Capstone Project Proposal

Stamatis Stamatiadis

September 29, 2020

Software Engineer Salary

Domain Background & Problem Statement

One of the most popular questions among graduate software engineers is the level of salary they should expect at their first job. The ecosystem of the software industry can be considered dynamic and very competitive. Therefore, the salary question persists and concerns even experienced professionals in the field. Due to the high demand for software solutions, new opportunities arise constantly and salary levels vary vastly [1].

On the other hand, companies try to compete for experience personnel while also trying to stay completive from an employer standpoint. Having an overview of the software market salary levels is considered necessary for every human-resources department.

This project is an attempt to answer the following question:

> What should be the expected annual salary for a software technologist given specific criteria that describe both the employer and the employee?

Datasets and Inputs

During September 2020, a YouTube channel named SocialNerds [2] released an anonymized dataset [3] of nearly 600 entries that describes salary levels of software engineers. The data was collected online during the summer of 2020 through a Google Forms questionnaire & commented upon on a video [4]. The participants are Greek software engineers working mostly for companies located in Greece or abroad.

The dataset includes the following information:

- Data & time when the form was submitted
- Years of past experience
- Type of software position/development

- Programming languages used in day to day development
- Company size
- Working remotelly or on premises
- Having a managerial role
- Devoting time to personal projects
- Home location
- Work location
- Sex
- Annual net salary

The following disadvantages were observed regarding the quality of the data. The dataset is in Greek (it will be translated for the needs of the capstone project) while the number of observations might be considered low (586 entries). However, even under those circumstances, the dataset is evaluated as worthwhile for further analysis and even model creation since it contains real, recent data.

Solution Statement

The suggested solution includes the creation of a machine learning model based on current market data. The model should accept various information about a job candidate (software technologist) and a job description (company) while the output of the model should be an estimation about the expected salary level.

The available dataset can serve as a representation of current market trends regarding salary levels in the software industry.

Benchmark Model

Up until now no benchmark model has been found. Furthermore, no past data similar to the dataset that is about to be used were discovered. As an alternative to a benchmark model, the exploratory analysis of the dataset might provide results that could be compared to previous research or articles that explore salary levels in the software industry.

Evaluation Metrics

The evaluation of the generated model will be conducted by withholding a part of the original dataset and use it for testing purposes. Moreover, since the problem can be characterized as a quantitative analysis problem, the evaluation of the model will be done with corresponding metrics (for example: mean square error, cross entropy loss, absolute error). The choice of metrics might change or be adjusted according to the algorithms or methods that will be used for the creation of the model (for example: linear regression).

Project Design

The proposed workflow for the solution of the defined problem can be outlined in the following steps (some steps might be removed or new will be added according to future needs):

1. Data preprocessing
    - Data translation from Greek to English
    - Data refinement
    - Feature selection
    - Dimensionality reduction
2. Exploratory data analysis
    - Visualizations
    - Data exploration
    - Basic statistical analysis
3. Model creation
    - Implement model with appropriate algorithms
        - Current candidates: AWS Linear Learner, XGBoost, regression estimators in scikit-learn [5], [6]
4. Model tuning
    - Hyperparameter evaluation
5. Model evaluation
    - Define appropriate metrics
    - Visualize results
    - Conclusions

Citations

[1] A. Loten, “Software Engineers’ Pay Is Rising Faster Outside Silicon Valley,” Wall Street Journal, 12-Feb-2020 [Online]. Available: https://www.wsj.com/articles/software-engineers-pay-is-rising-faster-outside-silicon-valley-11581550773. [Accessed: 29-Sep-2020]

[2] “SocialNerds - YouTube.” [Online]. Available: https://www.youtube.com/channel/UCd5jW000te6bExqYth
[Accessed: 29-Sep-2020]

[3] “[Original] Research: Software Engineer Salary (Responses),” Google Docs. [Online]. Available: https://docs.google.com/spreadsheets/d/1TVL6IfF9yaEKa3S6ma69pn-6o2YFxzUgEMTdiec8BpU/edit?usp=embed_facebook. [Accessed: 29-Sep-2020]

[4] “Software Engineer Salary, Research 2020, NerdCast - YouTube.” [Online]. Available: https://www.youtube.com/watch?v=e-83bz4RhQ4. [Accessed: 29-Sep-2020]

[5] “Choosing the right estimator — scikit-learn 0.23.2 documentation.” [Online]. Available: https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html. [Accessed: 29-Sep-2020]

[6] “Use Amazon SageMaker built-in algorithms - Amazon SageMaker.” [Online]. Available: https://docs.aws.amazon.com/sagemaker/latest/dg/algos.html. [Accessed: 29-Sep-2020]