Abstract—On April 10, 2006, Major General Bostick, the Commanding General of the United States Army Recruiting Command (USAREC), in conjunction with the Assistant Secretary of the Army for Manpower and Reserve Affairs, approved a new method of selecting individuals from within the Army for recruiting duty. Implementation currently waits for similar and current research concerning Drill Sergeants. Previously, the Army assigned successful noncommissioned officers (NCOs), regardless of inherent sales and marketing skills, into the recruiting force. The Army learned skills that make a successful combat leader do not always translate well into recruiting duty. This research began in 2001, after recruiting shortages, when the Army began researching recruiter selection methods. The result is an application that combines statistical learning with Industrial and Organizational (IO) psychology. The resulting selection model determines the better NCOs for service as detailed recruiters in the United States Army Recruiting Command, a 6000 plus sales force located worldwide. The application enhances IO psychology by providing a statistical prediction of job performance derived from psychological inventories and biographical data. The application uses a combination of statistical learning, variable selection methods, and IO psychology to determine the better prediction function approximation with variables obtained from the noncommissioned officer leadership skills inventory (NLSI) and biographical data. The application also creates a methodology for iteratively developing a statistical learning model. We learned that random forest models outperformed support vector regressions and stepwise regression for these data. A greedy algorithm enhanced model generalization by selecting a good subset of prediction variables. The model represents a multimodal relationship primarily between recruiter age, NLSI score, and, to a lesser degree, 34 other variables. The resulting model runs in $\mathbb{R}$ statistical language and is controlled within an Excel worksheet environment by using Visual Basic Application language and RExcel. The end product enables general utilization of a statistically elegant model, normally reserved for advanced researchers, engineers, statisticians, and economists.

Index Terms—Data mining application, Industrial and Organizational (IO) psychology, random forest, statistical learning.

I. INTRODUCTION

In 2001, the Secretary of the Army, the Honorable Louis Caldera, created and funded numerous recruiting initiatives (Army recruiting initiatives) designed to propel Army recruiting by modernizing the sales force, matching programs and incentives to the current youth market, and reducing barriers to Army enlistment. Part of the Army recruiting initiatives was the creation of a recruiter selection instrument. The initial intent of the recruiter selection instrument was to create and implement an Industrial and Organizational (IO) survey instrument that would accurately predict a noncommissioned officer’s (NCO) potential recruiting performance. Placed within the NCO education system, the instrument would survey the NCO population. Using a predicted performance evaluation, the Army could intelligently choose which NCO was best suited for recruiting duty.

Under these conditions, we demonstrate how to apply statistical learning/data mining and variable selection to assist in personnel selection. Because even the best models are not used if they are not presented well, we demonstrate one method that gained customer confidence and acceptance. Confidence and acceptance can be gained by employing and communicating a sound model-development system, applying industry accepted methods, and leveraging innovation. By coupling a model-development system with acceptable statistical learning and variable selection methods, we mathematically developed an accurate model containing cognitively accepted input variables. Innovation transferred the technological model use from statisticians, scientist, and engineers to common users.

This paper overviews the NLSI, provides a statistical learning/data mining methodology, examines the selected model (a random forest model), and discusses the model’s implementation.

II. NONCOMMISSIONED OFFICER LEADERSHIP SKILLS INVENTORY

A. NLSI Overview

In conjunction with Personnel Decisions Research Institute (PDRI), Army Research Institute (ARI) developed the NLSI to measure potential recruiter success. The NLSI is a psychological survey instrument that measures attributes and skills believed to be relevant to recruiter performance. Both ARI and PDRI have distinguished histories in analyzing and developing IO psychology instruments. Walter Borman, the Chief Executive Officer of PDRI, is one of the major IO psychology field leaders and is personally responsible for much of the current IO body of knowledge associated with developing and analyzing IO psychological inventories [3], [4], [25].

The physical construction of the NLSI consists of three parts and takes approximately 60–90 min to complete testing. The first two parts measure attributes related to recruiter performance. Examples include work orientation, leadership, and interpersonal skills. The third part measures situational judgment skills such as sales skill, social judgment, and leadership. The questions are designed within an elaborate process that begins with a job analysis. The NLSI instrument produces an NLSI score, which may have predictive potential for estimating recruiter productivity.
B. NLSI Data and Potential Generalization

The Army began testing NCOs detailed to recruiting at the Army recruiting course (ARC) at the Recruiting and Retention School (RRS) beginning in January 2002. Initially, the NLSI was given on paper. In January 2003, the NLSI began computerized administration. Over 4000 detailed recruiters have been tested with the NLSI from 2002 to 2005.

Because of these historical records, Army Recruiting Command can pair the NLSI results with the recruiter’s gross write rate (GWR) for each record. GWR is a recruiting performance metric and represents the average number of Army contracts a recruiter can sell in one month. This pairing allows us to develop a GWR prediction model using the NLSI, NLSI components, and biographical data.

Of the over 4000 NLSI records, we were able to precisely pair 1954 recruiter GWR to each record. The individual GWR were carefully selected to incorporate two years of production. We assumed an initial six months of learning, which were not included in the GWR calculation. The research customer, USAREC, demonstrated significant recruiter learning curves within recruiters’ first six months of production. The individual GWRs were calculated from a recruiter’s month seven through their 31 month of production.

Of these 1954 observations, we eliminated observations containing any blank entries and obtained 1913 complete observations and 260 input variables. The 260 input variables are the NLSI score, the NLSI components, and various biographical data such as age and gender.

From the 1913 observations, we constructed a training set and test set of data. Both data sets were randomly selected without replacement. The training data contained 1000 randomly chosen observations, while the test data contained 676 random chosen observations. The remaining 237 observations were withheld for flexibility, in case we required a validation set of data. Both the training and test data initially retained all 260 input variables. The training set of data is used to construct a model for predicting GWR. While the test set of data is used to assess the model’s prediction accuracy.

III. MODELING METHODOLOGY

A. Modeling by Statistical Learning

Because statistical learning is exploratory in nature, we developed a system that facilitates the discovery of the more useful function approximation paired with the best set of model variables. The system performs an iterative process until the best possible function approximation and variable set can be obtained from the data. The system processes are represented in Fig. 1 and are comprised of an input, system functions, and an output. The inputs are the data and the system requirements. In our application, the data were the NLSI test, components of the test, and NCO biographical data. The system requirement was one that accurately predicts potential recruiter performance. The system functions are conjecture, variable selection, statistical learning/data mining, and analysis. The functions are discussed in more detail within this section. The outputs are new intelligence about the problem and a model, in this case, a prediction model.

B. Conjecture

Initially, we begin with an assertion about the data that specifies which function approximations may accurately predict the response. Multivariate statistics may provide insight into relationships within the data. Preliminary data mining and statistical learning can also provide insight into data relationships and further provide an initial assessment if a function approximation is capable of accurately predicting the data. Either technique or a combination of the techniques provides more information about the data and potential statistical learning methods.

In subsequent iterations, data understanding is predominately obtained from analysis. Further function approximation assertions are determined by the analysis conducted in the previous cycle. More often, one function approximation dominates other function approximations at the conclusion of the first cycle. Depending on the data and the metric(s) used to evaluate function approximations, it is possible for more than one function approximation in subsequent iterations and occurs when tradeoffs exist between function approximations. Common to subsequent iterations is a narrower focus from the previous cycle until little or no gains in predictive accuracy can be obtained.

C. Variable Selection

Variable selection assists with model generalization. Generalization is the model’s ability to accurately predict new observations. Variable selection methods improve generalization primarily by removing as many noise variables or multicollinearity variables as possible. Noise variables do not contribute relevant information to the model and potentially cause large variances in prediction.

D. Statistical Learning and Data Mining

Statistical learning discovers a useful function approximation \( f(x) \) to the real function \( f(x) \) that underlies the predictive
relationship between the variables and the response. This process combines the data understanding obtained from the conjecture with variable selection to produce a more refined model.

In our application, we postulate that GWR is a function of NLSI, its components, and some biographical data. Mathematically, we write this as \( GWR = f(x, b) + \varepsilon \), for some function \( f(\cdot) \) with input vector \( x \) with some unknown parameters \( b \) and errors \( \varepsilon \). In the applications used later, \( f(\cdot) \) may be linear as with linear regression or nonlinear as in support vector regression (SVR) or random forest. Strictly, using the training data, we estimate \( f(\cdot) \) and GWR as \( \hat{GWR} = \hat{f}(x) = \hat{f}(x, \hat{b}) + \varepsilon \).

Hastie provides a comprehensive catalog of useful function approximations for regression [18]. Although not exhaustive, powerful regression functions can be obtained with linear regression, random forest, and SVR. Detailed expressions can be found in [5]–[7], [10], [18], and [27].

1) Linear Regression: Linear and multiple regressions are some of the most preferred methods used in prediction [24]. They are used frequently throughout science, economics, engineering, business administration, and the social, health, and biological sciences. Recently, linear regression is one of the principle analytical techniques used in Six-Sigma and Lean Six-Sigma [14]–[16]. Because of their widespread use, they are understood by many, and consequently, serve as a benchmark for comparing other function approximation predictions. Additionally, linear regression models have many associated variable selection methods, which also contribute to their popularity.

2) Support Vector Regression: SVR is a recent development in function approximation. The SVR is a robust regression technique. It has recently enjoyed tremendous success in social, health, and biological sciences. The SVR constructs a regression surface using observations rather than variables. It has some ability to predict anomalies, as well as the majority, within the data. A major concern with the SVR is a requirement for a large number of training observations in order to explain most future unseen observations. Smaller training data sets produce larger errors [18] and [19]. Because SVR predicts with observations (support vectors), variable selection methods are prone to including noise variables [10], [18], [19], [27].

3) Random Forest: In 2001, Breiman developed an ensemble regression method, which displayed remarkable accuracy [5] and [6]. An ensemble regression is a collection of several regressions in which each individual regression provides its independent prediction. A random forest is a collection of tree predictors (see [2], [8], [18], or [28] for tree predictor definitions and examples). Random forest accuracy is related to each tree’s strength and the dependence (correlation) between the trees (see Section II-B of [6]). Improving generalization requires sustaining strength while minimizing correlation between trees. Random features methods balance the strength and correlation requirements. Breiman’s research recommends using Adaboost, which randomly splits features at tree nodes, to maintain strength while reducing correlations [6]. The random forest prediction is the unweighted average vote over the entire collection of prediction trees.

Throughout this research, we grew a collection of 500 trees within our random forest. Each tree predicts a response and the collection of responses are averaged, yielding a predicted value.

E. Analysis

Comparing different function approximations can be challenging. This analysis compared function approximations with a common metric. The common metric obtains measurements strictly from test data. For this application, we developed a mean square error (MSE) that did not account for model parameter degrees of freedom. We used the form:

\[
MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}
\]

where \( y \) are the test data responses and \( \hat{y} \) are the function approximation’s predicted response over \( n \) test data observations.

F. Model Refinement

Model refinements occur in subsequent iterations of the processes. We focus our efforts in variable selection for the best function approximation determined by the analysis. The purpose remains prediction accuracy.

Refining variable selection methods depends on the function approximation. For linear regression, the initial variable selection will not change. However, for methods such as SVR or random forest, possibilities exist for improving the variable set. Applying an error metric through a greedy algorithm or a branch and bound algorithm can refine the variable selection and improve model generalization.

IV. RESULTS

Initial data analysis was conducted in R statistical language and Clementine software using all observations. Initial analysis indicated the relationship between NLSI score and GWR was nonlinear. As review, GWR is a recruiting performance metric and represents the average number of Army contracts that a recruiter can sell in one month. We did not include the first six months of recruiter production, which eliminates a recruiting learning curve. Our resulting response represents a steady-state production metric. From this initial analysis, tree methods and SVR demonstrated promise.

From the methodology’s conjecture stage, we explored three principle function approximations: random forest, SVR, and linear regression. We selected linear regression to serve as a benchmark, enabling comparisons. Two variable selection methods were employed. We used stepwise regression against a linear model and we used the natural variable selection provided by random forest. Because the variable selection methods were not native to SVR, we used the variable set provided by each method in two separate SVR models. Although there are variable selection methods for support vectors, they are often computationally exhaustive and typically include noise variables for smaller observations [19]. For these reasons, we chose to use the other variables for the SVR models. This resulted in the construction and analysis of four function approximations.
In the first iteration, we create one random forest, one linear regression, and two SVR models. SVR 1 contains the stepwise variables and SVR 2 contains the random forest variables. The models were trained on the training data. The resulting errors, resulting from the test data, are recorded in Table I. The errors are derived using (1) and random test data, separate from the training data. Each function approximation had one of two variable sets previously described.

The random forest model was chosen for its better MSE. The squared error represents better fit on test data. Subsequent model improvements involved determining if better subset of variables could improve generalization, the ability to accurately classify new and unseen data.

A greedy algorithm was developed to discover if a better subset of variables existed. The algorithm quickly examined the random forest variables within the ordered pair as determined by the random forest importance measure [5] and [6]. The better variable subset occurred with 36 variables and reduced the model’s MSE from 0.128537 to 0.127034 (test data (see Fig. 2)).

The resulting 36 variables were cognitively acceptable to senior Army leadership. The random forest model determined predicted GWR by predominately using the individual’s age and NLSI score. The remaining 34 variables were used to “fine-tune” the predicted recruiter production (GWR). The MSE increase occurring around the fifth variable most likely is not over fitting. The greedy algorithm ordered the variables by node purity. The increased MSE more likely reflects the variable’s importance when interacting with subsequent variables that follow it. Using the NLSI scales, the model used interpersonal relationships, followed by work motivation, emergent leadership, empathy, and hostility toward authority figures as the essential scales for refining the predicted GWR. Senior Army leadership accepted the model’s variables; the variables made sense and may have been those chosen by less computational means. Because the model selected reasonable and cognitive variables, the leadership readily accepted the model and immediately gained confidence in the model’s predictive power. The selected variables are displayed in Fig. 3.

In Fig. 3, the x-axis represents the random forest node purity. Unlike standard regression, random forest does not provide an estimated coefficient. The node purity is a measure of the variable’s importance in determining the estimated response. The vertical axis contains the more important variables from the variable selection. The axis is ordered from the most important variables to the least important variables within the subset. Age and NLSI scores are the most important and occupy the top. The remaining variables are NLSI test questions. Those variables are grouped into the scales, also contained on Fig. 3. The plot of the actual GWR versus the predicted GWR provides insight to the model’s generalization capacity (Fig. 4). The plot suggests that for predicted GWR greater than one (specifically $\hat{GWR} \geq 1.15$), the random forest model is conservative. The model is most likely to underestimate the true GWR. For those undesired instances when the estimated GWR is below this number, the model tends to overestimate and is, therefore, optimistic. The regression line that compares GWR with the predicted GWR is given by $GWR = -0.7345 + 1.6438 \hat{GWR}$, with an $r^2$ of 0.9648.

One important question concerns the independence between the recruiter’s age at the time of the NLSI test and the NLSI score. Intuition tells us that life experience and maturity should influence the NLSI score. If the intuition were correct, NCO would require multiple NLSI tests throughout their career. The pairwise scatter plot of NLSI score, age, and GWR suggests independence between NLSI score and age (see Fig. 5). Some of what the model cannot explain is the impact of the recruiters’ assigned mission, which would potentially cap

---

**Table I**

<table>
<thead>
<tr>
<th>Model</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forrest</td>
<td>0.128537</td>
</tr>
<tr>
<td>SVR 1</td>
<td>0.181814</td>
</tr>
<tr>
<td>SVR 2</td>
<td>0.186511</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>0.253944</td>
</tr>
</tbody>
</table>

---

**Fig. 2.** Random forest variable selection by using a greedy algorithm that searched for the better MSE within the set of ordered variables (better MSE occurred at 36 variables with an MSE of 0.127034).

**Fig. 3.** Random forest important variables and NLSI scale frequencies. Age and NLSI score are the predominate variables: 34 other variables assist with predicting GWR.
Fig. 4. Random forest models predicted GWR versus actual GWR. The plot suggests that the model is conservative for actual GWR greater than one and optimistic for actual GWR equal to and less than one.

Fig. 5. Pairwise scatter plots of age, NLSI score (NLSIPVSC), and GWR (GRS_AVG6). The scatter plot suggests a relationship between age and GWR as well as NLSI score and GWR. Additionally, age and NLSI score appear independent to each other.

V. NEW SYSTEM WITHIN A SYSTEM OF SYSTEMS

By adopting the recruiter selection model, recruiting command introduces a new technological system within an existing system of systems. The previous system contained a contract mission system and a recruiter placement system, both operating within a youth market system.

Because recruiter selection is joining two other established systems to improve Army recruiting, the best intentions of recruiter selection can be offset by not synchronizing the new system with other older systems. Even though the recruiter selection model can adequately select individuals, those individuals may not perform as promised if they are not positioned well within the market or provided an adequate contract mission for the youth labor market. By employing systems thinking, we also recognize that large-scale change in one system may produce catastrophic system effects as it interacts with the other two systems. Sage judgment introduces change carefully while monitoring the entire state of the system. To the Army’s credit, large-scale change is not a consideration. Even though the model potentially could reduce the size of the recruiter force for a stable contract mission, time is required to determine the state of the entire systems and to fully synchronize and reap the benefits of recruiter selection.

Total system risk is mitigated by deliberately implementing recruiter selection, synchronizing recruiter selection with the contract mission and recruiter placement systems, and by introducing small-scale change while reasonably measuring the state of the entire system.

VI. MODEL IMPLEMENTATION

Decision and policy makers prefer guidelines that are easily communicated and visualized. As an example within this application, it might be ideal to create a policy that selected individuals with NLSI scores between 55 and 65 with ages between 30 and 40. The resulting policy is easily understood and most can visualize the policy guidelines. However, in this circumstance, the policy as stated before would not achieve the desired outcome.

Fig. 6 illustrates the complex decision surface of recruiter selection. The peaks represent those individuals the Army would desire for recruiting duty. The valleys represent those who would
Fig. 7. Recruiter selection model’s initial graphical user interface providing links to technical information, CRAN Internet site, and the random forest selection model.

better serve the Army while remaining in operational units. The policy as stated in the previous paragraph would select both populations of individuals with little-to-no gain in contract production. The less performing individuals mitigate the better performing individual’s gains.

The response surface has a shape similar to a rugged mountain range. To navigate the mountains, without getting lost, you need a competent guide who is familiar with all the subtleties of the terrain. The random forest model is the competent guide. But, unlike many regression models, the random forest’s estimated parameters are not as transparent as those of linear regression. The consequence of this on a decision or policy maker can be apprehension and even fear. Because of these consequences, and other similar ones, many competent models are not implemented.

The model naturally selecting cognitively reasonable variables first reduced apprehension. Apprehension is further reduced with innovation. Transforming a complex model into a readily acceptable model requires some innovation. In this application, we used free and common software to develop a working application.

The random forest model operates with R statistical software, which is freely available on comprehensive R archive network (CRAN) [7]. For user comfort and ease of operation, we house the model in Excel and use Visual Basic Application (VBA) and RExcel, also available on CRAN, to call the R statistical language [1], [11], [29], [30]. By combining these technologies, we are able to provide a complex model normally reserved for engineers, statisticians, and scientists for general use. It is the combination and interoperability of this free or common software that provide innovation.

The recruiter selection model’s initial graphical user interface provides information about the model and includes general instructions, the CRAN Internet site, and the random forest model. Because the model is housed in Excel, users within the Army enthusiastically accepted the model (see Fig. 7).

The random forest prediction model is a VBA form with a series of buttons. By pressing the buttons sequentially from top to bottom, a random forest model is trained on historical data, new observations are introduced, a prediction is made, potential recruiters are sorted from best performing to lower performing, and a total GWR is calculated both individually and collectively. The model’s operation is over simplified by pressing VBA buttons. It is the ease of use that gains user acceptance of a complex model (see Fig. 8).

The Army’s Training and Doctrine Command plans to administer the NLSI by computer when soldiers attending professional schools. The typical tested soldier ranges in rank from specialist to staff sergeant, which is an ideal rank for recruiting duty consideration. These NCOs are young enough to identify with the youth labor market and mature enough to accept the additional responsibility and larger perspective.

The random forest model output on test data has an approximately Gaussian distribution shape. The goal is to obtain more personnel within the performance tail of the distribution. Just as in sales, selection becomes a numbers game. If we increase the number of NLSI administered, we can increase the performance of the detailed recruiting force. Fig. 9 demonstrates that as we administer more NLSI, the detailed recruiter production increases. Requiring less detailed recruiters for the contract mission indirectly implies enhanced production. A long-term increase of the detailed recruiting force performance could eventually spur detailed recruiter force reductions. Those NCOs not selected for detailed recruiting would remain in operational formations, where their leadership and experience are vitally cherished.

Model implementation is waiting for current and similar research that uses the NLSI for drill sergeant selection. Senior Army leadership desires multiuse of the NLSI and is correctly concerned with soldier acceptance of the NLSI. Leadership predicts soldiers will bias the NLSI if its singular use is recruiter selection. For the most part, many in the force hold a bias against recruiting duty. The bias typically reverses once a soldier is a
successful recruiter. Using the NLSI as a predictor for drill sergeant duty, which is preferred by the force, should mediate these concerns.

In the interim, the author trained USAREC analysts in R statistical language and provided all R code so that USAREC can update the function approximation. Life experience, including the United State’s global war on terror, potentially changes soldiers’ NLSI answers and results. Model updates should use this methodology to determine the better functional approximation and the better data and response pairs.

VII. CONCLUSION

The recruiter selection model is a case study on applying statistical learning and variable selection and gaining customer confidence and acceptance. The application also contributes to IO psychology by introducing the application of statistical learning theory to better select personnel for given occupations.

Through a logical model development methodology and the communication of that methodology, we were able to build customer confidence and acceptance. The methodology combined accepted professional standards with a systematic method toward model production. The resulting model not only predicted better than standard models obtained from packaged data mining applications (such as Clementine), but also contained acceptable variables. The variables become more acceptable when the model is able to choose variables we cognitively would have chosen without numerical methods. Confidence increases with improved generalization.

We demonstrated the innovation that can occur from operationally joining off-the-shelf software with free software. By combining R statistical language, RExcel, VBA, and Excel, we delivered the model in a user friendly and acceptable application package. We enhanced our customer’s confidence by rendering an eloquent model normally reserved for statisticians, engineers, and scientists simple for general use.

By introducing statistical learning theory and variable selection methods with IO psychology, we advanced the IO psychology body of knowledge. The advance centers on improving the statistical accuracy of IO models. By incorporating statistical learning, IO psychology advances beyond multivariate statistics and linear regression modeling.

Current and future research seeks to find a functional relationship between the NLSI and drill sergeant performance.

ACKNOWLEDGMENT

The author would like to thank Dr. W. Borman, Personnel Decisions Research Institute, and his staff for their support and pioneering spirit, which enabled the blending of Industrial and Organization psychology with statistical learning. He would like to appreciate their patriotism and passion for assisting the United States Army with recruiter selection. He also thanks the visionary leadership of the Army’s strategic leadership, most notably LTG(R) Cavin, MG Bostick, Dr. L. White of ARI, COL(R) Parlier, Ph.D., and Drs. Goerger for their editorial comment and support. He would also like to appreciate and thank the daily sacrifice of all men and women who serve in the United States Armed Forces.

REFERENCES


John B. Halstead (M’04) received the B.S. degree in mathematics of operations research from the United States Military Academy (USMA), West Point, NY, in 1986, the M.S. degree in operations research from Kansas State University, Manhattan, in 1997, and the Ph.D. degree in systems and information engineering from the University of Virginia, Charlottesville, in 2005. He is currently an Assistant Professor and the Core Engineering Sequence Program Director in the Department of Systems Engineering, USMA. He is an active duty Army Officer with a balance of operational leadership positions and senior staff positions. More recent duties included Strategic Planning Officer for the United States Accessions Command and Recruiting Command, Strategic Concepts Officer for United States Recruiting Command, and a Maneuver Observer Controller. His current research interests include statistical learning, multivariate statistics, prediction, and classification.