PREDICTING SALARY OUTCOME IN THE FIELD OF DATA SCIENCES WITH EXTREME GRADIENT BOOSTING ALGORITHM

Submitted to Compile Thesis

in the Department of Informatics Engineering, Faculty of Computer Science, UNSRI

By:

Muhammad Andry Erpapalemlah

NIM : 09021381823094

INFORMATICS ENGINEERING DEPARTMENT

FACULTY OF COMPUTER SCIENCE

UNIVERSITAS SRIWIJAYA

2023

Thesis Approval Sheet

PREDICTING SALARY OUTCOME IN THE FIELD OF DATA SCIENCES WITH EXTREME GRADIENT BOOSTING ALGORITHM

A THESIS

By:

Muhamad Andry Erpapalemlah
NIM : 09021381823094

Approved by

Advisor I	Advisor II
Nopi Yusliani, S. Kom., M.T	M. Qurhanul Rizqie, S.Kom., M.T., Ph.D.
NIP. 198211082012122001	NIP. 198712032022031006

Certified by,

Head of the Department of Computer Engineering,

Alvi Syahrini Utami, M.Kom
NIP. 197812222006042003

TANDA LULUS UJIAN KOMPREHENSIF SKRIPSI

Pada hari Selasa tanggal 25 Juli 2023 telah dilaksanakan ujian komprehensif skripsi oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama : Muhammad Andry Erpapalemlah
NIM : 09021381823094
Judul : Predicting Salary Outcome in the Field Of Data Sciences with Extreme Gradient Boosting Algorithm

Dan dinyatakan **LULUS.**

1. Ketua Penguji

Yunita, M.Cs.
NIP. 198306062015042002 .....

2. Penguji

Osvari Arsalan, M.T
NIP. 198806282018031001 .....

3. Pembimbing I

Novi Yusliani, M.T.
NIP. 198211082012122001 .....

4. Pembimbing II

M. Qurhanul Rizqie, S.Kom., M.T., Ph.D.
NIP. 198712032022031006 .....

Mengetahui,
Ketua Jurusan Teknik Informatika

Alvi Syahrini Utami, M.Kom.
NIP.197812222006042003

PAGE STATEMENT

The following that has been signed by the author:

Name : Muhammad Andry Erpapalemlah

NIM : 09021381823094

Thesis Tile : Predicting Salary Outcome in the Field Of Data Sciences with Extreme Gradient Boosting Algorithm

Software Check Results (iThenticate/Turnitin) : 11%

I declare that this project report is my own original work and not a result of copying/plagiarism. If any elements of copying/plagiarism are found in this project report, I am willing to accept academic penalties from Universitas Sriwijaya in accordance with the applicable regulations.

I hereby make this statement truthfully and without any coercion from anyone.

Palembang, 09 August 2023

Muhammad Andry Erpapalemlah
NIM. 09021381823094

Motto :

- Embrace the unknown and forge your path.
- Dream big, start small, achieve greatness.
- Cultivate resilience, reap endless possibilities.
- Learn from yesterday, live for today, hope for tomorrow.

I dedicate this work to **:**

- **To my extended family**
- **To all my friends in UNSRI**
- **To my major Faculty of Computer Science**
- **To Sriwijaya Universisty**

PREDICTING SALARY OUTCOME IN THE FIELD OF DATA SCIENCES WITH EXTREME GRADIENT BOOSTING ALGORITHM

by

Muhammad Andry Erpapalemlah
NIM. 09021381823094

ABSTRACT

In this study, the crucial aspect of job marketing was addressed, specifically the need for accurate salary predictions based on job seekers' skills. It is vital to minimize disparities between applicants' actual abilities and their salary expectations to ensure fairness and transparency in the hiring process. To tackle this challenge, the development of a salary prediction system tailored to data science jobs within the data science domain was proposed. This system would provide employers with valuable insights into the appropriate compensation they could offer to potential employees, considering their skill levels and expertise. To achieve this goal, the Extreme Gradient Boosting Algorithm was implemented into the system, leveraging its powerful predictive capabilities. Employing this algorithm, was aimed to enhance the accuracy and reliability of the salary predictions, ultimately facilitating better decision-making for both job seekers and employers. The findings from the Scenarios conducted show that the metric evaluation of the system is highly promising. The impressive mean absolute errors (MAE) of 0.321, 0.316, and 0.325 for Scenario 1, Scenario 2, and Scenario 3, respectively, indicate that the model's predictions are remarkably close to the actual salary values. Additionally, the mean absolute percentage errors (MAPE) of 2.856%, 2.797%, and 2.871% further confirm the system's exceptional accuracy in predicting salary outcomes for data science jobs

Keywords: Salary prediction, Job marketing, Decision-making, Extreme Gradient Boosting Algorithm

Advisor I

Novi Yusliani, S.Kom., M.T
NIP. 198211082012122001

Advisor II

M. Qurhanul Rizqie, S.Kom., M.T., Ph.D.
NIP. 198712032022031006

Approved by,
Head of the Department of Computer Engineering,

Alvi Syahrini Utami, M.Kom
NIP. 197812222006042003

ACKNOWLEDGEMENT

All the praise be to Allah ‘Azza Wa Jalla, who has given us all salvation and blessing in life. This thesis entitled “Predicting Salary Outcome in the Field of Data Sciences with Extreme Gradient Boosting Algorithm” could be finished to fulfill the bachelor's requirement at the Informatics Engineering, Faculty of Computer Sciences, Sriwijaya University.

First of all, the writer would like to express gratitude to (Late) Dr. Jaidan Jauhari, M.T., the Dean of Faculty of Computer Sciences, Alvi Syahrini Utami, M.Kom., the Head of the Informatics Engineering Department.

The writer would also like to express his deepest thanks to all who had helped, supported, and suggested while writing this thesis. The writer would like to express his deepest gratitude and appreciation to his thesis advisors, Novi Yusliani, S. Kom., M.T. and M. Qurhanul Rizqie, S.Kom., M.T., Ph.D. who have been willing to sacrifice their valuable time to encourage and to guide him throughout the completion of this thesis.

The writer would like to thank to all his friends at the Informatics Engineering who have supported him during his study.

The writer realizes that this thesis is far from perfect. However, the writer hopes this thesis will be useful for further researchers in the field of data sciences and machine learning.

Palembang, June 2023
The writer

Muhammad Andry Erpapalemlah

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Introduction

This chapter contains the background, problem statement, research objectives, benefits of the research, research limitations, writing systematics, and conclusions of this study. This chapter will also provide a general explanation of all the activities that will be carried out in this research.

1.2 Background

Numerous intriguing and lucrative employment options are offered by data science. Data scientists are increasingly vital in gaining insights, making data-driven decisions, and resolving difficult issues as a result of the exponential rise of data and its growing importance across industries. Large datasets are gathered, cleaned, and analyzed by data scientists using their expertise in statistical analysis, machine learning, and programming to find interesting patterns and trends (Frisse & Misulis, 2019; Ghosh, 2019). Models, algorithms, and predictive analytics are created and used to speed up corporate growth, simplify processes, and improve decision-making. Data analysts, data engineers, machine learning engineers, and data scientists are all job titles related to data science. These professionals address challenges including customer behavior analysis, fraud detection, risk assessment, and personalized suggestions. They work in variety of industries, including marketing, technology, finance, and healthcare (Machado, 2020). With the increasing demand for skilled data science professionals, pursuing a career in this

field offers enormous potential for growth, innovation, and impact. As Today's world generates massive amounts of data, businesses rely on it to advance. Massive amounts of unstructured data stored in the cloud must be processed and prepared before they can be used by industries. In the previous decade, data-driven technology has transformed our daily lives and businesses. Data science seeks to mine data for hidden potential and value. It has become an essential component of businesses at all levels. The insatiable demand for data analysis ensures a place for data scientists in businesses (Bose, 2022).

The explanation in first paragraph shows that data sciences are commonly used in businesses. They are used to transform a vast amount of data into insightful analyses and predictions, turning seemingly meaningless data into knowledge (Deshmukh, 2021). One example of their uses is that businesses use them when hiring new employees or changing jobs where the application of data science enables the selection of workers whose profiles are more compatible with and fit for the goals of the business (Rego et al., 2022). Several methods that can be used in data sciences. One of them is Extreme Gradient. Boosting This method is very useful for businesses with large scale time-series data. According to (Gregory, 2018), an extremely accurate customer churn model can be produced using extreme gradient boosting on a customer dataset with a wide range of temporal features, for example, the one related to salary outcome of the new employees. Predicting salary outcomes using Extreme Gradient Boosting looks very interesting to investigate because it will help to get the balance between applicants' actual ability and their employment salary expectation (Zhang & Cheng, 2019). It is then

the reason to do this study entitled "**Predicting Salary Outcome in the Field of Data Sciences Using Extreme Gradient Boosting Method"**, where the algorithms of extreme gradient boosting method (XGBoost) is used to predict the salary outcome from a data set.

1.3 Problem Formulation

Based on the explanation in the background, the research questions of this study are as follows:

1. How can the Extreme Gradient Boosting (XGBoost) algorithm be utilized to predict salary outcomes in the field of data science?
2. How is the performance evaluation of the Extreme Gradient Boosting (XGBoost) algorithm in predicting salary outcomes from data science salaries?

1.4 Research Objectives

Based on the research questions formulated in problem formulation, the objectives of this study are:

1. To predict the salary outcome from data science salaries using Extreme Gradient Boosting (XGBoost) algorithm.
2. To determine how the performance evaluation of the Extreme Gradient Boosting (XGBoost) algorithm is in predicting salary outcomes from data science salaries.

1.5 Benefits of Research

The benefits of this study are:

1. The findings of the study provide valuable preliminary insights for predicting salary outcomes in the field of data science. These findings are critical for making informed decisions, benchmarking, identifying influential factors addressing disparities and biases, facilitating career planning, and guiding recruitment and retention strategies. These insights enable individuals and organizations to make strategic decisions about compensation by analyzing relevant data such as education, experience, skills, and industry trends.
2. The findings of this study are expected to become a reference in the field of applied computer science, particularly in the field of machine learning.

1.6 Problem Limitation

This study's limitations are as follows:

1. The data that is used in this study is public data from Kaggle Data Science Salaries (Year 2023). (https://www.kaggle.com/datasets/arnabchaki/data-science-salaries-2023) The data consists of the salary starting from the year 2020 until 2023 and the region varies from different countries.

2. The method that is used in this research is XGBoost regressor from the XGBoost library. it has shown promising results but may have limitations in generalization, model complexity, and interpretability.

1.7 Systematization of Writing

The systematics of writing this thesis follows the standard thesis writing applied by the Faculty of Computer Science, Universitas Sriwijaya which consists of 6 chapters where each chapter can be described as follows:

CHAPTER I. INTRODUCTION

This chapter contains the background, problem formulation, research objectives, research benefits, problem restrictions, writing systematics, and conclusions in this research.

CHAPTER II. LITERATURE REVIEW

In this chapter, the fundamentals of the theory used in the research are discussed, such as the definitions of the XGBoost algorithm, Data Science, Job Marketing, software development types, Metrics Evaluation MAE and MAPE. As well as previous relevant research.

CHAPTER III. RESEARCH METHODOLOGY

In this chapter, the stages of doing this research are described. Each stage of the research plan will be thoroughly described, with reference to a framework. Project management for the research will be included at the end of this chapter.

CHAPTER IV. SOFTWARE DEVELOPMENT

In this chapter, the Rational Unified Process (RUP) method is described. The development of software will begin with the inception phase, followed by the elaboration phase, construction phase, and transition phase.

CHAPTER V. RESEARCH RESULTS AND ANALYSIS

In this chapter, the findings from the research experiments are discussed and analysed.

CHAPTER VI. CONCLUSION AND RECOMMENDATIONS

In this chapter, the conclusion and suggestion based on the research findings and discussion are presented.

1.8 Conclusion

It is concluded that this research will involve the development of a prediction system for Data Science Salaries data using the XGBoost algorithm, to predict the salary outcome, which can help job seekers and employers in making informed decisions.

REFERENCES

Adarbah, H. Y., & Goode, M. (2022). Key demand factors in professional business courses: A mixed-methods study. *Journal of Business, Communication & Technology*, *1*(2), 44-53.

Anwar, A. (2014). A review of rup (rational unified process). *International Journal of Software Engineering (IJSE)*, *5*(2), 12-19.

Azmi, S. S., & Baliga, S. (2020). An overview of boosting decision tree algorithms utilizing AdaBoost and XGBoost boosting strategies. *Int. Res. J. Eng. Technol*, *7*(5).

Bose, P. (2022). *Future Scope of Data Science: Career, Jobs, and Skills*. https://www.analytixlabs.co.in/blog/scope-of-data-science/

De Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean absolute percentage error for regression models. *Neurocomputing*, *192*, 38-48.

Deshmukh, C. S. (2021). Role of Data Science in Reshaping the Business Sectors: Opportunities and Challenges for India. *Electronic Journal Of Social And Strategic Studies*, *2*, 406-434.

Frisse, M. E., & Misulis, K. E. (2019). Data Science In M. E. Frisse & K. E. Misulis (Eds.), *Essentials of Clinical Informatics* (pp. 219-223). Oxford University Press.

Ghosh, P. (2019). *Data Science 101*. Retrieved 17 May 2023 from https://www.dataversity.net/data-science-101/

Gregory, B. (2018). Predicting customer churn: Extreme gradient boosting with temporal data. *arXiv preprint arXiv:1802.03396*.

Hanif, I. (2020, 2-3 August 2019). Implementing extreme gradient boosting (xgboost) classifier to improve customer churn prediction. Proceedings of the 1st International Conference on Statistics and Analytics, ICSA 2019, Bogor.

Hayashi, C. (1998). What is data science? Fundamental concepts and a heuristic example. Data Science, Classification, and Related Methods: Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan, March 27–30, 1996,

Jabeur, S. B., Mefteh-Wali, S., & Viviani, J.-L. (2021). Forecasting gold price with the XGBoost algorithm and SHAP interaction values. *Annals of Operations Research*, 1-21.

Karunasingha, D. S. K. (2022). Root mean square error or mean absolute error? Use their ratio as well. *Information Sciences*, *585*, 609-629. https://doi.org/https://doi.org/10.1016/j.ins.2021.11.036

Kotlia, A. (1998). The concept of the labor market. *Problems of Economic Transition*, *41*(3), 53-65.

Li, X. (2023). A comparative study of statistical and machine learning models on near-real-time daily emissions prediction. *arXiv preprint arXiv:2302.01152*.

Lubis, A. R., Prayudani, S., Fatmi, Y., & Lubis, M. (2021). MAPE accuracy of CPO Forecasting by Applying Fuzzy Time Series. 2021 8th International

Conference on Electrical Engineering, Computer Science and Informatics (EECSI),

Machado, C. S. M. (2020). *System for fraud detection: customer segmentation and predictive analysis: via-verde Portugal*

Mo, H., Sun, H., Liu, J., & Wei, S. (2019). Developing window behavior models for residential buildings using XGBoost algorithm. *Energy and Buildings*, *205*, 1-19. https://doi.org/https://doi.org/10.1016/j.enbuild.2019.109564

Rego, J. M., Souza, H. P. D., Oliveira, J. M. L. D., & Costa, R. R. D. L. (2022). The use of Data Science in hiring functions or transition of positions by companies. *International Journal of Advanced Research (IJAR)*, *10*(11), 1158-1164. https://doi.org/http://dx.doi.org/10.21474/IJAR01/15778

Schneider, P., & Xhafa, F. (2022). *Anomaly Detection and Complex Event Processing Over IoT Data Streams: With Application to EHealth and Patient Data Monitoring*. Academic Press.

Willmott, C. J., Matsuura, K., & Robeson, S. M. (2009). Ambiguities inherent in sums-of-squares-based error statistics. *Atmospheric Environment*, *43*(3), 749-752.

Wu, J., Guo, X., Fang, M., & Zhang, J. (2022). Short term return prediction of cryptocurrency based on XGBoost algorithm. 2022 International Conference on Big Data, Information and Computer Network (BDICN),

Zhang, J., & Cheng, J. (2019). Study of Employment Salary Forecast using KNN Algorithm. 2019 International Conference on Modeling, Simulation and Big Data Analysis (MSBDA 2019),

Zhou, Z.-H. (2012). *Ensemble methods: foundations and algorithms*. CRC press.