# KLIPS Data Analysis Report

Anar Abbas 20180821

## Introduction

In this research, we will analyze the data taken from the Korean Labor and Income Panel Study (KLIPS) which was launched in 1998. For this research, we will use the dataset from the 7th wave of the KLIPS in 2004. The goal of this this study, is to measure the participants' experiences of hiring discrimination.

The main objectives of this research is to answer the following questions:

- **Question 1**
- Is there a difference in under-reporting of hiring discrimination between males and females?
- **Question 2**
- Is there an association between the experience of hiring discrimination and health

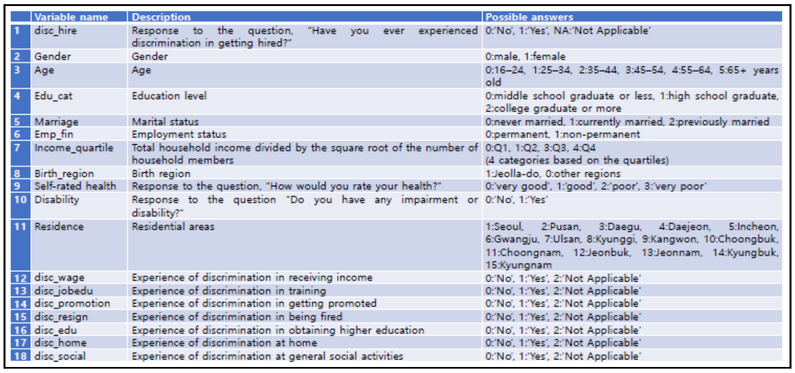
In order to answer the first question we will build a prediction model to predict the response of the 97 workers who responded “Not Applicable” to the question regarding hiring discrimination. Then we will compare the results of the predicted responses between males and females.

To answer the second question, we will compare the distribution of self-rated health among the fours groups of people whose response was “No’, “Yes”, “NA” but predicted as “No” and “NA” but predicted as “YES”.

## Methods

Data description:

The dataset that we will be analyzing includes the answers of 3,576 participants who were waged workers at the time of the survey to the question “Have you ever experienced discrimination in getting hired?”. Out of 3,576 participants, 97 of them responded “Not applicable” while the rest of them answered by “Yes” or “No”. This data includes 18 variables for each participant and each variable is given by some numeric variable corresponding to one's answer. For example, “disc_hire” variable is equal to 0,1, or NA corresponding to logical answers ‘No’, ‘Yes’, and ‘Not applicable’, respectively. In the below table, you can see all the variables, their descriptions and possible answers:

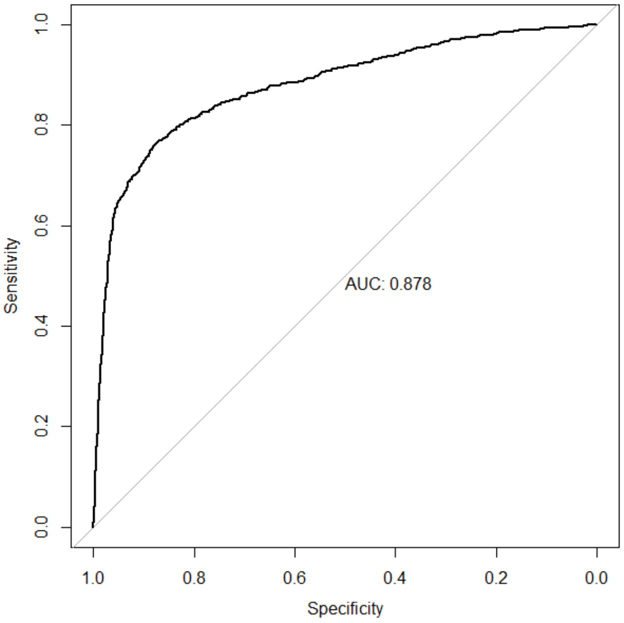


| | Variable name | Description | Possible answers |
|---|---|---|---|
| 1 | disc_hire | Response to the question, "Have you ever experienced discrimination in getting hired?" | 0:'No', 1:'Yes', NA:'Not Applicable' |
| 2 | Gender | Gender | 0:male, 1:female |
| 3 | Age | Age | 0:16–24, 1:25–34, 2:35–44, 3:45–54, 4:55–64, 5:65+ years old |
| 4 | Edu_cat | Education level | 0:middle school graduate or less, 1:high school graduate, 2:college graduate or more |
| 5 | Marriage | Marital status | 0:never married, 1:currently married, 2:previously married |
| 6 | Emp_fin | Employment status | 0:permanent, 1:non-permanent |
| 7 | Income_quartile | Total household income divided by the square root of the number of household members | 0:Q1, 1:Q2, 3:Q3, 4:Q4<br>(4 categories based on the quartiles) |
| 8 | Birth_region | Birth region | 1:Jeolla-do, 0:other regions |
| 9 | Self-rated health | Response to the question, "How would you rate your health?" | 0:'very good', 1:'good', 2:'poor', 3:'very poor' |
| 10 | Disability | Response to the question "Do you have any impairment or disability?" | 0:'No', 1:'Yes' |
| 11 | Residence | Residential areas | 1:Seoul, 2:Pusan, 3:Daegu, 4:Daejeon, 5:Incheon, 6:Gwangju, 7:Ulsan, 8:Kyunggi, 9:Kangwon, 10:Choongbuk, 11:Choongnam, 12:Jeonbuk, 13:Jeonnam, 14:Kyungbuk, 15:Kyungnam |
| 12 | disc_wage | Experience of discrimination in receiving income | 0:'No', 1:'Yes', 2:'Not Applicable' |
| 13 | disc_jobedu | Experience of discrimination in training | 0:'No', 1:'Yes', 2:'Not Applicable' |
| 14 | disc_promotion | Experience of discrimination in getting promoted | 0:'No', 1:'Yes', 2:'Not Applicable' |
| 15 | disc_resign | Experience of discrimination in being fired | 0:'No', 1:'Yes', 2:'Not Applicable' |
| 16 | disc_edu | Experience of discrimination in obtaining higher education | 0:'No', 1:'Yes', 2:'Not Applicable' |
| 17 | disc_home | Experience of discrimination at home | 0:'No', 1:'Yes', 2:'Not Applicable' |
| 18 | disc_social | Experience of discrimination at general social activities | 0:'No', 1:'Yes', 2:'Not Applicable' |

To answer the research questions we have mentioned in the Introduction part we will train prediction models using the following six models: logistic regression, random forest, penalized logistic regression with lasso penalty, k-nearest neighbors, support vector machine with radial basis kernel functions and single-layer neural network.

1. Logistic regression model

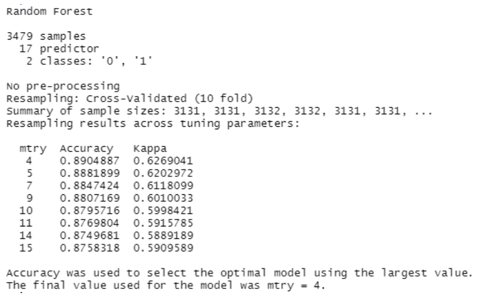
We first train prediction model using Logistic regression with **glm** function with family argument equal to binomial. After training we see that the accuracy for the training data is 0.892 when the threshold for the binary prediction is 0.5. To see how this model performs with different threshold we plot the ROC curve using **pROC** library:



The AUC (Area Under Curve) is 0.878.

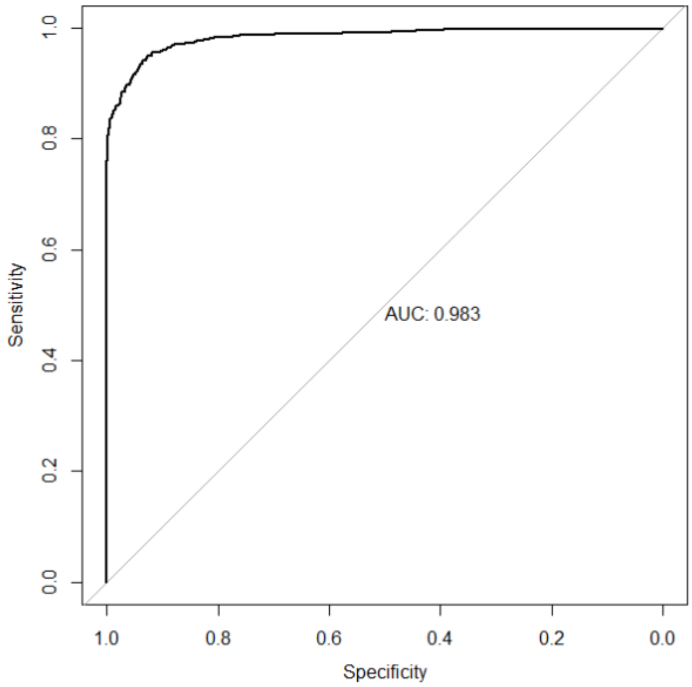
2. Random Forest model

We train random forest model with 10-fold cross-validation using the trainControl function of **caret** library. We use ntree=500 and tuneLength = 10 for predict function. The cross validation gives that the best value for mtry (Number of variables randomly sampled as candidates at each split) is equal to 4.



```
Random Forest

3479 samples
  17 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3131, 3131, 3132, 3132, 3131, 3131, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   4    0.8904887  0.6269041
   5    0.8881899  0.6202972
   7    0.8847424  0.6118099
   9    0.8807169  0.6010033
  10    0.8795716  0.5998421
  11    0.8769804  0.5915785
  14    0.8749681  0.5889189
  15    0.8758318  0.5909589

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 4.
```

The ROC plot is as below:



The AUC value is **0.983**.

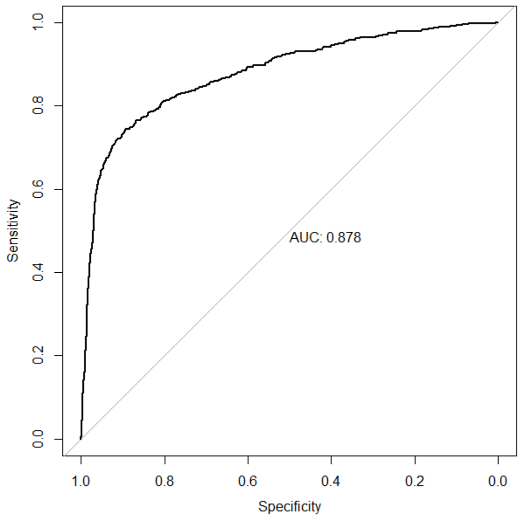
### 3. Penalized logistic regression with lasso penalty

We perform 10-fold cross validation to find out the best lambda. For this, we initialize a grid

```
grid <- 10^seq(10, -2, length = 100)
```

to perform cross validation for lambda values ranging from $10^{10}$ to $10^{-2}$.

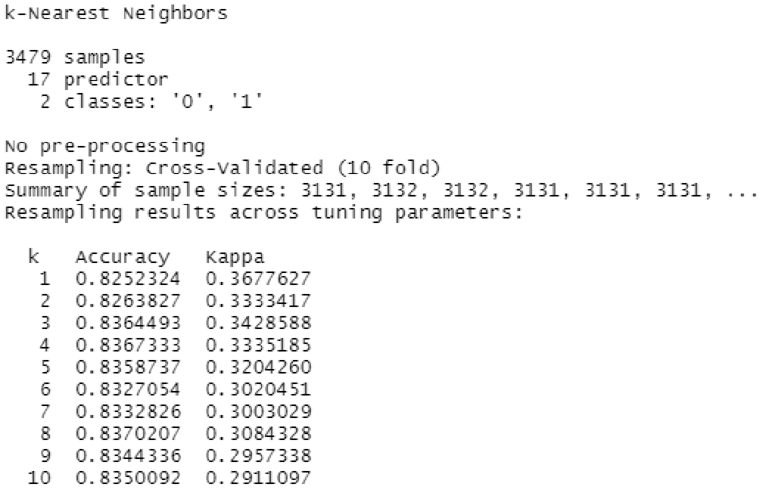
After the cross validation we see that cv.out$lambda.min = 0.003141718 which is almost zero. For that reason this models performs the same as logistic regression without regularization and we get the following ROC:



The AUC values is again **0.878**.

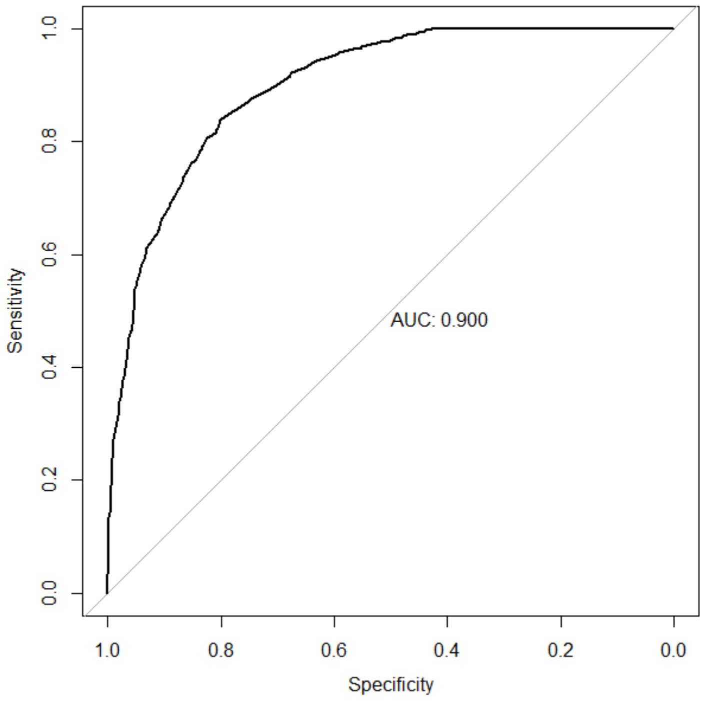
4. K-nearest neighbors

We again use the **caret** library for cross - validation where we input a grid with integer values for k ranging from 1 to 10.



```
k-Nearest Neighbors

3479 samples
  17 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3131, 3132, 3132, 3131, 3131, 3131, ...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   1  0.8252324  0.3677627
   2  0.8263827  0.3333417
   3  0.8364493  0.3428588
   4  0.8367333  0.3335185
   5  0.8358737  0.3204260
   6  0.8327054  0.3020451
   7  0.8332826  0.3003029
   8  0.8370207  0.3084328
   9  0.8344336  0.2957338
  10  0.8350092  0.2911097
```

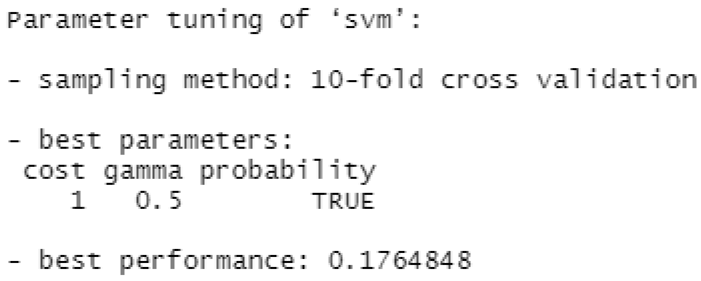
From the cross-validation results we see that the best values for k is 8. To see how this model performs for different thresholds for binary prediction we again plot the ROC curve:



The AUC value is **0.900**.

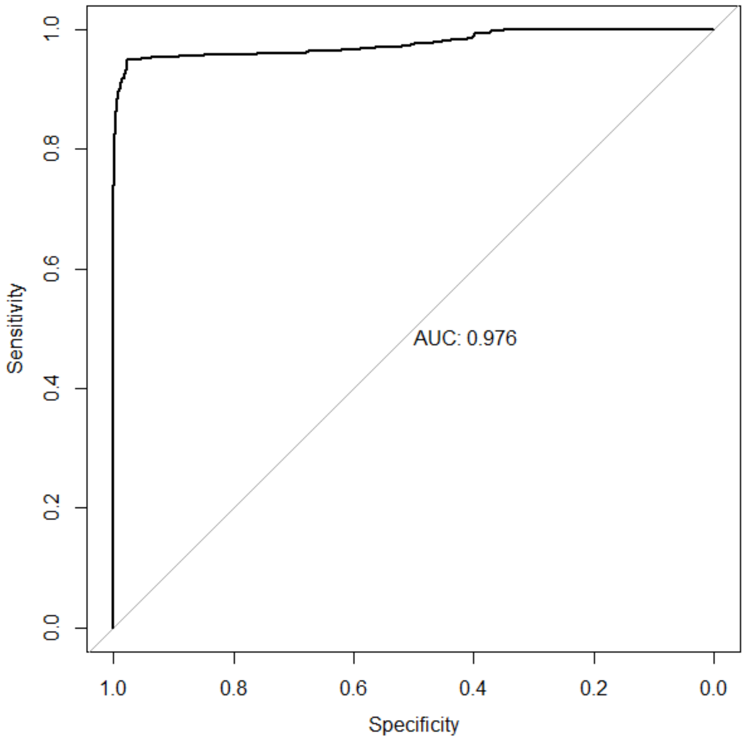
5. <u>Support vector machine with radial basis kernel functions</u>

To perform cross-validation we train the model with cost values = c(0.1, 1, 10, 100, 1000) and gamma values = c(0.5, 1, 2, 3, 4).



```
Parameter tuning of ‘svm’:

- sampling method: 10-fold cross validation

- best parameters:
 cost gamma probability
    1   0.5        TRUE

- best performance: 0.1764848
```

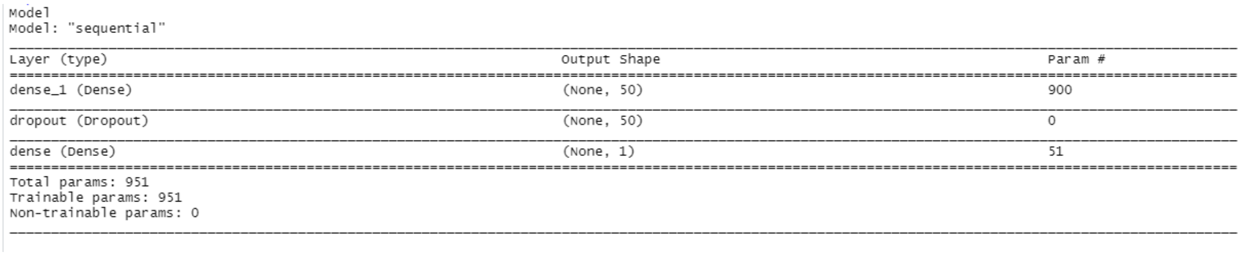
We see from the above plot that model with cost =1 and gamma = 0.5 gives the best performance. We plot the ROC curve with this model:



AUC values is equal to **0.976**.

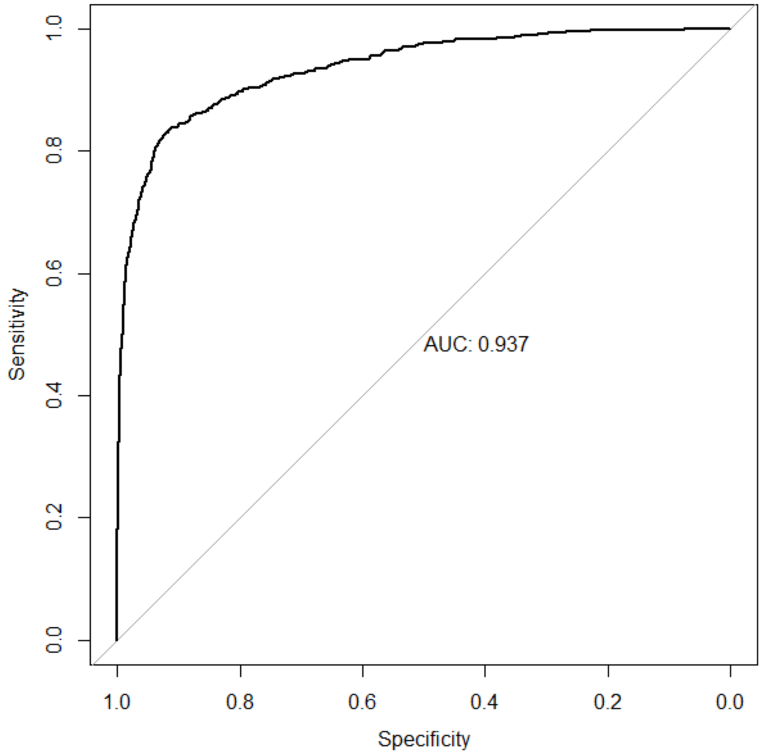
6. <u>Single layer neural network</u>

We train our single layer neural network with the following architecture:



```
Model
Model: "sequential"
________________________________________________________________________________________________________________________________
Layer (type)                                                    Output Shape                                          Param #
================================================================================================================================
dense_1 (Dense)                                                 (None, 50)                                            900
________________________________________________________________________________________________________________________________
dropout (Dropout)                                               (None, 50)                                            0
________________________________________________________________________________________________________________________________
dense (Dense)                                                   (None, 1)                                             51
================================================================================================================================
Total params: 951
Trainable params: 951
Non-trainable params: 0
________________________________________________________________________________________________________________________________
```

The ROC curve is given below:



The AUC value is **0.937**.

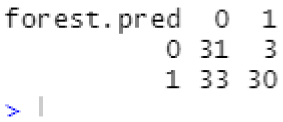
We see from the above results that, Random Forest model gives the highest AUC value, **0.983**. For this model we find the best prediction model with the following line of code

coords(test_roc, "best", ret = "threshold")

Which uses the “Youden index” which is simply the point with the cutoff with max(specificity + sensitivity). The above line outputs 0.119

## Results

Observing the results we got in the Method part, we see that Random Forest model with mtry = 4 performs the best, so we will use this model with the threshold 0.119 to make the predictions. And the following is the table for the prediction for the disc_hire and gender.

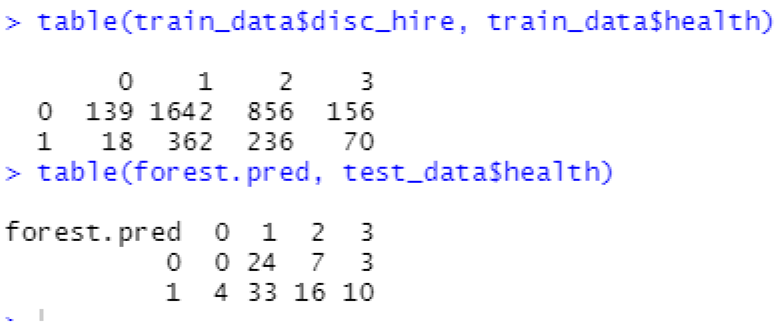


```
forest.pred  0  1
          0 31  3
          1 33 30
> 
```

We calculate Pr( pred = 1| gender = Female ) to be 30/33 = **0.91** whereas
Pr( pred = 1| gender = Male) = 33/64 = **0.52**.
This indicates that the random forest model that we have trained indicates that it very high likely for women to respond Yes to the hiring discrimination question.

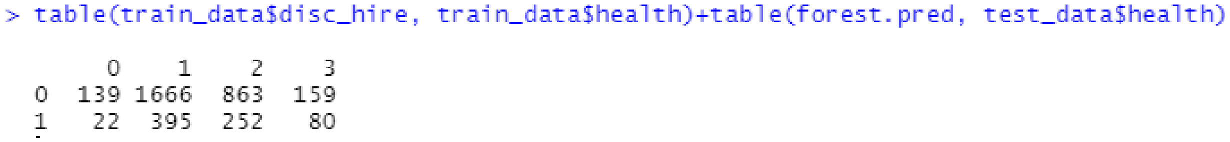
- **There is a difference in under-reporting of hiring discrimination between males and females**

To answer the second research question, we get the tables for training and test data:



```
> table(train_data$disc_hire, train_data$health)
   
       0    1    2    3
  0  139 1642  856  156
  1   18  362  236   70
> table(forest.pred, test_data$health)
           
forest.pred  0  1  2  3
          0  0 24  7  3
          1  4 33 16 10
```

And add them to get the following table:



```
> table(train_data$disc_hire, train_data$health)+table(forest.pred, test_data$health)

       0    1    2    3
  0  139 1666  863  159
  1   22  395  252   80
```

Again we calculate the following probabilities:

Pr(disc_hire = 1 | health = 0) = 0.129
Pr(disc_hire = 1 | health = 1) = 0.220
Pr(disc_hire = 1 | health = 2) = 0.276
Pr(disc_hire = 1 | health = 3) = 0.449

The above probabilities show that, people with worse self rated health are more likely to respond Yes to the discrimination question.

- **There is an association between the experience of hiring discrimination and health**

## Conclusion

In this study we trained 6 statistical models in order to find out answers to the 2 research questions we asked at the very beginning of this report. From the results we have obtained we conclude that, there is a difference in under-reporting of hiring discrimination between males and females since our statistical model has higher probability for females to respond Yes to the hiring discrimination question. Also, we conclude that there is an association between the experience of hiring discrimination and health, since the probability that  people to answer yes to the hiring discrimination increases as the self rated health increases ( 0 - very good, 3 - very poor ).

This research could be further improved if there were more data points since it is not very big and our test data was significantly lower than the train data (train data : 3479, test data: 97). Also to have faster predictions, especially for models like neural networks and support vector machines, we could try reducing the size of the feature space e.g. with PCA algorithm.