**MBA 442B: MACHINE LEARNING ALGORITHMS – I**

**CIA-1**

**DOMAIN-SPECIFIC MODEL BUILDING**

**SUBMITTED BY**

**MERLIN SABU**

**(2327331)**

**SUBMITTED**

**TO**

**PROF. HELEN JOSEPHINE V.L**



**PROJECT EXECUTION**

1. **BUSINESS UNDERSTANDING**
2. **Problem Identification**

The human resources (HR) department of a company faces challenges in determining fair and competitive salaries for employees. The current salary structure may not fully reflect the diverse factors that influence employee compensation. This can lead to issues such as employee dissatisfaction, high turnover rates, and difficulties in attracting top talent. By predicting monthly income based on various factors, the company aims to ensure a fair and competitive salary structure that prevents employee retention and enhance motivation.

1. **Variables**

|  |  |
| --- | --- |
| **Variables** | **Description** |
| **Department** | Gives the department information an employee is working |
| **Education Field** | Describes the field of education an employee came from. |
| **Job Role** | Specifies the job role in the company. |
| **Education** | Give the specialized level of education. |
| **Hourly Rate** | Rate of work per hour. |
| **Over Time** | Whether the employee worked overtime or not. |
| **Job Involvement** | Rate of involvement (out of 4) in his/her job role. |
| **Job Level** | Level of job (out of 5) in the organization. |
| **Job Satisfaction** | Rate of satisfaction level (out of 4) of the current job which can impact their productivity. |
| **Num Companies Worked** | Number of companies the employee worked before. |
| **Percent Salary Hike (%)** | Gives the percentage of salary hike. |
| **Performance Rating** | Describes the rate of performance in their work out of 4. |
| **Total Working Years** | Gives the total number of years the employee had experience. |
| **Years At Company** | Gives the years at the current company. |
| **Years In Current Role** | Give the years he/she was into current role. |
| **Monthly Income** | Specifies the monthly income employees are got. |

1. **Objectives**

The primary objective is to develop a predictive model for monthly income by understanding the the key factors influencing employee compensation to ensure equitable and competitive salary structures.

Outcomes: -

* Able to identify key determinants
* Predict Monthly Income
* Ensure Fair Compensation and increase in employee satisfaction
* Improve Employee Retention

1. **DATA UNDERSTANDING**
2. **Data collection**

Data collection was done mainly by focusing on the HR domain of interest. The source was an online learning platform which offers courses to analyze data.

1. **Data exploration**

* Categorical variables – Department, Education Field, Job Role, Education
* Continuous variables - Hourly Rate, Over Time, Job Involvement, Job Level, Job Satisfaction, Number of Companies Worked, Percent Salary Hike (%), Performance Rating, Total Working Years, Years at Company, Years in Current Role, Monthly Income.

Here for addressing the business problem, **Monthly income is considered as dependent variable** and rest all independent variables.

1. **Assessing data quality –** checked for missing values, duplicates and data inconsistencies.
2. **DATA PREPERATION**
3. **Data cleaning**

* Missing value analysis – there were no missing or blank data points to treat and therefore there was no need of data imputation to do.
* Variable standardization – there were some categorical variables with (Yes/No) characteristic for the variable Over Time. And it was converted to binary numbers as 1 and 0.
* Feature selection – it was done very carefully as the original dataset contained more than thirty variables. Out of that in order to maintain the accuracy and relevancy of certain factors/variables and dependent variable, some variables are deleted and a total of 16 variables are identified for the project execution.

1. **MODELLING**

**MULTIPLE LINEAR REGRESSION**

fullmodel <- lm(Monthly.Income ~ ., data = training)

summary(fullmodel)

Output summary interpretation: - here a multiple linear regression with monthly income as the response variable is done.

* **Residuals**

Min: -3975.1

1Q (First Quartile): -627.1

Median: -3.3

3Q (Third Quartile): 651.3

Max: 4365.7

These values provide a summary of the residuals (the differences between the observed and predicted values). Ideally, residuals should be symmetrically distributed around zero, indicating a good fit.

* **Intercept** value indicates the expected Monthly Income when all independent variables are zero.
* **Very strong significant variables** observed are – Job Role (Manager, Research Director), Job level and Total working years

**Moderate significant –** Job Role(laboratory technician, research scientist, Education (Bachelors and Master’s degree).

* **Multiple R-squared**: 0.9498 - 94.98% of the variability in Monthly Income is explained by the model.
* **Adjusted R-squared**: 0.9471 - Adjusted R-squared of 0.9471 indicates that the model is very good at predicting Monthly Income.
* **F-statistic**: 355.5 on 30 and 564 DF (p-value < 2.2e-16) - A high F-statistic and a p-value < 2.2e-16 indicate that the model is statistically significant.

The model indicates that factors like job role, job level, and total working years significantly impact monthly income.

**Step -wise regression**

Step-wise regression is performed to get a low AIC value with much more relevant as well as dimensionless data in order get a good fit model.

* step\_model <- stepAIC(fullmodel, direction = "both")

summary(step\_model)

Step: **AIC=8316.25** - Lower AIC values indicate a better model fit.

So here from final step AIC, the summary output interpretations are as follows:

* The final reduced AIC value of 8316.25 gives a summary as:

**Model Fit Statistics**

Residual standard error: 1071 on 580 degrees of freedom

Multiple R-squared: **0.9488 –** 94% of the variability in monthly income is explained by the significant variables.

Adjusted R-squared: **0.9475**

F-statistic: 767.2 on 14 and 580 DF, p-value: **< 2.2e-16.**

The model explains about **94.75%** of the variability in Monthly Income (Adjusted R-squared = 0.9475), which indicates a very good fit.

**Significant Predictors:**

* **Job Role**: Positions such as Manager **(coeff.** 4185.500, **p-value** - < 2e-16) and Research Director (**coeff.** 3780.891, **p-value -** < 2e-16) significantly increase Monthly Income, while roles like Laboratory Technician and Research Scientist have a negative impact.
* **Job Level**: Higher job levels are strongly associated with increased Monthly Income (**coeff.** 2873.959, **p- value** - < 2e-16)
* **Total Working Years**: Each additional year of working experience is associated with an increase in Monthly Income.
* **Percent Salary Hike**: Marginally significant, indicating a positive relationship with Monthly Income.
* Also, the DW statistic of 1.9102 and the p-value of 0.1356 suggest that there is no significant autocorrelation in the residuals of regression model. This is a good indication that model's residuals are not correlated over time, which is an assumption of the classical linear regression model.
  + print(predictions)

488 795 66 711 173

2274.601 ,6065.916, 16487.732, 15266.548, 2708.560

* These values are predicted ones for the first five rows of employees selected is in the range from **$2,274.60 to $16,487.73**. This indicates a significant variation in the predicted salaries for these employees.

**RIDGE REGRESSION**

> # Find the best lambda via cross-validation

> ridge\_reg1 <- cv.glmnet(X\_train, Y\_train, alpha = 0)

> bestlam <- ridge\_reg1$lambda.min

> print(bestlam)

[1] **448.5908 -** The best lambda value found through cross-validation is approximately 448.59. This is the optimal regularization parameter for Ridge regression model, balancing the trade-off between bias and variance to minimize prediction error on unseen data. It minimizes the cross-validated error. It is the best compromise between fitting the training data well and avoiding overfitting, thus ensuring better generalization to new data.

mse <- mean((Y\_cv - ridge.pred)^2)

> print(paste("Mean Squared Error:", mse))

[1] "**Mean Squared Error**: 1533067.59827185 - This is the average squared difference between the actual values and the predicted values in the cross-validation set.

The R² value of approximately **0.931 means that 93.1%** of the variance in the Monthly Income is explained by the Ridge regression model. Despite the high R², the high MSE indicates that the actual prediction errors are large. This suggests that while the model explains the variance well, its predictions may still be off by significant amounts in absolute terms.

**LASSO REGRESSION**

* The best lambda value found through cross-validation is **62.12473**. This value represents the degree of regularization applied in the Lasso regression model. A higher lambda value would impose more regularization, potentially leading to more features being shrunk to zero, while a lower lambda would impose less regularization.
* The **MSE** of the Lasso regression model on the validation set is 1,330,644.00.
* The **R² value** for the Lasso regression model is **0.9401**. This indicates that approximately **94.01%** of the variance in Monthly Income is explained by the Lasso regression model. This is a very high proportion, suggesting that the model fits the data well.

**Comparison**

* **MSE**: Lasso regression has a lower MSE (1,330,644.00) compared to Ridge regression (1,533,067.60), suggesting better predictive accuracy on the validation set.
* **R² Value**: Lasso regression has a higher R² value (0.9401) compared to Ridge regression (0.93096), indicating that Lasso regression explains a larger portion of the variance in Monthly Income.
* **Model Performance**: The Lasso regression model has performed better than the Ridge regression model in terms of both MSE and R², suggesting that it may be more effective in capturing the relationships between the predictors and the target variable in this dataset.
* **Feature Selection**: Lasso regression performs feature selection by shrinking some coefficients to zero, which can result in a more interpretable model with fewer predictors. This aspect is beneficial when dealing with high-dimensional data.

**BUSINESS IMPLICATIONS**

* Salary Policies - Insights from the model can inform salary policies, ensuring competitive compensation for key roles and levels to retain talent.
* Career Development - The importance of job level and total working years highlights the need for clear career progression paths and professional development opportunities.
* Performance and Retention - Understanding the impact of job roles and performance ratings on salary can help in designing better performance evaluation and reward systems.

**Democratizing the Solution**

* Training - Providing training sessions for HR staff to understand how to use the model effectively and interpret its outputs.
* Transparency - Ensure the model's workings and decision rules are transparent to build trust and facilitate acceptance among employees and stakeholders.
* Integration - Integrate the model into existing HR systems for seamless use in daily operations, such as salary reviews and performance evaluations.