information

*Article*

# Statistical Machine Learning Regression Models for Salary Prediction Featuring Economy Wide Activities and Occupations

**Yasser T. Matbouli [1,*] and Suliman M. Alghamdi [2]**

[1] Department of Industrial Engineering, Faculty of Engineering-Rabigh Branch, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[2] Advanced Resources for Information Technology Co., Riyadh 13524, Saudi Arabia
* Correspondence: ymatbouli@kau.edu.sa

**Abstract:** A holistic occupational and economy-wide framework for salary prediction is developed and tested using statistical machine learning (ML). Predictive models are developed based on occupational features and organizational characteristics. Five different supervised ML algorithms are trained using survey data from the Saudi Arabian labor market to estimate mean annual salary across economic activities and major occupational groups. In predicting the mean salary over economic activities, the Bayesian Gaussian process regression ML showed a marked improvement in $R^2$ over multiple linear regression (from 0.50 to 0.98). Moreover, lower error levels were obtained: root-mean-square error was reduced by 80% and mean absolute error was reduced by almost 90% compared to multiple linear regression. However, the salary prediction over major occupational groups resulted in artificial neural networks performing the best in terms of both $R^2$, with an improvement from 0.62 in multiple linear regression to 0.94 and errors were reduced by approximately 60%. The proposed framework can help estimate annual salary levels across different types of economic activities and organization sizes, as well as different occupations.

**Keywords:** machine learning regression; salary prediction; gaussian process regression; artificial neural networks; economic activities; occupational groups



**Citation:** Matbouli, Y.T.; Alghamdi, S.M. Statistical Machine Learning Regression Models for Salary Prediction Featuring Economy Wide Activities and Occupations. *Information* **2022**, *13*, 495. https://doi.org/10.3390/info13100495

Academic Editor: Agnes Vathy-Fogarassy

Received: 12 September 2022
Accepted: 9 October 2022
Published: 12 October 2022



## 1. Introduction

In this paper, we develop a holistic framework to predict labor salary for all job titles in the Saudi Arabian economy, across all economic activities and organizational sizes, by training limited survey data with statistical machine learning (ML). The performance of five ML algorithms is evaluated. Extensive salary surveys are needed across many occupations and sectors to estimate financial compensation for a given occupation, within a particular industry, of specific organization size.

The prediction framework presented in this paper is general and can be applied in any economy with standardized groups of occupations and economic activities. The full list of occupations is derived from the International Standard Classification of Occupations (ISCO-08) [1], which can be mapped to any national occupations classification. We obtained the list of all economic activities within an economy from the International Standard Industrial Classification of All Economic Activities (ISIC) [2].

We intend to predict spatial salary across all economic activities and occupations, and not temporal changes in salary. Thus, the mean annual salaries used in the prediction model were obtained from historical national surveys. Hence, salary estimates are predicted across economic activities for similar occupations, or for different occupations within the same economic sector. Nevertheless, the new approach allows more confounding variables to be considered when predicting salary. The systematic methodology enables fine-grained salary prediction with limited survey data. Independent variables in the prediction model are defined based on both theoretical and statistical rational covering both organizational characteristics and occupational features.

Salary is the most important element of labor compensation and is an essential part of an organization's strategy. As part of developing pay scales, organizations collect information about how much is paid for labor in similar organizations for similar occupations through the process of benchmarking. Data on market salary is used in benchmarking pay scales against competitors, or to avoid instances of inefficient pay. Likewise, job seekers can also benefit from information about pay scales for their prospective occupations. Labor can optimize their skills development by targeting highly paying sectors.

The new approach is formulated in a way that allows future improvement in the framework. The methodology presented in this paper is fresh and some concepts in defining job titles and economic activities were adopted from the standards of international organizations, which makes it easy to adapt the newly introduced framework universally across different international job markets. In the following sections, a review of the relevant background on how organizations develop their compensation policy is presented, together with a short survey of some predictive techniques. In the methodology section, a detailed rational of the prediction model is discussed, followed by the application of the prediction approach to the Saudi Arabian labor market to demonstrate applicability of the prediction and its performance.

## 2. Background

To select the intrinsic features that predict salary using ML, in this section, we explore how organizations develop their pay scales. We survey studies on predicting pay levels using multiple linear regression models and review the statistical ML regression models.

### *2.1. How Salary Scales Are Developed*

There are two main activities associated with the development of salary scales. First, understanding what the development of salary scale entails and how is it evaluated relatively within an organization. Second, appreciating what the data show on market pay for similar work, and how an organization's strategy determines how it benchmarks its pay scale with similar occupations in the labor market.

#### 2.1.1. Job Analysis and Evaluation

Jobs are analyzed in terms of the work involved in conducting the job. The analysis involves breaking down the job into the smallest units of work which together cover all tasks performed by the incumbent. Then, a job evaluation is performed to rank jobs within an organization for the development of pay scales [3]. There are four traditional job evaluation frameworks [4] which are considered useful [5]: the classification method [6], the ranking method [7], the point method [8], and the factor comparison method [4,9]. According to Heneman (2003) [5], work evaluation aims to pay workers based on technical and behavioral competencies, market pricing, and compensation bands based on organizational classification of common characteristics of jobs. However, competency based pay is not common, according to the same source.

#### 2.1.2. Salary Surveys and Benchmarking

For organizations to price labor within a single market, they seek labor compensation data of similar organizations in order to design compensation policies that are realistic while maintaining shareholder value maximization. Labor compensation data can be sourced from professional associations, industrialists, public databases, and third-party companies specialized in salary surveys [10]. The use of third party organizations to estimate market pricing of labor is widely adopted to avoid potential conflict of interest that may arise from the agency problem [11–14]. Compensation and benefits (C&B) policies are usually approved at the highest level of governance for private and public sector organizations. Thus, salary benchmarking is an essential exercise in designing and reviewing salary structures in public and private organizations. Salary benchmarking serves many purposes, for example, organizations that desire to remain competitive in retaining and attracting

talent perform salary benchmarking to adjust their salary scales. It also helps in identifying instances of overpayment or inefficient compensation policy. Nevertheless, obtaining data to perform salary surveys can be expensive and somewhat biased [15].

### *2.2. Salary Predictive Models*

Salary prediction models in literature are mostly concerned with the problem of unequal pay based on gender, race, or other biases that are not related to job content or job performance [16–20]. Statistical models such as multiple regression models are used to investigate instances where equal pay policies are violated or observed [21]. Hence, the purpose of this study is not equity analysis or bias detection, but to describe and predict salaries for individual job titles across all economic activities. Perhaps the first study that aimed to develop a holistic salary benchmarking is one by Meng et al. (2018) [22], in which a matrix factorization approach is applied to predict salaries for similar companies (i.e., similar business domain and company size) and similar jobs based on job descriptions. The work by Meng et al. [22], is particularly interesting because it used salary data from online recruitment websites and did not rely on private surveys. However, the grouping of companies based on similarity to deduce economic sectors is not comprehensive. Furthermore, job titles similarities and mapping based on job descriptions may not produce a consistent list of job titles in an entire economy. According to Pfeffer et al. (1987) [23], there are three main categories of confounding variables that predict salary levels. Firstly, organizational characteristics [23,24], such as size of the organization, or type of economic activity. Secondly, there are occupational characteristics, and these include features related to the job functions of the incumbent [25] as well as skills required to perform the work. Finally, there are economic consideration such as inflation and labor supply. To predict salary levels for specific occupations in any organization within a given year, characteristics of the occupation and characteristics of the organization are more relevant [23]. Earlier work on salary determinants focused on the characteristics of the occupation, such as specialization and skills level required for the job. As a result, it had been assumed that only individual attributes related to the incumbent, such as education or training level, determine the salary in accordance with the human capital approach [25]. It was later that the impact of organization characteristics, such as size and sector [23,24] were also recognized as determinants of salary levels.

### *2.3. Statistical Machine Learning (ML)*

ML approaches aim to teach computer systems to infer knowledge or predict quantitative output by examples [26]. Examples represents input-output based on set of data. A predictive model can be applied for classification, regression, or clustering. Classification and regression require supervised learning where the ML algorithms are given a set of data where input and output are known [27]. In classification, the output is a finite discrete set of variables while regression is for predicting continuous response variables. In this study, the objective is to predict salary values using ML regression.

This is perhaps the first study to apply ML to predict the price of labor holistically for an entire economy across major occupational groups. However, using ML for regression in financial forecasting is not new. Multiple surveys of ML have been applied in the field of finance [28,29]; more than 20 ML algorithms have been applied in classification, clustering, and forecasting of prices of financial assets [28,29]. For example, ref. [30] has applied support vector machine to predict stock price based on 69 features.

In this study, in addition to multiple linear regression (MLR) [27,31–33], we apply artificial neural networks (ANN) [34,35], tree-based regression (TBR) [27,36,37], support vector regression (SVR) [27,31,36,38,39], and Bayesian-based ML using the Gaussian process regression (GPR) [40,41]. It is currently more popular to use non-linear ML algorithms. However, MLR can be easier to interpret and applied in real life. Thus, standard linear algorithm is used as a baseline in the prediction model against which other non-linear approaches are

evaluated. Should improvements in goodness-of-fit and improvements in errors warrants using a non-linear approach, then it will plausible to use a non-linear approach.

## 3. Methodology

This section is divided into three parts. First, a holistic framework to develop a comprehensive set of all occupations and economic activities is introduced. Then, prediction models with selected features are developed. Finally, five ML prediction models are applied to real world salary data for Saudi labor market.

### *3.1. Sets of Economic Activities and Occupations*

As discussed earlier in Section 2.2, there are three categories of features that estimate a salary: first organizational features, second occupational characteristics, and the attributes of the incumbent. The most distinctive part of organizational features, is type of economic activity [23,24]. To capture the characteristics of all organizations in an economy for a prediction model, a finite list of all economic activities is adopted from the International Standard Industrial Classification of All Economic Activities, Revision 4 (ISIC4) [2], which provides a comprehensive and finite list of all types of economic activities. Likewise, a finite list of all occupations in an economy is compiled based on the International Standard Classification of Occupations (ISCO-08) [1]. Finally, the attributes of the incumbents [25] include educational level [25], and experience. However, other attributes such as immigration status and gender have also shown to impact salary levels. In the prediction model, we choose to ignore incumbent characteristics in favor of occupational features to avoid duplicating features. For example, complex work naturally requires higher educational or experience level. In contrast, simple work that requires low skilled labor does not pay highly skilled labor wage to a educated incumbent.

#### 3.1.1. Finite List of Economic Activities

ISIC4 [2] is an internationally adopted classification as part of standardized national accounts reporting, which is published by local and international statistical commissions. The first version of ISIC appeared in 1948 and the current fourth (ISIC4) revision was published in 2008 [2].

In the ISIC4 classification, there are four hierarchal categories: sections, divisions, groups, and classes. There are 21 sections as shown in Table 1. The smallest unit is class. The section with the largest number of groups and classes is manufacturing. Each category contains unique set of groups and classes without redundancies. The complete set has over 760 economic activities describing all possible industrial categories in any economy. Countries in national statistics may use a sub-set of ISIC4 as it fits their own diversity of economic activities. At the division level, as shown in Table 1, there are 88 types of economic activities. Sections in Table 1 are the broadest level in which there are 21 types of economic activities. In our prediction model, we use the list of economic activities as divisions.

#### 3.1.2. Finite List of Occupations in an Economy

Based on surveys, administrative records, and census data, the International Labor Office, which is part of the International Labor Organization (ILO), developed ISCO-08 [1]. ISCO-08 provides the most comprehensive list of occupations based on specialization and skills level as summarized in Table 2. ISCO-08 was formally adopted in 2007, published in 2012, and was last revised in 2018 [1]. It is still current and cover all occupations. Invention of new job titles is logically feasible; however, in our prediction model we assume that occupations in any given economy consist of a finite list based on ISCO-08 or the national occupation classification of a particular economy. Nevertheless, national occupation classification is more relevant to apply for any specific country. ILO recommends [1] that countries map their own national classification into ISCO-08.

At the most detailed level of ISCO-08 classification, there are 436 unit groups (see Table 2). However, there are only 10 major groups of occupations with a broad four skills level as summarized in Table 2. For a granular salary model prediction, it is suggested that unit groups are used to describe occupational features. However, for broader salary estimates, minor groups can also be featured in the prediction model to capture the occupational characteristics.

**Table 1.** Broad structure of ISIC4 classification of economic activities [2].

| Section | Description | Divisions | Groups | Classes |
|---|---|---|---|---|
| A | Agriculture, forestry and fishing | 3 | 13 | 38 |
| B | Mining and quarrying | 5 | 10 | 14 |
| C | Manufacturing | 24 | 69 | 137 |
| D | Electricity, gas, steam and air conditioning supply | 1 | 3 | 3 |
| E | Water supply; sewerage, waste management and remediation activities | 4 | 6 | 8 |
| F | Construction | 3 | 8 | 11 |
| G | Wholesale and retail trade; repair of motor vehicles and motorcycles | 3 | 20 | 43 |
| H | Transportation and storage | 5 | 11 | 20 |
| I | Accommodation and food service activities | 2 | 6 | 7 |
| J | Information and communication | 6 | 12 | 23 |
| K | Financial and insurance activities | 3 | 10 | 18 |
| L | Real estate activities | 1 | 2 | 2 |
| M | Professional, scientific and technical activities | 7 | 14 | 14 |
| N | Administrative and support service activities | 6 | 19 | 26 |
| O | Public administration and defense; compulsory social security | 1 | 3 | 7 |
| P | Education | 1 | 5 | 8 |
| Q | Human health and social work activities | 3 | 9 | 12 |
| R | Arts, entertainment and recreation | 4 | 4 | 10 |
| S | Other service activities | 3 | 5 | 17 |
| T | Activities of households as employers | 2 | 3 | 3 |
| U | Activities of extraterritorial organizations and bodies | 1 | 1 | 1 |

**Table 2.** Major occupation groups and skill Level as classified by ISCO-08 [1].

| Major Groups | Skill Level | Sub-Major Groups | Minor Groups | Unit Groups |
|---|---|---|---|---|
| 1 Managers | 3+4 | 4 | 11 | 31 |
| 2 Professionals | 4 | 6 | 27 | 92 |
| 3 Technicians and Associate Professionals | 3 | 5 | 20 | 84 |
| 4 Clerical Support Workers | 2 | 4 | 8 | 29 |
| 5 Services and Sales Workers | 2 | 4 | 13 | 40 |
| 6 Skilled Agricultural, Forestry, and Fishery Workers | 2 | 3 | 9 | 18 |
| 7 Craft and Related Trades Workers | 2 | 5 | 14 | 66 |
| 8 Plant and Machine Operators, and Assemblers | 2 | 3 | 14 | 40 |
| 9 Elementary Occupations | 1 | 6 | 11 | 33 |
| 0 Armed Forces Occupations | 1+2+4 | 3 | 3 | 3 |

### *3.2. Prediction Models*

The purpose of our prediction is to estimate average salaries across all economic activities and occupation group. Hence, our intention is not to predict future salary trends, but spatial salary levels across occupations and organizations. Thus, we examined several regression models using statistical ML for multiple linear regression: tree-based ML, support vector machine, and Bayesian ML based on GPR. In the following two sections, the performance of each regression model is given based on root-mean-square error (RMSE), R-squared ($R^2$), and mean absolute error (MAE). Let $Y$ be the mean annual salary for any occupation in any economic sector, $X$ represents the confounding variables that determine

the level of $Y$, $n$ is the number of observations in training data, and $v$ is the number of confounding variables predicting $Y$, then:

$$Y = f(X_{ij}) + \epsilon \tag{1}$$

where $i = \{1, 2, 3, \ldots, n\}$ is the $i^{th}$ observation, and $j = \{1, 2, 3, \ldots, v\}$ is the $v^{th}$ confounding variables. Note that $\epsilon$ represents random errors that is irreducible by improving $f(X_{ij})$. $X$ is a matrix of size $n \times v$.

In Table 3, the design of the salary survey of actual salary data is summarized. The mapping function to prepare the data for ML supervised training is given in the next column. The survey data needs two mapping functions: First is mapping incumbent job titles to occupational groups at unit level. This is facilitated by the index provided by ILO within its ISCO-08 publication, which maps real world job titles (i.e., commonly used in job markets names) with the classified occupational groups [1]. Second is mapping organizational characteristics to a division within ISIC4 [2] classification for economic activities. In addition, based on number of employees, the size of the organization is categorized to one of three levels: small, medium, and large. The total annual salary for the incumbent data is the output of all categorical variables and a single continuous variable that is mean or median salary per sector for a given year, or even for the entire economy for the same year.

**Table 3.** Salary survey for prediction model.

| # | Survey Data | | Model Mapping for Training Data | Type of Variable |
|---|---|---|---|---|
| 01 | Incumbent job title | $\Longrightarrow$ | ISCO-08 [1] Major occupational groups → sub-major groups → minor groups → unit groups | Discrete set of occupations |
| 02 | Demographic information about the incumbent | $\Longrightarrow$ | Such as gender, ethnic group, age, education level, experience, and etc. | Ignored, not included in the prediction model |
| 03 | Name of company/organization | $\Longrightarrow$ | ISIC4 [2] Economic Sections → Divisions | Discrete set of economic activities |
| 04 | Description of product/services provided | | | |
| 05 | Total number of employees | $\Longrightarrow$ | Size of the company | $\{small, medium, large\}$ |
| 06 | Total Annual Salary | $\Longrightarrow$ | Use total annual salary as output for training data, calculate salary mean, median, first and third quartiles as inputs | continuous variable |

### *3.3. Prediction Model for Saudi Labor Market Salary*

In this section the prediction model is implemented using observational data obtained solely from national statistics. An overview of Saudi Labor Market is given, followed by describing training data used in the prediction model. Finally, a discussion of results and insights obtained.

#### 3.3.1. Overview of Saudi Labor Market

The Saudi labor market is unique in its composition, with expatriates constituting about 75% of the labor market in 2020 [42,43] (note: labor market data exclude workers in police and armed forces). This feature is also exemplified in the size of remittance from the Saudi economy. According to the World Bank, in absolute dollars terms, remittance paid by labor from Saudi Arabia ranked only second to the United States in 2020 with more than $34 billion paid [42].

#### 3.3.2. Training Data

The Saudi national classifications of economic activities is based on ISIC4 [2] and the Saudi General Authority for Statistics has been publishing salary data across many variables such as age group, ethnicity, gender, size of the organization, and type of economic

activities [43]. The set of economic activities have been updated and expanded in 2010 to include 83 sectors (Section O, Division 84, which is "Public administration and defense; compulsory social security" is excluded from the Saudi survey data). The data from year 2000 to 2009 used a different classification of economic activities [43]. Therefore, we used data from 2010 to 2017 [43] for training. There are approximately 2000 rows of training data across variables such as category of economic activity, mean annual salary overall, size of organization, and the mean salary for each row as a response variable. Figure 1 shows the box-plot of mean annual salary across 83 economic sectors as classified in the national standards in accordance with ISIC4. It can be inferred from the plot that the average pay in extraction of crude petroleum and natural gas is the highest in the economy, which is not surprising given the fact that it is the main driver of national revenues for the Saudi economy.


**Figure 1.** Box plot of mean annual salaries across Saudi national economic activities (2010–2017) [43].

When it comes to occupational groups, publicly-available data show the average salary across major occupational groups only, i.e., ten major groups which exclude armed forces. The availability of data is limited to only two years (2018 and 2019, see Figure 2) with about 80 rows. Each row of data contains information about the average overall

salary, major occupation group, gender, nationality (a binary category of either Saudi or non-Saudi/expatriate) and the response variable being the mean salary for given categorical variables.

Having two different sizes for the training data sets presents an opportunity to empirically compare the performance of the five ML regression models used in this paper. Do some models perform better with large training data sets? Does performance suffer when only limited data is available? Moreover, we randomly selected a sample representing 15% of each data set to be excluded from the training data to use as test data for both models. The purpose is to test the models with data that the training algorithms have never seen. Thus, the consistency of the performance across validation and test data can be verified.

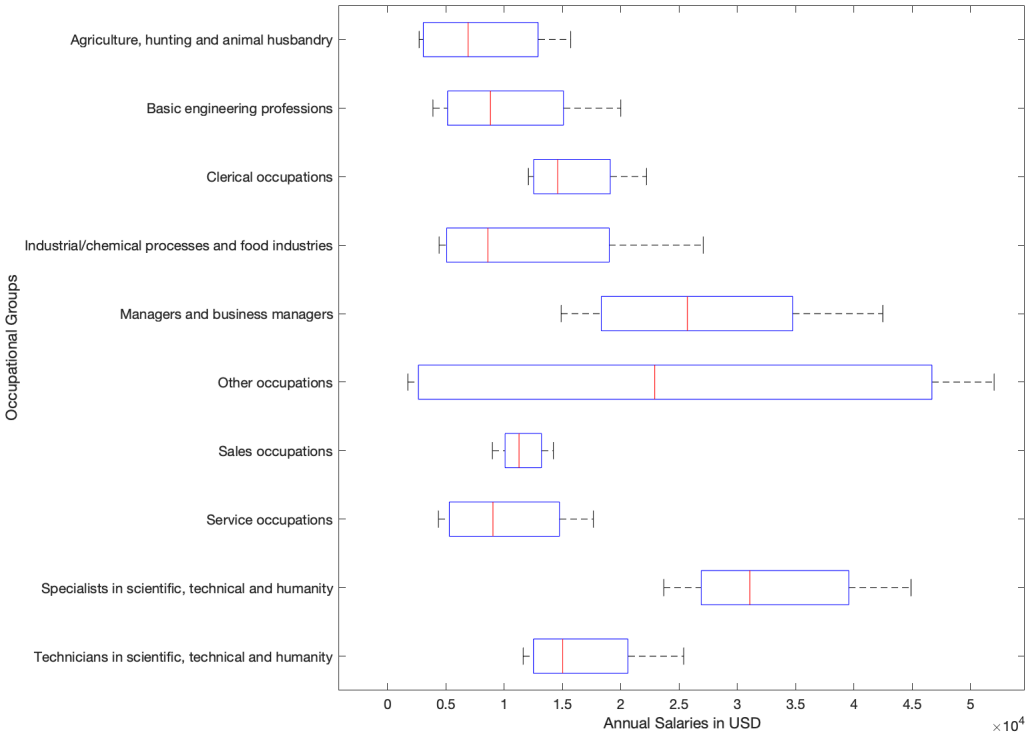


**Figure 2.** Box plot of mean annual salaries across major occupational groups (2018–2019).

#### 3.3.3. Mean Annual Salaries Across Economic Activities

In the prediction model to estimate means of annual salaries across economic activities, there is a single quantitative variable, which is the mean of annual salary across the economy in a given year $x_1$. Additionally, there are two categorical variables which are type of economic activity $x_2$, and size of the organization $x_3$. The set of categorical variables for the type of economic activity: $x_2$ contains 83 possible categories as given in the national classification according to ISIC4. The size of the organization is given in three levels: $x_3$ = small, organizations with less than 5 employees; medium, organizations with 5 to 19 employees; and large, organizations with more than 19 employees. These clusters of organization sizes is based on how the survey data is presented. Let $Y_A$ be the mean annual salary given organizational features, $x_{ij}$ represents the confounding variables that determine the level of $Y_A$, $n$ is the number of observations in training data, and $v = 3$ is the number of features predicting $Y_A$, then:

$$Y_A = f(x_{ij}) + \epsilon \tag{2}$$

where $i = \{1, 2, 3, \ldots, n\}$ is the $i^{th}$ observation, and $j = \{1, 2, 3\}$ is the $v^{th}$ confounding variables based on features in $A$. Also $\epsilon$ represents random errors that is irreducible by $f(x_{ij})$.

#### 3.3.4. Mean Annual Salaries across Occupations Groups

Similarly, the same regressions models were tested to estimate the mean annual salary across the ten major occupations groups. Likewise, there is a single quantitative variables, $x_4$, which the mean annual salary across occupational groups in a given year. Moreover, there is a single categorical variable, $x_5$, which is the set of major occupational groups according to ISCO-08. Let $Y_G$ be the mean annual salary given occupational features, $x_{ij}$ represents the confounding variables that determine the level of $Y_G$, $n$ is the number of observations in training data, and $v = 2$ is the number of features predicting $Y_G$, then:

$$Y_G = f(x_{ij}) + \epsilon \tag{3}$$

where $i = \{1, 2, 3, \ldots, n\}$ is the $i^{th}$ observation, and $j = \{4, 5\}$ is the $v^{th}$ confounding variables based on features in $G$. Also $\epsilon$ represents random errors that is irreducible by $f(x_{ij})$.

## 4. Results and Discussions

Several regression models based on statistical ML were implemented to predict mean annual salaries across economic activities, and the results, presented in Table 4, show a marked improvement in using non-linear ML over MLR. To fine-tune each regression model, we optimized hyperparameters by testing different numbers of layers and nodes in ANN, varying the number and size of leaves in TBR, and using different types of kernel functions in SVM and Bayesian-based regressions. Only best performing models are shown in Tables 4 and 5 along with the specific model hyperparameters. Best-performing Economic activity variables account for up to 98% of the variability in the response variable (see $R^2$ for Bayesian ML). RMSE and MAE were the least for the Gaussian process regression. In Figure 3, predicted versus actual response variables are plotted showing that the values are mostly consistent with a few data points with large differences between true and predicted values. Moreover, the residuals plot, as appears in Figure 4, the width of the band for residual values is mostly constant with a few exceptions. The improvements in the model are stable across all regression models because the performance of test data in the same Table 4 is shown to at least meet or exceed the performance of validation data. Note that test data represents about 15% of the original observations and were excluded at random from the training data.

**Table 4.** Performance of prediction models across economic activities.

| Regression Model | Validation Data | | | Test Data | | | Model Hyperparameters |
|---|---|---|---|---|---|---|---|
| | $R^2$ | RMSE | MAE | $R^2$ | RMSE | MAE | |
| MLR | 0.50 | 8706 | 3999 | 0.55 | 8799 | 4397 | Least squares method to estimate coefficients |
| ANN | 0.97 | 1964 | 627 | 0.99 | 1392 | 560 | Three layers with 10 nodes each |
| TBR | 0.97 | 2134 | 585 | 0.97 | 2088 | 578 | Minimum leaf size of 4 |
| SVM | 0.93 | 3209 | 641 | 0.99 | 1437 | 538 | Cubic kernel function |
| Bayesian-based GPR | 0.98 | 1882 | 481 | 0.99 | 1496 | 526 | Squared exponential isotropic kernel |

On the other hand, ANN performed the best when predicting mean annual salary across major occupational groups with an $R^2 = 0.94$, as shown in Table 5. However, Bayesian ML using the Gaussian process regression still performed well and even exceeding ANN model in performance for the test data (refer to Table 5). For the best performing ANN model, predicted versus actual salary values are presented in Figure 5. Nevertheless, the differences in the performance of nonlinear regression models are more significant here than the performance of the regression across economic activities. More specifically, TBR and SVM performed particularly poorly with the smaller training data set as illustrated in Table 5. Also, the difference in goodness-of-fit between validation data and test data is large for TBR, where $R^2$ dropped from 0.47 to a low 0.12. This may indicate that TBR

models require larger training data set. In addition, in Figure 6 the residuals plot for the ANN model shows an approximately constant deviation. In this case, despite having a smaller training data set, the results show plausible prediction model accuracy when using the bayesian based model or ANN.

**Table 5.** Performance of prediction models across occupations groups.

| Regression Model | Validation Data | | | Test Data | | | Model Hyperparameters |
|---|---|---|---|---|---|---|---|
| | $R^2$ | RMSE | MAE | $R^2$ | RMSE | MAE | |
| MLR | 0.62 | 7710 | 5056 | 0.62 | 6575 | 3493 | Least squares method to estimate coefficients |
| ANN | 0.94 | 3133 | 2036 | 0.91 | 3169 | 2311 | One layer of with 100 nodes |
| TBR | 0.47 | 9125 | 6203 | 0.12 | 9989 | 6439 | Minimum leaf size of 12 |
| SVM | 0.64 | 7442 | 4062 | 0.75 | 5382 | 3571 | Gaussian kernel function |
| Bayesian-based GPR | 0.87 | 4509 | 2299 | 0.91 | 3172 | 1927 | Rational quadratic kernel function |

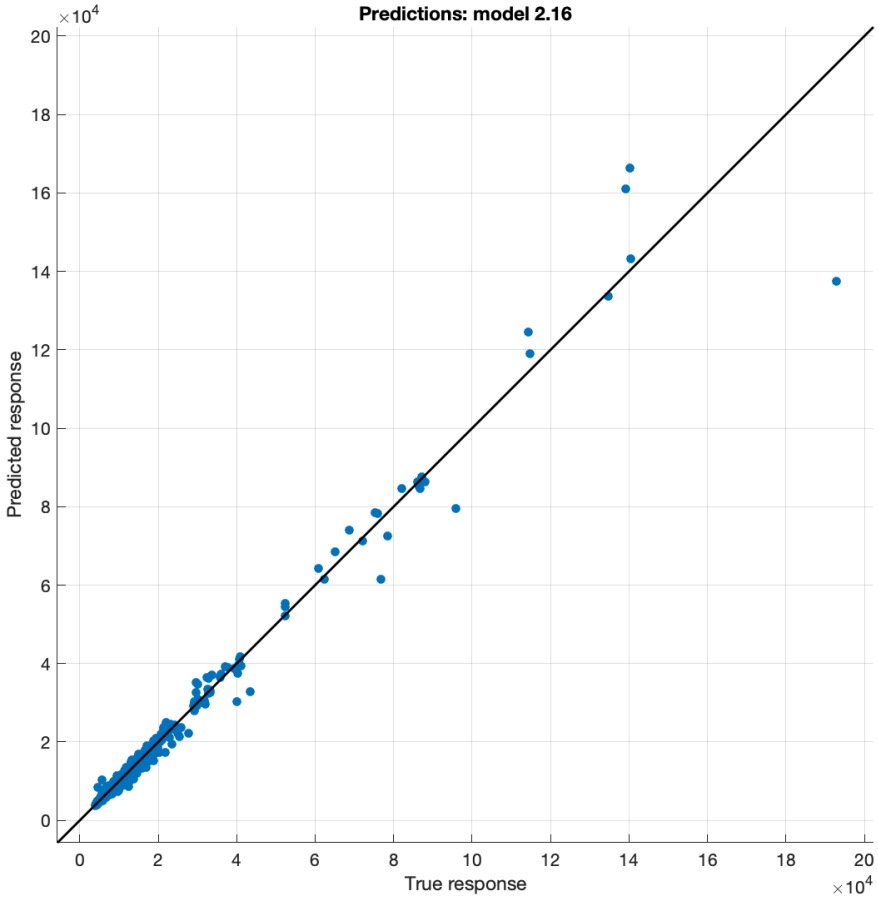


**Figure 3.** Predicted vs. actual annual mean salary across economic activities using Bayesian ML.

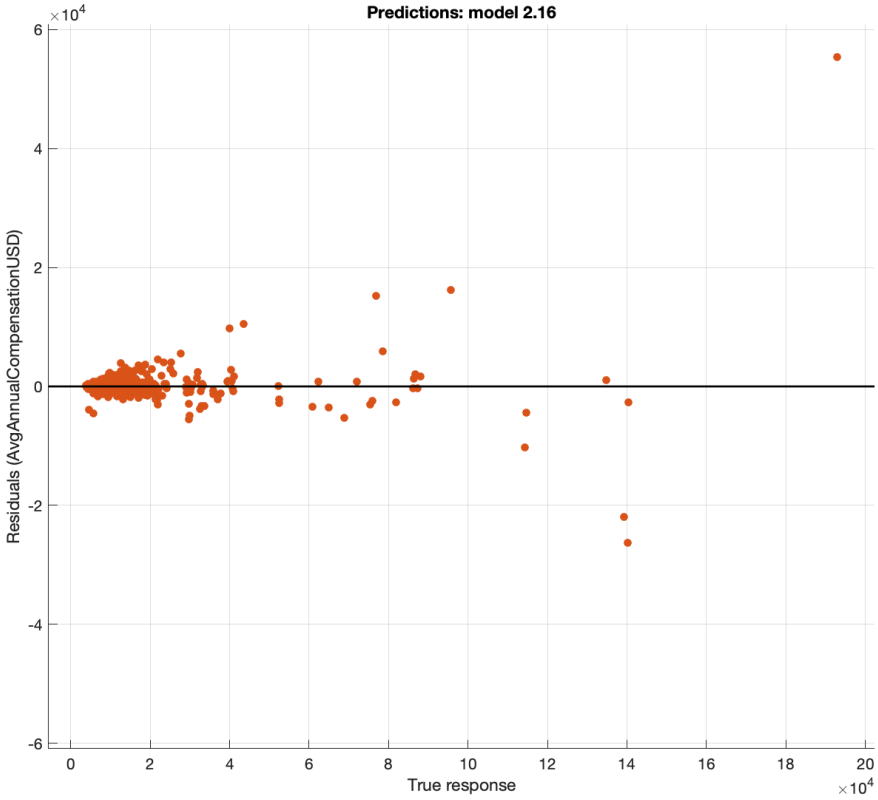


**Figure 4.** Residual errors plot for Bayesian ML regression across economic activities.

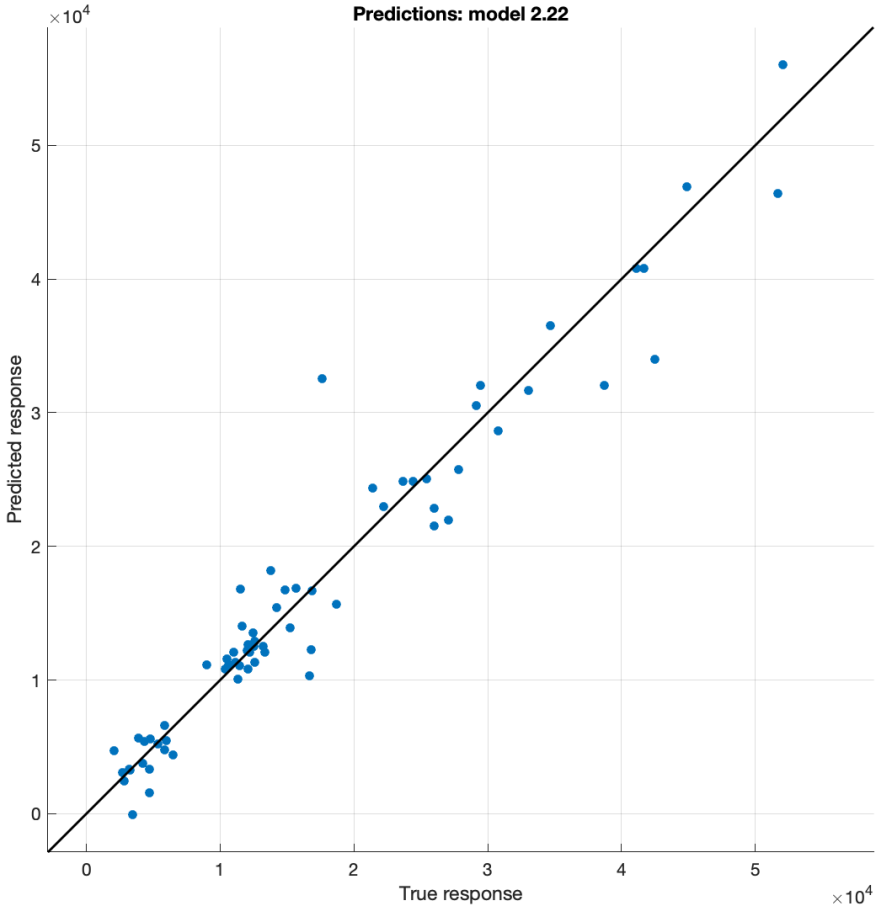


**Figure 5.** Predicted vs. actual annual mean salary across occupations groups using ANN.

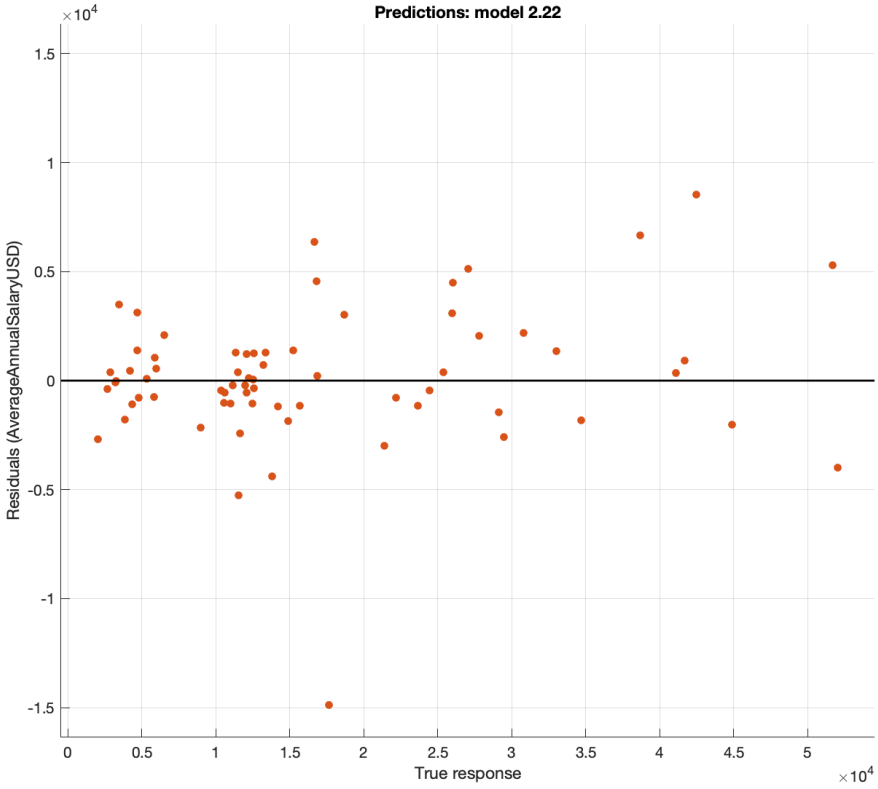


**Figure 6.** Residual errors plot for ANN regression across occupations groups.

## 5. Conclusions

In this paper, a framework to predict salary spatially across economic activities and occupations is developed, presented, and demonstrated using training data from the Saudi labor market. Our approach uses the international classification of economic activities and occupations as provided in ISIC4 and ISCO-08. Such use of international classifications makes this framework generalizable to other labor markets. Results suggests that MLR models do not provide the best fit; instead, the use of non-linear ML markedly improved the goodness of fit for the regression models. More specifically, the use of Bayesian ML by applying GPR performed with large and limited training data performed among the best. However, ANN also performed well when training data was limited.

Use of statistical ML, can both reduces the cost of salary benchmarking and improves accuracy especially when estimating salary levels for similar occupations in different industries, or when estimating different occupations within the same sector. The use of MLR models did not produce accurate predictions.

**Author Contributions:** Conceptualization, S.M.A.; methodology, Y.T.M. and S.M.A.; validation, S.M.A.; formal analysis, Y.T.M.; investigation, S.M.A.; resources, S.M.A.; data curation, S.M.A.; Software, Y.T.M.; writing—original draft preparation, Y.T.M.; writing—review and editing, S.M.A. and Y.T.M.; visualization, Y.T.M.; supervision, Y.T.M.; project administration, S.M.A. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** Public databases [42,43] were used and are referenced.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| GPR | Gaussian Process Regression |
| MAE | Mean absolute error |
| ML | Machine Learning |
| $R^2$ | Coefficient of determination |
| RMSE | Root-mean-square error |
| SVR | Support Vector Regression |

## References


1. International Labour Office. *International Standard Classification of Occupations (ISCO-08)*; International Labour Office: Geneva, Switzerland, 2012.
2. Department of Economic and Social Affairs, Statistics Division. *International Standard Industrial Classification of All Economic Activities (Revision 4)*; Department of Economic and Social Affairs, Statistics Division: New York, NY, USA, 2008.
3. Ingster, B. Job Analysis, Documentation, and Job Evaluation. In *The Compensation Handbook: A State of the Art Guide to Compensation Strategy and Design*; Berger, L.A., Berger, D.R., Eds.; McGraw Hill: New York, NY, USA, 2008.
4. Benge, E.J.; Burk, S.L.; Hay, E.N. *Manual of Job Evaluation: Procedures of Job Analysis and Appraisal*; Harper and Brothers: New York, NY, USA, 1941.
5. Heneman, R.L. Job and Work Evaluation: A Literature Review. *Public Pers. Manag.* **2003**, *32*, 47–71. [CrossRef]
6. Griffenhagen, E.O. *Classification and Compensation Plans as Tools in Personnel Administration: Notes on Definitions, Principles, and Developments*; Reprinted in Office Management Series; Office Management: New York, NY, USA, 1926.
7. Armstrong, M.; Cummins, A.; Hastings, S.; Wood, W. *Job Evaluation A Guide to Achieving Equal Pay*; Kogan Page Limited: London, UK, 2005.
8. Lott, M.R. *Wage Scales and Job Evaluation*; Ronald Press Company: New York, NY, USA, 1926.
9. Hay, E.; Purves, D. A New Method of Job Evaluation: The Guide–Profile Method. *Pers. Am. Manag. Assoc.* **1954**, *31*, 72–80.
10. York, D.; Brown, T. Salary Surveys. In *The Compensation Handbook: A State of the Art Guide to Compensation Strategy and Design*; Berger, L.A., Berger, D.R., Eds.; McGraw Hill: New York, NY, USA, 2008.
11. Dyl, E.A. Corporate control and management compensation: Evidence on the agency problem. *Manag. Decis. Econ.* **1988**, *9*, 21–25. [CrossRef]
12. Roth, K.; O'Donnell, S. Foreign Subsidiary Compensation Strategy An Agency Theory Perspective. *Acad. Manag. J.* **1996**, *39*, 678–703.
13. Bebchuk, L.A.; Fried, J. *Executive Compensation as an Agency Problem*; Technical report; National Bureau of Economic Research: Cambridge, MA, USA, 2003.
14. Chen, C.X.; Lu, H.; Sougiannis, T. The Agency Problem, Corporate Governance, and the Asymmetrical Behavior of Selling, General, and Administrative Costs. *Contemp. Account. Res.* **2012**, *29*, 252–282. [CrossRef]
15. Fitzpatrick, I.; McMullen, T.D. Benchmarking. In *The Compensation Handbook: A State of the Art Guide to Compensation Strategy and Design*; Berger, L.A., Berger, D.R., Eds.; McGraw Hill: New York, NY, USA, 2008.
16. Wiler, J.L.; Rounds, K.; McGowan, B.; Baird, J.; Choo, E.K. Continuation of Gender Disparities in Pay Among Academic Emergency Medicine Physicians. *Acad. Emerg. Med.* **2019**, *26*, 286–292. [CrossRef] [PubMed]
17. Gutiérrez-Martínez, I.; Saifuddin, S.M.; Haq, R. *The United Nations Gender Inequality Index*; Edward Elgar Publishing: Cheltenham, UK, 2021.
18. Bloyer, K. Gender Inequity in the Tech Industry Workplace. In *Proceedings of the Perspectives on Critical Issues*; Saint Mary's University: Halifax, NS, Canada, 2021; Volume 4, pp. 215–224.
19. Acheson, J.; Collins, M. The gender pay gap in Revenue. *Administration* **2021**, *69*, 45–75. [CrossRef]
20. Singh, P.; Pattanaik, F. Unequal Reward for Equal Work? Understanding Women's Work and Wage Discrimination in India Through the Meniscus of Social Hierarchy. *Contemp. Voice Dalit* **2020**, *12*, 19–36. [CrossRef]
21. Johnson, C.B.; Riggs, M.L.; Downey, R.G. Fun with Numbers: Alternative Models for Predicting Salary Levels. *Res. High. Educ.* **1987**, *27*, 349–362. [CrossRef]
22. Meng, Q.; Zhu, H.; Xiao, K.; Xiong, H. Intelligent Salary Benchmarking for Talent Recruitment: A Holistic Matrix Factorization Approach. In Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM), Singapore, 17–20 November 2018; pp. 337–346. [CrossRef]
23. Pfeffer, J.; Davis-Blake, A. Understanding Organizational Wage Structures: A Resource Dependence Approach. *Acad. Manag. J.* **1987**, *30*, 437–455.
24. Carroll, G.R.; Mayer, K.U. Job-Shift Patterns in the Federal Republic of Germany: The Effects of Social Class, Industrial Sector, and Organizational Size. *Am. Sociol. Rev.* **1986**, *51*, 323–341. [CrossRef]
25. Mincer, J. The Distribution of Labor Incomes: A Survey With Special Reference to the Human Capital Approach. *J. Econ. Lit.* **1970**, *8*, 1–26.
26. Kubat, M. *An Introduction to Machine Learning*, 2nd ed.; Springer: New York, NY, YSA, 2017.

27. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning with applications in R*, 2nd ed.; Springer: New York, NY, USA, 2021.
28. Krollner, B.; Vanstone, B.; Finnie, G. Financial time series forecasting with machine learning techniques: A survey. In Proceedings of the 18th European Symposium on Artificial Neural Networks (ESANN 2010), Bruges, Belgium, 28–30 April 2010; pp. 25–30.
29. Obthong, M.; Tantisantiwong, N.; Jeamwatthanachai, W.; Wills, G. A survey on machine learning for stock price prediction: Algorithms and techniques. In Proceedings of the 2nd International Conference on Finance, Economics, Management and IT Business, Online, 5–6 May 2020; pp. 63–71.
30. Leung, C.K.S.; MacKinnon, R.K.; Wang, Y. A Machine Learning Approach for Stock Price Prediction. In Proceedings of the 18th International Database Engineering & Applications Symposium, Porto, Portugal, 7–9 July 2014; Association for Computing Machinery: New York, NY, USA, 2014; pp. 274–277. [CrossRef]
31. Kavitha, S.; Varuna, S.; Ramya, R. A comparative analysis on linear regression and support vector regression. In Proceedings of the 2016 Online International Conference on Green Engineering and Technologies (IC-GET), Coimbatore, India, 19 November 2016; pp. 1–5. [CrossRef]
32. Shen, R.; Zhang, B.-W. The research of regression model in machine learning field. *MATEC Web Conf.* **2018**, *176*, 01033. [CrossRef]
33. Maulud, D.; Abdulazeez, A.M. A review on linear regression comprehensive in machine learning. *J. Appl. Sci. Technol. Trends* **2020**, *1*, 140–147. [CrossRef]
34. Specht, D.F. A general regression neural network. *IEEE Trans. Neural Netw.* **1991**, *2*, 568–576. [CrossRef] [PubMed]
35. Nghiep, N.; Al, C. Predicting housing value: A comparison of multiple regression analysis and artificial neural networks. *J. Real Estate Res.* **2001**, *22*, 313–336. [CrossRef]
36. Mitchell, T. *Machine Learning*; McGraw-Hill: New York, NY, USA, 1997.
37. Hastie, T.; Tibshirani, R.; Friedman, J.H.; Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, NY, USA, 2009.
38. Drucker, H.; Burges, C.J.C.; Kaufman, L.; Smola, A.; Vapnik, V. Support Vector Regression Machines. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 2–5 December 1996; Mozer, M., Jordan, M., Petsche, T., Eds.; MIT Press: Cambridge, MA, USA, 1996; Volume 9.
39. Awad, M.; Khanna, R. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*; Apress: Berkeley, CA, USA, 2015; pp. 67–80. [CrossRef]
40. Williams, C.K.; Rasmussen, C.E. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006.
41. Rasmussen, C.E.; Nickisch, H. Gaussian processes for machine learning (GPML) toolbox. *J. Mach. Learn. Res.* **2010**, *11*, 3011–3015.
42. The World Bank. *Personal Remittances Paid in USD*; The World Bank: Washington, DC, USA, 2020.
43. General Authority for Statistics Saudi Arabia (GaStat). *Annual Economic Surveys, Business Establishments (2003–2017)*; General Authority for Statistics Saudi Arabia (GaStat): Riyadh, Saudi Arabia, 2020.