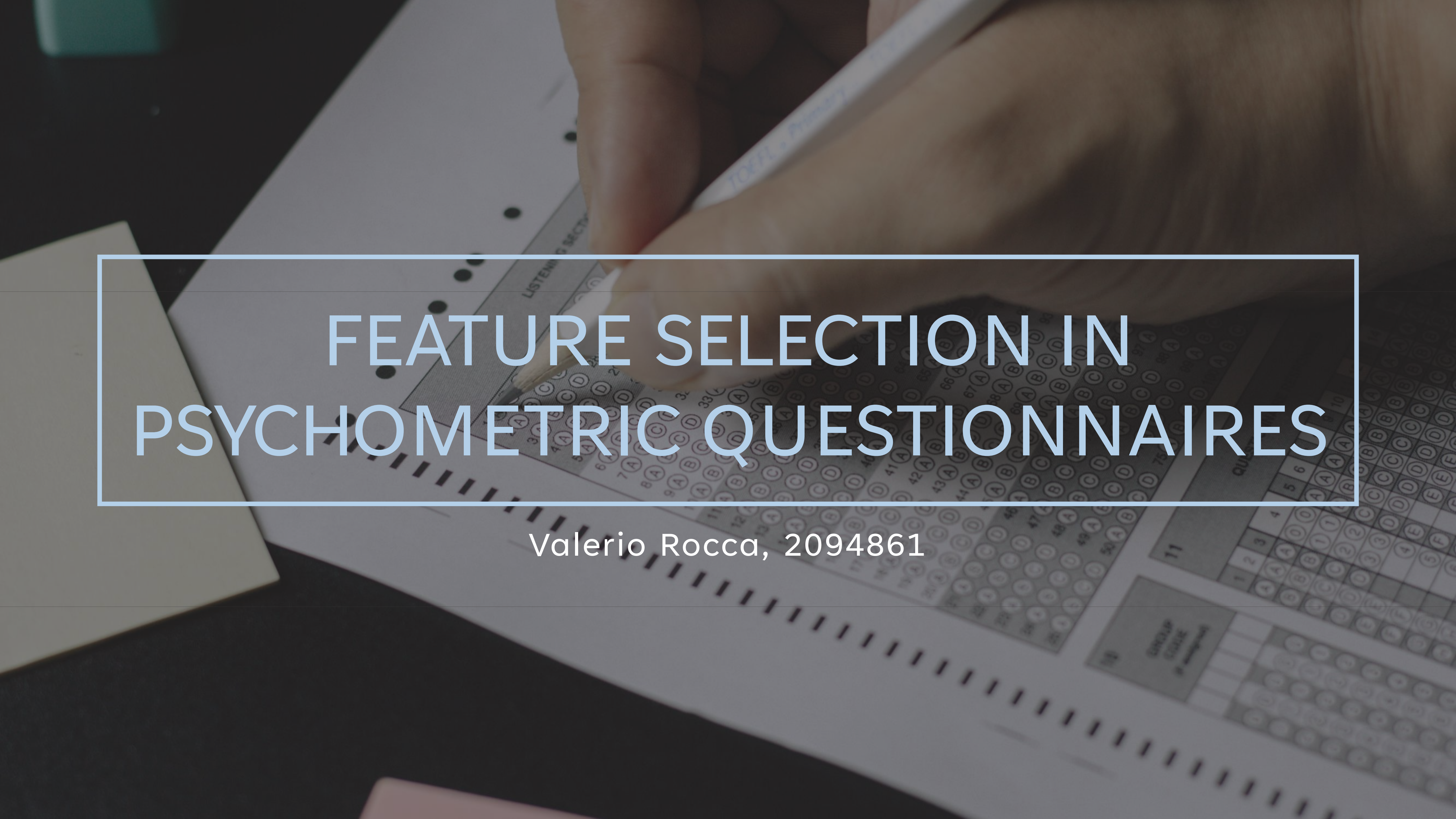


FEATURE SELECTION IN PSYCHOMETRIC QUESTIONNAIRES

Valerio Rocca, 2094861

THE NEED FOR FEATURE SELECTION

PARTICIPANTS-TO-QUESTIONS RATIO

It can be low in psychometric questionnaires

CURSE OF DIMENSIONALITY

Machine Learning models can suffer from a low ratio

FEATURE SELECTION

Extract only the most important features

PROJECT'S GOAL

Explore feature reduction techniques to solve the Curse of Dimensionality problem

THE DATASET

PID-5

Self-report questionnaire designed to assess «Big Five» personality traits

HOW IT WAS OBTAINED

412 participants answered the questionnaire twice: once honestly, once by pretending to have a mental disorder

TASK

Binary classification over honesty/dishonesty on original and feature-selected datasets

DATASETS OBTAINED THROUGH PCA

X

Original dataset

220 features

X_{20}

20% most important components

44 features

X_{OPT}

3 most important components

3 features

MACHINE LEARNING ARCHITECTURES TESTED

LOGISTIC REGRESSION

No regularisation, L1 regularization, L2 regularization

FEED-FORWARD NETWORK.

Tuning of learning rate, hidden layer size and dropout probability

K-NEAREST NEIGHBOURS

Tuning of K

RANDOM FOREST

Tuning of max-tree depth

SUPPORT VECTOR MACHINE

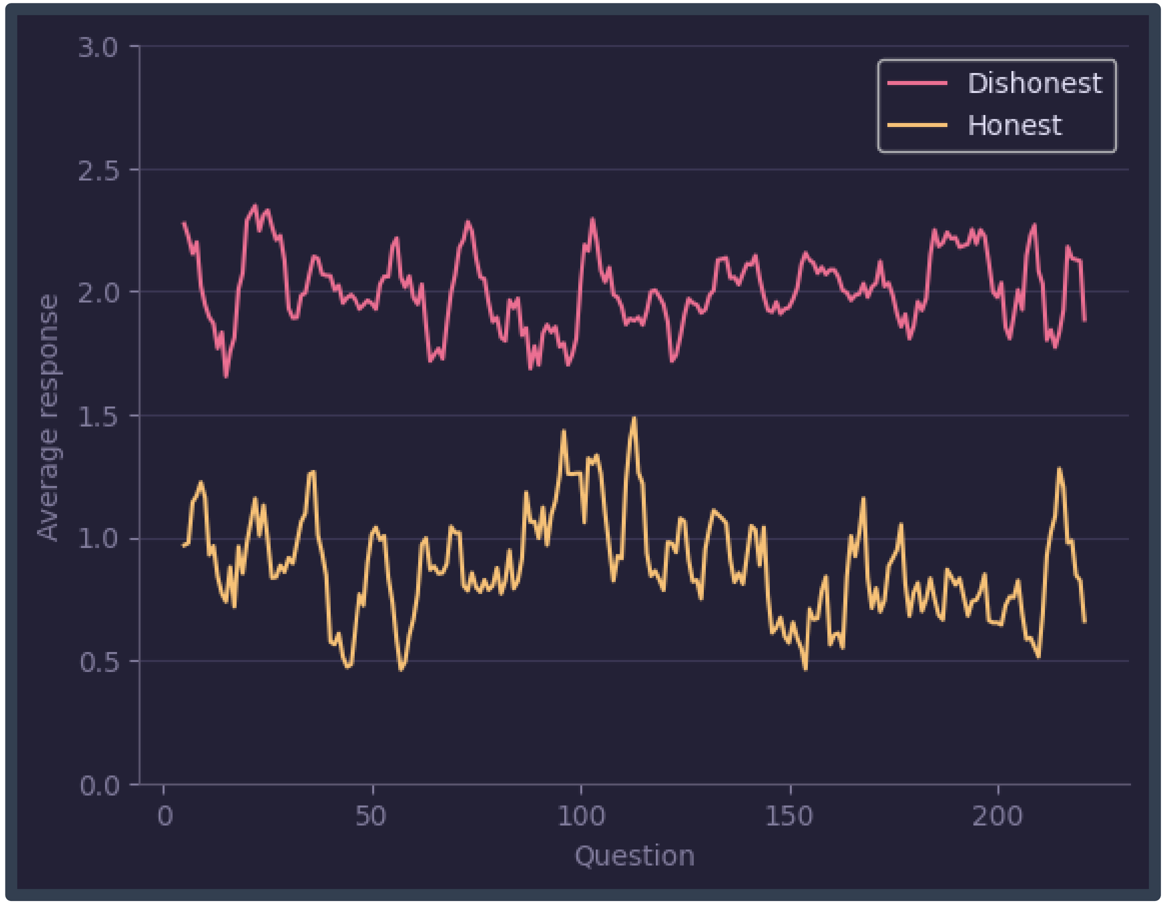
Tuning of the penalty parameter

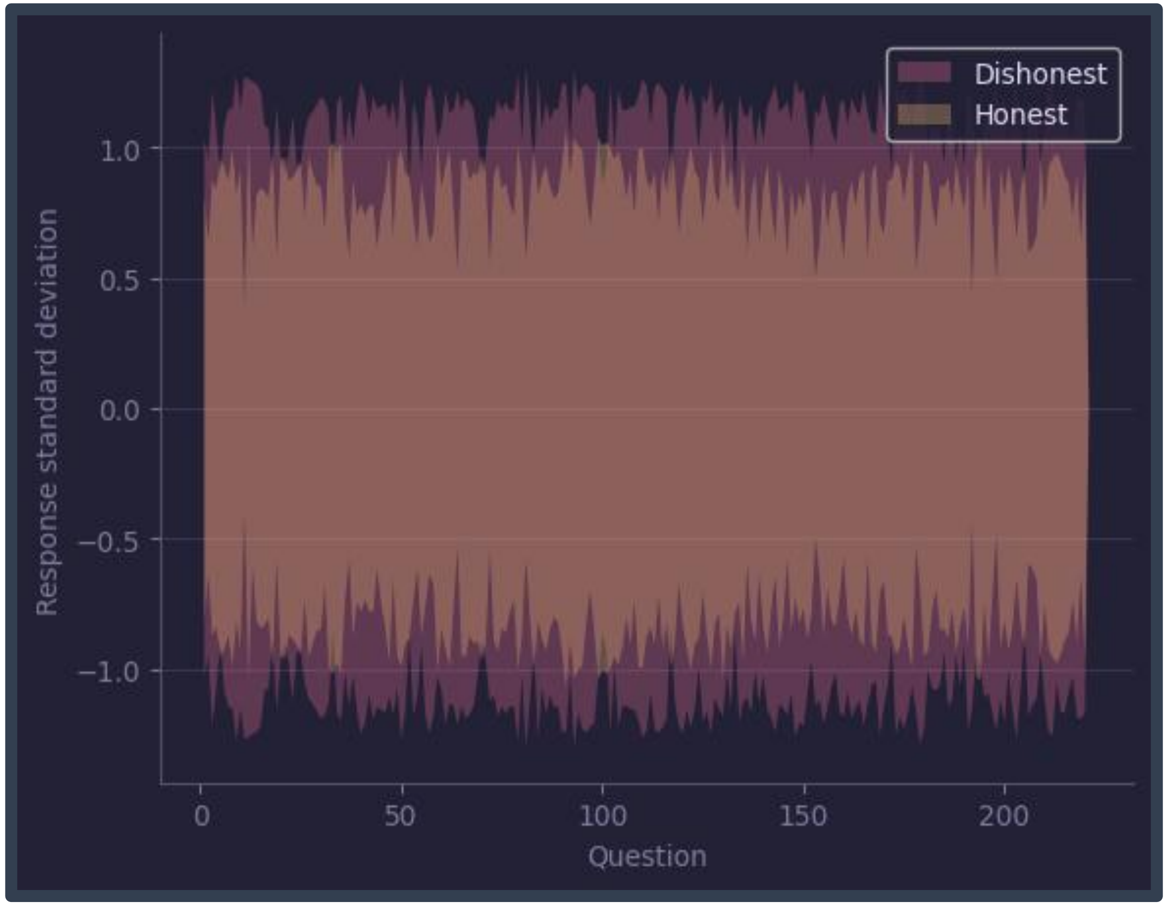
NOTE

Tuning was performed through Grid Search using 5-Fold Cross Validation

EXPLORATORY DATA ANALYSIS

Dishonest answers linked to higher variablity and higher Likert values





RESULTS AND CONCLUSIONS

TEST ACCURACIES

	X	X_{20}	X_{OPT}
Logistic Regression (no regularization)	95.2%	94.5%	94.5%
Logistic Regression (L1 regularization)	96.1%	95.5%	94.5%
Logistic Regression (L2 regularization)	96.4%	96.4%	94.5%
K-NN	93.3%	94.8%	94.8%
Random Forest	97.3%	97%	94.8%
Support Vector Machine	97.3%	96.7%	95.5%
Feed-Forward NN	96.7%	97.3%	94.8%
Average	**96.0%**	**96.0%**	**94.8%**

KEY FINDINGS

MODEL-AGNOSTIC FEATURE SELECTION

Accuracies are comparable despite PCA application

MODEL-DEPENDENT FEATURE SELECTION

Strong feature reduction on X through L1 regularisation does not lead to a lower accuracy

CONCLUSIONS

Models do not seem affected by the low Participants-to-Questions ratio

Nonetheless, feature selection does not impact classification quality