Analyze & Predict Hourly Wages

Abstract

Increasing education: What it will and will not do for earnings and earnings inequality? Are the earnings of those with a college degree more while the wages of those with lower levels of education stagnated or fallen? This project is also to explore Gender pay gap. These form the basis for the project study, to determine whether Age, Gender, Years of education play a key role in determining the wages of a person. In this project I have used the regression to explore the strengths in relationship, describe data and explain relationship between dependent and one or more independent variables as a primary step. The results were analyzed and Naïve Bayes was used for classification. Further study was done using Support Vector Machine classification and plots were plotted to better analyze the results. High accuracy, nice theoretical guarantees regarding overfitting, and with an appropriate kernel makes SVM a fitting choice .Regression ,Naïve Bayes and SVM were used as accuracy was very important and the only way to select the best was by cross-validation.

Introduction

Wage differentiation is a widespread phenomenon. Even wage segregation is common: certain jobs are "socially" attributed to a certain gender, which in turn might strive to widen the jobs it is "allowed" to perform. Wage structure for occupations, industries and regions is subject to very slow long-term change.

Elderly and youthful employees sometimes experience age discrimination in the workplace. Ageism, is stereotyping and discriminating against individuals or groups on the basis of their age. Does pay vary based on age?

Women have made tremendous strides during the last few decades by moving into jobs and occupations previously done almost exclusively by men, yet Pay equity may be affected by the segregation of jobs by gender and other factors. The gender wage gap fluctuates and changes depending on age and years of education, but across the board, on average, do women make less than men?

The cost of education these days would make anyone squirm, but is it worth it? People with less education in high-paying occupations can out-earn their counterparts with advanced degrees. But within the same industry, workers with more schooling usually land better paychecks.

To be able to answer some of these questions I have taken data from the Survey of Labor and Income Dynamics (SLID) which uses a mixed collection mode that combines survey data with

data from administrative sources. A greater insight and nature on the influence of age, gender and education is obtained.

Code with documentation

1. loading required packages

Uncomment the follwing install commands to install packages if needed

```
#install.packages("klaR")
#install.packages("e1071")
#install.packages("dplyr")
#install.packages("data.table")
#install.packages("e1071")
#install.packages("caret")
#install.packages("klaR")
#install.packages("randomForest")

require(klaR)
require(e1071)
require(dplyr)
require(ggplot2)
require(data.table)
require(car)
require(randomForest)

library(plyr)
library(dplyr)
library(data.table)
library(e1071)
library(caret)
library(ggplot2)
```

2. Load dataset as a data frame

*The following script assumes that **SLID-Ontario.txt** file is present in the current working directory*

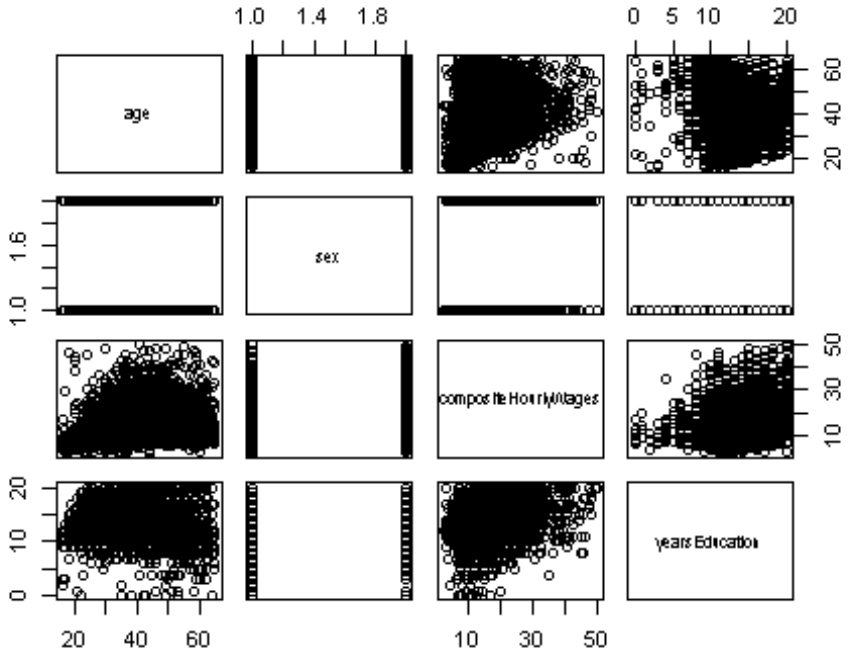
```
# Cleanup workspace.
rm(list=ls())

# Load data file as a data frame.
wages.data <- read.table("SLID-Ontario.txt", header = TRUE)
wages.data <- na.omit(wages.data)
```

2.1 View structure of the data

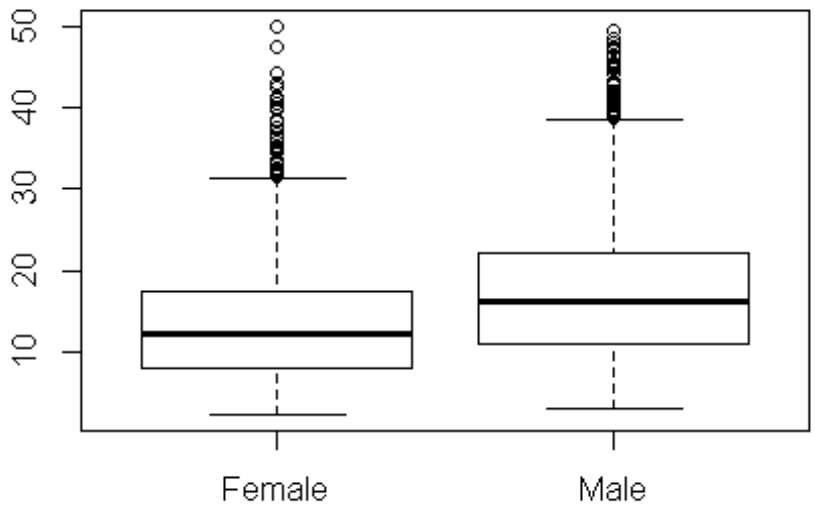
```
summary(wages.data)
```

```
##      age          sex      compositeHourlyWages yearsEducation
##  Min.   :16.00   Female:2007   Min.   : 2.30       Min.   : 0.00
##  1st Qu.:28.00   Male  :1990   1st Qu.: 9.25       1st Qu.:12.00
##  Median :36.00                 Median :14.13       Median :13.00
##  Mean   :36.96                 Mean   :15.54       Mean   :13.21
##  3rd Qu.:46.00                 3rd Qu.:19.75       3rd Qu.:15.00
##  Max.   :65.00                 Max.   :49.92       Max.   :20.00
```

```
pairs(wages.data)
```




```
# Plot and visually inspect if gender plays a role in wages.
plot(wages.data$sex, wages.data$compositeHourlyWages)
```



It seems like gender plays a role in hourly wages. This can be further analyzed using more advanced techniques.

3. Analysis

3.1 Regression model

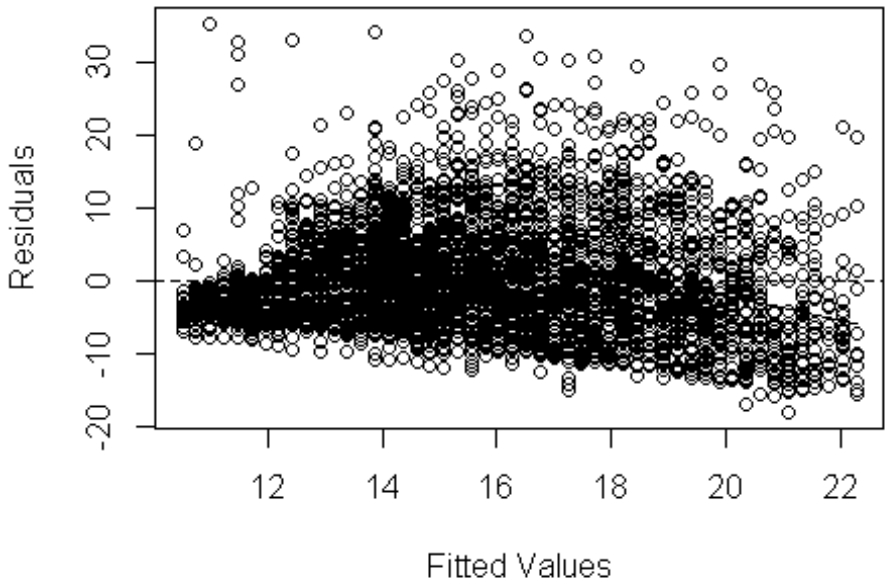
3.1.1 Age vs Hourly wage

Determine how hourly wage changes over age

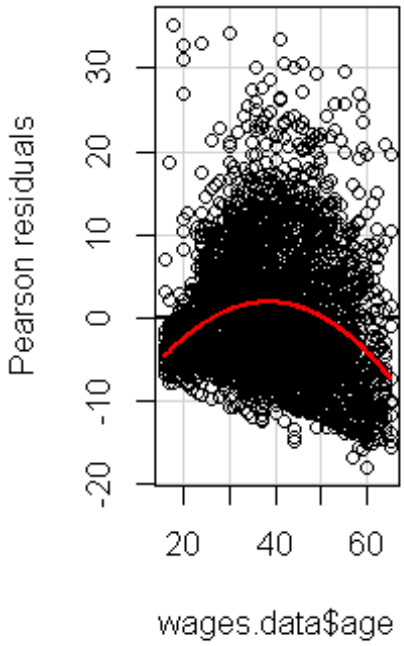
```
# Build a simple regression model to analyze how age determines hourly wage.
wageAndAge.lmfit    = lm(wages.data$compositeHourlyWages ~ wages.data$age)
summary(wageAndAge.lmfit)

##
## Call:
## lm(formula = wages.data$compositeHourlyWages ~ wages.data$age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.939  -4.810  -1.487   3.914  35.151
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

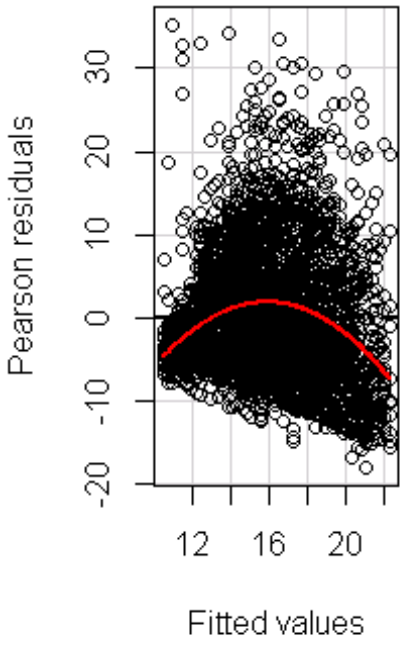
```
## (Intercept)    6.683133   0.374440   17.85   <2e-16 ***
## wages.data$age 0.239770   0.009636   24.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.312 on 3995 degrees of freedom
## Multiple R-squared:  0.1342, Adjusted R-squared:  0.134
## F-statistic: 619.1 on 1 and 3995 DF,  p-value: < 2.2e-16
```

```
plot(fitted(wageAndAge.lmfit), residuals(wageAndAge.lmfit),
 xlab = "Fitted Values", ylab = "Residuals")
 abline(h=0, lty=2)
 lines(smooth.spline(fitted(wageAndAge.lmfit), residuals(wageAndAge.lmfit)))
```




```
residualPlots(wageAndAge.lmfit)
```






```
##              Test stat Pr(>|t|)
## wages.data$age   -17.872       0
## Tukey test       -17.872       0
```

The values are reasonably densed around the center line and the model can be used to perform intial analysis.

It seems like hourly wages increase over age in general but tend to decline over late years

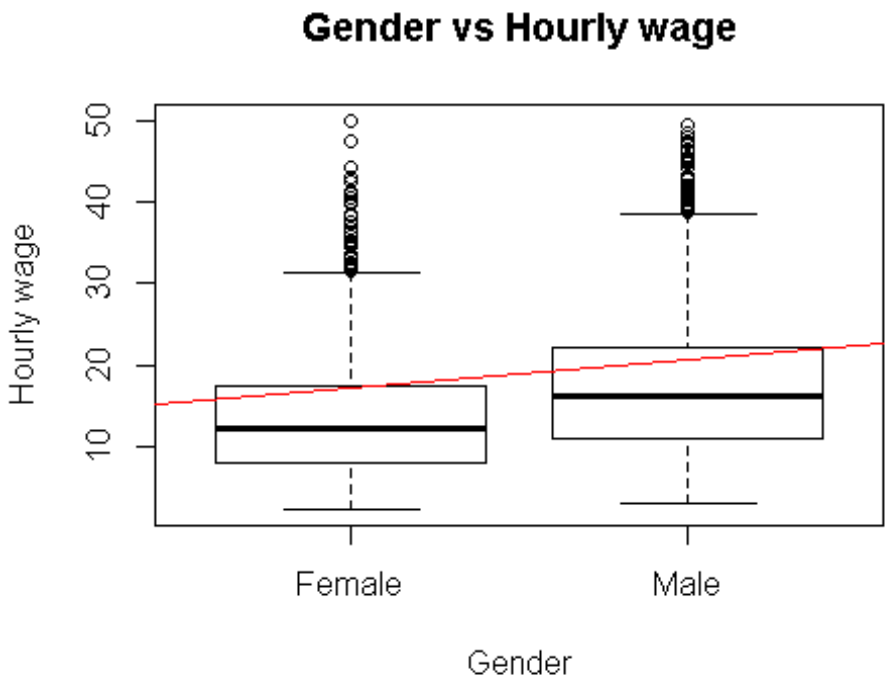
3.1.2 Gender vs Hourly wage

Determine if gender plays a role in hourly wages

```
# Build a simple regression model to analyze how gender determines hourly wage.
genderAndEducation.lmfit    = lm(wages.data$compositeHourlyWages ~ wages.data$sex, data =
wages.data)
summary(genderAndEducation.lmfit)

##
## Call:
## lm(formula = wages.data$compositeHourlyWages ~ wages.data$sex,
##     data = wages.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.263  -5.961  -1.331   4.159  36.079
```

```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)       13.8411     0.1712   80.85   <2e-16 ***
## wages.data$sexMale   3.4215     0.2426   14.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.67 on 3995 degrees of freedom
## Multiple R-squared:  0.04742,    Adjusted R-squared:  0.04718
## F-statistic: 198.9 on 1 and 3995 DF,  p-value: < 2.2e-16
```

```
plot(wages.data$compositeHourlyWages~wages.data$sex, data=wages.data, main="Gender vs Hourly wage", xlab = "Gender", ylab = "Hourly wage")
abline(genderAndEducation.lmfit, col="red")
```

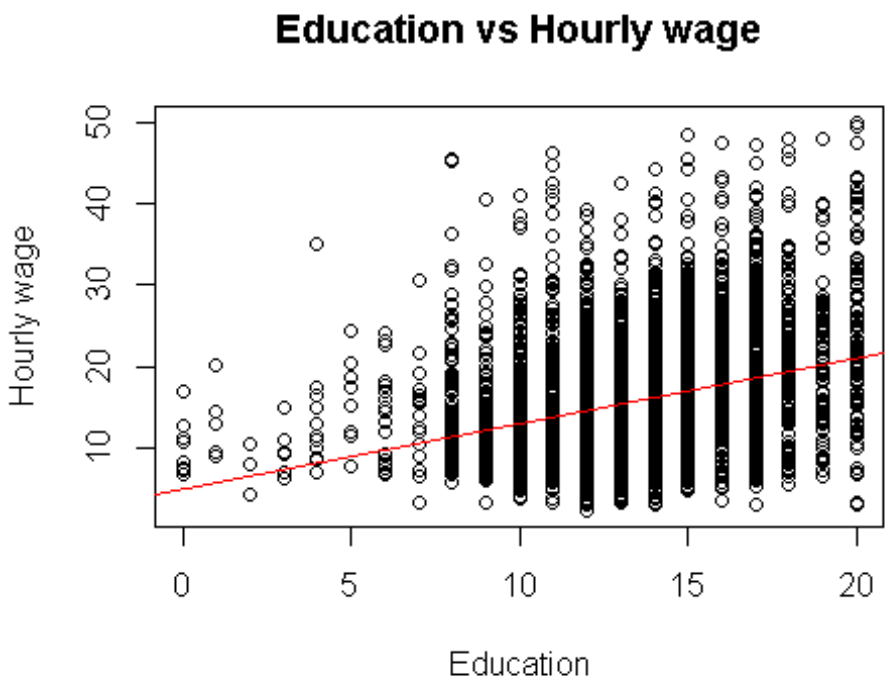


It seems gender plays a significant role in hourly wages. Men tend to getting paid more then women.

3.1.2 Years of education vs Hourly wage

```
# Build a simple regression model to analyze how years of education determines hourly wage.

wageAndEducation.lmfit  = lm(wages.data$compositeHourlyWages ~ wages.data$yearsEducati
```

```
on, data = wages.data)
summary(wageAndEducation.lmfit)

##
## Call:
## lm(formula = wages.data$compositeHourlyWages ~ wages.data$yearsEducation,
##     data = wages.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.939  -5.792  -0.981   4.133  34.197
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               4.79189    0.52681   9.096   <2e-16 ***
## wages.data$yearsEducation  0.81386    0.03886  20.943   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.459 on 3995 degrees of freedom
## Multiple R-squared:  0.09893,    Adjusted R-squared:  0.0987
## F-statistic: 438.6 on 1 and 3995 DF,  p-value: < 2.2e-16

plot(wages.data$compositeHourlyWages~wages.data$yearsEducation, data=wages.data, main
="Education vs Hourly wage", xlab = "Education", ylab = "Hourly wage")
abline(wageAndEducation.lmfit, col="red")
```



It seems that as number of years of education increases pay in general increases

3.2 Naive-Bayes

Perform further analysis using Naive-Bayes algorithm

3.2.1 Manipulate data

Add range of wages to existing data

```
min.wage = floor(min(wages.data$compositeHourlyWages))
max.wage = ceiling(max(wages.data$compositeHourlyWages))
num.of.levels = 10 # Add ten wage ranges.
level.range = ceiling((max.wage - min.wage) / num.of.levels)

# Add new levels to dataset.
wages.data$wageRange <- ""
x <- min.wage
y <- level.range

repeat {
  wages.data[wages.data$compositeHourlyWages >= x & wages.data$compositeHourlyWages <
y ,]$wageRange = paste(x,"To",y)
  x <- y
  y <- x + level.range
```

```
 if(y > max.wage){
   break
 }
}

wages.data.new <- wages.data
wages.data.new$compositeHourlyWages <- NULL # Remove composite data.
```

3.2.1 Naive-Bayes classification model

Create trainig and testing data

```
set.seed(1234)
index <- createDataPartition(wages.data.new$wageRange, p = .8, list = FALSE)
wages.training.data <- wages.data.new[index, ]
wages.testing.data <- wages.data.new[-index, ]
```

Create model

```
nb.model <- naiveBayes(as.factor(wageRange)~., data = wages.training.data)

# exploring the nb object:
names(nb.model)

## [1] "apriori" "tables"  "levels"  "call"

nb.pred <- predict(nb.model, newdata=wages.testing.data, laplace=3)

## Warning in data.matrix(newdata): NAs introduced by coercion
```

Predictions for gender vs wages

```
table(nb.pred, wages.testing.data$sex)

##
## nb.pred    Female Male
##   10 To 15    189   57
##   15 To 20     29  213
##   2 To 5        0    0
##   20 To 25     12   24
##   25 To 30      0    0
##   30 To 35      0    0
##   35 To 40      0    0
##   40 To 45      0    0
##   45 To 50      0    0
##   5 To 10     171  100
```

Predictions for age vs wages

```
table(nb.pred, wages.testing.data$age)
```

```
##
## nb.pred    16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
##   10 To 15  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  2  3  2  3  4 11
##   15 To 20  0  0  0  0  0  0  0  3  2  3  3  5  2  5  5 12  7  8 11  6  9
##   2 To 5    0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   20 To 25  0  0  0  0  0  0  0  0  0  0  0  0  0  1  1  0  1  0  1  1  2
##   25 To 30  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   30 To 35  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   35 To 40  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   40 To 45  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   45 To 50  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   5 To 10  10 14 19 15 15 13 18 18 12 21 14 16 17 14 13 11 10  3 10  8  0
##
## nb.pred    37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57
##   10 To 15 14  7 12 15  7 13  5  6  6  8 11 12  5  8  8  9 11  9  3  4 11
##   15 To 20 12 11  8 16 10  9  5 10 12  9  9  8 10  4  6  1  6  5  0  5  2
##   2 To 5    0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   20 To 25  1  1  2  1  1  2  1  0  1  2  3  4  0  2  1  1  1  0  1  1  0
##   25 To 30  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   30 To 35  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   35 To 40  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   40 To 45  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   45 To 50  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   5 To 10   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##
## nb.pred    58 59 60 61 62 63 64 65
##   10 To 15  8  5  7  5  6  1  3  2
##   15 To 20  0  0  0  1  0  0  1  1
##   2 To 5    0  0  0  0  0  0  0  0
##   20 To 25  0  1  1  0  0  1  0  0
##   25 To 30  0  0  0  0  0  0  0  0
##   30 To 35  0  0  0  0  0  0  0  0
##   35 To 40  0  0  0  0  0  0  0  0
##   40 To 45  0  0  0  0  0  0  0  0
##   45 To 50  0  0  0  0  0  0  0  0
##   5 To 10   0  0  0  0  0  0  0  0
```

Predictions for years of education vs wages

```
table(nb.pred, wages.testing.data$yearsEducation)
```

```
##
## nb.pred     0  1  2  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##   10 To 15  1  0  0  3  1  2  3 19 16 34 16 66 27 31 12 14  1  0  0  0
##   15 To 20  1  0  0  3  0  0  0  0  3 15 10 36 27 22 26 21 50 16 10  2
##   2 To 5    0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   20 To 25  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 10  7 19
##   25 To 30  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   30 To 35  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   35 To 40  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   40 To 45  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   45 To 50  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   5 To 10   1  1  1  0  0  1  0  0  9 16 30 68 42 46 23 19  9  3  2  0
```

Naive Bayes also confirms that gender plays a role in determining hourly wages

3.2.2 Support Vector Machine model

Build model

```
set.seed(5678)

fit.Control <- trainControl(method = "repeatedcv",
  number = 10,
  repeats = 10)

svm.model <- train(wageRange ~ ., data = wages.training.data,
              method = "svmLinear",
              trControl = fit.Control,
              verbose = FALSE)
svm.model
```

```
## Support Vector Machines with Linear Kernel
##
## 3202 samples
##    3 predictor
##   10 classes: '10 To 15', '15 To 20', '2 To 5', '20 To 25', '25 To 30', '30 To 35', '35 To 40', '40 To
45', '45 To 50', '5 To 10'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 10 times)
## Summary of sample sizes: 2883, 2882, 2882, 2881, 2879, 2880, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.3783627  0.1759849
##
```

```
## Tuning parameter 'C' was held constant at a value of 1
##

svm.pred <- predict(svm.model, newdata=wages.testing.data, laplace=3)
```

Predictions for gender vs wages

```
table(svm.pred, wages.testing.data$sex)

##
## svm.pred   Female Male
##   10 To 15    193   21
##   15 To 20     27  282
##   2 To 5        0    0
##   20 To 25      0    0
##   25 To 30      0    0
##   30 To 35      0    0
##   35 To 40      0    0
##   40 To 45      0    0
##   45 To 50      0    0
##   5 To 10     181   91
```

Predictions for age vs wages

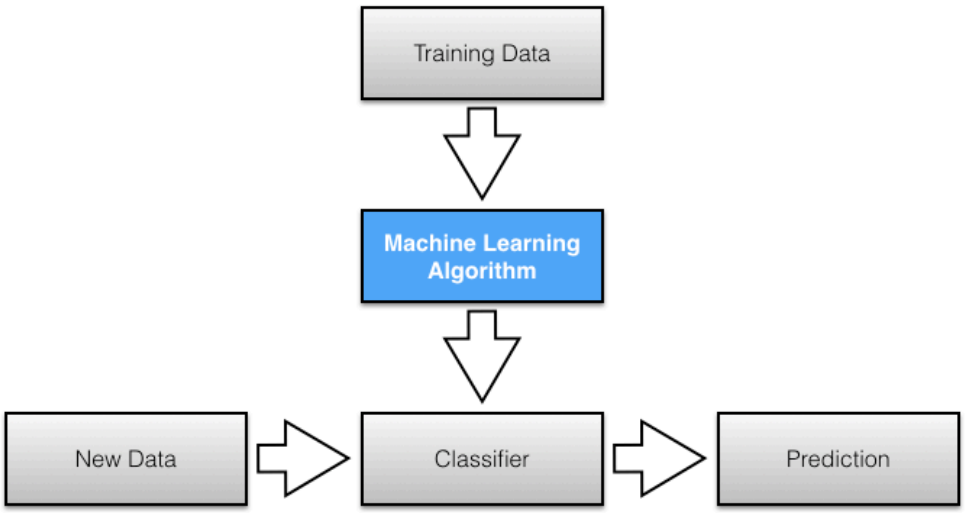
```
table(svm.pred, wages.testing.data$age)

##
## svm.pred   16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
##   10 To 15  0  0  0  0  0  0  0  0  0  0  0  0  0  2  3  2  2  2  2  6  7
##   15 To 20  0  0  0  0  0  1  2  3  4  4  4  6  3  5  3 12  9  8 12  5 11
##   2 To 5    0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   20 To 25  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   25 To 30  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   30 To 35  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   35 To 40  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   40 To 45  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   45 To 50  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   5 To 10  10 14 19 15 15 12 16 18 10 20 13 15 16 13 13 11 10  3 11  8  4
##
## svm.pred   37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57
##   10 To 15 13  8 12 17  7 14  5  8  5  8 11 12  5  7  7  7  9  7  0  1  9
##   15 To 20 13  8 10 15 10 10  6  8 13 11 12 12 10  7  8  4  9  7  4  9  4
##   2 To 5    0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   20 To 25  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   25 To 30  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   30 To 35  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

```
##   35 To 40  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   40 To 45  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   45 To 50  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   5 To 10   1  3  0  0  1  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0
##
## svm.pred   58 59 60 61 62 63 64 65
##   10 To 15  3  2  1  3  3  1  2  1
##   15 To 20  5  4  7  3  3  1  2  2
##   2 To 5    0  0  0  0  0  0  0  0
##   20 To 25  0  0  0  0  0  0  0  0
##   25 To 30  0  0  0  0  0  0  0  0
##   30 To 35  0  0  0  0  0  0  0  0
##   35 To 40  0  0  0  0  0  0  0  0
##   40 To 45  0  0  0  0  0  0  0  0
##   45 To 50  0  0  0  0  0  0  0  0
##   5 To 10   0  0  0  0  0  0  0  0
```

Predictions for years of education vs wages

```
table(svm.pred, wages.testing.data$yearsEducation)

##
## svm.pred    0  1  2  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##   10 To 15  1  0  0  4  1  2  3  6  8 23 13 54 27 31 12  7 12  2  6  2
##   15 To 20  0  0  0  0  0  0  0 13  9 24 13 44 28 24 29 31 40 24 11 19
##   2 To 5    0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   20 To 25  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   25 To 30  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   30 To 35  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   35 To 40  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   40 To 45  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   45 To 50  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   5 To 10   2  1  1  2  0  1  0  0 11 18 30 72 41 44 20 16  8  3  2  0
```

Results



Basic steps –

Data is loaded and explored to observe using the summary commands. Data is visually observed by plotting plots to see the dependencies in data.

Regression Analysis Step -

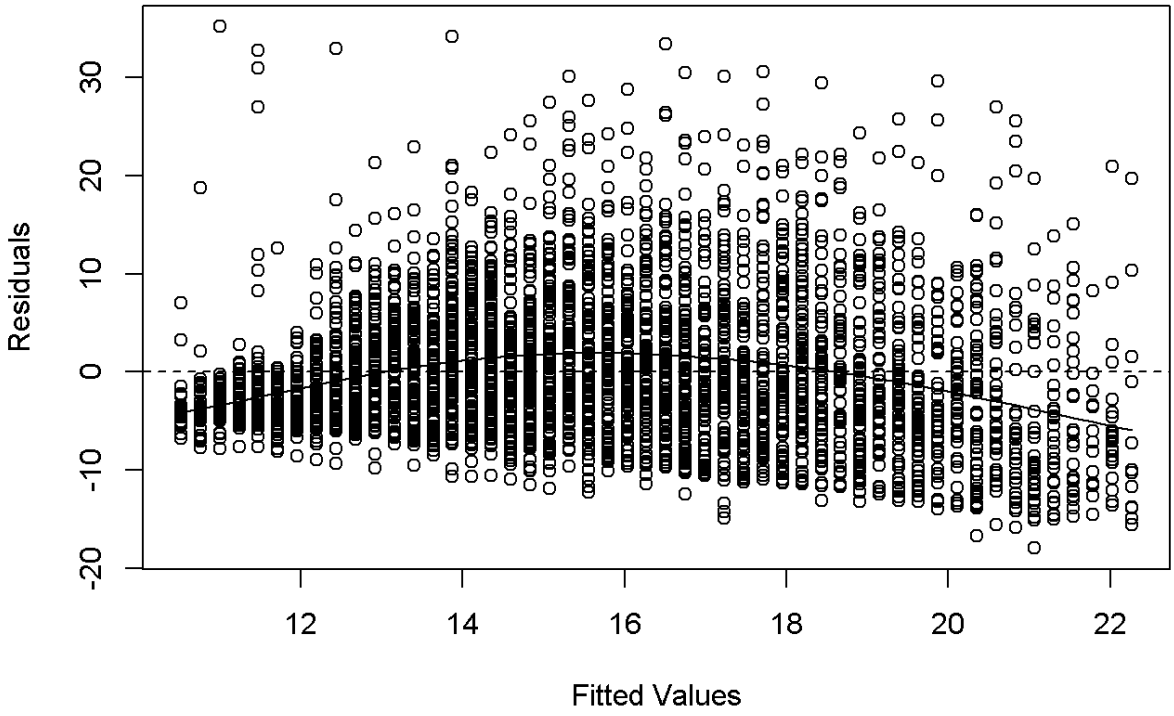
As a first step towards analyzing the data Regression analysis was performed as the first step. It is a statistical technique that attempts to explore and model the relationship between two or more variables. Regression analysis is to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships.

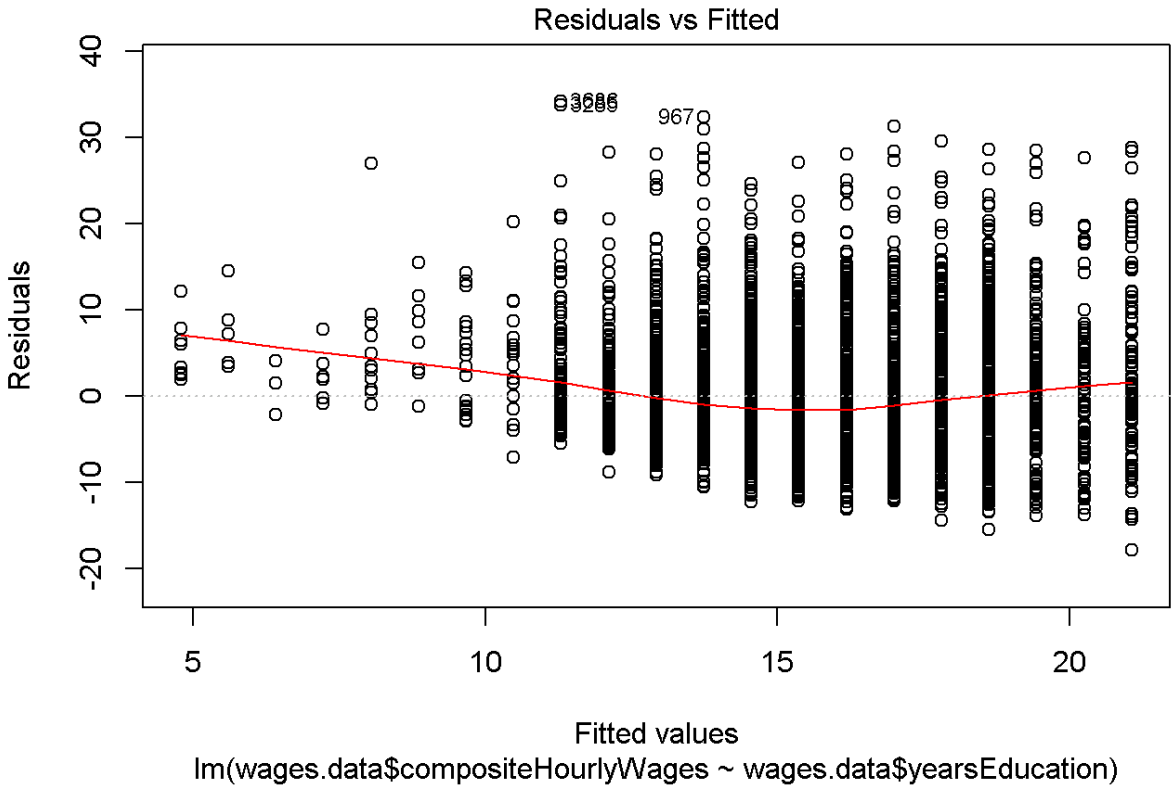
The predictor variables used are Gender, Years of education and and Age.

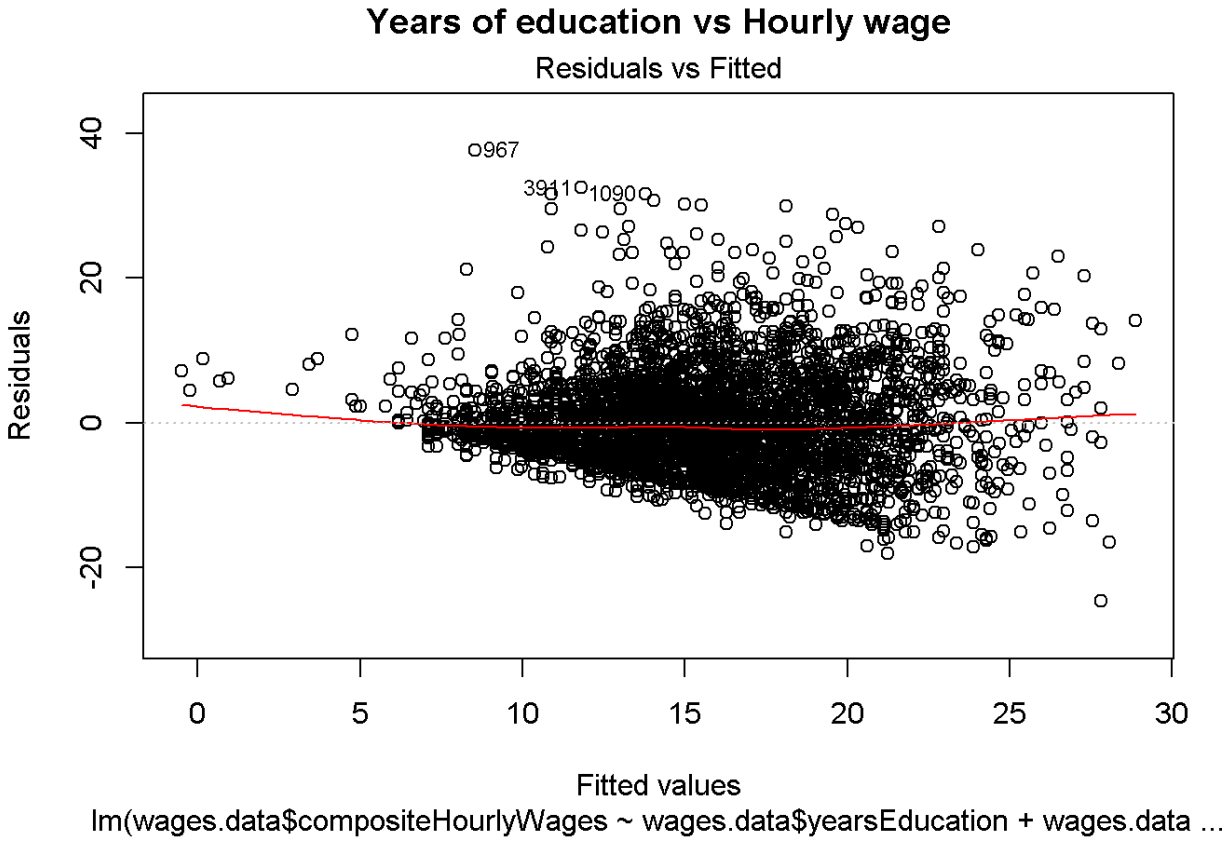
- Age vs Hourly wage
- Gender vs hourly wage
- Years of education vs hourly wage

- All three predictions indicated *higher significance legend* and that there is a higher relationship between wage and age, gender vs wage and also years of education vs wage.
- *Standard error* is also less than estimated coefficient.

- The *Degrees of Freedom* is the difference between the number of observations included in your training sample and the number of variables used in model (intercept counts as a variable).The numbers are higher suggesting that it may not be a normal distribution.
- The *R Squared* values ideally have to be in a higher range with 1 being the best .The ranges obtained for various predictors are 0.09893, 0.2585, and 0.04742.
- The visual plots also determine that there are lot of outliers in the graphs.




Residuals vs Fitted
Residuals
40
30
20
10
0
-10
-20
5
10
15
20
967
Fitted values
lm(wages.data$compositeHourlyWages ~ wages.data$yearsEducation)



Naïve Bayes Algorithm For Classification

Second step performed is Naïve Bayes.The Naive Bayes Classifier technique is based on the Bayesian theorem and is particularly suited when the dimensionality of the inputs is high.Naive Bayes classifiers can handle an arbitrary number of independent variables whether continuous or categorical.

New range of wages are added to the existing data, a Naïve Bayes classification model is created. Training data and testing data is divided .The data is divided as 80% training data and 20% testing data and model is created. The results observed are observed on testing data.

For wage vs gender

- It was observed that there were *more females in the lower wage range than males*.
- nb.pred Female Male
- ## 10 To 15 189 57
- ## 15 To 20 29 213
- The predictions were based on testing data.

For the number of years of education vs wages

- On the testing data it has been observed that for 15- 20 years of education the pay was higher when compared to 5-10 years of education and 10 – 15 years of education.

For Age vs wages

- The age did not seem to affect the pay scale.

Support Vector Machine

The third step was to perform Support Vector machine model to cross validate the results obtained through Naive Bayes model. The idea of support vector machine is to create a hyper plane in between data sets to indicate which class it belongs to. The challenge is to train the machine to understand structure from data and mapping with the right class label, for the best result, the hyper plane has the largest distance to the nearest training data points of any class.

The kernel used was linear Kernel.

Predictions for Gender vs wage

- Males are compared with the females and Males are on the higher wages scale when compared to females. Around 193 females were around the lower paying scale when only 21 males were present in the same 10-15 range.
- In the next pay range from 15-20 282 males were present whereas only 27 females were present.

Predictions For Age vs Wages

- The age did not affect the pay scale.

Predictions for Wages vs Education

- The results in SVM predict clearly that 15-20 years of education clearly had higher wages than 10-15 years
- 5-10 years has few outliers such as 44 and 72 but in general has a lesser payscale than 10-15 and 15-20 years of education.

Cross Validating SVM and Naïve Bayes

SVM and Naïve Bayes have produced similar results proving that hourly wages are influenced by Gender and Years of education but not by age. The higher numbers indicate a gender inequality and the choice of higher education being important.

Discussion.

- The major points in the study was to find out if wages were influenced by age ,gender and years of education.
- Naïve Bayes has produced results displaying that gender pay gap inequality and the year in higher education is visible .
- SVM has also produced similar results favoring number of years of education.
- Surprisingly SVM and Naïve Bayes have produced results that age is not a major factor when t comes to hourly wages.
- Though they are few outliers the majority of the results prove that as years of education increase there is an increase in pay scale.
- Thus this study prove that on this data set gender equality years of education do matter but age stereotypes does not exist.
- The limitations of this study are that the training data set lacked values in some age ranges making it to difficult to predict for certain age ranges the pay scale.
- Why are these important? These findings are essentially important to prove and to misprove some concepts. Age was a misleading factor proving that age did not matter.
- For further research investigate various models and predict over more recent and bigger data sets to predict more accurately.
- Thus to conclude the earnings inequality is visible gap and years of education does matter and age does not lay a vital role.

References

http://www.statcan.gc.ca/pub/75f0011x/75f0011x2013001-eng.htm- data source

http://www.iwpr.org/initiatives/pay-equity-and-discrimination

http://blog.echen.me/2011/04/27/choosing-a-machine-learning-classifier/

http://stackoverflow.com/questions/3648917/interpreting-naive-bayes-results

http://sebastianraschka.com/Articles/2014_naive_bayes_1.html

https://tom.host.cs.st-andrews.ac.uk/ID5059/L15-LeungSlides.pdf

http://blog.yhat.com/posts/r-lm-summary.html

https://www3.nd.edu/~dial/publications/hall1998decision.pdf

http://www.aauw.org/research/the-simple-truth-about-the-gender-pay-gap/