

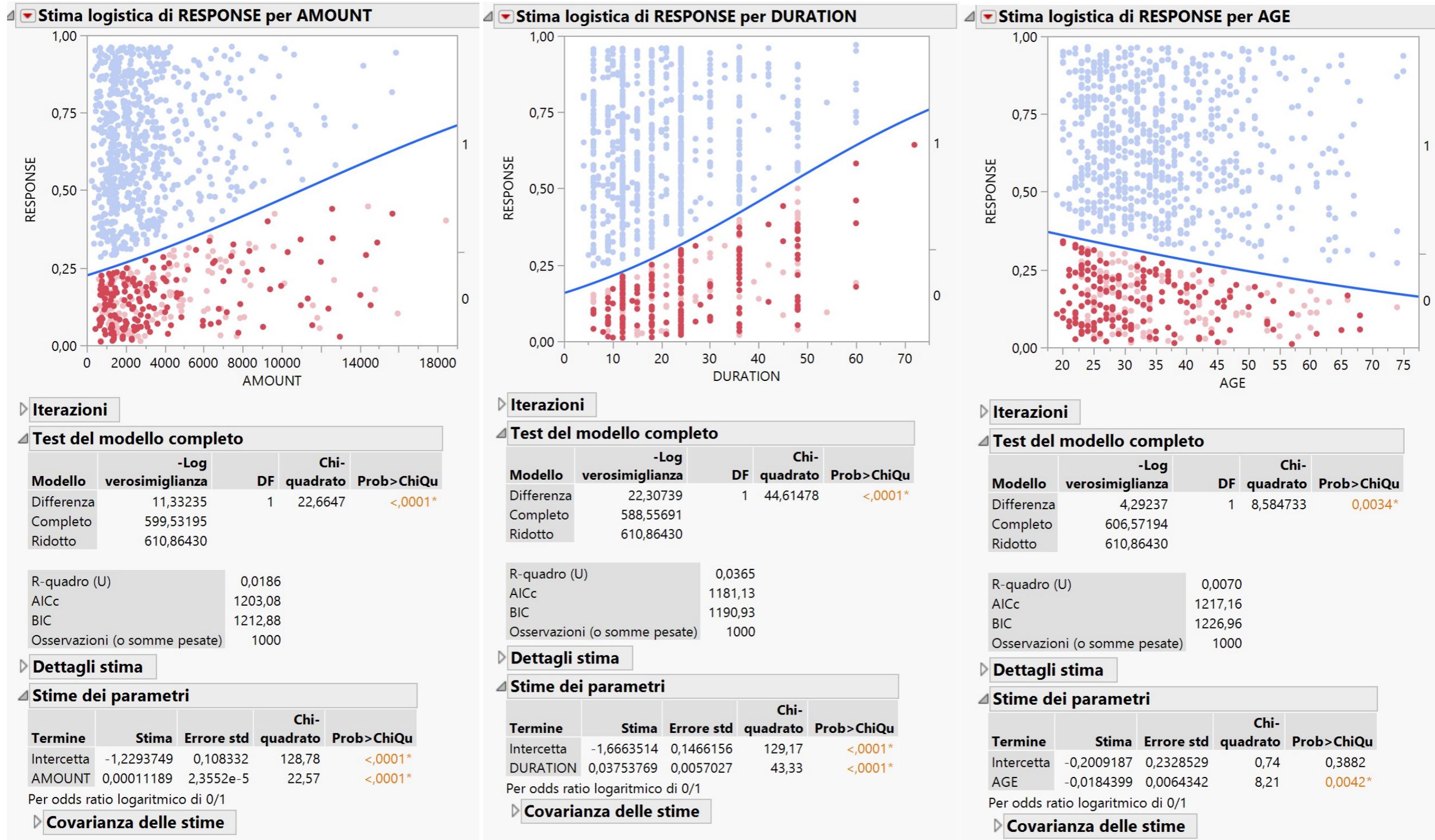
GROUP #4: GOTTA CAMILLA – IERARDI ANDREA –
LAZZARA FRANCESCO – LENI THOMAS – LEPEK
ALEKSANDRA

TASK 3: GERMAN CREDIT ANALYSIS

TASK 3 EXECUTION PLAN

- ▶ Logistic Model Outcomes
- ▶ Bootstrap Model Outcomes
- ▶ Partitioning Model Outcomes
- ▶ Comparison Analysis of Models
- ▶ Summary & Conclusion

LOGISTIC MODEL OUTCOMES



BOOTSTRAP MODEL OUTCOMES

Forest di bootstrap per RESPONSE

Specifiche

| | | | |
|--|-------------|---|-----|
| Colonna target: | RESPONSE | Righe di training: | 742 |
| Colonna di validazione: | Validazione | Righe di validazione: | 258 |
| Numero di alberi nella foresta: | 100 | Righe di test: | 0 |
| Numero di termini campionati per partizione: | 7 | Numero di termini: | 30 |
| | | Campioni bootstrap: | 742 |
| | | Numero minimo di partizioni per albero: | 10 |
| | | Dimensione minima di partizione: | 5 |

Statistiche generali

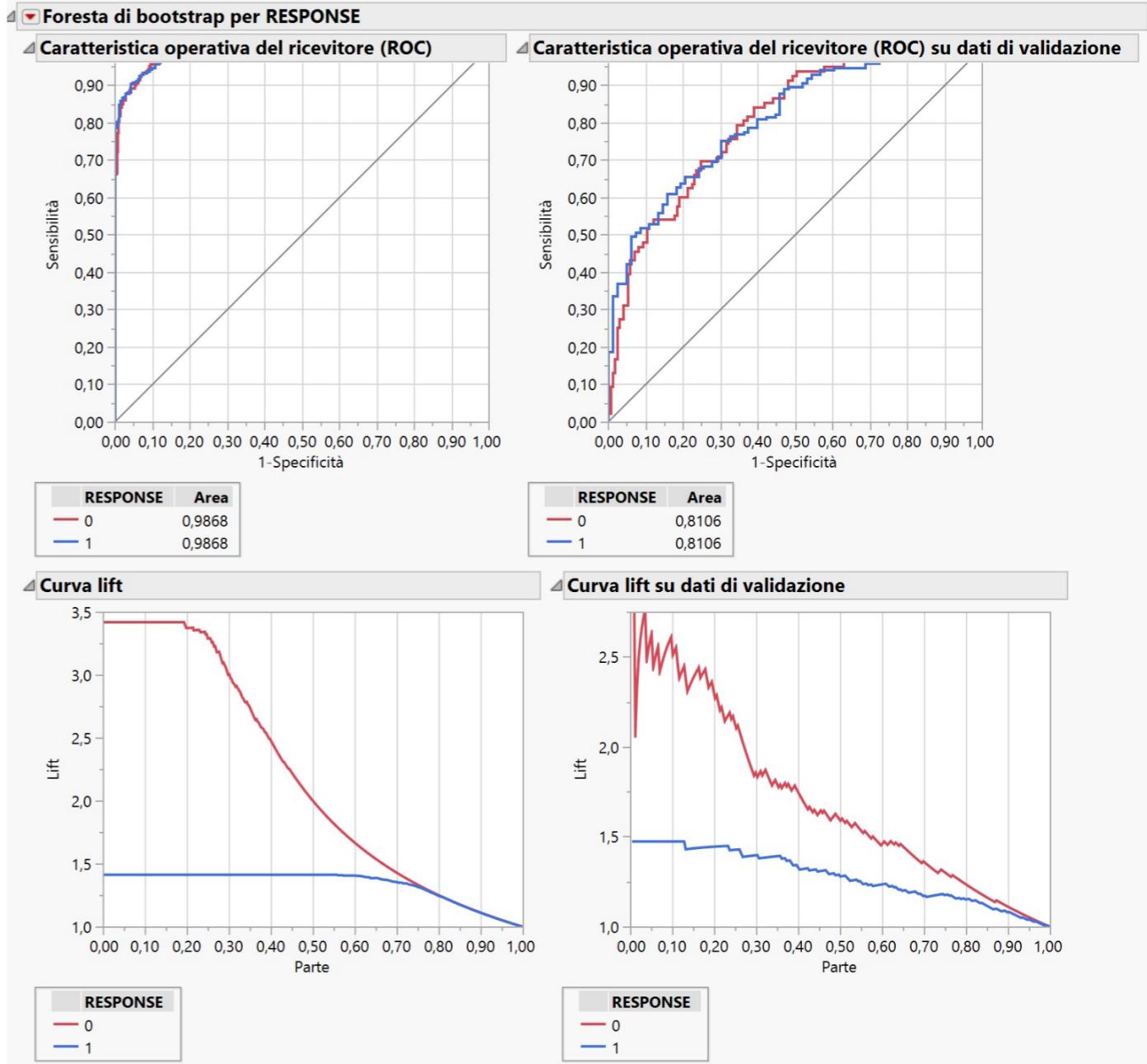
| Misura | Training | Validazione | Definizione |
|---------------------------------|----------|-------------|---|
| R-quadro entropia | 0,4488 | 0,2017 | $1 - \text{Loglike}(\text{model})/\text{Loglike}(0)$ |
| R-quadro generalizzato | 0,5969 | 0,3129 | $(1 - L(0)/L(\text{model}))^{(2/n)} / (1 - L(0)^{(2/n)})$ |
| Media -Log p | 0,3331 | 0,5014 | $\sum -\text{Log}(p[j]) / n$ |
| RMSE | 0,3088 | 0,4092 | $\sqrt{\sum (y[j] - p[j])^2 / n}$ |
| Scarto medio ass | 0,2619 | 0,3504 | $\sum y[j] - p[j] / n$ |
| Tasso errori di classificazione | 0,1024 | 0,2558 | $\sum (p[j] \neq p_{\text{Max}}) / n$ |
| N | 742 | 258 | n |

Matrice di confusione

| | | Training | | Validazione | |
|--|--|-----------|--------------------|-------------|--------------------|
| | | Osservati | Conteggio previsto | Osservati | Conteggio previsto |
| | | RESPONSE | 0 1 | RESPONSE | 0 1 |
| | | 0 | 141 76 | 0 | 24 59 |
| | | 1 | 0 525 | 1 | 7 168 |

Validazione cumulativa

Riepiloghi per albero



PARTITION (DECISION TREE) MODEL OUTCOMES

Dettagli stima

| Misura | Training | Validazione | Definizione |
|---------------------------------|----------|-------------|---|
| R-quadro entropia | 0,2055 | 0,1029 | $1 - \text{Loglike}(\text{model})/\text{Loglike}(0)$ |
| R-quadro generalizzato | 0,3135 | 0,1695 | $(1 - L(0)/L(\text{model}))^{(2/n)} / (1 - L(0)^{(2/n)})$ |
| Media -Log p | 0,4802 | 0,5635 | $\sum -\text{Log}(\rho_{ij})/n$ |
| RMSE | 0,3985 | 0,4381 | $\sqrt{\sum (y_{ij} - \rho_{ij})^2/n}$ |
| Scarto medio ass | 0,3191 | 0,3614 | $\sum y_{ij} - \rho_{ij} /n$ |
| Tasso errori di classificazione | 0,2358 | 0,2984 | $\sum (\rho_{ij} \neq \rho_{\text{Max}})/n$ |
| N | 742 | 258 | n |

Matrice di confusione

| Training | | Validazione | |
|-----------|--------------------|-------------|--------------------|
| Osservati | Conteggio previsto | Osservati | Conteggio previsto |
| RESPONSE | 0 1 | RESPONSE | 0 1 |
| 0 | 84 133 | 0 | 27 56 |
| 1 | 42 483 | 1 | 21 154 |

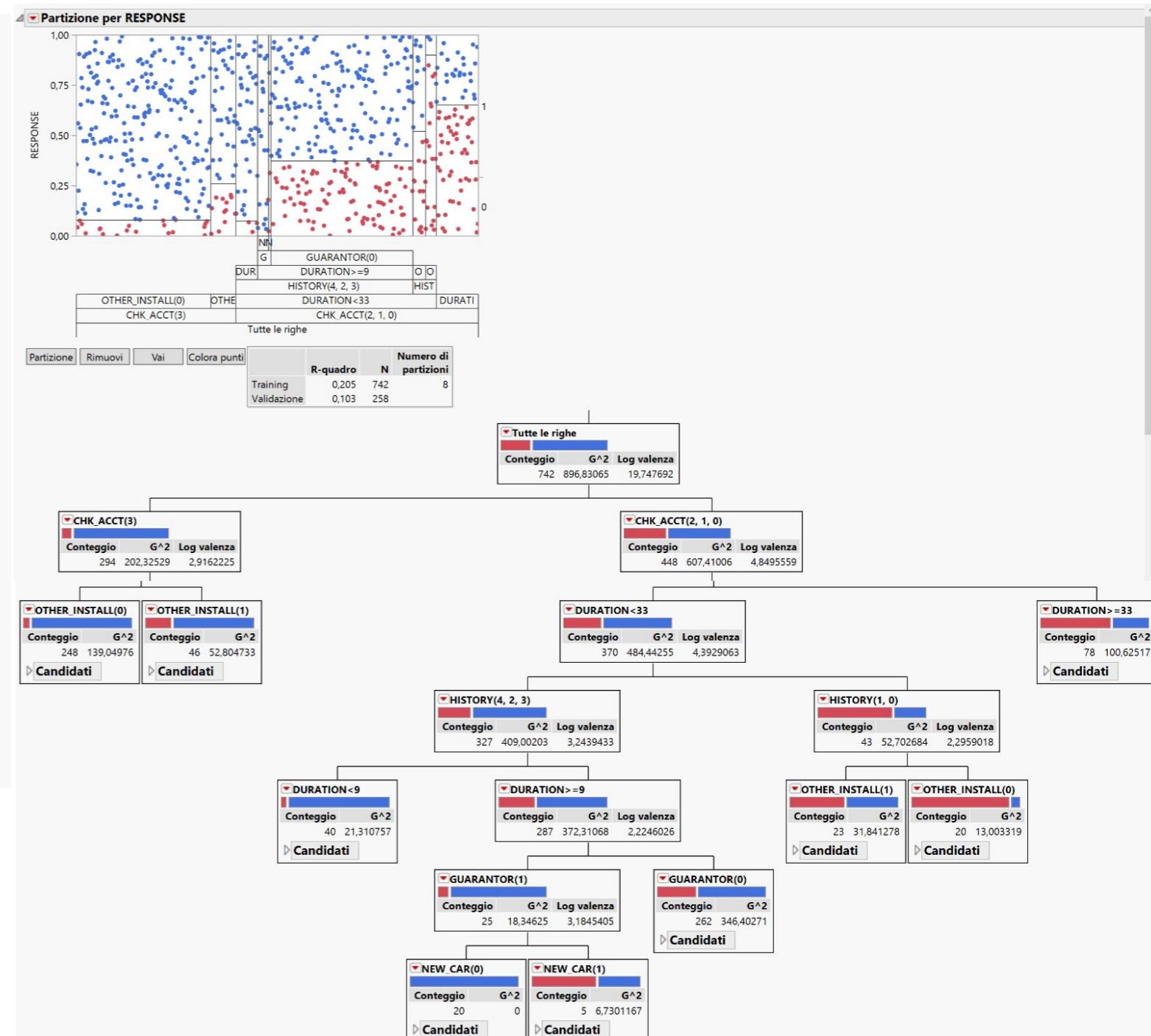
Matrice decisionale

| Training | | Validazione | | Matrice del profitto specificata | |
|-----------|---------------------------|-------------|---------------------------|----------------------------------|-----------|
| Osservati | Conteggio delle decisioni | Osservati | Conteggio delle decisioni | Effettiva | Decisione |
| RESPONSE | 0 1 | RESPONSE | 0 1 | | |
| 0 | 69 148 | 0 | 24 59 | 0 | -0,667 |
| 1 | 29 496 | 1 | 19 156 | 1 | 0 |

| Osservati | Tasso delle decisioni |
|-----------|-----------------------|
| RESPONSE | 0 1 |
| 0 | 0,318 0,682 |
| 1 | 0,055 0,945 |

Tasso errori di classificazione

0,2385 0,3023



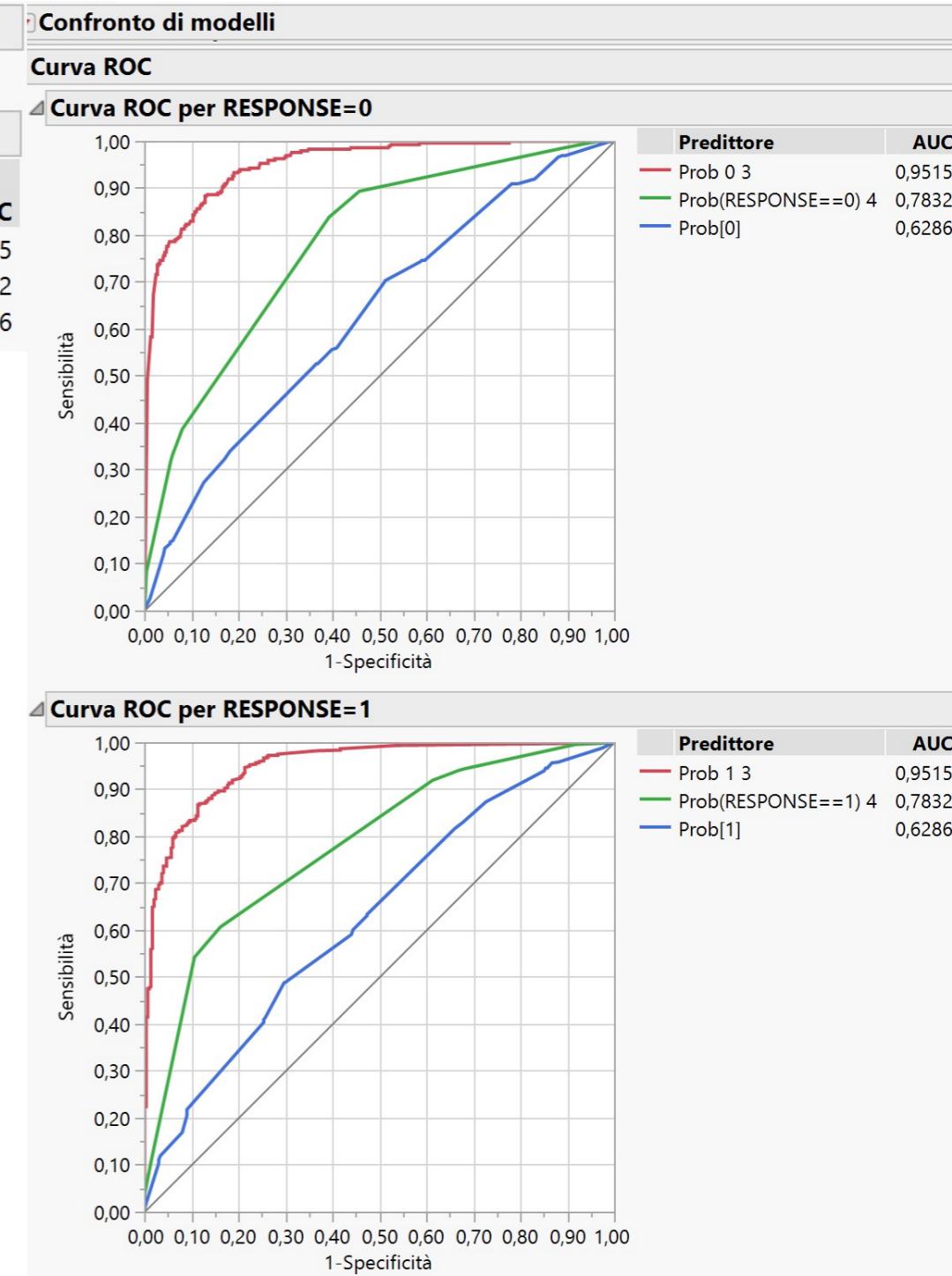
COMPARISON OF THE MODELS

Confronto di modelli

Predittori

Misure di stima per RESPONSE

| Creatore | R-quadro entropia | R-quadro generalizzato | Media - Log p | RMSE | Scarto medio ass | Tasso errori di classificazione | N | AUC |
|----------------------|-------------------|------------------------|---------------|--------|------------------|---------------------------------|------|--------|
| Foresta di bootstrap | ,2 ,4 ,6 ,8 | 0,3836 | 0,5305 | 0,3765 | 0,3376 | 0,2847 | 1000 | 0,9515 |
| Partizione | | 0,2055 | 0,3135 | 0,4802 | 0,3985 | 0,3191 | 742 | 0,7832 |
| Logistico | | 0,0365 | 0,0619 | 0,5886 | 0,4478 | 0,4008 | 1000 | 0,6286 |



SENSITIVITY & SPECIFICITY COMPARISON

| | Sensitivity | Specificity |
|----------------------------------|-------------|-------------|
| Partition | 38,7% | 92% |
| Performance of random forest 0,4 | 35% | 99,6% |
| Performance of random forest 0,6 | 83,9% | 96,4% |
| Performance of random forest 0,8 | 99,5% | 64,2% |
| Random forest | 65% | 100% |

CONCLUSION

WE CHOSE THE THREE MODELS AS: LOGISTIC, BOOTSTRAP FOREST AND PARTITION (DECISION TREE). THEN WE'VE EVALUATED THE PERFORMANCES OF THEM ALSO WITH THE ROC CURVE. IN CONCLUSION, BOOTSTRAP FOREST SEEMS TO FIT THE DATA IN THE BEST WAY AS IN THE ROC GRAPH IT IS THE FAREST LINE FROM THE DIAGONAL.