# Should they stay or should they return?

## Predicting product return analysis through Machine Learning models

**Master in BI & Big Data Analytics (2022/2023)**

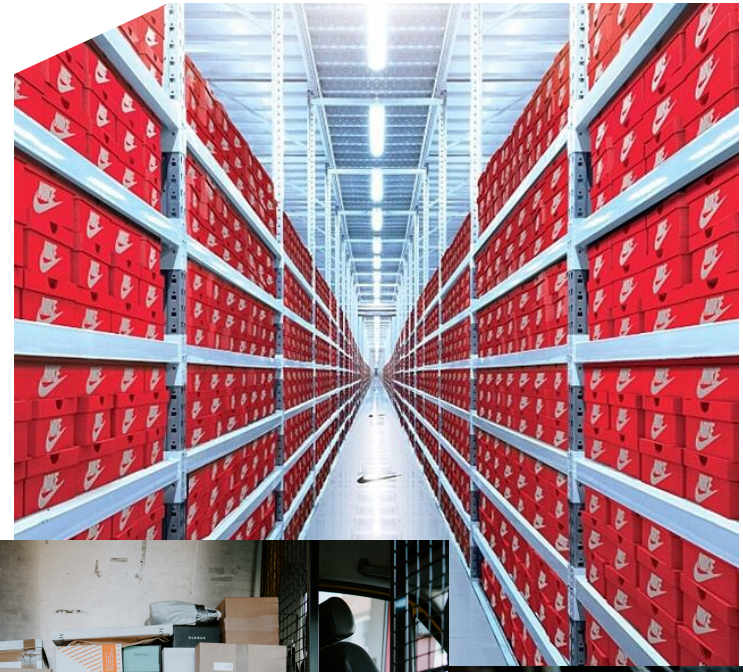
Statistical Modeling & Machine Learning – Team project work

**Team:**
**Samuela Ejelli**
**Stefano Girardini**
**Michele Iannotta**
**Fabio Jr Lorenzini**

Milano, October 2023

# Agenda

1. Business Scenario
2. Objectives
3. Approach overall
4. Exploratory Data Analysis
5. Models Selection and Validation
6. Assessment
7. Final Results & Next steps




# 1.Business Scenario

In the e-commerce era understanding returns volume is crucial for manufacturers and retailers.

Heavy return volume affect firms from several standpoints:

- **OPERATIONAL CHALLENGES**: returns require resources such as staff and space, leading to operational complexities.
- **PRODUCTION WORKFLOW IMPACT**: repairs of returned items can disrupt production workflows, affecting product types and workload levels.
- **FINANCIAL ISSUES**: predicting return volume helps estimate the financial costs and losses incurred.

According to a Gartner research (2015) **less than 50% of what is returned can be resold** at full price.

This is even more critical for online sales since proprietary and wholesale e-commerces growth rate is relentlessly increasing.

Having a reliable **predictive model tailored and tuned** for a specific firm can lead to a **significant money saving**.

# 2.Objectives

Considering a worldwide retail apparel and fashion company, here are the main goals of this project:

1. **Explore** annual (2022) **database** comprehensive of sales and returns (by product, customer, date, country, etc.)

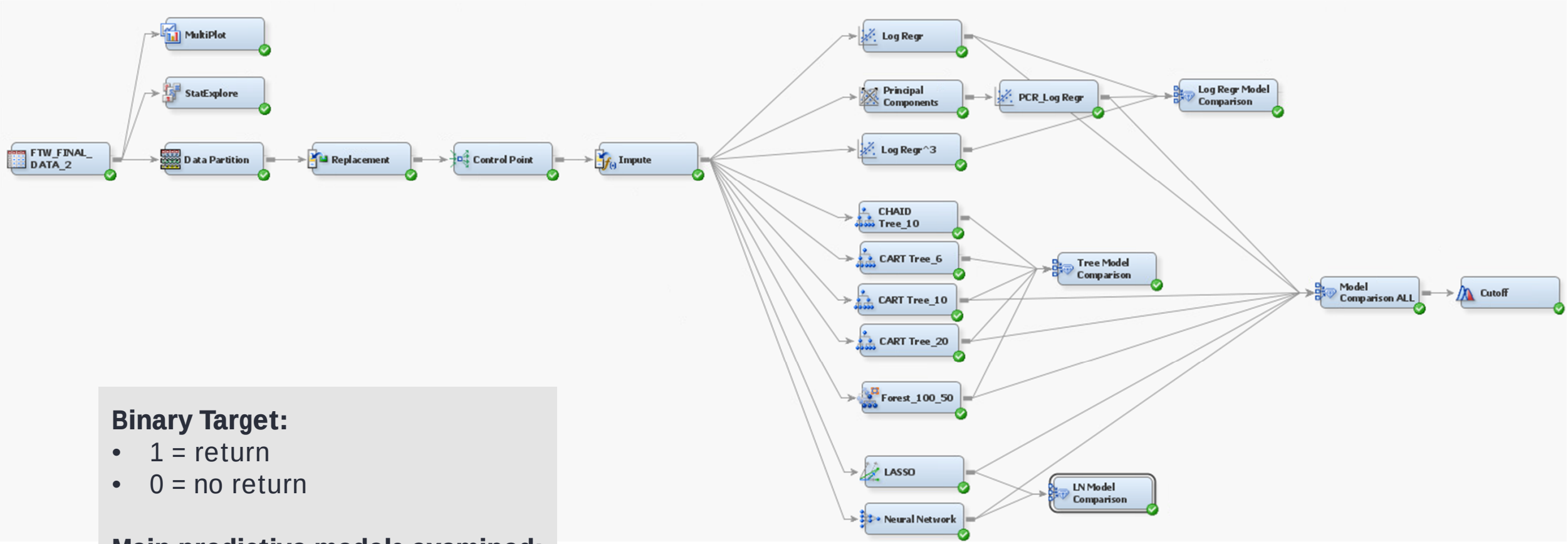
2. Analyze and find out most significant potential **return predictors**

3. Apply **pre-processing techniques** to improve performance

4. Apply and tune **ML models** through an iterative, adaptive approach

5. Declare **best model** to predict product returns

# 3.Approach overall




**Binary Target:**

- 1 = return
- 0 = no return

**Main predictive models examined:**

- Logistic regression
- Decision Trees
- Random Forest
- LASSO
- Neural Network

# 4.Exploratory Data Analysis | Pre-processing

The dataset comprises all orders and their respective returns (if any) recorded in Europe for an e-commerce site of a footwear and apparel brand.
The original dataset has around **7 million records**.

To simplify and **clean the dataset**, we:

- Narrowed our focus to only the '*Footwear*' category.
- **Removed irrelevant columns** for our project.
- **Merged rows** where a sale and its corresponding return occurred.
- **Filtered** out anomalous values and **managed NULLs**.
- Added a **binary target** column (*has_return*).
- Indexed the dataset.

# 4.Exploratory Data Analysis | Variables selection

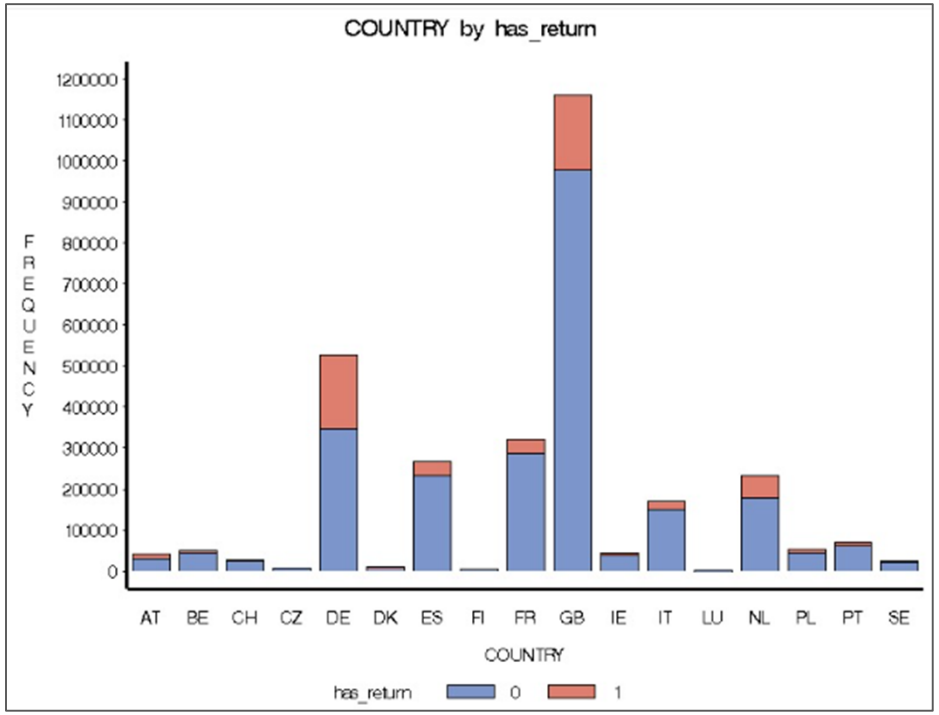
Dataset of 3.015.326 observations with the following variables:

- **Country:** Order/return country
- **MATNR:** Product code
- **SAL_ORDER_DATE:** Order date
- **SAL_QTY:** Product quantity in the order (for simplicity, only considered SAL_QTY = 1, majority of the dataset)
- **SAL_NET_AMOUNT_EUR:** Order value (excluding VAT)
- **RET_NET_AMOUNT_EUR:** Return value (excluding VAT)
- **RET_QTY:** Return quantity
- **Customer_number:** Customer code
- **Indigo:** Delivery method
- **Loyalty_enrollment_flag:** Customer enrolled in a loyalty program (true, false)
- **Size_department:** Item's gender/age type (unisex, men, women, kids)
- **Product_type:** Product category (footwear)
- **Shipping_method:** Shipping mode
- **BSTNK:** Order code
- **Has_return**: target variable
- **Return_description:** Short reason for return (already grouped in clusters, no free text)
- **Payment_type** (e.g. Klarna, CC,…)

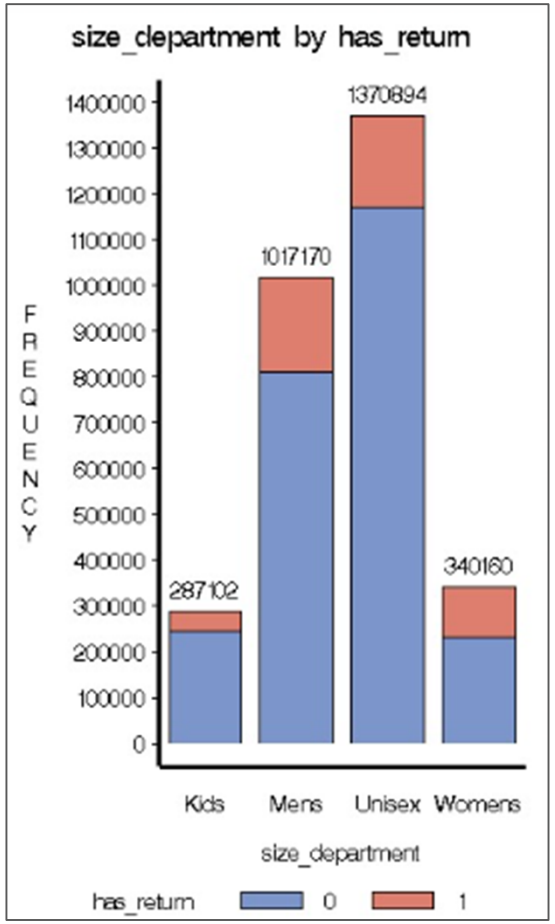
Variables: Input – ID – Rejected

# 4.Exploratory Data Analysis | Descriptive Statistics

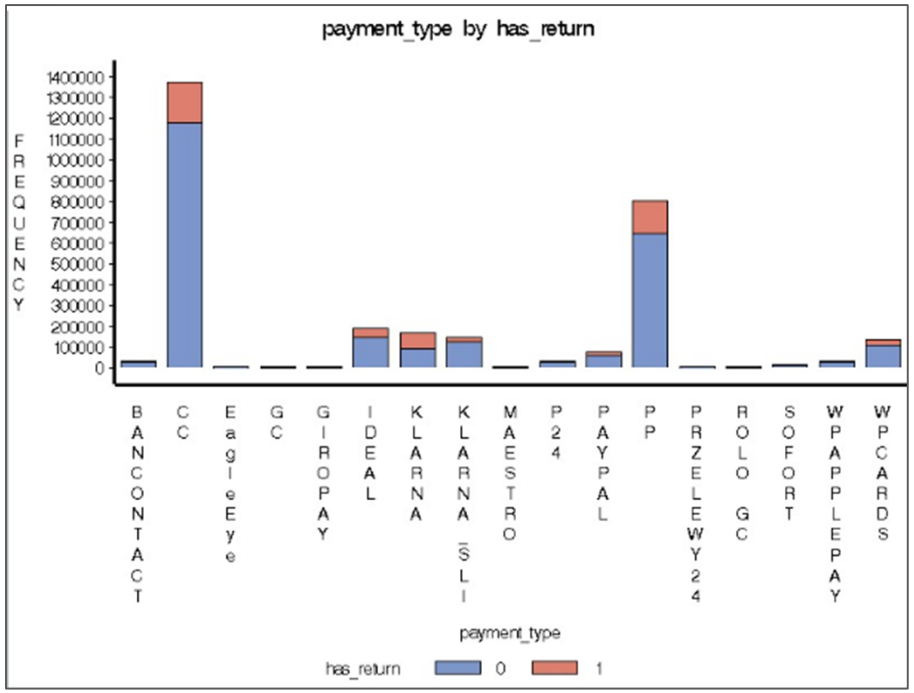
Germany is the country registering the highest number of returns (34%)




Women return more vs other age groups (32%)




Klarna payment method register 46% of returns

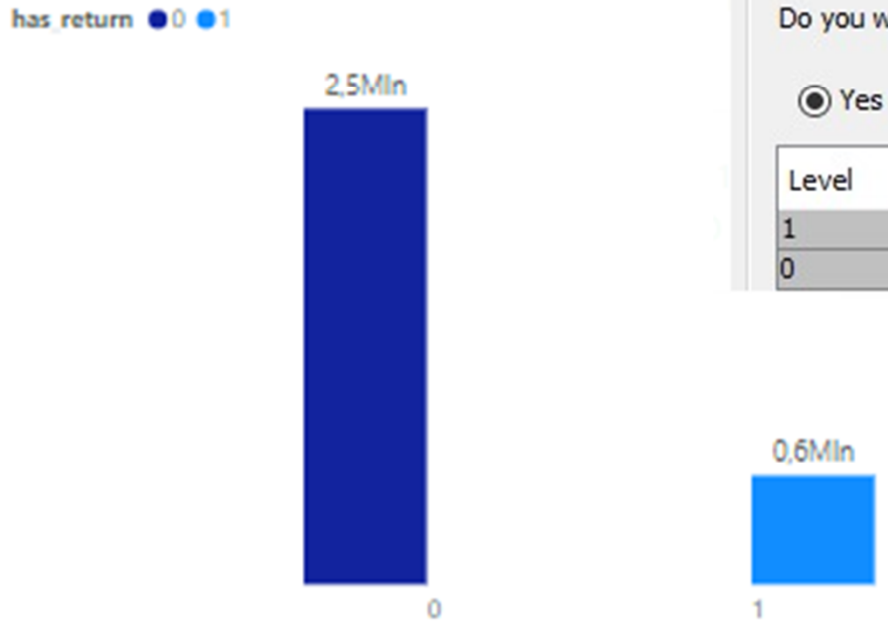



| Input | Chi-Square | Df | Prob |
|---|---|---|---|
| payment_type | 6481.3856 | 10 | <.0001 |
| COUNTRY | 5850.7949 | 16 | <.0001 |
| size_department | 279.4559 | 3 | <.0001 |
| shipping_method | 77.4992 | 5 | <.0001 |
| indigo | 21.4581 | 3 | <.0001 |
| loyalty_enrollment_flag | 1.3252 | 2 | 0.5155 |

# 4.Exploratory Data Analysis | True Priors & Data Partition

The dataset shows 19% returns, very optimistic percentages.

To recalibrate the priors, we assigned a probability of 25%, performing oversampling.




Decision Processing - FTW_FINAL_DATA_2

Targets | Prior Probabilities | Decisions | Decision Weights

Do you want to enter new prior probabilities?

(•) Yes ( ) No  Set Equal Prior

| Level | Count | Prior | Adjusted Prior |
|---|---|---|---|
| 1 | 559292 | 0.1855 | 0.25 |
| 0 | 2456034 | 0.8145 | 0.75 |

Due to high number of observations (3.015.326) an External Validation ("*Holdout method*") was adopted.

The dataset was partitioned into:

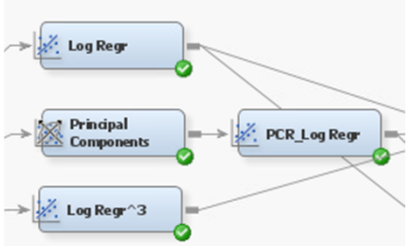
- **Training = 60%**
- **Validation = 30%**
- **Test = 10%**

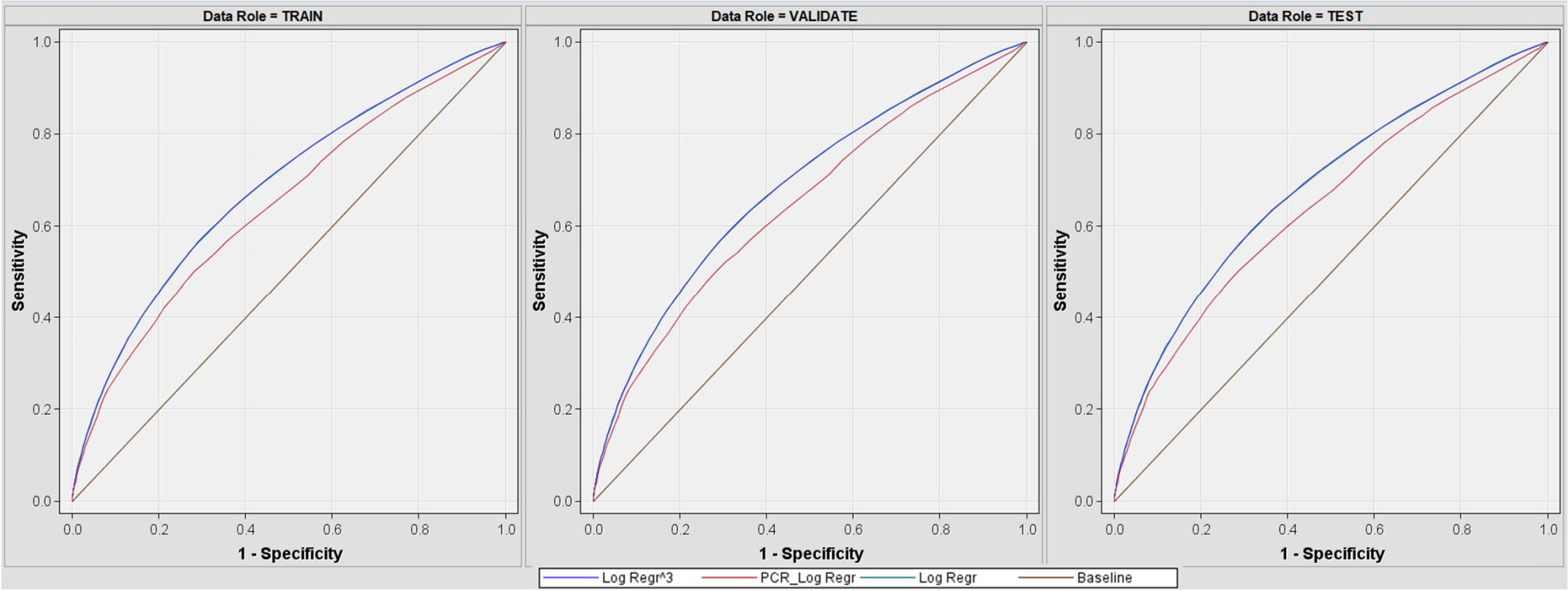
# 5.Model Selection and Validation | Logistic regression

- We did several logistic regressions: the one with PCR brought the worst result in the ROC chart, while the rest performed almost identically. The third-degree regression obtained a slightly lower ASE.

- Even by firmly pursuing the addition of complexity in our models, Overfitting was not obtained (at least for the sample selected by SAS).

| Model Description | Train: Average Squared Error | Valid: Average Squared Error | Test: Average Squared Error |
|---|---|---|---|
| PCR_Log Regr | 0.143745 | 0.14367 | 0.143937 |
| Log Regr^3 | 0.139588 | 0.139499 | 0.139723 |
| Log Regr | 0.139664 | 0.139554 | 0.139801 |

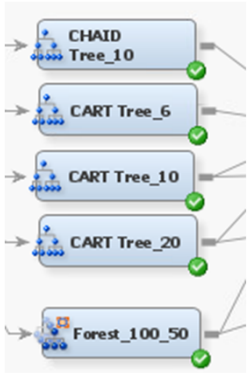


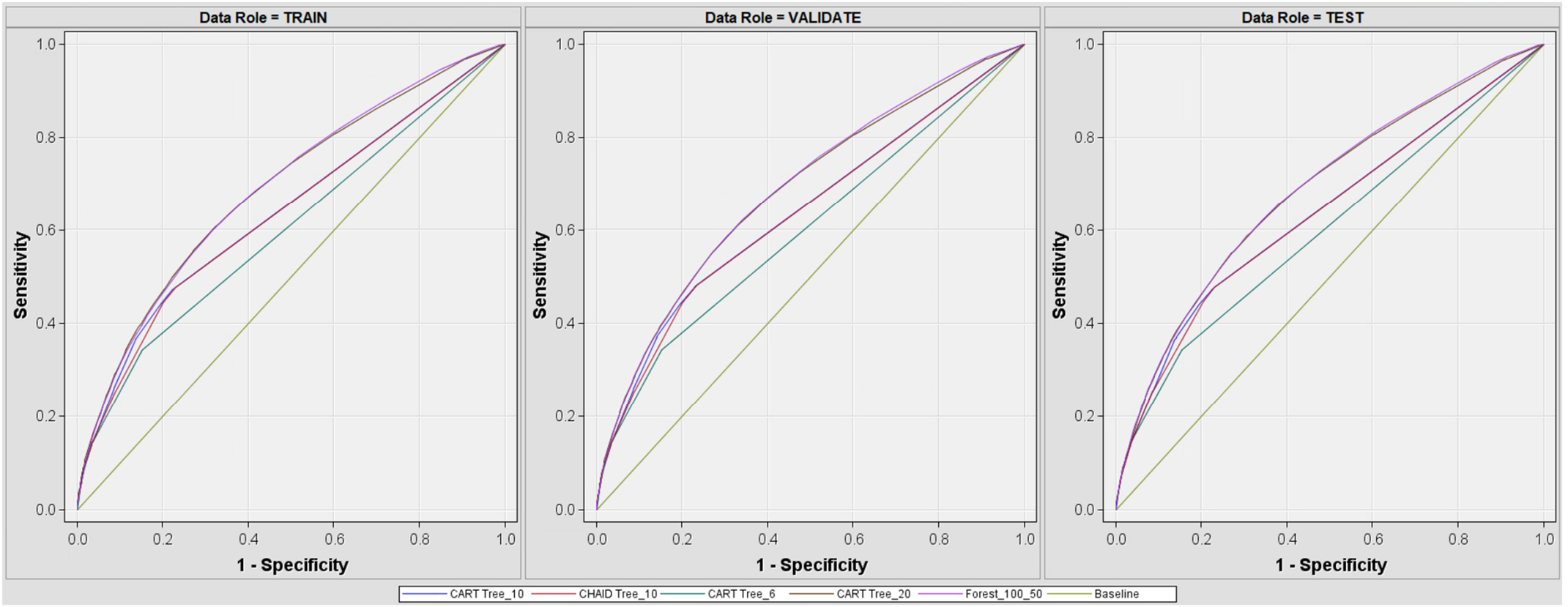



# 5.Model Selection and Validation | Decision Trees & Random Forest

- The decision trees presented a significant variance of results. We tried both trees with a "small" number of nodes (10), and a Random Forest that resulted in being the best model in the ROC chart.
- Each leaf could contain at least 5 observations, and we started with a minimum of 10 leaves. In another model, not reported, with a minimum of 60 we obtained slightly better results in ASE measure.

| Model Description | Train: Average Squared Error | Valid: Average Squared Error | Test: Average Squared Error |
|---|---|---|---|
| CART Tree 20 | 0.138397 | 0.138926 | 0.139275 |
| CART Tree 10 | 0.141289 | 0.141327 | 0.141505 |
| CHAID Tree 10 | 0.141868 | 0.141846 | 0.141973 |
| CART Tree 6 | 0.143914 | 0.143929 | 0.144099 |
| Forest 100 50 | 0.138962 | 0.139043 | 0.139225 |

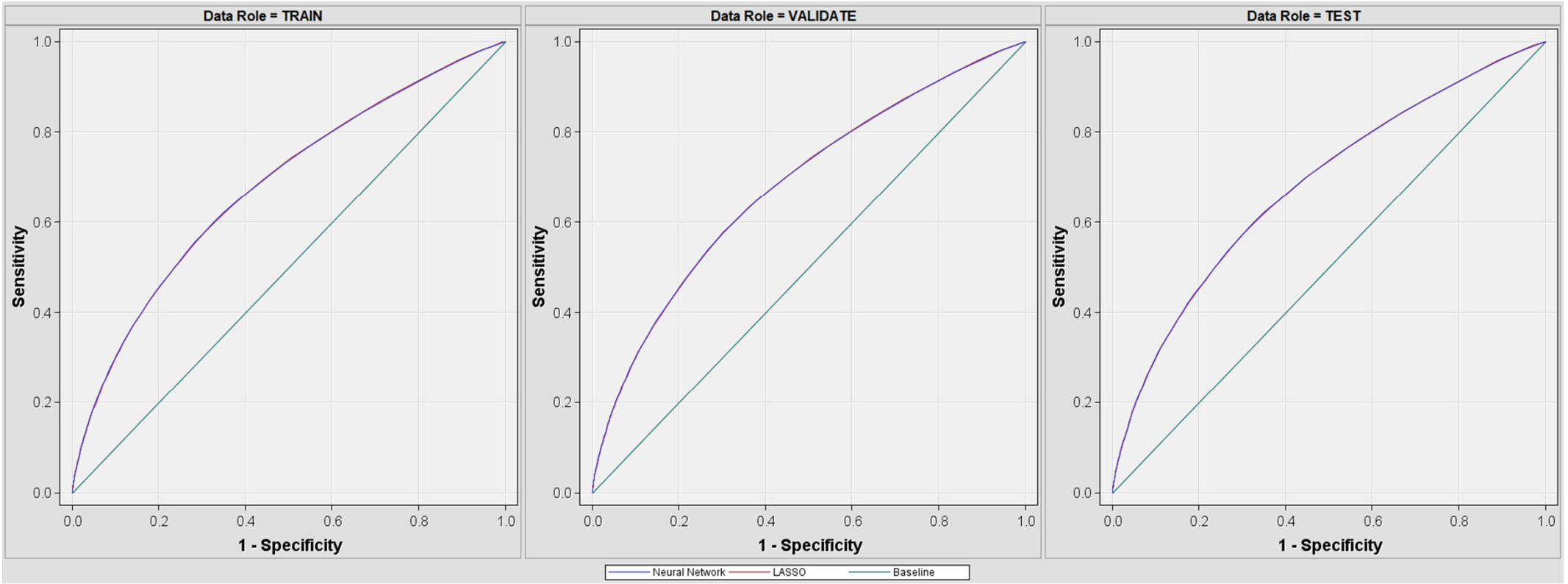


# 5.Model Selection and Validation | LASSO & Neural Network

- LASSO regression and Neural Network performed similarly.
  As the neural network setting the Multilayer Perceptor was set.
  Both did not outperform the Random Forest model in ROC chart with all models.

| Model Description | Train: Average Squared Error | Valid: Average Squared Error | Test: Average Squared Error |
|---|---|---|---|
| Neural Network | 0.139638 | 0.139539 | 0.139749 |
| LASSO | 0.139911 | 0.139793 | 0.140037 |

# 6.Assessment | ASE Validation

For each model here is reported the Average Squared Error for Train/Validation/Test partition and the related ratio between Validation vs Train and Test vs Train.

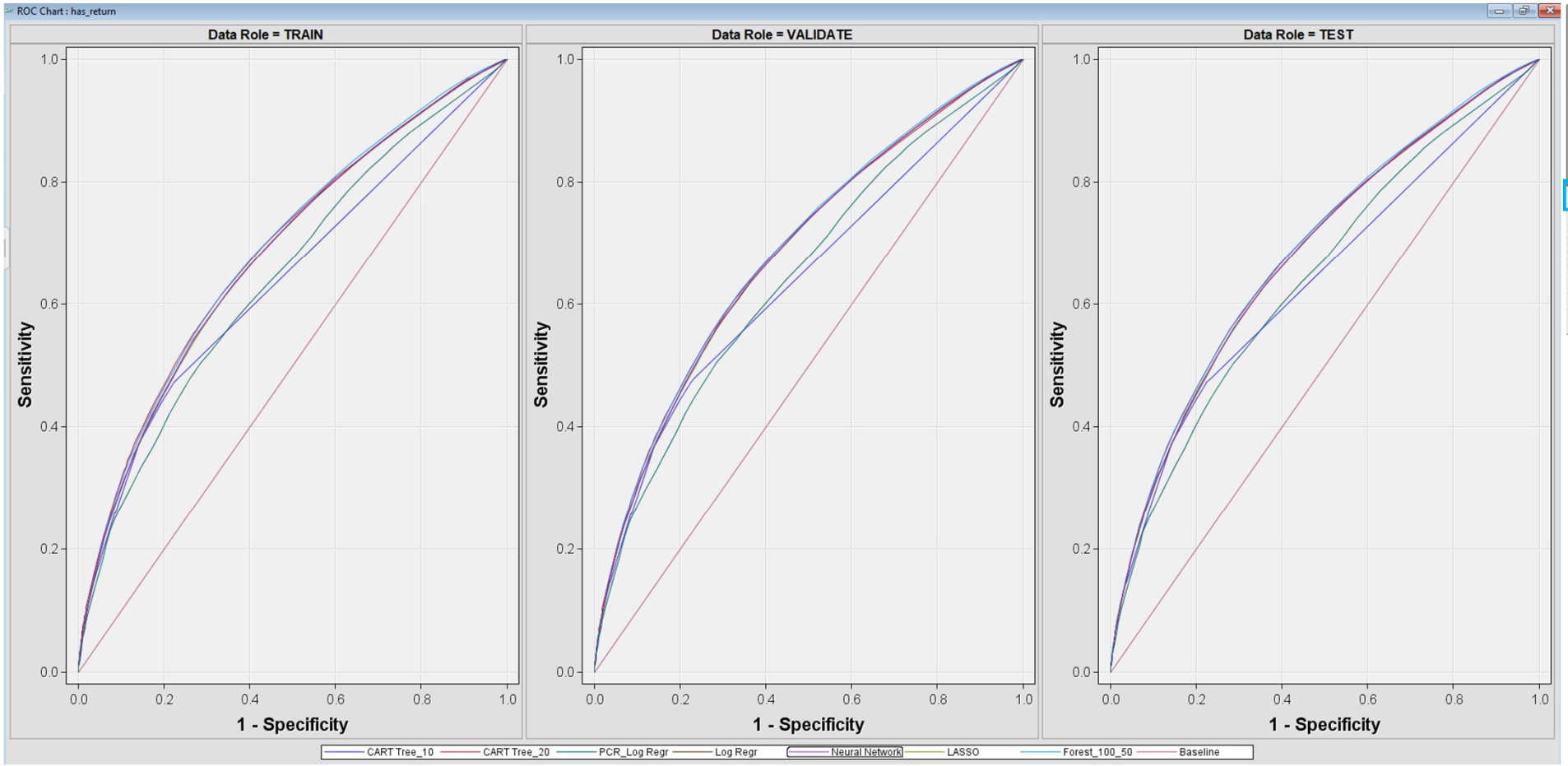
ASE represents only a quality index (goodness of fit), but not a classification method!

| | | ASE | | | (ASEval-ASEtrain)/ASEtrain | (ASEtest-ASEtrain)/ASEtrain |
|---|---|---|---|---|---|---|
| **MODEL** | **DETAIL** | **TRAIN** | **VALIDATION** | **TEST** | | |
| Logistic | | 0,139664 | 0,139554 | 0,139801 | -0,08% | 0,10% |
| Logistic gr3 | | 0,139588 | 0,139499 | 0,139723 | -0,06% | 0,10% |
| PCR + Logistic | | 0,143745 | 0,14367 | 0,143937 | -0,05% | 0,13% |
| LASSO | | 0,139911 | 0,139793 | 0,140037 | -0,08% | 0,09% |
| Dec Tree 1 | CART 6 | 0,143914 | 0,143929 | 0,144099 | 0,01% | 0,13% |
| Dec Tree 2 | CART 10 | 0,141289 | 0,141237 | 0,141505 | -0,04% | 0,15% |
| Dec Tree 3 | CART 20 | 0,138397 | 0,138296 | 0,139275 | -0,07% | 0,63% |
| Dec Tree 4 | CHAID 10 | 0,141868 | 0,141846 | 0,141973 | -0,02% | 0,07% |
| Random Forest | 100_50 | 0,138962 | 0,139043 | 0,139225 | 0,06% | 0,19% |
| Neural Network | 3 hidden layers \| Model Selection Criteria: Avg Error | 0,139638 | 0,139539 | 0,139749 | -0,07% | 0,08% |

# 6.Assessment | ROC curve

- ROC-AUC curve allows to understand the probability of correct classification of true positive events against false positive events for each threshold.
  According to ROC Index values most of the models seems quite similar.

- Looking at ROC-AUC diagrams **Random Forest** results the **best model** for each dataset (train/validate/test). In terms of overall performance it can be considered a 'fair' prediction model.

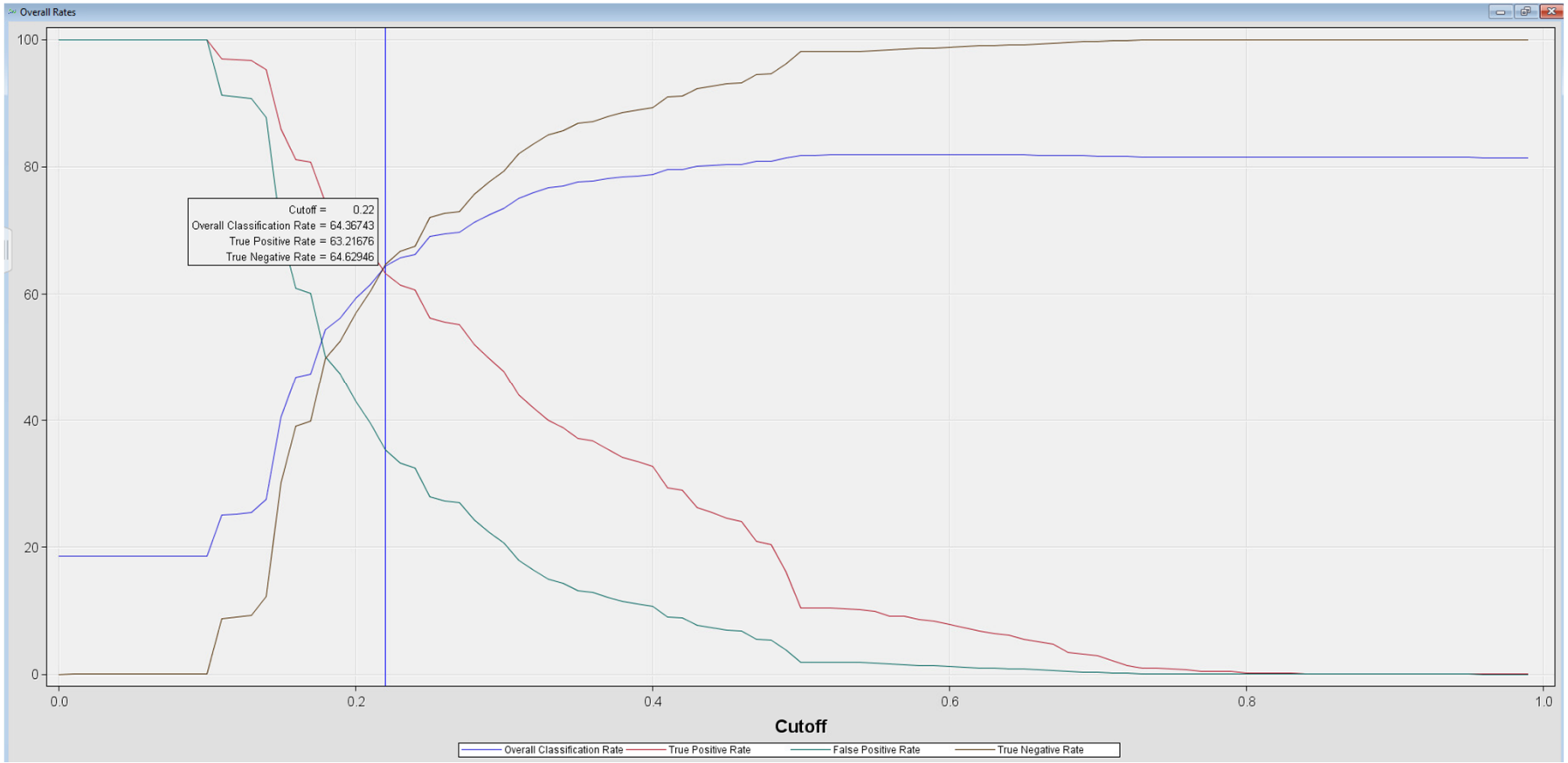



| Model Description | Train: Roc Index | Valid: Roc Index | Test: Roc Index |
|---|---|---|---|
| Forest 100 50 | 0.687 | 0.686 | 0.685 |
| LASSO | 0.68 | 0.681 | 0.68 |
| CART Tree 20 | 0.686 | 0.684 | 0.683 |
| CART Tree 10 | 0.637 | 0.637 | 0.636 |
| PCR Log Regr | 0.643 | 0.644 | 0.641 |
| Log Regr | 0.681 | 0.682 | 0.68 |
| Neural Network | 0.68 | 0.681 | 0.68 |

# 6.Assessment | Cutoff

Since for this case study the objective is to identify most of the 'true' positive events (i.e. has_return = 1), the Cutoff needs to focus on high level of True Positive Rate (TPR) (i.e. high Sensitivity), a good level of True Negative Rate (TNR) and the lowest False Positive Rate (FPR) possible.

→ **Cutoff = 0,22** allows to reach a **good trade-off** among TPR, FPR, TNR.




CUTOFF = 0,22

- True positive rate ~ 63,21
- False positive rate ~ 35,37
- True negative rate ~ 64,62

- Accuracy ~ 81,3%
- Precision ~ 49,2%

# 7.Final results & Next Steps

In the end, after the analysis of all models, the **Random Forest** resulted as the best model (Roc index ~ 0,687) since the area under the curve was the greatest.

The model **performance could be improved** by modifying the **dataset** and intervening in the **pre-processing phase**, through:

- Addition of potentially determinant variables to improve the model:
  - Product characteristics (color, size, season)
  - **Time variables** (e.g., day of the week, month)
  - **Demographic characteristics** of customers (if possible)
  - Consider other merchandise categories (perhaps on separate models)
  - Any **promotions** (which we excluded from the dataset)
  - Logistical costs related to returns (transportation and warehouse management) that we initially overlooked

- Implementation of the **cost and profit matrix** (feedback from with the finance department)

# THANKS

**Team:**
**Samuela Ejelli | *samuela.ejelli@icloud.com***
**Stefano Girardini | *stefanogirardini94@gmail.com***
**Michele Iannotta | *micheleiannotta273@gmail.com***
**Fabio Jr Lorenzini | *fabiojr.lorenzini@gmail.com* , *https://github.com/FabioJrLorenzini***