Identifying the Critical Features that Affect the Job Performance of Survey Interviewers

Fu Chang, Jeng-Cheng Chen, Chan-Cheng Liu, Chia-Hsiung Liu
Institute of Information Science
Academia Sinica
Taipei, Taiwan
{fchang, clement, cheng, chlu}@iis.sinica.edu.tw

Meng-Li Yang, Ruoh-Rong Yu
Research Center for Humanities and Social Sciences
Academia Sinica
Taiwan, Taiwan
{mengliya, yurr}@gate.sinica.edu.tw

Abstract—In an attempt to build a good predictor of the performance of survey interviewers, we propose a feature selection method that derives the features' strength (i.e., degree of usefulness) from various feature subsets drawn from a pool of all the features. The method also builds a predictor by using support vector regression (SVR) as the learning machine and the selected features as variables. Applying the method to a collection of 278 instances obtained from 67 interviewers participating in eight survey projects, we identified three critical features, experience and two attributional style variables, out of fifteen features. Compared with results of four existing methods, the proposed predictor produced the smallest predictive error. Furthermore, the three features utilized by our method were also identified as the most important features by the four compared methods.

Keywords—Adaptive multiple Feature subset (AMFES); Critical feature; Feature ranking; Feature selection; Support vector regression

I. INTRODUCTION

A person’s job performance is determined by more than one characteristic [1]. Quite a few studies, such as Ref [2-5], have considered the features that affect the job performance of survey interviewers. However, such research has not developed a clear theory or a method to identify the critical traits.

This paper reports the results of a study on the relative importance of interviewers’ features on their job performance. The features include their background information (e.g., age, gender, and experience) and personal traits, i.e., the five personality factors and the attributional styles. A key contribution of the study is the analytical method used. Most quantitative studies of the critical features of interviewers (or people who perform similar jobs, such as salespeople), employ factor analysis or linear regression techniques as analytical tools. In this paper, we propose a novel method for identifying the critical features that can be used to predict the performance of interviewers. The method belongs to the family of feature selection methods [6-8].

A. Salespeople’s Critical Features in Literatures

Since some of the critical features of good survey interviewers and good salespeople are the same, we borrow from the literature on the relationships between salespeople’s traits and their job performance. We then apply two psychological tests in the literature to determine whether the traits can be used to predict the job performance of survey interviewers. The tests are the Big-Five Personality Test [9] and the Attributional Style Questionnaire [10].

The relationship between the Big-Five and the job performance of salespeople has been widely researched. For example, it has been found that conscientiousness is highly related to, and therefore predictive of, a salesperson’s job performance [11-13]. Conscientious people are more likely to exercise self-control, follow the dictates of their conscience [9], and thus fulfill their obligations. Extraversion is also predictive of sales performance [14], as well as to job performance in occupations where interaction with others constitutes a significant part of the job [11, 15]. Extroverts are social, assertive, active, bold, energetic and adventurous [9, 16]. Many studies have found that both conscientiousness and extraversion are important traits that affect sales performance [11-13, 17].

The Attributional Style Questionnaire (ASQ) [10] is designed to assess how people interpret setbacks and failures because a person’s style affects how he/she reacts in such situations, which in turn determines the outcome or resolution of the problem. There are three attribution dimensions: permanence, pervasiveness, and personalization. Those who attribute unfavorable events to permanent, pervasive causes are considered pessimistic. The personalization dimension is related to an individual’s self-confidence [10].

The results of research in the business world also highlight the importance of attributional styles. The literature shows that insurance salespeople with optimistic dispositions are less likely to quit and sell more insurance policies than those with pessimistic attributional styles [18-20]. Pilot studies of salespeople in various industries, such as telecommunications, real estate, office products, auto sales, and banking, produced similar results [21]. Subsequently, Schulman [22, p.34] posited “optimism has an impact on sales productivity
regardless of the industry, whenever persistence is required to overcome adversity.”

In the ASQ\textsuperscript{1} \textsuperscript{10} version that we use, the three attributional styles are applied to both favorable and unfavorable events. Optimists include people who consider that negative (bad) events are due to non-permanent and specific factors, and those who believe positive (good) events are due to permanent and pervasive factors. However, in investigating which of the two types of optimists is more likely to succeed in business, studies in the US and the UK reported conflicting findings (for results in the US, see \cite{18}; for results in the UK see \cite{23-27}). Furthermore, Yang and Yu \cite{28} expressed concern over the interpretation of attributing favorable events to personal reasons due to cultural influences.

B. Feature Selection Methods

A feature selection method builds a predictor and selects features based on a certain objective. Various objectives can be set for feature selection. For example, we may want the predictor to achieve the best predictive power on the selected subset among all possible subsets. However, as this objective is very difficult to achieve \cite{29}. In this paper, we adopt the following approach. First, we rank all features based on their strength. Then, we examine all the subsets of the top-k ranked features and select the subset on which SVR achieves the best predictive power.

There are various ways to rank features. The filter method, for example, ranks features according to their individual properties, while the recursive feature elimination (RFE) method ranks a set of features in a backward manner. RFE first identifies the least useful feature as the lowest ranked feature, and proceeds in this manner with the remaining features. The filter method is fast; however, it may fail to recognize the value of a feature that is weak on its own, but strong when combined with some other features. The RFE method considers the interactive effects of features, but its computation process is slow when the number of features is large.

In this paper, we use the adaptive multiple feature subset (AMFES) method proposed by Chang et al. \cite{30, 31}. One difficulty encountered in evaluating features is the curse of dimensionality; that is, the higher the number of irrelevant features, the less accurate will be the evaluation results. To address the issue, we calculate the strength of features based on various subsets of features by examining the interaction of each feature with some but not all features, thereby reducing the effect of irrelevant features. We then rank the features based on the evaluation results.

To further reduce the effect of irrelevant features, we implement the ranking procedure in a number of stages. In the first stage, we evaluate and rank all features. Ranking has the effect of moving most, if not all, critical features to the top ranks, thereby reducing the number of irrelevant features in those ranks. In each subsequent stage, we input the features whose ranks in the previous stage were above the median rank. We thus deal with a set of top-ranked features, which contains fewer irrelevant features. Then, to improve the feature ranking, we re-rank these features in the same way as we did in the first stage.

II. DATA

In this section, we describe the performance measures and individual features. The performance measures reflect interviewers’ job effectiveness and the individual features reflect interviewers’ profiles and traits.

A. Individual features and performance measures

To obtain individual features, we collected interviewers’ personnel profiles. In addition, we asked each interviewer to self-administer the Big-Five personality test and the Attributional Style test. The 15 features of an interviewer are grouped into three categories: personnel profile, personality traits, and attributional styles, as shown in Table I.

<table>
<thead>
<tr>
<th>TABLE I. INDIVIDUAL FEATURES OF INTERVIEWERS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Personnel Profile</strong></td>
</tr>
<tr>
<td>Gen</td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>Edu</td>
</tr>
<tr>
<td>Exp</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

For the performance measures, we obtained both subjective and objective measures from eight telephone interview projects conducted between December 2008 and July 2009.\textsuperscript{1} Sixty-seven interviewers participated in the study. Most of them participated in more than one survey project and thus had multiple records. In total, there were 278 observations.

For performance, we had three objective measures and one subjective measure. Two of the objective measures relate to the number of successful interviews. They are the response rate (rr5, the fifth definition of the response rate) and the cooperation rate (coop3, the third definition of the cooperation rate) \cite{32}.\textsuperscript{3} The response rate is the proportion of interviews completed out of the total number of eligible cases; the cooperation rate is the number of interviews actually completed out of the number of cases contacted successfully. Both rates are between 0 and 1, with higher scores indicating better performance. The third objective measure, the number of errors (sheet), is the total number of errors an interviewer

\textsuperscript{1} In earlier versions of the ASQ, only unfavorable events were used to categorize test-takers’ attributional styles. However, in the version that we use, both favorable and unfavorable events are considered, and test-takers are required to choose an interpretation of the target event from two alternatives.

\textsuperscript{2} Information on all the personal features was obtained prior to starting all the survey projects used in the paper.

\textsuperscript{3} AAPOR gives several definitions for both the response rate and cooperation rate. We tried different definitions and obtained very similar results. For simplicity, here we present results of only the rr5 and coop3.
made in a survey project. The highest number we collected from the 278 instances was 21 and the lowest was 0.

The subjective instance is the supervisor’s evaluation (score). Such evaluation is routinely conducted and partially decides the interviewer’s pay. During each shift, the supervisors randomly monitor the interviewers’ telephone conversations with respondents. Based on the observations made during the monitoring process, the supervisor evaluates the interviewer’s performance on the following aspects: appropriateness of opening remarks, correctness of sampling within households, quality of tone and attitude, correctness in reading and interpreting the survey questions, probing when necessary, showing efforts to turn around refusals, correctness of answers recorded, and appropriateness in answering questions. Each aspect is evaluated on a 10-point scale. The scores for the aspects are summed to obtain the supervisor’s overall rating. The highest score we collected from the 278 instances was 80 and the lowest was 48.

B. Data Processing

As mentioned in the previous subsection, the ranges of the four performance measures are different. \( r \) and \( c \) range between 0 and 1, sheet ranges between 0 and 21 and grade is obtained as follows:

\[
\text{grade} = \frac{\text{score}}{\text{max score}} \times 100
\]

The four performance measures are different.

\[ B. \text{Data Processing} \]

Among the eight survey projects, some projects were more difficult than others (presumably, the larger the unsigned value of a coefficient, the more predictive will be the corresponding feature. Therefore, if \( f \) is a feature in \( S \) whose coefficient is \( w \), then weight \( t(f) = w^2 \) (cf. Guyon et al. [33] for a different justification).

To address the second question, we do not preset the value of \( m \). Instead, we add one feature subset at a time. Let \( \theta_m = (\theta_m(x_1), \theta_m(x_2), \ldots, \theta_m(x_d)) \) be the vector comprising \( d \) feature strengths derived from \( m \) feature subsets. We stop generating new feature subsets when

\[
\frac{||\theta_m - \theta_{m-1}||^2}{||\theta_{m-1}||^2} < 0.01,
\]

where \( ||\mathbf{x}|| \) is understood as the Euclidean norm of vector \( \mathbf{x} \).

B. The selection procedure

After the features have been ranked, the critical features can be found by identifying the \( F_k \) that has the most predictive power, where \( F_k \) comprises the top-\( k \) ranked features for \( k = 1, 2, \ldots, d \). Adopting a straightforward approach to solve
this problem can be time-consuming because there are \( d \) sets of top-\( k \) ranked features and \( d \) can be a very large number. A better approach involves creating some artificial features whose predictive powers are poor by design. We then put the original and artificial features together and rank them at the same time. The original features that are ranked behind the artificial features obviously have poor predictive powers, so we adopt the remaining features as critical features. Variations of this strategy were proposed [34-36]; our approach is similar to that of Tuv et al.

The method used to create artificial features is shown in Figure 1. We are given three vectors \( \mathbf{x}_1, \mathbf{x}_2, \) and \( \mathbf{x}_3, \) each of which occupies a row of the left-hand side matrix, referred to as \( \mathbf{L} \). Hence, there are three samples and four features in \( \mathbf{L} \). To create four artificial features, we permute the entries of each column in \( \mathbf{L} \). We then place the four original columns and the four permuted columns in the right-hand side matrix, referred to as \( \mathbf{R} \). The three vectors \( \mathbf{z}_1, \mathbf{z}_2, \) and \( \mathbf{z}_3, \) each of which occupies a row in \( \mathbf{R} \), comprise eight features; four of the features are original and the other four are artificial.

\[
\begin{align*}
\mathbf{x}_1 &= \begin{bmatrix} 1 & 4 & 8 & 12 \end{bmatrix} \\
\mathbf{x}_2 &= \begin{bmatrix} 3 & 6 & 9 & 10 \end{bmatrix} \\
\mathbf{x}_3 &= \begin{bmatrix} 2 & 5 & 7 & 11 \end{bmatrix} \\
\mathbf{z}_1 &= \begin{bmatrix} 1 & 4 & 8 & 12 & 3 & 6 & 7 & 11 \end{bmatrix} \\
\mathbf{z}_2 &= \begin{bmatrix} 3 & 6 & 9 & 10 & 2 & 5 & 8 & 12 \end{bmatrix} \\
\mathbf{z}_3 &= \begin{bmatrix} 2 & 5 & 7 & 11 & 1 & 4 & 9 & 10 \end{bmatrix}
\end{align*}
\]

Figure 1. We expand four features into eight features, four of which are original and the other four are artificial.

In the general setting, we transform the training data \((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\) to \((\mathbf{z}_1, y_1), (\mathbf{z}_2, y_2), \ldots, (\mathbf{z}_n, y_n)\), where \( \mathbf{x}_i \) has \( d \) components and \( \mathbf{z}_i \) has \( 2d \) components. The first \( d \) components of \( \mathbf{z}_i \) are identical to the \( d \) components of \( \mathbf{x}_i \) while the remaining \( d \) components of \( \mathbf{z}_i \) are artificial. Since the artificial components of \( \mathbf{z}_i \) are purposely made to disassociate with \( y_i \), they are useless for predicting \( y_i \).

The creation of artificial features involves a permutation of feature values, and therefore introduces a random factor to our setting. To minimize the statistical fluctuation caused by this factor, we require that 50 sets of transformed data must be created, resulting in 50\( d \) artificial features. After applying AMFES to the 50 sets of training data, for each original feature \( f \), we count how many artificial features are ranked above that feature. Denoting the latter number as \( n(f) \), we define the \( p \)-value of \( f \) as

\[
p(f) = \frac{n(f)}{50d}
\]

Finally, we adopt \( f \) as a critical feature if \( p(f) \) is less than a threshold. Usually, setting the threshold lower than or equal to 0.05 yields a good set of critical features.

IV. EXPERIMENTAL RESULTS

In the experiments, we compared the performance of AMFES with that of four alternative methods on the interviewers’ data. The first alternative was RFE [33], which finds and removes the least useful feature and proceeds recursively with the remaining features. The second alternative was correlation (CORR), which is a filter method that ranks features based on the correlation between the feature and the functional value [37, 38]. Like AMFES, RFE and CORR produce a ranked list of the features and use the same procedure to select features. The third alternative method was stepwise regression (STEPWISE), which uses the best individual variable in linear regression as the first critical feature. It then finds another critical feature that works the best with the first one. The procedure is repeated until a stop criterion is satisfied [39]. The last alternative was elastic net (E-Net), which selects features by solving a mathematical programming problem that adds two regularization terms to the usual objective of linear regression [40]. We should mention that, besides selecting critical features, each of the compared methods produces a predictor that calculates a predicted value for any given test object, and expresses the value as a linear function of the critical features.

In our experimental setting, there were 278 data points corresponding to the 278 time-persons involved in the eight survey projects. Each data point was represented by a 15-dimensional vector, whose features are shown in Table I. Since some interviewers participated in more than one project, their data points had the same vector representations. The functional value of each data point is the performance measure computed in Equation (1).

To obtain a test result, we divided the set of 278 data points into training and testing components. The training component consisted of 222 data points drawn at random from the data set; and the remaining data points were used for testing. First, we applied each of the compared methods to the training component to derive the respective predictors. We then computed the mean squared error (MSE) for each predictor applied to the test component. To obtain a stable test result, we selected 5 independent pairs of training and testing components from the data set. The results shown in Table II are the average scores, denoted as MSE scores, of the five methods for the 5 pairs. For AMFES, RFE, and CORR, the critical features were selected by setting a \( p \)-value at which the MSE score is smallest. No such \( p \)-value exists for STEPWISE and E-Net.

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE</th>
<th>Optimal P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMFES</td>
<td>0.3414</td>
<td>0.001</td>
</tr>
<tr>
<td>RFE</td>
<td>0.3595</td>
<td>0.001</td>
</tr>
<tr>
<td>CORR</td>
<td>0.3450</td>
<td>0.001</td>
</tr>
<tr>
<td>SVR Baseline</td>
<td>0.3647</td>
<td>N.A.</td>
</tr>
<tr>
<td>STEPWISE</td>
<td>0.4864</td>
<td>N.A.</td>
</tr>
<tr>
<td>E-NET</td>
<td>0.4003</td>
<td>N.A.</td>
</tr>
<tr>
<td>Linear Regression Baseline</td>
<td>0.5583</td>
<td>N.A.</td>
</tr>
</tbody>
</table>

The results in Table II show that AMFES achieved the smallest MSE score; CORR and RFE had a slightly higher MSE score than AMFES. All the three methods achieved the smallest MSE score at \( p = 0.001 \). Relative to the other methods, the three methods had comparable level of error. The MSE scores of STEPWISE and E-Net were similar, but were higher than those of AMFES, CORR, and RFE. We also show the MSE scores achieved by the SVR baseline and
by the linear regression baseline, where a baseline is a predictor that employs all available features as the variables.

To provide further insight into the predictors built by the five methods, we show the critical features derived by the methods as well their corresponding coefficients in Table III. Note that to derive the MSE scores of the five methods, we created 5 pairs of training and testing components. This strategy allowed us to obtain the average MSE for each method; however, it also created 5 different predictors. To obtain a unique predictor, we applied the five methods to the whole data set, i.e., the set of 278 data points. For AMFES, RFE, and CORR, we built the predictors using the p-value that yielded the smallest MSE score. We applied STEPSWISE and E-NET to the whole data set directly. It is noteworthy that, although STEPSWISE utilized the same three features as AMFES, the corresponding coefficients were different. This is because the two methods employed different learning machines to construct their predictors.

From Table III, we observe that AMFES and STEPSWISE yielded three critical features: exp, pmb, and psg; while RFE, CORR and E-NET yielded 7, 6 and 10 critical features respectively. Among the critical features, exp, pmb, and psg were obtained by all five methods. Note that exp belongs to the personnel profile category, while pmb and psg belong to the attributional style category. Furthermore, these three features were judged the most important ones by all the methods. We assessed the importance of a feature in terms of the unsigned value of its coefficient.

V. DISCUSSION

Our results are not totally consistent with those reported in the literature on salespeople. None of the five personality factors is related to the interviewers’ performance. This is different from the literature, but similar to the results reported by Yang and Yu [28], who found that extraversion was the only significant, but rather weak, predictor in their study of telephone interviewers.

In our study, we found that, to be a good telephone interviewer, the most important feature is interviewing experience and the most important attributional styles are pmb and psg. Good interviewers tend to have more interviewing experience. They are also less likely (the coefficient of a negative sign) to attribute negative events to a permanent cause (pmb), and less likely to attribute positive events to a personal cause (psg). It is reasonable to expect that interviewers with more experience will perform better, as research suggests that experienced interviewers achieve higher response rates [3]. The result of pmb is also reasonable. As telephone interviewers are constantly faced with rejection, a sense of frustration will soon discourage them from trying the next sample person if they do not quickly leave behind the frustration and interpret it as something that only happens occasionally. Yang and Yu also found that pmb is a highly significant predictor of each of the performance indicators [28].

However, the negative effect of psg was unexpected. According to research conducted in Western countries, such as the US and the UK, if psg has an effect, it should be positive. After some thoughts, we speculate that it may be the effect of Chinese culture, which is shared by a great majority of people in Taiwan. Because humility is an important virtue in the culture, Taiwanese people tend to be yielding and humble, and usually evaluate themselves lower than how other people think about them (e.g., Yik et al. [41]). They are likely to blame themselves for failures, and avoid taking credit for successes. In such an environment, attributing favorable events to personal factors may not be related to optimism at all. In contrary, people who are less likely to attribute favorable events to a personal cause may appear to be polite, respectful to others, and thus more likeable. These characteristics increase a person’s chances of obtaining an interview or of being rated highly by his/her supervisors.

VI. CONCLUSION

In this analytical study of the features that affect the performance of survey interviewers, we have proposed and evaluated a method called AMFES for ranking and selecting such features. To compile the set of features, we used information obtained from the personnel profiles of a group of interviewers and the results of two psychological tests. Our study yielded some interesting results. We found that AMFES could build a predictor that achieved the smallest predictive error compared to the errors recorded by four alternative methods. Significantly, the critical features identified by AMFES were also considered the most important features by all the alternative methods. One of the three features, exp, is certainly critical to the job performance of survey interviewers. The other two features, pmb and psg, are measured by the ASQ test, which is regarded as relevant to job performance in several professions. The negative coefficient associated with psg surprised us, since we expected that people scoring higher in psg should perform better in their jobs. This finding suggests that the relations between job performance and the various traits measured by ASQ are more complex than we thought. Thus, further investigations are needed to resolve the matter.

REFERENCES


