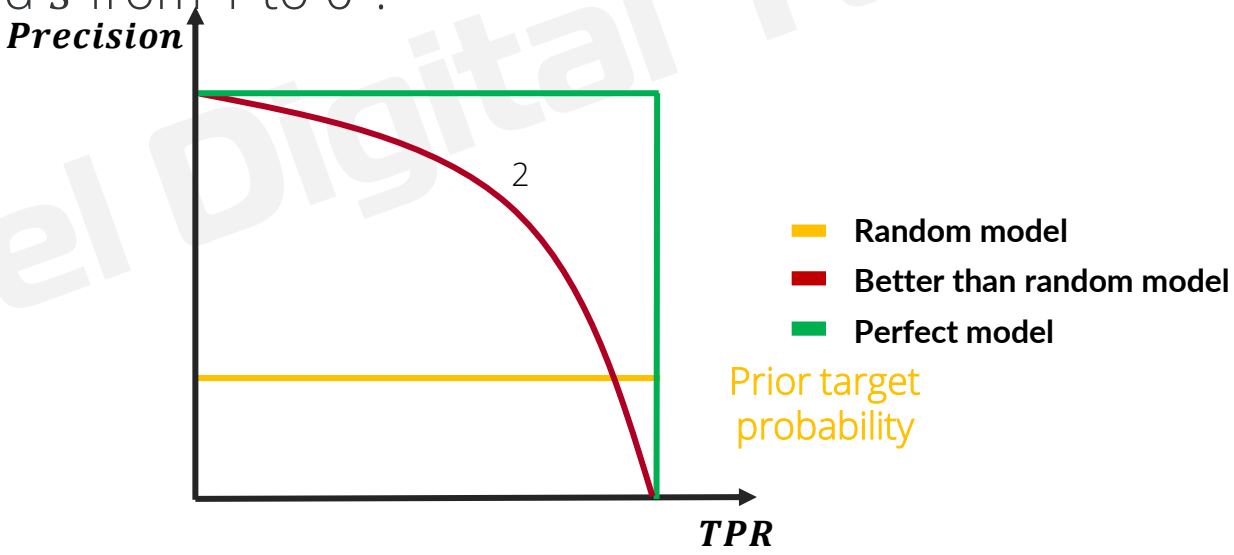
[2] While comparing models, we used to consider the ones which have a higher TPR at some cut-off values along the abscissa

Viettel Digital Talent 2023

4. Evaluation metrics

Common evaluation metrics

The Precision-Recall curve is a graphical representation plotting $Precision(s)$ against the $TPR(s)$, after sorting decreasingly the records by their scoring probabilities and varying the score threshold s from 1 to 0[1].



[1] While decreasing s, more records are positively classified so TPR increases but $Precision$ decreases; A random model is one whose $Precision$ corresponds to prior target probability

[2] While comparing models, we used to consider the ones which have a higher $Precision$ at some cut-off values along the abscissa

Viettel Digital Talent 2023

4. Evaluation metrics

Common evaluation metrics

Regression metrics[1]

Metrics include:

- $MSE = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2$
- $RMSE^2 = \sqrt{MSE}$
- $R^{2\,3} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} = \frac{SSR}{SST} = \frac{\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$

Viettel Digital Talent 2023

[1] There are many more regression metrics, but they are not covered since most marketing/sales use cases are classification-based

[2] While MSE refers to squared errors, $RMSE$, which simply is the average error, refers to errors and is more usually used for evaluation

[3] R^2measures the amount of variability in the data explained by the model; although this criterion is taught in schools, it never should be used for evaluation as it doesn't measure the goodness of fit (metric can be low/high although the model is correct/incorrect respectively)