DATA SCIENCE PROJECT

REPORT

By- Aditya Deo

**ABSTRACT :**

Using the dataset provided, we built a classification and visualization model for the same. Our classification model, uses SVM model which predicts the F1 score : 94%. Visualization model generates all the pdf of the graphs and relationship of the given question.

### **INTRODUCTION**

#### **Background:**

Cloud Counselage is the best company which helps the fresher in grooming their talents and get a deserving job in IT industry. Cloud Counselage believe in working ground-up to train and groom the youth of India into IT & Management, building capacities in terms of an online workforce PAN India to deliver projects globally. The biggest project they have delivered was for a giant MNC. It was a mix of IT & Management processes and involved working with 156 client offices across 120 countries worldwide. They generate their workforce PAN India by launching campaigns for their 'Online Career Programs' into IT & Management. They have launched an Online Internship Program for the college students to groom their skills and get ready for overcoming outside world obstacles and for this they require the deserving and eligible candidates for this program. Therefore, it is advantageous for them to accurately predict whether the applicant is deserving or not.

#### **Objectives :**

Obviously, Cloud Counselage would be very interested in accurate prediction of the well deserving and eligible candidates among all applicants, for helping the candidates who deserve for this opportunity.

**PROBLEM STATEMENT:**

Students from different cities from the state of Maharashtra had applied for the Cloud Counselage Internship Program. They have the dataset of consisting information of all the students. Using this data we want to get more insights and draw out more meaningful conclusions.

Following are the tasks interns need to perform :

1. Interns need to preprocess the data for missing values, unknown values, encoding categorical values.

2. Create a data visualization model to build graphs from the dataset answering the following questions:

1. The number of students applied to different technologies.
2. The number of students applied for Data Science who knew ‘’Python” and who didn’t.
3. The different ways students learned about this program.
4. Students who are in the fourth year and have a CGPA greater than 8.0.
5. Students who applied for Digital Marketing with verbal and written communication score greater than 8.
6. Year-wise and area of study wise classification of students.
7. City and college wise classification of students.
8. Plot the relationship between the CGPA and the target variable.
9. Plot the relationship between the Area of Interest and the target variable.
10. Plot the relationship between the year of study, major, and the target variable.

3. Identify the best binary classifier to classify data into “eligible/1” and “not eligible/0”.

**INPUT:**

DATASET provided during runtime.

**OUTPUT:**

1. Classification Model which tells F1 score
2. Visualization Model generates pdf of the all the visualization.

**METHODOLOGY:**

1. Exploratory Data Analysis
2. Data Cleaning
3. Pre-processing
4. Data Visualization
5. Model for Classification of Data
6. Evaluation of the Model

#### **DATA CLEANING:**

The data provided from the company has total 32 variables with 10000 observations. There are constant values in three of the columns such as, State has constant value “Maharashtra”, Course Type has constant value “Full-time”, and Current Employee Status has constant value “Student”.

The columns “City”, “College Name”, “University Name”, and “Degree” are highly correlated with each other and “Expected Graduation year” is highly correlated with “Which Year are you studying in?”.

Many columns were dropped which has null values in all the rows, some columns like , “First Name” , “Last Name” , “Email Address”, “Contact Number” and “Emergency Contact Number” were also dropped as it was least significant in prediction.

**MODELS USED:**

1. **Logistic Regression**

Logistic regression is a statistical method for predicting binary classes. It is one of the simplest and commonly used Machine Learning algorithms for Binary classification. It is easy to implement and can be used as the baseline for any binary classification problem. Logistic regression describes and estimates the relationship between one dependent binary variable and independent variables. The target variable has two possible classes in the dataset for eligibility being 1 Eligible or 0 Ineligible.

1. **Random Forest**

Random Forest, a simpler and powerful approach compared to the other non-linear classification algorithms. This algorithm works by selecting random samples from a given dataset, constructs a decision tree for each sample and get a prediction result from each tree, performs a vote for each predicted result. Select the prediction result with the most votes as the final prediction. Random forest is considered as a highly accurate and robust method because of the number of decision trees participating in the process. It does not suffer from the overfitting problem. The main reason is that it takes the average of all the predictions, which cancels out the biases.

1. **SVM**

SVM is a classification algorithm with a simple concept of finding an optimal hyperplane which helps in classifying new data points. It constructs a hyperplane in multidimensional space to separate different classes, generating an optimal hyperplane in an iterative manner, used to minimize an error. The main objective is to segregate the dataset in best possible way. SVM kernels take low dimensional input space and transforms into high dimensional space.

**RESULTS:**

|  |  |  |
| --- | --- | --- |
| **MODEL** | **Accuracy** | **F1score** |
| Logistic Regression | **0.77** | **0.71** |
| Random Forest | **0.99** | **0.99** |
| SVM | **0.97** | **0.94** |

**CONCLUSION :**

In this study, I analyzed the candidates scores and their skills based on which their eligibility was depended for the Internship. I identified CGPA, Written communication skills and Verbal communication skills are the most important features which affect the selection of the candidate. I built classification models to predict which candidates are eligible/ineligible. These models can be very useful in helping other companies in number of ways. For example, it could help identify the candidates who are more deserving and eligible for the posts available in the company.

Some of the analysis I observed as follows:

1. Most of the students with CGPA above 7.5 are eligible for the Internship.
2. Most of the eligible students are from the second, third and last year.
3. Most of the students applied for the Internship are from the Computer Engineering background.