XII CONGRESSO NAZIONALE degli ATTUARI

Cross-Selling

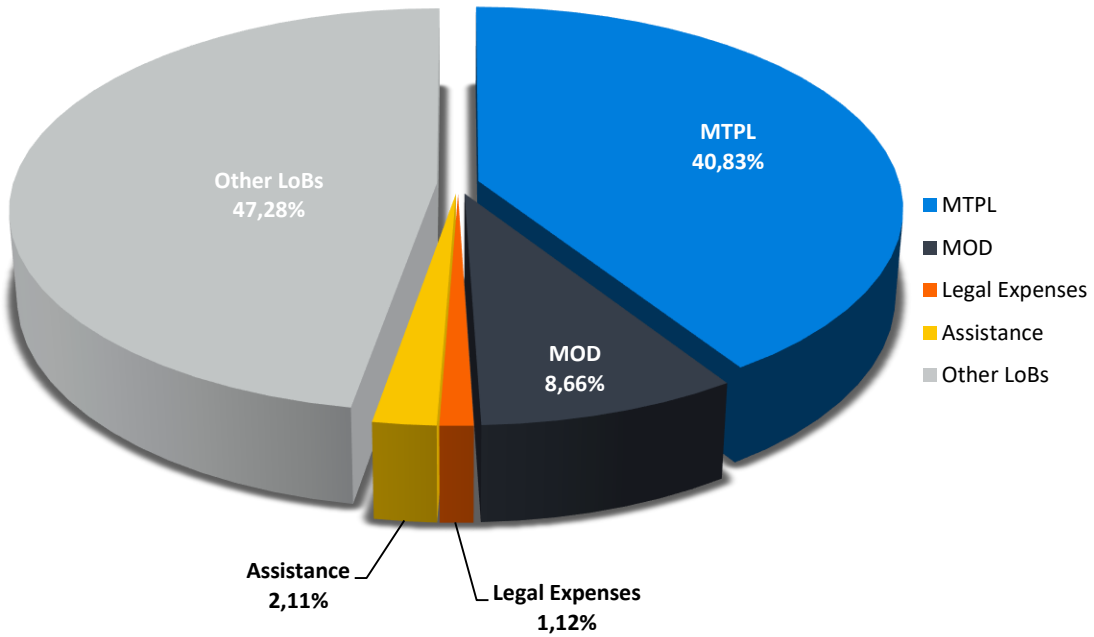
Machine Learning in campo assicurativo

Nicola Biscaglia – Alessandro Zanetti

23 novembre 2018

Mercato assicurativo italiano – *Non Life business*

% Premi contabilizzati (2017)

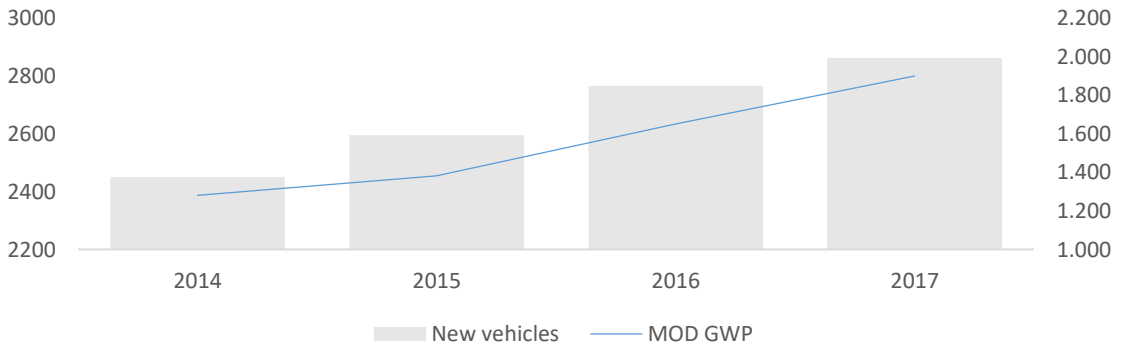


Fonte: ANIA Report 2017

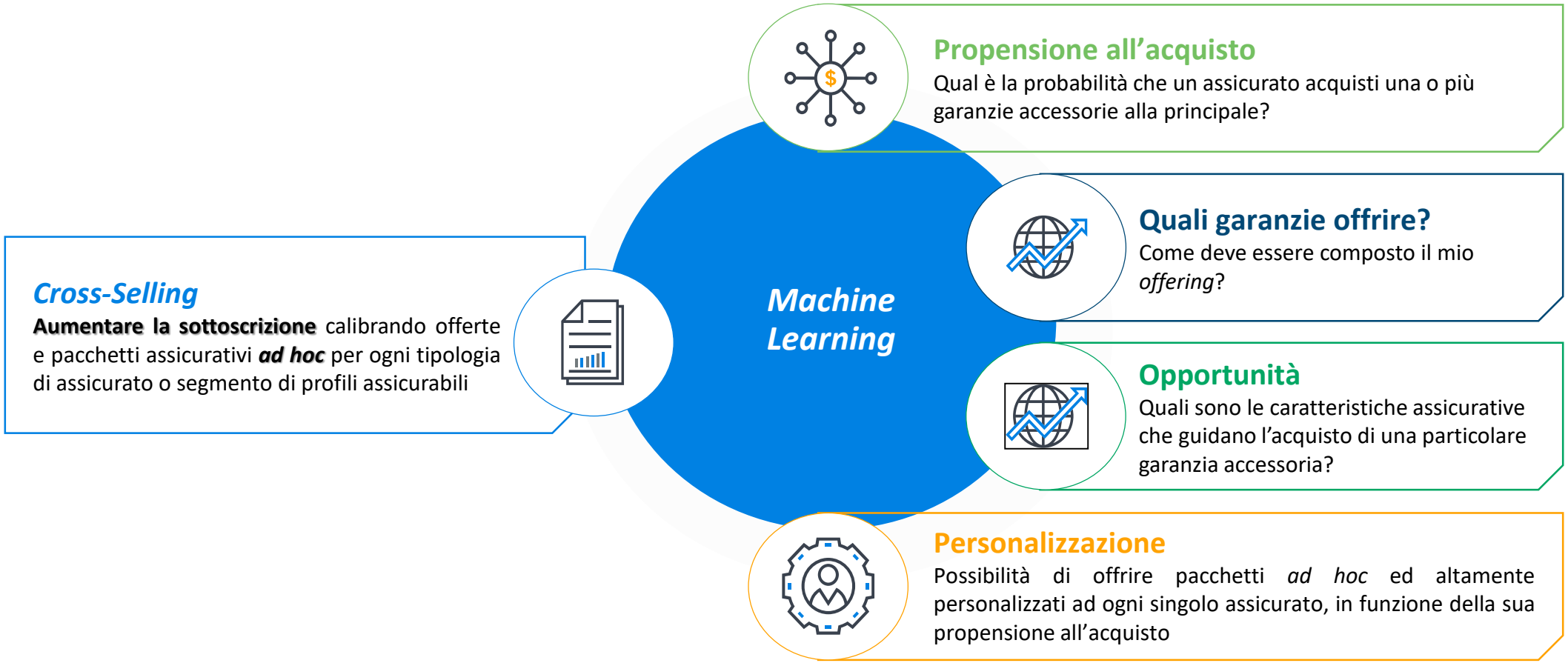
MTPL	2015	2016	2017	MOD	2015	2016	2017
Loss Ratio	72,1%	76,1%	75,9%	*Loss Ratio*	58,2%	57,4%	60,6%
Expenses Ratio	21,5%	21,4%	21,2%	*Expenses Ratio*	29,8%	30,5%	30,7%
Combined Ratio	93,6%	97.6%	97,1%	*Combined Ratio*	88,0%	87,9%	91,3%

MOD registra sempre un livello di profittabilità più alta del MTPL

I **volume del MOD sono cresciuti** negli ultimi anni, anche grazie alla forte connessione che queste coperture hanno con **le nuove immatricolazioni**, che risultano anch'esse in crescita



Perché fare *Cross-Selling*? (1)



Perché fare *Cross-Selling*? (2)

Cross Selling

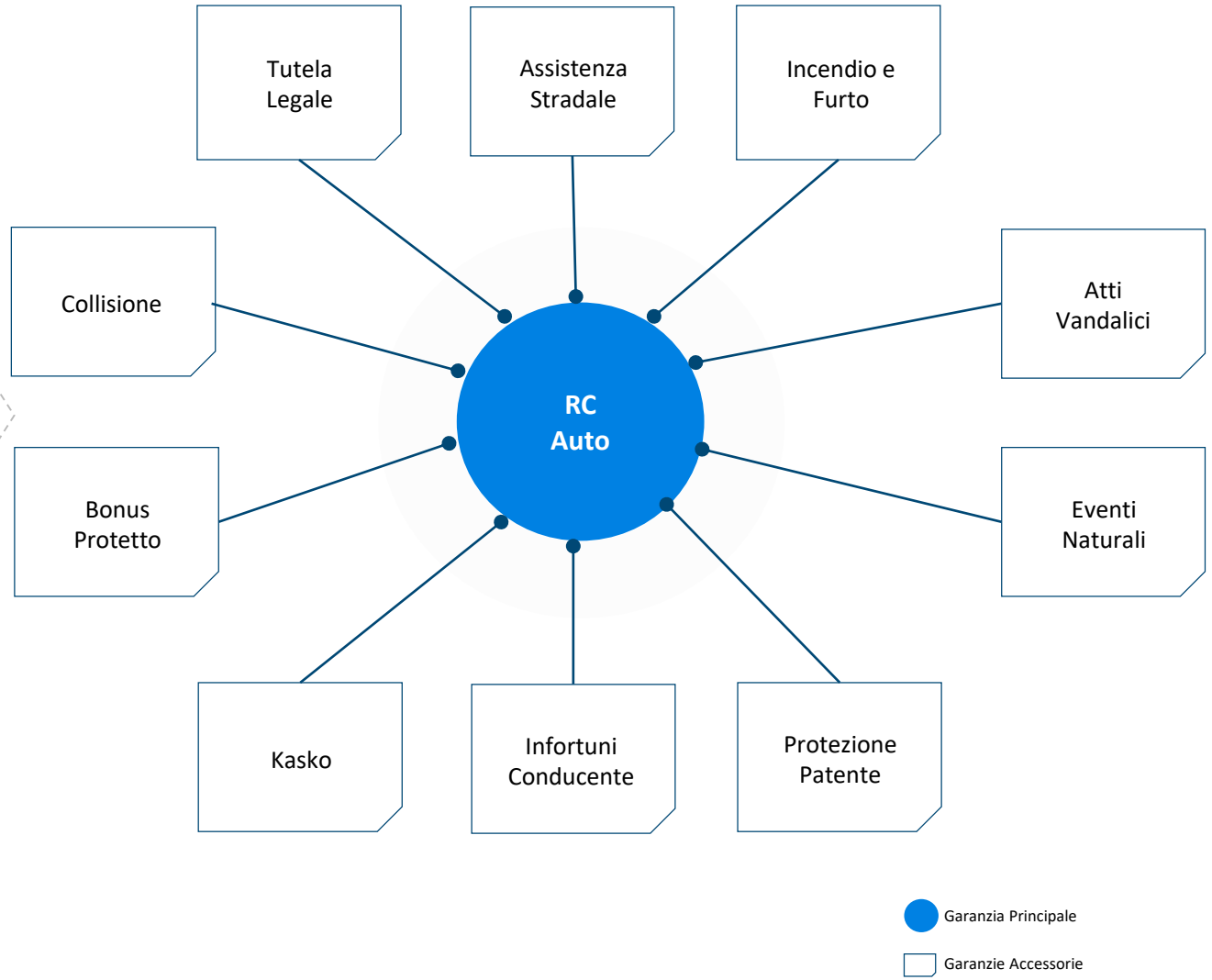
Attuare una efficace strategia commerciale finalizzata all'aumento dei volumi di sottoscrizione

Rinnovi

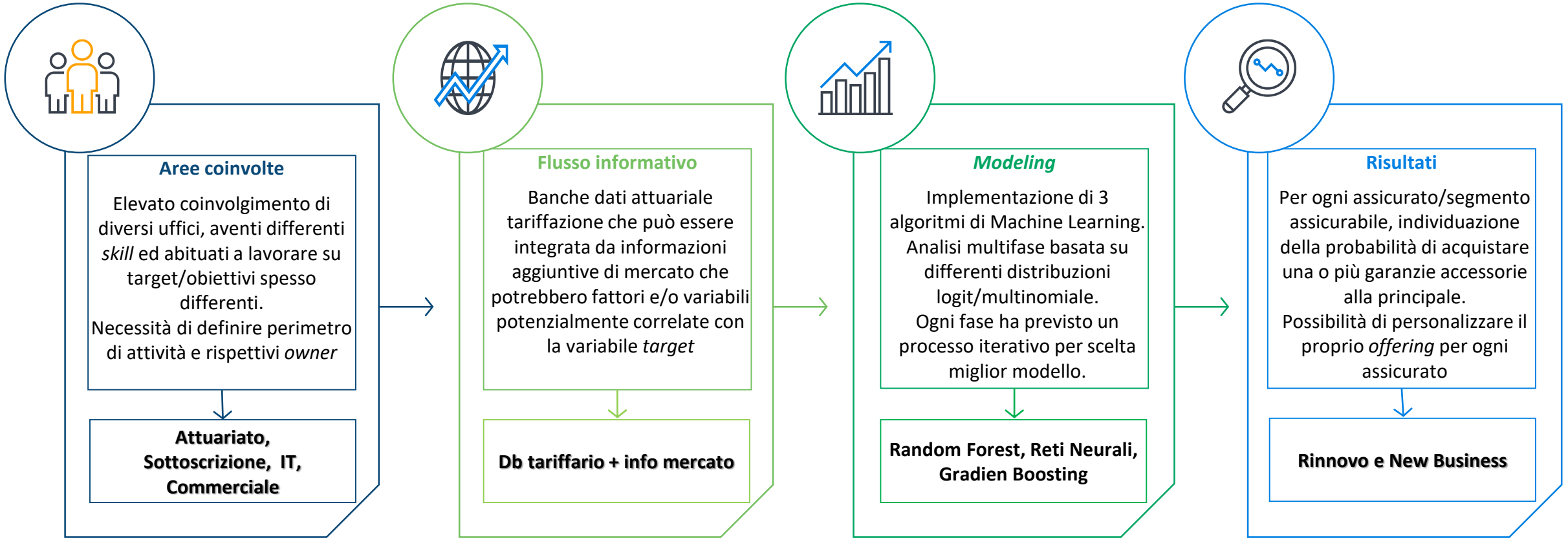
Individuare, per ogni singolo assicurato a scadenza, la sua propensione all'acquisto di una o più garanzie accessorie alla principale

New Business

Individuare segmenti di profili assicurabili, con caratteristiche assicurative e di propensione all'acquisto simili



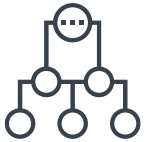
Cross-Selling: Workflow



Cross-Selling: Aree coinvolte

Compagnia - uffici

Information technology	Attuariale	Sottoscrizione	Commerciale
Predisposizione flusso dati per finalità di *pricing* attuariale. Deve supportare l'area attuariale al fine di definire tutte le informazioni da estrarre (*database* che prevedrà più rami ministeriali)	Implementazioni algoritmi Machine Learning. Sviluppo di un'analisi multifase che prevede processi iterativi finalizzato al calcolo della propensione all'acquisto. Attenzione sarà data alla valutazione della bontà di adattamento del modello	In funzione della propensione all'acquisto, l'area di sottoscrizione implementeranno le regole di sottoscrizione (i.e alcune garanzie non sono vendibili in certe regione, a veicoli immatricolati prima di una determinata data.	Implementerà gli sconti commerciali al fine di rendere l'*offering* maggiormente appetibile. Interessante capire l'interrelazione tra scontistica e propensione all'acquisto

Cross-Selling: Flusso informativo (1)

Come gestire il **flusso informativo** per gli algoritmi di ***Machine Learning***?

Apprendono direttamente dai dati

Ma gli algoritmi di Machine Learning come **«imparano dai dati»**?

Apprendimento supervisionato. Analizza contemporaneamente la relazione tra *input* e la variabile *target*→ Identifica una regola

Apprendimento non supervisionato. Viene fornito solamente gli *input*→ Risale a modelli e schemi nascosti identificando negli *input* una struttura logica che li classifica

Apprendimento semi supervisionato. Flusso dati ibrido, consistente di dati incompleti (alcuni hanno variabile *target*, altri no)→Identifica regole/schemi

Apprendimento per rinforzo. Ambiente dinamico (presenza di *input*) e mira a raggiungere un obiettivo→ Ricompensa se la prestazione è buona ed una punizione se non lo è

Come funzionano gli algoritmi di **apprendimento supervisionato**?

Tecniche di apprendimento finalizzate ad individuare una **insieme di regole o di compiti da risolvere** attraverso una serie di esempi concreti (costituiti da *input* e variabili *target*)

Quale **flusso informativo** per algoritmi **apprendimento supervisionato**?

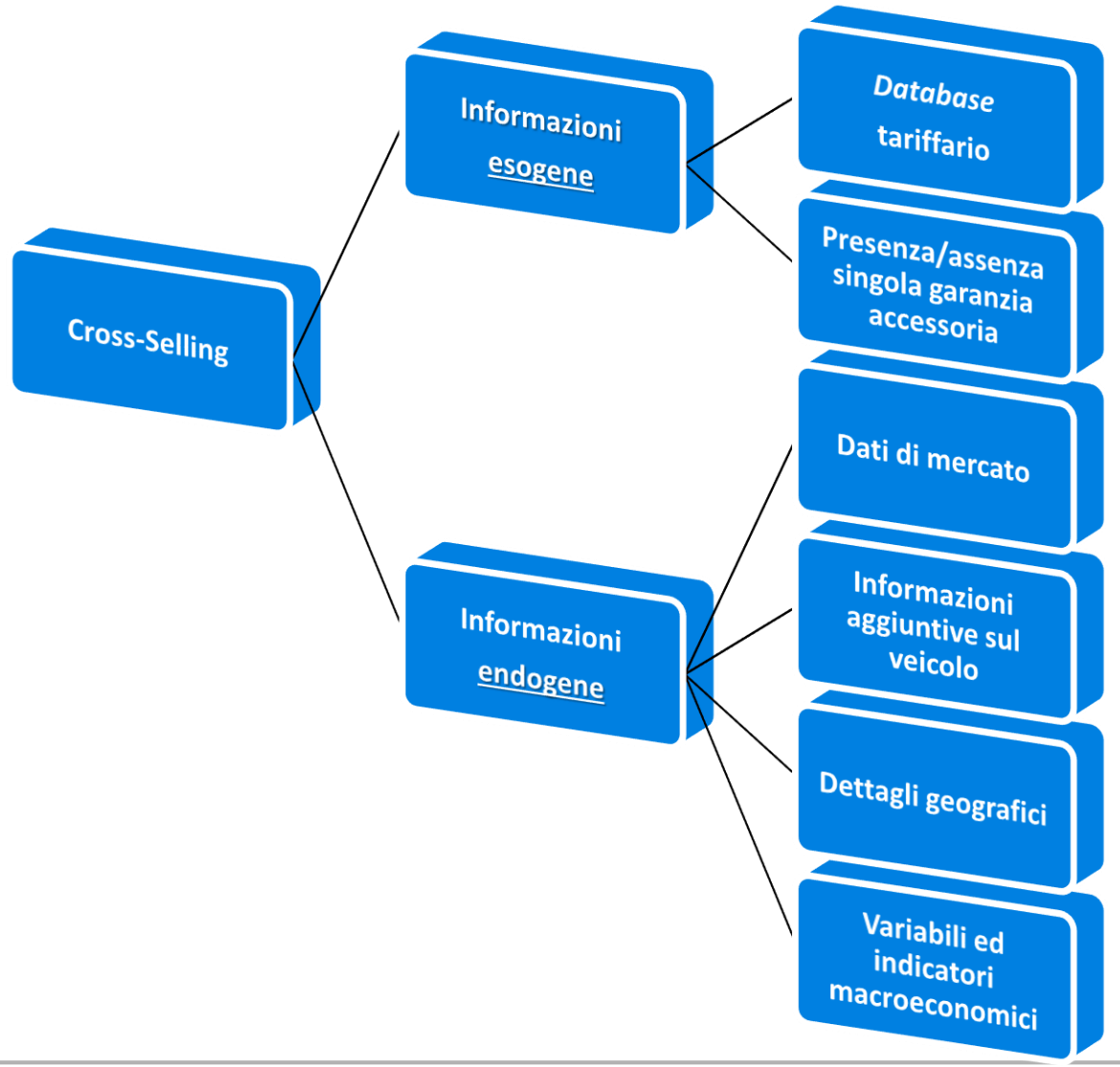
Prevede un flusso informativo costituito da variabili **indipendenti/esplicative** e dalla variabile **dipendente/risposta**

Cross-Selling: Flusso informativo (2)

```
> # RANDOM FOREST
> set.seed(42)
> # IMPUTE VALUES for the NA
> data.imputed<-rfImpute(hd~., data=data, iter=6) #hd=heart disease #iter=iteration #OOB=Out-Of-Bag error rate
ntree      OOB      1      2
  300:  17.49% 12.80% 23.02%
ntree      OOB      1      2
  300:  16.83% 14.02% 20.14%
ntree      OOB      1      2
  300:  17.82% 13.41% 23.02%
ntree      OOB      1      2
  300:  17.49% 14.02% 21.58%
ntree      OOB      1      2
  300:  17.16% 12.80% 22.30%
ntree      OOB      1      2
  300:  18.15% 14.63% 22.30%
> summary(data.imputed)
        hd           age        sex     cp         trestbps         chol       fbs     restecg
 Healthy  :164   Min.   :29.00   F: 97   1: 23   Min.   : 94.0   Min.   :126.0   0:258   0:151
 Unhealthy:139   1st Qu.:48.00   M:206   2: 50   1st Qu.:120.0   1st Qu.:211.0   1: 45   1:  4
                 Median :56.00           3: 86   Median :130.0   Median :241.0           2:148
                 Mean   :54.44           4:144   Mean   :131.7   Mean   :246.7
                 3rd Qu.:61.00                   3rd Qu.:140.0   3rd Qu.:275.0
                 Max.   :77.00                   Max.   :200.0   Max.   :564.0
    thalach      exang      oldpeak     slope   ca      thal
 Min.   : 71.0   0:204   Min.   :0.00   1:142   2:179   2:168
 1st Qu.:133.5   1: 99   1st Qu.:0.00   2:140   3: 65   3: 18
 Median :153.0           Median :0.80   3: 21   4: 38   4:117
 Mean   :149.6           Mean   :1.04           5: 21
 3rd Qu.:166.0           3rd Qu.:1.60
 Max.   :202.0           Max.   :6.20
>
> model<-randomForest(hd~., data=data.imputed, proximity=TRUE)
> model

Call:
 randomForest(formula = hd ~ ., data = data.imputed, proximity = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 16.83%
Confusion matrix:
           Healthy Unhealthy class.error
Healthy       142        22   0.1341463
Unhealthy      29       110   0.2086331
```



Cross-Selling: *Modeling*

Random Forest

Distribution free e basato su alberi di classificazione (1/0).

Il nostro processo iterativo ha testato circa 1.000 differenti RF, iterando **dimensione nodi finali**, **massimo numero di nodi finali** ed **il numero di variabili campionate** casualmente all'interno di ogni iterazione.

La scelta della RF migliore è stata basata sul livello di bontà di stima degli ***Out of Bag***

Boosting

Distribuzione logit (Bernoulli)

Il nostro processo iterativo ha testato circa 1.000 differenti *gradient boosting*, iterando il **numero di alberi**, il **tasso di apprendimento**, la massima profondità di ciascun albero ed il **numero minimo di osservazioni** all'interno dei nodi finali

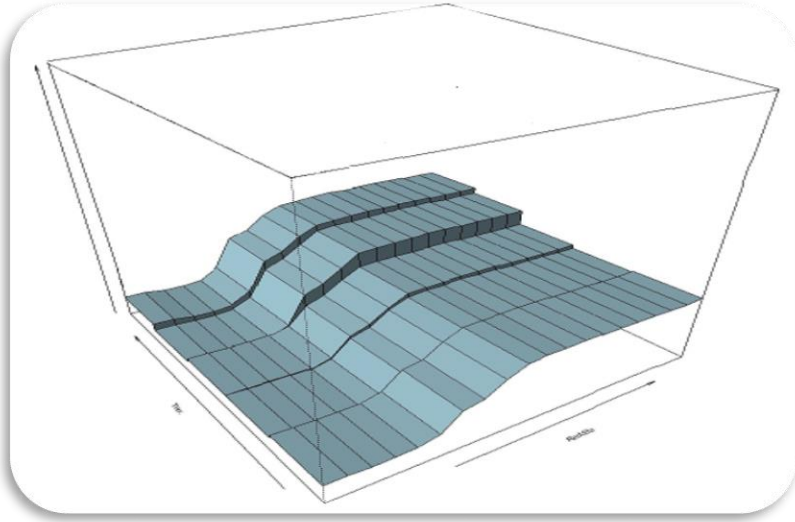
La scelta del GB migliore è stata basata sul livello di bontà della stima del ***validation set*.**

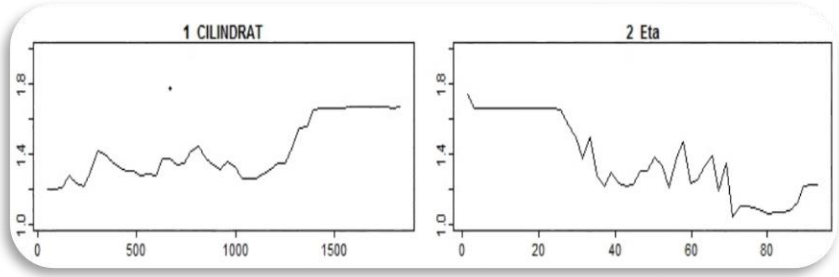
Reti neurali

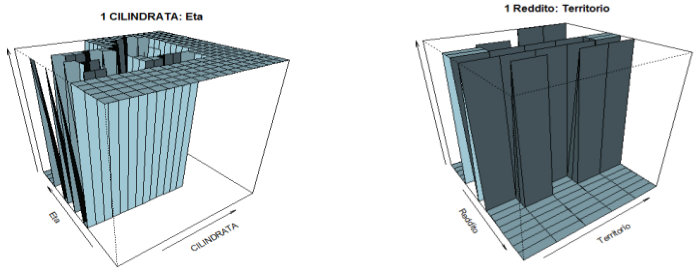
Distribuzione multinomiale (multinomial log-linear modello)

Il nostro processo iterativo ha testato circa 500 differenti reti neurali, iterando il numero di ***cross validation*** ed il numero di **interazioni**.

La scelta della NN migliore è stata basata sul livello di bontà di stima del **validation set**.

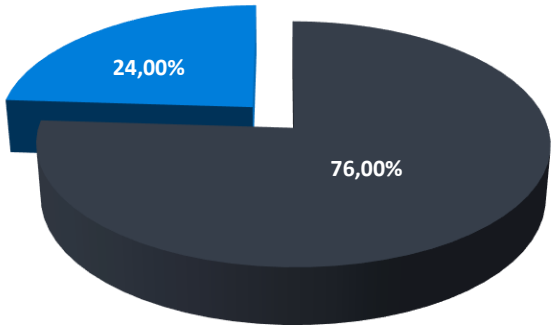






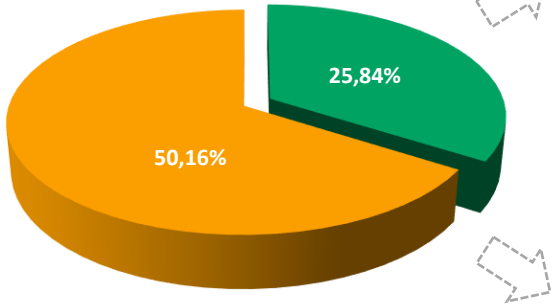
Cross-Selling: Risultati

Marca: Fiat
Modello: 500
Allestimento: Pop
Residenza: Milan
…
…
Anni assicurato: 4



Probabilità di acquistare **almeno una** garanzia accessoria

Probabilità di acquistare **solo** MTPL



Probabilità di acquistare **almeno una garanzia legata al valore del veicolo**

Probabilità di acquistare **solo garanzie non legate** al valore del veicolo