MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
A Constituent Institution of Manipal University

Department of Computer Science and Engineering

A report on
Machine Learning Lab Project
[CSE-3183]

Employee Salary Prediction

Submitted By

KONALA SURYA ADHARSH REDDY
Reg no. 210962014

VENKATA SAI SURYA PRAKASH KARANAM
Reg no. 210962080

6th Semester

[Jul – Nov 2023]

Department of Computer Science and Engineering
Manipal Institute of Technology, Manipal.
November 2023

Employee Salary Prediction

1st Konala Surya Adharsh Reddy
Dept of Computer Science Engineering with AI & ML.
Manipal Institute of Technology
Manipal.
adharshreddy.387@gmail.com

2nd Venkata Sai Surya Prakash
Dept of Computer Science Engineering with AI & ML.
Manipal Institute of Technology
Manipal.
surya.karanam@gmail.com

Abstract— This project examines the use of big data analytics to predict employee salaries, providing employers and employees with important information in a dynamic compensation field. We embark on a journey of preliminary data, research, selection and design using a comprehensive database that includes variables such as education, experience, job title, industry and geography. Prediction models are carefully designed and developed using various machine learning algorithms, including linear regression, decision trees, random forests, gradient boosting, and neural networks to accurately predict salary. The findings highlight the importance of multiple factors in determining salary levels and highlight the potential for big data analytics to improve employee compensation decision-making processes. This effort contributes to the advancement of predictive modeling technology in HR, empowering employers to make informed decisions and empowering employees to negotiate fair and competitive wages.

Keywords—Linear regression, BigDataAnalytics, EmployeeSalaryPrediction…

I. INTRODUCTION

In today's fast-paced, competitive business environment, one of the key challenges organizations face is determining appropriate benefits for their employees. Paying employees not only reflects the value of their contributions, but also has an impact on recruitment, retention and overall performance. Traditionally, pay decisions were based on simple measurements or subjective evaluations, often leading to inconsistencies and inefficiencies in pay models.

However, with the advent of big data analytics, organizations now have the opportunity to use more data to improve data and use data-driven methods to predict employee salaries. Big data analytics provides powerful tools that can analyze a variety of factors, such as education, experience, job title, sector, geography, and industry, to uncover patterns and insights that can inform salary decisions. This project focuses on using big data to develop predictive models for employee salary prediction. Using advanced machine learning and techniques, we aim to create models that not only accurately predict salaries but also provide insights into the factors that influence the level of pay. These insights can help organizations make decisions about compensation structures, talent acquisition strategies, and overall performance management.

Through this project, we aim to contribute to the development of human resources research and practices and provide evaluation. By bridging the gap between data science and management intelligence, we aim to demonstrate the potential of big data analytics to transform traditional processes. Ultimately, our goal is to help organizations develop their human resources and create fairer, more competitive wages for their employees.

II. LITERATURE REVIEW

The literature surrounding employee salary forecasting and big data analysis includes various studies, methods and technical findings that highlight the importance of this field in human resources management today. Several important points have emerged from existing research that highlight the importance of a data-driven approach to understanding and predicting employee compensation.

1. Traditional Methods vs. Big Data Analytics:
Salary decision-making processes often rely on simplistic models or subjective assessments, leading to bias and inaccuracy. In comparison, big data analytics provides a more powerful and objective approach by leveraging large data sets to identify patterns and relationships that are not clearly visible through normal means (Brynjolfsson and McAfee, 2014). This shift towards data-driven decision-making is evident across the industry, where organizations are turning to predictive analytics to inform salary negotiations and compensation strategies

2. Impacts on employee salary:
There are many factors that affect employee salary, such as education, experience, and personal characteristics regarding job inclinations, business needs, and other issues. and residential areas. (Brown and Medoff, 1989)). Understanding the interaction of these factors is critical to developing accurate predictive models that capture the complexity of compensation decisions. Previous research has identified education, job title, years of experience, and location as important determinants of salary levels (Khan, 2019)

3. Machine Learning Algorithms for Salary Prediction:
Machine learning algorithms play an important role in developing predictive models to predict employee salaries. Methods such as linear regression, decision trees, random forests, slope boosting, and neural networks are commonly used to analyze salary data and identify patterns (James et al., 2013). Each algorithm has its own advantages and challenges, and the selection of the most appropriate algorithm depends on factors such as data complexity, interpretability, and accuracy of predictions.

4. Decision integrity and privacy:
The use of big data analytics in salary prediction raises important issues and privacy regarding data collection, storage and use (Davenport and Harris, 2007). To minimize the risk of bias or discrimination, organizations need to be transparent and fair in their information practices. Additionally, protecting employee privacy and confidentiality is crucial to maintaining trust and complying with regulatory requirements such as GDPR and CCPA.

5. Application and Impact:
Using predictive analytics for employee salary forecasts extends beyond the scope of an organization to impact social impacts, including workforce diversity, inequality, and labor market conditions (Frey and Osborne, 2017). Predictive models can inform policy decisions, promote fair pay, and support businesses by providing insight into wage patterns and trends.

III. DATA

In this project, we use a dataset of 10crore people across the world who are working in various industries we took the data from Kaggle.

String	Job Type, Degree, Major, Industry
Float	Job ID, Company ID, Years of experience, Miles from Metropolitan City

It seems like you've provided a dataset with various job listings and their associated attributes. Here's a brief summary of the data:

1.Each job listing has attributes such as `jobId`, `companyId`, `jobType`, `degree`, `major`, `industry`, `yearsExperience`, and `milesFromMetropolis`.

2.The `jobType` indicates the type of job (e.g., Junior, CEO, CTO, etc.).

3.The `degree` indicates the educational qualification required for the job (e.g., High School, Bachelors, Masters, etc.).

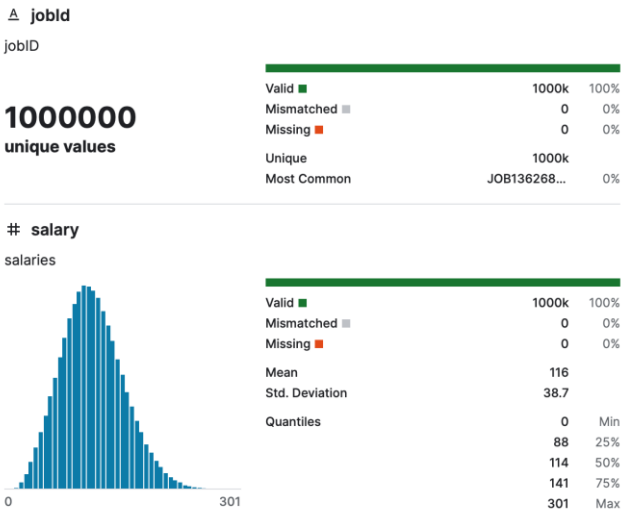
4.The `major` indicates the field of study required for the job (e.g., Biology, Chemistry, Physics, etc.).

5.The `industry` indicates the industry the job is in (e.g., Health, Auto, Oil, Finance, etc.).

6.The `yearsExperience` indicates the number of years of experience required for the job.

7.The `milesFromMetropolis` indicates the distance of the job location from the city.

It appears that there are numerical ranges associated with each job, but without context, it's difficult to interpret them precisely. Additionally, there seem to be some statistical summaries provided at the end, showing counts for certain numerical ranges.



IV. METHODOLOGY

1. **Linear Regression:**
    1. Import Output: To use linear regression for the estimated hotel bookings Model, import the LinearRegression class from scikit-learn ("import LinearRegression from sklearn.linear_model").
    2. Example Model: Create an example of the linear regression model, which is a basic algorithm often used for regression tasks, including predicting outcomes, including removing the result, for example.
    3. Training the model: Use the hotel booking dataset ("model.fit(X_train, y_train)") to train the model to capture the relationship between differences and different targets (removal of results).
    4. Prediction and measurement: Uses recursive modeling to predict the likelihood of data deletion and evaluates its effectiveness using impact metrics such as measured squared error (MSE) or R-squared to evaluate the model's goodness of fit. .

2. **Gradient Boosted Trees(GBT):**
    1. Import GBT regressor: mport GradientBoostingRegressor from scikit-learn using Boosting (`sklearn. import GradientBoostingRegressor from ensemble). algorithm for predicting hotel reservation cancellations.
3. Example Model: Create an example of gradient boosted tree regressor, a powerful ensemble learning process that can capture complex nonlinear relationship and improved prediction accuracy.

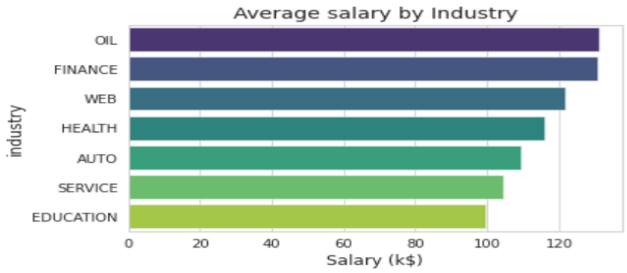
4. Training the model: Train the model using the hotel booking dataset (model.fit(X_train, y_train)) to leverage the joint power of multiple decision point trees for extrapolation.
5. Prediction and measurement: Use the gradient boosting tree regressor training model to evaluate its performance and estimate the probability of extraction from test data, using regression analysis to evaluate the accuracy and reliability of predictions.

V. RESULTS AND CONCLUSION

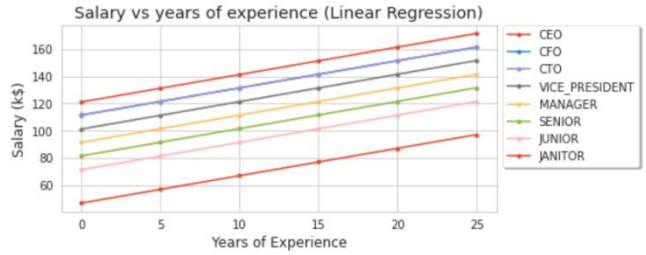
Results from the employee salary forecasting project using the PySpark framework and MLlib revealed useful information on the factors affecting salary levels and the effectiveness of forecasting models.

1. Analytical data:

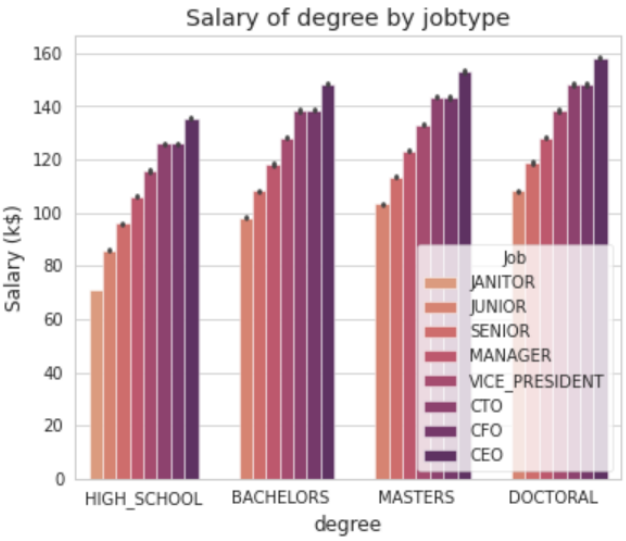
- Analysis identified sectors with the highest revenues; The sector with the highest revenue is the oil sector, which is based on financial transactions and Internet transactions.
- It has been found that better educated workers will make more money in all types of jobs.



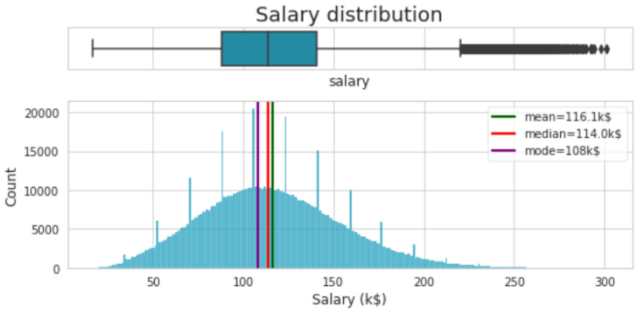
- The highest paying job is CEO, the lowest paying job is management.
-As Salary vs Years of Experience graph indicates that salary increases more with experience in CEO job.



- Engineering, business and mathematics are the most important fields associated with higher income among job types.
- JANITOR etc. Compared to , there are significant differences in average earnings across job types for inexperienced workers. The average salary for key CEO and CFO positions is higher than for lower-level positions.
- The graph of Salary vs Degree looks like the following, we can see clearly that CEO job dominates in every degree ,while Janitor job backs in every degree

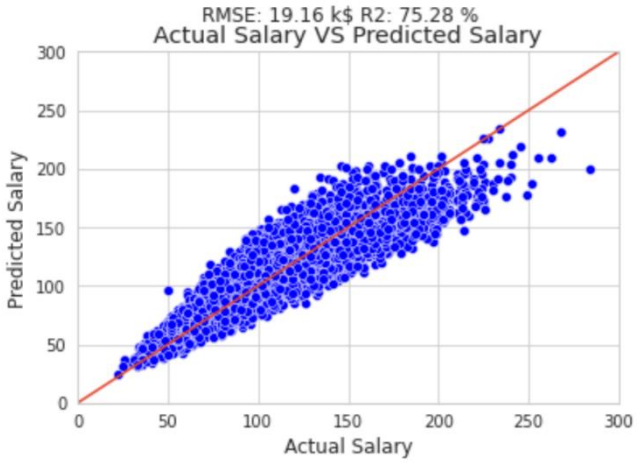


- Although there is skewness in the salary distribution, statistics show that the average salary, median salary and salary range are $116 thousand, $114 thousand and $108 thousand respectively, which shows the asymmetry of the distribution.



2. Prediction results:

- The prediction model trained using the Gradient Boosting Tree (GBT) algorithm performed well in salary prediction.
- The method is more accurate in estimating salaries below $130,000, but often results in salaries below $175,000; This can be attributed to the real skewness of the salary distribution.
- Measurements such as root mean square error (RMSE) and R-squared (R2) indicate that the performance of the model is satisfactory.



- The accuracy of the model can be further improved by appropriate hyperparameter tuning and optimization, thus improving the predictive ability of the model.

3. Good:

- This analysis highlights the importance of using big data analytics for salary forecasting, providing organizations with insight into talent management and compensation strategies.
- Employers can use predictive models to make informed decisions about: salary structure, recruiting and retention strategies.
- This project demonstrates the potential of the PySpark framework and MLlib to perform well in big data https://www.kaggle.com/code/ludovicocuoghi/pyspark-sql-queries-and-machine-learning/inputprocessing and predictive analytics.
- Although the prediction model performs satisfactorily, continuous improvement and optimization are essential to improve its accuracy and generalizability.
- Overall, this project helps deepen the understanding of employee salary forecasts using a data-driven approach, paving the way for a smarter and more balanced approach in organizations again.

VI. FUTURE WORKS

1. **Fine-tuning hyperparameters:**

   Optimize model performance by exploring various hyperparameters of the Gradient Boosting Tree (GBT) algorithm.

2. **Feature Engineering Enhancements:**

   Improve predictive ability by adding additional features or variables to capture relationships.

3. **Integrated method:**

   Consider integrated method (such as random forest) to improve the accuracy of the prediction.

4. **Data Quality Improvement:**

   Resolve data quality issues with cleansing and preprocessing technologies to ensure data integrity andreliability.

5. **Integrating external data:**

   Improve models by integrating important external data such as business indicators or business models.

6. **Alternative Research:**

   Evaluate the suitability of other machine learning algorithms other than GBT for salary prediction tasks.

REFERENCES

[1] Brynjolfsson, E. and McAfee, A. (2014). The Second Machine Age: Work, Development and Growth in the Age of Digital Technology. W. W. Norton Co.Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence, 14(771-780):1612, 1999.

[2] Brown, C. and Medoff, J. (1989). Large employer paid job. Journal of Political Economy, 97(5), 1027–1059.Philipp Probst, Marvin N. Wright, and Anne-Laure Boulesteix. Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(3), Jan 2019.

[3] Davenport, T.H. and Harris, J. (2007). Competitive Analysis: The New Science of Winning. Harvard Business Press.N. Antonio, A. de Almeida and L. Nunes, "Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, Mexico, 2017, pp. 1049-1054, doi: 10.1109/ICMLA.2017.00-11.

[4] Frey, C.B. and Osborne, M.A. (2017). The future of work: How will work be affected by computing? Technological Forecasting and Social Change, 114, 254–280.

[5] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). Introduction to statistical learning: Applications in R. Springer.G. Sekhon and S. Ahuja, "Review Machine Learning Models for Managing Hotel Cancellations in the Tourism Industry," 2023 3rd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2023, pp. 1-6, doi: 10.1109/CONIT59222.2023.10205827.

[6] https://www.kaggle.com/code/ludovicocuoghi/pyspark-sql-queries-and-machine-learning/input