A PROJECT REPORT
ON

Predicting Placement in Campus Recruitment

CREATED BY

Suyash Purushouttam Balshetwar

June-2020

Contents

1 Introduction **1**

2 Understand and Define the Problem **2**

3 Dataset Preparation and Preprocessing **4**

3.1 Data Collection 4

3.2 Data Visualization 5

3.3 Labelling 5

3.4 Data Selection 6

3.5 Data Preprocessing 6

3.6 Data Transformation 7

3.7 Dataset Splitting 7

4 Model Training **8**

5 Model Testing and Evaluation **9**

6 Improving Predictions with Ensemble Methods **10**

7 Model Deployment **11**

8 Conclusion **12**

9 Further Development **13**

Chapter 1

Introduction

Placement of the student one of the most important activity in educational institution. 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. Many a times the reputation of such institute is determined by the pay packages offered by recruiters to its students.

Students studying in final year of an Engineering college start feeling the pressure of the placement season with so much of placements activities happening around them. They feel the need to know where they stand and how they can improve their chances of getting placed. 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.

Chapter 2

Understand and Define the Problem

Campus placement chances are important criteria while selecting an educational institution by the student. 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.

So the main objective of this system is to Build a model that can be used to predict the probability that a randomly chosen student will be placed or not, to Identity the factors that are influencing the placement chances of a student and to study the nature of campus placements which is useful for both Students and Institution.

In educational field it is to increase learning process such as identifying, evaluating variables, extracting data set from the learning process. The campus placement of the students plays an important role in an educational institution. Prediction system could help in the academic planning of an institution.

Source of data is kaggel, data.gov etc Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

Tools required for system is jupyter notebook, pandas, numpy, matplotlib, seaborn, sklearn etc pandas: Python data analysis library enhancing analytics and modeling. matplotlib: Python machine learning library for quality visualizations. Jupyter notebook: collaborative work capabilities. Scikit-Learn: It is an open-source machine learning package. It is a unified platform as it is used for multiple purposes. It assists in regression, clustering, classification, dimensionality reduction, and preprocessing.

Chapter 3

Dataset Preparation and Preprocessing

3.1 Data Collection

Data set is collected from various sources such as kaggle, data.gov, UCI Machine Learning Repository etc. The name of dataset is Predicting Placement in Campus Recruitment and format of data set is csv(comma separated value). This data set consists of Placement data. It includes secondary and higher secondary school percentage and specialization. It also includes degree specialization, type and Work experience and salary offers to the placed students. Pandas dataframe is used to read the data set by using read_csv() function. Data set contain 215 rows and 14 columns.
This data set consists columns are as follow.
sscb- Secondary School Certificate Board.
hscb-Higher Secondary Certificate Board.
sl_no : serial number.
gender : male, female
ssc_p : 10th class percentage.
ssc_b : which board to passed out 10th.
hsc_p : 12th class percentage.

hsc_b : which board to passed out 12th.
hsc_s : which stream he choose (science,commerce,arts)
degree_p : Bachelor degree percentage.
degree_t : which stream choose for bachelor.
workex : It has a work experience or not.
etest_p : entrance test percentage.
specialisation : Master degree in Mkt&HR or Mkt&Fin.
mba_p : Master degree percentage.
status : He/She got placed or not in campus placement.
salary : placement packages.

3.2 Data Visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. In this project various data visualization technique is used such as barplot, boxplot, countplot and violinplot etc

3.3 Labelling

In this project all supervised machine learning algorithms are used so data set should be labeled Labeled data is a group of samples that have been tagged with one or more labels. After obtaining a labeled dataset, machine learning models can be applied to the data so that new unlabeled data can be presented to the model and a likely label can be guessed or predicted for that piece of unlabeled data. Target value of data set is 'status' and remaining are featured value.

3.4 Data Selection

In our project all column are selected for processing the data.

3.5 Data Preprocessing

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set. Their are various methods are applied in this step such as removing null values from data set and replace it with median, convert categorical column into numerical column using label encoder method.

3.6 Data Transformation

Data transformation is the process in which you take data from its raw, siloed and normalized source state and transform it into data that's joined together, dimensionally modeled, de-normalized, and ready for analysis. Standardization is a useful technique to transform attributes with a Gaussian distribution and differing means and standard deviations to a standard Gaussian distribution with a mean of 0 and a standard deviation of 1. We can standardize data using scikit-learn with the StandardScaler class.

3.7 Dataset Splitting

Separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. In our project 80 percent of data is trained and 20 percent of data is used for testing data. Analysis Services randomly samples the data to help ensure that the testing and training sets are similar. Data set is split by using train_test_split method which is provided by sklearn.model_selection library.

Chapter 4

Model Training

Machine learning works by finding a relationship between a label and its features. We do this by showing an object (our model) a bunch of examples from our dataset. Each example helps define how each feature affects the label. We refer to this process as training our model. In this project various machine leaning model are developed such as Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors and Support vector Machine by using sklearn library.

Chapter 5

Model Testing and Evaluation

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. Evaluation of model is calculated by using confusion matrix, Classification report and accuracy score. This technique is predefined in sklearn library. The accuracy of models is listed below.

1)Logistic Regression : 81.39%

2)Decision Tree: 93.02%

3)Random Forest : 95.34%

4)K-Nearest Neighbors : 81.39%

5)Support Vector Machine : 76.74%

Among the all model Random Forest model gives best accuracy.

Chapter 6

Improving Predictions with Ensemble Methods

In statistics and machine learning, ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. In my project there is no ensemble method is used.

Chapter 7

Model Deployment

The deployment of machine learning models is the process for making your models available in production environments, where they can provide predictions to other software systems. In this project model is deployed on new data. Deploy the model by giving the same inputs to all the models gender-1,ssc_p-98.00,ssc_b-0,hsc_p-93.34,hsc_b-0,hsc_s-1,degree_p-81.00,degree_t-1,workex-0,etest_p-77.0,specialisation-1,mba_p-98.1,salary-370000. The accuracy of all model is 1.

Chapter 8

Conclusion

Predicting the placement of a student gives an idea to the Placement Office as well as the student on where they stand. Not all companies look for similar talents. If the strengths and weaknesses of the students are identified it would benefit the student in getting placed. The placement Office can work on identifying the weaknesses of the students and take measures of improvement so that the students can overcome the weakness and perform to the best of their abilities.

In this regard to improve the student's performance, a work has been analyzed and predicted using the classification algorithms such as Logistic Regression, Decision Tree, Random forest, k-Nearest Neighbors and Support vector Machine algorithm to validate the approaches. The algorithms are applied on the data set and attributes used to build the model. From the analysis and prediction it's better if the Random Forest algorithm is used to predict the placement results.

Chapter 9

Further Development

Further work can be carried out by applying other algorithms that could lead to improvement in results, also increase the length of data set and add more columns(features) in dataset such as skill set of students etc. We can use ensemble methods for better performance of model.