Group 6: Madhu Atyam, Nisha Pardesi, Saranya Murugan, Parul Mogra and Dheena Chinnasamy

Project Final Report

Predictive Analytics (ISQS 6349)

Problem 1

- **Problem Statement**
  - **a.** What is the impact on work interference due to the mental health condition of the employees working in the US technology companies?
  - **b.** Below are the few survey questions from the dataset to determine the problem statement
    - Family History (FH): Do you have a family history of mental illness?
    - Treatment(T): Have you sought treatment for a mental health condition?
    - Work Interference (WI): If you have a mental health condition, do you feel that it interferes with your work? This is dependent variable.
    - Age: - Respondent age.
    - Male/Female: Respondent gender

- **Background (with reference)**
  - **a.** Mental Health Illness of the employee's data is 2014 survey data conducted by Open Sourcing Mental Illness (https://osmihelp.org/research/) organization across the countries, targeting the Tech companies.

- **Empirical Results and Discussion**
  - **a.** Regression Models: Logit Regression Model is used to predict the probabilities.

  - **b.** The dataset linked to the variables in the problem statement attributes are Timestamp – 27Aug 2014 to 30 Nov 2015, Country- United States, States- All states in USA and Tech companies.

  - **c.** Results and Interpretation
    Using General Linear Model and specifying the link function logit in R program to perform regression analysis.

```
Call:
glm(formula = y ~ family_history + treatment, family = binomial(link = "logit"),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7128  -0.7112   0.2261   0.3685   1.7312

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.2456     0.1601  -7.782 7.14e-15 ***
family_history   0.9987     0.2663   3.751 0.000176 ***
treatment        3.9010     0.3079  12.672  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 790.12  on 610  degrees of freedom
Residual deviance: 424.11  on 608  degrees of freedom
AIC: 430.11

Number of Fisher Scoring iterations: 6
```

Apart from the above variables more independent variables are considered and below are the results:

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 0.009506   1.486632   0.006 0.994898
data$data.Age              -0.012905   0.016979  -0.760 0.447223
data$data.family_history    0.969987   0.267153   3.631 0.000283 ***
data$data.treatment         3.913784   0.313095  12.500  < 2e-16 ***
data$data.Male             -0.823120   1.390147  -0.592 0.553777
data$data.Female           -0.836495   1.416052  -0.591 0.554706
```

Based on the above regression analysis, Age and gender doesn't have any significant effect on Work Interference (dependent variable).

Probabilities: - Finding the impact probability of independent variables on dependent variable.

I. **Work_Interference = -1.2456+(0.9987*FH)+(3.9010*T)**

Considering Treatment constant and the interpretation of FH.
For FH =0 : ln(p/1-p) = -1.234
FH = 1: ln (p/1-p) = -1.234+0.9887 = -0.246
β1 is change in log odds comparing having no FH and FH.

Transform odds into probability -> $P = e^y/1+e^y$
For FH =0 :- Probability = $e^{-1.234}$ / 1+ $e^{-1.234}$ = 0.23
For FH =1 :- Probability = $e^{-0.246}$ / 1+ $e^{-0.246}$ = 0.44

Difference in probability = 0.44-0.23 = 0.21
Interpretation: - *The company will have 21 percent points more Work Interference when employee has Family History compare to employees with no Family history.*

II. **Work_Interference = -1.2456+(0.9987*FH)+(3.9010*T)**

Considering Family History constant and the interpretation of Treatment.
For T =0 : ln(p/1-p) = -1.234
T = 1: ln (p/1-p) = -1.234+3.89 = 2.656
β2 is change in log odds comparing having no T and T.

Transform odds into probability -> $P = e^y/1+e^y$
For T =0 :- Probability = $e^{-1.234}$ / 1+ $e^{-1.234}$ = 0.23
For T =1 :- Probability = $e^{2.656}$ / 1+ $e^{2.656}$ = 0.93

Difference in probability = 0.93-0.23 = 0.71
Interpretation: - *The company will have 71 percent points more Work Interference when employee has taken Treatment compare to employees with no Treatment.*

III. **Work_Interference = -1.2456+(0.9987*FH)+(3.9010*T)**

Considering Family History and the interpretation of Treatment.
For FH and T = 1: ln (p/1-p) = -1.234+0.9887+3.89 = 3.644

Transform odds into probability -> $P = e^y/1+e^y$
For FH and T =1 :- Probability = $e^{3.644}$ / 1+ $e^{3.644}$ = 0.975
Interpretation: - *The company will have 97.5 percent points more Work Interference when employee has Family History and undergone Treatment.*

Using R "predict" function, the prediction results are predicted for all the observations of the variables. Syntax: - predlogit <- predict(logit, data, type= 'response').

It is necessary to compute the error percent and measure of fitness to understand the improvement. Therefore, from the prediction results confusion matrix is constructed to measure the error percentage.

Confusion Matrix: It states the error percent and it is 15%.

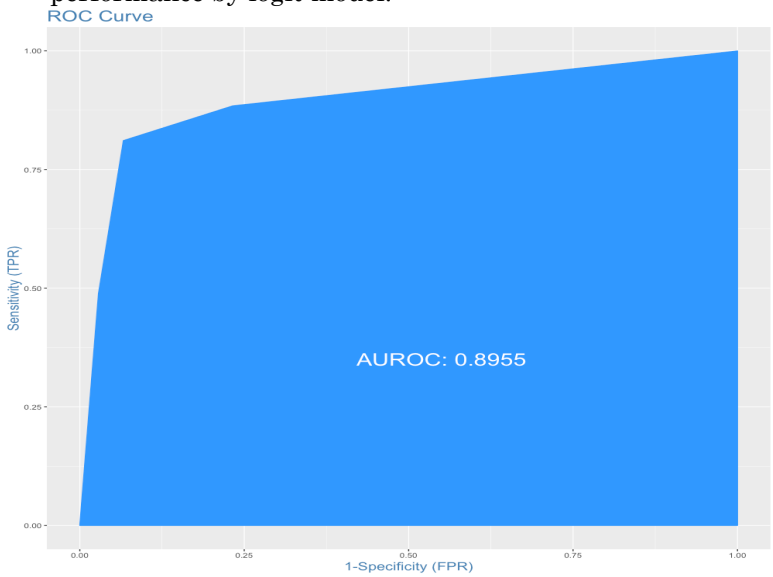
Predicted	Actual 0	1
0	197	75
1	14	322

Measure of Fitness:
Pseudo R2: 0.46

Chi Square- Pvalue logit 2.47e-79, since the P value is less than 5% CI, the fitness is considered to be good.

d. Strengths:

ROC(Receiver Operating Characteristics) curve: This curve determine the performance of the logit regression model. The area under the curve tending to 1 will determine the performance of the logit model. The more near to 1 is best considered as good performance and below 0.5 means we should reconsider our model. In this case, 0.895 area under curve value represents best performance by logit model.



X-axis of ROC:while, specificity is the percentage of 0's (actuals) correctly predicted. Specificity can also be calculated as 1 – False Positive Rate.

Yaxis of ROC- Sensitivity (or True Positive Rate) is the percentage of 1's (actuals) correctly predicted by the model,

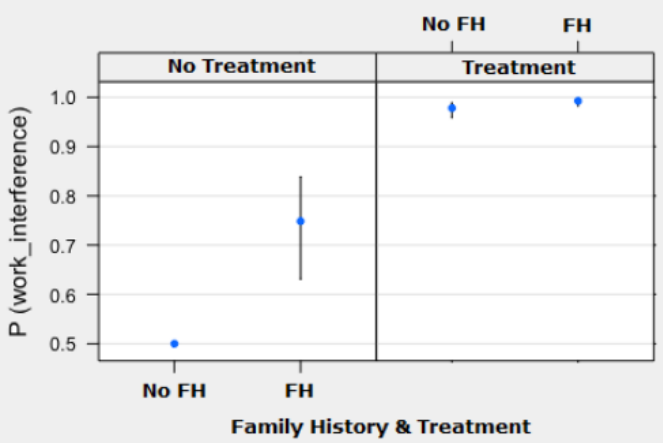
From the above curve it shows 89.55% that the performance is very good, and the analysis is true.

Limitations:

i. From the dataset, we might require more information about Treatment period, which is a limitation to analyze the effectiveness of the treatment taken by the employees.

- **Conclusion**
  - **a.** Managerial Implication / Recommendation

From the data it is evident that past treatment for mental illness has a pronounced effect on WI compared to the presence of the FH This is evident from the probability graph as shown, probabilities of WI given only FH ranges from 0.65 to 0.85 whereas the probability of WI given only past treatment ranges from 0.82 to 0.89. Irrespective of whether the employee has FH of mental illness, organization should encourage the employees to seek medical help if they suffer from uncoducive working environments.



  - **b.** What you learn from this project
    - i. Learnt to implement logit model for dataset where both dependent variable and independent variables are binary and Learned to transform the categorical data.
    - ii. Used new libraries in R:
      - pubh: inv_logit - to calculate inverse logit of regression model.
      - xyplot: to plot the boxplot of Independent variables probability against Dependent variables.
    - iii. Learnt to interpret the coefficients on independent variables in terms of point percentage of dependent variable.

==========

Problem 2

- **Problem Statement**
  - **a.** Whether the employers in USA are perceived to recognize the importance of mental health?
  - **b.** Below are the survey questions from the dataset to determine the behavior analysis of the company towards their employees
    - Leave(L): How easy is it for you to take medical leave for a mental health condition?
    - Mental vs physical(M&P): Do you feel that your employer takes mental health as seriously as physical health?
    - Supervisor(S): Would you be willing to discuss a mental health issue with your direct supervisor(s)?

- **Background (with reference)**
  - **a.** Mental Health Illness of the employee's data is 2014 survey data conducted by Open Sourcing Mental Illness (https://osmihelp.org/research/) organization across the countries, targeting the Tech companies.

- **Empirical Results and Discussion**
  - **a.** Regression Models: Logit Regression Model is used to predict the probabilities
  - **b.** The dataset linked to the variables in the problem statement are based on Timestamp –

27Aug 2014 to 30 Nov 2015, Country- United States, States- All states in USA and Tech companies.

c. Results and Interpretation:

– Using General Linear Model and specifying the link function logit in R program to perform regression analysis.

```
Call:
glm(formula = data$data.leave ~ data$data.mental_vs_physical +
    data$data.supervisor, family = binomial(link = "logit"),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4374  -0.8846  -0.6627   0.9380   1.8021

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -1.4042     0.1351 -10.391  < 2e-16 ***
data$data.mental_vs_physical   1.3295     0.2036   6.529 6.62e-11 ***
data$data.supervisor           0.6679     0.1933   3.454 0.000552 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 777.20  on 607  degrees of freedom
Residual deviance: 696.12  on 605  degrees of freedom
AIC: 702.12

Number of Fisher Scoring iterations: 4
```

– **Probabilities :-** Finding the impact probability of independent variables on dependent variable.

I. **Leave = -1.40 +(1.33*Mental vs physical)+(0.67*Supervisor)**

Considering Supervisor constant and the interpretation of Leave.
For Mental vs Physical =0 : ln(p/1-p) = -1.40
Mental vs Physical = 1: ln (p/1-p) = -1.40+1.33 = -0.07
β1 is change in log odds comparing having No Mental vs Physical and Mental vs Physical

For Mental vs Physical =0 :- Probability = $e^{-1.40}$ / 1+ $e^{-1.40}$ = 0.20
For Mental vs Physical =1 :- Probability = $e^{-0.07}$ / 1+ $e^{-0.07}$ = 0.48

Difference in probability = 0.48-0.20 = 0.28
Interpretation: - *The company will have 28 percent points more Leave when employee has* Mental vs Physical *compare to employees with no* Mental vs Physical.

II. **Leave = -1.40 +(1.33*Mental vs physical)+(0.67*Supervisor)**

Considering Mental vs Physical constant and the interpretation of Leave.
For Supervisor =0 : ln(p/1-p) = -1.40
Supervisor = 1: ln (p/1-p) = -1.40+0.67 = -0.74
β1 is change in log odds comparing having No Supervisor and Supervisor

For Supervisor =0 :- Probability = $e^{-1.40}$ / 1+ $e^{-1.40}$ = 0.20
For Supervisor =1 :- Probability = $e^{-0.74}$ / 1+ $e^{-0.74}$ = 0.32

Difference in probability = 0.32 - 0.2 = 0.12
Interpretation: - *The company will have 12 percent points more Leave when employee has* Supervisor cooperation *compare to employees with no* Supervisor cooperation

III. **Leave = -1.40 +(1.33*Mental vs physical)+(0.67*Supervisor)**

Considering Mental vs Physical and supervisor, the interpretation of Leave.

For Supervisor =1, Mental vs Physical =1 : ln(p/1-p) = 0.59

For Supervisor =1 and Mental vs Physical =1 :- Probability = $e^{-0.59}$ / 1+ $e^{-0.59}$ = 0.64

Interpretation: *- The company will have 64 percent points more Leave when employer consider mental vs physical and supervisor of employee is approachable.*

Using R "predict" function, the prediction results are predicted for all the observations of the variables. Syntax: - predlogit <- predict(logit, data, type= 'response').

It is necessary to compute the error percent and measure of fitness to understand the improvement. Therefore, from the prediction results confusion matrix is constructed to measure the error.

Confusion Matrix: It states the error percent and it is 26.97%.

Confusion Matrix:

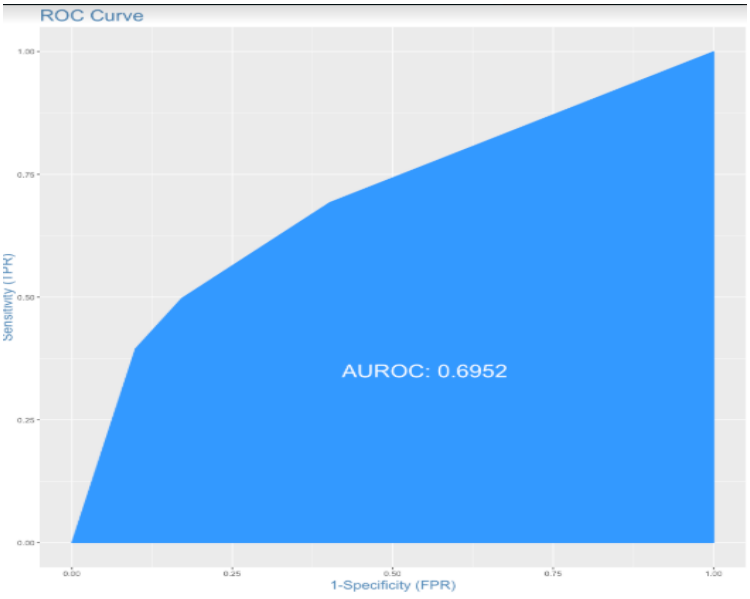
Predicted \ Actual	0	1
0	363	124
1	40	81

Measure of Fitness:
Pseudo R2: 0.11

Chi Square- pvalue logit 2.49e-18 since the P value is less than 5% CI, the fitness is considered to be good.

a. Strength :

ROC(Receiver Operating Characteristics) curve: This curve determine the performance of the logit regression model. The area under the curve tending to 1 will determine the performance of the logit model. The more near to 1 is best considered as good performance and below 0.5 means we should reconsider our model. In this case, 0.6952 area under curve value represents good performance by logit model.



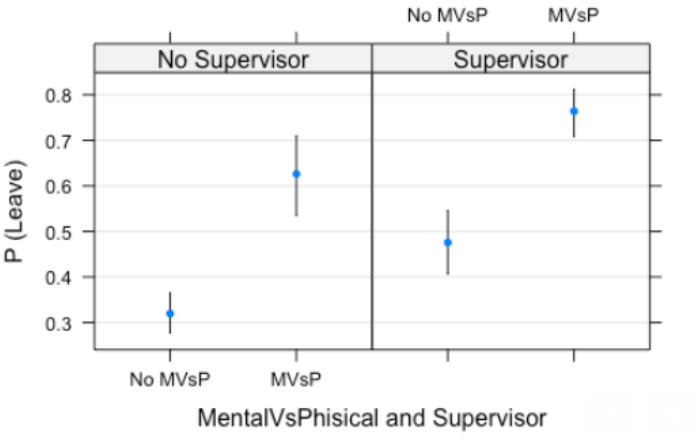
X-axis of ROC:while, specificity is the percentage of 0's (actuals) correctly predicted. Specificity can also be calculated as 1 – False Positive Rate.

Yaxis of ROC- Sensitivity (or True Positive Rate) is the percentage of 1's (actuals) correctly predicted by the model,

The above model has area under ROC curve 69.52%, which is good.

- **Conclusion**
  - **a.** Managerial Implication / Recommendation

From the population data it is evident that If the employer doesn't give equal cognizance to mental and physical health as well as employee finds it difficult to approach supervisor to discuss his mental illness, probability of getting leave is bleak (0.2 to 0.35) therefore the employer should be more sensitive towards the health of its employees be it mental or physical moreover the mental condition of the employee should be kept confidential. These two measures would motivate the employee to approach the management to seek help



  - **b.** What you learn from this project
    - i. Learnt to implement logit model for dataset where both dependent variable and independent variables are binary and Learned to transform the categorical data.
    - ii. Used new libraries in R:
      - pubh: inv_logit - to calculate inverse logit of regression model.
      - xyplot: to plot the boxplot of Independent variables probability against Dependent variables.
    - iii. Learnt to interpret the coefficients on independent variables in terms of point percentage of dependent variable.

==========

Problem 3

- **Problem Statement:**
  - **a.** To analyze the effectiveness of the health program for the employees working in USA and determine the employer offers privileges to their employees to tackle mental health issues?
  - **b.** Below are the few survey questions from the dataset to determine the problem statement.
    - Benefits(B): Does your employer provide mental health benefits?
    - Care options(CO): Do you know the options for mental health care your employer provides?
    - Wellness Program (WP): Has your employer ever discussed mental health as part of an employee program ?
    - Seek help(SH): Does your employer provide resources to learn more about mental health issues and how to seek help?
- **Background (with reference):**
  - **a.** Mental Health Illness of the employee's data is 2014 survey data conducted by Open Sourcing Mental Illness (https://osmihelp.org/research/) organization across the countries, targeting the Tech companies.

- **Empirical Results and Discussion:**
  - **a.** Regression Models: Logit Regression Model is used to predict the probabilities
  - **b.** The dataset linked to the variables in the problem statement are Timestamp – 27Aug 2014 to 30 Nov 2015, Country- United States, States- All states in USA and Tech companies.
  - **c.** Results and Interpretation
  - Using General Linear Model and specifying the link function logit in R program to perform regression analysis.

```
Call:
glm(formula = data$data.wellness_program ~ data$data.seek_help +
    data$data.benefits, family = binomial(link = "logit"), data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4663  -0.5528  -0.2410  -0.2410   2.6660

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -3.5248     0.3221 -10.945  < 2e-16 ***
data$data.seek_help  2.4588     0.2613   9.409  < 2e-16 ***
data$data.benefits   1.7236     0.3568   4.831 1.36e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 620.49  on 607  degrees of freedom
Residual deviance: 409.09  on 605  degrees of freedom
AIC: 415.09

Number of Fisher Scoring iterations: 6
```

```
Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)              -3.5800     0.3292 -10.874  < 2e-16 ***
data$data.seek_help       2.4282     0.2630   9.232  < 2e-16 ***
data$data.care_options    0.2402     0.2673   0.899    0.369
data$data.benefits        1.6441     0.3675   4.474 7.68e-06 ***
```

Based on the above regression analysis, care_options doesn't have any significant effect on Wellness Program (dependent variable).

- **Probabilities:-** Finding the impact probability of independent variables on dependent variable.

I. **Wellness program = -3.5248+(2.4588*SH)+(1.7236*B)**

Considering Benefits constant and the interpretation of SH.
For SH = 0 : ln(p/1-p) = -3.5248
SH = 1 : ln(p/1-p) = -3.5248+2.4588 = -1.066
β1 is change in log odds comparing having no SH

Probability Calculations:
For SH =0 :- Probability = $e^{-3.5248}$ / 1+ $e^{-3.5248}$ = 0.03
For SH =1 :- Probability = $e^{-1.066}$ / 1+ $e^{-1.066}$ = 0.26

Difference in probability = 0.26-0.03 = 0.23
Interpretation: - *The company will have 23 percent points more Wellness Program when employee has Seek Help compare to employees with no Seek Help.*

II. **Wellness program = -3.5248+(2.4588*SH)+(1.7236*B)**

Considering Benefits interpretation and constant for SH.

For B = 0 : ln(p/1-p) = -3.5248
B = 1 : ln(p/1-p) = -3.5248+1.7236 = -1.8012
β1 is change in log odds comparing having no Benefits.

Probability Calculations:
For B =0 :- Probability = $e^{-3.5248}$ / 1+ $e^{-3.5248}$ = 0.03
For B =1 :- Probability = $e^{-1.8012}$ / 1+ $e^{-1.8012}$ = 0.14

Difference in probability = 0.14-0.03 = 0.11
Interpretation: - *The company will have 0.11 percent points more Wellness Program when employee has Benefits compare to employees with no Benefits.*

III. **Wellness program = -3.5248+(2.4588*SH)+(1.7236*B)**
Considering Benefits and Seek Help for the interpretation.
For SH =1 and B =1
ln(p/1-p) = -3.5248+2.4588+1.7236 = 0.6576
Probability Calculations: Probability = $e^{-0.6576}$ / 1+ $e^{-0.6576}$ = 0.66
Interpretation: *- The company will have 66 percent points more Wellness Program when employee is aware of Benefits and seeks help.*

Using R "predict" function, the prediction results are predicted for all the observations of the variables. Syntax: - predlogit <- predict(logit, data, type= 'response').

It is necessary to compute the error percent and measure of fitness to understand the improvement. Therefore, from the prediction results confusion matrix is constructed to measure the error.

Confusion Matrix: It states the error percent and it is 14.69%.

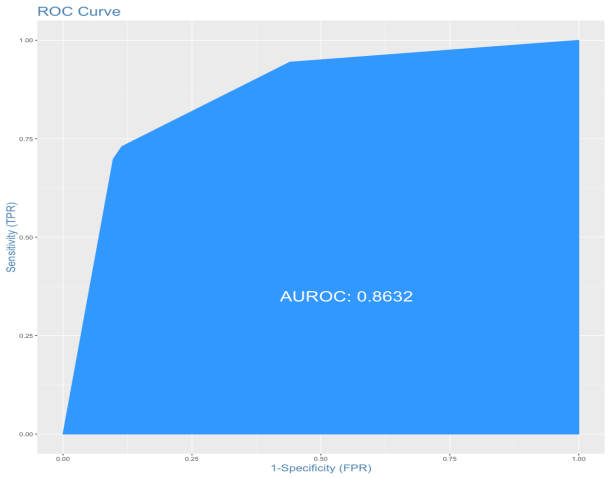
	Actual	
Predicted	0	1
0	435	38
1	47	88

Measure of Fitness:
Pseudo R2: 0.34

Chi Square- pvalue logit 1.24xe-46 since the P value is less than 5% CI, the fitness is considered to be good.

a. Strengths

– **ROC(Receiver Operating Characteristics) curve:** This curve determine the performance of the logit regression model. The area under the curve tending to 1 will determine the performance of the logit model. The more near to 1 is best considered as good performance and below 0.5 means we should reconsider our model. In this case, 0.8632 area under curve value represents good performance by logit model.



X-axis of ROC: while, specificity is the percentage of 0's (actuals) correctly predicted. Specificity can also be calculated as 1 – False Positive Rate.

Yaxis of ROC- Sensitivity (or True Positive Rate) is the percentage of 1's (actuals) correctly predicted by the model,

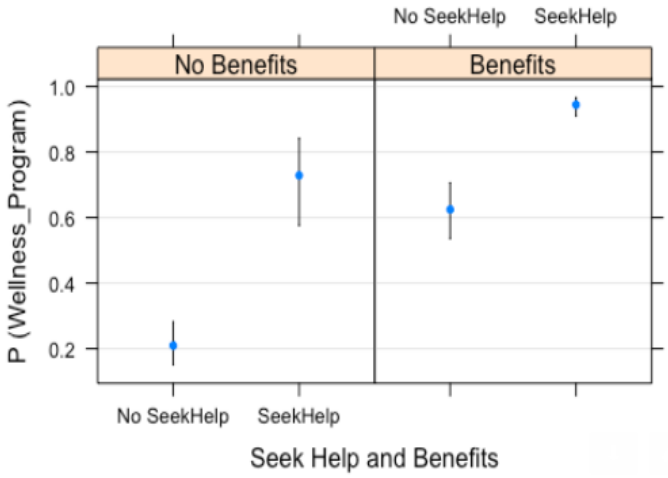
– Conclusion

a. Managerial Implication / Recommendation

When the organization doesn't have any predefined benefits for employees suffering from mental health condition and/or the organizations doesn't take initiates to make employees aware about the existence of wellness program, Overall effectiveness of the organization's wellness program is mitigated (probability = 0.15 to 0.23). In the benefit of the employees as well as the organization, wellness program should be well defined, and the employees should be made aware of the same.

This can be ensured by

1. Effective orientation programs.
2. Periodic quizes/quartely surveys to check employees awareness about the awareness among the employees



b. What you learn from this project

i. Learnt to implement logit model for dataset where both dependent variable and independent variables are binary and Learned to transform the categorical data.

ii. Used new libraries in R:
- pubh: inv_logit - to calculate inverse logit of regression model.
- xyplot: to plot the boxplot of Independent variables probability against Dependent variables.

iii. Learnt to interpret the coefficients on independent variables in terms of point percentage of dependent variable.