CHI SARANNO I PROSSIMI CLIENTI CHURNER E COSA LI SPINGE AD ABBANDONARE L'AZIENDA

Aurora Cerabolini	Andrea Malinverno	Corinna Strada	Mirko Tritella
839327	847340	839193	887196

1. Introduzione

Negli ultimi anni, i competitors che si affacciano sul mercato dell'intrattenimento sono sempre di più e con un'offerta che è sempre più ricca e di qualità.
Per ogni azienda che ambisce a rimanere competitiva, è cruciale riuscire ad implementare delle strategie in grado non solo di conquistare nuovi clienti, ma di mantenere quelli attuali nel tempo. Infatti, i costi necessari per conquistare nuovi clienti sono spesso superiori a quelli che devono essere sostenuti per mantenerli.
Per questi motivi, il focus delle aziende oggi è sempre più incentrato sulla customer retention, piuttosto che sulla customer acquisition.

L'abbandono di un cliente rappresenta un investimento significativo perso e l'azienda deve impiegare notevoli sforzi, sia in termini economici sia di tempo, per sostituirlo.
Essere in grado di prevedere la propensione dei clienti a rinunciare al servizio può dunque offrire enormi risparmi a un'azienda, nonché insights utili sulle strategie da implementare per massimizzare la retention anche tenendo conto del Customer Lifetime Value dei vari gruppi di clienti.

1.1 Obiettivi

Questo progetto ha lo scopo di identificare i clienti di un'azienda che stanno per disdire l'abbonamento, ovvero i potenziali churner, per poter fornire al management aziendale informazioni utili a sviluppare una strategia adatta. In particolare, le informazioni emerse dallo studio potranno dare indicazioni sulla numerosità dei potenziali churner, nonché la loro propensione a rinunciare ai servizi dell'azienda, e potranno essere utilizzate anche in collaborazione con settori aziendali quali quelli di Marketing e Vendite.

Sono stati quindi implementati vari modelli di machine learning, cioè un Classification Tree, un Random Forest, un modello Logistico, un Bagging Tree, un Naive-Bayes, un Gradient Boosting, un Neural Network Single-Layer Perceptron ed un Neural Network Multi-Layer Perceptron.

Successivamente, è stato individuato quello più performante, cioè quello in grado di predire il comportamento degli utenti in maniera migliore sulla base dei dati storici forniti dalla compagnia. Esso è risultato essere il Random Forest.
Questo modello permette non solo di individuare la predisposizione di un cliente a rinunciare al servizio, ma anche di capire quali sono le variabili che più hanno aiutato il modello ad effettuare la previsione.

1. Dataset

Il dataset, fornito da un'industria appartenente al settore dell'intrattenimento, contiene 102 variabili registrate su un totale di 330586 osservazioni: ogni osservazione fornisce informazioni sul comportamento di un utente su una pagina web che ha visitato.

Le variabili presenti nel dataset sono:

- *how_many_ok_urls*: numero di siti che il motore semantico è riuscito ad analizzare.
- *how_many_ko_urls*: numero di siti che il motore semantico non è riuscito ad analizzare.
- *external_id*: identificativo dell'utente.
- variabili dummy riguardanti il sistema operativo del browser utilizzato. Queste variabili sono: "*browser_android*", "*browser_chrome*", "*browser_chromium*", "*browser_edge*", "*browser_firefox*", "*browser_ie*", "*browser_opera*", "*browser_other*", "*browser_safari*", "*browser_unknown*".
- variabili dummy riguardanti il nome del browser. Queste variabili sono: "*browser_android*", "*browser_chrome*", "*browser_chromium*", "*browser_edge*", "*browser_firefox*", "*browser_opera*", "*browser_other*", "*browser_safari*", "*browser_unknown*", "*browser_ie*".
- variabili riguardanti un momento della giornata di un giorno feriale ("*feriale_morning*", "*feriale_afternoon*", "*feriale_evening*" e "*feriale_night*") in cui l'utente ha navigato. Per ogni osservazione, sommando i valori di queste variabili e di quelle riguardanti un momento della giornata durante il weekend si ottiene un valore pari a 1.
- variabili ("*weekend_morning*", "*weekend_afternoon*", "*weekend_evening*" e "*weekend_night*") in cui l'utente ha navigato. Per ogni osservazione, sommando i valori di queste variabili e di quelle riguardanti un momento della giornata durante un giorno feriale si ottiene un valore pari a 1.
- variabili che indicano la lunghezza dei testi presenti nelle pagine web che l'utente ha visitato. La variabile è scritta come *LminLength_maxLength*, dove *minLength* e *maxLength* rappresentano rispettivamente la lunghezza (in caratteri) minima e massima che deve avere il testo per essere registrato all'interno della variabile. Le variabili sono: "*L00_50*", "*L51_100*", "*L101_250*", "*L251_500*", "*L501_1000*", "*L1001_2500*", "*L2501_5000*", "*L5001_10000*". Per ogni osservazione, sommando i valori di queste variabili si ottiene un valore pari a 1.
- variabili dummy relative alla categoria di appartenenza del sito web. Queste variabili sono:
  "*categories_artandentertainment*", "*categories_automotive*", "*categories_business*", "*categories_careers*", "*categories_education*", "*categories_familyandparenting*", "*categories_finance*", "*categories_foodanddrink*", "*categories_healthandfitness*", "*categories_hobbiesandinterests*", "*categories_emotions*", "*categories_pets*", "*categories_homeandgarden*", "*categories_news*", "*categories_lawgovtandpolitics*", "*categories_realestate*", "*categories_religionandspirituality*", "*categories_science*", "*categories_sports*","*categories_intentions*", "*categories_technologyandcomputing*", "*categories_society*", "*categories_styleandfashion*", "*categories_uncategorized*", "*categories_travel*", "*categories_shopping*".
- variabili dummy che descrivono le categorie semantiche ad hoc per ogni sito web. Queste variabili sono: "*admants_disdettecontrattuali*", "*admants_appletv*", "*admants_chili*", "*admants_comparatoriprezzo*", "*admants_googleplayfilm*",

"*admants_netflix*", "*admants_mediaset*", "*admants_skygastronomia*", "*admants_novita*", "*admants_offerte*", "*admants_skyarte*" , "*admants_skycinema*", "*admants_skycompetitors*", "*admants_skyfamiglia*", "*admants_skyinternetfamiglia*", "*admants_skyinternet*", "*admants_skymusica*", "*admants_skysportformula1*", "*admants_skynews*", "*admants_skytelevisione*", "*admants_skyonline*", "*admants_skysport*", "*admants_skysportcalcio*", "*admants_skytecnologia*".

- variabili dummy relative al tipo di contratto attivo per ogni utente. Le variabili sono: "*CINEMA*", "*CALCIO*", "*SPORT*", "*SKY_FAMIGLIA*", "*FLG_MV*", "*FLG_MYSKYHD*", "*FLG_HD*", "*FLG_MYSKY*", "*FLG_SKY_ON_DEMAND*", "*STB_HD*", "*STB_MYSKYHD*", "*STB_MYSKY*", "*STB_SD*".
- *Pdisc*: variabile target riguardante gli utenti che hanno disdetto l'abbonamento (assume valore 0 se non ha disdetto e valore 1 se ha abbandonato).
- *DATA_RIF*: data di rilevazione dell'osservazione.

2. Data Preprocessing

3.1 Preparazione dei dati

Sono state eliminate la variabile *DATA_RIF* e le variabili contenenti solo valori pari a zero perché non utili all'analisi. Dopo questa operazione il dataset contiene 70 variabili.
Sono inoltre state identificate delle osservazioni duplicate e anch'esse sono state rimosse dal dataset.

Le osservazioni che contengono lo stesso *external_id* vengono aggregate tramite somma, per le variabili dummy, e media, per le variabili contenenti valori percentuali in modo che il totale sia sempre pari a 1. Successivamente, i valori aggregati maggiori di 1 nelle variabili dummy sono stati posti pari ad 1 per avere la variabile sempre espressa in termini di presenza/assenza.

3.2 Variabile target

Andando ad osservare come si distribuiscono le osservazioni nelle classi della variabile target *Psdisc*, si nota come solo il 2.9% degli utenti abbia disdetto il contratto. Questa classe è quindi da considerarsi rara e i modelli di classificazione potrebbero non essere in grado di classificare correttamente le unità appartenenti a tale livello del target.
Di conseguenza, una volta suddiviso il dataset iniziale in Training, Validation e Test, si effettuerà un maggiore bilanciamento delle osservazioni rare con le osservazioni non rare nel dataset di Training.

3.3 Studio della correlazione

Dopo aver controllato che non fossero presenti dati mancanti, le variabili sono state divise in numeriche e categoriali per eseguire un'analisi della collinearità.

La maggior parte delle variabili numeriche sono debolmente correlate tra loro. Si evidenzia un coefficiente di correlazione negativo pari a -0.5 tra le variabili *L00_50* e *L101_250* e un

coefficiente di correlazione positivo pari a 0.37 tra le variabili *L51_101* e *categories_technologyandcomputing*.
Non essendo stata riscontrata la presenza di elevata collinearità, non è stato necessario eliminare alcuna variabile numerica.

Le associazioni tra le variabili categoriali si valutano con il test Chi-quadrato e la coppia di variabili *os_other* e *browser_unknown* presenta un valore chi-quadrato normalizzato pari a 0.9. Essendo questa cifra superiore alla soglia di 0.8, che indica che l'associazione tra le due variabili non è dovuta al caso, è stato deciso di eliminare la variabile os_other dal dataset.

3.4 Zero Variance e Near Zero Variance

Nella tabella seguente non si riscontra la presenza di zero variance mentre si può notare la presenza di near zero variance in molte variabili numeriche: queste variabili non sono state eliminate in quanto potrebbe esserci una perdita di informazioni utili per l'analisi.

	freqRatio	percentUnique	zeroVar	nzv
how_many_ok_urls	1.417314	0.5529787	FALSE	FALSE
how_many_ko_urls	5.979839	0.2993131	FALSE	FALSE
feriale_morning	9.305726	8.3011145	FALSE	FALSE
feriale_afternoon	5.375996	8.0581715	FALSE	FALSE
feriale_evening	37.611168	5.8030599	FALSE	TRUE
feriale_night	39.710909	6.6722016	FALSE	TRUE
weekend_morning	55.104091	5.1324392	FALSE	TRUE
weekend_afternoon	45.388774	5.3040004	FALSE	TRUE
weekend_evening	148.768254	4.0739674	FALSE	TRUE
weekend_night	154.328094	4.0310771	FALSE	TRUE
L00_50	3.607438	7.8189048	FALSE	FALSE
L51_100	21.408217	6.3306108	FALSE	TRUE
L101_250	6.279219	7.3691692	FALSE	FALSE
L251_500	23.084061	5.4054054	FALSE	TRUE
L501_1000	44.971335	4.0356725	FALSE	TRUE
L1001_2500	142.078396	2.4683378	FALSE	TRUE
L2501_5000	1144.126866	1.0869632	FALSE	TRUE
L5001_10000	3865.686747	0.4727126	FALSE	TRUE
L10001_more	14117.260870	0.2515211	FALSE	TRUE
categories_artandentertainment	138.979535	19.0500407	FALSE	FALSE
categories_automotive	827.805732	8.9417121	FALSE	TRUE
categories_business	376.692308	9.3709216	FALSE	TRUE
categories_careers	1891.057692	4.5718015	FALSE	TRUE
categories_education	1746.800000	5.2825553	FALSE	TRUE
categories_familyandparenting	1971.725490	3.3898669	FALSE	TRUE
categories_finance	589.552972	11.9948899	FALSE	FALSE
categories_foodanddrink	451.210145	9.5679107	FALSE	TRUE
categories_healthandfitness	623.808717	7.5156090	FALSE	TRUE
categories_hobbiesandinterests	885.962712	8.5477339	FALSE	TRUE
categories_homeandgarden	2801.752294	3.1536638	FALSE	TRUE
categories_intentions	1519.743590	3.5454974	FALSE	TRUE
categories_lawgovtandpolitics	305.697638	14.5177597	FALSE	FALSE
categories_news	90.019413	14.5336903	FALSE	FALSE
categories_pets	4527.550725	2.4551643	FALSE	TRUE
categories_realestate	913.732899	6.6997739	FALSE	TRUE
categories_religionandspirituality	3180.969072	2.9416630	FALSE	TRUE
categories_science	102.931667	14.5487020	FALSE	FALSE
categories_shopping	142.990408	10.0452186	FALSE	FALSE
categories_society	276.970914	12.9746273	FALSE	FALSE
categories_sports	179.381463	20.4982017	FALSE	FALSE
categories_styleandfashion	943.256849	6.4265013	FALSE	TRUE
categories_technologyandcomputing	13.173006	23.1904882	FALSE	FALSE
categories_travel	454.130261	11.6128597	FALSE	FALSE
categories_uncategorized	153.526900	4.0237245	FALSE	TRUE

Terminate le analisi preliminari, è stato unito il dataset contenente solo le variabili numeriche a quello contenente solo le variabili di tipo factor tramite la variabile *external_id*.
Dopo aver ottenuto il dataset completo di tutte le variabili, è stata rimossa la variabile *external_id* in quanto si tratta di un identificativo non utile all'analisi.

3. Divisione del Dataset

Il dataset ricavato dopo gli step di preprocessing è stato diviso in ulteriori tre dataset attraverso un campionamento stratificato relativo alle classi della variabile target Pdisc.

- Training: i dati di training vengono utilizzati per addestrare i modelli e costituiscono il 60% delle osservazioni.
- Validation: i dati di validation costituiscono il 30% delle osservazioni e sono indipendenti da quelli di training. Con questi dati si valuterà la bontà classificativa dei modelli, si verificherà la presenza o meno di un overfitting e si confronteranno i modelli per scegliere quello vincente.
- Test: sui dati di test, che rappresentano il 10% delle osservazioni totali, verrà applicato il modello vincente per generare il target previsto.

Come è possibile vedere nella mostrata di seguito, in ognuno dei tre dataset le percentuali dei soggetti appartenenti ad ogni classe del target sono simili a quelle del dataset iniziale.

Dataset	% osservazioni classe 0	% osservazioni classe 1
Dataset iniziale	97.029%	2.971%
Dataset di Training	97.021%	2.978%
Dataset di Validation	97.071%	2.929%
Dataset di Test	96.946%	3.054%

4.1 Bilanciamento delle classi del target nel dataset di training

Per bilanciare le osservazioni rare con le osservazioni non rare è stata utilizzata la funzione SMOTE: tecnica ibrida che effettua un oversampling delle osservazioni rare creando dati sintetici (quindi non duplicando i dati ma creando delle osservazioni leggermente diverse dalle originali), ed eseguendo inoltre un undersampling delle osservazioni della classe non rara.
Dopo aver eseguito questa operazione, le percentuali dei soggetti appartenenti ad ogni classe del target è la seguente:

Classe 0	Classe 1
0.5714286	0.4285714

4. Addestramento dei modelli

Dal momento che lo scopo del progetto è quello di classificare correttamente i soggetti che disdicono l'abbonamento, cioè individuare i churner, i modelli sono stati addestrati sul dataset di Training cercando di massimizzare la specificity. In formule:

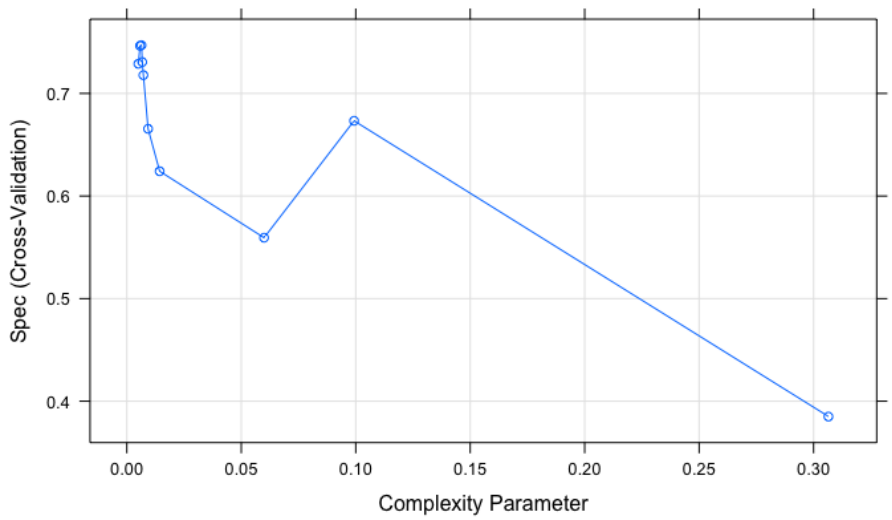
$$\text{specificity} = \frac{TN}{TN + FP}$$

I modelli sono quindi stati allenati con l'intento di individuare in maniera più precisa possibile i True Negative (TN) o, in altri termini, i clienti che abbandonano l'azienda.

Inoltre, per l'addestramento sul dataset di Training è stato utilizzato il metodo della cross-validation (10 fold) per ottenere delle metriche di performance robuste ed evitare il problema di overfitting, garantendo anche la replicabilità del modello.

5.1 Classification Tree

Con un parametro di complessità pari a 0.0064 si ottiene un Classification Tree che massimizza il valore della specificity, la quale risulta essere pari a 0.74:



La matrice di confusione è la seguente:

```
confusionMatrix(tree)
##
##            Reference
## Prediction   0    1
##        0    48.0 10.8
##        1     9.1 32.0
##
##  Accuracy (average) : 0.8002
```

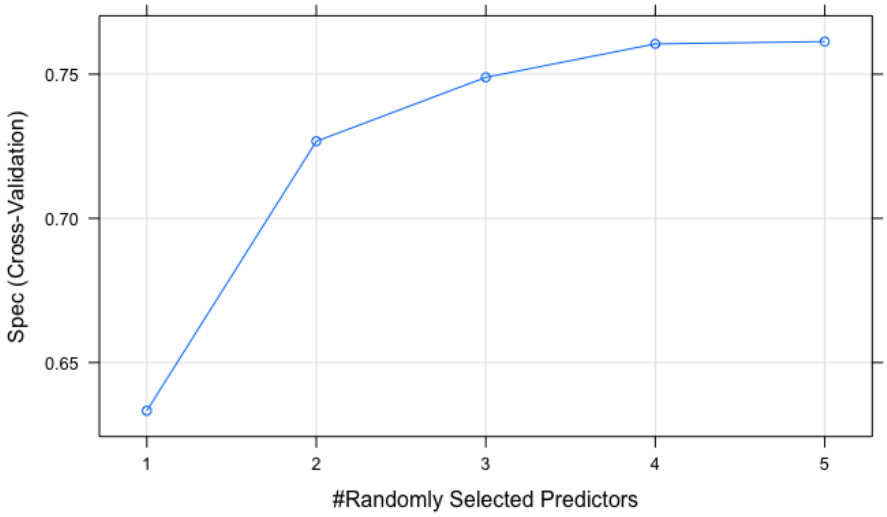
Si può vedere come i soggetti misclassificati come 1 sono il 9.1% mentre quelli misclassificati come 0 sono il 10.8%.

Questo modello può essere utilizzato come feature selector e sono stati creati nuovi dataset di training, validation e test inserendo solamente le variabili selezionate. Questi dataset verranno utilizzati per quei modelli che richiedono la feature selection.

Le variabili selezionate dal modello che risultano essere importanti per l'individuazione alla predisposizione al churn sono: "*SPORT*", "*L1001_2500*", "*CINEMA*", "*CALCIO*", "*FLG_MYSKYHD*" e "*categories_artandentertainment*".

5.2 Random Forest

Il modello che massimizza la specificity, ottenendo un valore di 0.76, ha un numero di covariate candidate ad ogni split point pari a 5:



La matrice di confusione è la seguente:

```
confusionMatrix(rf)
##
##             Reference
## Prediction   0     1
##        0     54.0  10.2
##        1      3.2  32.6
##
##  Accuracy (average) : 0.8661
```

Si può vedere come i soggetti misclassificati come 1 sono il 3.2% mentre quelli misclassificati come 0 sono il 10.2%.

Si decide di utilizzare il Random Forest come feature selector e, come era stato fatto anche per le variabili selezionate dal Classification Tree, sono stati creati nuovi dataset di training, validation e test.
Le variabili importanti selezionate dal modello per l'individuazione del churn risultano essere: "*how_many_ok_urls*", "*how_many_ko_urls*", "*feriale_morning*", "*feriale_afternoon*", "*feriale_evening*", "*feriale_night*", "*weekend_morning*", "*weekend_afternoon*", "*weekend_evening*", "*weekend_night*", "*L00_50*", "*L51_100*", "*L101_250*", "*L251_500*", "*L501_1000*", "*L1001_2500*", "*L2501_5000*", "*categories_artandentertainment*",

"*categories_automotive*", "*categories_business*", "*categories_sports*", "*categories_hobbiesandinterests*", "*categories_finance*", "*categories_foodanddrink*, "*categories_lawgovtandpolitics*", "*categories_news*", "*categories_science*", "*categories_travel*", "*categories_society*", "*categories_technologyandcomputing*", "*browser_chrome*", "*CINEMA*", "*SPORT*", "*SKY_FAMIGLIA*", "*FLG_MYSKYHD*", "*FLG_SKY_ON_DEMAND*", "*STB_HD*", "*STB_MYSKYHD*"

5.3 Modello logistico

Il Modello logistico addestrato sul dataset contenente le variabili selezionate dal Classification Tree ottiene un valore di specificity pari a 0.68.

La matrice di confusione è la seguente:

```
confusionMatrix(glm_tree)
##
##            Reference
## Prediction   0      1
##        0     47.0  13.8
##        1     10.1  29.1
##
##  Accuracy (average) : 0.7612
```

I soggetti misclassificati come 1 sono il 10.1% mentre quelli misclassificati come 0 sono il 13.8%.

Le variabili più importanti per questo modello sono: *"SPORT", "CINEMA"*, "*L1001_2500*" e *"CALCIO"*.

Il Modello logistico addestrato sul dataset contenente le variabili selezionate dal Random Forest restituisce un errore in quanto la variabile weekend_night causa separation. Questa variabile è stata quindi eliminata dal dataset per poter addestrare il modello.
Si ottiene un modello logistico con un valore di specificity di 0.7 e con la seguente matrice di confusione:

```
confusionMatrix(glm_rf)
##
##            Reference
## Prediction   0      1
##        0     47.1  12.7
##        1     10.0  30.1
##
##  Accuracy (average) : 0.7727
```

I soggetti misclassificati come 1 sono il 10% mentre quelli misclassificati come 0 sono il 12.7%.

Le variabili con un livello di importanza maggiore per l'individuazione della predisposizione al churn per questo modello sono: *"SPORT", "CINEMA"* e *"L1001_2500*"

5.4 Bagging Tree

Il modello Bagging ottiene un valore di specificity pari a 0.78 e la seguente matrice di confusione, dove è possibile osservare che i soggetti misclassificati come 1 sono il 4.3% mentre quelli misclassificati come 0 sono il 9.4%.

```
confusionMatrix(bagg)
##
##            Reference
## Prediction   0      1
##       0      52.9   9.4
##       1       4.3  33.5
##
##  Accuracy (average) : 0.8634
```

Le variabili che risultano maggiormente importanti all'individuazione del churn per il Bagging Tree risultano essere *"L1001_2500*", *"CINEMA",* "*weekend_afternoon*", "*FLG_MYSKY*", *"SPORT", "categories_sports"* e "*how_many_ok_urls"*.

5.5 Naive-Bayes

Il Naive Bayes che si ottiene ha un valore di specificity di 0.88. La matrice di confusione è la seguente:

```
confusionMatrix(naivebayes)
##
##            Reference
## Prediction   0      1
##       0       18.7   5.2
##       1       38.5  37.7
##
##  Accuracy (average) : 0.5633
##
```

I soggetti misclassificati come 1 sono il 38.5% mentre quelli misclassificati come 0 sono il 5.2%.

Le variabili che risultano maggiormente importanti per il modello Naivebayes sono: *"SPORT", "L1001_2500*", *"CINEMA", "FLG_MYSKYHD", "weekend_afternoon", "CALCIO", "categories_artandentertainment", "FLG_SKY_ON_DEMAND", "STB_MYSKYHD", "SKY_FAMIGLIA", "STB_HD", "categories_sport", "categories_shopping"* e *"categories_news".*

5.6 Gradient Boosting

Il modello Gradient Boosting che massimizza la specificity, la quale ha un valore pari a 0.77, ha i seguenti parametri:

- numero di iterazioni = 250
- massimo numero di nodi per albero = 5
- Shrinkage (Learning Rate) = 0.1

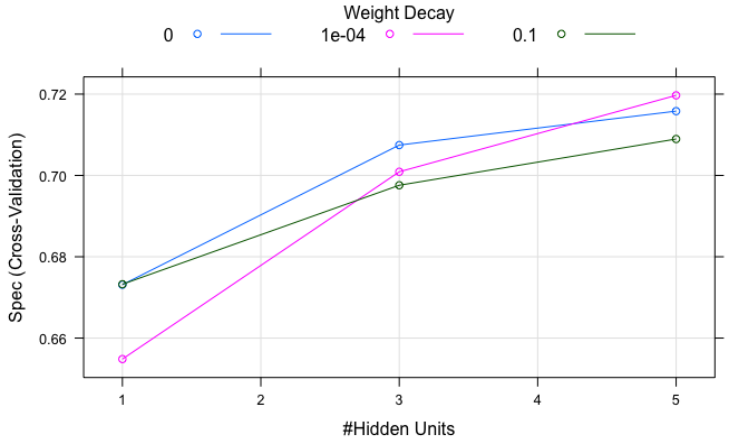
Per questo modello, si ottiene la seguente matrice di confusione, dove è possibile osservare che i soggetti misclassificati come 1 sono il 3.9% mentre quelli misclassificati come 0 sono il 9.7%.

```
confusionMatrix(gbm)
##
##            Reference
## Prediction   0      1
##        0    53.3   9.7
##        1     3.9  33.2
##
##  Accuracy (average) : 0.8644
```

Per questo modello, le variabili con un'importanza maggiore per individuare i clienti che disdicono l'abbonamento sono: *"SPORT", "L1001_2500"* e *"CINEMA"*

5.7 Neural Network

La Neural Network Single-Layer Perceptron addestrata sul dataset contenente le variabili selezionate dal Classification Tree, per massimizzare la specificity è composta da 5 neuroni all'interno dello strato nascosto ed ha un valore del parametro di decay pari a 1e-04:

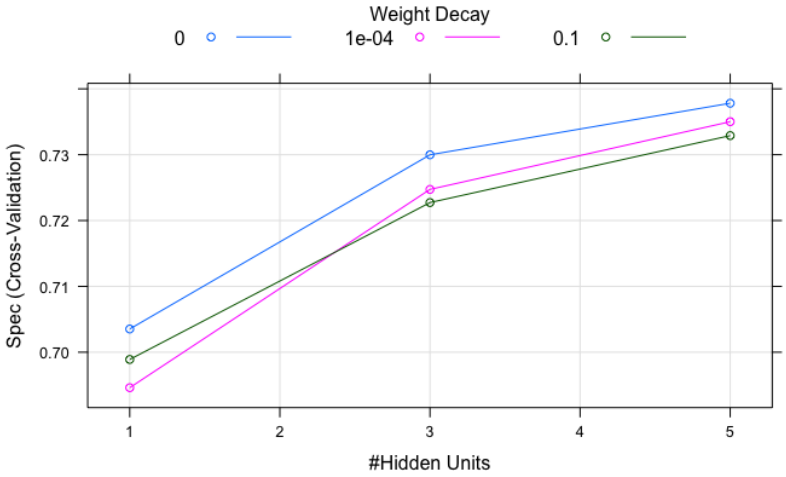


Questo modello ottiene un valore di specificity di 0.72 e la seguente matrice di confusione permette di osservare che i soggetti misclassificati come 1 sono l'8.6% mentre quelli misclassificati come 0 sono il 12%.

```
confusionMatrix(nnet_tree)
##
##            Reference
## Prediction   0      1
##        0    48.5  12.0
##        1     8.6  30.8
##
##  Accuracy (average) : 0.7938
```

Per questo modello, le variabili con un'importanza maggiore sono: *"L1001_2500", "categories_sports"* e *"weekend_afternoon".*

La Neural Network Single-Layer Perceptron addestrato sul dataset contenente le variabili selezionate dal Random Forest è composta da 5 neuroni all'interno dello strato nascosto ed ha un valore del parametro di decay pari a 0:

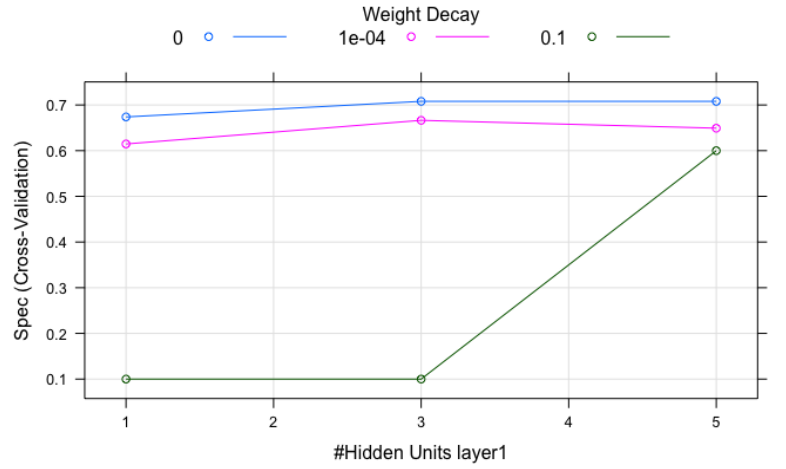


Il valore di specificity che si ottiene con questo modello è 0.73 e la matrice di confusione seguente mostra che i soggetti misclassificati come 1 sono il 7.2% mentre quelli misclassificati come 0 sono l'11.2%.

```
confusionMatrix(nnet_rf)
##
##            Reference
## Prediction   0     1
##        0    50.0  11.2
##        1     7.2  31.6
##
##  Accuracy (average) : 0.816
```

Le variabili maggiormente importanti per l'individuazione del churn sono: *"L2501_5000", "how_many_ko_urls", "STB_HD1"* e *"how_many_ok_urls"*

Il modello Neural Network Multi-Layer Perceptron addestrato sul dataset contenente le variabili selezionate dal Classification Tree ottiene una specificity di 0.7 ed è composto da un layer contenente 10 neuroni ed ha un valore del parametro di decay pari a 0 :

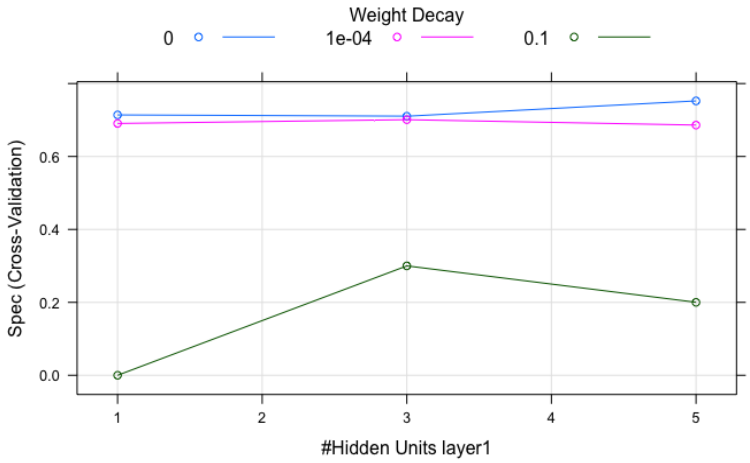


La matrice di confusione per questo modello mostra che i soggetti misclassificati come 1 sono il 9.2% mentre quelli misclassificati come 0 sono il 12.5%:

```
confusionMatrix(mlpML_tree)
##
##            Reference
## Prediction   0    1
##       0    48.0 12.5
##       1     9.2 30.3
##
##  Accuracy (average) : 0.783
```

Per questo modello, le variabili che risultano essere maggiormente importanti sono: *"SPORT", "L1001_2500"* e *"CINEMA"*

Il modello Neural Network Multi-Layer Perceptron addestrato sul dataset contenente le variabili selezionate dal Random Forest che massimizza il valore della specificity è composto da un layer contenente 5 neuroni ed ha un valore del parametro di decay pari a 0:



Il valore della specificity ottenuto è 0.75 e la matrice di confusione è la seguente:

```
confusionMatrix(mlpML_rf)
##
##            Reference
## Prediction   0    1
##       0    49.0 10.6
##       1     8.1 32.2
##
##  Accuracy (average) : 0.8129
```

È possibile osservare come i soggetti misclassificati come 1 sono l'8.1% mentre quelli misclassificati come 0 sono il 10.6%.

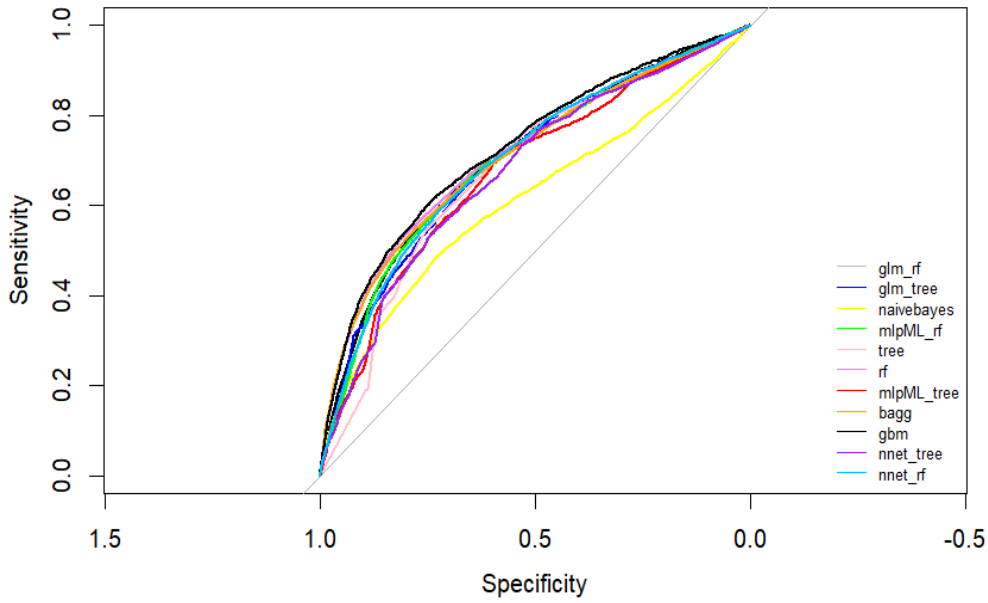
Le variabili più importanti per questo modello sono: *"SPORT", "L1001_2500*", *"CINEMA", "FLG_MYSKYHD", "weekend_afternoon", "categories_artandentertainment", "FLG_SKY_ON_DEMAND", "STB_MYSKYHD", "SKY_FAMIGLIA", "STB_HD", "categories_sport", "categories_news", "feriale_evening"* e *"weekend_night"*

5. Scelta del modello migliore

Per selezionare il modello migliore, sono state valutate le performance classificative dei modelli sul dataset di Validation.

6.1 Curve Roc e AUC

Con il target osservato e la probabilità prevista sul dataset di Validation dell'evento di interesse è possibile costruire le curve ROC per confrontare i modelli.
Le curve ROC mostrano come varia la sensitivity al variare del complemento a uno della specificity, ovvero quanto ogni modello classifica bene le osservazioni per tutte le possibili soglie.
Una curva ideale cresce repentinamente attorno all'origine e poi si assesta su una retta orizzontale e questo significa che, per ogni possibile soglia, il modello ha sempre correttamente classificato gli eventi e non ha mai misclassificato i non eventi.
Per valutare la bontà di una curva ROC si utilizza l'AUC, ovvero l'area sotto la curva, che corrisponde alla probabilità che un soggetto estratto a caso venga classificato correttamente dal modello indipendentemente dalla classe del target.

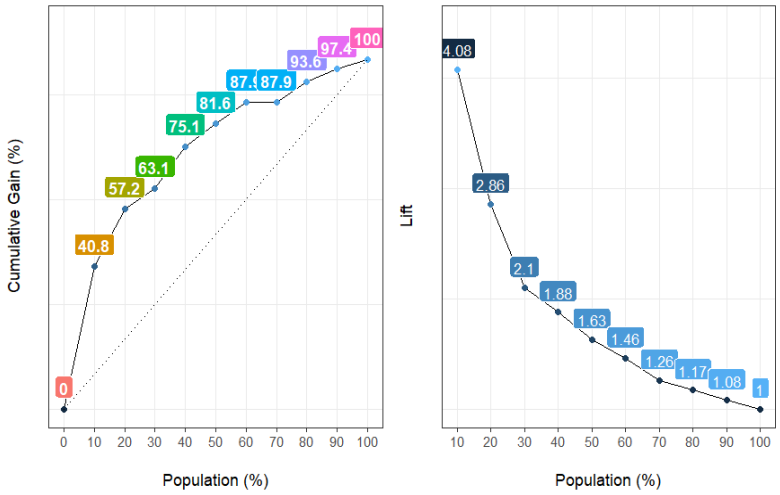


Le curve ROC dei modelli Random Forest, Bagging Tree e Gradient Boosting sono molto simili e in alcuni punti si sovrappongono. Il valore di AUC (Area Under The Curve) di questi tre modelli è rispettivamente pari a 0.7144, 0.7101 e 0.7255.
Per scegliere il modello migliore tra questi tre, si devono valutare le curve LIFT e le Cumulative Gains.

6.2 Cumulative Gains e Curve Lift

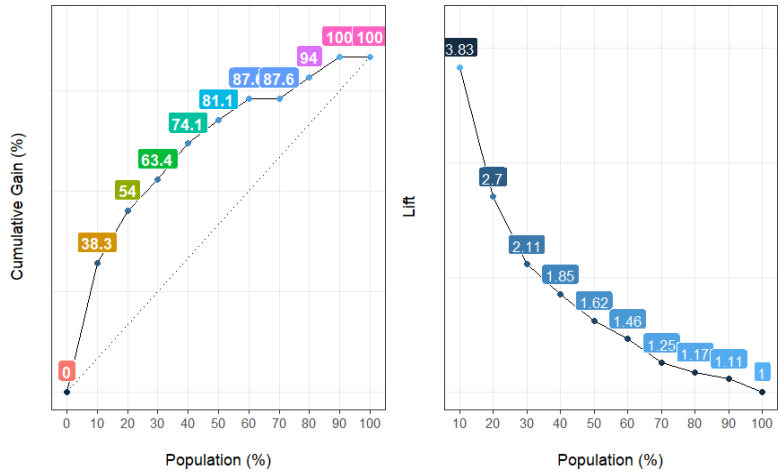
Nelle Cumulative Gains e nelle Curve LIFT, i soggetti vengono ordinati in ordine decrescente rispetto alle diverse probabilità previste per l'evento di interesse, in base alle quali si assegnano ai vari decili.
Un buon modello ha nei primi decili la maggior parte dei soggetti e le probabilità previste sono molto più alte rispetto a quelle degli ultimi decili.
Per scegliere il modello migliore è stato deciso di tenere in considerazione il secondo decile.

6.2.1 Cumulative Gain e Curva Lift del Random Forest



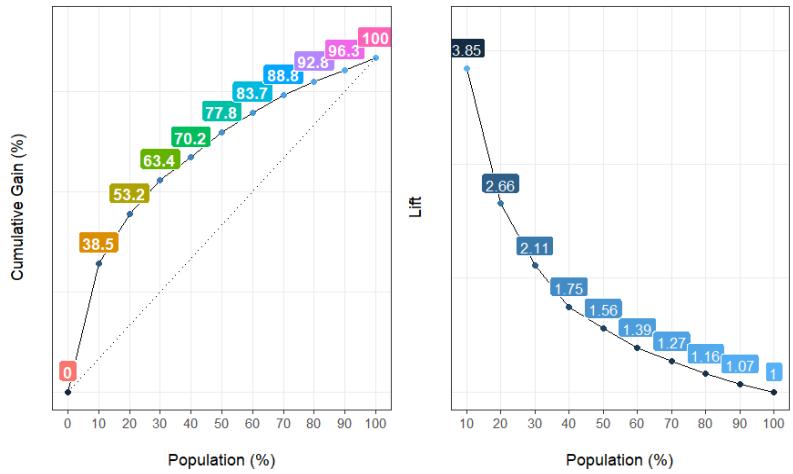
Osservando la Cumulative Gain, nel primo 20% della popolazione il modello riesce a raccogliere il 57.2% del totale dei casi di interesse. Nel grafico relativo alla curva LIFT si può osservare che, selezionando il primo 20% dei clienti, la probabilità dell'evento è 2.86 volte quella globale.

6.2.2 Cumulative Gain e Curva Lift del Bagging Tree



Nel grafico della Cumulative Gain, si osserva che questo modello nel secondo decile riesce a catturare il 54% del totale dei casi di interesse.
Osservando la curva Lift, nel secondo decile la probabilità dell'evento è 2.7 volte quella globale.

6.2.3 Cumulative Gain e Curva Lift del Gradient Boosting



Nel grafico della Cumulative Gain si nota che, selezionando il primo 20% dei soggetti, si riesce ad ottenere il 53.2% del totale dei casi di interesse.
La curva Lift mostra che nel secondo decile la probabilità dell'evento è 2.66 volte quella globale.

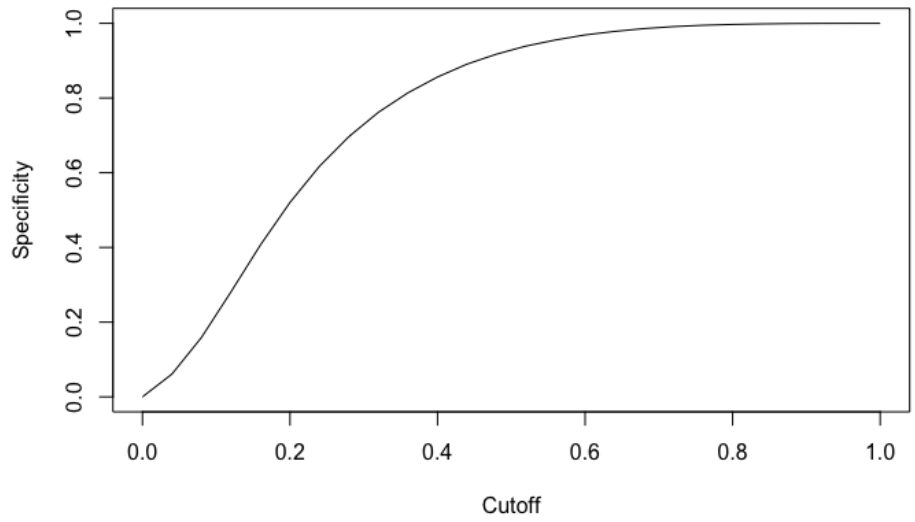
A seguito della valutazione delle curve Lift e delle Cumulative Gains, il Random Forest si è rivelato essere il modello che cattura il maggior numero di successi percentuali nei primi due decili perciò è il modello vincente da utilizzare sui nuovi soggetti.

6. Studio della soglia

In questo terzo step si sceglie il valore della soglia da applicare, nell'ultimo step, alle posterior che si ottengono sui dati di Test per classificare i nuovi soggetti.
Applicando la soglia di default (0.5) ai valori previsti dal Random Forest sui dati di Validation, si ottengono le metriche di performance e la matrice di confusione dalle quali si può osservare, paragonandole a quelle ottenute sul Training set, che il modello non overfitta e può essere generalizzato.

```
##
## Confusion Matrix and Statistics
##
##            Reference
## Prediction    0      1
##       0     89273   2021
##       1      5783    847
##
##              Accuracy : 0.9203
##                95% CI : (0.9186, 0.922)
##    No Information Rate : 0.9707
##    P-Value [Acc > NIR] : 1
##
##                 Kappa : 0.1433
##
##  Mcnemar's Test P-Value : < 0.0000000000000002
##
##            Sensitivity : 0.29533
##            Specificity : 0.93916
##         Pos Pred Value : 0.12775
##         Neg Pred Value : 0.97786
##              Prevalence : 0.02929
##          Detection Rate : 0.02554
## Detection Prevalence : 0.00865
##    Balanced Accuracy : 0.06771
##
```

Per scegliere il valore della soglia si osserva il grafico che mostra come cambia il valore della specificity per ogni possibile valore della soglia, vale a dire per tutti i valori delle posterior stimate sul dataset di Validation:



Sull'asse delle ordinate si sceglie il valore della specificity che risulta essere soddisfacente dal punto di vista classificativo e si proietta questo punto sull'asse delle ascisse in modo da ricavare la soglia di interesse per poter classificare i nuovi dati. La soglia scelta è pari a 0.6.

7. Dataset di test

Dopo aver scelta la soglia allo step precedente sul dataset di Validation, si applica il modello vincente al dataset di Test e si calcolano le posterior.
Infine, tramite la soglia scelta (0.6), si genera il target previsto per i soggetti presenti nel dataset di Test.
Si può notare come la percentuale della classe 1 sia molto simile a quella iniziale: questi risultati confermano che il modello scelto come migliore effettivamente ottiene una buona performance classificativa su nuovi dati.

Classe 0	Classe 1
0.97687029	0.02312971

8. Conclusione

Lo scopo di questo progetto era implementare un modello di machine learning in grado di identificare i potenziali churner, nonché individuare la loro propensione ad abbandonare l'azienda.

Tra i diversi modelli implementati, il modello migliore è risultato essere il Random Forest e la soglia di interesse per poter classificare i nuovi dati è stata scelta pari a 0.6: questo significa

che il modello identifica come predisposti al churn tutti quei soggetti che avranno una probabilità di churning superiore al 60%.
È possibile modificare il modello affinché la classificazione avvenga sulla base di soglie di probabilità diverse: ad esempio, se il management preferisce avere un approccio più pessimista, il valore di soglia per l'identificazione dei churner potrebbe essere portato al 50%. Al contrario, invece, se l'approccio da preferire fosse più ottimista, si potrebbe alzare il valore di soglia.

Le variabili che nel modello sono risultate maggiormente utili all'individuazione della predisposizione di un cliente al churn sono risultate essere: "*L1001_2500*", "*SPORT*", "*FLG_MYSKYHD*", "*CINEMA*", "*categories_artandentertainment*", "*categories_sports*", "*weekend_afternoon*", "*STB_MYSKYHD*", "*feriale_evening*", *"categories_shopping"*, "*how_many_ok_urls*", "*feriale_afternoon*", "*STB_HD*", "*L2501_5000*", "*L251_500*", "*feriale_morning*", "*how_many_ko_urls*" e "*categories_technologyandcomputing*".

Tra queste variabili, quelle che in media risultano essere importanti anche per gli altri modelli sono "*L1001_2500*", "*SPORT*", e "*CINEMA*".

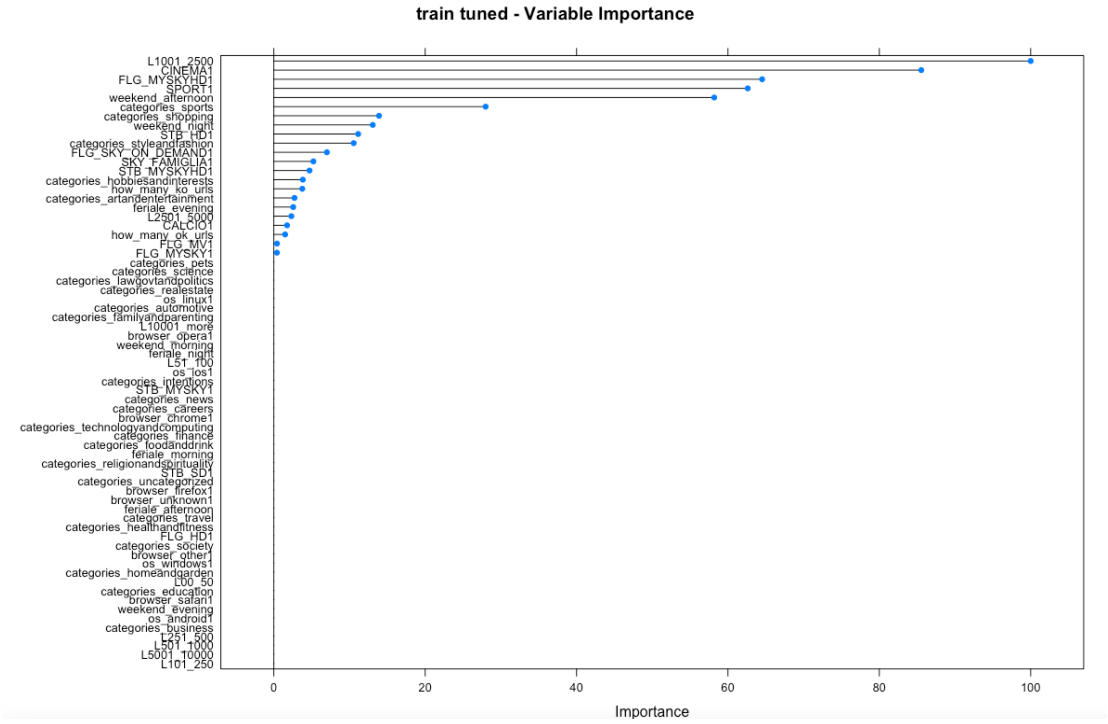
Questo significa che un cliente tende a non abbandonare se ha un contratto attivo relativo allo sport e al cinema. Per quanto invece riguarda la variabile "*L1001_2500*", sarebbe interessante cercare di approfondire le motivazioni che portano un cliente ad essere propenso ad abbandonare in base alla lunghezza dei testi presenti nelle pagine web che ha visitato

Sulla base dei risultati complessivamente emersi, possiamo consigliare al management aziendale di attuare delle strategie di fidelizzazione volte all'arricchimento dei diversi tipi di contratti, in particolar modo quelli relativi allo sport e al cinema.

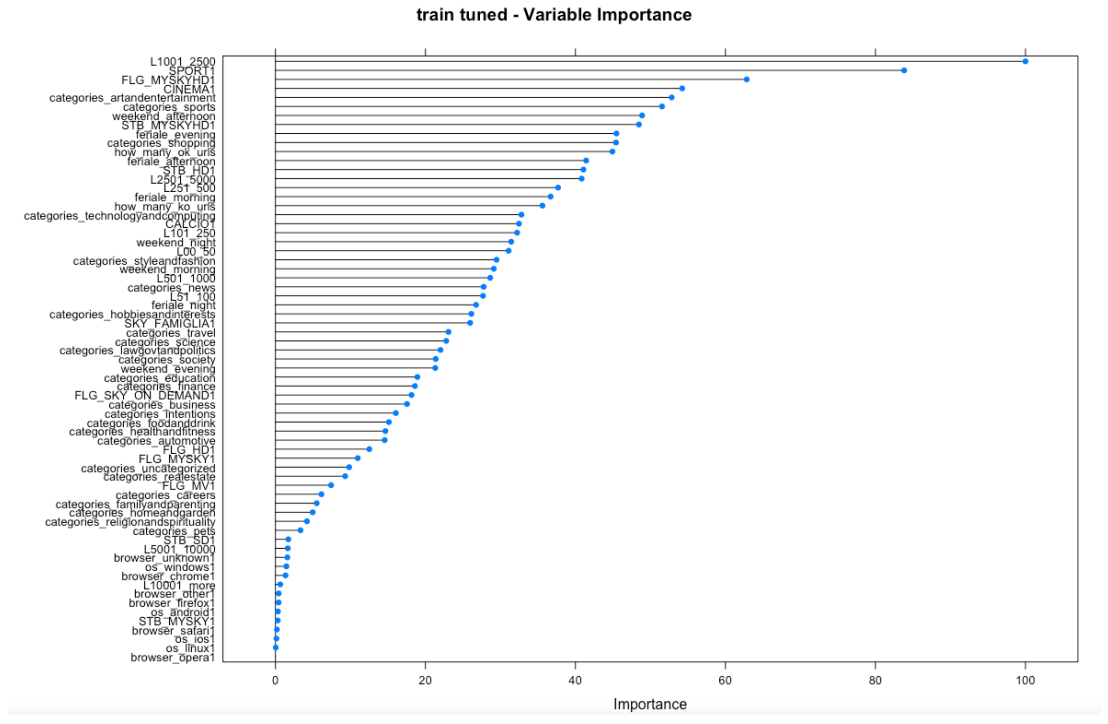
Per un futuro studio sulla propensione dei clienti al churn, potrebbe essere interessante avere a disposizione dei dati relativi ai movimenti del cliente all'interno della piattaforma, collegati al tipo di contratto attivo.
Inoltre, dal momento che molte variabili all'interno del dataset presentano near zero variance, per avere delle analisi più precise potrebbe essere utile avere a disposizione una raccolta dei dati più approfondita.

Appendice: importanza delle variabili nei singoli modelli

- Classification Tree

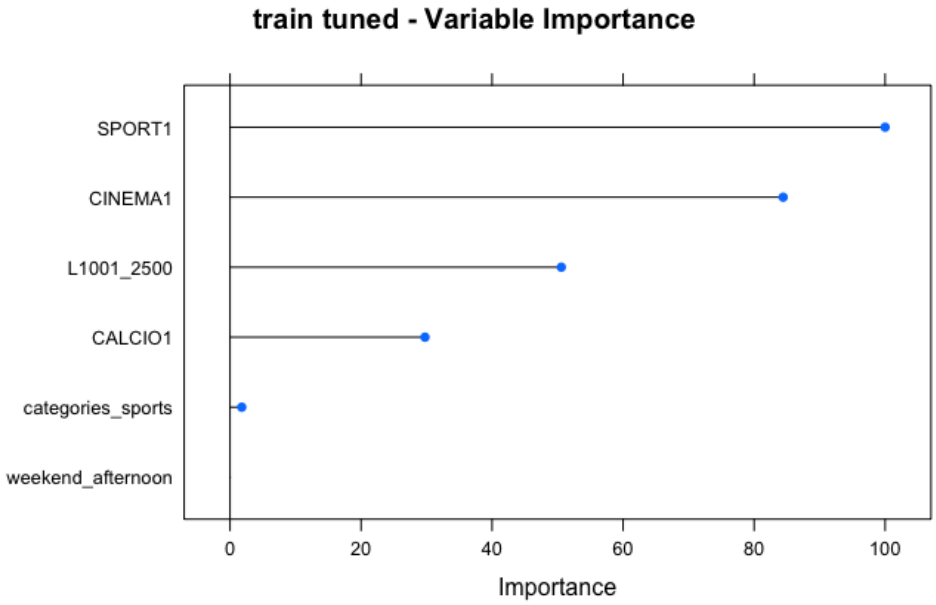


- Random Forest

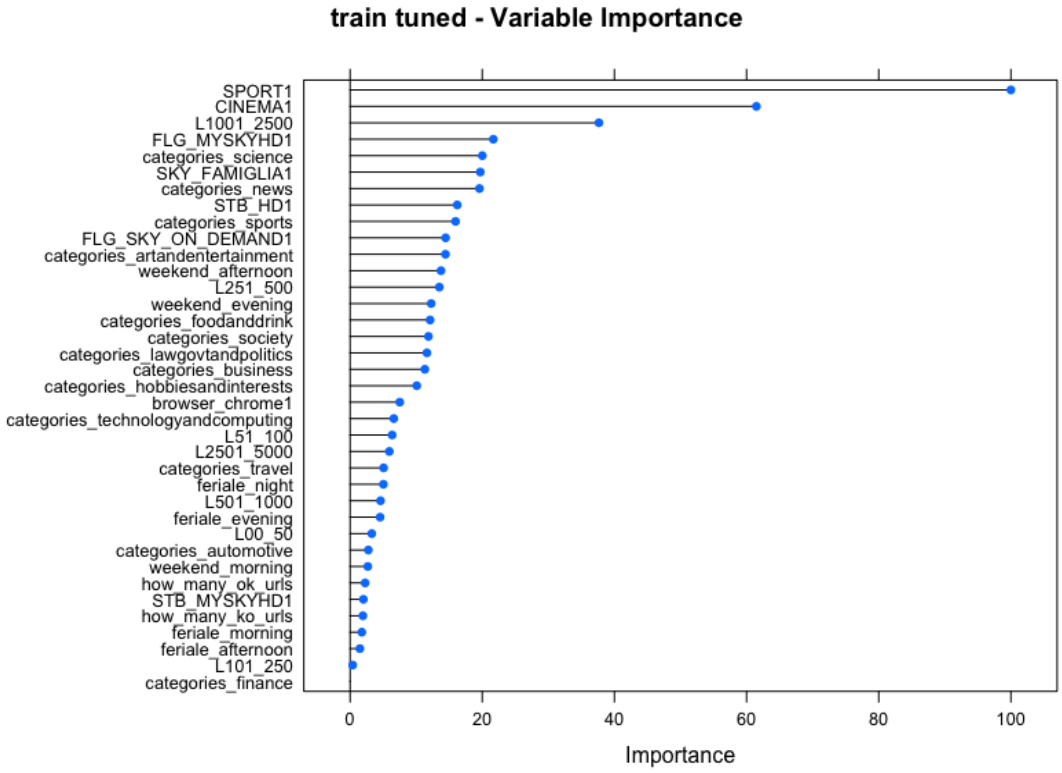


- Modello Logistico

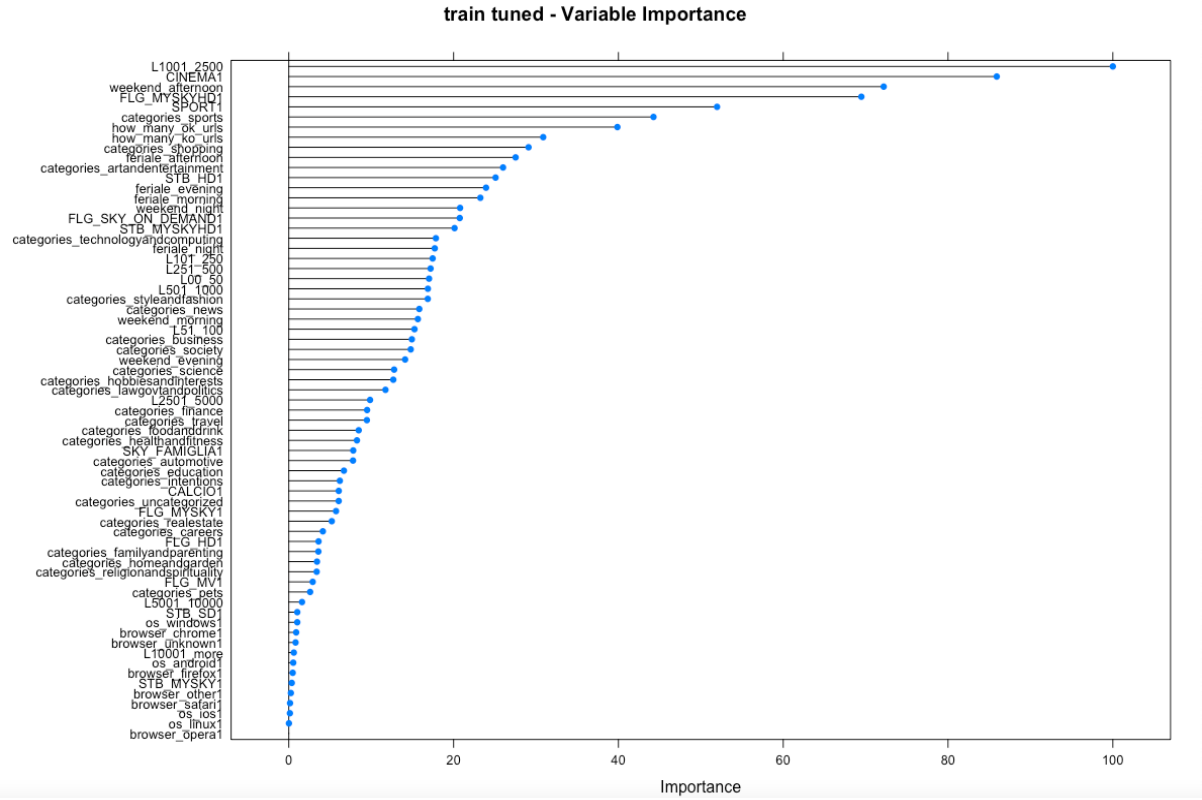
Il Modello logistico addestrato sul dataset contenente le variabili selezionate dal Classification Tree



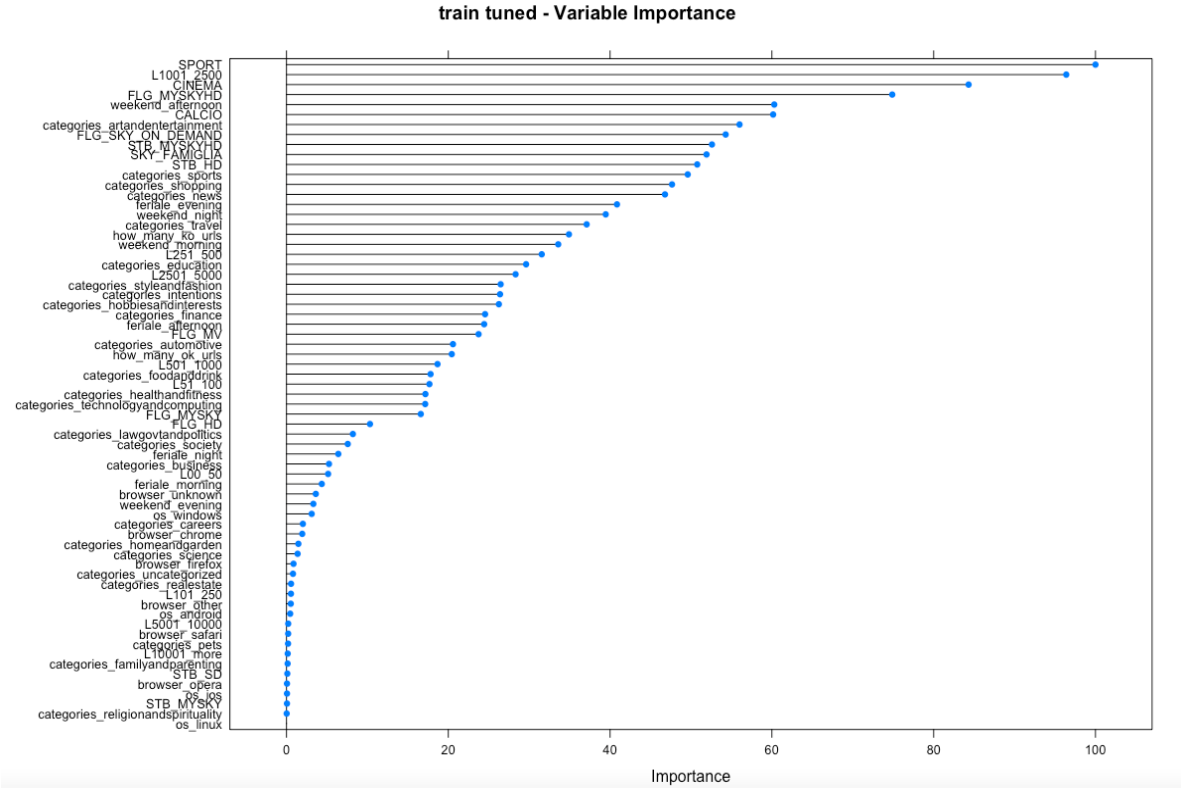
Il Modello logistico addestrato sul dataset contenente le variabili selezionate dal Random Forest



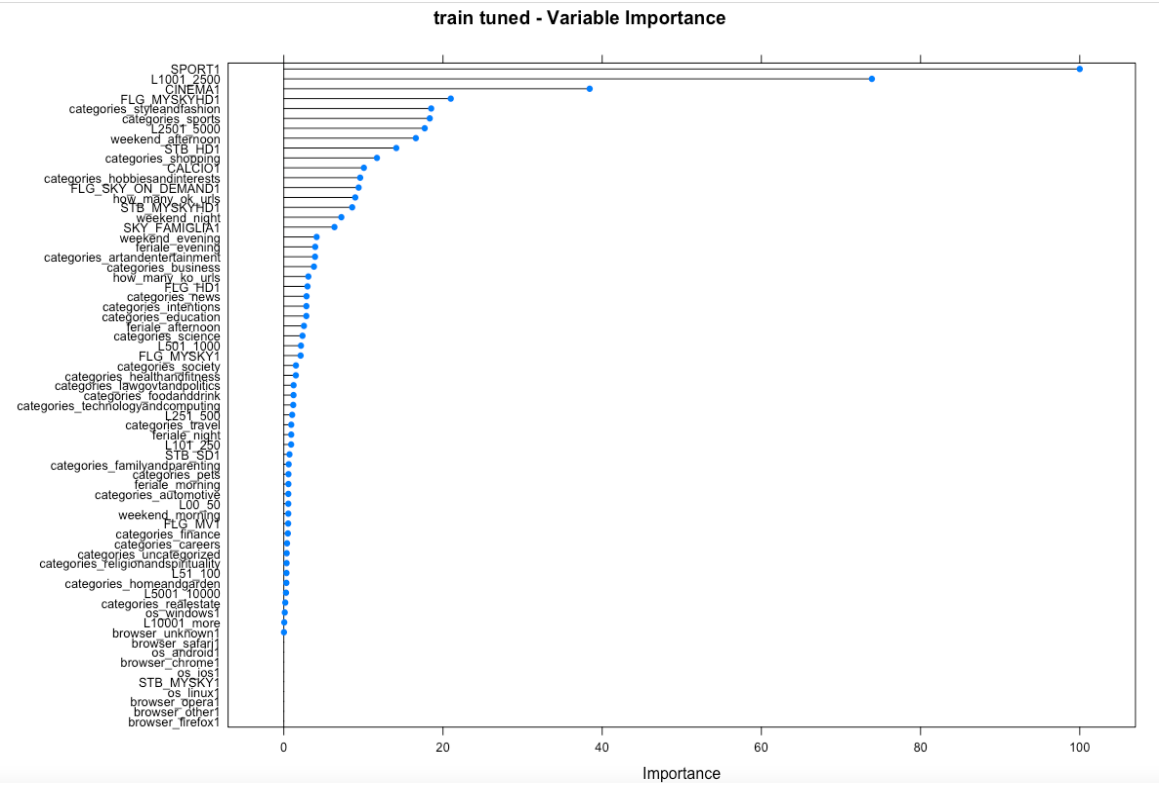
- Bagging Tree



- Naive Bayes

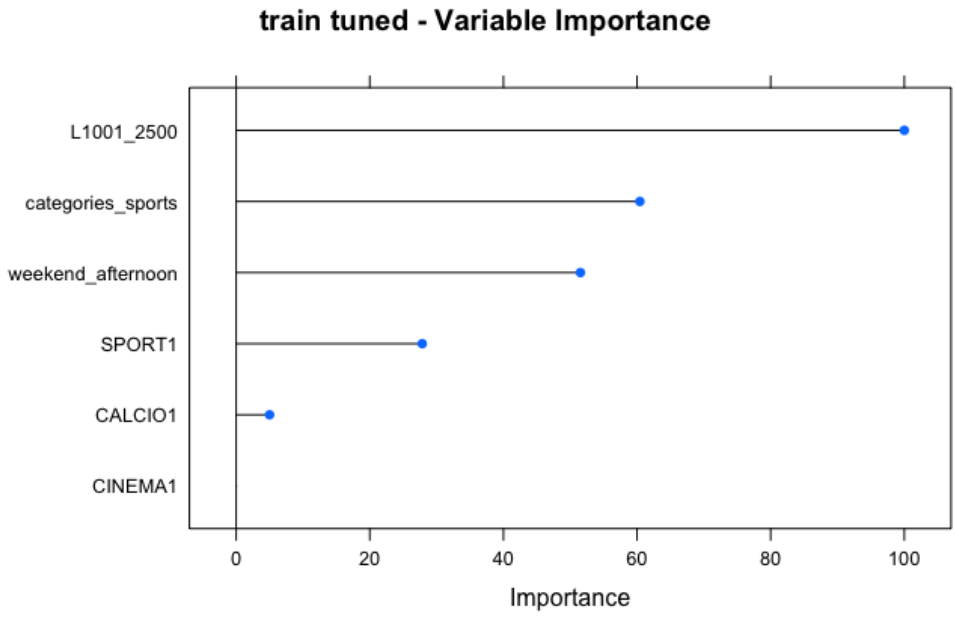


- Gradient Boosting

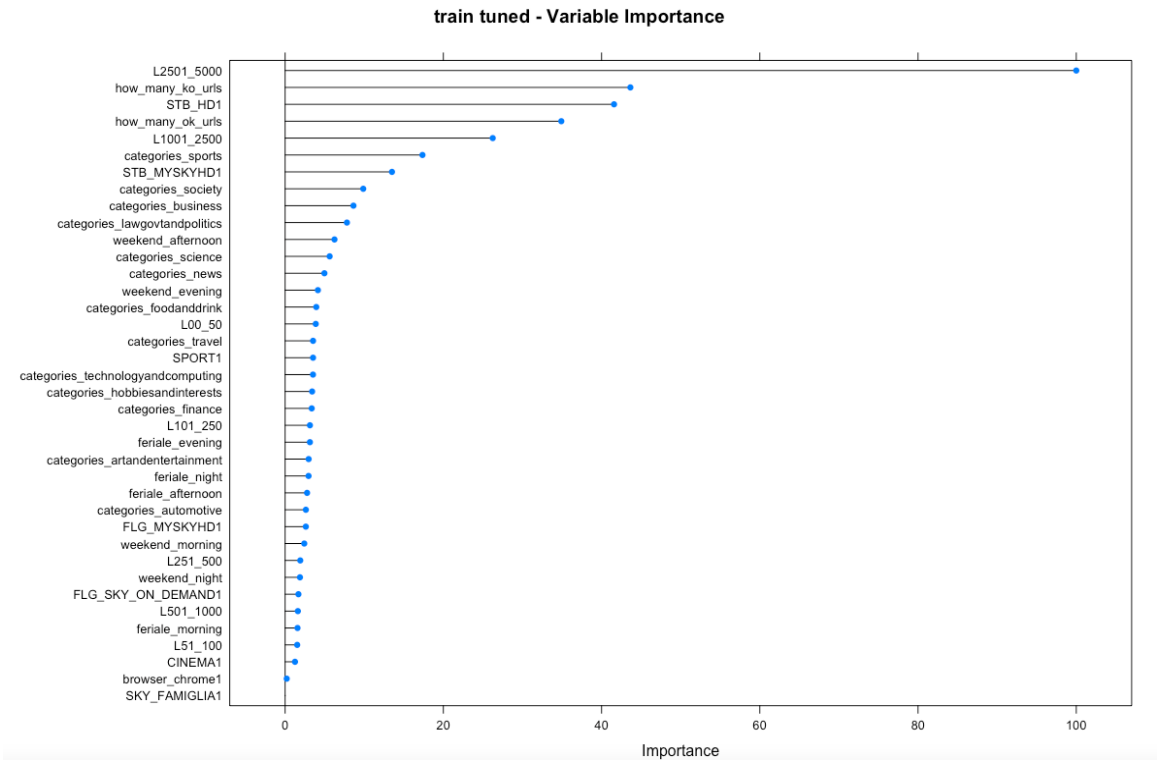


- Neural Network

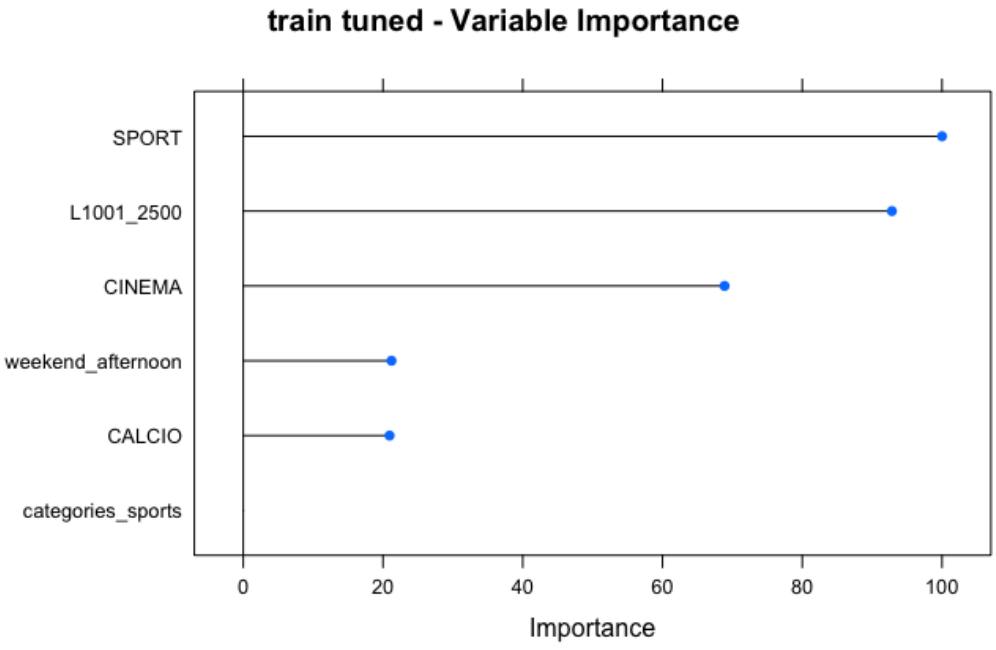
La Neural Network Single-Layer Perceptron addestrata sul dataset contenente le variabili selezionate dal Classification Tree



La Neural Network Single-Layer Perceptron addestrato sul dataset contenente le variabili selezionate dal Random Forest



Il modello Neural Network Multi-Layer Perceptron addestrato sul dataset contenente le variabili selezionate dal Classification Tree



Il modello Neural Network Multi-Layer Perceptron addestrato sul dataset contenente le variabili selezionate dal Random Forest