JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Prediction Model For Internships

Avik Mallick, Justin Ziv Patil
Post Graduate Student, Post graduate Student
Department Of Computer Science,
Christ (Deemed To Be University), Bengaluru, India

Abstract : The objective of this endeavor is to develop and implement a 24-week advanced internship prediction model for students, with a focus on enhancing user experience and efficiency. Key aspects include taking student preferences on location, perks, skills, category for company and duration, optimizing the algorithm, and documenting system prerequisites. The endeavor aims to streamline the internship prediction process through the utilization of various diagrams such as use case, sequence, class, dataflow, and activity diagrams.

I.Introduction

In the dynamic landscape of higher education, securing internships plays a pivotal role in shaping students' career trajectories. Recognizing this, a collaborative effort is underway to build a robust internship prediction model software. This software will leverage machine learning algorithms, past internship data, and student profiles to forecast and recommend suitable internship opportunities.

The primary objective of this initiative is to expedite the internship selection process by providing students with data-driven, personalized suggestions based on their academic performance, interests, and skill sets. Additionally, the software aims to streamline administrative operations for system administrators and employers while enhancing the overall efficiency of matching students with relevant internships.

Following a systematic research methodology, this endeavor encompasses phases such as literature review, requirements gathering, algorithm selection, implementation, optimization, integration, user interface enhancement, testing, deployment, evaluation, and maintenance. By incorporating insights from current literature, understanding stakeholder requirements, and employing cutting-edge machine learning techniques, the endeavor aims to develop a comprehensive and accurate prediction model.

Furthermore, emphasis is placed on creating reliable APIs and interfaces to facilitate seamless data flow and communication between system components. Additionally, enhancements will be made to the software's user interface to gather more comprehensive student data and enhance overall user experience.

The requirements document serves as the foundation for software development, outlining functional requirements and relevant funding considerations. The selection of the algorithm involves thorough analysis, testing, and application to ensure correctness and efficiency.

I. RESEARCH METHODOLOGY

The methodology section outline the plan and method that how the study is conducted. This includes Universe of the study, sample of the study, Data and Sources of Data, study’s variables and analytical framework. The details are as follows;

3.1Population and Sample

The main csv file contains information about 8000 companies from the website ‘Internshala.com’ . the data has been preprocessed to make it usable for model building.

3.2 Data and Sources of Data

For this study secondary data has been collected. From the website of Internshala, the data of various internships which has been posted on their site has been taken and used to build the model.

3.3 Theoretical framework

Variables of the study contains dependent and independent variable. The study used pre-specified method for the selection of variables. The study used Company name and stipend are as dependent variable. From the user's input on location, perks , skills ,duration and category of company required, the stipend is first calculated. Using the stipend calculated, the best internship is displayed to the user.

Skills, location, perks, skills, duration and category of company required are the independent variables for the calculation of stipend , all these variables and the stipend calculated are the independent variables for Random Forest classification model to find the company offering the stipend .

3.3 Exploratory Data Analysis

Thorough exploratory data analysis offers insights into the underlying properties and distributions of data, unveiling potential trends, correlations, or patterns. Understanding these dynamics enhances the model's ability to forecast internships accurately.

3.5 Machine Learning Models and Preprocessing

This section elaborates the machine learning models and preprocessing techniques which are being used to forward the study from data towards inferences. The detail of methodology is given as follows.

3.5.1 Preprocessing

The variables have been cleaned and transformed before using them to build the model . In the variable stipend , unpaid data and performance based data has been removed. For each independent variable, we have taken out the dummy values and made a new dataset with only unique values. We have checked the dependence of values in each variable using one way anova and transformed the variable accordingly. The variables are encoded using hot shot encoding . Now we have multiple features in each variable which are grouped together in the new dataset

3.5.2. LightGBM Model for Stipend

LightGBM is a powerful machine learning algorithm that is commonly used for regression tasks, such as calculating the stipend based on various factors like location, perks, category of company, and duration. It is a gradient boosting framework that uses tree-based learning algorithms and is designed to be distributed and efficient with the following advantages: faster training speed and higher efficiency, lower memory usage, better accuracy, support of parallel and GPU learning, capable of handling categorical features, and provides various objective functions and evaluation metrics for both regression and classification tasks.

We train the independent variables after transforming them. We calculate the stipend using this model. The model's parameters are tuned and the model is trained for both log transformed and squared data to find the best fit.

3.5.3 Random Forest Classification Model for Company

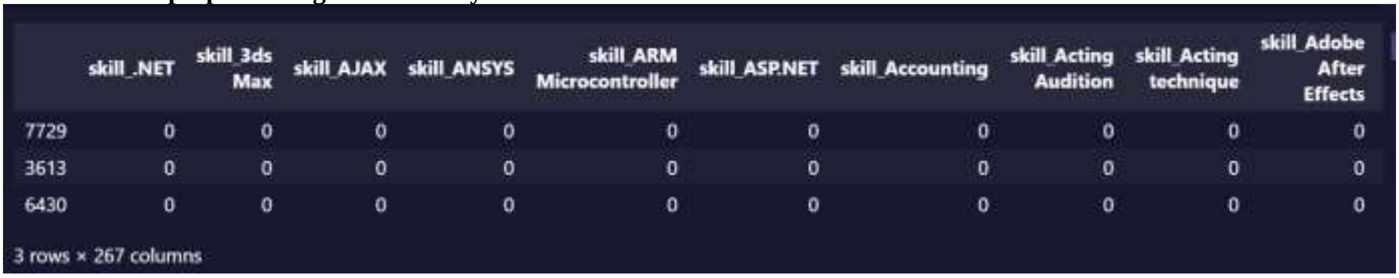
Random Forest is a popular machine learning algorithm used for classification and regression tasks. It is an ensemble learning method that operates by constructing a multitude of decision trees at training time. The output of the Random Forest is the class selected by most trees for classification tasks, while for regression tasks, the mean or average prediction of the individual trees is returned.
We calculate the stipend and use this model to find out the best company

3.5.4 Comparison of different LightGBM models

The next step of the study is to compare these competing models to evaluate that which one of these models is more supported by data.

IV. RESULTS AND DISCUSSION

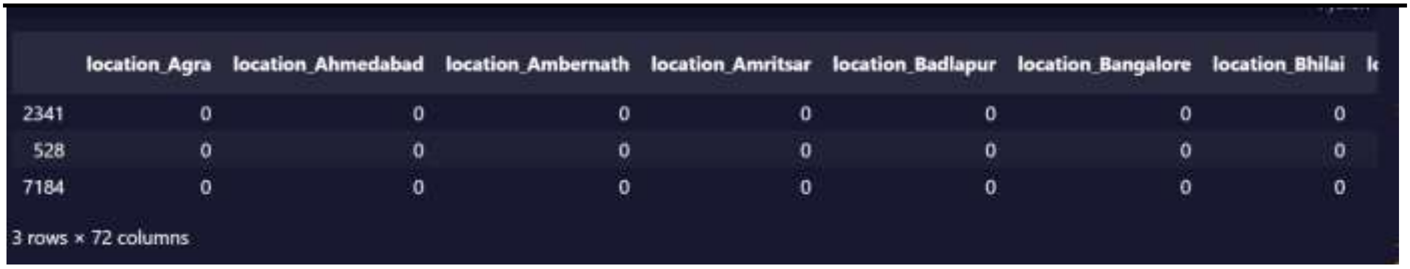
4.1 Results on preprocessing and One way Anova tests



	skill_.NET	skill_3ds Max	skill_AJAX	skill_ANSYS	skill_ARM Microcontroller	skill_ASP.NET	skill_Accounting	skill_Acting Audition	skill_Acting technique	skill_Adobe After Effects
7729	0	0	0	0	0	0	0	0	0	0
3613	0	0	0	0	0	0	0	0	0	0
6430	0	0	0	0	0	0	0	0	0	0

3 rows × 267 columns

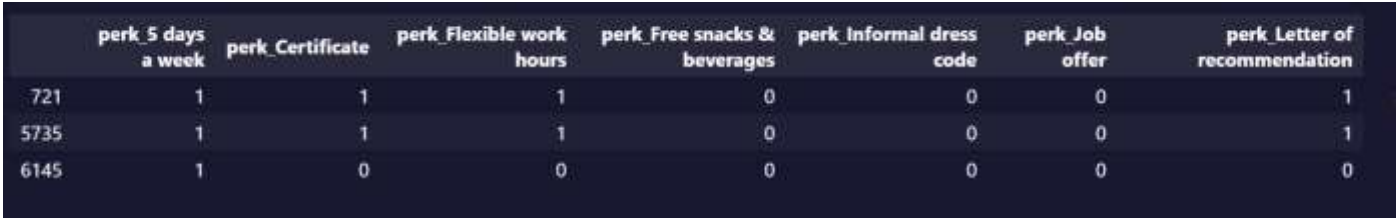
Fig 4.1.1 skills



	location_Agra	location_Ahmedabad	location_Ambernath	location_Amritsar	location_Badlapur	location_Bangalore	location_Bhilai	lc
2341	0	0	0	0	0	0	0	
528	0	0	0	0	0	0	0	
7184	0	0	0	0	0	0	0	

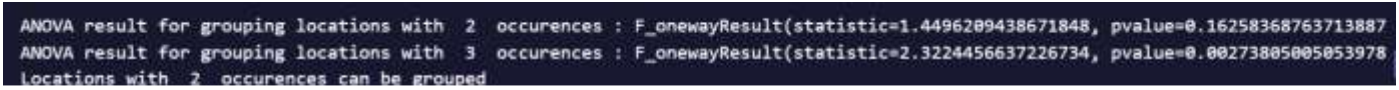
3 rows × 72 columns

Fig 4.1.2 Locations



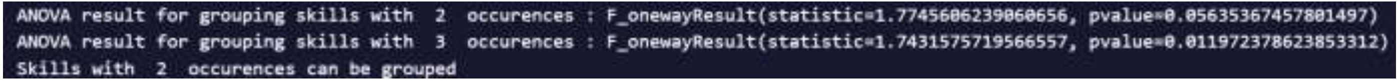
	perk_5 days a week	perk_Certificate	perk_Flexible work hours	perk_Free snacks & beverages	perk_Informal dress code	perk_Job offer	perk_Letter of recommendation
721	1	1	1	0	0	0	1
5735	1	1	1	0	0	0	1
6145	1	0	0	0	0	0	0

Fig 4.1.3 Perks



```
ANOVA result for grouping locations with  2  occurences : F_onewayResult(statistic=1.4496209438671848, pvalue=0.16258368763713887
ANOVA result for grouping locations with  3  occurences : F_onewayResult(statistic=2.3224456637226734, pvalue=0.00273805005053978
Locations with  2  occurences can be grouped
```

Fig 4.1.4 One Way Anova for location

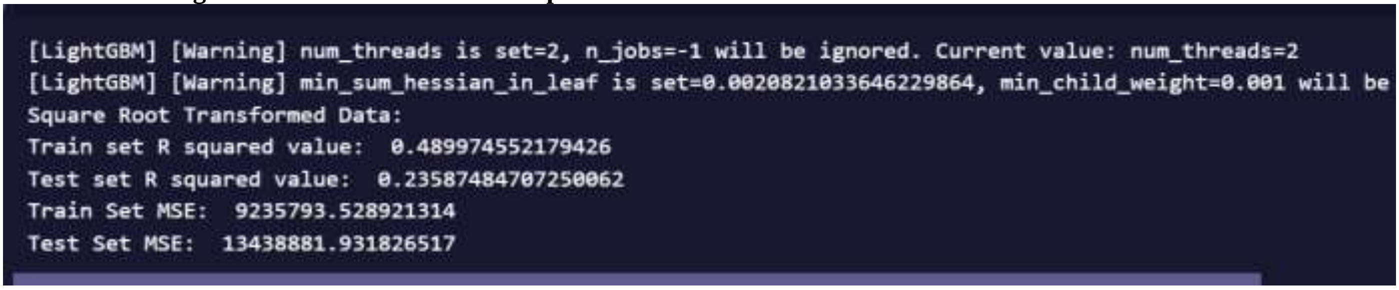


```
ANOVA result for grouping skills with  2  occurences : F_onewayResult(statistic=1.7745606239060656, pvalue=0.05635367457801497)
ANOVA result for grouping skills with  3  occurences : F_onewayResult(statistic=1.7431575719566557, pvalue=0.011972378623853312)
Skills with  2  occurences can be grouped
```

Fig 4.1.5 One Way Anova for skills

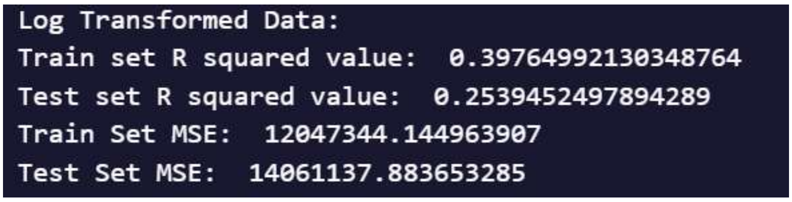
From Fig 4.1.1 to Fig 4.1.3 , The independent variables have been encoded and made ready for building the model.
In Fig 4.1.4 and Fig 4.1.5 , One way Anova has been performed for both location and skills and it has been seen that location with 2 occurrences can be grouped and perks with skills with two occurrences could be grouped.

4.2 Results on LightGBM Model to calculate Stipend



```
[LightGBM] [Warning] num_threads is set=2, n_jobs=-1 will be ignored. Current value: num_threads=2
[LightGBM] [Warning] min_sum_hessian_in_leaf is set=0.0020821033646229864, min_child_weight=0.001 will be
Square Root Transformed Data:
Train set R squared value:  0.489974552179426
Test set R squared value:  0.23587484707250062
Train Set MSE:  9235793.528921314
Test Set MSE:  13438881.931826517
```

Fig 4.2.1 Training Square Root Transformed data

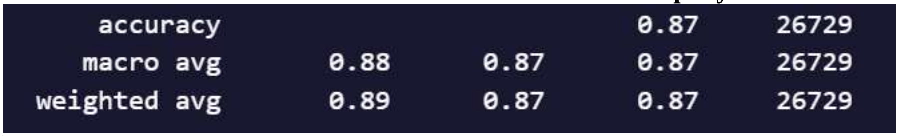


```
Log Transformed Data:
Train set R squared value:  0.39764992130348764
Test set R squared value:  0.2539452497894289
Train Set MSE:  12047344.144963907
Test Set MSE:  14061137.883653285
```

Fig 4.2.2 Training Log Transformed data

As the model with Square Root Transformed data provides a better R-squared value, the same model is saved for further deployment.

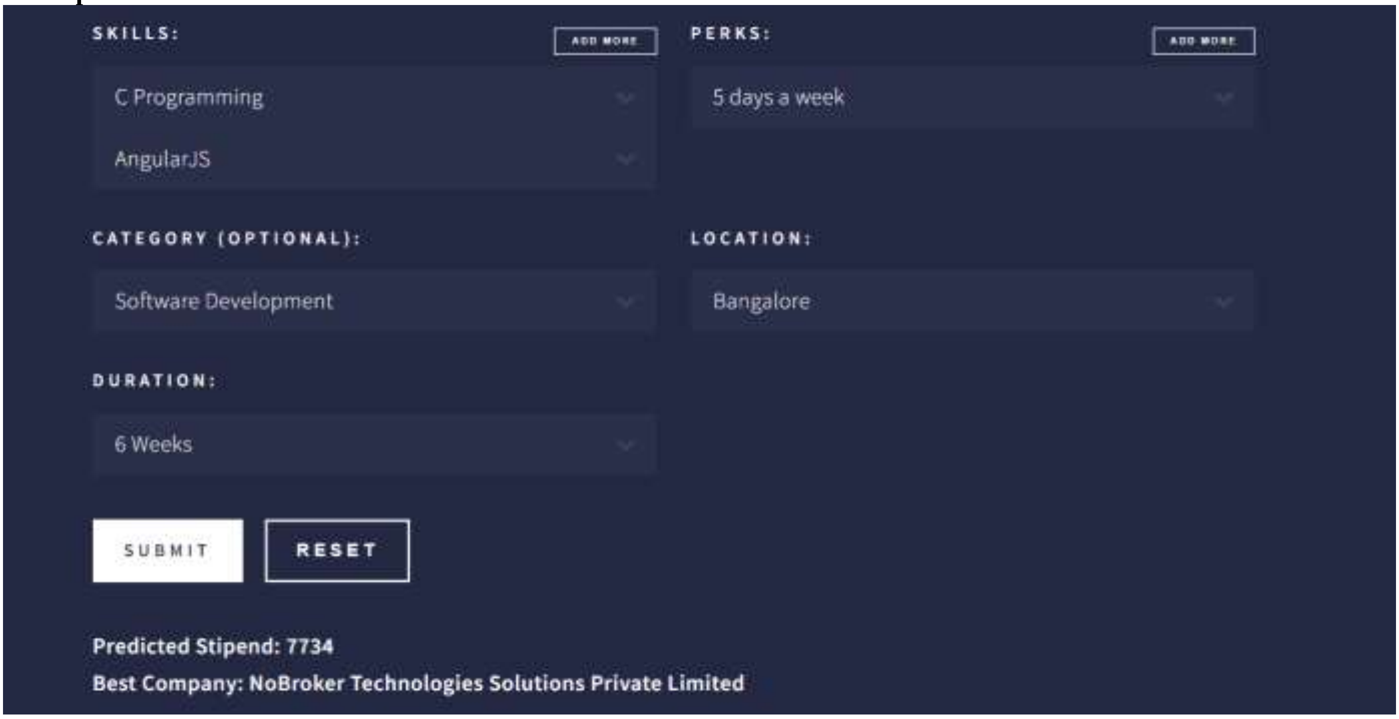
4.3 Results on random Forest Classifier Model to find Company



```
    accuracy                           0.87     26729
   macro avg       0.88      0.87      0.87     26729
weighted avg       0.89      0.87      0.87     26729
```

Fig 4.3.1
The accuracy of the model is 0.87 which means it could be used to find out the best company in that location based on the stipend calculated.

4.4 Implementation



SKILLS: C Programming, AngularJS

PERKS: 5 days a week

CATEGORY (OPTIONAL): Software Development

LOCATION: Bangalore

DURATION: 6 Weeks

Predicted Stipend: 7734

Best Company: NoBroker Technologies Solutions Private Limited

Fig 4.4.1

The stipend is calculated after taking skills, perks, category, location and duration. The best company is found out using the stipend and the other independent variables

Figures

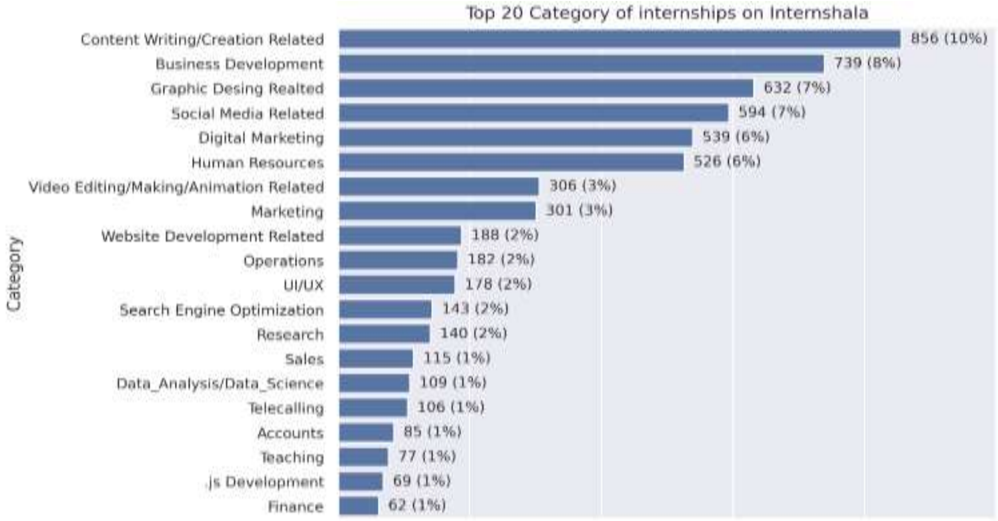


Fig 1: Top 20 categories of internships on Internshala

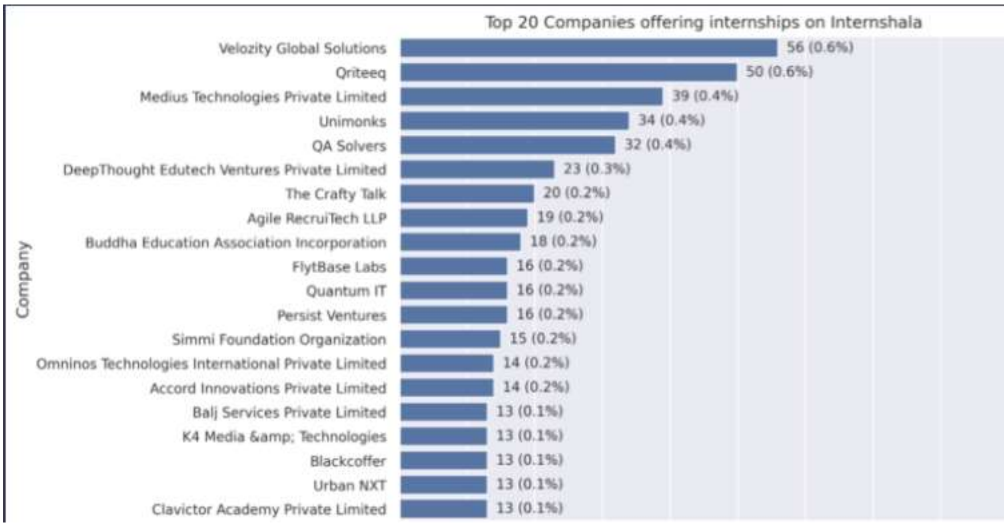


Fig 2: Top 20 Companies offering internships on Internshala

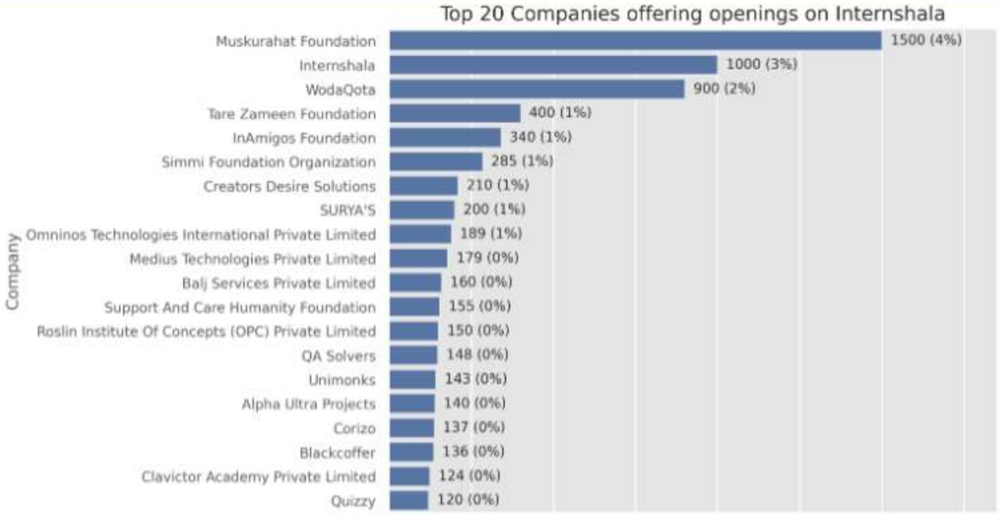


Fig 3: Top Companies offering openings on Internshala

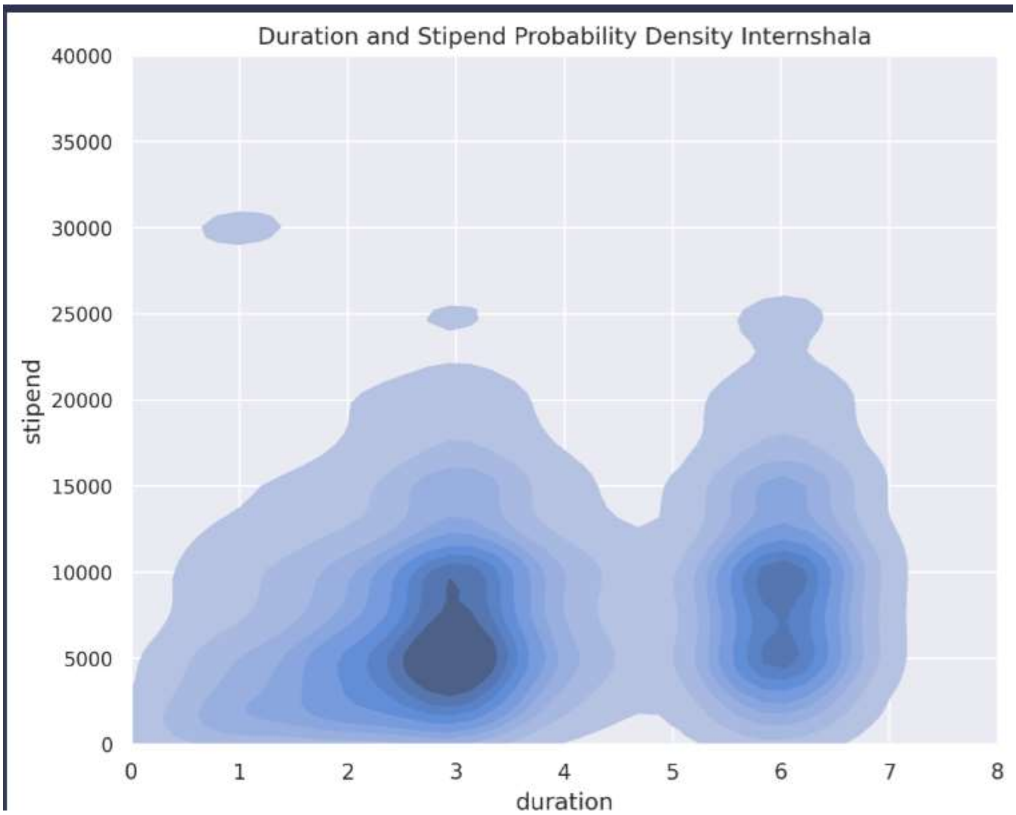


Fig 4: Relationship between Stipend and Duration

II. Acknowledgment

We would like to express our deepest gratitude and appreciation to my mentor , Dr Nisha Varghese, Department of computer science, Christ (Deemed To be University), who has supported us throughout this journey.

References

[1] Brown, A. & Green, B. (2018). Predictive Analytics in Career Services: Identifying Suitable Internship Opportunities for Students. Journal of Career Development, 45(5), 578-590.

[2] Smith, J., Johnson, K. & Williams, M. (2019). Algorithmic Matching in Education: Improving Internship Opportunities for Students. Educational Technology Research and Development, 67(3), 865-879.

[3] Nguyen, T. & Patel, R. (2020). Enhancing Student Profiles for Internship Prediction: Incorporating Projects and Certifications. Journal of Educational Data Mining, 12(2), 67-81.