

Salary Prediction for Software Developers

Data Exploration, Model Building
and Implementation

Presented by:

Sherly Sinha (04)

Paridhi Agarwal (35)

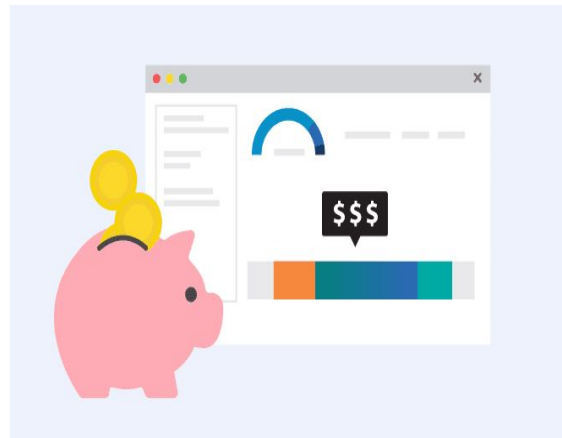
Under the guidance and supervision of,
Prof. Utpal Kumar Ray

Overview

- Need for a Salary Predictor
- Timeline
- Literature Review
- Implementation
- Results
- Conclusion

Need for a Salary Predictor!

- Budding software developers around the world struggle to understand their true market value. Traditional education system trains students for the skills required to begin their careers. However, it often fails to educate them about the job market.
- This project aims to solve this problem for millions of software developers around the world and help them make informed decisions before choosing their careers.
- In this project, we aimed to understand the salary distribution, community inclusiveness, and gender distribution of software developers around the world.



TIMELINE

Intro to Data Science

What is Data Science?
Applications, Steps, Data Scientists

Project Ideas

Brainstorm a list of different project ideas, Brief overlook of the dataset available in each case

Data Science Steps and Models

General steps to be followed,
Popular models with the idea behind them

Salary Predictor

Problem statement, Project Idea, Steps to be followed



Literature Review

Studied existing projects, systems similar to our project

Results

Exploratory data analysis, error percentages of models used

Data Preparation & Model Building

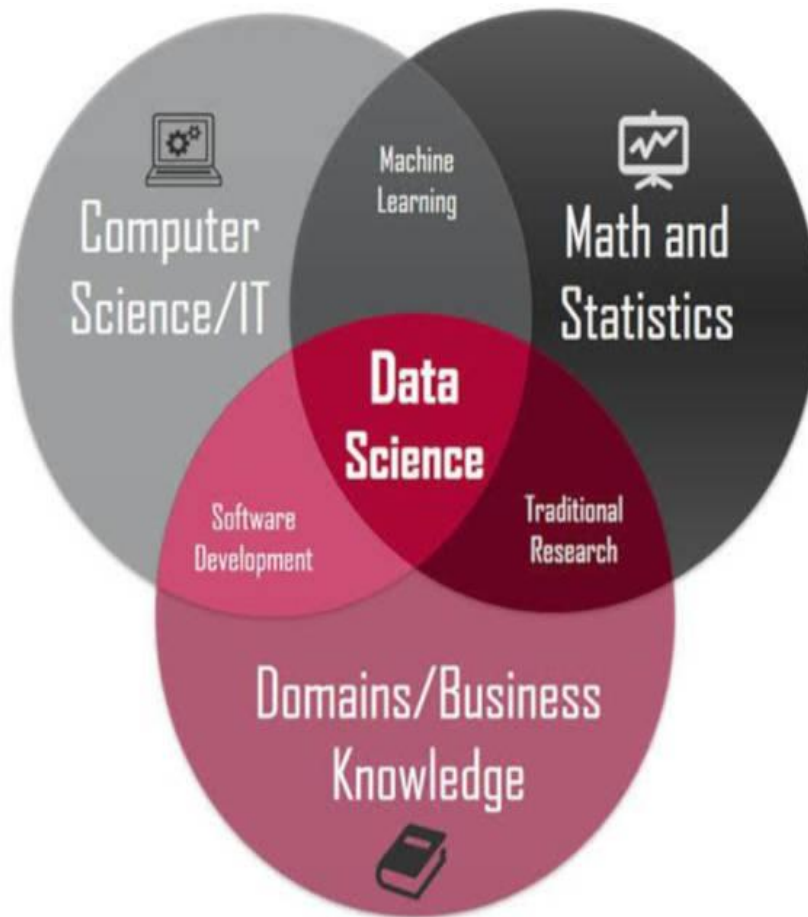
Data collection, Data preparation, Data exploration, Model building, model validation, etc

Further Implementation

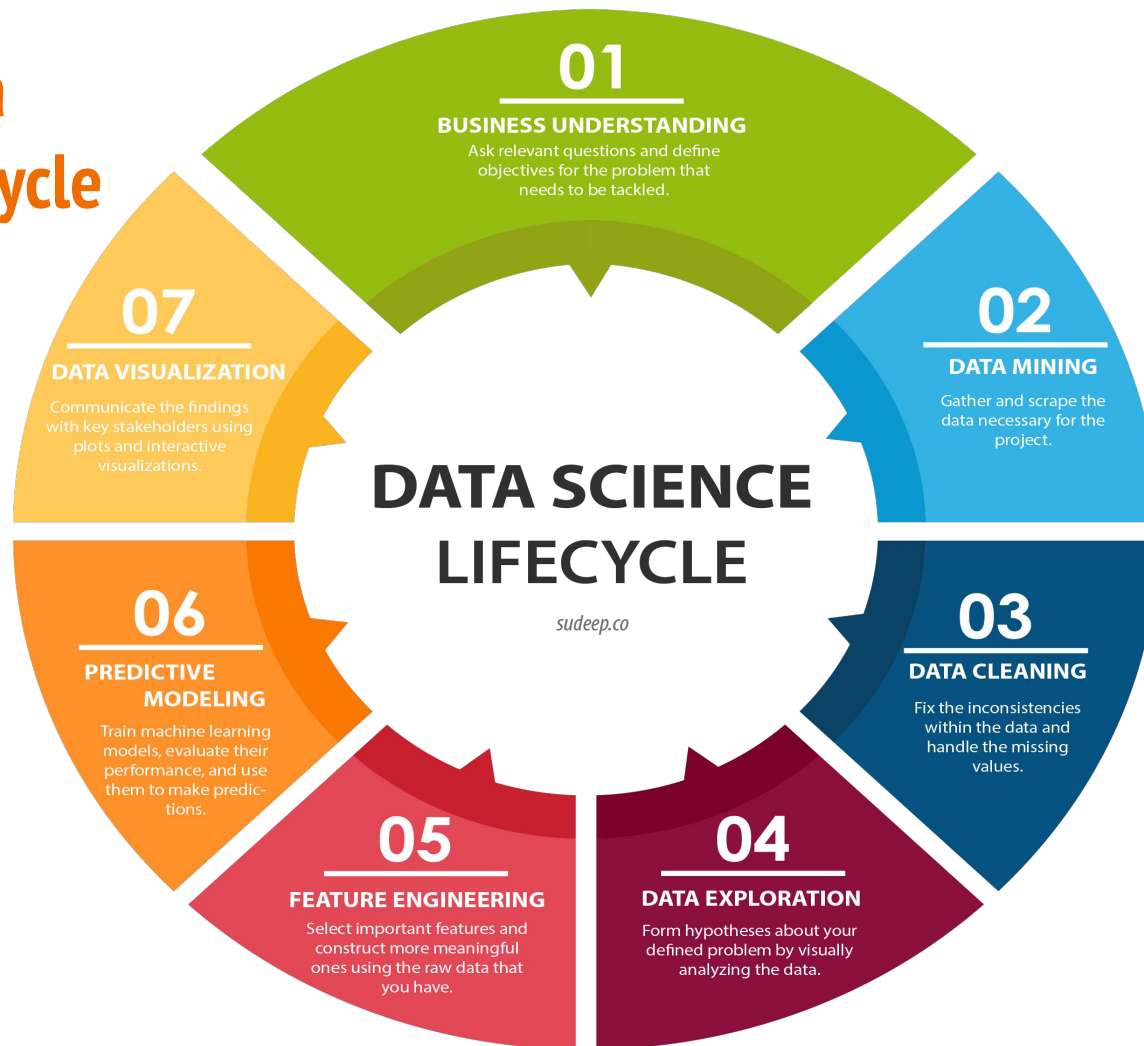
Exploring the impact of other algorithms on the predictions

Intro to Data Science

- What is Data Science?
 - Definition
 - Why the hype?
- Applications
 - Sports
 - Finance
- Python
 - Tools
 - Libraries
- Steps and Case study
 - Brief description of steps involved in data science
 - Ideas for case study



Steps in Data Science Lifecycle



General Steps in Data Science Project

1. **Data Gathering**

Need for data, Sources to collect data from

2. **Data Preparation**

Data pre-processing, profiling, cleansing, transformation

3. **Exploration**

Initial step in data analysis, visualisation

4. **Model Building**

Various important models, Idea behind choice

5. **Model Validation**

Testing a given dataset

6. **Model Deployment**

Application of model

Popular Models

Understanding important models along with the idea behind and examples

- Random Forest
 - SVM
 - Bayesian Predictors
 - K-Nearest Neighbours
 - Neural Networks
-

Project Ideas

- **Worldwide Salary Predictor for Software Developers**

Use the data that software developers around the world answered on a Stack Overflow survey, helpful in predicting salaries around the world, and learning about diversity in tech

- **Understanding most popular topics in Indian society**

This news dataset is a persistent historical archive of notable events in the Indian subcontinent from start-2001 to mid-2020, recorded in real time by the journalists of India.

- **Estimating drug economies in certain regions of the world**

Potentially find correlations between different drugs and from where/to they ship in the world to show correlations between types of drugs and where drug dealers that supply them are located.

- **Recommendation system based on user's interests**

We will use the data that explores the preferences, interests, habits, opinions, and fears of young people.

Salary Predictor for Software Developers

Problem Statement and Plan

- **Data Description**
Understanding of the given dataset
 - **Project**
Create a Machine Learning model that predicts the salary of developer based on other given data
 - **Steps**
Planning the steps involved
-

Literature Review

We present some existing projects or systems similar to our current project. We go through a complete thought process of the project subject and its viability.

1. **Salary as a predictor of salary: A 20-year study** Citation: Brenner, M. H., & Lockwood, H. C. .
Salary as a predictor of salary: A 20-year study. *Journal of Applied Psychology*, 49(4), 295–298
2. **Gender and Salary Inequity: Statistical Interaction Effects** Citation: *Social Work Research and Abstracts*, Volume 18, Issue 4,, Pages 24–27
3. **The Multifactorial Achievement Scale as a Predictor of Salary Growth** Citation: [Social Behavior and Personality: an international journal](#), Volume 23, Number 2, 1995, pp. 159-162(4)
4. **Faculty salary as a Predictor of Student Outgoing Salaries from MBA Programs** Citation: Karla H. Hamlen, William A. Hamlen
5. **Predicting compensation for job seekers** Citation: Fazio, Kilpatrick and Baker, Stanford University
6. **Salary Prediction in the IT job market - a Spanish case study** Citation: Martin, Mariello and Battiti
7. **Salary Prediction using Regression Techniques** Citation: Rupashri Barik, JIS College of Engineering

Literature Review Summary

S.No	Paper title	Technologies	Conclusion	Issues	Author	Year	Publisher
1.	Salary after a few years in an organization as a predictor for salary at a later date. (A 20 Year Study)	The salaries were combined to yield yearly distributions, 1 for beginning salary and 1 for each yr. of experience. The resulting distributions were intercorrelated.	Correlations between equidistant yr. became larger as tenure increased, and the variance of salaries increased with increasing tenure.	Limited sample size that overoptimizes on an unrelated field	Brenner, M. H., & Lockwood, H. C.	1965	<i>Journal of Applied Psychology</i> , 49(4), 295-298
2.	Gender-related inequity in the salaries of social workers	Simple statistical model that compares the salaries between the 2 genders	There was a huge gap between the salaries for men and women.	Outdated data that focuses on a single city	Hide Yamatani	1982	<i>Social Work Research and Abstracts</i> , Volume 18, Issue 4, Winter 1982, Pages 24-27

3.	The Multifactorial Achievement Scale as a Predictor of Salary Growth.	Multivariate linear regression.	Significant correlations were obtained between most of the subscales, an index of salary growth and work motivation.	A single factor that accounted for 61.4 percent of the variance, rather than all factors.	Orphen, Christopher	1995	Social Behavior and Personality: an international journal, Volume 23, Number 2, 1995, pp. 159-162(4)
4.	Faculty salary as a Predictor of Student Outgoing Salaries from MBA Programs	Hierarchical Linear Regression Analysis. (Program characteristics as control variable; Faculty salary as the predictor variable; Avg outgoing salary as the dependent variable)	Higher faculty salaries were associated with higher starting salaries for MBA students upon graduation.	Correlation might not signify causation	Karla H. Hamlen, William A. Hamlen	2015	Journal of Education for Business, Volume 91, 2015 -Issue 1
5.	Predicting compensation for job seekers.	Simple regression model , Tree based regression model	Predicting the log of salary values is substantially more accurate than predicting the original value.	Regularization with small values of λ delivers small gains in accuracy, no major boosts.	Megan Fazio , Jen Kilpatrick, Darren Baker	2016	CS 229 Project Final Report Autumn 2016

6.	Salary Prediction in the IT job market - a Spanish case study	Classification Models (SVM, Multi Layer Perceptron, Random Forests, AdaBoost)	Ensembles based on the above models behave better and lead to a best case accuracy of ~84%.	Multi Layer Perceptron performs poorly due to limited data in the sample.	Nacho Martin, Andrea Mariello, Robert Battiti	2018	International Journal of Computational Intelligence Systems 11(1):1192
7.	Salary of a person after a certain year.	Linear Regression	Used only 2 dimensions to predict salary - which is clearly insufficient for such a complex prediction task	Need a larger dataset and more data points.	Rupashri Barik	2020	SSRN Electronic Journal

Data Science Steps for Salary Predictor

1. Data Collection: StackOverflow 2019 Developer Survey



- **What is StackOverflow?**

StackOverflow features questions and answers on a wide range of topics in computer programming and is used by millions of developers worldwide.

- **What's in this Dataset?**

- Responses of 100,000 developers who took the 30-minute survey released by StackOverflow
- Contains information about their professional coding habits, tools used and salary.
- 129 attributes

2. Data Preparation

We performed data cleaning to remove inconsistencies within the data and handle the missing values.

- Remove empty values
- Remove outliers
- Data transformation

3. Data Exploration

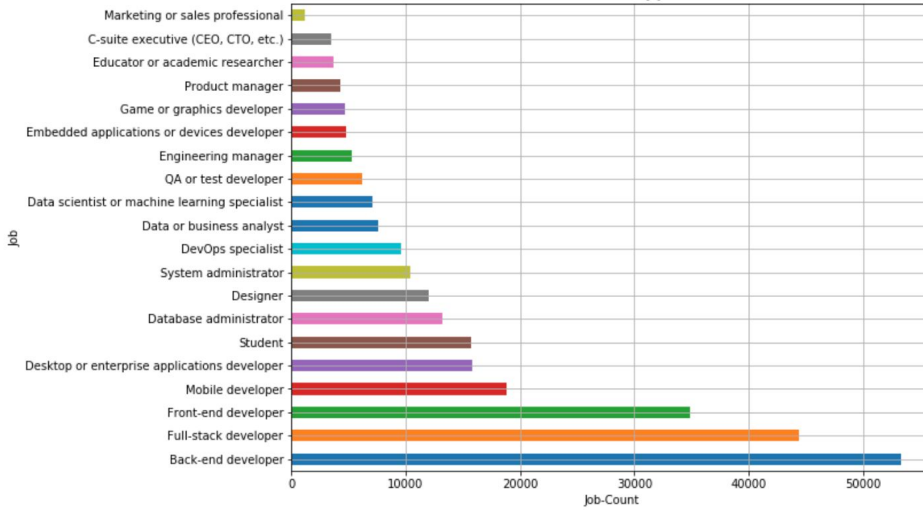
We explored:

- Job count
- Salaries
- Gender ratio
- Countries
- Experience

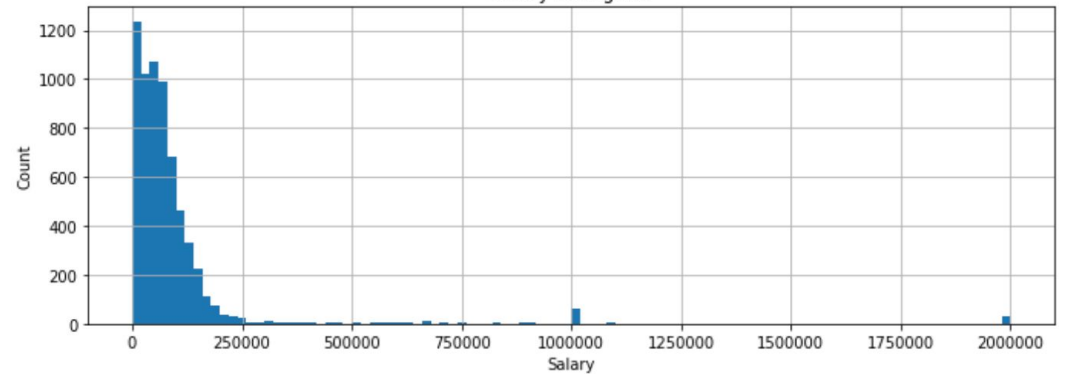
We performed data exploration to form hypotheses about our defined problem. We also visually analysed the data using Matplotlib, Plotly and Seaborn libraries.

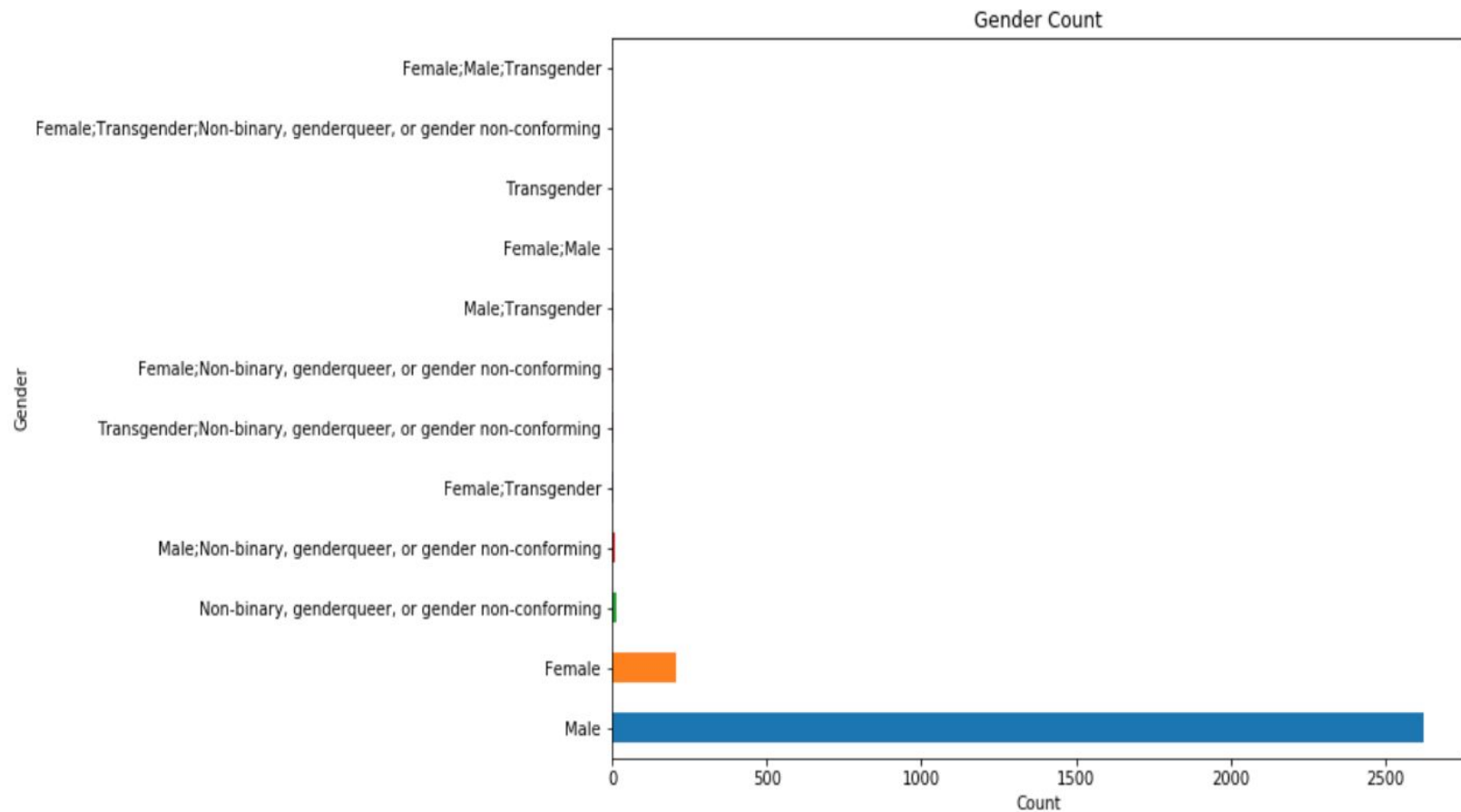
- **Matplotlib** -The matplotlib Python library is used to create high-quality graphs, charts, and figures.
- **Plotly** - The Plotly Python library is an interactive, open-source plotting library that supports over 40 unique chart types covering a wide range of statistical, financial, scientific, and 3-D use-cases.
- **Seaborn** - Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Stack Overflow Survey Job-Count

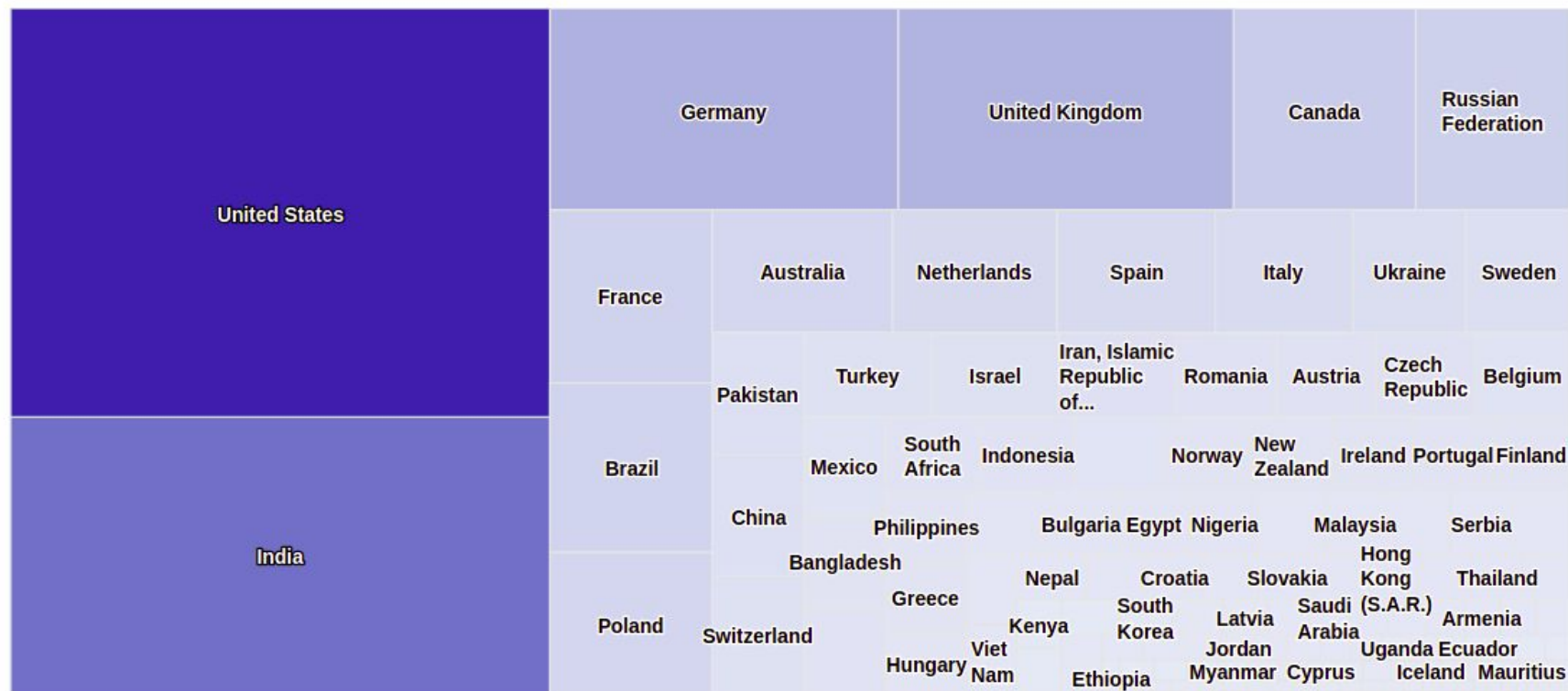


Salary Histogram



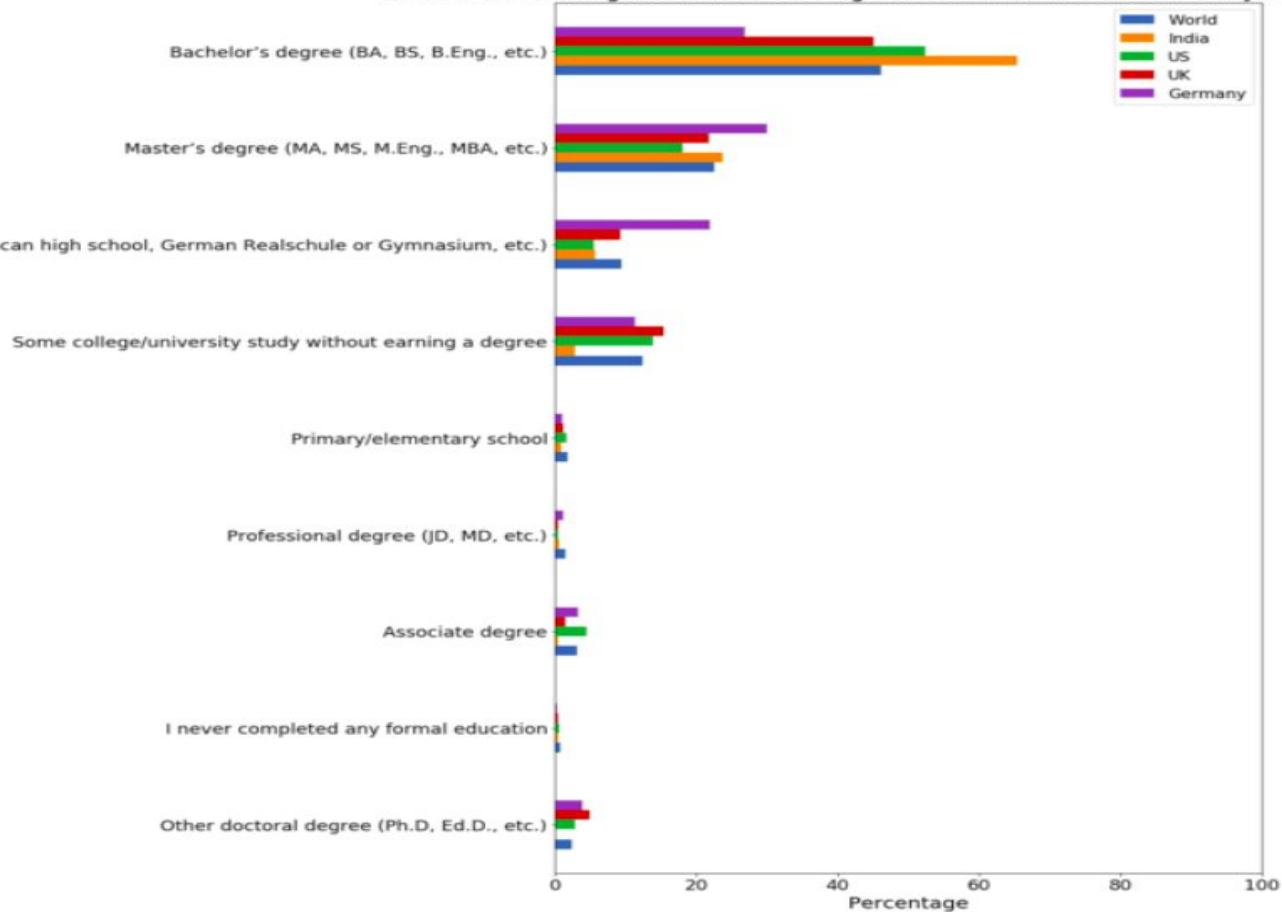


Countries from where overall respondents come from



Which of the following best describes the highest level of formal education that you've completed?

Highest level of formal education



Highest level of education of the responders

Model Building

We divided the dataset into 2 parts:

- Train set: used for training the models
- Test set: used for testing the accuracy of our trained models

Using this dataset, we trained 2 regression models:

1. Linear regression
2. Gradient Boosting regression

We trained them on the features to predict salaries.

Model 1: Linear Regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data.

One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.


Model 2: Gradient Boosting

Gradient boosting involves three elements:

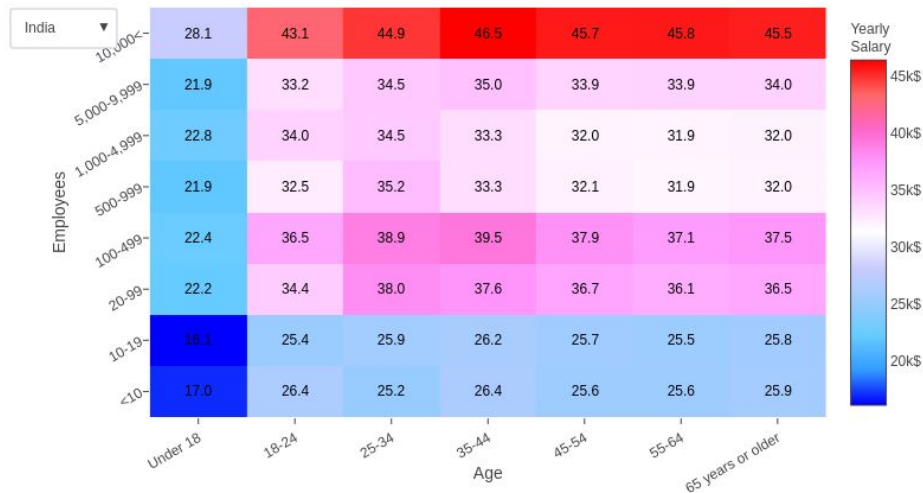
- A loss function to be optimized.
- A weak learner to make predictions.
- An additive model to add weak learners to minimize the loss function.

Results

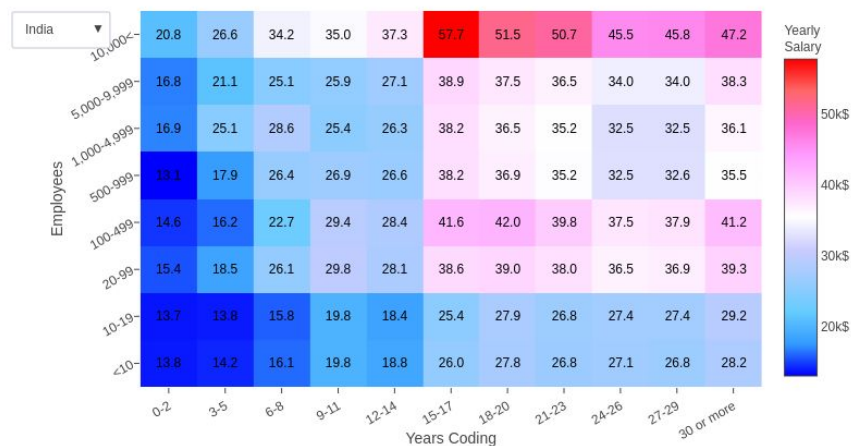
Here are the results that they achieved:

Model Used	Error Percentage (on test set) (%)
Linear regression	63.9%
Gradient Boosting regression	41.8% 

Median Salary Of A Data Scientist In India



Median Salary Of A Data Scientist In India



CONCLUSION

In terms of data,

- We can see that the most popular jobs for programmers is in web development as - front-end developer, backend developer and full-stack developer.
- We can also observe that a large majority of software developers around the world earn between 10,000 USD to 25,000 USD.
- The data also shows that the top 5 countries that produce software developers are the United States, United Kingdom, Germany, India and Canada.
- The data also points towards the gender disparity in the software industry - where there are at least 5 times more male software developers than female developers.

In terms of the model,

- We created a salary prediction model using the data mentioned before, that tries to predict the salary of an individual based on trends seen in the industry.
- We used 2 algorithms - linear regression and gradient boosting.
- The first model that used linear regression performed significantly worse than the second model that used gradient boosting.
- The second model gave 22.1% more accurate results as compared to the first.
- This was expected because gradient boosting is a more sophisticated algorithm than linear regression.

What's Next!

- We can try to further explore the impact of other types of algorithms on the predictions.
- We can use Random Forest algorithm, Support Vector Machines and neural network based models.

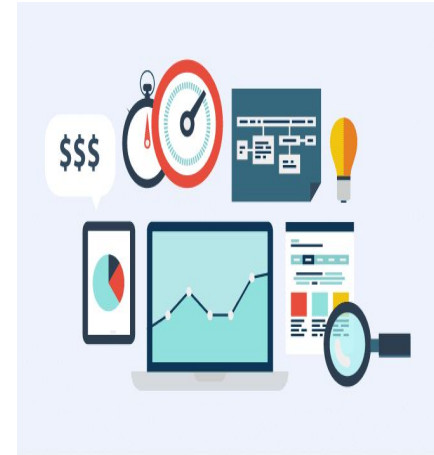


Salary Predictor System

Age <input type="text" value="30"/>	Gender <input checked="" type="radio"/> male <input type="radio"/> female
Career category <input type="text" value="ICT"/>	Availability to start work <input type="text" value="Single"/>
Educational degree <input type="text" value="Master"/>	Living place <input type="text" value="Bangkok"/>
GPA <input type="text" value="3.15"/>	Disability <input checked="" type="radio"/> No <input type="radio"/> Yes
Work experience (years) <input type="text" value="5"/>	<input type="button" value="PREDICT"/>

Your predictive salary is

33018 baht



Thank You!

Through this project, we aimed to understand the salary distribution, community inclusiveness, companionship and gender distribution of software developers around the world.

We tried to create a machine learning model that would help software developers understand their true worth.

