A Data Mining-Based Model for Assessing Guangzhou's Higher Vocational Colleges 'New Energy Automobile Majors' Vocational Skills

Ling He¹ and Hashima Hamid²

Abstract

New Energy Vehicles (NEVs) have evolved the rules in the Automobile Sector (AS), and Higher Vocational Colleges (HVC) must adapt in order to provide students with the skills they require to be successful within this rapidly evolving industry. For the purpose of measuring the real-world abilities of students participating in Renewable Energy (RE) vehicle programs at the HVC in Guangzhou, China, this study develops a unique model. The approach employs algorithms for data mining to enhance the accuracy and accessibility of results through the use of Random Forest (RF) and Generalized Additive Models (GAM) in a layering architecture. By combining GAM’s detailed study of the features' impact on job performance with RF’s accurate feature selection and the theory of evolution, researchers can investigate non-linear relationships and discover several things about the distinct functions performed by distinct personality traits and skills. In endurance validation tests, the hybrid model obtained an acceptable 88% F1 score, 90% recall, 86% precision, and 88% accuracy. The findings show the positive aspects of using modern data-driven methods to more closely match educational institutions with the constantly evolving needs of the AS. This might improve students' skills and job marketability. Along with solving an imbalance in the HVC training market, this study provides an adaptable framework that can be used in different areas of research.

Keywords: Vocational Education, Automotive Industry, Random Forest, Data Mining, Generalized Additive Models

INTRODUCTION

In the industry of Higher Vocational Colleges (HVC) development, the rapid development of the Automobile Sectors (AS)—particularly with the introduction of NEVs—presents an individual set of difficulties as well as possibilities [1]. These students participate in new energy automobile classes at HVC in Guangzhou, China, and the goal of this paper is to develop an evolutionary and objective method to assess the job-related skills of these students. Given the pressing requirement for synchronizing educational findings with the ever-changing requirements of modern AS, the significance of such a study has been brought to the attention. Training students, guaranteeing they are constantly applicable, and assuring that they are knowledgeable in the most modern technologies are all methods of achieving this objective.

The AS is currently experiencing an evolution between electric and hybrid vehicles, which has led to a refresh in the knowledge and skills that must continue to be acquired in this sector [2]. The change is going on as the worldwide attempt to create green modes of transport becomes increasingly common. When it comes to this future revolution, Guangzhou, China, which acts as a significant center in the AS in the People's Republic, is at the cutting edge of that transformation. The location is an ideal setting for cutting-edge types of HVC training that have been designed to adapt to the constantly changing needs of the transport industry [3]. This is because the region focuses an enormous value on sustainable transportation approaches and green technologies. Considering these circumstances, it is hard to emphasize the importance of Guangzhou's HVC educational institutions because they serve an essential part in the learning process of the foreseeable future of AS experts [4].

When it pertains to the capacity to adapt to swiftly evolving advances in technology, traditional HVC educational systems frequently lag below. This ultimately results in a gap between the skills that students learn and those that are required in the manufacturing sector [5]. As a result of this disparity, undergraduates may
not have been sufficiently ready for the complex realities of the businesses they select to study, which can have an adverse result on their future employment opportunities and academic achievement [6].

Implementing an advanced model for assessing the professional skills of students specializing in new energy AS inventions at Guangzhou's HVC is the main objective of the work that is being recommended. The analysis will be performed using data mining methods. The Random Forest (RF) and Generalized Additive Models (GAM) techniques are the two principal methodologies that are integrated into this model. Each of these techniques performs different roles within a hybrid system intending to improve the accuracy of predictions and the framework's usability. Additionally, traditional academic assessments generally fail to accurately evaluate the balance of technical expertise, interpersonal skills, and job experience essential for success in developing Renewable Energy (RE) AS [7]. This is a drawback of these assessments. Over the period of the last few years, there have been attempts made to deal with these problems through the use of adaptable and updated curriculum that integrate modern technologies. On the contrary, these attempts usually do not fully encompass the hands-on implementation of theoretical knowledge or the evaluation of interpersonal skills such as versatility, problem-solving, and collaborative behaviour, which are features that are becoming progressively vital in job scenarios that are inspired by innovative thinking and cooperative behaviour [8-9].

Feature Selection (FS) and data analysis are the two primary uses for radio frequency (RF). This can be done by taking advantage of its Ensemble Learning (EL) method, allowing the effective identification of the importance of different features via the development of several types of Decision Trees (DT). This method can be used to determine and change the most significant elements of job ability testing, which in turn allows for the dataset to be exposed to further research. This concludes with the application of GAM with the aim of performing more investigations into the links that exist between FE and real-life outcomes. The use of identical functions provides GAM with a chance to model every factor autonomously. The observation of non-linear innovations, which is crucial for understanding the processes by which specific variables impact vocational skill levels, becomes feasible as an outcome. In the next phase, a key distinction is made available to the knowledge of the concept. This phase helps education organizations comprehend the complicated impact that specific skills and features have on the academic success of students.

The success of a reliable model can be reached over any number of metrics, including accuracy, precision, recall, and F1-score, when RF and GAM perform their tasks concurrently within an ensemble framework. To provide an accurate context, the Stacked Model (SM) showed outstanding results in holdout validation tests, which led to an accuracy of 88%, a precision of 86%, a recall of 90%, and an F1 score of 88%. The outcomes of the present research indicate that the hybrid model has been successful in precisely and reliably assessing Vocational Skills (VS), consequently verifying its practical application across academic and job training settings.

The article is organized as follows: Section 2 details the methods involved in the work, Section 3 presents the problem statement and study area, Section 4 proposes methodology, Section 5 presents the experiment analysis, and Section 6 concludes the work.

Methods Used

Random Forest (RF)

RF is an EL method for classification and regression that operates by constructing a multitude of DTs at training time [10-12]. For classification tasks such as ours, the output of the RF is the class selected by the majority of the trees.

Mathematical Representation: Let $X$ be the input feature matrix, where each row represents an observation, and each column represents a feature. $Y$ represents the vector of target values corresponding to each observation. The RF model is denoted as $RF(X,Y)$, and it operates by creating a set of DT $\{T_1, T_2, ..., T_n\}$, where each tree $T_i$ is trained on a random subset of the data.

Feature Importance Computation: The importance of a feature in a RF is measured based on how much the accuracy of the trees decreases when the feature is excluded, EQU (1).
A Data Mining-Based Model for Assessing Guangzhou’s Higher Vocational Colleges’ New Energy Automobile Majors’ Vocational Skills

\[ F_I = \frac{1}{n} \sum_{i=1}^{n} \Delta \text{accuracy} (T_i, j) \]  

where \( F_I \) is the importance of feature \( j \), \( n \) is the number of trees, and \( \Delta \text{accuracy} (T_i, j) \) is the decrease in the accuracy of tree \( T_i \) when feature \( j \) is lost from the model.

**Generalized Additive Models (GAM)**

A GAM is a statistical model that allows for the flexibility to model each variable independently through smooth functions, thus capturing nonlinear trends or patterns. When used in the present instance, it improves knowledge of how unique features impact professional skill levels.

**Mathematical Representation:** EQU (2) is a formula that represents the GAM that was used for the features \( X' \) that were chosen and processed.

\[ \text{GAM}(X', Y) = \beta_0 + \sum_{i=1}^{p} f_i(x'_i) + \epsilon \]  

Here, \( \beta_0 \) is the intercept, \( f_i(x'_i) \) signifies the smooth function implied by the \( i^{th} \) feature transformed, and \( \epsilon \) symbolizes the GAM’s error time. The functions \( f_i \) are typically valued using splines or similar methods to agree with flexibility in the method.

**Smooth Function Estimation:** Individually function \( f_i \) is valued using methods such as cubic splines or kernel smoothing, confirming that the model can capture complex, non-linear associations between features and the target. The decision-making process of smoothing variables can often be determined by cross-validation. This is performed to avoid overfitting while providing a valid approximation to the data.

**PROBLEM DEFINITION**

Professional educators and industry stakeholders practice significant problems when measuring VS in the context of new energy AS training education. Because of the ever-changing character of the manufacturing sector, which is marked by frequent changes in regulatory requirements and technological advances, one of the main problems is the need for skilled workers. It can be challenging to design syllabus and tests that will be appropriate in the future due to the unpredictability of the context [13-15]. Furthermore, the skill sets that must be acquired are frequently complex and multifaceted, requiring integrating scientific knowledge and hands-on skills, resulting in the difficulty of using traditional assessment practices. Moreover, there is the challenge of synchronizing educational results with the demands of the marketplace, which can vary significantly from one company to the next and evolve as the business environment grows. Also, the unreliable nature of measuring soft skills, which include characteristics like adaptability, problem-solving, and collaborative behaviour, causes heterogeneity in tests, which can impact the reliability and consistency of analyses.

**Operational Definitions**

As it applies to the new energy AS, "Vocational Skill" includes all-embracing knowledge and skills crucial to attaining fulfillment in this industry. To begin with, the term "Technical Skills" encompasses skills and expertise. In order to operate and service electric automobiles, one needs a fundamental knowledge of electronic platforms, battery innovation, and modern engines. This includes the ability to demonstrate proficiency in these areas. Another reason why soft skills are so valuable is that they make it simpler to apply scientific knowledge among team members and across distinct divisions. Soft skills include the ability to work together, strong interpersonal skills, and creative problem-solving. One last significant factor is industry availability, which shows that an undergraduate or graduate has the skills for success in employment. In addition, it demands technical and soft skills but needs an awareness of the rules, practices, and common goals within the business sector. In order to make sure that students possess the skills for working successfully from the inception of their professional lives, it is essential to perform an accurate assessment encompassing practical job placements, project-driven education results, and opinions of business collaborators. This review is essential in order to determine industry preparedness measures.
Study Area: Guangzhou's HVC and the New Energy AS

Guangzhou, situated in China and famous for its significant role in AS, provides a unique ambience for hydrogen-powered vehicles, especially within the domain of new energy automobiles. Because the city aimed to build a centre for green technology and sustainable transportation solutions, it is a perfect location for this research. In this rapidly evolving field, the Human Resources and Development Center (HVC) in Guangzhou is at the cutting edge of providing education to future generations of experts.

Geographic and Economic Context: Guangzhou in China, which is located in the Pearl River Delta, a business and financially developing area, has the advantage of being close to vital technology and production hubs. Several investments have been made to improve the supply chains and manufacturing infrastructure of New Energy Vehicles (NEVs) in this part of the country. Such investments were made possible by regulations adopted by national and local governments to progress market demand for NEVs. The city's primary focus on decreasing greenhouse gases while supporting ecological sustainability enhances its potential as an investigation location for testing VS in the new energy AS. This is because the city is actively promoting environmental education and RE.

Educational Landscape: An effective platform for both theoretical and practical training is presented by the HVC in Guangzhou, China, which has a strong relationship with the local business community. For the objective of serving students with hands-on training in fields such as AS design, battery manufacturing, and electronic system management, these educational institutions have been provided with modern amenities, which include particular laboratories and training sessions. The learning material has been updated with the most recent improvements in technology and business standards because of partnerships with renowned AS.

Target Population: Students who have registered in such courses usually come from many different communities, comprising fresh graduates of high school and experts who are in the early stages of their professional lives and interested in focusing on NEV. This research's primary focus of interest is learners now enrolled in new energy AS courses at several HVCs in Guangzhou. In order to meet the market need that the AS has for highly trained workers skilled in cutting-edge technologies such as battery-powered drive systems, self-driven technology, and improved automotive testing, these courses have been intended.

Relevance to Industry: Given the Chinese government's aggressive targets for NEV sales and production, there is a high demand for skilled workers who can contribute to innovation and efficiency in this sector. The VS of graduates from these programs are critical to the industry’s growth and sustainability. By assessing these abilities effectively, educational institutions can tailor their offerings to meet industry needs better, thereby enhancing job readiness and employment prospects for graduates.

METHODOLOGY

Data Collection

In order to provide a profound overview of VS in the new energy AS, the data collection for the present research comprises the collection of various types of data from a selection of sources. The methods used for collecting data were explicitly chosen. This work was initiated with academic records from the HVC repositories, which embrace course grades $G$, entry logs register $A_t$, and performance system of measurement in specific subjects $P_s$. For the objective of assessing the sum of academic understanding that students have collected, these factors are required. In addition, the results of applications are deemed of the highest significance because they show the use of basic ideas for practical problems. This embraces project grades $G_p$, tutor response $F_t$, and peer evaluations $R_p$. Moreover, job feedbacks provide a direct view into the students' performance in real-world backgrounds, including valuations $E_p$, supervisor reports $S_r$, and skill rankings $S_k$ provided by business associates. Lastly, valuations of soft skills such as teamwork $T_w$, communication $C_m$, and problem-solving $P_s$ are composed through studies using identical tools like Google Forms/SurveyMonkey. These calculations help measure the personal skills crucial for success in the business.
Data Preprocessing

These factors are required in order to achieve the objective of measuring the average level of learning knowledge that students have attained. In addition, the results from using these tools are deemed of the highest importance because they show how to apply fundamental principles to problems addressed in real-life situations.

Lost values in a feature $x'$ are assigned by manipulative the mean $\mu'$ for constant variables and the approach Mo for definite variables, ensuring no loss of vital data due to partial data. Outliers, known using the Inter-Quartile Range (IQR) method, are indifferent if they drop elsewhere $1.5 \times IQR$ from the quartiles $Q_1$ and $Q_3$, confirming the robustness of the test analysis.

For the objective of providing similarity across multiple levels, the data is balanced after it has been processed. This is particularly important when analysing metrics such as grades, which can vary from a single lesson to the next. Individually, score $x$ is identical, employing the Z-score technique, EQU (3).

$$z = \frac{x - \mu(X)}{\sigma(X)}$$

(3)

where $\mu(X)$ and $\sigma(X)$ are the mean and standard deviation of the dataset $X'$, correspondingly. This normalization helps impartial evaluations across different data types.

The next step is processing the data, which involves the classification and design of features. The classifications, such as manager suggestions, have been encoded statistically to be integrated into the model for prediction. The technique for developing new features requires determining an aggregate result, indicated as CS, from several distinct weighted features, indicated as $F_i$. These features may include weighted ratings of technical and soft skills, as shown in EQU (4).

$$CS = w_1 \times f_1 + w_2 \times f_2 + \cdots + w_n \times f_n$$

(4)

The primary objective of this score is to present an overall assessment of VS through integrating every data point. Ultimately, data from different sources, such as educational databases, business relationships, and SQ, is merged into a single dataset signified by the letter 'D' through undergraduate identifiers signified by the letter 'ID'. In order to lay an environment for the development of effective models, this synthesis guarantees that the data associated with each student is complete and suitable for study.

Learning Model: SM Combining RF and GAMs

The following section on technique aims to present a summary of a specific technique which fuses the accurate predictive capabilities of RF with the interpretative strengths of GAM in order to evaluate the VS of students involved with the new energy AS. Within the context of educational evaluation and policy development, this method prioritises boosting the accuracy of predictions while also improving model transparency.

Stage 1: FS and Transformation Using RF

The study begins with the deployment of a RF model, denoted as $RF(X,Y)$, where $X$ signifies the input features matrix-comprising academic results $G$, entry logs $A_t$, and performance system of measurement $P_s$-and $Y$ is the fixed target variable indicating several levels of VS. It is feasible to find out which features of the dataset are most accurate through the use of the RF model, which is employed to calculate the value of each feature. This is stated by the function $= RF_{importance}(X,Y)$, frequent a ranked list of features based on their rank scores. The key features that have been proven to be the most reliable are then transformed in order to boost their predictive skill. This comprises scaling and hypothetically producing communication relationships, signified by $X' = tranform (F_I, X)$, where $X'$ is the matrix of transformed features.

Stage 2: Applying GAM on FS

Upon refining the feature set, a GAM is fitted to the selected and transformed features. This stage is expressed as $GAM(X', Y)$, with each feature $x'_i$ in $X'$ modelled independently using a smooth, non-linear function $f_i$. The GAM is structured explicitly as follows: EQU (5).
\[ Y = \beta_0 + \sum_{i=1}^{p} f_i(x'_i) + \epsilon \]  

(5)

In this equation, \( \beta_0 \) is the intercept, \( f_i(x'_i) \) are the smooth functions applied to each feature, and \( \epsilon \) is the error term. These smooth functions are typically estimated using cubic splines.

**Stage 3: Final Prediction Using Stacking**

For the final prediction, a stacking model is employed, which uses the outputs from both the RF and GAM as inputs to a secondary regression model. This model is denoted as EQU (6).

\[ \hat{Y}_{\text{stack}} = \text{stack} (\hat{Y}_{RF}, \hat{Y}_{GAM}) \]  

(6)

Here, \( \hat{Y}_{RF} \) and \( \hat{Y}_{GAM} \) are the predictions from the RF and GAM, respectively.

The decision to use a stacking approach was based on the need to blend the predictive strengths of both RF and GAM optimally. Stacking allows the model to not only capitalize on the individual strengths of each model but also to learn how to integrate these predictions best. The secondary model, typically a logistic regression due to the categorical nature of \( Y \), effectively adjusts the weights assigned to each primary model's predictions based on their performance on a validation set.

**Algorithm:** Assess VS using RF, GAM, and Stacking

**Inputs:**
- X: Input feature matrix with academic and soft skills assessments
- Y: Target variable vector (VS levels)

**Outputs:**
- Y_Hat_Stack: Predicted VS levels

**Procedure**

// Step 1: Data Preprocessing

For Each feature ‘x’ in X:
- Replace missing values with mean (continuous) or mode (categorical)
- Remove outliers where features>1.5*IQR from Q1 and Q3

Normalize X to X_Norm using:

\[ X_{\text{Norm}} = \frac{(X - \text{Mean}(X))}{\text{std}(X)} \]

// Step 2: FS using RF

Train RF on X_Norm and Y

Compute feature importance FI for all features in X

Select top k features based on FI scores to form X'

// Step 3: Apply GAM on FS

Fit a GAM for each feature ‘x’ in X':

\[ \text{GAM}_\text{Model} = \text{GAM}(X', Y) \]

\[ \hat{Y}_{\text{Hat}_\text{GAM}} = \text{GAM}_\text{Model}.\text{Predict}(X') \]

// Step 4: Obtain Predictions from RF and GAM

\[ \hat{Y}_{\text{Hat}_\text{RF}} = \text{RF}.\text{Predict}(X_{\text{Norm}}) \]  

// Using all features
Step 5: Stacking for Final Prediction

Initialize a secondary model, e.g., logistic regression.
Train the stacking model on concatenated outputs (Y_Hat_RF, Y_Hat_GAM):

\[
Y_{\text{Hat\_Stack}} = \text{Stack\_Model.Predict([}Y_{\text{Hat\_RF}}, Y_{\text{Hat\_GAM}}])
\]

Return

\[
Y_{\text{Hat\_Stack}} \quad // \text{Final predicted VS levels}
\]

Feature Selection

FS is critical to building a predictive model, especially in complex fields such as VS training for the new energy AS. The selection process is directed by statistical significance, relevance to the VS outcomes, and predictive power within the model. Below, we detail the criteria for selecting features, followed by a table summarizing the selected features and their corresponding scores.

Criteria for FS

**Relevance to VS:** Features must directly measure or relate to VS essential in the new energy AS, such as technical knowledge, project execution capabilities, and soft skills like teamwork and communication.

**Statistical Significance:** Features are evaluated for their statistical relationships with the target variable, VS levels, using methods like correlation coefficients for continuous variables and chi-squared tests for categorical variables.

**Feature Importance from RF:** The RF algorithm assesses feature importance based on the decrease in model accuracy when a feature is omitted, providing a quantitative measure of each feature's contribution to model performance.

**Redundancy Reduction:** Features providing overlapping information are identified, and efforts are made to minimize multicollinearity by excluding or combining these features to enhance the model's efficiency and accuracy.

**Practical Considerations:** Features that are reliable, consistently measurable, and easily interpretable are prioritized to ensure the model's applicability and usability in real-world settings.

Table 1 below lists the features selected for inclusion in the model, their importance scores derived from the RF, their correlation with the VS (where applicable), and notes on their practical relevance and statistical significance.

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Importance Score (RF)</th>
<th>Correlation with Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technical Skill Assessment</td>
<td>0.85</td>
<td>0.78</td>
</tr>
<tr>
<td>Project Execution Score</td>
<td>0.75</td>
<td>0.70</td>
</tr>
<tr>
<td>Internship Performance</td>
<td>0.65</td>
<td>0.65</td>
</tr>
<tr>
<td>Teamwork Capability</td>
<td>0.60</td>
<td>0.55</td>
</tr>
<tr>
<td>Attendance Records</td>
<td>0.55</td>
<td>0.50</td>
</tr>
<tr>
<td>Advanced Propulsion Knowledge</td>
<td>0.50</td>
<td>0.45</td>
</tr>
</tbody>
</table>

The features listed in Table 1 above are used in both the RF and GAM:

**RF:** All FS contribute to building robust DT, with each tree considering random subsets of these features to ensure diverse perspectives in the model ensemble.

**GAM:** Features especially critical for interpretability, such as 'Technical Skill Assessment' and 'Project Execution Score', are modelled using smooth functions to elucidate how these features affect the VS predictions.
EXPERIMENTAL SETUP, TRAINING THE MODEL, AND EVALUATION METRICS

The study's setup, the training methods for the hybrid model that employed RF and GAM, and the measurement techniques that were adopted to assess the model's performance are all explained in this section. The robustness and accuracy of the model are validated through employing well-established evaluation metrics.

Hardware and Software Environment

**Hardware:** The experiments are conducted using a server with an Intel Xeon Processor with 32 GB RAM and a 1 TB SSD for efficient data handling and computation.

**Software:** The analysis is performed using Python 3.8, utilizing libraries such as Scikit-Learn for ML models, Pandas for data manipulation, and Stats models for advanced statistical techniques. The GAM fits are conducted using the PyGAM package, which broadly supports spline functions and model diagnostics.

**Dataset Description:** The dataset comprises data from 600 students enrolled in new energy automobile programs across several HVCs in Guangzhou. The features include technical skills assessments, project execution scores, internship performances, teamwork capabilities, attendance records, and advanced propulsion knowledge, as outlined in the feature selection section.

**Training the Model**

**Random Forest Training**

**Data Partitioning:** The dataset is split into a training set (70%) and a test set (30%) to ensure the model is tested on unseen data.

**Parameter Tuning:** Hyperparameters for the RF model, such as the number of trees, depth of trees, and minimum samples per leaf, are optimized using a grid search with cross-validation on the training set to balance between bias and variance.

**Model Fitting:** The RF model is trained on the training set using the optimized parameters to predict VS levels.

**Generalized Additive Model Training**

**FS and Transformation:** Based on the feature importance scores from RF, the top features are selected and prepared for modelling in GAM.

**Smoothing Parameter Selection:** The smoothness of the function for each feature in the GAM is selected using automatic selection methods like Generalized cross-validation (GCV) within the training process to prevent overfitting.

**GAM Fitting:** The GAM is fitted on the same training set using the FS and smoothing parameters to provide a flexible but interpretable model of the relationships.

Table 2 provides values for the parameters used to train the RF and GAM models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>Number of Trees</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Max Depth</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Min Samples Split</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Min Samples Leaf</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Bootstrap</td>
<td>True</td>
</tr>
<tr>
<td></td>
<td>Criterion</td>
<td><strong>gini</strong> for classification</td>
</tr>
<tr>
<td></td>
<td>Smoothing</td>
<td>Automatic (GCV)</td>
</tr>
<tr>
<td>GAM</td>
<td>Splines Type</td>
<td>Cubic splines</td>
</tr>
<tr>
<td></td>
<td>Max Iterations</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Scale</td>
<td>Dependent on data distribution</td>
</tr>
<tr>
<td></td>
<td>Link Function</td>
<td>Logit for binary outcomes</td>
</tr>
</tbody>
</table>

Table 2: Training Parameters
A Data Mining-Based Model for Assessing Guangzhou’s Higher Vocational Colleges’ New Energy Automobile Majors’ Vocational Skills

Evaluation Metrics

Model Evaluation

Accuracy: The primary metric for assessing the performance of the RF and GAM models separately and in the final SM. Accuracy measures the proportion of correctly predicted instances out of all predictions.

Precision, Recall, and F1-Score: These metrics provide a deeper understanding of model performance, mainly in how well the model identifies students with specific levels of VS. Recall is a metric of the ability to identify all positive cases. In contrast, precision is a metric of the accuracy of optimistic predictions. The F1-score measures both precision and recall in a 1:1 ratio.

AUC-ROC Curve: The Area Under the Receiver Operating Characteristics (ROC) curve is applied primarily for the SM to assess how well it performs across different classification thresholds. This analysis provides knowledge about the trade-offs between sensitivity (TPR) and specificity (FPR).

Model Validation

K-Fold Cross-Validation: This is used mainly to validate the RF and GAM models during the tuning phase stage. This ensures that the model's accuracy remains similar across different data sections.

Holdout Validation: The final result of the validation step requires the use of the test set, which includes a minimum of 30% of the data that the model hadn't previously seen. This procedure is performed in order to simulate practical use and assess the capacity for generalization of the framework.

Results for Performance Metrics

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>82%</td>
<td>80%</td>
<td>85%</td>
<td>82.5%</td>
</tr>
<tr>
<td>GAM</td>
<td>78%</td>
<td>75%</td>
<td>80%</td>
<td>77.5%</td>
</tr>
<tr>
<td>SM</td>
<td>86%</td>
<td>84%</td>
<td>88%</td>
<td>86%</td>
</tr>
</tbody>
</table>

Figure 1: Result comparison for performance metrics
There are different views on how to use each model to assess VS, which are made apparent by the test data provided in Table 3 and Figure 1 for the RF, GAM, and SM. The following is an in-depth review based on the data points that were stated:

Excellent results are shown by the RF model, which has an accuracy of 82%, signifying that it successfully accurately classifies the VS in many cases. If the model provides a prediction that a student demonstrates a particular degree of VS, it is accurate 80% of the time, determined by the precision of 80%. In addition, finding that the model has a recall rate of 85% illustrates that it is more than adept at identifying the relevant cases that correlate to a specific VS grade. The F1-score, an appropriate ratio of precision and recall, is now 82.5%, demonstrating that the accuracy level is good in the two metrics.

With an accuracy of 78%, GAM presents an acceptable ranking, considering it is true that it is marginally less successful than the RF in this particular case. This decreased accuracy, along with the precision of 75% and recall of 80%, indicates that although GAM is usually reliable, it may miss or inaccurately label more instances than the RF while maintaining a high accuracy level. Additional proof that the hypothesis is true is presented by the F1-score of 77.5%, which shows a more balanced performance regarding precision and recall than RF.

Across all metrics, the SM provides outstanding results. This model aggregates the results from both the RF and GAM data. It improves the individual models, showing its higher capacity to accurately determine VS, with an accuracy of 86%, while also superior to each model. When it comes to predicting high VS, the model with the best accuracy rate, which is 84%, means that it is the most accurate of the models. Furthermore, the recall rate of 88% is the best, suggesting it is the most successful technique to determine all actual cases of achieving the levels described. The F1-score of 86% illustrates a high level of accuracy and consistency in the predictive model, proving that the addition of the two models increases the overall result by focusing on the best features of each model correctly.

### Five-Fold Cross-Validation Results

<table>
<thead>
<tr>
<th>Fold Number</th>
<th>RF</th>
<th>GAM</th>
<th>SM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold 1</td>
<td>81%</td>
<td>77%</td>
<td>85%</td>
</tr>
<tr>
<td>Fold 2</td>
<td>83%</td>
<td>79%</td>
<td>87%</td>
</tr>
<tr>
<td>Fold 3</td>
<td>82%</td>
<td>76%</td>
<td>86%</td>
</tr>
<tr>
<td>Fold 4</td>
<td>80%</td>
<td>78%</td>
<td>84%</td>
</tr>
<tr>
<td>Fold 5</td>
<td>84%</td>
<td>80%</td>
<td>88%</td>
</tr>
<tr>
<td>Average</td>
<td>82%</td>
<td>78%</td>
<td>86%</td>
</tr>
</tbody>
</table>

---

He and Hamid
The outcomes from the five-fold cross-validation method are presented in Figure 2 and Table 4 for the RF, GAM, and SM. These findings provide an in-depth view of the reliability and repeatability of every model across an assortment of data sets. The following section is an analysis that depends on the accuracy statistics that have been provided for each fold:

The RF model exhibits an efficiency that is relatively consistent across all five folds, with an accuracy between 80% to 84% and an average of 82%. This coherence indicates that the RF model is effective, as it performs reliably across an extensive selection of data sets without dealing with significant variations from one set to another. A sign that the model has excellent generalization skills is that it can preserve an accuracy higher than 80% across all folds.

GAM presents an average accuracy of 78%, which is an approximate higher degree of accuracy, ranging from 76% to 80%. It's feasible that this difference illustrates how sensitive GAM is to specific features of each data set. This might occur because GAM primarily emphasises normalizing and decoding the complicated relationships within the data. While the performance is more modest and less reliable than RF, it nonetheless provides an appropriate degree of accuracy, proving it is helpful when clarity is just as essential as predictive accuracy.

The accuracy levels for the SM vary from 84% to 88%, with a median of 86%. This implies that the SM is superior to the RF and the GAM in every dimension. The effectiveness of integrating the predictive features of RF and GAM is proven by the greater accuracy across all folds from this combination of tools. The SM not only reaches an enhanced level of accuracy but also exhibits outstanding stability and reliability, which implies that it successfully uses the best features of both models to maximize its performance.

**AUC-ROC Performance**

In order to assess the True Positive Rate (TPR) (sensitivity) in contrast with the False Positive Rate (FPR) (1-specificity) over several threshold preferences, the AUC-ROC is an important statistic. With the support of this measurement, one may learn the model's ability to distinguish between classes in a timely fashion. The total efficacy of each model is outlined in Table 5, which depends on the AUC-ROC scores of distinct models.

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>0.89</td>
</tr>
<tr>
<td>GAM</td>
<td>0.84</td>
</tr>
<tr>
<td>SM</td>
<td>0.92</td>
</tr>
</tbody>
</table>

**Figure 2: Cross-Validation Results**

The figure shows the cross-validation results for the RF, GAM, and SM models. The accuracy percentage is plotted against the fold number, demonstrating the reliability and repeatability of each model.
For the RF, the GAM, and the SM, the AUC-ROC scores that are provided in Figure 3 offer an in-depth review of the power of each model to distinguish between classes practically, especially between different levels of VS. A deeper examination of the AUC-ROC metrics for these models is provided below:

This demonstrates that the RF model has a significant capacity to identify between the classes with an acceptable level of accuracy. The AUC-ROC for the prediction is 0.89. An AUC-ROC value of less than 1 means the model performs highly adequately, while a rating of 0.89 for RF indicates high classification skills. This shows that the framework balances the trade-offs between sensitivity (TPR) and specificity (FPR), enabling it to make accurate predictions across many benchmarks.

The AUC-ROC for GAM is 0.84, less than that of RF; however, it demonstrates a high prejudiced performance. Based on this statistic, it can be concluded that whereas GAM can be used to distinguish between classes, it may be slightly less successful than RF when it comes to dealing with the sensitivity-specificity trade-off. On the other hand, the most significant benefit of GAM lies in its potential to present accessibility in the context of classification. This is a vital feature in cases where it is just as essential to know the impact of predictors on results as it accurately predicts the results directly.

With an AUC-ROC of 0.92, the SM provides the highest selective performance among the three models. The success of combining and improving the predictive skills of both RF and GAM is demonstrated by this outstanding result, which indicates performance. The SM provides a substantial boost in both sensitivity and specificity, which features a high AUC-ROC value. This proves that the SM can be particularly successful at accurately identifying participants across many decision levels. This particular model makes effective use of the
A Data Mining-Based Model for Assessing Guangzhou’s Higher Vocational Colleges’ New Energy Automobile Majors’ Vocational Skills strengths that are possessed by its foundational models in order to attain a prediction that is more precise as well as accurate.

**Holdout Validation Results**

To assess the degree to which the models can be applied to new data that wasn't observed before, this section must provide the results of each model on the holdout set. Regarding this issue, the main focus is vital metrics like accuracy, precision, recall, and F1 score.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>AUC-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>85%</td>
<td>83%</td>
<td>87%</td>
<td>85%</td>
<td>0.90</td>
</tr>
<tr>
<td>GAM</td>
<td>80%</td>
<td>78%</td>
<td>83%</td>
<td>80.5%</td>
<td>0.85</td>
</tr>
<tr>
<td>SM</td>
<td>88%</td>
<td>86%</td>
<td>90%</td>
<td>88%</td>
<td>0.93</td>
</tr>
</tbody>
</table>

**Figure 4: Results for holdout validation**

RF provides excellent work with an accuracy of 85%, which illustrates its substantial skill to accurately define VS, as demonstrated in the findings that appear in Fig. 4 and Tab. 6. Based on the precision of 83% and the recall of 87%, it indicates that RF not only provides a high percentage of results that are important, but it also performs highly to detect the most significant number of cases that result in positive. The F1 score of 85%, which is balanced, is more proof that it is successful. With an AUC-ROC score of 0.90, RF is a highly effective model that shows strong identification between classes at a selection of threshold levels.

Even though GAM is slightly less advanced than RF in metrics, it still provides promising findings with an accuracy of 80%. The fact that this model obtains a precision of 78% and a recall of 83% suggests that it is relatively accurate in successfully identifying the TP cases, given that there is some scope for greater efficiency in minimizing the FP. The F1 score of 80.5% reveals an adequate balance between precision and recall; however, it is not the best feasible balance. AUC-ROC values of 0.85 indicate that GAM is effective in classification tasks; however, it is not as effective as RF in this context.

Regarding how they perform, the SM, which incorporates the positive features of RF and GAM, is the most successful choice throughout all metrics. It attains a maximum precision of 88%, which suggests that it has high generalization skills. In addition, it has an overall precision rate of 86% and a recall rate of 90%, which indicates
its advantage in providing accurate outcomes and effectively finding the most positives. The reality that the model can attain an F1 score of 88% indicates that it achieves a perfect balance between precision and recall, thereby validating its reliability. Furthermore, an AUC-ROC of 0.93 illustrates its superior filtering skill, which renders it the most reliable model for predicting VS.

CONCLUSION AND FUTURE WORK

This investigation aimed to assess students' vocational skills registered in new energy automobile courses at higher vocational colleges in Guangzhou, China. The findings of this investigation effectively demonstrated the effectiveness of a hybrid model involving Random Forest and Generalized Additive Models. A powerful tool for accurately assessing technical and soft skills has been developed by combining these two effective data mining methods within a stacking model. In addition, the understanding of how these skills impact vocational skills has been made more accessible as a result of this fusion. Verifying its value in a real-world learning environment is the model's excellent performance, proven by its accuracy, precision, recall, and F1 score. This approach assists in developing an employment pool that efficiently meets the problems that will be prevalent in the future hiring process by connecting the gap between the skills provided in universities and those required by the developing automotive sector. The method goes beyond standard review techniques and includes an in-depth technical and soft skills analysis. This is performed to account for the complicated nature of the business's demand for awareness.

Further, recognising that this model has worked effectively in a place like Guangzhou indicates the likelihood that it could be extended and applied to different locations and professions. It presents educational policymakers and universities with a valuable framework that permits them to develop curricula and methods of instruction that correspond closely with the business world's requirements, improving the value and efficacy of vocational courses throughout the globe.

REFERENCES

K. Peng, C. Hua and Z. Heng, "Design and Production of University Students' All Terrain Two Drive Baja Racing," 2022 5th World Conference on Mechanical Engineering and Intelligent Manufacturing (WCMEIM), Ma'anshan, China, 2022, pp. 199-205, doi: 10.1109/WCMEIM56910.2022.10021379.
A Data Mining-Based Model for Assessing Guangzhou’s Higher Vocational Colleges ‘New Energy Automobile Majors’ Vocational Skills


