# **DATA SCIENCE PROJECT**

# JAPESH GAWALE

### **1. Introduction**

#### **1.1 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.

#### **1.2 Problem:**

Cloud Counselage is looking for tech-savvy college-going students who want to make their career in IT. The intern would participate in various technical activities vis-a-vis the latest technologies like Cloud Computing, Artificial Intelligence, Machine Learning, Data Science, Blockchain, etc. This project aims to predict whether the applicant is eligible and well deserving for this program or not.

#### **1.3 Interest:**

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.

### **2. Data acquisition and cleaning**

#### **2.1 Data sources:**

The dataset is provided from the Cloud Counselage Pvt. Ltd. of all the applicants applied for the IP program.

#### **2.2 Data cleaning:**

The data provided from the company has total 29 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”. There is a high cardinality in “First Name” column with 5793 distinct values, “Last Name” columns with 9035 distinct values and “DOB” column with 1819 distinct values.

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?”, “First Name” and “Last Name” columns are uniformly distributed and “Email Address”, “Contact Number” and “Emergency Contact Number” columns have unique values.

There are three columns “link to LinkedIn profile”, “Link to updated Resume (Google/ One Drive link preferred)” and “Certifications/Achievement/ Research papers” which contains null values in all observations, so these three columns are dropped as there is no information which can be used for prediction.

### 3. Exploratory Data Analysis

#### **3.1 The number of students applied for different technologies:**

The total numbers of students applied for this Internship are 10000 students for different technologies and there are total 16 different types of technologies in this program. So, the number of students applied for different technologies is visualized in bar chart shown below.

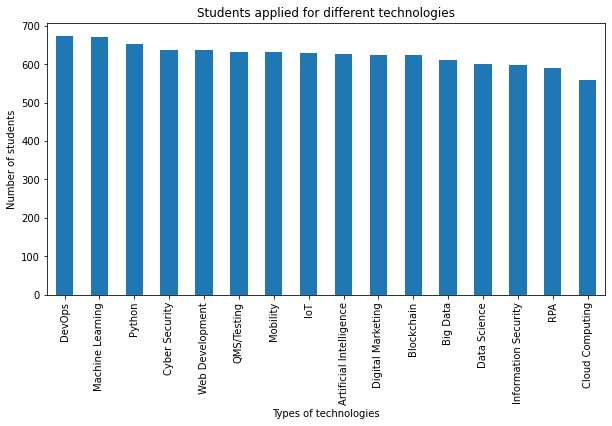


Figure 1. Number of students applied for different technologies.

#### **3.2 The number of students applied for Data Science who knew “Python” and who didn’t:**

The total numbers of students applied for Data Science are 601 among which the students who knew Python are 66 students and who didn’t know are 535 students. The visualization is shown below.

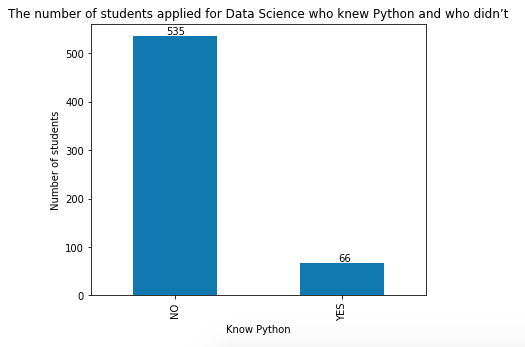


Figure 2. Number of students applied for Data Science who knew Python and who didn’t.

#### **3.3 The different ways students learn about this program:**

There are 9 different ways students can learn about this program. The numbers of students learnt from different type of ways are visualized below.

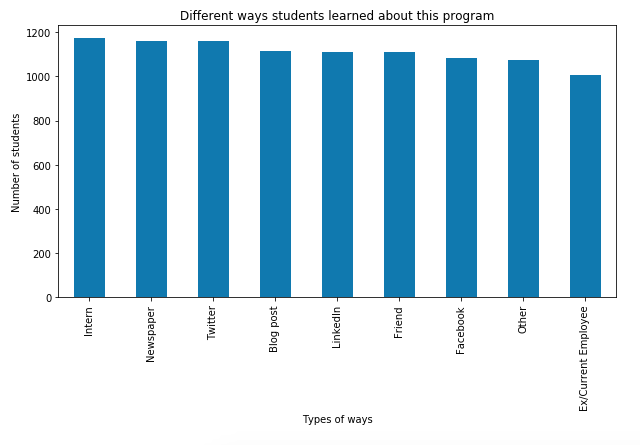


Figure 3. Different ways student can learn about this program.

#### **3.4 Students who are in fourth year and have a CGPA greater than 8.0:**

The total numbers of students who are in fourth year are 2477 and out of which 1697 students have CGPA greater than 8.0. The visualization of this shown below.

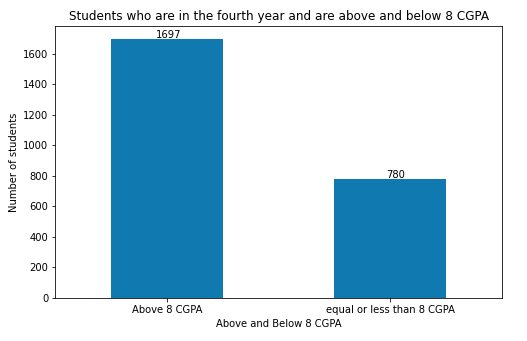


Figure 4. Students who are in fourth year and have CGPA greater than 8.0

#### **3.5 Students who applied for Digital Marketing with verbal and written communication score greater than 8.0:**

The number of students who applied for Digital marketing is 624, out of which the number of students who have verbal and communication score greater than 8.0 are 92. The visualization of this is shown below.

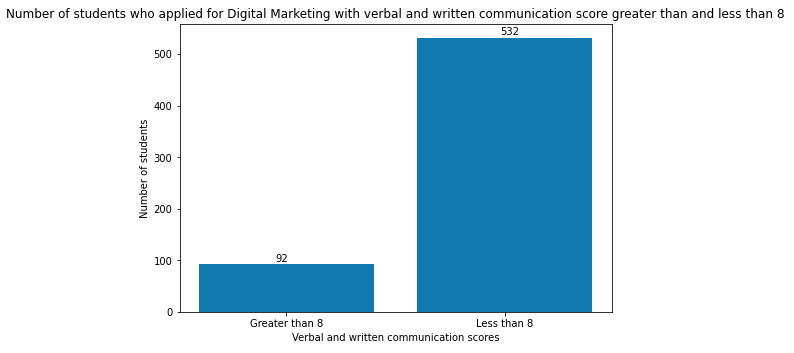


Figure 5. Students applied for Digital marketing with verbal and written communication scores greater than 8.0.

#### **3.6 Year wise and Area of Study wise classification of students:**

There are three area of study in this data, “Computer Engineering”, “Electrical Engineering” and “Electronics and Telecommunication Engineering” with four years of study. The visualization of this is shown below with the number of students applied respectively.

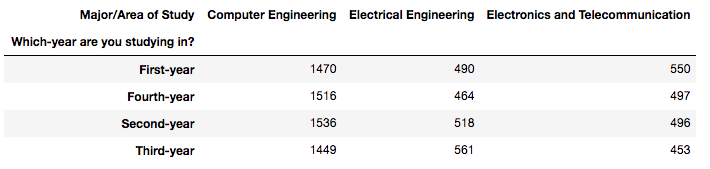


Figure 6. Number of students in different fields.

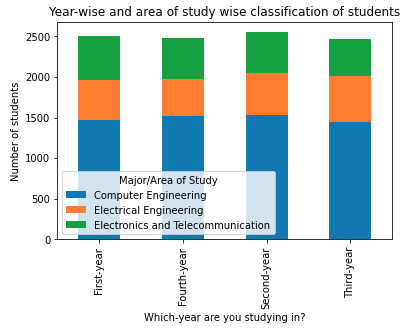


Figure 7. Year-wise and area of study wise classification of students.

#### **3.7 City and college wise classification of students:**

The students applied from six different cities, “Mumbai”, “Navi Mumbai”, “Pune”, “Kolhapur”, “Sangli”, and “Solapur” and from different colleges of this cites. The visualization of this is shown with the number of students applied from these cities with respective to their colleges is shown below.

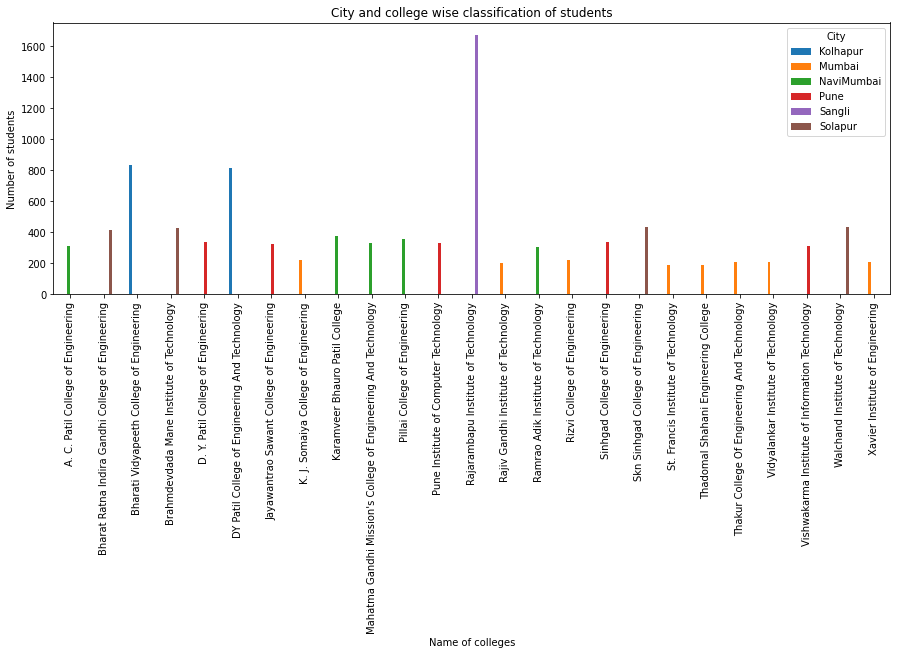


Figure 8. City and college wise classification of students.

#### **3.8 Relationship between the CGPA and target variable:**

The target variable consists of two values of eligible and ineligible students for this Internship and the CGPA is distributed in 6 intervals with 0.5 difference from 7 to 10, so this visualization is of number of students with different CGPA among which how many are eligible and not. The visualization of this is shown below.

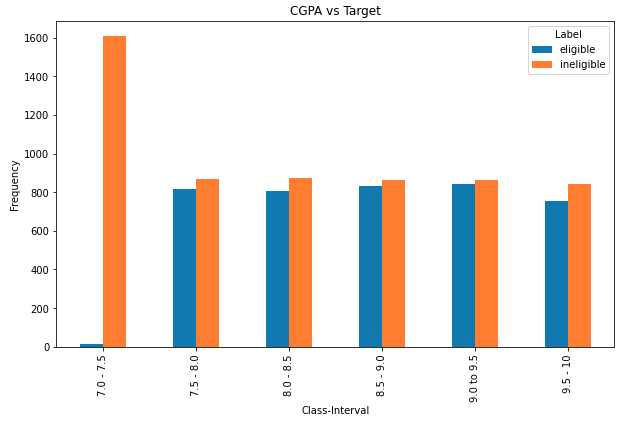


Figure 9. Relationship between CGPA and target variable

#### **3.9 Relationship between area of interest and target variable:**

As discussed before there are 16 different area of interests and in target variable there are two values, so this visualization is of number of students who applied for different technologies and among them how many are eligible and not. The visualization of this is shown below.

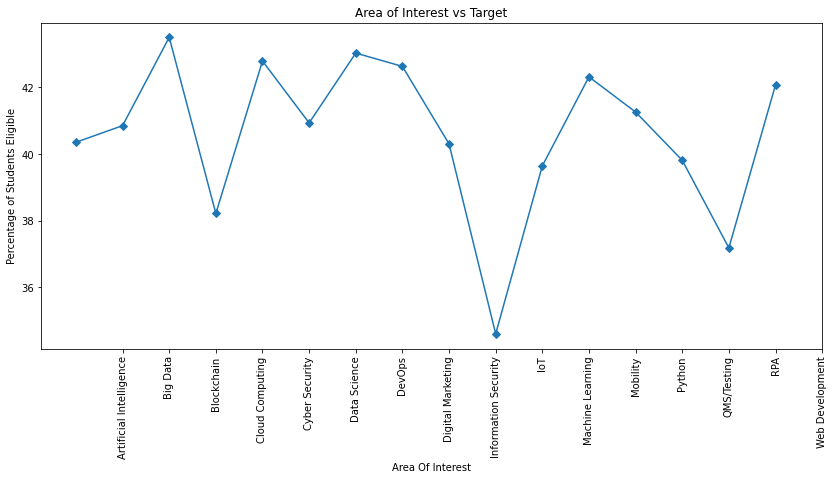


Figure 10. Relationship between area of interest and target variable.

#### **3.10 Relationship between year of study, major and the target variable:**

As discussed before there are three area of study in this data, “Computer Engineering”, “Electrical Engineering” and “Electronics and Telecommunication Engineering” with four years of study, so this visualization is showing the relationship between these two and the target variable as how many students are eligible and not among these. The visualization of this is shown below.

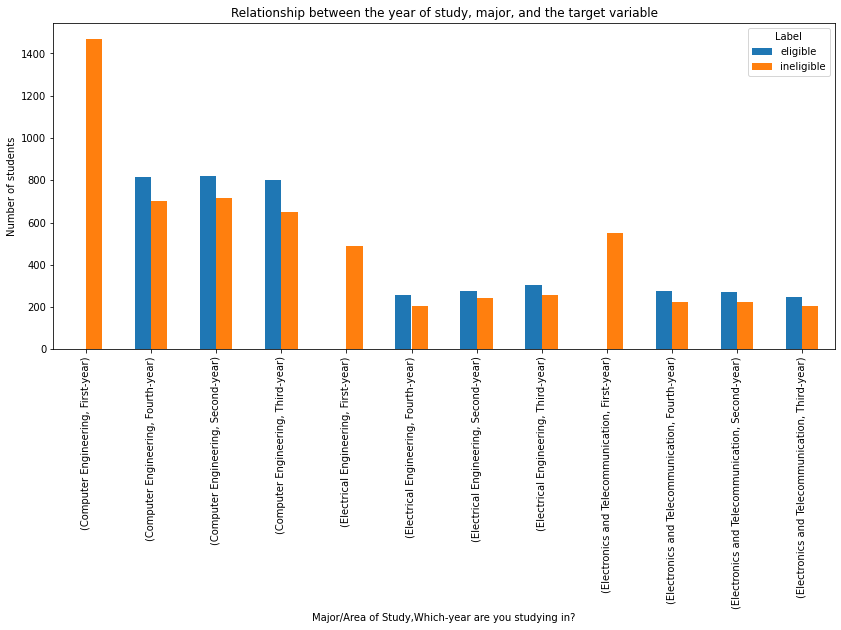


Figure 11. Relationship between year of study, major and the target variable.

### **4. Predictive Modelling**

#### **4.1 Applying Classification Models:**

I applied Decision tree, Random forest, Logistic regression and KNN models to the dataset. In training data, the 12 more correlating columns are used and the train test split is 80:20 after that the scaling of data is done, as scaling of data before the train test split can lead to leakage of information and can make model over learn. The best model are Decision tree and Random forest which gives the accuracy of 100%, but I will be going with Decision tree as Decision tree is simple model and uses less resources than Random forest. The observations of all the models are shown below.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Sr. No. | Model Names | Accuracy | F1 Score | True Positive | False Positive | True Negative | False Negative |
| 1. | Decision tree | 1.0 | 1.0 | 1208 | 0 | 0 | 792 |
| 2. | Random forest | 1.0 | 1.0 | 1208 | 0 | 0 | 792 |
| 3. | Logistic regression | 0.83 | 0.79 | 1026 | 145 | 182 | 647 |
| 4. | KNN | 0.80 | 0.75 | 1001 | 184 | 207 | 608 |

Table 1. Model Predictions

I also evaluated the models using their ROC curves, the ROC curve of all models is shown below.

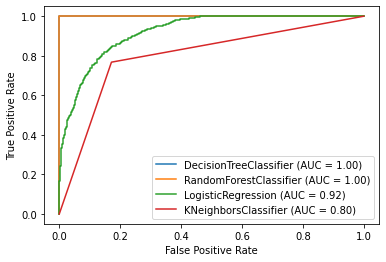


Figure 12. ROC curve of all models

#### **4.2 Using feature importance:**

I used feature importance for finding out the important columns, to get the best model. So, in training data we took the top three important columns and the train test split is same 80:20. The best models are Decision tree and Random forest which is able to predict with 81.2%. The observations of all the models are shown below.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Sr. No. | Model Names | Accuracy | F1 Score | True Positive | False Positive | True Negative | False Negative |
| 1. | Decision tree | 0.812 | 0.77 | 994 | 162 | 214 | 630 |
| 2. | Random forest | 0.812 | 0.77 | 994 | 162 | 214 | 630 |
| 3. | Logistic regression | 0.71 | 0.61 | 961 | 328 | 247 | 464 |
| 4. | KNN | 0.80 | 0.75 | 1004 | 194 | 204 | 598 |

Table 2. Model predictions after using feature importance

I also evaluated the models using their ROC curves on the important features, the ROC curve of all models is shown below.

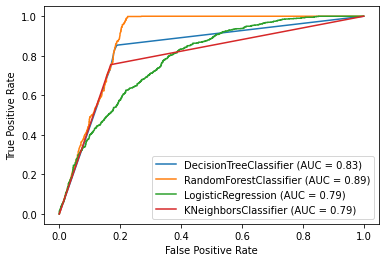


Figure 13. ROC curve of all models after using feature importance.

### **5. Conclusion**

In this study, I analysed 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. Highest selection percentage is for Blockchain and the lowest is for the IoT.
4. Most of the students applied for the Internship are from the Computer Engineering background.
5. Using Random forest, every time we were able to get almost correct predictions.

### **6. Future directions**

I was able to achieve 100% accuracy with normal model and 81% accuracy with the important features in the classification problem. However, there was still significance variance that could not be predicted by the models in this study. I think the models could use more improvements on capturing player’s individual traits. For example, two candidates might have similar CGPA, but one might have more personality traits, mental strength or other qualities. The future performance of these two candidates might be different.