SALARY SIGHT: Tech Salary Prediction GithubLink

Namratha S Kumar, Nandini S Nair, Pooja Manjunatha, Priya Varahan

MOTIVATION

Embarking on this project is crucial for those venturing into the tech sector. Understanding the murky waters of work-life balance and expected compensation is crucial groundwork for anyone entering this fast-paced environment. Essential insights were acquired into the industry's pulse by analyzing data from Stack Overflow, a platform respected by engineers throughout the world. These insights aren't just about projecting income trends; they're about comprehending the terrain one is about to enter. With this knowledge, individuals can make smarter choices and ensure that the career paths align with one's ambitions and realities. This initiative is about more than just data exploration; it's also about empowering individuals to embark on their technological journeys with clarity and purpose.

BACKGROUND

Lothe et al. underlined the fact that salary is often the most important factor determining employee retention and turnover. Higher salaries tend to encourage employees to stay with a company, but lower or stagnant pay may push them to look elsewhere. This emphasizes how important it is to precisely assess and estimate compensation levels. Also, individual characteristics such as, educational background, and work experience have a substantial impact on compensation variances. Salary prediction not only helps individuals determine their salary expectations, but it also allows businesses to match with their employees' compensation preferences. Machine Learning has made it easier to acquire large salary datasets, allowing prediction approaches to be applied to address these challenges.

The project's background is rooted in the pressing need to comprehend and forecast salary trends, particularly within the dynamic landscape of the tech industry. Numerous studies, including those conducted by Guanqi Wang, Satpute et al., Mishra et al., and Khongchai et al., have established methods for predicting employee salaries using advanced data mining and machine learning techniques, providing valuable insights into the factors influencing salary projections, such as college degrees, years of experience, demographic characteristics, and technological competencies.

LITERATURE REVIEW

A number of studies have extensively explored the use of machine learning techniques to forecast compensation for employees. Wang's research meticulously investigates various regression models, such as Multiple Linear, Ridge, Elastic-Net, Lasso, and Polynomial Regression, ultimately affirming Polynomial Regression as the most effective method ($R2 = 0.95$, $RMSE = 4.7$), while emphasizing the significance of advanced education. Satpute et al.'s study leverages socio-demographic data and years of experience to advocate for Random Forest Regression over Multiple Linear Regression ($R2 = 0.8990$ vs. 0.7201), shedding light on the intricate factors influencing wage projections. Similarly, Mishra et al. focus on facilitating students' ability to predict future salaries, with Decision Tree (J48) exhibiting superior performance in cross-validation. Additionally, Khongchai et al. tailor a specialized salary prediction system for graduate students, favoring Random Forest over Decision Trees (ID3, C4.5), achieving an impressive accuracy of 90.50%.

These studies collectively delve into machine learning's application in forecasting employee compensation, offering nuanced insights into the complexities of salary prediction. Guanqi Wang's work highlights Polynomial Regression's efficacy and emphasizes the impact of advanced education on salary outcomes, while acknowledging challenges like multicollinearity. Satpute et al.'s research underscores Random Forest Regression's superiority, underlining the importance of socio-demographic factors and addressing biases for fair salary predictions. Mishra et al.'s focus on student salary estimation bridges academic and professional realms, offering insights into career planning and educational policy implications. Khongchai et al.'s tailored prediction system for graduates showcases machine learning's potential in inspiring academic excellence and discusses its scalability beyond specific demographics. Together, these studies advocate for personalized salary predictions, address challenges, and contribute to informed decision-making in workforce planning and educational support.

METHODOLOGY

Data Collection

The data source used for this project is the Stack Overflow Developer survey (SODS) dataset, which is sourced from Stack Overflow's annual online survey of software developers all over the world. This dataset consists of 84 variables spread across seven sections that represent the developer community's perceptions, employment details, technological preferences, and demographics. The survey was carried out in 2023 and almost 91,000 software engineers in 185 countries took part in this.

The raw data weighs approximately 158.6 MB and is readily available on the Stack Overflow website in CSV format, ensuring transparency and reproducibility due to its public availability. The dataset is predominantly made up of categorical data, with only a small number of numerical variables, alongside NaN values necessitating data cleaning and transformations. To guarantee compliance with privacy laws and to preserve data integrity, thorough privacy and data consent checks are carried out before analysis. It serves as a valuable resource for comprehending the developer community and has the potential to greatly influence salary prediction in the technology industry. Figure 1 shows the top 3 rows of the dataset

	ResponseId	Q120	MainBranch	Age	Employment	RemoteWork	CodingActivities	EdLevel	LearnCode	LearnCodeOnline	...	Frequency_1	Frequency_2	Frequency_3	TimeSearching	TimeAnswering	ProfessionalTech	Industry
0	1	I agree	None of these	18-24 years old	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	2	I agree	I am a developer by profession	25-34 years old	Employed, full-time	Remote	Hobby;Contribute to open-source projects;Boots...	Bachelor's degree (B.A., B.S., B.Eng., etc.)	Books / Physical media;Colleague;Friend or fam...	Formal documentation provided by the owner of ...	...	1-2 times a week	10+ times a week	Never	15-30 minutes a day	15-30 minutes a day	DevOps function;Microservices;Automated testin...	Information Services, IT, Software Development...
2	3	I agree	I am a developer by profession	45-54 years old	Employed, full-time	Hybrid (some remote, some in-person)	Hobby;Professional development or self-paced l...	Bachelor's degree (B.A., B.S., B.Eng., etc.)	Books / Physical media;Colleague;On the job tr...	Formal documentation provided by the owner of ...	...	6-10 times a week	6-10 times a week	3-5 times a week	30-60 minutes a day	30-60 minutes a day	DevOps function;Microservices;Automated testin...	Information Services, IT, Software Development...

3 rows × 84 columns

```
Total count of rows: 89184
Total count of columns: 84
```

Figure 1

Exploratory Data Analysis

During dataset exploration, an attempt to understand variable characteristics and distributions was made. A significant focus was on developing the job satisfaction meter, which is required for assessing professionals' contentment in their roles. This entailed collecting responses to job satisfaction-related survey questions, transforming them quantitatively, and calculating an overall job satisfaction score for each responder. Based on quantiles, three categories were identified from Figure 3: "Not Satisfied" (1.0 - 2.875), "Neutral" (2.875 - 3.625), and "Satisfied" (3.625 - 5.0), which provided useful insights on professional satisfaction levels.

The EDA revealed non-normal distributions as shown in Figure 2 with varied skewness between features. 'ConvertedCompYearly' (annual compensation) which is our target variable deviated significantly from normalcy, with a high right skewness (94.74) and a Kolmogorov-Smirnov test returning a p-value of 0. The interquartile range (IQR) technique revealed 4.59% outliers. Similarly, 'YearsCodePro' and 'JobSatisfaction' exhibited significant skewness, indicating the presence of outliers.

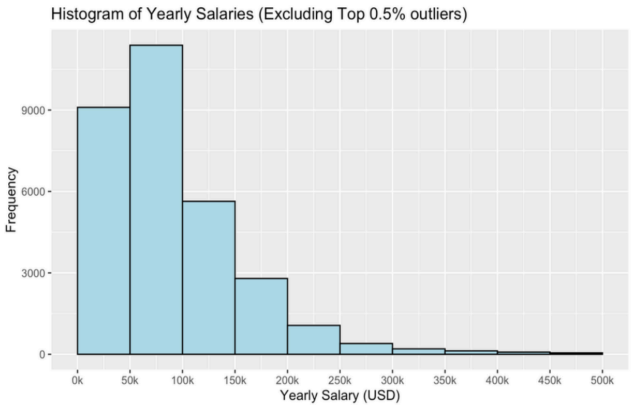

Figure 2

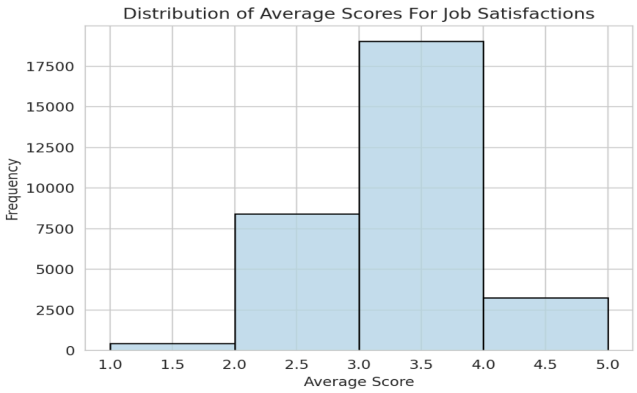

Figure 3

Data Cleaning

During the initial data cleaning process, columns having a high number of missing values (60%-70%)were detected and eliminated to reduce potential biases in the study. The target feature had missing values that were eliminated which accounted for about 50% of the dataset, to ensure data integrity. Preprocessing steps were applied to all the 'HaveWorkedWith' columns, which calculated the number of times each responder had worked with specific technologies. This required parsing and counting techniques to extract useful information, hence increasing the dataset's granularity. Columns with large quantities of missing data were removed to ensure a clean dataset for analysis. To enhance the reliability of the following studies, outliers in the salary data were removed using the interquartile range (IQR) approach. After these steps, the dataset was cut down to 48,019 rows and 35 columns

Data Transformation on Target Feature

Following the data cleaning process, many changes were performed on the dataset to prepare it for analysis. Geographical data from 183 countries was mapped to six continents. This classification aids subsequent analysis incorporating regional differences. Next, categorical features were categorized as nominal, ordinal, or interval depending on their nature, and missing values were handled using appropriate imputation techniques such as mode, median, and mean. Logarithmic transformations were applied to the target feature as well as other skew-related features.

Feature Selection

Feature selection was carried out using ANOVA (Analysis of Variance) with a significance level of 0.05 in order to identify important categorical features. By examining how compensation varies across different groups defined by individual categorical features, features were identified where these differences were statistically significant. This method ensures that only the most influential features are maintained for further analysis, hence increasing the model's predictive power and interoperability. Out of the 15 features initially chosen by ANOVA, seven were chosen for their major influence, while others, such as 'Ease of Survey', were deemed unimportant. Pearson's correlation analysis revealed eight features with considerable correlations, which expanded the feature selection. This careful selection procedure resulted in a final dataset of 15 features, ensuring a balanced yet targeted approach to modeling the data while precisely capturing its underlying patterns.

Data Transformation on Selected Features

As part of data transformation, categorical variables were transformed into numerical representations to make them more compatible with machine learning algorithms. One-hot encoding split the 'Employment' column into binary columns, which effectively captured its categories. Label encoding assigned distinct numerical labels to features like 'Age', 'EdLevel' etc., Custom mappings encoded 'JobSatisfaction' into numerical values. Histogram plots and skewness measurements were used to examine the distributions of numerical data, while outlier detection techniques revealed extreme values using the interquartile range (IQR). The consistent feature scaling is ensured by using min max scaling on features. The final dataset consisted of 30,208 rows and 15 variables after transforming the selected features. Figure 4 shows the encoded values.

	YearsCodePro	JobSatisfaction	CodingLanguageNum	OfficeStackAsyncNum	OpSysProfessionalNum	NEWCollabToolsNum	WorkExp	DatabaseNum	PlatformNum	OfficeStackSyncNum	avg_score	ConvertedCompYearly	LogComp	Full-time Employment	Age_encoded
0	9.0	0	3.0	7.0	3.0	2.0	10.0	1.0	3.0	10.0	3.625	285000.0	12.560244	1	1
1	23.0	0	2.0	2.0	2.0	2.0	23.0	0.0	5.0	3.0	3.500	250000.0	12.429216	1	3
2	7.0	1	7.0	1.0	3.0	4.0	7.0	2.0	2.0	5.0	4.125	156000.0	11.957611	1	1
3	4.0	2	3.0	3.0	1.0	2.0	6.0	4.0	5.0	4.0	2.875	23456.0	10.062882	1	1
4	21.0	2	6.0	4.0	1.0	2.0	22.0	4.0	3.0	4.0	2.500	96828.0	11.480691	1	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
30531	24.0	0	2.0	3.0	1.0	4.0	25.0	2.0	0.0	4.0	3.000	50719.0	10.834056	1	2
30532	2.0	0	11.0	4.0	2.0	14.0	3.0	6.0	2.0	4.0	3.250	16917.0	9.736074	1	0
30533	2.0	0	4.0	0.0	2.0	4.0	2.0	4.0	0.0	2.0	3.125	15752.0	9.664723	1	0
30534	9.0	2	6.0	2.0	1.0	3.0	9.0	1.0	1.0	1.0	2.375	64254.0	11.070599	1	1
30535	9.0	0	3.0	3.0	1.0	2.0	9.0	1.0	0.0	3.0	3.625	61041.0	11.019301	1	1

Figure 4

MODELING AND MODEL DETAILS

For the regression modeling focused on salary prediction ,the evaluation is performed on four models: Linear Regression,Support Vector Regression,Random Forest Regression and XG Boost. Linear regression model is used as the base model .The transformed dataset consisting of 14 features is used for modeling and is divided into train ,test and validation sets to evaluate model performance. The detailed feature extraction process is provided to the models with correct information to identify the pattern and make accurate salary predictions.Hyper parameter for each model is optimized to achieve best performance .The models are evaluated using Root Mean Squared Error(RMSE),Mean Absolute Error(MAE) and R-squared(R^2) metrics.

The dataset is split into train, test and validation set in the ratio 70:15:15 for unbiased performance evaluation. Figure 5 shows the dataset split .

```
Training set size: (21144, 14) (21144,)
Validation set size: (4532, 14) (4532,)
Test set size: (4532, 14) (4532,)
```

Figure 5

All the models are trained on the training set and are being evaluated on a validation set to check if the model is underfit or overfit.Training and the validation loss of all the four models were compared. Figure 6 represents the loss graph for the XG Boost model. Additionally, the train, test and validation loss were calculated to ensure the evaluation of model performance. Figure 7 shows the average MSE loss of train, validation and test set.

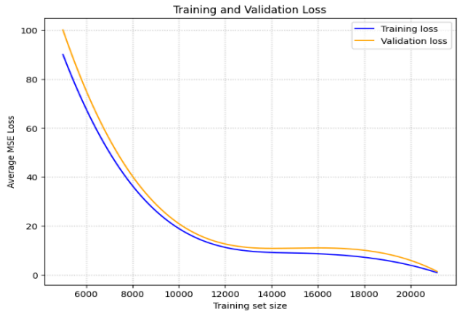

Figure 6

Training Set Size	Average MSE Loss (Training)	Average MSE Loss (Validation)	Test MSE Loss
6000.0	95.0	98.0	96.5
8000.0	70.0	75.0	72.5
10000.0	50.0	55.0	52.5
12000.0	35.0	38.0	36.5
14000.0	20.0	22.0	21.0
16000.0	10.0	12.0	11.0
18000.0	5.0	6.0	5.5
20000.0	2.0	3.0	2.5
22000.0	1.0	2.0	1.5

Figure 7

EXPERIMENTS AND RESULTS

Hyper parameters

L1 (Lasso) regularization is used for the linear regression model to prevent overfitting and 5-fold cross-validation is performed to find the best alpha(regularization strength) value.
The hyper parameters used for SVM models are C(regularization parameter),gamma, kernel and the best parameters were found using GridSearchCV.The hyperparameters used for the Random Forest model were n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features and the best parameters were found using K-fold cross- validation. The hyperparameters used for the XGBoost model are n_estimators, learning_rate, and max_depth. The metrics were evaluated before and after hyper parameter tuning . Below table compares each model

Models	Hyper Parameters Used	Before Hyper parameter tuning	After hyper parameter tuning
Linear Regression	L1 regularization	RMSE: 0.085	RMSE:0.085
		MAE: 0.073	MAE:0.073
		R^2: 0.52	R^2: 0.52
Support Vector Regression	C(regularization parameter), gamma, kernel	RMSE:0.075	RMSE:0.075

		MAE:0.668	MAE:0.065
		R^2:0.52	R^2:0.57
Random Forest Regression	n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features	RMSE:0.045	RMSE:0.04
		MAE:0.037	MAE:0.035
		R^2:0.72	R^2:0.75
XG Boost	n_estimators, learning_rate, and max_depth	RMSE:0.025	RMSE:0.02
		MAE:0.016	MAE:0.015
		R^2:0.876	R^2:0.886

RESULTS

The models were compared using the evaluation metrics and XGBoost Model outperforms all the other models with a highest R2 score of 0.88,RMSE of 0.02 and MAE of 0.015. The best model is used for predicting the salary. Figure 9 shows the snippet of the predicted salary where the user inputs all the14 features and the corresponding salary is predicted corresponding salary is predicted and Figure 9 shows the metrics of all models

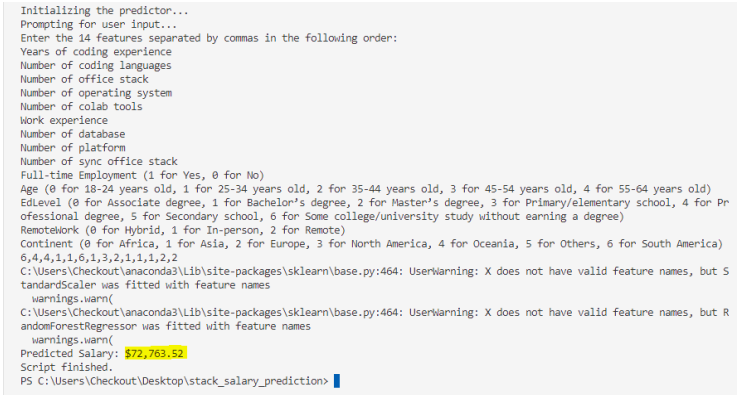

Figure 8

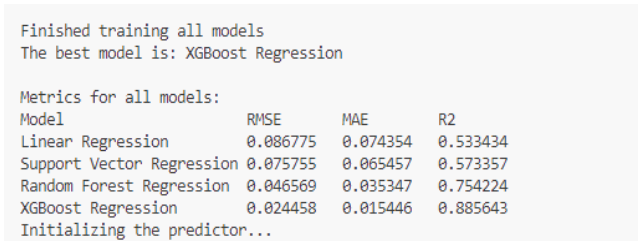

Figure 9

Acceptable Score

To assess if RMSE is acceptable , the target variable's scale is compared to RMSE for all the models. Fig shows a graph for the comparison of target variable scale and RMSE .The RMSE is significantly smaller than the range of the target variable, which suggests a good performance.

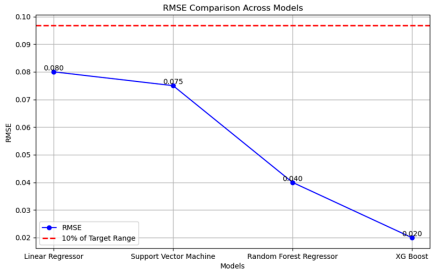

Figure 10

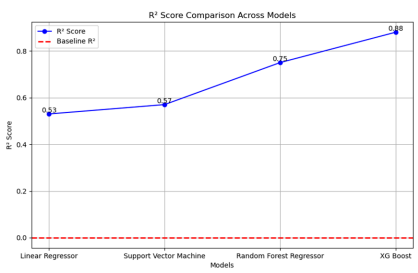

Figure 11

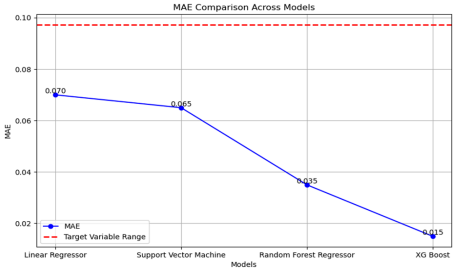

Figure 12

To assess if R2 score is acceptable the baseline model(Mean of target variable) is compared. Fig shows the graph for the comparison. The R2 score is significantly higher than the baseline .The MAE score is also compared to the target variable score and it is lower than the range of target variable.Figure 10 to 12 shows the graph for the acceptable RMSE, MAE and R2 score.

Coefficient of Variance(COV)

When evaluating the accuracy of a model, the Coefficient of Variation (COV) in salary prediction regression facilitates comparisons between various parameters or expected salaries by illustrating the relative variability of prediction errors. The COV for all the models is calculated as the ratio of the standard deviation to the mean of the predicted salary. Figure 14 and 15 shows the calculated COV for the XGBoost regression model and the graph shows comparison of COV of predicted salary with the model.A lower coefficient of variation (COV) implies that the model's predictions are more consistent.

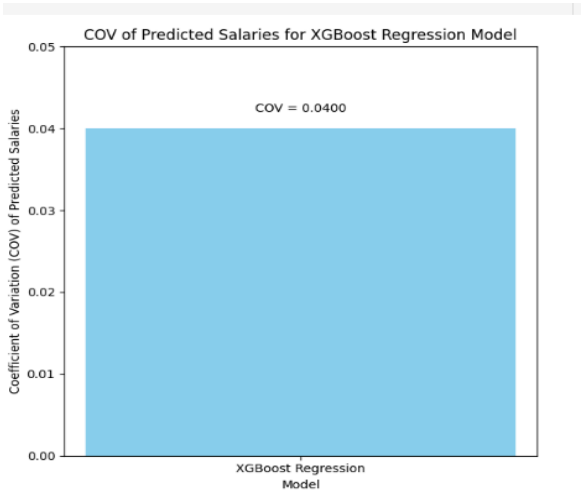

Figure 13

```
Mean of Predicted Salaries: 11.03499726572466
Standard Deviation of Predicted Salaries: 0.4411821279025582
Coefficient of Variation (COV) of Predicted Salaries: 0.03998026617305066
```

Figure 14

DISCUSSION & FUTURE IMPROVEMENT

As proven by the RMSE, MAE, and R-Squared scores, the predictive model developed aims to calculate a developer's yearly compensation with satisfactory accuracy. The model's primary strength lies in its inclusiveness, considering a wide range of variables during the feature selection process. This inclusiveness adds depth to the salary prediction and offers insights into how different features influence the salary. The evaluation metrics used confirm that the model effectively captures the relationship between the input features and the target variable.

The study of more complex models like polynomial regression, has the potential to further capture the non-linearity of the data, resulting in more accurate predictions. Improving the robustness of the model can be accomplished by addressing data imbalance issues by the application of strategies such as oversampling, undersampling, or the utilization of alternative performance metrics. Lastly, the incorporation of additional data sources and external datasets can provide a more comprehensive view of the variables that influence salary, thus enhancing the overall forecasting of the model.

REFERENCES

Babasaheb S. Satpute; Raghav Yadav; Pramod K. Yadav (2023, October 6). *Machine Learning Approach for Prediction of Employee Salary using Demographic Information with Experience*. IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/10353537

Guanqi Wang (2022, August 1).*Employee Salaries Analysis and Prediction with Machine Learning*. IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/9943146

M. D. Lothe, P. Tiwari, N. Patil, S. Patil and V Patil, "Salary Prediction using Machine Learning", *International Journal of Advance Scientific Research and engineering Trends*, vol. 6, no. 5, pp. 199-202, 2021.

Pornthep Khongchai; Pokpong Songmuang (2016). *Random Forest for salary prediction system to improve students' motivation*. IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/7907533

Praveen Mishra; Shivansh Srivastava; Priyanshi Gupta; Atul Anand; Subhash Chandra Gupta (2021, December 17). *A Comparative study of Machine learning Algorithms for salary estimation*. IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/9725775

Stack Overflow. (2023). Stack Overflow Developer Survey 2023. Retrieved from https://insights.stackoverflow.com/survey