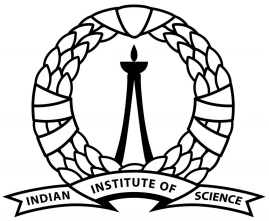

Student Placement Prediction

Team Legion

Team Number 8

A report submitted in partial fulfilment
of the requirements for the course of
E0:259 Data Analytics

Indian Institute of Science, Bangalore

Proponents:

Y S Sai Nitish	Y Rithvik	G Ramesh Babu	K Kalyan Reddy
(21677)	(21090)	(20950)	(21361)

Contents

Chapter 1

Introduction, motivation & Context

1.1 Introduction

Welcome to our exploration of student placement dynamics at IISc. In the past 5 years, we've accumulated a wealth of data on students, their courses, skillsets, and placement outcomes. In this journey, we delve into the intricate process of data preprocessing and prediction using a decision tree model.

1.2 Motivation

The drive behind our project is to unravel the factors influencing student placement. By understanding patterns in skillsets, courses, and academic performance, we aim to enhance the precision of our predictions, ultimately contributing to better-informed decisions in student career paths.

1.3 Context

In the dynamic landscape of student placement, our endeavor is rooted in leveraging machine learning techniques to make sense of the multifaceted data we've gathered. Through meticulous preprocessing and the application of decision tree modeling, we aim to uncover actionable insights that will not only benefit students but also inform strategic decisions in academic program development and career guidance.

Chapter 2

Methods Adopted

2.1 Overview of Data

***O*ur Dataset encompasses various facets of each student's profile:**

- Skillset
- Courses Completed
- Program
- Discipline
- CGPA
- Job Role Offered
- Range of Salary Offered

2.2 Pre-Processing Data

2.2.1 CGPA

- The CGPA, initially appearing clean, revealed a subtle complexity upon closer inspection in our dataset. Notably, a variation existed, with some candidates displaying a maximum GPA of 8.0, while others consistently had it as 10.0.
- To establish uniformity, our initial step involved scaling all values to 10.0.
- Further scrutiny unveiled entries where candidates recorded CGPA as 'x / 10' or 'y / 8', deviating from the standard format of 'x' or 'y' in these instances.
- Additionally, certain data points exhibited irregularities, such as spaces within numerical values, like '7. 10' instead of the conventional '7.10'. This necessitated meticulous preprocessing to ensure data consistency and accuracy.

2.2.2 Program & Discipline

- The subsequent preprocessing step involves encoding the categorical feature 'Program.' We applied a consistent encoding schema, representing various program options with numerical labels:

```
PROGRAM_OPTIONS = {
    'M.Tech': '1',
    'M.Tech(Res)': '2',
    'PhD(Eng)': '3',
    'MSc(Res)': '4',
    'BSc(Res)': '5',
    'M.Mgmt': '6',
    'M.Des': '7',
    'PhD(Sci)': '8',
    'PhD': '9',
    'Post Doc': '10'
}
```

- Following a similar ordinal encoding approach, we encoded the categorical feature **'Discipline'**.
- Notably, placement activities involved candidates from 60 distinct disciplines, reflecting the diverse academic backgrounds contributing to our dataset.
- This systematic encoding ensures numerical representation for these categorical features, facilitating the subsequent stages of our analysis.

2.2.3 Pooling Job Roles

- In an intermediary step, we categorized all 'Job Roles' into 13 broad categories to streamline our analysis.
- These categories, namely 'aero,' 'mech,' 'chem,' 'sde,' 'ds,' 'analytics,' 'misc,' 'vlsi,' 'comm,' 'ee,' 'rob,' 'mm,' and 'des,' provide a structured framework for understanding and grouping diverse roles.
- This classification proves instrumental in extracting insights into the specific courses candidates have undertaken to secure placements in distinct fields of work. The deliberate categorization of job roles facilitates a more focused examination of the academic pathways that align with varied professional trajectories.

2.2.4 Skillset Processing

- One of the pivotal factors influencing placement success is a candidate's Skillset. However, directly utilizing self-claimed skillsets presents challenges due to variations in both skill types and their quantity among students. To tackle this, we adopted a refined approach.
- Similar to courses, we extracted skillsets based on the pooling of roles. Instead of treating the entire skillset as a single feature, we categorized it into 13 distinct pools.
- Each of these 13 features corresponds to skills commonly associated with specific roles. This strategic approach allows for a more nuanced analysis of a candidate's suitability for a role.

2.2.5 Courses Processing

- Another crucial determinant of successful placement is the candidate's 'Courses Taken.' However, the diversity in courses and their varying numbers among students poses a challenge in direct utilization.
- To address this, we adopted a systematic approach. Instead of treating the entire set of courses as a single feature, we divided them into 13 distinct pools. Each of these 13 features corresponds to courses that align with specific roles.
- Building on the earlier identified role categories, we constructed a dictionary where each class name serves as a key.

- The associated values comprise courses taken by candidates placed in roles falling under that class. Leveraging text processing techniques on the course names allows for flexibility during testing.
- Even if a course name is not precisely matched to the dataset, we can map it to a particular class of roles suitable for the candidate. This method enhances the robustness and adaptability of our model during evaluation.

2.2.6 Final Data

The processed and refined dataset now comprises features in a total of 29 dimensions, encompassing:

- CGPA
- Program
- Discipline
- Skills (13 dimensions)
- Courses (13 dimensions)
- This structured representation captures the diverse aspects of a candidate's profile, providing a comprehensive foundation for subsequent analysis and modeling.

2.3 Visualizing Clean Data

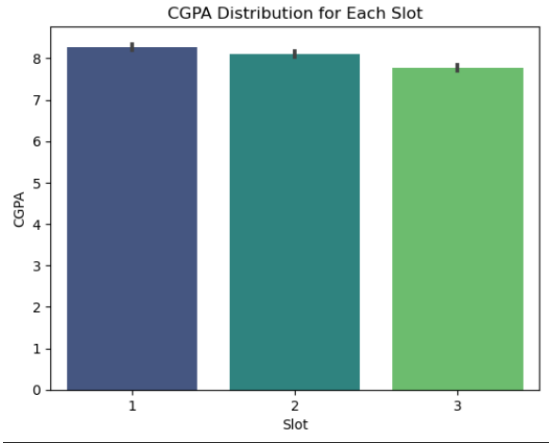

Figure 2.1: CGPA on clean data

Feature	Value
Program	M.Tech
Discipline	Computer Science and Automation
CGPA	9.7
Courses	Computer Vision, Game Theory, Reinforcement Learning, Advanced Image Processing, Linear Algebra, Pattern Recognition and Neural Networks, Stochastic Models, Deep Learning, Natural Language Processing
Skills	Python, C, PyTorch, TensorFlow, Pandas

Transformed Data Point: $[9.7, 51, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 0]$

2.4 Model Evaluation & Interpretation

- We employed 3 different machine learning models namely Logistic Regressor, Decision Tree & Random Forests for the purpose of predicting the Slot in which a candidate will probably get placed.
- We used a training to test split ratio of 90:10.
- A Logistic Regressor with a maximum of 10,000 iterations has been trained on the dataset.
- Extended the analysis to regression tasks(CTC Prediction).
- Trained a Decision tree Regression & Random Forest Regression models and evaluated them using Mean Squared Error & R-Squared.

2.5 Model Visualization

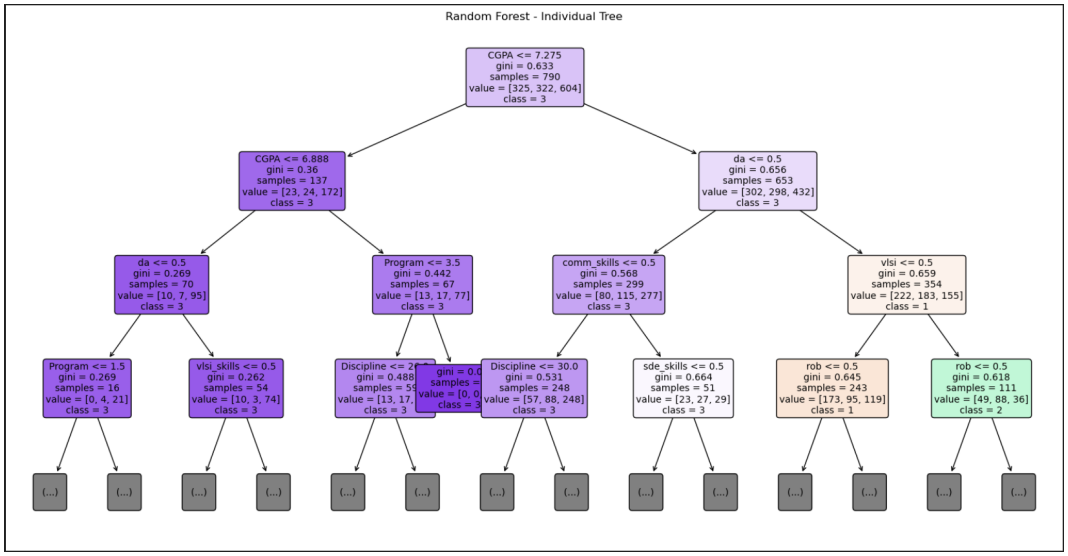

Figure 2.2: Individual Tree in the Random Forest Visualized for a depth of 3

Chapter 3

Results

3.1 Classification Results

3.1.1 Decision Tree

Confusion Matrix

	Predicted		
Actual	Class 1	Class 2	Class 3
Class 1	16	6	5
Class 2	7	17	6
Class 3	8	13	62

Classification Report

	Precision	Recall	F1-Score	Support
Class 1	0.52	0.59	0.55	27
Class 2	0.47	0.57	0.52	30
Class 3	0.85	0.75	0.79	83
Accuracy			0.68	

3.1.2 Random Forest

Confusion Matrix

	Predicted		
Actual	Class 1	Class 2	Class 3
Class 1	15	6	6
Class 2	9	18	3
Class 3	6	6	71

Classification Report

	Precision	Recall	F1-Score	Support
Class 1	0.50	0.56	0.53	27
Class 2	0.60	0.60	0.60	30
Class 3	0.89	0.86	0.87	83
Accuracy			0.74	

3.1.3 Logistic Regression

Confusion Matrix

	Predicted		
Actual	Class 1	Class 2	Class 3
Class 1	15	3	9
Class 2	11	10	9
Class 3	5	4	74

Classification Report

	Precision	Recall	F1-Score	Support
Class 1	0.48	0.56	0.52	27
Class 2	0.59	0.33	0.43	30
Class 3	0.80	0.89	0.85	83
Accuracy		0.71		

3.2 Feature importance Predictions

3.2.1 Shapley Additive Explanations performed on Random Forests

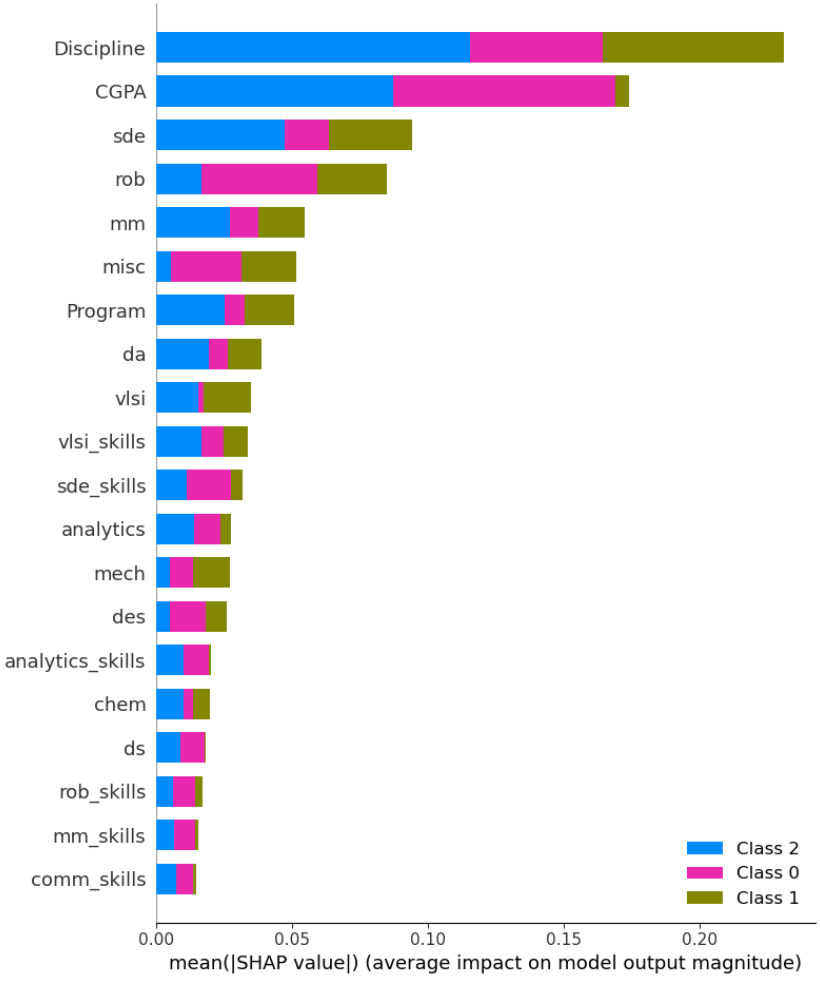

Figure 3.1: SHAP analysis

3.2.2 Mean Decrease Impurity - Decision Trees

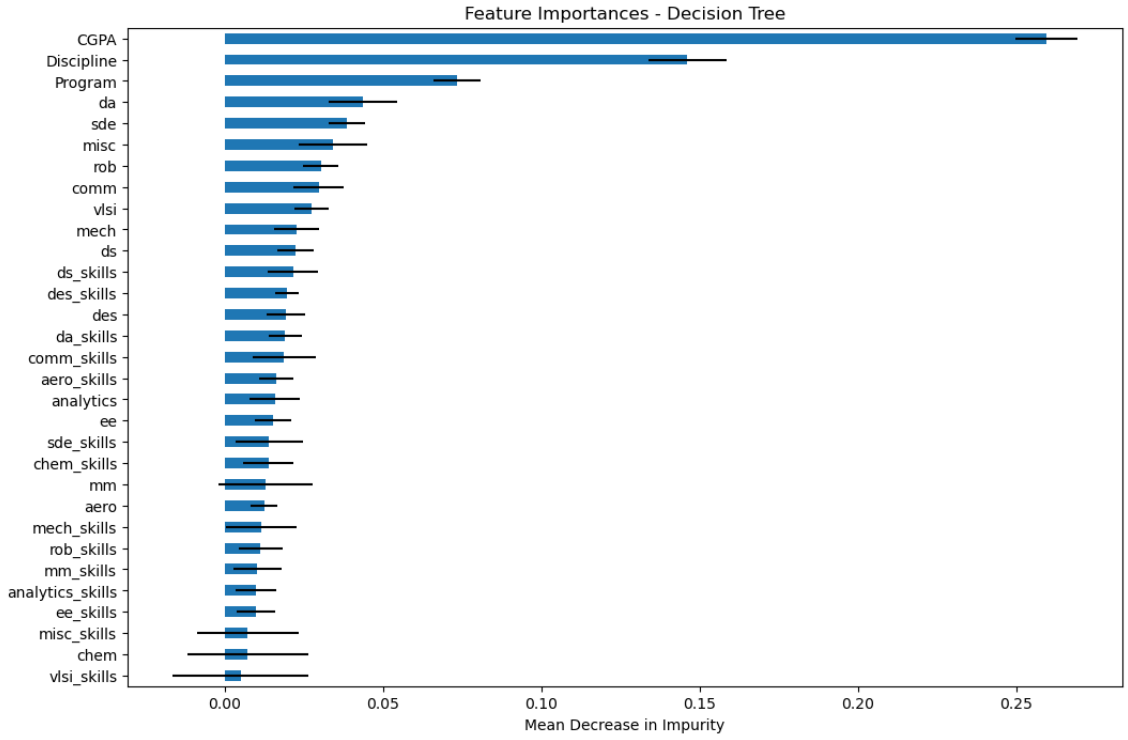

Figure 3.2: Feature Importance Analysis Results

3.2.3 Mean Decrease Impurity - Random Forest

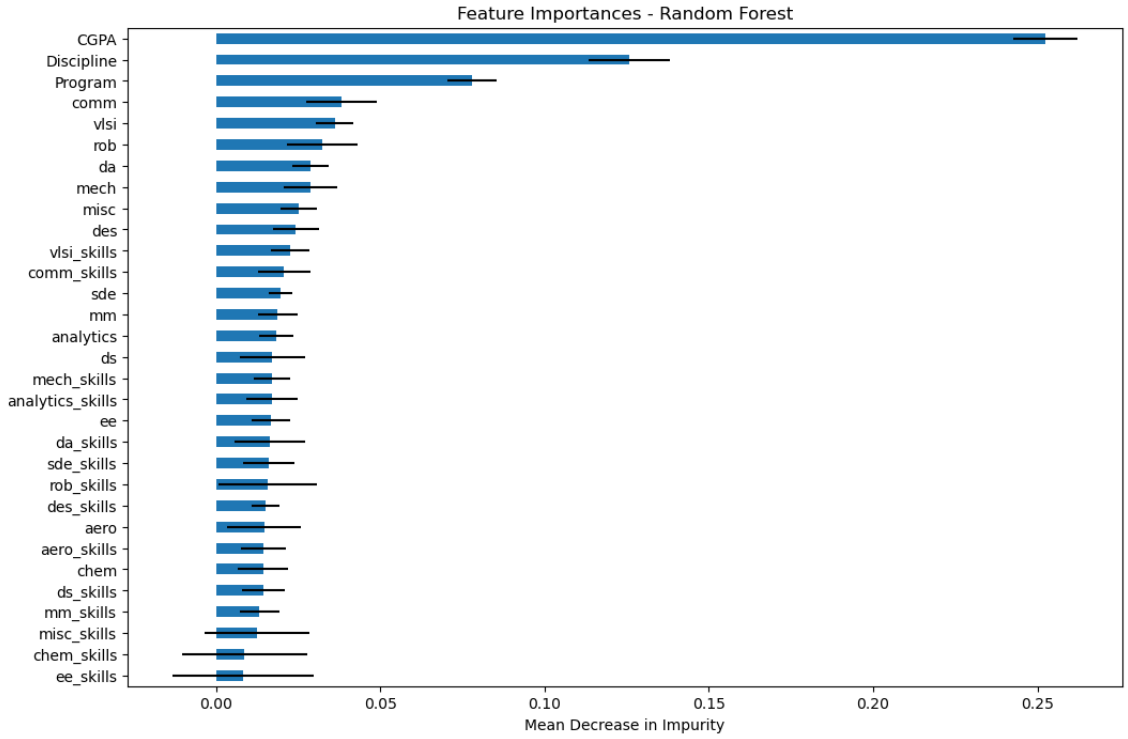

Figure 3.3: Feature Importance Analysis Results

- CGPA Dominance: CGPA emerges as the most influential factor, with a substantial feature importance of 25.2%, underscoring its pivotal role in slot allocation.
- Discipline and Program Impact: Academic discipline 12.6%) and program type 7.8%) are noteworthy contributors, indicating the influence of educational background on slot assignment.
- Skills Significance: Specific skills, such as those related to aero, analytics, and communication, exhibit varying but discernible impacts, emphasizing the importance of both technical and soft skills.
- Diversity in Skill Preferences: While some skills like robotics (3.2%) and software development (1.9%) hold notable importance, others, such as data analytics (1.6%) and miscellaneous skills (1.2%), contribute distinctly.

3.2.4 Mean Decrease Impurity on a subset of features - Random Forests

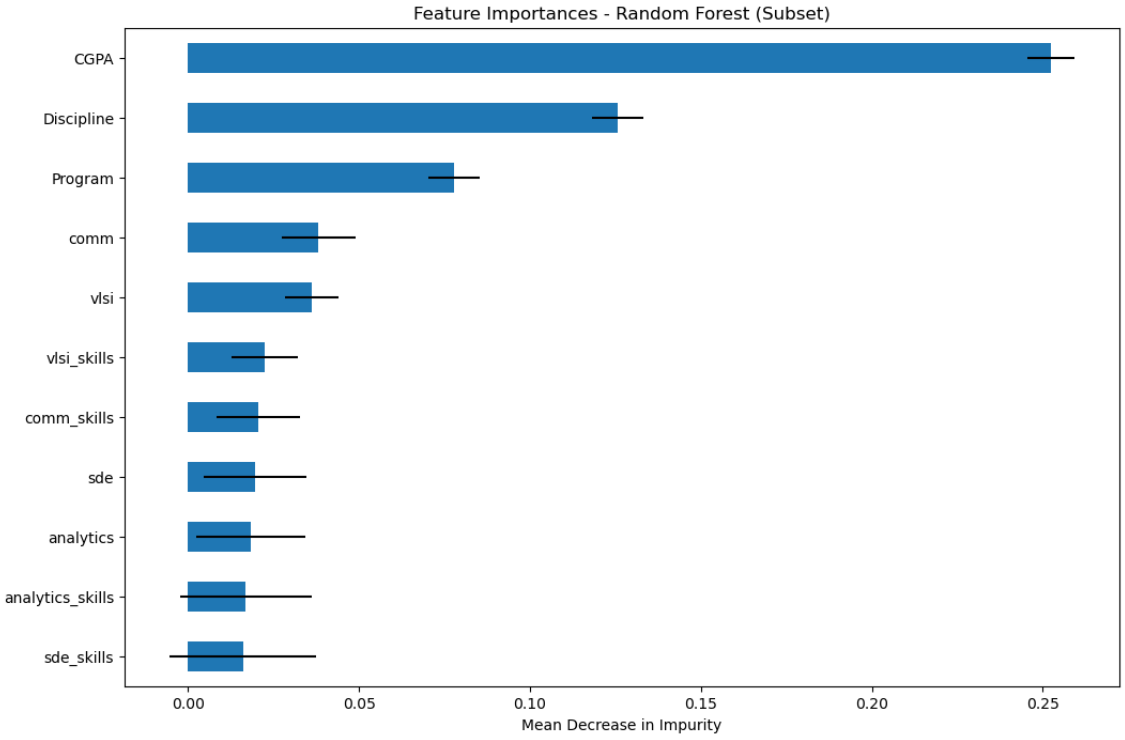

Figure 3.4: Feature Importance Analysis Results

3.3 Decision Tree Regression & Random Forest Regression Evaluations

- Decision Tree Regression Mean Squared Error: 1227236.4713890587
- Decision Tree Regression R-squared: -0.20694099338151917
- Random Forest Regression Mean Squared Error: 1050495913071.9984
- Random Forest Regression R-squared: 0.15817096864858948

3.4 Predictions on test points

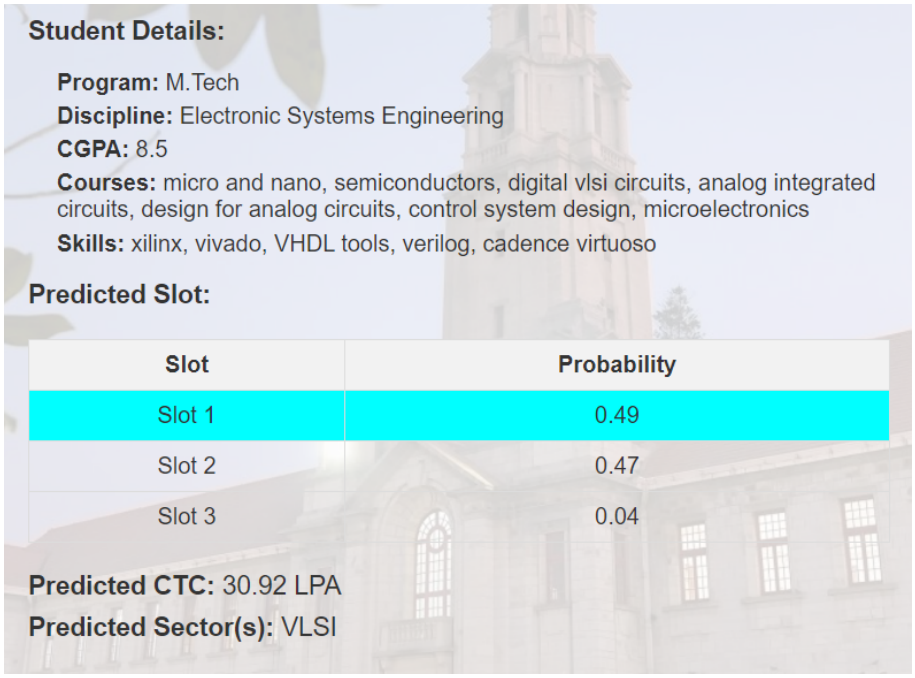

Figure 3.5: [Prediction 1

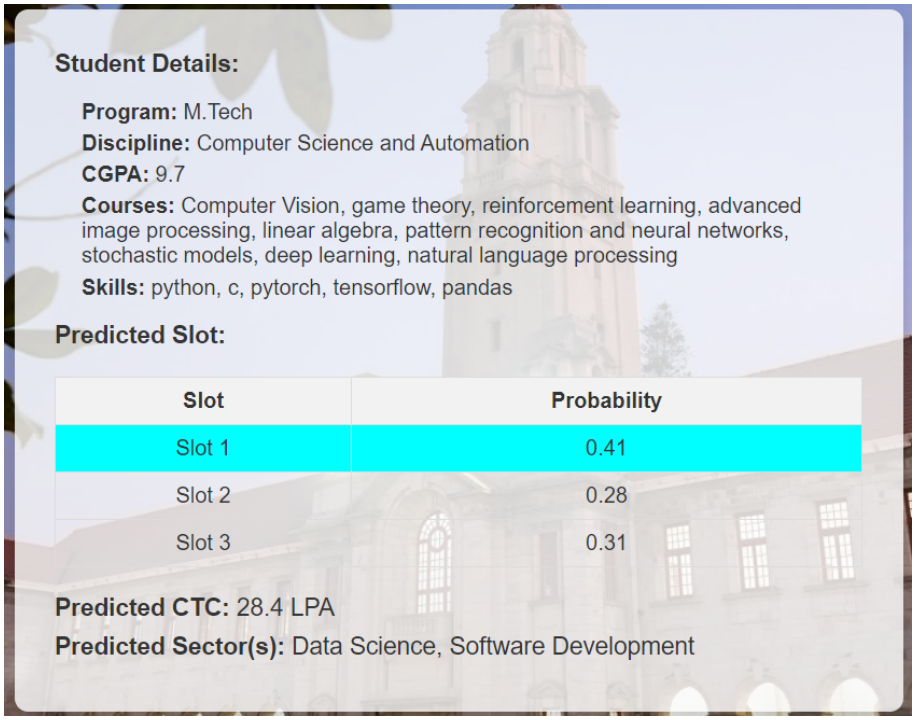

Figure 3.6: Prediction 2

Chapter 4

Conclusions

- Model Evaluation: The Random Forest classification model outperformed Decision Tree and Logistic Regression, indicating its robustness in predicting student slot categories. However, there is room for improvement in Logistic Regression's precision-recall trade-offs.
- Regression Challenges: The regression models faced challenges in accurately predicting Cost to Company (CTC), evident from high Mean Squared Error values. Further refinement is crucial for enhancing the predictive capabilities of these models.
- Feature Significance: Feature importances from the Random Forest model highlighted the critical role of CGPA, academic discipline, and specific skills in slot allocation. This insight emphasizes the importance of academic performance and diverse skill sets in the placement process.
- Slot Allocation Nuances: Despite achieving reasonable accuracy, the models might benefit from a more nuanced understanding of slot allocation factors. Additional features or advanced techniques could enhance prediction accuracy.
- Interpretability with Feature Importances: The Mean Decrease in Impurity (MDI) based feature importances offer transparency into the factors influencing model predictions, aiding stakeholders in making informed decisions.
- CTC Prediction Challenges: The Decision Tree Regression's negative R-squared value indicates a poor fit to the data, reflecting challenges in predicting CTC accurately. The Random Forest Regression, while marginally better, requires further refinement.
- The dataset exhibits a significant class imbalance, primarily comprising candidates from the EECS division, with approximately 1100 out of 1400 candidates falling within this category.
- Future Model Refinement: Future work should focus on refining model parameters, exploring new features, and employing advanced techniques to address the limitations observed in both classification and regression models.
- Holistic View: The combination of classification and regression approaches provides a holistic view of student placement, considering both categorical slot allocation and quantitative CTC values.
- Overall Insights: The project offers valuable insights into the complex dynamics of student placement, paving the way for more informed decision-making in the allocation process. Continuous improvement and adaptation are essential for optimizing model performance.

4.1 Future scope of work

- To improve predictions for minority classes, such as candidates from departments like mechanical, management, chemical, aerospace, etc., strategies need to be employed to address

this imbalance effectively by addressing the core issue of class imbalance by enhancing the dataset with either new features and more data points.

- Considering the current imbalance, enhancing the predictive performance for underrepresented classes becomes crucial. One potential approach is to incorporate additional features, specifically leveraging scores from widely conducted online placement exams. This inclusion can contribute valuable insights into the candidates' abilities, especially those from departments with fewer representations in the dataset.
- Furthermore, it's important to acknowledge that while CGPA, Discipline, and Program provide foundational information, other features derived from candidates' courses and self-reported skills may not entirely capture a candidate's true competence during the placement season. Exploring and incorporating more robust indicators of skill and aptitude could contribute to a more comprehensive and accurate predictive model.