IP – Data Science Live Project Report

Rahil Merchant

Abstract

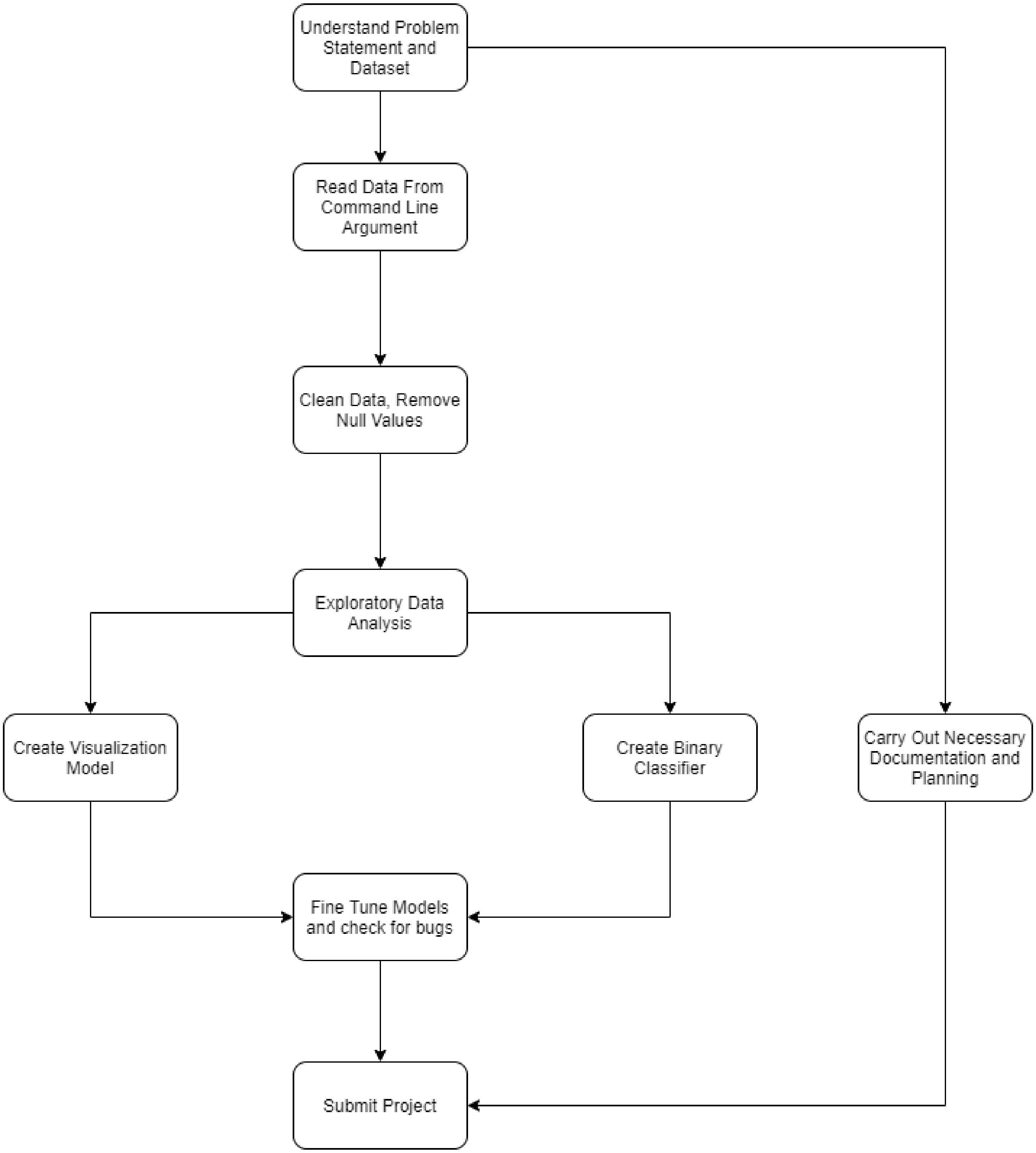
Using the sample dataset provided, I have made developed Data Visualization and Classification models that output an auto-generated pdf and the F1 Score of the best performing model. The pdf will show statistical analysis strive to gain useful insights as well as provide us with some knowledge about the dataset provided. This project can be used by Cloud Counselage to efficiently screen internship applications and reduce workload on staff by using an automated classification process.

Problem Statement

Students from different cities from the state of Maharashtra had applied for the Cloud Counselage Internship Program. We have the dataset consisting of information of all the students. Using this data, we want to get more insights and draw out more meaningful conclusions. Interns are expected to build a data visualization model and find the best data segmentation model using the student’s dataset. Following are the tasks interns need to perform:

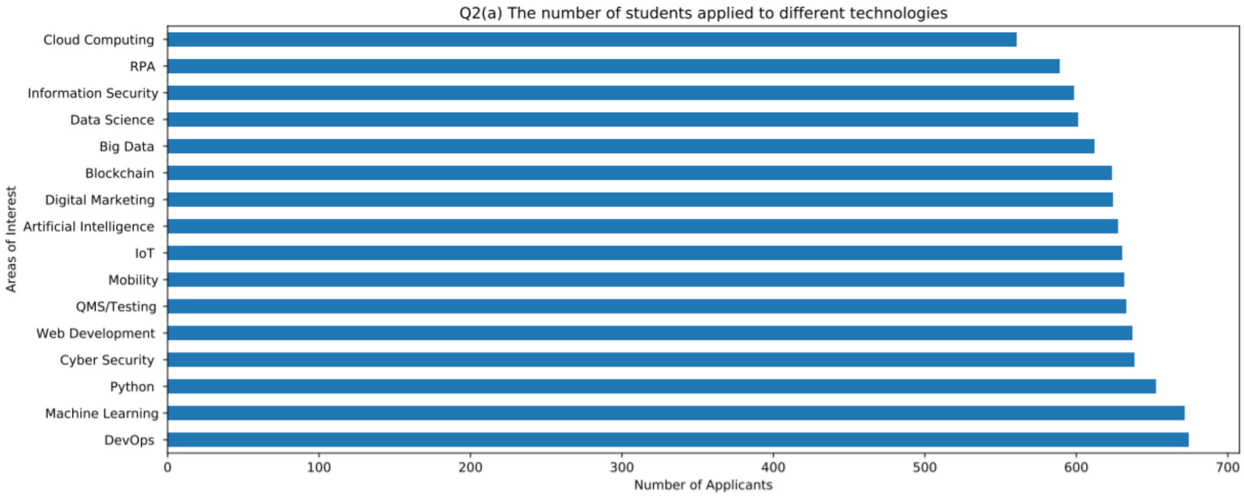
- Interns need to preprocess the data for missing values, unknown values, encoding categorical values.
- Create a data visualization model to build graphs from the dataset answering the following questions:
    - The number of students applied to different technologies.
    - The number of students applied for Data Science who knew ’’Python” and who didn’t.
    - The different ways students learned about this program.
    - Students who are in the fourth year and have a CGPA greater than 8.0.
    - Students who applied for Digital Marketing with verbal and written communication score higher than 8.
    - Year-wise and area of study wise classification of students.
    - City and college wise classification of students.
    - Plot the relationship between the CGPA and the target variable.
    - Plot the relationship between the Area of Interest and the target variable.
    - Plot the relationship between the year of study, major, and the target variable.
- Identify the best binary classifier to classify data into “eligible/1” and “not eligible/0”

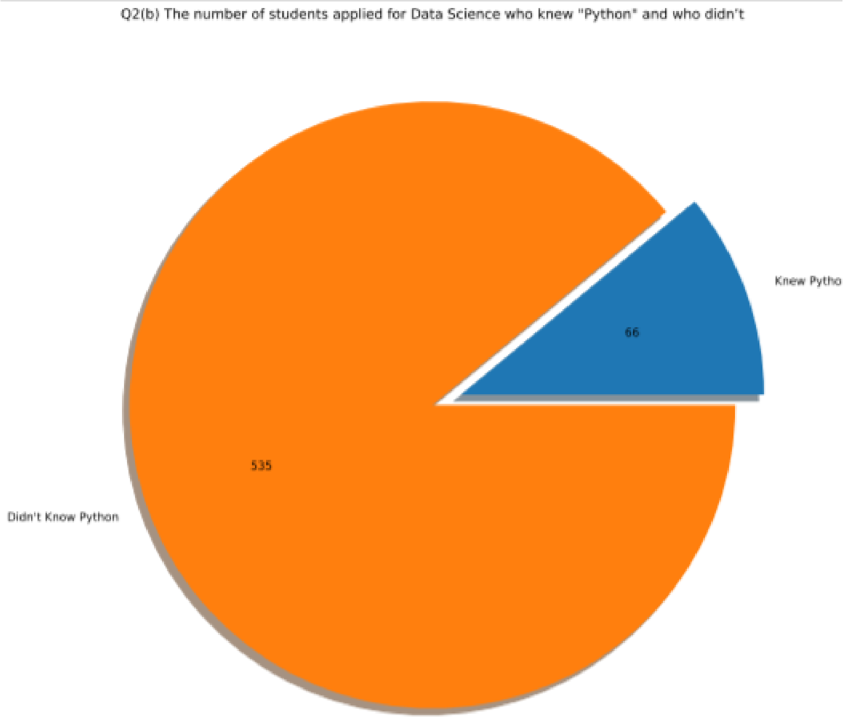
Design

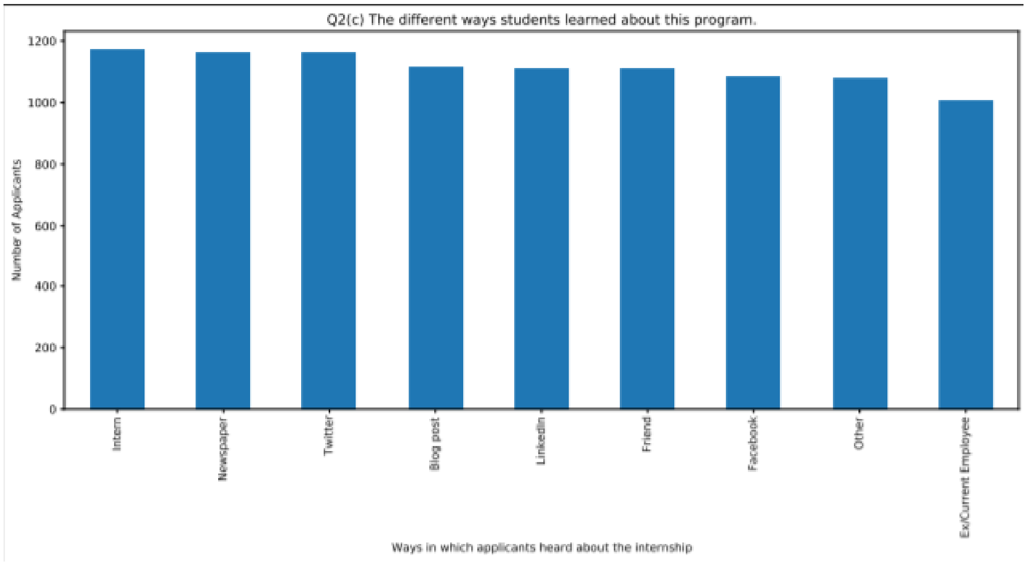

Understand Problem Statement and Dataset
Read Data From Command Line Argument
Clean Data, Remove Null Values
Exploratory Data Analysis
Create Visualization Model
Create Binary Classifier
Carry Out Necessary Documentation and Planning
Fine Tune Models and check for bugs
Submit Project

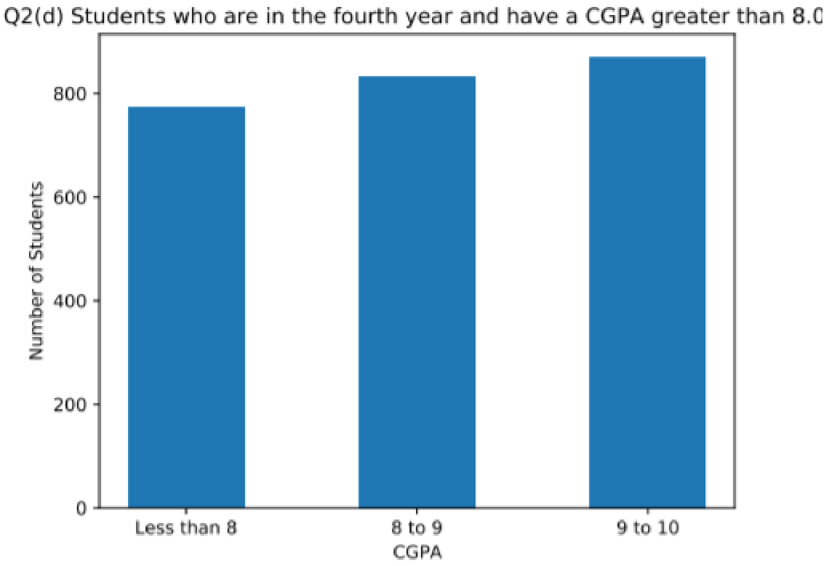
Visualizations Created

The following visualizations were created with the help of matplotlib and seaborn libraries. These strive to answer the questions asked in the problem statement. I

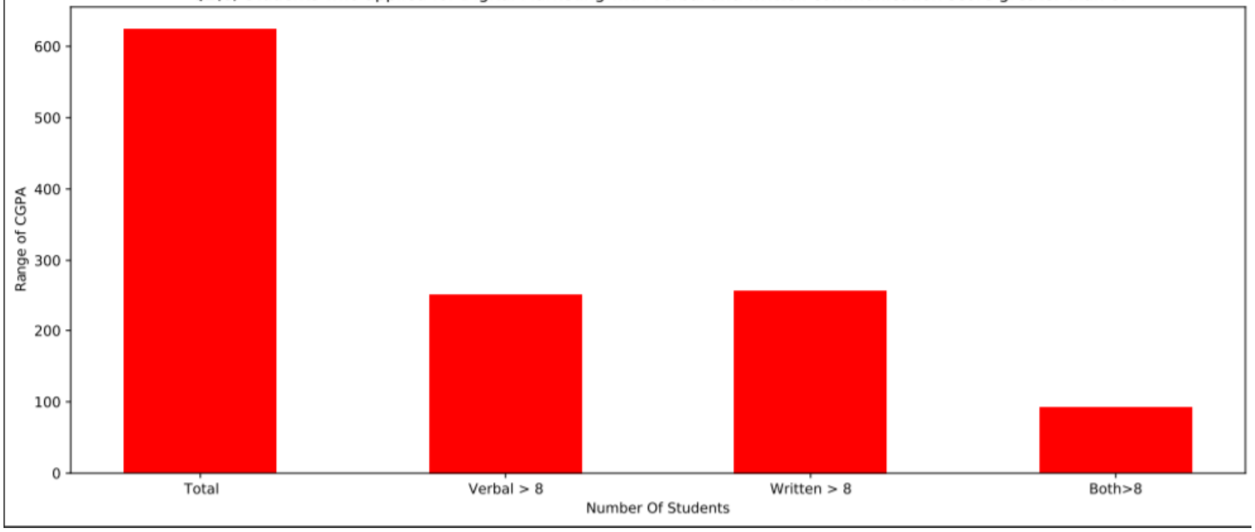




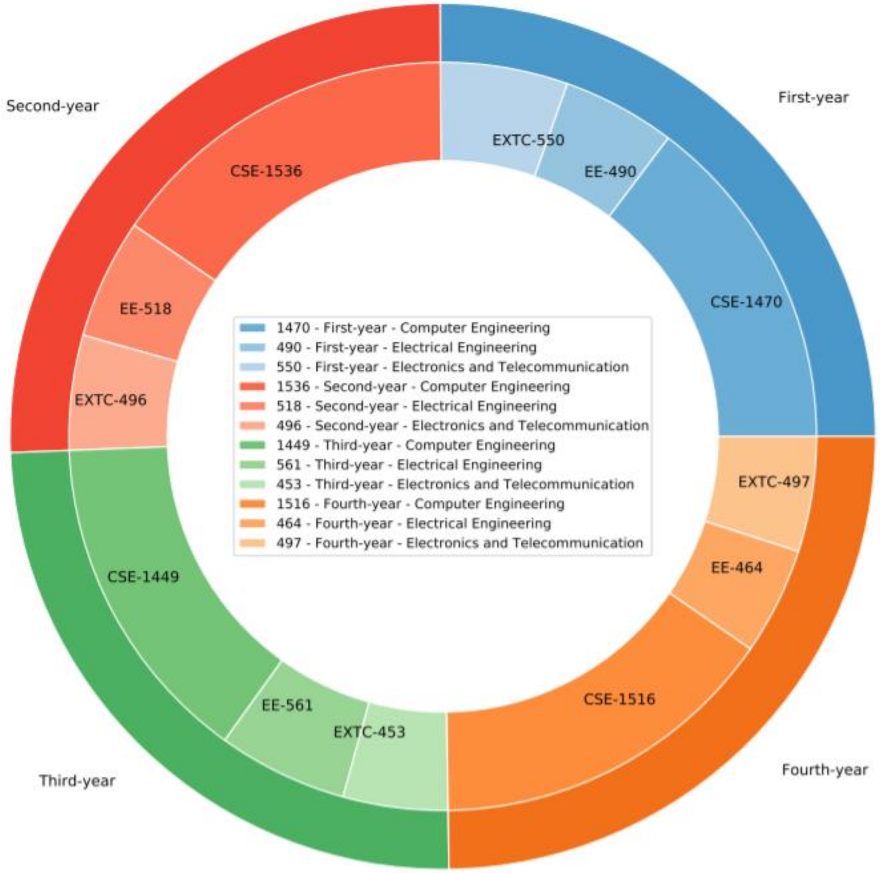

Q2(c) The different ways students learned about this program.
Number of Applicants
0
200
400
600
800
1000
1200
Intern
Newspaper
Twitter
Blog post
Linkedin
Friend
Facebook
Other
Ex/Current Employee
Ways in which applicants heard about the internship


Q2(d) Students who are in the fourth year and have a CGPA greater than 8.0
Number of Students
0
200
400
600
800
Less than 8
8 to 9
9 to 10
CGPA

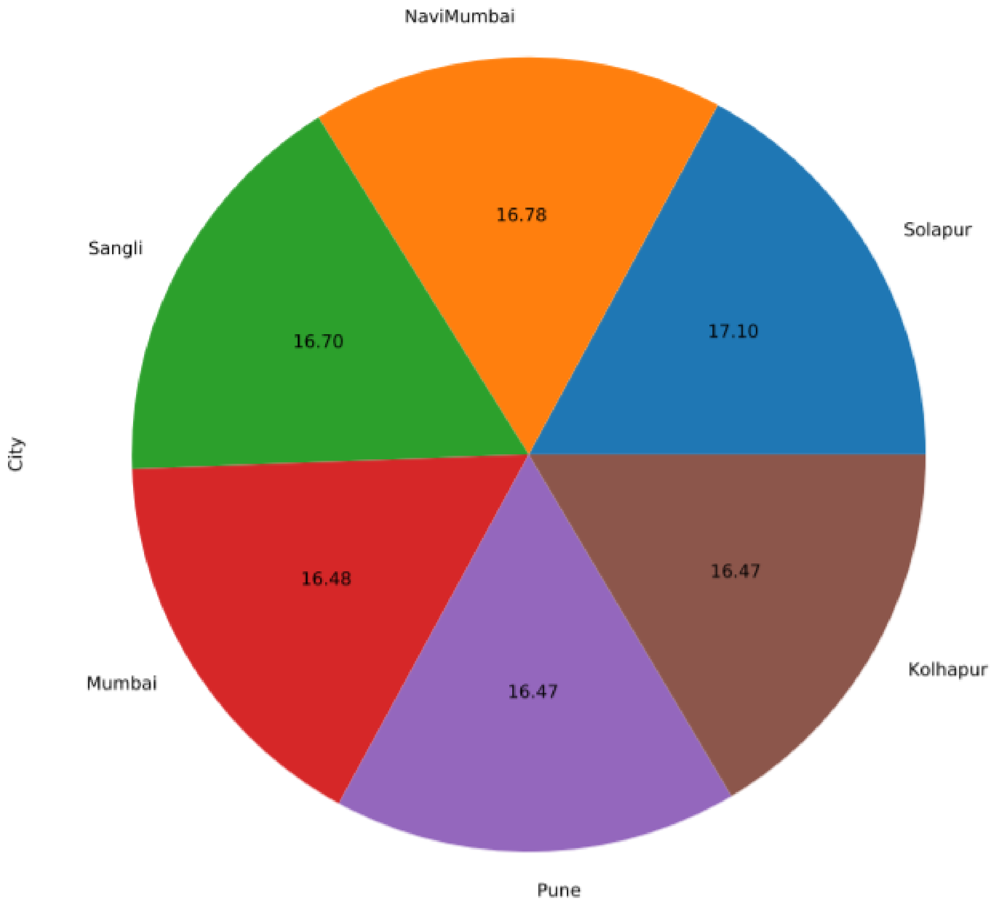
Q2(e) Students who applied for Digital Marketing with verbal and written communication score greater than 8.



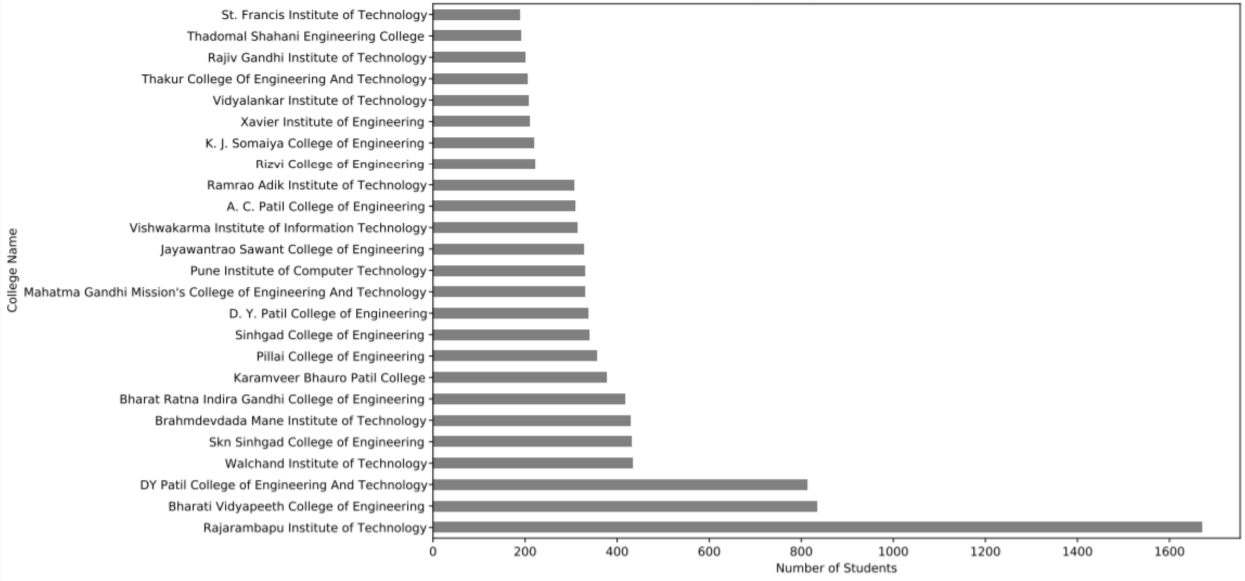
Q2(f) Year-wise and area of study wise classification of students.

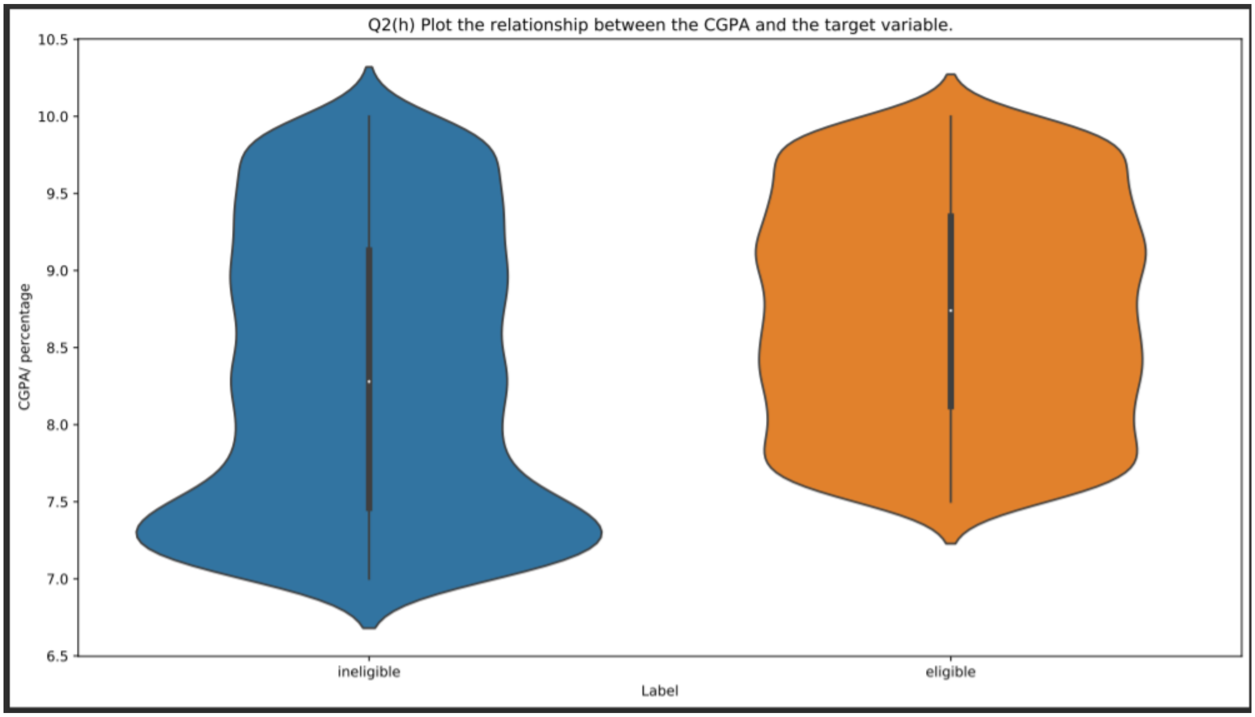


Q2(g) City-wise Distribution

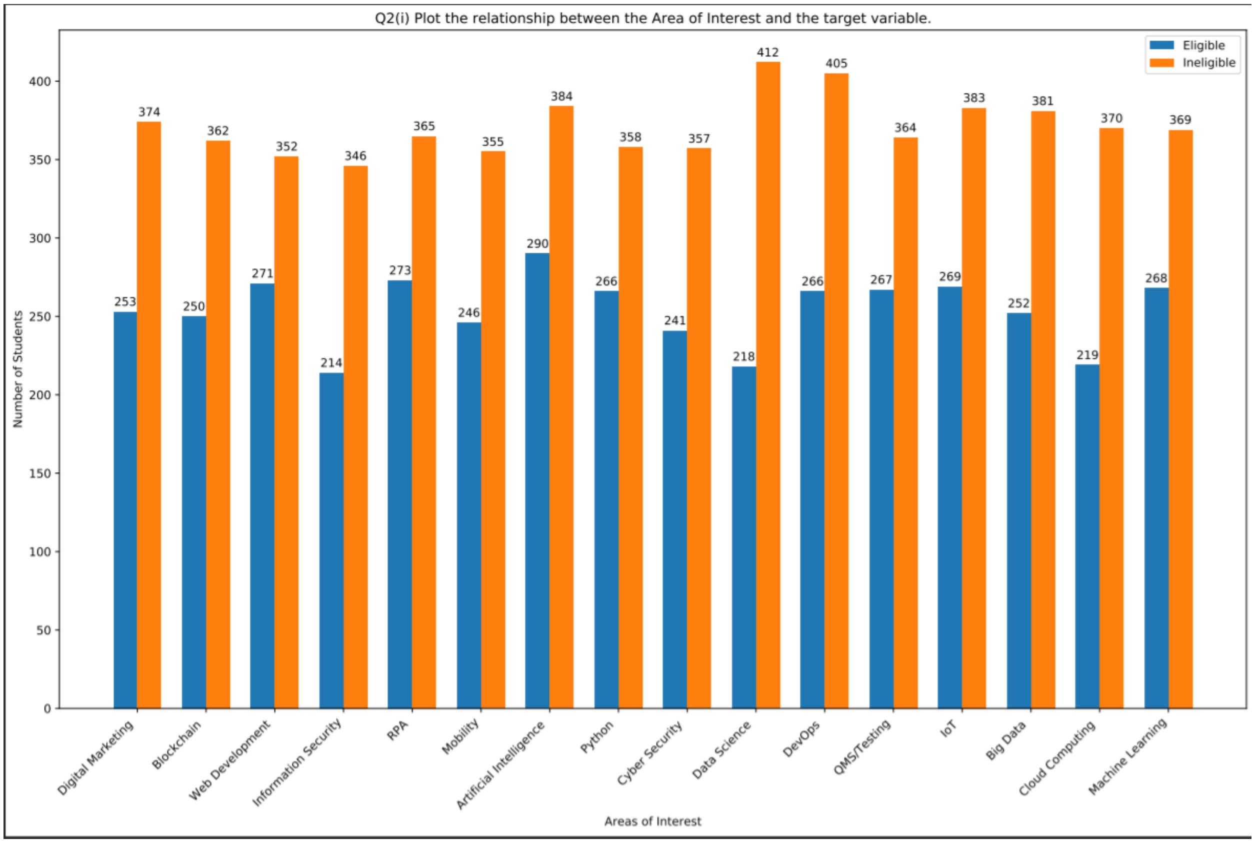


Q2(g) College-wise Distribution



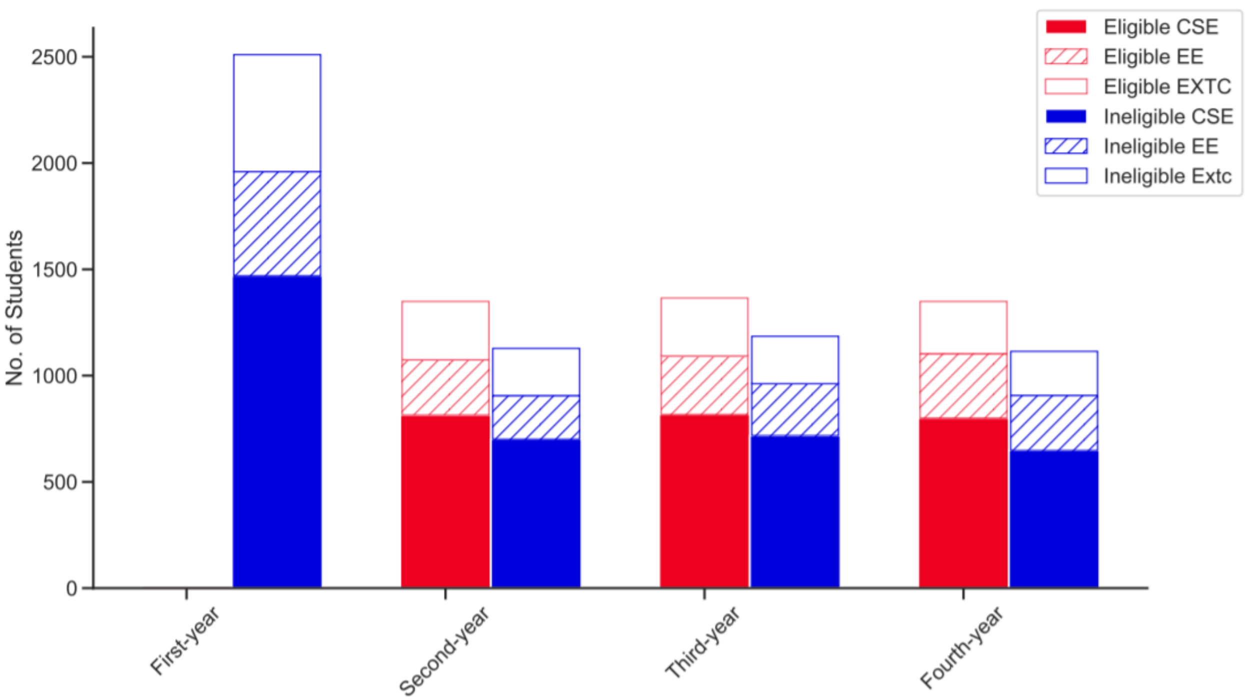


Q2(h) Plot the relationship between the CGPA and the target variable.



Q2(i) Plot the relationship between the Area of Interest and the target variable.

Q2(h) Plot the relationship between the year of study, major, and the target variable.



Classification

The following process was implemented:

- Null values and redundant columns were filtered out.
- Following this, variables with strong correlation to the target variable were identified.
- All categorical values were either Binary Encoded or One- Hot Encoded
- Standard Scaler was used for scale the GPA and Verbal and Written skills assessments. This was done so that models such as Support Vector Machine and K-Nearest Neighbors rely of scaled values and provide skewed results otherwise.
- The following algorithms were tested:
    - XGBoost
    - AdaBoost
    - SVM
    - KNN
    - Logistic Regression
    - Random Forest
    - Decision Trees
- The highest F1-score attained after cross validation was 1.0 achieved by XGBoost, AdaBoost and Random Forest

Acknowledgement

I would like to thank Cloud Counselage for this great opportunity to undergo training as well as get some invaluable experience working on a live project.

I would also like to thank Mr. Nirbhey Singh Pahwa, Mr. Harish Sapte and Mr. Jayanth GS for their valuable guidance throughout the tenure of the internship.