# DB13

*by* Vamshikrishna Namani

**Submission date:** 09-Mar-2023 03:07PM (UTC+1000)
**Submission ID:** 2032749520
**File name:** DB13.docx (251.57K)
**Word count:** 2296
**Character count:** 12674

# Salary Prediction

*Peddireddy Venkata Rohith*
*Student*
*Department of Computer Science and Engineering*
*Narasaraopet, India*
*rohithpeddireddy24@gmail.com*

***Abstract*— The Salary Prediction project aims to predict the salaries of employees based on various factors such as job title, years of experience, location, education level, and industry. The project uses machine learning algorithms to analyze a dataset of historical salary and employee information to develop a model that can accurately predict future salaries.The dataset used in this project contains information such as job title, years of experience, location, education level, and industry for thousands of employees. The dataset is cleaned and preprocessed to remove any missing or irrelevant data.Various machine learning algorithms such as linear regression, decision tree, and random forest are used to develop the salary prediction model. The performance of each algorithm is evaluated using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared.The final model is then deployed to make salary predictions for new employees based on their job title, years of experience, location, education level, and industry. The model can be used by companies to help with salary negotiations, employee retention, and workforce planning. The salary prediction project has the potential to help job seekers, employers, and policymakers make informed decisions about salaries and compensation packages. By providing accurate predictions of salaries, this model could help job seekers negotiate better salaries, assist employers in setting appropriate salaries for their employees, and help policymakers make informed decisions about minimum wages and labor policies.Keywords*—**MachineLearning , LinearRegression, Random Forest, Logistic regression, Flask**

## I. INTRODUCTION

A salary prediction project is a data science project that aims to build a model to predict the salary of a job position based on various factors such as education, years of experience, job title, industry, location, and other relevant factors.The project involves collecting and analyzing data from various sources such as job listings, industry reports, and online surveys. Once the data is collected, it is cleaned, preprocessed, and transformed into a format that can be used for modeling.The next step is to choose an appropriate machine learning algorithm to build the model. This may involve trying out different algorithms and selecting the one that gives the best performance. The model is trained on a subset of the data and evaluated on a separate set to ensure that it can accurately predict salaries for new job positions.Once the model is developed and validated, it can be used to predict salaries for new job positions. This can be useful for job seekers who want to negotiate their salaries or for companies that want to ensure they are offering competitive salaries to their employees.

Overall, a salary prediction project is an exciting and challenging data science project that can provide valuable insights into the job market and help people make informed decisions about their careers.and random forests to train the model.Once we have developed the model, we will evaluate its performance by testing it on a separate set of data. This will enable us to determine the accuracy of the model and identify any areas for improvement.Overall, a salary prediction project can be a valuable tool for employers and employees alike. By accurately predicting salaries, employers can ensure they are offering fair compensation and employees can negotiate their salaries more effectively

## II LITERATURE REVIEW

There have been several studies on salary prediction in recent years, focusing on different aspects of the problem. Some studies have looked at the impact of various factors on salary, such as education level, job experience, and industry sector. Other studies have explored the use of different machine learning algorithms for salary prediction.Predicting Salary from Job Posting Text" by Ahmad A. Tafti and Arman Cohan (2018) - This paper explores the use of natural language processing (NLP) techniques to predict salaries from job postings. The authors use various machine learning algorithms,such as linear regression and support vector machines, and compare their performance.Predicting Salaries with Machine Learning" by Arjun Adhikari (2019) - This article provides a step-by-step guide on how to build a salary prediction model using Python and scikit-learn. The author explains various preprocessing techniques, feature engineering, and model selection.A Deep Learning Model for Predicting Job Salaries" by Kai Shu et al. (2018) - This paper proposes a deep learning model for

predicting salaries based on job descriptions. The authors use a combination of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to capture both the semantic and syntactic information in the text.Predicting Employee Salary Using Machine Learning Techniques" by Ayesha Waheed et al. (2019) - This paper presents a study on predicting employee salaries using various machine learning algorithms, such as decision trees, random forests, and gradient boosting. The authors evaluate the performance of these algorithms on a real-world dataset.Predicting Salaries for Job Postings Using Machine Learning" by Brian Dew et al. (2018) - This article discusses a project in which the authors built a salary prediction model using job postings and salary data from Glassdoor. The authors use various NLP techniques, such as word embeddings and topic modeling, to extract features from the job postings.These papers and articles demonstrate the wide range of techniques and approaches that can be used for salary prediction, from traditional machine learning algorithms to deep learning models and NLP techniques

## III.EXISTING SYSTEM

The existing system,is used to predict the salary, based on the machine learning algorithms and data mining algorithm were widely used. Salary prediction is a popular application of machine learning in the automotive industry. In this project, the goal is to predict the salary based on various factors , such as work, age ,education martial status ,occupation, experience and other factors.The major drawback of this existing system is they need more attributes in order to predict the cars price . More comparison techniques must be used to get the result more efficiently.

Some of the challenges in car price prediction include dealing with outliers, handling missing data, and selecting the most relevant features. To address these challenges, various techniques such as data imputation, feature selection, and regularization are used in the existing systems of salary prediction projects in machine learning to get more accuracy .

## II. PROPOSED SYSTEM

The proposed system is based on the different factors, and features also with the help of experts knowledge the salary prediction has been done accurately to predict best salary. These are the features useful to develop a efficient and effective model which predicts the salary according to the user inputs In this paper, we applied different models and techniques and methods in order to achieve higher accuracy of the precision of the salary prediction.

## III. DATASET AND DATA VISUALIZATION

We have used the dataset available in Kaggle which consists different types of employment. The dataset is split into training and testing datasets . The training data is 80% of the total dataset, validation data is 10% of dataset and testing data is other 10% of the dataset.

Data visualization is an important tool for data analysis and communication that enables us to visually represent complex datasets and identify patterns and relationships within the data. Visualizations can take various forms, such as scatter plots, line charts, bar charts, histograms, heat maps, box plots, tree maps, and many more.

It is The choice of visualization technique depends on type of data and the specific insights being communicated. The goal of data visualization is to communicate complex information clearly and effectively, making it easy for the audience to understand the key insights and trends in the data.

Data visualization can take many forms, including, graphs, maps, charts and other visual aids. It is often used in fields such as business, science, engineering, medicine, and social sciences to present data in a way that is accessible and easy to understand.

| age | workclass | fnlwgt | education | education-num | marital-status | occupation | relationship | race | sex |
|---|---|---|---|---|---|---|---|---|---|
| 39 | State-gov | 77516 | Bachelors | 13 | Never-married | Adm-clerical | Not-in-family | White | Male |
| 50 | Self-emp-not-inc | 83311 | Bachelors | 13 | Married-civ-spouse | Exec-managerial | Husband | White | Male |
| 38 | Private | 215646 | HS-grad | 9 | Divorced | Handlers-cleaners | Not-in-family | White | Male |
| 53 | Private | 234721 | 11th | 7 | Married-civ-spouse | Handlers-cleaners | Husband | Black | Male |
| 28 | Private | 338409 | Bachelors | 13 | Married-civ-spouse | Prof-specialty | Wife | Black | Female |

Fig.1 : Data Set

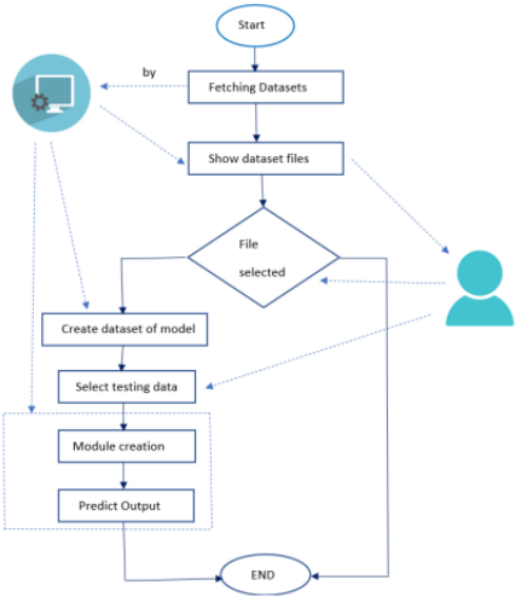


Fig.2: System architecture diagram

## IV. PREPROCESSING

Data preprocessing is an important step in the data analysis process, where raw data is transformed into a format that is suitable for analysis. Here are some common techniques used in data preprocessing:

Data Cleaning: It is the process of removing or correcting any errors or inconsistencies in the data.This can include removing duplicates, correcting misspelled values, or imputing missing data.

Data Transformation: It is the process of converting data from one format to another, such as converting categorical data to numerical data. This can also include scaling data to a common range or normalizing data have a mean of zero and standard deviation of one.

Data Reduction: Data reduction involves reducing the amount of data to be analyzed. This can include identifying and removing irrelevant features or reducing the resolution of data by aggregating it into larger groups.

Overall, data preprocessing is a major step in data analysis process as the data is consistent, accurate and in a format that can be easily can be analyzed.

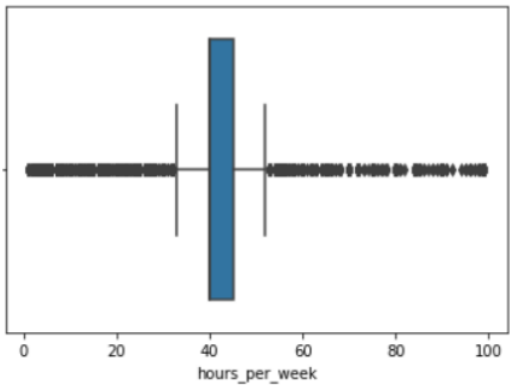


Fig.3: Removing outliers

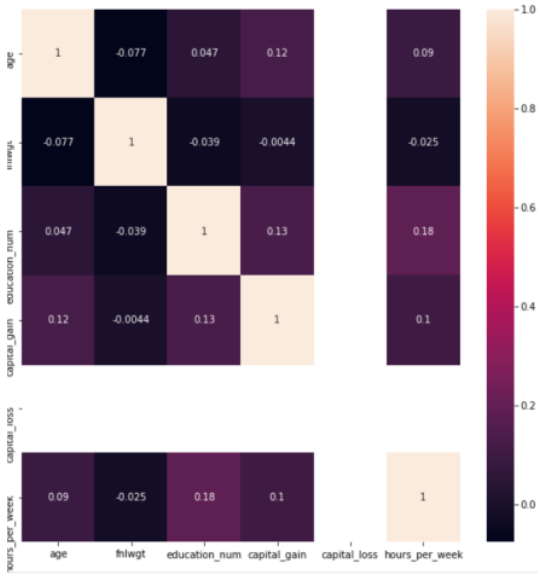


Fig .4: Co-relation

## V. METHODOLOGY AND IMPLEMENTAION

System analysis of car price prediction involves understanding the various components and processes involved in the system, and how they work together to predict the price of a car. Here are some key components of a car price prediction system:

Data Collection: The system needs to collect data from various sources, such as historical sales data, market trends, and car specifications. This data is then used to train the prediction model.

Data Preprocessing: The raw data collected from various sources may not be in a format that is suitable for analysis. Therefore, data preprocessing techniques such as data mining ,data cleaning, integration and transformation need to be applied to ensure that the data is accurate and consistent.

Feature Selection: The system needs to identify which features of the car are relevant for predicting its price. This can be done using statistical techniques or machine learning algorithms.

Prediction Model: The prediction model is trained using the preprocessed data and selected features. In the Various algorithms such as regression models, decision trees, can be used to build the prediction model.

Model Evaluation: The prediction model needs to be evaluated to assess its accuracy and effectiveness. This can be done using metrics such as mean squared error, root mean squared error, or R-squared.

Deployment: The prediction model is then deployed into a production environment, where it can be used to predict the price of a car based on its specifications.

Overall, a car price prediction system requires a combination of data collection, preprocessing, feature selection, prediction modeling, model evaluation, and deployment. The accuracy and effectiveness of the system depend on the quality and quantity of data collected, the effectiveness of preprocessing techniques, and the choice of machine learning algorithms used in the prediction model.

Different types models are used to find best accuracy:

Those are:

1.Linear Regression: It is a supervised and statistical Machine Learning algorithm used to predict a continuous output variable (also known as a dependent variable) based on one or more input variables (also known as independent or predictor variables). The relationship between the input variables and output variable is assumed to be linear, meaning that the relationship can be represented by a straight line. The algorithm tries to find the best fitting line (known as the regression line) that passes through the data points, minimizing the difference between the predicted and actual values of the output variable.

The equation for a simple linear regression can be written as:

$y = k0 + k1 * x$

where y is output variable, x is input variable, k0 is the intercept, and k1 is the coefficient of the input variable.

2.Random forest:

Random Forest is a supervised and versatile algorithm that can be used for classification and regression problems. It is also relatively easy to use, requires minimal data pre processing, and can handle both numerical and categorical data. Additionally, Random Forest has the ability to handle missing data and outlier values, making it a popular choice for many machine learning applications.

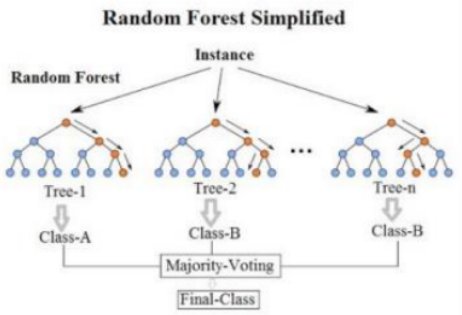


Fig 5: Random Forest Model

3.Logistic regression: Logistic Regression is a supervised and statistical method used for binary classification problems, where the outcome variable takes only two values (0 or 1). The goal of logistic regression is to find out the best fitting line (or hyperplane in higher dimensions) that separates the two classes.

The logistic function is given by:

$p = 1 / (1 + \exp(-z))$

where p is the probability of the positive class, z is the linear combination of the input features and their corresponding weights, and exp is the exponential function.

Logistic Regression works by optimizing the weights of the input features to maximize the likelihood of the observed data given the model parameters. This optimization is usually done using maximum likelihood estimation or gradient descent.

Logistic Regression is a popular algorithm due to its simplicity and interpretability. It can handle both categorical and continuous input features and is robust to noise and outliers. Additionally, it can be easily extended to handle multi-class classification problems using techniques such as One-vs-All and Softmax regression.

## VIII.RESULT AND ANALYSIS

The accuracy of the different model is shown below:

| Algorithm | Accuracy |
|---|---|
| Linear Regression | 83.20 |
| Random Forest | 85.37 |
| Logistic Regression | 92.47 |

The above table shows the accuraces of different models which are created by using the mentioned machine learning algorithms. Among all above models ,the model which is created by using Logistic Regression algorithm got good accurance .So we consider it as the final model.

This section explains final output which detects used salary

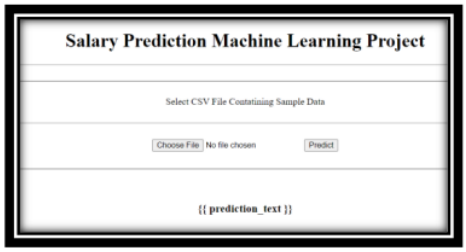


Fig.7: Home Page

IX.CONCLUSION:

We have used 3 algorithms like Linear Regression, Random Forest , Logistic Regression in order to predict the house price. The accuracy varies for different algorithms. The accuracy for Random Forest algorithm is 85.37% when. The accuracy of Linear Regression algorithm is 83.20% when correlation and information gain are applied. The highest accuracy for Logistic Regression using is 92.47%

DB13

ORIGINALITY REPORT

| 14% | 8% | 7% | 8% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | "Computer Networks and Inventive Communication Technologies", Springer Science and Business Media LLC, 2023<br>Publication | 2% |
| 2 | Submitted to Liverpool John Moores University<br>Student Paper | 2% |
| 3 | Submitted to University of Sunderland<br>Student Paper | 1% |
| 4 | Submitted to Rochester Institute of Technology<br>Student Paper | 1% |
| 5 | www.mdpi.com<br>Internet Source | 1% |
| 6 | www.researchgate.net<br>Internet Source | 1% |
| 7 | Tomáš Effenberger, Radek Pelánek, Jaroslav Čechák. "Exploration of the robustness and generalizability of the additive factors model", Proceedings of the Tenth International | 1% |

| | | |
|---|---|---|
| | Conference on Learning Analytics & Knowledge, 2020<br>Publication | |
| 8 | github.com<br>Internet Source | 1 % |
| 9 | Submitted to London Business School<br>Student Paper | 1 % |
| 10 | Submitted to Imperial College of Science, Technology and Medicine<br>Student Paper | 1 % |
| 11 | Prayitno Abadi, Umar Ali Ahmad, Yuichi Otsuka, Punyawi Jamjareegulgarn et al. "Modeling Post-Sunset Equatorial Spread-F Occurrence as a Function of Evening Upward Plasma Drift Using Logistic Regression, Deduced from Ionosondes in Southeast Asia", Remote Sensing, 2022<br>Publication | 1 % |
| 12 | www.sas.com<br>Internet Source | 1 % |
| 13 | medinform.jmir.org<br>Internet Source | <1 % |
| 14 | "Data Mining and Data Warehousing", Studies in Computational Intelligence, 2007<br>Publication | <1 % |
| 15 | mdpi-res.com<br>Internet Source | <1 % |

| | | |
|---|---|---|
| 16 | scholarworks.umt.edu<br>Internet Source | <1% |
| 17 | dokumen.pub<br>Internet Source | <1% |
| 18 | fau.digital.flvc.org<br>Internet Source | <1% |
| 19 | repository.uel.ac.uk<br>Internet Source | <1% |
| 20 | www.hindawi.com<br>Internet Source | <1% |

| | | | |
|---|---|---|---|
| Exclude quotes | On | Exclude matches | Off |
| Exclude bibliography | On | | |