PROJECT REPORT

PROJECT TITLE:

PREDICTING PLACEMENT IN CAMPUS RECRUITMENT

By: ADARSH ASHWINI SHUKLA

INTRODUCTION:

Placement of students is one of the most important objective of an educational institution. Reputation and yearly admissions of an institution invariably depend on the placements it provides it students with. That is why all the institutions, arduously, strive to strengthen their placement department so as to improve their institution on a whole. Any assistance in this particular area will have a positive impact on an institution's ability to place its students.

This will always be helpful to both the students, as well as the institution. In this study, the objective is to analyse previous year's student's data and use it to predict the placement chance of the current students. This model is proposed with an algorithm to predict the same. Data pertaining to the study were collected form the same institution for which the placement prediction is done, and also suitable data pre-processing methods were applied.

This proposed model is also compared with other traditional classification algorithms such as Decision tree and Random forest with respect to accuracy, precision and recall. From the results obtained it is found that the proposed algorithm performs significantly better in comparison with the other algorithms mentioned.

UNDERSTAND AND DEFINE PROBLEM:

Placements are considered to be very important for each and every college. The basic success of the college is measured by the campus placement of the students. Every student takes admission to the colleges by seeing the percentage of placements in the college. Hence, in this regard the approach is about the prediction and analyses for the placement necessity in the colleges that helps to build the colleges as well as students to improve their placements.

In Placement Prediction system predicts the probability of a undergrad students getting placed in a company by applying classification algorithms such as Decision tree and Random forest. The main objective of this model is to predict whether the student he/she gets placed or not in campus recruitment. For this the data consider is the academic history of student like overall percentage, backlogs, credits. The algorithms are applied on the previous year's data of the students.

DATASET PREPARATION AND PREPROCESSING:

In this stage of project implementation, focus is put on data collection, data selection, data preprocessing, and data transformation.

DATA COLLECTION:

The sample dataset is been already provided by Cognifront and it is in CSV format. In this dataset there are total **15** features and **215** data-points. This dataset contains student's data that has **education information** of each student starting from school to degree completion. So we have student educational info like ssc marks, hsc marks, degree percentage, mba percentage etc.

DATA VISUALIZATION:

The amount of data used in ML projects is large in size. When the large data is plotted i.e. visualized, it makes it easy to understand and analyze

We have used different functions that are available in Seaborn and Matplotlib, such as countplot, boxplot, etc.

LABELING:

Regression and Classification type of prediction is done using supervised machine learning technique. In this technique, the data points are labeled i.e. target values of the data points are known. If the data is not labeled, it needs to be done which takes lot of efforts and time. Many a times, the labeling work is outsourced i.e. given to the outside agency

DATA SELECTION:

All the collected data may not be useful. You have to select the subset of the data which is relevant and important for the project in hand.

DATA PREPROCESSING:

The purpose of preprocessing is to convert raw data into a form that is useful in training and testing the ML model. The structured and clean data produces more precise results. In short, good quality data when fed to the ML model, it produces better results. The Preprocessing technique includes data formatting, cleaning, and sampling techniques.

DATA FORMATTING: The data may come from different sources. Hence, it needs to be standardized.

DATA CLEANING: In this procedure, the noise in the data is removed and inconsistencies are fixed. The missing values in the data are filled with mean and median attributes. The outliers in the data are either removed or corrected.

DATA TRANSFORMATION:

In this stage, the data is transformed into the form which is appropriate for machine learning. The scaling and normalization is usually used to transform the data.

SCALING: The different attributes in the dataset may have different ranges i.e. data values may vary over different values. Scaling is used to correct this problem.

FEATURE EXTRACTION: Some of the existing features are combined to create new features which are useful for ML modeling.

DATASET SPLITTING:

The given dataset is split into three parts: training, testing, and validation sets. The ration of training and testing sets is typically 80 to 20 percent. The 20 percent of the training set is further split as a validation set.

MODEL TRAINING

In this stage, the training data is fed to the ML algorithm to build and train a model. The purpose of training is to develop a model.

MODEL TESTING AND EVALUATION

The goal of this step is to develop the simplest, reliable and efficient model. This requires model tuning. Depending on your project, you may use a number of algorithms to test and ultimately select the best model.

IMPROVING PREDICTIONS WITH ENSEMBLE METHODS

In most cases, the data scientists create and train one or more models. Then they select the best performing one. Like Random Forest, data scientists also like to combine (ensemble) various models for prediction. Ensemble methods provide better results.

There are three ways to combine models:

STACKING: In this case, usually used to combine models of different types. The aim of this method is to reduce generalization error.

BAGGING: In this case, the models of the same type are combined in sequential manner. The training dataset is split into subsets. Then the models are trained on each of these subsets. Ultimately, the prediction is based on combining the result using mean or majority voting. The bagging reduces model overfitting.

BOOSTING: In this case, the data scientists use subsets of data to train moderately performing models. The prediction is based on the majority voting principle. Every next model is trained on a subset received from the performance of the previous model (particularly the emphasis is put on misclassified data points).

MODEL DEPLOYMENT

When the reliable model is selected and validated, the model is put into production. Model Deployment means putting the model in use (production).

In most cases, the deployment is done by translating the Model written in Python language to another languages like Java, C, C++, PHP etc. Then the Alpha and Beta testing is done.

There are various ways of deploying the model. The actual deployment depends on the ML Team size and the IT infrastructure available with the company/business

BATCH BASED DEPLOYMENT: In this type, the prediction is done in batch of observations rather than on continuous basis

CONCLUSION:

The campus placement activity is incredibly a lot of vital as institution point of view as well as student point of view. In this regard to improve the student's performance, a work has been analyzed and predicted using the classification algorithms Linier Regression, Logistic Regression, Decision Tree, Random forest, SVM, KNN and K-means algorithm to validate the approaches. The algorithms are applied on the data set and attributes used to build the model.

The accuracy obtained after analysis for are:

1. By Linear regression: 0.5211346131181207

2. By Logistic regression: 0.8837209302325582

3. By Decision Tree: 0.9069767441860465

4. By Random forest: 0.9302325581395349

5. By SVM algorithm: 0.8837209302325582

6. By KNN algorithm: 0.5211346131181207

Hence, from the above said analysis and prediction its better if the **Random Forest algorithm** is used to predict the placement results