Salary Prediction Analysis

Venkata Rao Guduru

06/29/2022

## Summary

The paper will discuss the implementation of a logistic regression model that is essential in predicting employees’ salaries after a certain period. In addition, this paper has identified the various factors that help determine salary growth after a set period. With the project, students are provided with assistance related to the salary they may expect when they complete their studies. The project also provides insights into the talents they need to achieve to ensure they can achieve their professional goals. Insights into the association between family background and salary have also been provided in this project. For the salary analysis and prediction, the project has involved several stages. The first stage includes downloading data from the UCI machine learning website, which has been imported into the R environment. The second stage includes data processing and cleaning. The third stage includes performing exploratory data analysis and visualization. The fourth stage includes building the logistic model, and the last stage involves predicting and identifying the factors considered for salary allocation. The results of this project answer the research question, does salary allocation depend on years of experience or the talent of individuals. It is concluded that salary is more dependent on the type of employer.

## Introduction

Data mining has become one of the most popular trends recently, especially in identifying patterns and knowledge from data. This paper implements data mining techniques to identify patterns and insights regarding salary allocation and make predictions from the identified patterns by developing a logistic regression model. This project intends to provide students with knowledge concerning different courses depending on salary allocation as predicted. The logistic regression model has been created in this project for predictions. This paper is divided into the literature review, theory, data, methodology, results, implications, and conclusion sections.

## Literature Review

In a study by Pawha & Kamthania (2019), the authors proposed a salary prediction system that college students could use to enhance the student’s motivation. Through the use of the Decision Tree technique, these authors were able to generate a seven-feature prediction model. In another study, Bhuller et al. (2017) implemented an ordinary least squares regression model to predict the salaries of students based on family background and profiles. In another study by Shwartz et al.(2016), the authors created a hierarchical linear regression model where the salary was used as the output variable and profile as a fixed parameter. However, in their system, two problems were identified. The first problem is one needed a lot of statistical knowledge to understand the predictions, and the second problem is that the system was only personalized only for use for the students group. Therefore, with these problems, this project has implemented a dynamic prediction model that can be used by any group, not just the student group, and made sure that the results of the predictions are understandable by anyone who does not have any statistical knowledge.

## Theory

H1: There is no dependency on the type of employer and salary.

## Data

The dataset has been obtained from the UCI machine learning site(<https://archive.ics.uci.edu/ml/machine-learning-databases/adult/>).

## X age type\_employer fnlwgt education education\_num marital  
## 1 1 39 State-gov 77516 Bachelors 13 Never-married  
## 2 2 50 Self-emp-not-inc 83311 Bachelors 13 Married-civ-spouse  
## 3 3 38 Private 215646 HS-grad 9 Divorced  
## 4 4 53 Private 234721 11th 7 Married-civ-spouse  
## 5 5 28 Private 338409 Bachelors 13 Married-civ-spouse  
## 6 6 37 Private 284582 Masters 14 Married-civ-spouse  
## occupation relationship race sex capital\_gain capital\_loss  
## 1 Adm-clerical Not-in-family White Male 2174 0  
## 2 Exec-managerial Husband White Male 0 0  
## 3 Handlers-cleaners Not-in-family White Male 0 0  
## 4 Handlers-cleaners Husband Black Male 0 0  
## 5 Prof-specialty Wife Black Female 0 0  
## 6 Exec-managerial Wife White Female 0 0  
## hr\_per\_week native.country income  
## 1 40 United-States <=50K  
## 2 13 United-States <=50K  
## 3 40 United-States <=50K  
## 4 40 United-States <=50K  
## 5 40 Cuba <=50K  
## 6 40 United-States <=50K

Get structure of the data.

str(sal\_emp)

## 'data.frame': 32561 obs. of 16 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ age : int 39 50 38 53 28 37 49 52 31 42 ...  
## $ type\_employer : chr " State-gov" " Self-emp-not-inc" " Private" " Private" ...  
## $ fnlwgt : int 77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...  
## $ education : chr " Bachelors" " Bachelors" " HS-grad" " 11th" ...  
## $ education\_num : int 13 13 9 7 13 14 5 9 14 13 ...  
## $ marital : chr " Never-married" " Married-civ-spouse" " Divorced" " Married-civ-spouse" ...  
## $ occupation : chr " Adm-clerical" " Exec-managerial" " Handlers-cleaners" " Handlers-cleaners" ...  
## $ relationship : chr " Not-in-family" " Husband" " Not-in-family" " Husband" ...  
## $ race : chr " White" " White" " White" " Black" ...  
## $ sex : chr " Male" " Male" " Male" " Male" ...  
## $ capital\_gain : int 2174 0 0 0 0 0 0 0 14084 5178 ...  
## $ capital\_loss : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ hr\_per\_week : int 40 13 40 40 40 40 16 45 50 40 ...  
## $ native.country: chr " United-States" " United-States" " United-States" " United-States" ...  
## $ income : chr " <=50K" " <=50K" " <=50K" " <=50K" ...

Get summary of the dataset.

summary(sal\_emp)

## X age type\_employer fnlwgt   
## Min. : 1 Min. :17.00 Length:32561 Min. : 12285   
## 1st Qu.: 8141 1st Qu.:28.00 Class :character 1st Qu.: 117827   
## Median :16281 Median :37.00 Mode :character Median : 178356   
## Mean :16281 Mean :38.58 Mean : 189778   
## 3rd Qu.:24421 3rd Qu.:48.00 3rd Qu.: 237051   
## Max. :32561 Max. :90.00 Max. :1484705   
## education education\_num marital occupation   
## Length:32561 Min. : 1.00 Length:32561 Length:32561   
## Class :character 1st Qu.: 9.00 Class :character Class :character   
## Mode :character Median :10.00 Mode :character Mode :character   
## Mean :10.08   
## 3rd Qu.:12.00   
## Max. :16.00   
## relationship race sex capital\_gain   
## Length:32561 Length:32561 Length:32561 Min. : 0   
## Class :character Class :character Class :character 1st Qu.: 0   
## Mode :character Mode :character Mode :character Median : 0   
## Mean : 1078   
## 3rd Qu.: 0   
## Max. :99999   
## capital\_loss hr\_per\_week native.country income   
## Min. : 0.0 Min. : 1.00 Length:32561 Length:32561   
## 1st Qu.: 0.0 1st Qu.:40.00 Class :character Class :character   
## Median : 0.0 Median :40.00 Mode :character Mode :character   
## Mean : 87.3 Mean :40.44   
## 3rd Qu.: 0.0 3rd Qu.:45.00   
## Max. :4356.0 Max. :99.00

Check for any missing values in the data.

sapply(sal\_emp,function(x) sum(is.na(x)))

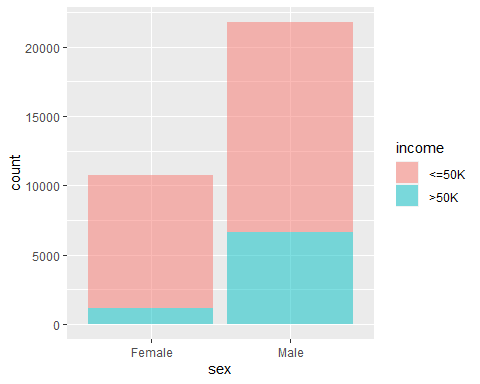
## X age type\_employer fnlwgt education   
## 0 0 0 0 0   
## education\_num marital occupation relationship race   
## 0 0 0 0 0   
## sex capital\_gain capital\_loss hr\_per\_week native.country   
## 0 0 0 0 0   
## income   
## 0

The data is clean and there are no missing values in the data.

## Methodology

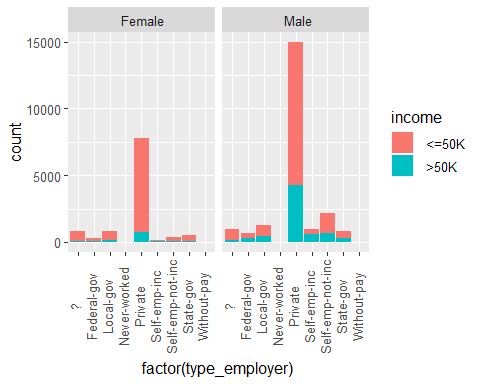
Since the data is clean, the next thing is perform analysis and visualization. First, we perform analysis on the distribution of salary in terms of gender.

ggplot(sal\_emp, aes(sex)) + geom\_bar(aes(fill= income), alpha=0.5)



An analysis of the distribution salary for the type of employer and gender. This is done by creating a stacked barchart.

ggplot(sal\_emp, aes(factor(type\_employer))) + geom\_bar(aes(fill=income)) + facet\_grid(.~ sex) + theme(text = element\_text(size=12),axis.text.x = element\_text(angle=90, vjust=0.2))



Next, we perform an analysis of on the distribution of salary for different types of employees.

table(sal\_emp$type\_employer)

##   
## ? Federal-gov Local-gov Never-worked   
## 1836 960 2093 7   
## Private Self-emp-inc Self-emp-not-inc State-gov   
## 22696 1116 2541 1298   
## Without-pay   
## 14

sal\_emp$type\_employer <- as.character(sal\_emp$type\_employer)  
str(sal\_emp)

## 'data.frame': 32561 obs. of 16 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ age : int 39 50 38 53 28 37 49 52 31 42 ...  
## $ type\_employer : chr " State-gov" " Self-emp-not-inc" " Private" " Private" ...  
## $ fnlwgt : int 77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...  
## $ education : chr " Bachelors" " Bachelors" " HS-grad" " 11th" ...  
## $ education\_num : int 13 13 9 7 13 14 5 9 14 13 ...  
## $ marital : chr " Never-married" " Married-civ-spouse" " Divorced" " Married-civ-spouse" ...  
## $ occupation : chr " Adm-clerical" " Exec-managerial" " Handlers-cleaners" " Handlers-cleaners" ...  
## $ relationship : chr " Not-in-family" " Husband" " Not-in-family" " Husband" ...  
## $ race : chr " White" " White" " White" " Black" ...  
## $ sex : chr " Male" " Male" " Male" " Male" ...  
## $ capital\_gain : int 2174 0 0 0 0 0 0 0 14084 5178 ...  
## $ capital\_loss : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ hr\_per\_week : int 40 13 40 40 40 40 16 45 50 40 ...  
## $ native.country: chr " United-States" " United-States" " United-States" " United-States" ...  
## $ income : chr " <=50K" " <=50K" " <=50K" " <=50K" ...

unemp <- function(job) {  
 job <- as.character(job)  
 if (job =='Never-worked'| job=='Without-pay') {  
 return("unemployed")  
 } else {  
 return(job)  
 }  
}

sal\_emp$type\_employer <- sapply(sal\_emp$type\_employer, unemp)  
table(sal\_emp$type\_employer)

##   
## ? Federal-gov Local-gov Never-worked   
## 1836 960 2093 7   
## Private Self-emp-inc Self-emp-not-inc State-gov   
## 22696 1116 2541 1298   
## Without-pay   
## 14

Next we split the data to train set(80%) and test set(20%) for creating the model.

# Set a random seed  
set.seed(101)   
# SplitRatio = percent of sample==TRUE  
sample <- sample.split(sal\_emp$income, SplitRatio = 0.80)   
# Training Data  
train = subset(sal\_emp, sample == TRUE)  
# Testing Data  
test = subset(sal\_emp, sample == FALSE)

Build the logistic regression model.

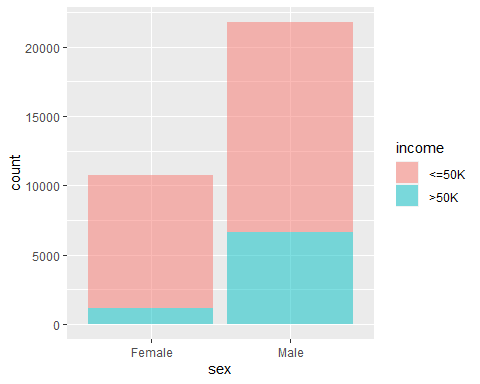
train$income <- as.factor(train$income)  
Train\_Model <- glm(income ~., family = binomial(logit), data =train)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

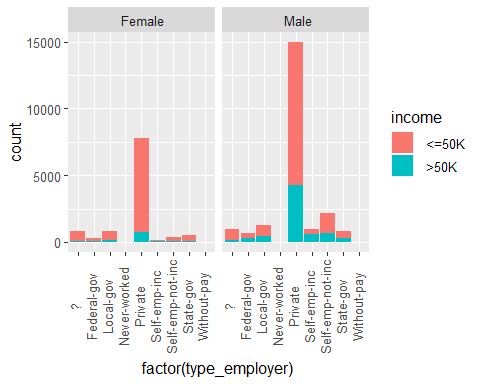
summary(Train\_Model)

##   
## Call:  
## glm(formula = income ~ ., family = binomial(logit), data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.3344 -0.5002 -0.1765 -0.0217 3.7353   
##   
## Coefficients: (2 not defined because of singularities)  
## Estimate Std. Error z value  
## (Intercept) -9.552e+00 5.047e-01 -18.926  
## X 2.338e-06 2.097e-06 1.115  
## age 2.631e-02 1.868e-03 14.081  
## type\_employer Federal-gov 1.056e+00 1.717e-01 6.149  
## type\_employer Local-gov 3.983e-01 1.575e-01 2.529  
## type\_employer Never-worked -1.032e+01 1.005e+03 -0.010  
## type\_employer Private 5.581e-01 1.401e-01 3.983  
## type\_employer Self-emp-inc 7.039e-01 1.685e-01 4.178  
## type\_employer Self-emp-not-inc 2.652e-02 1.536e-01 0.173  
## type\_employer State-gov 2.511e-01 1.699e-01 1.478  
## type\_employer Without-pay -1.430e+01 6.583e+02 -0.022  
## fnlwgt 6.454e-07 1.926e-07 3.351  
## education 11th 4.132e-02 2.318e-01 0.178  
## education 12th 4.462e-01 2.932e-01 1.521  
## education 1st-4th -7.213e-01 5.438e-01 -1.326  
## education 5th-6th -5.849e-01 3.931e-01 -1.488  
## education 7th-8th -6.568e-01 2.672e-01 -2.459  
## education 9th -3.561e-01 2.899e-01 -1.228  
## education Assoc-acdm 1.272e+00 1.952e-01 6.520  
## education Assoc-voc 1.383e+00 1.871e-01 7.391  
## education Bachelors 1.904e+00 1.738e-01 10.955  
## education Doctorate 3.062e+00 2.396e-01 12.783  
## education HS-grad 7.803e-01 1.689e-01 4.620  
## education Masters 2.176e+00 1.856e-01 11.726  
## education Preschool -2.107e+01 2.773e+02 -0.076  
## education Prof-school 2.778e+00 2.235e-01 12.430  
## education Some-college 1.069e+00 1.716e-01 6.229  
## education\_num NA NA NA  
## marital Married-AF-spouse 3.161e+00 6.271e-01 5.042  
## marital Married-civ-spouse 2.419e+00 2.958e-01 8.177  
## marital Married-spouse-absent 2.182e-01 2.502e-01 0.872  
## marital Never-married -4.909e-01 9.892e-02 -4.962  
## marital Separated -1.208e-01 1.860e-01 -0.649  
## marital Widowed -2.369e-02 1.828e-01 -0.130  
## occupation Adm-clerical 1.385e-01 1.112e-01 1.246  
## occupation Armed-Forces -1.012e+00 1.566e+00 -0.646  
## occupation Craft-repair 2.349e-01 9.470e-02 2.480  
## occupation Exec-managerial 9.347e-01 9.777e-02 9.560  
## occupation Farming-fishing -9.427e-01 1.619e-01 -5.824  
## occupation Handlers-cleaners -4.907e-01 1.620e-01 -3.029  
## occupation Machine-op-inspct -1.088e-01 1.186e-01 -0.917  
## occupation Other-service -7.320e-01 1.399e-01 -5.231  
## occupation Priv-house-serv -1.378e+01 1.829e+02 -0.075  
## occupation Prof-specialty 6.023e-01 1.051e-01 5.732  
## occupation Protective-serv 7.028e-01 1.461e-01 4.809  
## occupation Sales 4.085e-01 1.009e-01 4.047  
## occupation Tech-support 8.282e-01 1.320e-01 6.273  
## occupation Transport-moving NA NA NA  
## relationship Not-in-family 7.885e-01 2.931e-01 2.690  
## relationship Other-relative -2.993e-01 2.651e-01 -1.129  
## relationship Own-child -4.587e-01 2.859e-01 -1.605  
## relationship Unmarried 6.755e-01 3.109e-01 2.173  
## relationship Wife 1.424e+00 1.155e-01 12.332  
## race Asian-Pac-Islander 8.129e-01 3.203e-01 2.538  
## race Black 5.754e-01 2.773e-01 2.075  
## race Other 1.735e-01 4.056e-01 0.428  
## race White 7.503e-01 2.652e-01 2.829  
## sex Male 9.113e-01 8.952e-02 10.180  
## capital\_gain 3.289e-04 1.165e-05 28.240  
## capital\_loss 6.854e-04 4.238e-05 16.173  
## hr\_per\_week 3.101e-02 1.826e-03 16.984  
## native.country Cambodia 1.528e+00 6.540e-01 2.337  
## native.country Canada 2.886e-01 3.288e-01 0.878  
## native.country China -4.933e-01 4.528e-01 -1.089  
## native.country Columbia -1.931e+00 8.328e-01 -2.319  
## native.country Cuba 4.321e-01 3.707e-01 1.166  
## native.country Dominican-Republic -1.259e+00 1.073e+00 -1.174  
## native.country Ecuador -1.912e-01 8.171e-01 -0.234  
## native.country El-Salvador -3.100e-01 5.643e-01 -0.549  
## native.country England 5.836e-01 3.698e-01 1.578  
## native.country France 1.297e+00 6.031e-01 2.151  
## native.country Germany 6.358e-01 3.165e-01 2.009  
## native.country Greece -9.101e-01 6.650e-01 -1.369  
## native.country Guatemala 3.817e-01 8.079e-01 0.472  
## native.country Haiti 2.269e-01 7.105e-01 0.319  
## native.country Holand-Netherlands -1.215e+01 2.400e+03 -0.005  
## native.country Honduras -9.482e-01 2.597e+00 -0.365  
## native.country Hong 8.584e-02 6.972e-01 0.123  
## native.country Hungary 1.175e-01 8.045e-01 0.146  
## native.country India -2.092e-01 3.706e-01 -0.564  
## native.country Iran 4.462e-01 5.710e-01 0.781  
## native.country Ireland 4.507e-01 7.279e-01 0.619  
## native.country Italy 8.896e-01 3.746e-01 2.375  
## native.country Jamaica 3.045e-01 5.014e-01 0.607  
## native.country Japan 7.765e-01 4.793e-01 1.620  
## native.country Laos -1.279e+00 1.141e+00 -1.122  
## native.country Mexico -2.210e-01 2.848e-01 -0.776  
## native.country Nicaragua -1.255e+00 1.085e+00 -1.157  
## native.country Outlying-US(Guam-USVI-etc) -1.403e+01 5.871e+02 -0.024  
## native.country Peru -5.957e-01 8.720e-01 -0.683  
## native.country Philippines 5.691e-01 3.159e-01 1.801  
## native.country Poland -1.892e-01 5.006e-01 -0.378  
## native.country Portugal -2.968e-02 7.275e-01 -0.041  
## native.country Puerto-Rico -4.855e-01 4.749e-01 -1.022  
## native.country Scotland 2.784e-01 8.231e-01 0.338  
## native.country South -7.076e-01 4.884e-01 -1.449  
## native.country Taiwan 3.007e-01 5.188e-01 0.580  
## native.country Thailand -8.029e-01 9.738e-01 -0.825  
## native.country Trinadad&Tobago 1.392e-01 9.582e-01 0.145  
## native.country United-States 3.671e-01 1.535e-01 2.392  
## native.country Vietnam -1.258e+00 7.055e-01 -1.783  
## native.country Yugoslavia 1.017e+00 7.257e-01 1.402  
## Pr(>|z|)   
## (Intercept) < 2e-16 \*\*\*  
## X 0.264785   
## age < 2e-16 \*\*\*  
## type\_employer Federal-gov 7.82e-10 \*\*\*  
## type\_employer Local-gov 0.011454 \*   
## type\_employer Never-worked 0.991807   
## type\_employer Private 6.81e-05 \*\*\*  
## type\_employer Self-emp-inc 2.94e-05 \*\*\*  
## type\_employer Self-emp-not-inc 0.862940   
## type\_employer State-gov 0.139517   
## type\_employer Without-pay 0.982666   
## fnlwgt 0.000805 \*\*\*  
## education 11th 0.858526   
## education 12th 0.128149   
## education 1st-4th 0.184707   
## education 5th-6th 0.136841   
## education 7th-8th 0.013950 \*   
## education 9th 0.219298   
## education Assoc-acdm 7.03e-11 \*\*\*  
## education Assoc-voc 1.45e-13 \*\*\*  
## education Bachelors < 2e-16 \*\*\*  
## education Doctorate < 2e-16 \*\*\*  
## education HS-grad 3.84e-06 \*\*\*  
## education Masters < 2e-16 \*\*\*  
## education Preschool 0.939439   
## education Prof-school < 2e-16 \*\*\*  
## education Some-college 4.68e-10 \*\*\*  
## education\_num NA   
## marital Married-AF-spouse 4.61e-07 \*\*\*  
## marital Married-civ-spouse 2.90e-16 \*\*\*  
## marital Married-spouse-absent 0.383206   
## marital Never-married 6.96e-07 \*\*\*  
## marital Separated 0.516055   
## marital Widowed 0.896857   
## occupation Adm-clerical 0.212727   
## occupation Armed-Forces 0.518062   
## occupation Craft-repair 0.013134 \*   
## occupation Exec-managerial < 2e-16 \*\*\*  
## occupation Farming-fishing 5.75e-09 \*\*\*  
## occupation Handlers-cleaners 0.002453 \*\*   
## occupation Machine-op-inspct 0.359039   
## occupation Other-service 1.69e-07 \*\*\*  
## occupation Priv-house-serv 0.939924   
## occupation Prof-specialty 9.94e-09 \*\*\*  
## occupation Protective-serv 1.51e-06 \*\*\*  
## occupation Sales 5.18e-05 \*\*\*  
## occupation Tech-support 3.55e-10 \*\*\*  
## occupation Transport-moving NA   
## relationship Not-in-family 0.007142 \*\*   
## relationship Other-relative 0.258853   
## relationship Own-child 0.108598   
## relationship Unmarried 0.029791 \*   
## relationship Wife < 2e-16 \*\*\*  
## race Asian-Pac-Islander 0.011164 \*   
## race Black 0.038001 \*   
## race Other 0.668933   
## race White 0.004668 \*\*   
## sex Male < 2e-16 \*\*\*  
## capital\_gain < 2e-16 \*\*\*  
## capital\_loss < 2e-16 \*\*\*  
## hr\_per\_week < 2e-16 \*\*\*  
## native.country Cambodia 0.019454 \*   
## native.country Canada 0.380036   
## native.country China 0.275978   
## native.country Columbia 0.020418 \*   
## native.country Cuba 0.243815   
## native.country Dominican-Republic 0.240517   
## native.country Ecuador 0.815006   
## native.country El-Salvador 0.582720   
## native.country England 0.114470   
## native.country France 0.031489 \*   
## native.country Germany 0.044534 \*   
## native.country Greece 0.171107   
## native.country Guatemala 0.636616   
## native.country Haiti 0.749466   
## native.country Holand-Netherlands 0.995960   
## native.country Honduras 0.715021   
## native.country Hong 0.902015   
## native.country Hungary 0.883870   
## native.country India 0.572432   
## native.country Iran 0.434573   
## native.country Ireland 0.535834   
## native.country Italy 0.017573 \*   
## native.country Jamaica 0.543653   
## native.country Japan 0.105183   
## native.country Laos 0.261984   
## native.country Mexico 0.437808   
## native.country Nicaragua 0.247390   
## native.country Outlying-US(Guam-USVI-etc) 0.980936   
## native.country Peru 0.494545   
## native.country Philippines 0.071678 .   
## native.country Poland 0.705439   
## native.country Portugal 0.967450   
## native.country Puerto-Rico 0.306671   
## native.country Scotland 0.735195   
## native.country South 0.147373   
## native.country Taiwan 0.562235   
## native.country Thailand 0.409636   
## native.country Trinadad&Tobago 0.884529   
## native.country United-States 0.016743 \*   
## native.country Vietnam 0.074589 .   
## native.country Yugoslavia 0.160969   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 28759 on 26048 degrees of freedom  
## Residual deviance: 16310 on 25949 degrees of freedom  
## AIC: 16510  
##   
## Number of Fisher Scoring iterations: 15

## Results

From the analysis of the  salary distribution in terms of gender, there are more males than Females. In addition, more males earn less than 50k than those that earn above 50k. This is the same for females. There are more females who earn less than 50k compared to those who earn above 50k.  


The analysis on distribution salary for the type of employer and gender clearly shows that for private employer, they have the highest number of employees both for females and males. There are very few females in the self employment income. Federal-gov has the least males in employment.



## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :  
## prediction from a rank-deficient fit may be misleading

The model has an accuracy of 84%, which means it is pretty good to accurately predict the data.

print("The accuracy of the model")

## [1] "The accuracy of the model"

(6392+1414)/(528+881+1414+6392)

## [1] 0.8470971

## Implications

This project only applied the logistic regression model for the salary prediction. However, there needs to have more analysis and evaluation of the performance of other predictive models like Random Forest, Decision Tree, Support Vector Machines to ensure that the best and most suitable model in terms of accuracy and efficiency is used for salary prediction.

## Conclusion

A logistic regression model has been created in this project to provide salary predictions. The theory that there is no dependency between the type of employer and salary is not true. It is clear that private employers they have the highest number of employees who earn both less than 50k and above 50k.

# References

Bhuller, M., Mogstad, M., & Salvanes, K. G. (2017). Life-cycle earnings, education premiums, and internal rates of return. Journal of Labor Economics, 35(4), 993-1030.

Pawha, A., & Kamthania, D. (2019). Quantitative analysis of historical data for prediction of job salary in India-A case study. Journal of Statistics and Management Systems, 22(2), 187-198.

Shwartz, M., Burgess Jr, J. F., & Zhu, J. (2016). A DEA based composite measure of quality and its associated data uncertainty interval for health care provider profiling and pay-for-performance. European Journal of Operational Research, 253(2), 489-502.

Dataset: <https://archive.ics.uci.edu/ml/machine-learning-databases/adult/>