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

Pablo Rivas, Pamela Harper, John Cary, William Brown

To cite this version:

Pablo Rivas, Pamela Harper, John Cary, William Brown. ML-Based Feature Importance Estimation for Predicting Unethical Behaviour under Pressure. LatinX in AI Research at ICML 2019, Jun 2019, Long Beach, CA, United States. hal-02264914

HAL Id: hal-02264914
https://hal.archives-ouvertes.fr/hal-02264914
Submitted on 7 Aug 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
ML-Based Feature Importance Estimation for Predicting Unethical Behaviour under Pressure
Pablo Rivas, Pamela Harper, John Cary, William Brown

To cite this version:
Pablo Rivas, Pamela Harper, John Cary, William Brown. ML-Based Feature Importance Estimation for Predicting Unethical Behaviour under Pressure. LatinX in AI Research at ICML 2019, Jun 2019, Long Beach, CA, United States. hal-02264914

HAL Id: hal-02264914
https://hal.archives-ouvertes.fr/hal-02264914
Submitted on 7 Aug 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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 (López et al., 2018; Rahman et al., 2018; Xiao et al., 2018; Lin et al., 2017; Wei et al., 2016). Thus, this work aims to use ML to determine the relative importance of the feature set in our dataset.

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

2. Background and Methods
Dataset. In our previous study we administered a questionnaire to 336 subjects in the northeastern United States. The sample group included undergraduate students about to enter the workforce and graduate students who are currently employed. Table 1 in the appendix shows descriptive statistics about the features in the dataset.

Support Vector Machines for Regression (SVRs). If we define a positive constant C > 0 describing the trade off between the training error and define a penalizing term on the parameters of an SVR as ||w||^2 promoting sparser solutions on w. And if we further, let variables ξ_i and ξ_i^* be two sets of nonnegative slack variables that describe an ε-insensitive loss function; then we can define an SVR as a predictor over x with the objective min_{w,b,ξ} \frac{1}{2}||w||^2 + C \sum_{i=1}^{N} (ξ_i + ξ_i^*) with constraints y_i - w^T x_i - b \leq ε + ξ_i, w^T x_i + b - y_i \leq ε + ξ_i^*, and \xi, \xi^* \geq 0, for i = 1, 2, . . . , N, where D = {x_i,y_i}_{i=1}^N defines our data set. With this we trained an SVR to model \hat{f}(x) by successively selecting a single feature from the feature set to record the cross-validated R^2 coefficient using each, where \hat{R}^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}. Values of \hat{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 ranking 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. Another feature importance metric we can use with SVRs is in terms of its improvement or worsening of the \hat{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 \hat{R}^2 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, \Delta_R^k. The contribution of the k-th feature can be determined as \Delta_R^k = |\hat{R}^2| - |\hat{R}^2_k|. Figure 1 depicts the results of our \Delta_R^k analysis on the feature sets where it can be seen that the removal of the variables Employment, Education, and Pressure causes a positive \Delta_R^k, i.e., the model significantly drops its predictive ability if these features are removed.
**Random Forests for Regression** (RFs). 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). RFs with $M$ trees and $N$ samples usually have size in the order of $O(MN \log(N))$, and 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, we used RFs to determine feature importance. Features that are frequently and consistently closer to the root, i.e., that are more pure, are considered more important. Figure 1 shows the ranking of the features using RFs, which, consistently with SVRs, show that Pressure is highly predictive. Income, Education, and Religiosity seem to have adequate predictive power.

**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 a dependent variable (Lemmens & Croux, 2006). These plots have provided great insight in several areas, from the natural sciences (Isayev et al., 2017), to the legal studies (Berk et al., 2016). We considered Ethicality our dependent variable and chose the most predictive, as found with the earlier ML methods, variables for display and we used a standard Gradient Boosting Regressor (GBR) as our internal predictor (Peter et al., 2017). Figure 1 shows the partial dependence of Pressure (bottom left), the dependence of Income (bottom middle), and the interaction of both at the same time (bottom right). 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 2 (appendix) shows a three-dimensional version 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). This paper confirms previous findings by assessing the importance of features using 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 sets of features on each other and not necessarily on the dependent variable Ethicality.
References


A. Supplementary Material

Table 1. Descriptive statistics of the dataset.

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

Figure 2. 3D partial dependence plot of Pressure and Income.