Predictive Analytics Model of an Engineering and Technology Campus Placement

Conference Paper · April 2022

DOI: 10.1007/978-981-19-0863-7_21

Sachin Bhimrao Bhoite
MIT World Peace University, Pune

Anuradha Kanade
MITWPU

Punam Nikam
Dr. Vishwanath MIT World Peace University

Predictive Analytics Model of an Engineering and Technology Campus Placement

Dr. Sachin Bhoite[1], Dr. Anuradha Kanade[2], Punam Nikam[3],
Deepali Sonawane[4]

School of Computer Science, MIT WPU, Pune, Maharashtra,
India[1, 2.3.4]
Email: sachin.bhoite@mitwpu.edu.in

Abstract. Each student dreams to have the placement offer letter in their hand before concluding their final year of Engineering and Technology. Even the reputation of every Institute depends on the placement that they provide to their students. In this research Machine Learning (ML) techniques are applied on the dataset of previously placed students to predict the placement of upcoming batch students. In this work, we followed the Cross-Industry Standard Process (CRISP) methodology with the help of ML model building processes such as Feature selection, Label encoding, Feature scaling, Normalization and Standardization. We have selected the ML models for the prediction of the placement by experimenting and comparing suit of ML classification algorithms using the K-fold cross validation method and Ensemble Learning (EL) ML method. The suit of ML algorithms covers Logistic Regression, K-Nearest Neighbors', Decision Tree Classifier, Random Forest Classifier, Naive Bayes & Support Vector Machine classifiers. Under EL, we have tested Adaptive Boosting, Extreme Gradient Boosting (XGBoost) and Grid Search CV methods. EL methods are advanced and popular and hence it gives the best performance on a predictive modeling project. The performance XGBoost algorithm is best in predicting students' placement in the early stage than that of different algorithms employed in the study with the support of relevant input features. In the end, researchers have suggested "A free guide to notify the campus placement status (FGNCPS)" web module through which placement aspirant students will get to know the placement status in advance and as per prediction, unplaced students get time to improve their weaker areas.

Keywords: Cross-Industry Standard Process, K-fold cross validation, Machine learning, Ensemble Learning.

1 Introduction

According to the requirement of company, colleges must update their curriculum & provide necessary technical & practical knowledge to the students. It will help in fulfilling the requirement of skilled & qualified students of the industries. DM and machine learning (ML) scholars have studied classification problems most recurrently [5]. In which the value of a dependent variable can be predicted based on the values of other independent variables [6]. This paper aims to determine the features impacting on prediction of placement and also students will get to know the placement status and get help in improving their weaker areas in advance.

Basically, this model will help to make training and placement officers (TPO) work easy and increment the total number of placements. Hence, it will directly lead to increment in the rank of engineering and technology institutions. As our objective is to predict the placement of a student, in such a way that either he will get placement or not. It is a binary classification problem. To get good accuracy with minimum error, we have experimented with various classification ML algorithms with K-fold cross validation techniques and train and test the data splitting techniques. Value of K is tested for the better results though most of the time it has considered as 10. Also, we used EL techniques, which are comparatively faster and give better accuracy for classification projects.

2 Related work

The researchers have studied several connected national & international research papers, thesis to understand datasets, data pre-processing methods, features selection methods, type of algorithms used in the existing studies.

Authors in [1] performed a step-wise analysis based on specific statistical frameworks for the placement. The analysis concluded with student datasets including academic and selection subtleties are important for forecasting future selection possibilities. Authors in [2] proposed the campus placement prediction work using the classification algorithms Decision Tree and Random forest. The accuracy obtained after analysis for Random Forest is greater than the Decision tree. Authors in [3] used different ML algorithms to analyze students' admission preferences. They found Random forest classifier is a good classifier as its accuracy is very high. Authors in [4] used different ML models to analyze students' placement, they found AdaBoost classifier along with the Bagging and Decision Tree as Base Classifier gives high accuracy. The student placement analyzer recommendation system, built using classification rules-Naïve Bayes, Fuzzy C Means techniques, to predict the placement status of the student to one of the five categories, viz., Dream Company, Core Company, Mass Recruiters, Not Eligible and Not Interested in Placement. This model helps weaker students and provides extra care towards improving their performance henceforth [7]. Authors in [8] presented student career prediction using advanced ML techniques. In this paper Advanced ML algorithms like SVM, Random Forest decision tree, OneHot encoding, XG boost are used. Out of all SVM gave more accuracy with 90.3 percent and then the XG Boost with 88.33 percent accuracy.

Authors in [9] presented student placement and skill ranking predictors for programming classes using class attitude, psychological scales, and code metrics. They used Support Vector Machine with RBF Kernel (SVM), Support Vector Machine with Linear Kernel (SVML), Logistic regression (LR), Decision tree (DT), Random forest (RF) techniques. ML is used to predict placement results and the programming skill level. The researcher created a classification model with precision, recall, and F-measure.

Authors in [10] presented the study on educational data mining for student placement prediction using ML algorithms. ML algorithms are applied in the weka tool and R studio which are J48, Naïve Bayes, Random Forest, Random Tree, Multiple Linear Regression, binomial logistic regression, Recursive Partitioning, Regression Tree, conditional inference tree, Neural Network. In the weka tool random forest and random tree algorithms are giving 100 % accuracy on student placement dataset. Authors in [11] presented a survey on placement prediction system using ML. The author has suggested Ensemble methods, which is Machine Learning technique that combines several base models in order to produce one optimal predictive model.

3 Research Methodology

The proposed work was carried out by performing experiments on the pass out student's dataset with various ML algorithms.

3.1 Algorithms used

The objective of research needs to use classification methods. Hence, researchers have used following ML classification algorithms.

1. Logistic Regression
2. K-Nearest Neighbours
3. Decision Tree
4. Random Forest
5. Support Vector Machine
6. Naive Bays

Also, used following advanced EL algorithms.

7. Adaptive Boosting,
8. Extreme Gradient Boosting (XGBoost) and
9. Grid Search CV

4 Steps in Building Predictive Models Using ML

We followed the CRoss-Industry Standard Process (CRISP) methodology.

Understanding of problem and objectives of the research: Understanding dataset of already placed students and selection of the appropriate features for placement prediction.

Data Understanding: Data of already placed students were collected. All the attributes of the dataset were analyzed based on their importance and relevance based on the placement prediction. The point 5. About the Dataset of this topic explain details about the dataset.

Feature Engineering: In this phase, the data from multiple data sources were integrated into a one datasets. The next step is that the data was cleaned by removing unwanted columns, handling missing values, creating unique classes, performing transformation for numerical data, and all the cleaning activities on the data. The point 6. Feature engineering of this topic explain details about the same.

Experimenting: Number of ML algorithms were tested and experimented with parameter tuning mentioned in table 2 & table 3 to predict the college and its results are discussed in the point 9, Result and discussion.

Evaluation: Models developed were evaluated based on their performance for accuracy metric [13]. More information presented in the point 8.

Result & Discussion: Result and Discussion are discussed in point 9.

Implementation: Once the model evaluated it is used to evaluate on an unseen data. Which is discussed in point 10.

5 About the Dataset

Researchers have collected 16 engineering colleges' 9766 data records. No of columns in the dataset, per college were varied from 20 to 46. We merged the dataset by considering common and important column from our objective point of view in the excel file format and then converted it into a CSV file. Which is essential to read by the python code to implement ML algorithms.

6 Feature Engineering

In general, every ML algorithm takes some input data to generate desired outputs. This input data are called as features, which are usually presented in the structured columns. As per goal or objective algorithms require input features with some specific characteristic to get the desired output. Hence, there is a need of feature engineering. Feature engineering efforts mainly have two goals:

1. Generating the proper input dataset, as per the requirement of ML algorithm.
2. Improving the performance of ML models.

As per experience of researcher, we need to spend more than 70% of time on data preparation. The following steps are carried out to achieve the same.

1. Missing Values
2. Handling categorical data (Label Encoder)
3. Change data type
4. Drop columns

7 Feature Selection

Every time domain experts may not be available to decide independent features to predict the category of the target feature. Hence, before fitting model we must make sure that all the features that we have selected are contributing to the model properly and weights assigned to it are good enough so that our model gives satisfactory accuracy. For that, we have used 3 feature selection techniques: Univariate Selection, Recursive Features Importance, & Feature importance. We used python scikit-learn library to implement it.

Univariate Selection method shows highest score for following features.

Feature Name	Feature Score	Feature Name	Feature Score
Sem_IV_Aggregate Marks	311.517737	Sem_VI_Pending Back Papers	22.920021
Aggregate Present Marks	223.533006	Sem_V_Pending Back Papers	20.066178
Sem_III_Aggregate Marks	198.255768	Sem_III_Back Papers	16.854853
Sem_VI_Aggregate Marks	151.002450	Back Papers	12.203902
Sem_V_Aggregate Marks	147.823502	Pending Back Papers	7.822886
Sem_II_Aggregate Marks	142.692722	Sem_I_Back Papers	5.458014
Sem_I_ Aggregate Marks	138.860021	Sem_II_Back Papers	2.265504
College Name	55.624319	Sem_II_Pending Back Papers	0.701740
Sem_VI_Back Papers	53.893101	Sem_IV_Pending Back Papers	0.558951
SSC Aggregate Marks	42.917186	Sem_III_Pending Back Papers	0.200665

Defence Type	27.490582	Sem_I_Pending Back Papers	0.124713
Category	27.385665	Gender	12.518480
12th/Diploma marks	26.351629	Branch	0.103342

Table 1 Univariate Feature Selection for Placement Prediction

While using Recursive Feature Importance method following features are selected and remaining are rejected.
Selected Features: ['Pending Back Papers', 'Sem_III_Pending Back Papers', 'Sem_IV_Aggregate Marks', 'Sem_IV_Pending Back Papers', 'Sem_V_Pending Back Papers', 'Sem_VI_Back Papers']

Inbuilt class Feature importance comes with Tree Based Classifiers, we used Extra Tree Classifier from python scikit-learn library for extracting the top 7 features of the dataset.

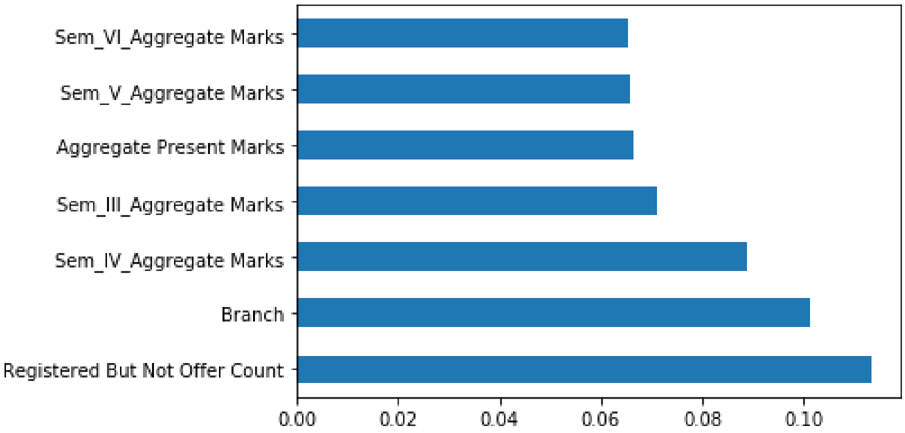

Fig. 1 Feature Selection using Feature Importance for Placement Prediction

Hence, as per all above methods and as per domain knowledge of researcher we have chosen 25 important features which are as follows to predict target feature ‘Job Offer’.
['Branch', 'Aggregate Present Marks', 'Back Papers', 'Pending Back Papers', 'Sem_I_ Aggregate Marks', 'Sem_I_Back Papers', 'Sem_I_Pending Back Papers', 'Sem_II_Aggregate Marks', 'Sem_II_Back Papers', 'Sem_II_Pending Back Papers', 'Sem_III_Aggregate Marks', 'Sem_III_Back Papers', 'Sem_III_Pending Back Papers', 'Sem_IV_Aggregate Marks', 'Sem_IV_Back Papers', 'Sem_IV_Pending Back Papers', 'Sem_V_Aggregate Marks', 'Sem_V_Back Papers', 'Sem_V_Pending Back Papers', 'Sem_VI_Aggregate Marks', 'Sem_VI_Back Papers', 'Sem_VI_Pending Back Papers', '12th/Diploma_Aggre_marks', 'SSC Aggregate Marks']

8 Experimentation

There are sufficient enough models are studied for each objective with K-fold Cross validation (K-FCV), Train-Test split (T-TS) method and tuning different parameters. In this process python sklearn library has played a very important role. So detail is mentioned in the table below.

Sr. No.	Name of Algorithm	Data splitting method used	Data splitting folds/ratio			Parameter tuned	No. of parameter Tested
1	Logistic Regression	K-FCV	3	5	10	label encoding	6 to 10
						one hot encoding	6 to 10
		T-TS	70:30	80:20	90:10	label encoding	6 to 10
						one hot encoding	6 to 10
2	Support Vector Machine (SVC)	K-FCV	3	5	10	estimator	6 to 10
						param_grid	6 to 10
		T-TS	70:30	80:20	90:10	estimator	6 to 10
						param_grid	6 to 10
3	Decision Tree	K-FCV	3	5	10	max_depth	6 to 10
						min_impurity_decrease	6 to 10
						max_leaf_nodes	6 to 10
						min_leaf_nodes	6 to 10
						max_feature s	6 to 10
		T-TS	70:30	80:20	90:10	max_depth	6 to 10
						min_impurity_decrease	6 to 10
						max_leaf_nodes	6 to 10
						min_leaf_nodes	6 to 10
						max_feature s	6 to 10
4		K-FCV	3	5	10	max_depth	6 to 10

	Random Forest					min_impurity_decrease	6 to 10
						max_leaf_nodes	6 to 10
						min_leaf_nodes	6 to 10
						max_feature s	6 to 10
		T-TS	70:30	80:20	90:10	max_depth	6 to 10
						min_impurity_decrease	6 to 10
						max_leaf_nodes	6 to 10
						min_leaf_nodes	6 to 10
						max_features	6 to 10
5	Gaussian NB	K-FCV	3	5	10		6 to 10
		T-TS	70:30	80:20	90:10		6 to 10
6	K Neighbors Classifier	K-FCV	3	5	10	leaf_size	6 to 10
						n_neighbors	6 to 10
		T-TS	70:30	80:20	90:10	leaf_size	6 to 10
						n_neighbors	6 to 10

Table 2 List of experiments with model combinations

Apart from above methods while doing parameter tuning we have used following ensemble algorithms.

Sr. No.	Name of the Algorithm	Data splitting method used
1	Ada Boost Classifier (DT)	T-TS
2	Extreme Gradient Boosting (XGBoost) Classifier	T-TS
3	Grid Search CV	T-TS

Table 3 List of experiments with advanced algorithms

After the discussion of the accuracy results researcher has suggested web module named “Free guide to notify the campus placement status (FGNCPS)” through which students will get to know their placement status in advance & also come to know to work more on weaker areas.

9. Result & Discussion

After cleaning all the data, removing all the noise, selecting relevant features and encoded it into ML form, the next step is building a predictive model by applying various ML techniques to find out the best model which gives us more accuracy for train and test both.

Sr. No.	Name of Algorithm	Train Accuracy	Test Accuracy
1	Logistic Regression	0.7251336898395722	0.7371794871794872
2	Support Vector Machine	0.7235294117647059	0.7393162393162394
3	Decision Tree Classifier	0.8165775401069518	0.782051282051282
4	Random Forest Classifier	0.823663101604278	0.7162393162393162
5	Gaussian NB	0.5294117647058824	0.49145299145299143
6	K Neighbors Classifier	0.8235294117647058	0.7606837606837606

Table 4 Results of placement prediction using ML techniques with K-fold cross validation.

Table 5 Results of placement prediction using Ensemble Learning

Sr. No.	Algorithm	Train Accuracy	Test Accuracy
1	AdaBoostClassifier(DT)	0.85	0.82
2	XGBoost	0.88	0.84
3	GridSearchCV	0.851336898395	0.82478632478632

Model selection for placement prediction: After implementing all above methods mentioned in the Table 4 and 5 we found XGBoost classifier is the best classifier to predict the campus placement.

We can see the result of placement prediction using ensemble classifier XGBoost with 0.88 training accuracy and with 0.84 testing accuracy. Which is comparatively very high. Hence we have chosen XGBoost classifier to implement the model.

10 Implementation

10.1 A free guide to notify the campus placement status (FGNCPS)

While predicting the campus placement of Engineering and Technology students, we have proposed following FGNCPS web module. The aspirant student has to submit some basic information which is nothing but selected input features to predict their placement status in the early stage of academics.

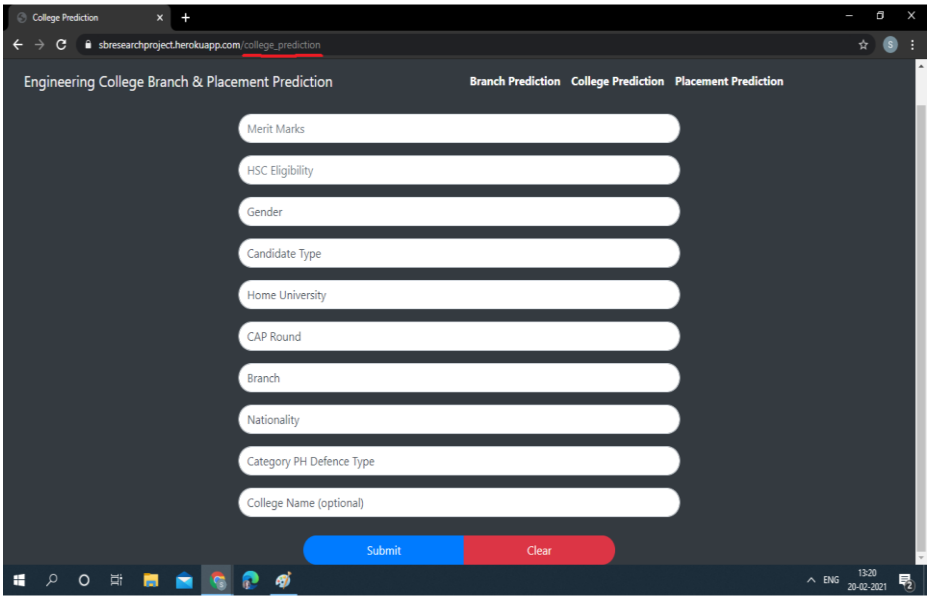

Fig. 2 Placement prediction web module

11 Conclusion

In this research, to predict campus placement of Engineering and Technology students, all the ML model building steps are rigorously implemented on the dataset. Python, various libraries played a vital role during whole this process. In this study, 25 input features are selected out of existing 46 features of the dataset. These features are very important, according to Univariate Selection, Recursive Features Importance, Lasso feature selection methods and researchers' domain knowledge. To predict the campus placement, suit of ML and EL methods are experimented and compared. This suit contains Logistic Regression, K-Nearest Neighbors', Decision Tree Classifier, Random Forest Classifier, Naive Bayes and Support Vector Machine classifiers. Under EL, we have experimented Adaptive Boosting, Gradient Boosting and GridSearchCV methods. After comparison of all algorithm's accuracy, we found that XGBoost classifier has greater accuracy for this project. Also, it has been observed that feature engineering is very important step in model building because after it, results have been more improved. At the end researchers have suggested, "A free guide to notify the campus placement status (FGNCPS)" web module for placement aspirant students.

References

1. Nikhil Kumar, Ajay Shanker Singh, Thirunavukkarasu K., E. Rajesh : ”Campus

Placement Predictive Analysis using Machine Learning", (2020) 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), ISBN: 978-1-7281-8337-4/20/$31.00 ©2020 IEEE

2. Pothuganti Manvitha, Neelam Swaroopa : "Campus Placement Prediction Using Supervised Machine Learning Techniques", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 9 (2019) pp. 2188-2191

3. Dhruvesh Kalathiya, Rashmi Padalkar, Rushabh Shah, Sachin Bhoite : Engineering College Admission Preferences Based on Student Performance, International Journal of Computer Applications Technology and Research, Volume 8–Issue 09, 379-384, (2019), ISSN:-2319–8656

4. Shubham Khandale, Sachin Bhoite : "Campus Placement Analyzer: Using Supervised Machine Learning Algorithms", International Journal of Computer Applications Technology and Research, Volume 8–Issue 09, 379-384, (2019), ISSN:-2319–8656, 358 – 362

5. Kabakchieva, D., Stefanova, K., Kisimov, V.: Analyzing University Data for Determining Student Profiles and Predicting Performance, 4th International Conference on Educational Data Mining (EDM 2011) (2011) The Netherlands, pp.347-348.

6. Min NIE1, Lei YANG1, Jun SUN1, Han SU1, Hu XIA1, Defu LIAN1, Kai YAN : Advanced Forecasting Of Career Choices For College Students Based On Campus Big Data _ Higher Education Press and Springer-Verlag Berlin Heidelberg (2017)

7. Apoorva Rao R1, Deeksha K C 2, Vishal Prajwal R3, Vrushak K4 , Nandini : 'Student placement analyzer: a recommendation system using machine learning', IJARIIE-ISSN(O)-2395-4396, (2018) Vol-4, Issue-3

8. Roy, K. S., Roopkanth, K., Teja, V. U., Bhavana, V., & Priyanka, J.: "Student Career Prediction Using Advanced Machine Learning Techniques". International Journal of Engineering & Technology (2018). 7, 26–29

9. Ishizue, Ryosuke & Sakamoto, Kazunori & Washizaki, Hironori & Fukazawa, Yoshiaki : "Student placement and skill ranking predictors for programming classes using class attitude, psychological scales, and code metrics", Research and Practice in Technology Enhanced Learning. (2018).13. 10.1186/s41039-018-0075-y.

10. Sreenivasa Rao, K. Swapna, N., Praveen Kumar, P.: "Educational data mining for student placement prediction using machine learning algorithms". International Journal of Engineering & Technology, [S.l.], v. 7, n. 1.2, p. 43-46, dec. (2017). ISSN 2227-524X.

11. Bangale, M., Bavane, S., Gunjal, A., Dandhare, R., & Salunkhe, S. D. (2019). A Survey on Placement prediction system using machine learning. IJSART - Volume 5 Issue 2 –FEBRUARY (2019) 5(2), ISSN [ONLINE]: 2395-1052

12. Syed Ahmed1, Aditya Zade2, Shubham Gore3, Prashant Gaikwad4, Mangesh Kolhal5 : 'Smart system for placement prediction using data mining', International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653 (2017) Volume 5