E0 259:Data Analytics

Course Project

Y Rithvik ,Ramesh
K Kalyan ,Y S Sai Nitish

Campus Placements Prediction from Student data

The Data

Dataset

- 1400 data points collected from OCCAP,IISc
- Campus Placements data for the past 5 years

Features of a student

- CGPA
- Discipline
- Programme
- Skills
- Courses

Target Variables

- Slot
- CTC

Data Pre-processing

1. **CGPA**: Achieved data uniformity by scaling all CGPA values to 10.0, addressing deviations like ('x / 10', 'y / 8').

2. Categorized **'Program'** and **'Discipline'** features using a consistent numerical encoding for efficient preprocessing.

3. In an intermediary step, we categorized all 'Job Roles' into the following 13 broad categories:
   1)'Data-Science' **2)**'Software-Development' **3)**'Design'
   4)'Electrical Engineering' **5)**'Aerospace' **6)**'VLSI'
   7)'Data Analytics' **8)**'Chemical Engineering' **9)**'Mechanical Engineering'
   10)'Misc' **11)**'Communications' **12)**'Robotics'
   13)'Multimedia systems'

4. Categorized diverse **'Skill sets'** and '**Courses Taken**' into 13 pools aligned with specific job roles.

Slot	*Slot 1*		1
Program	*M.Tech*		2
Discipline	*Computer Science and Automation*		52
Department	*Computer Science and Automation*		NA
Position	*System Software engineer*		‘sde’
CTC	4900807		4900807
CGPA	*8.1*		8.1
Skillset	AI Frameworks and Programming Languages:PyTorchTensorflow(Keras)C/C++MATLABSQLKnowledge Areas:Deep LearningMachine LearningComputer Vision :;		[0, 0 , 0 , 0 , 0, 0, 1, 0, 0,1, 0, 0, 1]
Courses	Linear Algebra and Probability Design and Analysis of Algorithms Machine Learning Game Theory Deep Learning for Natural Language Processing Data Analytics		[0, 0 , 0 , 0 , 1, 0, 1, 0, 0,1, 0, 1, 1]

Models and Results

Two Tasks:

1. Classification:
   - Slot
2. Regression
   - CTC

Classification:

- Random Forests
- Logistic Regression

Regression:

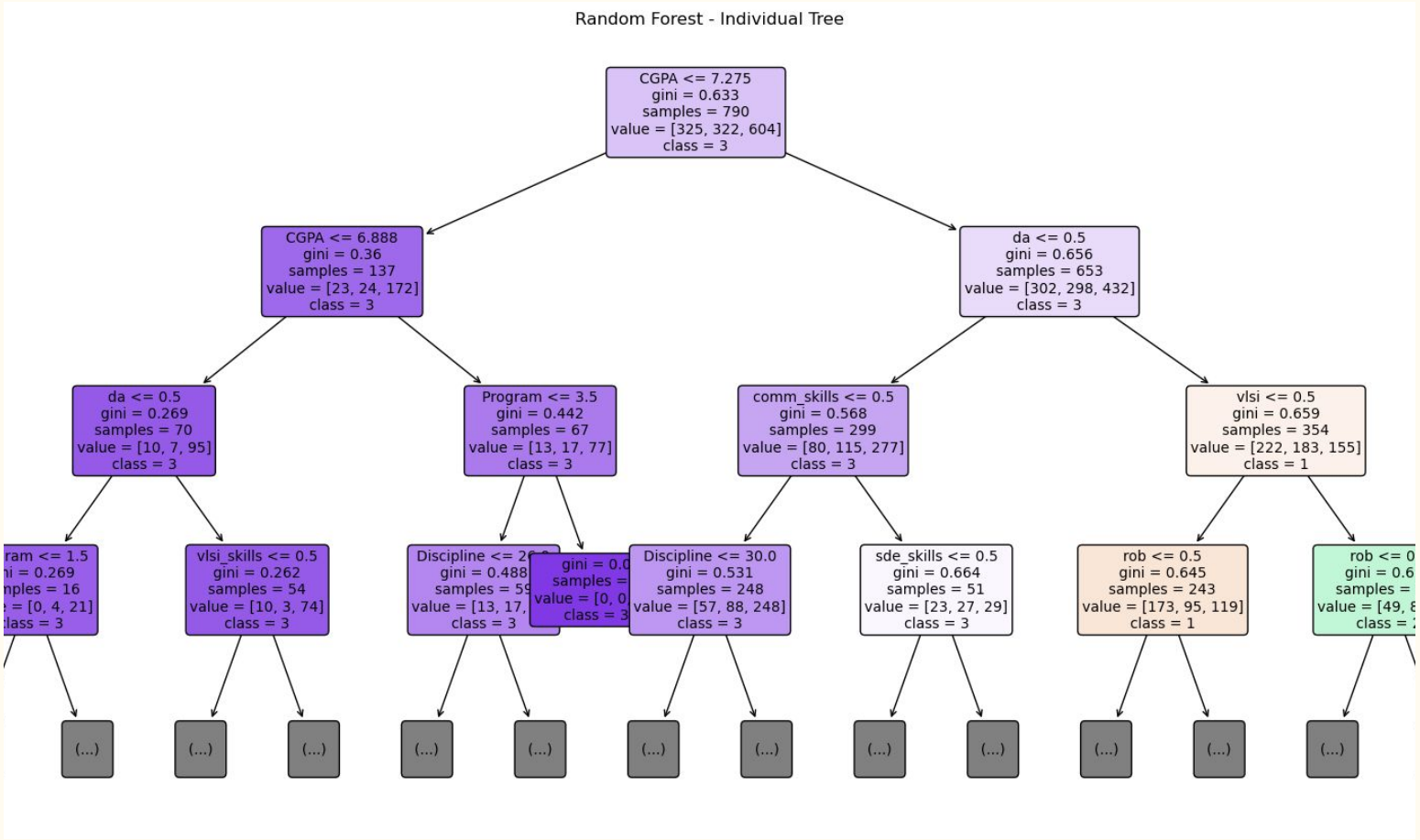
- Random Forests

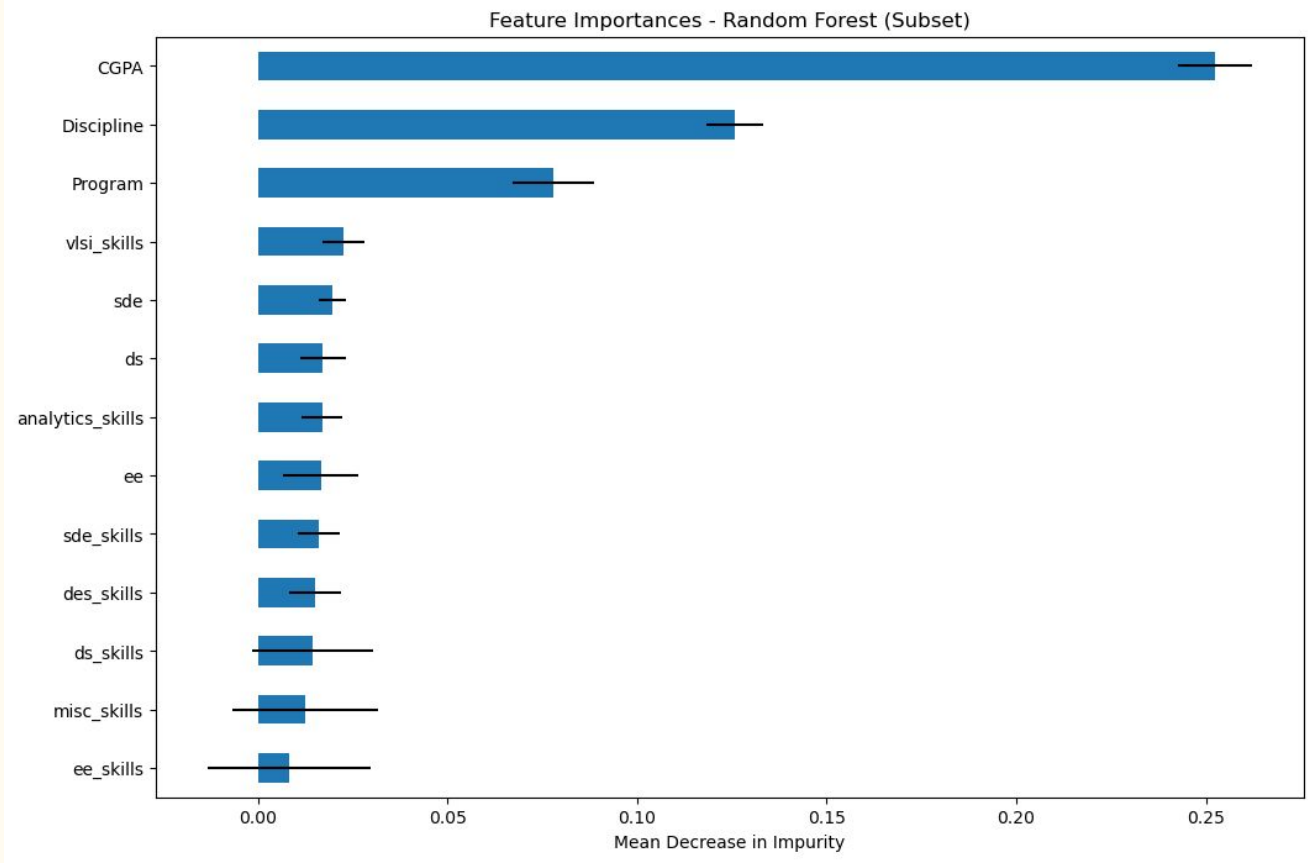
Slot Prediction

Random Forests

1. Train:Test split = **90:10**
2. Number of decision trees used : **100 with each tree of maximum depth 10**
3. Classifier predicts the probabilities of getting placed in slot 1,2,3.
4. Criterion : Gini Impurity
5. Achieved a test accuracy of 74%

A decision tree in a trained Random Forest




Feature Importances - Random Forest (Subset)
CGPA
Discipline
Program
vlsi_skills
sde
ds
analytics_skills
ee
sde_skills
des_skills
ds_skills
misc_skills
ee_skills
0.00
0.05
0.10
0.15
0.20
0.25
Mean Decrease in Impurity

Multi-class Logistic Regression

1. Train:Test split : **90:10**
2. Classifier predicts the probabilities of getting placed in slot 1,2,3.
3. Loss : Categorical Cross Entropy
4. Achieved a test accuracy of 71%

$$p(y_k = 1|\mathbf{x}) = \frac{e^{(\mathbf{w}_k\mathbf{x}+\mathbf{b}_k)}}{\sum_{j=1}^{K} e^{(\mathbf{w}_j\mathbf{x}+\mathbf{b}_j)}}$$

CTC Prediction

Random Forests Regression

1. Train:Test split : **90:10**
2. Number of trees used : **100 with each tree of maximum depth of 10**
3. Loss : Squared Loss
4. Regresses over CTC target variable

Some Predictions:

Student Details:

Program: M.Tech
Discipline: Computer Science and Automation
CGPA: 8.0
Courses: pattern recognition and neural networks, Computer Vision, machine learning, natural language processing, linear algebra, stochastic models, machine learning
Skills: python, deep learning, pytorch, pytorch, tensorflow

Predicted Slot:

Slot	Probability
Slot 1	0.68
Slot 2	0.29
Slot 3	0.03

Predicted CTC: 29.07 LPA
Predicted Sector(s): Data Science, Software Development

Student Details:

Program: M.Tech
Discipline: Artificial Intelligence
CGPA: 6.0
Courses: pattern recognition and neural networks, Computer Vision, machine learning, natural language processing, linear algebra, stochastic models, machine learning
Skills: python, deep learning, pytorch, pytorch, tensorflow

Predicted Slot:

Slot	Probability
Slot 1	0.03
Slot 2	0.04
Slot 3	0.93

Predicted CTC: 15.86 LPA
Predicted Sector(s): Software Development, Data Science

Some Predictions:

Student Details:

Program: M.Tech
Discipline: Electronic Systems Engineering
CGPA: 8.5
Courses: micro and nano, semiconductors, digital vlsi circuits, analog integrated circuits, design for analog circuits, control system design, microelectronics
Skills: xilinx, vivado, VHDL tools, verilog, cadence virtuoso

Predicted Slot:

Slot	Probability
Slot 1	0.49
Slot 2	0.47
Slot 3	0.04

Predicted CTC: 30.92 LPA
Predicted Sector(s): VLSI

Student Details:

Program: M.Tech
Discipline: Artificial Intelligence
CGPA: 8.3
Courses: machine learning, deep learning, pattern recognition and neural networks, natural language processing, linear algebra, computer vision
Skills: python, tensorflow, pytorch, numpy, pandas

Predicted Slot:

Slot	Probability
Slot 1	0.38
Slot 2	0.41
Slot 3	0.21

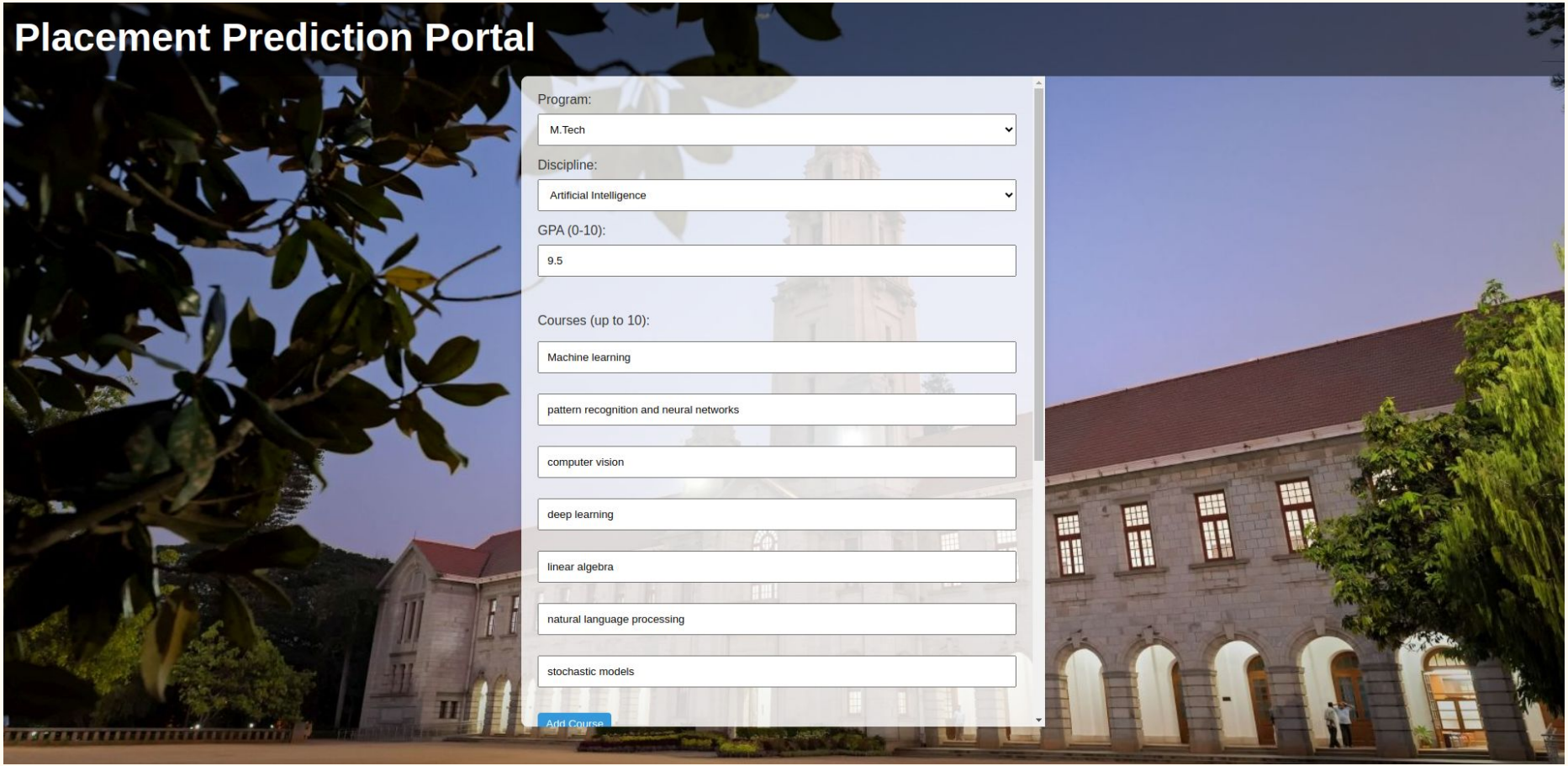
Predicted CTC: 24.07 LPA
Predicted Sector(s): Data Science, Software Development

Deployment

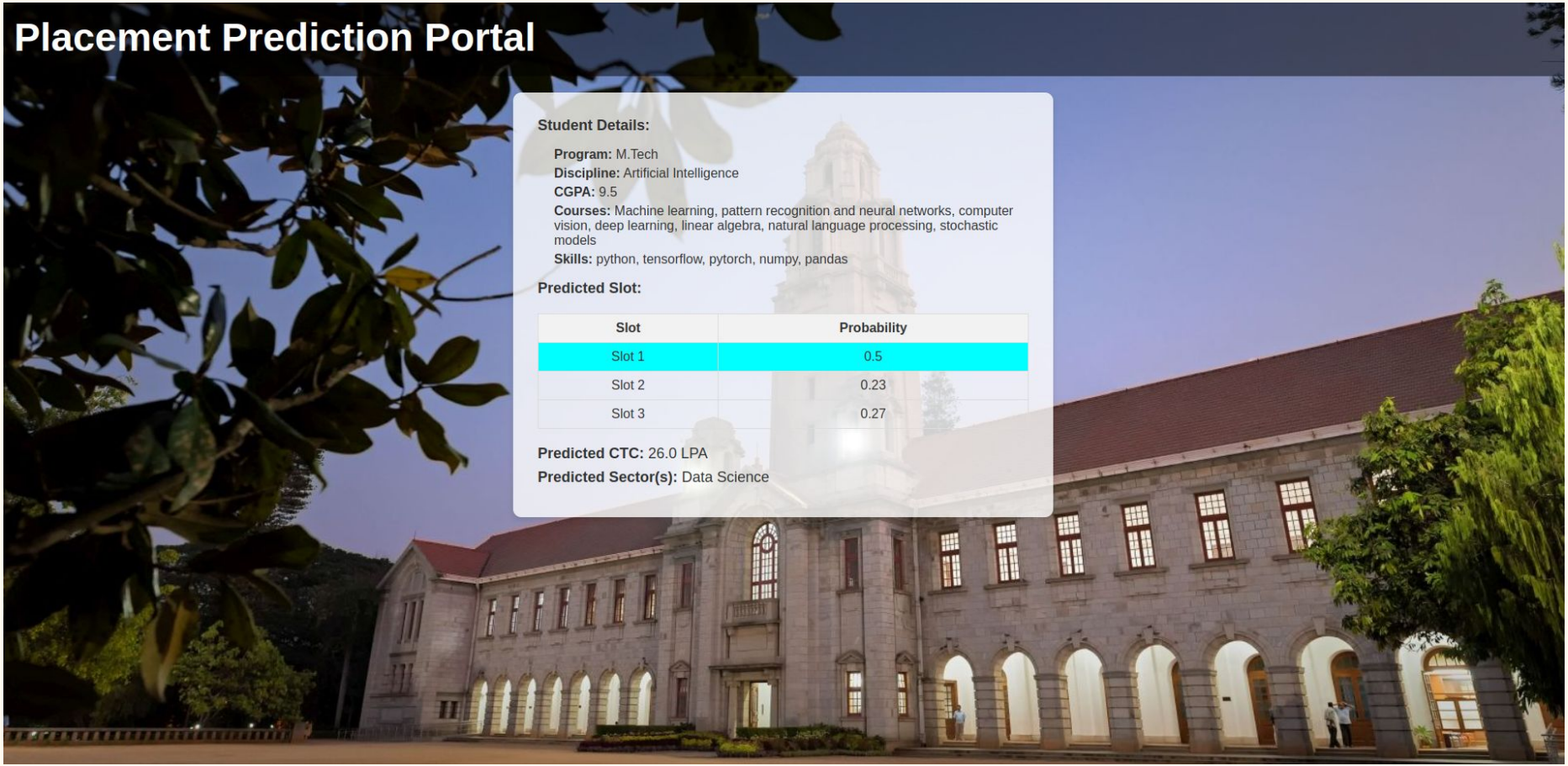
Placement Prediction Portal

Github

Placement Prediction Portal



Placement Prediction Portal



Student Details:

- **Program:** M.Tech
- **Discipline:** Artificial Intelligence
- **CGPA:** 9.5
- **Courses:** Machine learning, pattern recognition and neural networks, computer vision, deep learning, linear algebra, natural language processing, stochastic models
- **Skills:** python, tensorflow, pytorch, numpy, pandas

Predicted Slot:

Slot	Probability
Slot 1	0.5
Slot 2	0.23
Slot 3	0.27

Predicted CTC: 26.0 LPA

Predicted Sector(s): Data Science

Future work

1. We aim to gather more data for all the departments
2. Also, we have used transformed features in our model based on the courses and skills provided by the candidates, which may not be more relevant for our prediction
3. Hence, we aim to gather data with much more relevance to our tasks

Thank You