Explainable AI to Gain Insight into Big Five Personality Predictions from Financial Transaction Records

Yanou Ramon, Robert A. Farrokhnia, Sandra C. Matz, David Martens

INTRODUCTION

- Psychological profiling from digital footprints data
- Models built from sparse, high-dimensional data with many relevant features are "black box"
- Explainable AI is important to understand, validate and improve models for psychological profiling

METHODS: CASE STUDY

- ***Data:***

N=6,408 users of mobile app
Big Five personality survey data
578 pre-processed spending features

- ***Predictability of Personality:***

Decent accuracies to predict Big Five personality (*min=53.4%, max=61.8%*) *(Suppl. Material 1)*

- ***Explainable AI Techniques:***

Global: rule-extraction & feature importance ranking
Local: counterfactual explanation rules

"Local explanations reveal granular insights into why classifications are made. Our experiments show that individuals are classified as exhibiting a personality trait for reasons that reflect their unique financial spending behavior."

Example 1 of local explanation:
IF Person A spent less frequently in {Computer & Electronics}, {Insurance} and {Shops}, and more frequently in {Clothing} and {Restaurants} → THEN not predicted "High Neurotic"

Example 2 of local explanation:
IF Person B spent less frequently in {Tobacco} and {Shops}, and spent less money on {Subscription} and {Tobacco} → THEN not predicted "High Neurotic"

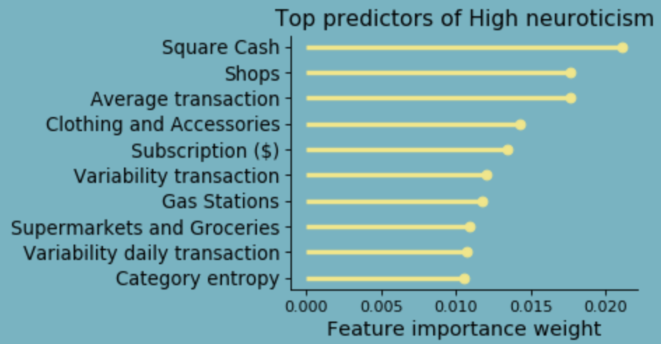

Fig.1: Global feature importance for "Neuroticism" model.

Local explanations differ from global explanations (see *Fig.1*). For example, '*Tobacco*' is ranked *73rd* out *578* features (not shown in *Fig.1*), but it is an important feature in example explanation 2.

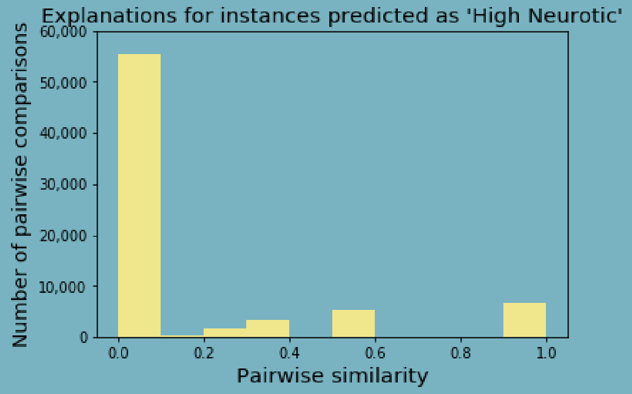

Fig.2: Similarity of local explanations for "Neurotic" predictions. A similarity of 1 indicates that two explanations are the same.

Fig.2 shows that people receive different explanations for predictions made about them. In the "Neuroticism" model, 91.1% of the explanations are unique.

SPSP ANNUAL CONVENTION 2022 February 16-19 San Francisco

RESULTS

- Local explanations for predictions are unique & concise *(Suppl. Material 2A)*
- Global explanation rules for predictions reflect overall classification behavior *(Suppl. Material 2B)*

DISCUSSION

- ***Local Explanations Useful When Modeling Digital Footprints Data:***

Insights into how data is used
Validation of individual predictions

- ***Implications of Explainable AI:***

For Academia:
Validation and improved insights
Robustness and replicability
For Industry:
Improved human-machine interaction
Transparency to data subjects (e.g., "Why am I seeing this ad?")

FUNDING & CREDITS

Funding by Research Foundation Flanders. Thanks to SaverLife for providing us with the data.

Universiteit Antwerpen · Columbia Business School · SaverLife · fwo