ML-Based Feature Importance Estimation for Predicting Unethical Behaviour under Pressure

Pablo Rivas [1] **Pamela J. Harper** [*2] **John C. Cary** [*3] **William S. Brown** [*4]

Abstract

We studied the utility of using machine learning algorithms in the estimation of feature importance and to visualize their dependence on *ethicality*. Through our analysis and partial dependence plot we found linear relationships among variables and gained insight into features that might cause certain types of ethical behaviour.

1. Introduction

As a result of numerous high-profile ethical lapses by corporations and their employees, research into the contributing factors of ethical conduct has grown. To that end we investigated several specific member attributes and behaviors that impact ethical conduct (Cary et al., 2018; Cary & Rivas, 2017). Our study examined the role of gender, income, and religiosity in shaping ethical conduct, and the degree to which perceptions of pressure might moderate these variables. Using standard statistical analysis such as linear regression and correlation coefficients, we determined that in addition to gender and religiosity, the perception of pressure is a factor in unethical behavior.

However, state-of-the-art machine learning (ML) algorithms have also proven to be robust in modeling features in datasets and utilizing intrinsic non-linear transformations over such features to determine the best way to utilize them. Thus, this work aims to use ML to determine the relative importance of the feature set in our dataset.

[*]Equal contribution [1]Department of Computer Science, School of Computer Science and Mathematics, Marist College, New York, USA [2]Department of Organization and the Environment, School of Management, Marist College, New York, USA [3]Economics, Accounting, and Finance Department, School of Management, Marist College, New York, USA [4]Management Department, School of Management, Marist College, New York, USA. Correspondence to: Pablo Rivas <Pablo.Rivas@Marist.edu>, Pamela J. Harper <Pamela.Harper@Marist.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

Table 1. Descriptive statistics of the dataset.

FEATURE	N	μ	σ	max	min
ETHICALITY	334	3.936	0.61	1.7	5
PRESSURE	334	2.700	0.78	1	5
AGE: 18-29	334	0.994	0.08	0	1
AGE: 30-49	334	0.003	0.06	0	1
AGE: 65+	334	0.003	0.06	0	1
SEX	334	0.554	0.50	0	1
M: NVR MARRIED	334	0.976	0.15	0	1
M: NOW MARRIED	334	0.018	0.13	0	1
M: SEPARATED	334	0.006	0.08	0	1
LVL EDUCATION	334	2.919	0.73	2	6
E: EMPLOYED	334	0.069	0.25	0	1
E: OUT OF WORK	334	0.006	0.08	0	1
E: SELF EMPLOYED	334	0.018	0.13	0	1
E: STUDENT	334	0.907	0.29	0	1
RELIGIOSITY	334	2.054	0.92	1	5
INCOME	334	5.180	1.26	1	6

2. Background and Methods

2.1. Dataset

In our previous study we administered a questionnaire to 336 business students of a small northeastern United States institution of higher education. The sample group included undergraduate students about to enter the workforce and graduate students who are currently employed. Table 1 shows descriptive statistics about the features in the dataset.

2.2. ML Algorithms

2.2.1. SUPPORT VECTOR MACHINES FOR REGRESSION

If we define a positive constant $C > 0$ describe the trade off between the training error and define a penalizing term on the parameters of a support vector machine for regression (SVR) as $||\mathbf{w}||_2^2$ promoting sparser solutions on $\mathbf{w}$. And if we further, let variables ξ_i and ξ_i^* be two sets of nonnegative slack variables that describe an $\epsilon-$insensitive loss function; then we can commonly define an SVR with the purpose of being a predictor over $\mathbf{x}$ whose objective function in its

primal form as follows:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*} \quad & \tfrac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \\ \text{s.t.} \quad & \begin{cases} y_i - \mathbf{w}^T\mathbf{x}_i - b \leq \epsilon + \xi_i \\ \mathbf{w}^T\mathbf{x}_i + b - y_i \leq \epsilon + \xi_i^* \\ \boldsymbol{\xi}, \boldsymbol{\xi}^* \geq \mathbf{0} \end{cases} \\ \text{for} \quad & i = 1, 2, \ldots, N. \end{aligned}$$

where $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$ defines our data set.

We trained an SVR to model $f(x)$ with our dataset by successively selecting a single feature from the feature set to record the cross-validated R^2 coefficient using each, where $R^2 = 1 - \frac{\sum(y-f(x))^2}{\sum(y-\bar{y})^2}$. Values of R^2 close to 1 or -1 would indicate high feature importance, while values close to 0 have low predictive value. Figure 1 depicts the results of raking the features using SVRs; this indicates that *Pressure* is one of the best predictors by itself; the rest of the features, individually, are arguably not good predictors.

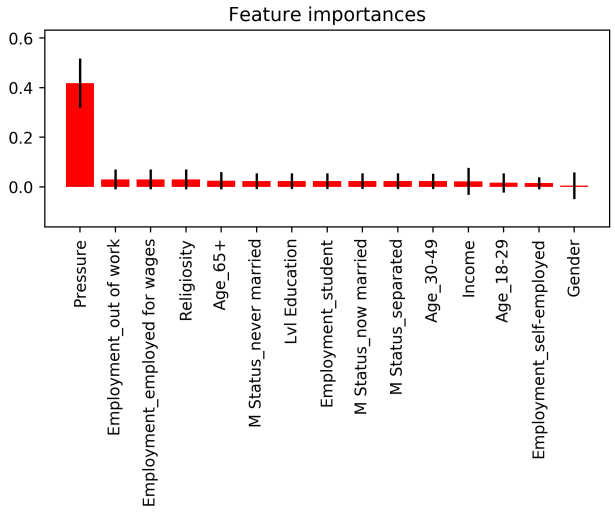

Figure 1. Feature importance by isolating features and training SVRs. The variable *Pressure* is the highest predictive value on R^2.

Another feature importance metric we can use with SVRs is in terms of its improvement or worsening of the R^2 coefficient. For this we systematically remove a specific feature and train with the rest to determine the contribution of such feature. First, we establish a baseline coefficient R^* which accounts for training with the full set of features and getting the cross validated score. Then for k-th feature we can quantify its level of contribution by observing the change with respect to the baseline, Δ_{R^*}. The contribution of the k-th feature can be determined as $\Delta_{R_k^*} = |R^*| - |R_k^2|$. Figure 2 depicts the results of our $\Delta_{R_k^*}$ analysis on the feature sets where it can be seen that the removal of the variable *Pressure* causes a positive $\Delta_{R_k^*}$, i.e., the model significantly drops its predictive capabilities if this feature is removed. It is followed by features related to *Marital Status* and *Employment*.

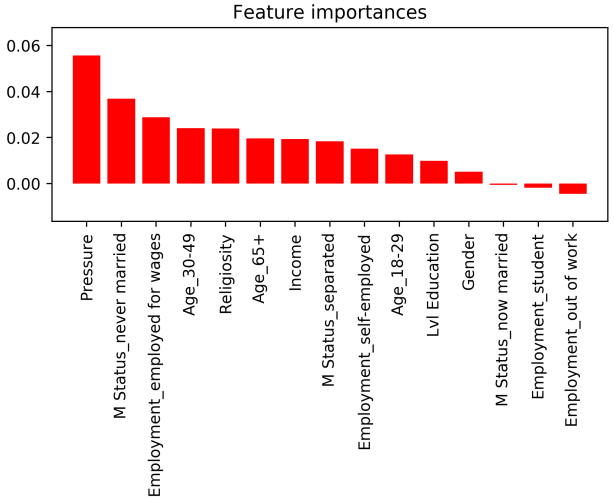

Figure 2. Analysis of $\Delta_{R_k^*}$. The removal of *Pressure* causes the model to drop its predictive value leading to the largest Δ_{R^*}

All our SVR experiments used Bayesian optimization to find the best set of hyper-parameters (Louppe, 2017).

2.2.2. Random Forests for Regression

Random Forests (RFs) belong the the ensemble category of supervised ML. The theory behind RFs indicates that each tree in the ensemble is constructed using bootstrapping to produce samples and to make usually small trees (Geurts & Louppe, 2011). For an RF model with M trees and N samples, the size is in the order of $O(MN\log(N))$ in the average case. One of the most interesting properties of RFs is that they have high bias and low variance, which made them popular in applications that require stability and automatic feature engineering (Soltaninejad et al., 2017; Pinto et al., 2018). Due to the latter, RFs can be used to determine feature importance by looking at features near the root of all trees. Features that are frequently and consistently closer to the root, i.e., that are more *pure*, are considered more important. Figure 3 shows the ranking o the features using RFs, which, consistently with SVRs, show that *Pressure* is highly predictive. Furthermore, *Income*, *Education*, and *Religiosity* seem to have adequate predictive power as well.

2.2.3. Partial Dependence Analysis

Partial dependence plots have been widely used to visually perceive the importance of features among themselves in order to assess their predictive power over single variables. These plots have provided great insight in several areas,

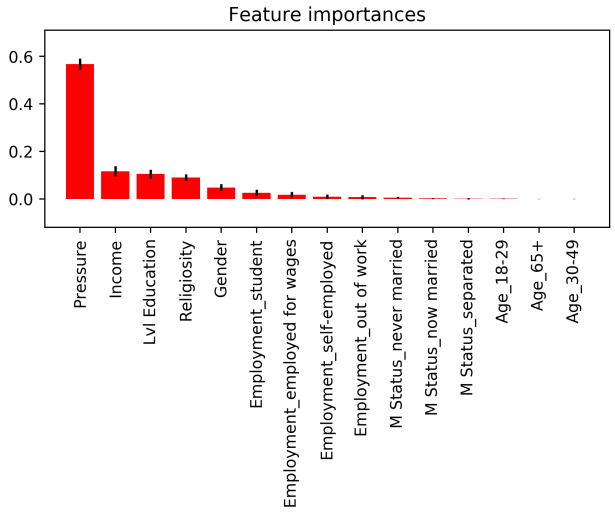

Figure 3. RFs feature ranking shows *Pressure* as the most important feature, followed by *Income*, *Education*, and *Religiosity*.

from the natural sciences (Isayev et al., 2017), to the legal studies (Berk et al., 2016).

Considering *ethicality* our dependent variable, y, a partial dependence plot will display the dependence between y and a single or a set of features, marginalizing any other features on a predictor (Lemmens & Croux, 2006). For this study we chose the most predictive variables to display, as found with the earlier ML methods, and we used a standard Gradient Boosting Regressor (GBR) as our new predictor (Peter et al., 2017). Figure 4 shows the partial dependence of *Pressure* on the left, while the dependence of *Income* is on the middle. The right side shows the interaction of both at the same time. Our partial dependence plots depict a linear relationship between *Pressure* and our independent variable; while *Income* shows a quasi-concave relationship that is more evident on the combined plot. Figure 5 shows the three-dimensional version of the contour plot of the dependence of both *Pressure* and *Income*.

3. Conclusions

In our previous studies we found that when pressure is introduced into a linear regression model, the ethicality of an individual is easier to predict with high statistical significance (Cary et al., 2018). Furthermore, the study presented here confirms our previous findings when we assessed the importance of features in much richer, state-of-the-art, ML models such as SVRs, RFs, and GBRs. However, until now we are able to visualize the quasi-linear dependence of *Pressure* with our dependent variable, ethicality, and further confirm the quasi-concave dependence behaviour of *Income*. The latter suggests that subjects in both the low end and high end of the income range are predictors of ethicality, while subjects whose income is in the middle range are not predictive on ethicality. Further studies will explore the predictive power of all features on each other and not necessarily on the dependent variable ethicality.

References

Berk, R. A., Sorenson, S. B., and Barnes, G. Forecasting domestic violence: A machine learning approach to help inform arraignment decisions. *Journal of Empirical Legal Studies*, 13(1):94–115, 2016.

Cary, J. and Rivas, P. Ethics under pressure? In *Proceedings of the 2017 Susilo Symposium, Boston University*, June 2017.

Cary, J. C., Brown, W. S., Harper, P. J., and Rivas, P. Ethics under pressure: A study of the effects of gender, religiosity, and income under the perception of pressure. In *Proceedinfs of the 25th International Vincentian Business Ethics Conference*, 2018.

Geurts, P. and Louppe, G. Learning to rank with extremely randomized trees. In *Proceedings of the Learning to Rank Challenge*, pp. 49–61, 2011.

Isayev, O., Oses, C., Toher, C., Gossett, E., Curtarolo, S., and Tropsha, A. Universal fragment descriptors for predicting properties of inorganic crystals. *Nature communications*, 8:15679, 2017.

Lemmens, A. and Croux, C. Bagging and boosting classification trees to predict churn. *Journal of Marketing Research*, 43(2):276–286, 2006.

Louppe, G. Bayesian optimisation with scikit-optimize. 2017.

Peter, S., Diego, F., Hamprecht, F. A., and Nadler, B. Cost efficient gradient boosting. In *Advances in Neural Information Processing Systems*, pp. 1551–1561, 2017.

Pinto, A., Pereira, S., Rasteiro, D., and Silva, C. A. Hierarchical brain tumour segmentation using extremely randomized trees. *Pattern Recognition*, 82:105–117, 2018.

Soltaninejad, M., Yang, G., Lambrou, T., Allinson, N., Jones, T. L., Barrick, T. R., Howe, F. A., and Ye, X. Automated brain tumour detection and segmentation using superpixel-based extremely randomized trees in flair mri. *International journal of computer assisted radiology and surgery*, 12(2):183–203, 2017.

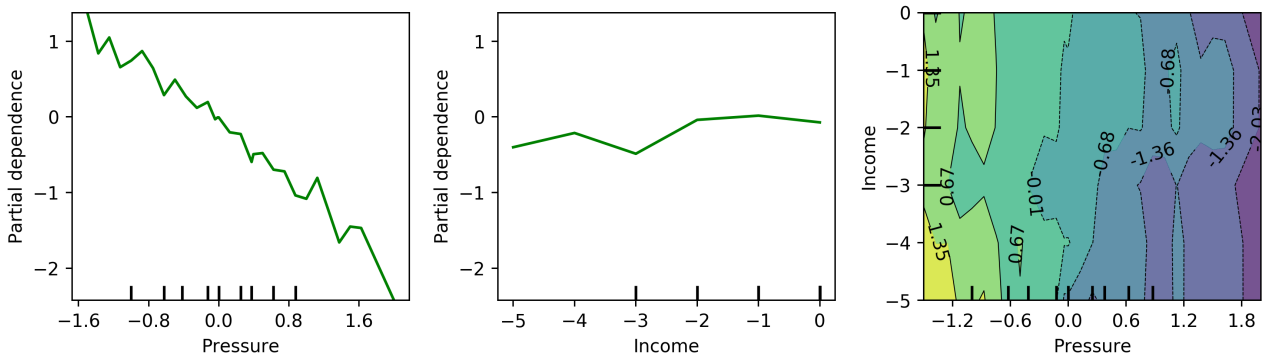

Figure 4. Partial dependence plots for *Pressure* (left), *Income* (middle), and the combination of both in a color-coded contour plot (right).

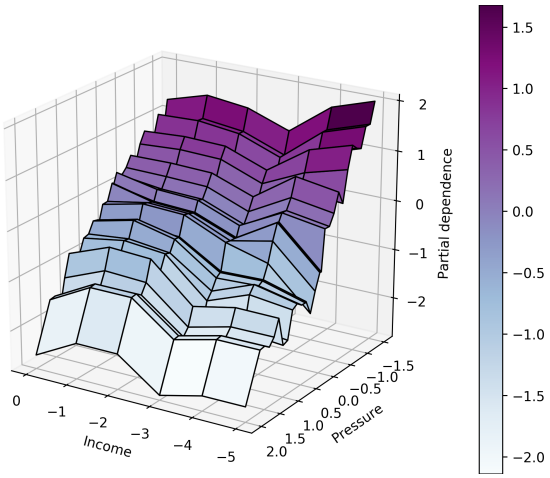

Figure 5. 3D partial dependence plot of *Pressure* and *Income*.