Issues in Information Systems
Volume 23, Issue 3, pp. 53-63, 2022

DOI: https://doi.org/10.48009/3_iis_2022_106

Utilizing Data Science and Analytics in Predicting Campus Placement

Caesar Jude Clemente, *Douglas College, clementec@douglascollege.ca*
Myungjae Kwak, *Middle Georgia State University*, *myungjae.kwak@mga.edu*

Abstract

Educational institutions are expected to make students marketable in their respective fields. Job placement exam is a tool to assess a student's readiness to face the industry's challenges. Previous studies have utilized machine learning algorithms to predict students' job placement. However, most of the past research was based on academic and non-academic performance metrics, not on a custom-made job placement exam. The training and test data used in the research were from computer science engineering students who took a job placement exam. The study examined the scores of job placement exam in the different subject areas. In this study, five machine learning methods were utilized to develop the predictive models. Of the five models explored, the random forest model got the highest accuracy, 90.85%, and an F measure of 91.59%. Feature selection using a forward algorithm was then employed to get the most influential predictors. The results showed that coding was deemed to be most important, followed by the aptitude score.

Keywords: Accuracy, F measure, Machine learning, Random Forest, Feature Selector

Introduction

One of the educational institutions' responsibilities is to prepare students for their transition to the professional world (Money et al., 2019). From this perspective, we can also argue that the ultimate stakeholders of universities and colleges are not students but their future employers. Thus, it is vital that educational institutions should prepare students to meet the rigors of the workplace intellectually, psychologically, and emotionally. Unfortunately, not every institution succeeds a hundred percent in this advocacy. For instance, in some places, it is reported that the number of graduating students does not qualify to the quality requirements of the IT industry (Samantha & Poojah, 2020). Another study also revealed that 10.3% of computer science students are unemployed six months after graduation (Smith et al., 2018). Clearly, there is a need for educational institutions to implement programs that will increase the probability of students finding a job immediately after graduation.

One of the programs that universities and colleges have adopted to improve job placement of graduating students is the campus placement program. The standard setup is inviting industry partners to come to the university and conduct a preliminary recruitment service (Trevor Yu & Cable, 2013). The event is typically known as a job fair. Students are asked to prepare their resumes and be ready for an on-the-spot interviews.

Other educational institutions that already have a close relationship with corporate partners usually do it more formally. They give a campus placement exam to graduating students. The exam serves as an initial filtering process before the student is sent to a partner company for an interview. In this setup, students recommended by the institution usually get an assurance of an initial interview, thus giving them an advantage over other candidates. If the student succeeds in getting the job, the university will state in their record that the student was “placed.”

A successful student placement program has many benefits. For the student, the obvious advantage is the chances of getting a job in the industry. For the institution, it is a source of pride to declare that it successfully found employers for students who are graduating or just about to graduate. The high percentage of students "placed" is a testament to industry partners' trust in the educational institution and the strength of the university program. Because of this, universities and colleges usually highlight this aspect when marketing the program to their future students.

It is, therefore, within the educational institution's interest to forecast the job placement of students. Predicting job placement has been done in many previous studies (Manvithha & Swaroopa, 2019; Harihar & Bhalke, 2020; Maurya et al., 2021; Shukla & Malviya, 2019). Most of the previous research has developed a binomial classification model of the campus placement and focused on different internal and external student attributes. This study concentrated on a customized campus placement exam, and comprehensively analyzed and compared a set of machine learning methods to find the most appropriate approach. In addition, it analyzed which attributes are the most influential predictors of each model.

Related Literature

At the heart of data analytics is the process of investigating datasets to get insights that can help organizations in their decision-making process (Runkler, 2020). Data mining is an integral part of data analytics as it is the technique by which data analysts can see patterns in data (Stedman & Hughes, 2021). Different data mining techniques are available and can be applied depending on the problem requirements. These techniques are association rule, sequence path, clustering, regression, classification, and neural network (Stedman & Hughes, 2021). Businesses have long utilized data mining to understand their customers better and leverage this knowledge to gain more customers. Currently, it is not just businesses that want to harness the power of data. Different industries now realize how data analytics can fulfill their organization's strategic objectives. Another terminology frequently mentioned in terms of data analytics is machine learning (ML). The relationship between machine learning and data analytics is that machine learning takes a step further when it comes to deriving knowledge from data. Machine learning does not just provide data analysis; it also learns from the data collected Gupta, 2022). ML’s capability to "learn" is why it is also considered a subset of Artificial Intelligence (Gupta, 2022).

Data analytics and machine learning is also now applied in education, specifically in campus job placement. For instance, in one study, the researchers used decision trees and random forests to predict job placement. Their research output demonstrated 84% accuracy for decision trees and 86% for random forest (Manvithha & Swaroopa, 2019). However, the paper did not identify which features were the most relevant. In another study conducted by Shukla and Malviya (2019), the researchers developed a model that seeks to improve the prediction of student job placement. Their paper utilized data mining techniques by combining clustering algorithms and Naïve Bayes. The result is a 95 to 98% accuracy. Features used were a mix of academic and non-academic attributes.

Another research in student placement prediction used multiple classifiers (Harihar & Bhalke, 2020). The results of Harihar and Bhalke (2020) study showed multiple layer perceptron (MLP), simple logistic and logistic model trees performed well, garnering an accuracy of 99.5%. The features used for the model were extracted from academic and non-academic evaluation (Harihar & Bhalke, 2020). Still, within the topic of multiple classifiers, authors in another study also developed multiple models to predict student placement. They focused on academic performance parameters such as grades on the tenth, twelve, upon graduation, and grades before graduation. Their results showed that support vector machine and logistic regressions got the highest mark having an area under the ROC curve (AUC) of .86. The research also identified the grades in the tenth as the top priority feature in the prediction model. (Maurya et al., 2021).

Another paper attempted constructing a binary logistic regression model to "predict campus placement "for Master of Business Administration students (Kumar et al., 2019, pp 2633). The study used six determinants based on past and present academic achievements. The model reflected 72.06 percent accuracy and identified four influential predictors based on their statistical significance. The chosen predictors are CGPA, specialization in postgraduate, specialization in undergraduate, and gender. Finally, in the study by Samantha and Poojah (2020), they proposed a job placement exam focusing on three areas: coding aptitude and technical skills. In their proposed system, the student will take the exam and be immediately given a prediction score. The authors concluded that the system would be a good indicator of students' readiness for work. The research, however, only focused on the proposed system. There was no accompanying implementation data or result.

Most of the aforementioned studies focused on past academic and non-academic features. Examples of these features are gender, grade point average (GPA), high school grades, interview results, or grades before graduation, to name a few. None of them focused on the result of a single job placement exam. This situation is expected as most universities and colleges do not give job placement exams. There are, however, universities and colleges that utilize these types of exams as their primary tool in choosing who to recommend for job placement. The most significant advantage of having a job placement exam over other job placement programs is that it can serve as an early warning system (Samantha & Poojah, 2020). Unlike the other programs where students are sent to the companies for preliminary evaluation, job placement exams can serve as a preliminary filtering system to detect students' weaknesses, thus providing administrators a chance to intervene and improve the situation. In addition, most of the previous research did not extract which features were the most influential. This study seeks to fill this gap by evaluating the effectiveness of a campus placement exam in predicting students who will successfully be employed and determining which features are the most effective predictors.

Research Questions

Like most of the previous research, the objective of this study is to develop a binomial predictive model of the campus placement. Unlike the previous studies, which focused on different internal and external student attributes, this research focused on a customized campus placement exam given to computer science engineering students. The study aims to answer the following research questions.

1. Can a job placement exam be used to predict employment for computer science engineering students? If yes, what machine learning techniques can be utilized to predict job placement?

2. Which job placement exam attributes are the most influential predictors of employment for computer science students?

Research Methodology

Dataset

The data utilized for this study was originally taken from a public dataset in Kaggle (Moitra, 2022). The researcher verified the information and obtained the approval of the original dataset author before using it for the research. The information in the dataset was anonymized and did not identify the students. The dataset was from a campus placement exam given to graduating students from twenty different disciplines. The dataset did not also provide any demographic description. Campus placement exams differ from past academic performance measures as the students take them for job placement purposes. They are not tied to any course or subject and, by nature considered to be summative exams.

For the purposes of this study, the data is filtered only to computer science engineering students as it has a sizable percentage of placed students. The following table shows the components of the exam.

Table 1-Exam Domains

Component	Description
Aptitude	Evaluates logical reasoning or personality to predict applicants' success on the job
English	Evaluates the students' capability to communicate using business level English.
Quantitative	Evaluates numerical reasoning and problem solving through math related problems.
Analytical	Evaluates general intelligence, critical thinking, problem solving, comprehension and synthesizing information.
Computer Fundamental	Evaluates knowledge of IT fundamental concepts
Coding	Test the students' programming logic formulation skills
Domain	Evaluates knowledge on a specific topic related to the discipline

An aptitude test is designed to test the student's cognitive ability or personality. Quantitative is for mathematical reasoning and analytical for comprehension and critical thinking. Computer Fundamental examines knowledge of basic IT principles and Coding is used test the student's programming capability. Domain exams evaluate specific knowledge and skill in a particular discipline.

Each component represents a column, and the data is in a percentage format. The predictive column is a binomial attribute of either 'placed' or 'not placed'. There were 2,542 computer science engineering students; however, only 263 of them were placed. To avoid an imbalanced dataset, only an equivalent of 263 none-placed students were taken to form the final dataset. The sampling of the 263 none-placed students was conducted using the random sampling. The final dataset comprised 526 data points.

Machine Learning Methods

The following machine learning methods were used for this study.

1. Logistic Regression. Logistic regression is a machine learning algorithm that classifies a categorical variable (yes or no, 0 or 1) based on a set of independent variables. Logistic regression belongs to the category of a generalized linear model. The logistic regression is popular because of the S shape of the logistic function, which shows the function value dramatically increases close to zero (0) to around one (1) as the domain value increases (Zumel & Mount, 2019; Kleinbaum et al, 2002).
2. Decision Trees. Decision trees are a tree-structured classifier where internal nodes represent the attributes of a dataset, the branches are the decision rules, and each leaf is the outcome. A decision tree predicts the value by learning the decision rules inferred from the dataset and one of the widely used techniques in data mining (Charbuty & Abdulazeez, 2021). In addition, it offers a means of identifying more important elements from predictor variables (Thomas & Galambos, 2004).

3. Random Forests. Random forests contain multiple decision trees based on the subsets of the data. This output takes the average results, hence significantly improving the accuracy of the prediction. Ensemble classification methods are machine learning algorithms that combine a set of classifiers to improve overall prediction accuracy (Yang, et al., 2010). Random forest is classified as an ensemble learning model as it combines more than one classifier to improve the model's performance (Akar & Güngör, 2012). A major benefit of using random forest is the capability to handle datasets with a large number of predictor variables (Speiser et al., 2019).
4. Support Vector Machine (SVM). SVM is a machine learning algorithm that finds the hyperplane between the two or more classes and used the hyperplane to linearly separate the data. The farther the data points from the line that separates the classes, the more accurate the classification (Noble, 2006). Therefore, the goal is to choose the best line that can correctly classify the data (Bambrick, n.d.).
5. Deep Learning. Deep learning is a machine learning algorithm based on neural networks that uses many layers to increase its learning capability. In deep learning, the algorithm works on the raw data and determines relevant features. The algorithm employs a nonlinear transformation of inputs to create a model and iterate it until an acceptable accuracy is achieved. The number of processing layers in which data is sifted refers to the model's "deepness" (Burns & Brush, 2021).

Validation Techniques

The research utilizes a K-fold cross-validation technique to prevent overfitting or underfitting. The cross-validation model is considered to be superior to the typical hold-out method, where only a part of the training data is used for validation. In the K-fold validation technique, the data is divided into k subsets, and each time one of the subsets is used as either a test or validation subset. In other words, every data item gets to be part of the validation and the training set. As a result, the entire process reduces bias and variance (Gupta, 2018). In determining the number of folds, the research adapted ten folds based on what previous studies have utilized and have been proven to reduce classification error (Marcot & Hanea, 2020).

Terms and Metrics

1. Confusion Matrix – A table that reflects the performance of a machine learning algorithm. It contains the true positive, which refers to the number of times the model correctly identified positive data, and the true negative (TN), which pertains to the number of times the model predicted negative data. The confusion matrix also shows the false positives (FP), the number of times the model misclassified a negative and false negative, the number of times the model misclassified a positive (Narkhede, 2021).
2. Accuracy – This is a summative metric that shows the ratio of correctly classified items over the total number of items. The following is the formula to compute accuracy (Harikrishnan, 2021).
$$\text{Accuracy} = (TN + TP) / (TN + FP + TP + FN)$$
3. Precision – This metric represents the ratio of the correctly classified positive items over the total predicted positives The following is the formula to compute precision (Harikrishnan, 2021).
$$P = TP / (TP + FP)$$
4. Recall - This metric represents the ratio of the correctly classified positive items over the total positive examples. The following is the formula to compute recall (Harikrishnan, 2021).
$$R = TP / (TP + FN)$$
5. *F*-score or *F*-measure – This metric is considered better than accuracy, especially if the data set is not balanced, as it also considers the precision and recall. F1 score is the harmonic mean of precision and recall. The following is the formula to compute the F1 score (Harikrishnan, 2021).

F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

Data Preparation and Tools

The raw data was in an excel format. Before it was fed to the model, data was prepared to ensure its viability for prediction. In the current sample set, only one data with a missing aptitude exam was found; hence, the record was removed. The attributes except for the placement status were in a percentage format with a percent sign at the end. The percent sign was erased to extract only the number values.

The attributes were examined for collinearity, as shown in the table below.

Table 2: Correlation Matrix of Features

Attributes	Aptitude	English	Quantitative	Analytical	Domain	Computer Fundamental	Coding
Aptitude	1.0	0.677	0.693	0.707	0.313	0.306	0.416
English	0.677	1.0	0.171	0.258	0.189	0.204	0.265
Quantitative	0.693	0.171	1.0	0.234	0.253	0.241	0.333
Analytical	0.707	0.258	0.234	1.0	0.203	0.188	0.254
Domain	0.313	0.189	0.253	0.203	1.0	0.317	0.379
Computer Fundamentals	0.306	0.204	0.241	0.188	0.317	1.0	0.357
Coding	0.416	0.265	0.333	0.264	0.379	0.357	1.0

Based on the above data, no attribute reached 1 or -1; hence all attributes were retained. RapidMiner, a data science software, was used to generate the machine learning models.

Results and Discussion

To determine if the results of the experimentation fulfilled the research objectives, we will map the results to each research question.

Research Question 1

Can a job placement exam be used to predict employment for computer science engineering students? If yes, what machine learning techniques based on a job placement exam can best predict job placement?

The following table shows the summary of models generated:

Table 3-Accuracy and F Score of the Models

Model	Accuracy	F Score
Logistic Regression	77.91%	74.03%
Decision Tree	89.13%	90.13%
Random Forest	90.85%	91.59%
SVM	79.62%	79.47%
Deep Learning	81.14%	81.75%

The above outcome clearly shows that it is possible to use a job placement exam to predict employment for computer science engineering students. These findings are confirmed by the high accuracy rate accrued by each model. Furthermore, the results have also demonstrated that machine learning algorithms such as the one listed in the table above can be utilized to develop predictive models for job placement. All models exhibited more than 70% accuracy and *F* score. Random forest got the highest accuracy (90.85%) and an *F* score (91.59%). Based on this output, Random Forest is the best predictive model. The above accuracy result is at par with previous studies, which have an accuracy rate of 80 to 99.5 percent (Manvithha & Swaroopa, 2019; Shukla & Malviya, 2019; Harihar & Bhalke, 2020; Kumar et al., 2019). Also, similar to previous studies, the results above also revealed random forest as the highest predictive model (Manvithha & Swaroopa, 2019).

Research Question 2

Which attributes are the most influential predictors of each model?

To answer the second research question, the study used a feature selector that extracted the most relevant features of the dataset. A deterministic forward selection algorithm was applied. The algorithm first created an initial population of items. A feature is added in an iterative method and creates a model. Every iteration, a feature is added until a new variable does not improve the model's performance anymore. As part of the process, the resulting model will extract the most influential predictors and only retain these predictors in the model (Kaushik, 2020). Below is the summary result of the feature selection.

Table 4-Accuracy and Features after Application of Selection Optimizer

Model	Accuracy before feature extraction	*F* measure before feature extraction	Accuracy after selection optimizer	*F* measure after selection optimizer	Features extracted
Logistic Regression	77.91	74.03	80.79	80.81	Aptitude, Quantitative, Computer Fundamentals and Coding
Decision Tree	89.13	90.13	89.73	90.41	Aptitude, English, Analytical, Domain and Coding
Random Forest	90.85	91.59	92.95	93.38	Aptitude, English, Analytical, computer fundamentals and Coding
SVM	79.62	79.47	80.62	80.53	Aptitude, Quantitative, Computer Fundamentals and Coding
Deep Learning	81.14	81.75	82.13	83.47	Aptitude, English, and Coding

There is a noticeable increase in the accuracy and *F* measure after selection optimization. This result is because only the most essential features per model were retained, and the model was again generated based on these features. From the extracted attributes, Coding is present in all models. This finding is not surprising considering most computer science graduates' first job is in software development. Therefore, universities and colleges must enhance subjects that emphasize logic formulation skills. Examples of these subjects are programming and software development courses.

The second feature that is present in five models is aptitude. Aptitude exams are standard recruitment assessments, yet some students fail in this subject. Educators should prepare students for aptitude exams by giving them practice exercises. Computer fundamentals and English are present in three models, making these subjects also important. English is a determining factor in deep learning. This result proves what some industry partners had readily observed: some computer science students are good in technical knowledge but deficient in written communication skills. Mathematical reasoning and comprehension skill attributes are only in two models.

These results suggest they are not as important as the previous features. However, what is surprising is the domain feature, which appeared only in one of the extracted models. One would expect that domain scores should matter as they specifically target the knowledge of the academic discipline. This finding suggests that administrators, industry partners, and other officials who look at these test results do not put much value in the domain attribute instead focus more on coding and aptitude scores.

The random forest model, which garnered the highest accuracy and *F* measure, retained all features except quantitative and domain scores. The above result also confirmed and validate the proposal of Samantha and Poojah (2020), that job placement exams can be utilized to predict job placement.

Conclusion and Recommendation

It is the responsibility of every educational institution to prepare students to enter the professional world. Job placement exams are tools that can assess students' readiness. The study demonstrated that machine learning algorithms could be successfully appropriated to predict job placement based on job placement exam results. In the study, random forest stood out from the other models by garnering an accuracy of 90.85% and an even higher F measure of 91.59%. The dataset feature that exhibited the most predictive value is coding, followed by the aptitude score.

The research output will be advantageous to school administrators as it will give them an insight into which exam domain is the most critical, thus instituting changes that will enhance students' skills in those areas. For example, academic administrators can include the summative report of the predictions as inputs for curriculum development. An example of this will be to include soft skills such as communicating in English as one of the learning objectives if it has been shown that students are weak in this area. Academic counselors can use the model to see which areas a student needs to improve, thus allowing them to give better advice. Counselors can add the model's result as a part of their intervention report, providing a concrete basis for developing the intervention plan. Educators will be given insights to develop instructional activities that will target specific students' weaknesses. Teachers can use the model result to focus more on specific areas. For example, teachers can give more practical exercises in programming to improve the student's scores in the coding part. The students will also significantly gain from the model as it will highlight their weaknesses and allow them to remedy the situation before graduation.

Future research recommendations would be to conduct similar research on other disciplines. Furthermore, a subsequent research step from the results of this study is to develop a remediation program for students who were deemed weak in essential areas. The remediation program results can be tested, and its effectiveness measured and later be also utilized to develop a predictive model.

References

Akar, Ö., & Güngör, O. (2012). *Classification of multispectral images using Random Forest algorithm*. Journal of Geodesy and Geoinformation, 1(2), 105-112.

Bambrick, N. (n.d.). *Support Vector Machines: A Simple Explanation*. KDnuggets. https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html

Burns, E., & Brush, K. (2021, March 29). *deep learning*. SearchEnterpriseAI. https://www.techtarget.com/searchenterpriseai/definition/deep-learning-deep-neural-network

Charbuty, B., & Abdulazeez, A. (2021). *Classification based on decision tree algorithm for machine learning*. Journal of Applied Science and Technology Trends, 2(01), 20-28.

Gupta, P. (2018). *Cross-Validation in Machine Learning - Towards Data Science*. Medium. https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f

Gupta, S. (2022). *Data science vs data analytics vs. machine learning vs. artificial intelligence*. (2022, February 25). Springboard Blog. https://www.springboard.com/blog/data-science/data-science-vs-data-analytics-vs-machine-learning-vs-artificial-intelligence/

Harikrishnan, H. N. (2021). *Confusion Matrix, Accuracy, Precision, Recall, F1 Score*. Medium. https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd

Harihar, D., & Bhalke, D. (2020). *Student Placement Prediction System using Machine Learning*. SAMRIDDHI, 12(2), 85–91.

Kaushik, S. (2020, October 19). *Feature Selection Methods | Machine Learning*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). *Logistic regression* (p. 536). New York: Springer-Verlag.

Manvithha, P., & Swaroopa, N. (2019*). Campus Placement Prediction Using Supervised Machine Learning Techniques*. International Journal of Applied Engineering Research, 14(9), 2188–2191.

Marcot, B. G., & Hanea, A. M. (2020). What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis*? Computational Statistics, 36(3), 2009–2031. https://doi.org/10.1007/s00180-020-00999-9*

Maurya, L. S., Hussain, M. S., & Singh, S. (2021). *Developing Classifiers through Machine Learning Algorithms for Student Placement Prediction Based on Academic Performance*. Applied Artificial Intelligence, 35(6), 403–420. https://doi.org/10.1080/08839514.2021.1901032

Moitra, S. (2022, February 14). *Campus Placement*. Kaggle. Retrieved February 23, 2022, from https://www.kaggle.com/sabyasachimoitra/campus-placement

Narkhede, S. (2021, June 15). *Understanding Confusion Matrix - Towards Data Science*. Medium. https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Noble, W. S. (2006). What is a support vector machine?. *Nature biotechnology*, *24*(12), 1565-1567.

Pereira, M., & da Silva, M. (2003). A Key Question for Higher Education: Who are the customers? *Proceedings of the 31st Annual Conference of the Production and Operations Management Society, POM-2003, April, 4–7, 2003, Atlanta, G*. https://pomsmeetings.org/ConfProceedings/001/Papers/SOM-19.3.pdf

Runkler, T. A. (2020). Data analytics. Springer Fachmedien Wiesbaden

Shukla, M., & Malviya, A. K. (2019). *Modified Classification and Prediction Model for Improving Accuracy of Student Placement Prediction*. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3351006

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert systems with applications*, *134*, 93-101.

Stedman, C., & Hughes, A. (2021, September 7). *Data mining*. SearchBusinessAnalytics. https://searchbusinessanalytics.techtarget.com/definition/data-mining#:%7E:text=Data%20mining%20techniques%20and%20tools,useful%20information%20in%20data%20sets.

Thomas, E. H., & Galambos, N. (2004). What satisfies students? Mining student-opinion data with regression and decision tree analysis. *Research in Higher Education*, *45*(3), 251-269.

Trevor Yu, K. Y. T., & Cable, D. M. (2013). *The Oxford Handbook of Recruitment*. Oxford University Press.

What Is Data Analytics? - 4 Ways to Use Data Analytics. (2021, December 14). Lotame. https://www.lotame.com/what-is-data-analytics/

Yang, P., Hwa Yang, Y., B Zhou, B., & Y Zomaya, A. (2010). *A review of ensemble methods in bioinformatics*. Current Bioinformatics, 5(4), 296-308.

Zumel, N., & Mount, J. (2019). *Practical Data Science with R* (2nd ed.). Manning.