Journal of
Data science and Modern Techniques

JScholar
Fostering Scholarly Communication

Research Article Open Access

Development of Salary Prediction Models for the Information Technology Industry

Berke Düzgün[1], Ceren Ulus[2] and Mehmet Fatih Akay[3*]

[1]Innovance, Department of Software Development, Istanbul, Turkey
[2]EFA Innovation Consultancy and Technology, Department of R&D, Adana, Turkey
[3]Cukurova University, Faculty of Engineering, Department of Computer Engineering, Adana, Turkey

[*]**Corresponding Author:** Mehmet Fatih Akay, Cukurova University, Faculty of Engineering, Department of Computer Engineering, Adana, Turkey, E-mail: mfakay@cu.edu.tr

Received Date: July 31, 2025 **Accepted Date:** August 11, 2025 **Published Date:** August 14, 2025

Citation: Mehmet Fatih Akay, Berke Düzgün, Ceren Ulus (2025) Development of Salary Predictions Models for the Information Technology Industry Salary Prediction Models. J Data Sci Mod Tech 2: 1-14

Abstract

In today's rapidly evolving technology and information age, every sector experiences constant workforce changes. In the Information Technologies (IT) sector, in particular, dynamic and rapidly changing working conditions lead to increased employee turnover. Personnel turnover impacts employee adaptation processes and, particularly when experienced personnel leave, leads to a significant loss of institutional knowledge in businesses. In this context, developing competitive salary systems and compensation policies aligned with employee expectations is a critical strategic imperative for businesses. Salary prediction-based on job positions and skill demographics stands out as an effective tool for determining wages that will enhance employee motivation. This study aims to develop salary prediction models using machine learning and ensemble learning methods. Models have been developed using machine learning-based Linear Regression (LR), Random Forest (RF), Ridge Regression (RR), Categorical Boosting (CatBoost), Support Vector Machines (SVM), Adaptive Boosting (AdaBoost) and Decision Tree (DT) as well as ensemble learning-based Voting and Stacking. After various preprocessing steps, such as encoding, a dataset consisting of approximately 10,000 rows has been used. Necessary measures have been taken during the training phase to prevent future information from leaking into the model, which could result in overestimated test performance and real-world failures. The performance of the developed models has been evaluated using Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Coefficient of Determination (R^2), training time (ms), and explainability criteria. The RF model achieved the highest success with a MAPE value of 2.60%. In contrast, the SVM and AdaBoost models exhibited lower predictive performance due to their longer training times and higher MAPE values.

Keywords: Machine Learning; Ensemble Learning; Salary Prediction; IT Sector; Support Vector Machines; CatBoost

Introduction

The IT sector is a broad sector that covers many areas such as computer hardware, software development, information processing systems and communication technologies. With the developing technology, the transaction volume and importance of this sector has increased rapidly. The IT sector has a dynamic structure that is constantly renewing and changing and operates in a rapid movement. Rapid developments in areas such as artificial intelligence, big data, cloud computing and the Internet of Things (IoT) in particular bring about continuous innovations in the sector. This dynamic structure necessitates the personnel working in the IT sector to constantly update themselves; it causes a high personnel circulation and constantly new career opportunities within the sector. IT personnel consist of individuals who constantly improve themselves in order to adapt to the rapid changes and innovations in technology and are open to learning new software, hardware and systems. These individuals have the ability to quickly adapt to current technological processes. However, this situation also brings about personnel turnover, which is a very important process for businesses. The process of an employee leaving the job and being replaced by another employee is called personnel turnover (personnel circulation). Personnel turnover plays an important role in critical factors such as cost optimization, sustainability and brand image for businesses. In the IT sector, the personnel turnover rate is generally high. High personnel turnover causes an increase in operating costs, decreases work efficiency and leads to loss of experience and knowledge within the organization. For this reason, businesses need to develop various strategies in order to reduce the personnel turnover rate. One of these strategies is to determine fair and market-appropriate salaries for employees during salary renewal periods [1].

Bonuses and salary increases provided to employees are important parameters in terms of increasing employee motivation and reducing the personnel turnover rate. It is of critical importance for employers to determine salary increases correctly, taking into account economic conditions and other relevant factors. In this process, both the goals of increasing employee loyalty and increasing their motivation should be met during the planning of salary increases. In addition, it should also be taken into account that the business keeps its costs under control and maintains its competitiveness. In this context, salary estimation stands out as a strategic action that helps businesses optimize their costs.

Accurate salary estimates increase employee satisfaction and loyalty by providing employees with clear, fair and competitive wages, while also helping to prevent high employee turnover. In addition, they contribute to the determination of appropriate salary ranges in recruitment processes, making it easier to bring in qualified employees. In this way, work efficiency is maximized and costs are kept at an optimum level.

This study aims to develop machine learning and ensemble learning-based salary estimate models. In this direction; salary estimate models have been developed using LR, RF, RR, CatBoost, SVM, AdaBoost, DT and Voting and Stacking methods from ensemble learning approaches.

This study is organized as follows: Section 2 includes relevant literature. Dataset generation is presented in Section 3. Methodology is presented in Section 4. Development of salary prediction models are presented in Section 5. Results and discussion are given in Section 6. Section 7 concludes the paper.

Literature Review

[2] aimed to examine how individual and demographic factors, such as professional experience, location, and education, influence salary trends; and to gain insights into the importance of these factors. For this purpose, three different Artificial Neural Network (ANN) models have been developed. A synthetic dataset of 1,000 samples with 6 features has been used to train the models. Model performance has been evaluated with test Loss values and the model containing a single hidden layer reached the lowest Loss value of 0.1415. The test Loss values of the model without hidden layer and the model with 2 hidden layers have been recorded as 0.1435 and 0.145, respectively. Characteristics such as high school education, doctorate, and years of experience show high contributions to salary estimates when measured by Permutation Importance. The results revealed how effective ANN's are in salary estimation, even with a synthetic dataset, it is possible to make general predictions

and outperform manual methods.

[3] aimed to apply ensemble learning methods through binary classification to salary prediction models which use DT, Logistic Regression (LR), Extreme Gradient Boosting, Light Gradient Boosting Machine, and RF methods. The models have been evaluated by Accuracy, Precision, Recall, and F1-Score metrics. By combining ensemble learning with the methods mentioned above, evaluation metric values have been enhanced. According to the results, it has been observed that ensemble learning significantly improves prediction performance, providing a more solid approach to salary forecasting. The study also emphasized the importance of model selection and evaluation metrics in performance optimization.

[4] presented a machine learning-based approach which uses professional and demographic features to predict the salary amounts of software developers. For this purpose, an interactive web application which includes a pre--trained regression model has been developed with Streamlit. The Stack Overflow Developer Survey 2020 anonymised responses have been used to train the model, which includes the most important numerical and categorical factors influencing pay outcomes. It has been observed that the program has a high degree of prediction capability, and it offers useful information to both individuals and businesses.

[5] aimed to analyze the of entry-level data science salaries in the United States using Multiple Linear Regression (MLR) on a dataset between 2020 and 2024. Key elements that affect income outcomes have been identified in the study, including job function, experience level, employment type, work arrangement, residency status, and firm size. Non-normality, heteroscedasticity, and multicollinearity by variable selection and data transformation problems have been handled, then forecast model has been finalized. The final model provided better interpretability and moderate predictive power. It has also presented valuable insights for employers, academics, and job seekers who want to comprehend and compare data science remuneration.

[6] developed a dynamic salary fluctuation trend prediction model that integrates numerous time series to increase the accuracy of forecasts and decrease errors. A multivariate time series data on wage variations has been preprocessed for outlier detection, removal, and filling. Based on this, the multivariate time series of salary has been trend extracted using the segmented aggregation approximation to produce the salary trend series. Lastly, a dynamic pay fluctuation trend prediction model has been built using Generative Adversarial Networks. According to the experimental findings, the suggested approach outperformed the comparison method and had a higher application value with a prediction error of less than 2.71%.

[7] aimed to develop a salary prediction system using machine learning methods which are LR, DT, Naive Bayes classifier, K-Nearest Neighbor and Support Vector Machine (SVM). The dataset has been collected from the 1994 census database which has 32,561 records of employee data. Prediction models have been implemented on both original train data and oversampled train data. According to the Accuracy value of the models, DT model performed better than the other models using the original train data.

[8] aimed to make job recommendations for science students according to their interests, talents and opportunities. For this purpose, an effective system has been introduced where Educational Data Mining is used to explore data from educational environments and to better understand students and how they learn. A dataset has been collected using employee and alumni data obtained from multiple sources.

[9] suggested an improved approach to salary prediction that uses a Principal Component Analysis system and a Deep Neural Network (DNN) model for the classification process to choose a subset of characteristics from all available data. After comparison with other traditional machine learning techniques like DT and RF. The proposed DNN model achieved a prediction error of only 5.1% for the DNN model, which is significantly lower than the 10.4% and 23.6% errors observed with DT and RF. This suggested that deep learning algorithms outperform traditional machine learning algorithms in salary classification and prediction tasks.

[10] investigated how factors like demography, academic success, personality traits, and test scores affect starting pay. Regression analysis has been conducted in this study using a machine learning approach. These processes

made use of the SVM, RF, and Naive Bayes algorithms. The results showed that skills like English, math ability, and the motivation to work hard and finish tasks well play an important role in the starting salaries of engineering graduates in the Indian job market. It has been shown that engineering major, college affiliation, faculty name, and academic achievement at the faculty level have been the most important determinants of starting pay.

[11] proposed a comprehensive, professional, and economy-wide framework for salary prediction. Based on professional and organizational factors, five different supervised machine learning algorithms have been trained. Dataset has been obtained from the Saudi Arabian labor market for estimating the average annual salary across economic activities and major occupational groups. According to the salary prediction results on economic activities, Bayesian Gaussian Process Regression showed a significant improvement in R^2 compared to MLR with a value of from 0.50 to 0.98. Additionally, Root Mean Square Error has been reduced by 80% and MAE decreased by nearly 90% compared to MLR. However, for salary prediction across major occupational groups, ANN demonstrated the best performance in terms of both R^2 metrics, resulting in an improvement from 0.62 to 0.94 compared to MLR, and errors have been reduced by approximately 60%.

[12] developed salary prediction models using three supervised machine learning techniques which have been Neural Networks (NN), RF, and LR. A dataset of over 20,000 incomes in the United States by utilizing the advantages of data science has been gathered and used to train the prediction models. The output of the three models has been analyzed and compared. According to the results, the NN outperformed the other machine learning models, and LR takes the shortest amount of time (0.363 seconds) with an Accuracy level of 83.2% to train the model.

[13] aimed to develop a salary prediction system that provides better assistance to college students regarding the salary that they can aspect after completing their course. For this purpose, a system that compares the student profile with those of graduated students has been developed using data mining techniques. Additionally, an experiment on student datasets using 10-fold Cross-Validation has been performed.

[14] aimed to predict salary data using linear and nonlinear machine learning models. 5 different machine learning models have been chosen, and a comparative experiment has been conducted-based on their usage in worldwide research. Models have been evaluated by using metrics commonly applied to time series-based models. Although it employs fewer parameters, Autoregression produced the best results. The Multilayer Perceptron, which is based on an ANN, came in second. Additionally, the Convolutional Neural Network and Autoregressive Integrated Moving Average models produced enough outcomes to be used for the forecast. Since the Moving Average model had the lowest success rate, it has been determined that it is the least suitable methos for solving this problem.

[15] aimed to predict individuals' yearly salary amounts. To achieve this, a computerized system that forecasts future salaries-based on historical data and generates graphical representations of salary growth over time has been developed. The system has been retrieving data from the organization's compensation database and uses it to create visual graphs. After analyzing relevant salary variables, the graphical representations and uses a prediction algorithm to estimate future salary values have been generated. Additionally, it has been also adapted to the use of the system for other estimation tasks beyond salary estimation.

Dataset Generation

The dataset has been created using the actual salaries of employees at Innovance. The dataset has been synthetically expanded to 10,000 rows. The dataset has no geographic scope and does not pose ethical concerns. As part of data preprocessing, missing data have been filled using mean, median and mode values. Outlier cleaning has been performed using Z-score and interquartile range (IQR) methods. Normalization has been applied to numerical variables with MinMaxScaler and StandardScaler methods. Encoding has been performed using LabelEncoder and OneHotEncoder techniques for categorical variables. Attributes and their descriptions are given in Table 1. Derived attributes and their descriptions are given in Table 2.

Table 1: Attributes and their descriptions

Attribute	Description
ID_Number	Unique identifier assigned to each individual.
Name Surname	Full name of the individual.
Gender	The individual's gender.
Age	The current age of the individual.
Email	The individual's email address.
Phone	The individual's contact phone number.
City	City of residence.
District	District within the city of residence.
Rol_Category	The category of the job role.
Rol	Specific job position.
Years_of_Experience	Total number of years the person has worked professionally (float).
Title	Combined seniority and role.
Education_Level	Highest level of education attained (e.g., Bachelor's, Master's, Ph.D.).
Field of Education	The field or major studied.
Graduation Year	Year the individual completed their most recent degree.
Start Date	Date the individual started their current job.
Last Promotion Date	Date of the most recent promotion in their career.
Company	Name of the company where the person works.
Company Size	Classes such as Startup, Small, Medium, Corporate.
Industry	Sector or industry the company operates in.
Work Style	Mode of working (e.g., full time, part time).
Remote Work Rate	Percentage of time worked.
Technologies Used	List of tools the individual uses.
Main Programming Language	The primary programming language they work with.
English Level	Proficiency in English (e.g., Beginner, Intermediate, Advanced, Native).
Other Languages	Any other languages the individual speaks and their proficiency.
Certifications	Professional certificates earned.
LinkedIn Profile	URL to their LinkedIn account.
GitHub Profile	URL to their GitHub account or portfolio.
Total Number of Projects	Count of completed or contributed projects.
Annual Leave Days	Number of leave days allotted per year.

Table 2: Derived attributes and their descriptions

Attribute	Description
Salary	Current annual or monthly salary.
Seniority	Level of experience or seniority.
Birth_year	The year the individual was born.
Technical Skill Score	0-100 points based on experience and certification.
Soft Skill Score	0-100 points based on role and seniority.

Feature Engineering

Some columns in the dataset have been excluded from the modeling phase because they have not been directly related to the estimation process or contained high variance. These columns have been categorized into four main groups: (i) variables containing identity and personal information (ID_No, First_Name, Last_Name, Email, Phone); (ii) user profile links and textual fields (LinkedIn_Profile, GitHub_Profile, Certificates); (iii) multi--component and free-text-based columns (Technologies Used, Main_Programming_Language, Other_Languages); and (iv) date information not directly included in the model (Employment_Start_Date, Last_Promotion_Date).

To ensure a more robust and balanced learning experience for the model, the following numerical variables have been normalized using the StandardScaler method. As a result of this process, each variable has been rescaled to a mean of 0 and a standard deviation of 1: Age, Years of Experience, Graduation Year, Year of Birth, Remote Work Rate, Total Number of Projects, Technical Skill Score, Soft Skill Score, and Annual Leave Days.

Ordinal categories have been converted to numeric form using LabelEncoder. The converted categories and their numerical values are shown in Table 3. The initially categorical variables City, district, role, role category, field of work, job type, company size and industry have been converted to one-hot encoding using pd.get_dummies(). This process resulted in a final feature vector of approximately 197 variables.

Table 3: The converted categories and their numerical values

Column	Description
Seniority	Intern (0) → Director (7)
Education_Level	High School (0) → PhD (4)
English_Level	Beginner (0) → Native Language (4)
Gender	Female (0), Male (1)

To ensure attribute compatibility and manage missing columns, the data collected from the user on the prediction screen has been aligned according to the column structure used in the training phase. Columns included in the training but not in the user data have been automatically filled with a value of zero (0) to ensure model consistency. The list of attributes used in the training process has been saved as the feature_columns.pkl file and reused within the application to maintain the same column structure during the model's prediction phase.

Methodology

Linear Regression

LR is a well-established and well-known modeling method widely used in statistics and machine learning. The

goal is to reveal the relationship between variables and use this relationship to make predictions. In this context, regression models analyze two basic types of relationships: If variables tend to increase together, the relationship is considered positive; if one increases while the other decreases, the relationship is considered negative. LR examines whether there is a statistically significant relationship between two or more variables. These variables play different roles in the model: The dependent variable is the variable to be predicted, while the independent variable, or explanatory variables, are used to explain the effects on the dependent variable. The mathematical representation of the model can be expressed as a linear line on a two-dimensional plane [16].

AdaBoost

AdaBoost is a method that aims to build a strong prediction model by sequentially combining weak learners. Weak learners are obtained by machine learning algorithms applied to randomly selected samples from the dataset. In each training cycle, observations are assigned specific weights, and these weights are used to learn the next prediction hypothesis. Misclassified examples are identified. These examples are then assigned higher weights by the next base learner. This process continues iteratively until the model has learned the target output with sufficient accuracy. The final output of the model is obtained by taking the weighted average or median of the predictions of the individual base learners [16].

Random Forest

Breiman RF is an ensemble learning method widely used for tasks such as classification, clustering, regression, and interaction detection. A single DT often suffers from high variance and bias, making it unreliable. However, RF overcomes these problems by combining multiple trees, providing more stable and accurate models. It constructs a forest by generating hundreds of random binary trees, each constructed using bootstrap samples and a random subset of variables at each node. It is-based on the Classification and Regression Trees (CART) methodology. For each tree, the out-of-bag (OOB) error rate is calculated using data not included in the bootstrap sample. Final predictions are made by majority voting across all trees. Variable importance is evaluated using metrics such as mean reduction in Gini impurity and mean reduction in accuracy. These metrics are commonly used for feature ranking and selection. To optimize model performance and reduce OOB error, two key parameters need to be tuned: the number of variables considered at each node and the total number of trees in the forest [17].

Ridge Regression

When L2 regularization is used during model training to prevent overfitting, this method is called RR. L2 regularization is an approach that essentially penalizes the magnitude of the model's weights along with the error term. This method aims to minimize the model's learning function and finds the weight vector that minimizes the sum of the squared difference between the inputs and outputs and the sum of the squared weights. Here, we use a matrix containing n data points, each with d features, and an n-dimensional vector containing the outputs corresponding to each data point. Lambda, a positive parameter, controls the magnitude of the model's weights. The larger the lambda value, the smaller the weights the model attempts to learn; this limits the model's complexity and helps prevent overfitting. When RR is combined with the kernel method, the samples are projected onto a high-dimensional space using a nonlinear transformation. Using these high-dimensional representations, the model attempts to learn the complex relationships in the data [18].

Categorical Boosting

CatBoost is an open-source, Gradient Boosting-based machine learning algorithm. It distinguishes itself from previous methods with its ability to process text, categorical, and numerical data, its fast learning process, GPU support, and diverse visualization capabilities. It can directly handle missing or categorical data without requiring an additional coding step during data preprocessing. The model's built-in functions are used to efficiently process categorical data and optimize CatBoost's settings. Particular attention is paid to the model's hyperparameters: depth, learning rate, and number of trees. Feature importance ratings are used to iteratively refine variables [19].

Voting Regressor

A voting ensemble is a machine learning ensemble

methodology that uses many methods in lieu of a single model to increase the system's performance. This approach can be applied to both classification and regression issues by combining the results of numerous methods. For regression issues, the ensembles for which are referred to as voting regressors (VRs), the estimators of all models are averaged to get a final estimate. There are two approaches to awarding votes: average voting (AV) and weighted voting (WV). In the case of AV, the weights are equivalent and equal 1. A disadvantage of AV is that all of the models in ensemble are accepted as equally effective; however, this situation is very unlikely, especially if different machine learning algorithms are used. WV specifies a weight coefficient to each ensemble member. The weight can be a floating--point number between 0 and 1, in which case the sum is equal to 1, or an integer starting at 1 denoting the number of votes given to the corresponding ensemble member [20].

Stacking Regressor

Stacked generalization, also known as stacking, has been proposed by Wolpert in 1992. This is a heterogeneous learning technique that combines several base learners to train a model, unlike homogeneous bagging and boosting methods, which directly combine the outputs of several learners to obtain the final prediction. Generally, stacking consists of several base learners (level 0) and a meta-learner (level 1), in which the outputs of the base learners serve as inputs to the meta-learner. Both the precision and diversity of the base learners affect the performance of a stacking algorithm. Diversity is a measure of the dependence or complementarity between learners [21].

Support Vector Machine

SVM is a common method used today, primarily for the regression and classification of small, high-dimensional, nonlinear samples. The SVM is based on the principle of minimum structural risk and the VC dimensionality of statistical learning theory. By using small sample data, learning is performed without introducing errors, and the model's accuracy is examined. The best universal capability is achieved by minimizing the deviation of the hyperplane from the sample points. Both linear and nonlinear regressions are included in SVMs. Important parameters affecting performance are the cost loss function (regularization parameter) and the kernel function, which measures the similarity between data points (i.e., between reflection values) [22].

Decision Tree

DT, also known as Classification and Regression Trees (CART) in the literature, are an effective method for solving not only classification problems but also regression problems. Regression trees are constructed recursively as a binary structure by selecting the most appropriate features and split points based on the minimum squared error criterion. Because the tree structure can be dynamically adapted to the distribution and characteristics of the dataset, there is no need to predetermine the functional form of the model. This flexibility allows it to work with both continuous and discrete variables simultaneously. However, as the structural complexity of the tree increases, problems such as overfitting or getting stuck in local minima are likely to be encountered during the model's learning process. Such situations can negatively impact the model's generalization capacity [23].

Development of the Salary Prediction Models

Within the scope of this study, salary prediction models have been developed using MLP, RF, Ridge, CatBoost, and ensemble learning methods such as voting and stacking. Optimal hyperparameter values have been found using Grid Search and Randomized Search. Cross-validation has been applied to test the models on different data subsets and analyze their generalization performance. The dataset has been split into 80% training and 20% test data; the learning performance of the models has been evaluated on the training data, and their generalization ability has been evaluated on the test data.

If future information is inadvertently leaked during the model training phase (for example, by including the mean of the entire dataset in the encoding process), test performance can appear artificially high, which can lead to serious failures in the real world. This problem is called data leakage. To avoid this problem, encoding has been fitted only to the training data, and only transformation has been applied to the test set. During the validation process, TimeSeriesSplit, which preserves time dependency, has been pre-

ferred over K-Fold. After each trial, variable importance levels have been examined to eliminate suspicious features that could cause leakage. In addition, modeling has been performed on a daily basis by creating special models that work separately for each day; this way, time-varying factors such as seasonal effects, campaigns and price change periods could be captured more precisely.

All models used have been taken from the sklearn package and optimized with appropriate hyperparameters. The hyperparameter values of the models are presented in Table 4.

Table 4: The hyperparameter ranges of the models

Models	Hyperparameter ranges
Voting	RR (“Alpha”: [0.1])RF (“N_Estimators”: [100],“Random_State”: [42])CatBoost (“Depth”:[4]“Iterations”:[100]“Learning_Rate”:[0.1])
Stacking	RR (“Alpha”: [0.1])RF (“N_Estimators”: [100], “Random_State”:[42])CatBoost(“Iterations”:[500], “Learning_Rate”:[0.1], “Depth”:[6], “Verbose”: [0], “Random_State”: [42])
Ridge	“Alpha”: [0.1 - 10]
DT	“Max_Depth”: [2 – 10] “Min_Samples_Split”: [2 – 10]
RF	“N_Estimators”: [50 - 150]“Min_Samples_Split”: [2]“Max_Depth”: [4 - 10]“Leaf_Node”: [250]
Grid Search CV (for RF)	“N_Estimators”: [50 - 150] “Max_Depth”: [4 - 8] “Min_Samples_Split”: [2 - 10]
SVM	“C”: [0.1 - 10]“Epsilon”: [0.01 - 0.2]
CatBoost	“Iterations”: [100 – 500]“Learning_Rate”: [0.01 – 0.3]“Depth”: [4 – 10]“I2_Leaf_Reg”:[3]“Random_State”: [42]“Verbose”: [0]“Early_Stopping_Rounds”: [50]
AdaBoost	“Max_Depth”:[4]“N_Estimators”:[100]“Learning_Rate”:[1.0]“Random_State”:[42]

Ensemble Modeling

The first approach utilized ensemble models, created by combining the outputs of multiple regression models. This structure aims to increase forecast accuracy by integrating the strengths of different model types. In the Voting Regressor method, the unweighted average of the forecast outputs of models such as Ridge, RF, and CatBoost has been taken. In the regression task, an unweighted (uniform) soft average method has been preferred instead of the hard voting approach commonly used in classification problems. In the Stacking Regressor approach, the final forecast has been obtained by combining the forecasts generated by Ridge, RF, and GB using a higher-level meta-model. The outputs of each base model have also been evaluated independently, and then the resulting forecast values have been classified according to pre-optimized ranges before proceeding to the final decision stage.

For each employee, the market gap (market_gap) has been calculated by referencing both the individual-level estimated salary and the relevant sector- and location-based market salary. In this analysis, the primary attributes considered in the model's learning process are the employee's number of skills, their most recent educational level, and the location-based market salary. The outputs of the model include predicted salary (predicted_salary), market salary (market_salary) and the difference category (difference_category) based on the difference between these two values, and this category consists of three classes: “low”, “market” and “high”.

Machine Learning-based Modeling

In this study, supervised learning approaches have been utilized in the field of machine learning as an alternative to traditional regression algorithms. The primary rationale for this choice has been the need to develop more flexi-

ble and robust models on tabular datasets with high dimensions and complex distributions. Supervised learning methods offer the advantages of more effectively modeling non-linear and complex relationships, learning the interactions between variables, processing categorical and numerical variables together, and optimizing the model's predictive performance based on specific metrics. For these reasons, the use of these methods has been deemed appropriate in this study.

Results and Discussion

The results obtained with the developed models have been evaluated using MAE, MAPE, R^2, Model Training Runtime (ms), Explainability metrics. Explainability refers to the extent to which the internal operating mechanisms of models are understandable and interpretable. The MAE, MAPE and R^2 values of the developed models are presented in Table 5. The Model Training Runtime (ms) and Explainability values of the developed models are presented in Table 6.

Table 5: The MAE, MAPE and R^2 values of the developed models

Model	MAE	MAPE	R^2
LR	2300	9.31%	0.936
RR	2150	9.28%	0.936
DT	0	0	0.1000
RF	1820	2.60%	0.993
AdaBoost	2400	10.31%	0.937
Voting	1400	5.66%	0.973
Stacking	1000	4.00%	0.975
CatBoost	1450	6.10%	0.910
SVM	1900	8.80%	0.890

Table 6: The Model Training Runtime (ms) and Explainability values of the developed models

Model	Model Training Runtime (ms)	Explainability
LR	170 ms	High
RR	185 ms	High
DT	190 ms	Intermediate (Overfit)
RF	2350 ms	Intermediate
AdaBoost	3250 ms	Low
Voting	5650 ms	Intermediate
Stacking	5000 ms	Intermediate
CatBoost	9700 ms	High
SVM	19500 ms	Low

- The superior model has been developed using RF, achieving a MAPE of approximately 2.60% and a R^2 of 0.993. The ability of RF to handle non-linear relationships, capture complex trait interactions,

and reduce overfitting through ensemble averaging provided the most successful result. Despite its high predictive performance, the model's explainability remained moderate, indicating limited interpretability of the decision-making process. Training time has been recorded at 2350 ms, indicating a moderate cost.

- The CatBoost model, with a MAPE of 6.10% and an R^2 of 0.910, showed competitive performance, though it exhibited lower predictive performance than RF. However, the training time has been significantly longer at 9700 ms.

- The LR and RR-based models exhibited the highest model interpretability and the shortest training times.

- The DT model suffered from overfitting, indicating poor generalization capacity.

- Both the AdaBoost and SVM models performed relatively poorly compared to the other approaches. The SVM model tends to be computationally expensive on large datasets and has shown lower predictive performance due to its high sensitivity to the choice of kernel and hyperparameters. On the other hand, the AdaBoost model is thought to sequentially adjust weights based on previous errors, overemphasizing noisy or uninformative examples and leading to poor generalization.

The highest accuracy performance has been achieved with the RF model, which stands out as the most reliable individual model in the project in terms of both prediction accuracy and generalizability. While the RR model provides fast training time and high explainability, it underperformed the other models in terms of accuracy. SVM, while a theoretically powerful approach, requires longer training times on large datasets and is more sensitive to hyperparameter settings. Furthermore, ensemble methods outperform individual models, but this presents a disadvantage in terms of interpretability.

A balance must often be struck between accuracy and interpretability during the model selection process. Complex models like RF and CatBoost offer high accuracy rates thanks to their ability to learn complex relationships among numerous variables. However, these models generally operate in a black-box fashion. In contrast, models like RR and LR, while having lower predictive accuracy, are highly interpretable because they clearly demonstrate the impact of each input on the outcome. This is particularly important in application areas such as employee communication, ethical assessments, or auditing processes, where justification for decisions is critical. The understandability, transparency, and explainability of the model are also crucial. In this context, it can be argued that in certain use cases, it may be more appropriate to choose models with lower accuracy but easy explanations over models that offer high accuracy but low interpretability.

When the results of the study have been evaluated in terms of practical applicability, HR departments can leverage the RF model to develop accurate salary forecasting systems that help establish data-driven, fair compensation policies. The RF model stands out as the most reliable model in terms of accuracy and generalizability, while linear models serve as interpretable baselines suitable for transparency-focused applications. In scenarios where explainability is more important than predictive accuracy, such as employee communication or auditing processes, models like RR may be preferred.

Conclusion

Increased employee turnover in the IT sector not only increases the risk of corporate knowledge loss for businesses but also complicates the onboarding process of new employees. To prevent this, providing employees with appropriate salary offers has become a strategic necessity for increasing employee motivation and optimizing operating costs. In this study, salary prediction models have been developed using machine learning-based LR, RF, RR, CatBoost, SVM, DT and AdaBoost and ensemble learning-based Voting and Stacking methods. When the results obtained with the developed prediction models have been examined, it has been observed that the RF model achieved the highest success rate with a MAPE value of 2.60%. In contrast, the SVM and AdaBoost models exhibited weaker prediction performance compared to other approaches. The

main difference of this study from similar studies in the literature is that it also evaluated ensemble learning approaches in addition to machine learning methods. While many studies focus solely on accuracy metrics such as MAE and MAPE, this study also analyzed critical operational performance metrics such as model training time (in milliseconds) and explainability. Furthermore, necessary precautions have been taken during the model training process to mitigate the risk of data leakage. This aimed to increase the generalizability of the obtained performance results to real--world scenarios. All these elements distinguish the study from the existing literature and make it a significant contribution.

References

1. Ji Y, Sun Y, Zhu H (2025) Enhancing Job Salary Prediction with Disentangled Composition Effect Modeling: A Neural Prototyping Approach. arXiv preprint arXiv:2503.12978.

2. Xu Michael (2025) Salary prediction using machine learning. Scholarly Review Journal.

3. Kumar AV, Swathi M, Mahesh N, Mahesh S, Ravishankar A (2025) Enhancing Salary Predictions with Ensemble Learning Techniques.

4. Malaiarasan S, Riyaz MA, Appadurai M (2025) Salary Prediction Using Machine Learning. International Journal of Scientific Research and Engineering Development. 8.

5. Deb D (2025) Data Science Job Salary Prediction Using Linear Regression.

6. Yan Y, Li X, Wang Y, Cai Y, Liao J (2025) Prediction Model for Dynamic Fluctuations in Salary by Integrating Multiple Time Series. International Journal of High Speed Electronics and Systems, 2540481.

7. Asaduzzaman A, Uddin MR, Woldeyes Y, Sibai FN (2024) A Novel Salary Prediction System Using Machine Learning Techniques. In 2024 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering: 38-43.

8. Dsouza OD, Goel S, Mallick A, Gilbile SP, Chitturi A Salary Estimator using Machine Learning.

9. Aminu H, Zambuk FU, Abdullahi A, Nanin ER, Yakubu, IZ Salary Prediction Model using Principal Component Analysis and Deep Neural Network Algorithm.

10. Saeed AKM, Abdullah PY, Tahir AT (2023) Salary Prediction for Computer Engineering Positions in India. Journal of Applied Science and Technology Trends. 4: 13-8.

11. Matbouli YT, Alghamdi SM (2022) Statistical machine learning regression models for salary prediction featuring economy wide activities and occupations. Information. 13: 495.

12. Kablaoui R, Salman A (2022) Machine learning models for salary prediction dataset using python. In 2022 International Conference on Electrical and Computing Technologies and Applications: 143-7.

13. Mishra P, Srivastava S, Gupta P, Anand A, Gupta SC (2021) A Comparative Study of Machine Learning Algorithms for Salary Estimation. In 2021 3rd International Conference on Advances in Computing, Communication Control and Networking: 1698-703.

14. Skledar P, Forecasting Univariate Time Series Salary Data with Machine Learning Models.

15. Das S, Barik R, Mukherjee A (2020) Salary prediction using regression techniques. Proceedings of Industry Interactive Innovations in Science, Engineering & Technology.

16. Shanmugasundar G, Vanitha M, Čep R, Kumar V, Kalita K, et al. (2021) A comparative study of linear, random forest and adaboost regressions for modeling non-traditional machining. Processes. 9: 2015.

17. Park S, Kim J (2019) Landslide susceptibility mapping-based on random forest and boosted regression tree models, and a comparison of their performance. Applied Sciences. 9: 942.

18. Nepal NK, Ghimire MP (2023) Predicting band gap of transition metal trihalides using machine learning. Journal of Nepal Physical Society. 9: 34-41.

19. Zhang F, Fleyeh H, Bales C (2022) A hybrid model-based on bidirectional long short-term memory neural network and Catboost for short-term electricity spot price forecasting. Journal of the Operational Research Society. 73: 301-25.

20. Erdebilli B, Devrim-İçtenbaş B (2022) Ensemble voting regression-based on machine learning for predicting medical waste: a case from Turkey. Mathematics. 10: 2466.

21. Zhang Y, Liu J, Shen W (2022) A review of ensemble learning algorithms used in remote sensing applications. Applied Sciences. 12: 8654.

22. Yuan H, Yang G, Li C, Wang Y, Liu J, et al. (2017) Retrieving soybean leaf area index from unmanned aerial vehi-

cle hyperspectral remote sensing: Analysis of RF, ANN, and SVM regression models. Remote Sensing. 9: 309.

23. Wang J, Li P, Ran R, Che Y, Zhou Y (2018) A short--term photovoltaic power prediction model-based on the gradient boost decision tree. Applied Sciences. 8: 689.