BANK CUSTOMER CHURN ANALYSIS

with SAS Model Studio

ECATERINA GURITANU AND FRANCESCO MILANESI

OBJECTIVE

Predict customer churn in a bank

HOW?

Using 5 Machine Learning models and determining the best-performing:

1. Logistic Regression
2. Decision Tree
3. Gradient Boosting
4. Random Forest
5. Neural Network

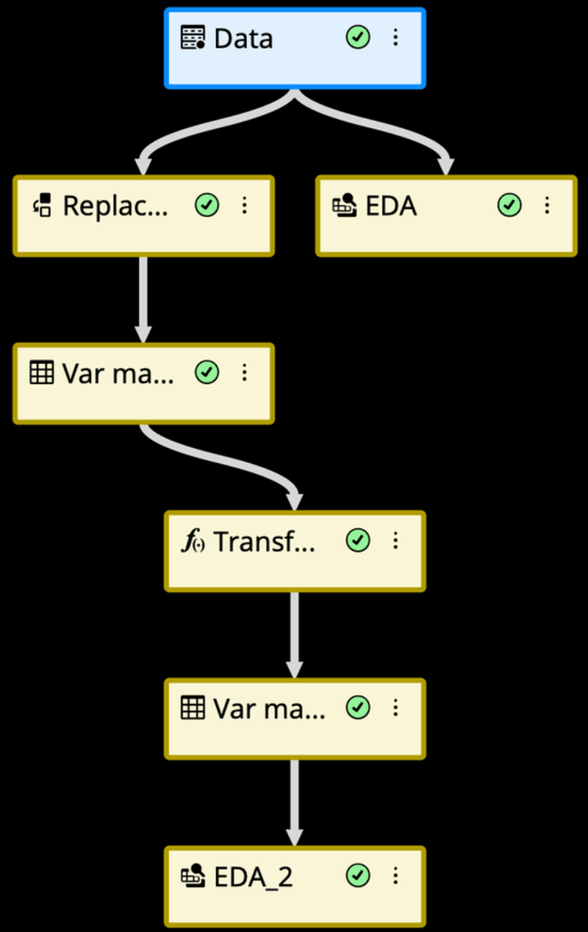
DATASET

Nome variabile ↑	☐	Tipo	Ruolo	Livello	Num. livelli	Mancanti	Minimo	Massimo
Age	☐	Numerico	Input	Intervallo	70	0.0000	18.0000	92.0000
Balance	☐	Numerico	Input	Intervallo	>254	0.0000	0.0000	250,898.0900
CreditScore	☐	Numerico	Input	Intervallo	>254	0.0000	350.0000	850.0000
CustomerId	☐	Numerico	ID	Intervallo	>254	0.0000	15,565,701.0000	15,815,690.0000
EstimatedSalary	☐	Numerico	Input	Intervallo	>254	0.0000	11.5800	199,992.4800
Exited	☐	Numerico	Input	Binario	2	0.0000	0.0000	1.0000
Gender	☐	Alfanumerico	Input	Binario	2	0.0000		
Geography	☐	Alfanumerico	Input	Nominale	3	0.0000		
HasCrCard	☐	Numerico	Input	Binario	2	0.0000	0.0000	1.0000
IsActiveMember	☐	Numerico	Input	Binario	2	0.0000	0.0000	1.0000
NumOfProducts	☐	Numerico	Input	Nominale	4	0.0000	1.0000	4.0000
RowNumber	☐	Numerico	ID	Intervallo	>254	0.0000	1.0000	10,000.0000
Surname	☐	Alfanumerico	ID	Nominale	>254	0.0000		
Tenure	☐	Numerico	Input	Nominale	11	0.0000	0.0000	10.0000

DATA PREPARATION

Nome variabile ↑	☐	Tipo	Ruolo	Livello	Num. livelli	Mancanti	Minimo	Massimo
Age	☐	Numerico	Input	Intervallo	70	0.0000	18.0000	92.0000
Balance	☐	Numerico	Input	Intervallo	>254	0.0000	0.0000	250,898.0900
CreditScore	☐	Numerico	Input	Intervallo	>254	0.0000	350.0000	850.0000
CustomerId	☐	Numerico	ID	Intervallo	>254	0.0000	15,565,701.0000	15,815,690.0000
EstimatedSalary	☐	Numerico	Rifiutato	Intervallo	>254	0.0000	11.5800	199,992.4800
Exited	☐	Numerico	Target	Binario	2	0.0000	0.0000	1.0000
Gender	☐	Alfanumerico	Input	Binario	2	0.0000		
Geography	☐	Alfanumerico	Input	Nominale	3	0.0000		
HasCrCard	☐	Numerico	Input	Binario	2	0.0000	0.0000	1.0000
IsActiveMember	☐	Numerico	Input	Binario	2	0.0000	0.0000	1.0000
NumOfProducts	☐	Numerico	Input	Nominale	4	0.0000	1.0000	4.0000
RowNumber	☐	Numerico	Rifiutato	Intervallo	>254	0.0000	1.0000	10,000.0000
Surname	☐	Alfanumerico	Rifiutato	Nominale	>254	0.0000		
Tenure	☐	Numerico	Input	Nominale	11	0.0000	0.0000	10.0000

BASE PIPELINE



Exploratory Data Analysis

- no missing values
- AGE is highly skewed
- platykurtic distributions

Replacement makes no replacements

Variable Management selects only the variables to transform

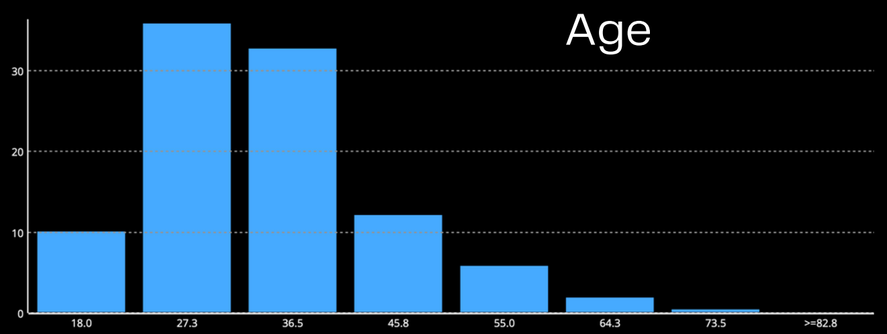
Transformation

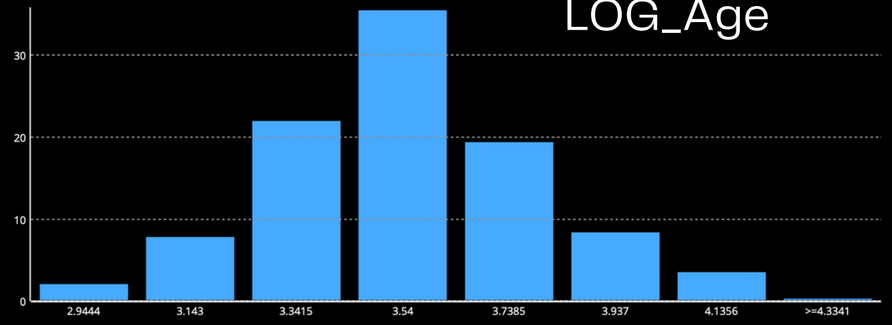
- applies log transformation to AGE
- applies one-hot encoding to categorical variables

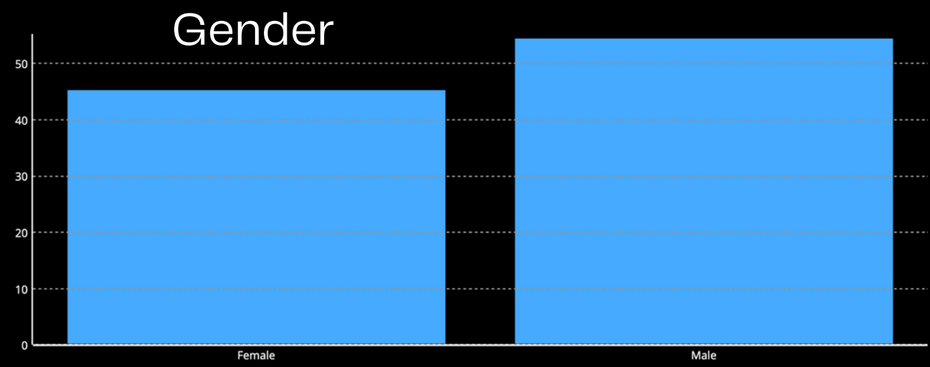
Variable Management reintroduces all variables

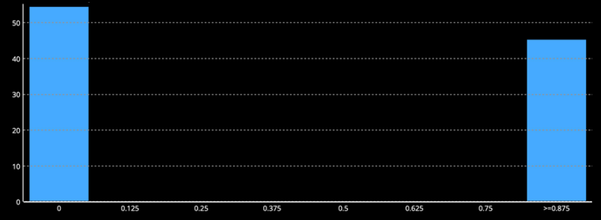
Exploratory Data Analysis

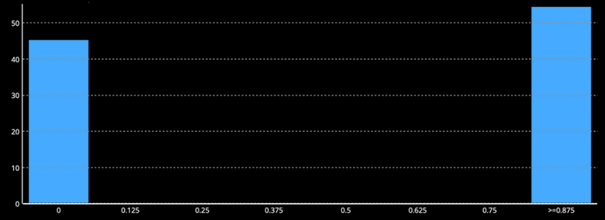
RESULTS OF TRANSFORMATION





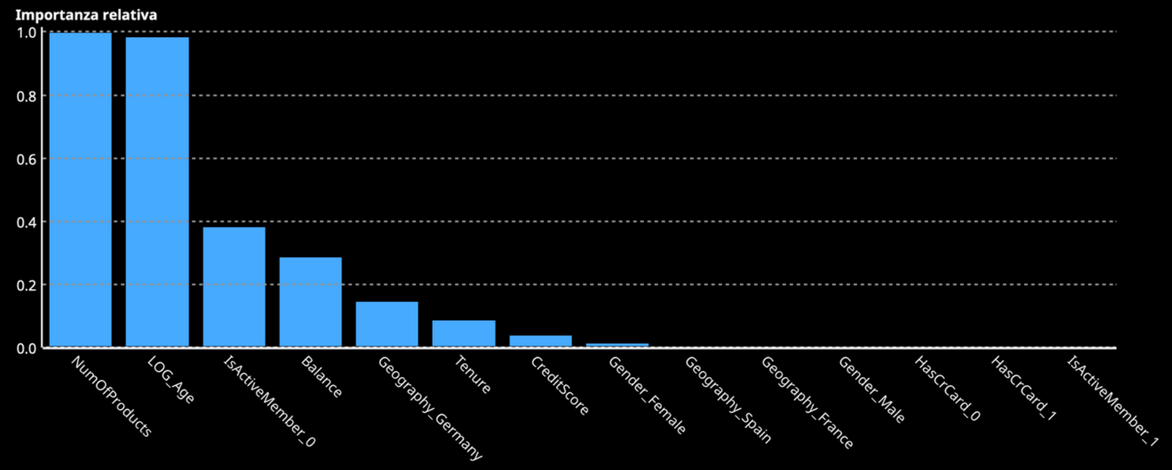






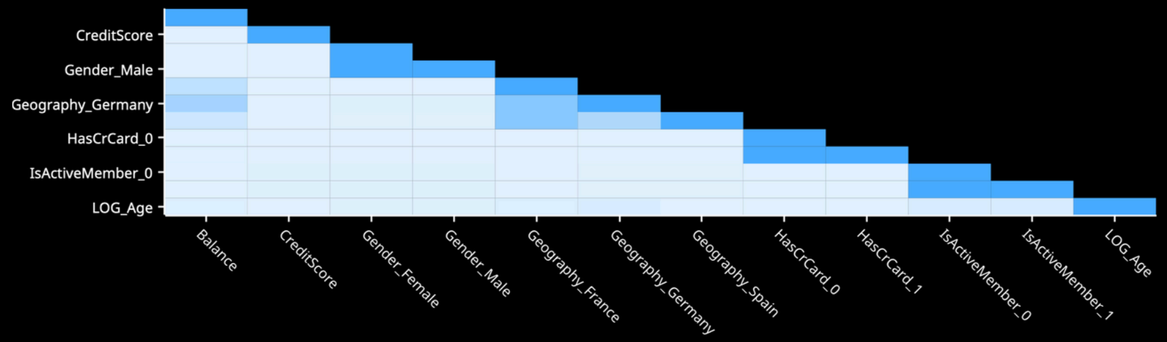
RESULTS OF EDA

Variable Importance

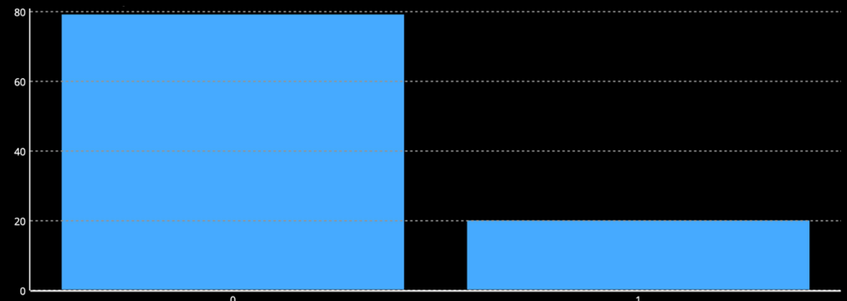


Most important variables: NumOfProducts, LOG_Age
Least important variables: Geography_Spain, Geography_France, Gender_Male, HasCrCard_0, HasCrCard_1, IsActiveMember_1

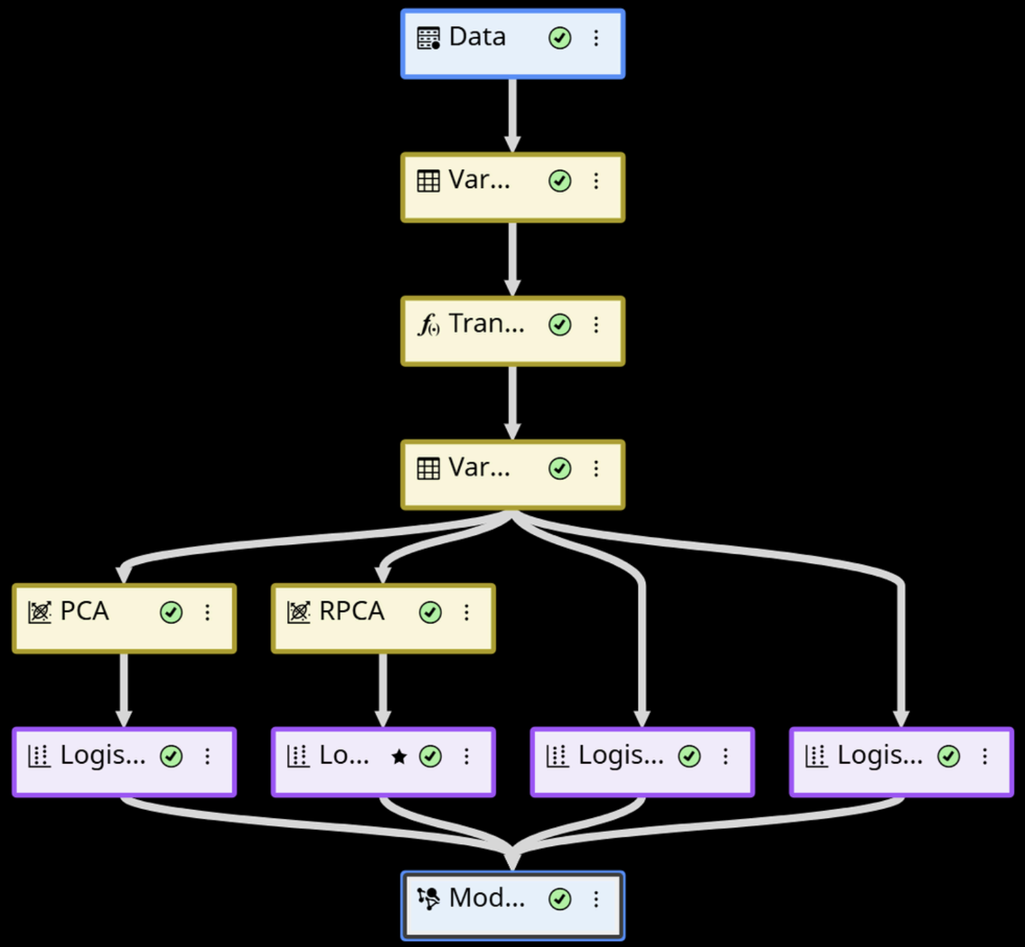
Pearson Correlation



Distribution of Exited



LOGISTIC REGRESSION

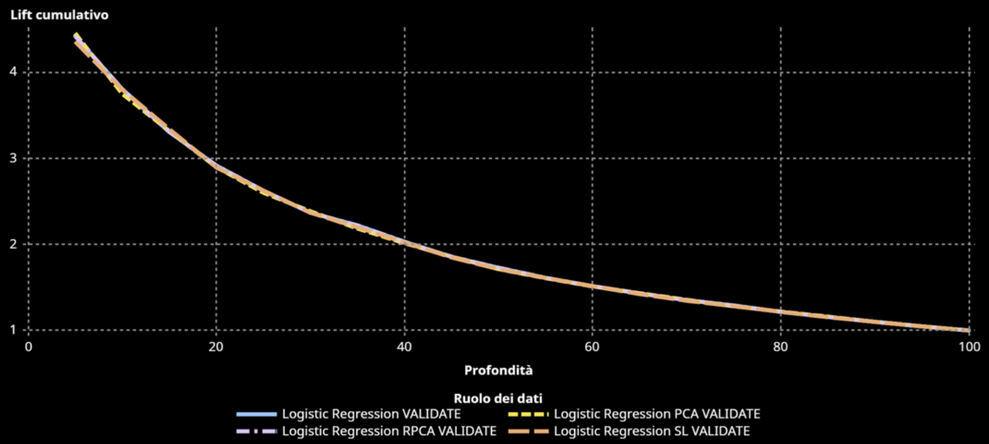


- Logistic regression after PCA
- Logistic regression after RPCA
- Logistic regression with default paramiters
- Logistic regression with significance level selection

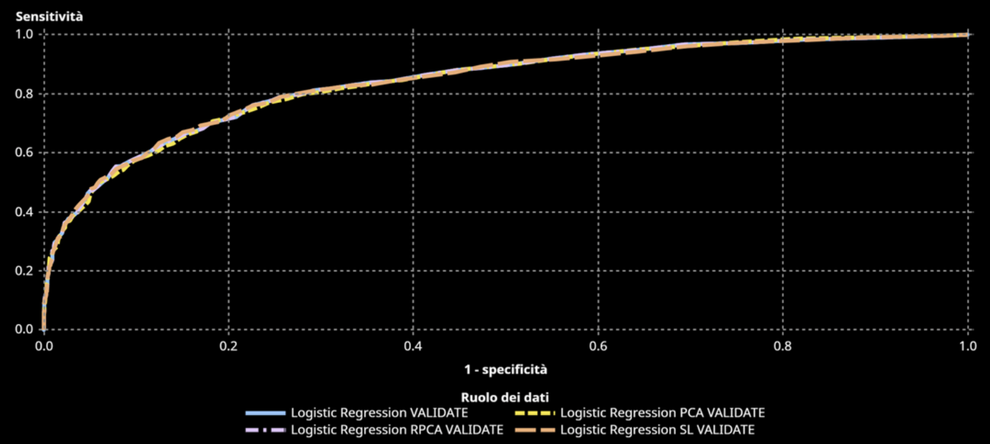
LOGISTIC REGRESSION - Model Comparison

Cha...	Nome	KS (Youden)	Accuratezza	Average Squar...	Area sotto ROC	Lift cumulativo	Percentuale ...	Score F1	Errore di classi...
★	Logistic Regression RPCA	0.5495	0.8500	0.1099	0.8458	3.6275	36.2745	0.5313	0.1500
	Logistic Regression PCA	0.5491	0.8510	0.1094	0.8479	3.7745	37.7451	0.5300	0.1490
	Logistic Regression	0.5495	0.8500	0.1099	0.8458	3.6275	36.2745	0.5313	0.1500
	Logistic Regression SL	0.5460	0.8460	0.1099	0.8446	3.6765	36.7647	0.5157	0.1540

Cumulative lift



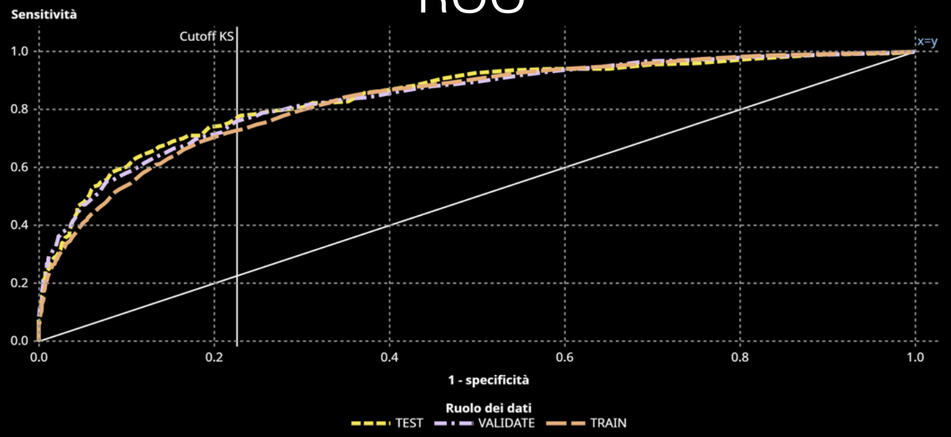
ROC



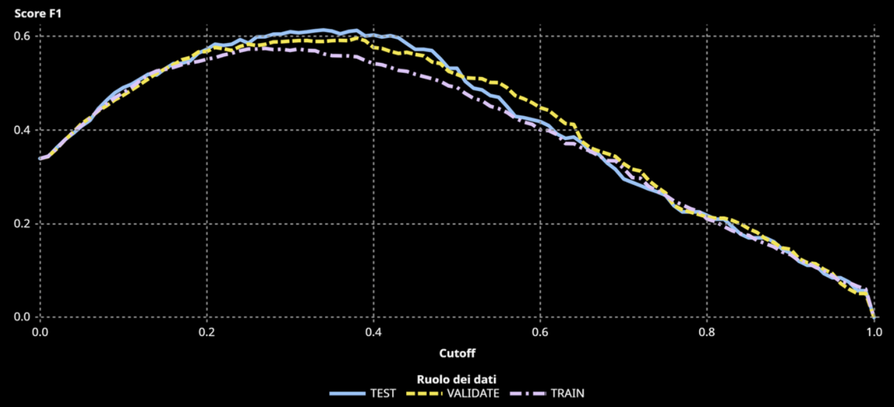
LOGISTIC REGRESSION - The Best Model

Effetto	Parametro	Valore t	Segno	Stima	Stima assol...	Errore stand...	Chi-quadrato	Pr > Chi-qu...
Intercept	Intercept	25.6780	-	-15.0747	15.0747	0.5871	659.3605	0.0000
LOG_Age	LOG_Age	22.3053	+	3.3722	3.3722	0.1512	497.5275	0.0000
IsActiveMember_0	IsActiveMember_0	14.1240	+	1.0368	1.0368	0.0734	199.4865	0.0000
Geography_Germany	Geography_Germany	8.5430	+	0.6987	0.6987	0.0818	72.9831	0.0000
Gender_Female	Gender_Female	6.9134	+	0.4844	0.4844	0.0701	47.7953	0.0000

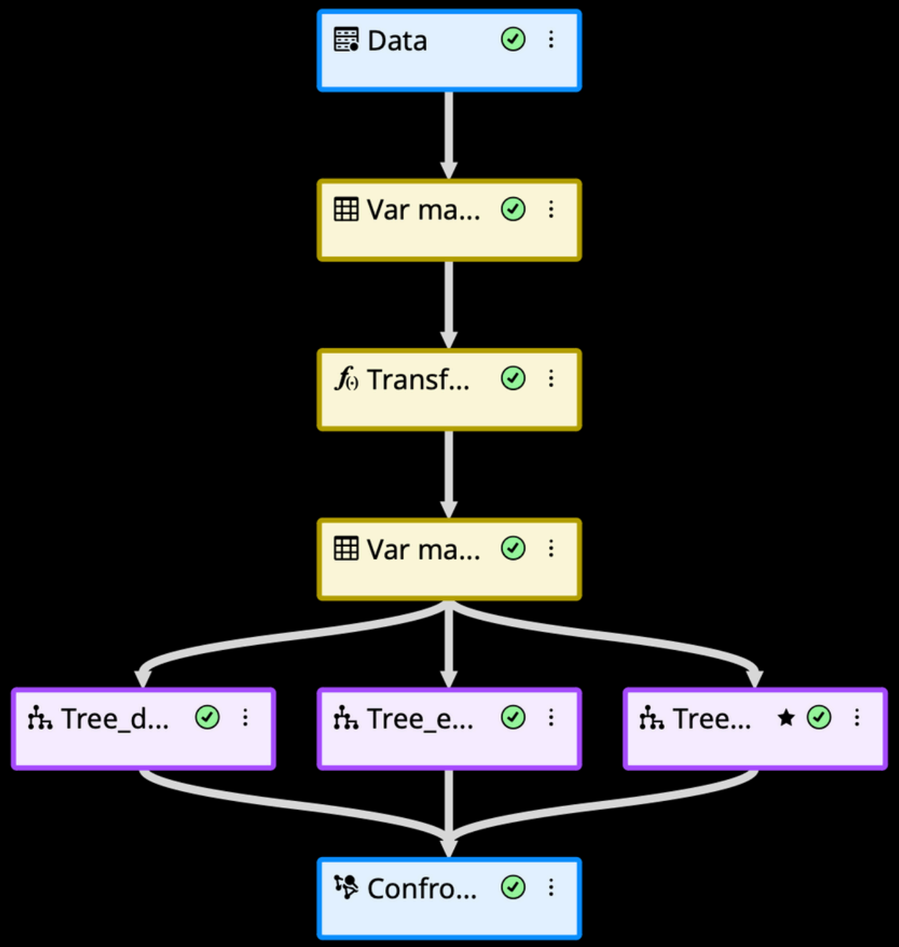
ROC



F1 Score



DECISION TREE

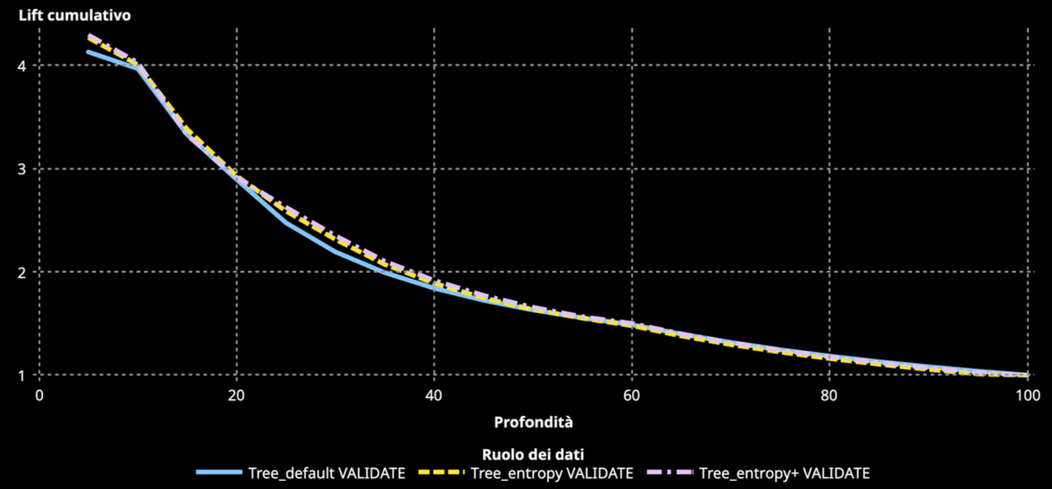


- Decision Tree with default parameters
- Decision Tree with entropy as splitting criteria
- Decision Tree with entropy as splitting criteria and minimum leaf size 6

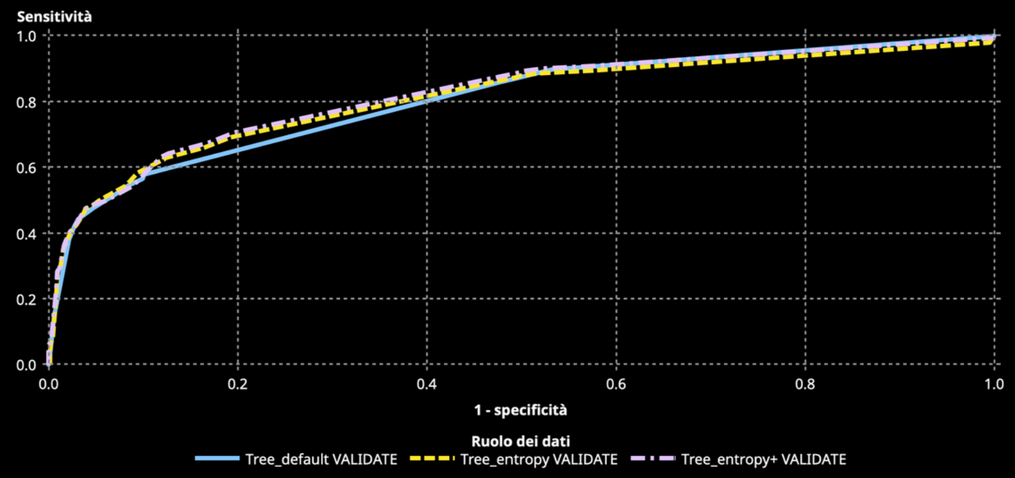
DECISION TREE – Model Comparison

Cha...	Nome	KS (Youden)	Accuratezza	Average S...	Area sotto ...	Lift cumula...	Percentual...	Score F1	Errore di classif...
★	Tree_entropy+	0.5290	0.8550	0.1136	0.8169	3.9608	39.6078	0.5723	0.1450
	Tree_default	0.4836	0.8570	0.1162	0.7897	3.9608	39.6078	0.5653	0.1430
	Tree_entropy	0.5253	0.8540	0.1143	0.8131	3.9837	39.8366	0.5731	0.1460

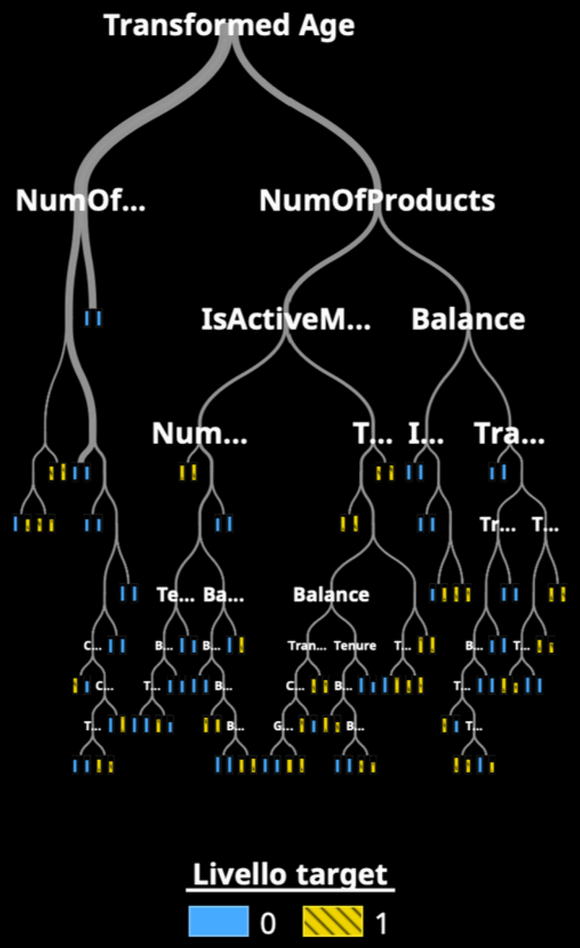
Cumulative lift



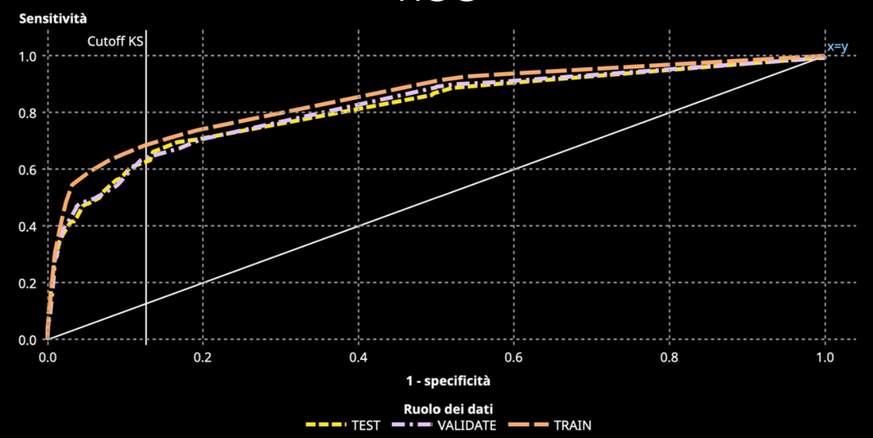
ROC



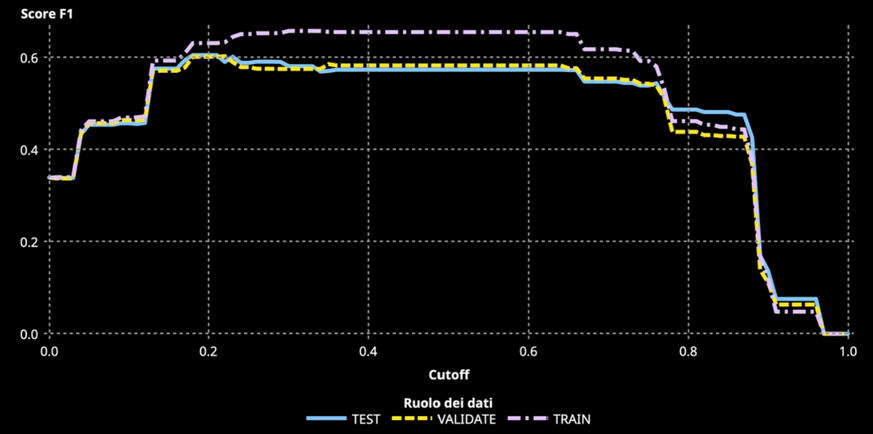
DECISION TREE – The Best Model



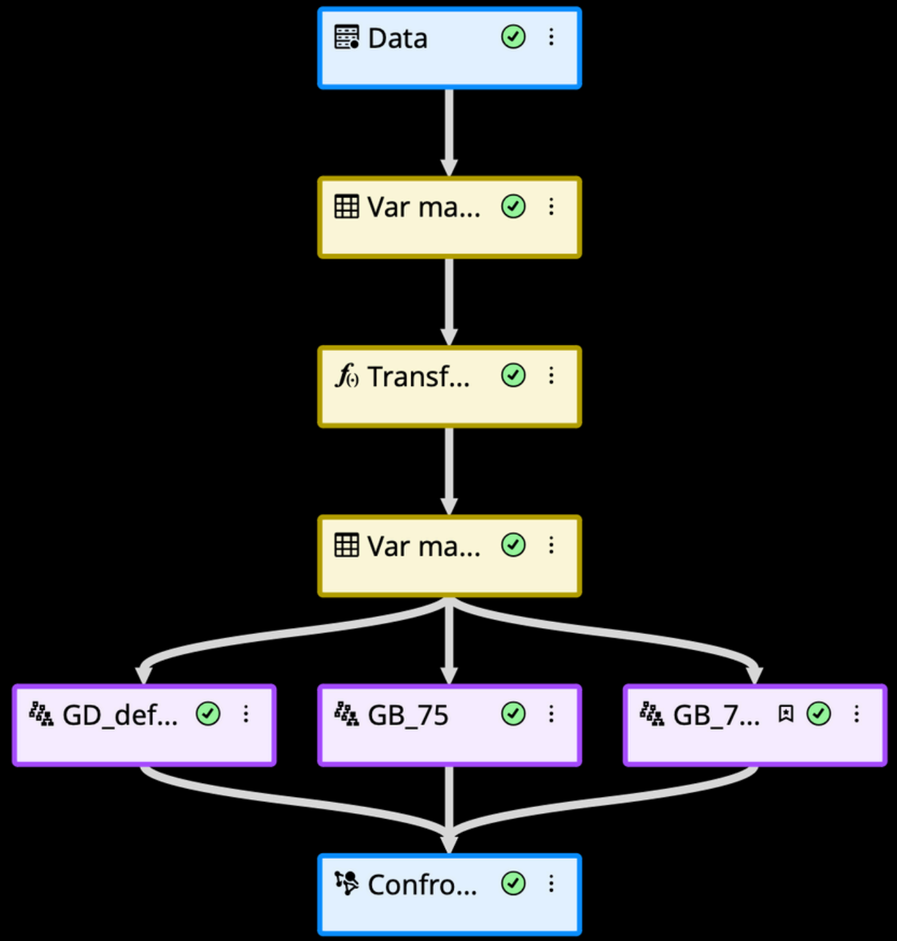
ROC



F1 Score



GRADIENT BOOSTING

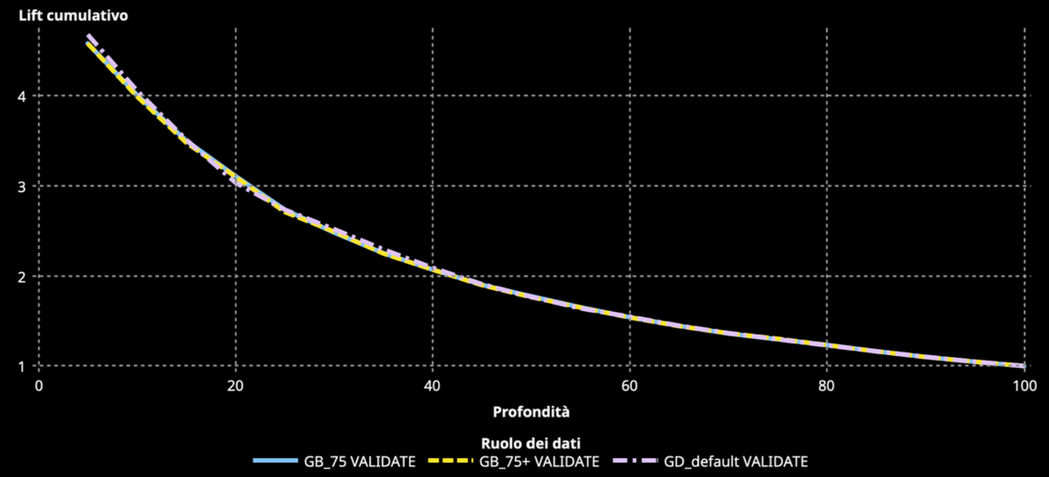


- Gradient Boosting with default parameters
- Gradient Boosting with 75 continuous groupings
- Gradient Boosting with 75 continuous groupings and 150 trees

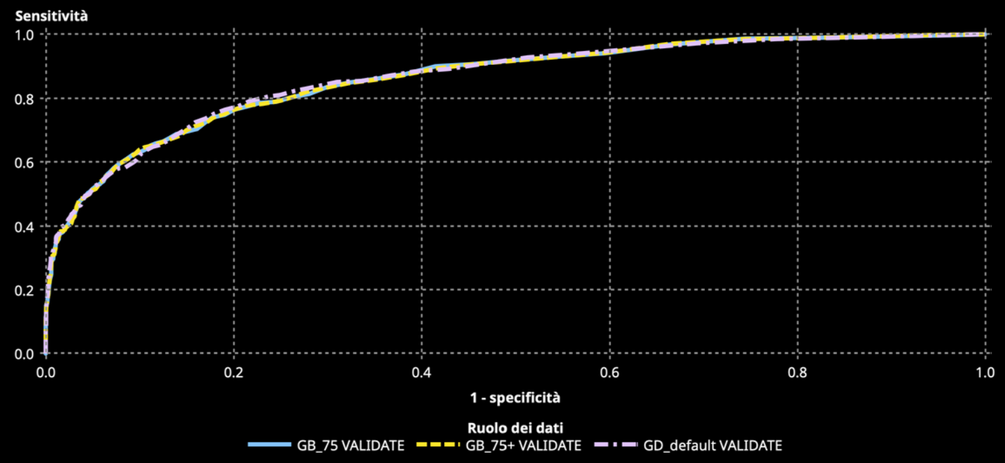
GRADIENT BOOSTING – Model Comparison

C.. ↑	Nome	KS (Youden)	Accuratezza	Average Sq...	Area sotto ...	Lift cumula...	Percentual...	Score F1	Errore di classifi...
★	GB_75+	0.5977	0.8640	0.1013	0.8643	3.9706	39.7059	0.6158	0.1360
	GB_75	0.5950	0.8590	0.1014	0.8640	3.9706	39.7059	0.5960	0.1410
	GD_default	0.5908	0.8630	0.1022	0.8630	3.8235	38.2353	0.5982	0.1370

Cumulative lift

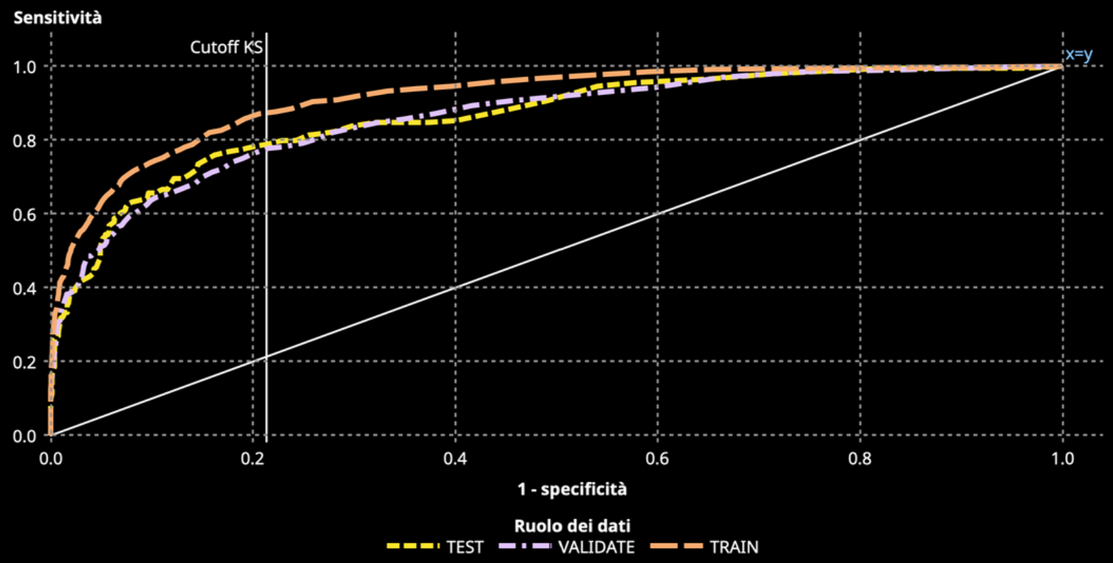


ROC

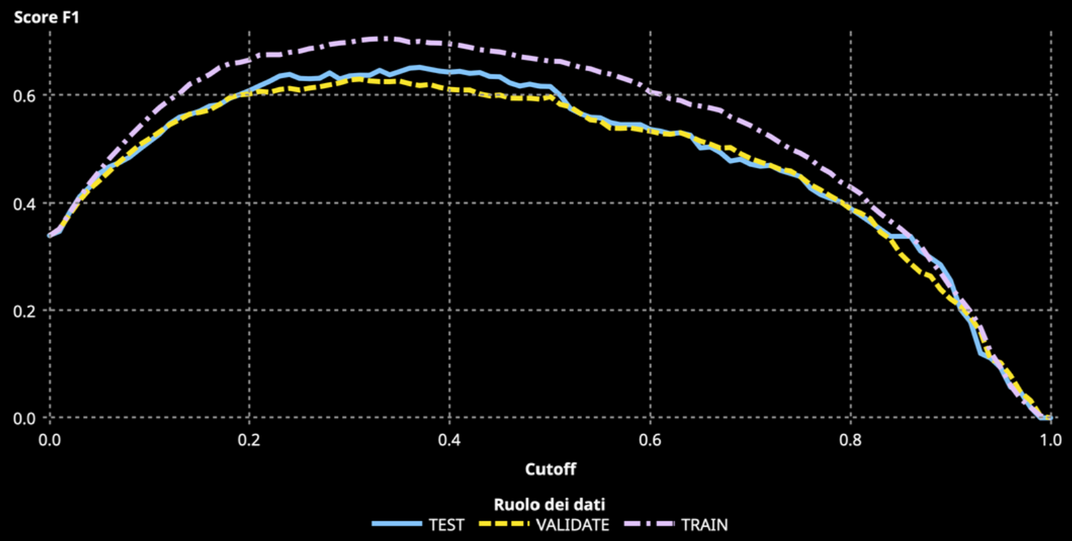


GRADIENT BOOSTING – The Best Model

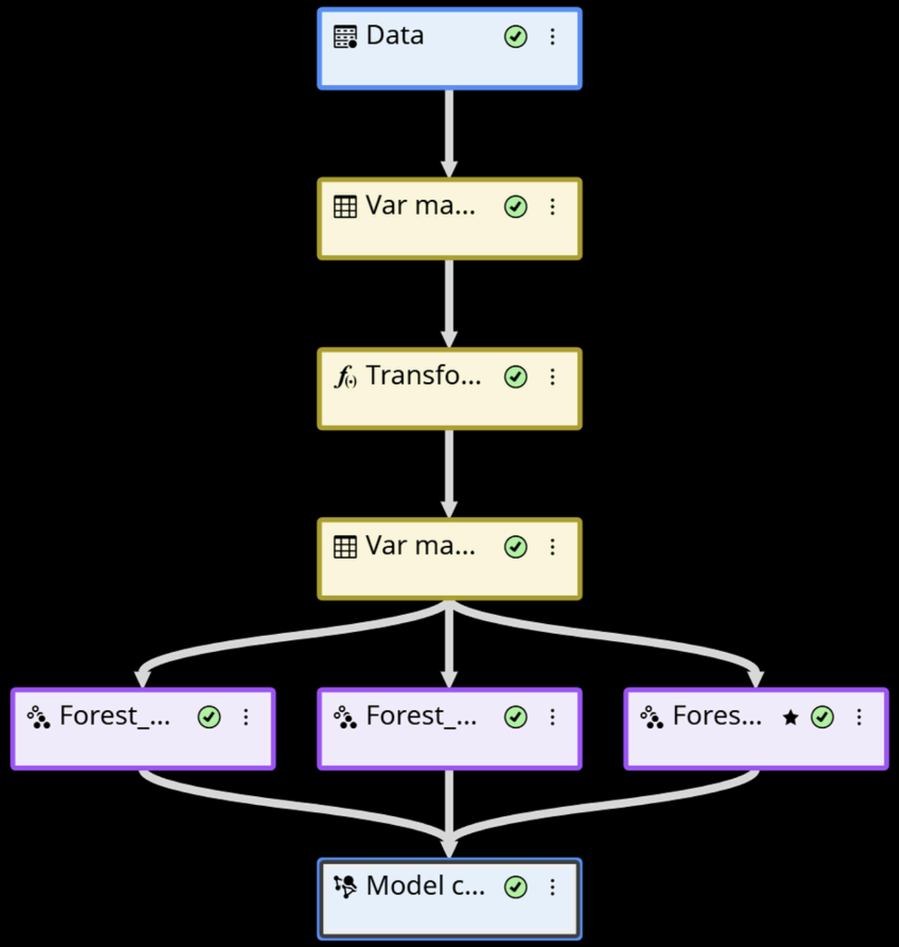
ROC



F1 Score



RANDOM FOREST

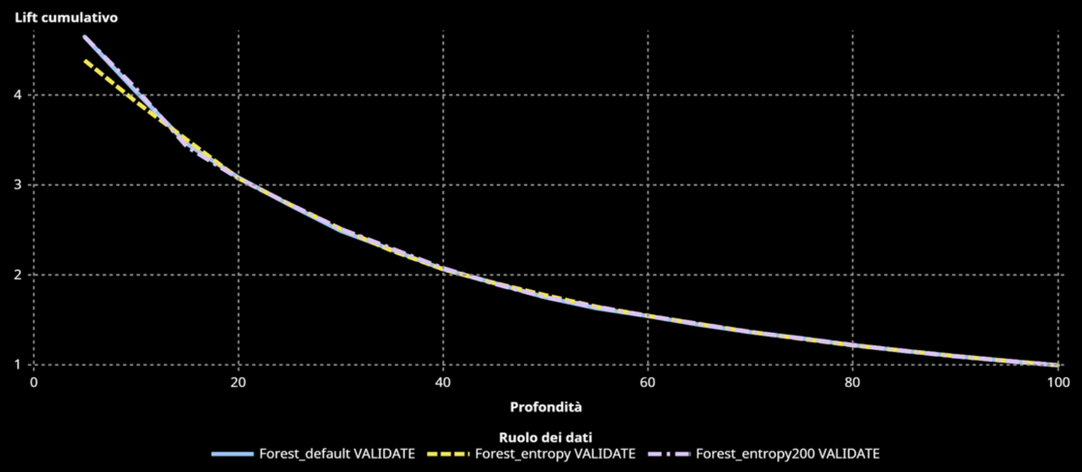


- Random forest with default parameters
- Random forest with entropy split criteria
- Random forest with entropy split criteria and 200 trees

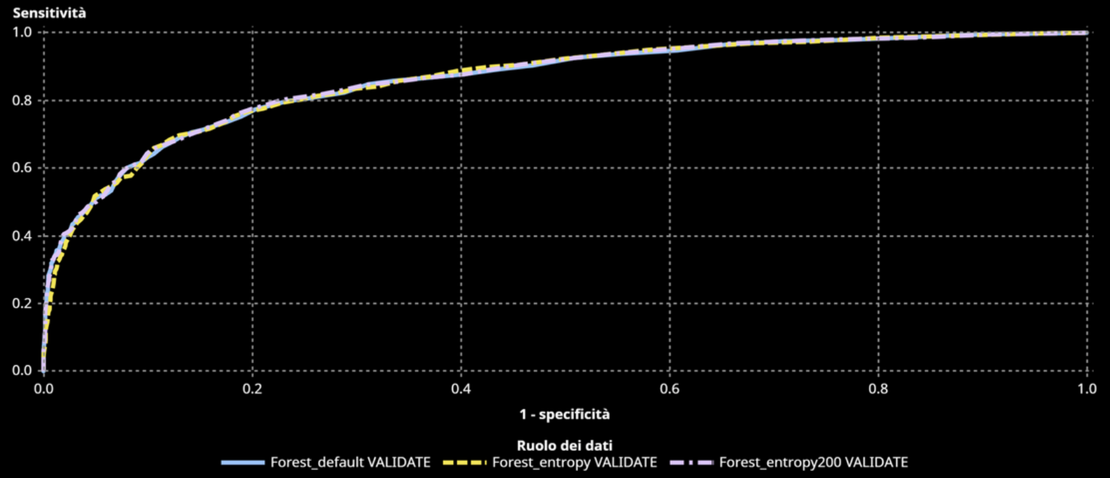
RANDOM FOREST – Model Comparison

Champion	Nome	KS (Youden)	Accuratez...	Average Squ...	Area sotto R...	Lift cumulativo	Percentuale ...	Score F1	Errore di cla...
★	Forest_entropy200	0.5643	0.8620	0.1039	0.8613	3.9216	39.2157	0.5688	0.1380
	Forest_default	0.5598	0.8630	0.1038	0.8591	4.0196	40.1961	0.5732	0.1370
	Forest_entropy	0.5586	0.8590	0.1045	0.8618	3.9706	39.7059	0.5635	0.1410

Cumulative lift

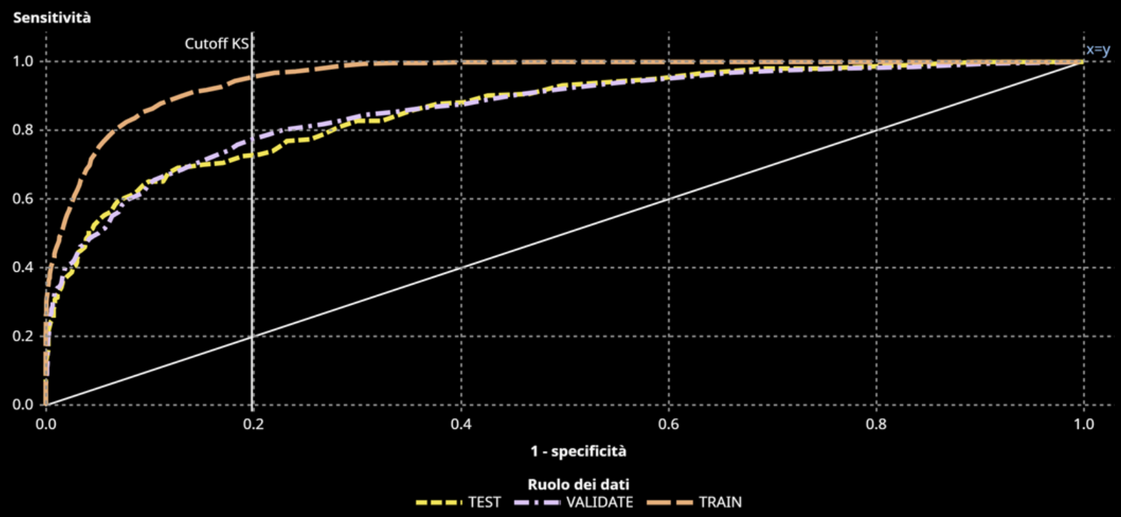


ROC

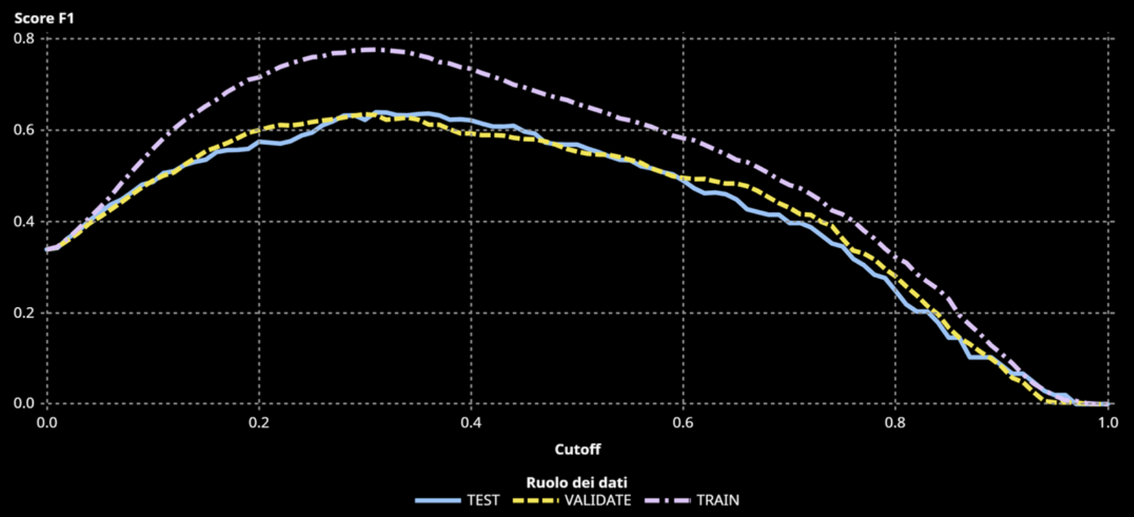


RANDOM FOREST – The Best Model

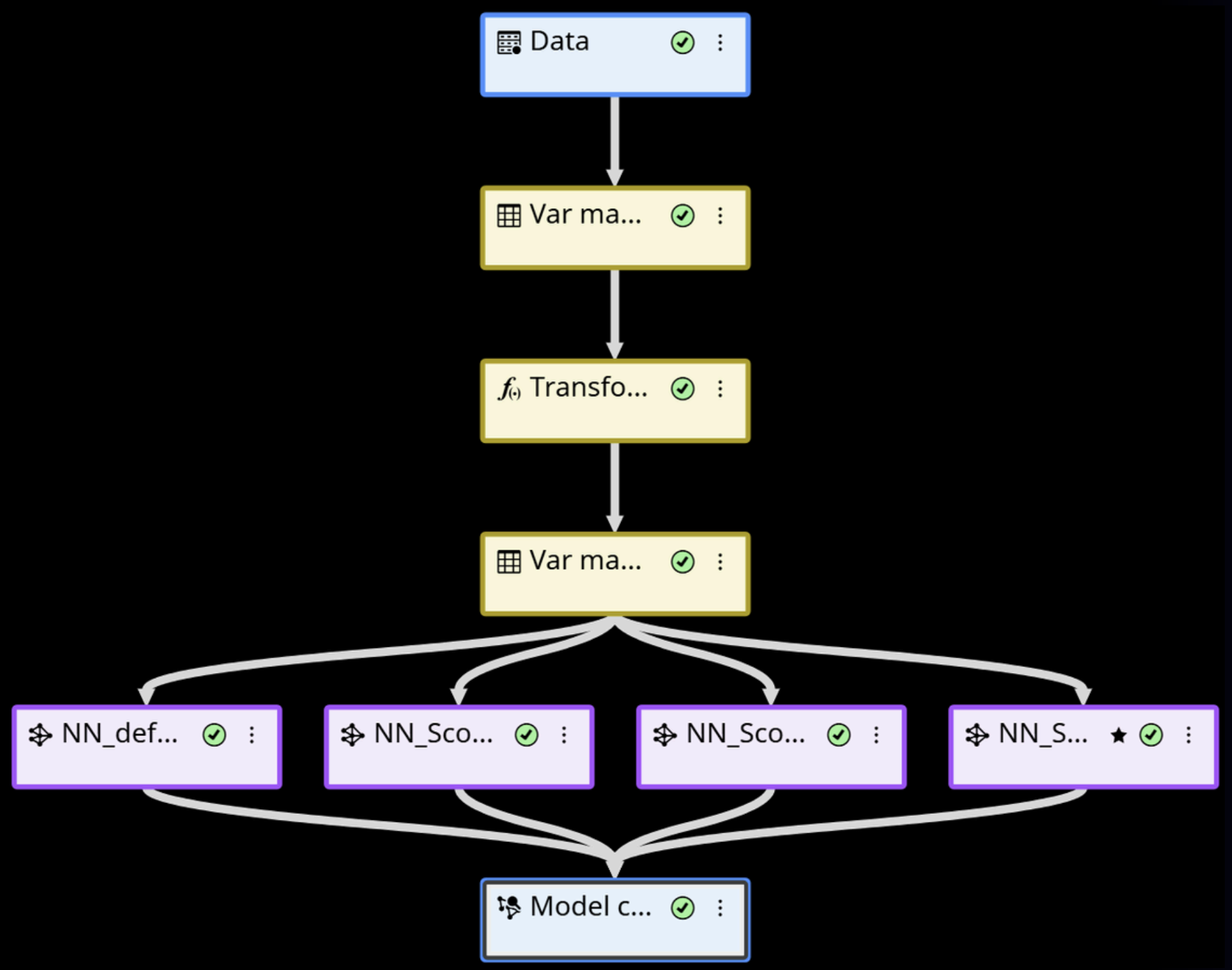
ROC



F1 Score



NEURAL NETWORK

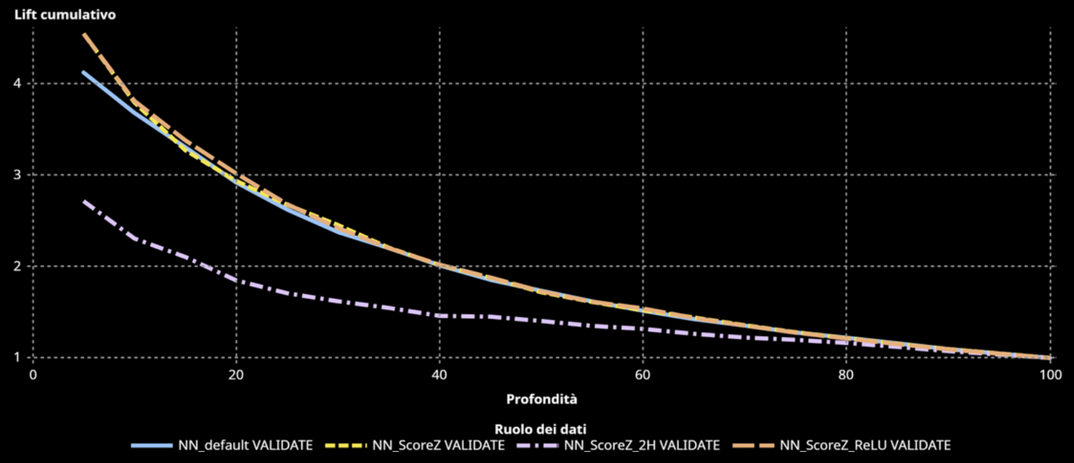


- Neural network with default parameters
- Neural network with ScoreZ standardization
- Neural network with 2 hidden layers
- Neural network with ReLU activation function

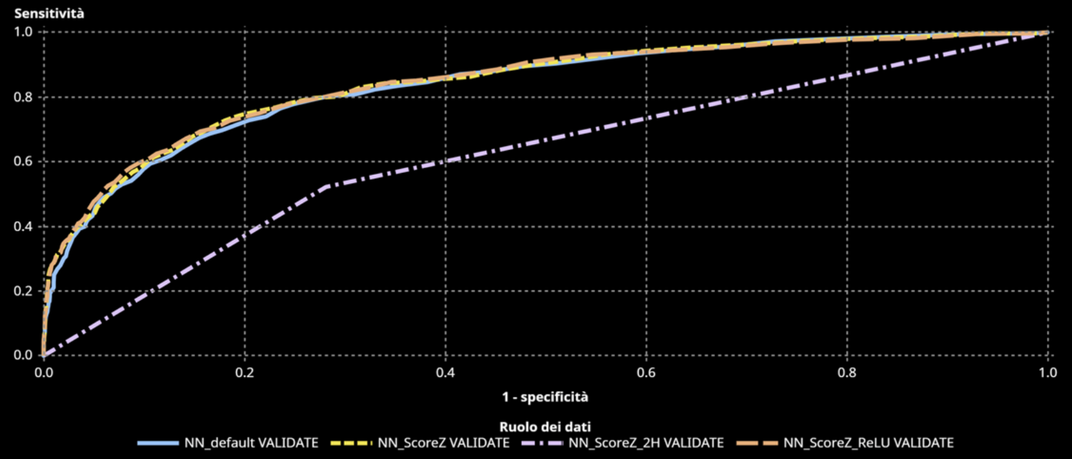
NEURAL NETWORK – Model Comparison

Champ...	Nome	KS (Youden)	Accuratezza	Average Squ...	Area sotto R...	Lift cumulati...	Percentuale ...	Score F1	Errore di cla...
★	NN_ScoreZ_ReLU	0.5722	0.8610	0.1086	0.8554	3.9216	39.2157	0.5559	0.1390
	NN_default	0.5418	0.8350	0.1139	0.8444	3.5294	35.2941	0.4407	0.1650
	NN_ScoreZ	0.5529	0.8600	0.1092	0.8533	3.9706	39.7059	0.5513	0.1400
	NN_ScoreZ_2H	0.1593	0.7960	0.1653	0.5796	2.0098	20.0980	0	0.2040

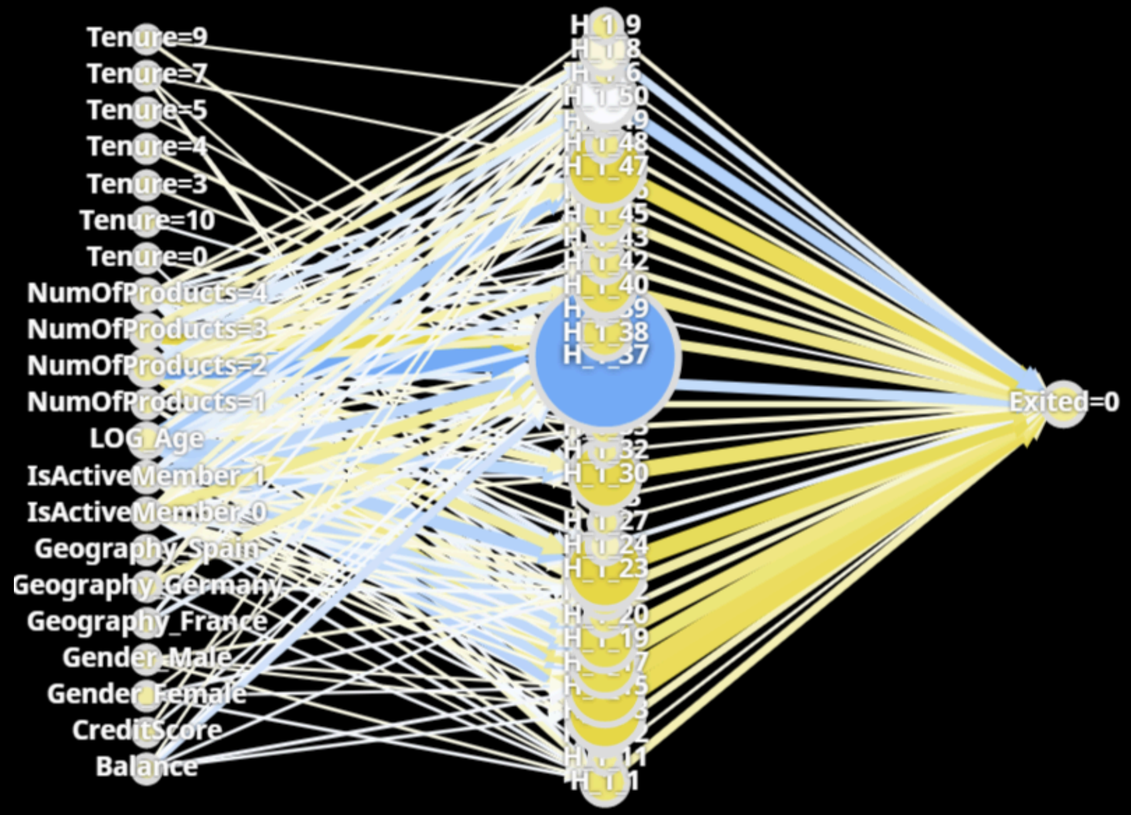
Cumulative lift



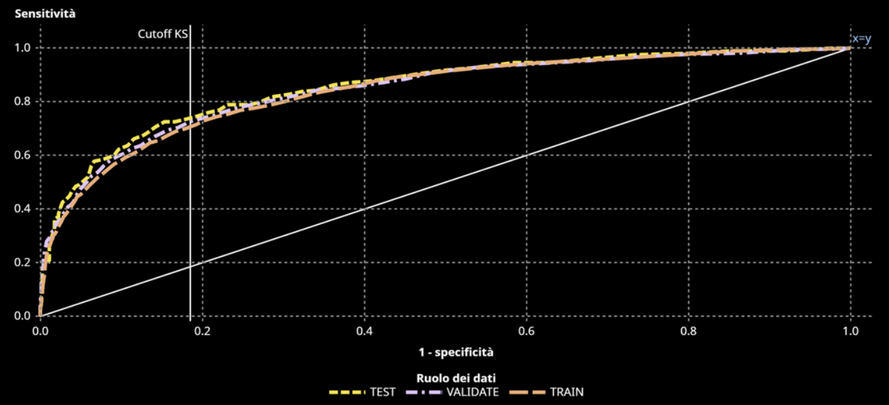
ROC



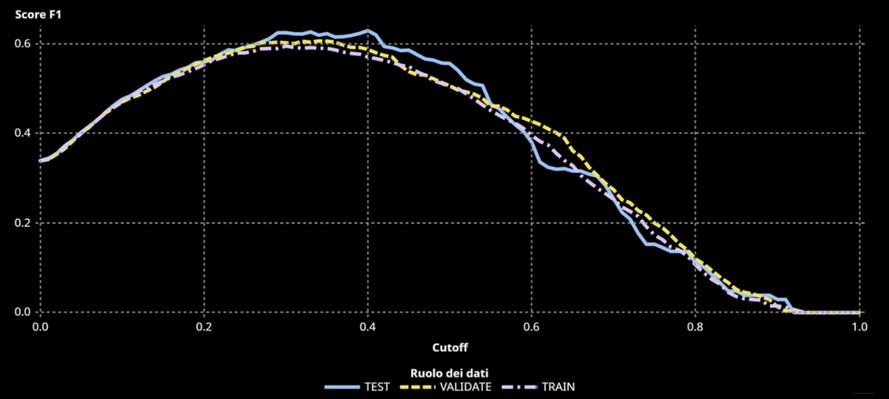
NEURAL NETWORK – The Best Model



ROC



F1 Score



PIPELINE COMPARISON

Champion ↓	Nome	Nome algoritmo	Nome della pipeline	KS (Youden)	Num. osservazioni
	GB_75+	Gradient boosting	Gradient Boosting	0.5977	1,000
	NN_ScoreZ_ReLU	Rete neurale	Neural Network	0.5722	1,000
	Forest_entropy200	Forest	Random Forest	0.5643	1,000
	Logistic Regression RPCA	Regressione logistica	Logistic Regression	0.5495	1,000
	Tree_entropy+	Albero decisionale	Decision Tree	0.5290	1,000

Best Model: Gradient Boosting with 75 continuous groupings and 150 trees

THANK YOU

For your attention!