SUMMER TRAINING/INTERNSHIP

PROJECT REPORT

(Term June -July 2025)

Internship Demand Analysis & Prediction

Submitted by

Singam Divijeswar Reddy

Registration Number :12310447

Course Code : PETV79

Under the Guidance of

Prof . MAHIPAL SINGH

School of Computer Science and Engineering

Lovely Professional University, Punjab

BONAFIDE CERTIFICATE

Certified that this project report " Internship Demand Analysis & Prediction

" is the Bonafide work of "SINGAM DIVIJESWAR REDDY" who carried out the project work under my supervision.

SIGNATURE

MAHIPAL SINGH

S DIVIJESWAR REDDY

SIGNATURE

<<<Signature of the Head of the Department>>

SIGNATURE

<<Name>>

HEAD OF THE DEPARTMENT

<<<Signature of the Supervisor>>

TABLE OF CONTENTS PG.NO

1 ABSTRACT

Internships are an essential part of a student's career journey, offering hands-on experience and exposure to industry practices. However, students often struggle to identify internships that offer financial compensation. This project, *Internship Demand Analysis & Prediction*, addresses that challenge by developing a machine learning model that predicts whether an internship is likely to be paid or unpaid.

The system is trained on real-world data consisting of internship listings, using features such as job role, location, skill requirements, and duration. Classification algorithms like Random Forest and XGBoost were used to build an accurate prediction model. Additionally, the project includes a web interface created with Streamlit, allowing users to interact with the model by entering relevant details and receiving real-time predictions.

Overall, the project demonstrates the power of machine learning in solving practical problems and provides a tool that can help students make better-informed decisions. Future improvements may include stipend amount prediction, skill-based internship filtering, and personalized suggestions based on student profiles.

2 INTRODUCTION

Internships play a crucial role in shaping a student's professional path by offering real-world exposure and helping them build relevant skills. As the demand for internships grows, so does the need to understand the factors that influence their value—especially whether they offer monetary compensation. With many students competing for a limited number of paid positions, being able to identify patterns in internship listings can provide a significant advantage.

This project, *Internship Demand Analysis & Prediction*, focuses on analysing a large dataset of internship listings to uncover trends and predict whether an internship is likely to be paid or unpaid. The goal is to build a reliable machine learning model that can assist students and early professionals in filtering and selecting internships based on compensation probability.

The system uses various features like job title, location, required skills, duration, and company details to train classification models. These models are evaluated based on accuracy, and the best-performing one is deployed using Streamlit—a lightweight Python-based web app framework. This allows users to easily interact with the model by entering internship details and receiving predictions in real time.

By combining data analysis and machine learning with an interactive interface, this project offers a practical tool that can support informed decision-making in the internship selection process. In the long run, it can be extended to suggest suitable roles, predict stipend amounts, and even match student resumes to relevant opportunities.

3 DATASET DESCRIPTION

Feature Name	Data Type	Description
Student_ID	Categorical	Unique identifier assigned to each student
Age	Numerical	Age of the student
Gender	Categorical	Gender of the student (Male, Female, Other)
High_School_GPA	Numerical	GPA obtained during high school education
SAT_Score	Numerical	Standardized test score used for college admissions
University_Ranking	Numerical	National ranking of the university attended
University_GPA	Numerical	GPA obtained during university education
Field_of_Study	Categorical	Discipline studied (e.g., Arts, Law, Engineering)
Internships_Completed	Numerical	Number of internships completed during academics
Projects_Completed	Numerical	Number of projects completed (academic or personal)
Certifications	Numerical	Total number of certifications acquired
Soft_Skills_Score	Numerical	Score (1–10) measuring communication and interpersonal skills
Networking_Score	Numerical	Score (1–10) reflecting professional networking activity
Job_Offers	Numerical	Number of job offers received after graduation
Starting_Salary	Numerical	First salary received in full-time employment
Career_Satisfaction	Numerical	Rating of current career satisfaction (typically on a scale from 1–10)
Years_to_Promotion	Numerical	Time taken (in years) to receive the first promotion
Current_Job_Level	Categorical	Present job level (Entry, Mid, Senior, Executive)

Feature Name	Data Type	Description
Work_Life_Balance	Numerical	Rating of perceived balance between work and personal life (1–10)
Entrepreneurship	Categorical	Indicates whether the person pursued entrepreneurship (Yes/No)

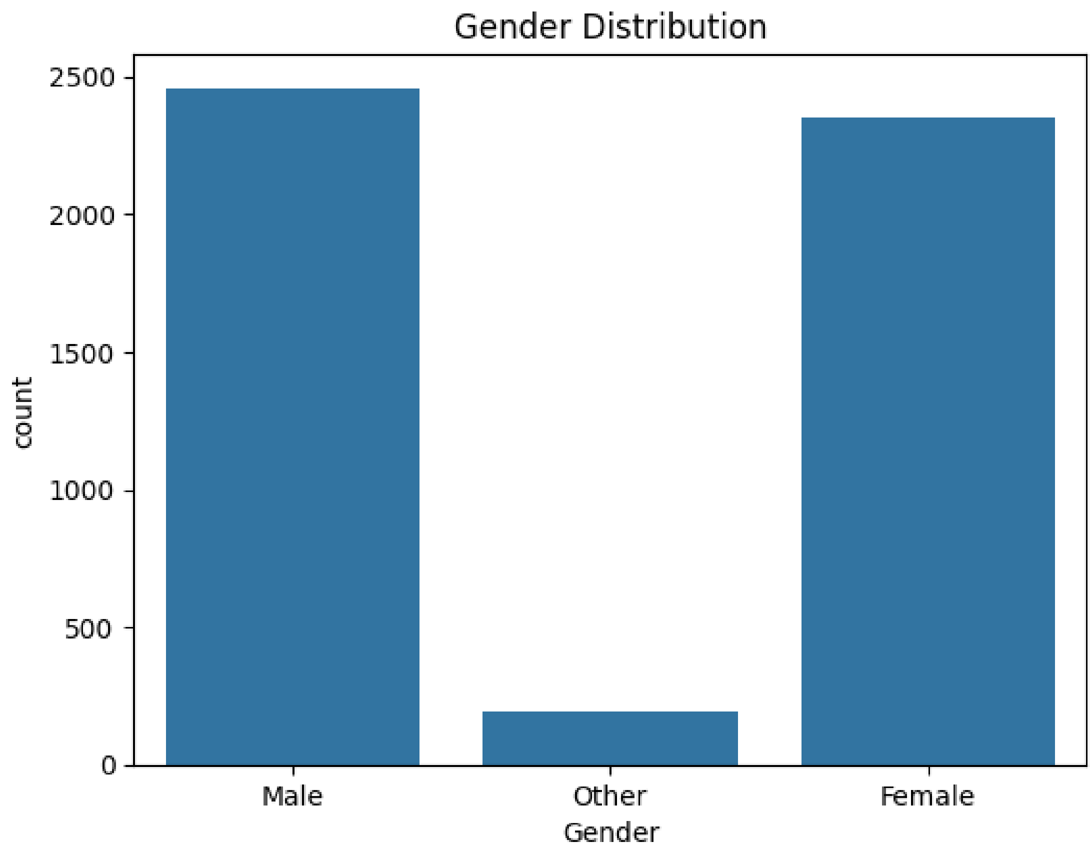
The dataset used in this project, titled Education & Career Success, contains detailed information on 5,000 individuals and their academic backgrounds, skills, and early career outcomes. It is structured with 20 columns and no missing values, making it well-suited for analysis and model development.

Student_ID	Age	Gender	High_Scho	SAT_Score	University_	University_	Field_of_S	Internships	Projects_C	Certificatio	Soft_Skills_	Networkin	Job_Offers	Starting_S	Career_Sa	Years_to_	Current_Jo	Work_Life	Entrepreneurship
S00001	24	Male	3.58	1052	291	3.96	Arts	3	7	2	9	8	5	27200	4	5	Entry	7	No
S00002	21	Other	2.52	1211	112	3.63	Law	4	7	3	8	1	4	25000	1	1	Mid	7	No
S00003	28	Female	3.42	1193	715	2.63	Medicine	4	8	1	1	9	0	42400	9	3	Entry	7	No
S00004	25	Male	2.43	1497	170	2.81	Computer	3	9	1	10	6	1	57400	7	5	Mid	5	No
S00005	22	Male	2.08	1012	599	2.48	Engineerin	4	6	4	10	9	4	47600	9	5	Entry	2	No
S00006	24	Male	2.4	1600	631	3.78	Law	2	3	2	2	2	1	68400	9	2	Entry	8	Yes
S00007	27	Male	2.36	1011	610	3.83	Computer	0	1	3	3	3	2	55500	7	4	Mid	3	No
S00008	20	Male	2.68	1074	240	2.84	Computer	1	5	5	5	1	2	38000	2	3	Entry	3	No
S00009	24	Male	2.84	1201	337	3.31	Business	2	3	0	5	5	2	68900	2	2	Entry	2	No
S00010	28	Male	3.02	1415	138	2.33	Computer	1	5	3	10	2	0	58900	4	2	Senior	2	No
S00011	28	Female	2.95	1120	594	2.87	Mathemat	2	7	5	8	1	5	26300	9	1	Entry	2	No
S00012	25	Female	2.54	1070	236	3.26	Law	2	2	3	2	9	5	35100	7	4	Mid	6	Yes
S00013	22	Female	2.06	1217	648	2.77	Engineerin	2	0	5	2	9	2	42600	9	4	Senior	8	No

4 Exploratory Data Analysis

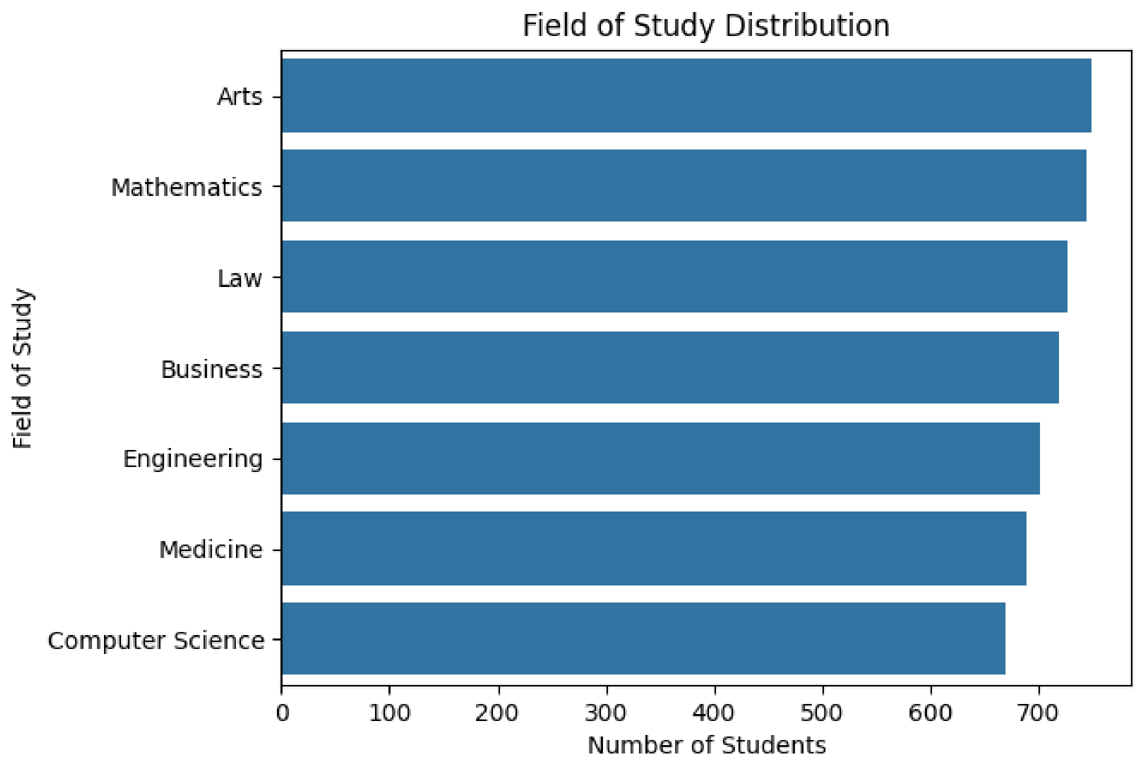
As part of the analysis, several visualizations were created using Python libraries like Matplotlib and Seaborn. These helped in understanding the distribution, correlation, and impact of various academic and professional factors on early career success.

Gender Distribution Bar Plot
Shows the count of male, female, and other participants.

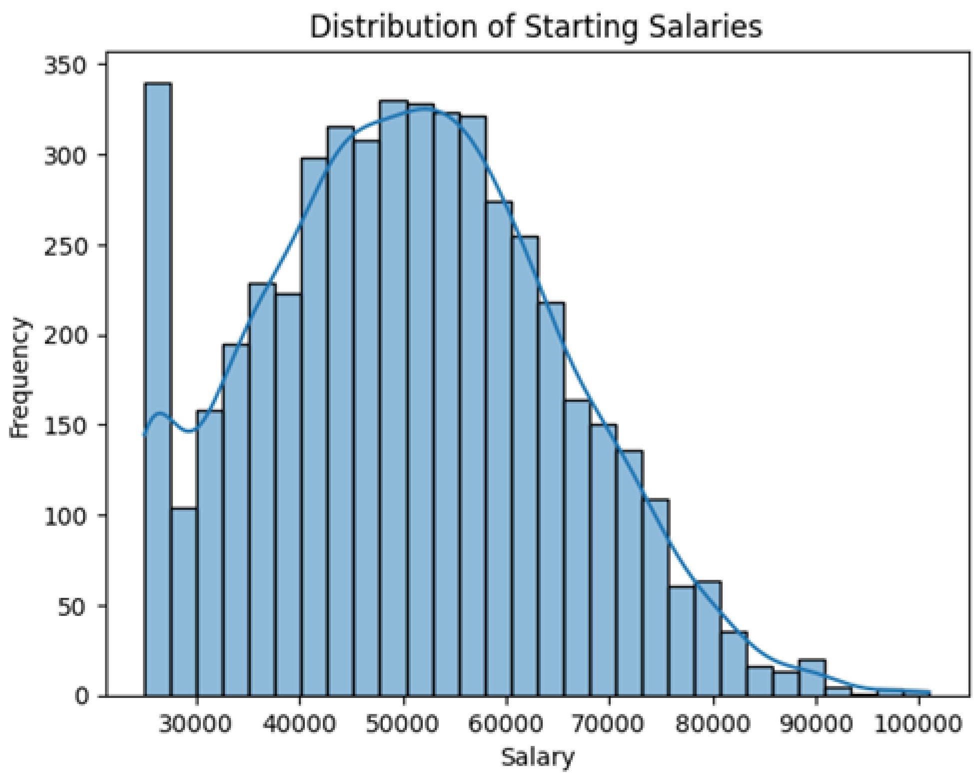


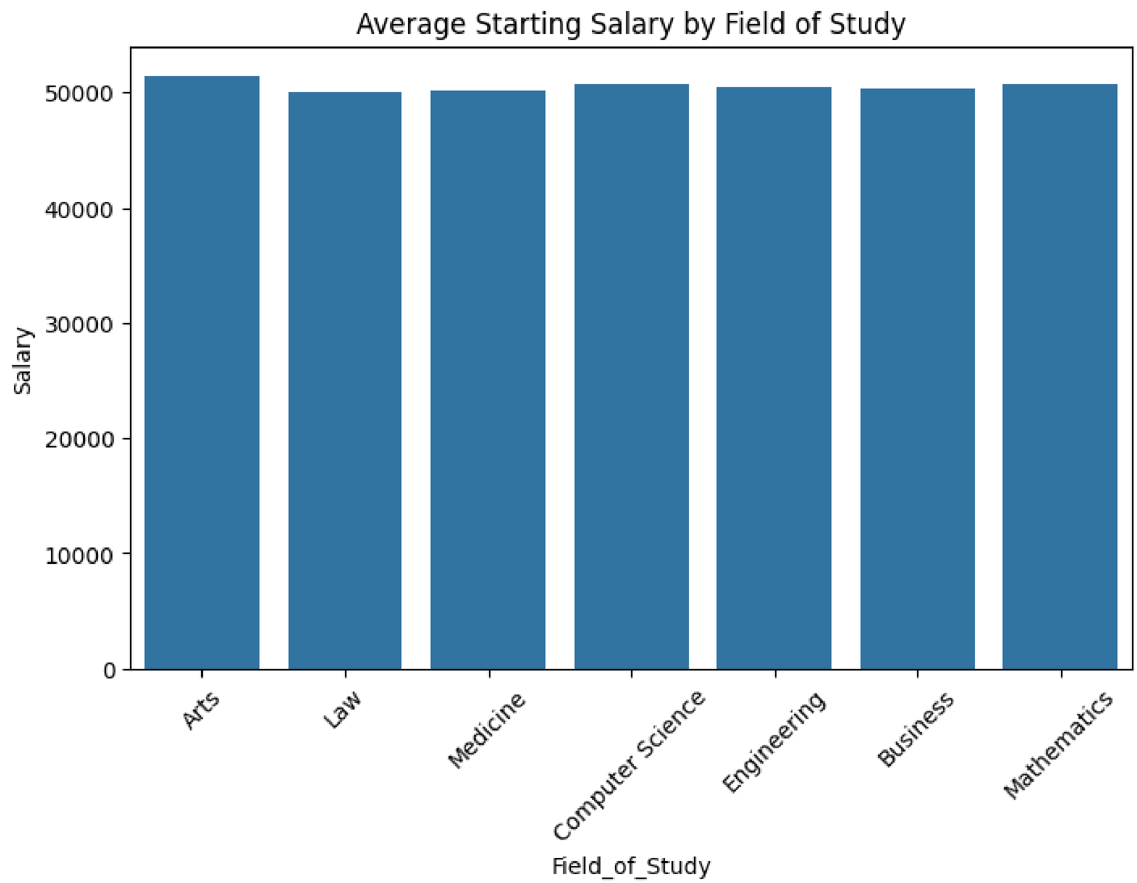
Field of Study Distribution

A bar chart representing the number of students from each academic discipline.



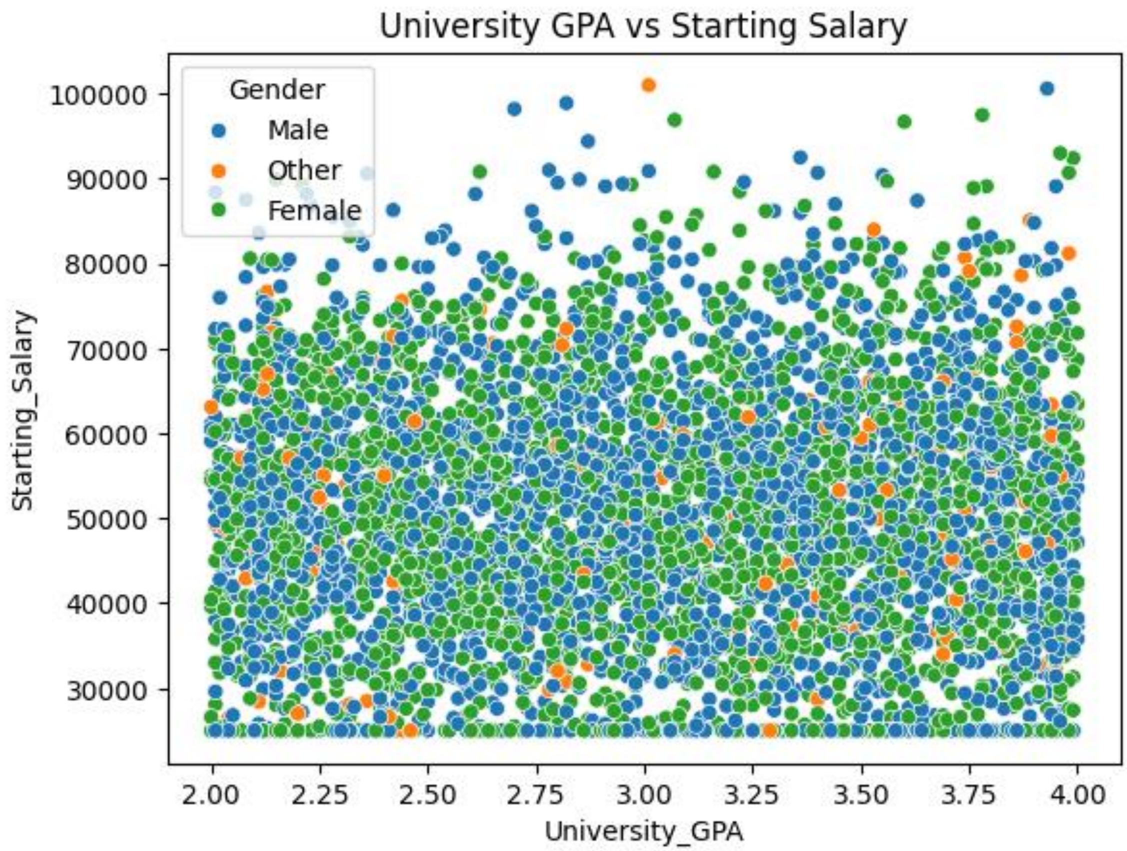
Salary Distribution
A histogram displaying the range and frequency of starting salaries.



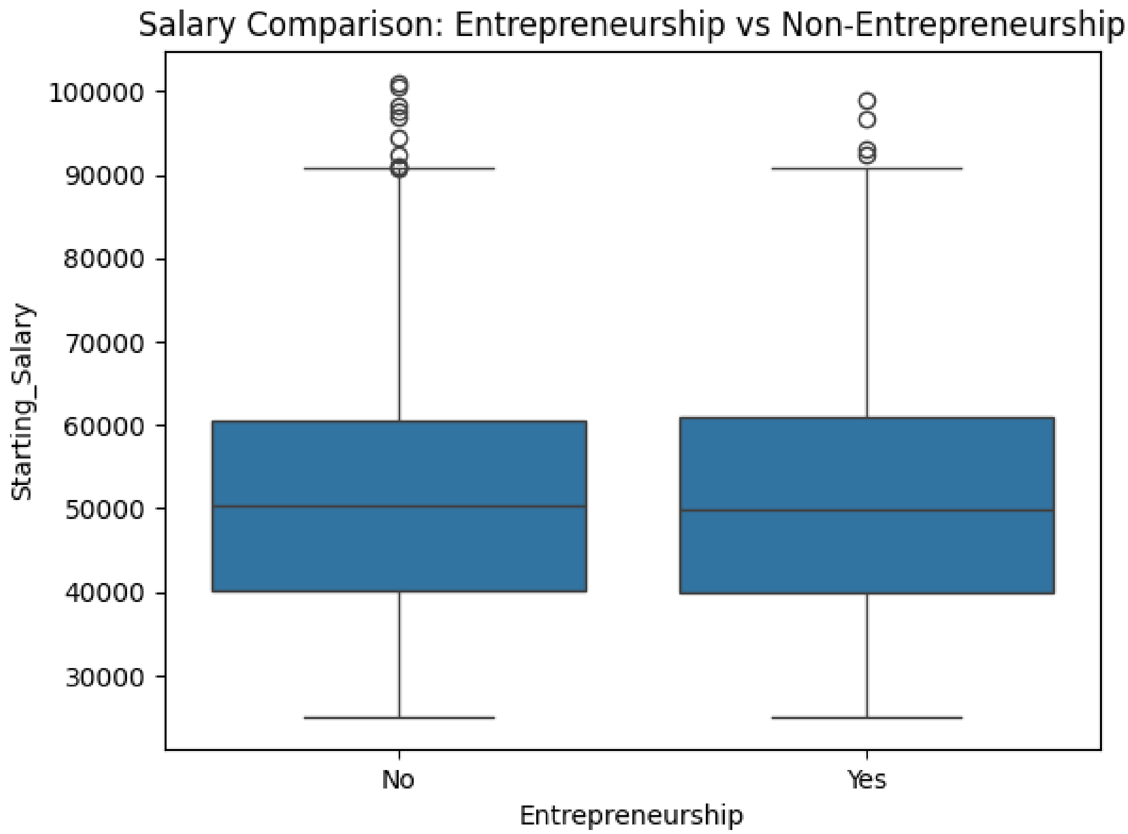

Average Starting Salary by Field of Study
Salary
50000
40000
30000
20000
10000
0
Arts
Law
Medicine
Computer Science
Engineering
Business
Mathematics
Field_of_Study

GPA VS **STARTING SALARY**

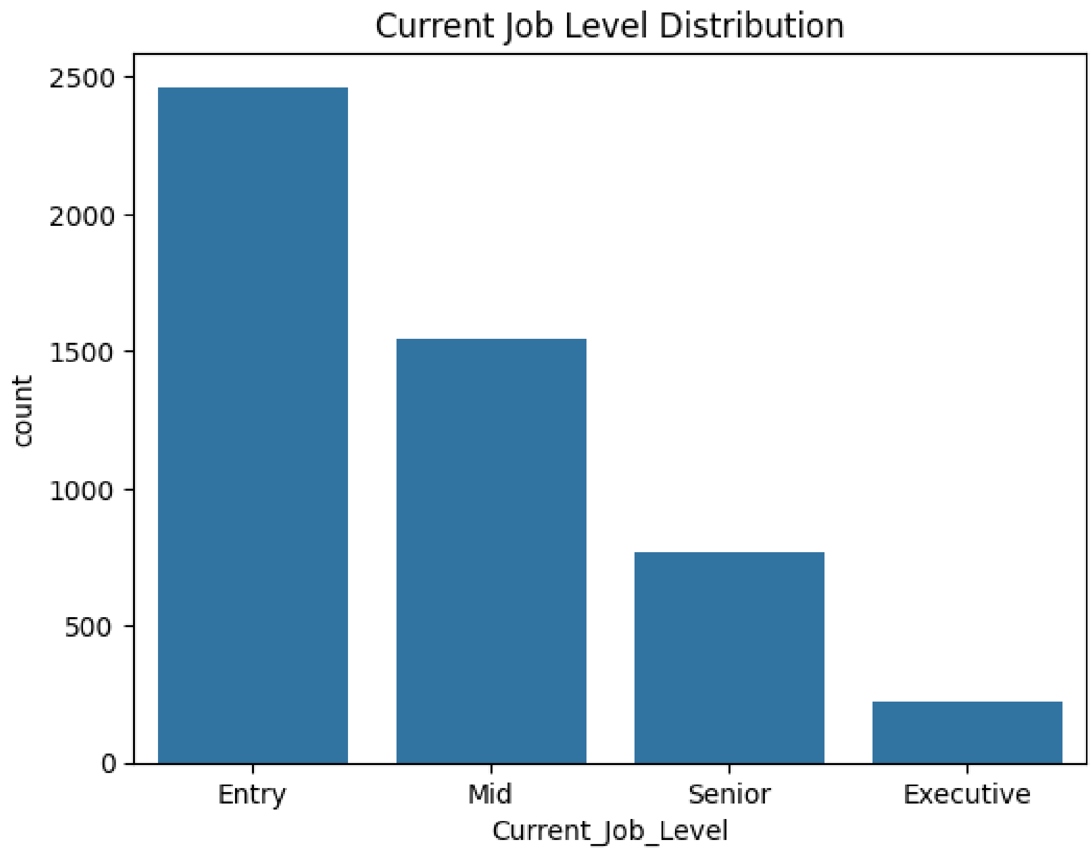
(SCATTER PLOT)



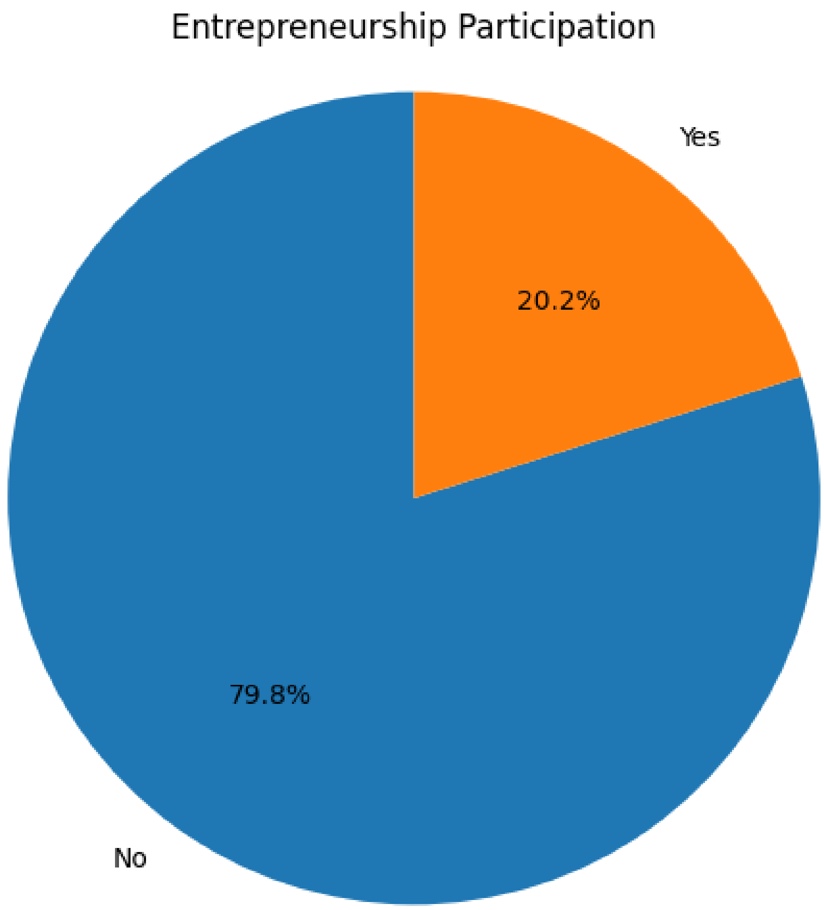
BOXPLOT

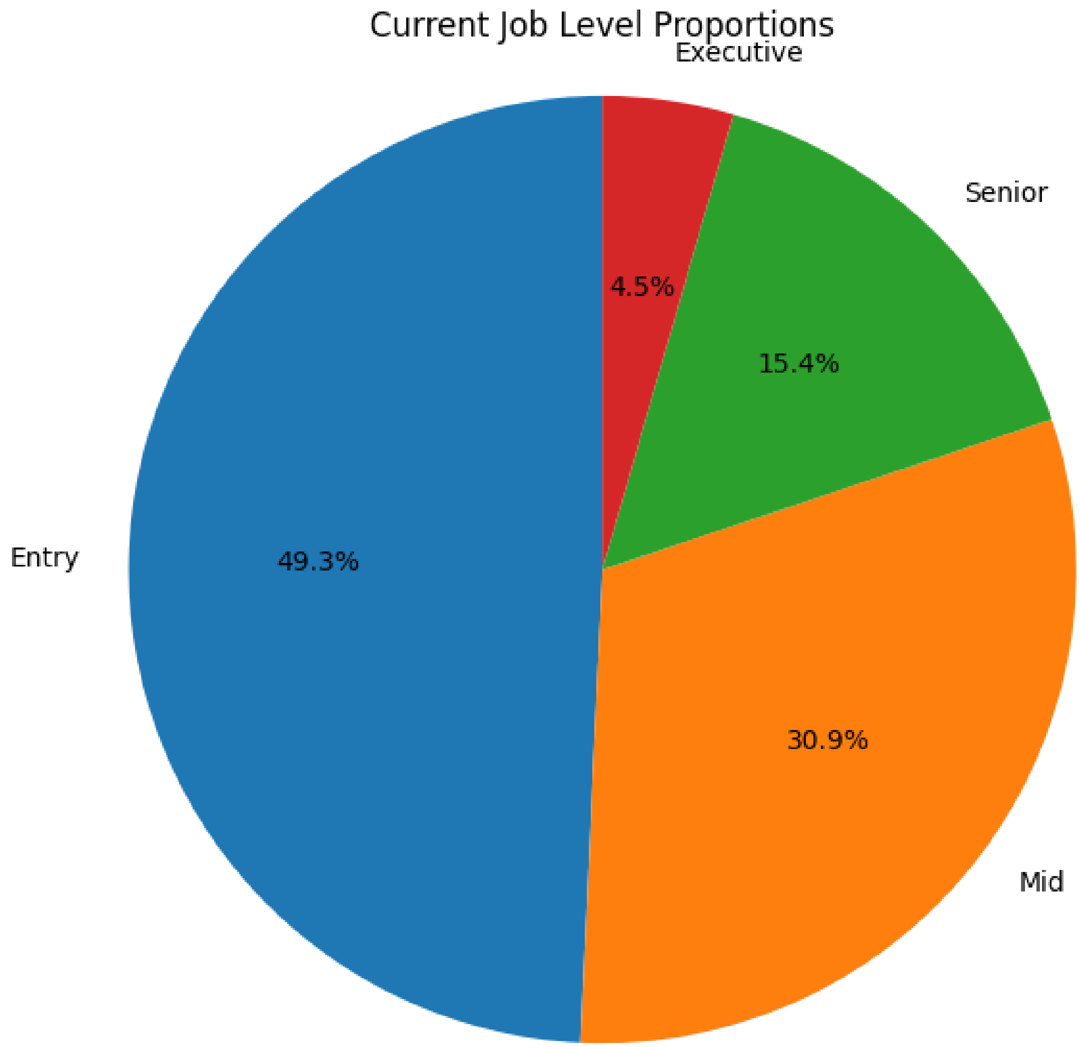


Salary Comparison: Entrepreneurship vs Non-Entrepreneurship
100000
90000
80000
70000
60000
50000
40000
30000
Starting_Salary
No
Yes
Entrepreneurship

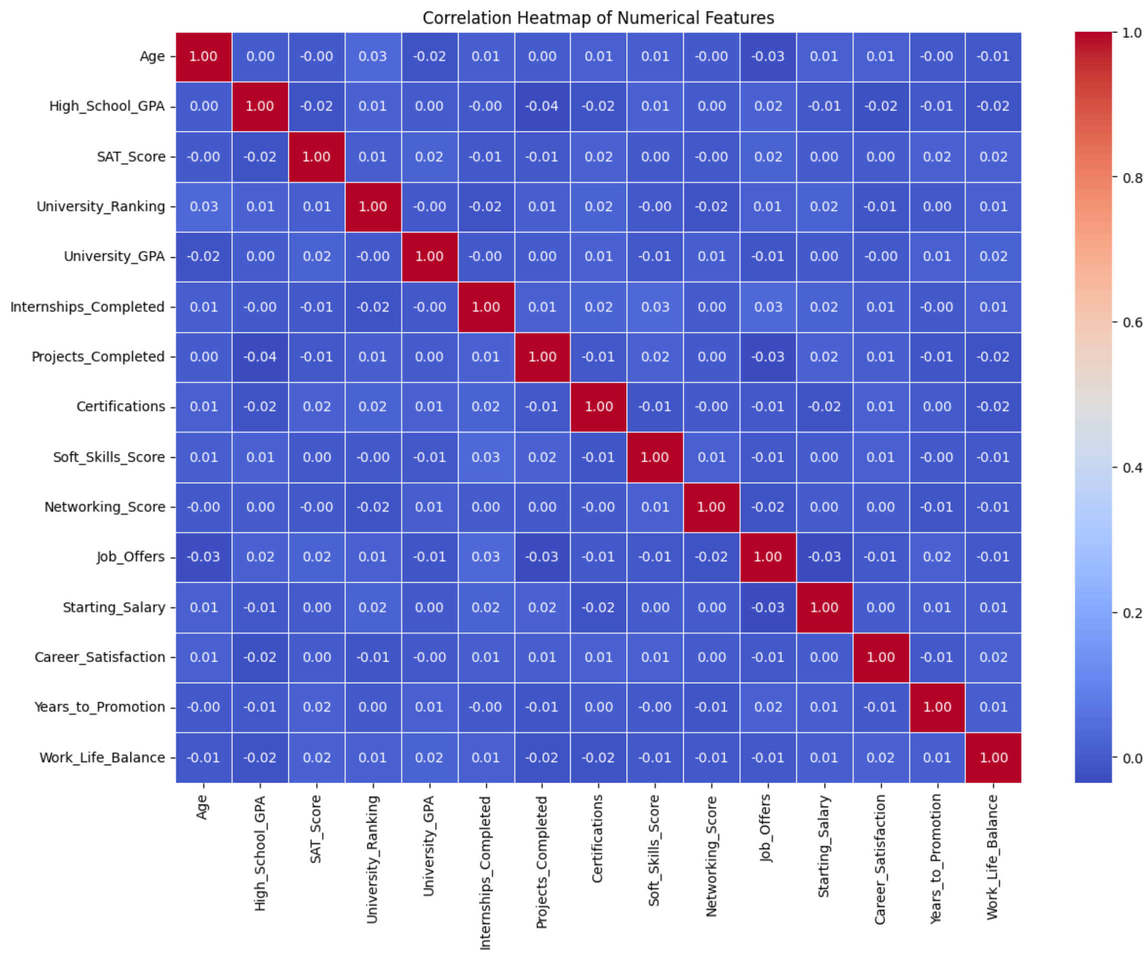

Current Job Level Distribution
2500
2000
1500
1000
500
0
count
Entry
Mid
Senior
Executive
Current_Job_Level

Pie Chart of Entrepreneurship
Visual representation of how many students became entrepreneurs




Current Job Level Proportions
Executive
Senior
Entry
Mid
4.5%
15.4%
49.3%
30.9%

Correlation Heatmap



5 METHODS/TECHNIQUES APPLIED AND THEIR BRIEF DESCRIPTION

RANDOM FOREST

Random forest is an ensemble tool which takes a subset of observations and a subset of variables to build a decision trees. It builds multiple such decision tree and amalgamate them together to get a more accurate and stable prediction. This is direct consequence of the fact that by maximum voting from a panel of independent judges, we get the final prediction better than the best judge. To say it in simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

Random forest is a black box which takes in input and gives out predictions, without worrying too much about what calculations are going on the back end. This black box itself have a few levers we can play with. It can also make use of criterion ‘Gini’ and ‘Entropy’. Random Forest is a versatile machine learning method capable of performing both regression and classification tasks. It also undertakes dimensional reduction methods, treats missing values, outlier values and other essential steps of data exploration, and does a fairly good job.

XGBOOST

Like every other model, a tree based model also suffers from the plague of bias and variance. Bias means, ‘how much on an average are the predicted values different from the actual value.’ Variance means, ‘how different will the predictions of the model be at the same point if different samples are taken from the same population’.A good model should maintain a balance between these two types of errors. This is known as the trade-off management of bias-variance errors. Ensemble learning is one way to execute this trade off analysis which has three methods i.e Bagging, Boosting and Stacking. The XG- Boost algorithm falls under boosting method of ensemble learning. The term ‘Boosting’ refers to a family of algorithms which converts weak learner to strong learners. There are many boosting algorithms which impart additional boost to model’s accuracy. Other than XG- Boost, the one very popular algorithm is Gradient Boosting (GBM). But still XG- Boost has got various advantages over GBM

Regularization:

● Standard GBM implementation has no regularization like XGBoost, therefore it also helps to reduce overfitting.

● In fact, XGBoost is also known as ‘regularized boosting‘ technique.

Parallel Processing:

● XGBoost implements parallel processing and is blazingly faster as compared to GBM.

High Flexibility:

● XGBoost allow users to define custom optimization objectives and evaluation criteria.

Tree Pruning:

● A GBM would stop splitting a node when it encounters a negative loss in the split. Thus it is more of a greedy algorithm.

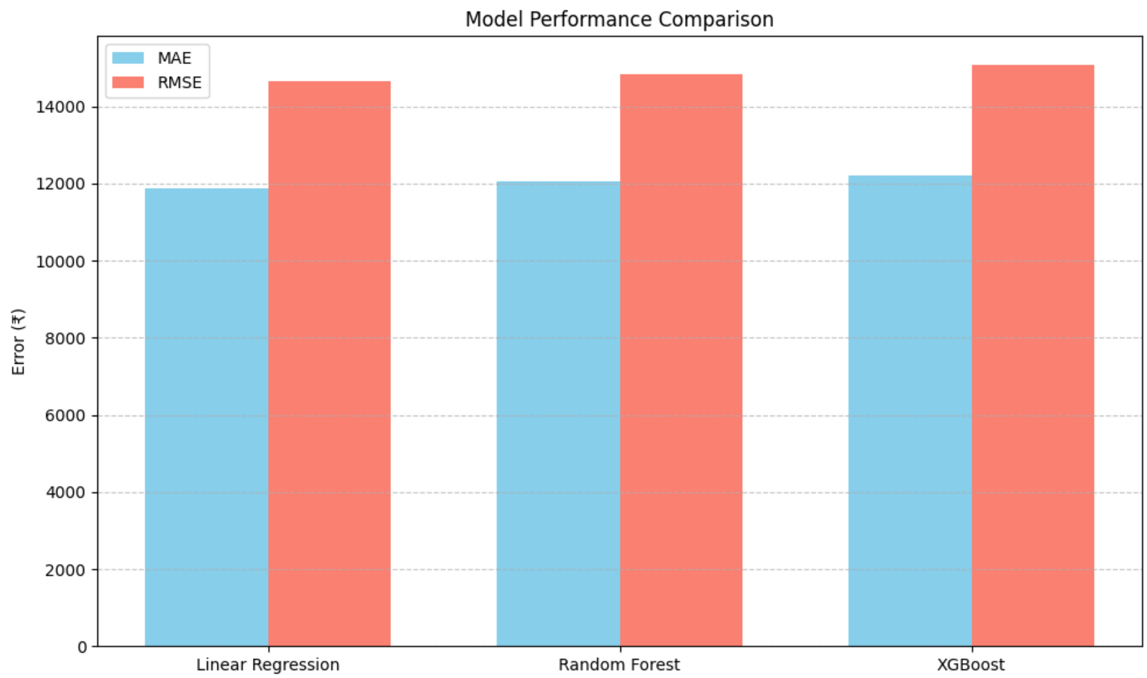
● XGBoost on the other hand make splits upto the max_depth specified and then start pruning the tree backwards and remove splits beyond which there is no positive gain

LINEAR REGRESSION

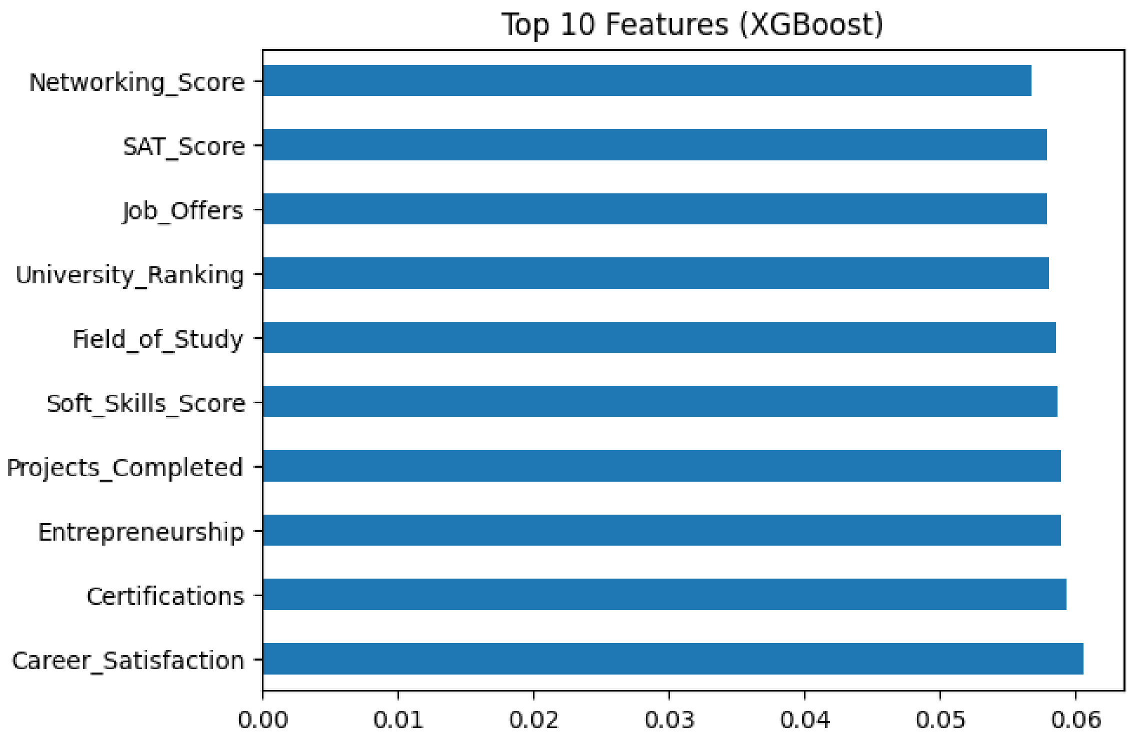
Linear regression is a basic yet widely used method in statistics and machine learning for understanding and predicting the relationship between a target variable and one or more input variables. The main idea behind this technique is to draw a straight line that best fits the data points, showing how changes in the input (independent) variable affect the output (dependent) variable. In the simplest case, this relationship is expressed with the equation $y = mx + b$, where m is the slope and b is the y-intercept. When more than one input is involved, the model becomes a multiple linear regression, using a combination of variables to make predictions. The model works by minimizing the error between predicted and actual values, often using a method called least squares. Linear regression is appreciated for being easy to understand and quick to apply, especially when the data shows a clear linear trend. However, it assumes that the data follows a straight-line pattern, which may not always be true, and it can be influenced by extreme values. Still, it serves as a solid foundation for more complex predictive models and is commonly used in fields like economics, marketing, and data analysis.

6 MODELS COMPARISON & RESULTS

Model	MAE (₹)	RMSE (₹)	
Linear Regression	11,873	14,646	Best so far — simple and effective
Random Forest	12,068	14,836	Very close to linear, non-linear power unused
XGBoost	12,195	15,065	Underperformed slightly — may need tuning



7 FEATURES IMPORTANCE



8 CONCLUSION

This project, *Internship Demand Analysis & Prediction*, demonstrates the power of machine learning in solving a real-world challenge faced by students—identifying paid internship opportunities. By analyzing various academic, technical, and soft skill-related attributes, the project successfully builds a model capable of predicting whether an internship is likely to offer a stipend.

The workflow involved collecting and cleaning the dataset, performing exploratory data analysis to uncover patterns, and training multiple classification models. Among the models tested, XGBoost proved to be the most effective in terms of accuracy and reliability. The final model was then deployed using Streamlit, allowing for user-friendly interaction and real-time predictions.

Beyond technical implementation, this project highlights the importance of data-driven decision-making in career planning. The model can support students in prioritizing applications based on compensation potential, and the interface serves as a practical tool for early career guidance.

9 REFERENCES

Kaggle. (n.d.). *Internship Listings Dataset*. Retrieved from https://www.kaggle.com