NOVA IMS

Information Management School

A Statistical and Machine Learning Approach for Assessing the Impact of Financial Research Reports on Clients' Trading Behavior

Diogo Tomás dos Santos Peixoto

Dissertation

Advisor: Professor Mauro Castelli

Date: 02nd February 2024

Instituto Superior de Estatística e Gestão da Informação
Universidade Nova de Lisboa

Introduction

Introduction

Who is BNP Paribas:

BNP Paribas is a market maker that buys and sells securities on its account.

How market makers make money:

The price difference between buying and selling a security to a client.

How clients invest:

Send a price request (RFQ) to ***n*** *market makers,* who reply with a price quote. After, they decide with whom to trade.

BNP Goal:

Having more clients request an (RFQ), and subsequently perform a trade with them. They believe research reports can help with this.

BNP PARIBAS

The bank for a changing world

CREDIT 360™
STRATEGY & ANALYTICS

DESK ANALYST
CREDIT TRADE IDEA EUROPE

Credit Trading Desk Analyst Commentary

Report Title: Twinkle Twinkle Falling Angel

[redacted]

Trade Ideas:

Sell HITTFP € 29, Buy A TIM € 30
Sell WIZZLN € 24, Buy ENAPHO € 24
Buy WNTRDE € 28 hybrids, Sell REPSM € 28 hybrids

[redacted]

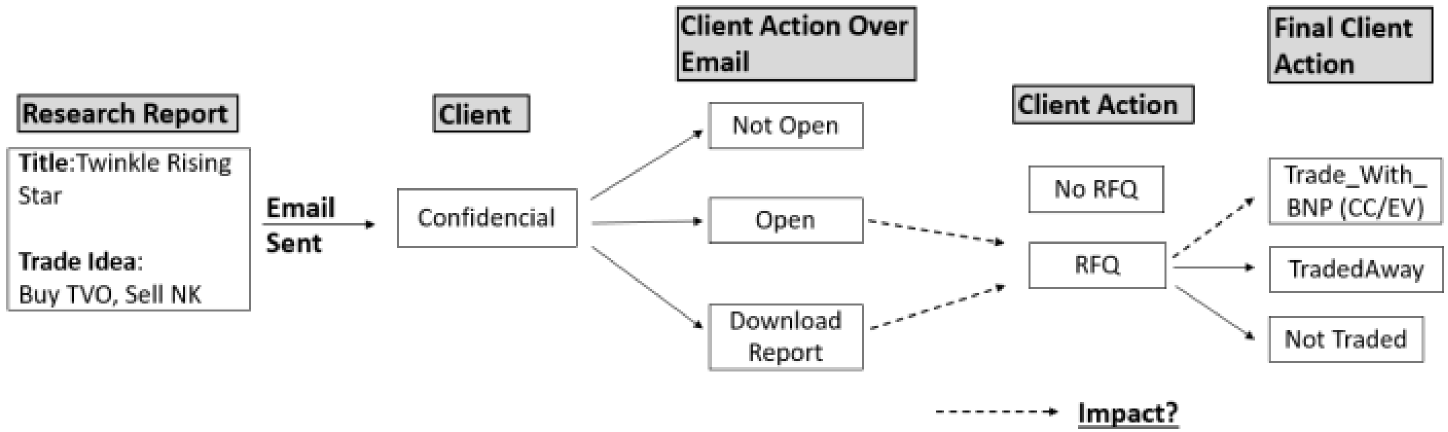
CREDIT 360™

Please refer to important information at the end of the commentary and MAR disclosure

Research report example

Research Problem

What is the impact that BNP Paribas research reports with <u>bond</u> trade suggestions have on clients' trade behaviour?



Business model studied flowchart

Literature Review

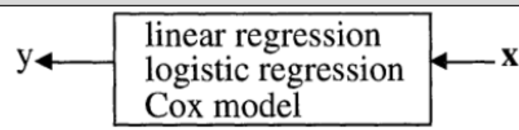
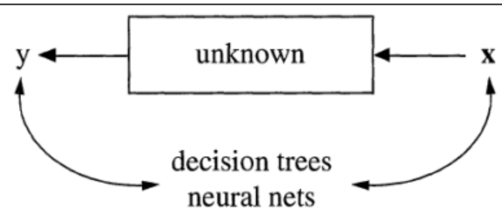
Literature Review

Data analyses can have **two** goals:

- **Explanation:**
    - Understand the relationship between the input and the response variables.
    - Usually taught in the economics and statistics field.
- **Prediction:**
    - Forecast responses to forthcoming input variables.
    - Usually taught in machine learning and data mining fields.

The research problem is an **<u>explanation task</u>**.

Historically, different models were used mainly for different purposes:

Cultures	Models	Model Evaluation
The data modelling culture		Goodness-of-fit tests and residual examination
The algorithm modelling culture		Measured by predictive accuracy

The two cultures to analyze data as defined by Breiman, L. (2001)

NOVA IMS Information Management School

Literature Review

The article by Shmueli and Koppius (2011) points out the following about linear regression models:

"…although it can be used for building an explanatory statistical model as well as a predictive model, the two resulting models will differ in many ways. The differences are (…)from the data used to estimate the model (e.g., variables included and excluded, form of the variables, treatment of missing data), to how performance is assessed (model validation and evaluation) ….".

Even for explanation tasks, the predictive analytics framework can be useful:

- A good predictive power is a good reason for accepting the explanation.
- Compare different models.

Which algorithms fit better under the explanatory task paradigm?:

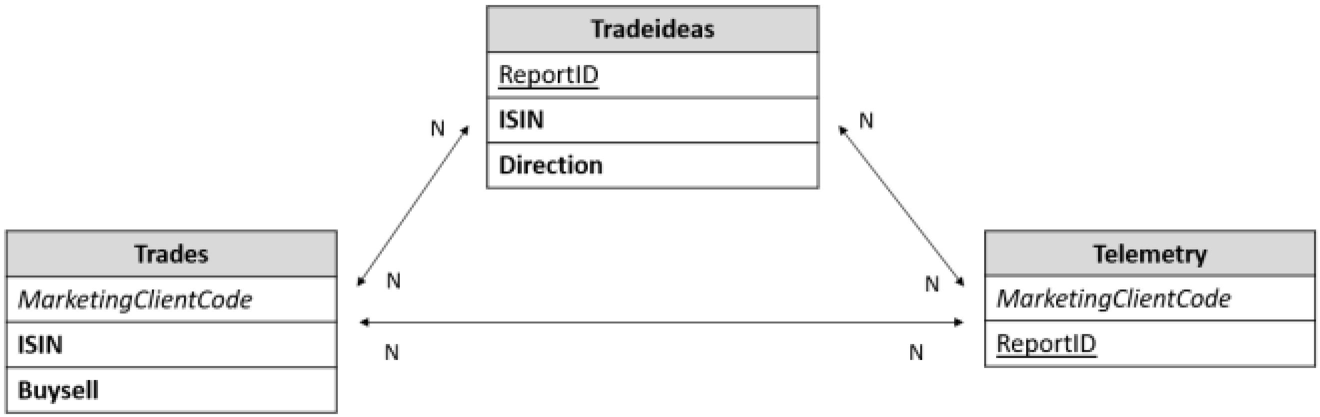
- Traditional machine learning algorithms like neural networks do not explain the results in a way humans can understand. They are more suitable for predictive tasks.
- **Regression models,** also called white boxes**,** provide results more explainable and transparent.
- Explainable machine learning methods are not reliable. They explain how the model works not the world.

Conclusion: Based on the nature of the research problem at hand (explanation task) and the literature review, the problem has been initially addressed with a **logistic regression** model with an explanatory statistical approach.

Data and Models

Data Assembly

- **Trades** – dataset with all the RFQs per client and bond security.
- **Tradeideas** – dataset with BNP research teams' advice to buy or sell bond securities.
- **Telemetry** – dataset with clients' records opening emails and downloading the research reports.



Data Sources and field identification to establish the data linkage

Data Assembly

TRADE_ID	Client Name	Trade Date	Bond Security Name	TELEMETRY_ID	Report Title	Send	Opened	Downloaded	RFQ
1004	Confidential	2022-02-28	SPMIM 3 3/4 09/08/23	2500	Saipem - Where there's a well there's a way	1	1	0	**YES**
3050	Confidential	2022-02-28	ROLLS 4 5/8 02/16/26	3156	Credit Strategy RV - Closing Long Cash/CDS basis on € ROLLS 4.625% 11/25-2/26	1	1	0	**YES**
	Confidential			719560	Colombia rates: 1y1y IBR payer	1	0	0	**NO**
	Confidential			721452	Chinese property pre-sales update 12 Nov 2021	1	1	0	**NO**

Data linkage output

Descriptive Statistics

Dependent Variable	Number of Observations	Total Number of Observations
RFQ=0	387746	421218
RFQ=1	33472	

Number of Observations

Dependent Variable	% Distribution
RFQ=0	92,05%
RFQ=1	7,95%

Dataset class distribution

Variable	Total Number of Distinct Values
Client	859
Report	606
ISIN*[1]	793
Ticker*[1]	320

Main variables distinct count

Models Comparison

Measurements	Standard Statistics (Traditional Regression)	Machine Learning
Goal?	Generally better for inferences about the relationship between variables and their significance	Generally better for predictions
Scientific Question?	How/why it happens?	What will happen?
Type of Data	Linear Data	Linear or non-linear
Training-Test Datasets?	No	Yes
Model Building Constraints	Must adhere to the threoretical model (e.g. in terms of forms, variables,specification).	Less constraints
Model evaluation	Explanatory power is measured by strength-of-fit measures	Predictive power is measured by accuracy of out-of-sample predictions

Comparison of standard statistic vs. machine learning approach

Standard Statistics
Logistic Regression

1. Feature Selection
2. Selection Model Criteria
3. Model Results

1. Feature Selection

Goal:

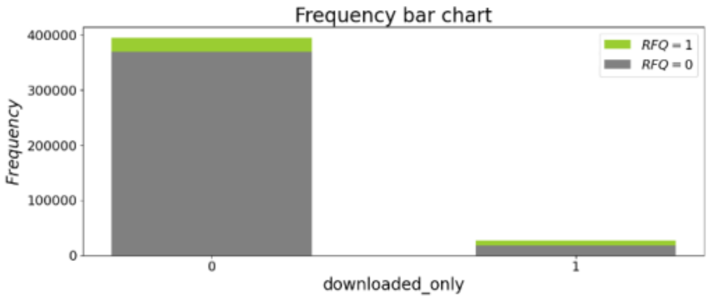
- Find a few meaningful features based on domain knowledge that make the mode interpretable and explain the data.
- Select independent variables that are reasonably associated with the response variable.

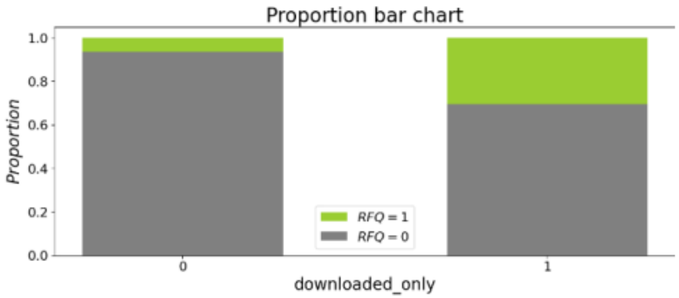
1. Chi-squared test: checks whether two categorical variables are independent.

$$H_0\text{: } \textit{the variables are independent} \quad vs \quad H1\text{: } \textit{the variables are dependent}$$

2. Built stacked bar charts using the contingency table frequencies – visualization aid.

Variable - *Downloaded_only*





3. Cramer's V test - quantify the association between variables

Cramer's V	Correlation Effect Interpretation
> 0.35	Large effect
0.21 < 0.35	Medium effect
<0.21	Small effect

Cramer's V interpretation

1. Feature Selection

Category	Profile	Variables Name	Meaning	Type
Independent Variables	Client Identification	T10_only	The top ten bank clients worldwide, which are also called Titanium 10.	Binary
		C100_only	The top ninety clients worldwide, after the Titanium 10.	
		Customer_Sector	The sector in the market the client belongs to.	Categorical Nominal
	Client Telemetry	Opened_only	The client opened the email sent with a report attached but had not downloaded it.	Binary
		Downloaded_only	The client downloaded the report attached to the email.	
	Client Historical Behaviour	Purchased_ISIN _Before_Only	Clients have bought that ISIN before the first time it was ever suggested in a report.	
		Purchased_Ticker _Before_Only	Clients have bought that Ticker before the first time it was ever suggested in a report but have not bought that ISIN before, meaning the variable "Purchased_ISIN_Before _Only" would be null.	
		Dowloaded _interaction	The ratio between the number of reports downloaded against the total reports received.	Categorical Ordinal
Dependent Variable	-	RFQ	Request for Quotation	Binary

Final features selected

NOVA IMS Information Management School

2. Selection Model Criteria

Akaike information criteria (AIC) – a statistical measure commonly used to compare models with different numbers of parameters.

Formula – AIC

$$AIC = -2ln(L_m) + 2q \quad (3.7)$$

Notation

- L_m is the value of the likelihood model fitted.
- q is the number of independent variables.

Model	Description	AIC
A	All IV(s)	167028
B	All IV(s) - *opened_only*	167162
C	All IV(s) - *purchased_ticker_before_only*	172090
D	*downloaded_only* + *purchased_isin_before_only*	176516
E	All IV(s) - downloaded_only + purchased_isin_before_only	210428

Notes: ▪ The mathematical signs (-) and (+) represent a variable excluded or included in the model, respectively.

3. Model Results

- Pseudo r-squared=0.2855. The model fits well with the dataset.
- Variables with the highest association with the outcome variable are download and purchased_isin_before.
- All the variables are statistically significant apart from the dummy "downloaded_interaction_medium".
- Marginal effects allow interpretation of the independent features on the probability scale.
- Downloading a report is associated with a 12% higher likelihood of requesting an RFQ.

Logit Marginal Effects

Dep. Variable:	rfq
Method:	dydx
At:	overall

	dy/dx	std err	z	P>|z|	[0.025	0.975]
t10_only	0.0583	0.002	28.130	0.000	0.054	0.062
c100_only	0.0212	0.001	25.429	0.000	0.020	0.023
opened_only	0.0099	0.001	11.507	0.000	0.008	0.012
downloaded_only	0.1200	0.002	54.779	0.000	0.116	0.124
purchased_isin_before_only	0.3618	0.003	126.715	0.000	0.356	0.367
purchased_ticker_before_only	0.0887	0.001	70.991	0.000	0.086	0.091
downloaded_interaction_small	-0.0087	0.001	-6.255	0.000	-0.011	-0.006
downloaded_interaction_medium	0.0007	0.001	0.503	0.615	-0.002	0.003
downloaded_interaction_high	0.0112	0.001	8.151	0.000	0.008	0.014
customersector_AM / INS / PENSION	0.0730	0.004	17.771	0.000	0.065	0.081
customersector_BANK / DEALER	0.0992	0.005	18.368	0.000	0.089	0.110
customersector_HEDGE FUNDS	0.0693	0.006	11.459	0.000	0.057	0.081
customersector_PRIVATE BANK / WM	0.1454	0.009	16.846	0.000	0.128	0.162

Logistic regression model average marginal effects

Logit Regression Results

Dep. Variable:	rfq	No. Observations:	421218
Model:	Logit	Df Residuals:	421204
Method:	MLE	Df Model:	13
Date:	Tue, 26 Sep 2023	Pseudo R-squ.:	0.2855
Time:	18:14:59	Log-Likelihood:	-83500.
converged:	True	LL-Null:	-1.1687e+05
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-6.0811	0.072	-83.925	0.000	-6.223	-5.939
t10_only	0.8421	0.025	33.462	0.000	0.793	0.891
c100_only	0.3649	0.014	25.972	0.000	0.337	0.392
opened_only	0.1759	0.015	11.663	0.000	0.146	0.205
downloaded_only	1.5033	0.021	71.339	0.000	1.462	1.545
purchased_isin_before_only	3.4470	0.024	146.428	0.000	3.401	3.493
purchased_ticker_before_only	1.5492	0.024	63.882	0.000	1.502	1.597
downloaded_interaction_small	-0.1572	0.026	-6.080	0.000	-0.208	-0.107
downloaded_interaction_medium	0.0104	0.024	0.434	0.664	-0.037	0.057
downloaded_interaction_high	0.1955	0.024	8.265	0.000	0.149	0.242
customersector_AM / INS / PENSION	1.2511	0.068	18.368	0.000	1.118	1.385
customersector_BANK / DEALER	1.4782	0.069	21.552	0.000	1.344	1.613
customersector_HEDGE FUNDS	0.9736	0.072	13.565	0.000	0.833	1.114
customersector_PRIVATE BANK / WM	1.7045	0.075	22.701	0.000	1.557	1.852

Logistic regression model results in Statsmodels library

Machine Learning

1. Classification Prediction Evaluation Metrics
2. Models Results

Note: Feature selection has been done previously based on:

- Domain Knowledge
- Statistical Tests
- Logistic Regression

NOVA IMS Information Management School

1. Classification Prediction Evaluation Metrics

Threshold metrics (e.g., accuracy and F-measure) – are used to minimize the number of classification errors.

Ranking metrics (e.g., ROC curve and AUC) - are concerned with evaluating classifiers based on how effective they are at separating classes.

Probability metrics (e.g., Brier Score and LogLoss) – these metrics measure the deviation from the true probability. They quantify the uncertainty in a classifier's predictions.

The article by Kuhn and Johnson (2013) states:
"…we desire that the estimated class probabilities are reflective of the true underlying probability of the sample. That is, the predicted class probability (or probability-like value) needs to be well-calibrated. To be well-calibrated, the probabilities must effectively reflect the true likelihood of the event of interest.".

Conclusion: The nature of the research problem (explanation task) leads to choosing the probability metrics.

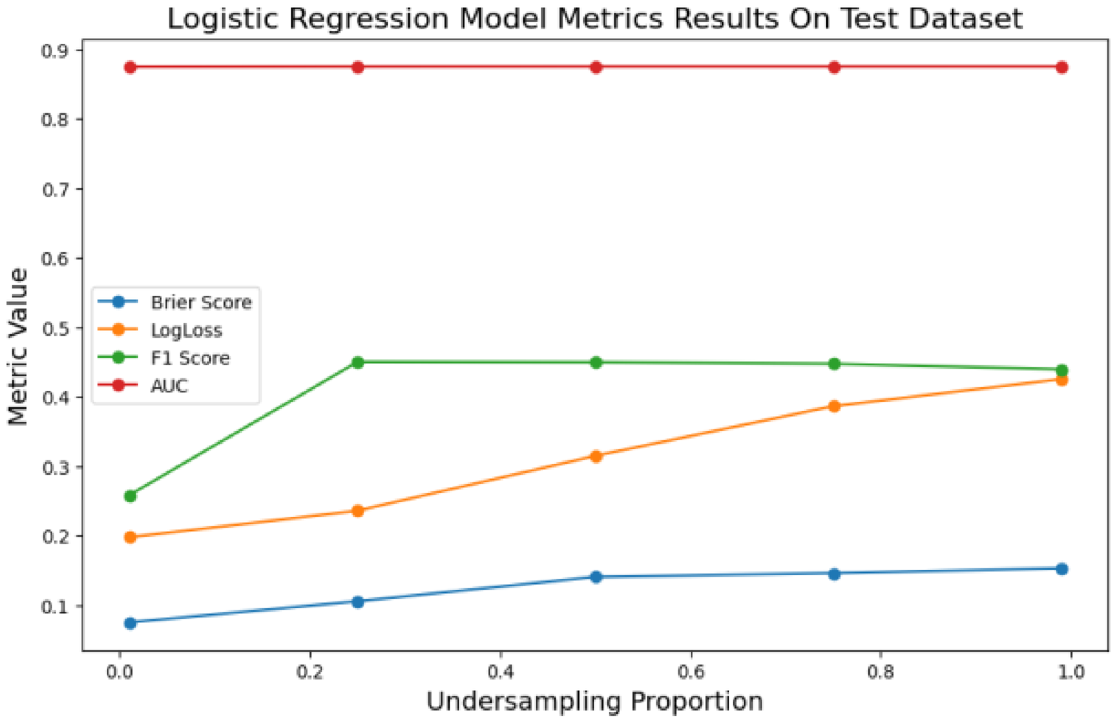
2. Model Results – Random Under Sampling

Dataset	Number of Observations	% Of Full Dataset
Train	336974	80%
Test	84244	20%

Train and test dataset number of observations

Random Under Sampling:

- **Improves** the threshold metric F1 score.
- **Worsens** the probability metrics Log Loss and Brier Score.



Logistic regression model metric results on the test dataset, trained with different class distributions

2. Model Results – LR vs. Random Forest

Why use the Random Forest?

- It is an interpretable algorithm that renders feature importance.
- It is an algorithm with a good performance/accuracy

```
purchased_isin_before_only          0.630442
downloaded_only                     0.144939
downloaded_interaction_high         0.038421
t10_only                            0.032058
purchased_ticker_before_only        0.031507
c100_only                           0.026275
opened_only                         0.018043
customersector_AM / INS / PENSION   0.017671
customersector_BANK / DEALER        0.016006
customersector_HEDGE FUNDS          0.013998
downloaded_interaction_small        0.012844
downloaded_interaction_medium       0.010072
customersector_PRIVATE BANK / WM    0.007722
```

Random Forest feature importance results

- Machine learning with a predictive analytics framework allows compare different model algorithms.
- LR and Random Forest provide similar results.

Model	Average Predicted (RFQ=1) *1	Metrics			
		Threshold	Ranking	Probability	
		F1-score	AUC	Brier Score	Log Loss
Logistic Regression	7,92%	0,26	0,87	0,057	0,197
Random Forest	7,91%	0,25	0,88	0,056	0,195

*1 – it is the mean of the predicted probabilities for each observation belonging to class 1.

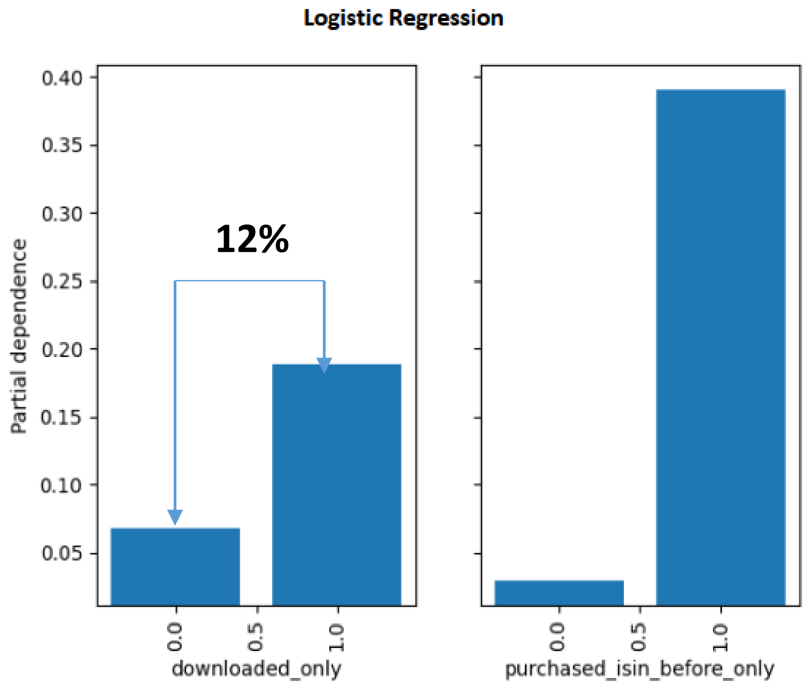
Classification metric results in the test dataset

2. Model Results – Logistic Regression

- Statsmodels vs. ScikitLearn libraries
- The partial dependence plot uses the marginal effect technique.
- The logistical regression mathematical concept remains the same, and as expected, the results too.

	Local	Global
Feature Effects	ICE LIME Counterfactuals Shapley Values SHAP	PDP ALE
Feature Importance	ICI	PI PFI SAGE

Model agnostic interpretation techniques



Logistic regression partial dependence plot results

Causal Inference

NOVA IMS Information Management School

Causal Inference

- Causality subject aims to answer how much a phenomenon X impacts an outcome Y.
- Causal inference could be solved with regression models if all the confounders were accounted.
- A confounding variable is a variable that correlates with both the treatment (download the report) and the outcome variable.
- In observation studies it is very difficult to make sure all confounders are accounted for.
- Example of confounders difficult to measure and include on this project:
    - Client satisfaction and trust in BNP Paribas
    - Other market maker’s reports with their Tradeideas
- **Conclusion:** the regression results must be presented with phrases like “is associated with” and “is likely to cause”, rather than statements that imply causation, such as “causes” or “results in”.

Other Results

Other Results

The results show that <u>downloading</u> or not a report **has no impact** on the client making a trade with BNP or another market maker competitor.

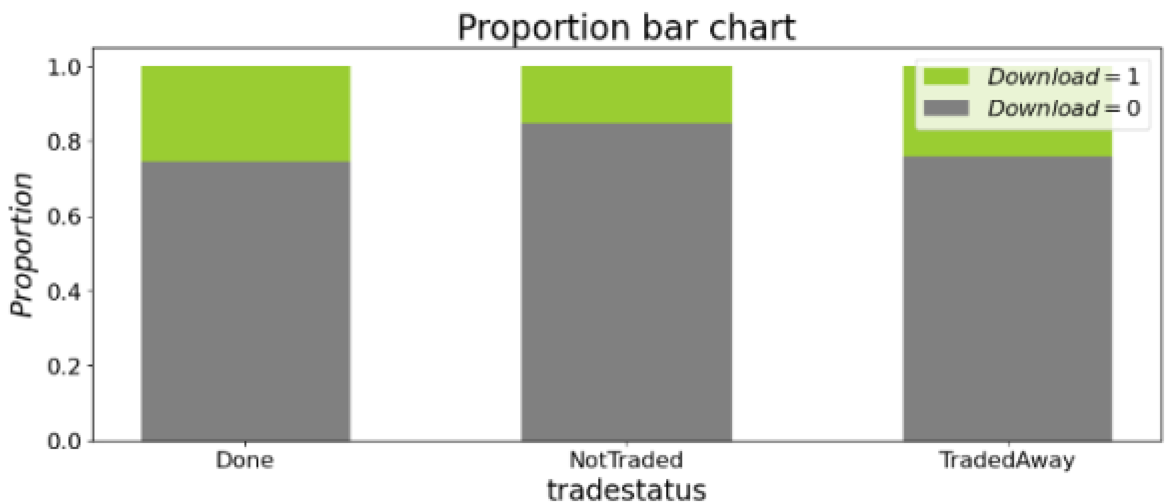


Figure 30: *Tradestatus* variable proportion bar chart against the variable *downloaded_only*

Conclusions

NOVA IMS Information Management School

Conclusions

- Historically, classical statistical education focuses on explanatory statistical modelling and statistical inference, while machine learning focuses on predictive tasks.
- Important to understand if the research problem is an explanatory or predictive task. It has an impact on:
  - The algorithms chosen and their evaluation metrics.
  - Feature selection.
- The clients who **download** the research report show, on average, a **12%** higher likelihood of requesting an RFQ.
- Clients who purchased a specific bond security before the first time it was ever suggested by the bank have **36%** higher chances of requesting an RFQ.
- The reports do not influence whether the client the client conducts a trade with BNP or with another market maker's competitor.
- The results should **not** be read from a causal perspective.

Thank You

Address: Campus de Campolide, 1070-312 Lisboa, Portugal
Phone: +351 213 828 610 Fax: +351 213 828 611

Acreditações e Certificações

Instituto Superior de Estatística e Gestão da Informação
Universidade Nova de Lisboa